Assessing a feature's trustworthiness and two approaches to feature selection

Franco Alessandro Ubaudi

THESIS

Submitted to the Faculty of Engineering and IT
University of Technology Sydney (UTS)
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Computing Sciences

2011

CERTIFICATE OF AUTHORSHIP/ORIGINALITY

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that this thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature sources used are indicated in the thesis.

Signature of Student

_____

Acknowledgements

I am in debt to many people—for without their help I would never have made it. First of all I wish to thank Dr. Paul J. Kennedy from the University of Technology Sydney (UTS), for you were my greatest source of encouragement and direction. Among other things Paul, you provided me with a high standard to shoot-for and although it was tuff going at times, your standards enabled me to grow even further in my abilities and I am truly grateful for that. I am also grateful for your *immense* patience and *continuous* words of encouragement. You are a great supervisor!

I owe a debt of gratitude to Adjunct Professor Daniel Catchpoole from The Children's Hospital at Westmead (CHW) for his encouragement, support and direction. In particular I am very grateful for the times you called me and provided timely words of encouragement. How many supervisors are there who chase up their students; but fortunately I had such a supervisor. There were times you even drove me home after a meeting, just so we could chat. Thanks also to Nicholas Ho at CHW, for helping me get my head-around *R*, for your practical support in building heat maps and using GSEA.

Thanks to Professor Simeon J. Simoff, head of the school of computing at the University of Western Sydney (UWS), for your support and encouragement. In particular you have a great ability to see the big picture and that often provided the needed focus on where to head, which was always valuable since it is so easy to be short-sighted when there is plenty of detailed work to do.

Thanks also to Jenny and Lyn for your practical support over the years. In particular Jenny, I wish to thank you for making a big deal when I completed my Masters degree and the surprise party, it was amazing. Thank you Lyn, for whenever I was down and needed someone to believe in me, you were always what I needed. I also owe a debt of thanks to Uncle Dick from the "Land of the 10 10". I have known two incredible men in my life and you are one of them. In addition to being a true man of God, you always chase me up and show interest in where I am at and you always provide words of encouragement. Your are an amazing man and I hope that some of you will rub off onto me.

I am also in debt to Blair, Kathryn and Nicola, for each of you were extremely understanding, patient and supportive to me during this long haul. I am so privileged in being your dad. I am proud of you guys and grateful for our friendship and the fun we have together. You are the best kids a dad

could ever hope for!

I wish to pay tribute to my parents Giovanni and Enza, for they sacrificed a great deal so I could have opportunities they never had. The memory of you two will always be with me, for you were great parents and I lacked *nothing* important. I will never forget you mum and dad, or what you taught me. May I always make both of you proud of me.

Thanks also to my work colleagues: Greg Edwards, Peter Stephanou, Melanie Wilmot, Aaron Ashe, Karen Parker, Peter Gaudron and the rest of you in Distribution Planning. For you guys were *always* flexible, supportive and a pleasure to work with.

But most of all I am entirely in debt to my God and savior, for he provided me with my abilities and the opportunities to use them. He opened and closed doors at the right time in my life and although I often failed to understand the timeliness of it all—looking back now, I realize more than ever that your timing and provision is always perfect. Among many other things, he provided me the opportunity to undertake this research and the perfect job at Ausgrid (formally Energy Australia), which in particular provided the needed flexibility and money to pay the bills. Thanks also for your word: for it provides me with a light showing the way and a secure foundation that will *never* fail me.

> Your word is a lamp to my feet and a light for my path.
> (Psalms 119:105 NIV)
>
> Therefore everyone who hears these words of mine and puts them into practice is like a wise man who built his house on the rock. The rain came down, the streams rose, and the winds blew and beat against that house; yet it did not fall, because it had its foundation on the rock. (Matthew 7:24-25 NIV)

Contents

List of Figures

List of Tables

Abstract

Improvements in technology have led to a relentless deluge of information that current data mining approaches have trouble dealing with. An extreme example of this is a problem domain that is referred to as"non-classical". Non-classical problems fail to fulfill the requirements of statistical theory: that the number of instances in the sample set be much greater than the number of dimensions. Non-classical problems are mainly characterized by many dimensions (or features) and few noise-affected samples.

Microarray technology provides one source of non-classical problems, which typically produces data sets with a dimensionality exceeding ten thousand and containing just a few hundred instances. A risk with such a data set is building a model that is significantly influenced by coincidental correlations between the inputs (or the model's features) and the output. A classical strategy for managing this risk is reducing the dimensionality without significantly affecting the correlation between the remaining features and the model's output. However this strategy does not explicitly consider the impact

of poor data quality (or noise) and having few data samples.

In order to actively manage noise—a feature selection strategy is needed that not only considers the correlation between the features and the output, but also the quality of the features. It is proposed that feature quality, or simply the feature's "trustworthiness", should be incorporated within feature selection. As the trustworthiness of a feature increases, it is expected that the ability to accurately extract the underlying structure of the data will also increase. Another characteristic of non-classical problems is significant feature redundancy (where information provided within one dimension is also present in one or other dimensions). This research postulates that the use of feature trustworthiness and redundancy provides an opportunity to actively reduce the noise associated with the selected feature set, while still finding features that are well correlated with the model's output.

Two fundamental contributions are provided by this thesis: the notion of feature "trustworthiness" and how trustworthiness can be integrated within feature selection. Trustworthiness provides a flexible approach for evaluating the quality of a feature's sample data and in certain cases, the quality of the test data. This flexibility encourages the use of prior knowledge about the specific problem and in particular, how the quality of the data is best estimated. Traditionally feature selection implicitly assumes that every instance of data, supplied by preprocessing, has the same quality. Trustworthiness

also provides an opportunity for incorporating a measure of the changes applied to the data set as a result of data cleaning.

Using an area of computational learning, a theoretical justification was constructed that showed the difficulty of building an accurate model for a non-classical problem. The justification showed how a modest data quality problem can result in insufficient sample data to permit successful learning. It also showed how selecting less noisy data, or sufficiently trustworthy features, can enable successful learning using the available data points.

This thesis presents two methodologies that incorporate a measure of data quality within feature selection: one methodology only uses training data, while the other also incorporates test data while evaluating feature trustworthiness. The two methodologies are contrasted with each other and with a traditional feature selection methodology, which does not consider data quality.

A number of data sets were used to test these methodologies, with the main data sets being: synthetic data, childhood leukaemia and chronic fatigue syndrome. In most cases the three feature selection methodologies achieved similar accuracy however there were clear differences in the features selected by each. Using heat maps to visualize the clarity of the separation of the class labels by the selected features—showed dramatic differences. The two methodologies that incorporate trustworthiness provided a clearer

separation, while the traditional methodology was substantially inferior and appeared to be heavily influenced by artifacts. Using Gene Set Enrichment Analysis (GSEA), a widely used resource for evaluating the biological meaningfulness of gene sets (Subramanian, Tamayo, Mootha, Mukherjee, Ebert, Gillette, Paulovich, Pomeroy, Golub, Lander, and Mesirov, 2005), showed that the two proposed methodologies selected genes that were more biologically meaningful than those selected by a traditional feature selection methodology.

The experiments also evaluated the sensitivity of trustworthiness to differences in the data set. By evaluating the trustworthiness of every feature, it was shown that considerable changes occurred across data folds. This result agrees with findings in the literature, such as (Ein-Dor, Kela, Getz, Givol, and Domany, 2005) and provides one explanation for the difficulty of modeling non-classical problems.

Contributions to knowledge

1. **Two approaches for estimating data quality**

   This thesis describes two approaches for estimating the quality of a feature's data set, one using a measure of signal-to-noise and the other, only a measure of noise. The latter approach permits the use of test data within the estimate of a feature's utility. In calculating the quality of a feature's data set, both approaches use prior knowledge about the data set.

2. **Definition and measure of feature trustworthiness**

   This thesis defines "trustworthiness" as an overall measure of the quality of a feature. Trustworthiness is required since each item of data associated with a feature has a unique level of quality. As a feature's trustworthiness increases, its utility approaches the estimated information it provides.

3. **Definition and measure of feature utility**

   Feature selection methodologies typically evaluate a feature's usefulness according to the feature's inherent information. However, noise may falsely increase its information; hence reduce the feature's trustworthiness. This thesis describes a measure of a feature's usefulness called "feature utility", which is a function of information provided by the feature and the noise present within its data set.

4. **Two feature selection methodologies based on feature utility**

   This thesis develops and evaluates two feature selection methodologies based on feature utility. The methodologies, "Feature Utility Ranking 1" and "Feature Utility Ranking 2", use respectively, a signal-to-noise ratio and a noise-only measure of quality.

   The methodologies were evaluated using computational learning theory, a series of synthetic data experiments and three biomedical data sets, which are characterized by high dimensionality, few samples and significant noise. The biomedical data experiments consist of leukaemia cell type classification, leukaemia treatment outcome prediction and chronic fatigue syndrome prediction.

5. **Mechanism for using data quality information collected during data cleaning**

   A mechanism for merging data cleaning outcomes within feature selection is explored. This approach allows the degree of change imposed by data cleaning to be incorporated in a feature's utility.

Publications

The work in this thesis has been published in:

Paul J. Kennedy, Simeon J. Simoff, Daniel R. Catchpoole, David B. Skillicorn, **Franco Ubaudi**, and Ahmad Al-Oqaily. Integrative Visual Data Mining of Biomedical Data: Investigating Cases in Chronic Fatigue Syndrome and Acute Lymphoblastic Leukaemia. In Simoff Simeon J., Michael H. Bohlen, and Mazeika Arturas, editors, *Visual Data Mining: Theory, Techniques and Tools for Visual Analytics*, volume 4404 of Lecture Notes in Computer Science, pages 367-388. LNCS, Springer, Berlin Heidelberg, 2008.

Franco A. Ubaudi, Paul J. Kennedy, Daniel R. Catchpoole, Dachuan Guo, and Simoff Simeon J. Microarray Data Mining: Selecting Trustworthy Genes with Gene Feature Ranking. In L. Cao, P. S. Yu, C. Zhang, and H. Zhang, editors, *Data Mining for Business Applications*, pages 159-168. Springer, 2009.

Chapter 1

Introduction

We are drowning in information and starving for knowledge.

Rutherford D. Roger

The topic of this thesis is feature selection (Guyon and Elisseeff, 2003) in multidimensional data, which involves ranking features according to the information they provide. For convenience, this type of feature selection will be referred to as "traditional feature selection". In the context of a "classical" problem domain, where a large number of samples are available and the number of features is small, traditional feature selection is straightforward, reliable and is generally manageable in terms of computational effort. Even if noise is present in the data set, the selection process can still succeed provided sufficient samples are available for discovering the underlying structure hidden within the data set.

1.1 Research problem

The advancement of technology has produced a "non-classical" problem domain which consists of few samples, a large number of features, the presence of significant noise and a large number of redundant features (Ding and Peng, 2003). In the case of a classification problem, non-classical problem domains may also consist of a class imbalance, which further increases the difficulty of selecting a subset of features that provide a reliable generalization of the problem being studied. An example of reliable generalization is a feature set that performs well against the test set and unlabeled data. Data with no class label or output value will be referred to as "unlabeled" data in this thesis. Some examples of non-classical problem domains are microarray data analysis (Hegde, Qi, Abernathy, Gay, Dharap, Gaspard, Earle-Hughes, Snesrud, Lee, and Quackenbush, 2000; Allison, Cui, Page, and Sabripour, 2006), the analysis of chemical composition by interpreting infra-red reflections (Borzenko, 2011), sound and speech analysis (Saraswathi and Geetha, 2006), face detection and recognition (Doukas and Maglogiannis, 2010), MRI brain imaging (Angelini, Jin, and Laine, 2005), automatic satellite image navigation (Emery, Baldwin, and Matthews, 2003) and text mining (Yang and Pedersen, 1997).

A data set consisting of one hundred samples and ten thousand features is

an example of a non-classical problem. The sparseness of such data impacts the reliability of the generalization achieved, even without the presence of noise. In the context of no noise, the sparse data set poses the same challenges to feature selection as for the classical domain, except for the task being more computationally intensive. However the addition of noise can introduce a fundamentally different challenge to the feature selection process. A threshold exists where the sparseness of the data set, the large number of features and the presence of significant noise, could result in the emergence of significant apparent information. The presence of apparent information, which differs from the information present in the underlying structure, results in the selection of features that perform well against the training set, but provide a poor generalization for data that was unseen by feature selection. As a result, the selected features perform poorly against the validation and test sets, since no mechanism exists within traditional feature selection to manage the impact of noise.

Consequently a feature selection methodology is needed that also considers the quality of the data being evaluated in the feature selection process. An option should also exist where the quality of test and unlabeled data are also considered during feature selection, thereby facilitating the selection of features that provide a better generalization than achievable by traditional feature selection.

1.1.1 Cost of prediction

Classical and non-classical problems have two associated costs: the cost of producing the data and the cost of prediction errors. The production of sample data may be expensive with respect to resources, particularly if there are many features. As a result, fewer samples may be available than desired or considered necessary. Methods or technology that are designed to produce data sets consisting of large numbers of features may be associated with significant noise. For example, the problem of selecting the most appropriate treatment for a seriously ill patient can involve ten thousand or twenty thousand gene expression measurements. Other domains have similarly high dimensional data.

There are four possible prediction outcomes for a binary classifier: true positive (TP), true negative (TN), false positive (FP) and false negative (FN) and each could have a different cost. In the case of selecting between two treatments that differ by their aggressiveness, predicting which treatment is most appropriate results in two favorable (TP and TN) and two undesired outcomes (FP and FN). Using patient well being as a measure of cost, the TP and TN outcomes have no cost since the most appropriate treatment was selected, while significant costs are associated with the FP and FN outcomes. A FP and a FN outcome have different costs since one is associated with

unnecessarily aggressive treatment, which although it results in survival, can be associated with long term health consequences due to the treatment's toxicity. While the highest cost outcome occurs when a patient dies because an insufficiently aggressive treatment was selected.

The presence of differing prediction costs demands feature selection methods that are better suited to non-classical problems and the above scenario highlights the importance of minimizing prediction errors. Even the additional effort of re-selecting features and rebuilding models, in response to the quality of unlabeled data may be justified, when significant costs are associated with prediction errors.

1.2 Aim of the research

The aim of this research is to investigate how a measure of data quality can be incorporated into the feature selection process and its possible benefits. The motivation for this aim comes from the need to better utilize sample data in a non-classical problem domain.

Current feature selection approaches manage redundancy in one of two ways. They either assume that redundancy is no different to irrelevance, or that redundancy provides an opportunity to bolster the information provided by one feature by incorporating other redundant features. However this approach risks incorporating excessive noise within the learning process, which

can be a significant problem if few samples are available. Since different levels of quality and therefore accuracy impact each feature, the developed methods in this thesis eliminate features that are of insufficient quality, so that remaining data is effective in the learning process. However each modeling problem has different quality issues. This thesis aims to develop a feature selection methodology that allows the user to incorporate prior knowledge into the measure of quality used. The use of quality in this way also provides a vehicle for incorporating the changes made to the data set during cleaning. *This is valuable since data preparation* (Zhang, Zhang, and Yang, 2003) *is traditionally completely decoupled from feature selection and model learning* (Mitchell, 1997). Consequently feature selection implicitly assumes that every item of data is of equal quality. However the degree of change applied to each item of data during cleaning can vary substantially and this change provides a measure of quality.

1.2.1 Objectives

The aims of this research have been decomposed into five objectives:

1. Define the impact of noise on the required samples for learning,

2. Develop a framework for estimating the quality of each dimension of a sample point,

3. Develop a framework for ranking features according to their utility,

4. Empirically evaluate two feature selection methodologies using synthetic and real-world non-classical data and

5. Develop a mechanism that allows changes in the data set caused by data cleaning to provide a measure of data quality and then an input to feature utility.

Objective 1 determines the impact of noise on the information contained within the data set, by developing an equation that describes the relationship between noise and the minimum number of samples required to achieve a model of arbitrary accuracy. This model is used to understand whether a problem from the non-classical domain can either fail to be learnable or provide an unacceptable level of accuracy, given the sample sizes that are typical of the domain.

Objective 2 is the development of a framework for estimating the quality of each dimension for each data point and to then combine these estimates of quality into an overall measure called a feature's "trustworthiness". The benefit of feature trustworthiness depends on the presence of a large number of features, many of which are expected to be redundant. This dependency on redundancy is a key prerequisite for the approach in this thesis and is a key differentiator when compared to traditional feature selection methodologies.

Given that sample sizes are typically fixed and limited in size, a rational strategy is to prefer features that are described by better quality data, rather than just the degree of information provided.

This thesis considers a feature's utility to be a function of the feature's trustworthiness and the information it provides. Objective 3 is the development of a framework for ranking features according to their utility and then producing two feature selection methodologies called Feature Utility Ranking 1 (FUR1) and Feature Utility Ranking 2 (FUR2). Feature Utility Ranking 1 determines data quality using a measure of signal-to-noise. It is limited to evaluating only the quality of the training set. In contrast, FUR2 determines data quality by only using a measure of the noise within the data set. This assessment of data quality enables FUR2 to evaluate the quality of the test set and unlabeled data. Feature Utility Ranking 2 is best suited to data sets consisting of a significant variation in data quality.

Objective 4 involves the empirical evaluation of FUR1 and FUR2 using a mixture of synthetic and real-world nonclassical data sets. The synthetic data set provides an environment where the amount of noise is controlled. The real-world data sets involve classification of childhood leukaemia samples, the prediction of leukaemia treatment outcome and classifying samples from a chronic fatigue syndrome research project. All the experiments performed for FUR1 and FUR2 are repeated using a traditional feature selection

methodology, in order to assess the potential benefits of using data quality.

Feature Utility Ranking 2 also fulfils objective 5—that changes in the data set caused by data cleaning can also provide a measure of quality. This allows the trustworthiness of each feature to be directly influenced by data cleaning, rather than feature selection assuming that every instance of data has the same quality.

One of the purposes of the objectives is to evaluate the merits of a signal-to-noise measure and a noise-only measure of quality. A signal-to-noise measure of quality recognizes the relationship between signal strength and noise. Hence the higher the ratio, the better the quality. However this measure limits the evaluation of quality to using the training data. Although a noise-only measure of quality fails to consider signal strength, it permits the use of test and unlabeled data, thereby enabling FUR2 to consider the quality of *all* data. It also allows the incorporation of quality obtained during data cleaning.

The noise component of a measure of quality involves the use of data that is not correlated with the prediction being made by the model. This use of noise provides considerable flexibility and enables the user to tailor the feature selection process to the specific problem.

1.2.2 Definitions

"Data quality" is defined in this thesis as the estimated accuracy of an *item of data*. There may be many different methods for estimating data quality within a sample set[1], since a different measuring instrument may be involved for each feature. However in this thesis, a single method was used for each of the experiments. Although a framework for measuring data quality is described in the proposed feature selection methodologies in sections 3.7 and 3.8, the user is required to populate the missing details through the application of domain knowledge.

Since it may not be possible to consider every factor that impacts data quality, the user is responsible for defining a measure that largely reflects the data's accuracy. Consider a simple example where the known accuracy of a measuring instrument is proportional to humidity. Given a humidity reading for every item of data, an estimate of its quality can be devised.

As each feature is associated with n estimates of data quality, since it has n data items, a scalar measure of accuracy for a feature is required in order to rank features according to their accuracy."Trustworthiness" is the proposed measure of a feature's accuracy. The proposed feature selection methodolo-

[1] "Data quality" is defined in this thesis as the estimated accuracy of an *item of data* and since a sample set is composed of n samples and p features, there are $n \times p$ items. In general there are p different methods for estimating data quality within a sample set, since a different measuring instrument may be involved for each feature.

gies describe a framework for calculating a feature's trustworthiness. The user is responsible for providing the method's missing details. Using the humidity example from above, a measure of trustworthiness might be the average or the standard deviation of the set of calculated quality values.

The input data of a classifier or a regression model is referred to as the "signal". Examples of signal data are temperature or the change in pressure due to a chemical reaction. This thesis proposes an approach for estimating the amount of noise present within a signal. Two approaches to measuring data quality are also proposed, the ratio of signal-to-noise and noise-only.

1.3 Research methodology

This research begins with a review of the literature to establish the challenges in selecting features and constructing models for non-classical problems. Gene expression (or activity) provides a suitable example of non-classical problems, since data sets consisting of one or two hundred samples, ten to twenty thousand features and substantial noise are common. Biomedical topics related to gene expression and two diseases are reviewed. Childhood leukaemia provides the main source of real-world data, which is strongly linked to genetics. Chronic fatigue syndrome is also used, since it is weakly linked with genetics and therefore provides a challenge to FUR with respect to discriminating between noise and a weak correlation between

disease and its classification. As microarrays (Southern, 2001) provide the source of gene expression data, a thorough review is made of the challenges in selecting features and building reliable models. Lastly a review of four aspects of data mining, consisting of noise, data pre-processing (Rahm and Do, 2000), feature selection and model learning is included.

There is a need to focus on the relationship between noise levels and sample sizes during model learning, in order to simplify the complex relationship between noise and feature selection. This suggests the importance of using an area of Computational Learning Theory called Probably Approximately Correct (PAC) learning theory (Valiant, 1984; Bshouty, Eiron, and Kushilevitz, 1999), which provides a theoretical framework for quantifying the impact of noise on sample size requirements and provides evidence for the benefit of this research.

The final part of the research methodology consists of substantial experimentation, using synthetic and real-world data sets. In addition, the principle method for evaluating the effectiveness of FUR consisted of comparing and contrasting results with a traditional feature selection methodology. Synthetic data is used since it provides an environment where the impact of noise can be controlled and assessed, while the real-world data sets provide an appreciation of the merits of the developed methodologies on challenging problems.

1.4 Scope

The focus of this investigation begins with the development of a generic approach for estimating the quality of individual dimensions, or features, of each data point. A generic approach is important since it provides an adaptable feature selection methodology, where prior domain knowledge determines the actual measure of quality used. Using these individual measures of quality, this research develops a measure of confidence, or trustworthiness in the information provided by each feature. Finally an approach for selecting features on the basis of their utility (or interaction between trustworthiness and information) is investigated.

A distinguishing characteristic of the developed feature selection methodology is its exploitation of feature redundancy. Unlike other feature selection methodologies in the literature, this research depends on removing features that are insufficiently trustworthy and replacing them with more trustworthy alternatives. By doing so, this approach overcomes the inability of traditional feature selection to distinguish between information generated by the underlying structure of the data set and information generated by noise. An extension to this feature selection methodology, which considers the quality of the test set and unlabeled data, is also investigated.

1.5 Limitations

Successful comparison of the traditional and the developed feature selection methodologies depends on a number of factors that are described in this section. The interaction between noise and the information hidden within the underlying structure of the data is an important factor in comparing the methodologies. Sufficient noise within the data set is required in order to cause enough detrimental features to be selected by the traditional methodology and used in the correspondingly built models. However the nature of this interaction is unpredictable and therefore sufficient data sets and experimentation is considered necessary to detect differences between the methodologies.

A number of features that are selected by traditional feature selection for use in a model, will contain apparent information. It is necessary that many of these features have redundant versions that were not used, since they provide less information. Although providing less, there must still be enough information to enable FUR to select these features and generate models that statistically outperform the traditional counterpart.

The measure of data quality used for experimental purposes must provide an accurate measure for the noise present within the data set. Similarly, the method used to convert data quality into feature trustworthiness must

provide sufficient sensitivity and summary of the quality issues impacting individual features.

A number of potential experimental variables were held constant in order to facilitate the effective comparison of the feature selection methodologies. For example, it is assumed that the trustworthiness and traditional feature selection thresholds used support an effective comparison.

1.6 Thesis structure

The thesis is structured into five chapters: introduction, literature review, theoretical framework, experiments and results and lastly the thesis conclusion.

Chapter 2, the literature review begins by describing the main source of real-world data used for experimental purposes. This consists of a brief description of molecular biology and two common diseases: childhood leukaemia and chronic fatigue syndrome. The main source of the disease data, which was generated using microarrays, is then reviewed. This review also describes the sources of noise and general difficulties in constructing accurate models using microarray data. Data mining forms the next major section of the literature review. First containing an orientation on the problem of noise and the role of data pre-processing, followed by a review of feature selection and theory on model learning. Although the impact of noise on appropri-

ate feature selection forms the topic of this thesis, the literature on model learning provides the best source of material for understanding the impact of noise and also generating a theoretical framework. The foundations of the framework, which consists of PAC Learning theory, enable the development of a model on the interaction between noise and classification accuracy.

Chapter 3, containing the theoretical framework moves from a discussion on the nature of noise, to using PAC theory to develop a theoretical relationship between noise and the number of samples required for a minimum level of classification accuracy of a model. This model clearly demonstrates the dramatic impact of noise on learning to a given level of accuracy and how quickly a limited sample size can render the problem un-learnable. A number of experiments are then used in order to test the developed theoretical relationship between noise, sample size and minimum accuracy. The experiments consist of incrementally adding noise to a publicly available data set, constructing a model and testing classification accuracy. The publicly available data set is a subset of the 1994 American Census. The experiments show how quickly noise can render a problem un-learnable and are also supportive of the developed theory.

Formal definitions for data quality and a number of other key terms are then developed before describing the developed feature selection methodology called "Feature Utility Ranking". The remainder of the chapter then builds

on the description of Feature Utility Ranking and develops variations called FUR1 and FUR2. The key difference between these two variations is the measure of data quality, where the former uses a measure of signal-to-noise and the latter, uses only a measure of noise.

Chapter 4, containing experiments and results, begins with a description of the experimental design and the data sets used. Firstly a synthetic data set is used, since it provides a controlled environment, particularly with respect to the amount of noise within it. The use of this data set mimics the American Census data experiments, except that FUR1 and FUR2 are used in addition to the traditional feature selection methodology. Another key difference is the development of regression models, rather than classification.

A series of real-world data set experiments are executed using the childhood leukaemia and chronic fatigue syndrome data sets. The leukaemia data set consists of two parts: classification by cell type and the prediction of treatment outcome. The chronic fatigue syndrome data sets are used to predict whether the patient has fatigue. The experimental approach consists of repeating the same experiment, for traditional, FUR1 and FUR2 methodologies. These experiments consist of comparing the classification accuracy of the constructed models and changes in the feature sets produced. Since the presence of redundant features is key requirement for the use of Feature Utility Ranking, experiments are also performed to access the degree of

redundancy within the individual data sets.

Chapter 5 presents concluding results on the research and a number of recommendations for further research. In addition there is an appendix containing a different interpretation of the synthetic data experimental results, a glossary of terms and a bibliography.

Chapter 2

Literature review

The overriding goal of this literature review is to establish the best practices for constructing classification models for domains characterized by: few samples, many dimensions and significant noise. Problems with these domain characteristics are termed "non-classical". In contrast, classical problem domains are characterized by many samples, few dimensions and less noise.

This goal considers the strengths and weaknesses of current approaches and facilitates the development of an approach to reduce the impact of noise. For the purposes of this thesis, noise is defined as "an external additive source that artificially causes variability within the data set".

Since non-classical problems are the focus of this thesis, a domain that provides an ideal example was selected—microarray data analysis. Microarrays and in particular two-channel cDNA microarrays, are known to be impacted by significant noise. *It is stressed that any non-classical problem, where an explicit measure of noise or data quality can be devised, can benefit*

from the contributions made by this thesis.

The chapter starts by reviewing related biomedical topics: basic molecular biology, including gene expression; two diseases (leukaemia and chronic fatigue syndrome); microarray basics; microarray technology; accepted challenges and issues regarding noise.

Finally the generic issues of classification model construction are established from the perspective of data preprocessing, feature selection and classification model construction. Using these generic issues it becomes clear that current practices assume: many samples, few dimensions and little or no noise—which is in conflict with the characteristics of a non-classical problem domain.

2.1 Biomedical background

This section presents an orientation on the biomedical domain related to the empirical content of this thesis. The following topics are presented: genomes, gene expression, cancer, leukaemia and chronic fatigue syndrome.

2.1.1 The cell

The cell is the most fundamental unit of the body and is the smallest structure capable of providing all the processes that define life in a multicellular organism. All cells store their genetic, or hereditary information, in

molecules of **d**eoxyribo**n**ucleic **a**cid (DNA). The genome corresponds to the complete set of genetic information, while the DNA is a medium that carries it. DNA consists of simple subunits based on four separate nucleotides **A**denine, **C**ytosine, **G**uanine and **T**hymine that provide the sugar-phosphate molecule and a nitrogen-containing side-group, which make up the double helical structure. All cells transcribe portions of their hereditary information into the same intermediary form, which is known as **r**ibo**n**ucleic **a**cid (RNA).

For humans, the DNA within a nucleus is divided into 48 pairs of chromosomes and consist of 3.2×10^9 nucleotides. Each chromosome consists of a single but enormously long strand of DNA molecules. This strand is made into a compact structure through folding and packing around proteins called "Histones". Other proteins, responsible for winding and unwinding the strand, are also involved in processes such as DNA replication, DNA repair and gene "transcription".

The DNA within the genome provides a blueprint for protein construction and each gene within the genome provides the necessary information for a particular protein (Alberts, Johnson, Lewis, Raff, Roberts, and Walter, 2002). Through the process of transcription DNA synthesizes mRNA, which is a copy of one strand of the DNA double helix. Then the process of "translation" uses the mRNA to generate protein. Proteins, the major constituent of cells, have a "chain" like structure, composed of over twenty different amino

acids. Many thousands of different proteins are known (Alberts et al., 2002, p.129). Proteins are essential components of the body and form structural material, like muscles, tissues and organs. Proteins are also important as regulators of functions as well as enzymes and hormones (McFerran, 1996). Enzymes are biological catalysts and control the rate of reactions. Hormones are responsible for modifying the structure or function of tissues, generally located far from the creation site of the hormone.

2.1.2 Cancer

In general terms, a cancerous cell consists of two characteristics: damaged genetic code and an inability to respond to the basic rules of behavior for multicellular organisms (Alberts et al., 2002). Normal cells possess mechanisms designed to minimize the opportunity for incorrect coding and rogue behavior, such as repairing errors in the the DNA and "apoptosis" or programmed cell death. However, cancerous cells fail to respond to these mechanisms and this results in the uncontrolled proliferation of cells.

One of the basic mechanisms of cell management in multicellular organisms is localization. However cancerous cells are invasive since they can break loose from their primary site, enter the bloodstream or lymphatic vessels and survive in other parts of the body. This ability to spread, or perform "metastases", combined with uncontrolled proliferation, often makes surgi-

cal treatment of cancer difficult or impossible. Hence, treatment of cancer requires the use of toxic drugs or radiation—to indiscriminately kill cells in the hope of eradicating *all* cancerous cells.

Classification of cancers

According to the international standard for the classification of cancer, defined by the *International Classification of Diseases for Oncology*[1], there are two classification methods: histological type and primary site. Histological classification deals with the type of tissue from which the cancer originated, while primary site classifies cancer according to the location where the cancer first developed.

Histological classification groups all cancers into five major categories: carcinoma, sarcoma, myeloma, leukaemia and lymphoma. Additionally there are cancers of mixed types.

A carcinoma is a malignant neoplasm of epithelial origin. Epithelial tissue is found throughout the body, it exists as skin; as a lining or covering of organs and internal passageways, such as the colon. Carcinomas account for 80% to 90% of all cancers. Most carcinomas affect organs capable of secretion: breasts, lungs, colon, prostate and bladder.

A sarcoma is a cancer of connective tissue. Examples of connective tissues

[1]Third Edition of the International Classification of Diseases for Oncology

are bones, tendons, cartilage, muscle, fat and lymphatic vessels. The most common sarcoma occurs as a painful mass on the bone. Sarcomas generally occur in young adults.

Myeloma is a cancer that originates within plasma cells of the bone marrow. These plasma cells are responsible for some of the proteins found in blood.

A lymphoma is any malignancy of the lymphatic system. The lymphatic system is a network of vessels, nodes and organs that purify bodily fluids and is responsible for the distribution of infection-fighting white blood cells called lymphocytes. Lymphomas consist of solid tumors that can also occur in the stomach, breast or brain.

2.1.3 Leukaemia

Leukaemias are cancers of sites in the bone marrow responsible for manufacturing blood cells. In Greek, the word leukaemia means "white blood". Blood cells are formed in the bone marrow, which is the soft spongy center of bones. New or immature blood cells are called blasts. Some blasts mature within the marrow, while others travel to other parts of the body to mature.

Blood consists of fluid called plasma and three types of cells: white blood cells or leukocytes, red blood cells or erythrocytes and platelets or thrombocytes. White blood cells play a roll in fighting infections and diseases.

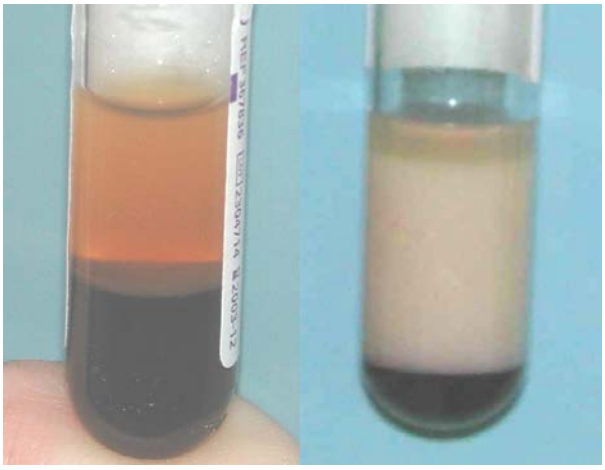


Figure 2.1: The left and right test tubes respectively contain blood from subjects without and with leukaemia. The Greek word for leukaemia means "white blood".

Red blood cells carry oxygen from the lungs to the body's tissues and transport carbon dioxide from the tissues back to the lungs. Platelets help the formation of blood clots which control bleeding.

By using a centrifuge, two major components of blood, consisting of white and read cells, are easily seen in the left test tube of figure 2.1. Blood cells are normally produced as required, however when leukaemia occurs—large numbers of abnormal blood cells are produced. In most types of leukaemia the abnormal cells are white blood cells, as seen in the right test tube.

Leukaemias are generally associated with overproduction of immature white blood cells. As white blood cells are responsible for managing infection, leukaemia patients are typically prone to infections. Leukaemia can also affect red blood cells, causing poor blood clotting and fatigue due to anemia.

Childhood leukaemia provides the main source of real-world data for this thesis.

Adaptive immune responses are carried out by white blood cells called lymphocytes. Two major responses provided by the immune system consist of antibodies and cell-mediated responses, which are carried out by two different lymphocytes called B-Cells and T-Cells. Antibody responses that are performed by B-Cells, secrete antibodies that are proteins called immunoglobulins. Cell-mediated immune responses involve T-Cells that react against foreign antigens (Alberts et al., 2002).

Classification of leukaemia

There are several types of leukaemia, which are grouped in two ways. Firstly according to the disease's rate of development, either "acute" or "chronic". Secondly by the type of blood cell affected, white or red blood cells.

Acute leukaemia involves very immature blood cells or blasts and because of their immaturity they are unable to carry out normal function. The number of blasts increases rapidly, as does the severity of the disease. In chronic leukaemia there are some blasts, although in general, blood cells are more mature and therefore perform a degree of normal function. The rate at which blasts are produced in chronic leukaemia is far less than for acute leukaemia and as a result, the development of chronic leukaemia is much more gradual.

Leukaemia affecting white blood cells or lymphoid cells, is called lymphocytic leukaemia. Myelogenous leukaemia involves red blood cells, which are also known as myeloid cells.

There are four subcategories of lymphocytic and myelogenous leukaemia. In the case of lymphocytic leukaemia there is Acute Lymphoblastic Leukaemia (ALL) and Chronic Lymphocytic Leukaemia (CLL), while for Myelogenous leukaemia there is Acute Myeloid Leukaemia (AML) and Chronic Myeloid Leukaemia (CML). Acute Lymphoblastic Leukaemia is most common for children, but it also affects adults, especially those over 64 years. Chronic Lymphocytic Leukaemia mostly affects adults over 55 years, although younger adults are sometimes affected, but children are almost never affected. Acute Myeloid Leukaemia, which is sometimes called acute nonlymphocytic leukaemia (ANLL), occurs in adults and children. Chronic Myeloid Leukaemia occurs mainly in adults, although a very small number of children are affected.

Treatment of leukaemia

An important step in the treatment of leukaemia is risk stratification, which involves classifying a patient according to the severity of the disease. The patient's classification determines the aggressiveness of the treatment employed (Ludwig, Haferlach, and Schoch, 2003; Pui, 2003; Greaves, 2002).

Treatment outcome is determined after five years has elapsed since diagnosis.

2.1.4 Chronic Fatigue Syndrome

Chronic Fatigue Syndrome (CFS) is an illness with a primary symptom of debilitating fatigue over a six month period (Afari and Buchwald, 2003). Currently diagnosis of CFS is generally made by clinical assessment of symptoms using a number of surveys measuring functional impairment, quantifiable measurements of fatigue and occurrence, duration and severity of the symptoms (Reeves, Wagner, Nisenbaum, Jones, Gurbaxani, Solomon, Papanicolaou, Unger, Vernon, and Heim, 2005). A primary goal of current research is to derive a definition of the syndrome, which goes beyond a clinical assessment of symptoms to an empirical diagnosis founded on an established biological lesion. The motivation for this kind of research is to gain a clearer understanding of the illness and to find empirical guidelines for its diagnosis.

The same level of detail is not given for CFS as for leukaemia. Instead CFS provides a data set with a weak correlation between classification and genetics, since CFS has a psychosocial origin (Krupp, Mendelson, and Friedman, 1991). Using the CFS data set tests whether FUR still operates correctly given a weak correlation between the model's inputs and output.

2.2 DNA Microarrays

DNA microarrays, developed in the 1990s, provide a quantitative analysis of patterns of gene behavior (Schena, Shalon, Davis, and Brown, 1995). A DNA microarray consists of an optical microscope-like support, upon which an array of microscopic DNA spots exist, with each spot representing an individual gene (Watson, Meng, Thompson, and Akil, 2000).

The number of DNA spots, or "genes represented" on an oligonucleotide microarray can vary substantially—from hundreds to tens of thousands and thereby permitting that number of simultaneous experiments. Given the number of genes, microarray experiments have provided behavior patterns on a large scale. Fundamentally, these experiments allow the determination of activity levels, or "expression", for each of the genes represented by the microarray spots, see figure 2.2.

Definition 1 *Expression measures a gene's activity level. In principle, the greater the activity—the more mRNA is produced. Expression acts like a throttle for subsequent biological function.*

2.2.1 Microarray basics

Microarray basics are described using a two-channel cDNA microarray. Each array spot contains genetic code, which itself is built using cDNA. This code

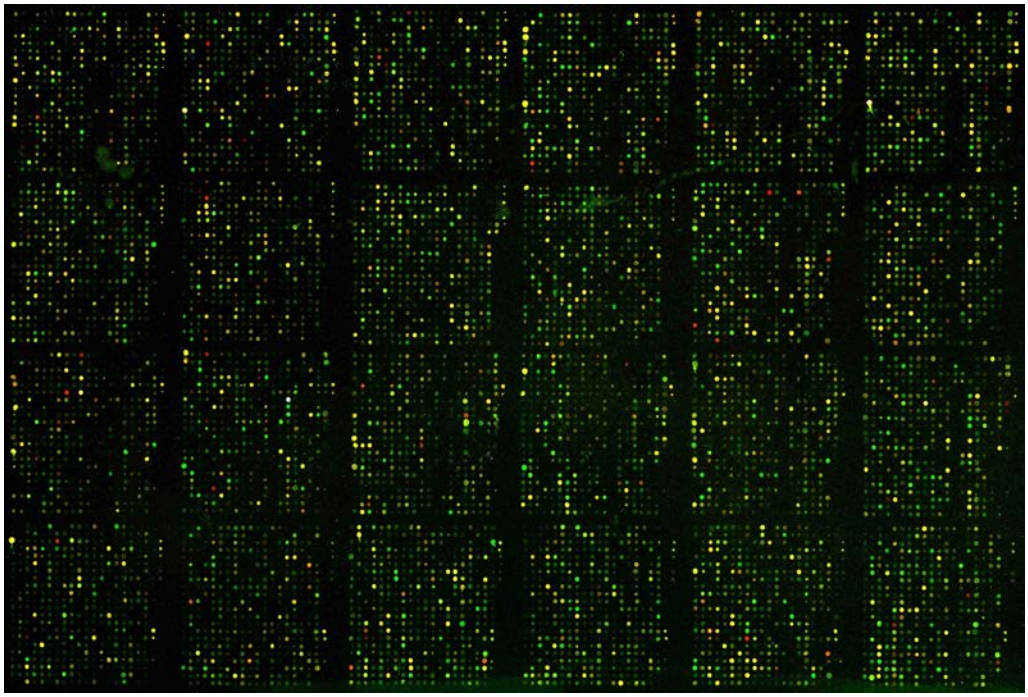


Figure 2.2: DNA microarray contains an array of DNA spots, which can vary from hundreds to tens of thousands. Shown is a two-channel microarray, where a spot's color provides a relative measure of expression for each gene.

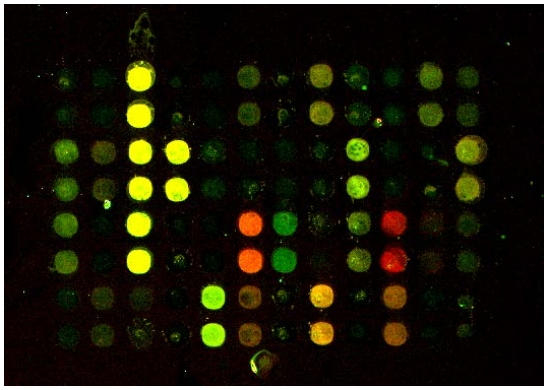


Figure 2.3: Close up of microarray DNA spots, showing a variety of colors that correspond to varying levels of expression.

represents a single strand of the DNA helix and contains enough code to uniquely identify the gene.

In the case of a two-channel microarray, the color of a spot represents the relative difference in expression between the two-channels. A red spot indicates greater expression by one channel, green indicates greater expression by the other, while equal expression results in a yellow spot, see figure 2.3.

A biological sample consisting of cells is required to use a microarray. Using this sample, mRNA is extracted and then hybridization to the microarray spot whose cDNA is its genetic complement. Therefore hybridization provides a method for binding mRNA with its corresponding gene, which is a spot with known location on the microarray. By tagging the mRNA with a fluorescent dye prior to hybridization—the quantity of mRNA is estimated by measuring the spot's fluorescence in response to laser excitation.

A gene's level of expression is considered to be proportional to the quantity of mRNA produced. A gene's expression level provides only presumed biological insight, since biological activity is ultimately indicated by protein activity, which is related to mRNA expression. A solution to this is two-channel microarrays; where one channel provides a "test" and the other a biological "control". For example, if the gene profile difference between cancerous and normal biological behavior is required; the test and control channels would be constructed from cancerous and normal biological samples respectively, see figure 2.4.

2.2.2 Microarray noise

Microarrays are affected by noise, which alters the "true" value of a data point and the degree of alteration is proportional to the noise. It is generally accepted that microarray "noise" *alone* is a significant hurdle to accurate data interpretation. It is responsible for substantial *variability* in measured gene expression—even for experiments that consist of a single biological sample (Tu, Stolovitzky, and Klein, 2002; Baldi and Brunak, 2001; Bolstad, Irizarry, Astrand, and Speed, 2003).

There are many sources of variability caused by noise and using the work of Parmigiani, Garrett, Irizarry, and Zeger (2003) we will describe a number of noise sources and how they impact microarray data. Microarray noise can

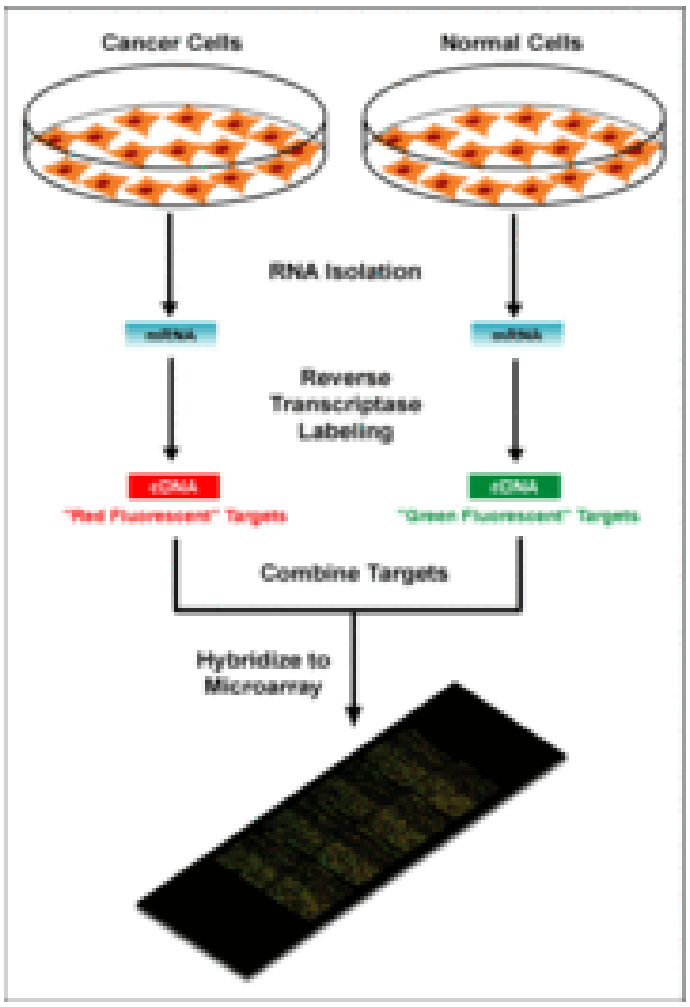


Figure 2.4: Using two-channel microarray technology to transform biological samples into cDNA. Once cDNA is constructed from the samples, the cDNA is combined and hybridized onto the microarray.

be decomposed into two areas: *biological* and *technical* noise sources (Aris, Cody, Cheng, Dermody, Soteropoulos, Recce, and Tolias, 2004).

With respect to a biological source of noise, Raser and O'Shea (2005) states "noise has multiple sources, including the stochastic or inherently random nature of the biochemical reactions of gene expression". Some causes of biological noise are: variations between patients or tumor locations, variations in the cellular composition of tumors and genomic instability of tumors due to heterogeneity of the genetic material (Aris et al., 2004). However this form of biological variability is either uncontrollable or of no significance in the context of microarrays—since *overall* production of mRNA is being measured. This leaves technical noise as the only source of (potentially manageable) variability.

Regarding technical variability, Parmigiani et al. (2003) see microarrays as a powerful tool but cautions that variability, which occurs throughout the measurement process, can obscure relevant biological signals. They classify these sources of variability into five areas: microarray manufacturing, mRNA preparation, hybridization, scanning and imaging. Each of these represent phases of data acquisition.

With respect to microarray manufacturing, a number of different sources are responsible for variations within the resulting data set. Since some of the sources are technology specific—we will focus on two-channel cDNA mi-

croarrays. The chemical processes of: amplification, purification and concentration of the DNA fragments are all responsible for variability. One reason for this is due to the physical size of DNA itself. This results in unavoidable variability in the number of fragments constructed and used. Spotting the DNA onto the microarray also results in variability in the quantity deposited, the amount actually bound to the array and the resulting spot shape. Spot shape is important as later processes, largely automated, require consistency in order to measure the full quantity of material while excluding any background signal. There are also systematic variations due to print tip inconsistencies and variations within the operation of robotics equipment used for manufacturing.

Similar issues impact the preparation of mRNA from biological samples. In addition, the preparation protocols used and inconsistencies within protocol execution are a source of variation. Humidity and temperature are also significant factors as different array batches can induce dissimilar biases and variations, while ambient temperature can affect the chemistry through reaction times and deterioration of arrays (Wildsmith and Elcock, 2001). In the case of two-channel cDNA microarrays and labeling, differential performance of dyes across sample classes can lead to biases within bonding and fluorescence.

Hybridization also suffers variability due to ambient temperature and

humidity. Edge effects, (ie. genes spotted near array edges) can impact uniformity of hybridization (Parmigiani et al., 2003), as can slight homogeneity of solutions, extraneous molecules, the binding of dust and cross hybridization of molecules whose sequences are insufficiently dissimilar. Washing an array after hybridization can result in the loss of material that has failed to bind sufficiently. Parmigiani et al. (2003) warns that although individual errors are small, "the compounding of their effects can be significant" and that the same sequence spotted in multiple locations can result in variations.

Noise is a significant issue for successful microarray data analysis (Aas, 2001) and clearly methods are needed to reduce the impact of noise, wherever possible. As will be shown later, there are numerous methods available but all depend upon one or more specific aspects of data analysis. For instance, some methods are integrated into data preprocessing and others within feature selection.

Managing noise

Our research goal is a technique that enables feature selection to not only consider the discriminative ability of individual genes—but also the degree of noise or variability present within the available data that describes each gene. Not all of the sources of variability described above are able to be quantified at this time; however there are techniques being developed that

are moving in that direction.

2.3 Data mining

The impact of noise and how it is managed during preprocessing, feature selection and model learning is presented in this section.

2.3.1 Noise

Noise can be found in many aspects of life, for example, verbal communication between people or telecommunications. Using the Cocktail Party Effect, Cherry (1953) showed that humans are able to tolerate some background noise and although many words may be missed, can still understand the concept being conveyed. In short, successful communication is more difficult in the presence of noise but may still be possible.

Sample data, such as microarray sample data, can be viewed as a linear combination of signal and noise, where signal is the quantity of interest. Accuracy in this case is determined by the degree of noise present. The problem which arises is how to define accuracy and has been considered in many disciplines for decades. Accuracy of information transmission can be measured via a signal to noise ratio (signal-to-noise). Here the greater the signal component compared to noise, the greater the accuracy and the less chance for transmission error. Analysis of microarray data is analogous to

the problem of accurately transmitting information held within the sample data to the classification model being constructed.

Wiener (1949) was perhaps the first to consider how best to estimate the values of the received signal while accepting inevitable noise. Shannon (1948, 1949) greatly added to this area of work on "information theory" (Cover and Thomas, 1991) and showed both the potential and challenges of error-free transmission of information. Of particular relevance regarding the work of Wiener and Shannon, was the necessity to understand the nature of noise to the point where noise itself can be modeled. Only when noise can be modeled can the original signal of interest be reliably decoded. These findings were affirmed by the work of Carlson (1975), who found that noise cannot be universally removed from all forms of information transfer via a single approach. Instead the management of noise requires awareness of the information transfer problem in question.

To our knowledge, the work to date within the area of microarray data analysis has focused on one of two approaches with regards to the management of noise. One approach ignores the effects of noise, while the other tries to remove it through some scheme. Ignoring noise can only succeed if the signal-to-noise ratio is sufficiently high to render the effects of noise insignificant. *However this approach risks constructing a model where the consequences of noise are unknown.* The approach of removal or reduction

of noise requires understanding the noise component. This thesis considers a novel approach to noise management, it considers the trustworthiness of the data. Rather than ignoring noise or seeking to extract signal which is unaffected by noise, our approach determines which features or dimensions within the data set can be trusted. Clearly approaches which ignore the presence of noise risk unpredictable consequences, while those that remove noise are dependent upon the accuracy of the noise model.

2.3.2 Preprocessing

Real world data is typically tainted by noise and other data quality issues. Data preprocessing can have a significant impact on improving the accuracy of the data and the models (Kotsiantis, Kanellopoulos, and Pintelas, 2006). Although preprocessing typically accounts for about 80% of the total modeling effort, it is often performed inadequately (Zhang et al., 2003; Bolstad, 2007). In other cases, preprocessing is treated as a "black box" process and is used without regard to the actual data or intricacies of the methods used (Bolstad, 2007).

Preprocessing is composed of three tasks: sampling, cleaning and transformation (Tan, Steinbach, and Kumar, 2006; Rahm and Do, 2000; Kotsiantis et al., 2006). Sampling refers to methods for selecting a subset that is representative of the entire data set, cleaning deals with removing errors or noise

from the data, while transforming, which includes normalization, converts data into a more appropriate form, for example, a data range that accommodates the algorithms to be used. We review those tasks, with a focus on "cleaning", since our interest is in managing data quality (Kotsiantis et al., 2006).

Sampling is required when far more data exists than is needed or its collection is prohibitive due to factors like cost. Consequently a subset of the data is used (Tan et al., 2006). One of the tasks of sampling is to provide training and test sets.

Once a completed microarray exists, the sampling task also consists of producing an image via laser scanning (Bolstad, 2007). This image is then processed to extract an estimate of each gene's expression, in the case of Oligonucleotide arrays or change of expression for cDNA arrays. Two subtasks: segmentation and summarization are required to obtain this expression reading (Parmigiani et al., 2003). Segmentation defines the parts of the array that contain expression images; one image per array spot. One approach to segmentation involves manually placing a mask over the array and only allowing spot images to be visible. Segmentation is generally imperfect as arrays typically consist of thousands of spots. Summarization is then used to produce a single expression reading for each array spot. Given there is human involvement during segmentation, quality issues are likely to emerge at

this early stage. Summarization is potentially problematic since artifacts can dramatically induce signal variations across a spot (Wang, Ghosh, and Guo, 2001). A large number of different methodologies exist for these subtasks, as well as substantial differences in the expression measured (Parmigiani et al., 2003).

Data that consists of outliers, missing information, or noise requires cleaning. Outliers are data points that fall outside of the expected or measurable range. In the case of microarray data, an example of outliers are data points located at the limits of the measurable range of spot intensity (Bolstad, 2007). Missing information occurs when a needed attribute is not supplied, a missing value for a data point exists, or few data points exist within a region of particular interest within the sample space. For microarray data, a missing attribute occurs when a gene of particular interest is not present on an array. Missing values can also occur due to problems with image resolution, the presence of dust, scratches and systematic artifacts during array printing (Berrar, Granzow, and Dubitzky, 2007). Outliers can also exist within the measurable range but are sufficiently different to other data points to arouse suspicion about their likely accuracy. Noise, in general, is responsible for these issues and data cleaning must determine how outliers should be adjusted or what value should be used for missing data.

Data cleaning requires knowledge of the distribution responsible for the

data in order to infer whether a data point is likely to occur (Walpole and Myers, 1978). However the distribution is generally not known and therefore must be estimated from the available data. This, in the case of a parameterized model, requires selection of a model and estimation of its parameters. This requires a sufficiently large sample that is sufficiently unaffected by noise. Determining what constitutes a sufficiently "large sample" and "sufficiently unaffected by noise" has been well studied in classical statistics (Adcock, 1997; Walpole and Myers, 1978; Vapnik, 2000; Hastie, Tibshirani, and Friedman, 2001). However domains like microarray data analysis fail to comply with the requirements of classical statistics (Sima and Dougherty, 2006; Tsai, Wang, Chen, and Chen, 2005; Dougherty, 2001), making preprocessing difficult due to the effects of a non-classical problem domain: high dimensionality, few samples and noise.

For quality purposes of microarrays, cleaning also depends upon visualization of each array (Parmigiani et al., 2003; Wang et al., 2001). Visual inspection of every microarray is essential for diagnosing the presence of artifacts. Some of the visual inspection methods used are (Parmigiani et al., 2003): boxplots for diagnosing print tip effects; relative expression plots; "gradient plots" and an MA plot, which is the distribution of the red/green intensity ratio ('M') plotted by the average intensity ('A'). The latter methods are used for identifying spatial biases. Cleaning also deals with background sub-

traction, which is needed to remove noise generated from images outside of the spot. However background correction is problematic since the intensity of the background can exceed that of the foreground, yielding a negative result, or increased variance from the combination of two intensity measures. Yang, Buckley, and Speed (2001a) as well as other researchers "have found that the background estimates produced by some of the most popular image-processing algorithms are not sufficiently reliable". Some researchers do not perform background correction in order to avoid these problems.

Transforming sample data generally alters the entire sample set or a subset, while cleaning tends to alter individual sample points. An example of a transformation is altering the range of a variable from $[-1000, +1000]$ to $[0, 1]$ in order to comply with the requirements of a learning algorithm (Tan et al., 2006). This type of transformation is generally called normalization. In the context of microarray data analysis, normalization is used to correct biases within and between arrays (Bolstad, 2007). Therefore normalization is used to adjust the arrays so that all have some common attributes, for example, the same average change in expression (Bolstad, 2007; Yang, Dudoit, Luu, and Speed, 2001b; Wang et al., 2001).

Transformations are an important tool in microarray analysis because systematic biases can exist, both within an array and across arrays (Parmigiani et al., 2003). Within-array normalization generally involves an iterative pro-

cess of visualization, identification of likely artifacts and their removal. However Tseng, Oh, Rohlin, Liao, and Wong (2001) and Yang and Speed (2002) have presented evidence that automated normalization is difficult. Some of the normalization algorithms used are: location normalization, which uses robust locally-weighted regression; scale normalization, which uses a robust estimate of scale, such as the median absolute deviation (Dudoit and Yang, 2003) and a Locally Weighted Scatterplot Smoothing (Loess) curve (Cleveland and Devlin, 1988), which estimates a smooth non-parametric curve (Bolstad, 2007). Normalization across arrays is simpler than within-array normalization. However some additional approaches are used, including standardized overall intensity; quantile normalization, which uses a distribution of intensities for each array, so all arrays achieve the same values at specified quantiles and print-tip-group normalization, which assumes that all log-ratios from different print-tip groups follow a normal distribution with zero mean and estimated variance scale factor (Yang et al., 2001b).

Data analysis consists of two phases: preprocessing and model construction (Kotsiantis et al., 2006; Zhang et al., 2003). Typically what is passed from preprocessing into model construction is an altered version of the raw data set and the nature of the changes that took place in producing the altered data set are only known by the preprocessing phase. Ideally a measure of the changes should also be passed onto the model construction phase (Lu,

Sam, and Sung, 1996). For instance, knowledge that a sample point was unaltered, since it was judged as already accurate, should be available in some form to model construction. Similarly, knowledge that a sample point was originally an outlier and therefore altered substantially should also be passed through to model construction. Without such knowledge model construction treats both sample points similarly.

Traditional data analysis reserves a subset of the entire data set for model construction, while the remaining data is used for model testing. Test and unlabeled data pose a significant problem for traditional data analysis, since their quality and the impact of quality issues are unknown by feature selection and model construction (Maindonald, 2006). Thus a question exists in how to manage differences in the quality of the training and test sets as well as the quality of unlabeled data.

In the case of microarray data, Bolstad (2007) believes that "preprocessing is a very important step" and although common techniques generally work well, it is not unusual for some arrays to be eliminated. He concludes that, "to some degree most standard normalization methods" depend on one of two assumptions: "the number of genes changing between conditions is small relative to the number of genes being measured on the microarray" and "an approximately equivalent number of genes are increasing in expression value as are going down in expression value between conditions". However

if "either of these assumptions are violated then there is a possibility that small changes in expression might be made undetectable, and large changes made smaller". If such problems are the product of noise, one is placed in a position where determining whether noise or biological function were responsible for the observed signal. The addition of a metric that indicates the likely quality or accuracy of the data could reduce such uncertainty.

Preprocessing activities are numerous and varied. As a result it appears impossible to compile a measure of the changes that have occurred, which could be used to guide model construction. Although feature selection is considered part of preprocessing, it tends to be performed at the end and like model construction, could benefit from knowledge of the changes that occurred. This research considers how a simple measure of the degree of change caused by pre-processing could be translated into a measure of data quality and in turn used by feature selection, so features that are more "trustworthy" are preferred.

2.3.3 Feature selection

Feature selection is reviewed with a focus on the microarray analysis domain. The principle interest here is the impact of high dimensionality and few samples on feature selection.

Feature selection can be defined as a process that selects a subset of

attributes so that the original feature space is optimally reduced according to an evaluation criterion (Liu and Yu, 2002). Due to the complexity and size of this optimization problem, generally a near optimal set is found, rather than the global optimum. Dash and Liu (1997) decompose the evaluation criterion into four steps: subset generation, subset evaluation, stopping criterion and result validation. Validation estimates the subset's likely effectiveness against the test set and unlabeled data. Using these four steps and the characteristics of the microarray analysis problem space, namely high dimensionality, few samples and noise, the challenges, best practices and weaknesses of feature selection are reviewed.

Feature subset generation

Generating a subset involves searching a feature space whose size is 2^P, where P is the total number of features. Two issues arise in this search: the size of the search space (P) and the time required to evaluate each candidate subsets. A candidate subset has dimensionality p, where $p < P$. This search is characterized by three different strategies: complete, heuristic and non-deterministic search (Liu and Motoda, 1998).

Complete search requires exhaustive exploration of the entire space, hence 2^P candidate subsets. However, even a feature space of moderate size can result in an intractable search problem (Kohavi and John, 1997) and search

in general has been shown to be NP-hard (Blum and Rivest, 1992). Search spaces in microarray data analysis, consist of approximately ten thousand features. So exhaustive search is not computationally feasible (Silva, Hashimoto, Kim, Barrera, Brandao, Suh, and Dougherty, 2005; Hua, Xiong, Lowey, Suh, and Dougherty, 2005). This high dimensionality requires a heuristic or non-deterministic search strategy.

A heuristic is a "rule of thumb" that generally provides a good solution (Shapiro, 1992; Russell and Norvig, 1995). Heuristic search strategies seek to find a good solution without employing exhaustive search. Two examples of exhaustive search are "hill climbing" and "best first search" (Cawsey, 1998). Although a heuristic strategy resolves the intractable problem of exhaustive search, it does so by introducing the risk of missing an optimal solution. Because of its dependence on a "rule of thumb", heuristic search has limited vision regarding the implications of a search path. A search path that appears promising may prove unsuccessful (Blum and Langley, 1997; Shapiro, 1992). Consequently they are unable to guarantee success. In addition, heuristics may not exist for some data types, for example data describing the quality or accuracy of a data point.

Another solution to intractability is non-deterministic search, which employs a random search strategy (Blum and Langley, 1997). Although intrinsically random in nature, non-deterministic search often finds good so-

lutions efficiently, but this generally occurs in small to medium sized search spaces. Additionally, this strategy is again unlikely to find the global maximum (Saeys, Inza, and Larranaga, 2007; Sun, Todorovic, and Goodison, 2008). Because of the random nature, repeated searches are unlikely to produce the exact same solution every time (Liu and Yu, 2002), even when all other parameters are held constant. Non-deterministic search is likely to be computationally efficient for large search spaces, such as those typically found within microarray data analysis. However efficiency is likely to improve if many near optimal solutions exist. Microarray data is known to contain numerous near optimal or good quality solutions (Sima and Dougherty, 2006; Hua et al., 2005; Ein-Dor et al., 2005).

Given the computational expense of executing complete and non-deterministic searches, the impact of very large problem spaces has been addressed in recent times by suboptimal search strategies or heuristic and non-deterministic methods (Sima and Dougherty, 2006; Hua et al., 2005).

Feature subset evaluation

Subset evaluation measures the optimality of a candidate feature subset (Saeys et al., 2007). Supervised approaches evaluate their optimality according to its ability to classify data (Dunham, 2003). An unsupervised approach either ignores class correlations or utilizes data that is uncorrelated to class

information (Guyon and Elisseeff, 2003). Examples of uncorrelated data in microarray data analysis are the standard deviation of spot pixel intensity and spot shape. However no examples were found in the literature where such uncorrelated data is used within feature selection or model construction in the microarray domain.

Subset evaluation in microarray data analysis is typically determined via classification performance on expression data (Hauskrecht, Pelikan, Valko, and Lyons-Weiler, 2007; Mutch, Berger, Mansourian, Rytz, and Roberts, 2002; Park, Yi, Lee, and Lee, 2005). The interest in this thesis is incorporating data quality information within feature selection, so this thesis investigates how to blend supervised and unsupervised methods together.

Subset evaluation can be categorized into filter, wrapper and embedded approaches (Saeys et al., 2007).

Filter feature selection techniques

Filter feature selection techniques determine feature relevance only through assessment of the sample data without building a classifier (Guyon and Elisseeff, 2003). As this method works independently of the classifier, it suffers from an inability to exploit any valuable biases within the classifier learning algorithm (Mitchell, 1997). Filter techniques consist of two fundamental approaches: univariate and multivariate assessment (Saeys et al., 2007).

Univariate methods assume feature independence when calculating effectiveness. This makes these techniques efficient as only the same number of iterations as features are required. However the assumption of independence makes them incapable of predicting interactions and may result in features being rejected that are valuable when combined. Multivariate methods resolve this problem, but through significant additional computational expense.

From the perspective of computational efficiency, filter techniques are preferred since they provide substantial dimensionality reduction prior to classifier construction. Univariate approaches are the most ideal due to the wealth of alternatives that exist, thus permitting tailoring to specific problems. Univariate techniques are also the least expensive computationally and have been shown to provide superior performance most of the time, both in general (Guyon and Elisseeff, 2003; Blum and Langley, 1997; Dash and Liu, 1997) and specifically in microarray data analysis (Liu and Motoda, 2008; Maciejewski, 2008; Hauskrecht et al., 2007).

A major drawback with univariate techniques is their inability to identify redundancy and microarray data is known to consist of significant redundancy (Saeys et al., 2007; Yang, Xiao, and Segal, 2004; Yeung, Bumgarner, and Raftery, 2005). Some researchers see redundancy as a benefit (Guyon and Elisseeff, 2003). Guyon and Elisseeff (2003) states that redundancy can enable "noise reduction and consequently better class separation". Since the

presence of duplicate information increases the likelihood of exposing underlying patterns.

However, most researchers consider redundancy a problem (Li and Yang, 2005; Ding and Peng, 2003; Yeung and Bumgarner, 2003). For Yu and Liu (2004), Xing, Jordan, and Karp (2001) and Xiong, Fang, and Zhao (2001), two of the benefits of eliminating redundancy are improved interpretability and classification accuracy. Another benefit is a reduction in computational expense.

This thesis views redundancy as an important ingredient for dealing with noise, particularly in the context of small sample sizes, but principally through the use of data quality information. The approach used efficiently removes many features that are "potentially" redundant, via unsupervised feature selection using quality information. This eliminates features that are of insufficient *trustworthiness* prior to processing by more computationally expensive methods, for example selecting features according to their information and then managing redundancy during model construction.

Univariate methods can be divided into parametric and model-free approaches, refer table 2.1. Two parametric techniques utilize a distribution that is assumed to be responsible for the given samples. The two sample t-test and Analysis of Variance (ANOVA) (Dawson and Trapp, 2001; Walpole and Myers, 1978), "are among the most widely used techniques in microar-

Table 2.1: Filter techniques for feature selection fall into three categories.

Category		
Univariate		
	Parametric	
		t-test
		Analysis of variance (ANOVA)
		Bayesian
		Regression
		Gamma
	Model-free	
		Wilcoxon rank sum
		Between-within classes sum of squares (BSS/WSS)
		Rank products
		Random permutations
		TNoM
Multivariate		
	Bivariate	
	Correlation-based feature selection (CFS)	
	Minimum Redundancy-Maximum Relevance (MRMR)	
	Uncorrelated Shrunken Centroid (USC)	
	Markov blanket	

ray studies" (Saeys et al., 2007). However their use, in their basic form, is considered unadvisable because of their strong assumptions (Saeys et al., 2007). For this reason a number of modifications of the t-test, altering the way variance is estimated, have been tried. These modifications are necessary because of the effects of few samples and noise. Bayesian frameworks using a t-test have also been tried (Baldi and Long, 2001; Fox and Dimmic, 2006). Although Gaussian assumptions have dominated microarray analysis (Saeys et al., 2007), other parametric approaches also exist, such as regression modeling (Thomas, M., Tapscott, and Zhao, 2001) and Gamma distribution models (Newton, M., Richmond, Blattner, and Tsui, 2004).

Model free univariate approaches are used because of uncertainty regarding the "true" underlying distributions for gene expression experiments. This uncertainty is accentuated by the typically few samples that makes validation of the selected distribution difficult (Saeys et al., 2007). Consequently model-free methods make less stringent distributional assumptions about the data. Some techniques that have proved useful are Wilcoxon rank-sum test (Thomas et al., 2001), the between-within classes sum of squares (BSS/WSS) (Dudoit, Fridlyand, and Speed, 2002b) and the rank products method (Breitling, Armengaud, Amtmann, and Herzyk, 2004). Another technique is random permutations, which randomly permutes the sample data to estimate a reference distribution of the statistic, so that a model-free

version of the parametric tests can be calculated (Efron, Tibshirani, Storey, and Virginia, 2001; Pan, 2003; Park, Pagano, and Bonetti, 2001; Tusher, Tibshirani, and Chu, 2001). A benefit of model-free techniques, in the context of few samples, has been the enhanced robustness against outliers.

The final filter approach reviewed uses multivariate techniques to evaluate feature-to-feature interactions. The bivariate, evaluates pairs of features (Bo and Jonassen, 2002). More advanced methods explore higher order interactions, such as correlation-based feature selection (CFS) (Wang, Tetko, Hall, Frank, Facius, Mayer, and Mewes, 2005; Yeoh, Ross, Shurtleff, Williams, Patel, Mahfouz, Behm, Raimondi, Relling, Patel, Cheng, Campana, Wilkins, Zhou, Li, Liu, Pui, Evans, Naeve, Wong, and Downing, 2002) and Markov blanket filter methods (Gevaert, De Smet, Timmerman, Moreau, and De Moor, 2006; Mamitsuka, 2006; Xing et al., 2001). Other methods are motivated by eliminating redundancy, such as the minimum redundancy-maximum relevance (MRMR) (Ding and Peng, 2003) and uncorrelated shrunken centroid (USC) (Yeung and Bumgarner, 2003) techniques.

Wrapper feature selection techniques

Wrapper feature selection techniques evaluate feature subsets according to their performance in the constructed model. The definition of "performance" is at the user's discretion but two examples are overall percentage accuracy

and false positive minimization. Wrapper techniques intrinsically exploit learning algorithm biases but with a loss of generality, since subset generation and evaluation is tailored to the model used (Saeys et al., 2007). Two styles of wrapper technique exist from the subset generation perspective: heuristic and non-deterministic.

Although heuristic guided search is simple, it risks overfitting and becoming stuck within a local optima. A major influence for overfitting is the exploitation of model bias. Due to the multimodal nature of the landscape within the problem space and a dependence on non-exhaustive search, heuristic guided search risks failing to locate the global optimum. Non-deterministic search suffers from increased risk of overfitting and increased computational expense, although it is more likely to find the global optimum (Saeys et al., 2007; Hua, Tembe, and Dougherty, 2009). However the additional computational expense can prove to be unmanageable. The increased risk of overfitting is partly due to chance, as randomization can result in artifact discovery rather than generalizing the behavior of the phenomenon. Examples of microarray wrapper technique implementations are: sequential search, genetic algorithms and estimation of distribution algorithms. As the name suggests, sequential search systematically searches the entire space (Inza, Larranaga, Blanco, and Cerrolaza, 2004; Xiong et al., 2001). The genetic algorithm and estimation of distribution techniques ran-

domly search the entire space using a heuristic (Jirapech-Umpai and Aitken, 2005; Li, Weinberg, Darden, and Pedersen, 2001; Ooi and Tan, 2003; Blanco, Larranaga, Inza, and Sierra, 2004)

Embedded feature selection techniques

The embedded feature selection technique takes the wrapper approach further by exploiting aspects of classifier construction in order to guide the generation of the next subset. An example of this is using a feature's weight to determine its discriminative value, where the magnitude of a weight is proportional to its estimated relevance. Embedded techniques are computationally less expensive than wrapper techniques, but they suffer from increased dependence on the classifier. This may be significant as flexibility is lost, since a more ideal bias offered by another classifier may not be available.

Examples of embedded feature selection techniques for microarray data are random forest, weight vector of support vector machines (SVM) (Burges, 1998; Schlkopf and Smola, 2002) and weights of logistic regression. The random forest classifier uses a committee of single decision trees to calculate the importance of each feature (Diaz-Uriarte and Alvarez de Andres, 2006; Jiang, Deng, Chen, Tao, Sha, Chen, Tsai, and Zhang, 2004). The weight vector of SVM uses the weights of each feature within linear classifiers like SVMs, to select features (Guyon, Weston, Barnhill, and Vapnik, 2002). Finally the

weights of logistic regression technique, uses the calculated feature weights in the logistic regression algorithm to select features (Ma and Huang, 2005).

Wrapper and embedded techniques both suffer from dependence on a classifier, overfitting, risk of failure, risk of selecting a local optimum and substantial computational complexity. These are accentuated in problems with small sample sizes and high dimensionality (Jain and Zongker, 1997; Dougherty, 2001; Sima and Dougherty, 2006; Tu et al., 2002).

As the techniques above depend entirely upon expression data—only training and validation data can be used during feature selection and model construction. Given the limited amount of sample data suitable for supervised learning, unsupervised learning may allow inclusion of test and unlabeled data. One option considered in this thesis is an unsupervised phase followed by a supervised phase during feature selection. The unsupervised phase uses the entire (training, validation, test and unlabeled) data set to rank the trustworthiness of individual features. A subset of the most trustworthy features can then be assessed by the supervised phase using traditional (training and validation) data.

Feature selection stopping criterion

A stopping criterion is needed in order to set limits for iterations of feature selection, particularly if a non-exhaustive search strategy is used (Dash and

Liu, 1997). A stopping criterion provides a mechanism for limiting resource consumption or providing a tradeoff between resources and the search for an optimal feature subset. Stopping criteria take many different forms, such as incremental improvement in subset optimality and a fixed number of features. Stopping criteria are not the focal point of this thesis, rather, selecting the features that are not just discriminative, but also trustworthy.

Feature selection result validation

Validation of a selected feature subset involves testing its effectiveness against the test set (Dash and Liu, 1997). The goal is to evaluate the generality of the selected subset by using data known as the "validation set". In order to evaluate generality, the validation data cannot be used during feature selection.

However the disadvantage of this validation process is that the quality of the validation and test set are not considered by this "traditional" feature selection process. As a result—on the basis of data quality—the selected features may perform poorly when applied to any data set other than the training set.

Feature selection and microarray data analysis

Feature selection in general (Guyon and Elisseeff, 2003; Jain and Zongker, 1997; Langley, 1994) and for microarray data analysis (Yu, 2008; Saeys et al.,

2007; Hauskrecht et al., 2007) depends on discriminating between classes of interest. Typically, the greater the change the better. For example an SVM (Furey, Cristianini, Duffy, Bednarski, Schummer, and Haussler, 2000; Brown, Grundy, Lin, Cristianini, Sugnet, Ares, and Haussler, 1999) searches for a maximum margin between the hyperplane, or decision boundary separating the classes (Schlkopf and Smola, 2002; Burges, 1998). In the case of a linear kernel, SVM learning prefers features that are proportional to larger class separation. So features that provide the greatest margin against classification error are implicitly preferred.

It is well known that executing feature selection on different subsets of the available data set often results in differences in the selected feature sets, particularly in the case of microarray data analysis (Jeffery, Higgins, and Culhane, 2006; Ein-Dor et al., 2005). It is not unusual for differences in the selected features to be significant, even though the only experimental variable is the data subset used (Sima and Dougherty, 2006; Kalousis, Prados, and Hilario, 2007; Quackenbush, 2001) and it is generally accepted that noise is responsible for such differences (Jeffery et al., 2006; Ein-Dor et al., 2005; Wang et al., 2001).

Although traditional feature selection techniques implicitly seek features that provide the greatest margin against error, it is clear that their approach is inadequate for microarray and similar data sets. The impact of noise

in a non-classical problem, requires an approach to feature selection that explicitly considers the impact of data quality. A possible variation on this approach is the inclusion of data quality information for the *entire* data set, therefore including the test set and unlabeled data.

In the spirit of work by Wang et al. (2001), this thesis investigates an approach for estimating data quality through using data "other" than just signal (or for example expression) data. The results of Wang et al. show that doing so provides a better estimate of data quality. An example of "other" data is spot shape, which determines whether signals, which are unrelated to the spot itself, are incorporated within the estimation of the spot's intensity.

Determining a measure of data quality involves the use of prior knowledge about the data itself, for example, that the shape of a spot influences the accuracy of the spot's measured intensity. Sima and Dougherty (2006) are proponents of the need for incorporating prior knowledge into the analysis of microarray data,

> [Our] conclusions lead to the further conclusion that we require feature-selection techniques that are not purely data driven. The search for features needs to be constrained and directed by the use of prior biological knowledge, ... Moreover, confidence in the correctness of prior assumptions should be integrated into the

design of techniques,

As suggested by Sima and Dougherty, a solution to the problem of few samples and noise is the use of prior knowledge. They focused on biological knowledge, while this thesis looks into using information that is indicative of the quality or accuracy of the expression data.

2.3.4 Model learning

Learning is "the phenomenon of knowledge acquisition in the absence of explicit programming" (Valiant, 1984) and this acquisition can be thought of as the process of building assertions, which in turn, can be characterized as "strongly" or "weakly" implied by the observations used (Micalski, 1983). In this thesis, observations are "sample data" or more precisely "training data". The set of assertions built are a model or hypothesis about the phenomenon of interest. Some assertions are "strongly" implied by the training data, while others are more "weakly" implied.

Learning is impacted by many factors (Vapnik, 2000; Duda, Hart, and Stork, 2001; Haykin, 1999), such as: sample size; complexity of the phenomenon; how representative the sample data is and the choice of learning algorithm, which determines what can be learnt and how quickly (Mitchell, 1997). Because of such complexity the field of computational learning theory has sought to develop generic theory in order to answer as many of these ques-

tions as possible (Anthony and Biggs, 1992; Mitchell, 1997; Haykin, 1999).

Probably approximately correct learning (PAC)

Valiant (1984) investigated learning in order to "shed light on the limits of what can be learned". But due to the complexity of the problem, he used a probabilistic setting to simplify the learning problem.

Valiant viewed learning as a search problem through all the hypotheses that could be learnt. A subset of the hypotheses provide an approximately correct model of the phenomenon being learnt, while a hypothesis might also exist that is an exact representation of the phenomenon. The remaining hypotheses are insufficiently accurate to be useful. Therefore the problem considered is estimating the sample size required to select, by a predetermined level of likelihood, a hypothesis that is at least sufficiently accurate—or approximately correct.

Valiant limited the learnable concepts to Boolean functions of a set of propositional variables, where a concept is modeled by a set of hypotheses. He restricted his research to noiseless sample data and assumed an arbitrary probabilistic distribution for searching the hypothesis space.

Valiant investigated the computational effort required for learning, since a search space might contain a very large number of candidate hypotheses and each hypothesis might require substantial computational effort to validate.

He showed that a concept can be learnt in a reasonable polynomial number of steps, although inherent algorithmic complexity appeared to set serious limits on the range of learnable concepts.

Valiant's work is attributed as the start of Probably Approximately Correct (PAC) learning theory, which provides a framework for estimating how much sample data is required to "probably" learn an "approximately correct" hypothesis or model. "Probably" refers to being able to construct a model with a correct concept or correctly learn with sufficient confidence (Haussler, 1995; Mitchell, 1997). In other words, "probably" refers to the likelihood of selecting an "approximately correct" model. The selected model must also be at least "approximately correct" for the range of all possible inputs. PAC theory provides a "bounds" on the minimum number of samples and computational effort required to learn a PAC "compliant" model (Haussler, 1995; Mitchell, 1997).

Valiant (1985) incorporated noise into his earlier work (Valiant, 1984) and showed that low error rates or noise levels can still allow a PAC compliant model to be learnt. Kearns and Li (1987) independently confirmed these findings and stated that a theoretical noise level barrier was 1/2, as "the errors in the ... process destroy all possible information".

Quinlan (1986a) was the first to study the impact of noise on real data using a context of building decision trees. He found that the effect of noise

was proportional to the information content provided by a feature. Thus the higher the information content provided, the more detrimental noise was. The experimental results suggest a complex and unpredictable outcome as noise increased. It was also shown that noise *always* degrades classification performance. Therefore the best strategy is to avoid noise and this thesis seeks to construct a methodology that can identify noise and avoid features that are associated with noise.

Quinlan showed that a model trained with noiseless data is inferior to a model trained with noisy data, when noisy unlabeled data is classified. This implies that data used to train a model must be of similar quality, or less quality, than the data to be classified.

Quinlan concluded with five recommendations, of which four relate to this thesis:

1. "It is important to eliminate noise affecting the class membership of the objects in the training data".

2. "It is not worthwhile expending effort to eliminate noise from the attribute values in the training set if there is going to be a significant amount of noise when the induced classification rule is used in practice."

3. "We are better off dispensing altogether with noisy, less important

attributes."

4. "The payoff in noise reduction increases with the importance of the attribute."

Quinlan's first recommendation is unsurprising and typically employed within data mining. However the second recommendation appears to be overlooked by the literature. For example, we are unaware of literature where the quality of test data is balanced with that of the training data. It is concluded that the above recommendations can only be exploited by assessing how much noise is present in *all* the data sets. For example, the quality of data to be classified must be comparable to that used to construct the model. One of the two feature selection methodologies proposed by this thesis explicitly considers noise levels across all the available data: training, validation, test and unlabeled. Regarding Quinlan's third recommendation, both of the proposed methodologies, described in this thesis, exclude noisy features. One of the proposed methodologies seeks to balance the measured noise level of each feature against the degree of information provided by the feature, therefore not using features whose information content is likely to be detrimentally affected by noise.

Kearns and Li (1987) saw a number of inherent strengths in PAC (Valiant, 1984), two of which are: "lack of assumptions on the probability distribu-

tion", which defines hypothesis selection and its inherent simplicity, or generality. However, according to Kearns and Li, a difficulty exists in justifying the use of error sources with a "nice form".

Kearns and Li (1987) investigated how PAC could be altered to manage, as he termed it, "malicious errors", or errors that are "generated by an adversary whose goal is to foil the learning algorithm". Although some success was had, it was found that developing sample size bounds for malicious errors was generally much easier that the ensuing computational learning expense. In the case of the "most interesting representation classes", the calculated "bound could always be achieved via a super-polynomial time exhaustive search learning algorithm". It was concluded that malicious errors could be bounded or learnt but learning was approximately NP-hard. This degree of learning difficulty is considered evidence for incorporating a measure of data quality within the feature selection process. Therefore the benefit of avoiding noisy, or at least excessively noisy features, could lead to a significant reduction in the computation effort required for feature selection and model learning. Given the presence of noise within microarray data and the associated issues of a non-classical problem domain, PAC provides an effective framework for estimating the worst case learning scenario.

If was also found that a noise model was needed in order to tolerate malicious errors and that positive and negative examples were needed for

efficiency reasons (Kearns and Li, 1987). Clearly learning is difficult in the presence of errors, particularly if the errors are of a malicious nature and that prior knowledge is essential for success. These findings suggest that learning a reliable model for microarray data is likely to be difficult. Worse still, microarray data is typically unbalanced as far more examples exist for one class than the other. However the outcomes of Kearns and Li suggests that prior knowledge offers a potential solution. A noise model is one example of prior knowledge. Another is a contribution of this thesis, that prior knowledge through measuring data quality can enable more effective feature selection and learning.

The literature is generally focused on large sample sizes, however the non-classical scenario, such as microarray data analysis is restricted to few samples and very high dimensionality. Angluin and Laird (1988) investigated how learning might cope in such a scenario and concluded—if a classifier can make random errors less than half the time and there is a "feasibly small number of examples" a "strategy of selecting the most consistent rule for the sample is sufficient". However they concluded that a noise model is essential, since one cannot assume an accurate model can be learnt *without understanding and managing noise.*

Ignoring noise for a moment, the curse of dimensionality also results in inaccuracies within the model being generated. For example, data sparseness

causes significant variance in the values being estimated and a bias exists with respect to the influence of individual samples, if localized learning methods are employed (Hastie et al., 2001, section(2.5)). But the presence of noise further exacerbates these inaccuracies and raises the importance of managing noise or, more generally, issues regarding data quality. In view of this, a measure of data quality would ideally consider anything that impacts the resulting accuracy of a model, for example, the sparseness of the sample data, or given the sample data, weaknesses in the learning technique.

Angluin and Laird (1988) also concluded that their basic ideas also extend to other types of random noise sources, other than Valiant's (Valiant, 1984) uniform distribution. However the ensuing "search problem associated with that strategy is intractable in general". But for particular classes of rules it may be possible to efficiently identify the target rule provided techniques specific to that class are used. Therefore one cannot in general expect to achieve a PAC compliant model without knowledge about the problem itself. Achieving PAC-identification requires finding a hypothesis, or model, that satisfies PAC requirements. For example a model for data errors or knowledge about error rates is required.

Predictability of error rates or possession of an accurate error model is more important for some problem domains than others. For some domains the error rate is relatively static, for instance using a specific measuring

instrument is likely to generate consistent errors. However for microarray analysis, the error rate is dependent on many factors. For example, individual arrays are in effect an individual measuring instrument and some arrays are additionally affected by environmental influences during their construction, such as dust.

PAC measures two aspects of learning: the required sample size to achieve PAC-identification and the computational effort needed to select a hypothesis (Haussler, 1990b). The sample size required is seen as a measure of "sample complexity", or the estimated minimum amount of data required to uncover the complexity of the concept to be learnt. Formally sample complexity is the smallest polynomial of the form $p(n, 1/\epsilon, 1/\delta)$, where n is the size of the instance space, ϵ is the probability of classification error for the required hypothesis and $1 - \delta$ is the probability of selecting that hypothesis.

Although PAC theory is algorithm independent, knowledge of the type of concept being learnt is needed to quantify the number of hypotheses within the search space and the computational search effort. For this reason there are many variants of the basic definition of PAC learnability (Haussler, 1990a), one being the introduction of syntactic complexity of the target concepts. This measures the number of symbols in the shortest description according to a concept description language. Other variants deal with differing levels of example errors. They have generally dealt with the com-

putational difficulties of learning, since many concept classes had previously proved intractable. However because of noise, all of these approaches depend on sufficient samples in order to achieve the emergence of the underlying pattern.

Kearns (1998) reviewed an extension of Valiant's (Valiant, 1984) learning model, which incorporates noise. The extension, devised by Angluin and Laird (1988), uses the simplest type of white noise to enable an algorithm to tolerate the highest possible noise rate. One algorithm for learning boolean conjunctions approached the information-theoretic barrier of 1/2. However Kearns concluded that little had been achieved in characterizing which classes could be efficiently learned and no general approaches had been devised for noise-tolerant learning.

As a result of the review by Kearns, he devised a "natural restriction on Valiant model algorithms that allows them to be reliably and efficiently simulated in the presence of arbitrarily large rates of classification noise" was devised (Kearns, 1998). Using the original noise-free algorithms for Valiant's model as a test base, efficient noise-tolerant algorithms were developed for almost every concept class. However a condition for this approach was the learning algorithm knowing an upper bound for the "noise rate". Therefore prior knowledge in the form of a noise model has become *indispensable*. It was considered desirable that this condition be removed or at least relaxed. But

relaxing this condition "severely limits the cases for which efficient learning is possible, and results in a perhaps overly pessimistic noise model (Sloan, 1988; Kearns and Li, 1987), unless the dependence of the noise on the input has natural structure that can be exploited by the learner" (Kearns and Schapire, 1990).

PAC theory has evolved sufficiently to cater for concept classes which once proved intractable. However the cost of advancement has, in effect, been safer bounds for sample size and computational estimates. Although these bounds are likely to guarantee success, the bounds are likely to be too pessimistic for many applications (Buntine, 1990; Sarrett and Pazzani, 1992).

The review shows that noise is detrimental to the construction and use of a model and that sufficient noise will prevent success. However, it was also shown that the use of prior knowledge may provide a viable approach to effectively dealing with noise. The key points on which the subsequent work of the thesis is based are:

- A point exists where the level of noise exceeds the benefit of a feature
- Existence of different noise levels between training and test data adds to the uncertainty of a classification
- For a set of redundant features, the feature that is least affected by noise is not only desirable, but potentially essential to success

- The use of prior knowledge has proven to be a significant benefit in handling noise

This research investigates a noise model that provides prior knowledge regarding the level of noise present in each feature. We develop a feature selection methodology that utilizes the estimated noise level and avoids features that are considered unlikely to benefit a model.

2.4 Approaches for handling noise and managing data quality

A general approach for managing noise is smoothing, which consists of three methods, binning, regression and clustering (Han and Kamber, 2006). Binning involves using information gained from the neighborhood of the instance being processed. Smoothing during regression involves a trading-off between training error and the preference for a simpler model. Note that smoothing is essentially the same as regularization, which is a popular approach referenced later in this section. Clustering provides an effective method for outlier detection and inherently offers information for repairing outliers. However the above methods require substantial sample data, for reasons similar to those presented in section 2.3.4, in order to uncover the underlying structure of the data when it is shrouded by noise.

Many approaches exist for cleaning data, but all involve making noisy

data more like the data that is considered to be free from noise, but this may not result in something that is necessarily correct (Han and Kamber, 2006). Using knowledge about the data is a recommended alternative for effective data repair (Han and Kamber, 2006; Hastie et al., 2001). In the spirit of this approach, we investigate using knowledge about the sources of noise which affect the data and consequently attribute a measure of quality or trustworthiness to the feature associated with the data. This measure of trustworthiness then provides input for feature selection to control the influence of the data on the basis of its quality.

In the context of non-classical problems, the curse of dimensionality, or the $p >> n$ problem, results in issues with model variance and overfitting when traditional methods are used (Hastie et al., 2001). As a result Hastie et al. (2001) concluded there is a need for new methods which use substantial regularization to produce simple models. They also asserted that the importance of feature selection increases proportionally with the dimensionality of the problem. Quinlan (1986b) also acknowledges the need for new methods when there is a substantial difference in the quality of the data present in the training and test sets. The importance of handling differences in quality increases as sample size decreases, since shrinking sample sizes is associated with greater uncertainty about the underlying structure of the data. Regularization depends on constructing a simple model that is expected to reflect

the underlying structure of the data, rather than attempting to build a more complex model using the training set and expecting it to perform similarly with test data (Hastie et al., 2001). Our approach manages noise by acknowledging the uncertainty associated with few noisy samples and then reducing uncertainty by using the most trustworthy feature from a set of redundant features.

2.5 Research gap

Traditional data mining and in particular feature selection, are ideally suited to problems that are characterized by many samples and few dimensions. Such problems, which this thesis referred to as being classical, generally provide enough samples to accurately determine the underlying signal. Even the presence of noise can often be overcome due to an abundance of samples. The existence of few features tends to also assist this situation by limiting the dimensionality of the search space, and thereby reducing the likelihood of model overfitting. This thesis considered a new type of problem that is referred to as non-classical, which is characterized by few samples, many dimensions, substantial redundancy and significant noise.

The non-classical characteristic of having few samples has resulted in considerable attention in the literature, since there is a concern that many problems may be un-learnable because of the interaction between the sam-

ples and noise (Dougherty, 2001). For example, considerable effort has been invested in trying to formulate, *a-priori*, how many samples are required for successfully learning individual problems (Sima and Dougherty, 2006).

Managing the non-classical characteristic of many features has also received considerable attention, since it results in the curse of dimensionality, the presence of substantial redundancy and an associated computational explosion. One of the known problems of many features is the variability of the selected feature set across data folds (Ein-Dor et al., 2005). The literature suggests a number of contributing factors for this variability (Tu et al., 2002), for example, the presence of noise within the data set and the enormous number of options provided by the available features. High dimensionality also increases the likelihood of overfitting.

Non-classical problems can also be characterized by substantial noise, which can be caused by the technology responsible for generating vast amounts of data, for example microarrays. The presence of substantial noise is of particular concern since every available instance of data is valuable. Given the existence of few samples and significant noise, many of the approaches proposed in the literature involve the use of prior knowledge. One approach involves the using a noise model (Raser and O'Shea, 2005), while another exploits a variability characteristic of the data itself (Baldi and Long, 2001). Sima and Dougherty (Sima and Dougherty, 2006) are strong proponents for

using prior knowledge in order to overcome the difficulties typically associated with non-classical problems. Another problem associated with noise was raised by Quinlan (Quinlan, 1986b), who proposed the need for a method to explicitly overcome differences in data quality between the training and the test sets.

The literature confirms that classical feature selection is focused on selecting features according to the signal within the data set and this could result in the problems identified above, such as overfitting and the variability of the selected features across data folds. It is also known that data pre-processing operates independently of feature selection and model building, hence the sole product of pre-processing used by feature selection is the adjusted data set. However feature selection and model building are not provided with any other information, such as which data instances were considered to be outliers.

The thesis addresses the identified research gap by developing an approach for evaluating the quality of individual features. This evaluation, which depends on a univariate approach, enables feature selection to eliminate features that are associated with poor quality data. Using a univariate approach helps reduce the computational expense of feature selection, particularly since this approach is executed prior to traditional feature selection. This elimination of poor quality data assists feature selection by presenting

data to the traditional feature selection step that is less affected by noise and therefore assists in the selection of a feature set that varies less across data folds. This also allows traditional feature selection to exploit feature redundancy, since redundant alternatives associated with lower quality are eliminated prior to this evaluation.

Two evaluation methods are proposed in this thesis: using a measure of signal-to-noise and a noise-only measure of quality. The second approach enables feature selection to also consider the quality of unlabeled samples, which addresses Quinlan's recommendations. The thesis also provides a method for exploiting quality information that is implicitly obtained during pre-processing.

2.6 Discussion

The goal of this review was to assess the effectiveness of current practices when the problem domain is characterized by: few samples, many dimensions and significant noise. and to determine how noise might be more effectively managed.

Microarray data analysis is an example of a domain that is seriously impacted by too few data samples. One impact of this limitation is overfitting—particularly if the dimensionality exceeds the number of samples (Maciejewski, 2008). Consequently, having few samples demands utilizing data to its

best advantage and for this reason, this research also investigated using test and unlabeled data during feature selection. In order to use test and unlabeled data we develop an unsupervised phase, which learns about the quality of the data or noise level present within *all* available data. This eliminates the problem of tainting supervised learning with knowledge about the test and unlabeled data.

Data with many dimensions necessitates feature selection in order to reduce model complexity, prevent overfitting and improve model performance. As high dimensionality is likely to lead to an intractable learning problem we use a filter-based feature selection approach. We also restrict our approach to univariate feature selection to prevent an intractable search problem and to maintain redundant features for later use. Although the presence of feature redundancy is necessary, it will not be explicitly considered while evaluating the quality of feature data or the information provided by a feature. Our review of molecular biology and microarray data has shown that large numbers of redundant features, or genes, typically exist. This redundancy provides an opportunity for selecting a redundant feature that is least affected by noise. This approach to feature selection is particularly important given that significant noise exists and that this noise varies from feature to feature.

PAC theory showed that noise should be avoided as far as possible. For

that reason we seek to avoid features that are known to be affected by noise. Just as the training and test data sets should be equally representative of each other from a classification perspective, PAC showed that the same also applies regarding noise. However traditional data mining provides no mechanism for achieving evenly distributed noise within every data set. As as result, we also investigate the benefit of utilizing test and unlabeled data within the unsupervised feature selection phase, therefore providing a mechanism for achieving noise uniformity across all data sets. PAC theory shows that prior knowledge is necessary to enable learning when noise is significant. Specifically, a noise model is often used within PAC as a mechanism to utilize prior knowledge about the noise. We also investigate using a noise model that permits noise level assessment for individual features.

Our approach seeks to reduce dimensionality by eliminating redundant features that are described by noisy sample data. By doing so, relevant features that are described by accurate sample data become candidates for class correlation analysis. This approach explicitly reduces the effects of noise by preferring features that correspond to accurate data.

By assessing the degree of noise present within sample data, through use of non-class correlated data, provides another potential benefit—the quality of test and unlabeled data can also be used to assess a feature's trustworthiness.

Chapter 3

Theoretical framework

This chapter presents the motivation and approach for incorporating a measure of the data's quality within the feature selection process. An overall measure of the quality for a feature's data is called the feature's "trustworthiness". Trustworthiness ultimately provides a ranking for a feature with respect to the quality of the data used to define the features behavior; the greater the quality of the feature's data, the higher the feature's trustworthiness.

Unlike classical feature selection, trustworthiness supports a mechanism that discriminates between a feature which provides actual information content and one that just appears to do so. This "apparent" information content only seems to provide information due to the effects of noise, hence providing a chance correlation with what is being learnt. The expression "information content" refers to the degree of correlation between the feature and what is to be learnt. For example, various gene expression variables and a known

classification for a tissue sample.

There are a number of sources of data that can be used to calculate a feature's trustworthiness and these are referred to as "confidence" data. Confidence data provides a measure of the quality of a feature's data and itself need not be correlated with the information content of the feature. One of many different sources of confidence data can be used and alternatively, one or more sources of confidence data can be combined. Note that the actual confidence data used depends on the problem in question. The following are examples of confidence data that could provide a measure of the quality of the feature's data: the age of the feature's data, the impact of weather, characteristics of the equipment used to generate the data, the experience of the equipment operator and the degree of change applied to feature data as a result of cleaning it. Although not all the presented examples may apply to microarray data, the theoretical framework is applicable whenever a data quality problem exists.

This chapter also describes the mechanism that trustworthiness supports, Feature Utility Ranking (FUR). "Feature utility ranking" provides a ranking for each feature using its trustworthiness, as well as information content, to perform Feature Selection (FS).

Traditional filter based FS is univariate, since it only considers a feature's information content, while FUR is bivariate since it also considers the qual-

ity of the features data. The effectiveness of FUR is related to the presence of redundant features. If two redundant features exist, the more trustworthy feature would be preferred. Therefore FUR is best suited to problems consisting of a large number of features and where significant redundancy is present.

Some justification for the value of trustworthiness and FUR is provided next. The approach begins with a theoretical and then a proof of principle for the impact of noise on successfully inferring the original message. In the interim of a formal definition of data quality, *data quality* and *noise* will be used interchangeable, since noise is responsible for a lack of data quality.

3.1 Determining the original message

"Noise" can be viewed as "any unintended signal perturbation" (Carlson, 1975, p.4) of the signal of interest and therefore obscuring its "true" nature. The problem of understanding what is being spoken within a noisy environment, such as a Cocktail Party, provides a familiar account of the impact of noise. The impact of noise can, intuitively, vary from nothing through to preventing every word from being heard correctly. Cherry conducted research into the problem of understanding what is spoken in a noisy environment; Cherry's research is commonly known as the "Cocktail Party Effect" (Cherry, 1953).

The Cocktail Party Effect can be viewed as the challenge of understanding individual words, but such an approach is naive since many words are likely to be lost because of noise. Cherry viewed this problem as one of seeking to understand the original intent of the speaker. Given that many words could be lost, the hearer seeks to infer the original intent using what is heard and other information sources such as facial expressions or hand gestures. This inference process consists of incrementally altering the understood intent in response to new information provided by a fragment given by the speaker. A fragment may consist of one or more words or facial expressions and hand gestures.

In the case of the hearer receiving a fragment that conflicts significantly with the currently understood intent, the hearer must decide whether a dramatic alteration is appropriate. The decision process may, among other things, involve a measure of "confidence" in the fragment that triggered the potential alteration. In the case of this research a fragment corresponds to an instance of data, while the entire "message" corresponds to the sample set. To not judge a fragment's confidence is to implicitly assume equal confidence for every fragment, which is the approach often used in FS methods; see section 2.6.

Consider that the message, in general, is fragmented across each instance within the sample set. Rationally, an instance that is significantly affected

by noise should have less influence in the process of inferring the message, while an instance that is unlikely to be affected by noise would have greater influence. As such, an instance that is significantly affected by noise should be associated with a low measure of confidence. One of the outcomes of this research is an approach for calculating an instance's confidence, which is synonymous to estimating the amount of noise present. Note that the approach used will seek to calculate a worst case value for confidence—-since the selection of the most trusted data is sought.

3.2 Theoretical impact of noise on sample size

This section conducts a quantitative assessment of the impact that noise has on model learning. This quantitative assessment is not part of the proposed feature selection methodology, but is provided as a justification for its development. The approach involves the comparison of noiseless and a noisy learning scenarios, which are otherwise identical. The noiseless scenario involves estimating the amount of noiseless sample data (N) that is required for constructing a model of predetermined accuracy (ϵ), while the noisy scenario estimates the required sample size ($\bar{N}$) to construct a model to the same level of accuracy.

Probably Approximately Correct (PAC) theory (Valiant, 1984; Haussler, 1995; Bshouty et al., 1999), originating from "Computational learning the-

ory" (Anthony and Biggs, 1992), is used for estimating sample requirements. Probably Approximately Correct theory addresses the question of how much sample data is required to learn the underlying structure held within the data (see section 2.3.4). According to the theory, the quantity of data required or "sample complexity", defines "how many training examples are needed for a learner to converge (with high probability) to a successful hypothesis" (Mitchell, 1997). Using sample complexity and a stochastic framework (Valiant, 1984; Haussler, 1995), PAC theory provides a minimum bound on sample data requirements for successful learning (Mitchell, 1997; Haykin, 1999). This minimum bound defines the smallest sample size required to achieve a hypothesis of predetermined accuracy ϵ. This accuracy could however, by chance, be achieved with fewer than the estimated samples.

3.2.1 Noiseless sample data

In the context of noiseless sample data, PAC theory states (Angluin and Laird, 1988, p. 346) that

$$N = \frac{1}{\epsilon} ln \frac{M}{\delta} \tag{3.1}$$

where N is sample size required to learn a probably approximately correct hypothesis (or PAC model); M is the number of possible hypotheses, hence a measure of the model's complexity; ϵ is the hypothesis error, or difference between the selected and the "true" hypothesis; and δ is the likelihood of

choosing the selected hypothesis. Note that ϵ and δ are in the range $(0, 1]$.

Equation (3.1) provides a *worst-case* estimate of the number of training samples (N) required, regardless of the learning algorithm or the underlying distribution of the sample population, for a given ϵ and δ (Angluin and Laird, 1988; Mitchell, 1997; Haykin, 1999).

3.2.2 Noisy sample data

In the context of noisy sample data, PAC states that an information-theoretic barrier of one-half exists (Kearns and Li, 1987; Angluin and Laird, 1988). This barrier is with respect to class information error during learning. If noise causes class information to be incorrect for 50% or more of training samples—learning is impossible.

Using a simple model of noise, which introduces random sampling errors regarding class information, Angluin and Laird (1988, p. 351), derived

$$\bar{N} = \frac{2}{\epsilon^2(1-2\eta)^2} \ln \frac{2M}{\delta} \tag{3.2}$$

where ϵ and δ are the same as for equation (3.1) and η is the sampling error rate.

3.2.3 Impact of noise

Using equations (3.1) and (3.2), we derive the function $\bar{N} = f(N, \epsilon, \delta, \eta)$, which relates the impact of noise (η) on sample size requirements ($\bar{N}$), com-

pared to a noiseless sample size (N). Therefore $f(\cdot)$ shows the increase in sample size requirements in order to maintain a PAC model when noise is added.

Using the axiom

$$\log xy \equiv \log x + \log y$$

(Grossman, 1977, p.286) and simple rearrangement of equation (3.2) gives

$$\bar{N} = \frac{2}{\epsilon(1-2\eta)^2}\left(\frac{1}{\epsilon}\ln\frac{M}{\delta} + \frac{1}{\epsilon}\ln 2\right)$$

Given that the first additive term in brackets is equivalent to N we now have

$$\bar{N} = \frac{2}{\epsilon(1-2\eta)^2}\left(N + \frac{1}{\epsilon}\ln 2\right) \tag{3.3}$$

which states the noisy sample size as a function of N, ϵ and η. Note that δ, which is the likelihood of choosing the selected hypothesis, is irrelevant. This equation also shows the information-theoretic barrier for noise in the component $(1-2\eta)$, which approaches zero as η approaches 0.5.

The $2/\epsilon(1-2\eta)^2$ portion of equation (3.3) predominately determines the increase in sample requirements due to noise. This portion is composed of three parts: the multiplier 2; the *hypothesis error multiplier*, or $1/\epsilon$ and the *noise multiplier*, or $1/(1-2\eta)^2$. An ϵ of 0.2, for example, results in a hypothesis error multiplier of 5, while an η of 0.2, for example, results in a noise multiplier of 2.8. An ϵ of 0.2 defines a hypothesis that provides one classification error in five to be acceptable, while any greater precision increases

sample size requirements dramatically. An η of 0.2 similarly declares one in five sample points to be in error due to noise. Combining all three into a single multiplier results in 28, which is a significant increase in sample size because of noise. Using these multipliers and equation (3.3) gives, for this example,

$$\bar{N} \approx 28 \times (N + 3) \tag{3.4}$$

The magnitude of the noise multiplier $1/(1 - 2\eta)^2$ is shown for various sampling error rates (η) in figure 3.1. The noise multiplier itself can become significant when $\eta \geq 0.1$ and sample size approaches the minimum bound set by the equation. A multiplier of 2 could result in insufficient data, which is likely in the case of microarray data analysis.

Although equation (3.4) is linear, the addition of noise has dramatically increased sample size requirements. For instance, if 50 noise free samples are required for successful learning, a sample size of 1,484 is required if 20% of the samples are incorrect because of noise. For a problem domain with limited sample data, for example a microarray data set with 100 sample points, the impact of noise is dramatic and may result in an inability to learn an acceptably accurate model.

The values for ϵ and η in the above example are considered conservative for an environment like microarray data analysis, which is known to contain

considerable noise and limited sample sizes (Raser and O'Shea, 2005; Baldi and Hatfield, 2002). Figure 3.1 shows various *noise multipliers* as a function of η and a constant ϵ of 0.2. Although the noise multipliers for low values of η are small, they could result in insufficient data when noise is added to a sample that is close to the minimum bound defined by equation (3.3), while much greater noise multipliers can occur as noise approaches the information-theoretic barrier.

Clearly data that is substantially impacted by noise is best avoided, while in cases of limited data, avoiding noisy data may become essential to success. Trustworthiness provides a mechanism for avoiding noisy data while still selecting features without compromising on information content.

3.3 Proof of principle for impact of noise

Using a real-word data set, a proof for the theoretical impact of noise on sample size is given.

3.3.1 Experimental design

A subset of the 1994 American Census data was used to construct a classifier that predicts whether an adult will earn more than $50,000 dollars per annum. This census data came from the UCI Machine Learning Repository[1] and has been used by numerous researchers since the first cited work by

[1] http://archive.ics.uci.edu/ml/datasets/Census+Income

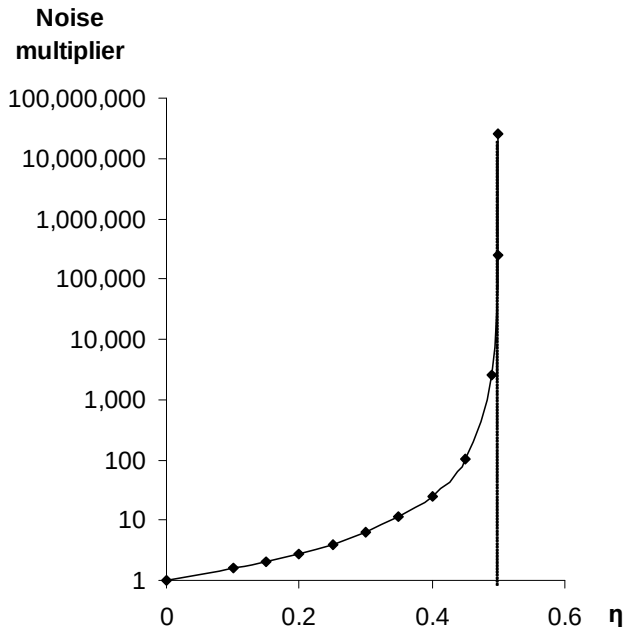


Figure 3.1: The magnitude of the noise multiplier is shown for various sampling error rates (η), which depicts the impact of noise on sample size.

Kohavi in 1996.

The original census data consists of 48,842 instances with fourteen attributes; six continuous and eight nominal. Removing instances with missing attributes yielded a final data set of 32,561 instances. The experimental goal was to measure the impact of noise on building a classifier that predicts whether an adult will earn more than $50,000 and relate these results to equation (3.3) where the noise used corresponds to the variable η.

The cleaned census data was labeled as the "noiseless data set". Five other data sets were constructed from the noiseless data set where each was given a noise level of 10%, 20%, 30%, 40% and 50%. Following PAC methodology, noise was *only* applied to the class value of a data instance by toggling its current state. The noise level determined the randomly selected percentage of instances whose class value was toggled.

Each experiment consisted of building a Naive Bayes Classifier using the same default parameters and ten-fold cross validation. The implementation of Naive Bayes was provided by the Weka machine learning environment (Witten and Frank, 2005; John and Langley, 1995). Each experiment consisted of using 46 different sample subsets, ranging in size from 100 to 30,000.

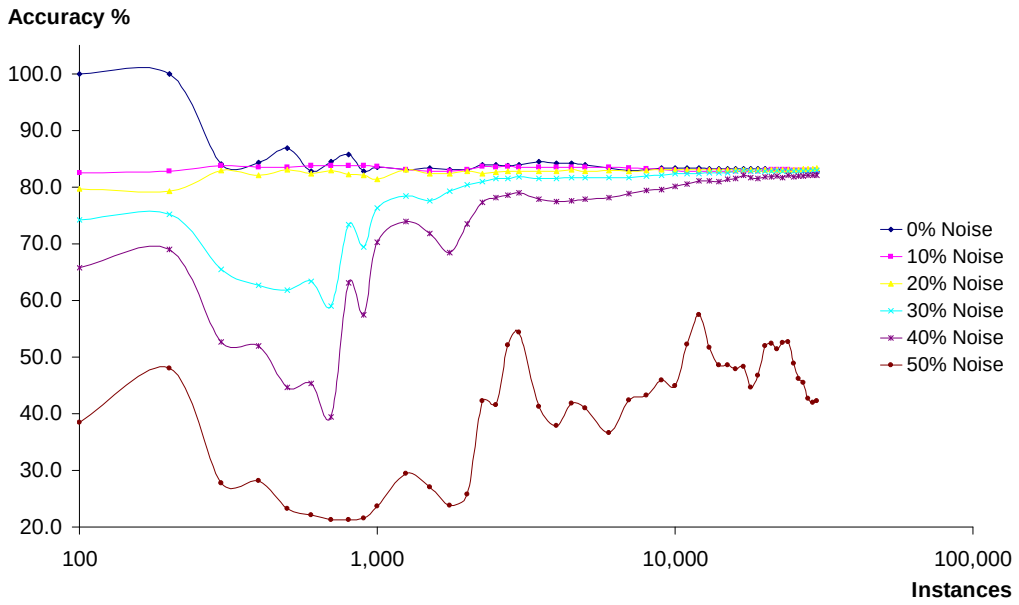


Figure 3.2: The results of the six Census data classifier experiments are shown. In general, noise has delayed achieving the classification accuracy of the noiseless data experiment.

3.3.2 Experimental results

The results of the six Census data classifier experiments, graphed in figure 3.2, show that increasing noise also increases sample size requirements, while a small amount of noise, 10% and 20% noise levels, was associated with accelerated learning. The expression *accelerated learning* is used to describe how a model achieved the accuracy of the noiseless model using fewer samples. The addition of 50% noise made learning impossible.

The noiseless experiment shows an erratic start to the learning process, which stabilized after a thousand instances. The following comments are

made using the noiseless experiment as a *reference* for successful learning and an arbitrary margin of 0.5%; refer to the line labeled 0% noise in figure 3.2. The ten percent noise experiment stabilized sooner that the reference, while the twenty percent noise experiment stabilized soon after the reference, at approximately 1,250 samples. A noise level somewhere between twenty and thirty percent resulted in a dramatic increase in sample size requirements. The thirty percent noise level experiment required 15,000 samples before achieving a performance similar to the reference, while the forty percent noise level experiment failed to match the reference given the available sample size. Finally a noise level of fifty percent prevented successful learning; as predicted by PAC theory.

An interesting pattern emerged around 700 instances in figure 3.2, which is particularly evident for 30% and 40% noise levels. This pattern clearly shows the presence of artifacts within the data and confirms that the sample sets are identical, except for different noise levels.

Equation (3.3) cannot be compared precisely with these experimental results for two reasons: the equation states the maximum sample requirements and the equation variable ϵ is unknown. Regardless of this, the impact of noise on sample size is clearly dramatic. From a non-classical domain perspective, this dramatic increase can prevent successful learning according to PAC theory.

3.3.3 Discussion

Small amounts of noise appear to accelerate model learning, however a point is reached where additional noise significantly increases the number of samples required. This suggests that a data set that is significantly affected by noise requires an approach for managing noise and seeking more from available data.

The PAC theory presented and equation (3.3), which was derived from the theory, states that a substantial increase in sample size is required as noise levels increase. This increase in sample requirements, in the case of a non-classical problem domain, could result in a requirement for sample size that exceeds the number of samples available.

The "proof of principle" has largely supported the above presented theory. In particular, increasing noise results in significantly increased sample requirements, although the experiment demonstrated that fewer samples were required for the 10% and 20% noise level data sets. In any case these results do not contravene the theory, since the theory defines the upper limit for the required sample size for successful learning, while some circumstances may serendipitously enable accelerated learning. Accelerated learning could be due to the nature of the samples provided or some property of the learning algorithm that benefits from the provided samples. However one can not

predict in advance whether accelerated learning will occur—therefore how much noise can be tolerated is determined by the available sample size and equation (3.3).

In the context of non-classical domains, the number of samples required, because of noise, may far exceed available data. In such a context there is a likelihood that poor model performance will be achieved and could explain the findings of Sima and Dougherty (2006), who state that "the inability to find a good feature set should not lead to the conclusion that good feature sets do not exist". It is also conceivable that insufficient samples, according to the presented theory, could result in learning that appears successful but is in fact a poor generalization of the phenomenon of interest; or was misled by apparent information. Therefore the above theory defines the minimum required number of samples in order to achieve successful learning or a model that can be rationally expected to be a good generalization. As a result, a non-classical problem, such as 10,000 features and 200 samples, which are affected by noise, is likely to fail these requirements. A potential solution to this problem, the subject of this thesis, is to exploit feature redundancy and evoke a preference for features whose sample data is least affected by noise.

3.4 Data quality

Noise is viewed as the influence of an external signal on sample data. Typically, sample data *contains* noise, which uniquely alters the sample data from its "true value"; the value that would have been measured in the absence of noise. Let the matrix **D** represent the sample data, **S** the true, but generally unknown value of the sample set and **E** as noise. Hence **D** is equal to the sum of **S** and **E**.

However sample data need not contain noise in order for its true value to be effectively altered. In certain contexts, time can effectively alter the true value, for example, rainfall data collected one hundred years ago may be less relevant, or effectively less accurate if rainfall patterns change over time.

Given that the accuracy of sample data may be compromised from the addition of noise, or externally through the impact of time, for example, the more general expression of "data quality" will be used.

3.5 Data mining

This section defines "Sample data", "Signal" and "Data quality".

3.5.1 Sample data

Sample data is collected from a phenomenon to be modeled and is represented by the matrix

$$\mathbf{D} = \begin{pmatrix} d_{11} & d_{12} & d_{13} & \dots & d_{1j} & \dots & d_{1p} \\ d_{21} & d_{22} & d_{23} & \dots & d_{2j} & \dots & d_{2p} \\ \vdots & \vdots & \vdots & \dots & \vdots & \dots & \vdots \\ d_{i1} & d_{i2} & d_{i3} & \dots & d_{ij} & \dots & d_{ip} \\ \vdots & \vdots & \vdots & \dots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & d_{n3} & \dots & d_{nj} & \dots & d_{np} \end{pmatrix} \tag{3.5}$$

where each row corresponds to an "instance" of sample data and each column to a different feature or dimension. The i^{th} instance of sample data and its j^{th} dimension if designated by d_{ij}, while the number of instances and dimensions are n and p respectively. The class label of each row is not included since its quality is not assessed and as will be shown later, permits the inclusion of unlabeled data points within the feature selection process.

Signal

Sample data $\mathbf{D}$ is often influenced by noise, which can result in $\mathbf{D}$ being different to the true "signal" or behavior exhibited by the underlying structure of the phenomenon. The true, but generally unknown signal is represented by the matrix

$$\mathbf{S} = \begin{pmatrix} s_{11} & s_{12} & s_{13} & \dots & s_{1p} \\ s_{21} & s_{22} & s_{23} & \dots & s_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ s_{n1} & s_{n2} & s_{n3} & \dots & s_{np} \end{pmatrix} \tag{3.6}$$

The extraneous signals that cause $\mathbf{D}$ and $\mathbf{S}$ to be different are generally termed noise. This difference is represented by

$$\mathbf{D} = \mathbf{S} + \mathbf{E} \tag{3.7}$$

where $\mathbf{E} \in \Re^{n \times p}$ contains the extraneous signals or noise. This equation suggests that successfully deducing $\mathbf{S}$ from $\mathbf{D}$ requires $\mathbf{E}$ to either be insignificant or known.

Data quality

Matrix $\mathbf{E}$ provides a measure of the quality of $\mathbf{D}$. If every entry in $\mathbf{E}$ is zero then $\mathbf{D}$ has perfect quality.

The quality of $\mathbf{D}$ is determined by equation (3.7) and $\mathbf{E}$ can be decomposed as

$$\mathbf{E} = \mathbf{E_{est}} + \mathbf{E_{unk}} \tag{3.8}$$

where $\mathbf{E_{est}}$ can be estimated and is ideally the more significant component of $\mathbf{E}$, while $\mathbf{E_{unk}}$ is unknowable or too complex to accurately estimate.

Determining $\mathbf{S}$ using equation (3.7) requires an accurate estimate of $\mathbf{E_{est}}$ and many techniques depend on such an approach. However in the context of a non-classical problem, this thesis argues that estimating $\mathbf{E_{est}}$ is difficult and therefore presents a novel approach. The approach depends on estimating a *worst case* for $\mathbf{E_{est}}$ and although such an estimate cannot be used in

equation (3.7) to calculate the true signal (**S**), it does provide a measure of the quality of **D**.

Training, test and unlabeled data

Training and test data are logically contained in separate data sets, $\mathbf{D_1}$ and $\mathbf{D_2}$ respectively. In addition there are the corresponding but unknown noise matrices: $\mathbf{E_1}$ and $\mathbf{E_2}$. In the context of traditional methods, Quinlan (1986b) showed than an unmanageable problem occurs when the noise contained in the test set is significantly different to that of the training set, since the selected features may prove to be quite ineffective when the model is applied. As reported in the literature review in section 2.4, there still does not appear to be any methods which manage differences in noise—or alternatively data quality. In particular there does not appear to be any methods that use data quality to manage the feature selection process.

Non-traditional data

Feature selection and model learning traditionally uses data that is related to the phenomenon being modeled; it provides information related to the underlying processes involved. This thesis investigates incorporating another source of data known as "non-traditional" data. This data does *not* provide any information about the phenomenon, but can provide a measure of the data's quality. For example, the intensity emitted by a microarray spot is

related to gene activity, while a measure of the spot's "roundness" is entirely unrelated. However the spot's roundness does influence the quality of the intensity reading.

3.6 Feature utility ranking methodology

In the context of non-classical problems, Sima and Dougherty (2006) declared the need for feature selection techniques that are "not purely data driven". By this they mean not entirely driven by the use of traditional data. The Feature Utility Ranking (FUR) methodology fulfills this need by incorporating within the feature selection process—the use of prior knowledge and information that is indicative of the data's quality. Quinlan (1986b) identified a more specific problem, where differences in the quality of the training and test sets is undetected by traditional techniques and can result in the selection of inappropriate features—a variant of FUR specifically addresses this problem.

This section describes the FUR methodology and two variants: FUR1 and FUR2. Although the methodology is designed to allow the use of different measures of quality, one specific measure, which uses a common method for estimating noise, was empirically evaluated. The key difference between FUR1 and FUR2 is the use of signal-to-noise and *only* noise respectively, in order to measure quality. Noise was evaluated using the standard deviation

of pixel intensity.

3.6.1 Methodology overview

Using equations (3.7) and (3.8), the measurable difference between $\mathbf{D}$ and $\mathbf{S}$ is accounted for by $\mathbf{E_{est}}$, which effectively provides a measure of quality for $\mathbf{D}$. However $\mathbf{E_{est}}$ must be estimated since $\mathbf{S}$ is unknown. Each element of $\mathbf{E_{est}}$ is estimated using the noise vector ($\mathbf{n}$), where the vector's elements are combined by a user defined function $f_1(\mathbf{n})$. This function and the noise vector are constructed using prior knowledge about the problem and the data set. Note that the noise vector used in chapter 4 contains a single element, which is the standard deviation in pixel intensity.

Once $\mathbf{E_{est}}$ is estimated, a measure of quality ($\mathbf{Q}$) for $\mathbf{D}$ is calculated using a process represented by the equations

$$\mathbf{Q} = f_2(\mathbf{D},\ \mathbf{E_{est}}) \tag{3.9}$$

and

$$\mathbf{Q} = f_3(\mathbf{E_{est}}) \tag{3.10}$$

where f_2 and f_3 represent the FUR1 and FUR2 methodologies respectively.

Ideally $\mathbf{Q}$ will be calculated such that the worst case quality of $\mathbf{D}$ is estimated, which parallels the reasoning used by the PAC theory presented in section 3.2; a model of determined accuracy will be achieved with the calculated sample size, however fewer samples may be required.

Both $\mathbf{Q}$ and $\mathbf{D}$ have the same dimensionality, since each element of $\mathbf{Q}$ contains a scalar measure for the quality of the corresponding element in $\mathbf{D}$. Equation (3.9), which corresponds to FUR1, calculates $\mathbf{Q}$ using a signal-to-noise approach, hence

$$\mathbf{Q} \propto \mathbf{D} \times \mathbf{E}_{\mathbf{est}}^{-\mathbf{1}}$$

where $\mathbf{D}$ is used since $\mathbf{S}$ is unknown. Equation (3.10), which corresponds to FUR2, calculates $\mathbf{Q}$ *only* using noise, hence

$$\mathbf{Q} \propto \mathbf{I} \times \mathbf{E}_{\mathbf{est}}^{-\mathbf{1}}$$

thereby enabling the inclusion of the test and unlabeled data sets in $\mathbf{E}_{\mathbf{est}}$.

The FUR methodology consists of the following main steps:

1. Build noise vector
2. Build quality matrix
3. Calculate feature trustworthiness
4. Select trusted features
5. Select discriminative features

where all the steps, except for 2, were implemented identically for the experiments conducted.

Step 1 : Build noise vector

A noise vector $\mathbf{n}_{ij}$ is associated with each element d_{ij} of $\mathbf{D}$ and contains one or more elements. Each element of $\mathbf{n}_{ij}$ provides a different measure of noise that impacts the accuracy of d_{ij}. In the case of microarrays, the following three elements might be used: standard deviation of pixel intensity, precision of the alignment of the mask and the intensity of the background. The noise vector for each d_{ij} contains the same number of elements.

Step 2 : Build quality matrix

Since the quality matrix $\mathbf{Q}$ has the same dimensionality as $\mathbf{D}$, each q_{ij} provides a measure of quality for d_{ij}. Each q_{ij} is a function of its corresponding noise vector. The user of the methodology is responsible for constructing the function that subsumes the noise vector into a scalar value contained in q_{ij}.

Step 3 : Calculate feature trustworthiness

Each column of $\mathbf{Q}$ (or $\mathbf{q}_j$) corresponds to a feature contained in $\mathbf{D}$. Each $\mathbf{q}_j$ is subsumed into a scalar by a user defined function and the resulting scalar is known as the feature's trustworthiness (t_j). The methodology encourages the user to define the function using prior knowledge of the modeling problem; a simple example is the median of $\mathbf{q}_j$.

Step 4 : Select trusted features

Using three sub-tasks, a subset of the most trustworthy features is produced. The first sub-task ranks all of the features according to their trustworthiness. Then a minimum trustworthiness threshold is determined and lastly a subset of features exceeding the threshold is produced.

Step 5 : Select discriminative features

The last step follows similarly to the previous, except that a traditional feature selection method is applied to the subset of trustworthy features produced in step 4. This step is responsible for selecting features according to their information content and results in set of features that are also trustworthy.

3.6.2 Build noise vector

The noise vector, or $\mathbf{n}_{ij}$, contains different measures of the quality of d_{ij}, with each element measuring a different aspect of d_{ij}. The noise vector *only* contains non-traditional data.

Consider an example where the goal is to accurately measure rainfall and $\mathbf{n}_{ij}$ contains:

1. Known accuracy of the measuring instrument, if multiple are used

2. Delay since precipitation ceased and the measurement was made

3. Average air temperature during the delay

4. Average wind speed during the delay

5. Average humidity during the delay.

Each of these measures, provides input for determining the amount of noise affecting the measured rainfall, or d_{ij}. Note, if the first measure was only used, it would provide a genuine measure of quality. However the second and third measures are of little or no benefit in isolation, while together their benefit increases as either increases. For example, if the measurement delay is significant and average temperature is conducive of evaporation, these two measures are particularly valuable. Therefore individual elements of $\mathbf{n}_{ij}$ may be valuable in isolation; dependent on each other; or provide an independent check, for example, two approximate sources of time.

In the case of FUR2, it is a requirement that *no* element of $\mathbf{n}_{ij}$ be correlated with the class label corresponding to the i^{th} row of $\mathbf{D}$, since the test set is also being evaluated. However, since it is intended that a noise vector only contain non-traditional data, the above requirement should be applied irrespective of the variant of FUR. This separation of traditional (d_{ij}) and non-traditional ($\mathbf{n}_{ij}$) components provides two benefits: the standardization of calculating quality in equations (3.9) and (3.10) and the ability of equation (3.10) to consider training, test and unlabeled data.

The noise vector provides a method for exploiting domain knowledge, as illustrated in the example of accurately measuring rainfall. Using domain knowledge enables the user to calculate the $\mathbf{E_{est}}$ component of equation (3.8). Domain knowledge about microarrays reveals numerous areas where quality issues may exist, refer section 2.2, such as the visual appearance of a spot, specifically regarding its shape and color.

A spot's shape is expected to impact accuracy since signal measurement assumptions exist regarding its size and shape. A restricted area exists within which a spot is expected to reside and it is in this area that intensity measurements are made. If an area of the array is measured that is not associated with the spot, then the spot's intensity reading will be flawed. Similarly, if a part of a spot is not measured at all, a flawed reading will be achieved.

The overall appearance of a spot's color also provides a perception of the likely quality of any intensity reading. Normally a spot's color ranges from red through to green and will also display a consistency in it's overall appearance, while any unexpected color or consistency is indicative of a quality issue. However a method for detecting such issues is generally not employed and therefore reliable and unreliable intensity readings are treated equally.

Whether different laser scanners are used also provides insight regarding potential quality issues. Different scanners have associated accuracies which impact the scanned image's intensity and is therefore a potential candidate

for an element of the noise vector. Other potential elements are the age of the array at the time of scanning; environmental factors which affect scanning accuracy, temperature and humidity; and operator experience.

Work by other researchers can also provide a source of methods for calculating a measure of confidence. With respect to microarrays, a method for overall spot quality was developed by Wang et al. (2001). They utilized five spot metrics: spot size, signal-to-noise ratio, local background variability, presence of excessive high local background and occurrence of photo intensity saturation, refer section 2.2. Feature Utility Ranking can be viewed as a generalization of that work, since any number of metrics can be used and because aspects other than his five spot metrics can be evaluated.

In the case of the experimental portion of this thesis, the same two-dimensional $\mathbf{n}_{ij}$ vector is used in FUR1 and FUR2, which incorporate equations (3.9) and (3.10) respectively. These two entries are used since d_{ij} is a ratio of the *control* signal and *test* signal strengths, hence μ_c/μ_t, resulting in a $\mathbf{n}_{ij}$ vector consisting of $(\sigma_c,\ \sigma_t)$, which are the standard deviation of pixel intensity for the *control* signal and *test* signal strengths.

3.6.3 Build quality matrix

The quality matrix

$$\mathbf{Q} = \begin{pmatrix} q_{11} & q_{12} & q_{13} & \cdots & q_{1p} \\ q_{21} & q_{22} & q_{23} & \cdots & q_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ q_{n1} & q_{n2} & q_{n3} & \cdots & q_{np} \end{pmatrix} \tag{3.11}$$

has the same dimensions as $\mathbf{D}$ with each element providing a measure of quality for the corresponding element in $\mathbf{D}$. As a consequence, each element of $\mathbf{Q}$ is independent of other elements. It is considered desirable that each q_{ij} be a worst case estimate in order to reduce the risk of attributing greater quality than exists.

The content of $\mathbf{Q}$ is determined according to the variant of FUR; for FUR1

$$q_{ij} = \frac{d_{ij}}{f_1(\mathbf{n}_{ij})} \tag{3.12}$$

and for FUR2

$$q_{ij} = \frac{1}{f_1(\mathbf{n}_{ij})} \tag{3.13}$$

where $f_1(\cdot\ )$ combines the elements of $\mathbf{n}_{ij}$ into a scalar. The function $f_1(\cdot\ )$ is determined by the user using domain knowledge.

3.6.4 Calculate feature trustworthiness

Trustworthiness of the j^{th} feature is a scalar value used for ranking features by their relative accuracy. The use of trustworthiness depends on two assumptions: that features are independent of each other with respect to their

individual accuracy (ie. no feature interaction) and that the calculation of trustworthiness is independent of any feature specific properties, therefore permitting the direct comparison of features.

The general formula for calculating trustworthiness is

$$t_j = h(\mathbf{q}_j) \tag{3.14}$$

where $h(\cdot\ )$ combines the elements in the j^{th} column of $\mathbf{Q}$, according to some predetermined requirement, into a scalar. Allowing a predetermined requirement enables the user, in conjunction with domain knowledge, to focus on specific data characteristics, such as statistical "accuracy" or "precision" (Cook and Weisberg, 1999) or some combination of the two. Accuracy refers to how closely a set of values are located, on average, to the expected value. Precision refers to the variation between individual values, for example standard deviation. Some possible predetermined requirements are: the mean, the standard deviation and statistical p-value. The predetermined requirement used in this research was the median value.

3.6.5 Select trusted features

Feature Utility Ranking depends on the presence of redundant features from which the most trustworthy features can be selected. The effectiveness of this operation is dependent on setting an appropriate threshold, which removes all but the most trusted features that also provide all the required information

for prediction.

A single trustworthiness threshold is used for the experiments presented in section 4.1.3. Briefly, the approach used seeks to remove the least trusted features without risking the loss of information required for effective modeling. This approach assumes that a single threshold is equally applicable to every feature and that the same distribution of redundant features exists. In simple terms and ignoring noise, redundant features form subsets where each member of a subset provides the same information content. It is an assumption that every feature that is required by the model, exists in an adequately large subset and that least one member of the subset will be located above the threshold. It is also assumed that a sufficient number, or ideally the majority of features, will fall below the threshold.

Multiple thresholds

A future version of FUR could employ a different threshold for each redundant subset, such that a single feature in the subset exceeds the threshold. The effective operation of this approach depends on the identification of redundant subsets (Ding and Peng, 2003; Olivetti, Veeramachaneni, and Avesani, 2008) and the corresponding removal of features.

3.6.6 Select discriminative features

Feature utility ranking does not prescribe the method used for selecting features on the basis of the information content they provide. Furthermore, the operation of FUR does not depend on whether any features are eliminated according to calculated information content; FUR depends on eliminating features that fall below a minimum threshold of trustworthiness.

3.7 Feature utility ranking 1

This section describes an implementation of a variant of the base methodology, called Feature Utility Ranking 1 (FUR1). The key characteristic of FUR1 is its approach to calculating a feature's trustworthiness, which involves a form of signal-to-noise ratio. Signal-to-noise ratio is calculated using the signal's strength and a measure of noise, which provides a measure of confidence that $d_{ij} = s_{ij}$. The motivation for using a signal-to-noise approach stems from communication theory and the minimization of errors (Shannon, 1948, 1949; Carlson, 1975). Fundamentally it states that the magnitude of noise that can be tolerated, without causing errors, increases as the magnitude of the signal increases. The use of signal-to-noise by FUR1 involves two aspects: the magnitude of the signal for a sample point and the magnitude of the difference between two sample points, where each corresponds to a

different class label. This variant evaluates both signal magnitudes for the purpose of calculating trustworthiness.

Another key characteristic of FUR1, which is inherited from FUR, is the mechanism used for filtering poor quality data. Consider a data point, which corresponds to a single row in matrix $\mathbf{D}$, one approach to managing poor data quality is removing an entire data point. However that approach reduces the number of samples for every feature. Feature Utility Ranking operates under the premise that data quality, in general, varies for each $\mathbf{d_{ij}}$. Consequently FUR evaluates the overall quality of data that is associated with a feature and whether the feature should be removed—this approach reduces dimensionality rather than the number of samples.

The mechanism used for filtering poor quality data by FUR addresses the difficulties associated with non-classical problems. The motivation for the approach used is explained using standard statistical theory. According to the "central limit theorem": a sample will provide accurate estimates of the population mean and standard deviation if the sample size is sufficiently large (Dawson and Trapp, 2001). Hence in the case of the classical problem domain, errors are, in effect, automatically reduced by virtue of large sample sizes (Walpole and Myers, 1978; Adcock, 1997). In other words, even if each sample point is affected by noise, the underlying behavior will still emerge. However a non-classical problem domain dictates the necessity for

new methods, since determining population statistics is more difficult due to small sample sizes and the effects of dimensionality (Tibshirani, 2001; Dudoit, Yang, Callow, and Speed, 2002a; Sima and Dougherty, 2006; Klebanov, Qiu, Welle, and Yakovlev, 2007). Consequently, the use of prior knowledge about the data's quality is being explored as one possible solution. The use of trustworthiness enables FUR1 to avoid features that are more likely to cause errors, while simultaneously providing a method for dimensionality reduction. In using this approach, it is assumed that the accuracy of an estimate for a feature's population parameters is proportional to the feature's trustworthiness.

Feature Utility Ranking 1 uses a two phase feature selection process, where each phase produces a list of features, as shown in figure 3.3. The first phase produces a list of features ranked according to trustworthiness. The second phase takes a subset of the features selected from phase 1 and ranks them according to their information content. The data set shown at the top left of the figure consists of two components: clinical and microarray data. These sets provide the data needed to group the signal, or in the case of microarrays, expression data according to replicate arrays and the biological sample donors. This data together with the clinical data provides the class labels for the biological samples.

In FUR1, the trustworthiness data set is constructed from two inputs:

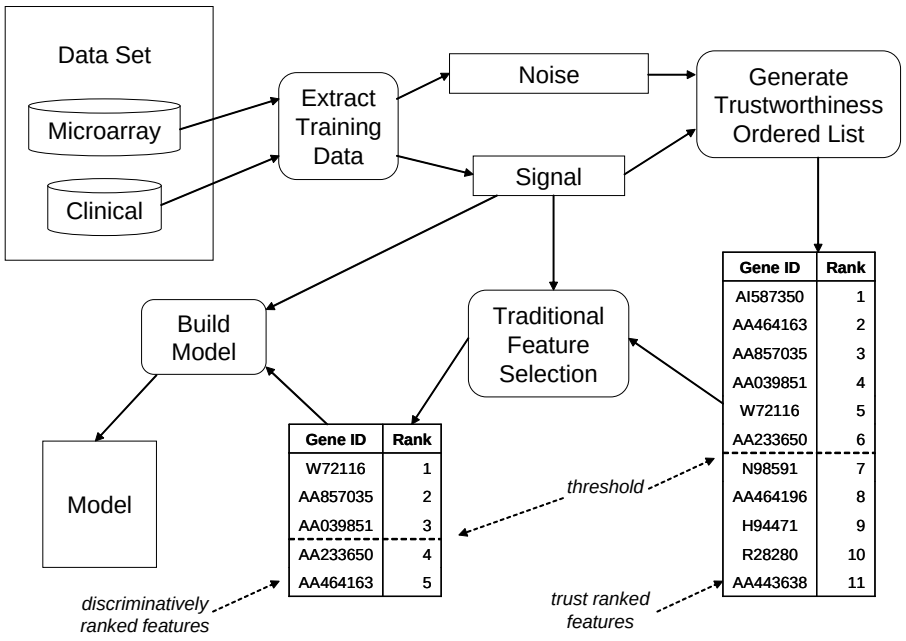


Figure 3.3: The Feature Utility Ranking 1 methodology, which consists of two feature selection phases, is shown. Data quality, which is evaluated in the first phase, is determined using signal data and an estimate of the noise affecting it. Traditional feature selection is used for the second phase.

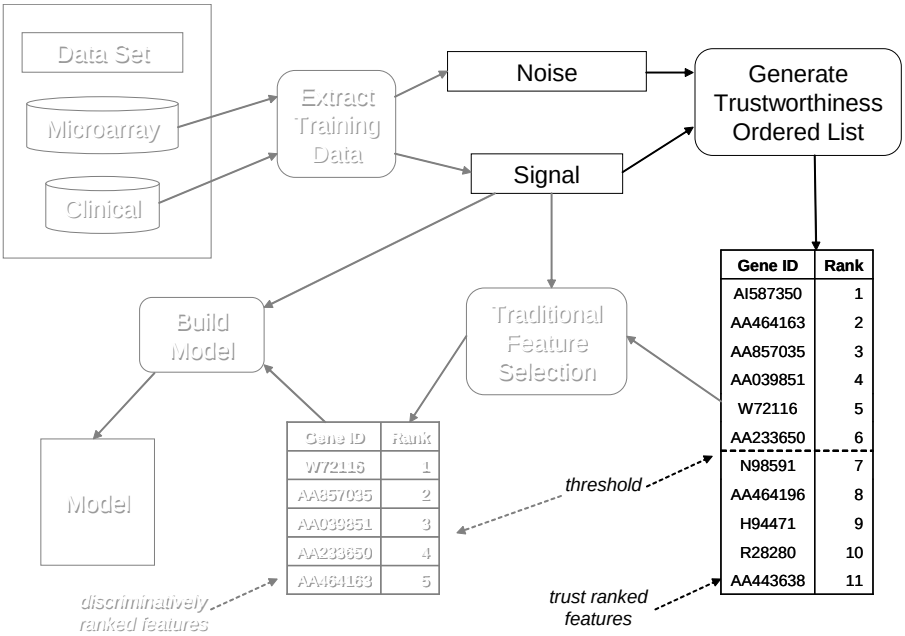


Figure 3.4: The first feature selection phase for Feature Utility Ranking 1 is shown, which selects a subset according to trustworthiness.

signal and quality data. Signal and quality data are combined in order to locate the most reliable signal data. This reliability is two fold: firstly preferring larger signal differences between class labels and secondly preferring features associated with larger signal.

3.7.1 First feature selection phase

The first feature selection phase, which consists of calculating data quality, calculating feature trustworthiness and selecting the most trusted features is summarized in figure 3.4 and described below.

Calculating quality

The first task is calculating a measure of noise in each of the data points corresponding to the feature whose trustworthiness is being calculated. Any appropriate measure of noise can be used. As described earlier, an appropriate measure for microarray data is the standard deviation in pixel intensity for a spot, which was used in the experiments described in chapter 4.

The quality matrix $\mathbf{Q}$ contains a measure of confidence for each corresponding entry in the data matrix $\mathbf{D}$; therefore q_{ij} is a measure of confidence in d_{ij}. Each q_{ij} is constructed using a *noise vector*, which contains one or more scalar values that provide a measure of confidence in d_{ij}. As described earlier, the noise vector $\mathbf{n}_{ij} = (\sigma_c,\ \sigma_t)$ was used, where σ_c and σ_t is the standard deviation of the pixel intensity for the respective *control* and *test* channels associated with d_{ij}. The implementation used for equation (3.12), which is used to calculate the quality value q_{ij}, is

$$q_{ij} = \frac{d_{ij}}{\log\left(\frac{(\sigma_c+\sigma_t)}{2} + |\sigma_c - \sigma_t|\right) + 1} \tag{3.15}$$

where $d_{i,j}$ is the signal reading being evaluated, while σ_c and σ_t are the noise vector entries for that signal reading.

The numerator and denominator of equation (3.15) respectively provide a form of signal-to-noise ratio that is used by FUR1. The denominator consists of two parts: a log function, which maps the typically large range of variation

in pixel intensity into a much shorter range; the addition of one prevents a divide by zero in the event of no variation in pixel intensity. Given that a small denominator is desired, the log function contains two parts: the average variation of pixel intensity and the difference between the two variation of pixel intensities. The two parts of the log function provide a simultaneous preference for a low average standard deviation of pixel intensity and similar standard deviations in pixel intensity.

Calculating trustworthiness

The calculation of trustworthiness is performed on individual features since the trustworthiness of a feature is assumed to be independent of any other feature. Although systematic errors can affect multiple features, individual features do not determine the quality of the data associated with other features. Consequently, the assumption of feature independence is considered reasonable and factors that determine the trustworthiness of a feature are considered solely by the accuracy of the data associated with that feature.

The calculation of the trustworthiness for the k^{th} feature, according to FUR1, is shown in Algorithm 1. Each iteration of the algorithm evaluates pairs of sample points, where each is located in a different class. Although the algorithm presented is limited to binary classification, a simple extension will permit the use of any number of classes.

Algorithm 1 Feature Utility Ranking 1 algorithm for calculating t_k (the trustworthiness of the k^{th} feature contained in $\mathbf{D}$; for the training data set.)

Require: $\mathbf{D}^+$ {Subset of $\mathbf{D}$ containing +ve classifications}
Require: $\mathbf{D}^-$ {Subset of $\mathbf{D}$ containing -ve classifications}
Require: $\mathbf{Q}$ {Quality matrix corresponding to the training data contained in $\mathbf{D}$}

```
1: m = numberOfRows(D+)
2: n = numberOfRows(D−)
3: for i = 1 to m do
4:    d+ = D+_ik {Scalar signal reading contained in D+}
5:    q+ = f(d+, E_ik) {Calculate quality of d+, refer equation (3.9)}
6:    for j = 1 to n do
7:       d− = D−_jk {Scalar signal reading contained in D−}
8:       q− = f(d−, E_jk) {Calculate quality of d−, refer equation (3.9)}
9:       t_(i−1)×n+j = | d+ − d− | × d+ × d− × q+ × q−
         {Add the trustworthiness of an intra class pair to vector t}
10:   end for
11: end for
12: t_k = medianOf(t) {Trustworthiness of k^th feature}
```

Algorithm 1 calculates the weighted trustworthiness of each sample intra-class label pair found in $\mathbf{D}$. This is illustrated in figure 3.5, which depicts an approach similar to that used for "distance measures" in clustering (Dunham, 2003, chap. 5). Using the distance measures paradigm, a "complete link" approach was followed (Dunham, 2003, p.130), although other approaches that are considered more appropriate to a specific problem could be used. The calculated weighted trustworthiness of each sample intraclass label pair is stored in vector $\mathbf{t}$ on line 9 of Algorithm 1.

Lines 1 and 2 calculate the number of data instances that correspond to +ve and -ve class labels respectively. The **for** loop, on line 3 of Algorithm 1,

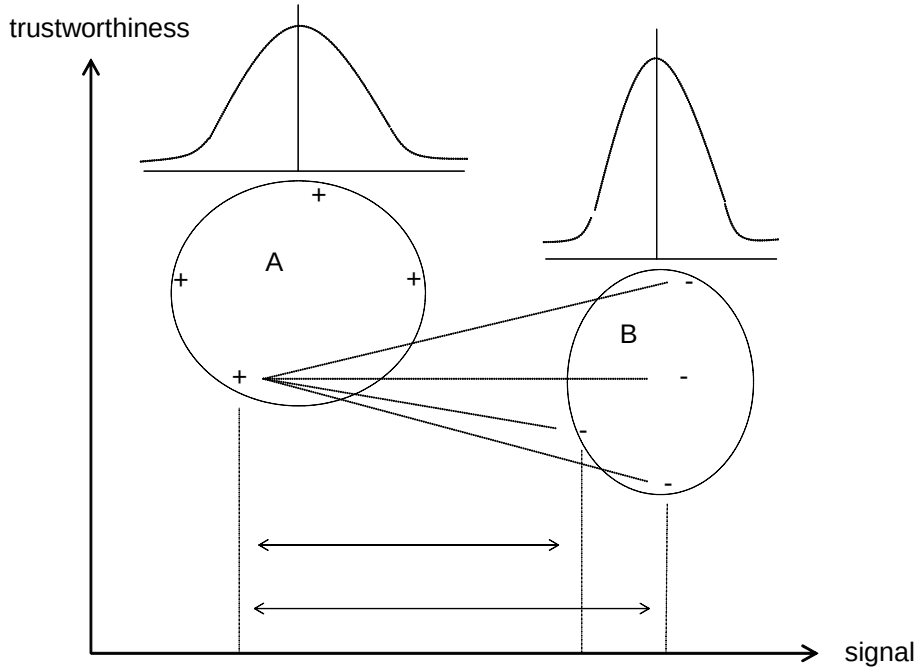


Figure 3.5: Calculating the trustworthiness for the k^{th} feature.

calculates the signal strength d^+ and a measure of quality q^+ in its accuracy, for a +ve classification sample point of the k^{th} feature. This calculation is repeated for a -ve classification, by the **for** loop on line 6. The formula on line 9 calculates the trustworthiness for the k^{th} feature, given the two sample points and is stored in $\mathbf{t}$. This formula evaluates the trustworthiness of the two intraclass sample points according to their individual magnitude, the difference in those magnitudes and a measure of confidence in each. Therefore the motivation is that the individual signal strengths are as large as possible in order to provide the greatest margin against ambient noise. Similarly, the greatest margin between the individual signals minimizes the opportunity for

class confusion, while the individual confidence in each signal reading acts as an overall weight.

The use of $numberOfRows(\mathbf{D}^+) \times numberOfRows(\mathbf{D}^-)$ iterations by FUR1 makes the calculation of trustworthiness an $O(n^2)$ algorithm (Duda et al., 2001), which compares favorably with other tasks involved in constructing the classifier. In addition, the use of trustworthiness reduces the computational effort for selecting features in the second phase according to their information content, which is particularly valuable if a multivariate approach were used.

The final task of Algorithm 1, on line 12, calculates the feature's trustworthiness by determining the median value stored in $\mathbf{t}$. A median function was used since it is less affected by outliers and more likely to provide an accurate assessment of a feature's overall trustworthiness.

Selecting a subset of features according to trustworthiness

This section describes the motivation for the method selected for choosing a subset of trustworthy features and the operation of that method.

The primary goal of the experiments is to determine the benefit of incorporating feature trustworthiness within a feature selection methodology and to compare FUR1 and FUR2. For this reason it was decided that the process of selecting the number of trustworthy features used would be the

same for all the FUR methodology experiments. It was also decided that the number of features removed would eliminate all of the most untrustworthy features regardless of the specific FUR experiment performed. As a result a single trustworthiness threshold would be applied to every FUR experiment, where the threshold is directly determined by the number of features, rather than indirectly, such as through the magnitude of the trustworthiness values calculated.

The selection of the trustworthiness threshold involved a review of all the trustworthiness curves for the experiments performed. These were produced by ranking features according to their trustworthiness. This revealed the presence of a characteristic shape for a trustworthiness curve, an example of which is shown in figure 3.6. The horizontal axis in the figure shows the trustworthiness rank of each feature; the most trusted feature has a rank of one. The vertical axis is the log transformed feature trustworthiness.

The trustworthiness curve in figure 3.6 shows three main parts: the left vertical section, which consists of the most trustworthy features that vary significantly in their trustworthiness; a relatively stable middle section, where the level of trustworthiness varies relatively gradually; a vertical right section containing the least trustworthy features that vary significantly in their trustworthiness. After analyzing the trustworthiness curves from all the experiments, a threshold of 7,000 features was always well clear of the right

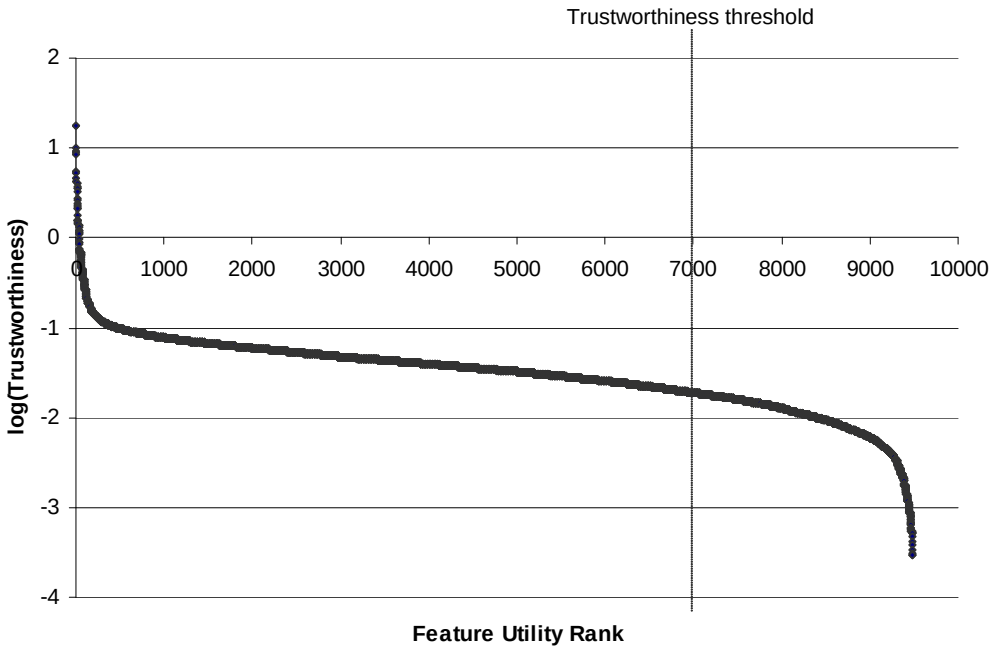


Figure 3.6: All the Feature Utility Ranking experiments produced the same characteristic trustworthiness curve, an example is shown. The same trustworthiness threshold was used in all the experiments.

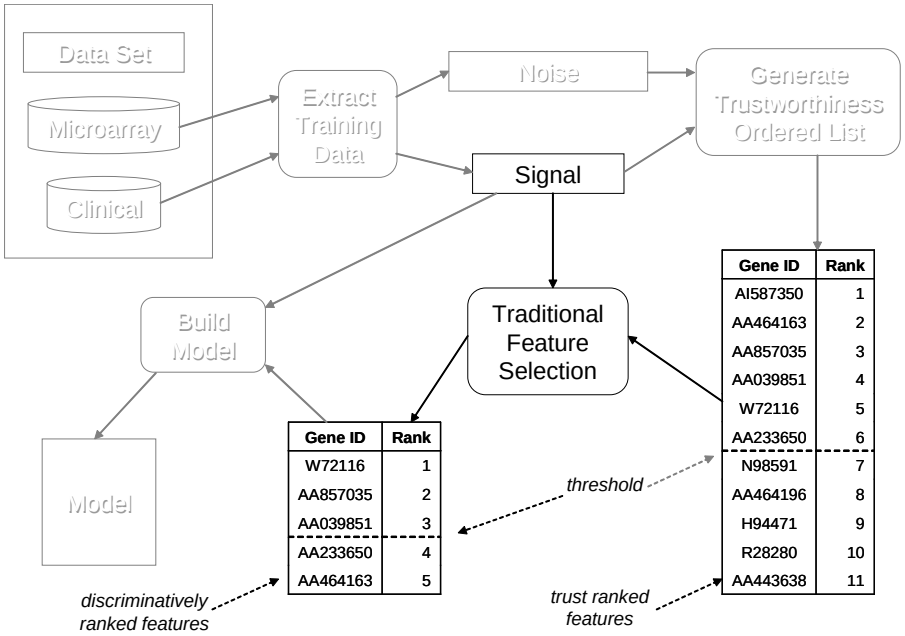


Figure 3.7: The second feature selection phase for Feature Utility Ranking 1, which constructs a subset of features ranked by information content.

vertical section and therefore applied in the experiments.

3.7.2 Second feature selection phase

The second feature selection phase begins with the subset of trustworthy features generated by the first phase and produces a list of features ranked according to their information content, as shown in figure 3.7.

For comparison, experiments were repeated using different numbers of features, either 16, 1 024 or 9 485 of the top n features ranked by information content. The motivation for a 16 feature set is to determine whether the use of trustworthiness can produce a difference in such a small feature set. A

1,024 feature set was considered sufficiently large to establish whether using trustworthiness could produce substantial differences in which features are used and their respective ranking. The 9,485 feature set, which is the entire feature set available, provides a baseline for comparing the gains provided by the individual feature selection methodologies.

Other than for the three feature set sizes used, the same method for calculating information content, namely InfoGain, was used in order to minimize unintended experimental variability.

3.8 Feature utility ranking 2

This section describes Feature Utility Ranking 2 (FUR2). Unlike FUR1, FUR2 incorporates the use of the test and unlabeled data sets within the calculation of feature trustworthiness. This refers to data that is *not* used during traditional feature selection. In this section, only differences between FUR1 and FUR2 will be described, since the architecture and operation of both methodologies is identical.

The motivation for incorporating test data within the calculation of trustworthiness is the expectation that different noise distributions exist for the training and test sets. In order to use the *entire* data set within the calculation of feature trustworthiness, the **D** matrix cannot be directly used. It however can be used by populating the noise vector with some measure of

correlation found within a column of the matrix. In order to compare the architectural differences of FUR1 and FUR2, the same measure of noise is used. The architectural differences are—using only noise versus signal-to-noise, in order to determine data quality; hence the *entire* data set can be used instead of only the training data.

Feature Utility Ranking 2 only uses noise estimates for all the available data, hence the training and test data sets are used by the first feature selection phase. Although the inclusion of test and unlabeled data results in reconstructing the classifier whenever new data comes to hand, it forms a protection against classification errors, which are a particular concern for non-classical problems.

Feature Utility Ranking 2 also consists of two feature selection phases composed of five main components, which are summarized in figure 3.8. The fundamental difference between FUR2 and FUR1 is the removal of signal (or expression) data from the calculation of trustworthiness.

3.8.1 First feature selection phase

The first feature selection phase, composed of three main components, shown in figure 3.9, begins with the entire feature set and produces a subset of features ranked according to their trustworthiness.

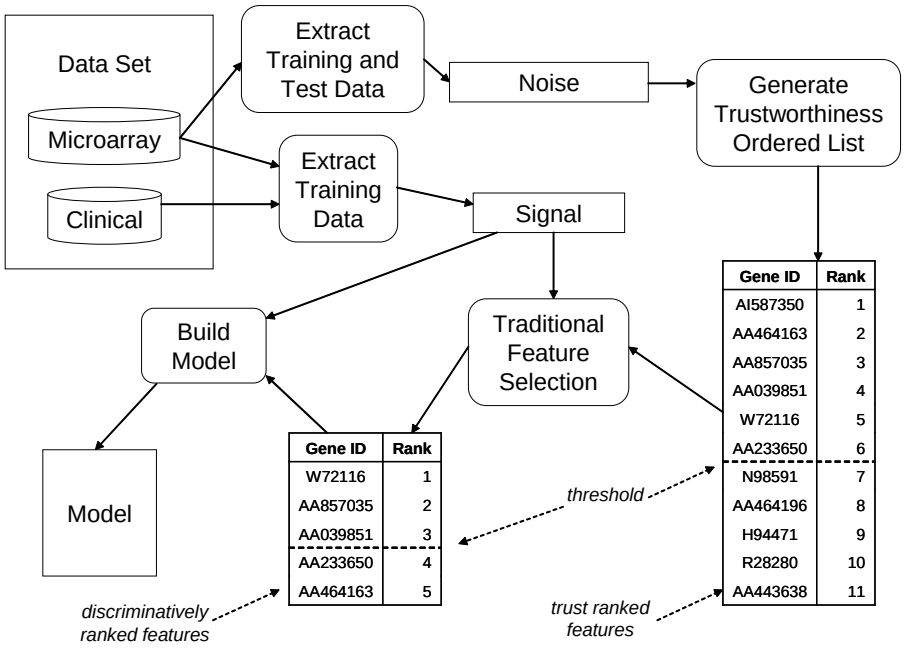


Figure 3.8: The Feature Utility Ranking 2 methodology, which consists of two feature selection phases, is shown. Data quality, which is evaluated in the first phase, is determined using an estimate of the noise affecting the signal data. Traditional feature selection is used in the second phase.

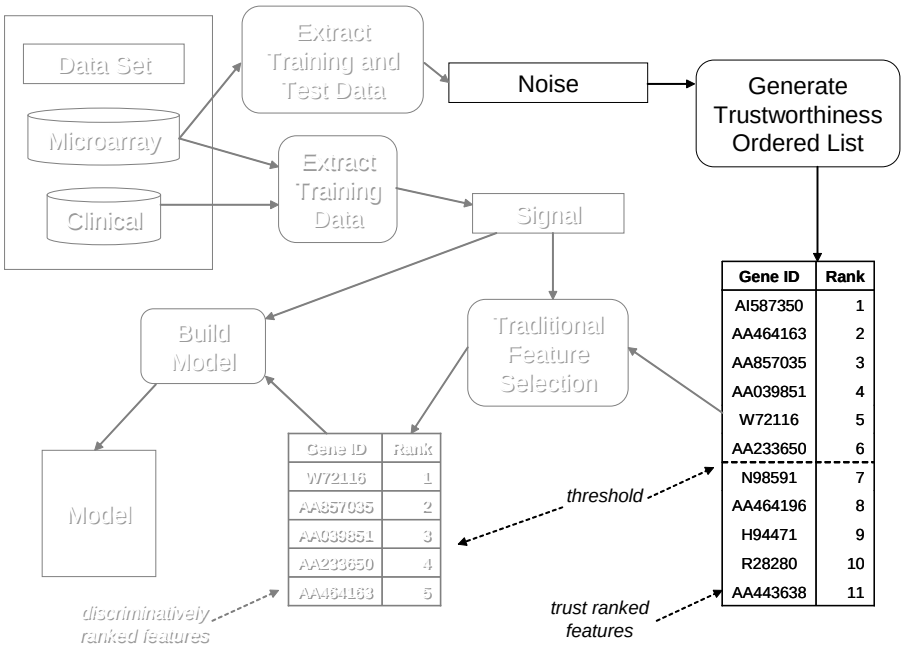


Figure 3.9: The first feature selection phase of Feature Utility Ranking 2, which selects a subset according to trustworthiness.

Calculating quality

The quality matrix **Q** still contains a measure of the confidence for each entry in **D**. Therefore q_{ij} is a measure of confidence in d_{ij} and each q_{ij} is constructed using the same noise vector as used by FUR1.

The implementation of equation (3.13) is

$$q_{ij} = \frac{1}{\log\left(\frac{(\sigma_c+\sigma_t)}{2} + |\sigma_c - \sigma_t|\right) + 1} \tag{3.16}$$

where σ_c and σ_t are same noise vector entries used by FUR1. Note that the only difference between this implementation and that for FUR1 is the absence of the signal component d_{ij} from the numerator.

Calculating trustworthiness

The calculation of trustworthiness, shown in Algorithm 2, follows the same approach used for FUR1, however, the exclusion of the direct use of **D** results in simplification. In addition, application of the assumption of feature independence, as in FUR1, produces an $O(n)$ algorithm resulting in a reduction in computational effort compared to FUR1.

The **for** loop on line 1 of Algorithm 2 iterates through each row of the **D** matrix for the k^{th} feature. Note that **D** includes training and test data. Using equation (3.10), line 2 calculates a measure of confidence for the corresponding entry $\mathbf{d}_{ik}$. Line 3 stores the calculated confidence measure in vector

Algorithm 2 Feature Utility Ranking 2 algorithm for calculating t_k (the trustworthiness of the k^{th} feature contained in **D**; for training, test and unlabeled data.)

Require: **Q** {Quality matrix corresponding to training, test and unlabeled data contained in **D**}

1: **for** $i = 1$ to numberOfRows(**D**) **do**
2: $q = f(\mathbf{E}_{ik})$ {Calculate quality of d_{ik}, refer equation (3.10)}
3: $\mathbf{t}_i = q$ {Add the trustworthiness of the i^{th} row to vector **t**}
4: **end for**
5: $t_k = medianOf(\mathbf{t})$ {Trustworthiness of k^{th} feature}

t and line 5 calculates the trustworthiness, which is the median value of the vector entries.

Selecting a subset of features according to trustworthiness

The same subset selection method as presented for FUR1 is used for selecting features, which are ranked by trustworthiness. A review of all the experiments performed showed that a FUR2 trustworthiness curve is very similar to that produced by FUR1, refer figure 3.6. Consequently the same trustworthiness threshold of 7,000 was used.

3.8.2 Second feature selection phase

The same approach as for FUR1 was used for the second feature selection phase, which is shown in figure 3.10.

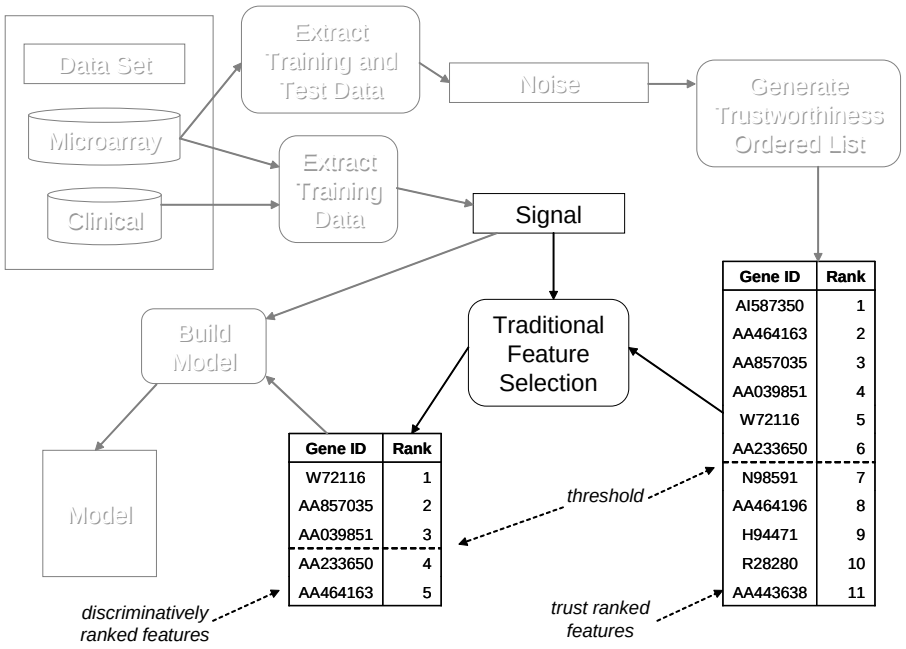


Figure 3.10: The second feature selection phase for Feature Utility Ranking 2 constructs a subset of features ranked according to information content.

3.9 Discussion

This chapter considered the problem of uncovering the underlying signal that is hidden within a noisy data set. In section 3.2, Probably Approximately Correct (PAC) theory was used to determine the relationship between the amount of noise present within the data set and the minimum amount of data required to guarantee the successful construction of a model of predetermined accuracy. The determined relationship, shown in equation (3.3) on page 88, showed that the number of training samples increases exponentially in response to increasing noise. Given a non-classical problem, it was shown in section 3.2.3, that increasing noise can quickly result in a problem where the required accuracy can not be achieved because of insufficient sample data. The proposed solution is eliminating features whose associated data quality falls below a predetermined level. By eliminating such features, it is argued that sufficient sample data will be available for constructing a model of required accuracy.

Using census data, it was shown in section 3.3 that noise generally results in a dramatic increase in the number of samples required to achieve a predetermined level of model accuracy. This finding is consistent with PAC theory, which acknowledges that an upper limit of training examples exists for a given amount of noise. This finding suggests that a complex relationship

exists between noise and the precise number of samples required in a specific learning problem. As a result a definition for data quality was provided in section 3.4, while section 3.5 described the relationship between data quality, noise and signal. Given the complex relationship between sample size, noise and required model accuracy, a measure of feature dependability or "trustworthiness" was developed. Using a process that parallels the use of entropy to rank features according to their discriminative information, one can also rank features according to their trustworthiness, or the quality of the data that describes them.

A feature selection methodology was proposed in section 3.6 that utilizes trustworthiness. This methodology, which is called Feature Utility Ranking (FUR), provides a framework for constructing a measure of noise affecting each feature and consequently calculating a measure of trustworthiness for the feature. Lastly the methodology provides a process for merging a feature's trustworthiness and a traditional measure of its discriminative ability, in order to produce a single ranked list.

Two versions of FUR, namely FUR1 and FUR2, were given in sections 3.7 and 3.8. Feature Utility Ranking 1 evaluates data quality using an accepted signal-to-noise paradigm, while Feature Utility Ranking 2 uses a noise-only measure. The use of noise-only enables FUR2 to calculate the quality of the training and the test sets, as well as unlabeled data. This ability to

consider the quality of the *entire* data set addresses a specific need raised by Quinlan (1986b). Although considering the quality of unlabeled data can result in the reapplication of feature selection and model construction, it provides a mechanism for choosing the most effective features in response to data quality. This ability to tailor the feature set to unlabeled data also caters for the situation where the need exists to manage the cost of different prediction errors.

Feature Utility Ranking addresses the call by Hastie et al. (2001) to use prior knowledge about the quality of the data. Feature Utility Ranking also provides a mechanism for incorporating tailored methods that assess the quality of individual features. This ability to assess individual features acknowledges that different factors or measuring instruments may be associated with each feature. It also provides the opportunity to incorporate the outcome of data cleaning into feature selection, rather than implicitly assuming that every item of data has the same quality.

Chapter 4

Experiments and results

This chapter begins by describing and justifying the experimental design used. Assumptions and the expected limitations of the experiments are provided, together with a description of the data sets involved, the third party tools and data pre-processing employed. Next the experimental results are presented and lastly a discussion and conclusion.

4.1 Experimental design

The precise impact of noise on feature selection and model learning is unpredictable, since it is determined by the interaction between a unique data set (**D**) and unique noise (**E**), refer section 3.4. This unpredictable nature has led PAC researchers to a statistical approach for defining a *worst-case bound* for the number of samples required for model learning in the presence of noise. Being a bounded statistical estimate, the number of samples required could be significantly less and could also be more than estimated. Given this diffi-

culty, a comparative experimental approach is used to identify whether any benefit can be derived from incorporating data quality information within the feature selection process.

4.1.1 Approach

The comparative approach consists of repeating an experiment such that the only difference is whether data quality information is used within the feature selection process. The experiment consists of feature selection, model construction and comparison of the models with respect to their classification accuracy and differences in the features used by each model.

The absence of data quality within the feature selection process is achieved through using a popular feature selection approach, which is referred as a *traditional methodology* and which assumes feature independence during the calculation of a feature's discriminative ability. A search strategy that considers feature interaction will be referred to as *multivariate feature selection*. Using a traditional feature selection methodology is considered appropriate when a non-classical problem domain applies, since the curse of dimensionality can easily make multivariate feature selection computationally expensive or intractable. The use of a traditional feature selection methodology is a benchmark to the limitations of a methodology that does not use data quality information.

The use of data quality within the feature selection process is provided by the two feature selection methodologies Feature Utility Ranking 1 (FUR1) and Feature Utility Ranking 2 (FUR2). The three feature selection methodologies (traditional, FUR1 and FUR2) are tested against three data sets: synthetic, leukaemia and chronic fatigue syndrome. Since the synthetic and leukaemia data sets are divided into two parts, a total of thirteen individual experiments are conducted.

Feature selection methodologies

Each of the three feature selection methodologies consist of two parts: an algorithm that ranks features according to their information content and a strategy for determining the number of features to be used. However the FUR methodologies also consist of an additional two parts which rank features according to their trustworthiness and determine how many features will be used. Regardless of the methodology used, the same number of features is used by the model to be constructed.

Feature ranking within the traditional feature selection methodology and within the second phase of FUR1 and FUR2, is provided by Information Gain (InfoGain), an information theoretic method. The implementation of InfoGain used was provided by Weka (Hall, Frank, Holmes, Pfahringer, Reutemann, and Witten, 2009).

The following strategy was used for determining the number of highest ranking features to be selected for use in the model. Since comparing the three feature selection methodologies was the experimental goal, three feature set sizes were used throughout the experimentation: 16, 1 024 and 9 485. The smallest feature set size of 16 was selected to easily determine if differences exist between the features selected by each of the three methodologies. The next feature set size of 1,024 was chosen to determine if differences still exist when a significant proportion of the available features are selected. The largest feature set size of 9,485, representing all the available features, provides the limiting performance achieved from using all the features and associated noise.

The same measure of information content for a feature was used for all the feature selection methodologies, in order to facilitate their comparison.

4.1.2 Comparison

Two methods were used for comparing models: classification accuracy, measured by the number of correct classifications and changes in the list of features selected.

Model construction

Regardless of the feature selection methodology, the same method for constructing models was used. Because a non-classical problem domain applies,

a model learning algorithm that is well suited to few samples, high class imbalance and noisy data was chosen. A class imbalance occurs when a substantial difference exists in the number of data instances that correspond to each class label. As a result, AdaBoost, as implemented in Weka (Hall et al., 2009), was selected for the model learning algorithm.

Adaboost has two key characteristics: boosting and the ability to adjust the learning process in response to data instances that are incorrectly classified. The idea of combining a number of weak classifiers to produce a single strong classifier is known as "boosting". Boosting can prove valuable when the classification accuracy of individual classifiers is only slightly better than random guessing. Such classifiers are the product of weak learning (Schapire, 1990). Combinations of several weak learners can result in a single strong learner (Freund, 1995). One influence of weak learning is class imbalance. AdaBoost manages this by associating a weight with each data instance. Weights are continually adjusted during the iterative learning process, which concludes when an acceptable level of classification accuracy is achieved (Freund and Schapire, 1999, 1997). Instances that are misclassified have their weight increased, which results in increased focus by the learning process; hence the learning effort attributed to each instance is proportional to the magnitude of its weight.

The base classifier used with AdaBoost was Decision Stump, which was

implemented in Weka (Hall et al., 2009) and provided an entropy based classifier.

Model comparison

Model comparison evaluates differences in the classification accuracy of models constructed using each methodology. The goal is the detection of differences in model accuracy as a result of incorporating data quality within feature selection.

The construction of a classification model depends on successive subset selection, particularly in the case of the FUR methodology, which contains an additional subset selection. Using arbitrary numbers for illustrative purposes, consider for example an original list consisting of 900 features. The application of the first FUR selection phase results in a list of 600 features, according to trustworthiness. This list is further reduced to 400 features as a result of the second feature selection phase, according to discriminative ability. This list of 400 features will be referred to as the prior list. Finally a model using 100 features is constructed. Since the model only contains 100 features, it is conceivable that these features were, in essence, located throughout the prior list of features, rather than congregated around the first 100 or so features. This conceivable situation is due to the assumption of feature independence made by InfoGain, which does not consider feature

redundancy. So although a feature may be highly ranked, it may not provide any new information since it is redundant, which results in the feature being excluded from use in the model.

Feature selection comparison

The other approach to model comparison consists of evaluating differences in the rank of each feature as determined by each methodology. In the absence of noise, features will be ranked identically by each of the methodologies. However ranking differences are expected if noise is present. For example, a feature affected by significant noise should be ranked more lowly by FUR in comparison to the traditional methodology.

The motivation for this comparison is to determine whether a FUR methodology will allocate a lesser rank, in comparison to the traditional methodology, if the visual appearance of one or spots is indicative of a quality issue.

4.1.3 Assumptions and limitations

Determining the superiority of FUR over traditional feature selection is impacted by the following *assumptions* and *limitations* of the experimental design: sufficient redundancy, an appropriate threshold and an appropriate measure of data quality.

Sufficient redundancy

Redundancy refers to sets of features, in the absence of noise, which provide a similar capacity to discriminate between classes in a data set. Consider a scenario where each member of these feature sets is affected by a different amount of noise. The goal of FUR is to select the least noise-affected feature from each set.

Feature Utility Ranking consists of two feature selection phases: selection by trustworthiness, followed by information content. As a result of these two phases it is expected that the most trustworthy features, in each redundant subset, will be selected. However some redundant subsets will also be eliminated entirely since they fall below the minimum information content threshold.

Feature utility ranking is dependent on the presence of a sufficiently large number of redundant features. Traditional feature selection chooses features that are, in an "information content" sense, considered to be the most informative. However, FUR is able to provide a benefit if the features it selects are superior to those selected by a traditional approach. In this context, a feature providing equal information content is "superior" if it also provides increased trustworthiness. Consequently sufficient feature redundancy is a prerequisite for FUR being able to locate a more trustworthy substitute for

a given level of information content.

Appropriate trustworthiness threshold

The number of features removed in FUR's first feature selection phase determines whether the second phase searches through a trustworthy list. A trustworthiness threshold that is too high will result in the loss of information and therefore a model that may be inferior to one generated by traditional feature selection. A trustworthiness threshold that is too low risks the unnecessary incorporation of noisy features in the constructed model.

Since the goal of the experiments is the comparison of three feature selection methodologies, two of which use trustworthiness, a threshold that provides equal advantage to each of the methodologies is desired. Therefore the same threshold will be used in every instance and the selected threshold will be expected to work similarly regardless of the methodology, the data set used, or the number of features selected in the second selection phase. A trustworthiness threshold of 7,000 was determined empirically in section 3.7.1.

Appropriate measure of data quality

Two assumptions were made regarding the method used for calculating data quality. The first is that the measure of quality used quantifies some aspect of the accuracy of the information provided by the data. The second assumption

is that the aspect being measured accounts for a significant proportion of the noise in the data set.

4.2 Data sets

Three data sets were used: synthetic, leukaemia and chronic fatigue syndrome. Synthetic data provides a controlled test, where the amount of noise present is known. The leukaemia data set provides a real-world example of a non-classical problem domain. The chronic fatigue syndrome data set is also a real-world data set, but one that provides a poor level of correlation with class labels.

4.2.1 Synthetic data set

It is reasonable to expect that a model constructed from noiseless data would correctly classify instances in the training set, except perhaps, for instances on the decision boundary. It is also reasonable to expect that test data from a similar distribution to the training data to also be correctly classified. However with the addition of random noise, the classification accuracy becomes unpredictable if the nature of the noise is also unknown (Quinlan, 1986b). Consequently, the use of synthetic data provides data sets where parameters defining the noise are known. Also, the use of synthetic data provides the ability to compare changes in classifier accuracy, in response to varying

amounts of noise.

Data preparation

Synthetic data was constructed to partially replicate the characteristics of a non-classical problem domain. Although the noisy data set only consists of 100 features, its construction guarantees significant redundancy, variability of the predictive significance of each feature and variable influence of noise. Whilst the synthetic data set does not have a similar number of features as in real world non-classical data sets, such as in microarray data, important characteristics of the non-classical domain are still fulfilled. At the same time, the limitation of 100 features enables closer inspection of the differences in the features used by each feature selection method, which would have been more difficult if, for example, 10,000 features were used.

As shown in figure 4.1, two synthetic data sets were created, noiseless and noisy and each was divided into ten folds. These folds, which result in numerous training and test sets, are affected by the same known amount of noise. The noiseless synthetic data, which was used to construct the noisy version, was also used for experiments.

The noiseless synthetic data was constructed using

$$y = 1 + x_1 + 2x_2^2 + \ldots + 10x_{10}^{10} \tag{4.1}$$

where each x_i representing a feature, was randomly selected from a uniform

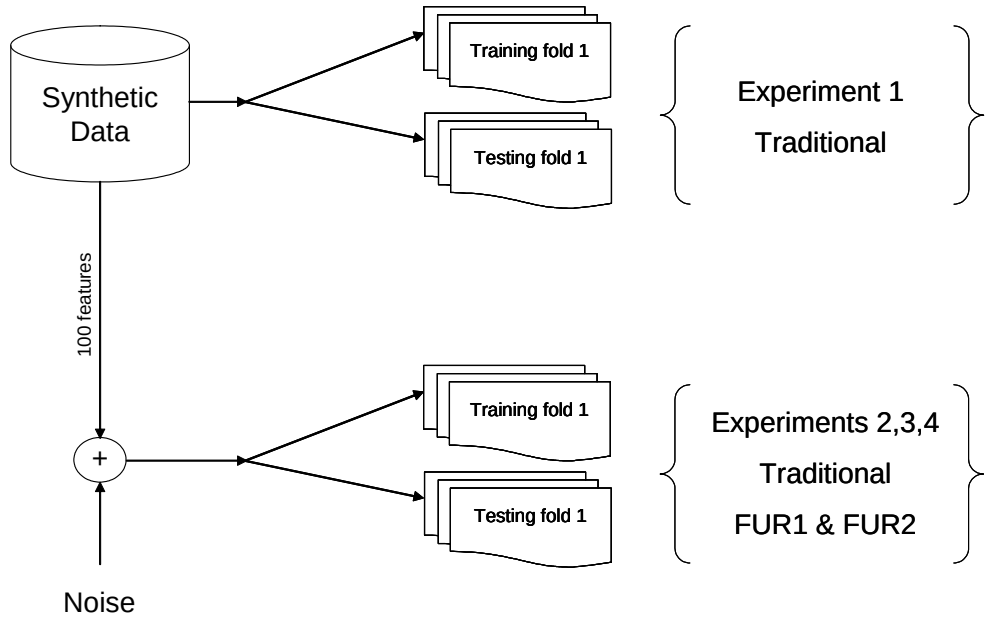


Figure 4.1: Synthetic data experiments consist of two data sets and four experiments. The noiseless data set is used with traditional feature selection. A noisy version of the noiseless data set is used to test all three feature selection methodologies: traditional, Feature Utility Ranking 1 and Feature Utility Ranking 2.

distribution ranging in $[0, 1]$. Given that each x_i in equation (4.1) has a different multiplier and exponent, the influence of each feature will be correspondingly different. This data set represents the underlying "signal" referred to as matrix $\mathbf{S}$ in equation (3.6). A data set consisting of 100 instances was generated.

The noisy data set in figure 4.1 was constructed by replacing each x_i with ten noisy versions of it. Hence the noisy data set contains a total of one hundred features,

$$\{a_1,\ a_2,\ \ldots\ a_{10},\ b_1,\ b_2,\ \ldots\ b_{10},\ \ldots\ j_1,\ j_2,\ \ldots\ j_{10}\} \tag{4.2}$$

which consists of ten subsets, each containing ten features. The ten subsets

$$\mathbf{a},\quad \mathbf{b},\quad \mathbf{c},\quad \mathbf{d},\quad \mathbf{e},\quad \mathbf{f},\quad \mathbf{g},\quad \mathbf{h},\quad \mathbf{i},\quad \mathbf{j} \tag{4.3}$$

are redundant versions of each other. For example $b_7, c_7, \ldots, j_7$ are redundant versions of feature a_7, where any differences are only due to noise.

The one hundred features, shown in (4.2), are calculated as follows. The symbol z_i representing one of the $a_i,\ b_i,\ \ldots\ j_i$ features, is calculated using the y and x_i values from the noiseless data set, as

$$z_i \quad = \quad \underbrace{x_i}_{\text{true value}} \quad \pm \quad \underbrace{w_{(j-1)\times 10\ +\ i} \quad \times \quad \epsilon}_{\text{noise}} \tag{4.4}$$

where j is the index of the feature subsets shown in (4.3). Whether the noise component is added or subtracted is randomly determined. For example, if

$j = 6$, then $z_i \equiv f_i$, which in the case of the ninth feature, is equivalent to

$$f_9 \quad = \quad x_9 \quad \pm \quad w_{59} \quad \times \quad \epsilon$$

The x_i parameter in equation (4.4) represents the true "signal" in $\mathbf{S}$, while z_i represents the measured sample data contained in $\mathbf{D}$, refer equation (3.7). The noise component in equation (4.4) represents $\mathbf{E}$ in equation (3.7) and is composed of two parts: noise that can be estimated ($\mathbf{E_{est}}$) and an unknown noise ($\mathbf{E_{unk}}$), refer equation (3.8).

Parameter w is used to simulate a constant bias, or a systematic error, such as found in a measuring instrument. For this reason, w is *constant* for each feature, for example, the same value for w_{59} is always applied to each instance of d_9. The value for w is randomly selected from a uniform distribution with a range of $[0, 1]$. The set of values used for each w are shown in figure 4.2. The w component in equation (4.4) is represented by $\mathbf{E_{est}}$, since w provides a constant error that should be detectable and manageable by either FUR1 or FUR2.

The value ϵ is also randomly selected from a uniform distribution with a range of $[0, 1]$, however the random selection occurs each time equation (4.4) is executed. In contrast to w, ϵ simulates an unpredictable perturbation that is more difficult to estimate and is best represented by $\mathbf{E_{unk}}$ in the case of FUR1.

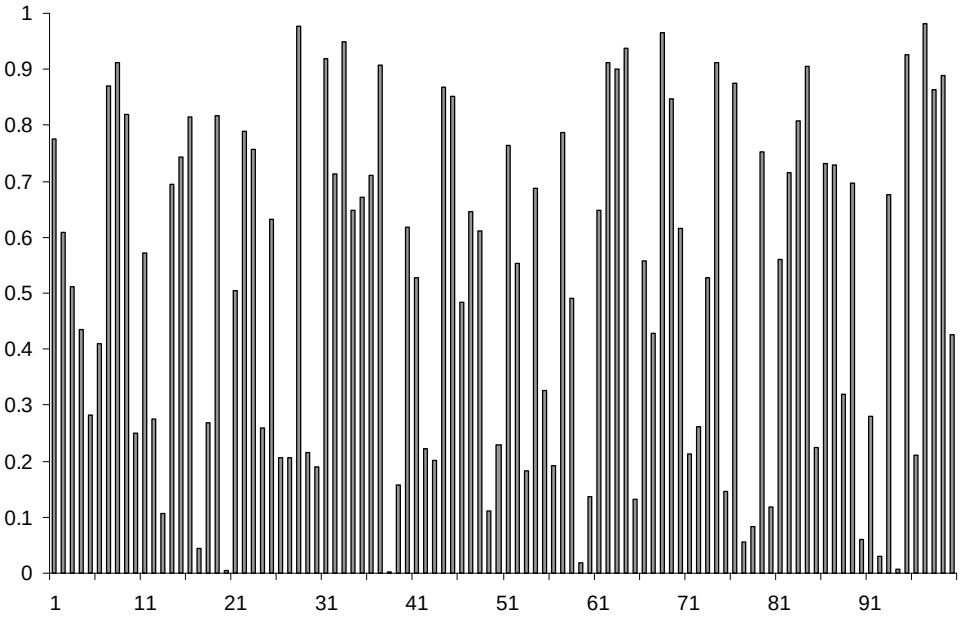


Figure 4.2: The distribution of w in equation (4.4), where the vertical and horizontal axes are w_k and k respectively and k ranges from $[1, 100]$.

Because w is a constant, its influence will be identical for the training and test data sets. As a result, FUR1 should effectively manage the impact of w, particularly considering its use of a signal-to-noise approach. In contrast, FUR2 is better placed to handle this error with respect to ϵ, since its impact could be different for the training and test sets. However in principle, ϵ could be managed more effectively by FUR1, if a measure of quality were able to determine bounds for the error instilled by ϵ, if the training set provided the same bounds as found in the test set.

4.2.2 Leukaemia data

Gene expression data, specifically collected for sub-typing Acute Lymphoblastic Leukaemia (ALL) patients, provides an example of a real-world data set. The clinical subtypes of this data set are: leukaemia cell type and treatment outcome. Two leukaemia cell types are considered: B-Cell and T-Cell leukaemia, refer section 2.1.3. Two treatment outcomes are considered: Complete Clinical Remission (CCR) and relapse. A patient is classified as attaining CCR if no relapse occurs within five years of completing treatment, otherwise a relapse classification is given.

The ALL expression data was provided by The Children's Hospital at Westmead in Sydney, Australia. Two-channel microarray data, consisting of 12,096 spots, was used to quantify the ALL gene expression and every sample

Table 4.1: Preprocessed childhood leukaemia data characteristics.

Classifier	Class	Arrays	Patients
immunophenotype	B ALL	120	47
	T ALL	44	14
outcome	CCR	115	58
	Relapsed	11	4

was processed within the same research laboratory. One array channel was constructed using bone marrow samples from a childhood leukaemia patient, while the other came from bone marrow samples from subjects that are clinically classified as normal. All of the leukaemia patients were subject to the same treatment protocol, namely BMC.

Data set characteristics

The characteristics of the leukaemia data set are summarized in table 4.1. The arrays column states the number of arrays for each patient subtype: B-Cell leukaemia, T-Cell leukaemia, CCR or relapse. The patients column states the number patients involved. In the case of B-Cell leukaemia, an average of 2.6 (or 120/47) replicate arrays were constructed from the same biological sample. Although the immunophenotype for each patient is known at the time of diagnosis, outcome is determined five years after the commencement of treatment. In the case of 38 patients, five years had not elapsed since commencing treatment, therefore fewer arrays were usable for classifying outcome that for immunophenotype classification.

Table 4.1 highlights two of the challenges provided by this real-world data set: few samples and a significant class imbalance, particularly for outcome classification. Another challenge is the curse of dimensionality, a 9,485 dimensional space populated with 126 samples, in the case of outcome classification. However, fewer than 126 samples are available for model training, since a proportion are required for validation and testing. With respect to class imbalance: a T-Cell immunophenotype is described by 44 sample points, or 27% of the samples; a relapse outcome is described by only 11 samples, or less than 9% of the samples.

Sources of raw data

Both clinical and microarray data was used to construct the sample data summarized in table 4.1. The clinical data source is used to construct two categorical output class features, as shown in figure 4.3. The first categorical feature uses the immunophenotype field to distinguish between cell types: B-ALL and T-ALL. The second categorical feature is constructed from the dateOfDeath field to establish whether the outcome was CCR or relapse; an empty field implies CCR. The microarray data source is used to construct numerical input features. These features are constructed from two data groups within the raw microarray dataset: expression and confidence.

The expression group provides features that are used for traditional fea-

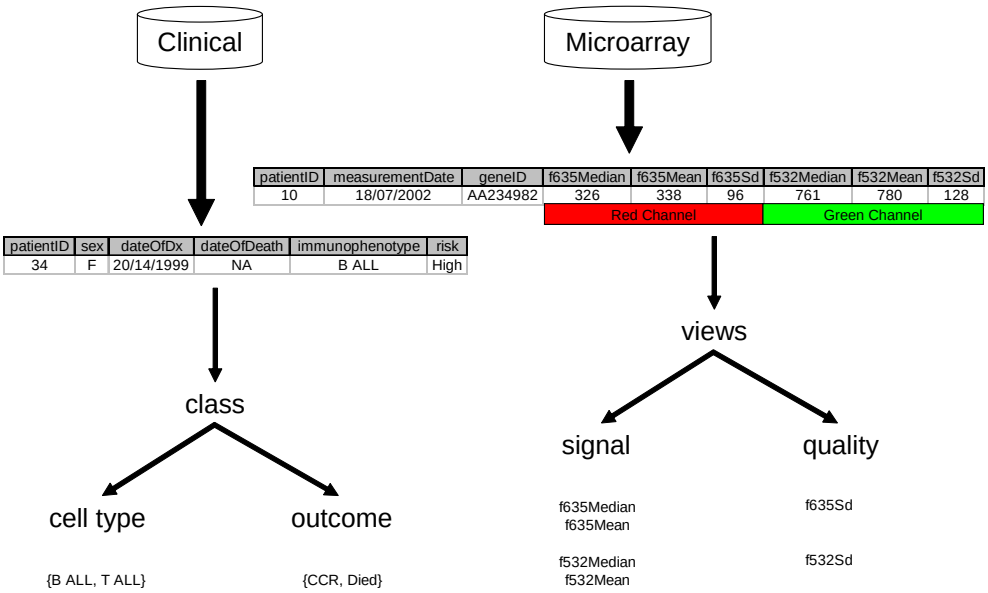


Figure 4.3: A subset of the raw childhood leukaemia biomedical data is shown. Each of the microarray spot channels provides the median and mean intensity for the spot and the standard deviation of pixel intensity within the spot. The subgroups *signal* and *quality* respectively refer to the expression readings and a measure of confidence in those readings.

ture selection. The median expression value is used to limit the influence of outliers. The f635Median and f532Median fields correspond respectively with the red and green channels of a microarray spot. A single expression feature for a spot is calculated by dividing the expression value of the test channel by the control channel. In the absence of dye swapping, the expression values of the test and control channels correspond to the f635Median and f532Median respectively. Some dye swap arrays exist in the raw microarray data set and their expression values were swapped.

Using FUR terminology, the confidence group corresponds to the raw microarray data used to determine a measure of "confidence" in the accuracy of an "expression" value. The two microarray data fields f635Sd and f532Sd are used as a measure of quality for the "test" and "control" channels respectively. Hence f635Sd and f532Sd are the entries within the noise vector $\mathbf{n_{ij}}$. The attributes f635Sd and f532Sd are the standard deviation of the pixel intensity readings for each channel of a microarray spot. In the case of FUR1 and FUR2, equations (3.9) and (3.10) respectively are used to calculate the confidence readings of each spot.

4.2.3 Chronic fatigue syndrome data

The second real-world data set, for chronic fatigue syndrome, consists of clinical and associated patient gene expression data. This data set was supplied

by the 2006 Critical Assessment of Microarray Data Analysis (CAMDA) Conference[1]. Using this data, patients were classified into one of two mutually exclusive classes: whether a patient did or did not have fatigue disease. For the purposes of this research, chronic fatigue syndrome and idiopathic fatigue disease were combined because of the limited sample size.

Given that biomedical expertise suggests that chronic fatigue syndrome is more likely to have a physiological, rather than genetic origin, the biological hypothesis tested was whether a genetic link exists. In testing this hypothesis it is assumed that the most appropriate genes are utilized and that the classification models being generated, as well that method used to generate them, are appropriate for discovering any genetic connection. Regardless of whether a genetic link exists with chronic fatigue syndrome, the underlying experimental goal is the comparison of traditional and FUR feature selection.

Data set characteristics

The characteristics of the CAMDA chronic fatigue syndrome data are summarized in table 4.2. Of 172 study members, 119 had been diagnosed as suffering from chronic fatigue syndrome, while 53 provided the non-fatigued control group. Single channel microarray technology was used, with a single array per study member and a total of 20,160 spots (or genes) printed on

[1]Further details about the 2006 CAMDA Conference are located at http://www.camda.duke.edu/camda06 and http://aboutmecfs.org/Rsrch/Camda06.aspx

Table 4.2: Preprocessed fatigue data characteristics.

Class	Patients
Fatigue	119
Non-fatigue	53

each array.

Table 4.2 shows similar characteristics to the leukaemia data described in section 4.2.2. There are very few samples and an even more significant curse of dimensionality issue, since the dimensionality has doubled. A class imbalance also exists, as 69% and 31% of samples represent the respective fatigued and non-fatigued classes.

Raw data sources

As with the leukaemia data, two data sources (clinical and microarray) were used to construct the sample data summarized in table 4.2. From the original data source, a field called "Intake Classific", which contains five categories, is used to define a binary feature: "Fatigued" and "Non-fatigued". Four of the "Intake Classific" categories (Ever CFS, Ever CFS-MDDm, Ever ISF, Ever IFS-MDDm) were subsumed into the fatigued class. As with the leukaemia data set, the expression and confidence subgroups were also constructed. The expression subgroup contains the "ARM Dens" field, which provides an artifact removed average density of the individual pixel densities. Unlike the leukaemia data set, the confidence subgroup contains a single field, SD, which

provides the standard deviation of the pixel density.

For the CAMDA data set, the FUR1 and FUR2 confidence equations (3.9) and (3.10) are set as follows. Expression value d_{ij} corresponds to ARM Dens and the noise vector contains a single entry since single channel microarray technology was used. Therefore equation parameters for σ_c and σ_t are set to the single standard deviation value SD. Parameter settings for Algorithm 1 and Algorithm 2 follows similarly.

4.3 Data preparation

Three major preprocessing steps were used for preparing the leukaemia and chronic fatigue syndrome data sets prior to feature selection and model construction: sample selection, cleaning and normalization.

4.3.1 Sample selection

Sample selection consists of four parts: patient selection, array selection, collecting dye-swap information and gene subset selection.

Constructing a predictive model requires selection of a list of patients with a known outcome. In the case of leukaemia, defining outcome requires a minimum of five years since completing treatment.

Array selection involves compiling a list of microarrays associated with the patient list and whether dye-swapping occurred. Gene subset selection

identifies which genes are present on all arrays. In addition, "technical control spots" are excluded during gene subset selection.

4.3.2 Cleaning

The second preprocessing step is cleaning, which involves detecting and repairing outliers. Outliers are intensity readings that are out of character with the majority. That is, any reading that may have been corrupted by the limitations of the measuring instrument, which in the context of microarrays, is a laser scanner.

Spot intensity readings are collected using an unsigned 16 bit integer, giving a measurable range of [0, 65 535]. A reading that corresponds to the minimum or maximum value results in information loss, since the true intensity reading is unknown. It is also possible that the sensitivity of the scanner may deteriorate as the minimum or maximum reading is approached. Therefore a reduced measurable range was deemed appropriate, beyond which any value would be considered an outlier.

This range was deduced empirically as follows. It was assumed that the intensity readings are normally distributed and any readings which deviated from this were outliers. With respect to the left and right tails of the distribution, in the case of the leukaemia data set, a total of 7 and 370 readings respectively had clearly deviated from the norm. Using these outliers, it was

concluded that the measurable range is [10, 65 000] and any readings outside this range are outliers. The same process was followed for the chronic fatigue syndrome data set.

Outliers were repaired by replacing them with the mean value of the corresponding gene for patients of the same class. This involved examining individual intensity readings for a gene. They were subdivided into two groups corresponding to the class, for example CCR and relapse. The mean expression reading was calculated for each subgroup, excluding readings that were determined to be outliers. Finally, the mean expression value was set to each outlier belonging to the same subgroup.

This description of cleaning highlights a fundamental problem with repairing outliers—that expression values are made to be artificially consistent with non-repaired values. A traditional feature selection methodology is unable to detect this and is therefore unable to take it into account in the selection process. However, FUR is able to adjust the selection process, since quality issues with repaired expression values are expected to be discernable by their measured variation of pixel intensity.

4.3.3 Normalization

Normalization is the process of adjusting the expression levels of every microarray spot, so that their values are directly comparable. All sample data

in the **D** matrix requires normalization.

This research uses a global baseline normalization method for the biomedical data sets, hence the hypothesis used was that every array should have the same average intensity. A normalization factor, consisting of $\frac{global\ average}{local\ average}$ was used to adjust each expression reading on each array. The average intensity of every spot on every array is the *global average*, while the average intensity for every spot on a single array is the *local average*.

In some cases, confidence values also require normalization. Spot intensity values, for example, should be normalized. Confidence values correspond to the data found in the **C** matrix. However, normalization may not be appropriate for confidence values that measure the consistency of a spot based on, for example, the variation of individual pixel intensities. In the case of this research, all intensity values, whether obtained from the **D** or **C** matrices, were normalized, while pixel intensity variation data was not.

4.3.4 Prepared data characteristics

The result of preparing the leukaemia and chronic fatigue syndrome data sets are described in this section.

Leukaemia

The leukaemia data set was generated from two-channel microarrays consisting of 12,096 spots, however, some spots were given the same gene name,

resulting in a total of 9,554 unique spot names.

Data preparation established that some genes are not present on every microarray and that the number of replicates of each gene varies for each microarray. As a result, it was determined that 9,485 genes were present on every microarray in the B versus T-Cell and treatment outcome data sets.

Chronic fatigue syndrome

The chronic fatigue syndrome data set was prepared using the same approach as for the leukaemia data set. The chronic fatigue syndrome data set was generated from single-channel microarrays and preprocessing determined that a total of 20,160 unique genes were located on every array.

4.4 Experiments using synthetic data

Three feature selection methodologies were evaluated with experiments using synthetic data: traditional feature selection, which does not consider data quality; feature utility ranking 1 and feature utility ranking 2.

Although a binary classifier is used for the purposes of the real-world data set experiments, FUR can also be used for the construction of regression models. To show this and because regression is considered a more demanding test with respect to accuracy, regression was used. The approach consists of four experiments: the first uses a noiseless version of the synthetic data

set, while the remaining three use a noisy version of the data, as shown in figure 4.1.

Regression provides a method for modeling a continuous target value. Support Vector Regression was used together with Sequential Minimization Optimization, or SMOreg (Smola and Scholkopf, 1998; Shevade, Keerthi, Bhattacharyya, and Murthy, 1999). The implementation of SMOreg used was provided by WEKA (Hall et al., 2009) and default values were used, except for an exponent of the polynomial kernel of three. Model evaluation consisted of stratified ten-fold cross validation.

The first experiment, labeled "experiment 1" in figure 4.1, involved constructing a regression model using all features and a noiseless version of the synthetic data set. This test provides an "idealized" reference for all models, since this model is constructed with noiseless data. The second experiment also involved all features and the noisy version of the synthetic data set. The third and fourth experiments used the same noisy synthetic data set, although a FUR trustworthiness threshold of 0.7 was used. In the case of FUR1, this threshold resulted, on average, in the removal of 10.6 features per fold.

The confidence value used by FUR, for calculating trustworthiness, was the known noise value w in equation (4.4). As a result, trustworthiness was calculated using the mean of the set of $signal \times confidence$ values, where

Table 4.3: The results of the synthetic data experiments, where *mean* and *std dev* are the mean and the standard deviation of the correlation coefficients for ten-fold cross validation.

Fold	Noiseless Traditional	Noisy Traditional	Noisy FUR1	Noisy FUR2
1	0.9189	-0.5105	-0.5222	-0.4305
2	0.9289	-0.0881	-0.0860	-0.2123
3	0.9201	-0.2602	-0.4630	-0.5343
4	0.8729	0.2874	-0.0888	-0.0321
5	0.9420	0.0365	-0.0237	-0.0785
6	0.9746	0.0281	0.0750	-0.1497
7	0.9998	-0.6137	-0.4185	-0.3447
8	0.7840	-0.0406	0.0527	0.0946
9	0.8256	-0.0293	-0.0413	0.0973
10	0.9119	0.3801	0.1610	0.5286
mean	0.9079	-0.0810	-0.1355	-0.1062
std dev	0.0618	0.2964	0.2303	0.2915

$confidence = w$ and for example, $signal = f_i$.

4.4.1 Experimental results

The results for the synthetic data experiments are shown in table 4.3, which contains the Pearson correlation coefficient between the predicted and true y values of equation (4.1) for each fold. The last two rows of the table show the mean and standard deviation for each experiment or column. The difference in mean correlation between the noiseless and noisy traditional regression experiments is indicative of the amount of noise added, refer figure 4.2.

Table 4.3 contains a number of negative correlation coefficients, which indicate an inverse relationship between the predicted and true y values. The

presence of negative correlation coefficients results in a canceling effect when the mean of the coefficients is calculated, hence sending the experimental mean toward zero. Taking the absolute value of the individual correlation values for each fold does not result in a very different outcome. An analysis of using the absolute value is presented in Appendix A.

A hypothesis test (single tailed t-test) was constructed to determine whether the mean correlation coefficient for pairs of experiments are statistically different. The hypothesis test used is

$H_0 : mean_1 = mean_2$

$H_1 : mean_1 < mean_2$

and the results are shown in table 4.4. This table shows the p-value for permutations of experimental pairs, for example, the alternative hypothesis (H_1) that FUR1 is statistically different to Noisy Traditional can only be accepted with 33.14% confidence. Other than for the first three rows, the null hypothesis (H_0) must be accepted that no statistical difference exists.

As described in chapter 3, noise can result in apparent information content, which causes some of a model's accuracy through chance. Hence to some extent, noise can serendipitously be consistent with the underlying structure of the phenomenon being modeled. Some of the accuracy, illustrated in table 4.3, may be due to apparent information content. Nonetheless, FUR1 outperformed FUR2 and FUR2 outperformed the traditional methodology,

Table 4.4: Using the *mean* correlation confidences in table 4.3, the p-value for permutations of the means are shown. The column labeled "Sig", when populated with an $*$, states the result is statistically significant. The last row for example, shows that the alternate hypothesis (H_1) that FUR2 is more accurate than FUR1 can only be accepted with 18.45% confidence.

$mean_1$	$mean_2$	p-value	Sig
Noiseless Traditional	Noisy Traditional	≈ 0	*
Noiseless Traditional	Noisy FUR1	≈ 0	*
Noiseless Traditional	Noisy FUR2	≈ 0	*
Noisy Traditional	Noisy FUR1	0.669	
Noisy Traditional	Noisy FUR2	0.858	
Noisy FUR1	Noisy FUR2	0.816	

although the differences are not statistically significant. In addition, the standard deviation of the means was least for FUR1, followed by FUR2 and then traditional, suggesting that FUR1 provided the most consistent results.

The synthetic data experiments provided an idealized scenario, in that much of the noise was known to FUR, through w in equation (4.4). Consequently, any shortcoming is expected to either be due to the method used for calculating a feature's utility, or an unmanageable consequence of noise.

That FUR1 outperformed FUR2, albeit marginally, was unexpected since FUR2 also evaluates the quality of the test set, which is considered to be important given the influence of the ϵ noise component. This outcome could be explained if the influence of ϵ is similar in both data sets. Alternatively, the impact of the correlation coefficient's sign may be responsible for the unexpected result, as suggested in the appendix.

Overall, these results suggest a possible difficulty in setting an appropriate trustworthiness threshold, for FUR1 and FUR2: too low a threshold would fail to filter out features affected by significant noise, too high a threshold would result in too much information loss. Setting the right threshold may require a search method, such as a Genetic Algorithm. Nonetheless, table 4.4 suggests that FUR1 and FUR2 did offer a benefit, although marginally.

4.5 Leukaemia data experiments

Two aspects of leukaemia were tested, firstly the discrimination between B-Cell and T-Cell Leukaemia and secondly discrimination between Complete Clinical Remission (CCR) and a death outcome.

4.5.1 Traditional feature selection

The results of the traditional feature selection experiments consist of: the ten highest ranking features, the classification accuracy of three different models, an assessment of the degree of redundancy present within the data sets and contrasting the B-Cell versus T-Cell feature ranks with the treatment outcome features.

B versus T-Cell

The ten highest ranking features for B-Cell versus T-Cell classification are shown in table 4.5. The third and fourth columns of the table, respectively,

contain the *Precision* and *Recall* values obtained for a classifier that uses the top i ranked features. For example, the *Precision* and *Recall* values shown for the feature ranked third are the result of a model built using the top three ranked features. For the results shown in table 4.5, between three and four classification errors occurred for each of the ten classification models, which shows that high accuracy was obtained using only a few features. Although feature trustworthiness is not part of traditional feature selection, the resulting trustworthiness rank for each feature is included in the last column. A rank of one and a rank of 9,514 correspond to the most and least trusted features respectively. Since the trustworthiness ranks indicate that the features in the table are amongst the 12% least trusted, the model's high accuracy may in part be due to the effects of noise.

A hierarchical cluster analysis, or heat map, of the same data is shown in figure 4.4. The red and green rectangle directly above the heat map depicts the relative number of B-Cell versus T-Cell samples, where B-Cells form the majority. Each row of the heat map corresponds to a feature (gene), while each column corresponds to a microarray, where a patient can be represented by more than one array. The color located in the intersection of a row and column corresponds to the relative change in gene expression for that patient. The baseline for the relative change is derived from a pool of patients deemed to be "normal ". Visually, the heat map shows a weak separation between

Table 4.5: The ten highest ranking features for the B versus T-Cell data set, according to traditional feature selection. The Feature Utility Ranking 1 value of trustworthiness is shown in the last column for comparison purposes.

Rank	Gene ID	*Precision*	*Recall*	Trustworthiness
1	AA055946	0.976	0.976	8,752
2	T64192	0.982	0.982	8,866
3	AA420981	0.982	0.982	9,056
4	N91921	0.976	0.976	9,328
5	AA469965	0.988	0.988	8,255
6	AA775257	0.988	0.988	9,079
7	AA481988	0.982	0.982	8,388
8	AA283629	0.982	0.982	8,886
9	AA243694	0.982	0.982	8,631
10	AW009739	0.983	0.982	8,776

the classes. Note that all the features contained in the heat map fell below the trustworthiness threshold, which may provide some explanation for the weak separation.

Although trustworthiness is not used by traditional feature selection, the FUR1 trustworthiness of the features is included in table 4.5 for reference purposes. Given that the total number of features is 9,485, every feature shown in the table is exceedingly untrustworthy. Although all the features shown in the table were selected by traditional feature selection, the use of a trustworthiness threshold of 7,000 prevented the use of any of those features by Feature Utility Ranking 1. Therefore according to trustworthiness, the features used by the traditional methodology were significantly affected by noise. Because feature N91921, with a trustworthiness rank of 9,328, is adja-

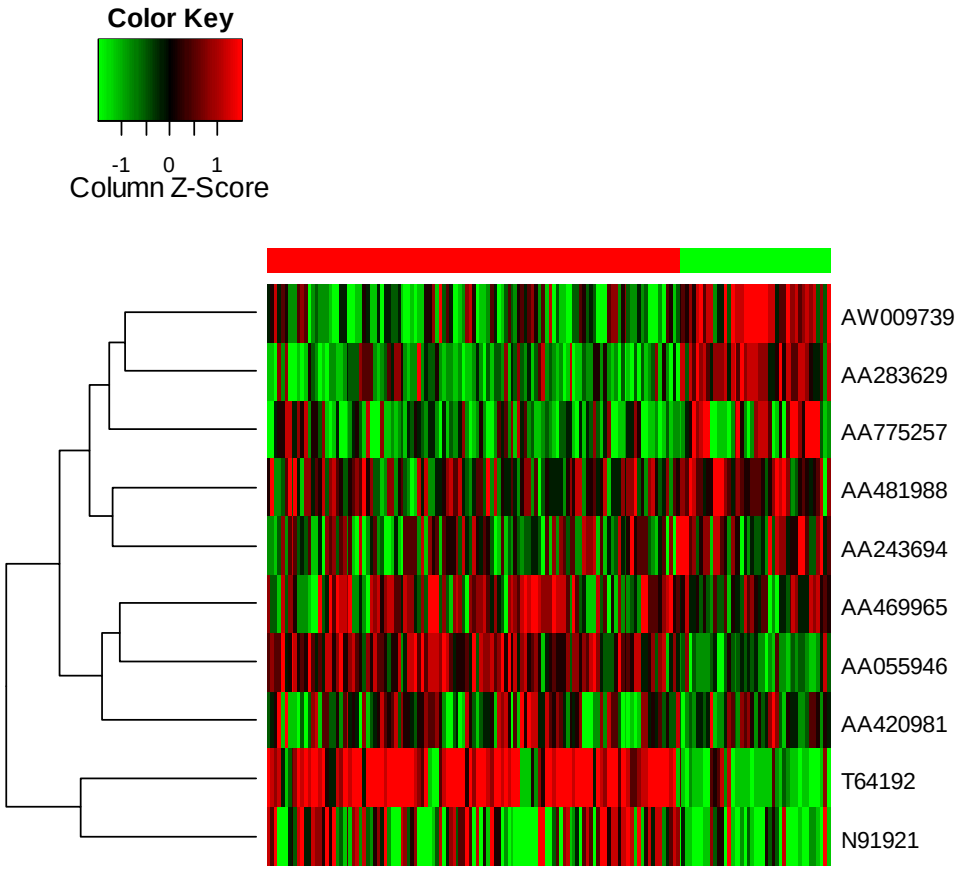


Figure 4.4: Hierarchical cluster analysis of traditional feature selection for the B versus T-Cell data set. All the features fell below the trustworthiness threshold.

Table 4.6: Classification accuracy of the B versus T-Cell data set, resulting from traditional feature selection.

Features	Traditional	*Precision*	*Recall*
16	96.97%	0.982	0.982
1,024	90.91%	0.982	0.982
9,485	98.78%	0.988	0.988

cent to the right vertical segment of figure 4.8, its trustworthiness is located with features that are of distinctly poor quality. However traditional feature selection ranked this feature as fourth with respect to information content.

Again using traditional feature selection, the experimental classification models containing 16, 1 024 and all features were built. The accuracy of these models, shown in table 4.6, reflects the strong correlation that exists between genetics and B versus T-Cell leukaemia. The *Precision* and *Recall* columns in the table reveal a model accuracy that is unbiased. Given the trustworthiness of the ten highest ranking features used in the 1,024 feature model and particularly the 16 feature model, the measured accuracy of those models is expected to be untrustworthy. Note that 9,485 is the total available features and that the figures quoted are stratified ten-fold cross validation averages.

One of the prerequisites of the FUR methodology is the presence of significant feature redundancy, so the next experiment aimed to measure redundancy. The approach consisted of iteratively building classifiers using fewer

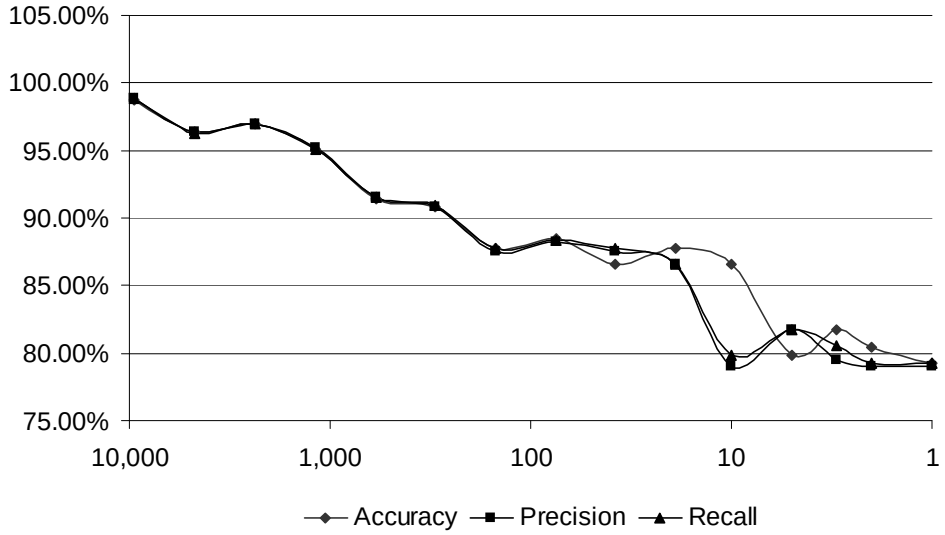


Figure 4.5: An indication of feature redundancy within the B versus T-Cell data set. The number of features used by the model and the resulting accuracy are shown on the x-axis and y-axis respectively. The features removed were randomly selected.

features. It began with the entire feature set (9,485), constructed a model and tested its accuracy. This process was repeated after randomly removing half the features. Results are shown in figure 4.5, which includes *Precision* and *Recall*, which demonstrates that accuracy was not caused by any class bias.

This result is indicative of substantial redundancy, since the classification accuracy, overall, reduced gradually from approximately 99% to 79% rather than abruptly, which would be expected if little redundancy existed.

Treatment outcome

The ten most highly ranked features of the treatment outcome data set are shown in table 4.7, ranked by information content. As suggested by the *Precision* and *Recall* values, accuracy was largely due to a bias toward the majority class (CCR), thus suggesting a poorer link between genetics and outcome when compared with the B versus T-Cell classification problem. The FUR1 trustworthiness is not used by the traditional methodology and is only shown in table 4.7 for comparison. In contrast to the B versus T-Cell data set, the top ten ranked features for the outcome data set are more trustworthy; only five features would have been rejected by FUR1. In addition, some of the features listed in the table are quite trustworthy, for example those ranked 4 and 2. A hierarchical cluster analysis of the same data is shown in figure 4.6. Note the red and green rectangle above the heat map shows the relative number of CCR and relapse outcomes, where CCR forms the majority. The features labeled with an '*' fell below the trustworthiness threshold. The class division within the heat map is difficult to determine, possibly because of the significant class imbalance, or because of the complexity of the mechanisms that determined treatment outcome.

Due to the complexity of the mechanisms that determine treatment outcome, there is expected to be far less redundancy in the data set, compared

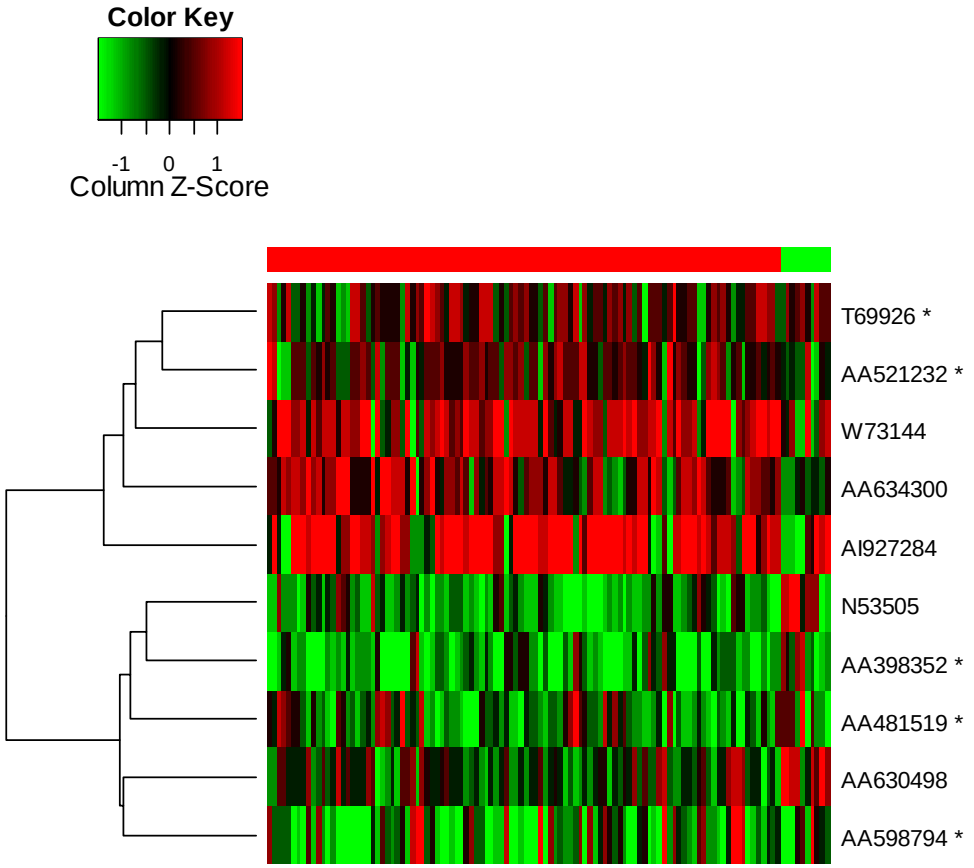


Figure 4.6: Hierarchical cluster analysis of traditional feature selection for the treatment outcome data set. Features marked with a '*' fell below the trustworthiness threshold, refer table 4.7.

Table 4.7: Using traditional feature selection and the treatment outcome data set, the ten highest ranking features for the first training fold are shown, together with the classification accuracy of the top i ranked features. FUR1 trustworthiness is only shown for comparison purposes.

Rank	Gene ID	*Precision*	*Recall*	Trustworthiness
1	AA598794	0.930	0.937	9,345
2	AA630498	0.909	0.929	1,541
3	N53505	0.909	0.929	2,511
4	AI927284	0.918	0.929	485
5	AA634300	0.918	0.929	6,192
6	T69926	0.921	0.907	7,066
7	AA481519	0.926	0.929	8,571
8	W73144	0.926	0.929	4,247
9	AA398352	0.914	0.921	9,286
10	AA521232	0.926	0.929	7,013

to the B versus T-Cell data set. This is evident in figure 4.7, since the approximate accuracy of the resulting models, over the course of continually halving the size of the feature set, fluctuated between 82% and 92%, although in one instance it dropped to about 66%. This overall behavior suggests that artifacts were mainly responsible for the achieved accuracy. On the basis of trustworthiness, table 4.7 states that all but one AI927284 of the traditionally selected features were associated with poor quality data. In addition these results do not provide any evidence for the existence of significant feature redundancy, which agrees with our understanding of the disease.

None of the same genes were listed in both top ten ranked features of the B versus T-Cell and treatment outcome classifiers. Within the top ten

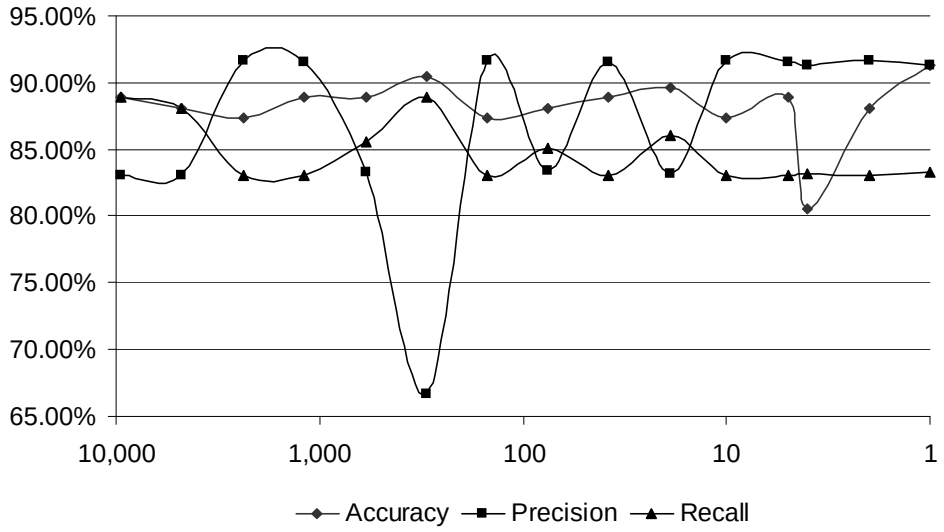


Figure 4.7: An indication of feature redundancy in the leukaemia treatment outcome data set, which was generated by repeatedly halving the feature set size; the features removed were randomly selected.

ranked features of the outcome classifier, the closest commonality occurred for the seventh feature, which matched the five hundred and thirty ninth ranked B versus T-Cell feature.

As shown in table 4.8, the treatment outcome classifier performed similarly to the B versus T-Cell model in terms of accuracy. Although this suggests a strong correlation between genetics and treatment outcome, closer inspection reveals a strong bias toward the majority class (CCR), which is supported by *Recall* being greater than *Precision*. On average and with respect to individual folds, relapse was incorrectly classified more often than correctly, which is not surprising given a class imbalance of 115 CCR to 11

Table 4.8: Classification accuracy for a treatment outcome classifier, where the feature ranks was determined by a Traditional Feature Selection. The figures quoted are stratified ten-fold cross validation averages.

Features	Traditional	*Precision*	*Recall*
16	92.06%	0.914	0.921
1,024	96.00%	0.889	0.905
9,485	88.89%	0.831	0.889

relapse, as shown in table 4.1 on page 151. This suggests a poor level of correlation between genetics and treatment outcome.

Discussion

Using a traditional feature selection methodology, quite different results were achieved for the Leukaemia B-Cell versus T-Cell and treatment outcome data sets. A very high classification accuracy was achieved for the B versus T-Cell data set and it was suggested that this accuracy was not due to the class imbalance that exists. The existence of substantial feature redundancy was confirmed for this data set, which may be exploited by FUR. However, whilst substantial redundancy exists, it may prove difficult for FUR to improve upon the high accuracy already achieved by the traditional approach.

The traditional feature selection methodology was unable to produce a genuinely accurate classifier using the treatment outcome data set. What appeared to be high classification accuracy, may be due to the significant class imbalance in the data sets. There did not appear to be significant

feature redundancy, but that may be due to the significant class imbalance.

4.5.2 Feature Utility Ranking 1

Feature Utility Ranking 1 was also applied to the Childhood Leukaemia data set.

B versus T cell

Using the **C** matrix view of the training data set, the first phase of FUR1 calculated the trustworthiness of every feature. The trustworthiness curve as shown in figure 4.8 has the same characteristic shape as the other experiments. The curve was constructed by ranking all the features (or genes) according to their trustworthiness. Using the same threshold as described in section 3.7.1, all genes ranked lower than 7,000 were removed from further consideration.

The trustworthiness curve shown in figure 4.8 consists of three major segments: the left vertical segment, the relatively flat middle segment and the right vertical segment. The left vertical segment, which contains a small number of features that are the most trustworthy, is associated with the greatest rate of change in trustworthiness. The turning point between the left and middle segments emphasizes that the left segment is "out of character" to the majority of features. This segment may be out of character since its features are unusually trustworthy or because of a shortcoming in the

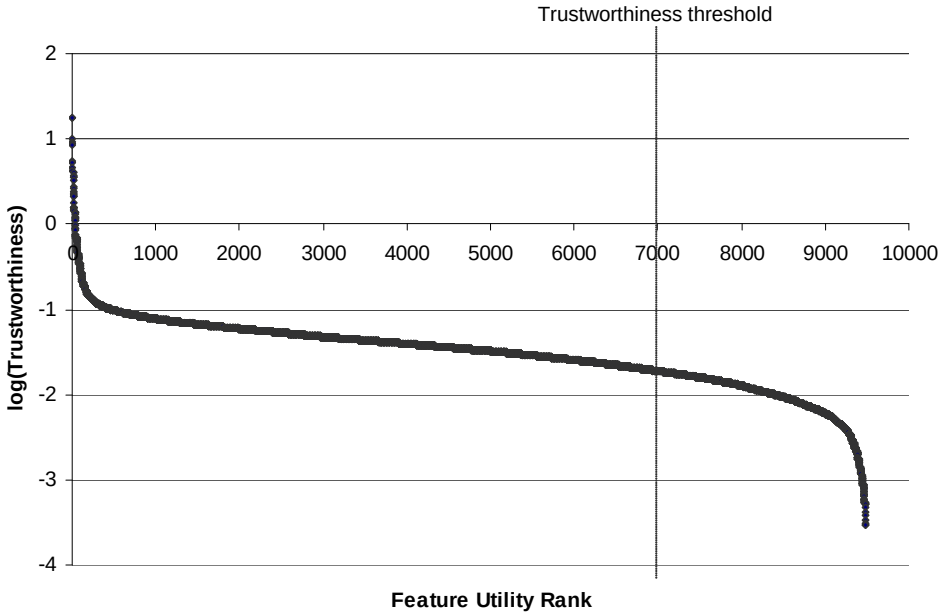


Figure 4.8: Feature Utility Ranking 1 trustworthiness for the B versus T-Cell data set.

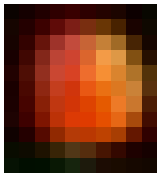


Figure 4.9: Example of a poor quality microarray spot, which belongs to a feature that fell below the trustworthiness threshold.

measure of quality used. The latter cause appears plausible, since very high expression can be associated with very little pixel variation as the intensity has become saturated. Spots have been found which have little pixel variation as the intensity has become saturated, an example exhibiting an unusually bright color is shown in figure 4.9. This spot, which belongs to a feature that fell below the trustworthiness threshold, has an unusual appearance. Other examples of poor spot quality are shown in figure 4.10, which comes from a research paper on microarray image quality (Guo, Cutri, and Catchpoole, 2004).

The middle segment of figure 4.8, contains the vast majority of features and is characterized by a gradual rate of change in trustworthiness. The right vertical segment contains few features and consists of a rapid rate of change. The right vertical segment is composed of untrustworthy features, rather than some shortcoming in the calculation of quality.

Figure 4.11 shows the expression ratio and calculated quality, or composite pixel deviation, of each sample point for the most trustworthy gene

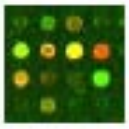
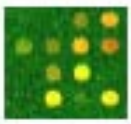
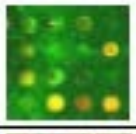
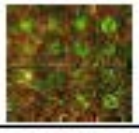
Category	Characteristic	Quality Assessment	Number	Example
A	bright spot with minimal background	complete reliable data	22	
B	bright spot and limited background	reliable data	20	
C	faint spot, high background or presence of minor artefacts	partial reliable data	13	
D	failed hybridization, excessive artefact	no reliable data	5	

Figure 4.10: Examples of varying microarray spot quality.

according to the first training fold. This fold, derived from 165 arrays, shows a small number of arrays deviated significantly from the norm, which is indicated by the peaks in the expression ratio curve adjacent to array indexes 49 to 61. Investigation into why these arrays were different revealed a correlation with the date the biological sample was bound to the array, as shown in table 4.9. No other arrays were located in this date range, except for those shown in the table. This correlation suggests two possible causes: the arrays themselves were faulty or their construction deviated in some way from the norm. The biological sample itself is not considered responsible for this deviation since the same sample appeared on five other arrays within this fold

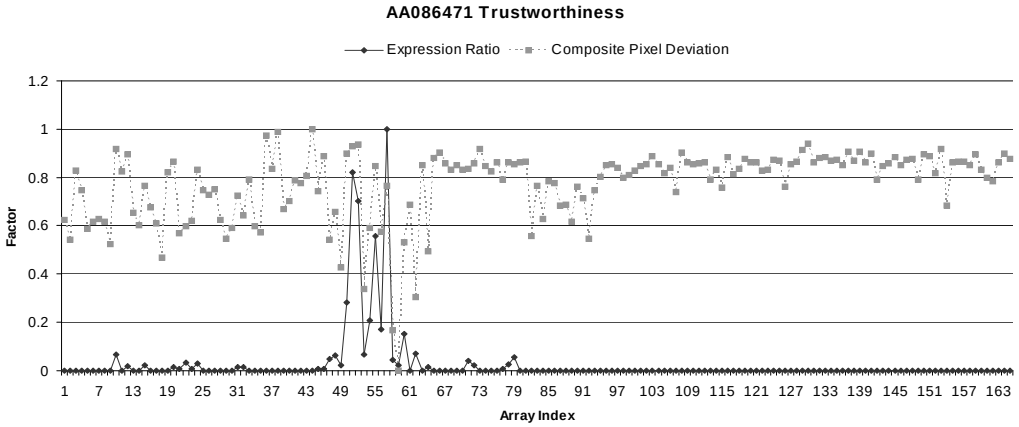


Figure 4.11: Characteristics of the *most* trustworthy gene AA086471 according to the first training fold and Feature Utility Ranking 1.

Table 4.9: The *most* trusted gene AA086471 within the first training fold revealed a correlation with array construction dates for six of the arrays.

Patient ID	Array ID	Construction Date
81	b	17/9/2003
82	b	17/9/2003
81	c	24/9/2003
37	b	30/9/2003
80	c	30/9/2003
37	c	2/10/2003

and all exhibited characteristics consistent with the majority. Note that the two curves in figure 4.11 were normalized so that they could be displayed together.

The standard deviation of pixel intensity for both the test and control channels, namely composite pixel deviation, is also shown in figure 4.11. For the most part, the variation in composite pixel deviation ranges between 0.4

and 1.0. Therefore the majority of spots possessed higher deviation than the eight arrays located in the array index range of 49 to 61, hence the implication that the majority of spots for AA086471 were of comparatively poorer quality. The exception to this behavior is the arrays associated with a higher than characteristic expression ratio.

The premise used is that pixel deviation is a measure of spot quality, specifically there is an inverse relationship between pixel deviation and spot quality. However when the intensity of a spot reaches the scanner's upper measurable limit, pixel saturation occurs and the inverse relationship breaks down. A correspondingly similar problem exists when the scanner's lower limit of measurability is reached. Clearly some of the data outliers evaded the cleaning process described in section 4.3.2. Either the data cleaning process or the measure of quality requires improvement. Suggestions for an improved measure of data quality are provided in further work in section 5.4. In light of the new relationship between spot intensity and pixel deviation when the limits of the scanner are reached, it clear that the majority of composite pixel deviation readings are actually normal.

It is noteworthy that the small peaks in the expression ratio curve of figure 4.11 correspond to T-Cell leukaemia, while the remaining sample points correspond to B-Cell. The data points located above 0.2 on the vertical axis are considered outliers. The gene in this figure AA086471, was given

a ranking of 993 by traditional feature selection and therefore was not used within the 16 feature model.

Figure 4.11 confirms that FUR1 did balance signal magnitude against signal noise in terms of expression ratio against pixel variation, except when pixel saturation occurs. As can be seen in most circumstances, peaks in expression ratio were associated with dips in composite pixel deviation, while lower expression ratio values were associated with higher pixel deviation. Therefore, on the basis of this figure, FUR1 did successfully employ a signal-to-noise approach for calculating a measure of quality for this feature.

A gene whose trustworthiness lies within the normal / majority portion of figure 4.8 will now be considered. The characteristic behavior of the $2{,}500^{th}$ most trustworthy gene, AI017417, is shown in figure 4.12. In comparison to figure 4.11, this gene contains nine arrays whose expression ratio instances lie in excess of 20% average expression in comparison to the six for AA086471. The composite pixel deviation curve has a lower offset, although its variations are not entirely dissimilar to that shown in figure 4.11. If composite pixel deviation is a valid measure of quality, figure 4.12 suggests that lower composite pixel deviation can compensate for an increased number of expression ratio outliers.

Comparing figures 4.12 and 4.11 suggests that trustworthiness is sensitive to changes in expression ratio and composite pixel deviation, since the differ-

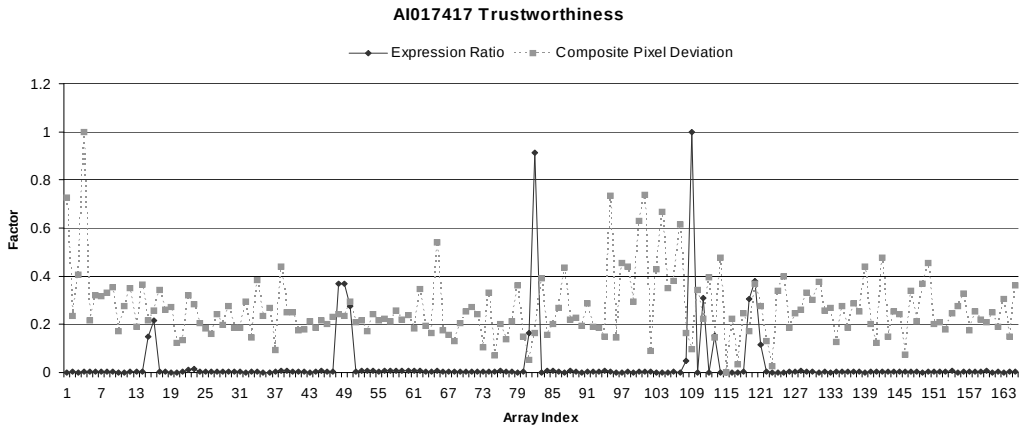


Figure 4.12: The 2,500th most trustworthy gene AI017417 for the first training fold, using Feature Utility Ranking 1.

ence in rank is 2,500 and few changes have occurred with respect to quality. This gene was ranked 8,781 by traditional feature selection.

As a final comparison, the *least* trustworthy gene AA873060 in the first training fold is shown in figure 4.13. Additional outliers are shown in the figure, thirteen in total, for which many possess higher amplitude than seen for the previous two genes. Again using the previous genes as a reference, figure 4.13 shows a more erratic composite pixel deviation curve. This gene was ranked 5,674 by traditional feature selection.

Because ten-fold cross validation is used, the trustworthiness of each gene may vary from fold-to-fold. This variability across the training folds provides a measure of quality for the training data and is also indicative of the quality of the test data. A simple test of variation is used to assess how a gene's

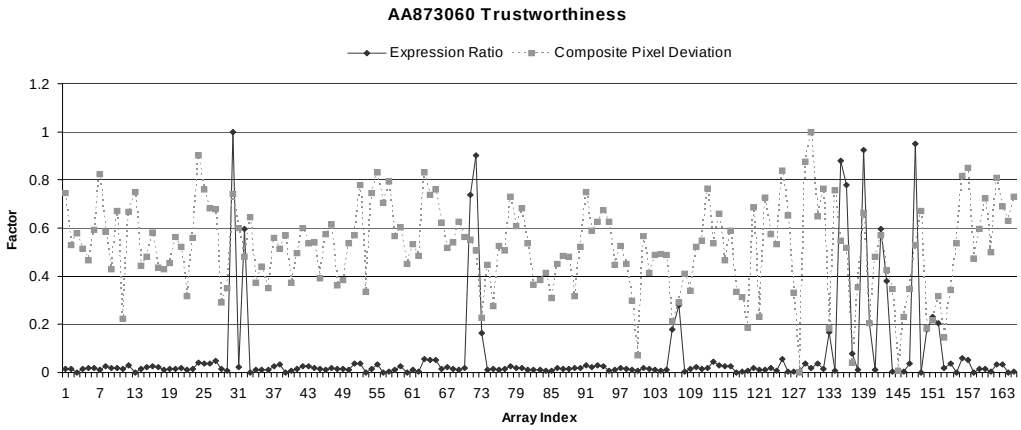


Figure 4.13: The *least* trustworthy feature AA873060 for the first data fold and using Feature Utility Ranking 1; the effective rank is 9,485.

rank varies across folds. Each gene was given a unique index that is an invariant across all folds and is set to be the gene's trustworthiness according to the first data fold. A simple example of this variation test, for two genes, is shown in figure 4.14, which shows dramatic differences between genes at index 59 and 195, corresponding to AI989344 and AA452916 respectively. Index 59 shows a gene that is generally very trustworthy in every data fold and has on average a trustworthiness of 59. However index 195 shows a gene whose trustworthiness varies dramatically from fold to fold; folds two, five, six and ten have quality issues. This variation supports the findings of other researchers who conclude that selecting a universal feature set is difficult in data sets with small sample sizes (Dougherty, 2001; Jain and Zongker, 1997) and suggests that FUR could be used to locate problematic arrays.

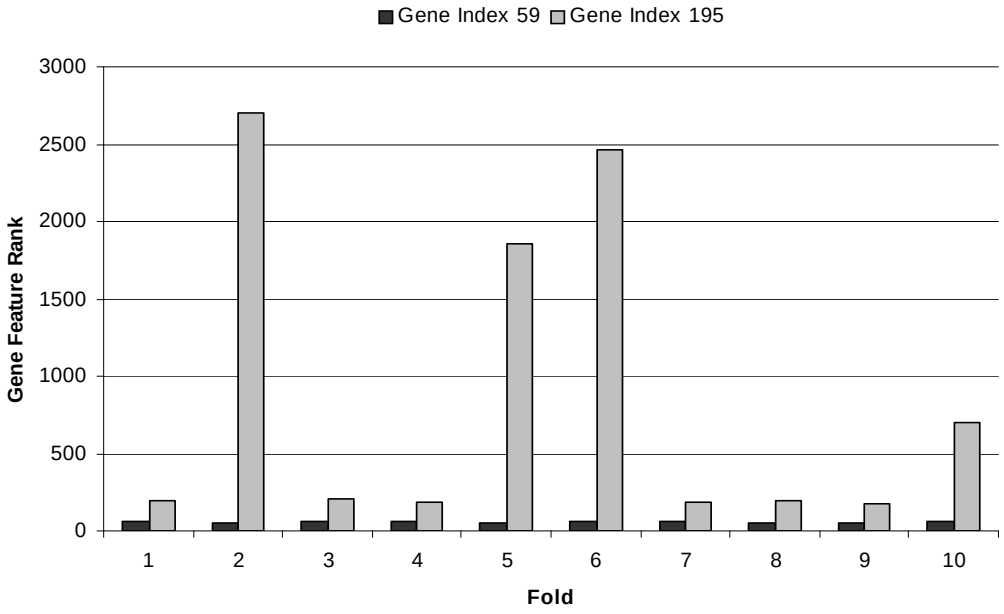


Figure 4.14: The variation of a feature's trustworthiness across 10 data folds, according to Feature Utility Ranking 1. Index 59 and 195 correspond respectively to AI989344 and AA452916.

The result of a complete variation test is shown in figure 4.15. The horizontal axis shows the index or trustworthiness of each feature according to the first fold, while the vertical axis shows the standard deviation of the indexes for all the folds. Note how the most and least trustworthy features, respectively located at the ends of the graph, have a very small standard deviation. The very small standard deviation for the most trustworthy is expected, since by its nature, exhibits very little variation in its quality. A similar argument also applies to the least trustworthy feature. A consistent increase and then tapering is also evident for the entire feature set, which is intuitively expected since moving from left to right is associated with decreasing trustworthiness and therefore increasing variation of quality amongst the folds. In addition, spikes are also visible, which indicate localized quality issues. Gene index 3,295 is responsible for the tallest spike.

The dramatic variation in trustworthiness shown above highlights a general problem—the trustworthiness of genes selected by FUR1 can be very different across folds; where different folds correspond to test data. This motivates the development of FUR2, which evaluates the trustworthiness of genes using *all* available data.

The second feature selection phase of FUR1 selects features according to their information content. The top ten ranked features are shown in table 4.10, together with their trustworthiness. A feature's trustworthiness

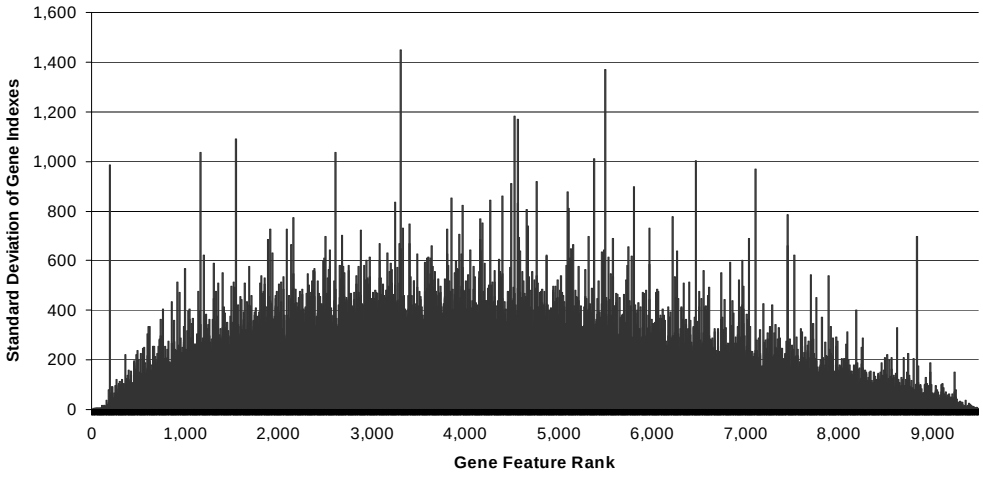


Figure 4.15: The trustworthiness of a feature can vary from fold-to-fold; shown is the variation in Feature Utility Ranking 1 trustworthiness for the B versus T-Cell data set, which was divided into ten folds.

and information content are always quoted as the rank of the gene in these results. The *Precision* and *Recall* columns show the classification accuracy for the top i features. The trustworthiness figures shown suggest that most of the features were significantly affected by noise. The top ranking feature in table 4.10 cleared the trustworthiness threshold by less than 300 and the average trustworthiness is 3,968, which raises the question whether the threshold is too high.

Hierarchical cluster analysis of this data is provided in figure 4.16. The boundary between the classes is clearly seen in the heat map and provides a stronger division than the traditional feature selection in figure 4.4 on page 169, thus suggesting that FUR1 results in improved generalization.

Table 4.10: Highest ranking features according to the second feature selection phase of Feature Utility Ranking 1. These results were obtained using the first fold of the B versus T-Cell data set.

Rank	Gene ID	*Precision*	*Recall*	Trustworthiness
1	AA486532	0.884	0.878	6,707
2	AI356451	0.862	0.865	4,792
3	H42728	0.839	0.845	5,083
4	R97095	0.883	0.885	1,990
5	AA775223	0.933	0.932	5,314
6	AA706022	0.959	0.959	1,281
7	H63077	0.959	0.959	47
8	H02333	0.959	0.959	4,550
9	R33103	0.953	0.953	6,653
10	AA451863	0.973	0.973	3,267

Note that none of the features selected by FUR1 were also selected by the traditional methodology.

Finally, three classifiers were constructed based on the three feature set sizes: 16 features, 1,024 features and 9,485 features, as shown in table 4.11. The results shown are stratified ten-fold cross validation averages. Compared to the results for traditional feature selection, shown in table 4.6 on page 170, the accuracy of the 16 feature models is almost identical, while the 1,024 feature model based on FUR1 was approximately 7% more accurate and the 9,485 feature models performed identically as expected. Differences in the features used to build the model using the traditional and FUR1 methodologies are dramatic. In the case of the 16 feature models, none of the features selected by the traditional methodology were located to the left

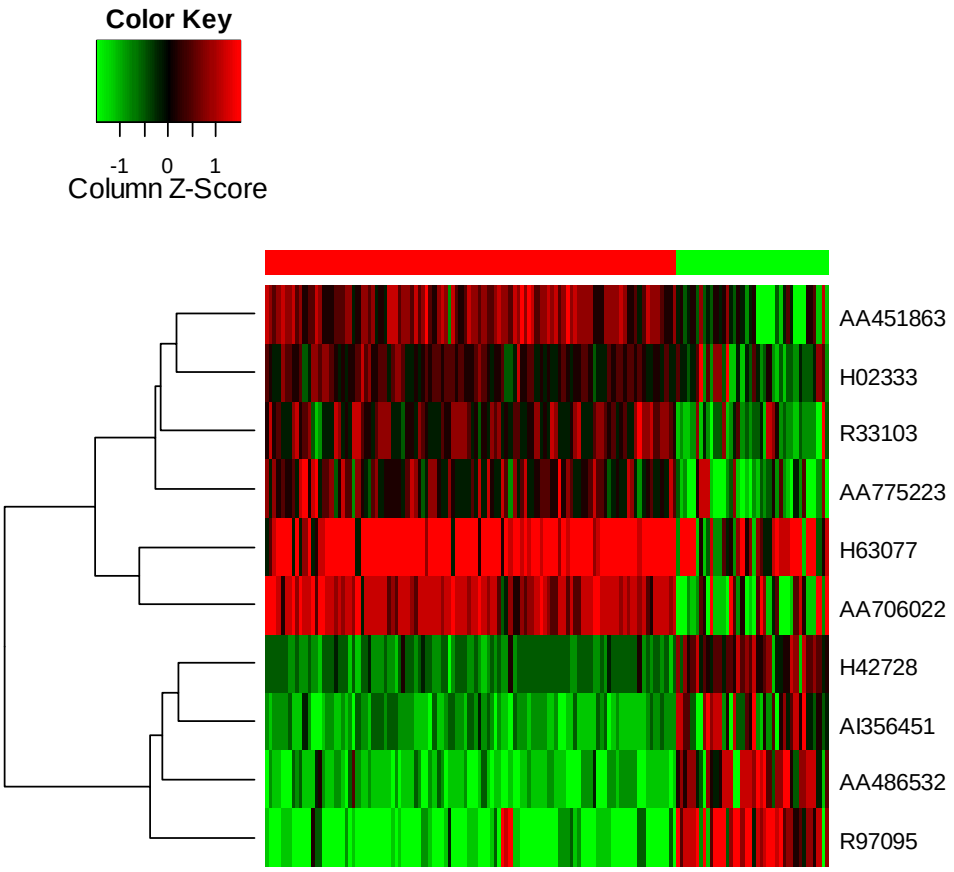


Figure 4.16: Hierarchical cluster analysis of the Feature Utility Ranking 1 selection for the B versus T-Cell data set.

Table 4.11: Classification accuracy for the B versus T-Cell data set according to Feature Utility Ranking 1. The accuracy achieved using traditional feature selection is also shown.

Features	FUR1	*Precision*	*Recall*	Traditional
16	96.46%	0.975	0.975	96.97%
1,024	97.01%	0.975	0.983	90.91%
9,485	98.78%	0.959	0.992	98.78%

of the trustworthiness threshold (see figure 4.8 on page 178) and were insufficiently trustworthy to use. The features used in the 1,024 models also follow a similar pattern.

That the average accuracy of the traditional and FUR1 generated models were very similar suggests that there is substantial feature redundancy. This redundancy enabled FUR1 to achieve a similar classification accuracy to that of the traditional methodology, despite using entirely different features. If FUR1 correctly evaluates data quality, then the accuracy obtained by the traditional methodology was due to apparent information content.

For the 1,024 feature model generated by the traditional methodology, only 412 of its features were located to the left of the trustworthiness threshold. This difference caused substantial changes in the feature sets of the traditional and FUR1 generated models. This suggests that FUR1 chose features, which not only performed well for the training and validation sets, but also for the test set. This is expected to be due to the fact that FUR1 identifies features that are also consistent with respect to quality.

It is worth noting that the trustworthiness threshold used eliminated approximately 26% of the entire feature set. Yet approximately 40% of the features used by the traditionally constructed model were rejected by FUR1 on the grounds of trustworthiness. This significant difference in proportions suggests, together with improved accuracy by the FUR1 generated model, that FUR1 did correctly identify a quality issue. Conversely, this suggests that the traditional methodology was either misled by apparent information content, or that features selected were associated with poor data quality within the test set.

The results presented are an ideal outcome for FUR1, since FUR1 either matched or improved upon the accuracy achieved by the traditional methodology.

Classifying treatment outcome

Following the basic approach used for the B versus T-Cell experiments, this section reports on classifying patients as achieving Complete Clinical Remission (CCR) or death. The accuracy of the models is expected to be less than the leukemic B and T-Cell experiments, due to the higher class imbalance in this data set and the fact that the relationship between treatment and outcome (CCR or death) are thought to be less strongly linked to gene expression.

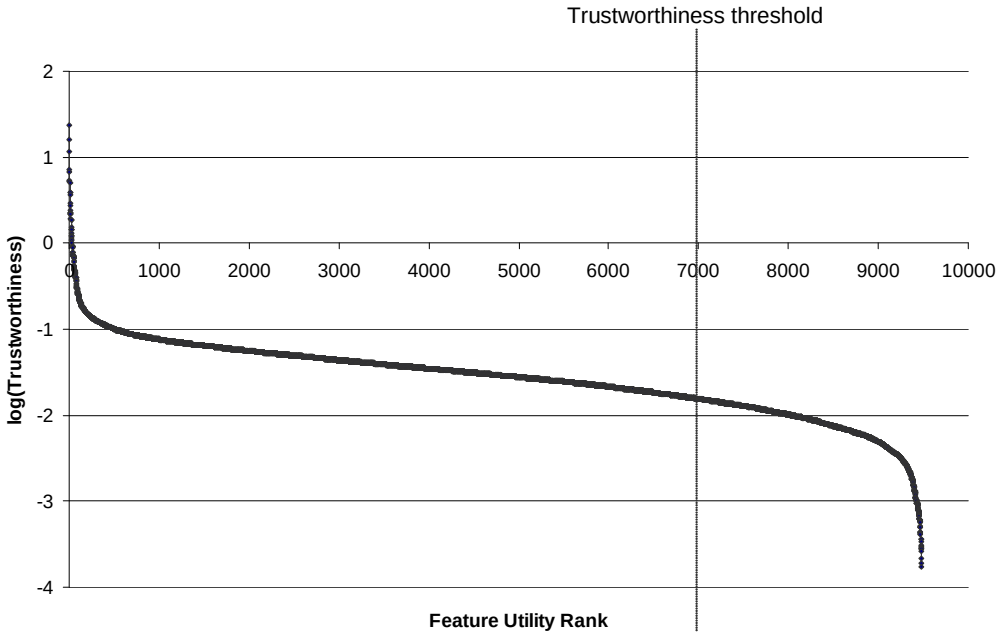


Figure 4.17: Feature trustworthiness, according to Feature Utility Ranking 1, for the treatment outcome data set.

The result of using the first feature selection phase, ranking according trustworthiness is shown in figure 4.17. The same characteristic shape as seen before is shown in the figure and the same threshold (7,000th most trustworthy feature) is used. There is little difference between this trustworthiness curve and the B versus T-Cell one in figure 4.8 on page 178. The vertical segments of this curve extend slightly further than the B and T-Cell one, suggesting the presence of problematic features at both extremes of the curve.

Although the characteristic shape of the trustworthiness curve remains

unchanged by moving from cell type (B versus T-Cell) to outcome (CCR versus death) classification, the trustworthiness of the features has changed dramatically, as shown in figure 4.18. This change is due to differences in the data set, which is caused by the introduction of time to the phenomenon being modeled. Although the leukaemic cell type is known at the point of diagnosis for every patient, time plays a part in determining a patient's outcome. For 38 patients, out of the 164, five years has not elapsed since the completion of treatment. If a relapse occurs within five years, the outcome is known in advance, but a CCR outcome cannot be determined until the end of this period; therefore preventing the inclusion of such patients within the data set.

The horizontal and vertical axes of figure 4.18 respectively show a feature's rank for the B versus T-Cell and treatment outcome data sets. Although these two data sets are fundamentally the same, they are not identical because some of the patients in the former could not be included in the latter, since the outcome of their treatment was unknown. As a result, there have been substantial changes in rank and many changes were dramatic; for example one feature, located in the bottom right section of the figure, changed rank from approximately 9,000 to 1,000. This suggests that trustworthiness is sensitive to changes in quality and that *changing the content of a data set can have dramatic implications.*

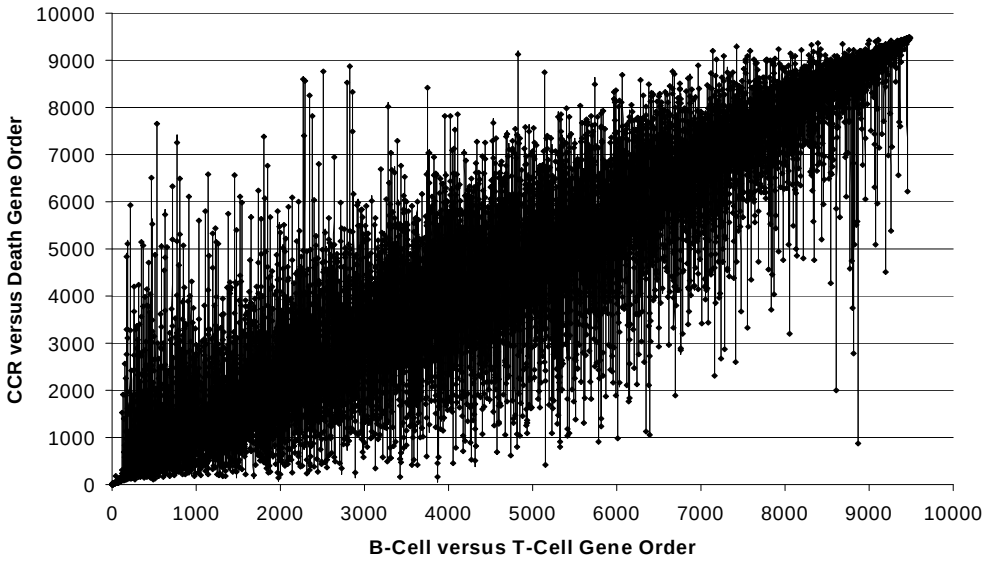


Figure 4.18: Changes in the trustworthiness of individual features caused by moving from the B versus T-Cell to treatment outcome data set; evaluated using Feature Utility Ranking 1.

The most trustworthy feature for the outcome data set, according to the first training fold, is shown in figure 4.19. The figure shows that a small number of arrays deviated significantly from the norm, which is consistent with the more extreme behavior noted on page 192. The same unexpected behavior, where a high expression ratio is correlated with low composite pixel deviation, is evident in this figure as for the B versus T-Cell data set. Investigation into why these arrays were different revealed a correlation for construction dates, as was the case for the B versus T-Cell analysis in section 4.5.2. Note that the most trustworthy feature AA663923 was not used in the 16 and 1,024 feature models, since it was ranked 3,201 with respect to

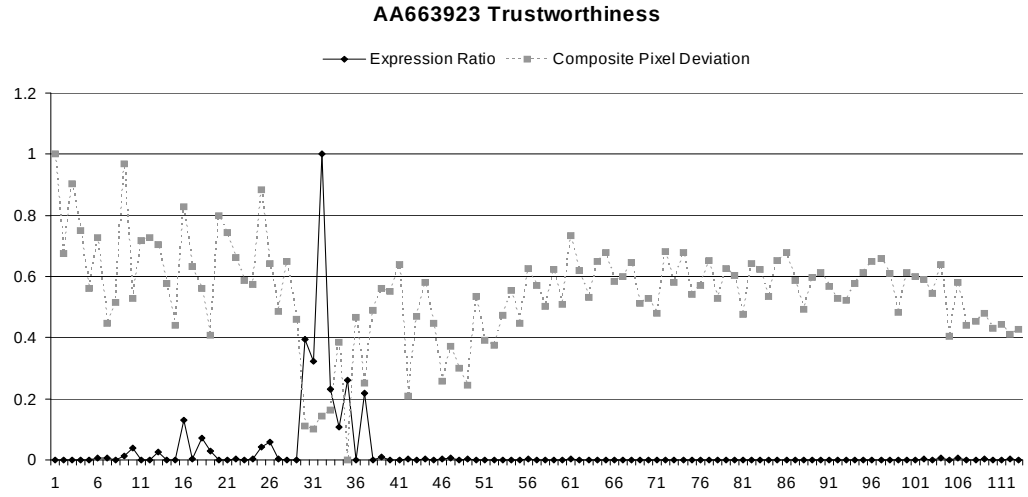


Figure 4.19: The most trustworthy feature for treatment outcome prediction, according to the first validation fold and Feature Utility Ranking 1.

its information content.

The second FUR1 phase, ranking features using their information content, was performed. The top ten features are listed in table 4.12, which shows that some of the features are particularly trustworthy, having a high rank. The *Precision* and *Recall* columns show the classification accuracy for the top i ranked features. The features shown correspond to the first training fold of the data set. Five of the features listed in the table were also selected from the first training fold for the traditional methodology experiments (see table 4.7 on page 174). However three of the features used within the traditionally built models are particularly untrustworthy, having ranks below 8,500.

The result of hierarchical cluster analysis of the data in table 4.12 is

Table 4.12: The ten highest ranking features for the second phase of Feature Utility Ranking 1, using the treatment outcome data set.

Rank	Gene ID	*Precision*	*Recall*	Trustworthiness
1	AA630498	0.906	0.921	1,541
2	N53505	0.897	0.912	2,511
3	AI927284	0.906	0.921	485
4	AA634300	0.950	0.947	6,192
5	W73144	0.942	0.939	4,247
6	N55205	0.942	0.939	29
7	N50880	0.905	0.912	4,488
8	AA029041	0.905	0.912	5,336
9	AA455108	0.905	0.912	3,966
10	H63077	0.911	0.921	59

shown in figure 4.20. This heat map provides a poor visualization of class separation as was the case for traditional feature selection shown in figure 4.6 on page 173. This poor visualization is most likely for the same reasons: class imbalance and the complex relationship between genetics and treatment outcome. However the FUR1 generated heat map subjectively provides improved color consistency, which may be due to a better generalization. Note that the heat map shows individual features whilst classification involves interactions between the features.

Finally three classification models, consisting of 16, 1 024 and 9 485 (or all) features, were built and tested. The accuracy of these models is shown in table 4.13. The 16 feature model, compared to the traditional feature selection model, was approximately 2% less accurate. Approximately half

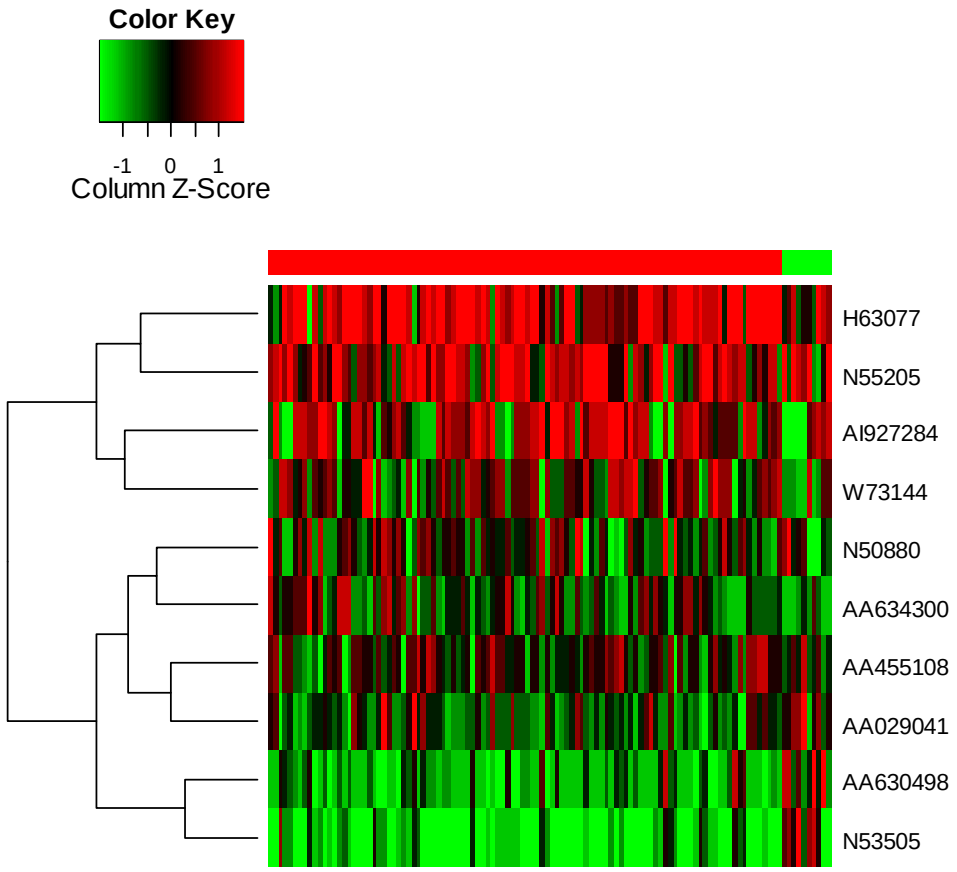


Figure 4.20: Hierarchical cluster analysis of the Feature Utility Ranking 1 selection for the outcome data set.

Table 4.13: The classification accuracy for models constructed using Feature Utility Ranking 1 and the treatment outcome data set.

Features	FUR1	*Precision*	*Recall*	Traditional
16	90.00%	0.929	0.963	92.06%
1,024	88.33%	0.920	0.954	96.00%
9,485	88.89%	0.914	0.972	88.89%

of the features used in the FUR1 generated models were also present in the traditionally built models, however their ranks were different. Therefore the decrease in accuracy of the FUR1 models was due to a loss of information content, caused by feature set differences. This reduction was either caused by a shortcoming in FUR1 or additional information found by the traditional approach. If trustworthiness does correctly identify the presence of noise, then the additional information found by the traditional approach was due to noise. The 1,024 feature models generated by the traditional and FUR1 methodologies performed similarly to the 16 feature versions, except that FUR1 accuracy fell by approximately an additional 6%. The 9,485 feature models performed identically, as expected.

4.5.3 Feature Utility Ranking 2

This section applies Feature Utility Ranking 2 (FUR2) to the leukaemia data set experiments. Feature Utility Ranking 2 uses the $\mathbf{C}$ matrix view of the entire data set, which encompasses the training and test data sets in the calculation of trustworthiness. This approach is possible since only a measure

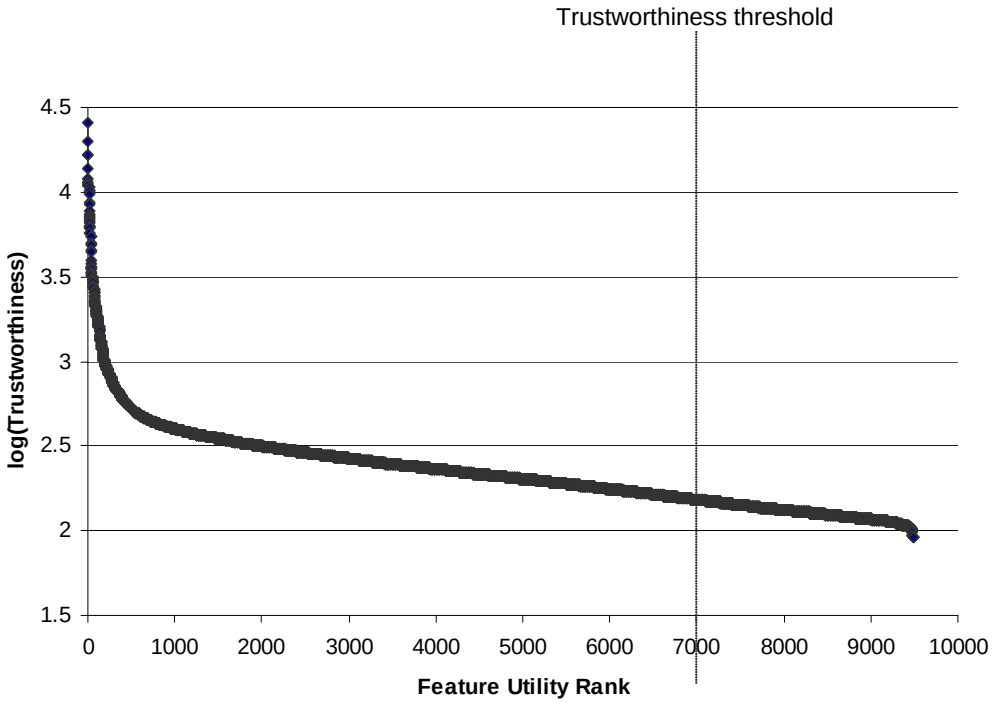


Figure 4.21: Feature Utility Ranking 2 trustworthiness for the entire B versus T-Cell data set.

of noise is used in calculating trustworthiness.

Classifying B versus T-cell

Using the **C** matrix of the entire data set, the first phase of FUR2 calculates feature trustworthiness for each feature. The resulting trustworthiness curve for the data set is shown in figure 4.21.

The trustworthiness curve still has the same characteristic shape of curves generated earlier with FUR1. This similarity is, at first, unexpected because of the absence of signal in the calculation of trustworthiness by FUR2. How-

ever, a relationship exists between expression levels approaching the measurable limit and pixel saturation, which most likely explains the similarity for the left vertical segment.

The right vertical segment of figure 4.21 is very short compared to FUR1. The almost non-existent right vertical segment shows that the number of unusually poor quality sample points is small using this different measure of trustworthiness. The relatively flat middle segment of the figure exhibits a more gradual decent than the FUR1 generated curves. Lastly, the offset of the curve is somewhat higher than for FUR1 curves. Although differences between FUR1 and FUR2 generated curves are evident, it is important to note that the relative position or rank of features is important, rather than any absolute value of trustworthiness.

By chance the top ten features selected by traditional feature selection, shown in table 4.5 on page 168, were located above the trustworthiness threshold. Consequently these features were attributed the same rank by FUR2. Therefore in order to discern a difference, the top twenty features, rather than the top ten, are listed in tables 4.14 and 4.15. The first difference in table 4.14 is in the eleventh position, where AI336342 fell below the trustworthiness threshold and was replaced by T66799 in the "FUR2 Selection" column. All the features in the "Traditional Selection" column marked with # were eliminated by FUR2 on the basis of trustworthiness. This resulted

Table 4.14: Differences between feature selection by the Feature Utility Ranking 2 and traditional methodologies for the B versus T-Cell data set. The features marked with a # only appear in the right column since they fell below the trustworthiness threshold, consequently four features, marked with a * in the left column, are their replacements.

Rank	FUR2 Selection	Traditional Selection
1	AA055946	AA055946
2	T64192	T64192
3	AA420981	AA420981
4	N91921	N91921
5	AA469965	AA469965
6	AA775257	AA775257
7	AA481988	AA481988
8	AA283629	AA283629
9	AA243694	AA243694
10	AW009739	AW009739
11	T66799	AI336342 #
12	R97095	T66799
13	H42728	R97095
14	R06417	AA933862 #
15	AI453185	H42728
16	H63077	AA668821 #
17	AI023136 *	N29639 #
18	AA709036 *	R06417
19	AA442984 *	AI453185
20	AA451863 *	H63077

in features ranked twenty one to twenty four by traditional feature selection, being ranked seventeen to twenty by FUR2, these features are marked with a *. The FUR2 selected features together with their resulting classifier accuracy are shown in table 4.15.

Figure 4.22 contains the result of hierarchical cluster analysis of the data in table 4.15. The heat map contains twenty features since the first ten were

Table 4.15: The twenty highest ranking features of the B versus T-Cell data set, according to the second phase of Feature Utility Ranking 2.

Features	Gene ID	*Precision*	*Recall*	Trustworthiness
1	AA055946	0.976	0.976	4,600
2	T64192	0.982	0.982	746
3	AA420981	0.982	0.982	2,600
4	N91921	0.976	0.976	244
5	AA469965	0.988	0.988	1,672
6	AA775257	0.988	0.988	6,550
7	AA481988	0.982	0.982	4,488
8	AA283629	0.982	0.982	4,807
9	AA243694	0.982	0.982	3,216
10	AW009739	0.982	0.982	4,640
11	T66799	0.976	0.976	4,600
12	R97095	0.988	0.988	746
13	H42728	0.988	0.988	2,600
14	R06417	0.988	0.988	244
15	AI453185	0.988	0.988	1,672
16	H63077	0.994	0.994	6,550
17	AI023136	0.994	0.994	4,488
18	AA709036	0.988	0.988	4,807
19	AA442984	0.988	0.988	3,216
20	AA451863	0.988	0.988	4,640

Table 4.16: Classification accuracy of the B-Cell versus T-Cell data set using Feature Utility Ranking 2.

Features	FUR2	*Precision*	*Recall*	Traditional	FUR1
16	90.39%	0.992	1.000	96.97%	96.46%
1,024	91.17%	0.983	0.992	90.91%	97.01%
9,485	98.78%	0.992	0.992	98.78%	98.78%

also selected by the traditional methodology. This suggests that traditional feature selection is closer in similarity to FUR2 than FUR1. Visually the separation between classes is evident, but perhaps not as effectively as for FUR1.

The results of constructing three different classification models are shown in table 4.16. The 16-feature FUR2-generated model was approximately 7% less accurate than obtained using a traditional methodology and approximately 6% less accurate than obtained with FUR1. This effective loss of accuracy could be due to the use of apparent information by the traditional methodology. The 1,024 feature FUR2 model achieved a marginal improvement over its traditional version, which could be due to the benefit of using more trustworthy features, not only with respect to the training set, but also the test set. The 9,485 feature model performed identically, as expected. In the case of the 16 and 1,024 feature models, FUR2 was less accurate than FUR1, which suggests that this type of problem benefits from using a signal-to-noise measure of quality.

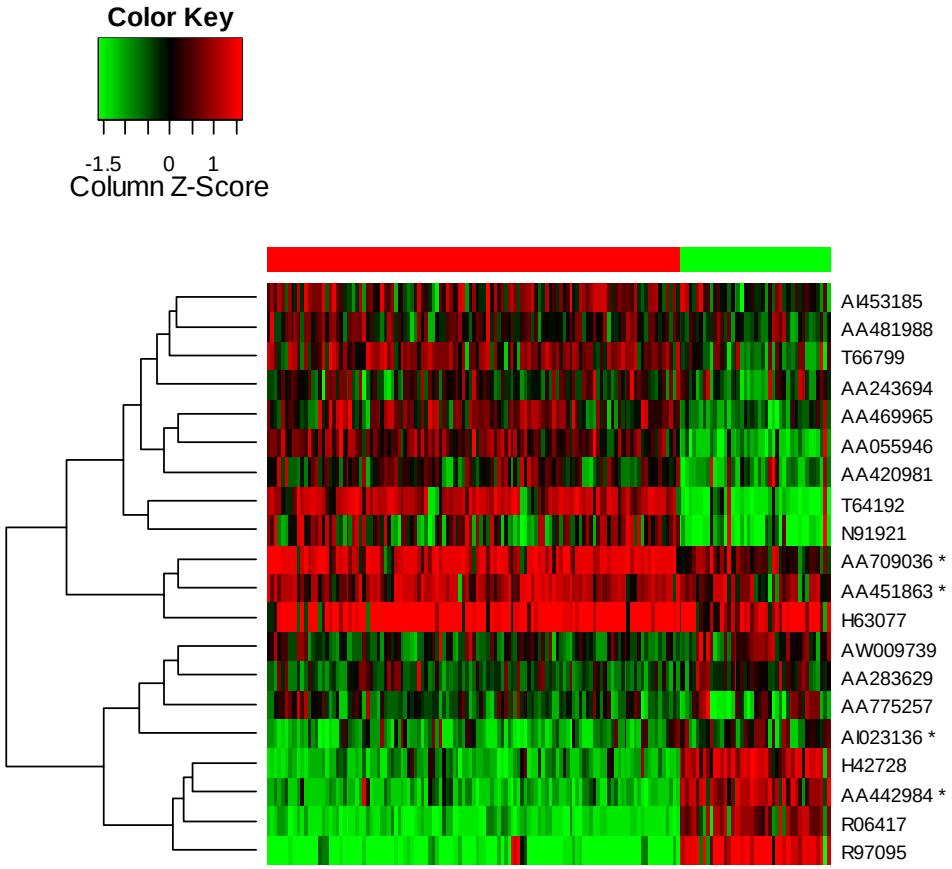


Figure 4.22: Hierarchical cluster analysis of the Feature Utility Ranking 2 selection for the B versus T-Cell data set. Features marked with a * were not rejected by traditional feature selection.

Classifying treatment outcome

Following the empirical approach used with the B versus T-Cell data set, this section reports on evaluating the FUR2 methodology using the leukaemia treatment outcome data set. The trustworthiness curve generated by the first feature selection phase of FUR2 is shown in figure 4.23. In comparison to the B versus T-Cell trustworthiness curve, there are no qualitative differences. However significant changes occurred in the trustworthiness rank of individual features, as a result of moving from the Leukemic Cell Type to Treatment Outcome data sets. The changes in trustworthiness of individual features are shown in figure 4.24.

The first difference between the FUR2 and the traditionally selected features occurred in position 162, which resulted in identical FUR2 and traditional sixteen feature models, refer table 4.17. In the case of the traditionally built 1,024 feature model, 304 features were eliminated on the basis of trustworthiness by FUR2. These results only refer to the first data fold.

The figures in table 4.17 are stratified 10-fold stratified cross validation averages. The 16 and the 9,485 feature FUR2 generated models performed identically to the traditional models. The FUR2 generated version of the 1,024 feature model was approximately 6% less accurate than the traditional model. This loss of accuracy could be due to a possible shortcoming with

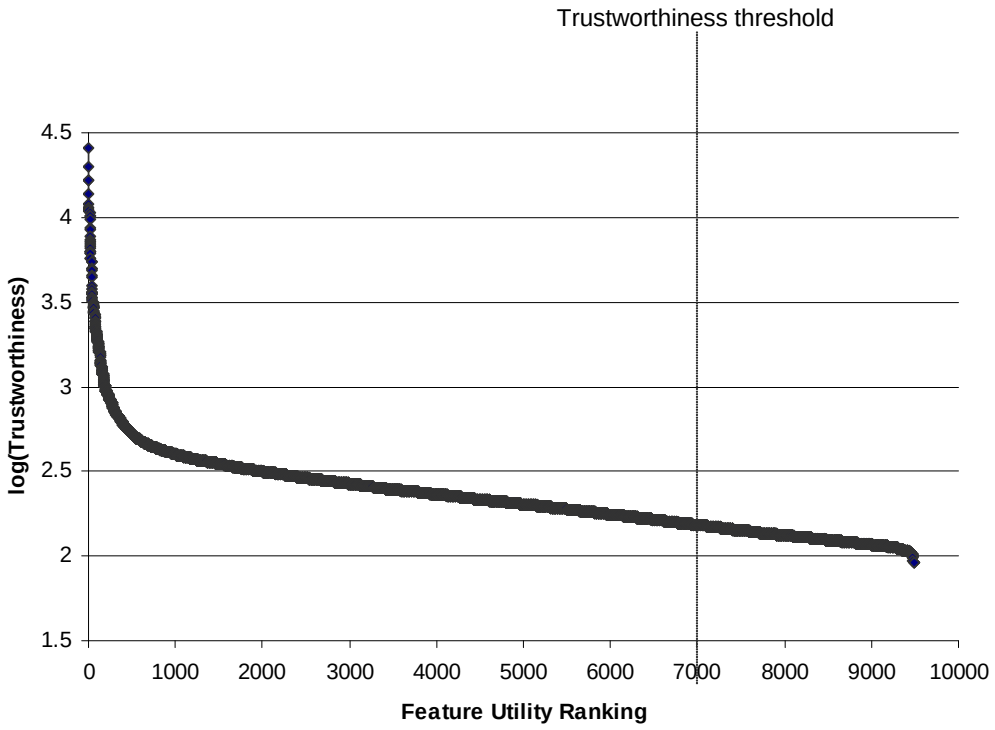


Figure 4.23: Feature Utility Ranking 2 trustworthiness for the entire treatment outcome data set.

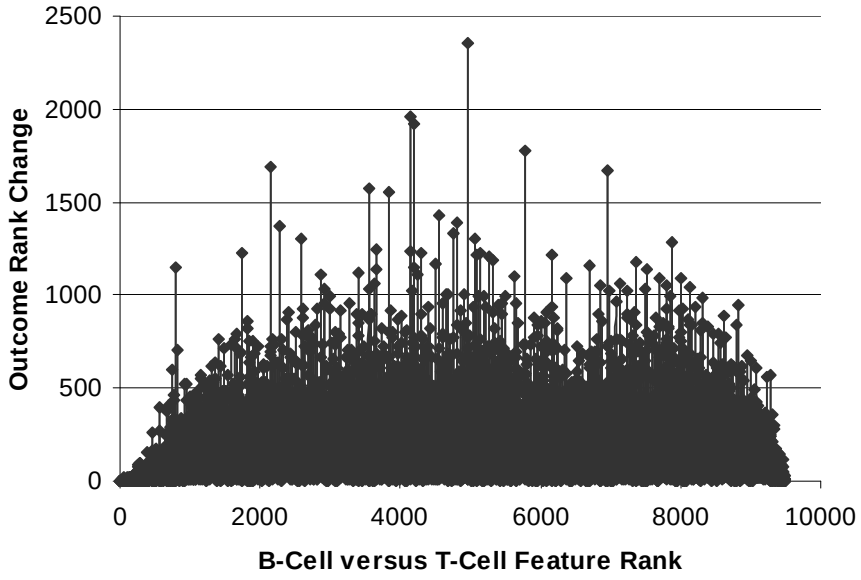


Figure 4.24: Changes in trustworthiness of individual features, as a result of moving from the B versus T-Cell to treatment outcome data set; evaluated using Feature Utility Ranking 2.

Table 4.17: Classification accuracy for treatment outcome data set, using Feature Utility Ranking 2.

Features	FUR2	*Precision*	*Recall*	Traditional	FUR1
16	92.06%	0.914	0.921	92.06%	90.00%
1,024	89.68%	0.860	0.897	96.00%	88.33%
9,485	88.89%	0.831	0.889	88.89%	88.89%

FUR2, or due to the choice of quality measure, or the use of apparent information by traditional feature selection. However in the case of the 16 and 1,024 feature models, FUR2 out performed FUR1, which may be due to significant differences in noise between the training and test sets.

4.6 Chronic fatigue syndrome experiments

This section uses the Chronic Fatigue Syndrome (CFS) gene expression data set to evaluate the effectiveness of Feature Utility Ranking 1 and 2. The experimental approach followed is a simplified version of that used for the leukaemia data set experiments in section 4.5.

The chronic fatigue syndrome data set exhibits similar characteristics to the leukaemia data set, such as few samples, a significant class imbalance and the presence of noise as outlined in section 4.2.2. These characteristics are prerequisites for using Feature Utility Ranking. Note that if every instance of the chronic fatigue syndrome data set was given the non-fatigued classification, a model accuracy of approximately 69% would be achieved.

Table 4.18: Traditional feature ranking of the first training fold for the chronic fatigue syndrome data set.

Rank	Gene ID
1	XM_065418
2	L32835
3	NM_033536
4	AC018758 (1)
5	NM_031954
6	BC002828
7	BC012427
8	BC002361
9	NM30474
10	AB023961

4.6.1 Traditional feature selection

Using traditional feature selection, all features of the chronic fatigue syndrome data set were ranked according to their information content. The ten highest ranking features are shown in table 4.18.

The result of hierarchical cluster analysis of table 4.18 is shown in figure 4.25. The red and green rectangle directly above the heat map shows the relative number of classes, where fatigued forms the majority. The difficulty in visually separating the classes in the heat map suggests poor generalization by the features, which is supported by the classifier accuracy results.

Using the list of ranked features, three different classifiers were constructed and evaluated with stratified ten-fold cross validation. Results are shown in table 4.19. The classification accuracy shown in the table is rela-

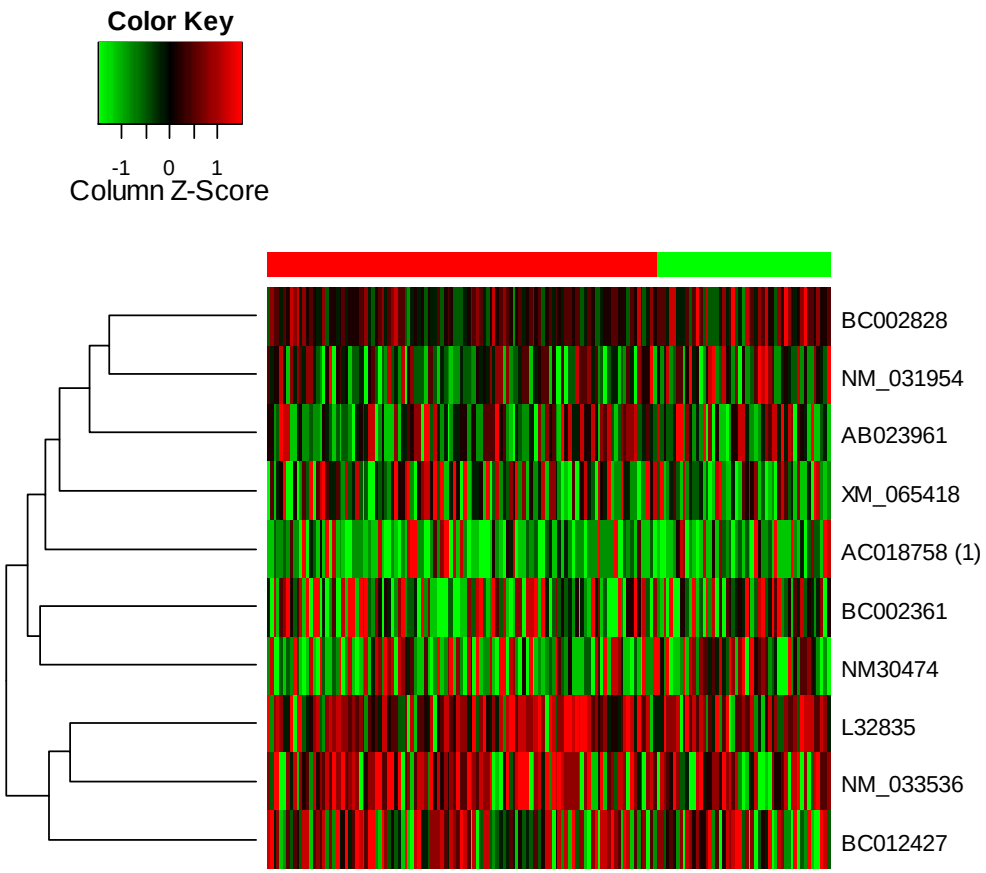


Figure 4.25: Hierarchical cluster analysis of traditional feature selection for the chronic fatigue syndrome data set.

Table 4.19: Classification accuracy of the chronic fatigue syndrome data set, resulting from traditional feature selection.

Features	Traditional	Precision	Recall
16	69.8%	0.860	0.672
1,024	72.7%	0.867	0.714
20,160	59.3%	0.696	0.731

tively poor and the precision and recall values are consistent with significant classification errors for both class labels.

The poor accuracy of the classifiers could be due to the following factors: an absence of relevant features, excessive noise within the data set, or using an inappropriate modeling technique. The first of these is considered most likely, since chronic fatigue syndrome is generally considered to have a psychosocial rather than genetic origin. There is also likely to be excessive noise within the data set. The third possibility of using an inappropriate modeling technique is less important because the same modeling technique is used throughout and the aim is to contrast feature selection methods, rather than classifiers.

4.6.2 Feature utility ranking 1

Using the **C** matrix view of the training data set, the trustworthiness of every feature was calculated using FUR1's first feature selection phase and graphed in figure 4.26. The same trustworthiness threshold of 7,000, as used for the leukaemia experiments, is also used and the same characteristic trustworthiness curve is evident. This characteristic shape implies the same type

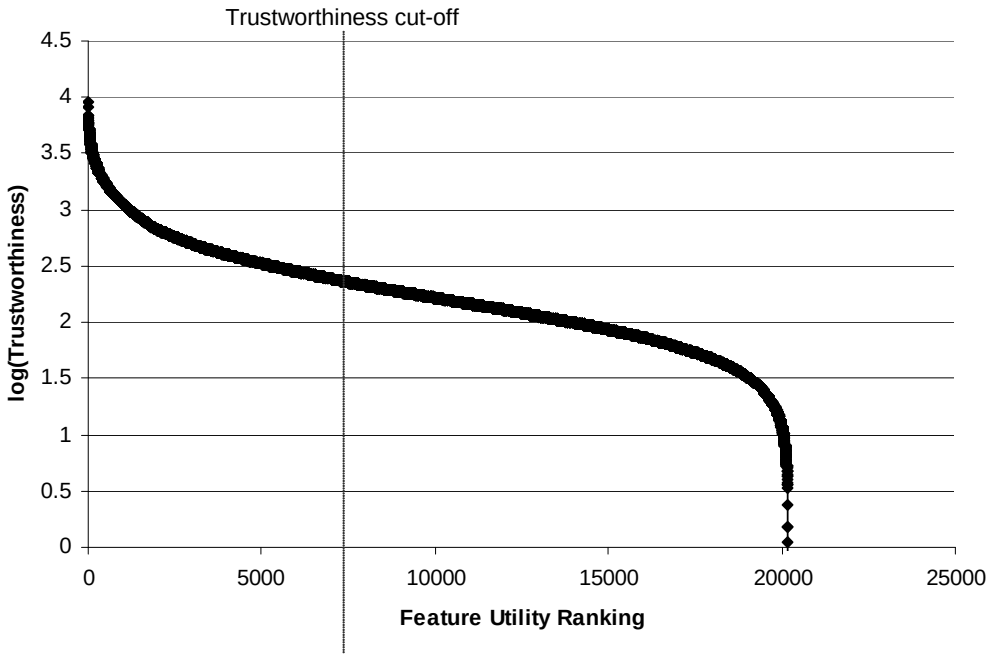


Figure 4.26: Feature trustworthiness, according to Feature Utility Ranking 1, for the chronic fatigue syndrome data set.

of distribution of quality exists in the leukaemia and chronic fatigue syndrome data sets, which is unsurprising since quality is expected to gradually vary, except at each end of the distribution. Application of the threshold eliminates a total of 13,160 features from the data set.

The second feature selection phase identified a total of 1,024 features, of which the first ten features, for the first training fold, are shown in table 4.20. Note that the three features marked with a * were are also present within the top ten traditionally selected features in table 4.18. The result of hierarchical cluster analysis of table 4.20 is shown in figure 4.27, which visually provides

Table 4.20: Feature Utility Ranking 1 selection of the chronic fatigue syndrome data set. Only the features marked with an * were also used by traditional feature selection.

Rank	Gene ID
1	AC018758 (1) *
2	BC002361 *
3	BC002828 *
4	AB005622
5	NM_007164
6	AK093161
7	AF261689
8	AF229126
9	AB009619
10	AB071198

Table 4.21: Classification accuracy of the chronic fatigue syndrome data set, resulting from Feature Utility Ranking 1.

Features	FUR1	*Precision*	*Recall*	Traditional
16	71.5%	0.807	0.773	69.8%
1,024	72.7%	0.816	0.782	72.7%
20,160	59.3%	0.696	0.731	59.3%

a similar degree of class separation as for using the traditional methodology.

Using the features ranked by FUR1, three classification models were constructed and evaluated using ten-fold cross validation, see table 4.21. Comparison of the FUR1 and traditionally built models reveals little difference. This result suggests a number of possible causes, for example, the absence of sufficiently relevant features, insufficient redundancy, or a weakness in FUR1.

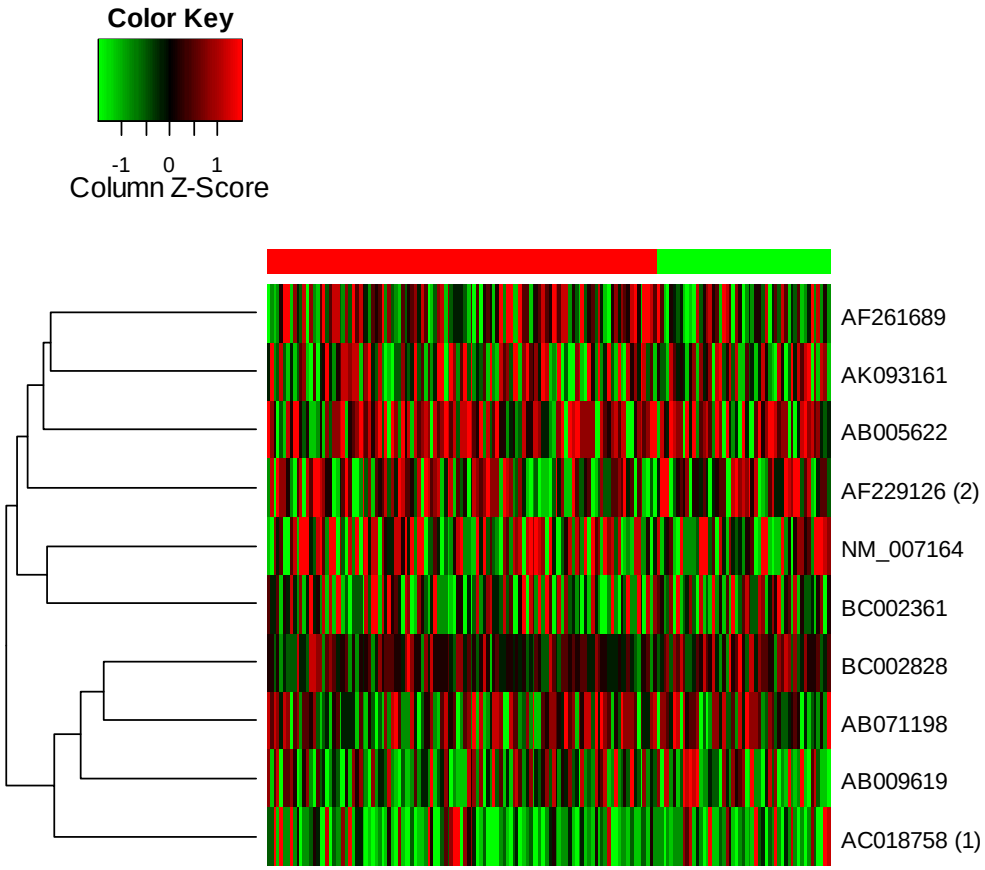


Figure 4.27: Hierarchical cluster analysis of Feature Utility Ranking 1 selection for the chronic fatigue syndrome data set.

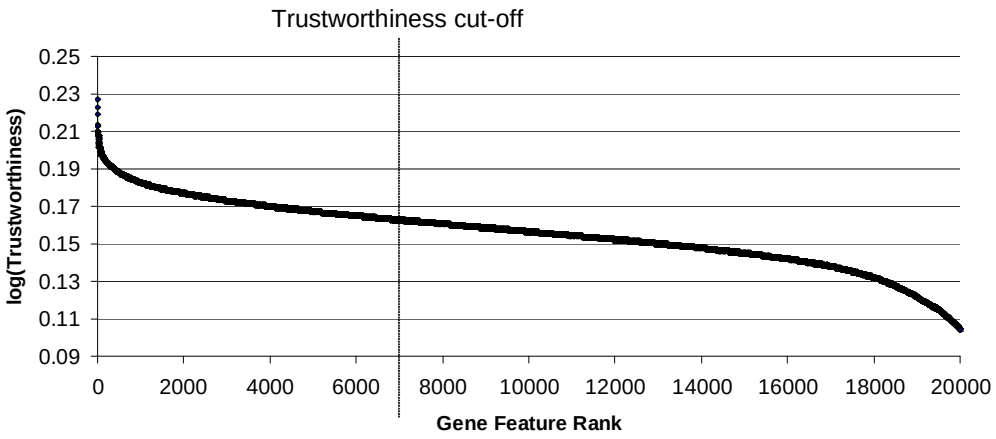


Figure 4.28: Feature Utility Ranking 2 trustworthiness of the chronic fatigue syndrome data set.

4.6.3 Feature utility ranking 2

Using the **C** matrix view of the entire data set, the first phase of FUR2 calculates the trustworthiness of every feature, which is graphed in figure 4.28. Compared to the FUR1 curve in figure 4.26, the FUR2 curve is more compressed in the vertical axis and the right vertical segment of the curve is much flatter. These differences suggest that data quality is more stable according to FUR2.

The second phase of FUR2 selects features above the trustworthiness threshold and ranks them by information content; the top twenty features are shown in table 4.22. Features marked with a * were also present within the top twenty traditionally ranked features. With respect to the traditionally selected features (table 4.18), fifteen features fell below the trustworthiness

Table 4.22: Feature Utility Ranking 2 selection of the chronic fatigue syndrome data set. Features marked with a * were also present within the top twenty traditionally ranked features.

Rank	Feature
1	NM_000478 *
2	NM_000625 *
3	BC005233 *
4	Blank(86) *
5	NM_016388
6	XM_091100
7	XM72411
8	AF491780
9	BC000241
10	NM_001567
11	XM_029925
12	L02320
13	BC001756
14	NM_002177
15	AC006023
16	AF548661
17	XM72615
18	BC030012
19	AF208862
20	NM38991 *

threshold and were eliminated from further use by FUR2.

The result of hierarchical cluster analysis of table 4.22 is shown in figure 4.29. Just as for every FUR2 analysis of the biomedical data in this thesis, twenty features were required to reveal differences with traditional feature selection. As for traditional and FUR1, no discernable class separation is evident in in the heat map shown in figure 4.29.

Using the list of features ranked by FUR2's second phase, three classifiers

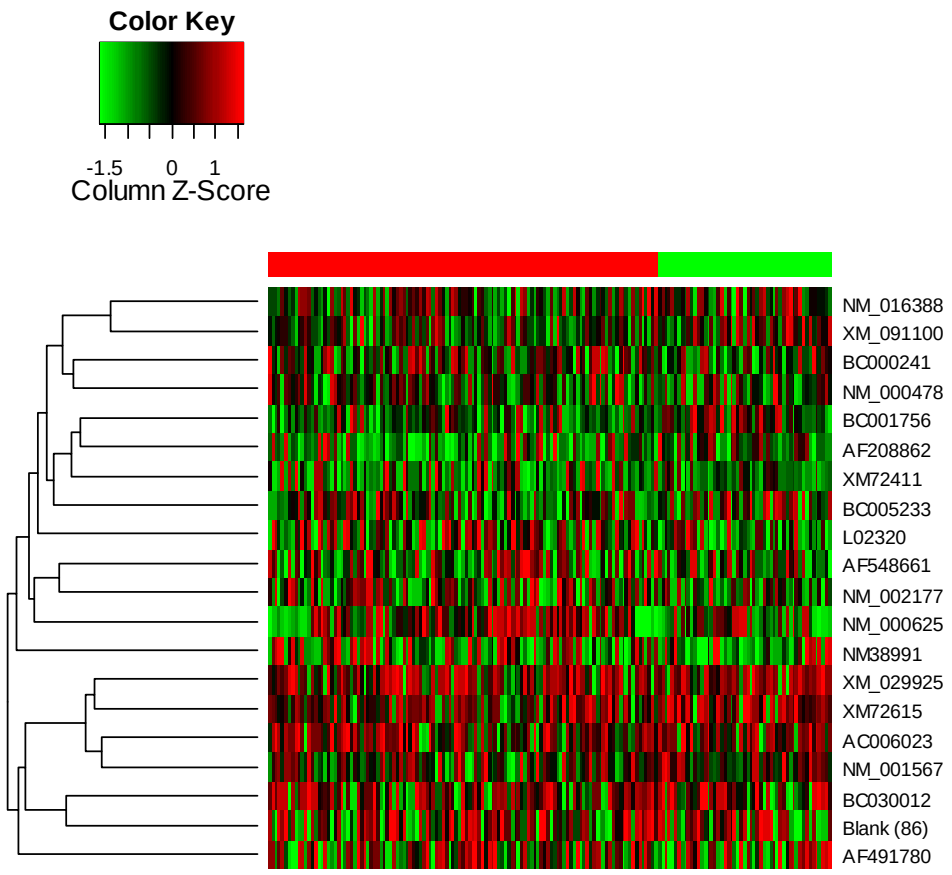


Figure 4.29: Hierarchical cluster analysis of Feature Utility Ranking 2 selection for the chronic fatigue syndrome data set.

Table 4.23: Classification accuracy of the chronic fatigue syndrome data set, resulting from Feature Utility Ranking 2.

Features	FUR2	*Precision*	*Recall*	Traditional	FUR1
16	69.8%	0.802	0.748	69.8%	71.5%
1,024	65.1%	0.776	0.697	72.7%	72,7%
20,160	59.3%	0.696	0.731	59.3%	59.3%

were constructed and evaluated using ten-fold cross validation, see table 4.23. The classifiers constructed using the Traditional methodology and FUR1 models (in table 4.19) are used to contrast these results. With respect to the sixteen feature model, FUR2 achieved the same accuracy, while the 1,024 feature FUR2 model weas approximately 7% less accurate. However FUR1 did outperform FUR2, which suggests that a benefit exists from incorporating signal within a measure of quality.

The poor classification results presented for FUR2 and the other two methodologies, supports the original expectation that chronic fatigue syndrome is of psychosocial origin, rather than genetic. The 16 feature FUR2 model achieved the same accuracy as the traditional model, while using different features. Again using different features, the 1,024 feature FUR2 model was approximately 7.5% less accurate than the traditional model. The decrease in accuracy by the 1,024 feature FUR2 may be a more accurate reflection of the latent information contained in the data and the improved accuracy of the traditional model may be due to apparent information.

4.7 Biological relevance

As a final analysis, Gene Set Enrichment Analysis (GSEA) was used to evaluate the biological "meaningfulness" of the genes selected by each of the three feature selection methodologies (Subramanian et al., 2005; Mootha, Lindgren, Eriksson, Subramanian, Sihag, Lehar, Puigserver, Carlsson, Ridderstrale, Laurila, Houstis, Daly, Patterson, Mesirov, Golub, Tamayo, Spiegelman, Lander, Hirschhorn, Altshuler, and Groop, 2003). GSEA uses a biological knowledge based approach for evaluating the meaningfulness of the selected genes. Given how GSEA works, a list used to classify a well understood biological phenomenon is preferable. Of the three real-world data sets evaluated in this research, the immunophenotype (or B versus T Cell) data set is the most ideal and accordingly used in this evaluation. Again since GSEA is dependent on prior knowledge, a list consisting of the largest possible number of genes is expected to reduce the chances of false associations with biological knowledge. As a result three gene lists, each corresponding to one of the three feature selection methodologies and containing 7,000 genes were evaluated. The value of 7,000 was determined by the trustworthiness threshold used in this research, hence the maximum number of trusted genes. To minimize experimental variability a limit of 7,000 was also applied to the traditionally generated list.

Three of the measures of meaningfulness used by GSEA are the Enrichment Score (ES), a nominal p value for the ES and a False Discovery Rate (FDR). The ES is a measure of the number of biologically meaningful gene sets located in the extremes of an evaluated gene list; hence a high ES score is associated with biologically meaningful genes being located either toward the start or end of the list, while a lower ES is associated with meaningfulness being located toward the center of the list. Each gene set corresponds to different biological meaning. The p value provides the statistical significance of the measured ES. This significance is calculated by contrasting the ES with the ES that is measured when the class labels in the same data set are randomly permuted. Since the calculation of a p value involves multiple hypotheses about biological meaning, a FDR is calculated in a way that normalizes the results for the individual hypotheses; this results in a more demanding assessment.

The GSEA results for the immunophenotype feature selection are presented in table 4.24. The table shows the number of biologically meaningful gene sets found that satisfy the following criteria. The first result column shows the number of identified gene sets, which correspond to a nominal p value of less than 1%. In the case of FUR2, 32 different sets were found. The second result column shows the number of identified gene sets that are associated with a FDR of less than 25%. According to this criterion, only

Table 4.24: Gene Set Enrichment Analysis estimation of significance and False Discovery Rate results for the B versus T Cell data set.

Feature Selection Methodology	Number of gene sets identified for the criteria:	
	Nominal p value $< 1\%$	FDR $< 25\%$
Traditional	6	0
FUR1	24	0
FUR2	32	24

FUR2 selected biologically meaningful genes; corresponding to 24 different sets. The first meaningful set identified by traditional feature selection occurred at a FDR of less than 50%.

Figure 4.30 shows the relationship between the number meaningful gene sets and the FDR. This figure shows a clear difference exits between the biological meaningfulness of the genes selected using FUR and those selected using a traditional methodology. FUR1's use of a signal-to-noise measure of data quality provided a clear improvement in biological meaningfulness. Although FUR2 only considers noise in the assessment of data quality, the inclusion of the test set in the calculation of data quality is expected to be responsible for the significant reduction in the FDR. These results suggest that noise is a significant issue and that a difference in the quality of the training and test set exists. It also suggests that the traditional feature selection methodology is principally influenced by noise, rather than biological meaningfulness.

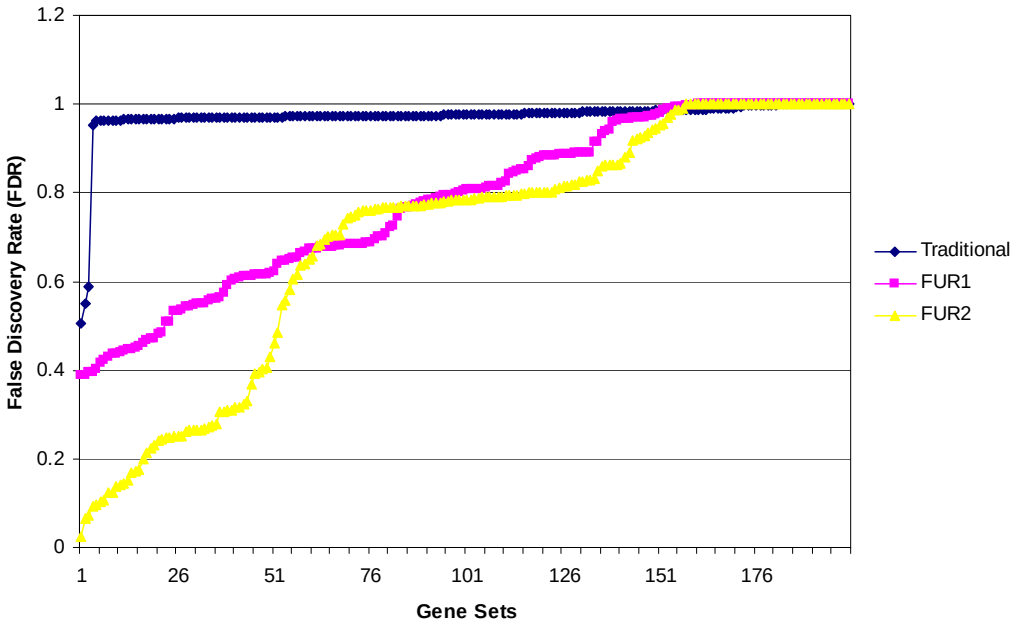


Figure 4.30: Gene Set Enrichment Analysis results for gene sets versus False Discovery Rate of the B versus T Cell data set.

4.8 Discussion

The evaluation of three feature selection methodologies was conducted in this chapter. The three methodologies consisted of Feature Utility Ranking 1 and Feature Utility Ranking 2, as well as a reference methodology, namely traditional feature selection. The evaluation used a synthetic and real-world data sets. Synthetic data was used to provide an environment where the degree of feature redundancy and noise was known. However the major portion of the experimentation involved three real-world data sets for classifying: leukaemia immunophenotype, or cell type; leukaemia treatment outcome and whether a patient has chronic fatigue syndrome. In the case of evaluating information content by each of the three methodologies, the same univariate method, consisting of infoGain, was used. The only experimental difference in using the three feature selection methodologies was the use of trustworthiness by FUR1 and FUR2. However FUR1 used signal-to-noise, while FUR2 used a noise-only approach for evaluating data quality.

The synthetic data set provided a simplified example of a non-classical data set consisting of significant redundancy and noise. This simplified data set was used in order to facilitate analysis. The real-world data sets provided a good example of a non-classical problem; they consisted of more than 9,000 features and less than 200 samples, which were affected by significant noise.

These data sets were all prepared using the approach described in section 4.3.

The evaluation of the methodologies consisted of classification accuracy, analysis of the variation of feature trustworthiness across data folds and two different measures of feature set quality. Heat maps provided the first measure of feature set quality, which enabled a qualitative appraisal of the clarity of class separation. The second measure of quality consisted of quantitatively appraising the biological meaningfulness of the selected features. The second measure of quality was only applied to leukaemia cell type classification, since it provides the best source of known gene meaningfulness with respect to the disease.

4.8.1 Classifier accuracy

Regression models were built using the synthetic data set and the results reported in tables 4.3 and 4.4, shown respectively on pages 163 and 165. Although the results are not statistically significant, both FUR1 and FUR2 provided a more accurate model than achieved with traditional feature selection. Feature Utility Ranking 1 gave marginally better accuracy than Feature Utility Ranking 2, which was initially unexpected given that the noise affecting the test set was known to be different than in the training set. This result shows that a signal-to-noise measure of data quality is preferable for this type of data set, which was significantly affected by noise. Given that equation 4.4

on page 147 determined the impact of noise for each feature in the data set, the experimental result appears consistent with the fact that individual features were predominantly affected by a constant noise factor. This constant noise factor would be effectively evaluated using a signal-to-noise measure of just the training set.

A summary of the real-world data set results are shown in table 4.25. Three of the experiments used the entire available feature set, which provided an accuracy reference since feature selection was not required. For five of the six remaining experiments, FUR provided an improvement or similar accuracy to that achieved by the traditional methodology. In one experiment, the traditional methodology provided a 6.5% improvement over FUR. These results suggest that FUR is able to provide similar or improved accuracy when compared to a traditional feature selection methodology. The one instance where FUR was 6.5% less accurate was the 1,024 feature treatment outcome data set. In this instance the traditional methodology was notably more accurate than for each of the other experiments. It is possible that apparent information was responsible for the accuracy achieved by the traditional methodology. But it is worth noting that the involvement of genetics in predicting treatment outcome is much more complex than for cell type classification, for which FUR performed well.

Table 4.25 shows that FUR1 provided high accuracy for cell type clas-

Table 4.25: Summary of the real-world data set experimental results. The columns contain the stratified ten-fold cross validation averages gained by the respective feature selection methodologies and * identifies the approximately highest accuracy.

Data set	Features	Traditional	FUR1	FUR2
B versus T-Cell	16	96.97% *	96.46% *	90.39%
B versus T-Cell	1,024	90.91%	97.02% *	91.17%
B versus T-Cell	9,485	98.78%	98.78%	98.78%
Treatment Outcome	16	92.06% *	90.00%	92.06% *
Treatment Outcome	1,024	96.00% *	88.33%	89.68%
Treatment Outcome	9,485	88.89%	88.89%	88.89%
Chronic Fatigue Syndrome	16	69.80%	71.50% *	69.80%
Chronic Fatigue Syndrome	1,024	72.70% *	72.70% *	65.10%
Chronic Fatigue Syndrome	9,485	59.30%	59.30%	59.30%

sification, which could be due to an essentially equal distribution of noise across the training and test sets, or the need to consider the signal-to-noise ratio, as was the case for the synthetic data experiments. Although FUR2 out performed FUR1 for the treatment outcome data set, traditional feature selection provided the best accuracy. This result may be explained by the presence of artifacts within the data set, refer figure 4.7 on page 175. Figure 4.7 showed that overall classification accuracy was unaffected by randomly reducing the size of the feature set, which is in stark contrast for the cell type data set result in figure 4.5 on page 171. This suggests that artifacts, hence noise, were responsible for the accuracy achieved by traditional feature selection. If this is correct, then this may explain why FUR2 provided better accuracy than FUR1, since it enabled the quality of the entire

data set to be considered. For that particular experiment, the management of noise by FUR resulted in poorer accuracy than achieved by the traditional methodology. This result may be caused by noise enabling the traditional methodology to serendipitously locate features, amongst the few samples, that provided better accuracy.

4.8.2 Variation of feature trustworthiness

It is well known that the reapplication of feature selection across different data folds, particulary for a non-classical data set, results in the selection of different features (Ein-Dor et al., 2005). A number of factors may be responsible for such a variation: the dimensionality of the data set, the small number of samples and the effects of noise. If noise is a major factor for this variation, the use of data quality within feature selection should provide a reduction in such variations.

The experiments identified substantial variations in trustworthiness across different data folds, refer figure 4.15 on page 188. Figure 4.15 shows the sort of variation in quality that traditional feature would have been subjected to. While FUR's knowledge of data quality enables it to select features of consistent quality, such as in figure 4.11 on page 181. This variation agrees with the issues raised by Ein-Dor et al. (2005) and shows that using trustworthiness reduces such variations. Figures 4.18 and 4.24, found on pages 195 and 208,

shows the sensitivity provide to FUR1 and FUR2 respectively, in response to changes in data quality.

A different perspective on the issue of data quality was presented using three different genes, which had a calculated rank of 1; 2,500 and 9,485 according to trustworthiness. Figures 4.11, 4.12 and 4.13, respectively shown on pages 181; 184 and 185, clearly show a deterioration in spot quality as the rank of the gene increases in numerical value. Trustworthiness also assists in locating poor quality arrays and poor quality spots, an example of which is shown in figure 4.9 on page 179.

4.8.3 Clarity of the class separation

Heat maps provide a qualitative basis for comparing the ability of a feature selection methodology to produce a clear separation between classes, refer for example to figures 4.4 and 4.16 (shown respectively on pages 169 and 190). Figures 4.4 and 4.16 were generated respectively using traditional and FUR1 feature selection and show that FUR1 provided a much clearer class separation. With respect to the entire set of heat maps generated, the class boundary provided by FUR1 and FUR2 is either similar or superior to that provided by traditional feature selection.

For the B versus T Cell classification experiments, traditional selection provided a weaker separation in comparison to FUR1, while FUR1's heat

map possibly provided the best example of class separation for the entire set of experiments. Overall FUR2 also provided good separation. However all of the methodologies had trouble generating a clear separation for leukaemia treatment outcome and chronic fatigue syndrome classification. This difficulty is possibly due to the complex nature of leukaemia treatment outcome prediction and the understanding that chronic fatigue syndrome is unrelated to molecular biology. It is suggested that this qualitative measure of class separation is proportional to the variation of quality in each data point. Since FUR inherently minimizes the variation in data quality by maximizing feature trustworthiness—it is expected that features which provide a more convincing separation will be preferred. Alternatively traditional feature selection is purely focused on classification accuracy and therefore is more likely to be negatively influenced by artifacts in the data set and consequently producing heat maps with poorer class separation.

4.8.4 Biological relevance

The goal of a traditional feature selection methodology is the identification of a feature set that maximizes the correlation between the features and the designated class labels. However this goal makes no distinction between the underlying structure of the data and noise. As a result, a traditional methodology may achieve greater accuracy than FUR, particulary in the

context of a non-classical problem—but at the expense of the meaningfulness of the selected features. If this line of thought is correct, then this argument provides a explanation for the similar accuracy achieved by the traditional and FUR methodologies. More importantly however, it also provides an explanation for the improved meaningfulness of the features selected by FUR. It also may explain a small loss in accuracy by FUR, but in exchange for extracting greater insight into the underlying structure of the data set. This last thought is clearly supported by the research findings, since in the case of leukaemia cell type classification, FUR1 achieved higher accuracy than FUR2—however FUR2 provided greater biological meaningfulness. While the traditional feature selection methodology achieved the highest accuracy, the features it selected provided little or no biological meaningfulness.

Chapter 5

Conclusion

This thesis investigated the impact of a non-classical problem domain on feature selection. In particular this thesis considered whether including a measure of *data quality* within the feature selection process will result in a more accurate model, where accuracy is measured by the number of correct classifications, or in the case of regression, reducing prediction errors. Among other things—the thesis showed that *only* focusing on accuracy may *not* result in the selection of meaningful features, with respect to the underlying structure of the phenomenon being modeled.

5.1 Problem significance

This thesis categorizes feature selection into two problem domains: classical and non-classical. The "classical problem domain" is characterized by many samples, few dimensions (or features) and a degree of noise, while the "non-classical problem domain" is characterized by few samples, many features

and significant noise.

From the literature reviewed, it is clear that traditional feature selection methodologies are well suited to the classical problem domain. Traditional feature selection ranks features according to the amount of information they provide and assumes that the underlying structure for each feature can be discovered, given the available sample data. However this thesis sought to explore whether non-classical problems can be a source of significant "apparent information", which consequently misleads traditional feature selection into identifying features that provide a poor generalization. Apparent information exists when the calculated information of a feature exceeds what is present in the underlying structure. The likelihood of apparent information occurring increases in response to decreasing sample size, increasing dimensionality and increasing noise within the data set.

The impact of noisy data on learning was theoretically considered in section 3.2, which showed that the required sample size for successful learning increases dramatically as noise increases. As a result, the number of samples required because of noise can render a non-classical problem unlearnable because of the number of samples that are typically available. This was explored and confirmed in section 3.3, using a subset of the 1994 American Census data set.

5.2 Research outcomes

As a result of investigating how a measure of data quality can be incorporated into the feature selection process. The developed approach involves splitting feature selection into two steps, where each ranks features according to a different measure of utility and then selects a subset. The *first* step ranks features according to their trustworthiness and produces a subset that consists of the n most trusted features. The *second* step ranks the n trusted features according to the amount of discriminative information they provide and selects the m most discriminative features, where $n > m$. Two variations of the feature selection process, or methodology that incorporates data quality, was developed and presented: Feature Utility Ranking 1 (FUR1) and Feature Utility Ranking 2 (FUR2). Finally the research consisted of empirically evaluating the developed methodologies by contrasting them with a traditional feature selection methodology.

The process of developing FUR1 and FUR2 was decomposed into five objectives. The *first* consisted of using PAC theory in order to evaluate the impact of noise on model learning, particularly in the context of a limited sample size. This showed that increasing noise can quickly prevent successful learning of a non-classical problem. As a result it was deemed necessary to avoid the elimination of sample data.

One of the proposed solutions highlighted in the literature, involves the use of a noise model in order to detect the underlying signal, in the absence of sufficient samples. Without the aid of a noise model, signal detection would otherwise be problematic in the context of a non-classical problem. Using PAC theory and the experiments conducted, concerns surfaced regarding the possible accuracy of a noise model using current approaches. This and other concerns provided a motivation to investigate the use of a simplified noise model that can be estimated more easily for a non-classical problem. The simplified model consists of a *worst case estimate* of the noise present within each feature. Using this type of noise model, an approach for calculating a feature's trustworthiness was developed in order to eliminate features that fall below a threshold of trustworthiness, rather than eliminating samples on the basis of their quality.

The use of a worst case estimate for noise is related to the *second* research objective, refer page 6, which consists of a framework for estimating data quality. Because of the characteristic few samples found in a non-classical problem, this research considered the use of a worst case noise model to be preferable in comparison to the noise models presented in the literature. The approach of using a worst case noise model depends on the presence of many features and the presence of significant redundancy, which is another characteristic of non-classical problems.

The *third* objective provided a framework for merging feature trustworthiness with a feature's discriminative ability, in order to produce a subset of features that are trustworthy and discriminative of class labels. This framework was necessary since one cannot directly compare the trustworthiness and discriminative ability of feature.

The *fourth* objective consisted of evaluating FUR1 and FUR2, as well as contrasting them against a traditional feature selection methodology. This evaluation involved the use of synthetic and real-world data sets. The primary real-world data sets consisted of gene expression data, which is an exemplary example of non-classical problems. Using heat maps, these experiments clearly showed qualitatively, that FUR1 and FUR2 produced a superior distinction between class labels in comparison to traditional feature selection; for example refer to figures 4.4 and 4.16, on pages 169 and 190 respectively. The experiments also showed that data quality, which is related to feature trustworthiness, varies considerably across data folds, see figure 4.18 on page 195.

The experiments also contrasted the biological meaningfulness of the genes that were selected by each of the methodologies. This showed a significant difference in the meaningfulness of the selected genes, refer table 4.24 on page 222 and figure 4.30 on page 223. The results suggest that the genes generated by traditional feature selection have no known biological relevance to

B versus T Cell classification. However FUR1, which uses a signal-to-noise measure of data quality, provided genes with known biological relevance. While FUR2's consideration of the quality of the test set, resulted in the selection of the most biologically meaningful genes. The empirical results also showed that accuracy is inversely proportional to biological meaningfulness. Therefore a small reduction in classification accuracy was associated with dramatically improved biological meaningfulness.

The *fifth* objective of the thesis was the development of a feature selection methodology that provides a mechanism for also evaluating data quality during data cleaning. This mechanism is particularly valuable because data cleaning is customarily decoupled from feature selection and therefore any insight gained about the quality of the data is *not* exploited by feature selection. Furthermore, information about how data is altered as a result of data cleaning is also a valuable source of quality information.

5.3 Contributions to knowledge

The five contributions to knowledge are: (i) two approaches for estimating data quality, (ii) a definition and measure of feature trustworthiness, (iii) a definition and measure of feature utility, (iv) two feature selection methodologies based on feature utility and (v) a mechanism for using data quality information collected during data cleaning.

5.3.1 Two approaches for estimating data quality

The first contribution to knowledge, in sections 3.6.3, 3.7.1 and 3.8.1, involves two approaches for estimating data quality. A signal-to-noise and a noise-only approach were developed in order to estimate the quality of individual instances of individual dimensions of a sample point. Using signal-to-noise acknowledges that the amount of noise that can be tolerated is determined by the strength of the signal. However the use of signal data prohibits evaluating the quality of the test and unlabeled data sets. Alternatively, the noise-only approach permits the use of the *entire* data set in the evaluation of data quality.

Experimentation in chapter 4 showed that the signal-to-noise approach generated the most accurate models in most circumstances. This result suggests the use of signal strength has an important role in evaluating data quality. But it is possible that the test data, which contained very few samples, serendipitously contained data whose quality was equal or superior to that of the training set. Nonetheless, there were instances when the noise-only approach provided the most accurate model. Regardless of the approach used for calculating data quality, the experiments showed that data quality varies considerably and this may explain why researchers, such as Ein-Dor et al. (2005), had difficulty in finding a unique feature set.

One intention of this contribution is encouraging the use of prior knowledge in determining how noise is measured and in the case of signal-to-noise, the method for combining signal and noise. This contribution also enables the use of quality measures developed by other researchers, in the case of microarrays for example, the quantitative quality control developed by Wang et al. (2001).

5.3.2 Definition and measure of feature trustworthiness

The second contribution to knowledge is the definition and measure of feature trustworthiness, as given in section 3.6.4. An overall measure of quality is required because each feature has its own measure of quality due to the sample data. A feature's "trustworthiness" is a scalar value that provides this overall measure of quality.

The motivation for trustworthiness comes from the need to better utilize the few samples that are typical of non-classical problems. Rather than focusing on eliminating poor quality samples, this thesis proposed eliminating features on the basis of their trustworthiness. The benefit of this approach is to maximize the number of samples used, while reducing dimensionality.

Trustworthiness also provides a measure for graphing the distribution of quality of a feature set. Section 3.7.1 introduces this distribution of feature

quality known as a "trustworthiness curve", which provides a basis for identifying which features should be eliminated. The experimental results, in chapter 4, contain a number of trustworthiness curves with the same characteristic shape. This shape shows that all the real-world data sets used in this thesis are composed of three groups of features: a small number that are exceedingly trustworthy, a group with a gradual change in trustworthiness and a small group providing a rapid decline in trustworthiness. In particular, the experimental results for Chronic Fatigue Syndrome were published in the book chapter Kennedy et al. (2008).

The characteristic shape of the trustworthiness curve was evident regardless of using a signal-to-noise or a noise-only measure of data quality. Consequently the shape appears to be primarily determined by the noise-only component for measuring data quality. However the best improvements in accuracy and feature meaningfulness, in the experimental results in section 4.8, were obtained with a signal-to-noise measure of quality.

Section 4.5.2 contains a variety of graphs showing the impact on feature trustworthiness due to changes in the data. The data set changes were caused by using stratified ten-fold cross validation. This shows that the measure of data quality and trustworthiness used were sensitive to changes in the data set and that the trustworthiness of a feature can vary dramatically. Specifically, the graphs were the product of two explorations: the quality

differences between three features, as while as the changes in trustworthiness for every feature across the ten data sets. The construction of these graphs also revealed issues with six microarrays and it was determined that these arrays were all constructed within two months of each other. This finding parallels the quantitative quality control work by Wang et al. (2001). The sensitivity of trustworthiness changes in the data set also agrees with the findings of many researchers, who have found that small changes in a data set often result in significant changes in the features selected. These and other findings infer that trustworthiness implicitly searches on the basis of feature meaningfulness, rather than just classification accuracy.

5.3.3 Definition and measure of feature utility

The third contribution to knowledge is the definition and measure of feature utility, given on page 82. Feature utility is calculated using two parameters: the information provided by the feature and its trustworthiness. Since feature utility is a scalar, features can therefore easily be ranked according to their information **and** their trustworthiness.

Of all the classifiers evaluated in chapter 4, all but one of the most accurate models were constructed using feature utility. This result supports the theses' assertion that *apparent information* can occur and that features which provide a better generalization are selected through the use of feature

utility.

5.3.4 Two feature selection methodologies based on feature utility

The fourth contribution to knowledge is the two feature selection methodologies, Feature Utility Ranking 1 (FUR1) and Feature Utility Ranking 2 (FUR2), described in sections 3.7 and 3.8 respectively. The common goal of these methodologies is to eliminate redundant features whose utility falls below a predetermined level. This elimination process involves two separate steps. The first eliminates features that fall below a trustworthiness threshold, followed by selecting a subset of features ranked according to the information they provide. The difference between the two methodologies is the measure of data quality: FUR1 uses signal-to-noise and FUR2 only uses noise.

The experiments in section 4.4 used a synthetic data set, which allowed the amount of noise to be controlled. The overall noise in the data set consisted of two parts: random noise and noise that was *consistent* for each feature. This provided an idealized environment in which the noise that was *consistent* and specific to each feature, was used as a direct measure of data quality. This experiment showed that Feature Utility Ranking, with minor changes, can be used for regression, as well as classification problems. Both

FUR1 and FUR2 provided a more accurate model than traditional feature selection, with FUR1 providing the highest accuracy. Although the results are not statistically significant, they still suggest that a measure of data quality results in improved prediction accuracy.

Experiments in sections 4.5 and 4.6 (using the leukaemia and chronic fatigue syndrome data sets) are generally consistent on whether FUR1 or FUR2 provides the highest accuracy, regardless of the number of features used. For the B versus T Cell data set, FUR1 provided higher accuracy than FUR2. For all of the B versus T Cell data set experiments, FUR1 either outperformed or approximately matched the traditional and FUR2 methodologies. In the case of the treatment outcome data set, FUR2 provided better accuracy than FUR1. For the 16 feature model, FUR2 matched the accuracy of the traditional methodology. However the traditional methodology provided the highest accuracy for the 1,024 feature model. FUR1 provided the best overall accuracy for the chronic fatigue syndrome data set. Experimental results for FUR1 were published in the book chapter Ubaudi et al. (2009).

5.3.5 Mechanism for using data quality information collected during data cleaning

The fifth contribution to knowledge, which stems from the fifth research objective in section 1.2.1, is a mechanism for using data quality information

collected during data cleaning. Based on the review of the literature, this thesis argued that data cleaning is decoupled from feature selection, other than supplying cleaned data. Consequentially, feature selection implicitly assumes that every instance of sample data has the same level of quality. However the degree of change applied by the cleaning process depends on individual sample points. Furthermore, changes can also vary for each dimension of a sample point and as a result, the change can be directly interpreted as a measure of noise. Therefore the feature selection methodologies developed in this thesis can easily be coupled with the data cleaning process.

5.4 Further work

The following areas for further work are suggested: (i) modeling the occurrence of apparent information, (ii) explicitly involving a measure of feature redundancy, (iii) exploring the use of data quality information collected during data cleaning, (iv) significantly increasing the number of experiments, (v) the development of methods for using a noise vector and (vi) a deeper investigation of feature relevance or meaningfulness; this is related to area (i).

5.4.1 Modeling occurrence of apparent information

A primary justification for Feature Utility Ranking is the occurrence of apparent information, which arises when the amount of information provided

by a feature exceeds that which is present within the underlying structure. A process similar to that followed in the "theoretical impact of noise on sample size" in section 3.2 and the "proof of principle for the impact of noise" in section 3.3, would provide an understanding of the frequency of apparent information occurring and when the increase becomes statistically significant. An approach would be the development of a theoretical model of the relationship between class label errors and a feature's measured information, for the purpose of determining the threshold of class label errors required to produce a statistically significant increase in measured information. Using a synthetic data set would be ideal for validating the theoretical model.

5.4.2 Explicitly involve a measure of feature redundancy

The amount of feature redundancy present within the data set is not explicitly considered by the version of Feature Utility Ranking described. One improvement would be to construct sets of redundant features for the data set and then determine a trustworthiness threshold for each set, thereby removing all but the most trusted feature in each set.

5.4.3 Exploring the use of data quality information collected during data cleaning

Empirically evaluate the use of data quality information collected during data cleaning. Use a linear regression model of a multi-dimensional synthetic data set containing noise. Build and compare two versions of the model where the first involves classical data cleaning and model construction using least-squares regression. Secondly repeat the process using Feature Utility Ranking and a measure of error for each dimension (trustworthiness), where the trustworthiness is the least-squared error. The entire experiment could be repeated using a more sophisticated model, such as LOWESS (locally weighted scatterplot smoothing).

5.4.4 Significantly increasing the number of experiments

Although the experiments reported in this thesis suggested that Feature Utility Ranking increased model accuracy, the results were not statistically significant. Therefore the experiments performed need to be extended so that a statistically significant outcome is facilitated. The suggested approach is to repeat the experiments described, except for using $N-1$ data folds. This approach will also provide a way for contrasting the degree of variation in the feature sets selected by the traditional and FUR methodologies.

5.4.5 Methods for using a noise vector

Section 3.6.2 provides a conceptual description for the use of a noise vector containing more than one unrelated entry. However an approach for combining the entries in the noise vector is required. In the case of the rainfall example in section 3.6.2, none of the vector entries can be compared and combined into a single scalar value representing a measure of noise. A possible solution involves the use of summary statistics, where for example, the likely error in each entry is calculated and then combined into a scalar using a geometric mean.

5.4.6 Deeper investigation of feature relevance

Section 4.7 showed that the features chosen by FUR were more biologically meaningful than those chosen by the traditional methodology. It was also found that this increase in meaningfulness was often associated with a small lose in accuracy when compared to the traditional methodology. Feature Utility Ranking's ability in this respect needs further testing using a synthetic data set, where noise is used to determine the degree of meaningfulness of each feature. Perform $N-1$ fold testing using the synthetic data, in order to establish whether a statistically significant difference exists between the three feature selection methodologies, with respect to the selection of meaningful features.

5.5 Concluding remarks

A long standing problem in data mining has been the issue of data quality and how to manage its impact on feature selection and model construction. In 1986 Quinlan performed a number of experiments specifically regarding the impact of noise on feature selection and model learning in a classical problem domain. As a result of his experiments, Quinlan identified two major issues that impact the accuracy of a model: (i) the impact of noise on the test data and (ii) differences in quality of the training, test and unlabeled data sets. Although Quinlan viewed these issues as a significant problem—he was unable to suggest a possible solution.

In 2005 and 2007, Ein-Dor et al. and Saeys et al., respectively, described a problem that has plagued researchers who build classifiers and prediction models in non-classical problem domains—the inability to select a unique feature subset. This is principally caused by problems with data quality. Since different subsets of the data are impacted by different quality issues, this causes changes in feature trustworthiness. The research presented in this thesis inherently addresses this issue since the selection of trustworthy features promotes the selection of a unique feature subset.

The findings of Quinlan, Ein-Dor et al. and Saeys et al. are evidence of the difficulty that exists with managing the impact of varying data quality

within the feature selection process. This thesis developed and tested two feature selection methodologies, specifically FUR1 and FUR2, that address both of the issues raised by Quinlan and the difficulty of identifying a unique feature set.

The approach developed in this thesis to assess the trustworthiness of features and enhance feature selection sets the stage for further improvements in classifier accuracy, the identification of a unique feature set and the selection of features that are better aligned with the underlying structure of data set. A primary source of improvement involves a method for selecting a trustworthiness threshold that eliminates features containing excessive amounts of apparent information. A further improvement is the identification of subsets of features, where each subset consists of redundant features and only the most trustworthy feature in each subset is used. The methods in this thesis will contribute to improved understanding in non-classical problem domains such as the gene expression microarray domain studied here, as well as sound and speech analysis, face detection and recognition and MRI brain imaging.

Although FUR achieved similar accuracy to traditional feature selection, there were clear differences in the selected feature sets. Heat maps, which provide a method for visualizing the inherent structure in the data, were one of the methods used to understand the nature of these differences. Using heat maps, it became apparent that FUR consistently provided a superior

delineation between class labels, which suggests that FUR was more effective in uncovering the underlying structure of the data and was therefore less affected by noise. Heat maps generated using traditional feature selection tended to lack any apparent structure. Furthermore it was shown that features selected by FUR were far more aligned with known biological meaning than the features obtained using a traditional feature selection methodology.

Given that traditional feature selection is entirely driven by the search for information hidden within the data set, the findings of this thesis are not surprising. A more accurate understanding of the phenomenon being modeled is achieved by a feature selection methodology that incorporates a measure of data quality.

Appendix A

Synthetic data experimental results

The synthetic data experimental results presented in section 4.4.1 used correlation coefficients with a changing numerical sign, since it was not clear whether the absolute value should be used. A brief reanalysis of that section is repeated here, except that the absolute value of correlation coefficients is used. Therefore the material presented within this section should be read in conjunction with section 4.4.1.

Table A.1 shows that FUR1 was outperformed by FUR2, as was expected. It is noteworthy that the difference between the experimental means, for the results presented in tables A.1 and 4.3, is approximately the same. So although FUR2 out performed FUR1, as shown in table A.2, the Null hypothesis must still be accepted that FUR1 and FUR2 are no different. In addition, the null hypothesis must also be accepted that neither version of FUR is superior to the noisy traditional alternative.

Table A.1: The alternate results of the synthetic data experiments, where *mean* and *std dev* are the mean and the standard deviation of the correlation coefficients for ten-fold cross validation.

Fold	Noiseless Traditional	Noisy Traditional	Noisy FUR1	Noisy FUR2
1	0.9189	0.5105	0.5222	0.4305
2	0.9289	0.0881	0.0860	0.2123
3	0.9201	0.2602	0.4630	0.5343
4	0.8729	0.2874	0.0888	0.0321
5	0.9420	0.0365	0.0237	0.0785
6	0.9746	0.0281	0.0750	0.1497
7	0.9998	0.6137	0.4185	0.3447
8	0.7840	0.0406	0.0527	0.0946
9	0.8256	0.0293	0.0413	0.0973
10	0.9119	0.3801	0.1610	0.5286
mean	0.9079	0.2275	0.1932	0.2503
std dev	0.0618	0.2067	0.1846	0.1833

Table A.2: Using the *mean* correlation confidences in table A.1, the p-value for permutations of the means are shown. The column labeled "Sig", when populated with an $*$, states the result is statistically significant. The last row for example, shows that the alternate hypothesis (H_1) that FUR2 is more accurate than FUR1 can only be accepted with 48.11% confidence.

$mean_1$	$mean_2$	p-value	Sig
Noiseless Traditional	Noisy Traditional	≈ 0	*
Noiseless Traditional	Noisy FUR1	≈ 0	*
Noiseless Traditional	Noisy FUR2	≈ 0	*
Noisy Traditional	Noisy FUR1	0.715	
Noisy Traditional	Noisy FUR2	0.807	
Noisy FUR1	Noisy FUR2	0.519	

Appendix B

Glossary

Benign tumor consists of neoplastic cells which have remained clustered together in a single mass.

Complete Clinical Remission (CCR) is achieved if a patient has not relapsed for a defined time period. In the context of childhood leukaemia, medical experience has shown that a patient is unlikely to relapse after an event free period of 5 years.

cDNA is a DNA molecule made as a copy of messenger RNA and therefore lacking the introns that are present in genomic DNA. cDNA clones represent DNA cloned from cDNA and a collection of such clones, usually representing the genes expressed in a particular cell type or tissue, is a cDNA library

Chromosome "is a structure composed of a very long DNA molecule and associated proteins that carries part (or all) of the hereditary informa-

tion of an organism. Especially evident in plant and animal cells undergoing mitosis or meiosis, where each chromosome becomes condensed into a compact rodlike structure visible under the light microscope"

Expression is the "production of an observable phenotype by a gene, usually by directing the synthesis of a protein"

Leukaemia Any of a group of malignant diseases in which increased numbers of certain immature or abnormal leucocytes are produced. This leads to increased susceptibility to infection, anaemia and bleeding. Other symptoms include enlargement of the spleen, liver and lymph nodes. Leukaemias may be acute or chronic depending on the rate of progression of the disease. They are also classified according to the type of white cell that is proliferating abnormally (see terminology; lymphoblast, myeloblast and myeloid leukaemias). Leukaemias are treated with radiotherapy or cytotoxic drugs.

There are two major classes of leukaemia, Acute Myeloid Leukaemia (AML) and Acute Lymphoblastic Leukaemia (ALL), and these classes are defined by the cell type involved.

ALL which corresponds to a lymphoblast, is an abnormal blood cell that is present in the blood and blood-forming organs in a type of leukaemia.

During the normal case, the leucocyte cell is a variety of white blood cell present within the lymph nodes, spleen, thymus gland, gut wall and bone marrow. These cells are involved in immunity and can be subdivided into B-lymphocytes and T-lymphocytes which are primarily responsible for cell-mediated immunity.

ALL B-Cell The B-lymphocyte, or ALL B cell for short, is a further subtype of ALL leukaemia. The B-lymphocyte cell produces circulating antibodies. Refer to glossary entry for B-Cell for further details.

ALL T-Cell The T-lymphocyte, or ALL T cell for short, is a further subtype of ALL leukaemia. The T-lymphocyte cells are primarily responsible for cell-mediated immunity. T-lymphocytes can be further differentiated into helper, killer or suppressor cells. Refer to glossary entry for T-Cell for further details.

AML is a variety of leukaemia in which the type of blood cell that proliferates abnormally originates within the blood-forming tissue of the bone marrow. This blood-forming tissue is responsible for a number of classes of blood cells.

Malignant tumor Tumors are *only* known as cancer when malignant, for

they then invade surrounding tissue.

Metastases are secondary tumor

mRNA or messenger RNA, or messenger RiboNucleic Acid.

Myeloid like, derived from, or relating to bone marrow.

Nucleotide compound that consists of a nitrogen-containing base linked to a sugar and phosphate group

Bibliography

Kjersti Aas. Microarray Data Mining: A Survey. Technical report, Norsk Regnesental, Norwegian Computing Center, January 2001.

C. J. Adcock. Sample size determination: A review. *The Statistician*, 46: 261–283, 1997.

N. Afari and D. Buchwald. Chronic Fatigue Syndrome: A review. *American Journal of Psychiatry*, 160:221–236, 2003.

Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Molecular Biology of the Cell*. Garland Science, New York, fourth edition, 2002.

David B. Allison, Xiangqin Cui, Grier P. Page, and Mahyar Sabripour. Microarray data analysis: from disarray to consolidation and consensus. *Nature reviews, Genetics*, 7(1):55–65, 2006.

Elsa Angelini, Yinpeng Jin, and Andrew Laine. State of the Art of Level Set Methods in Segmentation and Registration of Medical Imaging Modalities. In Jasjit S. Suri, David L. Wilson, and Swamy Laxminarayan, editors, *Handbook of Biomedical Image Analysis*, Topics in Biomedical Engineering. International Book Series, pages 47–101. Springer US, 2005.

Dana Angluin and Philip Laird. Learning From Noisy Examples. *Machine Learning*, 2(4):343–370, 1988.

M. Anthony and N. Biggs. *Computational learning theory: An introduction*. Cambridge University Press, Cambridge England, 1992.

Virginie M. Aris, Michael J. Cody, Jeff Cheng, James J. Dermody, Patricia Soteropoulos, Michael Recce, and Peter P. Tolias. Noise filtering and nonparametric analysis of microarray data underscores discriminating markers of oral, prostate, lung, ovarian and breast cancer. *BMC Bioinformatics*, 5 (185):9, 2004.

Pierre Baldi and Sören Brunak. *Bioinformatics: The Machine Learning Approach*. A Bradford Book: The MIT Press, Cambridge, Massachusetts, second edition, 2001.

Pierre Baldi and Wesley G. Hatfield. *DNA Microarrays and Gene Expression: From Experiments to Data Analysis and Modelling*. Cambridge University Press, Cambridge, 2002.

Pierre Baldi and Anthony D. Long. A Bayesian Framework for the Analysis of Microarray Expression Data: Regularized t-test and Statistical Inferences of Gene Changes. *Bioinformatics*, 17:509–519, 2001.

Daniel P. Berrar, Martin Granzow, and Werner Dubitzky. Introduction to Genomics and Proteomic Data Analysis. In Werner Dubitzky, Martin Granzow, and Daniel P. Berrar, editors, *Fundamentals of Data Mining in Genomics and Proteomics*, Science+Business Media, pages 1–37. Springer, New York, 2007.

Rosa Blanco, Pedro Larranaga, Inaki Inza, and Basilio Sierra. Gene selection for cancer classification using wrapper approaches. *International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI)*, 18(8):1373–1390, 2004.

A. L. Blum and R. L. Rivest. Training a 3-node neural network is NP-complete. *Neural Networks*, 5:117–127, 1992.

Avrim L. Blum and Pat Langley. Selection of Relevant Features and Examples in Machine Learning. *Artificial Intelligence*, 97:245–271, 1997.

Trond Hellem Bo and Inge Jonassen. New feature subset selection procedures for classification of expression profiles. *Genome Biology*, 3(4):11, 2002.

B. M. Bolstad, R. A. Irizarry, M. Astrand, and T. P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003.

Benjamin Bolstad. Pre-Processing DNA Microarray Data. In Werner Dubitzky, Martin Granzow, and Daniel P. Berrar, editors, *Fundamentals of Data Mining in Genomics and Proteomics*, Science+Business Media, pages 51–78. Springer, New York, 2007.

A. Borzenko. Analytical Instrumentation: A Guide to Laboratory, Portable and Miniaturized Instruments. *Analytical Chemistry*, 66(6):648–649, 2011.

Rainer Breitling, Patrick Armengaud, Anna Amtmann, and Pawel Herzyk. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Letters*, 573 (1-3):83–92, 2004.

Michael P. S. Brown, William Noble Grundy, David Lin, Nello Cristianini, Charles Sugnet, Manuel Jr. Ares, and David Haussler. Support Vector Machine Classification of Microarray Gene Expression Data. Technical Report UCSC-CRL-99-09, University of California, Santa Cruz, 12/06/1999 1999.

Nader H. Bshouty, Nadav Eiron, and Eyal Kushilevitz. PAC Learning with Nasty Noise. *Theoretical Computer Science*, 288:1–18, 1999.

W. Buntine. *A Theory of Learning Classification Rules*. Phd, University of Technology, Sydney, 1990.

Christopher J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

Bruce A. Carlson. *Communication Systems, An Introduction to Signals and Noise in Electrical Communication*. McGraw-Hill Education, second revised edition, 1975.

Alison Cawsey. *The essence of artificial intelligence*. The essence of computing series. Prentice Hall, Hertfordshire, 1998.

E. C. Cherry. Some experiments on the recognition of speech, with one and two ears. *Acoustic Society of America*, 25:975–979, 1953.

W. S. Cleveland and S. J. Devlin. Locally-weighted regression: An approach to regression analysis by local fitting. *American Statistics Association*, 83: 596–610, 1988.

Dennis R. Cook and Sanford Weisberg. *Applied Regression Including Computing and Graphics*. Probability and Statistics. Wiley Series in Probability and Statistics, 1999.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 1991.

M. Dash and H. Liu. Feature selection for classification. *Intelligent Data Analysis*, 1(3):131–156, 1997.

Beth Dawson and Robert G. Trapp. *Basic & Clinical Biostatistics*. Health Professions. McGraw-Hill Higher Education, Singapore, third edition, 2001.

R. Diaz-Uriarte and S. Alvarez de Andres. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7(3), 2006.

C. Ding and H. Peng. Minimum redundancy feature selection from microarray gene expression data. *IEEE Conference on Computational Systems Bioinformatics*, pages 523–529, 2003.

Edward R. Dougherty. Small sample issue for microarray-based classification. *Comparative and Functional Genomics*, 2:28–34, 2001.

Charalampos Doukas and Ilias Maglogiannis. A Fast Mobile Face Recognition System for Android OS Based on Eigenfaces Decomposition. In Harris Papadopoulos, Andreas Andreou, and Max Bramer, editors, *Artificial Intelligence Applications and Innovations IFIP Advances in Information and Communication Technology*, volume 339, pages 295–302. Springer Boston, 2010.

Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. Wiley Interscience Publication (John Wiley and Sons), second edition, 2001.

S. Dudoit, Y. H. Yang, M. J. Callow, and T. P. Speed. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, 12(1):111–139, 2002a.

Sandrine Dudoit and Jean Yee Hwa Yang. Bioconductor R Packages for Exploratory Analysis and Normalization of cDNA Microarray Data. In Giovanni Parmigiani, Elizabeth S. Garrett, Rafael A. Irizarry, and Scott L. Zeger, editors, *The Analysis of Gene Expression Data: Methods and Software*, pages 73–101. Springer-Verlag, New York, 2003.

Sandrine Dudoit, Jane Fridlyand, and Terry Speed. Comparison of discriminant methods for the classification of tumors using gene expression data. *Journal of American Statistical Association*, 97:77–87, 2002b.

Margaret H. Dunham. *Data Mining: Introductory and Advanced Topics*. Prentice Hall, New Jersey, 2003.

Bradley Efron, Robert Tibshirani, John D. Storey, and Tusher Virginia. Empirical Bayes Analysis of a Microarray Experiment. *Journal of American Statistical Association*, 96(456):1151–1160, 2001.

Liat Ein-Dor, Itai Kela, Gad Getz, David Givol, and Eytan Domany. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, 21(2):171–178, 2005.

W.J. Emery, D. Baldwin, and D.K. Matthews. Maximum Cross Correlation Automatic Satellite Image Navigation and Attitude Corrections for Open Ocean Image Navigation. *Geoscience and Remote Sensing*, 41:33–42, 2003.

R. Fox and M. Dimmic. A two-sample Bayesian t-test for microarray data. *BMC Bioinformatics*, 7(126):1–11, 2006.

Yoav Freund. Boosting a weak learning algorithm by majority. *Information and Computation*, pages 1–50, 1995.

Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Computer and System Sciences*, 55(1):119–139, 1997.

Yoav Freund and Robert E. Schapire. A Short Introduction to Boosting. *Japanese Society for Artificial Intelligence*, 14(5):771–780, 1999.

Terrence S. Furey, Nello Cristianini, Nigel Duffy, David W. Bednarski, Mich Schummer, and David Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10):906–914, 2000.

Oliver Gevaert, Frank De Smet, Dirk Timmerman, Yves Moreau, and Bart De Moor. Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. *Bioinformatics*, 22(14):e184–e190, 2006.

Mel Greaves. Childhood leukaemia. *BMJ*, 324(02/02/2002):283–287, 2002.

Stanley I. Grossman. *Calculus*. Academic Press Inc., New York, 1977.

Dachuan Guo, Belinda Cutri, and R. Daniel Catchpoole. The influence of RNA integrity, purity and cDNA labelling on glass slide cDNA microarray image quality. *Genome Biology*, 2004(5):13–38, 2004.

Isabelle Guyon and Andre Elisseeff. An Introduction to Variable and Feature Selection. *Machine Learning Research*, 3:1157–1182, 2003.

Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*, 46(1-3):389–422, 2002.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA Data Mining Software: An Update. *SiGKDD Explorations*, 11(1):10–18, 2009.

Jiawei Han and Micheline Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, San Francisco, second edition, 2006.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, New York, 2001.

Milos Hauskrecht, Richard Pelikan, Michal Valko, and James Lyons-Weiler. Feature Selection and Dimensionality Reduction in Genomics and Proteomics. In Werner Dubitzky, Martin Granzow, and Daniel P. Berrar, editors, *Fundamentals of Data Mining in Genomics and Proteomics*, Science+Business Media, pages 149–172. Springer, New York, 2007.

David Haussler. Probably Approximately Correct Learning. In *AAAI*, pages 1101–1108, 1990a.

David Haussler. Applying Valiant's learning framework to AI concept-learning problems. In Yves Kodratoff and Ryszard S. Michalski, editors, *Machine Learning: An Artificial Intelligence Approach, Volume III*, volume 3, pages 641–669. Morgan Kaufmann Publishers, Inc., San Mateo, California, 1990b.

David Haussler. Overview of the Probably Approximately Correct (PAC) Learning Framework. 1995.

Simon Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice-Hall, Inc., New Jersey, 2nd edition, 1999.

Priti Hegde, Rong Qi, Kristie Abernathy, Cheryl Gay, Sonia Dharap, Renne Gaspard, Julie Earle-Hughes, Erik Snesrud, Norman Lee, and John Quackenbush. A Concise Guide to cDNA Microarray Analysis. *Biotechniques*, 29(3):548–562, 2000.

Jianping Hua, Zixiang Xiong, James Lowey, Edward Suh, and Edward R. Dougherty. Optimal number of features as a function of sample size for various classification rules. *Bioinformatics*, 21(8):1509–1515, 2005.

Jianping Hua, Waibhav D. Tembe, and Edward R. Dougherty. Performance of feature-selection methods in the classification of high-dimension data. *Pattern Recognition*, (42):409–424, 2009.

Inaki Inza, Pedro Larranaga, Rosa Blanco, and Antonio J. Cerrolaza. Filter versus wrapper gene selection approaches in DNA microarray domains. *Artificial Intelligence in Medicine*, 31(2):91–103, 2004.

Anil Jain and Douglas Zongker. Feature Selection: Evaluation, Application, and Small Sample Performance. *IEEE Transactions on Pattern Analysis and Machine Learning*, 19(2):153–158, 1997.

Ian B. Jeffery, Desmond G. Higgins, and Aedin C. Culhane. Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. *BMC Bioinformatics*, 7(359):16, 2006.

Hongying Jiang, Youping Deng, Huann-Sheng Chen, Lin Tao, Quinying Sha, Jun Chen, Jung-Jui Tsai, and Shuanglin Zhang. Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes. *BMC Bioinformatics*, 5(81), 2004.

T. Jirapech-Umpai and S. Aitken. Feature selection and classification for microarray data analysis: evolutionary methods for identifying predictive genes. *BMC Bioinformatics*, 6(148), 2005.

George H. John and Pat Langley. Estimating Continuous Distributions in Bayesian Classifiers. In *Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345. Morgan Kaufmann, San Mateo., 1995.

Alexandros Kalousis, Julien Prados, and Melanie Hilario. Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowledge and Information Systems*, 12(1):95–116, 2007.

M. Kearns. Efficient Noise-Tolerant Learning From Statistical Queries. *Journal of the ACM*, 45(6):983–1006, 1998.

M. Kearns and M. Li. Learning in the presence of malicious errors. Technical Report TR-03-87, Center for Research in Computing Technology, Harvard University, 1987.

Michael J. Kearns and Robert E. Schapire. Efficient distribution-free learning of probabilistic concepts. In *31st Annual Symposium on Foundations of Computer Science*, pages 382–391, 1990.

Paul J. Kennedy, Simoff Simeon J., Daniel R. Catchpoole, David B. Skillicorn, Franco Ubaudi, and Ahmad Al-Oqaily. Integrative Visual Data Mining of Biomedical Data: Investigating Cases in Chronic Fatigue Syndrome and Acute Lymphoblastic Leukaemia. In Simoff Simeon J.,

Michael H. Bohlen, and Mazeika Arturas, editors, *Visual Data Mining: Theory, Techniques and Tools for Visual Analytics*, volume 4404 of *Lecture Notes in Computer Science*, pages 367–388. LNCS, Springer, Berlin Heidelberg, 2008.

Lev Klebanov, Xing Qiu, Stephen Welle, and Andrei Yakovlev. Statistical methods and microarray data. *Nature Biotechnology*, 25:25–26, 2007.

R. Kohavi and G. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97:273–324, 1997.

Ron Kohavi. Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid. In *Second International Conference on Knowledge Discovery and Data Mining*, pages 202–207. AAAI Press, 1996.

S. B. Kotsiantis, D. Kanellopoulos, and P. E. Pintelas. Data Preprocessing for Supervised Learning. *International Journal of Computer Science*, 1(2): 111–117, 2006.

L. B. Krupp, W. B. Mendelson, and R. Friedman. An overview of chronic fatigue syndrome. *J Clinical Psychiatry*, 52(10):403–410, 1991.

Pat Langley. Selection of Relevant Features in Machine Learning. In *AAAI Fall Symposium on Relevance*, pages 140–144, New Orleans, LA, 1994. AAAI Press.

Fan Li and Yiming Yang. Analysis of recursive gene selection approaches from microarray data. *Bioinformatics*, 21(19):3741–3747, 2005.

Leping Li, Clarice R. Weinberg, Thomas A. Darden, and Lee G. Pedersen. Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. *Bioinformatics*, 17(12):1131–1142, 2001.

H. Liu and L. Yu. Feature Selection for Data Mining. Technical report, Department of Computer Science and Engineering; Arizona State University, 2002.

Huan Liu and Hiroshi Motoda. *Feature Selection for Knowledge Discovery and Data Mining*, volume 454 of *The Springer Internal Series in Engineering and Computer Science*. Kluwer Academic Publishers, Boston, 1998.

Huan Liu and Hiroshi Motoda, editors. *Computational Methods of Feature Selection*. Data Mining and Knowledge Discovery Series. Chapman & Hall/CRC, Boca Raton, 2008.

Hongjun Lu, Yuan Sam, and Ying Lu Sung. On Preprocessing Data for Effective Classification. In *ACM SIGMOD: Workshop on Research Issues on Data Mining and Knowledge Discovery*, pages 1–7, 1996.

Wolf-Dieter Ludwig, Torsten Haferlach, and Claudia Schoch. *Classification of Acute Leukemias; Perspective 1.* Treatment of Acute Leukemias; New Directions for Clinical Research. Humana Press, Totowa, New Jersey, 2003.

S. Ma and J. Huang. Regularized ROC method for disease classification and biomarker selection with microarray data. *Bioinformatics*, 21:4356–4362, 2005.

Henryk Maciejewski. Quality of Feature Selection Based on Microarray Gene Expression Data. In M. Bukak, Geert van Albada, Jack Dongarra, and Peter Sloot, editors, *Computation Science - ICCS 2008, Part III, LNCS 5103*, LNCS, pages 140–147. Springer, Berlin Heidelberg, 2008.

David Maindonald. Data Mining Methodological Weaknesses and Suggested Fixes. In *Conferences in Research and Practice in Information Technology (CRPIT)*, volume 61, pages 9–16, 2006.

H. Mamitsuka. Selecting features in microarray classification using ROC curves. *Pattern Recognition*, 39:2393–2404, 2006.

Tanya McFerran, editor. *Oxford Dictionary of Nursing.* Market House Books, second edition, 1996.

R. S. Micalski. A Theory and Methodology of Inductive Learning. In R. S. Micalski, J. G. Carbonell, and T. M. Mitchell, editors, *Machine Learning: An Artificial Intelligence Approach.* Tioga, Palo Alto, Californina, 1983.

Tom M. Mitchell. *Machine Learning.* Computer Science. McGraw Hill, New York, 1997.

Vamsi K. Mootha, Cecilia M. Lindgren, Karl-Fredrik Eriksson, Aravind Subramanian, Smita Sihag, Joseph Lehar, Pere Puigserver, Emma Carlsson, Martin Ridderstrale, Esa Laurila, Nicholas Houstis, Mark J. Daly, Nick Patterson, Jill P. Mesirov, Todd R. Golub, Pablo Tamayo, Bruce Spiegelman, Eric S. Lander, Joel N. Hirschhorn, David Altshuler, and Leif C. Groop. PGC-1[alpha]-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*, 34(3):267–273, 2003.

David M. Mutch, Alvin Berger, Robert Mansourian, Andreas Rytz, and Matthew-Alan Roberts. The limit fold change model: A practical approach for selecting differentially expressed genes from microarray data. *BMC Bioinformatics*, 3(17):1–11, 2002.

M. A. Newton, Kendziorski C. M., C. S. Richmond, F. R. Blattner, and K. W. Tsui. On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology*, 8(1):37–52, 2004.

Emanuele Olivetti, Sriharsha Veeramachaneni, and Paolo Avesani. Active Learning of Feature Relevance. In Vipin Kumar, editor, *Computational Methods of Feature Selection*, pages 89–107. Chapman & Hall/CRC, Boca Raton, 2008.

C. Ooi and P. Tan. Genetic algorithms applied to multi-class prediction for the analysis of gene expression data. *Bioinformatics*, 19:37–44, 2003.

W. Pan. On the use of permutation in and the performance of a class of nonparametric methods to detect differential gene expression. *Bioinformatics*, 19:1333–1340, 2003.

P. J. Park, M. Pagano, and M. Bonetti. A nonparametric scoring algorithm for identifying informative genes from microarray data. In *Pacific Symposium on Biocomputing*, pages 52–63, 2001.

Taesung Park, Sung-Gon Yi, SeungYeoun Lee, and Jae K. Lee. Diagnostic plots for detecting outlying slides in a cDNA microarray experiment. *BioTechniques*, 38(3):463–471, 2005.

Giovanni Parmigiani, Elizabeth S. Garrett, Rafael A. Irizarry, and Scott L. Zeger. The Analysis of Gene Expression Data: An Overview of Methods and Software. In Giovanni Parmigiani, Elizabeth S. Garrett, Rafael A. Irizarry, and Scott L. Zeger, editors, *The Analysis of Gene Expression Data: Methods and Software*, pages 1–45. Springer-Verlag, New York, 2003.

Ching-Hon Pui, editor. *Treatment of Acute Leukemias, New Directions for Clinical Research*. Current Clinical Oncology. Humana Press, Totowa, 2003.

John Quackenbush. Computational analysis of microarray data. *Nature Reviews, Genetics*, 2(June):418–427, 2001.

J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986a.

J. Ross Quinlan, editor. *The Effects Of Noise On Concept Learning*, volume 2 of *Machine Learning: An Artificial Intelligence Approach*. Morgan Kaufmann Publishers, Inc., California, 1986b.

Erhard Rahm and Hong Hai Do. Data Cleaning: Problems and Current Approaches. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 23:1–11, 2000.

Jonathan M. Raser and Erin K. O'Shea. Noise in Gene Expression: Origins, Consequences, and Control. *Science*, 309(5743):2010–2013, 2005.

W. C. Reeves, D. Wagner, R. Nisenbaum, J. F. Jones, B. Gurbaxani, L. Solomon, D. A. Papanicolaou, E. R. Unger, S. D. Vernon, and C. Heim. Chronic Fatigue Syndrome - A clinically empirical approach to its definition and study. *BMC Medicine*, 3(19), 2005.

Stuart Russell and Peter Norvig. *Artificial Intelligence, A Modern Approach*. Pretence Hall, New Jersey, 1995.

Yvan Saeys, Inaki Inza, and Pedro Larranaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, 2007.

S. Saraswathi and T. Geetha. Time scale modification and vocal tract length normalization for improving the performance of Tamil speech recognition system implemented using language independent segmentation algorithm. *International Journal of Speech Technology*, 9(3):151–163, 2006.

W. Sarrett and M. Pazzani. Average case analysis of empirical and explanation-based learning algorithms. *Machine Learning*, 9(4):349–372, 1992.

Robert E. Schapire. Strength of Weak Learnability. *Machine Learning*, 5: 197–227, 1990.

M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science*, 270(5235):467–470, 1995.

Bernhard Schlkopf and Alexander J. Smola. *Learning with kernels : support vector machines, regularization, optimization, and beyond*. Adaptive computation and machine learning. MIT Press, Cambridge, Mass., 2002.

C. E. Shannon. A Mathematical Theory of Communication. *Bell Syst. Tech. J.,*, 27:379–423, 623–656, 1948.

C. E. Shannon. Communication in the Presence of Noise. In *Institute of Radio Engineers*, volume 37, pages 10–21. IEEE Journals, 1949.

S. C. Shapiro. *Encyclopedia of Artificial Intelligence*. Wiley, New York, 2nd edition, 1992.

S.K. Shevade, S.S. Keerthi, C. Bhattacharyya, and K.R.K. Murthy. Improvements to SMO Algorithm for SVM Regression. Technical Report CD-99-16, Control Division Dept of Mechanical and Production Engineering National University of Singapore, 1999.

Paulo J. S. Silva, Ronaldo F. Hashimoto, Seungchan Kim, Junior Barrera, Leonidas O. Brandao, Edward Suh, and Edward R. Dougherty. Feature selection algorithms to find strong genes. *Pattern Recognition Letters*, 26: 1444–1453, 2005.

Chao Sima and Edward R. Dougherty. What should be expected from feature selection in small-sample settings. *Bioinformatics*, 22(19):2430–2436, 2006.

Robert H. Sloan. Types of noise in data for concept learning. In *Proceedings of the first annual workshop on Computational learning theory*, pages 91–96, MIT, Cambridge, Massachusetts, United States, 1988. Morgan Kaufmann.

Alex Smola and Bernhard Scholkopf. A Tutorial on Support Vector Regression. Technical Report NC2-TR-1998-030, NeuroCOLT2 Technical Report Series, 1998.

E.M. Southern. DNA Microarrays. History and overview. *Methods in Molecular Biology*, 170:1–15, 2001.

Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS*, 102(43):15545–15550, 2005.

Yijun Sun, Sinisa Todorovic, and Steve Goodison. A Feature Selection Algorithm Capable of Handling Extremely Large Data Dimensionality. In *8th SIAM International Conference on Data Mining*, pages 530–540, Gainesville Florida, 2008. University of Florida.

Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining*. Pearson Education Inc, Boston, 2006.

Jeffrey G. Thomas, Olson James M., Stephen J. Tapscott, and Lue Ping Zhao. An Efficient and Robust Statistical Modeling Approach to Discover Differentially Expressed Genes Using Genomic Expression Profiles. *Genome Research*, 11:1227–1236, 2001.

Robert Tibshirani. Statistical challenges in the analysis of DNA microarray data. In *Conference of the CRM-SSC Prize in statistics*, page 35, Paris, 2001.

Chen-An Tsai, Sue-Jane Wang, Dung-Tsa Chen, and James J. Chen. Sample size for gene expression microarray experiments. *Bioinformatics*, 21(8): 1502–1508, 2005.

G. C. Tseng, M. K. Oh, L. Rohlin, W. Liao, and W. Wong. Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Research*, 29: 2549–2557, 2001.

Y. Tu, G. Stolovitzky, and U. Klein. Quantitative noise analysis for gene expression microarray experiments. *PNAS*, 99(22):14031–14036, 2002.

Virginia Goss Tusher, Robert Tibshirani, and Gilbert Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Science (PNAS)*, 98(9):5116–5121, 2001.

Franco A. Ubaudi, Paul J. Kennedy, Daniel R. Catchpoole, Dachuan Guo, and Simoff Simeon J. Microarray Data Mining: Selecting Trustworthy Genes with Gene Feature Ranking. In L. Cao, P. S. Yu, C. Zhang, and H. Zhang, editors, *Data Mining for Business Applications*, pages 159–168. Springer, 2009.

L. G. Valiant. A theory of the learnable. *Communications of the Association for Computing Machinery*, 27(11):1134–1142, 1984.

L. G. Valiant. Learning disjunctions of conjunctions. In *Ninth International Joint Conference on Artificial Intelligence*, pages 560–566, Los Angeles, CA, 1985. Morgan Kaufmann.

Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Statistics for Engineering and Information Science. Springer, New York, second edition, 2000.

Ronald E. Walpole and Raymond H. Myers. *Probability and Statistics for Engineers and Scientists*. Macmillan Publishing Co, New York, 2nd edition, 1978.

Xujing Wang, Soumitra Ghosh, and Sun-Wei Guo. Quantitative quality control in microarray image processing and data acquisition. *Nucleic Acids Research*, 29(15 e75):8, 2001.

Yu Wang, Igor. V. Tetko, Mark A. Hall, Eibe Frank, Axel Facius, Klaus F. X. Mayer, and Hans W. Mewes. Gene selection from microarray data for cancer classification - a machine learning approach. *Computational Biology and Chemistry*, 29(1):37–46, 2005.

Stanley J. Watson, Fan Meng, Robert C. Thompson, and Huda Akil. The "Chip" as a Specific Genetic Tool. *Society of Biological Psychiatry*, 48(12): 1147–1156, 2000.

N. Wiener. *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. 1949.

S. E. Wildsmith and F. J. Elcock. Micoarrays under the microscope. *Clinical Pathology*, (54):8–16, 2001.

Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Data Management Systems. Morgan Kaufmann, second edition, 2005.

Eric P. Xing, Michael I. Jordan, and Richard M. Karp. Feature selection for high-dimensional genomic microarray data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 601–608. Morgan Kaufmann, 2001.

M. Xiong, Z. Fang, and J. Zhao. Biomarker identification by feature wrappers. *Genome Research*, 11:1878–1887, 2001.

H. Yang and T. P. Speed. Design issues for cDNA microarray experiments. *Nature Genetics Reviews*, 3:579–588, 2002.

Y. H. Yang, M. J. Buckley, and T. P. Speed. Analysis of cDNA microarray images. *Briefings in Bioinformatics*, 2(4):341–349, 2001a.

Yee. Hwa. Yang, Sandrine. Dudoit, Percy. Luu, and Terrance. P. Speed. Normalization for cDNA Microarray Data. *SPIE BiOS*, 2001b.

Yee Hwa Yang, Yuanyuan Xiao, and Mark R. Segal. Identifying differentially expressed genes from microarray experiements via statistic synthesis. *Bioinformatics*, 21(7):1084–1093, 2004.

Yiming Yang and Jan O. Pedersen. A Comparative Study on Feature Selection in Text Categorization. pages 412–420. Morgan Kaufmann, 1997.

Eng-Juh Yeoh, Mary E. Ross, Sheila A. Shurtleff, Kent Williams, Divyen Patel, Rami Mahfouz, Fred G. Behm, Susana C. Raimondi, Mary V. Relling, Anami Patel, Cheng Cheng, Dario Campana, Dawn Wilkins, Xiaodong Zhou, Jinyan Li, Huiqing Liu, Ching-Hon Pui, William E. Evans, Clayton Naeve, Limsoon Wong, and James R. Downing. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, 1:133–143, 2002.

K. Yeung and R. Bumgarner. Multiclass classification of microarray data with repeated measurements: application to cancer. *Genome Biology*, 4 (R83), 2003.

Ka Yee Yeung, Roger E. Bumgarner, and Adrian E. Raftery. Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data. *Bioinformatics*, 21(10):2394–2402, 2005.

Lei Yu. Feature Selection for Genomic Data Analysis. In Huan Liu and Hiroshi Motoda, editors, *Computational Methods of Feature Selection*, Data Mining and Knowledge Discovery Series, pages 337–353. Chapman & Hall/CRC, Boca Raton, 2008.

Lei Yu and Huan Liu. Redunancy Based Feature Selection for Microarray Data. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 737–742, Seattle WA USA, 2004. ACM.

Shichao Zhang, Chengqi Zhang, and Qiang Yang. Data preparation for data mining. *Applied Artificial Intelligence*, 17:375–381, 2003.