Robust Classification of High Dimensional Unbalanced Single and Multi-label Datasets

A Thesis Submitted for the Degree of
Doctor of Philosophy

By
Ali Braytee

in

School of Software
UNIVERSITY OF TECHNOLOGY SYDNEY
AUSTRALIA
FEBRUARY 2018

© Copyright by Ali Braytee, 2018

UNIVERSITY OF TECHNOLOGY SYDNEY
SCHOOL OF SOFTWARE

The undersigned hereby certify that they have read this thesis entitled "**Robust Classification of High Dimensional Unbalanced Single and Multi-label Datasets**" by **Ali Braytee** and that in their opinions it is fully adequate, in scope and in quality, as a thesis for the degree of **Doctor of Philosophy**.

Dated: February 2018

Principal Supervisor: _____
A/P Paul Kennedy

CERTIFICATE

Date: **February 2018**

Author: **Ali Braytee**

Title: **Robust Classification of High Dimensional Unbalanced Single and Multi-label Datasets**

Degree: **Ph.D.**

I, Ali Braytee declare that this thesis, submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the School of Software/ Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise reference or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by an Australian Government Research Training Program Scholarship.

Signature of Author

Acknowledgements

My first acknowledgement is to God for being my source of spiritual strength. I am greatly indebted to my supervisor, A/P Paul Kennedy for his continuous encouragement, advice, help and invaluable suggestions. I owe my research achievements to his experienced supervision. Many thanks are also due to my co-supervisor, Dr. Wei Liu for his valued suggestions and constant support, and for the numerous conversations with him. I gratefully acknowledge the useful discussions with A/P Daniel Catchpoole and A/P Farookh Hussain. I appreciate the travel support for attending the international conferences which I received from the School of Software and the Centre for Artificial Intelligence.

I would like to thank my parents that I could never have done this without their faith, support, and constant prays. I also thank my wife Zaynab, my partner in life, who managed to be nothing but supportive, for your patience, and your faith, because you always understood.

Last but not least, the financial assistance of an Australian Postgraduate Award and UTS Research Excellence Scholarship is gratefully appreciated.

To my father and mother who instilled in me the virtues of perseverance and commitment. To my wife Zaynab, whose sacrificial care for me and our children enabled the hours of research, contemplation, and writing necessary to complete this thesis. To my children Hassan and Adam you are my greatest achievement, and you will always be my blessings from God. All together we have managed to survive throughout this challenging journey. I love you all.

Table of Contents

List of Tables

List of Figures

Abstract

Single and multi-label classification are arguably two of the most important topics within the field of machine learning. Single-label classification refers to the case where each sample is assigned to one class, and multi-label classification is where instances are associated with multiple labels simultaneously. Nowadays, research to build robust single and multi-label classification models is still ongoing in the data analytics community because of the emerging complexities in the real-world data, and due to the increasingly research interest in use of data analytics techniques in many fields including biomedicine, finance, text mining, text categorization, and images. Real-world datasets contain complexities which degrade the performance of classifiers. These complexities or open challenges are: imbalanced data, low numbers of samples, high-dimensionality, highly correlated features, label correlations, and missing labels in multi-label space. Several research gaps are identified and motivate this thesis. Class imbalance occurs when the distribution of classes is not uniform among samples. Feature extraction is used to reduce the dimensionality of data. However, the presence of highly imbalanced data in single-label classification misleads existing unsupervised and supervised feature extraction techniques. It produces features biased towards classification of the class with the majority of samples, and results

in poor classification performance especially for the minor class. Furthermore, imbalanced multi-labeled data is more ubiquitous than single-labeled data because of several issues including label correlation, incomplete multi-label matrices, and noisy and irrelevant features.

High-dimensional highly correlated data exist in several domains such as genomics. Many feature selection techniques consider correlated features as redundant and therefore need to be removed. Several studies investigate the interpretation of the correlated features in domains such as genomics, but investigating the classification capabilities of the correlated feature groups in single-labeled data is a point of interest in several domains. Moreover, high-dimensional multi-labeled data is more challenging than single-labeled data. Only relatively few feature selection methods have been proposed to select the discriminative features among multiple labels due to issues including interdependent labels, different instances sharing different label correlations, correlated features, and missing and noisy labels.

This thesis proposes a series of novel algorithms for machine learning to handle the negative effects of the above mentioned problems and improves the performance of the classifiers in single and multi-labeled data. There are seven contributions in this thesis. Contribution 1 proposes novel cost-sensitive principal component analysis (CSPCA) and cost-sensitive non-negative matrix factorization (CSNMF) methods for handling feature extraction of imbalanced single-labeled data. Contribution 2 extends a standard non-negative matrix factorization to a balanced supervised non-negative matrix factorization (BSNMF) to handle the class imbalance problem in supervised non-negative matrix factorization. Contribution 3 introduces an ABC-Sampling algorithm for balancing imbalanced datasets based on Artificial Bee Colony algorithm.

Contribution 4 develops a novel supervised feature selection algorithm (SCANMF) by jointly integrating correlation network and structural analysis of the balanced supervised non-negative matrix factorization to handle high-dimensional, highly correlated single-labeled data. Contribution 5 proposes an ensemble feature ranking method using co-expression networks to select optimal features for classification. Contribution 6 proposes a Correlated- and Multi-label Feature Selection method (CMFS), based on NMF for simultaneously performing multi-label feature selection and addressing the following challenges: interdependent labels, different instances sharing different label correlations, correlated features, and missing and flawed labels. Contribution 7 presents an integrated multi-label approach (ML-CIB) for simultaneously training the multi-label classification model and addressing the following challenges namely, class imbalance, label correlation, incomplete multi-label matrices, and noisy and irrelevant features.

The performance of all novel algorithms in this thesis is evaluated in terms of single and multi-label classification accuracy. The proposed algorithms are evaluated in the context of a childhood leukaemia dataset from The Children Hospital at Westmead, and public datasets for different fields including genomics, finance, text mining, images, and others from online repositories. Moreover, all the results of the proposed algorithms in this thesis are compared to state-of-the-art methods. The experimental results indicate that the proposed algorithms outperform the state-of-the-art methods. Further, several statistical tests including, t-test and Friedman test are applied to evaluate the results to demonstrate the statistical significance of the proposed methods in this thesis.

Publications

Below is the list of journal and conference papers associated with my PhD research:

1. **Braytee, A.**, Liu, W., Anaissi, A. & Kennedy, P. J. (2017), 'Correlated Multi-label Classification with Missing Labels and Class Imbalance', *Data Mining and Knowledge Discovery.* (Under review).

2. **Braytee, A.**, Liu, W., Catchpoole, D. R. & Kennedy, P. J. (2017), Multi-Label Feature Selection using Correlation Information, in 'Proceedings of the 26th ACM International on Conference on Information and Knowledge Management', ACM, (To appear). (ERA rank A).

3. **Braytee, A.**, Liu, W. & Kennedy, P. J. (2017), Supervised context-aware non-negative matrix factorization to handle high-dimensional high-correlated imbalanced biomedical data, in 'International Joint Conference on Neural Networks (IJCNN), 2017', IEEE, pp. 4512−4519. (ERA rank A).

4. **Braytee, A.**, Catchpoole, D. R., Kennedy, P. J. & Liu, W. (2016), Balanced supervised non-negative matrix factorization for childhood leukaemia patients, in 'Proceedings of the 25th ACM International on Conference on Information and Knowledge Management', ACM, pp. 2405−2408. (ERA rank A).

5. **Braytee, A.**, Liu, W. & Kennedy, P. (2016), A cost-sensitive learning strategy for feature extraction from imbalanced data, in 'International Conference on Neural Information Processing', Springer, pp. 78−86. (ERA rank A).

6. Anaissi, A., Goyal, M., Catchpoole, D. R., **Braytee, A.** & Kennedy, P. J. (2016), 'Ensemble feature learning of genomic data using support vector machine', *PloS one* **11**(6), e0157330.

7. **Braytee, A.**, Hussain, F. K., Anaissi, A. & Kennedy, P. J. (2015), ABC-sampling for balancing imbalanced datasets based on artificial bee colony algorithm, in 'IEEE 14th International Conference on Machine Learning and Applications (ICMLA), 2015,IEEE, pp. 594−599.

Table of Symbols

Symbols	**Description**
X	Data matrix
n, N	Samples or instances in the matrix
m, M	Features or dimensions in the matrix
p, k, K, R	approximate rank for factorization
U	Factorized matrix
V	Factorized matrix
r	Uniform random number between [0,1]
D	Problem dimensionality
FS	Food source (possible solution)
ABC	Artificial Bee Colony
NMF	Non-negative matrix factorization
PCA	Principle component analysis
GA	Genetic algorithm
PSO	Particle swarm optimization
ACO	Ant Colony Optimization
DE	Differential algorithm
SVM	Support vector machine
y	Class vector
α	Majority classes weight
C^+	Positive imbalance cost ratio
C^-	Negative imbalance cost ratio

W	Factorized matrix
H	Factorized matrix
X'	Weighted data matrix
ADASYN	Adaptive Synthetic Sampling Approach for Imbalanced Learning
CCPDT	Class Confidence Proportion Decision Tree
RU	Random undersampling
ML-CST	Maximum likelihood in cost-sensitive learning
Y, Z	Data matrices
A	Common factorized matrix between Y and Z
B, C	Factorized matrices
X^+	data matrix of positive samples
X^-	data matrix of negative samples
σ^+	weight for positive samples
σ^-	weight for negative samples
$LIMIT$	the number of times of using the possible solution
Abandoned	the existing solution that reach to maximum $LIMIT$
$FoodSourceSize$	the size of possible solutions
$Best_FS$	A solution with best fitness value
BIRF	A Balanced Iterative Random Forest
T	Arbitrary data matrix
D	Diagonal matrix
Tr	Trace operation of matrix
corr	Correlation function such as Pearson correlation
θ	Power of correlation coefficient
A	Adjacency matrix
n^-,N^-	is the number of negative samples
n^+,N^+	is the number of positive samples
$l_{2,1}$	norm regularization
l_1	norm regularization

β	control the contribution of $l_{2,1}$ norm
α	control the contribution of the network structure
AUC	Area under curve
WGCNA	Weighted correlation network analysis
HLR L1/2 L2	Combination of $l_{2,1}$ and l_2 norm regularization
SVD	Singular value decomposition
SVM-RFE	Feature ranking based on SVM
L, P	A matrix to absorb the different scales of matrices
Q	Feature combinations matrix
c	Number of clusters
Y	Multi-label matrix (Chapter 5)
B	Clusters of labels matrix (Chapter 5)
R	Graph Laplacian (Chapter 5)
S	Similarity matrix (Chapter 5)
G	Diagonal matrix (Chapter 5)
α	Control the contribution of label correlation (Chapter 5)
ϵ	Control the contribution of the network structure (Chapter 5)
γ	Control the contribution of sparseness of the model (Chapter 5)
$\hat{Y}$	New multi-label matrix
L	Similarity matrix (Chapter 5)
W	Label-specific features (Chapter 5)
V	Label regularization (Chapter 5)
α	Control the contribution of the new label matrix manifold (Chapter 5)
β	Control the contribution of the difference between the new and original label matrix (Chapter 5)
MLSMOTE	Multi-label SMOTE

Chapter 1

Introduction

Classification is defined as the problem of identifying the class or label of a new observation based on labeled training data. Several characteristics of data including high dimensionality, highly correlated features, label correlations, and imbalanced data with low number of samples cause difficulties for classification. The aim of this thesis is to develop a series of novel algorithms for machine learning to handle the mentioned problems and improve the performance of the classifiers. In machine learning, classification is divided into two types: single-label classification where each sample is assigned to one class or category, and multi-label classification where instances are associated with multiple target labels simultaneously. For example, in information retrieval, the document can cover multiple topics, and in bioinformatics, a gene can be associated with multiple functions. The proposed algorithms in this thesis handle the complexities and challenges of single and multi-labeled data from several domains including bioinformatics, text, video, and information retrieval.

Real-world datasets contain complexities as mentioned above which are considered as open challenges in data analytics. These challenges negatively affect the performance of classifiers and motivate the work in this thesis to build more robust

single-label and multi-label classification models which can work on many datasets from different domains.

The large number of dimensions or features is a common problem in several domains including bioinformatics (Saeys et al. 2007), text mining (Berry & Castellanos 2008), among others. These dimensions contain some potential features which assist in improving the performance of classification model. However, many other features are typically noisy or redundant.

In some fields such as genomics, high dimensionality is exacerbated in the presence of a low number of samples leading to an issue known as "the curse of dimensionality" (Bellman 2003). Consequently, building a classification model on high dimensional data often leads to over-fitting, where the classification model works well on training data, but fails to predict new observations. Two general approaches are proposed in the literature to reduce the number of dimensions: feature selection, which selects a subset of the most informative features without changing the original variables, and feature extraction, which builds a new set of features from the original data (Saeys et al. 2007). Correlation among the features is commonly observed in several domains (Horvath 2011, Ivliev et al. 2010). Many feature selection methods consider correlated features as redundant, needing to be removed (Guyon et al. 2002, Tibshirani 1996). However, some studies have explored the idea that correlated features are considered as potential predictors (Bondell & Reich 2008).

Finally, another problem is class imbalance, where the number of samples from one class significantly outnumbers the number of samples from the other class. This problem is observed in many fields including genomics, text mining, and others (Kotsiantis et al. 2006).

The standard classifiers in machine learning do not function well with respect to these data challenges. They impact both single and multi-label classification. However, multi-label classification is more challenging due to the correlation among the labels and the incomplete label space. Furthermore, proposed methods for high dimensional, class imbalanced and highly correlated data in single-labeled data are not necessarily appropriate for multi-labeled data.

This thesis proposes a series of novel algorithms to improve single and multi-label classification in the presence of datasets' complexities. The main objective from the proposed algorithms is to enhance the generalization of standard classifiers in order to accurately work on a large number of datasets from different domains.

1.1 Research Challenges

This section lists the main issues investigated in this thesis and presents the research questions.

- **Classification in single-labeled data in the presence of the class imbalance problem produces poor performance on the minor class samples**.

  Class imbalance in single-labeled data is caused by unequal distributions of the data between class labels (Chawla et al. 2004). It occurs due to a paucity of cases, for example, patients with a rare disease, or difficulties in collecting samples due to high cost or privacy. It occurs when the number of samples for a particular class, known as the majority or negative class, outnumbers the number of samples for other classes known as minority or positive classes (He & Garcia 2009). This problem greatly affects the performance of the standard

classifiers (He & Garcia 2009). Several algorithms addressing this problem are proposed in the literature to explore this problem at different levels including the data-level, the cost-sensitive level and the algorithm level. However, this problem is still considered as an open challenge in data analytics applications (Krempl et al. 2014). Recently a line of research investigated enhancing the classification capability of feature extraction methods by exploiting the existence of class labels (Huh et al. 2013).

Standard feature extraction methods such as principal component analysis (PCA) (Pearson 1901) and non-negative matrix factorization (NMF) (Lee & Seung 2001) are significantly affected by the class imbalance problem. For example, PCA seeks the orthogonal feature extractors that maximize the total variance. Therefore, the extracted features favor the majority class because there are more records than for the minority class. Moreover, NMF is a matrix factorization method for dimensionality reduction which decomposes the original matrix into two non-negative matrices by minimizing the squared error between them. In this approach, the factorized matrices are affected by the imbalance problem because the magnitude of the residual squared error of negative class instances is more than for the positive ones. Moreover, the performance of the supervised NMF method proposed by Huh et al. (2013) deteriorates due to the imbalance problem because it generates features that predict the majority class rather than the minority class, which is the class of interest. Therefore, these problems highlight the urgency of handling the class imbalance problem in feature extraction methods to select the non-biased potential features.

- **Multi-label classification performance is significantly affected by the label imbalance problem and incomplete label space**.

  The class imbalance issue is ubiquitous in multi-labeled data than single-labeled data. Typically, there are two different types of class imbalance in multi-label classification, which are referred to instance-class imbalance and label-class imbalance. First, the number of instances assigned to positive labels is less than to negative ones. This is a similar situation to the imbalance problem in binary classification. Second, is the label-class imbalance problem is where the number of labels assigned to different instances is significantly different between labels. Surprisingly, the class imbalance issue has not been extensively studied in multi-label classification context (Wu et al. 2016). Moreover, in addition to imbalanced data problem and particularly in real applications, it is difficult to find a complete label vector for each instance due to the unavailability of all labels in some applications. Furthermore, the label matrix may contain noisy labels during manual labelling. Therefore, it is important to handle the class imbalance problem and missing labels in label matrix to improve the performance of multi-label classification algorithms.

- **Feature selection methods consider correlated features as redundant features**.

  In the presence of high dimensional, highly correlated features, many feature selection techniques consider correlated features as redundant and need to be removed. Initially, researchers were unaware of the importance of correlated covariates in interpreting predictive models. Recent studies aim to interpret groups of highly correlated features to identify significant functional modules, to

improve classification accuracy, and to reflect the semantic components of these features (Horvath 2011, Bondell & Reich 2008). Therefore, it is an important problem to investigate the classification capabilities of the correlated feature groups due to their benefits in exploring semantic components in different fields including genomics and social networks.

- **Reducing high-dimensional multi-labeled data by selecting predictive features is a non trivial task due to the existence of label-label and feature-feature correlations**.

  Feature selection methods for single-labeled data are not appropriate for multi-labeled data due to several issues including the presence of multiple labels for each single sample, label correlations, and incomplete label space. First, the labels in a multi-label learning problem are correlated and interdependent, which is not the case in standard single-labeled data where the classes are mutually exclusive. For example, "political" and "election" documents or labels are more related than "sports" documents. Second, the interaction among features greatly contributes to finding a shared space for correlated labels. Third, the assumption of global label correlations among all instances is not correct in all cases. Mostly, the label correlations are shared only by a subset of instances. Surprisingly, relatively few feature selection methods in multi-label learning take label correlations into consideration. Therefore, there is a need to take into account the correlation information during the feature selection process to improve the performance of multi-label classification algorithms.

Based on the above research issues, the research questions of this thesis are specified as follows:

RQ1: Is it possible to propose appropriate methods to handle the class imbalance problem in single-labeled data?

RQ2: Is it possible to develop dimensionality reduction methods to reduce the high dimensional data that do not consider the correlated features as redundant?

RQ3: How to improve the multi-label classification performance in the presence of high dimensional multi-labeled data, label correlations, label imbalance and incomplete label space?

1.2 Contributions to Knowledge

This section presents seven contributions to knowledge based on the three research questions above as shown in Figure 1.1.

1. A cost-sensitive learning strategy for feature extraction to handle the class imbalance in single-labeled data.

2. A balanced supervised non-negative matrix factorization approach to generate features which are not biased to predict the majority class.

3. ABC-Sampling for balancing imbalanced datasets based on Artificial Bee Colony algorithm.

4. Supervised context-aware non-negative matrix factorization to handle high-dimensional, highly correlated single-labeled data.

5. Ensemble feature ranking method using co-expression networks to select optimal features for classification.

6. Multi-label feature selection using correlation information to select optimal features in multi-labeled data.

7. Improving multi-label classification in the presence of incomplete label space and the class imbalance problem.

These contributions to knowledge are described in detail below.

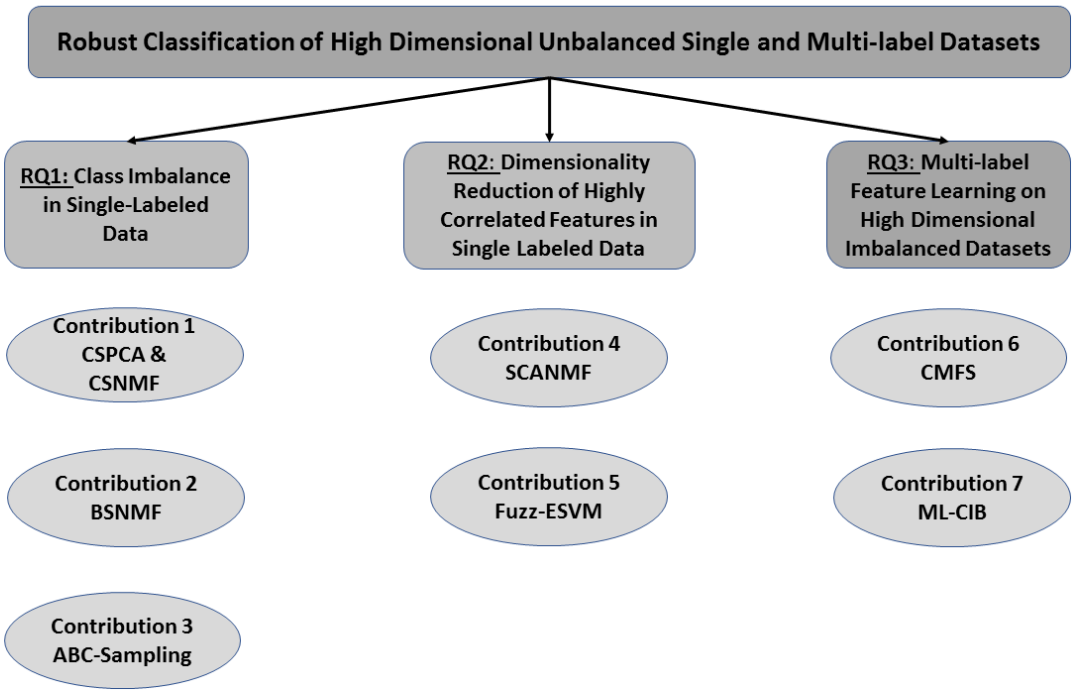


Figure 1.1: Thesis RQs and contributions structure

Contribution 1: A cost-sensitive learning strategy for feature extraction to handle the class imbalance in single-labeled data.

This thesis proposes a cost-sensitive learning strategy to address the imbalanced class problem that can be applied to existing feature extraction techniques such

as PCA and NMF. Integrating the cost-sensitive strategy to PCA and NMF results in two proposed methods: cost-sensitive PCA (CSPCA) and cost-sensitive NMF (CSNMF). The proposed methods embed the labelling information in the feature extraction methods to extract non-biased features towards majority samples (balanced features) which improve the accuracy of classification. Balanced features are the potential selected features from the original data which are not biased to the majority class and take into account the minority samples, which are the class of interest. Evaluation on multiple popular classifiers and benchmark datasets shows the high performance of the proposed methods in this thesis against the state-of-the-art. They can deal with different levels of imbalance and size of dataset. Finally, they can alleviate the imbalanced class problem without modifying the existing base classifiers, or changing the original information in the training datasets. The first contribution corresponds to RQ1 and is described in Section 3.1.

Contribution 2: A balanced supervised non-negative matrix factorization approach to generate features which are not biased to predict the majority class.

Supervised feature extraction is a growing research area in data analytics. Several supervised feature extraction methods have been suggested to enhance the classification of the extracted features. Non-negative matrix factorization attracted research attention due to the intuitive interpretation of its data decomposition in many fields. This thesis introduces a balanced supervised non-negative matrix factorization (BSNMF) based on the theory of coupled matrix factorization to extract a common balanced basis matrix between the positive

and negative samples. The proposed method exploits the class labels to lessen the inherent bias of the extracted components to the majority class. The proposed BSNMF algorithm is used to achieve feature selection in high-dimensional data. Experiments show that the proposed method outperforms the state-of-the-art methods. The second contribution corresponds to RQ1 and is presented in Section 3.2.

Contribution 3: ABC-Sampling for balancing imbalanced datasets based on Artificial Bee Colony algorithm.

This study proposes an undersampling approach and is based on artificial bee colony (ABC), a robust optimization method that is part of the family of swarm intelligence algorithms and simulates the foraging behaviour of honeybees. The proposed ABC-Sampling algorithm classifies imbalanced data by identifying the most informative majority examples. The output of the proposed algorithm is a balanced dataset. The proposed method is evaluated against a state-of-the-art methods by training a Support Vector Machine (SVM) classifier on the retrieved balanced dataset and evaluating on a test set. Evaluation at different levels of imbalance shows the superiority of the proposed method compared to state-of-the-art. The third contribution corresponds to RQ1 and is given in Section 3.3.

Contribution 4: Supervised context-aware non-negative matrix factorization to handle high-dimensional, highly correlated single-labeled data.

This research proposes a method to deal with correlated features during the feature selection process. First, it proposes supervised context-aware non-negative

matrix factorization which simultaneously incorporates the correlation structure of the features to reflect the semantic components along with the prior knowledge of the class labels and the original features to the data decomposition technique NMF. Then, for feature selection, $l_{2,1}$ -norm regularization is imposed in the basis matrix due to its robustness to outliers and to its sparsity criteria over rows of the matrix. The selected features are discriminative, non-biased, not misleading, and can be used to infer some semantic components due to using NMF in data decomposition. Evaluation on several high-dimensional high-correlated imbalanced real world datasets demonstrates promising results for the proposed algorithm compared to state-of-the-art. The fourth contribution corresponds to RQ2 and is described in Section 4.1.

Contribution 5 : **Ensemble feature ranking method using co-expression networks to select optimal features for classification**.

This thesis proposes an embedded feature ranking method to achieve feature selection in the presence of the correlated features. It follows a backwards feature elimination method and is divided into two phases. First it ranks features within the correlated groups, then, it selects a certain number of these features from different groups, aggregate, and ranks them again to select the top features between the groups. Therefore, the proposed method enhances the correlation inside the groups and lessens the correlation within groups. Also, the algorithm does not treat the correlated features as redundant features which need to be removed. Evaluation shows the superiority of the proposed method compared to other state-of-the-art feature selection methods based on classification accuracy on high-dimensional, highly correlated datasets. Moreover, the proposed

method demonstrates the capability for interpretation in a biological dataset showing the importance of integrating the correlated features in the proposed method. The fifth contribution corresponds to RQ2 and is presented in Section 4.2.

Contribution 6 : **Multi-label feature selection using correlation information to select the optimal features in multi-labeled data**.

This thesis proposes a Correlated- and Multi-label Feature Selection method in the context of NMF which is able to address problems in multi-label classification including high-dimensional multi-labeled data and label correlations. In particular, the original data and label matrices are decomposed into a lower-dimensional space, based on defining three kinds of interactions: label-label, label-feature and sample-label interaction. Typically, the low rank feature matrix finds relevant features able to capture the dependency among multiple labels. The low rank structure can generate the local label correlations based on the similarity between instances. Furthermore, the proposed method defines a new complete label matrix which addresses missing and noisy labels. This algorithm is a first attempt at exploiting correlation information of labels, features and samples together to find relevant features shared across multiple labels. Evaluation on high-dimensional real-world multi-labeled datasets demonstrates that the proposed method outperforms the state-of-the-art methods. The sixth contribution corresponds to RQ3 and is given in section 5.1

Contribution 7 : **Improving multi-label classification in the presence of incomplete label space and the class imbalance problem**.

This thesis proposes an integrated approach addressing three challenges in multi-label classification, namely the class imbalance problem, incomplete label matrix and generating label correlations from boolean matrix. Relevant features are retained in the feature matrix by imposing l_1-norm regularization. A new continuous label matrix is constructed by considering three factors: (a) the new label matrix should be consistent with the original label matrix, (b) handling the imbalance class problem in the original label matrix, and (c) the new label matrix uses the semantic instance assumption which means if the two instances x_i and x_j are similar, then their labels are correlated. A label regularization is proposed to handle the imbalanced label issue in the new predicted label matrix. An integrated approach that combines these components is formulated as a constrained optimization problem and solved using the accelerated proximal gradient method. Evaluation on high-dimensional multi-labeled dataset show that the proposed algorithm outperforms the state-of-the-art based on several evaluation metrics. The seventh contribution corresponds to RQ3 and is presented in Section 5.2.

1.3 Research Significance

This thesis proposes novel algorithms to deal with several data challenges in single and multi-label classification problems. These algorithms have the potential to significantly enhance the ability of existing classifiers to work with complex data in

different domains leading to benefits in the community.

The benefits of the proposed algorithms are important in domains including genomics, text, and image processing. In genomics, several experiments have been conducted on genomics data including a childhood leukaemia dataset from The Children's Hospital at Westmead, Australia. There is an urgent demand to predict the treatment outcome of children. This can be posed as a classification problem with a single label, e.g. relapse or not, or multi-label such as the patient relapse and mortality. This thesis proposes several algorithms to build an accurate single-label and multi-label classification model which may be considered as crucial step in assisting clinicians in determining a successful treatment.

In the domain of text analytics, the proposed algorithms in this thesis show their efficacy to categorize text documents such as news stories generated by Reuters into multiple categories or groups, a problem known as multi-label text categorization. Due to the enormous growth of online text documents, it is impossible to manually categorize these documents. The proposed algorithms improve the performance of automatic multi-label text categorization for the text documents. These benefits are reflected in the experiments of the proposed methods in Chapters 3, 4, and 5.

1.4 Thesis Structure

This doctoral thesis is presented as follows:

Chapter 1 states the research problems and motivations, research questions, and research contributions and publications.

Chapter 2 presents background concepts and previous works related to the research

topics of the class imbalance problem in single and multi-labeled data, high-dimensionality in single and multi-labeled data, feature selection, feature ranking, highly correlated features, highly correlated labels, and missing labels space. Furthermore, this chapter presents brief theory underlying feature extraction methods including PCA and NMF. Also, it demonstrates an understanding of classifiers used in this thesis such as SVM, K-nearest neighbour (KNN), and Naive Bayes.

Chapter 3 presents three algorithms to handle the class imbalance problem in single-labeled data, namely handling class imbalance with the classic feature extraction methods PCA and NMF, addressing the class imbalance problem with supervised NMF, and balancing the imbalanced data using artificial bee colony algorithm. Each algorithm describes and motivates the research problem, presents the proposed methods, conducts and discusses the experiments and concludes with contribution to knowledge.

Chapter 4 proposes two methods to handle high-dimensional, highly correlated features: supervised context-aware non-negative matrix factorization and an embedded feature ranking method. Supervised context-aware non-negative matrix factorization simultaneously incorporates the correlation structure of the features with the prior knowledge of the class labels and the original features into the NMF data decomposition technique. An embedded feature ranking method is proposed to achieve feature selection in the presence of correlated features. The series of research problem, methods, experiments, and contribution to knowledge are described in both methods.

Chapter 5 presents two algorithms: a Correlated- and Multi-label Feature Selection method and a multi-label classification method for multi-labeled imbalanced data. The former is proposed in the context of NMF which is able to select important features by simultaneously addressing three kinds of interaction: label-label, label-feature and sample-label interaction. The later addresses three challenges in multi-labeled data, namely the multi-label imbalance problem, incomplete label matrix and generating label correlations from boolean matrix. The proposed methods are evaluated and compared to the state-of-the-art and conclude with contribution to knowledge.

Chapter 6 concludes the thesis, putting the work into a broader context. It discusses the strengths and weaknesses of the proposed contributions, and describes future directions of research.

Chapter 2

Literature Review and Background

This chapter reviews the literature related to this thesis. Section 2.1 presents a brief revision of principal component analysis (PCA) and non-negative matrix factorization (NMF) and reviews the previous work on handling the class imbalance problem in single-label classification. Furthermore, it describes the methods that address imbalanced data in supervised feature extraction methods. Section 2.2 reviews the dimensionality reduction algorithms that handle the correlated features. Section 2.3 presents the previous methods that handle high-dimensional imbalanced multi-labeled data. Research gaps are highlighted in the respective sections. Finally, Section 2.4 describes the standard classifiers used for evaluation in this thesis together with their advantages and disadvantages.

2.1 Imbalanced Data in Single-Label Classification

Class imbalance is a common problem for predictive modelling in many domains including bioinformatics, text mining, and finance (Kotsiantis et al. 2006). It occurs when the distribution of classes is not uniform among samples and results in a biased

prediction of learning towards majority classes. This problem has been explored in the literature, however, it is still considered as an open challenge in data analytics in many domains including genomics, fraud detection, and credit scoring (Krempl et al. 2014). Recent research is explored to extend the standard feature extraction methods to supervised feature extraction methods. However, most studies do not pay attention to the class imbalance problem in these methods. Furthermore, as Hoens et al. (2012) describe, there is still a need for a more rigorous evaluation of the proposed methods on real-world data, and to explore the impact of the imbalanced data problem which is exacerbated in the presence of other problems, such as high-dimensional data or a low number of samples.

This section reviews the work on three main approaches for handling imbalanced data in single-label classification: a class imbalance in standard feature extraction methods, imbalanced data in supervised NMF and sampling data using the artificial bee colony algorithm. Furthermore, feature extraction methods, NMF and PCA are presented.

2.1.1 Feature Extraction

This section presents a brief description of the feature extraction methods used in this thesis, namely NMF and PCA. Feature extraction or dimensionality reduction is an approach to construct new features from the original features. It has been used in many domains (Saeys et al. 2007) to reduce the number of dimensions in high-dimensional datasets and to visualize the data in two or three dimensions.

A Review of PCA

Principal component analysis (Pearson 1901) is one of the most popular feature extraction techniques. It uses an orthogonal transformation to construct a low-dimensional representation of the data known as principal components. The linear transformation aims to maximise the global variance of the data as well as to minimise the least squares error of the transformation. The first principal component represents the direction of the largest variance of the data; the remaining principal components have smaller variance and are orthogonal to the preceding ones. Consider a data matrix $X \in \mathbb{R}^{m \times n}$ where m columns are the dimensions or features and each of the n rows represents the instances or observations. The first loading w_1 is computed by

$$w_1 = \max \sum_{ij} (X_{ij} \cdot w_{1j})^2 \tag{2.1.1}$$

where i and j are the indices of the rows and columns of X respectively, and w_1 is the first principal component of p dimensions and $p << m$.

A Review of NMF

Non-negative matrix factorization is a matrix factorization technique where the original and factorised matrices have a property of non-negative elements (Lee & Seung 1999). Given a data matrix $X \in \mathbb{R}^{m \times n}$, NMF seeks to find two matrices U and V whose product can approximate the original matrix X, described as

$$X \approx UV^T \tag{2.1.2}$$

where $U \in \mathbb{R}^{m \times k}$,$V \in \mathbb{R}^{n \times k} \geq 0$ and $k < min(n, m)$.

To evaluate the quality of reconstructing the original matrix X by multiplying U

and V, a square of Euclidean distance is proposed as a cost function to quantify the difference between two matrices:

$$\mathbf{O} = \left\|X - UV^T\right\|^2 = \sum_{i,j} \left[X_{ij} - (UV)_{ij}\right]^2 \tag{2.1.3}$$

The objective of the cost function $\mathbf{O}$ in Eq. 2.1.3 is to minimise the residual sum of squares of the matrices. The objective function $\mathbf{O}$ has two main concerns. Firstly, It is not convex in both approximate matrices U and V simultaneously. Secondly, it is an inverse problem which does not have a unique solution. In order to overcome these concerns and solve the objective function in Eq. 2.1.3, Lee & Seung (2001) propose two iterative update algorithms, which they proved could find the local minima of the objective function $\mathbf{O}$. The two iterative algorithms are described as

$$v_{jk} \leftarrow v_{jk} \frac{\left(X^T U\right)_{jk}}{\left(VU^T U\right)_{jk}}, \qquad\qquad u_{ik} \leftarrow u_{ik} \frac{(XV)_{ik}}{\left(UV^T V\right)_{ik}}$$

where u_k is the column vector of the feature matrix U, and v_j is the compressed version of the data point with respect to feature matrix U. The constructed data point x_j is approximated by a linear combination of the column vectors in U multiplied by the component v_j in V matrix.

$$x_j = U v_j$$

A good approximation of the feature vectors in U will assist in discovering the structure of the data. NMF follows the notion of the parts-based local representation of data, which means that combining the parts will lead to forming the whole (Lee & Seung 1999). Also, it leads to learning the relationships among the parts.

The inherent non-negativity constraint on matrices allows the factorised matrices to reconstruct the original matrix by adding their parts. Furthermore, this property provides an elegant interpretation in many fields, unlike other matrix factorization methods such as PCA. For instance, in image reconstruction, the pixels in a greyscale image with positive intensities can be meaningfully interpreted. Also, the non-negative coefficients of the feature matrix U, which is known as a metagene matrix in DNA gene expression data, is easily considered as the contribution of the genes (Devarajan 2008).

This thesis investigates in Section 2.1.2 the research gaps of the effects of the class imbalance problem in the standard feature extraction methods PCA and NMF.

2.1.2 Class Imbalance in Feature Extraction Methods

Feature extraction techniques are considered preprocessing methods to reduce the number of dimensions in the dataset. In the case of imbalanced data, the extracted features tend to be biased to predict the majority class samples potentially leading to poor performance on classification (Wang et al. 2012). The unsupervised PCA algorithm seeks orthogonal feature extractors that maximise the total variance. Therefore, the extracted features favor the majority class because the number of majority samples is more than the minority class. On the other hand, the unsupervised NMF has recently been shown to be a very effective matrix factorization technique in approximating high dimensional data (Lee & Seung 2001). It is a vector space method that uses matrix factorization to find two non-negative reduced-dimension matrices W and H (Lee & Seung 2001). The factorised matrices W and H may be affected by the imbalance problem and the basis matrix W may be biased to represent the majority

class samples, because the magnitude of the residual squared error of negative class instances is much more than the positive ones. Based on the literature review for this thesis, there are no previous studies that address the imbalanced class problem in the most widely used feature extraction approaches namely PCA and NMF.

In this section, a number of existing algorithms are reviewed which explore the class imbalance problem in single-label classification. The surveys, Japkowicz (2000) and Weiss (2004), provide three strategies: data sampling, cost-sensitive learning and algorithm-level that have been proposed in the literature to alleviate the negative impact of the class imbalance problem on classification. The data sampling strategy has attracted much research attention due to its simplicity. The main target of this approach is to balance the dataset. Random undersampling (RUS) is the simplest form of the undersampling approach; it randomly removes examples from the majority class until the desired ratio is reached. However, eliminating samples from a dataset which originally suffers from a small number of samples will not be an appropriate solution in real applications (Yen & Lee 2009). An intelligent algorithm has been proposed in the undersampling approach, RUSBoost (Seiffert et al. 2010) which presents a novel hybrid of undersampling and boosting. However, RUSBoost is sensitive to some parameters such as the number of boosting iterations and the level of sampling which may affect its generalisation.

Similar to undersampling methods, the oversampling approach presents a simple method called random oversampling (ROS) which adds some duplicated instances on minority class samples. However, this technique is more prone to overfitting (Kubat & Matwin 1997). Several more advanced algorithms for oversampling data have been proposed, such as the Synthetic Minority Over-sampling Technique (SMOTE)

(Chawla et al. 2002), Borderline-SMOTE (Han et al. 2005), Adaptive Synthetic Sampling (ADASYN) (He et al. 2008). SMOTE creates synthetic minor instances. Borderline-SMOTE over-samples only the minority samples which are the closest to the decision boundary because they are most likely to be misclassified. Lastly, ADASYN involves the density of the minority sample in the original distribution to generate the synthetic minority samples. In fact, these techniques increase the training time of models, and they are based on k-nearest neighbour, which may not be useful in very high dimensional datasets (Drummond & Holte 2003).

Another strategy for dealing with the imbalanced data problem is cost-sensitive learning. Several methods modify a threshold for an existing classifier. For example, Duda et al. (2012) define the threshold equal to the ratio of misclassification costs. Elkan (2001) attaches the proportion of positive and negative class samples with the misclassification cost. Moreover, some work has been devoted to integrating the cost-sensitive term in standard support vector machines (Masnadi-Shirazi & Vasconcelos 2010). To this end, Dmochowski et al. (2010) propose a weighted maximum likelihood in cost-sensitive learning. They propose a sigmoid approximation of the risk to minimise the loss based on asymmetric misclassification cost value. The common disadvantage of cost-sensitive methods is that there is no clear strategy on shifting the threshold of any existing cost-sensitive classifier. Furthermore, it may need the involvement of domain experts to determine the costs. Finally, the algorithm-level strategy modifies existing learning algorithms to tend towards the minority samples. For example, the Hellinger distance decision tree (HDDT) (Cieslak & Chawla 2008) and Class Confidence Proportion Decision Tree (CCPDT) (Liu et al. 2010) have been proposed to modify the standard decision trees algorithm. However, these algorithms

solve the class imbalance problem at the classifier level.

Based on the literature review, the class imbalance problem is not explicitly handled in the standard feature extraction. Some approaches use the concept of projections to alleviate the imbalanced class problem by enhancing the diversity. For example, Rotation Forest (Rodriguez et al. 2006) is an ensemble of decision trees based on PCA to generate different groups of extracted features for each base classifier.

This thesis proposes a cost-sensitive strategy to handle the imbalanced data in PCA and NMF in Section 3.1.

2.1.3 Imbalanced Data with Supervised NMF

Supervised feature extraction methods have received considerable attention in the data mining community due to their capability to improve classification performance compared to unsupervised dimensionality reduction methods. They are proposed to enhance the classification of factorised matrices by utilizing class labels (Huh et al. 2013). The most popular methods in supervised feature extraction are Supervised PCA (Chen et al. 2008), Supervised NMF (Huh et al. 2013), and Discriminant Non-Negative Matrix Factorization (DNMF) (Jia et al. 2015). Although these methods are useful in dimensionality reduction, they are not appropriate for imbalanced datasets because the factored components are biased towards predicting the majority class samples. This leads to poor performance in classification, and the class imbalance problem is exacerbated when dealing with high-dimensional data. Moreover, the resulting principal components of PCA and Supervised PCA (Chen et al. 2008) often

do not have a clear intuitive interpretation. In contrast NMF, because of the non-negativity constraint on its entries, provides an elegant interpretation of the data decomposition in many fields.

Based on the literature review, there is no one method address the imbalanced class problem in the supervised NMF. Based on this drawback, this thesis proposes a balanced supervised NMF algorithm in Section 3.2.

2.1.4 Sampling Data using the Artificial Bee Colony Algorithm

In addition to addressing the research gaps of imbalanced data in unsupervised and supervised feature extraction methods, this thesis investigates the class imbalance issue in swarm intelligence algorithms. Swarm intelligence algorithms simulate the behaviour of natural biological systems. They are composed of a population of single agents that interact with each other. They solve optimization problems by simulating the global behaviour of the internal agents (Bonabeau et al. 1999). First, a description of the standard artificial bee colony algorithm is presented, followed by a review of the previous methods in the data sampling approach in heuristic search algorithms.

Artificial Bee Colony (ABC) Algorithm

Artificial Bee Colony (ABC) was proposed for solving optimization problems. It belongs to the swarm intelligence algorithms which simulate the behaviour of natural biological systems. It is based on the foraging behaviour of real honeybees, which are classified into two kinds: employed and unemployed foragers. Employed bees exploit food sources and bring information about nectar to the hive to communicate with

unemployed bees. Unemployed bees are of two types: onlookers, that wait in the hive for shared nectar information; and scouts, that search for new food nearby (Jeanne 1986). When employed bees bring nectar information from the food source to the hive, they aim to communicate with onlookers to choose the best quality nectar among the food sources. Bee communication occurs through dance with the direction and duration related to the distance and direction of the food. The amount of loaded nectar refers to food quality. Onlooker bees watch the employed bees to choose the best food source based on nectar quality. Food sources are abandoned once employed bees have fully exploited them, at which time the employed bees become scouts.

In ABC, food sources represent potential problem solutions with each initially exploited by one artificial bee. The nectar information in the food source is the value of the solution. The bee colony is divided in half between employed and onlooker bees. Food sources are abandoned after several iterations, i.e. when the search does not find better quality neighbouring solutions. Employed bees become scouts and start to search for new solutions. The original algorithm, as shown in Figure 2.1, has the following stages: initialisation, employed bee, onlooker, and scout stage.

Initialisation The ABC algorithm starts by randomly creating N food sources (i.e. potential solutions) with food source i characterised by a vector $FS_i \in \mathbb{R}^D$ with D, the problem dimensionality. FS_{ij}^{min} and FS_{ij}^{max} define the minimum and maximum values of the jth element of FS_i. Food sources are initialised as

$$FS_{ij}^{'} = FS_{ij} + r(FS_i^{max} - FS_j^{min}) \tag{2.1.4}$$

where $r(0, 1)$ is uniform random number in [0, 1].

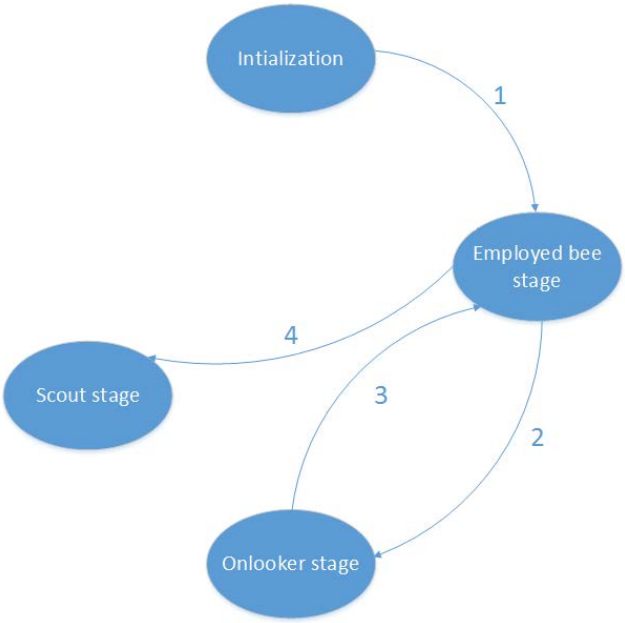


Figure 2.1: ABC search stages

Employed bee stage This stage associates a single employed bee with food source i and finds the neighbourhood of the food source (i.e. other potential solutions) using

$$FS_{ij}^{'} = FS_{ij} + \varphi_{ij}(FS_{ij} - FS_{kj}) \tag{2.1.5}$$

where $\varphi \in [-1, 1]$ is a uniform random number, i indexes the N food sources, j indexes the elements of FS_i and $k \neq i$ is a uniform randomly selected FS other than i. The fitness function of FS_i is

$$fitness_i = \begin{cases} (1 + f_i)^{-1}, & \text{if } f_i \geqslant 0 \\ 1 - f_i, & \text{if } f_i < 0 \end{cases} \tag{2.1.6}$$

where f_i is the cost (objective) function being maximised.

Onlooker stage Employed bees share their fitness with onlookers who select the maximum fitness among all food sources. The best result is recorded as a global

solution. The probability of selecting food source i is

$$p_i = \frac{fitness_i}{\sum_{i=1}^{N} fitness_i} \tag{2.1.7}$$

Scout stage Scout bees explore new food sources. Each employed bee has a limit on the number of times it can use the same food source. The count increments if the fitness value of the current food source is better than the neighbourhood one. Employed bees convert to scouts upon reaching the limit.

Sampling based on Swarm Intelligence Algorithms

The research by Yang et al. (2009) applies particle swarm optimization to optimize an imbalanced dataset. However, their algorithm is tested on a small amount of biomedical data only. Another study by Yu et al. (2013) proposes a method to select majority samples for undersampling but, in addition to being computationally intensive, and being restricted to DNA microarray data. Schiezaro & Pedrini (2013) apply ABC optimization for feature selection, but this thesis presents the first algorithm to balance imbalanced data using the ABC optimization algorithm. Moreover, it can easily be adapted to oversampling.

The main advantages of ABC over the other heuristic search algorithms are listed as follows. ABC has fewer parameters compared to other heuristic search techniques such as the genetic algorithm (GA) (Holland 1992), the particle swarm optimization algorithm (PSO) (Kennedy 2011) and the differential evolution algorithm (DE) (Storn & Price 1997). Apart from the maximum evaluation number and population size, ABC has only one control parameter (limit). But, PSO and GA each have three control parameters: cognitive and social factors, and inertia weight; and crossover

rate, mutation rate, and generation gap, respectively. Also, DE has two parameters: crossover and the scaling factor. Another issue is that the ABC algorithm addresses the diversity in the search space better than the GA and DE algorithms. Genetic algorithm and DE use a mutation operator that randomly modifies a part of the current solution. However, ABC balances between the local search process and the global search process. It uses a similar process to GA and DE to slightly modify a part of the current solution that is useful for a local search and the speed of convergence. On the other hand, it removes the whole solution and generates a new random solution using a scout bee. This mechanism increases the diversity of solutions of ABC, and it avoids premature convergence in the search. Finally, the ABC algorithm may replace the global best solution if it reaches the maximum exploitation limit by a new random solution. But, in the case of PSO and DE, and in some cases of GA, they keep the best solutions in the population which always contributes to producing new solutions (Karaboga & Basturk 2007).

Due to the many advantages of ABC compared to the other optimization methods, this thesis adapts the standard ABC algorithm to handle the class imbalance problem, as discussed in Section 3.3.

2.2 Dimensionality Reduction with Highly Correlated Features

Feature selection is considered a prerequisite step for the analysis of high-dimensional datasets. It reduces the number of dimensions by selecting a certain number of best features which can explain the differences between different classes (Guyon & Elisseeff

2003). Many benefits are gained by performing feature selection, including the ability to better understand data which has more informative features, a reduction in the complexity and computation time of the learning model, and the removal of noisy features. There are three types of feature selection approaches as defined by Saeys et al. (2007): filter, wrapper and embedded approaches, the main difference between them being that the filter approach is independent of any classification algorithm. However, wrapper and embedded approaches use the classification algorithm in the feature selection process. Wrapper feature selection evaluates the goodness of features using the classification algorithm and the embedded approach performs feature selection during the learning process.

In machine learning, many feature selection algorithms have been proposed to select the important features and eliminate the unimportant ones. However, most of these existing algorithms follow an individual feature ranking approach which discards the existence of the correlated features including SVM-RFE (Guyon et al. 2002), LASSO (Tibshirani 1996), and others. This section reviews the previous work related to feature selection algorithms in the presence of highly correlated features.

Feature selection for predictive models in the presence of high-dimensional and imbalanced data with many highly correlated covariates is a challenging problem that affects many disciplines. Initially, researchers were unaware of the importance of correlated covariates in interpreting predictive models. However, recent studies have been conducted to interpret groups of highly correlated features to identify significant functional modules, to improve classification accuracy, and to reflect on the semantic components of these features. As a motivating example, Bondell & Reich (2008) investigated the best soil characteristics in the Appalachian Mountains of

North Carolina. The authors identified 15 soil characteristics as potential predictors, of which half were highly correlated. Another example from genomics demonstrates that groups of highly interconnected genes are often closely linked to specific functional categories (Stuart et al. 2003). Moreover, they are of biological interest since the correlated genes have a significant relationship with clinical traits such as disease status, survival time or ageing (Horvath 2011, Ivliev et al. 2010). Several approaches have been proposed in the literature including penalized regression methods and supervised clustering methods.

2.2.1 Penalized Regression Methods

Variable selection in regression is a very challenging problem when there is a large number of highly correlated variables. Since Tibshirani (1996) proposed the popular least absolute shrinkage and selection operator (LASSO) method, many other methods based on regularization have been developed for model selection, particularly in the presence of correlated variables in high dimensions. LASSO imposes l_1 regularization on the predicted coefficients. It is a shrinkage and variable selection model due to shrinking several coefficients to zero. LASSO is not appropriate if there are groups of highly correlated variables because it tends to randomly select any variable from each group and discards the rest. This results in difficulties in interpreting variables and leads to unstable outcomes (Tibshirani et al. 2005). Several penalized regression methods have been proposed for grouped predictors, most of which are extensions of LASSO (Bondell & Reich 2008, Zou & Hastie 2005, Tibshirani et al. 2005). They introduce a new penalty term in addition to l_1 to take into account the correlated features. For example, Tibshirani et al. (2005) proposed Elastic Net which in addition to

LASSO, uses ridge regression regularization (l_2 norm) to average the highly correlated predictors (Zou & Hastie 2005). Unlike Elastic Net, Fused LASSO (Tibshirani et al. 2005) makes an assumption on the hierarchical structure of the features to impose a certain type of ordering. The additional penalty term of Fused LASSO next to l_1-norm encourages sparsity in the differences of the feature's coefficients. However, such ordering of the features does not necessarily exist in many applications. Bondell & Reich (2008) proposed OSCAR, which groups correlated features and assigns equal weights to features in the same group by applying a linear combination of sparse norm l_1 and pairwise l_∞ norm penalties. A recent feature selection method (HLR) is proposed by Huang, Liu & Liang (2016). HLR is similar to the Elastic Net method, but it adds the regularization term $l_{1/2}$ in addition to ridge regression regularization l_2. However, assigning equal weights for any two features whether they are nearby or away will lead to superfluous bias, particularly for large coefficients (Yang et al. 2013). Furthermore, a common drawback for Elastic Net, OSCAR and HLR is that neither take into account the correlation structure among features (Bühlmann et al. 2013). Moreover, the penalized regression methods do not take into consideration the class imbalance problem.

Other sparse model methods are proposed to reduce the number of dimensions including group lasso (Meier et al. 2008) and fused SVM (Rapaport et al. 2008). They suffer from correlation bias during the feature weighting process, because they assign weights based on group size (Toloşi & Lengauer 2011). Therefore, features which belong to a big group may be assigned weights. Furthermore, they are considered as parametric methods which need to set some parameters beforehand, which is not guaranteed to hold in practical applications (Conn et al. 2015).

Inspired by the regularization idea for feature selection, this thesis proposes an algorithm in Section 4.1 to incorporate the correlation structure of the features with the original data and $l_{2,1}$-norm regularization simultaneously in an iterative optimization method.

2.2.2 Supervised Clustering Methods

Identifying groups of correlated features using cluster analysis is proposed in several studies. For example, in genomics, techniques have been proposed to perform supervised gene clustering along with sample classification (Dettling & Bühlmann 2004, Park et al. 2007). Park et al. (2007) first apply hierarchical clustering on the predictors, then generate super-genes which are considered as the average of grouped predictors. After clustering, it uses LASSO to select the best subset of features from the grouped predictors. However, as discussed earlier, combining features beforehand can decrease the group's predictive strength (Bondell & Reich 2008).

Interestingly, several extensions of NMF have been proposed to produce predictive features for classification. Fisher-NMF incorporates the Fisher constraint in the NMF objective function (Jia & Turk 2004). However, the objective function may not be convex, which is not guaranteed to converge. Recently, a Supervised-NMF method (Huh et al. 2013) was presented to adopt graph embedding with NMF to improve classification using the generated features. In a general sense, all of the previous work tried to exploit the availability of labels to matrix factorization techniques in order to improve classification accuracy. But none of these techniques handle the class imbalance problem and they do not incorporate co-expression networks to find group predictors. Recent work, called the fuzzy forests method, has been proposed by Conn

et al. (2015) which uses recursive feature elimination random forests to select the features from the correlated feature blocks. Fuzzy forests depends on random forest feature selection which has a high computational complexity in terms of running time compared to the feature selection method using the support vector machine (Anaissi et al. 2016). Furthermore, the fuzzy forests method does not take into account the imbalanced data problem which may generate features which are biased towards the majority class. A drawback of supervised clustering methods is that they do not identify the correlated features to improve the classification performance along with the interpretation

The proposed fuzzy ensemble feature learning algorithm in this thesis discussed in Section 4.2 selects the top features by increasing the correlation within the modules and decreasing it between the modules. Also, it can handle the class imbalance problem along with feature ranking.

2.3 High-Dimensional Imbalanced Data in Multi-Label Classification

Multi-label classification (MLC) is the problem of classifying instances into multiple categories or classes simultaneously. For example, in text categorization, a document can belong to multiple topics (Ueda & Saito 2003); similarly in genomics, a gene or protein is related to multiple functional labels (Barutcuoglu et al. 2006). This should not be confused with multi-class classification, where one class is to be predicted for each instance. Multi-label classification methods fall into two categories: algorithm

adaptation and problem transformation (Cherman et al. 2011). Algorithm adaptation methods directly extend specific classifiers to handle multi-labeled data. Problem transformation methods, on the other hand, transform the multi-label problem into more than one single-label problem. Several algorithms have been proposed in the literature which follow the problem transformation approach including binary relevance (BR) (Boutell et al. 2004), classifier chain (CC) (Read et al. 2011), label powerset (LP) (Tsoumakas & Katakis 2006), and others. Binary relevance is a simple method which builds independent binary classifiers for each label, and the predicted label for the new instance is determined by aggregating the predicted results from all classifiers. This method does not take into consideration label correlations. Classifier chain is an extension of BR. However, it takes into account dependencies among labels by adding the labels into the feature space of the data. Particularly, given a set of labels L, an instance x_i from dataset X is a set of features, and y_i is the subset of the labels. Classifier chain builds a number of classifiers of size $|L|$. For example, to build the classifier c_j of label l_j, the instance x_i is transformed into $(x_i, l_1, \ldots, l_{j-1})$. Label powerset (LP) maps the multi-label problem into a single-label multi-class classification problem. It modifies the label space of the data into a set of combinations between the labels, which is the power set of all labels. This method takes into consideration label dependency. However, it is not appropriate for large scale label space due to the exponential number of possible label sets, as the number of label sets reaches up to $2^{|L|}$.

Multi-labeled data is similar to single-labeled data in that it also contains several complexities such as high dimensional data, the class imbalance problem, incomplete label space, and others (Jian et al. 2016, Wu et al. 2016). However, these problems are

more challenging in multi-labeled data for several reasons, including label correlations, missing labels, feature correlations, and a large number of labels. This section presents the previous work for high-dimensional multi-labeled data, imbalanced multi-labeled data, and incomplete label space.

2.3.1 High-Dimensional Multi-Labeled Data

In past decades, many single-label feature selection methods have been proposed to reduce the number of dimensions of large datasets in several domains (Zhou & Tuck 2007, Tibshirani 1996). However, these methods are not suitable for multi-label high-dimensional data.

Multi-label learning, similar to single-label learning, suffers from high-dimensional data which contains noisy and irrelevant features. In the literature, several methods transform the multi-label problem into single-label sub-problems, and make use of ReliefF or information gain to measure the features' importance (Spolaôr et al. 2013, Lee & Kim 2015, Tsoumakas et al. 2011). However, none of these methods take into consideration label correlations. A recent study by Jian et al. (2016) exploits the label correlations to select the discriminative features. However, this study assumes global label correlations among instances. Furthermore, it uses Latent Semantic Indexing (LSI) (Dumais 2004) to decompose the multi-labeled output space, which generates mixed sign values in low-dimensional space. Therefore, the low-dimensional space cannot be interpreted, and the structure of the original matrix is unrecoverable.

Interestingly, several methods have been proposed as multi-label feature extraction techniques. They do not preserve the original features, but they project the data to a new space. These methods, including the method called MLLS proposed by Ji et al.

(2008), extract the common features among multiple labels. Another study by Zhang & Zhou (2010) proposed MDDM to transform the original data into a reduced space by maximizing the dependence between the original features and related labels. The MLLS and MDDM methods take into account the interaction between data and labels, however, they do not incorporate the label correlations in their methods. Another feature extraction technique, "Multi-Label Learning by Exploiting Label Correlations Locally" (ML-LOC), exploits label correlations locally (Huang et al. 2012). A method "Sub-Feature Uncovering with Sparsity method" (SFUS) (Ma et al. 2012) uncovers the shared feature subspace, but it does not explicitly exploit the label correlations in feature selection.

Based on the highlighted drawbacks, this thesis proposes in Section 5.1 a multi-label feature selection method by using the correlation information. In contrast to the drawbacks of the previously mentioned methods, the proposed method differs from the others in the following aspects: (1) it is considered a comprehensive method which simultaneously addresses the most important challenges in multi-label learning in one approach by incorporating the correlation information of features, labels and samples; (2) it is a feature selection method which preserves the physical meaning of the data by selecting the discriminative features to reduce the space; and (3) it is a variant of the NMF method which allows the interpretation of low-dimensional space, and the structure of the original data can be recovered and explained.

2.3.2 Class Imbalance and Incomplete Label Space

This section reviews the existing approaches related to MLC methods that handle a noisy and incomplete label matrix, and imbalanced multi-labeled data. Several approaches integrate label correlation with the feature selection process, namely MIFS, "Correlated multi-label feature selection", and "Feature selection for multi-label naive Bayes classification method" (Jian et al. 2016, Gu et al. 2011, Zhang et al. 2009), and a number of methods which extract the low dimensional space shared among all labels by projecting the original data matrix with the correlated labels (Huang, Li, Huang & Wu 2016, Ji et al. 2008). However, these approaches exploit the label correlations as prior knowledge, which is typically constructed from an incomplete original label matrix. In contrast, a small number of projection approaches have been proposed that tackle the noisy label problem by projecting the original label matrix to a low dimensional space (Hsu et al. 2009, Zhou et al. 2012), but the major drawback of these approaches is that the label projection process is achieved before it can be embedded in the model training steps. A recent study has been proposed to learn a new supplementary label matrix (Xu et al. 2014), but the new label matrix depends only on the original label correlations and it does not handle the imbalanced class problem.

Class imbalance in multi-label learning has not been extensively handled in the literature, as stated recently by Wu et al. (2016). Typically, the methods proposed in previous studies to handle class imbalance can be divided into three main categories: data sampling strategy, cost-sensitive strategy and algorithmic adaptations. These three categories are inspired by the class imbalance problem in single-label learning. The data sampling strategy has attracted research attention due to its simplicity.

It aims to modify the distribution of the dataset to make it balanced, and this is achieved by using undersampling and oversampling techniques. The former reduces the number of majority class samples, while the latter increases the number of minor class samples (He & Garcia 2009).

A simple undersampling algorithm for multi-label learning was proposed by Dendamrongvit & Kubat (2009) to classify multi-label text datasets. The authors considered the problem as a binary class problem by handling each label separately. Instances that belong to a certain label are considered positive labels and others are regarded as negative labels. The number of majority samples is then reduced to lower the bias in each label. Another undersampling method was presented in Charte et al. (2015*a*), which randomly reduces the number of majority samples. An inverse random undersampling (BR-IRUS) technique which was originally proposed for single-label classification was applied by to multi-labeled datasets (Tahir et al. 2012). The idea of this technique is to train multiple classifiers on balanced data by using all the minority samples and a subset of the majority samples. However, these techniques may exclude some important information by randomly removing the majority samples. Also, they are sensitive to the number of retained samples, which may affect the generalization of the method. Two techniques based on random oversampling have been proposed, namely, LP-ROS and ML-ROS (García et al. 2012). Both techniques randomly duplicate the instances associated with the minority class, but random oversampling is more prone to over-fitting. Several more advanced over-sampling algorithms have been proposed such as, SMOTEUG (Giraldo-Forero et al. 2013) and MLSMOTE (Charte et al. 2015*b*), which are based on the classic SMOTE method in single-label learning. Briefly, the SMOTE method creates synthetic minor instances

to balance the dataset. SMOTEUG proposed a method to identify the seed samples in order to clone them using the SMOTE algorithm. However, only one minority label is considered in SMOTEUG to find the seed samples. In MLSMOTE, the set of samples from each minority label is processed instead of synthetic samples from one minority label being generated. Typically, over-sampling techniques increase the model training time due to the addition of more samples, which may not be appropriate in the case of large datasets. Over-sampling has several disadvantages: (1) they are considered as simple techniques, which ignores the intrinsic nature of the multi-labeled datasets; (2) they modify the distribution of the dataset, which may degrade the performance of the classifiers; (3) they do not incorporate the correlation between labels, which helps to predict new labels, and identifies missing and noisy labels.

Compared to single-label learning, cost-sensitive learning studies in MLC are very few. A study conducted by Dembczynski et al. (2013) maximizes the imbalance specific metric F1-macro. However, this algorithm is dependent on the specific evaluation metric. Also, it is difficult in cost-sensitive learning strategies to set the values of the cost matrix (Guo et al. 2017). Several algorithmic adaptation methods have been proposed recently to deal with the imbalanced class problem in multi-label learning. For example, Tepvorachai & Papachristou (2008) introduces a method of forming a smaller subset of balanced data from the imbalanced multi-label training data, which was used to train a single artificial neural net (ANN) classifier. The subset of data is initialized by clustering the imbalanced data into multiple clusters, and an equal number of data are collected from each cluster. During the training phase of the

ANN model, samples are incrementally added and removed to improve training performance. This method is closer to being an ensemble method because it does not explicitly modify the learning model. Also, it is necessary to set a predefined parameter for the number of clusters, which is not considered a trivial task. Another method based on ANN has been proposed, called IMIMLRBF (Li & Shi 2013). The IMIMLRBF method takes into account the number of samples per label when it selects the number of units in the hidden layer, and the weights associated with the hidden layer are adjusted according to the individual bias of each label. In the proposal by Chen et al. (2006), the authors used min-max modular classifiers which divide the problem randomly into several smaller binary SVM classifiers. This decomposition is achieved by a clustering method or by using principal component analysis. The proposed methods in the algorithmic adaptation strategy are designed to be algorithm dependent as previously mentioned. A recent method is "Towards Class-Imbalance Aware Multi-Label Learning" (COCOA) (Zhang et al. 2015), in which the classification decision is predicted by combining a binary-class imbalance of the current class with a multi-class classifier for other classes. However, the authors do not consider the extreme imbalance ratio in their study; also, noisy and mislabelled classes are ignored.

Very recent work, called "Constrained Submodular Minimization for Missing Labels and Class Imbalance in Multi-label Learning" (MMIB), was proposed by Wu et al. (2016) to jointly handle the class imbalance problem and missing labels in multi-label learning. However, the MMIB method is conducted separately from the model training steps, which implies a high possibility that the new label representations may not be aligned with the multi-label trained model. Also, the computational

complexity of the training model will be high as a result of handling missing labels and class imbalance separately from the training phase.

It can be seen from this review that, the existing multi-label feature selection algorithms in the literature do not take the imbalanced class problem into consideration, which leads to the selection of features that favor the majority label. In popular multi-label feature selection methods "Multi-label informed feature selection" (MIFS) (Jian et al. 2016) and "Web image annotation via subspace-sparsity collaborated feature selection" (SFUS) (Ma et al. 2012), the former exploits the label correlations to extract the common features among multiple labels, and the latter discovers the shared feature space in multi-label learning with feature selection. The common drawback is that neither takes into account the imbalanced class problem. They both also they use the original noisy and incomplete labels to exploit the label correlations.

The proposed integrated multi-label approach, discussed in this thesis in Section 5.2, ML-CIB, simultaneously handles the incomplete label and imbalanced class problem in MLC and predicts a new label matrix which accurately defines the contribution of the label, as shown in the example in Figure 5.6. Furthermore, the proposed method can be used as a multi-label feature selection method which reduces the dimensions of large datasets by selecting balanced features to improve the accuracy of MLC models.

2.4 Classifiers used in this Thesis

Several popular classifiers are used in this thesis to evaluate the performance of the proposed methods along with state-of-the art methods. This section reviews the classifiers used, namely the support vector machine (SVM), k-Nearest neighbour (kNN),

and Naive Bayes.

2.4.1 Support Vector Machine (SVM)

The support vector machine is recognized as one of the top algorithms in machine learning (Wu et al. 2008). It was proposed by Vapnik (2000) for classification and regression analysis. SVM has been extensively used in different applications (Catania et al. 2012, Manimala et al. 2012, Libbrecht & Noble 2015). The idea of SVM is to find an optimal hyperplane from a set of available hyperplanes in order to separate between a set of points. The chosen hyperplane must maximise its margin, which is the distance between the hyperplane and the nearest points from each class, which are known as support vectors (Chang & Lin 2011). SVM is a supervised learning algorithm which builds a model based on a given set of samples with known labels or categories, and the model predicts the label or category of the new observation. There are two types of SVM: binary and multi-class SVM. Binary classification is where the output of the new samples is mapped to one of two classes or categories. Multi-class SVM, on the other hand, is where the input is mapped to one of a set of multiple classes (more than two classes). This thesis evaluates the algorithms based on binary SVM.

In linear binary SVM, to determine the optimal hyperplane, there is no need to use all training points, but rather, only a small set of these points known as support vectors. Support vectors lie on the margin of the hyperplane. Given a set of training points $(x_1, y_1), \ldots, (x_i, y_i), \ldots, (x_N, y_N)$, where x_i is the input sample that is labeled by $y_i \in \pm 1$.

$$\begin{aligned} \hat{w}, \hat{b} = \underset{w,b,\xi \geq 0}{\operatorname{argmin}} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{N} \xi_i \\ \text{s.t. } y_i(w^T x_i + b) \geq 1 - \xi \quad i = 1 \dots N \end{aligned} \tag{2.4.1}$$

where w is the weight vector and b is the bias. For example, in the case of two dimensional data, the hyperplane $y = w^T x + b$ is a straight line as shown in Figure 2.2. The SVM score is the value of a function $w^T x + b$, the score of a point $x_i \geq +1$, then the class of this point is assumed to be a positive class $y_i = 1$. However, if the score is ≤ -1, then it will be classified as a negative class. The slack variable ξ is used to relax the constraints when the above assumption is not valid. The distance between the closest points from different classes is $\frac{1}{\|w\|}$, therefore maximizing the margin hyperplane $\frac{1}{\|w\|}$ is equivalent to minimizing $\frac{1}{2} \|w\|^2$. Based on the training model which finds the best $\hat{w}$ and $\hat{b}$, the new point x_j is classified as

$$\begin{aligned} w^T x_j + b \geq 0 \quad \text{positive class} \\ w^T x_j + b < 0 \quad \text{negative class} \end{aligned} \tag{2.4.2}$$

A support vector machine can deal with non-linearly separable data, where it needs a non-linear classifier to separate between different classes. An SVM linear classifier can be transformed to an SVM nonlinear classifier by using the kernel trick. Kernel methods are able to transform the data from its original space to feature space by computing the inner products in feature space which are known as kernel trick. In the case of SVM, point x is replaced by $\phi(x)$ where $\phi : x \to f$ is a nonlinear function, therefore, the nonlinear decision boundary is

$$y = w^T \phi(x) + b \tag{2.4.3}$$

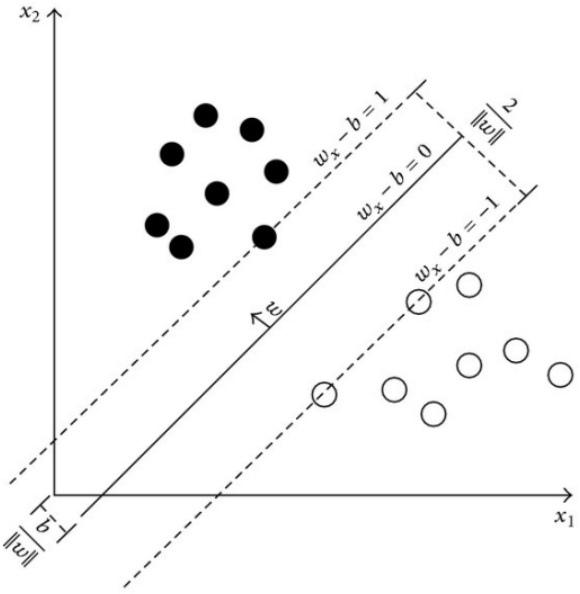


Figure 2.2: Binary SVM classifier from Koo et al. (2013)

Table 2.1: The popular kernel methods

Kernel name	Kernel function
Linear kernel	$k(x_i, x_p) = x_i^T x_p$
Polynomial kernel	$k(x_i, x_p) = (\eta x_i^T x_p + c)^d$
Gaussian kernel	$k(x_i, x_p) = \exp\left(\|x_i - x_p\|^2 / 2\sigma^2\right)$
Sigmoidal kernel	$k(x_i, x_p) = \tanh(\eta x_i^T x_p + c)$

Kernel trick is used to efficiently compute the inner product between $\phi^T(x_i)\phi(x_p)$, which is much faster than the standard inner product. There are several types of kernel functions as shown in Table 2.1 (Kavzoglu & Colkesen 2009), where x_i and x_p are two samples, η is the slope, c is the constant term, and d is the polynomial degree.

2.4.2 k-Nearest Neighbour (k-NN)

The k-Nearest neighbour classifier (k-NN) (Cover & Hart 1967) is selected as one of the top ten algorithms in data mining (Wu et al. 2008). The idea of this algorithm is

explained in the following example. Suppose a bank wants to know whether a client will rent or own a house, so instead of calculating the credit rating for a certain person, the algorithm searches for the most similar case and uses it to predict the decision of the client. The k-NN classifier uses a distance/similarity function to classify the new points based on its neighbours. The parameter k defines the number of nearest neighbours that must be considered to identify the class of the new sample. Assuming $k = 3$, the algorithm will select the closest three points and classify the new point based on the majority class of the three neighbours. It calculates the Euclidean distance between the new point and the existing points. The k-NN algorithm is simple, useful and very easy to understand. k-NN has been extensively used in several domains including genomics (Diaz et al. 2009), text mining (Khan et al. 2010), and others. However, it has the following disadvantages (Cunningham & Delany 2007):

- It is sensitive to noise and irrelevant features.
- It depends on the distance which may not work in the case of high dimensions and a low number of samples.
- It weights all the features equally, including noisy and irrelevant data.
- It is sensitive to imbalanced class distribution which may mislead the prediction decision.

2.4.3 Naive Bayes

The Naive Bayes classifier is a probabilistic classifier based on Bayes' rule, which describes the probability of an event based on prior knowledge of events that have already occurred which might be related to the event in question (Bishop 2006).

For example, if loans are related to a borrower's salary, then, using Bayes' rule, a borrower's salary can be used to increase the chance of the borrower to successfully applying for a loan.

Naive Bayes builds a binary classification model using a set of training points with known class membership. It performs the classification by computing the score of the new point in every class. If the score of class 1 is greater than class 2, then it will be assigned to class 1, and vice-versa. Given a binary class $C_k \in (0, 1)$, and $X \in \mathbb{R}^{n \times p}$, where n is the number of points in each class, and p is the number of features, Naive Bayes assumes that the features are independent in order to compute the likelihood as a conditional independence probability defined as

$$P(C = C_k|X_1, \ldots, X_p) = P(C = C_k) \prod_{j=1}^{p} P(X_j|C = C_k) \tag{2.4.4}$$

Therefore, to classify a new sample X^{new}, a posterior probability for X_{new} is computed as

$$C^{new} = \underset{C_k}{\operatorname{argmax}}\, P(C = C_k) \prod_{j=1}^{p} P(X_j^{new}|C = C_k) \tag{2.4.5}$$

Naive Bayes has been successfully used in several applications (Patil & Sherekar 2013, Liu et al. 2013). It has several advantages including that it is easy to train, easy to update with new training data, and simple to understand. However, the assumption of independent features with respect to the class is not reasonable in many datasets which contain highly correlated features (Wu et al. 2008).

2.5 Research Gaps

This chapter reviews the current research based on three questions: class imbalance on single-labeled data, highly correlated high-dimensional data, high dimensional and class imbalance multi-labeled data. This thesis proposes several algorithms to address the research gaps which are identified in this chapter and are listed as follows.

The current class imbalance methods in single-labeled data fail to address the class imbalance in unsupervised and supervised feature extraction methods, namely PCA and NMF. For highly correlated high-dimensional single-labeled data, the current feature selection methods either consider correlated features as redundant or they do not explicitly investigate the classification capabilities of the correlated feature groups. For multi-labeled data, current multi-label feature selection fails to take into account the correlation information during the feature selection process to improve the performance of multi-label classification algorithms. Lastly, as reviewed in this chapter, relatively few current studies address the multi-label imbalance problem which suffers from several drawback as reviewed in Section 2.3.2. This thesis proposes a method to handle the class imbalance problem and missing labels in the label matrix to improve the performance of multi-label classification algorithms.

Chapter 3

Class Imbalance in Single-Labeled Data

This chapter introduces several approaches to handle the imbalance class problem in binary classification models for single-labeled datasets. The class imbalance issue is caused by unequal distributions of the data between class labels (Chawla et al. 2004). It occurs due to a paucity of cases, for example, patients with a rare disease, or difficulties in collecting samples due to high cost or privacy. The imbalanced class problem is considered a crucial issue in machine learning and data analytics for two reasons. Firstly, learning from an imbalanced dataset leads to poor classification because classical data mining algorithms tend to favor classifying examples as belonging to the majority class (negative class). Thus, these base learning algorithms are less capable of classifying the instances of the minority class (positive class), which are usually the class of interest. Secondly, it is a common problem in many real-world domains including those related to biomedicine, finance data and others. Based on the literature review of the class imbalance approaches presented in chapter 2, there is no unique algorithm or method that is optimal in handling the imbalanced class problem and it is still considered to be an open problem in the data analytics field (Krempl

et al. 2014). To solve this problem, this chapter proposes three novel algorithms to: (1) address the imbalanced data issue in the traditional feature extraction techniques; (2) handle the class imbalance problem in supervised non-negative matrix factorization; and (3) handle the imbalance class problem by proposing a wrapper method to select the optimal instances. The rest of this chapter is organised as follows. Section 3.1 introduces a cost-sensitive learning strategy to handle the imbalanced data in classical feature extraction methods, namely PCA and NMF. Section 3.2 presents the Balanced Supervised NMF algorithm (BSNMF) to handle the class imbalance problem in the supervised NMF method. Finally, section 3.3 introduces the ABC-Sampling algorithm for balancing imbalanced datasets based on the Artificial Bee Colony algorithm. In each section, to validate the proposed algorithms, sub-sections describe the used datasets and experiments, compare the state-of-the-art algorithms, and give a conclusion.

This chapter, encapsulating **Contributions 1, 2 and 3** of this thesis, is an extended description of my publications ((Braytee, Liu & Kennedy 2016), (Braytee, Catchpoole, Kennedy & Liu 2016), and (Braytee et al. 2015)).

3.1 Class Imbalance on Classical Feature Extraction Methods

This section covers the theory, implementation and experiments for a cost-sensitive approach for classical PCA and NMF feature extractions which are able to solve the imbalanced class problem without modifying the existing base classifiers, or changing

the original information of the training datasets. Based on the literature review, this study is unaware of any previous work that addresses the imbalanced class problem in the most widely used feature extraction approaches, namely NMF and PCA.

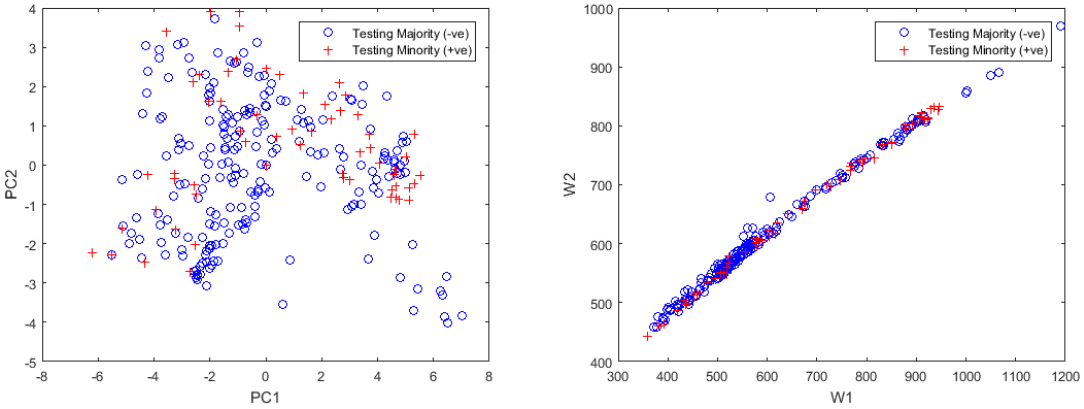
3.1.1 A Motivating Example

A training dataset is comprised of instances $\mathbf{X} \in \mathbb{R}^{n \times m}$ with n instances and m features. In a binary case, these instances belong to two different classes which are known as the majority (negative) and minority (positive) class respectively. In this section, a vehicle dataset is used and the description of the characteristics of this dataset is given in Table 3.1.

A major motivation for proposing a cost-sensitive feature extraction method is that classical PCA and NMF relatively ignore the minority samples during the feature extraction which negatively affects the performance of classifiers and leads to an overlap class problem that increases with the imbalance ratio.

Figures 3.1a and 3.1b show the distribution of test samples on original vehicle imbalanced dataset after a classical PCA and NMF is applied on it respectively. It clearly shows that the classifiers will find it very difficult to distinguish between classes due to the overlap class problem.

This thesis assumes that integrating the weighted label information in feature extraction methods can often improve the performance of supervised learning algorithms, as shown in section 5.2.3, because it generates extracted features that consider both classes and corrects the overlap class problem by separating the two classes.



(a) Testing data after projecting the instances using PCA

(b) Testing data after projecting the instances using NMF factors

Figure 3.1: Applying PCA and NMF on imbalanced vehicle data leads to the overlapping problem

3.1.2 Methods

In this section, the theoretical analysis of the proposed cost-sensitive PCA (CSPCA) and cost-sensitive NMF (CSNMF) is presented.

Imbalance Cost Ratio

Different cost ratios are used in all training examples for the majority and minority classes:

$$C_i = \begin{cases} C_i^- = \frac{(1-\alpha)}{N_-}, & \text{if } y_i = -1, \text{(Negative class)} \\ C_i^+ = \frac{(1+\alpha)}{N_+}, & \text{if } y_i = +1, \text{(Positive class)} \end{cases} \tag{3.1.1}$$

for $i = 1, ..., N$ samples, N_- and N_+ are the total number of negative and positive samples respectively. Parameter $0 \leqslant \alpha < 1$ weights the majority class. If $\alpha = 0$, the majority class is weighted by the ratio of the two class sizes in the training data. In the case where the α value is close to 1, it gradually represents the learning from the positive examples only.

Cost-sensitive Principal Component Analysis (CSPCA)

Principal component analysis (PCA) is one of the most popular feature extraction technique as reviewed in section 2.1.1. In the case of the class imbalance issue, the spread of data is dominated by the majority samples, because when the directions of the principal axes (components) of both classes are different, the reduced space found by PCA represents the majority space and under-represents the minority one.

Geometrically, the first step in PCA is to centre the data by subtracting the mean of the data from all points. However, in the case of imbalanced data, the global mean may be shifted to the majority samples space. Moreover, PCA computes the covariance matrix of the data which captures the variance of the dataset. But, in the case of highly skewed data, the covariance matrix mostly represents the variance of majority class samples, and the largest variance direction of the data may be captured mostly from the majority space.

Therefore, a cost-sensitive PCA (CSPCA) technique is proposed to improve the computations of the principal components with consideration to the imbalanced class issue. In the binary case, assume that the negative and positive samples are discounted by imbalance cost ratios C^- and C^+ respectively. The weighted first principal component becomes:

$$\mathbf{w}_1 = \arg\max \sum_{i:y_i=-1,j} \left(C_i^- \mathbf{X}_{ij} \cdot \mathbf{w}_{1j}\right)^2 + \sum_{i:y_i=+1,j} \left(C_i^+ \mathbf{X}_{ij} \cdot \mathbf{w}_{1j}\right)^2 \tag{3.1.2}$$

where C_i^- and C_i^+ are defined in Eq. 3.1.1, and j is the dimension index. Using different cost ratios for the negative and positive classes leads to lessening the dominant effect of the negative samples on the extracted features.

Cost-sensitive Non-negative Matrix Factorization (CSNMF)

Non-negative matrix factorization (Lee & Seung 2001) is a matrix factorization technique under the constraint that the values of the input matrix are non-negative. It can be described by the following factorization form

$$X^{n\times m} \simeq W^{n\times p} H^{p\times m} \tag{3.1.3}$$

where n is the number of observations, m is the dimension of the data, an integer p such that $p < \min(m, n)$ and $X \in \mathbb{R}^{+n\times m}$, $W \in \mathbb{R}^{+n\times p}$, $H \in \mathbb{R}^{+p\times m}$. The size of the matrices W and H is much smaller than X.

To find the approximate matrix factorization (3.1.3), an optimization function is defined by Lee & Seung (2001) to *minimize* $\left(\|X - WH\|^2\right)$ with respect to W and H.

In the case of imbalanced data, a cost-sensitive NMF (CSNMF) is proposed, which injects the classical unsupervised NMF with labelling information to take into consideration the class imbalance problem. The proposed CSNMF function modifies the original matrix to alleviate the effectiveness of the negative samples, and a new matrix X' is defined by

$$X' = \begin{bmatrix} \left[C_i^- X_i\right], \text{ if } y_i = -1 \\ \left[C_i^+ X_i\right], \text{ if } y_i = +1 \end{bmatrix} \tag{3.1.4}$$

where C^- and C^+ are defined in (3.1.1), $X' \in \mathbb{R}^{n\times m}$. CSNMF aims to find two non-negative matrices $W = [w_{ip}] \in \mathbb{R}^{n\times p}$ and $H = [h_{jp}] \in \mathbb{R}^{m\times p}$ whose products can estimate the balanced matrix X'.

The objective function is the Euclidean distance between two matrices. It can be written as:

$$O = \left\|X' - WH^T\right\|^2 = \left(X' - WH^T\right)\left(X' - WH^T\right)^T \tag{3.1.5}$$

The objective function O is convex with respect to W and H separately. Lee & Seung (2001) proposed iterative multiplicative update rules to minimise the error of O in Eq. 3.1.5

$$h_{jp} \leftarrow h_{jp} \frac{(X'W)_{jp}}{(HW^TW)_{jp}} \tag{3.1.6}$$

$$w_{ip} \leftarrow w_{ip} \frac{(X'^TH)_{ip}}{(WH^TH)_{ip}} \tag{3.1.7}$$

The convergence of the objective function with X' matrix is similar to the classical NMF which is proposed by Lee & Seung (2001).

Flowchart of the Weighting Strategy

Figure 3.2 shows the stages of implementing the proposed weighting strategy approach. The training set is weighted before applying the feature extraction PCA or NMF, then classifiers are built on the features extracted from the training data.

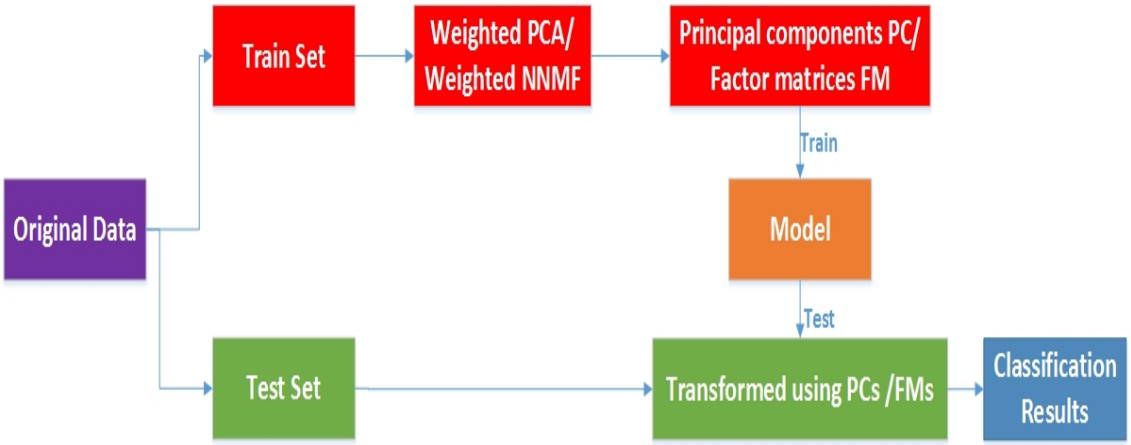


Figure 3.2: Weighting strategy flowchart

3.1.3 Experiments and Datasets

This section analyses and compares the performance of CSPCA and CSNMF against classical PCA and NMF, the existing ADASYN algorithm (He et al. 2008), random undersampling (RU), maximum likelihood cost-sensitive (ML-CST) (Dmochowski et al. 2010) and the Class Confidence Proportion Decision Tree (CCPDT) (Liu et al. 2010). Four classifiers are used as instance based learning approaches namely the support vector machine (SVM), Naive Bayes, decision trees (CART), and k-NN .

Experimental Framework

Firstly, a set of parameter values is defined for the classifiers that have been used in all the experiments. Without loss of generality, $k = 5$ neighbors in k-NN classifier is selected. The number of principal components is chosen to represent 90% of the variance of the original data, and the desired rank for CSNMF equals the number of class labels because it is highly related to the cluster structure of the data (Carmona-Saez et al. 2006). Also, the balance factor $\alpha = 0$ is set for imbalance cost ratio C in (3.1.1) to represent equal class proportions in the training set.

The proposed methods in this thesis are evaluated on 12 datasets, each with two classes. These datasets are selected from the UCI repository (Lichman 2013) and KEEL-datasets (Alcalá-Fdez et al. 2010). Details of the datasets are shown in Table 3.1. The AUC metric is used to measure the accuracy in the experiments. This is a widely used metric for evaluations on imbalanced data. Estimates of AUC were averaged over five-fold cross-validation.

Table 3.1: Characteristics of the datasets. Column #IR is the imbalance ratio (i.e., Neg/Pos), #Attr is the number of features.

Dataset	#Instance	#Attr	#IR	Dataset	#Instances	#Attr	#IR
Diabetes	768	20	1.86	Blood	748	5	3.2
Glass	214	9	6.38	Colon	62	2000	1.82
Yeast	1484	8	28.1	Survival	306	3	2.77
Spambase	4601	57	1.53	Adeno	86	76	5.34
Breast	198	34	3.21	Ecoli	336	7	8.6
Vehicle	846	18	2.99	Abalone	4174	8	129.43

3.1.4 Results and Analysis

This section analyses the behaviour of classical PCA and NMF with existing algorithms that are proposed to handle the imbalanced class issue against the proposed CSPCA and CSNMF methods. Firstly, the state-of-the-art methods are composed of two methods from the data sampling techniques, namely ADASYN (He et al. 2008) and random undersampling, one method from the cost-sensitive level, namely ML-CST (Dmochowski et al. 2010), and one method from the algorithm level, namely CCPDT (Liu et al. 2010).

The results in Tables 3.2 to 3.11 show that the performance of the classifiers substantially improves for the cost-sensitive versions of PCA and NMF. The proposed CSPCA and CSNMF using base classifiers outperformed the classical PCA and NMF along with the state-of-the-art methods. Therefore, the proposed method can solve the imbalanced class issue at the feature extraction level without the need to change the data distribution or modify the existing algorithms. To evaluate the statistical significance of the proposed algorithms, a t-test is conducted between the vector of the results of the proposed methods (Base) against the compared methods for each

classifier, under the null hypothesis that the AUC on the vectors of the used methods is not significantly different. As shown on the bottom line of Tables 3.2 to 3.11, the *p-values* reject the null hypothesis, as most values of the proposed methods (base) are lower than 0.01. This indicates that the proposed methods CSPCA and CSNMF improve the performance of the classifiers and this improvement is statistically significant.

Effectiveness of CSPCA and CSNMF on Classification Performance

In this section, the classification performance of the extracted features from the proposed CSPCA and CSNMF are compared to the baseline methods (classical PCA and NMF), along with data-sampling level methods, a cost-sensitive level method, and an algorithm level method. First, the effects of the proposed CSPCA and CSNMF algorithms are analysed with baseline and data-sampling level methods. The initial experiments are conducted on the baseline (classic PCA and NMF) on the imbalanced training set. The following experiment balances the imbalanced dataset by using sampling algorithms, namely ADASYN and random undersampling (RU) methods, then it applies the classical PCA and NMF to construct the reduced features. Finally, the balanced features are extracted from the proposed CSPCA and CSNMF on the imbalanced dataset. The projected test data based on the extracted features above is classified using the base classifiers, namely SVM, Naive Bayes, decision trees and k-NN. Tables 3.2 to 3.9 show the superiority of the proposed CSPCA and CSNMF method over the existing data sampling techniques. The best average values per each approach are highlighted in bold.

Second, the proposed CSPCA and CSNMF method are compared with the cost-sensitive level method namely, ML-CST. ML-CST is a dependent method which uses logistic regression as a base classifier. In this experiment, the features are constructed using PCA and NMF, then ML-CST is used to train a logistic regression model which classifies the projected test samples. For a fair comparison, the logistic regression classifier is used to train the extracted features from the proposed CSPCA and CSNMF. The results in Table 3.10 show that the AUC measures of the proposed algorithms are better than the-state-of-the-art methods in almost all datasets except one.

Finally, the last experiment is conducted to compare the proposed CSPCA and CSNMF with the algorithm level method, namely CCPDT, by using decision trees as a base classifier. Table 3.11 shows the quality of using the proposed solution to apply feature extraction on imbalanced datasets, as there is a statistically significant difference between the results of the proposed methods, the standard PCA and NMF, and the other compared algorithms. One may also observe the generalization of the proposed methods by improving the performance of classification on a set of well-known classifiers.

As highlighted earlier, CSPCA and CSNMF methods improve the performance of classification for two reasons. Firstly, they tackle the overlaping class problem by shrinking the majority samples using the imbalance cost ratio. Secondly, the extracted features of CSPCA and CSNMF consider the minority samples, even in highly imbalanced datasets.

Table 3.2: CSPCA SVM classification performance on imbalanced datasets compared to the algorithms. The best AUC are highlighted in bold.

Dataset	PCA	ADASYN- PCA	RU-PCA	CSPCA
Diabetes	0.695	0.706	0.576	**0.762**
Spambase	0.869	0.868	0.872	**0.904**
Breast	0.562	0.586	0.556	**0.620**
Ecoli	0.860	0.828	0.857	**0.879**
Abalone	0.669	0.627	0.666	**0.683**
Survival	0.607	0.546	0.536	**0.611**
Blood	0.676	0.676	**0.677**	0.660
Glass	0.945	0.933	0.900	**0.951**
Yeast	0.850	0.831	0.839	**0.871**
Vehicle	0.634	0.620	0.628	**0.681**
Colon	0.512	0.490	0.517	**0.531**
Adeno	0.579	0.525	0.497	**0.610**
t-test	3.4×10^{-3}	9.7×10^{-5}	5.3×10^{-3}	Base

Table 3.3: CSPCA Naive Bayes classification performance on imbalanced datasets compared to the algorithms. The best AUC are highlighted in bold.

Dataset	PCA	ADASYN- PCA	RU-PCA	CSPCA
Diabetes	0.641	0.694	0.633	**0.71**
Spambase	**0.857**	0.703	0.815	0.841
Breast	0.503	0.556	0.557	**0.611**
Ecoli	0.722	0.857	0.864	**0.881**
Abalone	0.500	0.645	0.692	**0.711**
Survival	0.515	0.532	0.510	**0.590**
Blood	0.546	0.655	0.651	**0.741**
Glass	0.914	**0.922**	0.928	0.912
Yeast	0.514	0.825	0.821	**0.842**
Vehicle	0.508	0.619	0.638	**0.650**
Colon	0.532	0.510	0.552	**0.640**
Adeno	0.542	0.414	0.597	**0.620**
t-test	1×10^{-3}	3.1×10^{-3}	2×10^{-3}	Base

Table 3.4: CSPCA Decision Trees classification performance on imbalanced datasets compared to the algorithms. The best AUC are highlighted in bold.

Dataset	PCA	ADASYN- PCA	RU-PCA	CSPCA
Diabetes	0.624	0.645	0.616	**0.689**
Spambase	0.811	0.802	0.819	**0.836**
Breast	0.515	0.561	0.542	**0.591**
Ecoli	0.563	0.775	0.849	**0.852**
Abalone	0.503	0.522	0.529	**0.615**
Survival	0.514	0.474	0.498	**0.600**
Blood	0.577	0.575	0.606	**0.641**
Glass	0.821	0.867	0.882	**0.891**
Yeast	0.543	0.683	0.808	**0.812**
Vehicle	0.555	0.565	0.595	**0.625**
Colon	0.587	0.392	0.590	**0.630**
Adeno	0.578	0.518	0.538	**0.565**
t-test	4.2×10^{-3}	7.5×10^{-4}	1.4×10^{-3}	Base

Table 3.5: CSPCA 5-NN classification performance on imbalanced datasets compared to the algorithms. The best AUC are highlighted in bold.

Dataset	PCA	ADASYN- PCA	RU-PCA	CSPCA
Diabetes	0.632	0.633	0.536	**0.671**
Spambase	0.855	0.828	0.848	**0.862**
Breast	0.550	0.590	0.495	**0.620**
Ecoli	0.580	0.823	0.844	**0.856**
Abalone	0.500	0.573	0.621	**0.641**
Survival	0.509	0.481	0.458	**0.570**
Blood	0.609	0.626	0.623	**0.681**
Glass	0.930	0.850	0.925	**0.941**
Yeast	0.535	0.721	0.834	**0.880**
Vehicle	0.546	0.581	0.613	**0.600**
Colon	0.600	0.472	0.615	**0.631**
Adeno	0.476	0.589	0.503	**0.591**
t-test	6.7×10^{-3}	1.1×10^{-3}	4.2×10^{-3}	Base

Table 3.6: CSNMF SVM classification performance on imbalanced datasets compared to the other algorithms. The best AUC are highlighted in bold.

Dataset	NMF	ADASYN- NMF	RU-NMF	CSNMF
Diabetes	0.695	0.699	0.691	**0.721**
Spambase	0.627	0.665	0.636	**0.825**
Breast	0.573	0.484	0.516	**0.597**
Ecoli	0.868	0.855	0.840	**0.869**
Abalone	0.674	0.577	0.670	**0.693**
Survival	0.577	0.523	0.533	**0.621**
Blood	0.679	0.566	0.559	**0.756**
Glass	0.911	0.832	0.790	**0.925**
Yeast	0.844	0.833	0.817	**0.890**
Vehicle	0.645	0.637	0.600	**0.740**
Colon	0.752	0.772	0.752	**0.798**
Adeno	0.537	0.620	0.613	**0.723**
t-test	5.5×10^{-3}	1.1×10^{-4}	1.9×10^{-4}	Base

Table 3.7: CSNMF Naive Bayes classification performance on imbalanced datasets compared to the other algorithms. The best AUC are highlighted in bold.

Dataset	NMF	ADASYN- NMF	RU-NMF	CSNMF
Diabetes	0.670	0.693	0.686	**0.743**
Spambase	0.580	0.583	0.582	**0.622**
Breast	0.547	0.484	0.516	**0.583**
Ecoli	0.850	0.853	0.833	**0.882**
Abalone	0.500	0.669	0.616	**0.712**
Survival	0.604	0.489	0.592	**0.621**
Blood	0.553	0.561	0.544	**0.680**
Glass	0.536	0.760	0.808	**0.870**
Yeast	0.548	0.727	0.726	**0.781**
Vehicle	0.621	0.647	0.643	**0.660**
Colon	0.687	0.780	0.672	**0.791**
Adeno	0.500	0.475	0.471	**0.660**
t-test	1.8×10^{-3}	6.5×10^{-4}	2.5×10^{-4}	Base

Table 3.8: CSNMF Decision Trees classification performance on imbalanced datasets compared to the other algorithms. The best AUC are highlighted in bold.

Dataset	NMF	ADASYN- NMF	RU-NMF	CSNMF
Diabetes	0.612	0.613	0.630	**0.750**
Spambase	0.709	0.711	0.699	**0.770**
Breast	0.526	0.507	0.523	**0.630**
Ecoli	0.684	0.748	0.882	**0.890**
Abalone	0.497	0.543	0.623	**0.651**
Survival	0.566	0.558	0.550	**0.680**
Blood	0.589	0.624	0.611	**0.710**
Glass	0.807	0.760	0.770	**0.830**
Yeast	0.593	0.753	0.734	**0.771**
Vehicle	0.617	0.578	0.617	**0.651**
Colon	0.692	0.760	0.727	**0.751**
Adeno	0.520	0.471	0.582	**0.620**
t-test	3.8×10^{-5}	6.4×10^{-5}	2.6×10^{-4}	Base

Table 3.9: CSNMF 5-NN classification performance on imbalanced datasets compared to the other algorithms. The best AUC are highlighted in bold.

Dataset	NMF	ADASYN- NMF	RU-NMF	CSNMF
Diabetes	0.631	0.643	0.658	**0.700**
Spambase	0.722	0.724	0.723	**0.800**
Breast	0.520	0.500	0.519	**0.591**
Ecoli	0.733	0.831	0.850	**0.880**
Abalone	0.500	0.616	0.682	**0.691**
Survival	0.530	0.563	0.535	**0.591**
Blood	0.598	0.628	0.615	**0.674**
Glass	0.788	0.800	0.782	**0.791**
Yeast	0.497	0.789	0.795	**0.821**
Vehicle	0.591	0.581	0.647	**0.691**
Colon	0.777	0.727	0.672	**0.791**
Adeno	0.460	0.570	0.586	**0.630**
t-test	1.3×10^{-3}	5.9×10^{-5}	1.9×10^{-4}	Base

Table 3.10: CSPCA and CSNMF on logistic regression classifier compared to ML-CST. The best AUC are highlighted in bold.

Logistic regression	PCA ML-CST	CSPCA	NMF ML-CST	CSNMF
Diabetes	0.604	**0.640**	0.625	**0.680**
Spambase	0.729	**0.810**	0.716	**0.720**
Breast	0.523	**0.656**	0.500	**0.580**
Ecoli	0.541	**0.728**	0.500	**0.610**
Abalone	0.499	**0.660**	0.500	**0.730**
Survival	0.543	**0.610**	0.500	**0.690**
Blood	0.510	**0.610**	0.500	**0.581**
Glass	**0.915**	0.660	0.500	**0.590**
Yeast	0.496	**0.574**	0.500	**0.570**
Vehicle	0.511	**0.634**	0.630	**0.691**
Colon	0.590	**0.664**	0.602	**0.642**
Adeno	0.510	**0.610**	0.500	**0.650**
t-test	4.3×10^{-2}	Base	2.9×10^{-4}	Base

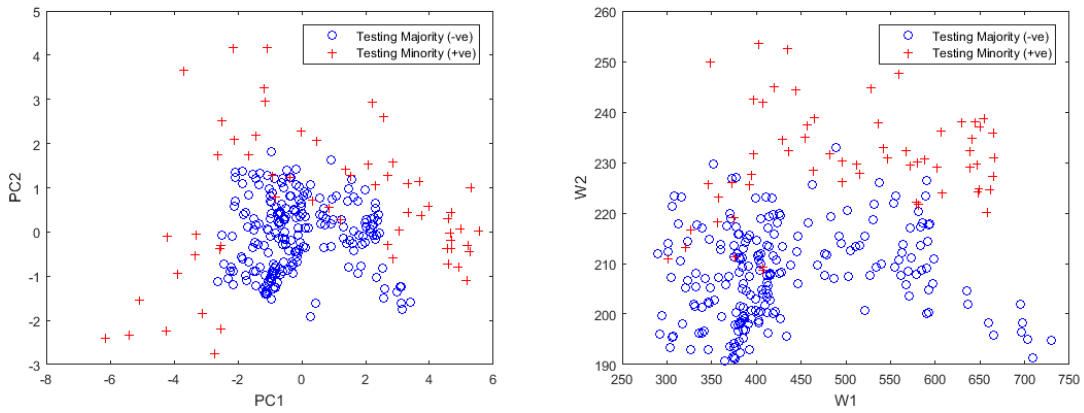
Table 3.11: CSPCA and CSNMF CART classification performance compared to CCPDT. The best AUC are highlighted in bold.

Decision Trees	PCA-CCPDT	CSPCA-DT	NMF-CCPDT	CSNMF-DT
Diabetes	0.676	**0.689**	0.704	**0.750**
Spambase	0.918	**0.836**	0.756	**0.770**
Breast	0.487	**0.591**	0.487	**0.630**
Ecoli	0.498	**0.852**	0.498	**0.890**
Abalone	0.481	**0.615**	0.481	**0.651**
Survival	0.492	**0.600**	0.492	**0.680**
Blood	0.721	**0.641**	0.552	**0.710**
Glass	0.854	**0.891**	0.767	**0.830**
Yeast	0.491	**0.812**	0.634	**0.771**
Vehicle	**0.652**	0.625	**0.671**	0.651
Colon	0.472	**0.630**	0.528	**0.751**
Adeno	0.433	**0.565**	0.443	**0.620**
t-test	3.3×10^{-2}	Base	9.5×10^{-4}	Base

Revisiting the Motivating Example

This section visualizes the improvements of classification after applying the two proposed methods, CSPCA and CSNMF, on the same dataset that was used in section 3.1.1. Figure 3.3, showing the projected test samples using CSPCA and CSNMF respectively, demonstrates that classes can be better distinguished than in Figure 3.1, showing the projected test samples using traditional PCA and NMF respectively.

Figure 3.3 shows how the classes can be better distinguished compared to Figure 3.1.



(a) Testing data after projecting the instances using CSPCA

(b) Testing data after projecting the instances using CSNMF factors

Figure 3.3: Applying CSPCA and CSNMF on an imbalanced vehicle dataset improves the classification performance.

3.2 Handling Class Imbalance on Supervised Non-Negative Matrix Factorization

Supervised feature extraction methods are an extension of standard unsupervised feature extraction methods which exploit the existence of label information to use the extracted features in classification. Several approaches have been proposed including supervised PCA (Chen et al. 2008) and supervised NMF (Huh et al. 2013) as reviewed in section 2.1.3.

This section presents a Balanced Supervised NMF (BSNMF) method which aims to exploit the class labels to lessen the inherent bias of the extracted components to the majority class. Then, the resulting BSNMF feature matrix is used to achieve a feature ranking. Based on literature review in Section 2.1.3, the researcher is unaware of any previous studies that address the imbalanced class problem using supervised feature extraction approaches. The organization of this section incorporates the definition of the method, implementation and experiments related to the BSNMF method.

3.2.1 Definition of the Method

Non-negative matrix factorization is an unsupervised matrix factorization method which is designed for dimensionality reduction and reconstruction, but not for classification. Several approaches have been proposed as shown in chapter 2 to extend NMF by constructing predicted features which can be used for classification. These approaches utilize the existence of class labels to improve the classification capability of the extracted features. However, the current supervised NMF methods in the literature fail to handle the class imbalance problem, because the extracted features

are biased to predict the majority class samples that lead to poor performance on classification. The following section introduces a Balanced Supervised NMF method to handle the class imbalance issue in supervised NMF methods.

BSNMF using Coupled Matrix Factorization

Coupled matrix factorization aims to capture the sparse factors of jointly analysed multiple matrices with one mode in common. This approach has previously been used in several data fusion studies (Yokoya et al. 2012). Without any loss of generality, assume matrices $Y \in \mathbb{R}^{N\times M}$ and $Z \in \mathbb{R}^{N\times E}$ have the first mode in common, which means in matrices Y and Z, the number of rows is the same which equal to N. The objective of coupled matrix factorization is to find three matrices $A \in \mathbb{R}^{N\times R}$, $B \in \mathbb{R}^{M\times R}$ and $C \in \mathbb{R}^{E\times R}$ by minimizing the cost function as follows

$$F\left(A,B,C\right) = \left\|Y - AB^T\right\|^2 + \left\|Z - AC^T\right\|^2$$

where matrix A is the common factorized matrix between Y and Z, and $\|.\|$ denotes the Frobenius norm for matrices. N is the number of samples, M and E represent the dimensions of matrices Y and Z respectively, and R is the number approximate rank.

This section uses the idea of coupled matrix factorization to handle the imbalance problem in supervised non-negative matrix factorization. The proposed approach aims to enhance the classification capability of NMF on imbalanced datasets. In the binary case, the existence of the class labels is exploited to extract the shared mode (basis matrix U) from the positive and negative matrices, X^+ and X^- respectively by using the idea of coupled matrix factorization.

Given a data matrix $X = [x_{ij}] \in \mathbb{R}^{M\times N}$, the original matrix X is decomposed into

two matrices as follows:

$$\begin{cases} X_i^-, & \text{if } y_i = -1, \text{(Negative class)} \\ X_i^+, & \text{if } y_i = +1, \text{(Positive class)} \end{cases}$$

where X^- corresponds to the data points classified as a negative class and X^+ corresponds to the positive class points. The proposed BSNMF method aims to find the shared feature matrix U and the other two matrices V^+ and V^- by minimizing the objective function

$$\begin{aligned} \mathbf{O}\left(U, V^-, V^+\right) = \left\|X^- - U(V^-)^T\right\|^2 + \left\|X^+ - U(V^+)^T\right\|^2 \\ s.t.\ U, V^-, V^+ \geq 0 \end{aligned} \tag{3.2.1}$$

However, in the imbalanced class problem, the magnitude of the error of the negative samples (the first term of Eq. 3.2.1) dominates the magnitude error of positive samples. This is because the number of negative samples is bigger than the positive ones. Therefore, the objective function is discounted in Eq. 3.2.1 with the class ratio of the samples. The objective function is changed to

$$\mathbf{O}\left(U, V^-, V^+\right) = \sigma^- \left\|X^- - U(V^-)^T\right\|^2 + \sigma^+ \left\|X^+ - U(V^+)^T\right\|^2 \tag{3.2.2}$$

where $\sigma^- = \frac{N}{N^-}$, $\sigma^+ = \frac{N}{N^+}$, N is total number of samples, and N^- and N^+ denote the total number of negative and positive samples respectively.

The objective function $\mathbf{O}$ in Eq. 3.2.2 is convex in U, V^- and V^+ separately. Therefore, three iterative procedures are generated which are able to find the local minima.

By expanding Eq. 3.2.2, the objective function is written as:

$$\begin{aligned} \mathbf{O}\left[U, V^-, V^+\right] = \sigma^- \left[X^-(X^-)^T - 2X^-V^-U + U(V^-)^TV^-U^T\right] \\ +\sigma^+ \left[X^+(X^+)^T - 2X^+V^+U + U(V^+)^TV^+U^T\right] \end{aligned} \tag{3.2.3}$$

The three factor matrices U, V^-, and V^+ are randomly initialized to non-negative values between 0 and 1, then they are refined by means of gradient descent approaches. So, Lagrange multipliers ψ, ϕ and ε for constraints U, V^-, and V^+ are proposed. The Lagrange form is

$$\begin{aligned}\mathbf{L} = \sigma^- \left[X^-(X^-)^T - 2X^-V^-U + U(V^-)^TV^-U^T\right] \\ +\sigma^+ \left[X^+(X^+)^T - 2X^+V^+U + U(V^+)^TV^+U^T\right] \\ +\psi U + \phi V^- + \varepsilon V^+\end{aligned} \tag{3.2.4}$$

The partial derivatives of $\mathbf{L}$ with respect to U, V^-, and V^+ are

$$\begin{aligned}\frac{\partial \mathbf{L}}{\partial U} = & -2\sigma^-X^-(V^-)^T + 2\sigma^-U(V^-)^TV^- \\ & -2\sigma^+X^+(V^+)^T + 2\sigma^+U(V^+)^TV^+ + \psi\end{aligned} \tag{3.2.5}$$

$$\frac{\partial \mathbf{L}}{\partial V^-} = -2\sigma^-X^-U^T + 2\sigma^-V^-UU^T + \phi \tag{3.2.6}$$

$$\frac{\partial \mathbf{L}}{\partial V^+} = -2\sigma^+X^+U^T + 2\sigma^+V^+UU^T + \varepsilon \tag{3.2.7}$$

The following equations are formulated based on Karush-Kuhn-Tucker conditions where $\psi_{ik}u_{ik} = 0$, $\phi_{jk}vm_{jk} = 0$ and $\varepsilon_{mk}vp_{mk} = 0$,

$$\begin{aligned}&-\left[\sigma^-X^-(V^-)^T\right]_{ik}u_{ik} + \left[\sigma^-U(V^-)^TV^-\right]_{ik}u_{ik} \\ &-\left[\sigma^+X^+(V^+)^T\right]_{ik}u_{ik} + \left[\sigma^+U(V^+)^TV^+\right]_{ik}u_{ik} = 0\end{aligned} \tag{3.2.8}$$

$$-\left[\sigma^-X^-U^T\right]_{jk}vm_{jk} + \left[\sigma^-V^-UU^T\right]_{jk}vm_{jk} = 0 \tag{3.2.9}$$

$$-\left[\sigma^+X^+U^T\right]_{mk}vp_{mk} + \left[\sigma^+V^+UU^T\right]_{mk}vp_{mk} = 0 \tag{3.2.10}$$

These equations lead to three updating rules denoted as follows

$$u_{ik} \leftarrow u_{ik}\frac{\left[\sigma^{-}X^{-}(V^{-})^{T}\right]_{ik}+\left[\sigma^{+}X^{+}(V^{+})^{T}\right]_{ik}}{\left[\sigma^{-}UV^{-}(V^{-})^{T}\right]_{ik}+\left[\sigma^{+}UV^{+}(V^{+})^{T}\right]_{ik}} \tag{3.2.11}$$

$$vm_{jk} \leftarrow vm_{jk}\frac{\left[X^{-}U^{T}\right]_{jk}}{\left[V^{-}UU^{T}\right]_{jk}} \tag{3.2.12}$$

$$vp_{mk} \leftarrow vp_{mk}\frac{\left[X^{+}U^{T}\right]_{mk}}{\left[V^{+}UU^{T}\right]_{mk}} \tag{3.2.13}$$

The convergence analysis and proof that objective function **O** is non-increasing under the updating rules in Eq. 3.2.11, 3.2.12 and 3.2.13 are described in Appendix A of this thesis.

3.2.2 Experiments and Datasets

Experiments are conducted on seven high-dimensional imbalanced microarray datasets. The characteristics of these datasets are summarised in Table 3.12. The main dataset in this case study is a childhood leukaemia dataset collected using the U133A Affymetrix platform from the Children's Hospital at Westmead (TB-CHW). It is important to note that the main purpose of these experiments is to show the strength of the proposed BSNMF method to identify non-biased informative genes which are most closely correlated with patient relapse.

Selecting the Important Genes using the Proposed BSNMF

Non-negative matrix factorization has been applied recently to computational biology for problems including molecular pattern discovery, disease prediction, and others (Ottoboni et al. 2012) (Li & Ngom 2013). It has been successfully used for gene ranking analysis in microarray gene expression data (Devarajan 2008). However,

Table 3.12: Datasets, column #IR is the imbalance ratio (Neg/Pos), #Attr is the number of features, and #Inst is the number of instances.

Dataset	#Attr	#Inst	#IR	Source
Childhood leukaemia	22277	60	2.85	TB-CHW
Colon cancer	2000	60	1.81	(Alon et al. 1999)
Lung cancer	12533	181	4.8	(Gordon et al. 2002)
Global Cancer Map	16064	281	2.11	(Reich et al. 2006)
Lymphoma Outcome	7130	78	3.05	(Reich et al. 2006)
ALL/AML	7129	73	2.92	(Golub et al. 1999)
Breast cancer	4869	77	1.33	(Van De Vijver et al. 2002)
Diffuse large b-cell lymphomas (DLBCL)	5470	77	3.05	(Reich et al. 2006)

unsupervised NMF has limits in identifying relevant genes which can be used for classification due to the fact that sample label information is not used. Furthermore, as shown in chapter 2, supervised NMF methods including supervised NMF (Huh et al. 2013), and DNMF (Jia et al. 2015), are not able to handle the class imbalance problem in the generated components.

This section uses the common shared basis matrix U which is generated from the proposed BSNMF method to achieve a gene ranking. The samples in the coefficient matrices V^- and V^+ are more predictive than those generated by imbalanced NMF methods. The proposed method provides a more efficient and flexible approach to gene ranking, and is comparable in terms of computational time yo supervised and unsupervised NMF methods.

The proposed BSNMF feature ranking is described as following: given a gene expression matrix of M genes and N samples which is decomposed using BSNMF, a sparse shared basis matrix U is obtained which carries the original gene contributions in its entries. Therefore, the differentially expressed features can be recognized using the basis matrix U, which can be expressed as

$$U = \begin{bmatrix} u_{11} & u_{12} & \dots & u_{1K} \\ \vdots & \vdots & \vdots & \vdots \\ u_{m1} & u_{m2} & \dots & u_{MK} \end{bmatrix}$$

The columns of the shared basis matrix U represent the contribution of each original gene in the factors V^- and V^+. The values of the basis matrix U are called gene values which define the gene ranks. To compute the rank of the original genes, the shared basis matrix U is summed by rows as

$$Score_j = \left[\sum_{k=1}^{K} u_{1k}, \dots\dots, \sum_{k=1}^{K} u_{mk} \right]^T \tag{3.2.14}$$

where K is the desired rank, $m = 1, \dots, M$ is the number of features in U. As $Score_j \to 0$, feature j is considered less important. Secondly, the score list obtained in Eq. 3.2.14 is sorted in descending order. Finally, without loss of generality, the most differentially expressed genes are picked out from the sorted list by imposing a threshold on entries of *Score* list.

In this section, to estimate the significance of the gene scores in the proposed BSNMF algorithm, a permutation test is defined based on a test proposed by Jia et al. (2015). A null hypothesis is set to check if the total added columns of matrix U are non-discriminative to the sample classification. The elements of U are randomly shuffled, and the permutated $Score^{ob}$ is computed for every iteration b, where $B = 1000$ times. The P-value (p) for an observed $Score^{ob}$ vector is

$$p = \frac{1}{MB} \sum_{b=1}^{B} \sum_{j=1}^{M} I\left(\left| Score^{ob} \right| < \left| Score_j^b \right| \right) \tag{3.2.15}$$

where M is the number of features, $Score_j^b$ is the permutated score for each feature, and $I(.)$ is an indicator function, returning 1 if true and 0 otherwise.

Figure 3.4 shows the stages of the BSNMF algorithm. In the first stage, the imbalance class problem is addressed on the training set by using the supervised BSNMF method. Then, it ranks the features in matrix U, and finally the classification model is built by using the selected features from the training data.

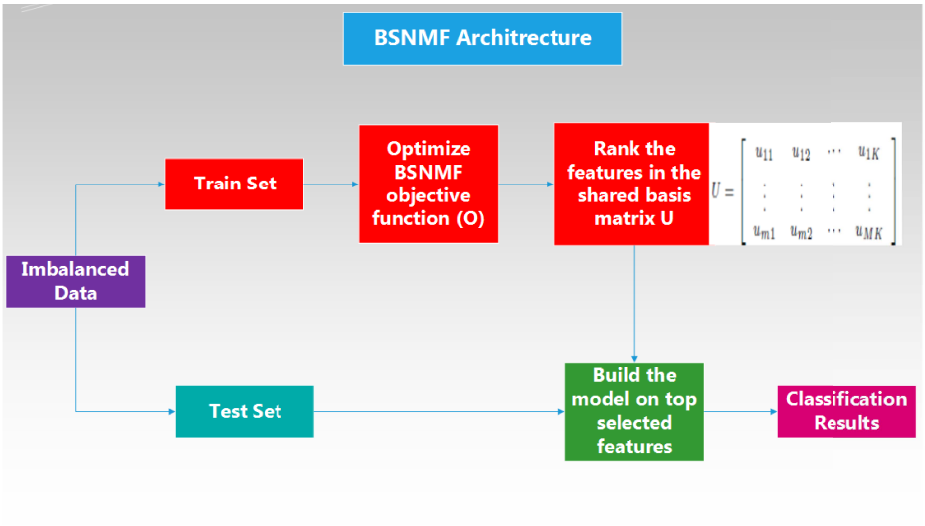


Figure 3.4: BSNMF feature ranking flowchart

3.2.3 Results

The performance of the proposed BSNMF method in this thesis is compared to standard NMF, and supervised DNMF (Jia et al. 2015). Without loss of generality, linear SVM is used with a soft margin loss function as an instance-based learning approach and AUC as an evaluation metric. Estimates of AUC are averaged over leave-one-out cross-validation. Moreover, the dataset entries are normalized between [0,1]. The results in Table 3.13 show that the performance of the SVM classifier is substantially improved by the BSNMF algorithm. Moreover, the proposed BSNMF with base classifier outperforms the classical NMF and the state-of-the-art method with different

approximate rank k.

A t-test is conducted between BSNMF (base) against the compared methods, under the null hypothesis that the AUC of the used methods is not significantly different. The p-values of Table 3.13 reject the null hypothesis, as most values of BSNMF (base) are lower than 0.05. The results in Table 3.13 show that the performance of the SVM classifier is substantially improved by using the BSNMF method. It outperforms the state-of-the-art methods. Furthermore, the proposed method selects the important genes in the childhood leukaemia dataset which are able to predict the relapse class. Also, the results indicate that its performance is exacerbated in the presence of imbalanced and high-dimensional data. Therefore, BSNMF addresses the class imbalance issue in supervised NMF without the need to change the data distribution or modify the existing classifiers. Furthermore, it was observed that gene ranking in DNMF (Jia et al. 2015) is designed to work only on a rank equal to 2.

3.3 Addressing Class Imbalance using Artificial Bee Colony Algorithm

This section covers the method and implementation of the ABC-Sampling algorithm in detail followed by its evaluation with experiments.

The ABC-Sampling algorithm is based on an optimization method called Artificial Bee Colony (ABC) (Karaboga & Basturk 2007) which aims to ameliorate the effects of the imbalanced learning problem on different datasets. The standard artificial bee colony algorithm is reviewed in section 2.1.4. ABC-Sampling uses a forward search strategy in order to find the most informative samples which contribute to

Table 3.13: AUC evaluation results on imbalanced datasets. The best AUC are highlighted in bold and (✓) sign to show the base method is statistically significant.

Dataset	Rank (K)	NMF	DNMF	BSNMF
Childhood leukaemia	2	0.75	0.50	**0.85**
	4	0.75	-	**0.90**
	6	0.55	-	**0.80**
	8	0.75	-	**0.75**
	10	0.65	-	**0.90**
Colon cancer	2	0.75	0.90	**0.85**
	4	0.85	-	**0.85**
	6	0.81	-	**0.90**
	8	0.85	-	**0.95**
	10	0.85	-	**0.90**
Lung cancer	2	0.88	0.50	**0.89**
	4	0.79	-	**0.99**
	6	0.91	-	**0.98**
	8	0.98	-	**0.93**
	10	0.95	-	**0.97**
Global Cancer Map	2	0.77	0.60	**0.78**
	4	0.73	-	**0.87**
	6	0.84	-	**0.79**
	8	0.81	-	**0.85**
	10	0.79	-	**0.86**
Lymphoma Outcome	2	0.84	0.50	**0.97**
	4	0.81	-	**0.84**
	6	0.91	-	**0.86**
	8	0.81	-	**0.86**
	10	0.78	-	**0.83**
ALL/AML	2	0.77	0.50	**0.80**
	4	0.78	-	**0.99**
	6	0.89	-	**0.92**
	8	0.72	-	**0.88**
	10	0.96	-	**0.97**
Breast cancer	2	0.51	0.61	**0.73**
	4	0.60	-	**0.63**
	6	0.63	-	**0.80**
	8	0.70	-	**0.70**
	10	0.70	-	**0.76**
DLBCL	2	0.84	0.75	**0.88**
	4	0.81	-	**0.87**
	6	0.94	-	**0.95**
	8	0.81	-	**0.90**
	10	0.81	-	**0.94**
t-test	2	0.01 ✓	0.003 ✓	Base
	4	0.009 ✓	-	Base
	6	0.01 ✓	-	Base
	8	0.07	-	Base
	10	0.02 ✓	-	Base

achieving high results in classifying data. The proposed algorithm can be applied for undersampling or oversampling the imbalanced dataset. However, this thesis only evaluates the undersampling strategy. Several strategies and methods have been proposed to solve the imbalanced class problem with accompanying debates on the merits of different strategies. Undersampling is often preferred because no extra information is added to the data, but there is no consensus in the literature on an optimal strategy (Batista et al. 2004). Particularly, ABC-Sampling finds the optimal instances or samples in the majority set of the dataset in order to retain them during the undersampling process.

3.3.1 Sampling Strategy

The ABC-Sampling method is divided into three main phases as shown in Figure 3.5

Dataset Partition

In the first phase, the given imbalanced dataset is divided into a training and testing set. Without loss of generality, the training set is 2/3 of the data and the testing set is 1/3. The training set is used by the second component to train the data. However the testing data is untouched and it is retained for the evaluation phase.

Splitting the Training Set

This phase uses two strategies for cross-validation to optimise the proposed ABC-Sampling algorithm namely leave-one-out and k-fold. Leave-one-out is used in biomedical data because it contains a small number of samples. The 3-fold cross validation strategy is used on non-biomedical data.

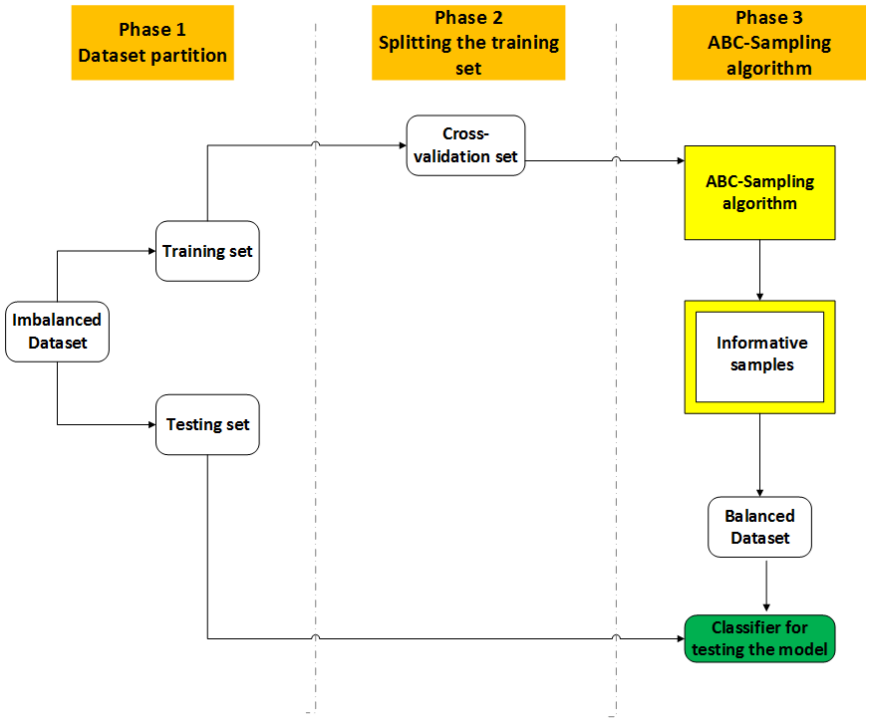


Figure 3.5: ABC-Sampling phases

ABC-Sampling System

In the last phase, ABC-Sampling modifies the standard ABC algorithm to select the best majority samples during the undersampling process to balance the imbalanced data. ABC-Sampling uses different cross-validation sets from the previous phase and optimises the fitness function. The new fitness function of the proposed ABC-Sampling is the classification accuracy of the selected majority samples with the minority samples. In particular, ABC-Sampling is composed of three stages. Firstly, the possible solutions in the proposed optimization algorithm are represented by a bit vector which is known as a food source. Each food source randomly initiates a bit vector of size D, where D is the size of the majority samples in the training set. The value of the positions in the vector represents the presence or absence of the

majority sample in the subset. For example, in Figure 3.6, the first and D^{th} samples are part of the subset. Several types of bees in this algorithm represent the food sources (possible solutions) in different occasions. For example, in this problem, the scout bee is the set of selected majority samples which is not able to improve the fitness function in several iterations. Secondly, the ABC-Sampling algorithm selects the samples which are present in every generated subset and combines them with the minority samples. Then, it evaluates the fitness value which is the classification metric AUC and F-measure for the current subset. The output of the ABC-Sampling algorithm is a list of the most informative majority samples sorted by the number of times selected. Finally, the algorithm applies the undersampling strategy to produce a balanced dataset by selecting a number of the top majority samples equal to the number of minority ones. Furthermore, in the case of biomedical data, the datasets suffer from a low number of samples. Therefore, the ABC-Sampling algorithm does not lose the excluded majority points but it adds them to the testing set to evaluate the algorithm.

1	0	0	.....	0	1
Sample1	Sample2	Sample3		Sample D-1	Sample D

Figure 3.6: Samples bit vector

3.3.2 ABC-Sampling Algorithm

This sections presents in detail the optimization process of ABC-Sampling algorithm. It consists of the following steps:

1. *Create M food sources and initialize parameters*: The value M is half of the samples in the training set as recommended by the standard ABC algorithm, but it

can be any other value. Randomly initialize the D bit values for each food source. Initialize the maximum number of iterations of the algorithm and the limit counter for abandoning the food sources.

2. Compute fitness: Submit the food source samples to a classifier and evaluate accuracy as the fitness value. In this paper, SVM is used to evaluate the selected samples using AUC and F-measure metrics, but other approaches are reasonable. For very high-dimensional datasets, a feature selection using the Balanced Iterative Random Forest method (Anaissi et al. 2013) is applied to reduce the number of features.

3. Find neighbouring food and compute their fitness: Determine the neighbouring food sources using Eq. 2.1.5. The food source values are represented by bits and the perturbation frequency φ is a real value. The real-valued neighbour values are converted to bits through a sigmoid function or any other approach. The fitness function of the neighbour samples is computed as in step 2. If the fitness is higher than the current food source, replace the neighbour samples with the food source set. Otherwise, the limit variable for the food source is incremented. If this reaches the maximum, the algorithm eliminates the current food source and increments the abandoned variable to create a new food source by the scout bees.

4. Share results and record the best: The employed bees share the resulting fitness values with onlookers, and the onlookers select the top average of AUC and F-measure fitness values from the food sources and re-execute step 3. The maximum fitness value of the onlookers is recorded as the global best value in the best majority samples set.

5. Launch scout bees: The scout bees set is the number of the exhausted food sources that need to be abandoned. The food source has a parameter counter

$LIMIT_z$, which is updated during the search. If the value of this parameter reaches the maximum limit ($LIMIT_{MAX}$), the food source will be abandoned and replaced by a new food source produced by a new scout bee. Scout bees are created by Eq. 2.1.4 and are added to the original set of food sources.

6. Test for termination: Check if the maximum number of iterations is reached and if so return the best informative samples set. Finally, the algorithm applies the undersampling strategy to produce a balanced dataset by selecting a matching number of the top majority samples to the number of minority ones.

Algorithm 1 shows the detailed steps for selecting the optimal instances in ABC-Sampling.

3.3.3 Experiments and Datasets

Datasets

The proposed ABC-Sampling algorithm has been evaluated on ten imbalanced datasets, each with two classes (see Table 3.14). Apart from the childhood leukaemia dataset, these datasets are obtained from the UCI Machine Learning Repository (Lichman 2013). As the first three datasets, childhood leukaemia, colon and breast, are small and high-dimensional, the leave-one-out cross-validation is used. Three-fold cross validation is used for the others to reduce computation time and the variance of estimators.

The main specifications of these datasets are summarised in Table 3.14. The imbalance ratio is defined as the number of majority samples which are the samples belonging to the dominant class over the number of minority samples. The algorithms are evaluated on biomedical and non-biomedical datasets. The biomedical data comprise

Input: training set
Output: *Rankset* of majority samples
Initialization `// Step 1 in section 3.3.2`
$FoodSourceSize$=$TrainingSamples/2$, FS_Set, $MaxIterations$, $LIMIT_{MAX}$,$LIMIT$, *Abandoned*
FS_Acc: is a vector to store the fitness value of the possible solution set
minoritySet: is the minority samples in training set
for $i = 1 : FoodSourceSize$ **do**
  Initialize position of FS_Set_{ij} by Eq. 2.1.4 for $j \in [0, 1]$
end
`// Step 2 in section 3.3.2`
for ***each*** FS_i *in* FS_Set **do**
  Create an internal set $FS_i \bigcup minoritySet$;
  Train classifier and evaluate the fitness function;
  Save the fitness value in FS_Acc;
end
for $t = 1 : MaxIterations$ **do**
  `// Step 3 in section 3.3.2`
  for ***each*** FS_i *in* FS_Set **do**
    [FS_Set,*Abandoned*]=**ProcessNeighbours** (FS_i,FS_Set,FS_Acc_i,$LIMIT$,$LIMIT_{MAX}$)
  end
  $FS\_Set \bigcup minoritySet$ Train classifier and evaluate the fitness function;
  Save the fitness value in FS_Acc;
  `// Step 4 in section 3.3.2`
  Memorize best global food source ;
  $Global = \max FS\_Acc$;
  $Best\_FS = FS\_Set[Global]$;
  $Rankset = Rankset \bigcup Best\_FS$;
  `// Step 5 in section 3.3.2`
  for $v = 1 : Abandoned$ **do**
    Initialize position of FS_Set_{ij} by Eq. 2.1.4 for $j \in [0, 1]$
  end
end
Return *Rankset* of majority samples;

Algorithm 1: ABC-Sampling algorithm

Function *ProcessNeighbours* (FS_i,$fsSet$,FS_Acc_i,$LIMIT$,$LIMIT_{MAX}$)
  Determine the *neighbour* by Eq. 2.1.5;
  if $neighbour \neq currentFS$ **then**
    $neighbour\_Acc$=Evaluate the fitness function of *neighbour*;
    if $FS\_Acc_i < neighbour\_Acc$ **then**
      Replace FS_i with *neighbour*;
    else
      Increment the $LIMIT$ of FS_i;
      if $LIMIT \geq LIMIT_{MAX}$ **then**
        Increment the *Abandoned*;
        $fsSet = fsSet - FS_i$;
      end
    end
  end
  Return $fsSet$;

Algorithm 2: Generate and evaluate neighbours

five datasets: childhood leukaemia, colon (Alon et al. 1999), breast (Van De Vijver et al. 2002), blood and survival (Lichman 2013). The childhood leukaemia dataset was generated from the U133A affymetrix microarray gene expression platform and was collected from The Children's Hospital at Westmead. It has 60 samples with a significant imbalance ratio of 1.85. Another affymetrix microarray gene expression dataset used in this study is the colon dataset (Alon et al. 1999) which contains 62 samples with an imbalance ratio of 1.82. The third biomedical dataset used in the experiments is the breast dataset which contains 77 samples and an imbalance ratio of 1.34. Blood was generated by the Blood Transfusion Service Center in Taiwan, and Survival was produced from the survey conducted on the survival of patients who had undergone surgery for breast cancer. Diabetes was generated from the study of diabetes in the Pima Indian population. The non-biomedical data comprises Spambase,

Australian credit approval, and ionosphere (Lichman 2013). Spambase is a large collection of spam and non-spam emails which is collected from postmaster and personal emails. Australian credit approval and ionosphere are not highly imbalanced, so they were artificially sampled at 1:5 (each positive sample versus 5 negative samples), 1:10 and 1:15, to produce more imbalance.

Table 3.14: Datasets characteristics to evaluate ABC-Sampling

Dataset	#Samples	#Attributes	Imbalance ratio
Childhood leukaemia	60	22277	1.85
Colon	62	2000	1.82
Breast	77	4869	1.34
Blood	748	5	3.16
Survival	306	3	2.77
Diabetes	768	8	1.86
Spambase	4601	57	1.537
Australian Credit Approval	414	14	4.97, 10, 14.73
Ionosphere	247	34	5, 9, 15

Evaluation and Parameter Selection

Feature selection is applied to high-dimensional biomedical datasets to select the most informative features using a Balanced Iterative Random Forest (BIRF) method (Anaissi et al. 2013). This technique robustly selects a small number of informative genes. Evaluation metrics are vital in measuring learning performance. Learning from imbalanced datasets requires the use of Area Under the ROC Curve (AUC), F-measure or similar as they do not assume each class is a similar size. The ABC-Sampling algorithm defines parameters based on parameter selection experiments, as shown in Table 3.17 and Figure 3.7. The proposed algorithm is compared to learning from imbalanced datasets, random undersampling and particle swarm optimization

(PSO) (Yang et al. 2009). The PSO technique searches for an optimal subset of the majority samples and combines them with the minority samples to build a balanced classification model. Random undersampling reduces the number of majority samples by selecting a random subset of the majority class equal to the number of minority samples and combines both in the training dataset (Yen & Lee 2009).

3.3.4 Results

Performance of ABC-Sampling

In this section, two experiments are described. ABC-Sampling is applied on the six biomedical datasets that are either high or low dimensional data. The results are measured on independent test sets to accurately report the generalization of the proposed algorithm. The obtained results are compared to three state-of-the-art methods: PSO (Yang et al. 2009), random undersampling (labeled "RU") and learning from the original imbalanced dataset ("baseline"). As shown in Table 3.15, the results demonstrate the superiority of the proposed ABC-Sampling method over the state-of-the-art undersampling techniques using two different measure metrics AUC and F-measure. Also, the results show the importance of undersampling methods in improving the classification performance over the evaluated datasets. They attain the highest values compared to classifying the original imbalanced datasets. The second experiment evaluates an Australian credit approval and an ionosphere datasets that are artificially imbalanced in different ratios as shown in Table 3.16. The experiment is also conducted over the large dataset Spambase. ABC-Sampling outperforms the other methods over all the tested levels of imbalance. ABC-Sampling is a strong alternative to existing methods for balancing datasets and leads to excellent learning

outcomes. Furthermore, it shows its scalability by achieving good results on large and highly imbalanced datasets.

Parameter Sensitivity

A study is performed to investigate the parameter sensitivity of the proposed ABC-Sampling algorithm. The ABC optimization algorithm has fewer control parameters compared to other heuristic search algorithms such as PSO and ant colony optimization (ACO). Apart from the maximum number of iterations, ABC has only one control parameter ($LIMIT_{MAX}$). The $LIMIT_{MAX}$ parameter preventing the algorithm from staying in the local optima. As shown in Table 3.17, the optimal value of parameter $LIMIT_{MAX}$ is 10, based on the AUC and F-measure of the testing set. Furthermore, it is observed that if the value of the $LIMIT_{MAX}$ parameter is less than 10, accuracy is decreased due to the early convergence into the local optimum.

Another parameter $MaxIterations$ is analysed. This specifies the maximum number of iterations of the algorithm. Seven datasets are evaluated using a different number of iterations to choose the optimal value based on two factors: the highest AUC and the smallest number of iterations. As shown in Figure 3.7, iteration 250 attains the highest accuracy compared to the other iteration values. The plots at iteration 500 have similar performance to iteration 250, but using more iterations until termination will increase the computation time of the algorithm.

Time Complexity

Heuristic search algorithms may not present better results compared to brute force search methods. However, the time complexity to evaluate all the subsets of the

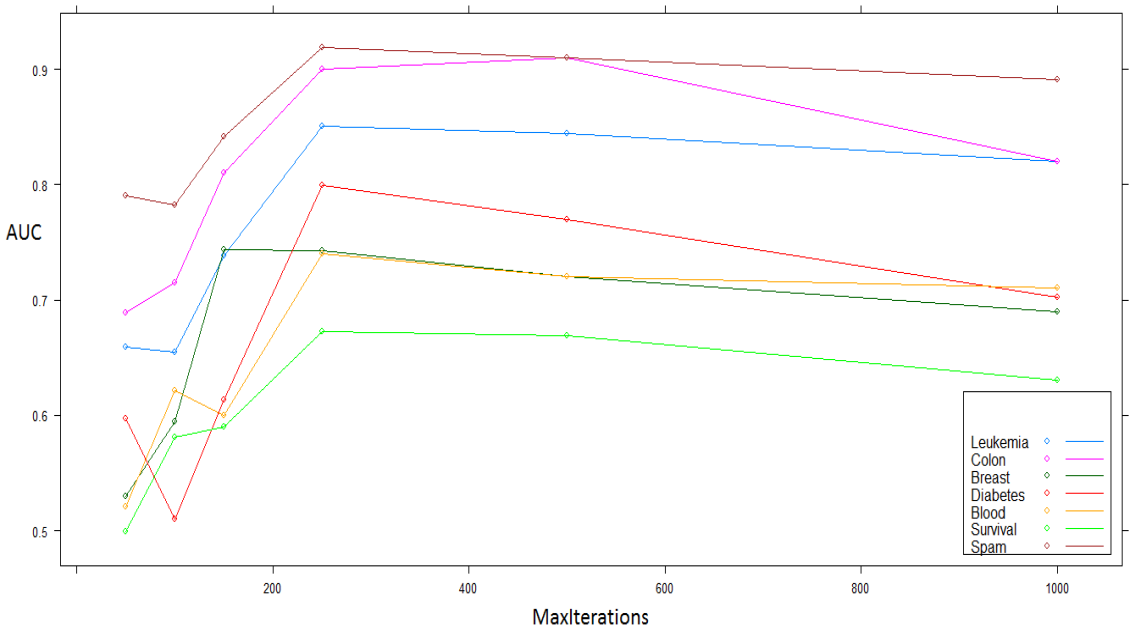


Figure 3.7: ABC-Sampling parameter selection ($MaxIterations$)

brute force search is O(2^n) where n is the number of data points, which is practically impossible. Therefore, the time complexity of the proposed algorithm is studied in terms of the number of fitness function evaluations.

In ABC-Sampling, there is more than one fitness function evaluation for each food source during an iteration. During the initialization and employed phases, the fitness function must be calculated for the whole population size N of food sources. By the same token, during the onlooker stage, it needs to evaluate N fitness functions for the top food sources. Therefore, the overall number of fitness function evaluations in the initialization stage is N and in the employed and onlooker stages, it is $2TN$ where T is the number of iterations. In the scout bees stage, the proposed algorithm selects the food sources that exceed $LIMIT_{MAX}$ trials (labeled as t) and abandons them and replaces them with new food source. Hence the individual needs $t/2$ iterations

to exceed the $LIMIT_{MAX}$ trials. As a result, the highest computational time of evaluating ABC-Sampling algorithm is described as follows

$$N\left(1+3T-\frac{t}{2}\right) \tag{3.3.1}$$

ABC-Sampling has a lower complexity than the standard ABC algorithm as it avoids the execution of the fitness function when the neighbour and current food sources are identical, as shown in Algorithm 1.

Table 3.15: Evaluation results of seven methods on imbalanced datasets. Values are on the test sets. The best AUC and F-measure are highlighted in bold.

Methods	Imbalanced Datasets						
	Childhood leukaemia	Colon	Breast	Diabetes	Blood	Survival	Spambase
ABC-Sampling							
AUC	**0.851**	**0.9**	**0.743**	**0.799**	**0.74**	**0.73**	**0.919**
F-measure	**0.755**	**0.813**	**0.698**	**0.743**	**0.74**	**0.681**	**0.918**
PSO							
AUC	0.812	0.891	0.732	0.791	0.721	0.691	0.892
F-measure	0.698	0.793	0.621	0.641	0.489	0.577	0.864
RU							
AUC	0.724	0.882	0.72	0.751	0.668	0.631	0.892
F-measure	0.611	0.774	0.613	0.612	0.471	0.481	0.886
Baseline							
AUC	0.792	0.81	0.583	0.621	0.632	0.592	0.77
F-measure	0.628	0.76	0.498	0.58	0.42	0.432	0.73

3.4 Contribution and Conclusion

This chapter addresses **Contributions 1, 2 and 3** of this thesis by developing a series of algorithms for handling class imbalance in single-labeled data. This chapter proposes novel algorithms to handle the imbalance problem at two levels: feature extraction and the wrapper method. The first contribution proposes a cost-sensitive

Table 3.16: Classification performance on imbalanced datasets. Values are on the test sets. The best AUC and F-measure are highlighted in bold

	Aus. Credit Approval				Ionosphere			
Ratio	ABC-Sampling	PSO	RU	Baseline	ABC-Sampling	PSO	RU	Baseline
1:5								
AUC	**0.87**	0.821	0.53	0.69	**0.84**	0.712	0.58	0.71
F-measure	**0.853**	0.781	0.44	0.58	**0.829**	0.691	0.44	0.66
1:10								
AUC	**0.86**	0.813	0.45	0.673	**0.71**	0.692	0.46	0.61
F-measure	**0.84**	0.798	0.39	0.591	**0.703**	0.61	0.42	0.587
1:15								
AUC	**0.763**	0.71	0.35	0.615	**0.73**	0.683	0.391	0.62
F-measure	**0.732**	0.681	0.332	0.592	**0.719**	0.642	0.378	0.612

learning strategy to address the imbalanced class problem that can be applied to well-known feature extraction techniques such as PCA and NMF. The proposed strategy embeds weighted label information in classical feature extraction methods namely PCA and NMF, in order to extract non-biased features. It improves the AUC of classification and reduces the overlapping between the classes. The obtained results show the high quality of the proposed methods on multiple popular classifiers and benchmark datasets compared to the state-of-the-art methods. They can deal with different levels of imbalance and different sized dataset.

The second contribution proposes a method called balanced supervised non-negative matrix factorization (BSNMF) based on the theory of coupled matrix factorization. This algorithm represents the second contribution, as listed in section 3.2. The aim of the proposed method is to extract a common balanced basis matrix between the positive and negative samples. The proposed method exploits the class labels to lessen the inherent bias of the extracted components towards the majority class. The BSNMF

Table 3.17: ABC-Sampling parameter selection. The best AUC and F-measure are highlighted in bold

Methods	$LIMIT_{MAX}$ value					
	5	10	15	20	30	50
Childhood leukaemia						
AUC	0.615	**0.851**	0.78	0.771	0.851	0.719
F-measure	0.595	**0.755**	0.71	0.677	0.745	0.628
Colon						
AUC	0.63	**0.9**	0.79	0.891	0.8	0.713
F-measure	0.593	**0.813**	0.72	0.691	0.71	0.66
Breast						
AUC	0.512	**0.743**	0.69	0.679	0.669	0.618
F-measure	0.5	0.698	**0.699**	0.631	0.61	0.539
Diabetes						
AUC	0.589	**0.799**	0.712	0.681	0.68	0.699
F-measure	0.511	**0.743**	0.623	0.663	0.661	0.632
Blood						
AUC	0.533	0.74	**0.744**	0.623	0.613	0.613
F-measure	0.413	0.74	**0.741**	0.5	0.51	0.592
Survival						
AUC	0.51	**0.73**	0.689	0.691	0.66	0.541
F-measure	0.413	**0.681**	0.531	0.533	0.543	0.534
Spambase						
AUC	0.82	**0.919**	0.9	0.91	0.889	0.89
F-measure	0.795	**0.918**	0.87	0.893	0.873	0.882

algorithm achieves feature selection in high-dimensional data. The experiments show that the proposed method outperforms the state-of-the-art methods in all datasets.

Lastly, the third contribution, proposes a method called ABC-Sampling based on the undersampling strategy. This algorithm represents the third contribution. The aim of the proposed method is to tackle highly imbalanced datasets. It returns a balanced dataset to a classifier which is composed of the best informative majority samples and minority samples. The results of the experiments showed that the proposed algorithm outperformed the state-of-the-art. Furthermore, the proposed

algorithm is independent of any type of dataset. In addition, the proposed algorithm is scalable and works with different imbalance ratios.

Based on the classification results of the three approaches above, ABC-Sampling outperforms the other two methods in Contributions 1, 2. For example, in the common datasets colon and breast, the AUC of ABC-Sampling is 0.9 and 0.743 respectively which is better than the other proposed approaches CSPCA, CSNMF and BSNMF.

Chapter 4

Dimensionality Reduction of Highly Correlated Features in Single-Labeled Data

This chapter proposes two approaches for dimensionality reduction in the presence of highly correlated features in single label datasets. The existing feature selection techniques are used to identify a subset of the most useful features, and may consider the rest as unimportant, redundant or noisy. In the presence of highly correlated features, many variable selection methods consider correlated features as redundant and need to be removed. This chapter proposes two novel algorithms for dimensionality reduction in datasets with highly correlated features without considering them as redundant or noisy features. It is organised into two sections. Section 4.1 proposes a supervised context-aware non-negative matrix factorization method using adjacency networks, and section 4.2 introduces a fuzzy ensemble feature selection method using co-expression networks. To validate the proposed algorithms, sub-sections describe the used datasets, detail the experiments, and compare the proposed algorithms with the state-of-the-art, before presenting the conclusion.

This chapter encapsulates **Contribution 4 and 5** of the thesis, and it is an

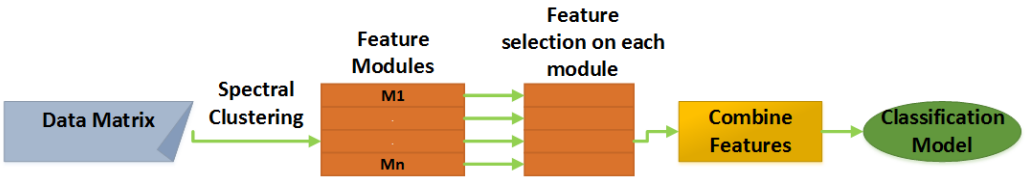
extended version of two publications (Braytee, Liu & Kennedy 2017) and (Anaissi et al. 2016).

4.1 Supervised Context-Aware NMF to Handle High-Dimensional Highly Correlated Data

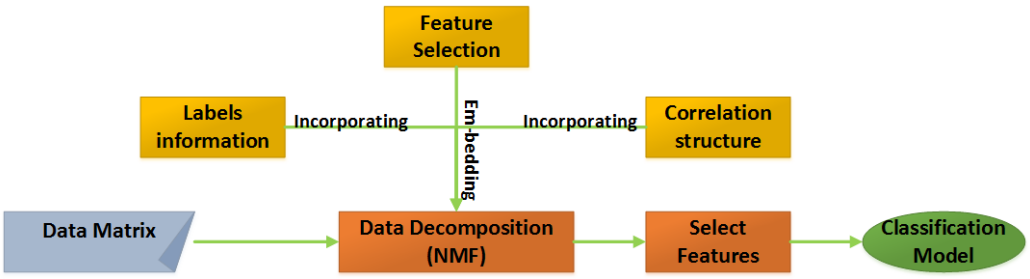
In this section, a supervised context-aware non-negative matrix factorization (SCANMF) is proposed for interpretation and classification in a variety of domains. The method incorporates the correlation structure using correlation networks which reflect the semantic components along with prior knowledge of the class labels and the original features of the data decomposition technique (NMF) as shown in Figure 4.1b. The method is a variant of the classic NMF. Non-negative matrix factorization is a parts-based representation which allows only additive combinations of a non-negative basis (Lee & Seung 2001). The primary property of NMF is the non-negativity constraint which facilitates an elegant interpretation in many applications. In this algorithm, the Balanced Supervised NMF method discussed in section 3.2 is integrated with SCANMF to lessen the inherent bias of the extracted basis vectors to the majority class when dealing with the class imbalance problem.

In light of this, correlation networks are proposed to build the adjacency (similarity) matrix to capture the semantic components from the correlation structure. Correlation networks are defined as undirected graphs consisting of nodes and edges. Nodes are the features, and an edge connects a pair of features if both are correlated.

The method proposed in this study is formulated as a constrained optimization problem integrating the correlation structure which reflects the semantic structure of the local discriminative information. Thus, decomposing the co-expression network in the optimization method using NMF will lead to groups of features that naturally co-function (Kuang et al. 2012). Also, for feature selection, $l_{2,1}$-norm minimization is incorporated into the basis matrix due to its robustness to outliers, and to its sparsity criteria over rows of the matrix. Feature selection is directly embedded into the supervised NMF which simultaneously takes into account the correlation structure, label information and imbalanced proportion of classes without using several separate methods as shown in Figure 4.1b. Consequently, an efficient iterative algorithm is proposed to solve the optimization problem. Extensive experiments are conducted on imbalanced highly correlated datasets to evaluate the proposed method.



(a) Existing supervised clustering approaches based feature selection



(b) Supervised Context-Aware NMF (SCANMF) based feature selection

Figure 4.1: Differences between the existing and the proposed feature selection-based approach to determine the meaningful features

4.1.1 Definition of the Method

Some notations used throughout this section are given in the following. Briefly, assume a data matrix $X = [x_{ij}] \in \mathbb{R}^{d \times N}$, where d and n represent the number of dimensions and samples of the matrix respectively. For an arbitrary matrix $T \in \mathbb{R}^{n \times d}$, the $l_{2,1}$-norm is defined as

$$\|T\|_{2,1} = \sum_{i=1}^{n} \sqrt{\sum_{j=1}^{d} T_{ij}^2} = 2\mathrm{Tr}\left[T^T D T\right]$$

where t_i is the ith row vector of T, D is a diagonal matrix with $D_{ii} = 1/2 \left\|t_i\right\|_2 + \epsilon$, $\left\|t_i\right\|_2$ is very close to 0 and ϵ is a small positive constant. $\|T\|_F$ is the Frobenius norm of T and Tr(T) is the trace operation of matrix T if T is square.

To select the discriminative features using supervised NMF, SCANMF is proposed to jointly exploit the prior knowledge of the labels with the correlation structure of attributes to capture the semantic components simultaneously. Thus, the proposed method uses the concept of coupled matrix factorization (a.k.a., collective matrix factorization) to extract the shared basis matrix from the two components in the method, namely the balanced supervised NMF and the correlation network. Furthermore, $l_{2,1}$-norm regularization is imposed on the shared basis matrix to be sparse in rows.

Correlation Networks

The interactions between nodes such as attributes can be represented in a network. Analysing the interactions in complex correlation networks may capture useful information about the network modules which are the clusters of interconnected nodes.

Several research areas explore these networks, such as social interaction networks, gene co-expression networks, protein-protein interactions, and others. In the case of gene co-expression networks, the network modules may be relevant to a trait or disease type (Horvath & Dong 2008). The co-expression networks are built based on node values. The network nodes correspond to features. The similarity s_{ij} between the features i and j is defined as the absolute value of the Pearson correlation coefficient between their values

$$s_{ij} = |corr(x_i, x_j)|$$

The similarity network is transformed to an adjacency matrix which represents the strength of the connections between nodes. Two thresholding procedures are proposed to construct the adjacency matrix: hard and soft thresholding. Hard thresholding leads to an unweighted adjacency network. It defines parameter $\varepsilon \in (0, 1)$, and constructs the adjacency matrix as follows

$$a_{ij} = \begin{cases} 1 & s_{ij} \geq \varepsilon \\ 0 & \text{otherwise} \end{cases}$$

The two features are correlated if the link strength is equal to 1. Another method of constructing a weighted adjacency matrix is by using a soft thresholding strategy. This strategy is generally better than the hard thresholding strategy due to the continuous nature of the adjacency values. Also, it does not lead to loss of information due to the use of a fixed threshold. The adjacency value between two features is defined as a power of the correlation coefficient s_{ij}

$$a_{ij} = s_{ij}^{\theta} \tag{4.1.1}$$

where $\theta \geq 1$.

SCANMF Formulation

The supervised context-aware non-negative matrix factorization method is an extension of original NMF based on coupled matrix factorization. The objective function of SCANMF takes into account three main issues: handling the problem of imbalanced classes, integrating the network structure and, for feature selection, and imposing an $l_{2,1}$-norm on the basis matrix. The adjacency matrix is adopted for the correlated features in the objective function to extract feature modules from the shared matrix $U \in \mathbb{R}^{d \times k}$, where k is the approximate rank. Finally, $l_{2,1}$-norm imposes sparseness on the shared matrix U. It is appropriate for controlling sparsity because it takes into consideration the module effects on shared matrix U. Thus, the objective function is defined as follows

$$\begin{aligned} \mathbf{O} = & \sigma^{-} \left\| X^{-} - U(V^{-})^{T} \right\|_{F}^{2} + \sigma^{+} \left\| X^{+} - U(V^{+})^{T} \right\|_{F}^{2} \\ & + \alpha \left\| A - UT^{T} \right\|_{F}^{2} + \beta \left\| U \right\|_{2,1} \\ & s.t.\ U, V^{-}, V^{+}, T \geq 0 \end{aligned} \tag{4.1.2}$$

where $A \in \mathbb{R}^{d \times d}$ is the adjacency matrix of the features in the original matrix X, and $\sigma^{-} = \frac{n}{n^{-}}$, $\sigma^{+} = \frac{n}{n^{+}}$, n is total number of samples, and n^{-} and n^{+} denote the total number of negative and positive samples respectively. $X^{-} \in \mathbb{R}^{d \times n^{-}}$ corresponds to the data points classified as negative class and $X^{+} \in \mathbb{R}^{d \times n^{+}}$ corresponds to the

positive class points. The factorised matrix $T \in \mathbb{R}_{+}^{d\times k}$ is the factorised matrix of A. As shown in Algorithm 6, once the basis matrix U is learned, SCANMF selects the top s ranked features by sorting the features in U according to $\|(u_i)\|_2$ in descending order. Moreover, the parameters α and β are used to control the contribution of correlation network and sparsity on feature matrix U respectively.

The objective function $\mathbf{O}$ in Eq. 4.1.2 is convex in U, $V^- \in \mathbb{R}^{n^-\times k}$, $V^+ \in \mathbb{R}^{n^+\times k}$ and T separately. Therefore, four iterative procedures are introduced which can find the local minima.

The objective function can be written as

$$\begin{aligned}\mathbf{O} =&\sigma^- \left[\mathrm{Tr}(X^-(X^-)^T) - 2\mathrm{Tr}(X^-V^-U) + \mathrm{Tr}(U(V^-)^TV^-U^T)\right] + \\ &\sigma^+ \left[\mathrm{Tr}(X^+(X^+)^T) - 2\mathrm{Tr}(X^+V^+U) + \mathrm{Tr}(U(V^+)^TV^+U^T)\right] + \\ &\alpha \left[\mathrm{Tr}(AA^T) - 2\mathrm{Tr}(AUT) + \mathrm{Tr}(UT^TTU^T)\right] + \beta\mathrm{Tr}(U^TDU)\end{aligned} \tag{4.1.3}$$

The four factor matrices U, V^-, V^+, T are initialized to non-negative values and then refined by means of gradient descent. D is the diagonal matrix with $D_{ii} = \frac{1}{2\|a^i\|_2+\epsilon}$. So, Lagrange multipliers ψ, ϕ, ε and τ are proposed for constraints U, V^-, V^+ and T respectively. The Lagrange form as follows

$$\begin{aligned}\mathbf{L} =&\sigma^- \left[\mathrm{Tr}(X^-(X^-)^T) - 2\mathrm{Tr}(X^-V^-U) + \mathrm{Tr}(U(V^-)^TV^-U^T)\right] + \\ &\sigma^+ \left[\mathrm{Tr}(X^+(X^+)^T) - 2\mathrm{Tr}(X^+V^+U) + \mathrm{Tr}(U(V^+)^TV^+U^T)\right] + \\ &\alpha \left[\mathrm{Tr}(AA^T) - 2\mathrm{Tr}(AUT) + \mathrm{Tr}(UT^TTU^T)\right] + \beta\mathrm{Tr}(U^TDU) + \\ &\mathrm{Tr}(\psi U) + \mathrm{Tr}(\phi V^-) + \mathrm{Tr}(\varepsilon V^+) + \mathrm{Tr}(\tau T)\end{aligned} \tag{4.1.4}$$

The partial derivatives of $\mathbf{L}$ with respect to U, V^-, V^+ and T are:

$$\frac{\partial \mathbf{L}}{\partial V^-} = -2\sigma^- X^- U^T + 2\sigma^- V^- U U^T + \phi \tag{4.1.5}$$

$$\frac{\partial \mathbf{L}}{\partial V^+} = -2\sigma^+ X^+ U^T + 2\sigma^+ V^+ U U^T + \varepsilon \tag{4.1.6}$$

$$\frac{\partial \mathbf{L}}{\partial T} = -2\alpha A U^T + 2\alpha T U U^T + \tau \tag{4.1.7}$$

$$\begin{aligned}\frac{\partial \mathbf{L}}{\partial U} = &-2\sigma^- X^- (V^-)^T + 2\sigma^- U (V^-)^T V^- \\ &- 2\sigma^+ X^+ (V^+)^T + 2\sigma^+ U (V^+)^T V^+ \\ &+ 2\alpha A T^T + 2\alpha U T^T T + 2\beta D U + \psi\end{aligned} \tag{4.1.8}$$

The following equations are formulated based on Karush–Kuhn–Tucker conditions where $\psi_{ik} u_{ik} = 0$, $\phi_{jk} vm_{jk} = 0$, $\varepsilon_{mk} vp_{mk} = 0$ and $\tau_{pk} t_{pk} = 0$,

$$-\left[\sigma^- X^- U^T\right]_{jk} vm_{jk} + \left[\sigma^- V^- U U^T\right]_{jk} vm_{jk} = 0 \tag{4.1.9}$$

$$-\left[\sigma^+ X^+ U^T\right]_{mk} vp_{mk} + \left[\sigma^+ V^+ U U^T\right]_{mk} vp_{mk} = 0 \tag{4.1.10}$$

$$-\left[\alpha A U^T\right]_{pk} t_{pk} + \left[\alpha T U U^T\right]_{pk} t_{pk} = 0 \tag{4.1.11}$$

$$\begin{aligned}&-\left[\sigma^- X^- (V^-)^T\right]_{ik} u_{ik} + \left[\sigma^- U (V^-)^T V^-\right]_{ik} u_{ik} \\ &-\left[\sigma^+ X^+ (V^+)^T\right]_{ik} u_{ik} + \left[\sigma^+ U (V^+)^T V^+\right]_{ik} u_{ik} \\ &-\left[\alpha A (U)^T\right]_{ik} u_{ik} + \left[\alpha U (T)^T T\right]_{ik} u_{ik} + \left[\beta D U\right]_{ik} u_{ik} = 0\end{aligned} \tag{4.1.12}$$

The four updating rules are denoted as

$$vm_{jk} \leftarrow vm_{jk}\frac{\left[X^{-}U^{T}\right]_{jk}}{\left[V^{-}UU^{T}\right]_{jk}} \tag{4.1.13}$$

$$vp_{mk} \leftarrow vp_{mk}\frac{\left[X^{+}U^{T}\right]_{mk}}{\left[V^{+}UU^{T}\right]_{mk}} \tag{4.1.14}$$

$$t_{pk} \leftarrow t_{pk}\frac{\left[AU^{T}\right]_{pk}}{\left[TUU^{T}\right]_{pk}} \tag{4.1.15}$$

$$u_{ik} \leftarrow u_{ik}\frac{\left[\sigma^{-}X^{-}(V^{-})^{T}+\sigma^{+}X^{+}(V^{+})^{T}+\alpha A(T)^{T}\right]_{ik}}{\left[\sigma^{-}UV^{-}(V^{-})^{T}+\sigma^{+}UV^{+}(V^{+})^{T}+\alpha U(T)^{T}T+\beta DU\right]_{ik}} \tag{4.1.16}$$

Input: Data Matrix X $\in \mathbb{R}_{+}^{d\times n}$
Parameters: k, α , β and MaxIteration
Initialization:
Construct the Adjacency Matrix A $\in \mathbb{R}_{+}^{d\times d}$
Initialize U $\in \mathbb{R}_{+}^{d\times k}$, V^{+} $\in \mathbb{R}_{+}^{n^{+}\times k}$, V^{-} $\in \mathbb{R}_{+}^{n^{-}\times k}$, set identity matrix D $\in \mathbb{R}_{+}^{d\times d}$, MaxIteration and $t = 0$
optimization:
while *MaxIteration*$> t$ **do**
Update V^{-} using Eq. 4.1.13 ;
Update V^{+} using Eq. 4.1.14 ;
Update T using Eq. 4.1.15 ;
Update U using Eq. 4.1.16 ;
Update Diagonal D as:
$\begin{bmatrix} \frac{1}{2\|(u_t)_1\|_2} & & \\ & \cdots & \\ & & \frac{1}{2\|(u_t)_d\|_2} \end{bmatrix}$;
end
Output: Sort all the features in optimized matrix U according to $\|(u_i)\|_2$ in descending order.
Return the top s ranked features.

Algorithm 3: SCANMF for Feature Selection

4.1.2 Experiments and Datasets

In this section, the performance of the proposed algorithm SCANMF is evaluated and compared with state-of-the-art classification algorithms. In the experiments, initially the algorithm selects the top s ranked features then utilizes one of several supervised learning algorithms, such as SVM, Naive Bayes, and CART to classify the samples based on the top ranked features.

Datasets

As shown in Table 4.1, the experiments are conducted on seven publicly available datasets, in addition to Affymetrix childhood leukaemia which is described in Section 3.3.3. ALL/AML is a gene expression data to classify acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) (Golub et al. 1999). Breast cancer is another gene expression dataset associated with breast cancer patients who developed metastases within 5 years or remain metastases free after 5 years (Van De Vijver et al. 2002). The complex structure of the evaluated datasets such as high-dimensionality, class-imbalance, highly correlated data with a low number of samples serves as a good test bed for a comprehensive evaluation.

Experimental Settings

To demonstrate the effectiveness of SCANMF for feature selection, its performance is compared with the following state-of-the-art feature selection methods which are reviewed in chapter 2:

1. LASSO is a penalised regression method which is used for variable selection (Tibshirani 1996).

Table 4.1: Datasets, column #IR is the imbalance ratio (Neg/Pos)

Dataset	#Attributes	#Instances	#IR	Source
Childhood Leukaemia	22277	60	2.85	TB-CHW
Colon cancer	2000	60	1.81	(Alon et al. 1999)
Lung cancer	12533	181	4.8	(Gordon et al. 2002)
Global Cancer Map	16064	281	2.11	(Reich et al. 2006)
Lymphoma Outcome	7130	78	3.05	(Reich et al. 2006)
ALL/AML	7129	73	2.92	(Golub et al. 1999)
Breast cancer	8141	295	2.49	(Van De Vijver et al. 2002)
DLBCL	5470	77	3.05	(Reich et al. 2006)

2. Elastic Net is a variable selection method that uses the (l_2 norm) to shrink the correlated covariates toward each other (Zou & Hastie 2005).

3. Gene ranking using discriminant non-negative matrix factorization (DNMF) exploits the existence of the labels (Jia et al. 2015).

4. Fuzzy Forest is a feature selection method to handle the highly correlated high-dimensional datasets (Conn et al. 2015).

The performance of the selected features is evaluated in relation to the different algorithms regarding classification accuracy by the widely used evaluation metric AUC. Three classifiers are used, namely SVM, Naive Bayes, and CART. Estimates of AUC were averaged over five-fold cross-validation. The adjacency matrix using soft thresholding is built, and it sets $\theta = 1$ according to Eq. 4.1.1. For a fair comparison between the algorithms, the results in Tables 4.2, 4.3 and 4.4 used $s = 10\%$ of the top ranked features.

Without loss of generality, some parameters need to be set in advance to conduct the experiments. The optimal values for the parameters are selected by using a grid-search strategy as reported in the following subsection 4.1.3.

4.1.3 Results and Analysis

The performance of the proposed SCANMF algorithm is evaluated against state-of-the-art algorithms on eight datasets. Three different popular classifiers are used to evaluate the selected features and to demonstrate the generality of SCANMF. The detailed results on the datasets are summarized in Tables 4.2, 4.3 and 4.4. The results show the performance of the classifiers substantially improves by using the SCANMF feature selection method. The proposed SCANMF outperforms the compared feature selection algorithms on almost all eight datasets. From the results in Tables 4.2, 4.3 and 4.4, the following observations are made. First, compared to LASSO, the results show the importance of using correlated features to enhance the AUC and not to consider them as redundant features. Thus, the algorithms can exploit the correlation to improve the classification performance and the interpretation of the results. Second, by jointly using NMF and correlation networks together, the selected features may have an interpretation value which may discover some genetic pathways. Finally, SCANMF achieves better classification results compared to the other methods due to the robustness characteristics of the proposed method, which takes into consideration both the correlation structure of the features, and the class imbalance problem with $l_{2,1}$-norm regularization. Moreover, Figure 5.4 showing the convergence curves of SCANMF against the number of iterations, demonstrates that the proposed optimization algorithm is efficient and converges quickly. T-tests are conducted between the vector results of SCANMF (Base) against state-of-the-art methods for each classifier, under the null hypothesis that the AUC on vectors of the used methods is not significantly different. As shown in the last column of Tables 4.2, 4.3, and 4.4, the p-values reject the null hypothesis, as most values of the SCANMF methods (base) are

Table 4.2: SVM classification results (AUC) of different feature selection algorithms on several datasets. The best AUC results are highlighted in bold.

	LASSO	Elastic Net	DNMF	Fuzzy Forest	**SCANMF**
Childhood leukaemia	0.55	0.60	0.50	0.75	**0.86**
Global Cancer Map	0.60	0.63	0.60	0.77	**0.84**
ALL/AML	0.62	0.72	0.50	0.80	**0.84**
Breast cancer	0.66	0.76	0.61	0.69	**0.76**
DLBCL	0.72	0.78	0.75	0.82	**0.98**
Lung Cancer	0.71	**0.91**	0.50	0.83	0.90
Colon Cancer	0.74	0.78	**0.90**	0.78	0.86
t-test	1.4×10^{-4}	2.1×10^{-2}	2.4×10^{-3}	1.7×10^{-2}	Base

Table 4.3: Naive Bayes classification results (AUC) of different feature selection algorithms on several datasets. The best AUC results are highlighted in bold.

	LASSO	Elastic Net	DNMF	Fuzzy Forest	**SCANMF**
Childhood leukaemia	0.43	0.45	0.50	0.76	**0.83**
Global Cancer Map	0.61	0.65	0.60	0.74	**0.76**
ALL/AML	0.68	0.71	0.50	0.73	**0.82**
Breast cancer	0.69	0.69	0.61	0.66	**0.72**
DLBCL	0.79	0.79	0.75	0.78	**0.88**
Lung Cancer	0.78	0.84	0.50	0.80	**0.85**
Colon Cancer	0.71	0.73	**0.90**	0.73	0.85
t-test	1.4×10^{-2}	3.7×10^{-2}	8.4×10^{-3}	2.0×10^{-2}	Base

lower than 0.05. This indicates that the proposed method SCANMF has significantly improved the performance of the classifiers.

Parameter Sensitivity

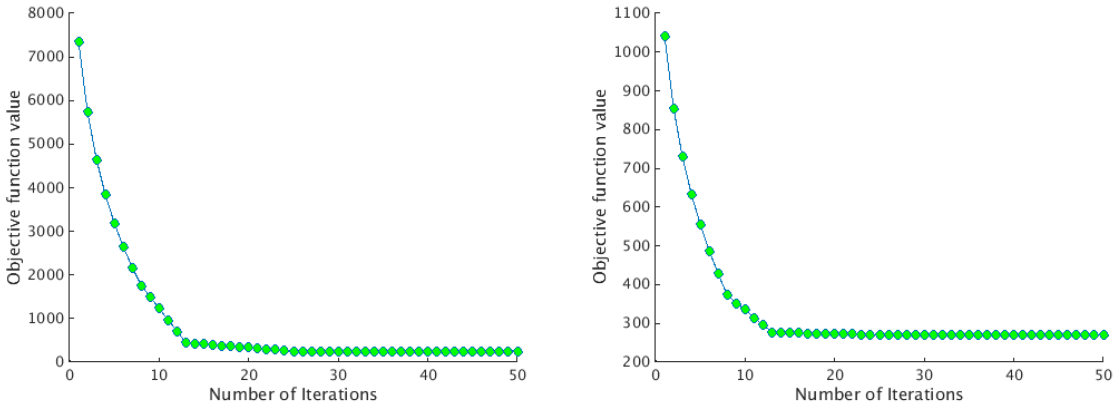
The proposed SCANMF is similar to other feature selection algorithm in that it requires parameters to be set in advance. For SCANMF, they are α, β and the latent dimension k. In this subsection, the sensitivity of the proposed algorithm to these

Table 4.4: CART classification results (AUC) of different feature selection algorithms on several datasets. The best AUC results are highlighted in bold

	LASSO	Elastic Net	DNMF	Fuzzy Forest	**SCANMF**
Childhood leukaemia	0.47	0.45	0.50	0.77	**0.82**
Global Cancer Map	0.58	0.59	0.60	**0.78**	0.74
ALL/AML	0.70	0.73	0.50	0.71	**0.83**
Breast cancer	0.67	0.67	0.61	0.67	**0.69**
DLBCL	0.78	0.82	0.75	0.79	**0.95**
Lung Cancer	0.75	0.89	0.50	0.79	**0.90**
Colon Cancer	0.74	0.74	**0.90**	0.75	0.89
t-test	1.1×10^{-2}	6.4×10^{-2}	2.1×10^{-3}	6.1×10^{-2}	Base

parameters is discussed on the selected features regarding classification performance. First, parameters α and β control the contribution of the network structure of the features and the $l_{2,1}$-norm regularization respectively. These two parameters are tuned using a grid-search from $\{0.1, 0.2, 0.4, 0.6, 0.7, 0.9, 1\}$. As shown in Figure 4.3, the performance changes with respect to different datasets. A common observation from Figure 4.3 is that the AUC values are changed by selecting different parameter values which indicates the contribution of the SCANMF components in the performance of the selected features. Specifically, the results in Figure 4.3a and 4.3b are responsible for the contribution of the network of correlated features in the objective function. It suggests that very small α degrades the AUC which demonstrates the importance of considering the network structure to find the optimal features. It shows that a medium value of α achieved good results. The results obtained in Figure 4.3c and 4.3d show the importance of incorporating $l_{2,1}$ on the feature basis matrix to reduce irrelevant features. The best results are achieved when $\beta \in [0.4, 0.7]$. Finally, the impact of different values of latent dimension k is explored. The results obtained on

the given datasets, as shown in Figure 4.4, demonstrate that small k and particularly $k = 2$ achieves better results than a large k, potentially indicating a relation between the latent dimension and the number of labels in the dataset (Witten & Tibshirani 2010).

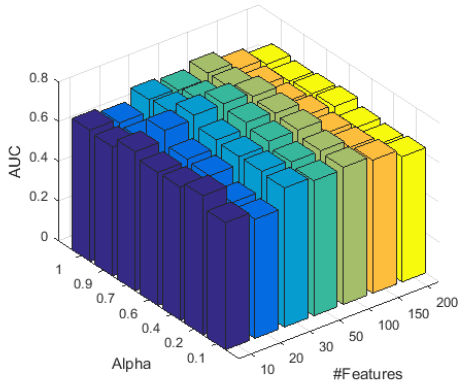


(a) Childhood Leukaemia dataset (b) Lung Cancer

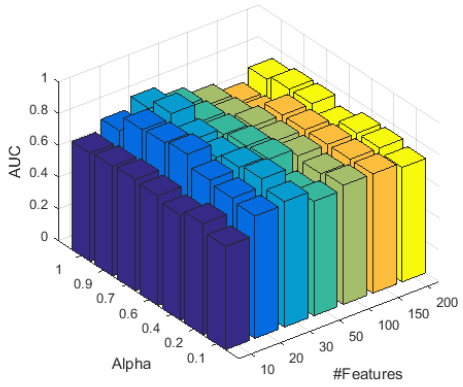
Figure 4.2: Convergence curves of the SCANMF algorithm

4.2 Fuzzy Ensemble Feature Learning for Highly Correlated high-dimensional Data

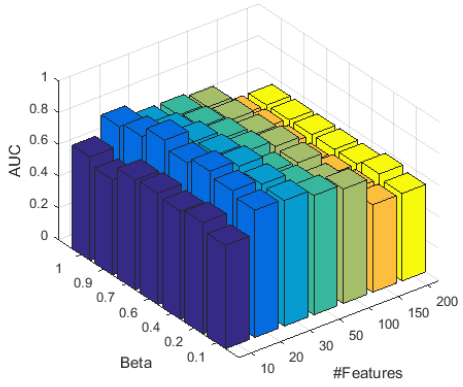
In the era of high-throughput technologies, the term "big data" is coined to reflect the amount of the data increasingly being generated in many fields. The available data exceeds the ability of the existing machine learning algorithms to analyse it. The complexities and challenges of data in some fields are reflected in the generated datasets. One of these types of complex structures is high-dimensional data which have a relatively low number samples, known as "the curse of dimensionality" problem



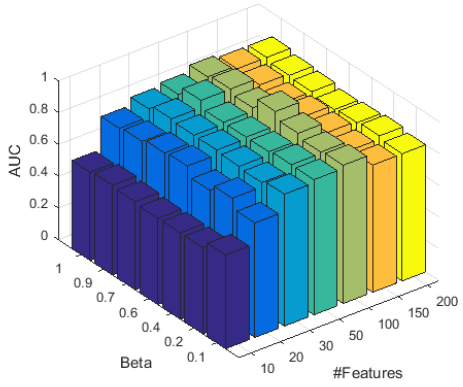
(a) Childhood Leukaemia



(b) Lung Cancer

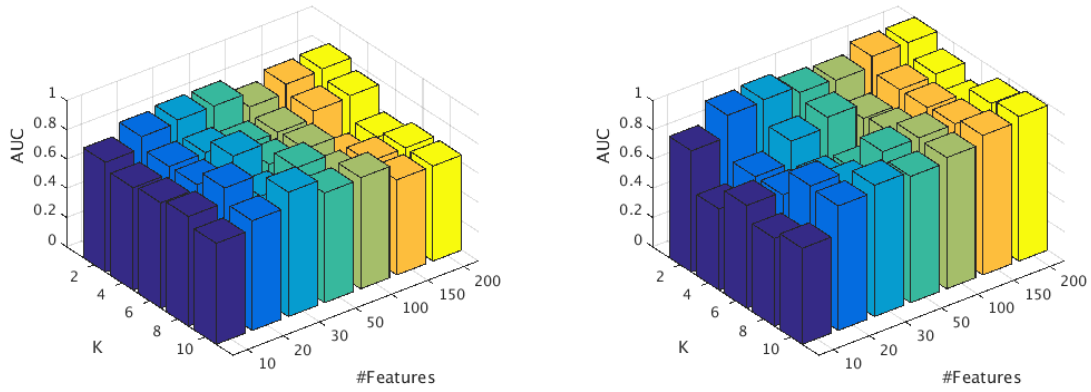


(c) Childhood Leukaemia



(d) Lung Cancer

Figure 4.3: Classification performance of SVM (AUC) using SCANMF with different α, β and feature numbers



(a) Childhood Leukaemia (b) Lung Cancer

Figure 4.4: Classification performance of SVM (AUC) using SCANMF with different k and feature numbers

or $p >> n$. The problem of the curse of dimensionality has become increasingly common in several domains, especially in biomedicine and genomics applications. Furthermore, the dilemma is exacerbated by the presence of highly correlated features and the imbalanced data problem. Surprisingly, a few feature ranking algorithms based on correlated features are proposed in the literature, as reviewed in section 2.2.

Recently, I co-authored an ensemble SVM (ESVM-RFE) algorithm (Anaissi et al. 2016) for individual feature ranking in high-dimensional data. The ESVM-RFE uses the ensemble strategy with the SVM (Boser et al. 1992) classifier as the base learning model. As reviewed in chapter 2, binary SVM uses a decision boundary to separate two classes, defined by solving a quadratic optimization problem. The decision boundary is specified by a subset of critical training samples named support vectors that lie on the edge. Ensemble techniques have the advantage of handling the problem of the curse of dimensionality and reducing the potential of over-fitting the training data. The ESVM-RFE follows the ensemble and bagging concepts of random forest and

adopts a backwards elimination strategy. Also, it handles the problem of imbalanced datasets by constructing roughly balanced bootstrap samples or bootstrap samples biased to the minority class.

In this section, a novel fuzzy ensemble feature learning algorithm (Fuzz-ESVM) is proposed to handle the correlated features in high-dimensional data. The Fuzz-ESVM algorithm consists of three components: first, it generates isolated feature modules based on the network structure of the data. Each module contains the correlated features, and the correlation between the modules is low. Second, the previous study of the ESVM-RFE algorithm (Anaissi et al. 2016) is used to select the most important features within each module. Finally, the selected features are aggregated and again, ESVM-RFE is applied to select the optimal features of the modules.

4.2.1 Methods

In this section, Fuzz-ESVM is proposed as a feature module learning framework. Sub-section 4.2.1 shows how to cluster the co-expression networks to generate the feature modules. Sub-section 4.2.1 reviews the ESVM-RFE algorithm, and finally, presents the proposed Fuzz-ESVM algorithm.

Clustering Co-Expression Networks

A very widely used method to cluster the co-expression biological networks is hierarchical clustering and in particular, the weighted gene co-expression network analysis (WGCNA) algorithm. The WGCNA is initially developed to find the relevant biological modules by detecting a network of highly correlated genes (Bin & Steve 2005). The gene co-expression network generated by WGCNA can be clustered into groups

of highly interconnected nodes.

The WGCNA uses a similarity function such as Pearson correlation to construct a correlated similarity network between the genes. Then, the similarity network is transformed into an adjacency network by taking the absolute value of the similarity network entries and raising it to the power β. This step indicates the strong correlation among genes and rejects the weak ones. Scale-free topology criterion is used to choose the best value of parameter β. Next, the modules are identified by searching for strongly connected genes which is known as high topological overlap. After constructing the topological overlap network for all pairs of genes, the hierarchical clustering algorithm uses this information to identify the modules of correlated genes. The WGCNA has the advantage that it does need to set the number of clusters in advance.

Review of ESVM-RFE

The ESVM-RFE ranks the features by constructing an ensemble of SVM models in each iteration of SVM-RFE using a random bootstrap subset from the training set. Then, it aggregates all the feature rankings as an ensemble vote. The least important features are eliminated based on multiple votes in each iteration. This process is repeated until a specified number of features is reached.

The Proposed Fuzz-ESVM based on the Co-Expression Feature Network

The proposed Fuzz-ESVM does not consider the correlated features as redundant which must be removed. For example, in microarray gene expression data, genes

that have either similar genomic locations or molecular functions are assumed to co-function and are highly correlated (Toloşi & Lengauer 2011). The correlation issue negatively impacts the classical feature selection algorithms which follow an individual feature ranking process.

The Fuzz-ESVM algorithm aims to achieve a feature selection in the presence of the correlated features. It follows a backwards feature elimination method. The flowchart of the Fuzz-ESVM algorithm is shown in Figure 4.5. It is divided into two phases: an intra-screening phase and an inter-screening phase. The intra-screening phase is composed of multiple steps: firstly, it constructs a feature co-expression network which captures the correlation between the features. This step can be achieved using WGCNA or any other graph clustering method. Then, the modules M_i of the correlated features are extracted from the feature co-expression network using hierarchical clustering. The screening phase is applied on each module M_i to filter out the unimportant features. It is known as intra-screening because it operates on each module independently. For each correlated feature module M_i, ESVM-RFE is used to generate weights for the features. This is described in Algorithm 4. It starts with the entire set of features in the module, and in each iteration, an ensemble SVM is trained by taking bootstrap samples from the training dataset. The feature weights are estimated using the absolute value of coefficients of the support vectors for each SVM model. The estimated feature weights are aggregated from the ensemble of SVM models and ranked in decreasing order to remove the least important features with the small weights. Features are eliminated over multiple iterations on each module until a specific threshold of the number of selected features is reached.

The selected features from each module in the intra-screening phase are aggregated

and passed as an input to the inter-screening phase. The inter-screening phase in the Fuzz-ESVM algorithm is to capture the interaction among the modules. It aggregates all the surviving features from the previous phase and applies one more ESVM-RFE to select the global surviving features.

The proposed Fuzz-ESVM algorithm is an appropriate solution to reduce the correlation bias in the presence of imbalanced data. It differs from other feature selection algorithms in the following ways: first, it makes use of the ensemble SVM to reduce the influence of correlation bias, because in each iteration, the ranking decision is generated from multiple SVM models on different bootstraps and it is not related to the module size. Second, feature ranking in each iteration is achieved on equal bootstrap samples to mitigate the effect of the imbalance class problem. Third, it uses WGCNA to estimate the network structure of the data, and consequently, estimate the correlated features. Finally, stability is targeted by achieving multiple perturbations of constructing SVM models in each iteration.

4.2.2 Experiments and Datasets

In this section, the experimental evaluations on high-dimensional, highly correlated datasets are reported. This chapter analyses and compares the classification performance of the proposed Fuzz-ESVM against the state-of-the-art algorithms namely SVM-RFE (Guyon et al. 2002), Fuzzy Forests (Conn et al. 2015), and Hybrid L1/2 L2 regularization (HLR) (Huang, Liu & Liang 2016). SVM-RFE is evaluated as a baseline method, Fuzzy Forests as an ensemble feature ranking algorithm for correlated features, and HLR from the point of view of a sparse model for dimensionality reduction in the presence of correlated features. It is important to note that the main

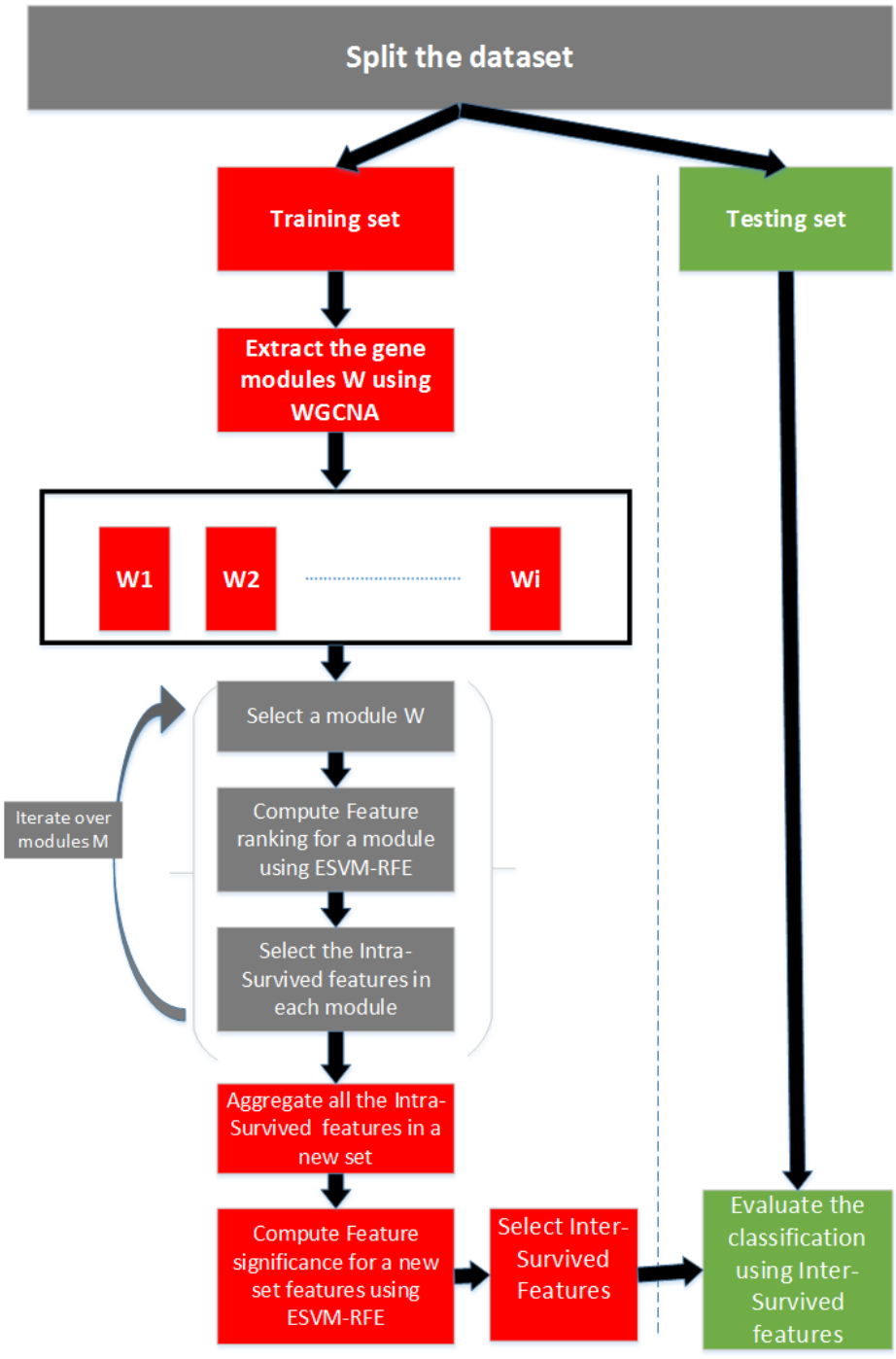


Figure 4.5: The flowchart of Fuzz-ESVM

Input: training data X
Class labels y
parameter: *inter-d* ; `// Number of selected features between modules`
intra-d ; `// Number of selected features from each module`
b ; `// Size of ensemble SVM in each iteration`
E ; `// The proportion of features to eliminate at each iteration`
bagSize ; `// Roughly balanced bootstrap from training dataset`

$modules \leftarrow$ `WGCNA(`X`)` ; `// the interconnected features from each modules using WGCNA package in R`
$l \leftarrow \text{length}(modules)$;
for $i \leftarrow 1$ **to** l **do**
  $M \leftarrow modules_i$;
  $data \leftarrow X(, M)$;
  ; `//` M `is the correlated genes in each module`
  $intra\text{-}features \leftarrow$ `ESVMRFE(`$data$, y, b, E, $intra\text{-}d$,$bagSize$`)`;
  $intra\text{-}Set \leftarrow intraSet \cup intra\text{-}features$;
end
$selectedData = trainingdata[, intra\text{-}Set]$;
$inter\text{-}features \leftarrow$ `ESVMRFE(`$selectedData$,y,b,E,$inter\text{-}d$,$bagSize$`)` ;
`// inter-features: the surviving features between the modules using Algorithm ESVM-RFE`
Output: $inter\text{-}features$

Algorithm 4: Fuzzy ensemble feature learning algorithm (Fuzz-ESVM)

Function *ESVMRFE (data, class, b, E, d, bagSize)*
  Input: $surviveIndexes = seq(1 : ncol(data))$
    $n = nrow(data)$
  for $d \leftarrow 1$ **to** $length(surviveIndexes)$ **do**
    $m = length(surviveIndexes)$;
    $survive = m - m \times E$; `// survive: number of features to select in the current iteration`
    $ensRes = matrix(n, b)$; `// ensRes: feature's weight of each SVM model`
    for $i \leftarrow 1$ **to** b **do**
      $bag \leftarrow bootstrap(data, bagSize)$;
      $bagClass \leftarrow bootstrap(class, bagSize)$;
      $model \leftarrow svm(bag[, survivingIndexes], bagClass)$;
      $weightVector \leftarrow transpose(model\$coefs) \% * \% model\$SV$;
      `// Compute the weight vector`
      $featureWeight \leftarrow weightVector * weightVector$; `// Compute ranking criteria`
      $ensRes \leftarrow merge(ensRes, featureWeight)$; `// Accumulate feature's weight`
    end
    $totalWeight = rowSum(ensRes)$; `// Aggregate feature's weight`
    $sortedWeight \leftarrow sort(totalWeight)$; `// Sort the total feature's weight by decreasing order`
    $sortedIndexes \leftarrow index(sortedWeight)$;
    $surviveIndexes \leftarrow surviveIndexes[sortedIndexes[1 : survive]]$;
    `// Eliminate features with smallest weight`
  end
  Output: $selectedData = data[, surviveIndexes]$

Algorithm 5: ESVM-RFE for feature learning

purpose of these experiments is to evaluate the potential of the proposed Fuzz-ESVM algorithm to improve the classification performance in the presence of a large number of correlated features.

Without loss of generality, linear SVM is used as a classifier to evaluate the performance of the selected features from the compared algorithms. The performance is measured by the widely used metric AUC under the receiver operating characteristic (ROC) analysis. The optimal tuning parameters of the Fuzz-ESVM, Fuzzy Forests, HLR and SVM-RFE approaches were identified by five-fold cross-validation on the training set. The datasets are divided at random such that approximately 75% is used as a training set and 25% as a test set. The datasets are z-score normalized.

Datasets

The experiments are conducted on one dataset collected from The Children's Hospital at Westmead, and four public datasets. The details of these datasets are summarised in Table 4.5. The common characteristics of these datasets are highly dimensional, highly correlated, have a small number of samples and some of them are imbalanced. A stratified random sampling function (stratified) in R is applied on the evaluated datasets to split the data into a training and testing set, with a quarter of the dataset considered as a testing set and the reminder as a training set. The Childhood Leukaemia dataset is described in Section 3.3.3. The DLBCL-FSCC dataset is a clinical dataset on patients who received cyclophosphamide, adriamycin, vincristine and prednisone (CHOP)-based chemotherapy, and the class outcome of each patient is either cured or fatal (Shipp et al. 2002). The prostate cancer dataset is a microarray gene expression dataset which consists of 6033 genes and 102 samples. The class

Table 4.5: Datasets

Dataset	#Attributes	#Instances	Source
Childhood Leukaemia	22277	60	TB-CHW
DLBCL-FSCC	7129	77	(Shipp et al. 2002)
Prostate cancer	6033	102	(Singh et al. 2002)
ALL/AML	7129	73	(Golub et al. 1999)
Breast cancer	8141	295	(Van De Vijver et al. 2002)

outcome is whether a patient suffers prostate cancer or is healthy (Singh et al. 2002). ALL/AML and Breast datasets are described in Section 4.1.2.

4.2.3 Results and Discussion

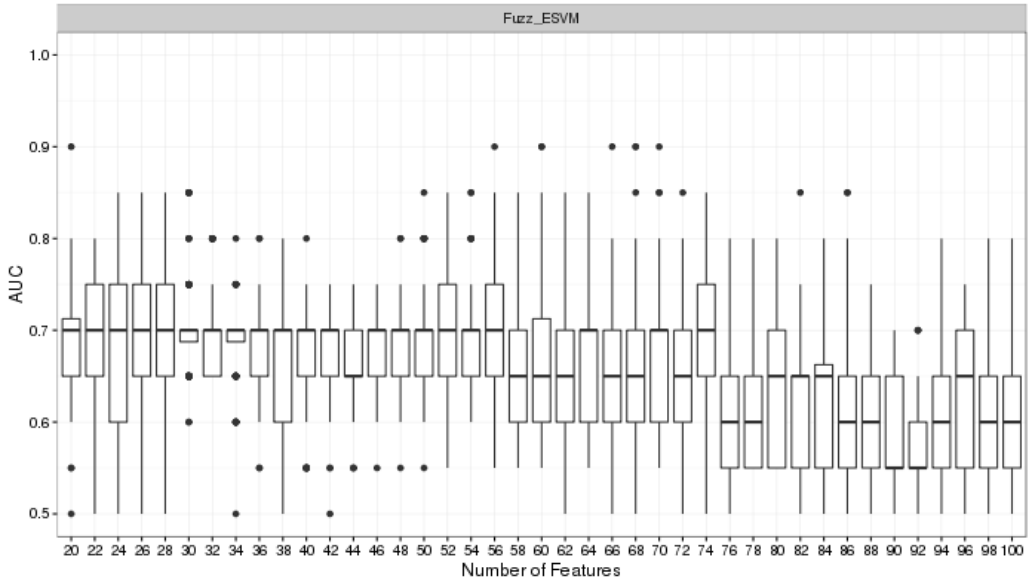
The goal of this section is to evaluate the performance of the selected features from the compared algorithms on the real-world datasets. In the following experiments, for a fair comparison of all algorithms, the AUC accuracy is estimated using the .632+ bootstrap method (Ambroise & McLachlan 2002) with 100 bootstrap samples. For each bootstrap sample, AUC accuracy is obtained on the test dataset.

Figure 4.8 to Figure 4.15 show the AUC evaluated on the test dataset across a different number of features. The figures present the results of up to 100 features because the evaluated datasets contain a small number of samples which needs a small number of features to avoid over-fitting. As shown in Figure 4.8 to Figure 4.15, the proposed Fuzz-ESVM algorithm outperforms the state-of-the-art feature selection methods in most feature sets in all datasets. The AUC classification performance is further investigated based on the best number of selected features. As shown in Table 4.6, several statistical measures are included, namely minimum, maximum, first quartile, third quartile, median and mean on 100 bootstrap samples on the test data.

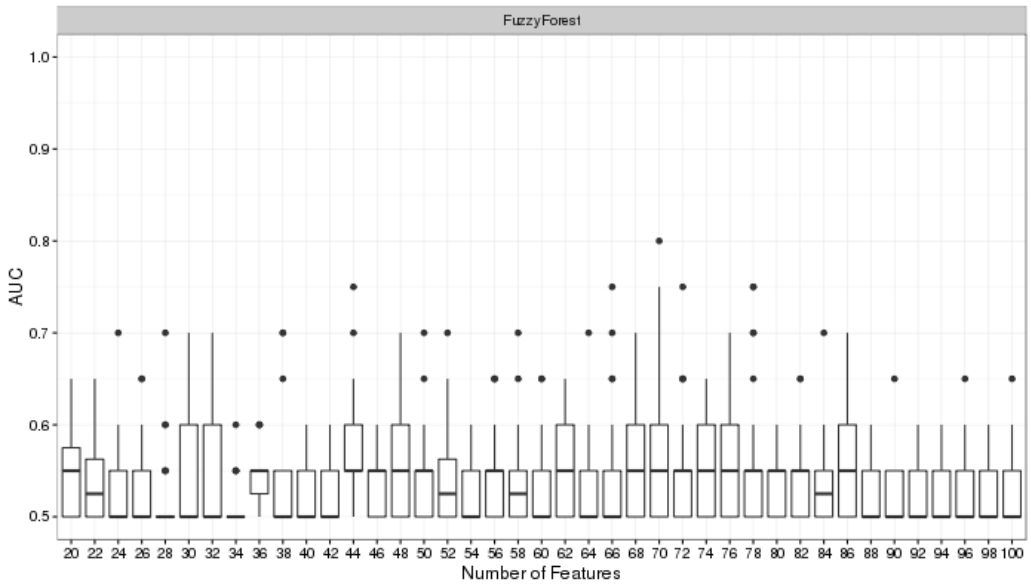
For example, as shown in Table 4.6 for the DLBCL-FSCC dataset, the best AUC is achieved for the compared algorithms: Fuzz-ESVM, Fuzzy Forest, HLR and SVM-RFE is 50, 74, 74, and 78 features respectively. This clearly shows that the proposed algorithm achieves better results than the compared algorithms using different statistical measures. Furthermore, the proposed algorithm Fuzz-ESVM obtained the best accuracy results compared to the others with a small number of features in most datasets, which leads to less computational complexity during the training process. It can also be observed that the classification results of the Fuzz-ESVM algorithm tend to be stable after increasing the selected features above approximately 50 features in the evaluated datasets. This indicates the stability and capability of Fuzz-ESVM to select a lower percentage of features and realise good accuracy results.

On the other hand, the experimental results indicate that the feature selection methods that handle correlated features such as the proposed Fuzz-ESVM, Fuzzy forest, and HLR, perform better than SVM-RFE which handles the individual feature selection method. Therefore, it demonstrates the importance of handling correlated features in high-dimensional datasets to improve the performance of the classifiers. Finally, a statistical t-test is also conducted between the vector results of the proposed algorithm against state-of-the-art methods under the null hypothesis that AUC on vectors of the used method is not significantly different to Fuzz-ESVM. The p-value is lower than 0.05 which rejects the null hypothesis.

A further investigation is made of the selected features using the proposed Fuzz-ESVM from the ALL/AML dataset to see if the proposed algorithm can define separated clusters based on ALL and AML class outcomes. To do this, Singular Value Decomposition (SVD) is applied to the original ALL/AML training set using all the

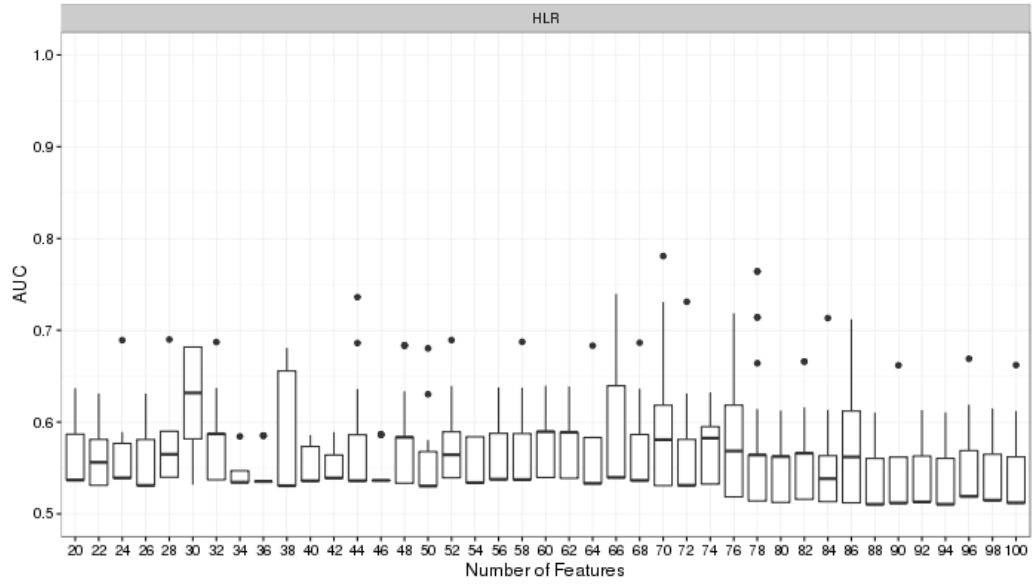


(a) Fuzz-ESVM

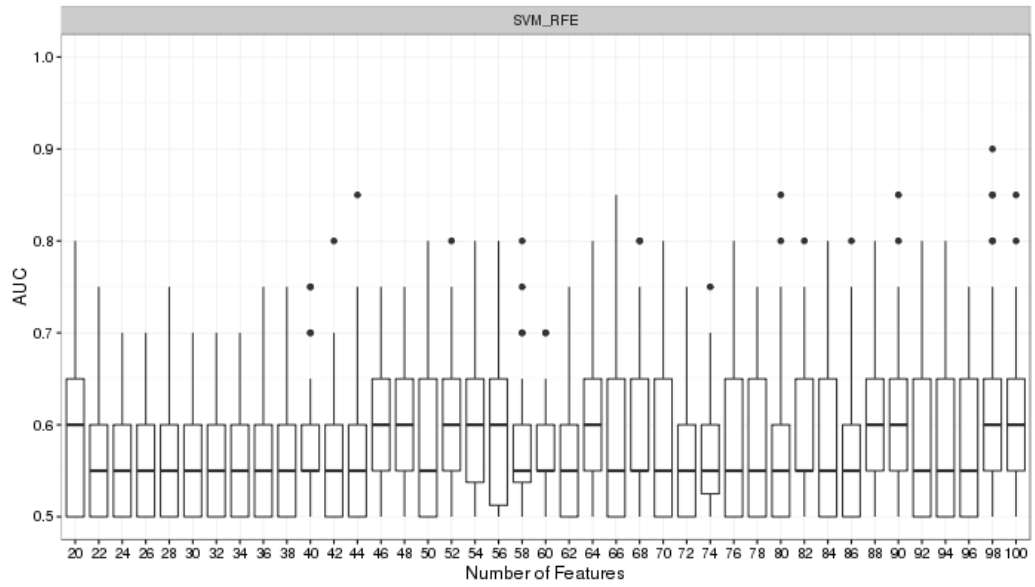


(b) Fuzzy Forest

Figure 4.6: Classification performance comparison between algorithms evaluated on Childhood Leukaemia dataset using the 0.632+ bootstrap method with 100 bootstrap samples across a different number of features

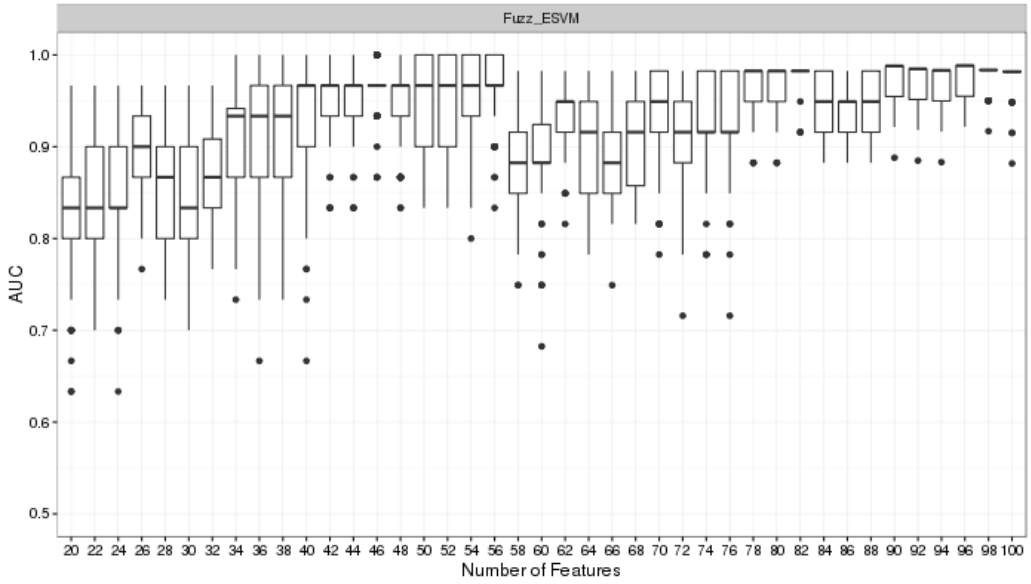


(a) HLR

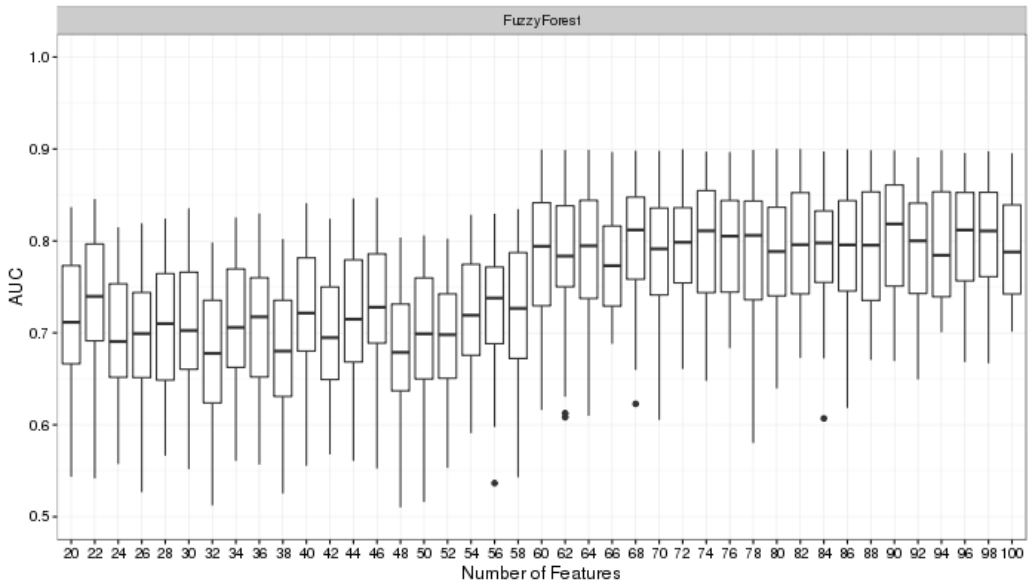


(b) SVM-RFE

Figure 4.7: Classification performance comparison between algorithms evaluated on Childhood Leukaemia dataset using the 0.632+ bootstrap method with 100 bootstrap samples across a different number of features (continuation of Figure 4.6)

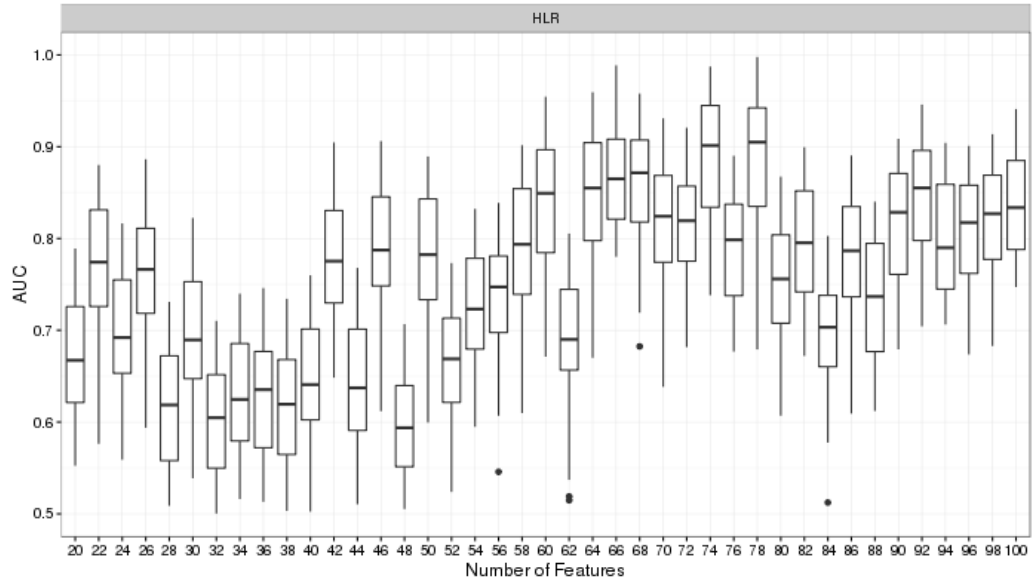


(a) Fuzz-ESVM

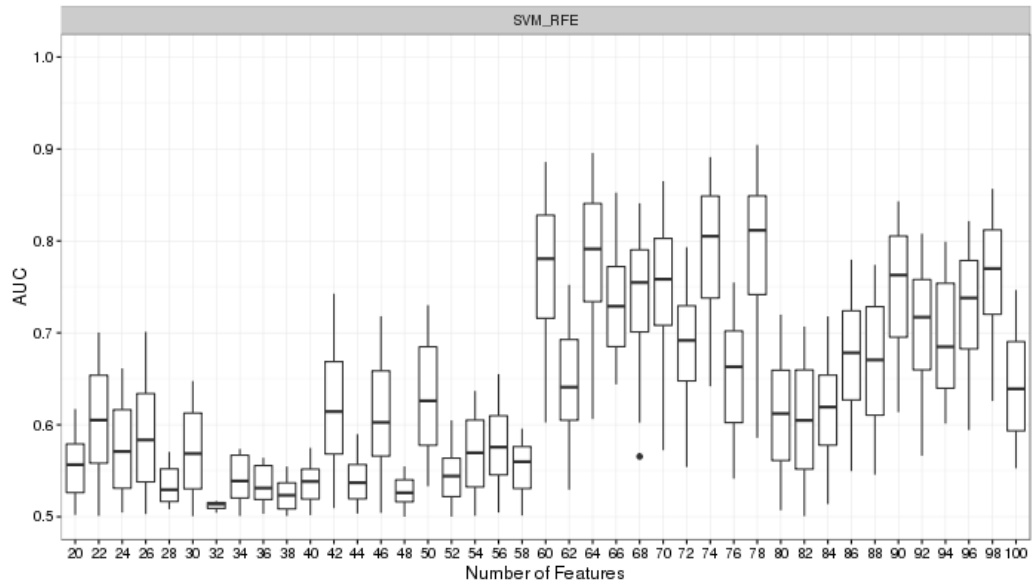


(b) Fuzzy Forest

Figure 4.8: Classification performance comparison between algorithms evaluated on the DLBCL-FSCC dataset using the 0.632+ bootstrap method with 100 bootstrap samples across a different number of features

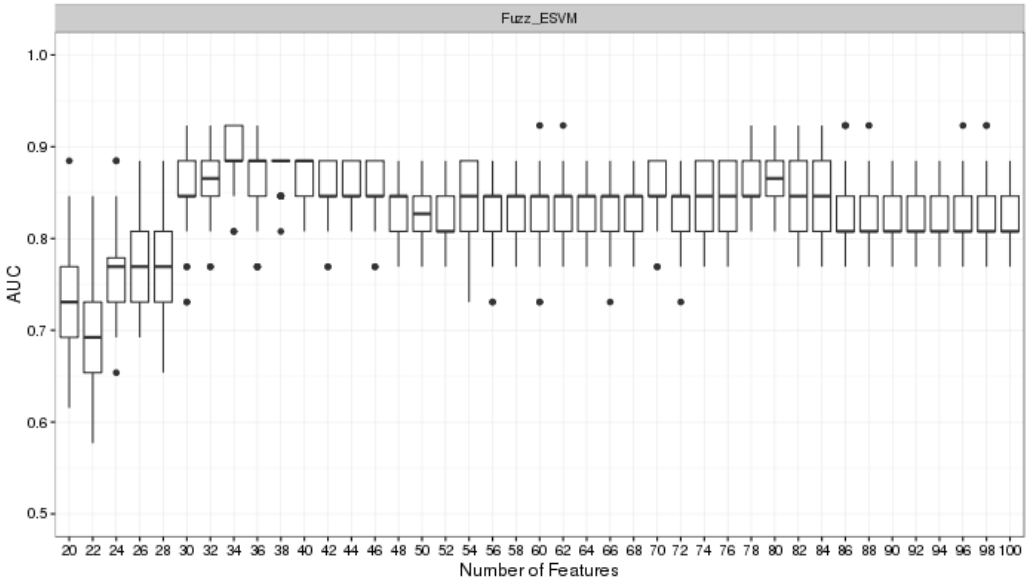


(a) HLR

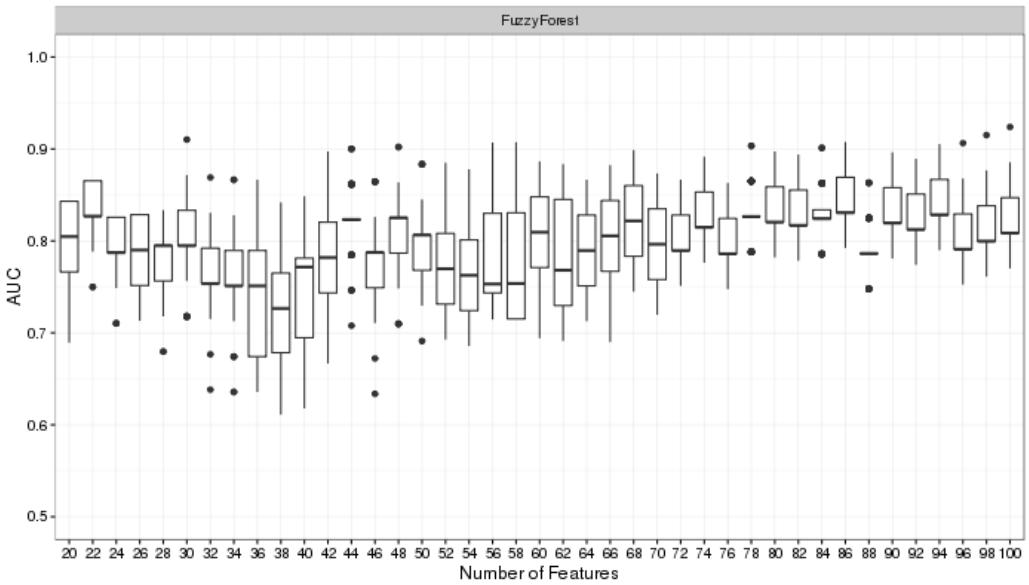


(b) SVM-RFE

Figure 4.9: Classification performance comparison between algorithms evaluated on the DLBCL-FSCC dataset using the 0.632+ bootstrap method with 100 bootstrap samples across a different number of features (continuation of Figure 4.8)

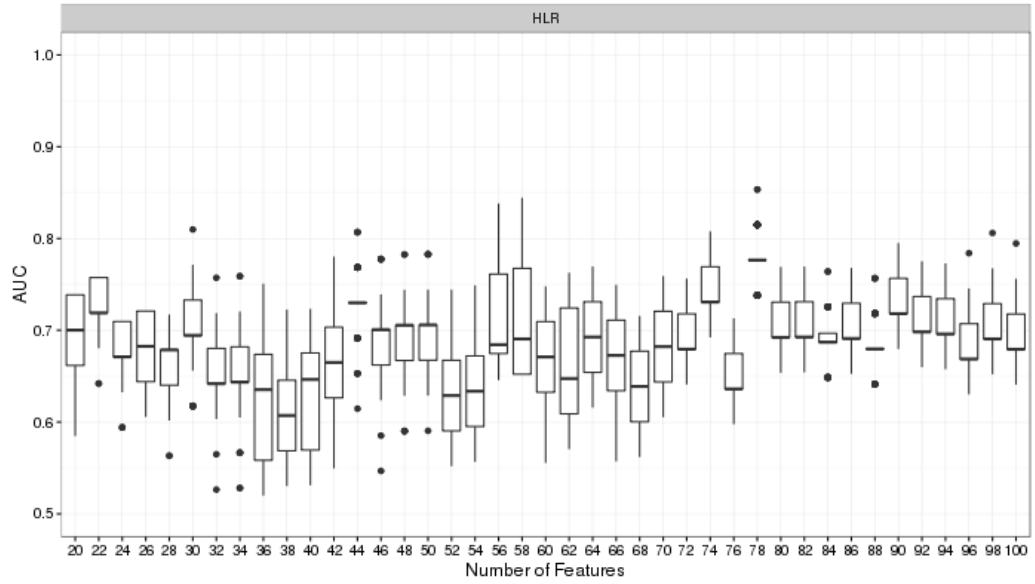


(a) Fuzz-ESVM

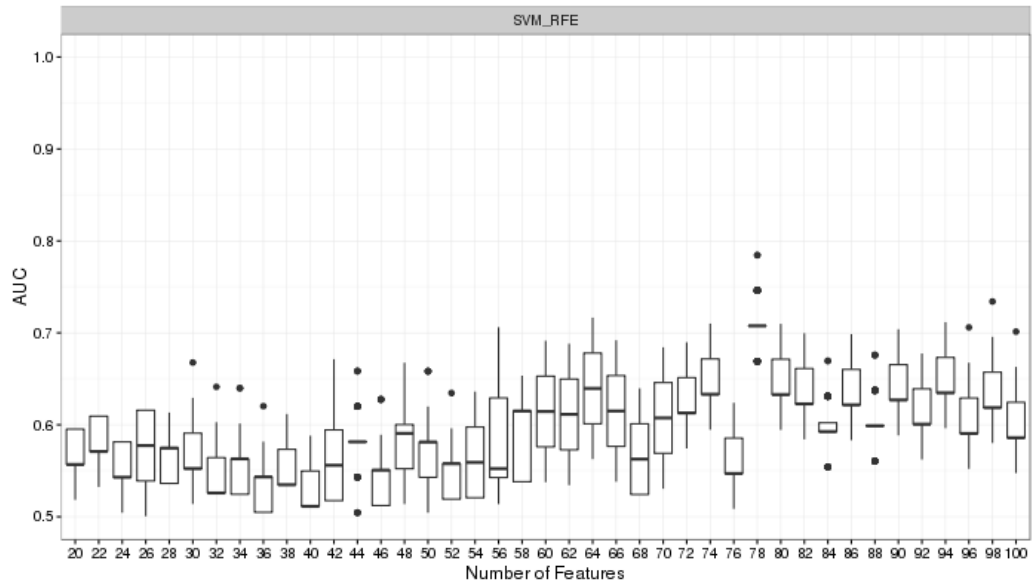


(b) Fuzzy Forest

Figure 4.10: Classification performance comparison between algorithms evaluated on the prostate cancer dataset using the 0.632+ bootstrap method with 100 bootstrap samples across a different number of features

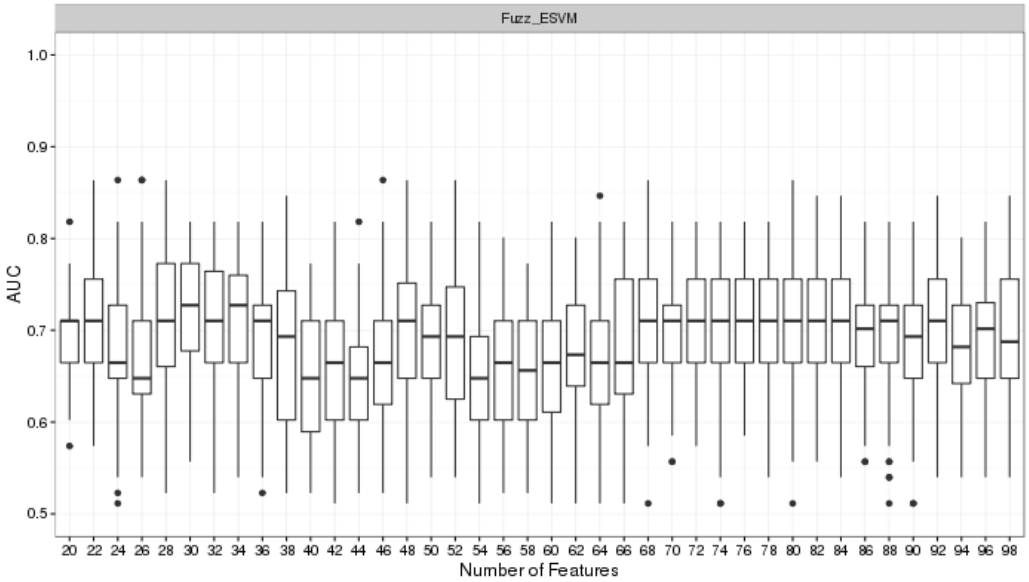


(a) HLR

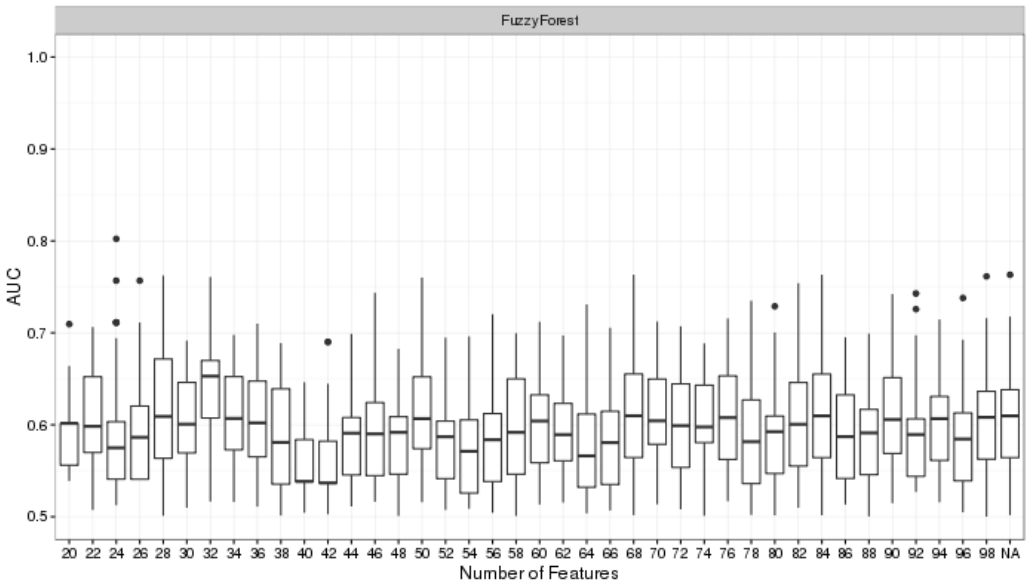


(b) SVM-RFE

Figure 4.11: Classification performance comparison between algorithms evaluated on the prostate cancer dataset using the 0.632+ bootstrap method with 100 bootstrap samples across a different number of features (continuation of Figure 4.10)

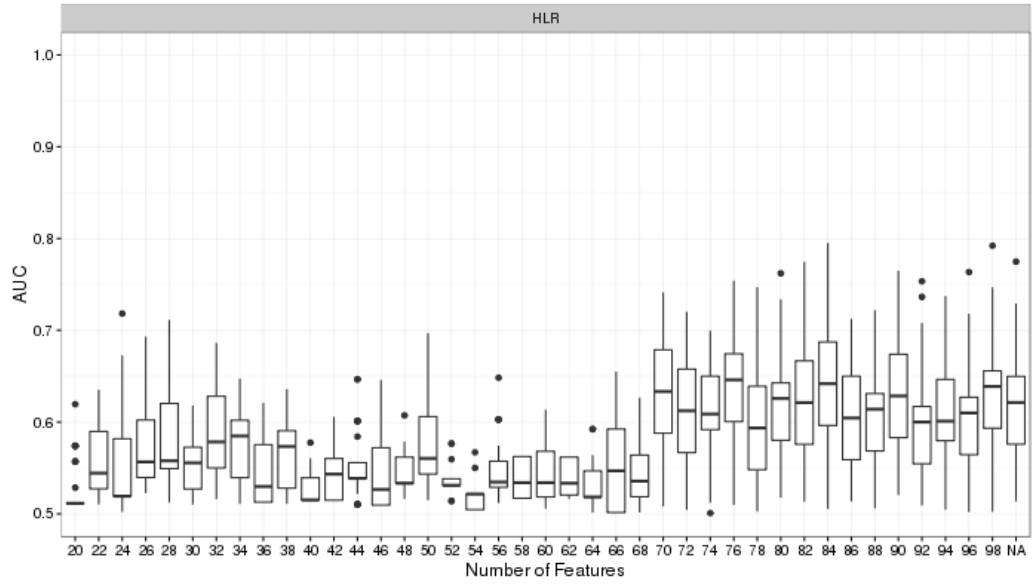


(a) Fuzz-ESVM

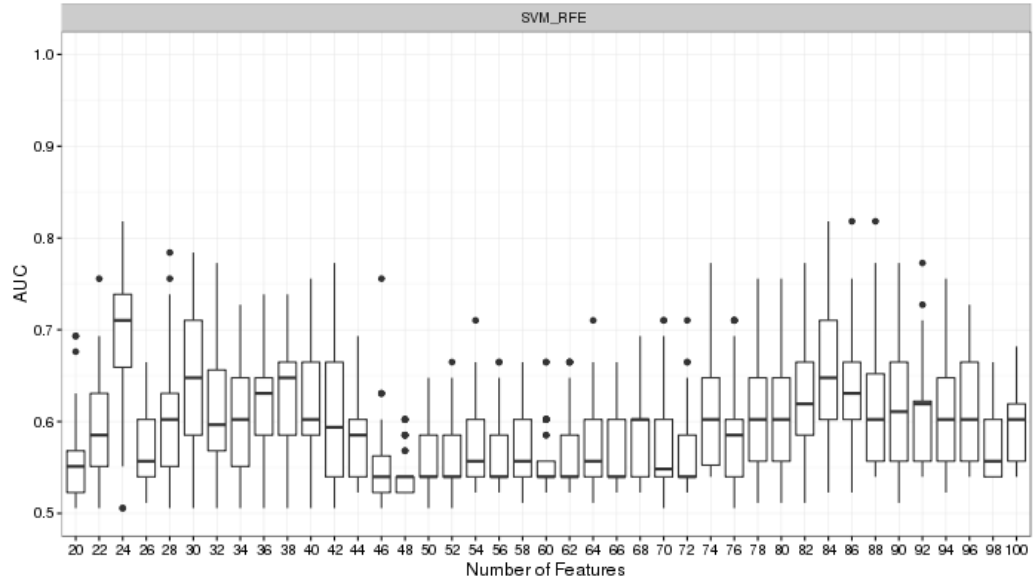


(b) Fuzzy Forest

Figure 4.12: Classification performance comparison between algorithms evaluated on the breast cancer dataset using the 0.632+ bootstrap method with 100 bootstrap samples across different number of features

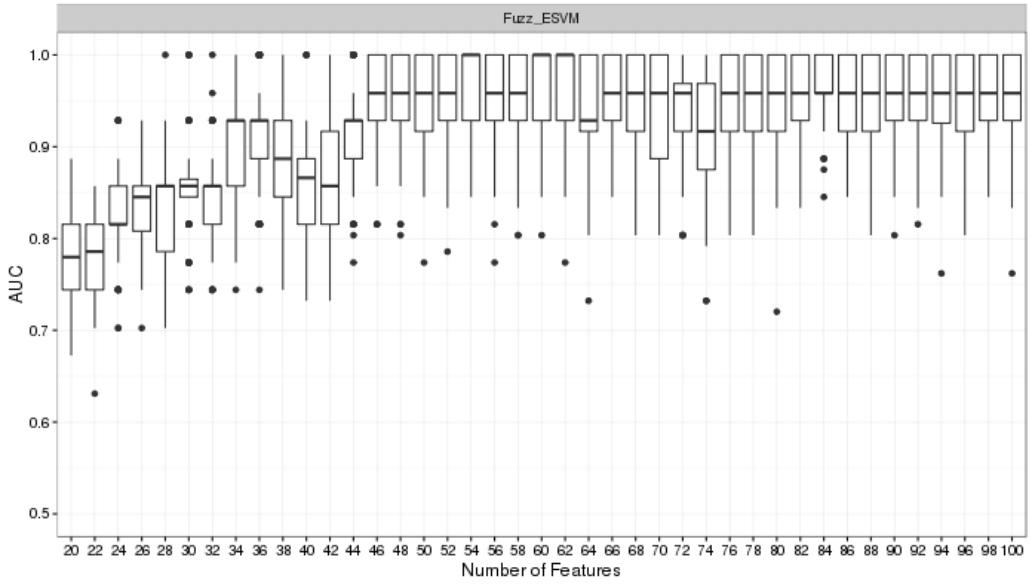


(a) HLR

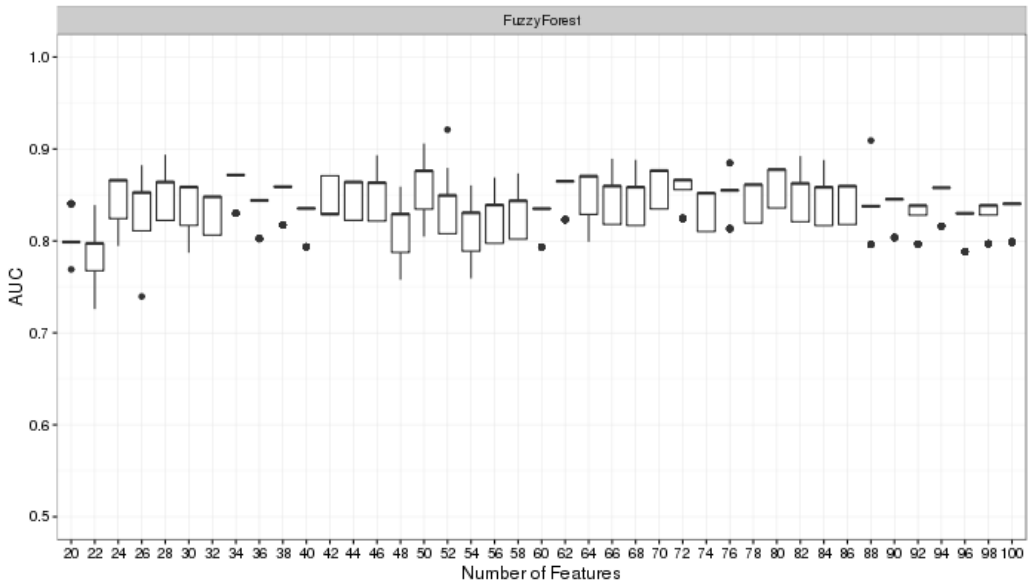


(b) SVM-RFE

Figure 4.13: Classification performance comparison between algorithms evaluated on the breast cancer dataset using the 0.632+ bootstrap method with 100 bootstrap samples across a different number of features (continuation of Figure 4.12)

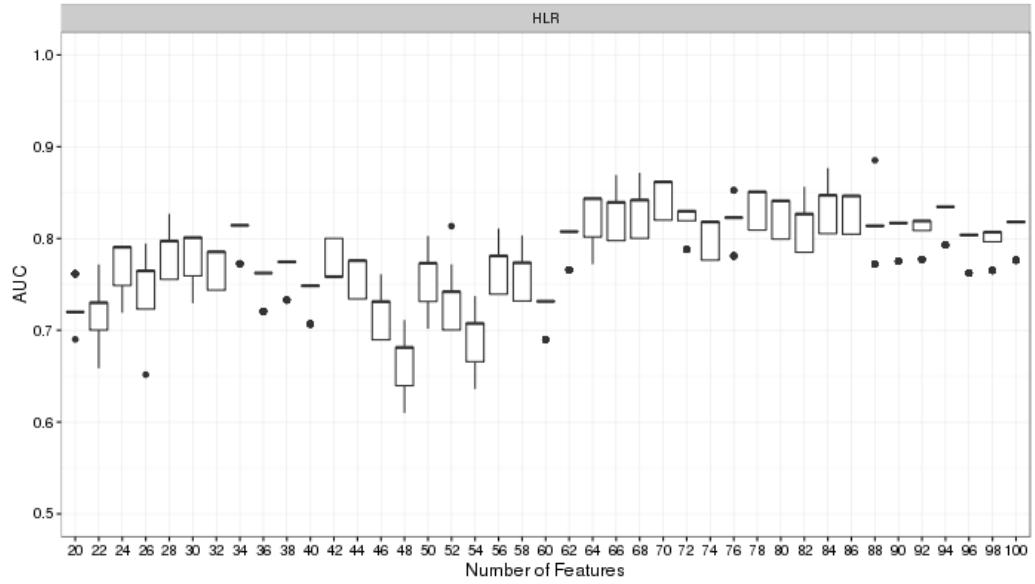


(a) Fuzz-ESVM

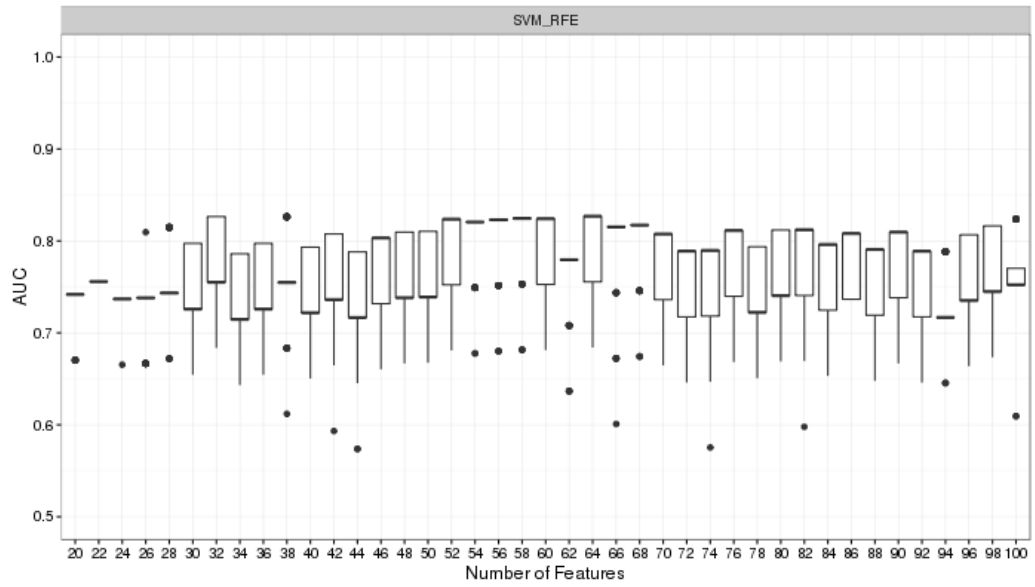


(b) Fuzzy Forest

Figure 4.14: Classification performance comparison between algorithms evaluated on the ALL/AML dataset using the 0.632+ bootstrap method with 100 bootstrap samples across a different number of features



(a) HLR



(b) SVM-RFE

Figure 4.15: Classification performance comparison between algorithms evaluated on the ALL/AML dataset using the 0.632+ bootstrap method with 100 bootstrap samples across a different number of features (continuation of Figure 4.14)

Table 4.6: The quartile and mean values of AUC accuracies of the compared algorithms on the evaluated datasets at the best number of features.

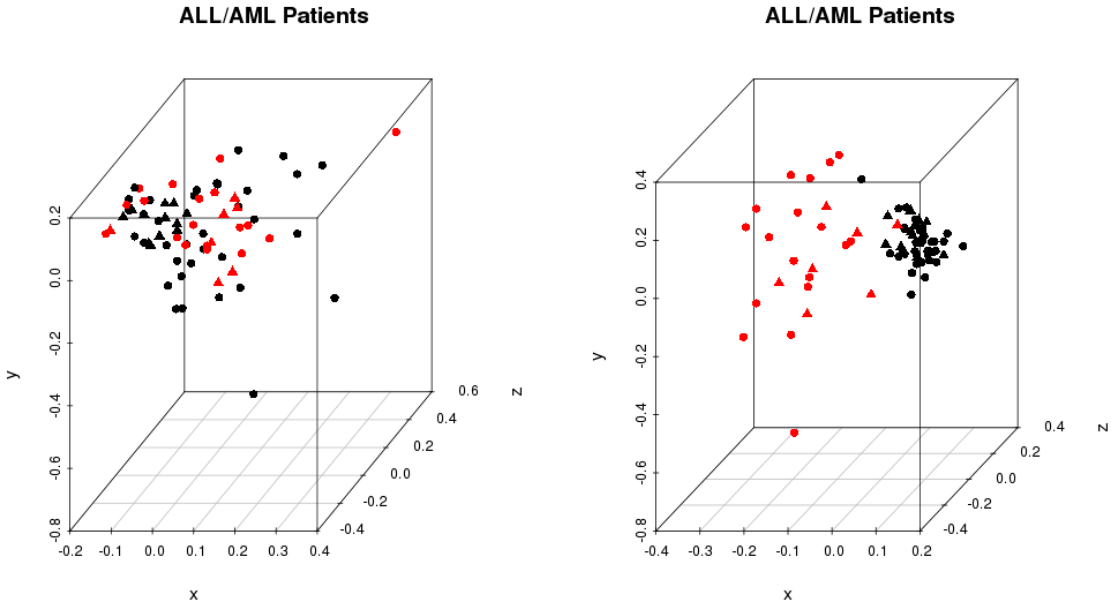
Dataset	Method	Min	1st Qu	Median	Mean	3rd Qu	Max	Best Features
Childhood leukaemia	SVM-RFE	0.250	0.400	0.500	0.509	0.600	0.800	64
	Fuzzy Forest	0.250	0.387	0.450	0.448	0.500	0.700	48
	HLR	0.081	0.331	0.381	0.370	0.431	0.681	30
	Fuzz-ESVM	**0.500**	**0.600**	**0.700**	**0.692**	**0.750**	**0.850**	24
DLBCL -FSCC	SVM-RFE	0.585	0.741	0.811	0.797	0.848	0.904	78
	Fuzzy Forest	0.647	0.743	0.810	0.800	0.854	0.897	74
	HLR	0.679	0.834	0.904	0.891	0.942	0.997	74
	Fuzz-ESVM	**0.833**	**0.900**	**0.966**	**0.939**	**1.00**	**1.00**	50
Prostate	SVM-RFE	0.562	0.601	0.639	0.639	0.678	0.716	64
	Fuzzy Forest	0.744	0.783	0.821	0.829	0.860	0.898	68
	HLR	0.690	0.730	0.730	0.745	0.769	0.807	74
	Fuzz-ESVM	**0.807**	**0.884**	**0.884**	**0.886**	**0.923**	**0.923**	34
Breast	SVM-RFE	0.505	0.659	0.710	0.701	0.738	0.818	24
	Fuzzy Forest	0.471	0.544	0.624	0.618	0.669	0.760	30
	HLR	0.462	0.570	0.633	0.618	0.678	0.741	70
	Fuzz-ESVM	**0.556**	**0.677**	**0.727**	**0.716**	**0.772**	0.818	30
ALL/AML	SVM-RFE	0.683	0.754	0.754	0.784	0.826	0.826	32
	Fuzzy Forest	0.804	0.834	0.876	0.859	0.876	0.905	50
	HLR	0.800	0.800	0.841	0.827	0.841	0.871	68
	Fuzz-ESVM	**0.815**	**0.928**	**0.958**	**0.953**	**1.00**	**1.00**	46

features and project the testing samples using the first three principal components. Then, the testing samples are visualised with different shapes for the ALL and AML samples. A similar process is applied on the training set with the top 46 features selected by the proposed Fuzz-ESVM feature selection algorithm. Without loss of generality, the top 46 features are used in these figures. As shown in Figure 4.16, it is clear that the clusters of the ALL and AML classes in Figure 4.16b are well separated compared to the clusters in Figure 4.16b, which overlap. This example confirms that efficacy of the proposed algorithm to select the optimal features in the presence of complex datasets which importantly, are able to explain the differences between different classes.

Case study: Application to Analyse of the Glioblastoma Dataset

This study demonstrates the biological significance of the proposed Fuzz-ESVM by analysing the microarray gene expression glioblastoma dataset (Nutt et al. 2003). Glioblastoma is a malignant brain tumor mainly located in the cerebral hemispheres of the brain. It is considered a highly cancerous tumor because it can grows rapidly. The average survival time of the patients from the time of diagnosis is 15 months. The main cause of glioblastoma is still unknown and treating this cancer is not trivial because it contains many types of cells.

The main goal of this analysis is to identify the genes and corresponding pathways associated with malignant glioblastoma. The glioblastoma dataset is generated by the Affymetrix human genome U95Av2 array (Nutt et al. 2003). It contains data for 50 patients with expression values for 12625 probes. The patients are divided into between four diagnostic classes: 14 samples are classic glioblastoma (CG), 7 samples



(a) SVD on ALL/AML dataset with all features

(b) SVD on ALL/AML dataset using the top 46 features selected by Fuzz-ESVM

Figure 4.16: SVD on ALL/AML dataset to show the clusters of ALL and AML patients. Black=ALL, red=AML, circle=training samples, and triangle=testing samples

are classic oligodendroglioma (CO), 14 samples are non-classic glioblastoma (NG), and the rest are non-classic oligodendroglioma (NO).

The Fuzz-ESVM algorithm is applied to identify the genes and the associated pathways which are related to malignant glioblastoma. Fuzz-ESVM identified the 50 genes connected by KEGG pathways. The results from the algorithm suggest that several pathways might be related to malignant glioblastoma. The largest subnetwork is composed of 12 genes involving ribosome pathways. Ribosome proteins have been reported in several types of human malignancies (Vaarala et al. 1998). Moreover, it is observed that increasing expressions for ribosome proteins is associated with poor patient survival (Vaarala et al. 1998). The detected ribosomes of glioblastoma are detailed in Table 4.7. Importantly, it is shown that the expression of the human S3 ribosomal gene is associated with colorectal cancer (Pogue-Geile et al. 1991). Furthermore, the expression of S6 ribosomal protein genes in leukaemia blast cells (Ferrari et al. 1990), and ribosomal protein L19 is a prognostic marker for human prostate cancer (Bee et al. 2006). Therefore, there is a suggestion that the ribosomal protein genes play a role in tumors. Some ribosomal proteins have a negative impact on cell proliferation such as L17 (Smolock et al. 2012). Moreover, Nagao-Kitamoto et al. (2015) observed that "S24 were identified to positively regulate migration of osteosarcoma cell, glioma cells and colorectal cancers cells".

The other small subnetworks of genes are involved in asthma, allograft rejection, autoimmune thyroid disease pathways. They contain genes mainly about the major histocompatibility complex genes. Zagzag et al. (2005) suggest that the fundamental feature of glioblastoma cells is the invasion surrounding brain tissue. They find in their study that decreasing expressions of major histocompatibility complex antigens

permit glioma cells to invade the surrounding brain (Zagzag et al. 2005). In summary, these results demonstrate that the proposed method can identify KEGG pathways which are potentially relevant to malignant glioblastoma. This argument is well supported by the previous studies.

Table 4.7: The correlated genes and KEGG pathways

Pathway	Probes	Gene Description
Ribosome	2016_s_at	Ribosomal protein L10
	32440_at	Ribosomal protein L17
	32435_at	Ribosomal protein L19
	32341_f_at	Ribosomal protein L23
	33485_at	Ribosomal protein L4
	33116_f_at	Ribosomal protein S12
	33619_at	Ribosomal protein S13
	347_s_at	Ribosomal protein S23
	32315_at	Ribosomal protein S24
	34570_at	Ribosomal protein S27
	1653_at	Ribosomal protein S3
	35125_at	Ribosomal protein S6
Asthma	37344_at	Major histocompatibility complex, class II, DM alpha
	41609_at	Major histocompatibility complex, class II, DM beta
	38833_at	Major histocompatibility complex, class II, DP alpha 1
	36773_f_at	Major histocompatibility complex, class II, DQ beta 1

4.3 Contribution and Summary

This chapter addresses **Contributions 4 and 5** of this thesis by developing a two novel algorithms to select the best features in the presence of highly correlated features. The first approach is supervised feature selection (SCANMF), which simultaneously combines supervised NMF with feature selection. The proposed SCANMF jointly integrates prior knowledge of the labels with handling the class imbalance

problem, and the correlation structure of the data. Furthermore, $l_{2,1}$-norm regularization is imposed to constrain the basis matrix to be sparse in rows and remove noisy features. Due to the nature of the proposed algorithm, the selected features are discriminative, non-biased, smooth, and can infer some semantic components due to using NMF in data decomposition and correlation network clustering. The proposed method can select discriminative features which are not biased to the majority class and have the potential to yield better interpretation results in its application. Extensive experiments are conducted on high-dimensional highly correlated imbalanced real-world datasets demonstrating promising results for the proposed algorithm. The second algorithm is a novel fuzzy ensemble feature ranking method which selects the optimal features in the presence of correlated features. It does not consider the correlated features as redundant, rather it selects the top correlated features from each feature module using the ESVM-RFE algorithm. Then, it aggregates all features from different modules and again applies ESVM-RFE to rank the combined features. The proposed algorithm can improve the classification accuracy in the presence of very complex datasets. These datasets contain high-dimensional, highly correlated features and a low number of samples. Extensive experiments are conducted on different datasets. The proposed algorithm outperformed the-state-of-the-art methods and a small study has been presented to interpret the selected features in the glioblastoma dataset.

Chapter 5

Multi-label Feature Learning on High-Dimensional Imbalanced Datasets

This chapter introduces two approaches: multi-label dimensionality reduction using correlation information and a multi-label classification model to handle missing labels and class imbalance. In machine learning, multi-label learning deals with instances associated with multiple target labels simultaneously, unlike single-label learning where each sample is assigned to one class or category. Multi-label learning is applicable in many real-world applications, such as bioinformatics (Barutcuoglu et al. 2006), information retrieval (Ueda & Saito 2002), image processing (Boutell et al. 2004) and others. For example, in information retrieval, the document can cover multiple topics, and in bioinformatics, a gene can be associated with multiple functions (Barutcuoglu et al. 2006). Multi-labeled data, similar to single-labeled data, suffer from the curse of dimensionality which leads to drawbacks, including increasing computational costs, increasing the complexity of the learning models, and deteriorating performance of the classifiers. In past decades, numerous single-label feature selection

methods have been proposed to decrease the number of dimensions of data by eliminating redundant and irrelevant features. However, these methods are not appropriate to apply to multi-labeled data. Label correlations have been thoroughly used in the multi-label learning studies in the last few years, and they significantly contribute to predicting labels in multi-label classification. However, integrating label correlations in multi-label learning may not effectively work if it does not take into account some important challenges such as incomplete and noisy label space and the imbalanced class problem. This chapter proposes two methods to address these problems: (1) a multi-label feature selection method in section 5.1 using correlation information; and (2) a multi-label classification method in section 5.2 which handles the incomplete label space and class imbalance. In each section, the proposed algorithm is validated with experiments, and comparing behaviour with benchmark algorithms.

This chapter encapsulates the **Contribution 6 and 7** of the thesis, and it is an extended version of my publications (Braytee, Liu, Catchpoole & Kennedy 2017) and "Correlated Multi-label Classification with Missing Labels and Class Imbalance" which under-review of journal of data mining and knowledge discovery.

5.1 Multi-label Feature Selection using Correlation Information

Building a multi-label feature selection algorithm is a non-trivial task due to several challenges. First, the labels in a multi-label learning problem may be correlated and interdependent, which is not in the case in traditional single-labeled data where the

classes are mutually exclusive. For example, genes are often grouped together in families or categories based on similar sequences or structures which have particular common functions. Also, in information retrieval, "political" and "election" documents or labels are more related than "sports" documents. In another example from image annotation, "road" and "car" are highly likely to appear together in the same image, while "shark" does not usually appear with the "road". Therefore, if any image is annotated with label "road", it has a high probability of being labeled "car" as well. Second, interaction among features greatly contributes to finding a shared space for correlated labels. Returning to the example above, correlated terms such as "party", "government" and "constitution" might be important in distinguishing the correlated labels "political" and "election" from others such as "sports" documents. Furthermore, in genomics, gene products often work in specific pathways or established mechanisms. One feature (i.e gene) may interact with multiple other features that link it to different pathways and upstream/ downstream mechanisms. Third, the assumption of global label correlations among all instances may not be correct in all cases. Mostly, the label correlations are shared only by a subset of instances. Returning to the image example above, if "road" and "car" labels are correlated in some instances, it is not correct to assume that "road" and "car" labels exist together in all instances. Last, in real applications, it is difficult to obtain complete label information for an instance. Moreover, it is common for an instance to contain flawed (noisy) labels due to human annotation errors (Xu et al. 2014).

Surprisingly, only relatively feature selection methods in multi-label learning, such as MIFS (Jian et al. 2016) and SFUS (Ma et al. 2012), take label correlations into consideration. However, the former does not accurately capture real label relations

to alleviate the negative effects of missing and flawed labels. Also, interpreting its low-dimensional space is impossible due to mixed signs vectors. The latter does not explicitly consider the label correlations in feature selection, and the other existing methods either transform the problem to single-label problems (Spolaôr et al. 2012, Tsoumakas et al. 2011), or propose feature extraction methods which change the physical meaning of the original data (Zhang & Zhou 2010). In this section, a Correlated- and Multi-label Feature Selection method (CMFS) is proposed in the context of NMF which is able to address the aforementioned challenges. Importantly, the proposed algorithm imposes a non-negativity constraints on the original and low-dimensional matrices, which facilitates meaningful interpretation of the results. In particular, the original data and label matrix are decomposed into low-dimensional reduced space, based on defining three kinds of interactions: label-label, label-feature and sample-label interaction. Typically, the low rank feature matrix generally finds the relevant features which capture the dependency among multiple labels. The low rank structure is able to capture the local label correlations based on the similarity between instances. Furthermore, the proposed method defines a new complete label matrix which is able to address the missing and flawed labels. Based on literature review in chapter 2, this is the first attempt to exploit the correlation information of labels, features and samples together to find the relevant features that are shared across multiple labels.

5.1.1 Definition of the Method

In multi-label learning, given a data matrix $X \in \mathbb{R}^{n \times d}$ where the rows are the samples, a sample x_i in X is associated with a set of output labels $y_i \subseteq \{1, 2, \cdots, l\}$, where y_i

is a binary vector of labels. For example, $y_i = \{2, 4\}$ means that x_i is labeled by the 2^{nd} and 4^{th} labels. For an arbitrary matrix $T \in \mathbb{R}^{n \times d}$, the $l_{2,1}$-norm is defined as

$\|T\|_{2,1} = \sum_{i=1}^{n} \sqrt{\sum_{j=1}^{d} T_{ij}^2} = 2\mathrm{Tr}\left[T^T D T\right]$ where t_i is the ith row vector of T, D is a diagonal matrix with $D_{ii} = 1/(2\left\|a^i\right\|_2 + \epsilon)$, and $\left\|a^i\right\|_2$ is very close to 0 and ϵ is a small positive constant (Nie et al. 2010). $\|T\|_F$ is the Frobenius norm of T and $\mathrm{Tr}(T)$ is the trace operation of matrix T if T is a square matrix.

The Correlated- and Multi-label Feature Selection Approach

The CMFS feature selection approach simultaneously incorporates in its objective function the following information: label correlations, feature interactions and sample similarities. The CMFS approach is considered a variant of NMF. Whereas NMF has been applied in many applications (Pascual-Montano et al. 2006, Braytee, Catchpoole, Kennedy & Liu 2016), the proposed approach simultaneously decomposes the original data and the multi-labeled data to a low-dimensional space. Hence, it alleviates the negative effect of noisy features and flawed labels due to human annotated labels.

5.1.1.0.1 Label and Feature Correlations The first part of CMFS method takes into consideration three kinds of interactions: 1) the interaction between the labels themselves in each sample known as label correlations; 2) the interaction between the features themselves to find typical feature combinations which have a semantic meaning based on its application; 3) the interaction between labels and features which is represented by the relation between the combined features and correlated labels. This interaction results in finding related shared features which are important to predict the correlated labels. Mathematically, CMFS decomposes the original data as well as the multi-labeled output space into non-negative three-factor decomposition.

The objective function optimizes

$$\begin{aligned} \min_{V,L,Q,P,B} \quad & \|X - VLQ\|_F^2 + \alpha \|Y - VPB\|_F^2 + \beta \|L - P\|_F^2 \\ \text{s.t} \quad & \{V, L, Q, P, B\} \geq 0 \end{aligned} \tag{5.1.1}$$

Motivated by an orthogonal non-negative matrix the tri-factorizations method (Ding et al. 2006), in the first component, $X \in \mathbb{R}_+^{n \times d}$, $V \in \mathbb{R}_+^{n \times c}$, $L \in \mathbb{R}_+^{c \times c'}$ and $Q \in \mathbb{R}_+^{c' \times d}$. In CMFS, $c = c'$ due to simultaneously cluster the rows and columns. So, the number of row clusters (c) is equal to the number of column clusters (c'). The factor L absorbs the different scales of matrices X, V, Q. The factor Q represents the feature combinations, and V is shared between two components, which represents the coefficients' low-dimensional space. The second component in Eq. 5.1.1 represents the decomposition of the multi-label output matrix into three factors: $Y \in \mathbb{R}_+^{n \times l}$, $P \in \mathbb{R}_+^{c \times c}$ and $B \in \mathbb{R}_+^{c \times l}$. The factor Y is the multi-label matrix for the training samples, and B represents the clusters of labels which defines the label correlations. The third component minimizes the differences between factors L and P, which results in capturing the combined prominent features for the correlated labels.

5.1.1.0.2 Incorporating Sample Similarities Incorporating interactions between the samples and the labels in the objective function results in the missing and flawed labels being handled in low-dimensional space. Also, it enhances the locality of the label correlations among the similar instances. This work is inspired by the study of Huang et al. (2012), that assumes that if samples x_i and x_j in X are strongly correlated, this means that they share the same subset of label correlations. Consequently, the local influence of the correlated labels is the same for all similar instances. Therefore, the proposed method is reformulated as

$$\begin{aligned} \min_{V,L,Q,P,B} \quad & \|X - VLQ\|_F^2 + \alpha \|Y - VPB\|_F^2 + \beta \|L - P\|_F^2 + \varepsilon \mathrm{Tr}(R(VPB)^T VPB) \\ \text{s.t} \quad & \{V, L, Q, P, B\} \geq 0 \end{aligned} \tag{5.1.2}$$

where R is called the graph Laplacian, $R = G - S$, $S_{ij} \in \mathbb{R}_+^{n \times n}$ represents the similarity matrix between samples x_i and x_j in feature space, and G is a diagonal matrix whose entries are the column sum of S. Several methods can be used to compute the similarity matrix S_{ij} such as cosine similarity. Therefore, based on the last component of Eq. 5.1.2, if the similarity between two samples x_i and x_j is large, this leads to very similar label vectors for $(VPB)_i$ and $(VPB)_j$, which share the same subset of labels. Also, the similarity between labels in B is large, which forms the clusters of labels or label correlations. Therefore, the factor $P \in \mathbb{R}_+^{c \times c}$ represents the local influence of label correlations in B based on the similar instances. Furthermore, the approximated multi-label matrix VPB has complete label information by accurately capturing the relation between samples and label correlations. In addition, $l_{2,1}$-norm regularization is imposed on feature matrix Q to ensure its sparsity on rows. The final objective function for CMFS method can be stated as

$$\begin{aligned} \min_{V,L,Q,P,B} \quad & \|X - VLQ\|_F^2 + \alpha \|Y - VPB\|_F^2 + \beta \|L - P\|_F^2 \\ & + \varepsilon \mathrm{Tr}(R(VPB)^T VPB) + \gamma \|Q\|_{2,1} \\ \text{s.t} \quad & \{V, L, Q, P, B\} \geq 0 \end{aligned} \tag{5.1.3}$$

Finally, CMFS algorithm sorts all the features in factorized matrix Q according to $\|q_i\|_2$ in descending order and returns the top ranked features.

5.1.1.0.3 CMFS Optimization Algorithm The objective function in Eq. 5.1.3 has non-negativity constraints on the original and factorized matrices. Lagrangian multipliers $\psi, \iota, \varphi, \rho, \varrho$ is introduced to enforce the constraints for V, L, Q, P, B, respectively, Eq. 5.1.3 can be reformulated to express the Lagrangian form as,

$$\begin{aligned}\mathcal{L} =&\mathrm{Tr}\left[(X - VLQ)(X - VLQ)^T\right] + \alpha\mathrm{Tr}\left[(Y - VPB)(Y - VPB)^T\right] \\ &+ \beta\mathrm{Tr}\left[(L - P)(L - P)^T\right] + \varepsilon\mathrm{Tr}(R(VPB)^T VPB) + \gamma\mathrm{Tr}\left(Q^T DQ\right) \\ &+ \mathrm{Tr}(\psi V) + \mathrm{Tr}(\iota L) + \mathrm{Tr}(\varphi Q) + \mathrm{Tr}(\rho P) + \mathrm{Tr}(\varrho B)\end{aligned} \tag{5.1.4}$$

where Eq. 5.1.4 uses the matrix properties $\mathrm{Tr}(UV) = \mathrm{Tr}(VU)$ and $\mathrm{Tr}\left(U^T\right) = \mathrm{Tr}(U)$.

The objective function in Eq. 5.1.3 is non-convex with respect to all V, L, Q, P, B together. Therefore, to optimize the objective function, an alternating projected gradient method is proposed, which iteratively updates one variable and fixes the others. Since the objective function is convex by updating one variable and fixing the others, the partial derivatives of $\mathcal{L}$ w.r.t $V, L, Q, P,$ and B are:

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial V} =& - XQ^T L - \alpha YB^T P + VLQQ^T L^T + \alpha VPBB^T P^T + \\ & \varepsilon RVPBB^T P^T + \psi \\ \frac{\partial\mathcal{L}}{\partial Q} =& - (VL)^T X + (V^T VLL^T + \gamma D)Q + \varphi \\ \frac{\partial\mathcal{L}}{\partial B} =& - \alpha(VP)^T Y + \alpha V^T VPP^T B + \varepsilon V^T RVPP^T B + \varrho \\ \frac{\partial\mathcal{L}}{\partial L} =& - Q(X^T V) - \beta P + LV^T VQQ^T + \beta L + \iota \\ \frac{\partial\mathcal{L}}{\partial P} =& - \alpha B(Y^T V) - \beta L + \alpha PV^T VBB^T + \beta P + \varepsilon V^T RVPBB^T + \rho\end{aligned} \tag{5.1.5}$$

The equations are formulated based on Karush–Kuhn–Tucker conditions where $\psi_{ik}v_{ik} = 0$, $\varphi_{jk}q_{jk} = 0$, $\iota_{kk}l_{kk} = 0$, $\rho_{kk}p_{kk} = 0$, and $\varrho_{mk}b_{mk} = 0$, then these iterative

algorithms are obtained as follows

$$
\begin{aligned}
v_{ik} &\leftarrow v_{ik}\frac{[XQ^TL+\alpha YB^TP+\varepsilon SVPBB^TP^T]_{ik}}{[VLQQ^TL^T+\alpha VPBB^TP^T+\varepsilon GVPBB^TP^T]_{ik}}\\
q_{jk} &\leftarrow q_{jk}\frac{[(VL)^TX]_{jk}}{[(V^TVLL^T+\gamma D)Q]_{jk}}\\
b_{mk} &\leftarrow b_{mk}\frac{[\alpha(VP)^TY+\varepsilon V^TSVPP^TB]_{mk}}{[\alpha V^TVPP^TB+\varepsilon V^TGVPP^TB]_{mk}}\\
l_{kk} &\leftarrow l_{kk}\frac{[Q(X^TV)+\beta P]_{kk}}{[LV^TVQQ^T+\beta L]_{kk}}\\
p_{kk} &\leftarrow p_{kk}\frac{[\alpha B(Y^TV)+\varepsilon V^TSVPBB^T+\beta L]_{kk}}{[\alpha PV^TVBB^T+\varepsilon V^TGVPBB^T+\beta P]_{kk}}
\end{aligned}
\tag{5.1.6}
$$

5.1.2 Experiments and Datasets

Experiments are conducted on ten high-dimensional multi-labeled datasets, which are available in the Mulan[1] and LAMDA[2] repositories. The characteristics of these datasets are summarized in Table 5.1. The main dataset used is an Affymetrix childhood leukaemia dataset which described in Section 3.3.3. To evaluate the effectiveness and efficiency of the selected features by the proposed and compared methods, a binary relevance model (BR) is used which decomposes a multi-label classification problem into several independent binary classification problems. The proposed CMFS intends to use the BR strategy because it does not take into consideration any label correlations, which demonstrates the strengths of the compared algorithms in addressing correlated multi-label learning. Linear SVM is used to learn the binary classifiers (SVM-BR). On each dataset, 5-fold cross validation is conducted and the mean value of the multi-label accuracy measures is recorded. The proposed method uses two common evaluation metrics to validate the performance of the methods i.e.

[1]http://mulan.sourceforge.net/datasets-mlc.html
[2]http://lamda.nju.edu.cn/Data.ashx

Table 5.1: Characteristics of the evaluated datasets

Dataset	#Features	#Labels	#Instances
Childhood Leukaemia	22277	2	183
RCV1V2 (S1)	23679	103	6000
RCV1V2 (S2)	23571	103	6000
RCV1V2 (S3)	23631	103	6000
Bibtex	1836	159	7395
Medical	1449	45	978
Genbase	1185	27	662
Enron	1001	53	1702
Image	294	5	2000
MIMLtext	243	7	7119

Micro-F1 and average precision (Tsoumakas et al. 2009). The evaluation metrics in multi-label classification are substantially different those applied in binary or multi-class classification.

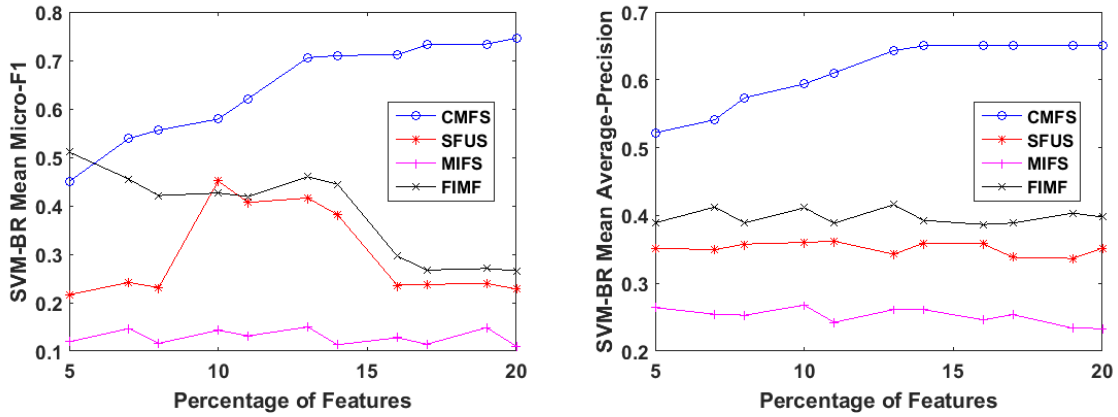
The proposed CMFS approach is compared to three state-of-the-art multi-label feature selection methods: Multi-Label Informed Feature Selection (MIFS) (Jian et al. 2016), Sub-Feature Uncovering with Sparsity (SFUS) (Ma et al. 2012) and Fast Multi-Label Feature Selection (FIMF) (Lee & Kim 2015). To ensure a fair comparison, the parameters of the methods are tuned by 5-fold cross validation and report their best results. The regularization parameters of the compared methods are tuned by searching the grid $\{10^{-2}, 0.1, 0.3, 0.5, 0.7, 0.9, 1\}$, and the approximate rank c is tuned by $\big\{2,\ 0.25|L|,\ 0.5|L|,\ 0.75|L|,\ |L|\big\}$, where $|L|$ is the total number of labels.

5.1.3 Results and Analysis

Experiments are conducted to evaluate the performance of the proposed CMFS algorithm compared to the state-of-the-art multi-label feature selection methods. The experimental results are reported in Tables 5.2, 5.3, Figure 5.1, and Figure 5.2.

The classification results are evaluated using Micro-F1 and average precision criteria, where the larger the values, the better the performance. As shown in Tables 5.2, 5.3, Figure 5.1 and 5.2, the classification results are generated based on the percentages between 5% to 20% of the top-ranked features. The results show that the performance of the classifier substantially improves by using CMFS compared to other methods. Generally, it is observed that CMFS algorithm performs better than the other compared algorithms on almost all evaluated datasets. Particularly, in Figure 5.1 and Figure 5.2, it is clearly shown that the proposed algorithm significantly outperforms the state-of-the-art methods on Childhood Leukaemia and RCV1V2 high-dimensional multi-labeled datasets. Furthermore, it can be observed that the classification results of CMFS tend to be stable after increasing the selected features above $10\% - 15\%$. This indicates the stability and capability of the proposed algorithm to select a lower percentage of features and realize better classification results. Therefore, these results demonstrate the great benefits of taking into account feature and local label correlations in the context of NMF to select the relevant features.

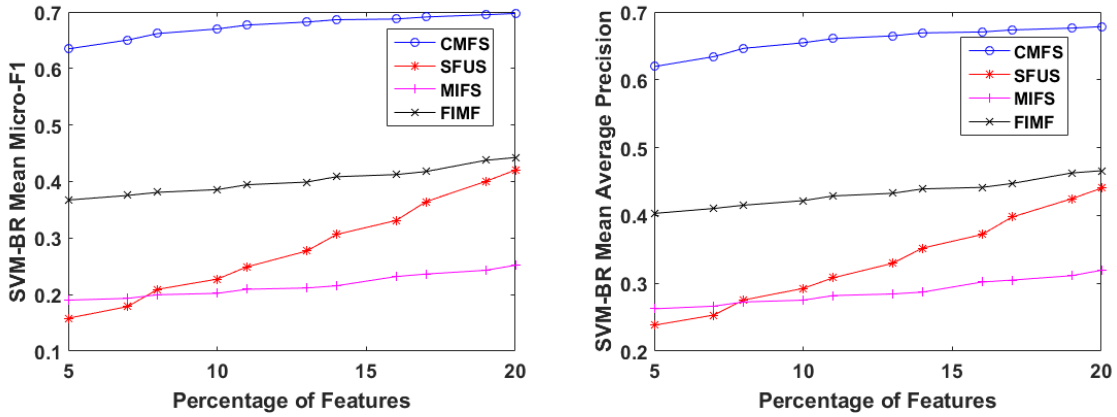
To perform an appropriate statistical comparative analysis, a Friedman test is conducted on the experimental results, which is a preferable statistical test between different algorithms over multiple datasets (Demšar 2006). As shown in Tables 5.2 and 5.3, CMFS algorithm achieves the best performance in most cases by investigating the ranks of the Friedman test. Furthermore, the p-values in most cases are less than 0.05 which rejects the null hypothesis that all the compared algorithms have an equal performance. Consequently, the Nemenyi test (Demšar 2006) is used to evaluate whether the proposed algorithm performs significantly differently. This can be achieved by checking if the average ranks of the compared algorithms differ by at



(a) Micro-F1 average (b) Average Precision average

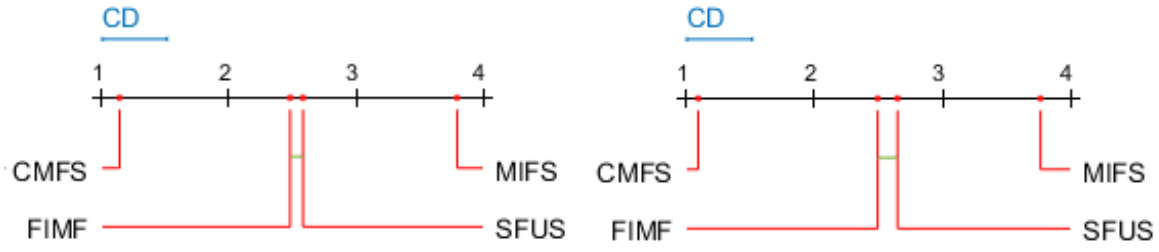
Figure 5.1: Comparison of four feature selection algorithms on Childhood Leukaemia dataset

least one critical difference (CD) over all the datasets. In Figure 5.3, any algorithm is significantly different if its rank is located outside the interval of CD. Figure 5.3, visually shows that CMFS algorithm performs greatly better than SFUS, FIMF and MIFS in both evaluation criteria. Moreover, it can be observed that SFUS and FIMF seem to have an equivalent performance. As a result, this confirms that the proposed method is superior across different types of datasets, different domains and different evaluation metrics from a statistical point of view.



(a) Micro-F1 average (b) Average Precision average

Figure 5.2: Comparison of four feature selection algorithms on the RCV1V2 (S1) dataset



(a) Micro-F1 (b) Average Precision

Figure 5.3: Comparison of all compared algorithms on each evaluation criteria using the statistical Nemenyi test.

Table 5.2: SVM-BR classification results of each compared algorithm (mean) on the multi-labeled datasets. The ranks in the parentheses are computed by the Friedman test to identify the best performing algorithm.

Micro-F1 evaluation criterion									
		Multi-Labeled Datasets							
% of top features	Algorithm	RCV1V2 (S2)	RCV1V2 (S3)	Bibtex	Medical	GenBase	Enron	Image	MIMLtext
5%	CMFS	**0.54 (1)**	**0.52 (1)**	**0.56 (1)**	**0.73 (1.4)**	**0.99 (1.4)**	**0.63 (1)**	**0.50 (1)**	0.45 (2.4)
	SFUS	0.14 (4)	0.17 (4)	0.49 (2)	0.66 (2.6)	0.98 (2.4)	0.46 (3.4)	0.41 (2.4)	0.50 (1.8)
	MIFS	0.26 (3)	0.25 (2.6)	0.40 (4)	0.02 (4)	0.05 (4)	0.45 (3.6)	0.29 (3.8)	0.24 (4)
	FIMF	0.35 (2)	0.26 (2.4)	0.46 (3)	0.69 (2)	0.98 (2.2)	0.55 (2)	0.36 (2.8)	**0.51 (1.8)**
	P-value	0.001	0.003	0.001	0.013	0.013	0.003	0.007	0.021
10%	CMFS	**0.62 (1)**	**0.58 (1)**	**0.62 (1)**	**0.79 (1)**	**0.99** (1.8)	**0.66 (1)**	**0.57 (1)**	**0.67 (1.4)**
	SFUS	0.19 (4)	0.23 (3.6)	0.56 (2)	0.70 (2.4)	0.99 **(1.4)**	0.44 (3.2)	0.45 (2.2)	0.63 (2.6)
	MIFS	0.32 (2.8)	0.30 (2.6)	0.46 (4)	0.05 (4)	0.05 (4)	0.41 (3.8)	0.36 (3.8)	0.35 (4)
	FIMF	0.36 (2.2)	0.29 (2.8)	0.51 (3)	0.68 (2.6)	0.98 (2.8)	0.54 (2)	0.40 (3)	0.66 (2)
	P-value	0.002	0.013	0.001	0.003	0.007	0.002	0.005	0.010
15%	CMFS	**0.64 (1)**	**0.62 (1)**	**0.65 (1)**	**0.80 (1)**	**0.99** (1.8)	**0.68 (1)**	**0.59 (1.2)**	**0.76 (1.2)**
	SFUS	0.30 (3.8)	0.31 (2.8)	0.59 (2)	0.73 (2)	0.99 **(1.4)**	0.48 (2.8)	0.49 (1.8)	0.69 (2.8)
	MIFS	0.37 (2.8)	0.34 (2.8)	0.51 (3.8)	0.08 (4)	0.05 (4)	0.41 (4)	0.43 (3.6)	0.40 (4)
	FIMF	0.38 (2.4)	0.28 (3.4)	0.55 (3.2)	0.67 (3)	0.98 (2.8)	0.52 (2.2)	0.42 (3.4)	0.72 (2)
	P-value	0.007	0.021	0.002	0.001	0.007	0.002	0.005	0.005
20%	CMFS	**0.67 (1)**	**0.67 (1)**	**0.66 (1)**	**0.80 (1)**	**0.99** (1.8)	**0.69 (1)**	**0.62 (1.2)**	**0.82 (1)**
	SFUS	0.38 (3.8)	0.36 (2.8)	0.60 (2)	0.75 (2)	0.99 **(1.4)**	0.49 (2.8)	0.54 (1.8)	0.72 (3)
	MIFS	0.41 (2.8)	0.37 (2.8)	0.55 (3.8)	0.10 (4)	0.05 (4)	0.44 (4)	0.47 (3.6)	0.44 (4)
	FIMF	0.42 (2.4)	0.32 (3.4)	0.57 (3.2)	0.68 (3)	0.98 (2.8)	0.51 (2.2)	0.43 (3.4)	0.75 (2)
	P-value	0.007	0.021	0.002	0.001	0.007	0.002	0.005	0.001

Table 5.3: SVM-BR classification results of each compared algorithm (mean) on the multi-labeled datasets. The ranks in the parentheses are computed by the Friedman test to identify the best performing algorithm.

Average precision evaluation criterion									
		Multi-Labeled Datasets							
% of top features	Algorithm	RCV1V2 (S2)	RCV1V2 (S3)	Bibtex	Medical	GenBase	Enron	Image	MIMLtext
5%	CMFS	**0.54 (1)**	**0.52 (1)**	**0.49 (1)**	**0.72 (1.6)**	**0.99 (1.6)**	**0.53 (1)**	**0.70 (1)**	0.61 (2.2)
	SFUS	0.21 (4)	0.25 (3.8)	0.45 (2)	0.65 (2.4)	0.98 (2.4)	0.37 (3.4)	0.60 (2.4)	0.62 (2)
	MIFS	0.31 (3)	0.30 (2.8)	0.40 (4)	0.12 (4)	0.11 (4)	0.36 (3.6)	0.53 (3.8)	0.43 (4)
	FIMF	0.39 (2)	0.30 (2.4)	0.42 (3)	0.69 (2)	0.98 (2)	0.48 (2)	0.58 (2.8)	**0.64 (1.8)**
	P-value	0.001	0.007	0.001	0.018	0.018	0.003	0.007	0.026
10%	CMFS	**0.61 (1)**	**0.58 (1)**	**0.55 (1)**	**0.78 (1)**	**0.99 (1.8)**	**0.58 (1)**	**0.75 (1)**	**0.77 (1.4)**
	SFUS	0.26 (4)	0.29 (3.6)	0.52 (2)	0.68 (2.6)	0.99 (1.8)	0.36 (3.2)	0.64 (2.6)	0.71 (2.6)
	MIFS	0.36 (3)	0.34 (2.6)	0.44 (4)	0.13 (4)	0.11 (4)	0.33 (3.8)	0.58 (3.6)	0.53 (4)
	FIMF	0.41 (2)	0.32 (2.8)	0.46 (3)	0.68 (2.4)	0.98 (2.4)	0.47 (2)	0.60 (2.8)	0.76 (2)
	P-value	0.001	0.013	0.001	0.003	0.021	0.002	0.013	0.010
15%	CMFS	**0.63 (1)**	**0.62 (1)**	**0.58 (1)**	**0.79 (1.2)**	**0.99 (1.8)**	**0.61 (1)**	**0.76 (1)**	**0.84 (1)**
	SFUS	0.35 (3.6)	0.35 (2.8)	0.56 (2)	0.72 (1.8)	0.99 (1.8)	0.42 (2.6)	0.66 (2)	0.76 (2.8)
	MIFS	0.40 (2.6)	0.37 (2.6)	0.48 (3.8)	0.15 (4)	0.11 (4)	0.32 (4)	0.61 (3.6)	0.56 (4)
	FIMF	0.39 (2.8)	0.31 (3.6)	0.50 (3.2)	0.68 (3)	0.98 (2.4)	0.45 (2.4)	0.62 (3.4)	0.79 (2.2)
	P-value	0.013	0.013	0.002	0.002	0.021	0.003	0.003	0.002
20%	CMFS	**0.65 (1)**	**0.66 (1)**	**0.60 (1)**	**0.79 (1)**	**0.99 (1.8)**	**0.63 (1)**	**0.78 (1)**	**0.89 (1)**
	SFUS	0.41 (3.6)	0.39 (2.6)	0.57 (2)	0.75 (2)	0.99 (1.8)	0.42 (3)	0.70 (2)	0.79 (3)
	MIFS	0.43 (3)	0.40 (2.8)	0.51 (3.8)	0.15 (4)	0.11 (4)	0.36 (3.8)	0.65 (3.4)	0.59 (4)
	FIMF	0.45 (2.4)	0.35 (3.6)	0.52 (3.2)	0.69 (3)	0.98 (2.4)	0.46 (2.2)	0.62 (3.6)	0.82 (2)
	P-value	0.010	0.013	0.002	0.001	0.021	0.005	0.003	0.001

Parameter Sensitivity and Convergence Analysis

The proposed CMFS approach has four important parameters: α, ε, γ and c. These parameters control the contribution of the following components: label correlation, sample similarities to find local labels, sparseness of the model, and approximate rank of the NMF algorithm, respectively. This section investigates the impact of these parameters on the classification accuracy of the proposed algorithm. A several experiments are conducted on the Enron dataset using the average precision metric. First, parameters α, ε and γ are tuned using a grid search from $\{10^{-2}, 0.1, 0.3, 0.5, 0.7, 0.9, 1\}$. The parameters are tuned iteratively by optimizing one parameter and fixing the others at a certain time. The common observation from the results in Figure 5.5a, 5.5b and 5.5c indicates that the classification results are not very sensitive to changes in the values of parameters α, ε and γ. The best classification performance was achieved when $\alpha = 0.1$, $\varepsilon = 0.3$ and $\gamma = 0.3$. Finally, this thesis explores the effect of rank c. As shown in Figure 5.5d, small c achieves better results than large c in the Enron dataset, and particularly when $c = 2$.

The convergence curves of CMFS are shown in Figure 5.4. It can be observed that the optimization function of CMFS converges rapidly and has almost 20 iterations. This demonstrates the efficacy and the speed of the proposed algorithm.

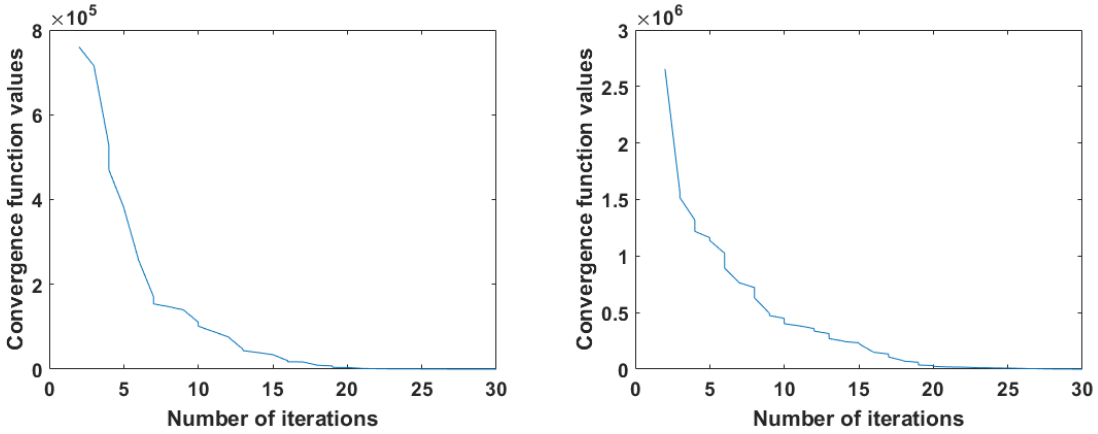
Computational Time Analysis

Typically, the computational complexity of the feature selection methods in single or multi-label high-dimensional data is high (Liu & Motoda 2007). To demonstrate the efficiency of the proposed CMFS method, the computation time of each method is reported on all datasets in Table 5.8. This is the time taken by each method

Table 5.4: The computation time (sec) of different algorithms on the high-dimensional datasets

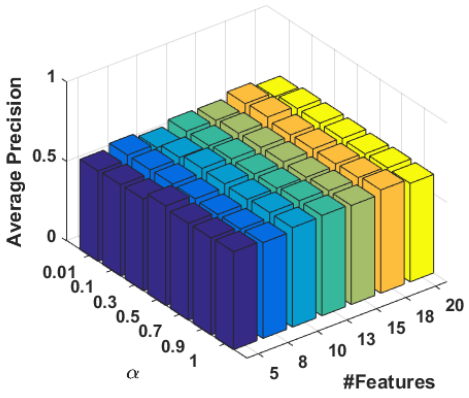
Dataset	CMFS	SFUS	MIFS	FIMF
Childhood Leukaemia	799	4689	1356	3956
RCV1V2 (S1)	827	3105	1460	4899
RCV1V2 (S2)	839	3256	1501	5122
RCV1V2 (S3)	864	3355	1472	5312

to converge to the optimal solution. The experiments are conducted on five core Linux server with each core at 3.10-GHZ, and a total memory of 96-GB. The times recorded in Table 5.8 show the optimal parameter setting in each method. It can be observed that the proposed algorithm is the fastest compared to the state-of-the-art algorithms on the most datasets. This indicates the efficiency of the proposed optimization method to converge quickly to the optimal solution.

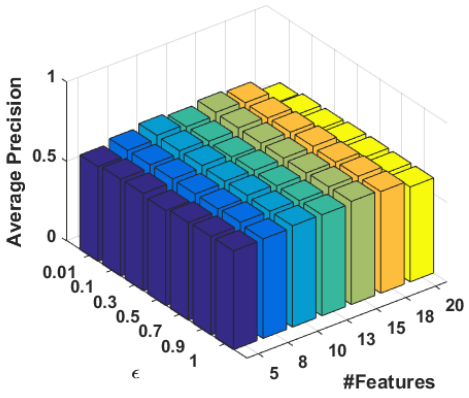


(a) Enron dataset (b) MIMLtext dataset

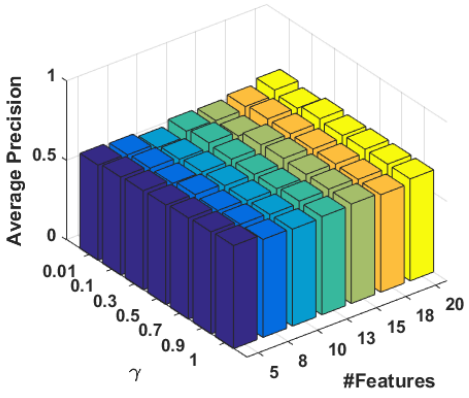
Figure 5.4: Convergence curves of CMFS algorithm



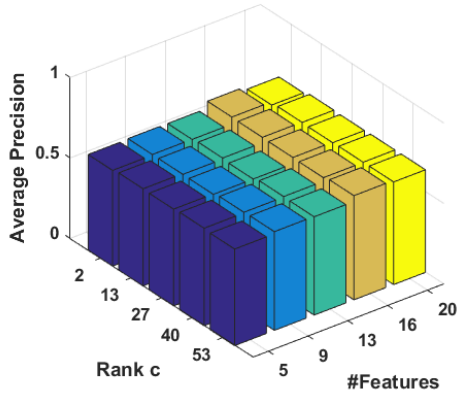
(a) Optimize α parameter



(b) Optimize ε parameter



(c) Optimize γ parameter



(d) Optimize rank c parameter

Figure 5.5: The average precision results on CMFS on the Enron dataset w.r.t different parameters

5.2 Handling Class Imbalance and Incomplete Label Space in Multi-Label Classification

In multi-label classification (MLC), the correlated labels are exploited to improve the label inference and assignment (Zhang & Zhou 2014, Huang, Li, Huang & Wu 2016). For instance, in image annotation as shown in Figure 5.6, it is not a simple task to predict the label of the image whether it is a "sea turtle" or "land turtle" because they both look very similar. However, by exploiting the correlation information between the labels, if the image will be tagged as "fish", then it is most probable that the "turtle" in Figure 5.6a will be annotated as "sea turtle", and similar case for Figure 5.6b, if the image is annotated as "rabbit", then the "turtle" in Figure 5.6b will be annotated as "land turtle". Label correlations have been thoroughly used in multi-label learning studies in the last few years which significantly contributes to predict labels in MLC. However, integrating label correlations in multi-label learning may not effectively work if MLC method does not take into account some important challenges such as incomplete and noisy label space, and class imbalance problem.

In real applications, it is particularly difficult to find a complete label vector for each instance due to the unavailability of all labels in some applications. Furthermore, the label matrix may contain noisy labels during manual labeling. Therefore, constructing a label matrix correlation from an incomplete and noisy label matrix is not appropriate and may not represent the real dependencies between the labels. The other important challenge is the structure of the label matrix as Boolean data. Several previous studies have applied existing similarity metrics, including Manhattan distance and Euclidean distance which are mainly built for numerical data (Gu

(a) Sea turtle (b) Land turtle

Figure 5.6: Using label correlations to predict the type of the turtle in image annotations

et al. 2011, Huang, Li, Huang & Wu 2016). Using existing categorical similarity metrics such as overlap or frequency based measure is too simplistic for capturing the dependencies between the categorical labels. Therefore, proposing a new predictive numerical label matrix has twofold benefit. First, the method converts the categorical data to numerical data which specifically defines the contribution of the label in the example. For instance, in Figure 5.6, the turtle in Figure 5.6a is bigger than the turtle in Figure 5.6b, and this will be reflected as a continuous value in the example. In the case of categorical labels, however, the label shows whether present or absent in the example. Second, it generally captures the global correlations between all labels in the new complete label matrix.

The most challenging case in MLC is the class imbalance problem. In standard binary and multi-class learning models, the class imbalance is ubiquitous in real machine learning applications, which degrades classification accuracy (Prati et al. 2015). The class imbalance issue is even greater in MLC due to the existence of the incomplete label matrix. Typically, there are two different types of class imbalance in

MLC, which are: instance-class imbalance and label-class imbalance. In the first, the number of instances assigned a positive label is much less than the number assigned a negative label, which is similar to the imbalance problem in binary classification. The second is the label-class imbalance problem in which the number of labels assigned to different instances is significantly disparate. Surprisingly, the class imbalance issue has not been extensively studied in the MLC context (Wu et al. 2016). Finally, in the context of multi-label feature selection, the class imbalance problem and incomplete label matrix lead to the selection of biased features to the majority class label, which degrades the performance of the MLC.

In this section, an integrated approach (ML-CIB) is proposed which addresses the three aforementioned challenges: the class imbalance problem, the incomplete label matrix and the generation label correlations from a multi-label Boolean matrix. This approach learns a new label matrix along with capturing the new label correlations, while simultaneously training the multi-label model. In addition, l_1 regularization norm is imposed on the feature matrix to retain the relevant features. A new continuous label matrix is constructed by considering three factors: (a) the new label matrix should be consistent with the original label matrix, (b) handling the imbalanced class problem in the original label matrix, and (c) the new label matrix uses the semantic instance assumption which means if the two instances x_i and x_j are similar, then their labels are correlated. Motivated by work in Wang et al. (2008), the proposed algorithm proposes a label regularization to handle the imbalanced label issue in the new predicted label matrix. The new predicted label matrix captures richer information than the original matrix because it is complete and handles the imbalanced class problem. An integrated approach is formulated to combine these

components as a constrained optimization problem. The proposed objective function is non-smooth due to the existence of the l_1 norm regularization term. Therefore, this study uses the accelerated proximal gradient method to solve the proposed optimization problem, and present a variant of a proposed method ML-CIB-FS to achieve a supervised feature selection on high-dimensional multi-labeled datasets, taking into account the imbalanced class problem and incomplete label matrix. It is considered as a first attempt to handle the imbalanced class problem in the multi-label feature selection approach. Consequently, experiments are conducted on regular and large-scale imbalanced multi-labeled datasets, which clearly verify that the results of the proposed approach favourably outperform the state-of-the-art algorithms.

5.2.1 Definition of the Method

This section presents a multi-label classification method which simultaneously estimates the label-specific features W using linear regression by taking into consideration the accurate label correlations and class imbalance problem. The proposed ML-CIB is optimized using the accelerated approximal optimization method.

Inferring the Accurate Label Correlation

Previous studies, as shown in Section 2.3, have observed that if the labels in MLC are strongly correlated, this indicates that they share a common discriminative features, which is significant for improving the classification accuracy of MLC models. However, it is not appropriate to capture the label correlations from incomplete, incorrect and flawed labels. Moreover, the original label space has logical labels, which restricts the existing MLC algorithms from capturing the accurate label correlations. To tackle

this problem, a new label matrix $\hat{Y} \in \mathbb{R}^{n \times l}$ is proposed, estimated during the training process, in which n is the number of instances and l are the labels. The new label space $\hat{Y}$ is extended to Euclidean space, where the Boolean label vector in each instance is changed to a real value. The new label space has several benefits, such as allowing the discovery of more semantic information about the labels, precisely depicting the contribution of the labels in each instance by using numerical values, and capturing accurate label correlations by exploring the label manifold using the traditional similarity metrics in Euclidean space.

The new label matrix is predicted using two assumptions. First, the smoothness assumption on the instance level (Chapelle et al. 2009), which assumes that instances close to each other are highly likely to have similar labels. Consequently, the local topological structure is extracted of the original data matrix $X \in \mathbb{R}^{n \times d}$ in the feature space by computing the similarity matrix $L \in \mathbb{R}^{n \times n}$, n and d are the number of samples and features in X respectively.

$$L = XX^T \tag{5.2.1}$$

Computationally, each label $\hat{y}_i$ can be optimally constructed using the topological structure of X, the value of $\hat{Y}_i^j \neq 0$ if x_i and x_j are similar. The approximation of the $\hat{Y}$ matrix can be solved by minimizing

$$\min_{\hat{Y}_i} \hat{Y}_i^T L_i \hat{Y}_i$$

The other factor in constructing the new label matrix is label consistency, since this thesis assumes that the proposed new label matrix $\hat{Y}$ must be consistent with the original label matrix $Y \in \mathbb{R}^{n \times l}$ by controlling the difference between the two matrices $\hat{Y}$ and Y. In the proposed method ML-CIB, the least square loss function is adopted

to learn the label-specific features W by inducing sparsity on W using l_1 regularization norm. Interestingly, the features in W are influenced by more accurate correlations between all labels, which contributes to finding better predictive features for MLC.

The assumptions above are formulated as

$$\min_{W,\hat{Y}} F(W,\hat{Y}) = \frac{1}{2}\left\|XW-\hat{Y}\right\|_F^2 + \alpha\hat{Y}^T L\hat{Y} + \beta\left\|\hat{Y}-Y\right\|_F^2 + \gamma\left\|W\right\|_1 \qquad (5.2.2)$$

Label Regularization for Class Imbalance

The class imbalance problem is considered to be a significant challenge in multi-label learning which degrades the classification accuracy. The original label matrix Y consists of j-th label vectors $\hat{y}_j$ which are assigned to the instances in the training data X, and the number of negative instances in each j-th label vector is greatly larger than the number of positive instances. In case of linear regression, for instance, the feature values of W in Eq. 5.2.2 are influenced by the incorporation of the label correlation as noted above. Therefore, the features generated in W will be biased towards predicting the majority labels due to the class imbalance problem among the instances in each binary label vector $\hat{y}_j$. The proposed ML-CIB can be used as a feature selection method by imposing l_1-norm regularization term on the feature matrix W. The values in W can be used to rank the features, and the largest value features will be selected to train the j-th binary classifiers for each label vector $\hat{y}_j$. Thus, integrating the label correlations without regularizing the binary class imbalance problem in each classifier will degrade the classification accuracy of the generated classifiers.

Motivated by the work in Wang et al. (2008), the matrix $V \in \mathbb{R}^{l\times l}$ is introduced as a label regularization to control the influence of the majority labels on the instances. The matrix V is constructed from the current label matrix Y as

$$V_{ij} = \frac{\sum_{i,j \in Y}^{l} I_{Y_i}(Y_j)}{n} \tag{5.2.3}$$

where

$$I_{Y_i}(Y_j) = \begin{cases} 1 & \text{if } Y_i = Y_i \\ 0 & \text{if } Y_i \neq Y_i \end{cases} \tag{5.2.4}$$

The proposed label regularization is used to normalize the new label matrix $\hat{Y}$, which is defined as $\hat{Y} = YV$. The higher values of the label regularization matrix make a larger contribution to the construction of the new label matrix $\hat{Y}$, and in estimating the balanced predictive features in W. The objective function is modified as follows

$$\min_{W,\hat{Y}} F(W,\hat{Y}) = \frac{1}{2}\left\|XW - \hat{Y}\right\|_F^2 + \alpha \hat{Y}^T L \hat{Y} + \beta \left\|\hat{Y} - YV\right\|_F^2 + \gamma \|W\|_1 \tag{5.2.5}$$

By optimizing Eq. 5.2.5, ML-CIB jointly trains a multi-label classification model and exploits accurate label correlations through a new predicted label matrix by exploring the label manifold and controlling the class imbalance problem.

Predicting Labels for Testing Samples

The proposed ML-CIB method predicts the labels of the testing samples using a linear regression method based on a threshold τ without using a specific classifier. Assume a test data $X_{ts} \in \mathbb{R}^{k \times d}$, where k is the number of testing samples. The score matrix $S \in \mathbb{R}^{k \times l}$ is computed for the test samples by $S = X_{ts}W$, and the proposed method classifies the test samples by proposing $Y_{ts} \in \mathbb{R}^{k \times l}$ as follows

$$Y_{ts} = \begin{cases} 0 & \text{if} S < \tau \\ 1 & \text{otherwise} \end{cases}$$

Multi-label Feature Selection for Class Imbalance and Incomplete Label Space (ML-CIB-FS)

In machine learning, feature selection and feature extraction techniques are affected by the class imbalance problem. Feature selection including the wrapper and embedded strategies, gives more weight to the features that predict the majority class, because they depend on the performance of the classification models, which are sensitive to the class imbalance problem. Feature extraction techniques such as NMF and principal component analysis generate biased features that favor the majority class (Braytee, Liu & Kennedy 2016). This subsection proposes the first multi-label feature selection algorithm (ML-CIB-FS) to handle the multi-label imbalance problem and incomplete label matrix. ML-CIB-FS is a variant of the proposed ML-CIB method. It sorts all the features in factorized matrix W according to $\|w_i\|_2$ in descending order and returns the top ranked features.

5.2.2 Optimization Procedure for ML-CIB

The minimization problem in Eq. 5.2.5 is non-convex. The objective function is partially convex with respect to each variable by fixing the other variables. It alternately minimizes F w.r.t W and $\hat{Y}$, one variable at a time while fixing the other variable at the most recent value. Furthermore, the objective function is non-smooth due to the presence of l_1 norm regularization. This study uses the accelerated proximal gradient (APG) method to solve the objective function due to two reasons. First, l_1 regularization norm makes the optimization function non-smooth. Second, the APG method generally has a faster convergence than other methods such as alternating non-negative least square methods (ANLS) (Beck & Teboulle 2009).

The proximal gradient method as proposed by Nesterov (2005) is formulated as follows. Consider the following convex optimization problem

$$\min_{a \in \chi} F(a) = f(a) \tag{5.2.6}$$

where χ is a convex set and f is a smooth function with Lipschitz continuous gradient L_f,

$$\|\nabla f(a) - \nabla f(b)\|_2 \leq L_f \|a - b\|_2, \text{ for all } a, b \in \chi$$

The projected gradient method for Eq. 5.2.6 is

$$a^{t+1} = \left(a^t - \alpha_t \nabla f(a^t)\right)$$

where α_t is the step size at specific iteration t, and the new iteration a^{t+1} is a minimizer of the linearised proximal regularization of function f at point a^t,

$$a^{t+1} = \min_{x \in \chi} f(a^t) + \left\langle \nabla f(a^t), a - a^t \right\rangle + \frac{L_f}{2} \left\| a - a^t \right\|_2^2$$

. The iteration is updated

$$a^{t+1} = \min_{x \in \chi} f(b^t) + \left\langle \nabla f(b^t), a - b^t \right\rangle + \frac{L_f}{2} \left\| a - b^t \right\|_2^2 \tag{5.2.7}$$

where b^t is a point at which f is linearised. The closed form of Eq. 5.2.7 is

$$a^{t+1} = \left(b^t - \frac{1}{L_f} \nabla f(b^t)\right)$$

This method is guaranteed to solve Eq. 5.2.6 with the appropriate value of b^t. The optimal solution is achieved by letting $a^t = b^t$.

The objective function of the proposed method in Eq. 5.2.5 is solved as described by Nesterov (2005). It updates both the W-subproblem and the $\hat{Y}$-subproblem at each iteration by

$$W^{t+1} = W^t - \frac{1}{L_{W^t}} \nabla_W F(W^t, \hat{Y}^t) + g(W)$$
$$\hat{Y}^{t+1} = \hat{Y}^t - \frac{1}{L_{\hat{Y}^t}} \nabla_{\hat{Y}} F(W^{t+1}, \hat{Y}^t)$$

where L_{W^t} and $L_{\hat{Y}^t}$ are Lipschitz constants of gradient $\nabla_W F(W, \hat{Y}^t)$ and $\nabla_{\hat{Y}} F(W^{t+1}, \hat{Y})$ by F as a function w.r.t W and $\hat{Y}$ respectively, $g(W) = \gamma \|W\|_1$. In the algorithm, $L_{W^t} = XX^T$ and $L_{\hat{Y}^t} = I + \alpha L + \beta I$.

The proximal gradient method achieves a promising convergence result, but it converges relatively slowly. The Nesterov-type acceleration technique is used to speed up the algorithm (Nesterov 2005). The acceleration technique defines a positive extrapolation weight ω. The idea behind this technique is if the current iteration of a^t approaches an optimal point a^* than the previous iteration a^{t-1}, then ω_t is closer to a^* than a^t. ω is defined as $\omega_t = \frac{k_{t-1}-1}{k_t}$, where $k_t = \frac{(1+\sqrt{1+4k_{t-1}^2})}{2}$. Therefore, the setting of subproblems W and $\hat{Y}$ is changed as follows

$$Z_W^t = W^t + \omega_t (W^t - W^{t-1})$$
$$Z_{\hat{Y}}^t = \hat{Y}^t + \omega_t (\hat{Y}^t - \hat{Y}^{t-1})$$

The updating rules are modified as follows

$$W^{t+1} = Z_W^t - \frac{1}{L_{W^t}} \nabla_W F(Z_W^t, \hat{Y}^t) + \gamma \|W\|_1 \tag{5.2.8}$$
$$\hat{Y}^{t+1} = \hat{Z}_{\hat{Y}}^t - \frac{1}{L_{\hat{Y}^t}} \nabla_{\hat{Y}} F(W^{t+1}, \hat{Z}_{\hat{Y}}^t) \tag{5.2.9}$$

where $\nabla_W F(Z_W^t, \hat{Y}^t) = XX^T W - XY^T$ and $\nabla_{\hat{Y}} F(W^{t+1}, \hat{Z}_{\hat{Y}}^t) = \hat{Y}(I + \alpha L + \beta I) - XW - \beta YV$ by taking the partial derivative w.r.t W and $\hat{Y}$.

In the updating rule Eq. 5.2.8, $g(W)$ is the l_1 norm regularization term, then in every iteration W^{t+1} is solved using the soft-threshold operator S_γ , where $\gamma > 0$.

W^{t+1} is defined as

$$W^{t+1} = S_\gamma(W^{t+1}) = Z_W^t - \frac{1}{L_{W^t}} \nabla_W F(Z_W^t, \hat{Y}^t) + \gamma \|W\|_1 \tag{5.2.10}$$

where the soft-threshold operation S_γ is as follows

$$S_\gamma[w_{ij}] = \begin{cases} w_{ij} - \gamma & \text{if } w_{ij} > \gamma \\ w_{ij} + \gamma & \text{if } w_{ij} < -\gamma \\ 0 & \text{otherwise} \end{cases}$$

Algorithm 6 summarizes the overall optimization process of the objective function using the accelerated proximal gradient algorithm.

Input: Training data $X \in \mathbb{R}^{n \times d}$
label matrix $Y \in \mathbb{R}^{n \times l}$
label regularization $V \in \mathbb{R}^{l \times l}$
Parameters: α, β, γ and MaxIteration
Initialization:
$\varepsilon > 0$: to avoid divide by zero
$\hat{Y}_1 = YV$
$W_1 = (X^T\hat{Y})/(X^TX + \varepsilon)$
k_0, $k_1 = 1$ and $t = 0$
while *MaxIteration*$> t$ **do**
  $\omega_t = \frac{k_{t-1}-1}{k_t}$;
  $Z_W^t = W^t + \omega_t(W^t - W^{t-1})$;
  $W^{t+1} = Z_W^t - \frac{1}{L_W}\nabla_W F(Z_W^t, \hat{Y}^t)$;
  $W^{t+1} = S_\gamma(W^{t+1})$;
  $Z_{\hat{Y}}^t = \hat{Y}^t + \omega_t(\hat{Y}^t - \hat{Y}^{t-1})$;
  $\hat{Y}^{t+1} = \hat{Z}_{\hat{Y}}^t - \frac{1}{L_{\hat{Y}}}\nabla_{\hat{Y}} F(W^{t+1}, \hat{Z}_{\hat{Y}}^t)$;
  $k_{t+1} = \frac{(1+\sqrt{1+4k_t^2})}{2}$;
  $t = t + 1$;
end
Output: Coefficient matrix $W \in \mathbb{R}^{d \times l}$
New label matrix $\hat{Y} \in \mathbb{R}^{n \times l}$

Algorithm 6: ML-CIB algorithm

5.2.3 Experiments and Datasets

In this section, experiments are conducted on multi-labeled datasets from different applications to demonstrate the efficacy of the proposed method. The experiments are structured in two phases based on the assessments of the proposed method. First, ML-CIB is introduced as a multi-label classification method which uses linear regression to classify the test points. Second, the proposed method is evaluated as a feature selection method. Several state-of-the-art multi-label learning methods for imbalanced datasets and multi-label feature selection methods are selected for comparison. This section is divided into four subsections: datasets, evaluation metrics, experimental configuration, experimental results and parameter sensitivity.

Datasets

The experiments are conducted on fifteen multi-label benchmark datasets. These datasets are collected from various domains to empirically show the generalization of the proposed method and to serve as a solid basis for the results analysis. The characteristics of the datasets are summarized in Table 5.5, which shows that these datasets cover a broad range of applications, including music, image, text, biology and video. They are available online from the Mulan[3] and LAMDA[4] repositories. The average class imbalance ratio $AvgImR$ is computed for the datasets by $\sum_{j=1}^{l} AvgImR_j$, and $ImR_j = \max\left(D_j^+, D_j^-\right) / \min\left(D_j^+, D_j^-\right)$, where D_j^+ and D_j^- are the number of positive and negative instances respectively w.r.t label j (Zhang et al. 2015). In addition, $MinImR$ and $MaxImR$ are the minimum and maximum imbalance ratio among all labels.

[3]http://mulan.sourceforge.net/datasets-mlc.html
[4]http://lamda.nju.edu.cn/Data.ashx

Table 5.5: Characteristics of the evaluated datasets

Dataset	Domain	#Features	#Labels	#Inst	$MinImR$	$MaxImR$	$AvgImR$
Bibtex[5]	Text	1836	159	7395	6.09	144	87.69
Medical	Text	1449	45	978	2.67	977	328.06
Enron	Text	1001	53	1702	1.00	1701	136.86
Genebase	Biology	1185	27	662	2.87	661	143.45
Image	Images	294	5	2000	2.44	3.89	3.11
MIMLtext	Text	243	7	7119	2.15	11.08	5.71
CAL500	Music	68	174	502	1.04	99.40	22.34
Corel5k	Images	499	374	5000	3.46	4999	845.28
Emotions	Music	72	6	593	1.24	3	2.32
Education	Text	550	33	5000	2.17	4999	568.77
Languagelog	Text	1004	75	1459	7.53	1458	334.11
MediaMill	Video	120	101	43907	1.74	1415.4	331.43
RCV1V2 (S1)	Text	23679	103	6000	1.12	5999	436.9
RCV1V2 (S2)	Text	23571	103	6000	1.12	5999	436.9
RCV1V2 (S3)	Text	23631	103	6000	1.12	5999	436.9

Evaluation Metrics

Evaluating the classification performance of the multi-label learning models by using standard evaluation metrics is not appropriate. A number of evaluation metrics have been proposed specifically for multi-label learning (Zhang & Zhou 2014). Seven evaluation metrics are used in this study, comprising Hamming Loss, One-Error, Coverage, Ranking Loss, Average Precision, Subset Accuracy, and Micro-F1 (Schapire & Singer 2000, Ghamrawi & McCallum 2005, Godbole & Sarawagi 2004). In the first four evaluation metrics (Hamming Loss, One-Error, Coverage, and Ranking Loss), the smaller the value, the better the performance. However, In the other four evaluation metrics, the reverse is true.

Given a multi-label test set $S = \{(x_i, Y_i)\}_{i=1}^{n}$, where $Y_i \in \{0, 1\}^l$ is the ground truth labels of the test point x_i, l the labels, n is the number of test points, $|Y_i|$ is the size label set for instance x_i, and $h(x_i)$ is the multi-label classifier. This section defines the evaluation metrics as follows

Subset Accuracy evaluates the correctly classified examples among all labels.

$$\text{Subset Accuracy} = \frac{1}{n}\sum_{i=1}^{n} I\left[h(x_i) = Y_i\right]$$

where $I\left[.\right]$ is the indicator function.

Hamming Loss evaluates the frequency of an example-label pair being misclassified, i.e. a non-relevant label is predicted for the example, or relevant is missed.

$$\text{Hamming Loss} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{l}\sum_{j=1}^{l} I\left[h(x_i)_j \neq Y_{ij}\right]$$

Average Precision is the average of the proportion of relevant labels ranked higher than a particular labels.

$$\text{Average Precision} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\left|Y_i^l\right|}\sum_{y_j \in Y_i^l}\frac{\left|\left\{y_q \in Y_i^l : R_i(y_q \leq R_i(y_j))\right\}\right|}{R_i(y_j)}$$

where $R_i(y_j)$ is the predicted rank of the class label y_j for an instance x_i.

One-Error is the fraction of the examples whose most ranked label is irrelevant.

$$\text{One-Error} = \frac{1}{n}\sum_{i=1}^{n} I\left[\min_{y_j \in Y_i} R_i(y_j \notin Y_i)\right]$$

Coverage is the metric that on average computes the number of steps are needed to go down in the ranked predicted list to cover all the relevant labels for the instance.

$$\text{Coverage} = \frac{1}{n}\sum_{i=1}^{n}\max_{y \in Y} R_i(y) - 1$$

Ranking Loss computes the fraction of the incorrectly ordered label pairs.

$$\text{Ranking Loss} = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|(y_a, y_b) : R_i(y_a) > R_i(y_b), (y_a, y_b) \in Y_i^l \times \widehat{Y_i^l}\right|}{\left|Y_i^l\right|\left|\widehat{Y_i^l}\right|}$$

where $\widehat{Y}_i = Y \setminus Y_i$, and (y_a, y_b) is the pair class label for instance x_i.

Micro-F1 evaluates every label separately using F1 measure and averages over all labels.

$$\text{Micro-F1} = \frac{2\sum_{j=1}^{l}\sum_{i=1}^{n} y_{ij}h(x_i, y_j)}{\sum_{j=1}^{l}\sum_{i=1}^{n} y_{ij} + \sum_{j=1}^{l}\sum_{i=1}^{n} h(x_i, y_j)}$$

Experimental Settings and the State-of-the-art

Extensive experiments are conducted in this section to demonstrate the benefits of the ML-CIB method. Experimental setting is divided into two stages. First, MLC is achieved by using a linear regression method based on a threshold τ without using a specific classifier. Second, ML-CIB is used as a supervised feature selection method (ML-CIB-FS), which reduces the dimensionality of the large-scale datasets by performing a feature selection on the matrix W. To ensure a fair comparison between the compared methods, the classification results of every metric are averaged based on five-fold cross-validation.

In the first stage of MLC, $\tau = 0.5$ is set. The MLC performance of the proposed algorithm is compared with four state-of-the-art methods which are specifically proposed to handle the class imbalance problem in MLC. These methods are: MLSMOTE (Charte et al. 2015*b*), BR-IRUS (Tahir et al. 2012), COCOA (Zhang et al. 2015), and MMIB (Wu et al. 2016). Also a baseline binary relevance (BR) method (Boutell et al. 2004) is utilised in these experiments. A linear SVM classifier is used for binary classification for BR, CC and MLSMOTE.

1. Binary relevance: Independently trains one binary classifier for each label.

2. MLSMOTE[6] : Multilabel Synthetic Minority Over-sampling Technique, which generates synthetic samples for imbalanced multi-labeled datasets.

3. BR-IRUS: Inverse random under-sampling technique, which trains an ensemble of classifiers in each label by using a random subset of the majority samples along with all minority samples.

4. COCOA[7]: Cross-Coupling Aggregation method, which proposes a predictive model that combines one binary-class imbalance learner for a certain label with several multi-class imbalance learning models.

5. MMIB[8]: Proposes a constrained sub-modular minimization optimization function to handle the missing labels and class imbalance problem in multi-label learning.

In the second stage, ML-CIB-FS is presented as a supervised feature selection method. The effectiveness of selected features is evaluated from the compared algorithms by checking the classification performance. To ensure a fair comparison between all algorithms, the binary relevance (BR) and classifier chains (CC) strategies are used to decompose the multi-label classification problem into binary sub-problem classifiers. Without loss of generality, this study uses a linear SVM to train the binary sub-problem classifiers. The results, based on different numbers of features vary from 5% to 20%. ML-CIB-FS is benchmarked against the following state-of-the-art multi-label feature selection methods.

[6]The java application is avaialble at http://simidat.ujaen.es/papers/MLSMOTE
[7]The source code is obtained from http://cse.seu.edu.cn/people/zhangml/Resources.htm
[8]The source code is obtained from https://sites.google.com/site/baoyuanwu2015/Publications

1. MIFS (Jian et al. 2016): Multi-Label Informed Feature Selection which exploits the label correlations to extract the common features of multiple labels.

2. SFUS (Ma et al. 2012): Sub-Feature Uncovering with Sparsity, which jointly discovers the shared feature space in multi-label learning with feature selection.

5.2.4 Results and Analysis

MLC results

Tables 5.6 and 5.7 report the classification results of the algorithms using the different evaluation metrics. These results are the average values of each compared algorithm using five-fold cross-validation. Several datasets with different characteristics are used in the experiments to investigate the generalization of the algorithms. As shown in Table 5.5, a regular and large-scale datasets are used in terms of the number of instances and dimensions. Further, several datasets with a small, regular and extreme imbalance ratio are used in the experiments. The experimental results in Tables 5.6 and 5.7 indicate the following observations:

- First, the state-of-the-art algorithms perform better than the baseline method (BR) in most of the evaluation metrics. This indicates the importance of handling the class imbalance and the incomplete label matrix to improve the performance of MLC.

- Second, the results produced in Tables 5.6 and 5.7 by the ML-CIB algorithm are compared with the results obtained by the-state-of-the-art algorithms, which show that the proposed method outperforms the compared algorithms in most benchmark datasets using various evaluation metrics.

- Third, theoretically, the missing labels for the multi-label matrix increase the imbalance ratio of the datasets. The results show that addressing the problem of missing labels along with the class imbalance problem improves the performance of MLC, and this is shown in the results of ML-CIB and MMIB against the results of the other algorithms.

- Fourth, the Ranking Loss and One-Error metrics rank the irrelevant labels higher than the relevant labels by definition (Zhang & Zhou 2014). As shown in Tables 5.6 and 5.7, the algorithm achieves the best results in the ranking loss metric, because it handles the incomplete label matrix by predicting the relevant labels using the label manifold during the training process.

- Fifth, the Micro-F1 used F1-score, which is appropriate in the presence of the class imbalance problem. The reported results in these two metrics show that ML-CIB achieves better performance than the other algorithms in most benchmark datasets, which indicates the superiority of the proposed algorithm in handling the class imbalance problem.

- Finally, the proposed algorithms work efficiently on the extreme levels of class imbalance ratio ($AvgImR > 50$) without being restricted by the exclusion of any class label, as in (Zhang et al. 2015). The results in Tables 5.6 and 5.7 show that ML-CIB achieves better performance than the others on extreme imbalance ratio datasets such as Languagelog, MediaMill, RCV1V2(S1) and Corel5k.

This study conducts a Friedman test, which is suitable for statistically demonstrating the difference in performance between multiple learning methods in various datasets. The Friedman test is performed with 95% confidence, under the hypothesis that the

evaluation metric results between the two types of methods would not be significantly different. As shown in Table 5.7, the p-value is lower than 0.05 in the most cases, which rejects the hypothesis and indicates that the proposed method outperforms the other methods. Also, the ranks are reported on each dataset of the compared algorithms, and a (✓) sign for the Friedman test suggests that the methods being compared are significantly different. As shown in the results in Table 5.7, the ML-CIB method ("Base") is significantly better than the compared methods BR, MLSMOTE, BR-IRUS, COCOA and MMIB. As highlighted earlier, the ML-CIB algorithm improves the performance of multi-label classification for two reasons. First, it tackles the class imbalance problem including the high level of imbalanced ratios. Second, it predicts a new label matrix from the existing incomplete label matrix which alleviates the negative effects of the rare appearance of many class labels.

Table 5.6: Multi-label classification results of each compared algorithm (mean) on the regular-scale multi-labeled datasets. Rank between (.). The best results are highlighted in bold.

Evaluation criteria	Algorithm	Bibtex	Medical	Enron	Genebase	Image	MIMLText	CAL500	Corel5k	Emotions
Hamming Loss ↓	MLCIB	**0.005** (1)	**0.007** (1)	0.054 (2)	0.003 (3)	**0.142** (1)	**0.039** (1)	0.324 (6)	**0.005** (1)	**0.097** (1)
	BR (Baseline)	0.015 (4)	0.010 (3)	0.059 (4)	**0.001** (1)	0.176 (5)	0.051 (3)	0.137 (3)	0.011 (3)	0.199 (5)
	MLSMOTE	0.014 (2)	0.008 (2)	0.058 (3)	0.001 (1)	0.175 (4)	0.046 (2)	0.135 (2)	0.008 (2)	0.197 (4)
	BR-IRUS	0.031 (6)	0.020 (6)	0.110 (6)	0.006 (5)	0.224 (6)	0.067 (6)	0.190 (5)	0.022 (6)	0.227 (6)
	COCOA	0.017 (5)	0.015 (5)	0.074 (5)	0.002 (2)	0.160 (3)	0.059 (5)	0.188 (4)	0.014 (4)	0.176 (3)
	MMIB	0.014 (3)	0.012 (4)	**0.051** (1)	0.005 (4)	0.153 (2)	0.055 (4)	**0.098** (1)	0.018 (5)	0.174 (2)
One-Error ↓	MLCIB	0.199 (3)	**0.043** (1)	**0.086** (1)	**0.000** (1)	0.140 (2)	**0.027** (1)	0.106 (4)	**0.077** (1)	**0.103** (1)
	BR (Baseline)	0.232 (5)	0.072 (3)	0.112 (4)	0.001 (2)	0.155 (5)	0.040 (3)	0.001 (1)	0.469 (5)	0.129 (5)
	MLSMOTE	0.229 (4)	0.068 (2)	0.110 (3)	0.001 (2)	0.153 (4)	0.037 (2)	**0.001** (1)	0.467 (4)	0.128 (4)
	BR-IRUS	0.448 (6)	0.134 (6)	0.207 (6)	0.001 (2)	0.197 (6)	0.052 (6)	0.002 (2)	0.787 (6)	0.149 (6)
	COCOA	0.193 (2)	0.087 (5)	0.135 (5)	0.001 (2)	0.141 (3)	0.049 (5)	0.004 (3)	0.429 (3)	0.113 (2)
	MMIB	**0.172** (1)	0.076 (4)	0.095 (2)	0.003 (3)	**0.135** (1)	0.046 (4)	0.002 (2)	0.154 (2)	0.115 (3)
Ranking Loss ↓	MLCIB	**0.101** (1)	**0.030** (1)	**0.092** (1)	**0.000** (1)	**0.174** (1)	**0.035** (1)	**0.175** (1)	0.324 (2)	**0.216** (1)
	BR (Baseline)	0.597 (5)	0.214 (3)	0.489 (4)	0.009 (3)	0.503 (5)	0.159 (3)	0.772 (4)	0.885 (5)	0.454 (5)
	MLSMOTE	0.594 (4)	0.211 (2)	0.486 (3)	0.006 (2)	0.502 (4)	0.155 (2)	0.771 (3)	0.882 (4)	0.450 (4)
	BR-IRUS	0.99 (6)	0.396 (6)	0.889 (6)	0.013 (5)	0.631 (6)	0.201 (6)	0.99 (5)	0.979 (6)	0.507 (6)
	COCOA	0.489 (3)	0.248 (5)	0.580 (5)	0.013 (5)	0.450 (3)	0.177 (5)	0.996 (6)	0.810 (3)	0.396 (3)
	MMIB	0.438 (2)	0.222 (4)	0.405 (2)	0.010 (4)	0.434 (2)	0.169 (4)	0.545 (2)	**0.293** (1)	0.394 (2)
Coverage ↓	MLCIB	**0.199** (1)	**0.039** (1)	**0.250** (1)	**0.008** (1)	**0.195** (1)	**0.070** (1)	0.737 (2)	0.562 (5)	0.350 (3)
	BR (Baseline)	0.477 (5)	0.147 (3)	0.453 (4)	0.0183 (3)	0.276 (5)	0.116 (3)	0.871 (4)	0.439 (4)	0.359 (5)
	MLSMOTE	0.474 (4)	0.143 (2)	0.448 (3)	0.014 (2)	0.275 (4)	0.114 (2)	0.867 (3)	0.437 (3)	0.355 (4)
	BR-IRUS	0.917 (6)	0.273 (6)	0.821 (6)	0.026 (6)	0.347 (6)	0.149 (6)	0.982 (5)	0.736 (6)	0.404 (6)
	COCOA	0.391 (3)	0.170 (5)	0.537 (5)	0.022 (4)	0.248 (3)	0.128 (5)	0.994 (6)	0.401 (2)	0.314 (2)
	MMIB	0.352 (2)	0.151 (4)	0.377 (2)	0.022 (5)	0.242 (2)	0.122 (4)	**0.615** (1)	**0.145** (1)	**0.310** (1)
Subset Accuracy ↑	MLCIB	**0.240** (1)	0.540 (4)	**0.145** (1)	**0.999** (1)	0.282 (5)	0.502 (6)	**0.195** (1)	**0.191** (1)	0.283 (3)
	BR (Baseline)	0.143 (5)	0.662 (2)	0.108 (4)	0.975 (3)	0.384 (4)	0.725 (2)	0.000 (6)	0.007 (6)	0.263 (5)
	MLSMOTE	0.153 (4)	**0.671** (1)	0.117 (3)	0.985 (2)	0.393 (3)	**0.733** (1)	0.008 (4)	0.015 (4)	0.270 (4)
	BR-IRUS	0.002 (6)	0.107 (6)	0.012 (6)	0.641 (6)	0.279 (6)	0.534 (5)	0.010 (3)	0.112 (2)	0.223 (6)
	COCOA	0.163 (3)	0.554 (5)	0.081 (5)	0.970 (4)	0.416 (2)	0.655 (4)	0.015 (2)	0.091 (3)	**0.292** (1)
	MMIB	0.173 (2)	0.647 (3)	0.118 (2)	0.968 (5)	**0.430** (1)	0.691 (3)	0.008 (5)	0.008 (5)	0.291 (2)
Average Precision ↑	MLCIB	**0.670** (1)	**0.970** (1)	**0.738** (1)	**0.999** (1)	**0.880** (1)	**0.990** (1)	**0.679** (1)	**0.399** (1)	**0.951** (1)
	BR (Baseline)	0.372 (5)	0.783 (3)	0.457 (4)	0.990 (2)	0.720 (5)	0.887 (3)	0.292 (4)	0.098 (5)	0.740 (5)
	MLSMOTE	0.379 (4)	0.792 (2)	0.466 (3)	0.999 (1)	0.727 (4)	0.895 (2)	0.300 (3)	0.106 (3)	0.747 (4)
	BR-IRUS	0.021 (6)	0.126 (6)	0.079 (6)	0.650 (5)	0.531 (6)	0.656 (6)	0.180 (5)	0.023 (6)	0.647 (6)
	COCOA	0.432 (3)	0.658 (5)	0.366 (5)	0.984 (3)	0.789 (3)	0.803 (5)	0.180 (5)	0.099 (4)	0.833 (3)
	MMIB	0.463 (2)	0.765 (4)	0.529 (2)	0.980 (4)	0.813 (2)	0.848 (4)	0.372 (2)	0.220 (2)	0.836 (2)
Micro-F1 ↑	MLCIB	0.453 (3)	**0.870** (1)	**0.671** (1)	**0.964** (1)	**0.757** (1)	0.760 (5)	0.358 (2)	**0.439** (1)	**0.778** (1)
	BR (Baseline)	0.416 (5)	0.804 (3)	0.513 (4)	0.950 (5)	0.577 (5)	0.853 (2)	0.330 (4)	0.157 (5)	0.646 (4)
	MLSMOTE	0.424 (4)	0.812 (2)	0.521 (3)	0.959 (4)	0.585 (4)	**0.862** (1)	0.338 (3)	0.165 (3)	0.655 (3)
	BR-IRUS	0.024 (6)	0.132 (6)	0.089 (6)	0.651 (6)	0.425 (6)	0.632 (6)	0.202 (6)	0.043 (6)	0.563 (5)
	COCOA	0.483 (2)	0.675 (5)	0.414 (5)	0.963 (2)	0.631 (3)	0.771 (4)	0.205 (5)	0.163 (4)	0.728 (2)
	MMIB	**0.519** (1)	0.785 (4)	0.597 (2)	0.961 (3)	0.650 (2)	0.814 (3)	**0.420** (1)	0.357 (2)	0.728 (2)

Table 5.7: Multi-label classification results of each compared algorithm (mean) on the large-scale multi-labeled datasets. Rank between (.). The best results are highlighted in bold.

Evaluation criteria	Algorithm	Education	Languagelog	MediaMill	RCV1V2(S1)	RCV1V2(S2)	RCV1V2(S3)	Avg. rank	Friedman
Hamming Loss ↓	MLCIB	**0.030** (1)	**0.011** (1)	**0.006** (1)	**0.008** (1)	0.011 (2)	0.013 (3)	1.73	**Base**
	BR (Baseline)	0.037 (4)	0.018 (4)	0.031 (4)	0.013 (3)	0.012 (3)	0.012 (2)	3.4	✓0.020
	MLSMOTE	0.032 (2)	0.014 (3)	0.028 (3)	0.009 (2)	**0.008** (1)	**0.011** (1)	2.26	0.070
	BR-IRUS	0.047 (5)	0.024 (5)	0.061 (5)	0.019 (5)	0.017 (6)	0.018 (5)	5.53	✓0.004
	COCOA	0.035 (3)	0.012 (2)	0.013 (2)	0.013 (3)	0.016 (5)	0.013 (3)	3.6	✓0.020
	MMIB	0.032 (2)	0.014 (3)	0.028 (3)	0.016 (4)	0.015 (4)	0.017 (4)	3.06	0.070
One-Error ↓	MLCIB	**0.051** (1)	0.313 (3)	**0.024** (1)	0.025 (3)	**0.016** (1)	**0.019** (1)	1.66	**Base**
	BR (Baseline)	0.076 (4)	0.417 (5)	0.107 (5)	0.023 (2)	0.025 (3)	0.024 (3)	3.66	✓0.0045
	MLSMOTE	0.071 (3)	0.415 (4)	0.106 (4)	**0.019** (1)	0.021 (2)	0.021 (2)	2.8	✓0.0045
	BR-IRUS	0.092 (5)	0.473 (6)	0.206 (6)	0.030 (5)	0.033 (5)	0.031 (6)	5.26	✓0.0008
	COCOA	0.065 (2)	**0.203** (1)	0.036 (2)	0.026 (4)	0.025 (3)	0.027 (4)	3.06	0.07
	MMIB	0.065 (2)	0.257 (2)	0.090 (3)	0.023 (2)	0.026 (4)	0.029 (5)	2.66	0.438
Ranking Loss ↓	MLCIB	**0.222** (1)	**0.205** (1)	**0.169** (1)	**0.294** (1)	**0.289** (1)	**0.279** (1)	1.06	**Base**
	BR (Baseline)	0.764 (5)	0.718 (5)	0.536 (5)	0.316 (5)	0.314 (5)	0.306 (4)	4.4	✓0.0001
	MLSMOTE	0.762 (4)	0.714 (4)	0.533 (4)	0.315 (4)	0.309 (4)	0.304 (3)	3.4	✓0.0001
	BR-IRUS	0.912 (6)	0.810 (6)	0.988 (6)	0.354 (6)	0.367 (6)	0.329 (6)	5.53	✓0.0001
	COCOA	0.629 (3)	0.347 (3)	0.173 (2)	0.299 (2)	0.294 (2)	0.283 (2)	3.46	✓0.0001
	MMIB	0.628 (2)	0.442 (2)	0.433 (3)	0.303 (3)	0.313 (3)	0.323 (5)	2.73	✓0.0008
Coverage ↓	MLCIB	0.285 (2)	**0.151** (1)	**0.195** (1)	**0.253** (1)	**0.256** (1)	**0.249** (1)	1.53	**Base**
	BR (Baseline)	0.300 (4)	0.359 (5)	0.625 (5)	0.277 (5)	0.277 (4)	0.274 (4)	4.2	✓0.0008
	MLSMOTE	0.298 (3)	0.358 (4)	0.622 (4)	0.276 (4)	0.274 (3)	0.273 (3)	3.2	✓0.0008
	BR-IRUS	0.358 (5)	0.408 (6)	0.976 (6)	0.312 (6)	0.324 (6)	0.295 (6)	5.86	✓0.0001
	COCOA	0.249 (1)	0.174 (2)	0.201 (2)	0.261 (2)	0.260 (2)	0.254 (2)	3.06	✓0.0201
	MMIB	**0.249** (1)	0.222 (3)	0.507 (3)	0.264 (3)	0.278 (5)	0.289 (5)	2.8	0.070
Subset Accuracy ↑	MLCIB	**0.297** (1)	**0.384** (1)	**0.298** (1)	**0.644** (1)	**0.643** (1)	**0.666** (1)	1.93	**Base**
	BR (Baseline)	0.195 (5)	0.225 (5)	0.070 (5)	0.437 (5)	0.436 (4)	0.439 (4)	4.33	✓0.020
	MLSMOTE	0.203 (4)	0.233 (4)	0.077 (3)	0.445 (4)	0.445 (3)	0.448 (3)	3.13	✓0.020
	BR-IRUS	0.150 (6)	0.189 (6)	0.005 (6)	0.379 (6)	0.357 (6)	0.404 (6)	5.46	✓0.0008
	COCOA	0.220 (3)	0.334 (2)	0.110 (2)	0.456 (2)	0.460 (2)	0.470 (2)	3.26	0.07
	MMIB	0.221 (2)	0.306 (3)	0.075 (4)	0.452 (3)	0.433 (5)	0.412 (5)	3.33	0.07
Average Precision ↑	MLCIB	**0.729** (1)	**0.469** (1)	**0.878** (1)	**0.843** (1)	**0.881** (1)	**0.851** (1)	1	**Base**
	BR (Baseline)	0.297 (4)	0.168 (5)	0.476 (5)	0.709 (5)	0.708 (4)	0.715 (4)	4.46	✓0.0001
	MLSMOTE	0.305 (3)	0.177 (4)	0.485 (4)	0.718 (4)	0.718 (3)	0.722 (3)	3.13	✓0.0001
	BR-IRUS	0.232 (5)	0.138 (6)	0.045 (6)	0.621 (6)	0.585 (5)	0.661 (6)	5.73	✓0.0001
	COCOA	0.342 (2)	0.247 (2)	0.793 (2)	0.748 (2)	0.753 (2)	0.769 (2)	3.2	✓0.0001
	MMIB	0.342 (2)	0.224 (3)	0.561 (3)	0.739 (3)	0.708 (4)	0.679 (5)	2.93	✓0.0001
Micro-F1 ↑	MLCIB	**0.412** (1)	**0.376** (1)	**0.882** (1)	**0.883** (1)	0.793 (2)	**0.853** (1)	1.53	**Base**
	BR (Baseline)	0.338 (5)	0.205 (5)	0.518 (5)	0.752 (5)	0.755 (5)	0.756 (4)	4.4	✓0.0045
	MLSMOTE	0.346 (4)	0.213 (4)	0.525 (4)	0.760 (4)	0.764 (3)	0.765 (3)	3.26	✓0.0045
	BR-IRUS	0.267 (6)	0.172 (6)	0.048 (6)	0.659 (6)	0.624 (6)	0.702 (6)	5.53	✓0.0001
	COCOA	0.391 (3)	0.303 (2)	0.861 (2)	0.792 (2)	**0.803** (1)	0.813 (2)	2.93	0.1967
	MMIB	0.392 (2)	0.275 (3)	0.612 (3)	0.784 (3)	0.757 (4)	0.718 (5)	2.66	0.1967

ML-CIB-FS Feature Selection Results

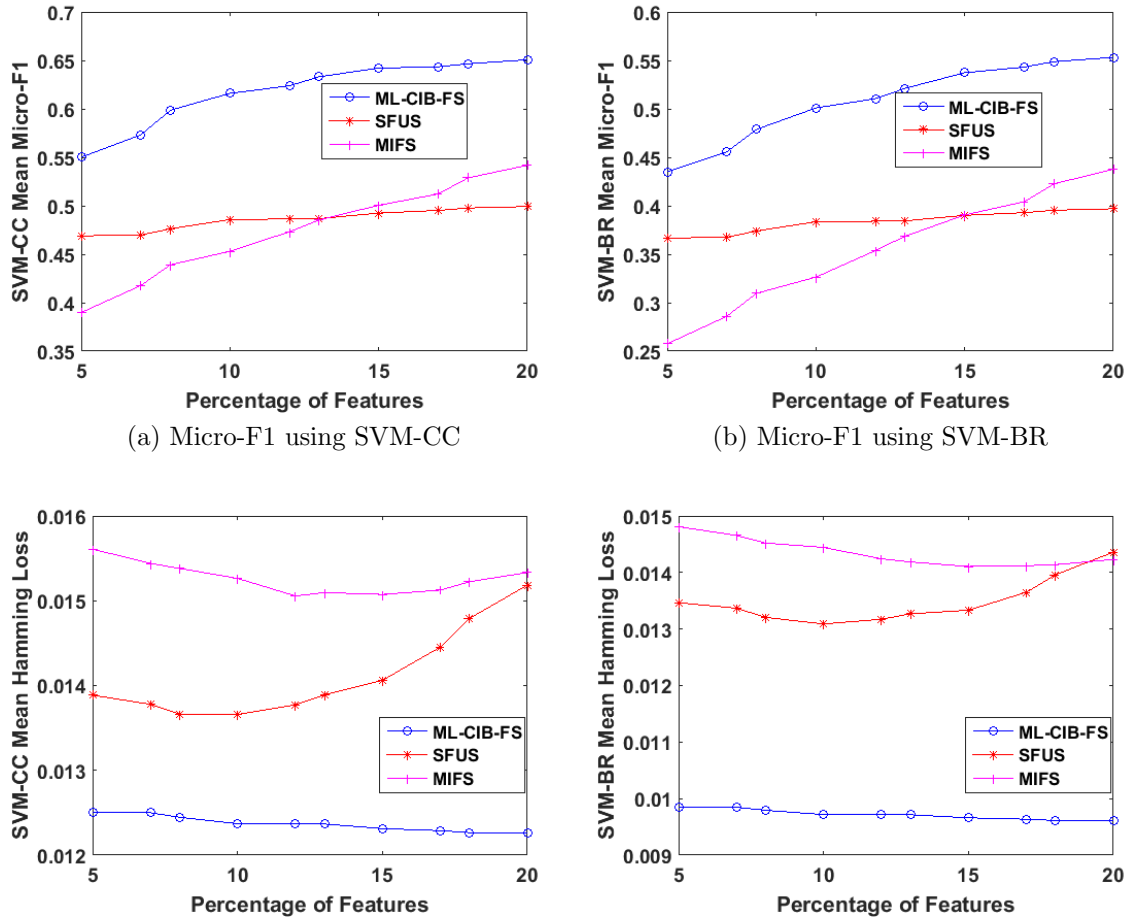
This thesis introduces ML-CIB-FS as a feature selection method for multi-label learning, which uses ML-CIB to rank the features. As mentioned above, ML-CIB handles the class imbalance and incomplete label matrix in order to improve the multi-label classification performance. Also, ML-CIB imposes l_1-norm regularization on the multi-label feature matrix W as defined in Eq. 5.2.5. Therefore, the top ranked features in W are those features that make a greater contribution to classifying the multi-label sample. The features with the largest values are selected as the most predictive balanced features, which are able to lessen the inherent bias to the majority labels due to class imbalance problem and incomplete label matrix. The ML-CIB-FS algorithm sorts all the features in W according to $\|w_i\|$ in descending order and returns the top ranked features.

Several experiments are conducted to assess the importance of the predicted features selected by the compared algorithms to enhance the multi-label classification. The features are selected from the top 5% to 20% of features for all algorithms. The selected features are evaluated on high-dimensional multi-labeled datasets, using two evaluation metrics, Micro-F1 and Hamming Loss, to evaluate the advantages of addressing the class imbalance problem and incomplete label matrix respectively in feature selection multi-label methods. As aforementioned, Micro-F1 is based on F1-measure and is used in the class imbalance problems, and Hamming Loss is used to check the fraction of the wrongly predicted labels. The results are reported in Figures 5.7 through 5.11. It is observed that the ML-CIB-FS algorithm outperforms the two compared algorithms in all benchmarked datasets. This indicates the importance of

handling the class imbalance problem in feature selection multi-label learning methods. Furthermore, the performance of the MIFS and SFUS algorithms very close, but ML-CIB-FS clearly demonstrates better performance as shown in the figures. The results in Figures 5.10 and 5.11 are really promising, as ML-CIB-FS significantly outperforms MIFS and SFUS in high-dimensional datasets such as RCV1V2 (S1) and RCV1V2 (S2) where the number of dimensions of each is > 40000. Moreover, as shown in the results, the proposed algorithm achieves better results using the Hamming Loss metric, which indicates the importance of selecting features based on a new complete label matrix $\hat{Y}$. The last observation in the reported results is the stability of the results after 10% of the features have been selected,, which demonstrates the ability of the ML-CIB-FS algorithm to achieve better results with fewer features. Using fewer features will reduce the computational time of training the classification model.

A statistical comparative analysis is performed by using the Nemenyi test (Demšar 2006) to check the performance of the proposed algorithm ML-CIB-FS against the other algorithms. According to the Nemenyi test, two methods perform differently if they vary by at least one critical difference (CD). This test is an important statistical test across different algorithms over multiple datasets. Based on the results of the Friedman mean ranks for all algorithms, Figure 5.12 shows that the rank results of the ML-CIB-FS algorithm performed differently to the SFUS and MIFS algorithms, because their ranks are located outside the interval of CD. Furthermore, it can be observed that SFUS and MIFS in all cases achieve equivalent performance due to their connection by a colored line, which indicates that there is no significant difference between connected algorithms. As a result, ML-CIB-FS is superior across different

types of high-dimensional datasets in different domains. The results indicate the importance of handling the class imbalance problem in multi-label feature selection.



(a) Micro-F1 using SVM-CC

(b) Micro-F1 using SVM-BR

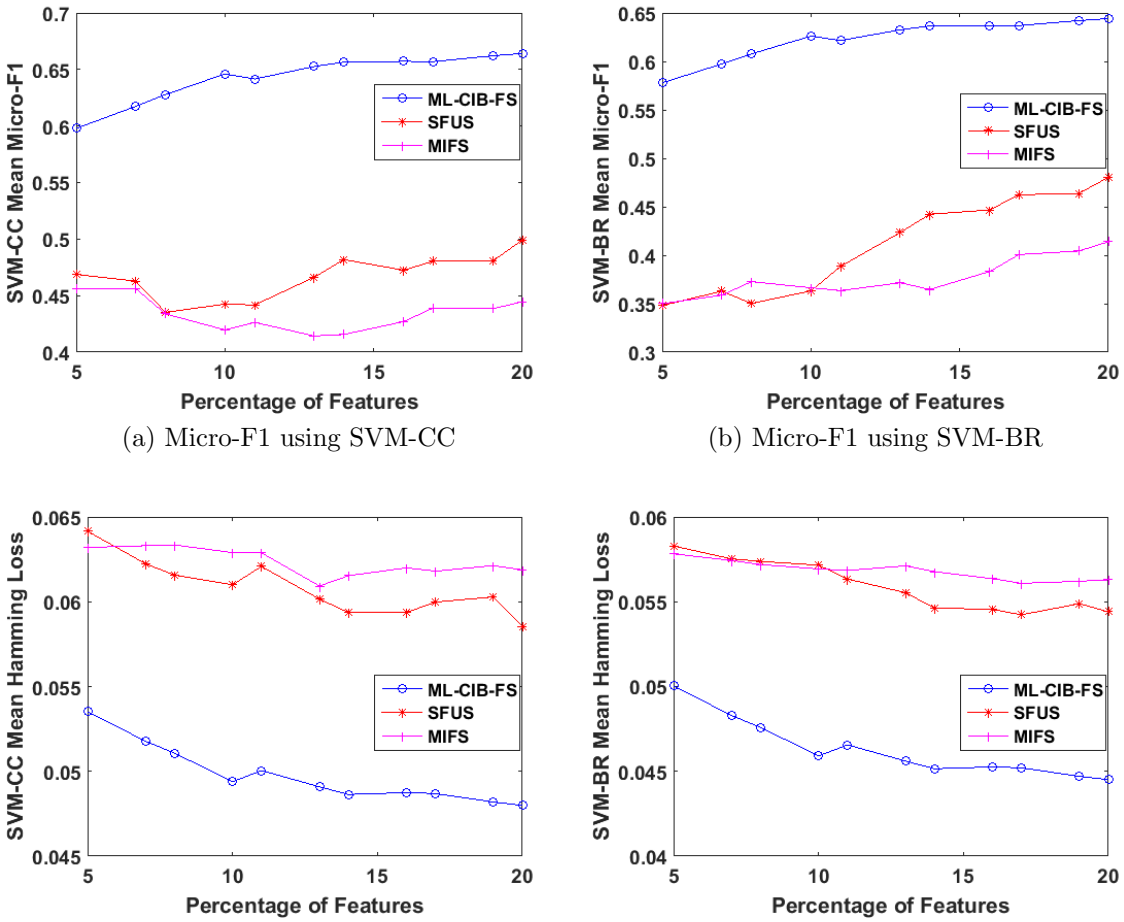
(c) Hamming Loss using SVM-CC

(d) Hamming Loss using SVM-BR

Figure 5.7: ML-CIB-FS average classification results on bibtex dataset

Time Complexity Analysis

This subsection reports the computation time of each method on all datasets in Table 5.8 to demonstrate the efficiency of the proposed method compared to the state-of-the-art methods. This is the time taken by each method to converge to the optimal



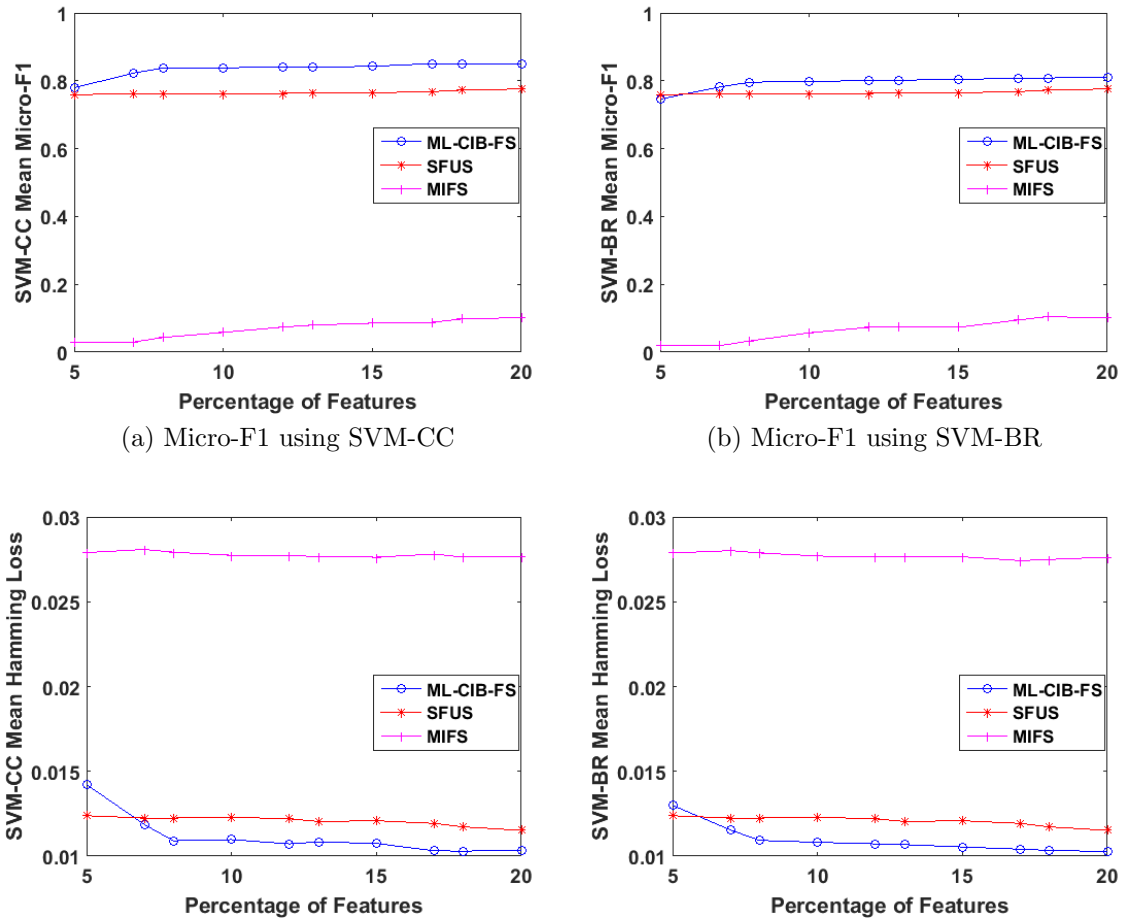
(a) Micro-F1 using SVM-CC

(b) Micro-F1 using SVM-BR

(c) Hamming Loss using SVM-CC

(d) Hamming Loss using SVM-BR

Figure 5.8: ML-CIB-FS average classification results on Enron dataset



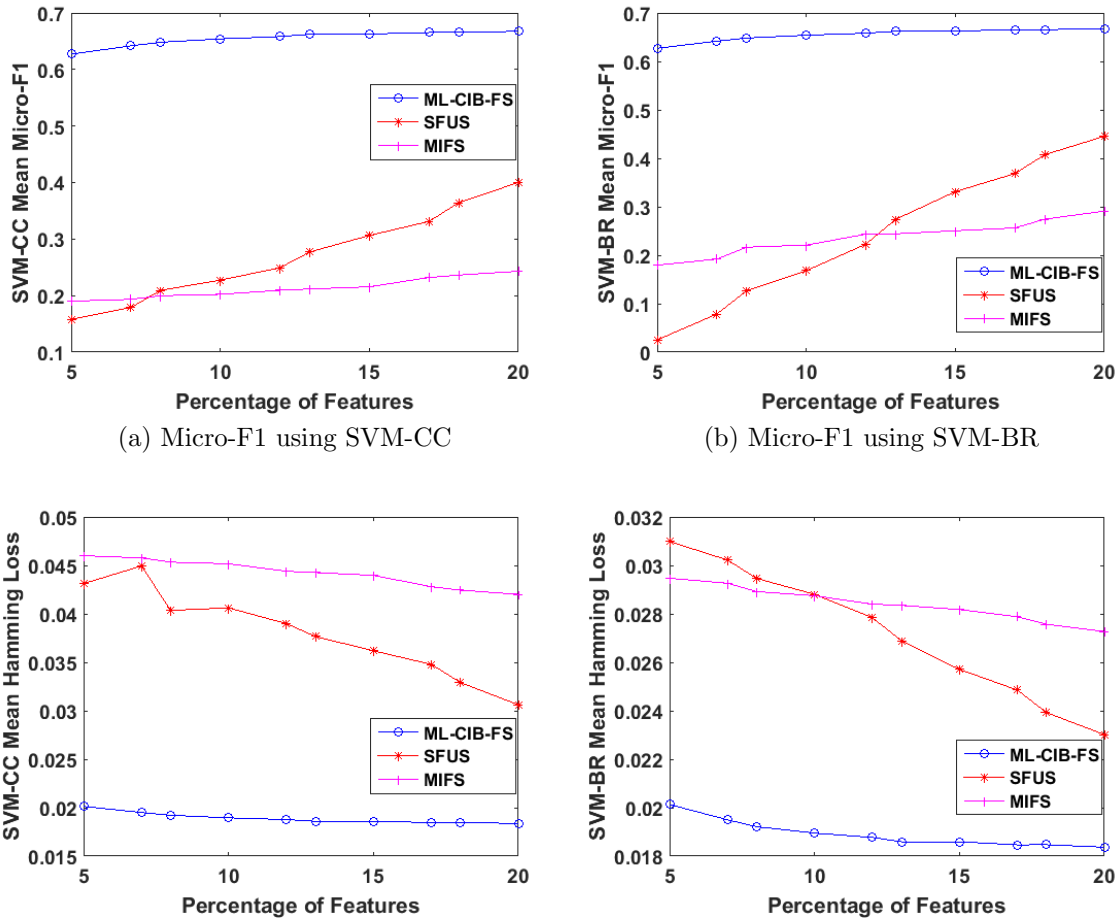
(a) Micro-F1 using SVM-CC

(b) Micro-F1 using SVM-BR

(c) Hamming Loss using SVM-CC

(d) Hamming Loss using SVM-BR

Figure 5.9: ML-CIB-FS average classification results on Medical dataset



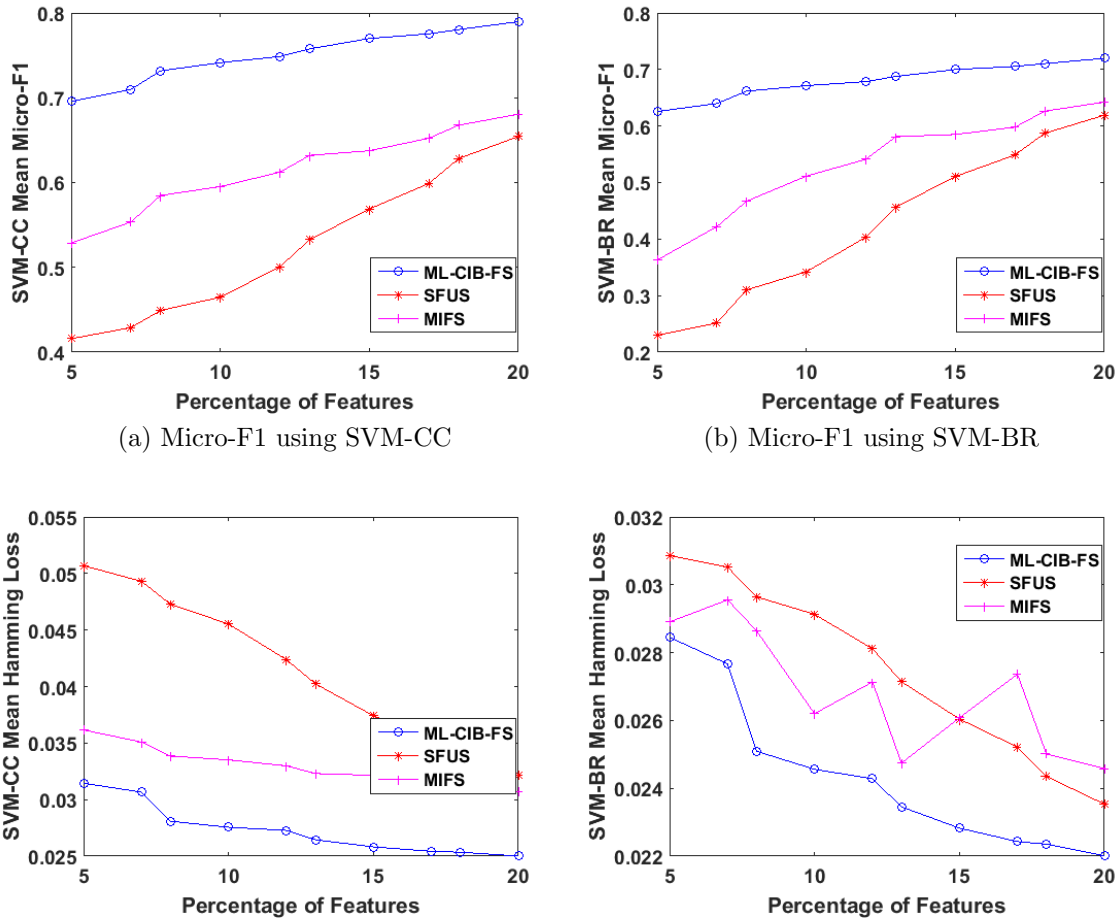
(a) Micro-F1 using SVM-CC

(b) Micro-F1 using SVM-BR

(c) Hamming Loss using SVM-CC

(d) Hamming Loss using SVM-BR

Figure 5.10: ML-CIB-FS average classification results on RCV1V2(S1) dataset



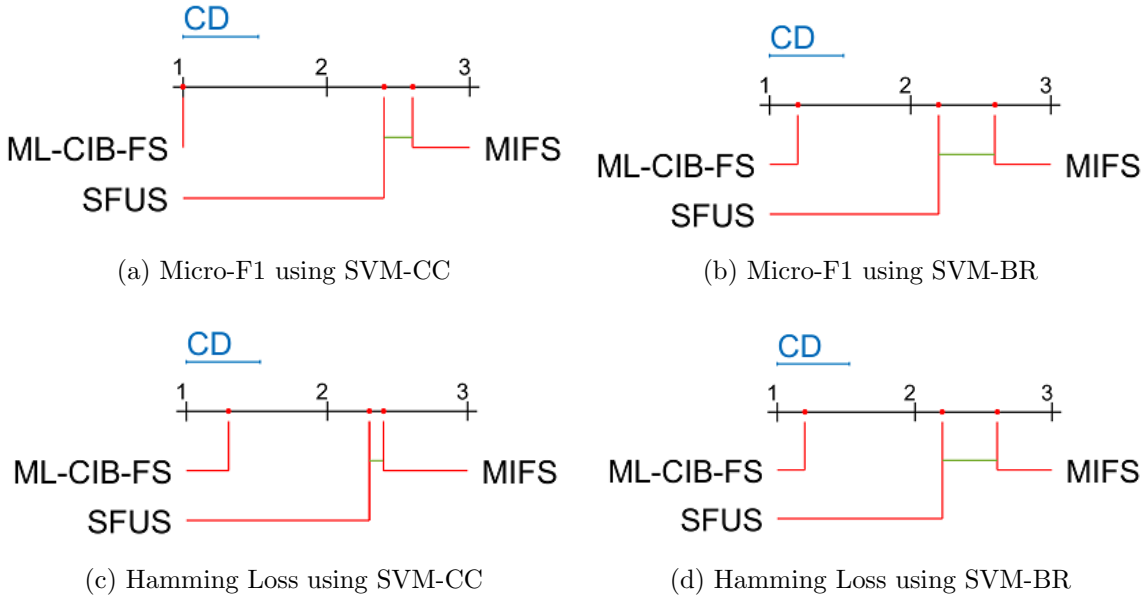
(a) Micro-F1 using SVM-CC

(b) Micro-F1 using SVM-BR

(c) Hamming Loss using SVM-CC

(d) Hamming Loss using SVM-BR

Figure 5.11: ML-CIB-FS average classification results on RCV1V2(S2) dataset



(a) Micro-F1 using SVM-CC

(b) Micro-F1 using SVM-BR

(c) Hamming Loss using SVM-CC

(d) Hamming Loss using SVM-BR

Figure 5.12: Comparison of ML-CIB-FS against the state-of-the-art algorithms using Nemenyi test.

solution. To ensure fair comparison, all experiments are conducted on a five core Linux server with 3.10-GHZ cores, and a total memory of 126-GB. The running time in Table 5.8 is for the optimal parameter setting in each method. The proposed algorithm is observed as the fastest algorithm compared to the state-of-the-art algorithms over most datasets. This indicates the efficiency of the proposed optimization method in converging quickly to the optimal solution.

ML-CIB performs matrix multiplication and matrix inversion computation at initialization, requiring $\mathcal{O}(d^3 + dnl)$ to compute W and $\mathcal{O}(l^2 n)$ to compute $\hat{Y}$. At each iteration, it needs $\mathcal{O}(d^2 l)$ for ∇_W and $\mathcal{O}(ndl + nl^2)$ for $\nabla_{\hat{Y}}$.

Table 5.8: The computation time (sec) of different algorithms

Dataset	MLCIB	BR	MLSMOTE	BR-IRUS	COCOA	MIMB
Bibtex	576.7	633.5	789.2	711.5	945.2	1659.2
Medical	6.7	90.2	268.2	68.2	241.2	299.3
Enron	13.5	14.1	20.2	13.1	16.9	21.5
Genebase	2.9	4.5	12.1	9.6	8.5	9.5
Image	5.9	6.1	16.5	9.4	9.5	12.1
MIMLtext	196.8	190.2	269.2	199.8	313.5	295.4
CAL500	15.3	26.3	44.5	18.2	56.5	48.7
Corel5k	539.3	501.2	874.1	485.3	699.5	581.2
Emotions	1.1	12.8	16.4	13.9	14.7	11.5
Education	72.3	174.2	283.2	95.2	89.8	299.1
Languagelog	73.5	108.8	194.1	81.1	95.3	106.9
MediaMill	25520.3	36841.1	66541.2	34215.5	46439.1	76520.3
RCV1V2 (S1)	19541.5	25142.2	43259.1	31451.1	61254.1	53546.5
RCV1V2 (S2)	18649.5	24518.8	43494.8	30999.1	62351.1	53875.2
RCV1V2 (S3)	19549.5	24894.3	44001.5	32019.1	62111.1	52896.4

Convergence Analysis and Parameter Sensitivity

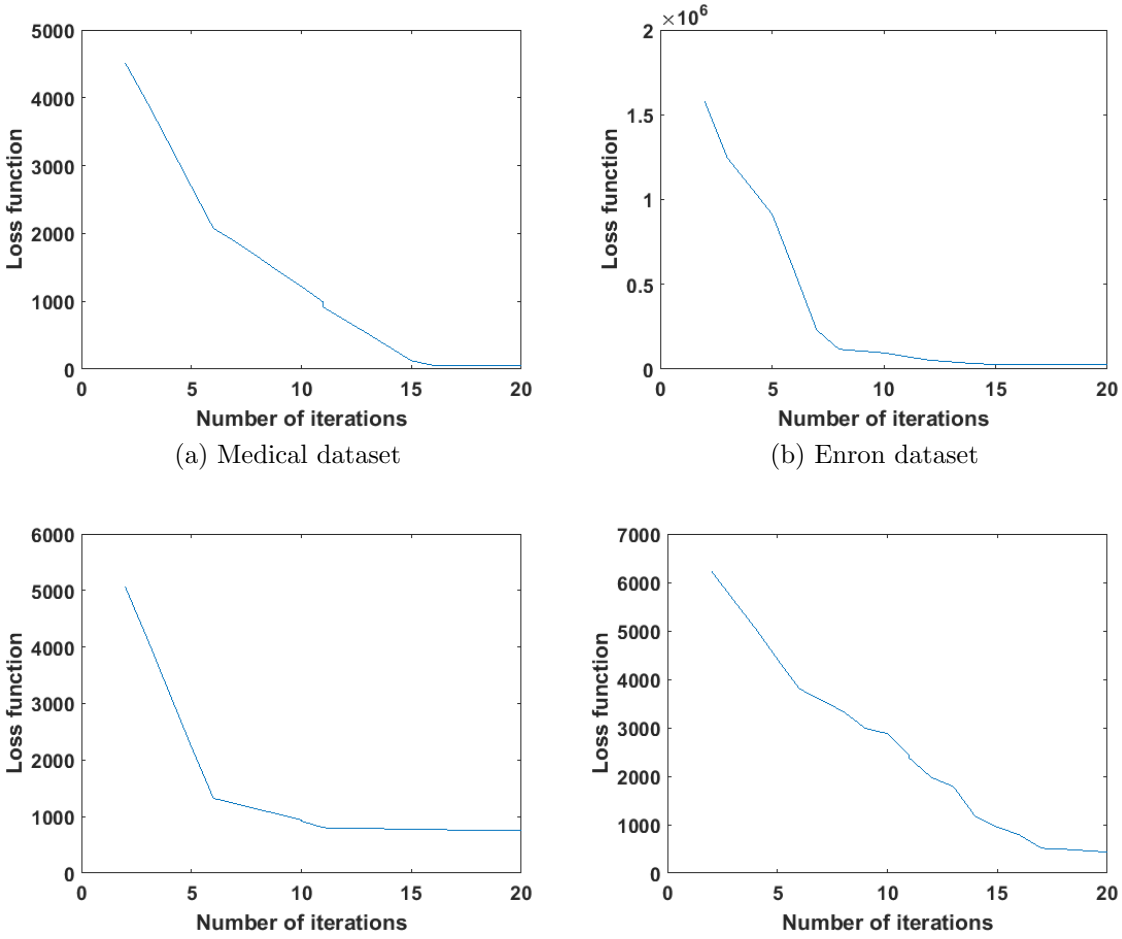
The ML-CIB algorithm uses the accelerated proximal gradient which monotonically decreases its objective function in Eq. 5.2.5 until convergence. Based on the convergence curves in Figure 5.13, it can be seen that the optimization function rapidly converges after 15 to 20 iterations, which demonstrates the efficacy of the proposed algorithm.

This section conducts an analysis to examine the sensitivity of the parameters in the proposed algorithm on four datasets namely, Medical, Enron, Emotions and MIMLtext. ML-CIB has three important parameters: α, β and γ. These parameters control the contribution of the following components: the new label matrix manifold, the difference between the new and original label matrix and the sparsity of the feature matrix W respectively. They are tuned using a "grid-search" strategy from 0.1 to 1

with step size 0.1. This study reports the results the results in terms of the Micro-F1 metric over the Medical, Enron, Emotions and MIMLtext datasets, averaged over five-fold cross validation. The experimental results are shown in Figures 5.14 through 5.17. The results evaluate all the combinations of the grid search line for the three parameters; however, to enable the results to be visualized, one parameter is fixed and show the results of the remainder. The following results are observed: (1) the ML-CIB algorithm is not sensitive to its parameters with wide ranges; (2) in most cases, the classification accuracy is increased by increasing the value of the parameter γ, which can be interpreted as a result of imposing sparsity over the feature matrix W; (3) increasing the values of the parameters α and β leads to an improvement in classification performance, which indicates the importance of the contribution in predicting a new label matrix during the training phase.

5.3 Contribution and Conclusion

This chapter addresses **Contributions 6 and 7** of this thesis by developing a two novel algorithms in multi-label learning. The former proposes a novel multi-label feature selection method which incorporates the correlation information of the features, labels and samples into a unified approach. The CMFS makes use of NMF for data and label decomposition to find the low-dimensional space which guides the feature selection process. Then, to select the predictive features that are shared across multiple labels, it exploits the label and feature correlations. Furthermore, it employs the data similarities to handle local label correlations. Extensive experiments are conducted on high-dimensional real-world multi-labeled datasets. The outcomes show that the proposed algorithm CMFS outperformed the state-of-the-art methods on



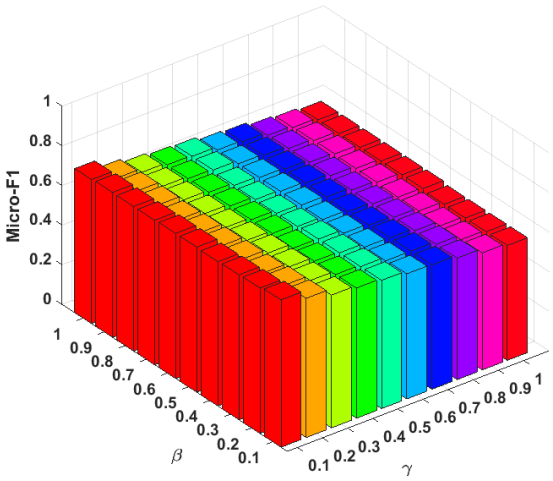
(a) Medical dataset

(b) Enron dataset

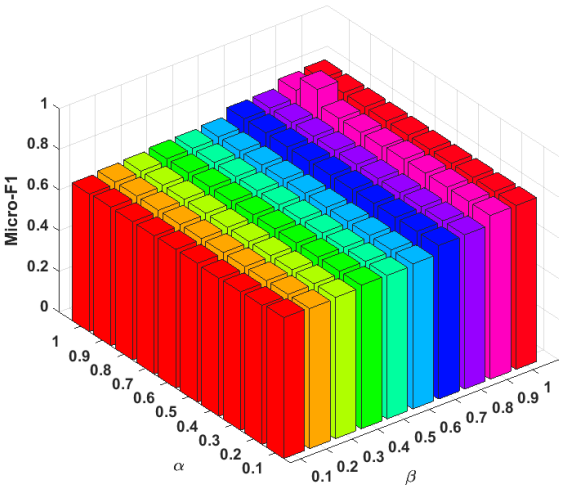
(c) Emotions dataset

(d) MIMLtext dataset

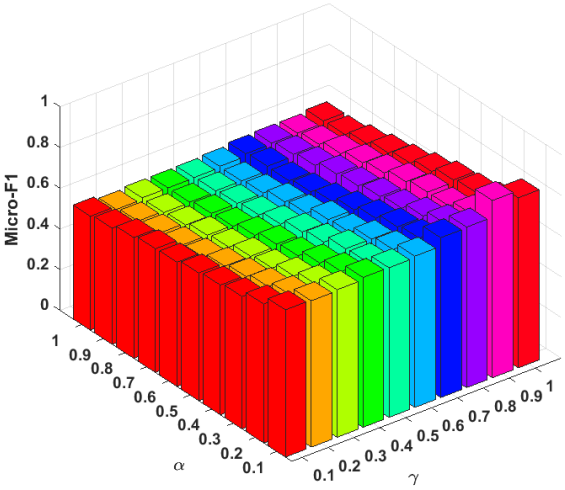
Figure 5.13: Convergence analysis for the proposed algorithm ML-CIB



(a) Optimize β and γ while keeping $\alpha = 0.1$

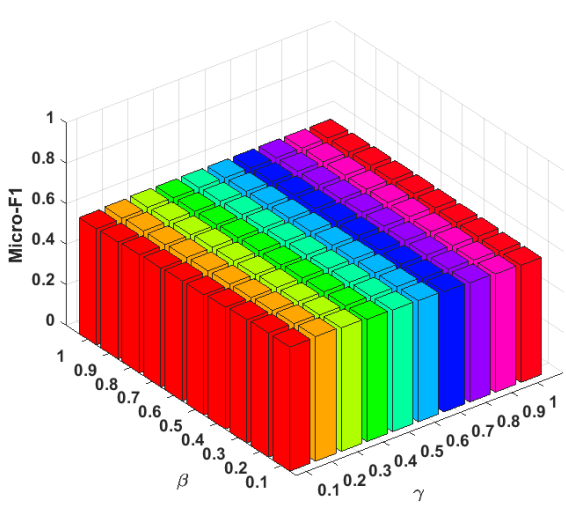


(b) Optimize α and β while keeping $\gamma = 0.1$

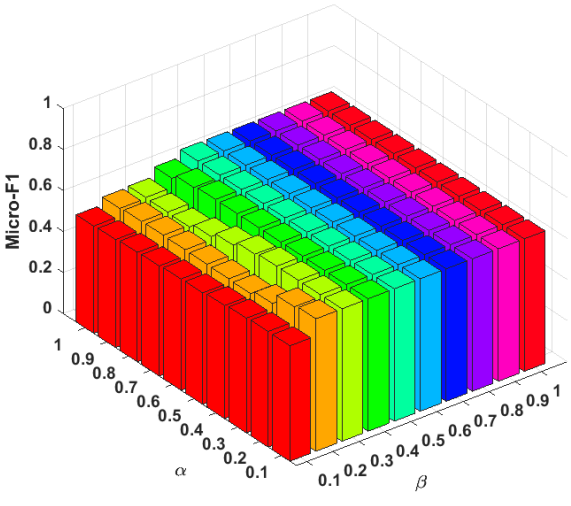


(c) Optimize α and γ while keeping $\beta = 0.1$

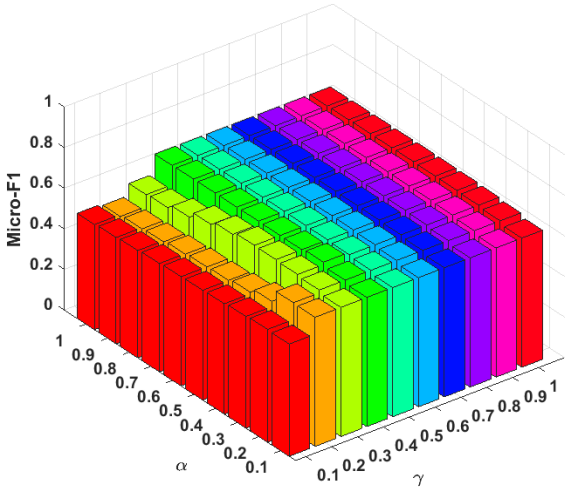
Figure 5.14: Micro-F1 measure results of ML-CIB on the Medical dataset w.r.t different parameter values



(a) Optimize β and γ while keeping $\alpha = 0.1$

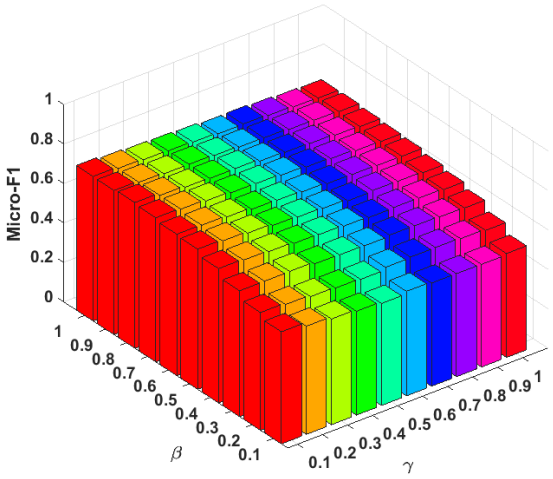


(b) Optimize α and β while keeping $\gamma = 0.1$

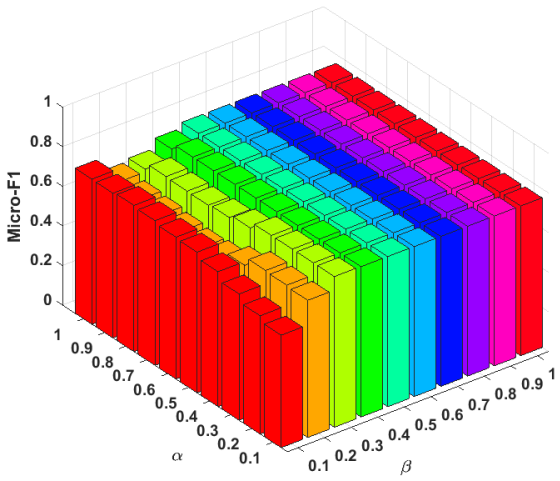


(c) Optimize α and γ while keeping $\beta = 0.1$

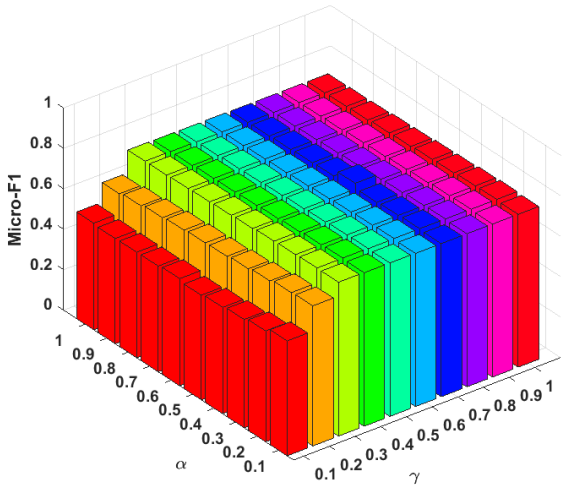
Figure 5.15: Micro-F1 measure results of ML-CIB on the Enron dataset w.r.t different parameter values



(a) Optimize β and γ while keeping $\alpha = 0.1$

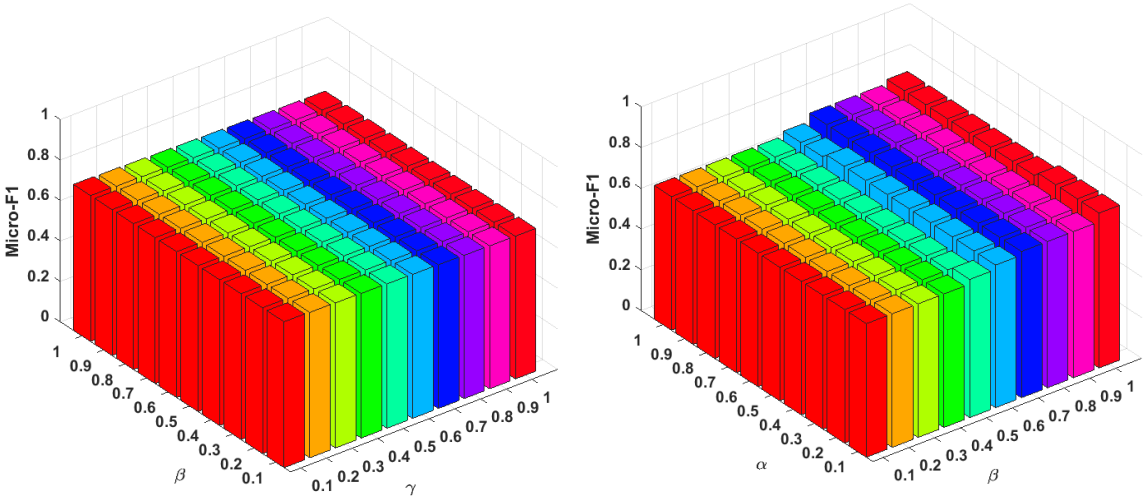


(b) Optimize α and β while keeping $\gamma = 0.1$



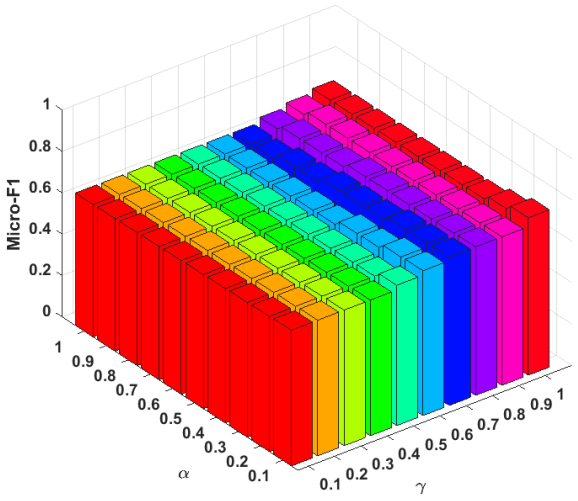
(c) Optimize α and γ while keeping $\beta = 0.1$

Figure 5.16: Micro-F1 measure results of ML-CIB on the Emotions dataset w.r.t different parameter values



(a) Optimize β and γ while keeping $\alpha = 0.1$

(b) Optimize α and β while keeping $\gamma = 0.1$



(c) Optimize α and γ while keeping $\beta = 0.1$

Figure 5.17: Micro-F1 measure results of ML-CIB on the MIMLtext dataset w.r.t different parameter values

most of datasets from different domains. The latter develops a novel an integrated multi-label classification approach which handles three challenges, namely, class imbalance, incomplete multi-label matrix, and the generation of label correlations from a Boolean label matrix. The proposed method is able to learn a new multi-label matrix along with simultaneously capturing the new label correlations while training the multi-label model. It presents a label regularization to handle the class imbalance problem, and l_1 regularization norm is incorporated in the model to select the relevant features. A variant of the proposed method is presented to evaluate multi-label classification performance using the selected features. Extensive experiments are conducted on regular-scale and large-scale imbalanced multi-labeled data, which clearly shows the superiority of the proposed approach over the state-of-the-art methods. It is considered as the first attempt to handle the imbalance class problem in the multi-label feature selection approach.

Chapter 6

Conclusion

Complexities in real-world datasets degrade the performance of standard supervised learning methods for single and multi-labeled data. These problems are still considered as ongoing issues in the data mining community, including high dimensionality, highly correlated features, class imbalance, missing values in multi-label matrix, and label correlations. Current research on single-label classification handling for complex datasets is not well aligned with the new approaches to improving classification performance including supervised feature extraction and exploiting feature correlations. Moreover, current studies in multi-label classification are insufficient because of the existing correlation among features, labels and samples along with missing labels. These make single-label methods inappropriate in the multi-label context. Thus, there is an urgent demand for researching and handling these challenges when dealing with patient's treatment and prognosis for cancers.

The research in this thesis is motivated by the fact that existing classifiers fail to handle the challenges in complex datasets for single and multi-label classification. As reviewed in the literature (chapter 2), performance of the existing classifiers is affected for the following reasons:

- Classification of single-labeled data in the presence of the class imbalance problem produces poor performance on the minor class samples.
- Multi-label classification performance is significantly affected due to label imbalance problem and incomplete label matrix.
- Feature selection methods consider correlated features as redundant features, and they do not investigate the prediction capability of correlated features.
- Reducing high-dimensional multi-labeled data by selecting predictive features is not a trivial task due to the existence of label-label and feature-feature correlations.

In light of these reasons, this thesis asked the following research questions:

RQ1 : Is it possible to propose appropriate methods to handle the class imbalance problem in single-labeled data?

This question is answered by proposing three novel algorithms as follows: a cost-sensitive learning strategy in Section 3.1, a balanced supervised non-negative matrix factorization approach in Section 3.2, and ABC-Sampling in Section 3.3.

RQ2 : Is it possible to develop dimensionality reduction methods to reduce the high dimensional data that do not consider the correlated features as redundant?

Two novel algorithms are proposed to answer this question, namely, supervised context-aware non-negative matrix factorization in Section 4.1, and ensemble feature ranking method using co-expression networks in Section 4.2.

RQ3 : How to improve the multi-label classification performance in the presence

of high dimensional multi-labeled data, label correlations, label imbalance and incomplete label space?

The multi-label classification is improved by developing two novel algorithms, namely, multi-label feature selection using correlation information in Section 5.1, and multi-label classification in the presence of incomplete label space and the class imbalance in Section 5.2.

Based on these research questions, this research makes seven contributions as follows:

Contribution 1: A cost-sensitive learning strategy for feature extraction to handle the class imbalance in single-labeled data

Section 3.1 describes a cost-sensitive learning strategy for feature extraction to handle the class imbalance in single-labeled data. The proposed method presents a weighting variable for the labelling information to be integrated with the existing feature extraction, namely PCA and NMF. This weighting strategy aims to alleviate the bias of the selected features towards the majority class in single-labeled data, which increases the performance of the classification process. The performance of the proposed method in this thesis was evaluated on datasets from different domains and compared to state-of-the-art methods. The results demonstrate statistical superiority of the proposed method compared to the state-of-the-art.

Contribution 2: A balanced supervised non-negative matrix factorization approach to generate features which are not biased to predict the majority class.

Section 3.2 presents a balanced supervised non-negative matrix factorization to generate features which are not biased to predicting the majority class. This study addresses the class imbalance problem in a supervised NMF method using the coupled matrix factorization approach. Moreover, it presents a feature ranking method based on the factorised basis matrix. The selected features are able to predict the minority class which is the class of interest. The experimental results suggest that the proposed method is effective in improving the classification performance on a complex dataset such as genomics dataset. Furthermore, the classification results outperform the state-of-the-art.

Contribution 3: ABC-Sampling for balancing imbalanced datasets based on Artificial Bee Colony algorithm.

Section 3.3 proposes a wrapper method called ABC-Sampling for balancing imbalanced datasets based on Artificial Bee Colony algorithm. This research extends the robust optimization method artificial bee colony to select the optimal majority samples and remove the rest. The optimal majority samples are those samples that achieve the best classification performance in the fitness function of the optimization method. The proposed method is an undersampling approach which reduces the number of majority samples to balance the imbalanced data. Evaluation on different levels of imbalance ratio demonstrating that the proposed method achieves a high classification performance compared to the state-of-the-art.

Contribution 4: Supervised context-aware non-negative matrix factorization to handle high-dimensional, highly correlated single-labeled data.

Section 4.1 develops a supervised context-aware non-negative matrix factorization to handle high-dimensional, highly correlated single-labeled data. The proposed method simultaneously integrates the correlation structure of the features using the adjacency matrix along with the prior knowledge of the class labels and the original features to the data decomposition technique NMF. Then, $l_{2,1}$ -norm regularisation is incorporated into the optimization function for feature selection, due to its robustness to outliers. The promising experimental results on seven high-dimensional high-correlated imbalanced real-world datasets demonstrate that the proposed method is superior compared to the state-of-the-art.

Contribution 5: Ensemble feature ranking method using co-expression networks to select optimal features for classification.

Section 4.2 proposes an ensemble feature ranking method using co-expression networks to select optimal features for classification. The proposed algorithm is composed of two phases. First, to capture the correlated features, it uses the WGCNA method to cluster the correlated features into groups and ranks the features within the groups using the ESVM-RFE method. Second, it selects the most appropriate number of features from each group, aggregates and ranks them by using ESVM-RFE to select the best features between the feature groups. The proposed approach is an embedded feature ranking method to achieve feature selection in the presence of the correlated features. Evaluation shows the statistical superiority of the proposed method compared to the state-of-the-art feature selection methods based on the classification accuracy on high dimensional, highly correlated datasets. Moreover, the proposed

method demonstrates the capability of interpretation for a biological dataset showing the importance of integrating the correlated features in the proposed method.

Contribution 6: Multi-label feature selection using correlation information to select the optimal features in multi-labeled data.

Section 5.1 presents a multi-label feature selection method using correlation information to select the optimal features in multi-labeled data. This approach handles high-dimensional multi-labeled data taking into consideration feature-feature correlations and label-label correlations. The proposed method is an optimisation method in the context of NMF. The low rank feature matrix finds relevant features able to capture dependencies between multiple labels. Furthermore, the proposed method defines a new label matrix which predicts the missing labels for samples. Classification results on high-dimensional real-world multi-labeled datasets shows that the proposed method outperforms the state-of-the-art.

Contribution 7: Improving multi-label classification in the presence of incomplete label space and the class imbalance problem.

Section 5.2 develops a multi-label classification method in the presence of incomplete label space and the class imbalance problem. The proposed method is an integrated approach addressing three challenges in multi-label classification, namely the class imbalance problem, incomplete label matrix and generating label correlations from boolean matrix. The l_1 regularisation norm is imposed for feature selection. The three components are combined in an

optimization method which is solved using the accelerated proximal gradient method. The proposed algorithm was evaluated on high-dimensional imbalanced multi-labeled datasets. The results demonstrate the statistical superiority of the proposed method compared to the state-of-the-art. Furthermore, the proposed method is used as feature selection method. It outperforms the state-of-the-art feature selection methods.

To sum up, the proposed methods of addressing the data mining challenges including high dimensionality, highly correlated features, label correlations, and imbalanced data have potential benefits for improving the single and multi-label classification in real-world datasets. The proposed methods are applicable on real-world datasets from different domains including biomedicine, text, and images.

6.1 Future Research Directions

This thesis presents a series of novel algorithms to improve the classification performance of single and multi-labeled data. The following research plan seeks to apply the proposed methods on specific domains and overcome limitations of these methods.

6.1.1 Interpretation of Correlated Features in Biology

One potential research direction is to investigate the benefits of the proposed methods that exploit feature correlations by applying them on specific domains including SCANMF (Section 4.1), Fuzz-ESVM (Section 4.2) and CMFS (Section 5.1). In genomics, gene products often work in specific pathways. One gene may interact with multiple other genes that link it to different specific pathways and upstream/ downstream mechanisms. Hence, the groups of highly interconnected genes are often

closely linked to specific functional categories. In future research, it is reasonable to evaluate the proposed methods in this thesis namely SCANMF, Fuzz-ESVM, and CMFS on biomedical data to highlight the biological interpretations. Particularly, to infer the semantic components of the combined genes to find common pathways associated with the disease.

6.1.2 Investigating the Features for Multi-Layers of Non-Negative Matrix Factorization

Another research direction is to investigate feature selection on multi-layer non-negative matrix factorization which provides a hierarchical feature selection process. Complex data contains multiple layers, in each layer, there are a small number of optimal features that can be selected which contribute to predict the output. For example, to predicting a human face, a face image contains several attributes in different layers such as pose features, expression features and identity features. Therefore, it is important to select features for pose, expression and identity layer.

6.1.3 Interpretation of Correlated Features in Social Network

The proposed methods in this research that are based on NMF decomposition can be applied in the social network domain due to the non-negativity constraint on NMF which facilitates an elegant interpretation.

6.1.4 Incorporating External Sources to Capture Label Correlation

This thesis only considers the correlation between labels based on the multi-label matrix. This constrains the method to noisy labels due to manual labeling. It may

be interesting in future work to integrate external sources of labels such as ontologies of document categorization in text mining to assist in predicting multiple labels of new observations along with rectifying the erroneous labels on the existing multi-label matrix.

6.1.5 Extending Multi-Label Feature Selection Method to Consider Local Label Correlation

The research in this thesis partially addresses the local label correlations in multi-labeled data, where the labels are correlated in some samples, but not in all. It will be an important research direction to investigate explicitly the local label correlations issue by proposing or extending the proposed methods in this thesis.

6.1.6 Extending ABC-Sampling to oversampling strategy

This thesis proposes ABC-Sampling using undersampling strategy. In future work, the proposed method will be extended to work with oversampling strategy and it will present the suggestions by comparing the results with undersampling strategy.

6.2 Conclusion

To conclude, handling the data mining challenges in single and multi-label classification, including high-dimensionality, highly correlated features, class imbalance, missing labels, and label correlations is becoming a significant research area. It helps scientists to properly identify groups of new observations in specific domains. This thesis developed a series of novel algorithms to address the fundamental challenges of single and multi-labeld data to provide the research community with novel methods to solve these challenges.

Appendix

The Proof of Convergence of BSNMF

This section presents the convergence analysis and proving that objective function **O** is non-increasing under the updating rules in Eq. 3.2.11, 3.2.12 and 3.2.13.

Theorem: *The objective function* ***O*** *of BSNMF in Eq. 3.2.2 is non-increasing under the update steps in Eq. 3.2.11, 3.2.12, and 3.2.13.*

To prove this Theorem, the update steps in Eq. 3.2.12, and 3.2.13 follow the similar procedure to prove their convergence as described in the original NMF paper (Lee & Seung 2001). This section proves that the objective function in Eq. 3.2.2 is non-increasing under the update the part related to U only in Eq. 3.2.11.

The auxiliary function is used to prove the theorem above which based on Expectation-Maximization algorithm (Dempster et al. 1977).

Definition: $G(u, u')$ is an auxiliary function for $F(u)$ if the conditions below are satisfied

$$G(u, u') \geq F(u), \quad G(u, u) = F(u)$$

The following Lemma makes the auxiliary function very useful.

Lemma 1: If G is an auxiliary function for F, therefore F is non-increasing under

the following update step

$$u^{t+1} = \min_{u} G\left(u, u^{t}\right) \tag{6.2.1}$$

Proof: $F\left(u^{t+1}\right) \leq G\left(u^{t+1}, u^{t}\right) \leq G\left(u^{t}, u^{t}\right) = F\left(u^{t}\right)$

In the following, the update step is proved in Eq. 3.2.11 for U is exactly the update in Eq. 6.2.1 with a proper auxiliary function.

Our objective function $\mathbf{O}$ for BSNMF algorithm in Eq. 3.2.2 as follows:

$$\begin{aligned}\mathbf{O} = \sigma^{-}\left\|X^{-} - U(V^{-})^{T}\right\|^{2} + \sigma^{+}\left\|X^{+} - U(V^{+})^{T}\right\|^{2} \\ = \sum_{i=1}^{M}\sum_{j=1}^{N^{-}}\sigma_{ij}^{-}\left(x_{ij}^{-} - \sum_{k=1}^{K}u_{ik}v_{jk}^{-}\right) + \\ \sum_{i=1}^{M}\sum_{j=N^{-}+1}^{N}\sigma_{ij}^{+}\left(x_{ij}^{+} - \sum_{k=1}^{K}u_{ik}v_{jk}^{+}\right)\end{aligned}$$

Consider element u_{ab} in U, F_{ab} is used to demonstrate the part of $\mathbf{O}$ which is only relevant to u_{ab}, it is easy to generate the first order derivative from the update step at point point u_{ab}:

$$\begin{aligned}F'_{ab} = \frac{\partial \mathbf{O}}{\partial U} = (-2\sigma^{-}X^{-}(V^{-})^{T} + 2\sigma^{-}U(V^{-})^{T}V^{-} \\ -2\sigma^{+}X^{+}(V^{+})^{T} + 2\sigma^{+}U(V^{+})^{T}V^{+})_{ab}\end{aligned} \tag{6.2.2}$$

and the second order derivative:

$$F''_{ab} = \frac{\partial F'}{\partial U} = (\sigma^{-}V^{-}(V^{-})^{T})_{bb} + (\sigma^{+}V^{+}(V^{+})^{T})_{aa} \tag{6.2.3}$$

The update step is element-wise, so, it is sufficient to show that each F_{ab} is non-increasing under each step in Eq. 3.2.11.

Lemma 2:

$$\begin{aligned}G\left(u, u_{ab}^{t}\right) = F_{ab}\left(u_{ab}^{t}\right) + F'_{ab}\left(u_{ab}^{t}\right)\left(u - u_{ab}^{t}\right) + \\ \left(u - u_{ab}^{t}\right)^{2}\frac{\left[\sigma^{-}U(V^{-})^{T}V^{-} + \sigma^{+}U(V^{+})^{T}V^{+}\right]}{u_{ab}^{t}}\end{aligned} \tag{6.2.4}$$

$G\left(u, u_{ab}^{t}\right)$ is an auxiliary function of F_{ab} on the part which relevant to u_{ab}.

Proof: It is obvious that $G\left(u, u\right) = F\left(u\right)$, it needs only to prove that $G\left(u, u_{ab}^{t}\right) \geq F_{ab}\left(u\right)$. To achieve this, the Taylor series expansion of $F_{ab}(u)$ is compared as

$$\begin{aligned} F_{ab}\left(u\right) &= F_{ab}\left(u_{ab}^{t}\right) F_{ab}'\left(u_{ab}^{t}\right) + \left(u - u_{ab}\right)^{2} F_{ab}''\left(u_{ab}^{t}\right) \\ &= F_{ab}\left(u\right) = F_{ab}\left(u_{ab}^{t}\right) F_{ab}'\left(u_{ab}^{t}\right) + \\ &\left(u - u_{ab}\right)^{2} \left[\left(\sigma^{-} V^{-} (V^{-})^{T}\right)_{bb} + \left(\sigma^{+} V^{+} (V^{+})^{T}\right)_{aa}\right] \end{aligned} \tag{6.2.5}$$

To find $G\left(u, u_{ab}^{t}\right) \geq F_{ab}\left(u\right)$ with Eq. 6.2.4 is equivalent to

$$\begin{aligned} &\frac{\left[\left(\sigma^{-} U V^{-} (V^{-})^{T}\right)_{ab} + \left(\sigma^{+} U V^{+} (V^{+})^{T}\right)_{ab}\right]}{u_{ab}^{t}} \\ &\geq F_{ab}'' = \left(\sigma^{-} V^{-} (V^{-})^{T}\right)_{bb} + \left(\sigma^{+} V^{+} (V^{+})^{T}\right)_{aa} \end{aligned} \tag{6.2.6}$$

We have

$$\left(\sigma^{-} U V^{-} (V^{-})^{T}\right)_{ab} = \sum_{l=1}^{K} u_{al}^{t} \left(\sigma^{-} V^{-} (V^{-})^{T}\right)_{lb} \geq u_{ab}^{t} \left(\sigma^{-} V^{-} (V^{-})^{T}\right)_{bb} \tag{6.2.7}$$

and

$$\left(\sigma^{+} U V^{+} (V^{+})^{T}\right)_{ab} = \sum_{l=1}^{K} u_{al}^{t} \left(\sigma^{+} V^{+} (V^{+})^{T}\right)_{lb} \geq u_{ab}^{t} \left(\sigma^{+} V^{+} (V^{+})^{T}\right)_{aa} \tag{6.2.8}$$

Thus, Eq. 6.2.6 holds on $G\left(u, u_{ab}^{t}\right) \geq F_{ab}\left(u\right)$, the convergence of Theorem is now indicated as:

Proof of Theorem: The results obtained below by replacing $G\left(u, u_{ab}^{t}\right)$ in Eq. 6.2.1 with Eq. 6.2.4:

$$\begin{aligned} u_{ab}^{t+1} &= u_{ab}^{t} - u_{ab}^{t} \frac{F_{ab}'\left(u_{ab}^{t}\right)}{\left[\sigma^{-} U (V^{-})^{T} V^{-} + \sigma^{+} U (V^{+})^{T} V^{+}\right]_{ab}} \\ &= u_{ab}^{t} \frac{\left[\sigma^{-} X^{-} (V^{-})^{T}\right]_{ab} + \left[\sigma^{+} X^{+} (V^{+})^{T}\right]_{ab}}{\left[\sigma^{-} U V^{-} (V^{-})^{T}\right]_{ab} + \left[\sigma^{+} U V^{+} (V^{+})^{T}\right]_{ab}} \end{aligned} \tag{6.2.9}$$

Since Eq. 6.2.4 is an auxiliary function, F is non-increasing under the update rule in Eq. 3.2.11.

Bibliography

Alcalá-Fdez, J., Fernández, A., Luengo, J., Derrac, J., García, S., Sánchez, L. & Herrera, F. (2010), 'Keel data-mining software tool: Data set repository', *Journal of Multiple-Valued Logic and Soft Computing* **17**(2-3), 255–287.

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D. & Levine, A. J. (1999), 'Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays', *Proceedings of the National Academy of Sciences* **96**(12), 6745–6750.

Ambroise, C. & McLachlan, G. J. (2002), 'Selection bias in gene extraction on the basis of microarray gene-expression data', *Proceedings of the National Academy of Sciences* **99**(10), 6562–6566.

Anaissi, A., Goyal, M., Catchpoole, D. R., Braytee, A. & Kennedy, P. J. (2016), 'Ensemble feature learning of genomic data using support vector machine', *PloS One* **11**(6), e0157330.

Anaissi, A., Kennedy, P. J., Goyal, M. & Catchpoole, D. R. (2013), 'A balanced iterative random forest for gene selection from microarray data', *BMC Bioinformatics* **14**(1), 261.

Barutcuoglu, Z., Schapire, R. E. & Troyanskaya, O. G. (2006), 'Hierarchical multi-label prediction of gene function', *Bioinformatics* **22**(7), 830–836.

Batista, G. E., Prati, R. C. & Monard, M. C. (2004), 'A study of the behavior of several methods for balancing machine learning training data', *ACM SIGKDD Explorations Newsletter* **6**(1), 20–29.

Beck, A. & Teboulle, M. (2009), 'A fast iterative shrinkage-thresholding algorithm for linear inverse problems', *SIAM Journal on Imaging Sciences* **2**(1), 183–202.

Bee, A., Ke, Y., Forootan, S., Lin, K., Beesley, C., Forrest, S. E. & Foster, C. S. (2006), 'Ribosomal protein l19 is a prognostic marker for human prostate cancer', *Clinical Cancer Research* **12**(7), 2061–2065.

Bellman, R. (2003), *Dynamic Programming*, Dover Books on Computer Science Series, Dover Publications.
URL: *https://books.google.com.au/books?id=fyVtp3EMxasC*

Berry, M. W. & Castellanos, M. (2008), *Survey of text mining II*, Vol. 6, Springer.

Bin, Z. & Steve, H. (2005), 'A general framework for weighted gene co-expression network analysis', *Statistical Applications in Genetics and Molecular Biology* **4**(1), 1128.

Bishop, C. M. (2006), *Pattern Recognition and Machine Learning*, Springer.

Bonabeau, E., Dorigo, M. & Theraulaz, G. (1999), *Swarm intelligence: from natural to artificial systems*, Oxford University Press.

Bondell, H. D. & Reich, B. J. (2008), 'Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar', *Biometrics* **64**(1), 115–123.

Boser, B. E., Guyon, I. M. & Vapnik, V. N. (1992), A training algorithm for optimal margin classifiers, *in* 'Proceedings of the Fifth Annual Workshop on Computational Learning Theory', ACM, pp. 144–152.

Boutell, M. R., Luo, J., Shen, X. & Brown, C. M. (2004), 'Learning multi-label scene classification', *Pattern Recognition* **37**(9), 1757–1771.

Braytee, A., Catchpoole, D. R., Kennedy, P. J. & Liu, W. (2016), Balanced supervised non-negative matrix factorization for childhood leukaemia patients, *in* 'Proceedings of the 25th ACM International on Conference on Information and Knowledge Management', ACM, pp. 2405–2408.

Braytee, A., Hussain, F. K., Anaissi, A. & Kennedy, P. J. (2015), ABC-Sampling for balancing imbalanced datasets based on artificial bee colony algorithm, *in* 'IEEE 14th International Conference on Machine Learning and Applications (ICMLA), 2015', IEEE, pp. 594–599.

Braytee, A., Liu, W., Catchpoole, D. R. & Kennedy, P. J. (2017), Multi-label feature selection using correlation information, *in* 'Proceedings of the 26th ACM International on Conference on Information and Knowledge Management', ACM.

Braytee, A., Liu, W. & Kennedy, P. (2016), A cost-sensitive learning strategy for feature extraction from imbalanced data, *in* 'International Conference on Neural Information Processing', Springer, pp. 78–86.

Braytee, A., Liu, W. & Kennedy, P. J. (2017), Supervised context-aware non-negative matrix factorization to handle high-dimensional high-correlated imbalanced biomedical data, *in* 'International Joint Conference on Neural Networks (IJCNN), 2017', IEEE, pp. 4512–4519.

Bühlmann, P., Rütimann, P., van de Geer, S. & Zhang, C.-H. (2013), 'Correlated variables in regression: clustering and sparse estimation', *Journal of Statistical Planning and Inference* **143**(11), 1835–1858.

Carmona-Saez, P., Pascual-Marqui, R. D., Tirado, F., Carazo, J. M. & Pascual-Montano, A. (2006), 'Biclustering of gene expression data by non-smooth non-negative matrix factorization', *BMC Bioinformatics* **7**(1), 1.

Catania, C. A., Bromberg, F. & Garino, C. G. (2012), 'An autonomous labeling approach to support vector machines algorithms for network traffic anomaly detection', *Expert Systems with Applications* **39**(2), 1822–1829.

Chang, C.-C. & Lin, C.-J. (2011), 'LIBSVM: a library for support vector machines', *ACM Transactions on Intelligent Systems and Technology (TIST)* **2**(3), 27.

Chapelle, O., Scholkopf, B. & Zien, A. (2009), 'Semi-supervised learning', *IEEE Transactions on Neural Networks* **20**(3), 542–542.

Charte, F., Rivera, A. J., del Jesus, M. J. & Herrera, F. (2015*a*), 'Addressing imbalance in multilabel classification: Measures and random resampling algorithms', *Neurocomputing* **163**, 3–16.

Charte, F., Rivera, A. J., del Jesus, M. J. & Herrera, F. (2015*b*), 'MLSMOTE: Approaching imbalanced multilabel learning through synthetic instance generation', *Knowledge-Based Systems* **89**, 385–397.

Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. (2002), 'SMOTE: synthetic minority over-sampling technique', *Journal of Artificial Intelligence Research* pp. 321–357.

Chawla, N. V., Japkowicz, N. & Kotcz, A. (2004), 'Special issue on learning from imbalanced data sets', *ACM SIGKDD Explorations Newsletter* **6**(1), 1–6.

Chen, K., Lu, B.-L. & Kwok, J. T. (2006), Efficient classification of multi-label and imbalanced data using min-max modular classifiers, *in* 'International Joint Conference on Neural Networks (IJCNN), 2006', IEEE, pp. 1770–1775.

Chen, X., Wang, L., Smith, J. D. & Zhang, B. (2008), 'Supervised principal component analysis for gene set enrichment of microarray data with continuous or survival outcomes', *Bioinformatics* **24**(21), 2474–2481.

Cherman, E. A., Monard, M. C. & Metz, J. (2011), 'Multi-label problem transformation methods: a case study', *CLEI Electronic Journal* **14**(1), 4.

Cieslak, D. A. & Chawla, N. V. (2008), Learning decision trees for unbalanced data, *in* 'Machine Learning and Knowledge Discovery in Databases', Springer, pp. 241–256.

Conn, D., Ngun, T., Li, G. & Ramirez, C. (2015), 'Fuzzy Forests: Extending random forests for correlated, high-dimensional data'.

Cover, T. & Hart, P. (1967), 'Nearest neighbor pattern classification', *IEEE Transactions on Information Theory* **13**(1), 21–27.

Cunningham, P. & Delany, S. J. (2007), 'k-nearest neighbour classifiers', *Multiple Classifier Systems* pp. 1–17.

Dembczynski, K., Jachnik, A., Kotlowski, W., Waegeman, W. & Hüllermeier, E. (2013), 'Optimizing the F-Measure in multi-label classification: Plug-in rule approach versus structured loss minimization.', *International Conference on Machine Learning* **28**, 1130–1138.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977), 'Maximum likelihood from incomplete data via the em algorithm', *Journal of the royal statistical society. Series B (methodological)* pp. 1–38.

Demšar, J. (2006), 'Statistical comparisons of classifiers over multiple data sets', *Journal of Machine Learning Research* **7**(Jan), 1–30.

Dendamrongvit, S. & Kubat, M. (2009), Undersampling approach for imbalanced training sets and induction from multi-label text-categorization domains, *in*

'Pacific-Asia Conference on Knowledge Discovery and Data Mining', Springer, pp. 40–52.

Dettling, M. & Bühlmann, P. (2004), 'Finding predictive gene groups from microarray data', *Journal of Multivariate Analysis* **90**(1), 106–131.

Devarajan, K. (2008), 'Nonnegative matrix factorization: an analytical and interpretive tool in computational biology', *PLoS Comput Biol* **4**(7), e1000029.

Diaz, N. N., Krause, L., Goesmann, A., Niehaus, K. & Nattkemper, T. W. (2009), 'Tacoa–taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach', *BMC Bioinformatics* **10**(1), 56.

Ding, C., Li, T., Peng, W. & Park, H. (2006), Orthogonal nonnegative matrix t-factorizations for clustering, *in* 'Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', ACM, pp. 126–135.

Dmochowski, J. P., Sajda, P. & Parra, L. C. (2010), 'Maximum likelihood in cost-sensitive learning: Model specification, approximations, and upper bounds', *The Journal of Machine Learning Research* **11**, 3313–3332.

Drummond, C. & Holte, R. C. (2003), C4.5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling, *in* 'Workshop on Learning from Imbalanced Datasets II', Vol. 11, Citeseer.

Duda, R. O., Hart, P. E. & Stork, D. G. (2012), *Pattern classification*, John Wiley & Sons.

Dumais, S. T. (2004), 'Latent semantic analysis', *Annual Review of Information Science and Technology* **38**(1), 188–230.

Elkan, C. (2001), The foundations of cost-sensitive learning, *in* 'International Joint Conference on Artificial Intelligence', Vol. 17, Citeseer, pp. 973–978.

Ferrari, S., Manfredini, R., Tagliafico, E., Rossi, E., Donelli, A., Torelli, G. & Torelli, U. (1990), 'Noncoordinated expression of S6, S11, and S14 ribosomal protein genes in leukemic blast cells', *Cancer Research* **50**(18), 5825–5828.

García, V., Sánchez, J. S. & Mollineda, R. A. (2012), 'On the effectiveness of preprocessing methods when dealing with different levels of class imbalance', *Knowledge-Based Systems* **25**(1), 13–21.

Ghamrawi, N. & McCallum, A. (2005), Collective multi-label classification, *in* 'Proceedings of the 14th ACM International Conference on Information and Knowledge Management', ACM, pp. 195–200.

Giraldo-Forero, A. F., Jaramillo-Garzón, J. A., Ruiz-Muñoz, J. F. & Castellanos-Domínguez, C. G. (2013), Managing imbalanced data sets in multi-label problems: a case study with the SMOTE algorithm, *in* 'Iberoamerican Congress on Pattern Recognition', Springer, pp. 334–342.

Godbole, S. & Sarawagi, S. (2004), Discriminative methods for multi-labeled classification, *in* 'Pacific-Asia Conference on Knowledge Discovery and Data Mining', Springer, pp. 22–30.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A. et al. (1999), 'Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.', *Science* **286**(5439), 531–537.

Gordon, G. J., Jensen, R. V., Hsiao, L.-L., Gullans, S. R., Blumenstock, J. E., Ramaswamy, S., Richards, W. G., Sugarbaker, D. J. & Bueno, R. (2002), 'Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma', *Cancer Research* **62**(17), 4963–4967.

Gu, Q., Li, Z. & Han, J. (2011), Correlated multi-label feature selection, *in* 'Proceedings of the 20th ACM International Conference on Information and Knowledge Management', ACM, pp. 1087–1096.

Guo, H., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H. & Bing, G. (2017), 'Learning from class-imbalanced data: Review of methods and applications', *Expert Systems with Applications* **73**, 220–239.

Guyon, I. & Elisseeff, A. (2003), 'An introduction to variable and feature selection', *Journal of Machine Learning Research* **3**(Mar), 1157–1182.

Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. (2002), 'Gene selection for cancer classification using support vector machines', *Machine Learning* **46**(1), 389–422.

Han, H., Wang, W.-Y. & Mao, B.-H. (2005), Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning, *in* 'Advances in Intelligent Computing', Springer, pp. 878–887.

He, H., Bai, Y., Garcia, E. A. & Li, S. (2008), ADASYN: Adaptive synthetic sampling approach for imbalanced learning, *in* 'IEEE International Joint Conference on Neural Networks (IJCNN), 2008', IEEE, pp. 1322–1328.

He, H. & Garcia, E. A. (2009), 'Learning from imbalanced data', *IEEE Transactions on Knowledge and Data Engineering* **21**(9), 1263–1284.

Hoens, T. R., Polikar, R. & Chawla, N. V. (2012), 'Learning from streaming data with concept drift and imbalance: an overview', *Progress in Artificial Intelligence* **1**(1), 89–101.

Holland, J. H. (1992), *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*, MIT press.

Horvath, S. (2011), *Weighted network analysis: applications in genomics and systems biology*, Springer Science & Business Media.

Horvath, S. & Dong, J. (2008), 'Geometric interpretation of gene coexpression network analysis', *PLoS Computational Biology* **4**(8), e1000117.

Hsu, D. J., Kakade, S. M., Langford, J. & Zhang, T. (2009), Multi-label prediction via compressed sensing, *in* 'Advances in neural information processing systems', Vol. 22, pp. 772–780.

Huang, H.-H., Liu, X.-Y. & Liang, Y. (2016), 'Feature selection and cancer classification via sparse logistic regression with the hybrid l1/2+ 2 regularization', *PloS One* **11**(5), e0149675.

Huang, J., Li, G., Huang, Q. & Wu, X. (2016), 'Learning label-specific features and class-dependent labels for multi-label classification', *IEEE Transactions on Knowledge and Data Engineering* **28**(12), 3309–3323.

Huang, S.-J., Zhou, Z.-H. & Zhou, Z. (2012), Multi-label learning by exploiting label correlations locally., *in* 'Association for the Advancement of Artificial Intelligence', pp. 949–955.

Huh, S.-i., Gupta, M. D. & Xiao, J. (2013), 'Supervised nonnegative matrix factorization'. US Patent 8,498,949.

Ivliev, A. E., AC't Hoen, P. & Sergeeva, M. G. (2010), 'Coexpression network analysis identifies transcriptional modules related to proastrocytic differentiation and sprouty signaling in glioma', *Cancer Research* **70**(24), 10060–10070.

Japkowicz, N. (2000), Learning from imbalanced data sets: a comparison of various strategies, *in* 'AAAI Workshop on Learning from Imbalanced Datasets', Vol. 68, Menlo Park, CA, pp. 10–15.

Jeanne, R. L. (1986), 'The evolution of the organization of work in social insects', *Monitore Zoologico Italiano-Italian Journal of Zoology* **20**(2), 119–133.

Ji, S., Tang, L., Yu, S. & Ye, J. (2008), Extracting shared subspace for multi-label classification, *in* 'Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', ACM, pp. 381–389.

Jia, Y. W. Y. & Turk, C. H. M. (2004), Fisher non-negative matrix factorization for learning local features, *in* 'Proceedings Asian Conference on Computer Vision (ACCV)', pp. 27–30.

Jia, Z., Zhang, X., Guan, N., Bo, X., Barnes, M. R. & Luo, Z. (2015), 'Gene ranking of RNA-Seq data via discriminant non-negative matrix factorization', *PloS One* **10**(9), e0137782.

Jian, L., Li, J., Shu, K. & Liu, H. (2016), Multi-label informed feature selection., *in* 'Proceedings of the 25th International Joint Conference on Artificial Intelligence', pp. 1627–1633.

Karaboga, D. & Basturk, B. (2007), 'A powerful and efficient algorithm for numerical function optimization: artificial bee colony (abc) algorithm', *Journal of global optimization* **39**(3), 459–471.

Kavzoglu, T. & Colkesen, I. (2009), 'A kernel functions analysis for support vector machines for land cover classification', *International Journal of Applied Earth Observation and Geoinformation* **11**(5), 352–359.

Kennedy, J. (2011), Particle swarm optimization, *in* 'Encyclopedia of Machine Learning', Springer, pp. 760–766.

Khan, A., Baharudin, B., Lee, L. H. & Khan, K. (2010), 'A review of machine learning algorithms for text-documents classification', *Journal of Advances in Information Technology* **1**(1), 4–20.

Koo, C. L., Liew, M. J., Mohamad, M. S. & Mohamed Salleh, A. H. (2013), 'A review for detecting gene-gene interactions using machine learning methods in genetic epidemiology', *BioMed Research International* **2013**.

Kotsiantis, S., Kanellopoulos, D., Pintelas, P. et al. (2006), 'Handling imbalanced datasets: A review', *GESTS International Transactions on Computer Science and Engineering* **30**(1), 25–36.

Krempl, G., Žliobaite, I., Brzeziński, D., Hüllermeier, E., Last, M., Lemaire, V., Noack, T., Shaker, A., Sievi, S., Spiliopoulou, M. et al. (2014), 'Open challenges for data stream mining research', *ACM SIGKDD Explorations Newsletter* **16**(1), 1–10.

Kuang, D., Ding, C. & Park, H. (2012), Symmetric nonnegative matrix factorization for graph clustering, *in* 'Proceedings of the 2012 SIAM international conference on data mining', SIAM, pp. 106–117.

Kubat, M. & Matwin, S. (1997), Addressing the curse of imbalanced training sets: one-sided selection, *in* 'The Fourteenth International Conference on Machine Learning', Vol. 97, Nashville, USA, pp. 179–186.

Lee, D. D. & Seung, H. S. (1999), 'Learning the parts of objects by non-negative matrix factorization', *Nature* **401**(6755), 788–791.

Lee, D. D. & Seung, H. S. (2001), Algorithms for non-negative matrix factorization, *in* 'Advances in Neural Information Processing Systems', pp. 556–562.

Lee, J. & Kim, D.-W. (2015), 'Fast multi-label feature selection based on information-theoretic feature ranking', *Pattern Recognition* **48**(9), 2761–2771.

Li, C. & Shi, G. (2013), 'Improvement of learning algorithm for the multi-instance multi-label rbf neural networks trained with imbalanced samples.', *Journal of Information Science and Engineering* **29**(4), 765–776.

Li, Y. & Ngom, A. (2013), 'The non-negative matrix factorization toolbox for biological data mining', *Source Code for Biology and Medicine* **8**(1), 10.

Libbrecht, M. W. & Noble, W. S. (2015), 'Machine learning in genetics and genomics', *Nature Reviews Genetics* **16**(6), 321.

Lichman, M. (2013), 'UCI machine learning repository'.
URL: *http://archive.ics.uci.edu/ml*

Liu, B., Blasch, E., Chen, Y., Shen, D. & Chen, G. (2013), Scalable sentiment classification for big data analysis using naive Bayes classifier, *in* 'IEEE International Conference on Big Data, 2013', IEEE, pp. 99–104.

Liu, H. & Motoda, H. (2007), *Computational methods of feature selection*, CRC Press.

Liu, W., Chawla, S., Cieslak, D. A. & Chawla, N. V. (2010), A robust decision tree algorithm for imbalanced data sets, *in* 'Proceedings of the 2010 SIAM International Conference on Data Mining', SIAM, pp. 766–777.

Ma, Z., Nie, F., Yang, Y., Uijlings, J. R. & Sebe, N. (2012), 'Web image annotation via subspace-sparsity collaborated feature selection', *IEEE Transactions on Multimedia* **14**(4), 1021–1030.

Manimala, K., Selvi, K. & Ahila, R. (2012), 'Optimization techniques for improving power quality data mining using wavelet packet based support vector machine', *Neurocomputing* **77**(1), 36–47.

Masnadi-Shirazi, H. & Vasconcelos, N. (2010), Risk minimization, probability elicitation, and cost-sensitive SVMs., *in* 'International Conference on Machine Learning', Citeseer, pp. 759–766.

Meier, L., Van De Geer, S. & Bühlmann, P. (2008), 'The group lasso for logistic regression', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**(1), 53–71.

Nagao-Kitamoto, H., Setoguchi, T., Kitamoto, S., Nakamura, S., Tsuru, A., Nagata, M., Nagano, S., Ishidou, Y., Yokouchi, M., Kitajima, S. et al. (2015), 'Ribosomal protein s3 regulates gli2-mediated osteosarcoma invasion', *Cancer Letters* **356**(2), 855–861.

Nesterov, Y. (2005), 'Smooth minimization of non-smooth functions', *Mathematical Programming* **103**(1), 127–152.

Nie, F., Huang, H., Cai, X. & Ding, C. H. (2010), Efficient and robust feature selection via joint 2, 1-norms minimization, *in* 'Advances in Neural Information Processing Systems', pp. 1813–1821.

Nutt, C. L., Mani, D., Betensky, R. A., Tamayo, P., Cairncross, J. G., Ladd, C., Pohl, U., Hartmann, C., McLaughlin, M. E., Batchelor, T. T. et al. (2003), 'Gene expression-based classification of malignant gliomas correlates better with survival than histological classification', *Cancer Research* **63**(7), 1602–1607.

Ottoboni, L., Keenan, B. T., Tamayo, P., Kuchroo, M., Mesirov, J. P., Buckle, G. J., Khoury, S. J., Hafler, D. A., Weiner, H. L. & De Jager, P. L. (2012), 'An RNA profile identifies two subsets of multiple sclerosis patients differing in disease activity', *Science Translational Medicine* **4**(153), 153ra131–153ra131.

Park, M. Y., Hastie, T. & Tibshirani, R. (2007), 'Averaged gene expressions for regression', *Biostatistics* **8**(2), 212–227.

Pascual-Montano, A., Carmona-Saez, P., Chagoyen, M., Tirado, F., Carazo, J. M. & Pascual-Marqui, R. D. (2006), 'bioNMF: a versatile tool for non-negative matrix factorization in biology', *BMC Bioinformatics* **7**(1), 1.

Patil, T. R. & Sherekar, S. (2013), 'Performance analysis of naive Bayes and j48 classification algorithm for data classification', *International Journal of Computer Science and Applications* **6**(2), 256–261.

Pearson, K. (1901), 'LIII. on lines and planes of closest fit to systems of points in space', *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **2**(11), 559–572.

Pogue-Geile, K., Geiser, J., Shu, M., Miller, C., Wool, I. G., Meisler, A. I. & Pipas, J. M. (1991), 'Ribosomal protein genes are overexpressed in colorectal cancer: isolation of a cDNA clone encoding the human s3 ribosomal protein.', *Molecular and Cellular Biology* **11**(8), 3842–3849.

Prati, R. C., Batista, G. E. & Silva, D. F. (2015), 'Class imbalance revisited: a new experimental setup to assess the performance of treatment methods', *Knowledge and Information Systems* **45**(1), 247–270.

Rapaport, F., Barillot, E. & Vert, J.-P. (2008), 'Classification of arrayCGH data using fused SVM', *Bioinformatics* **24**(13), i375–i382.

Read, J., Pfahringer, B., Holmes, G. & Frank, E. (2011), 'Classifier chains for multi-label classification', *Machine Learning* **85**(3), 333–359.

Reich, M., Liefeld, T., Gould, J., Lerner, J., Tamayo, P. & Mesirov, J. P. (2006), 'Genepattern 2.0', *Nature genetics* **38**(5), 500–501.

Rodriguez, J. J., Kuncheva, L. I. & Alonso, C. J. (2006), 'Rotation forest: A new classifier ensemble method', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **28**(10), 1619–1630.

Saeys, Y., Inza, I. & Larrañaga, P. (2007), 'A review of feature selection techniques in bioinformatics', *Bioinformatics* **23**(19), 2507–2517.

Schapire, R. E. & Singer, Y. (2000), 'Boostexter: A boosting-based system for text categorization', *Machine Learning* **39**(2-3), 135–168.

Schiezaro, M. & Pedrini, H. (2013), 'Data feature selection based on artificial bee colony algorithm', *EURASIP Journal on Image and Video Processing* **2013**(1), 1–8.

Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J. & Napolitano, A. (2010), 'RUSBoost: A hybrid approach to alleviating class imbalance', *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans* **40**(1), 185–197.

Shipp, M. A., Ross, K. N., Tamayo, P., Weng, A. P., Kutok, J. L., Aguiar, R. C., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G. S. et al. (2002), 'Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning', *Nature Medicine* **8**(1), 68–74.

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D'Amico, A. V., Richie, J. P. et al. (2002), 'Gene expression correlates of clinical prostate cancer behavior', *Cancer Cell* **1**(2), 203–209.

Smolock, E. M., Korshunov, V. A., Glazko, G., Qiu, X., Gerloff, J. & Berk, B. (2012), 'Ribosomal protein l17, rpl17, is an inhibitor of vascular smooth muscle growth and carotid intima formation', *Circulation* p. 112.

Spolaôr, N., Cherman, E. A., Monard, M. C. & Lee, H. D. (2012), Filter approach feature selection methods to support multi-label learning based on relieff and information gain, *in* 'Advances in Artificial Intelligence, SBIA', Springer, pp. 72–81.

Spolaôr, N., Cherman, E. A., Monard, M. C. & Lee, H. D. (2013), Relieff for multi-label feature selection, *in* 'Brazilian Conference on Intelligent Systems (BRACIS), 2013', IEEE, pp. 6–11.

Storn, R. & Price, K. (1997), 'Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces', *Journal of Global Optimization* **11**(4), 341–359.

Stuart, J. M., Segal, E., Koller, D. & Kim, S. K. (2003), 'A gene-coexpression network for global discovery of conserved genetic modules', *Science* **302**(5643), 249–255.

Tahir, M. A., Kittler, J. & Yan, F. (2012), 'Inverse random under sampling for class imbalance problem and its application to multi-label classification', *Pattern Recognition* **45**(10), 3738–3750.

Tepvorachai, G. & Papachristou, C. (2008), Multi-label imbalanced data enrichment process in neural net classifier training, *in* 'IEEE International Joint Conference on Neural Networks (IJCNN), 2008', IEEE, pp. 1301–1307.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. & Knight, K. (2005), 'Sparsity and smoothness via the fused lasso', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**(1), 91–108.

Toloşi, L. & Lengauer, T. (2011), 'Classification with correlated features: unreliability of feature ranking and solutions', *Bioinformatics* **27**(14), 1986–1994.

Tsoumakas, G. & Katakis, I. (2006), 'Multi-label classification: An overview', *International Journal of Data Warehousing and Mining* **3**(3).

Tsoumakas, G., Katakis, I. & Vlahavas, I. (2009), Mining multi-label data, *in* 'Data Mining and Knowledge Discovery Handbook', Springer, pp. 667–685.

Tsoumakas, G., Spyromitros-Xioufis, E., Vilcek, J. & Vlahavas, I. (2011), 'Mulan: A JAVA library for multi-label learning', *Journal of Machine Learning Research* **12**(Jul), 2411–2414.

Ueda, N. & Saito, K. (2002), Parametric mixture models for multi-labeled text, *in* 'Advances in Neural Information Processing Systems', pp. 721–728.

Ueda, N. & Saito, K. (2003), 'Parametric mixture models for multi-labeled text', *Advances in Neural Information Processing Systems* pp. 737–744.

Vaarala, M. H., Porvari, K. S., Kylloenen, A. P., Mustonen, M. V., Lukkarinen, O. & Vihko, P. T. (1998), 'Several genes encoding ribosomal proteins are over-expressed in prostate-cancer cell lines: confirmation of l7a and l37 over-expression in prostate-cancer tissue samples', *International Journal of Cancer* **78**, 27–32.

Van De Vijver, M. J., He, Y. D., Van't Veer, L. J., Dai, H., Hart, A. A., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J. et al. (2002), 'A gene-expression signature as a predictor of survival in breast cancer', *New England Journal of Medicine* **347**(25), 1999–2009.

Vapnik, V. (2000), *The nature of statistical learning theory*, Springer Science & Business Media.

Wang, J., Jebara, T. & Chang, S.-F. (2008), Graph transduction via alternating minimization, *in* 'Proceedings of the 25th International Conference on Machine Learning', ACM, pp. 1144–1151.

Wang, J., You, J., Li, Q. & Xu, Y. (2012), 'Extract minimum positive and maximum negative features for imbalanced binary classification', *Pattern Recognition* **45**(3), 1136–1145.

Weiss, G. M. (2004), 'Mining with rarity: a unifying framework', *ACM SIGKDD Explorations Newsletter* **6**(1), 7–19.

Witten, D. M. & Tibshirani, R. (2010), 'A framework for feature selection in clustering', *Journal of the American Statistical Association* **105**(490), 713–726.

Wu, B., Lyu, S. & Ghanem, B. (2016), Constrained submodular minimization for missing labels and class imbalance in multi-label learning., *in* 'Association for the Advancement of Artificial Intelligence', pp. 2229–2236.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Philip, S. Y. et al. (2008), 'Top 10 algorithms in data mining', *Knowledge and Information Systems* **14**(1), 1–37.

Xu, L., Wang, Z., Shen, Z., Wang, Y. & Chen, E. (2014), Learning low-rank label correlations for multi-label classification with missing labels, *in* 'IEEE International Conference on Data Mining (ICDM), 2014', IEEE, pp. 1067–1072.

Yang, P., Xu, L., Zhou, B. B., Zhang, Z. & Zomaya, A. Y. (2009), 'A particle swarm based hybrid system for imbalanced medical data sampling', *BMC Genomics* **10**(Suppl 3), S34.

Yang, S., Yuan, L., Lai, Y.-C., Shen, X., Wonka, P. & Ye, J. (2013), Feature grouping and selection over an undirected graph, *in* 'Graph Embedding for Pattern Analysis', Springer, pp. 27–43.

Yen, S.-J. & Lee, Y.-S. (2009), 'Cluster-based under-sampling approaches for imbalanced data distributions', *Expert Systems with Applications* **36**(3), 5718–5727.

Yokoya, N., Yairi, T. & Iwasaki, A. (2012), 'Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion', *IEEE Transactions on Geoscience and Remote Sensing* **50**(2), 528–537.

Yu, H., Ni, J. & Zhao, J. (2013), 'ACOSampling: An ant colony optimization-based undersampling method for classifying imbalanced DNA microarray data', *Neurocomputing* **101**, 309–318.

Zagzag, D., Salnikow, K., Chiriboga, L., Yee, H., Lan, L., Ali, M. A., Garcia, R., Demaria, S. & Newcomb, E. W. (2005), 'Downregulation of major histocompatibility complex antigens in invading glioma cells: stealth invasion of the brain', *Laboratory Investigation* **85**(3), 328–341.

Zhang, M.-L., Li, Y.-K. & Liu, X.-Y. (2015), Towards class-imbalance aware multi-label learning., *in* 'International Joint Conference on Artificial Intelligence', pp. 4041–4047.

Zhang, M.-L., Peña, J. M. & Robles, V. (2009), 'Feature selection for multi-label naive Bayes classification', *Information Sciences* **179**(19), 3218–3229.

Zhang, M.-L. & Zhou, Z.-H. (2014), 'A review on multi-label learning algorithms', *IEEE Transactions on Knowledge and Data Engineering* **26**(8), 1819–1837.

Zhang, Y. & Zhou, Z.-H. (2010), 'Multilabel dimensionality reduction via dependence maximization', *ACM Transactions on Knowledge Discovery from Data (TKDD)* **4**(3), 14.

Zhou, T., Tao, D. & Wu, X. (2012), 'Compressed labeling on distilled labelsets for multi-label learning', *Machine Learning* **88**(1-2), 69–126.

Zhou, X. & Tuck, D. P. (2007), 'MSVM-RFE: extensions of SVM-RFE for multiclass gene selection on DNA microarray data', *Bioinformatics* **23**(9), 1106–1114.

Zou, H. & Hastie, T. (2005), 'Regularization and variable selection via the elastic net', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**(2), 301–320.