Case-Based Retrieval Framework for Gene Expression Data

Ali Anaissi[1], Madhu Goyal[1], Daniel R. Catchpoole[2], Ali Braytee[1] and Paul J. Kennedy[1]

[1]Center for Quantum Computation and Intelligent Systems, Faculty of Engineering and Information Technology, University of Technology Sydney, Broadway, New South Wales, Australia. [2]The Tumour Bank, Children's Cancer Research Unit, The Children's Hospital at Westmead, Westmead, New South Wales, Australia.

APSTRACT

BACKGROUND: The process of retrieving similar cases in a case-based reasoning system is considered a big challenge for gene expression data sets. The huge number of gene expression values generated by microarray technology leads to complex data sets and similarity measures for high-dimensional data are problematic. Hence, gene expression similarity measurements require numerous machine-learning and data-mining techniques, such as feature selection and dimensionality reduction, to be incorporated into the retrieval process.

METHODS: This article proposes a case-based retrieval framework that uses a *k*-nearest-neighbor classifier with a weighted-feature-based similarity to retrieve previously treated patients based on their gene expression profiles.

RESULTS: The herein-proposed methodology is validated on several data sets: a childhood leukemia data set collected from The Children's Hospital at Westmead, as well as the Colon cancer, the National Cancer Institute (NCI), and the Prostate cancer data sets. Results obtained by the proposed framework in retrieving patients of the data sets who are similar to new patients are as follows: 96% accuracy on the childhood leukemia data set, 95% on the NCI data set, 93% on the Colon cancer data set, and 98% on the Prostate cancer data set.

CONCLUSION: The designed case-based retrieval framework is an appropriate choice for retrieving previous patients who are similar to a new patient, on the basis of their gene expression data, for better diagnosis and treatment of childhood leukemia. Moreover, this framework can be applied to other gene expression data sets using some or all of its steps.

KEYWORDS: case base reasoning, gene expression, machine learning, data mining, dimensionality reduction, feature weighting

CITATION: Anaissi et al. Case-Based Retrieval Framework for Gene Expression Data. *Cancer Informatics* 2015:14 21–31 doi: 10.4137/CIN.S22371.

RECEIVED: December 01, 2014. **RESUBMITTED:** January 18, 2015. **ACCEPTED FOR PUBLICATION:** January 22, 2015.

ACADEMIC EDITOR: J.T. Efird, Editor in Chief

TYPE: Methodology

FUNDING: The Cancer Institute NSW provided funding to DRC and PJK for data analysis. The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: ali.anaissi@uts.edu.au

Background

Case-based reasoning (CBR)[1] is considered a most trustworthy methodology for building intelligent systems for the storage and retrieval of past experiences to solve new problems. It is described by the following four processes: retrieve, reuse, revise, and retain. Each case in the case base is characterized by multiple features or attributes. When a new case is presented to the CBR system, the system measures the similarity between features of this query instance and those of other cases in the case base to retrieve the most similar one from the case base. The most important assumption in CBR is that similar experiences can guide future reasoning, problem solving, and learning.[2] CBR has been extensively used in the biomedical field and successfully extended to many applications. For instance, Marling and Whitehouse[3] have used CBR in the care of patients with Alzheimer's disease. The authors report that CBR finds effective treatment by matching patients to treatments that were effective for similar patients in the past. Similarly, Lieber and Bresson[4] propose a CBR system for breast cancer treatment, and they claim that their system suggests appropriate solutions for new patients. Diaz et al[5] use a CBR system for cancer classification based on microarray data.

The process of retrieving cases from a case base similar to a query case is regarded as a primary and fundamental step in CBR,[6–8] and the similarity measurement between cases plays a very important role in this process. The most widely used methods for similarity measures are distance-based functions that calculate the distance between cases using some or all of the attributes constituting the cases. However, distance measurement is not directly applicable on cases with many attributes, such as gene expression data sets, due to the curse of dimensionality. The curse of dimensionality[9] affects similarity searching in high-dimensional space because it makes nearest neighbor searching senseless because the anticipated distance between the points (cases) converges to zero as the dimension goes to infinity.[10]

Four main issues affect data mining when calculating similarity between cancer patients on the basis of their gene expression data: irrelevant features, high dimensionality, relative importance of features, and imbalanced classes.

Gene expression data sets are complex and often highly dimensional.[11] Many of these dimensions (genes) are irrelevant to a specific trait of interest. Moreover, gene expression microarray data sets often consist of a limited number of patients (hundreds) relative to the large number of gene expression values (thousands of genes). This problem is compounded in the case of imbalanced data sets wherein there is a big difference between the numbers of data points in each target class. In addition, some data sets have the class variable defined based on a risk assessment done on the basis of clinical judgment, such as risk factor. This clinical judgment of patient's risk category might be different from that of other data sets wherein the class variable is often clearly defined by histology (eg, cancer tissue is distinguished from noncancer tissue by microscope biopsies). Clinical judgment results in some variation in assignment of risk among clinicians, which lowers the expectation of very high predictive accuracies using a simple feature selection algorithm with a nearest-neighbor classifier. With this type of data, the ability to successfully distinguish between patients based on their gene expression data and to explore the neighborhood space of patients to find how patients are similar or dissimilar to others requires a more sophisticated use of data-mining and machine-learning techniques in the similarity measurement process.

Data-mining techniques such as feature selection and feature weighting have previously been successfully combined with the CBR system[12–15] but separately. Arshadi and Jurisica[13] propose a CBR system for ultra-high-dimensional biological data sets. The authors apply spectral clustering followed by feature selection to preprocess the data. The main problem of this approach is that *k*-means clustering technique is applied in high-dimensional space without dealing with the curse of dimensionality.[9] The Euclidean distance measurement often performs poorly as the dimensionality of the analyzed data increases.[10] Moreover, the authors evaluate the system on two simple publicly available microarray data sets that cover leukemia and lung cancer samples. They report improvements in classification accuracy of approximately 20% from 65% to 79% for Leukemia and from 60% to 70% for Lung cancer. Similarly, Diaz et al[5] apply only the feature selection algorithm in the retrieval stage, and clustering techniques are applied during the reuse and prediction stage. In this article, we show how this data-mining technique (feature selection) can further improve the retrieval process of a CBR system when combined with other data-mining techniques such as dimensionality reduction and feature weighting. This study explores problems of retrieving similar cases in the CBR system, dealing with extremely complex gene expression data sets wherein the case base has relatively few imbalanced-class cases each having thousands of features. For some gene expression data sets such as Golub's leukemia data set[16] and the Lung cancer data set,[17] the nearest-neighbor classifier – with the help of feature selection algorithm – can accurately retrieve cases that are similar to a query case from the case base. However, similarity measurement becomes a big challenge in the case of extremely complex gene expression data sets such as an imbalanced-classes data set or a data set with the class variable defined based on a risk assessment done on the basis of clinical judgment.

This study proposes a CBR system to help clinicians and biologists in their prediction of risk of relapse in childhood leukemia sufferers by comparing them to previous patients based on their gene expression measurements. The main focus of this article is to develop a case-based retrieval framework for *k*-nearest-neighbor classifier (*k*NN) with a weighted feature-based similarity that is able to retrieve similar patients from a case base of acute lymphoblastic leukemia (ALL) patients based on their gene expression data. By observing the treatment and outcome for the retrieved similar patients, more reliable decisions about this new patient can be made. CBR is particularly applicable to this problem domain and can be used to propose new solutions or evaluate solutions to avoid potential problems such as relapse of ALL. CBR can yield better diagnosis and treatment for childhood leukemia sufferers by suggesting the previous medical treatment that accomplished the desired result, to enable curing of new patients. The assumption here is that patients with similar gene expression profiles will react similarly to therapy and should be treated in like manner.

Methods

Case-based retrieval framework. This article presents a novel case-based retrieval framework that involves several computational intelligence techniques. The purpose of developing this conceptual framework is to show that data-mining techniques such as feature selection and dimensionality reduction have several positive effects on the gene expression data sets in terms of alleviating the curse of dimensionality and enhancing the similarity measurement; that the effectiveness of weighting the features is important in the distance measurement; and that an improvement in the case-based retrieval process can be adopted by applying oversampling techniques in case of imbalanced gene expression data sets. This framework is composed of two modules: Module 1 for training and Module 2 for retrieval (Fig. 1).

Module 1 concerns the preprocessing steps of the training data set. The purpose of this module is to preprocess the training data set and to handle the complexities of the gene expression data sets using methods such as feature selection, dimensionality reduction, and feature weighting.

Module 2, on the other hand, concerns case retrieval for a new query case. The purpose of this module is to use the outputs of the training process, including a list of the selected features, to preprocess the query before retrieving similar previous cases from the case base.

Module 1: preprocessing the training data set. The first module of the case-based retrieval framework preprocesses

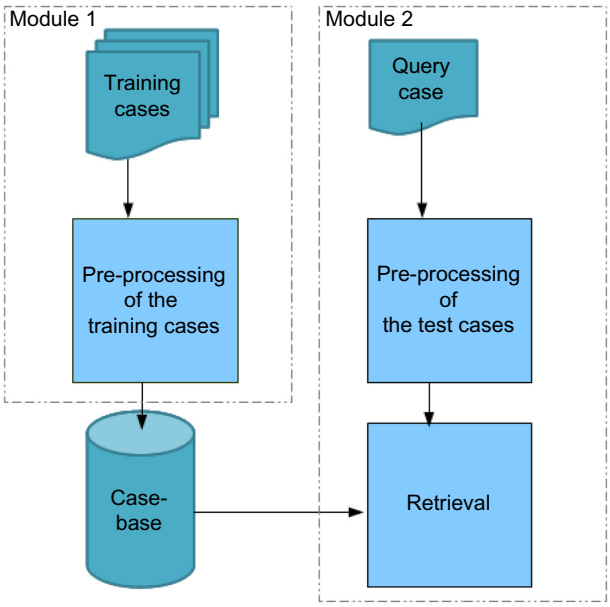

Figure 1. Case-based retrieval framework.

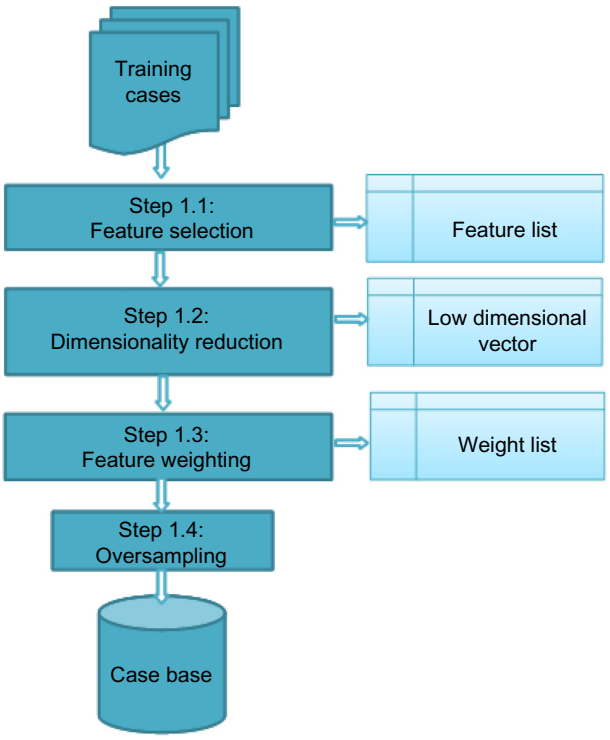

Figure 2. Preprocessing the training data set in the case-based retrieval framework.

the training data set and is composed of four steps: feature selection, dimensionality reduction, feature weighting, and oversampling (Fig. 2). The first step of this framework – feature selection – is essential for any gene expression data set. However, the need for the other steps is based on the complexity of the data set. For example, the oversampling step is only required in case of class-imbalanced data sets.

The first step aims to select a subset of genes representing the most relevant features to a specific output of interest. The second step applies dimensionality reduction algorithms on the data set to reduce the impact of the curse of dimensionality. Once the data set is processed by the dimensionality reduction algorithm and is transformed to a lower dimensional space, the data set is presented to the feature-weighting process to identify the relative importance of each feature for similarity and classification optimization. The final step applies an oversampling technique to increase the number of samples of the minority classes and to reduce the effect of the imbalanced classes. Figure 2 shows the steps in the training process.

Step 1.1: feature selection. The presence of too many features in a data set adversely affects similarity measurement and classification performance if many of these features are irrelevant to the specific trait of interest. Hence, selection of a subset of genes that are relevant to a trait of interest is crucial and plays a vital role for building a successful gene expression similarity measurement model.

The feature selection algorithm balanced iterative random forest (BIRF)[18] is initially applied to the training cases to select relevant features. On the basis of the performance of the BIRF reported in the study by Anaissi et al,[18] BIRF is an appropriate choice to select genes from imbalanced high-throughput gene expression microarray data. In the report by Anaissi et al,[18] BIRF was evaluated on four cancer microarray data sets: a childhood leukemia data set collected from The Children's Hospital at Westmead, a Colon cancer data set, the NCI-60 data set, and a Lung cancer data set. Significant results were achieved using BIRF in comparison to the other state-of-the-art methods such as support vector machine–recursive feature elimination,[19] random forest (RF),[5] and naive Bayes[20] classifiers.

The output of this step is the training data set, with a subset of genes that are strongly associated with the output of interest. The selected genes are stored as a list of genes, labeled "Feature list" in Figure 2, to be used later for processing the test data samples.

Step 1.2: dimensionality reduction. Dimensionality reduction aims to transform a high-dimensional data set into a lower-dimensional one representing the most important variables underlying the high-dimensional data. In contrast to feature selection, dimensionality reduction aims to extract new features from the original set of features. Although the number of features of the data set is reduced after removing the irrelevant genes, the data set may still exist in a high-dimensional space, and similarity measurement may still suffer from the curse of dimensionality.[21] Principal component analysis (PCA) is applied to the training set samples. This method is characterized by its simplicity, and it is a non-parametric method. Moreover, the low-dimensional vector learned by PCA can be applied to out-of-sample data points to get their low-dimensional embedding. The outputs of this step are the training data set in a low-dimensional space and

a vector, labeled "Low-dimensional vector" in Figure 2, to be used later in processing the test data samples.

Linear kernel methods[22] are used along with PCA to save considerable amounts of computation time in finding the effective principal components. This is because the number of attributes or features is very large, much higher than the number of samples, in gene expression data sets. In normal PCA, the size of the covariance matrix is $m \times m$, where m is the number of attributes. However, while using kernel methods, the size of the kernel matrix is $n \times n$, where n is the number of observations or samples. The idea behind kernel PCA (KPCA) is to find the directions or components for which the data set has maximum variance in the feature space. This is achieved by finding the eigenvalues with the corresponding eigenvectors for the kernel matrix of the data set. Dimensionality reduction is then achieved by choosing the largest eigenvalues obtained by KPCA to represent the data in fewer dimensions.

Dimensionality reduction based on KPCA takes as input $\mathbf{X} \varepsilon R^{n \times m}$ and produces output $\mathbf{Y} \varepsilon R^{n \times d}$, where m and d are the dimensionality of the input and output data sets, respectively, and n is the number of points. The question in this process is this: what is the minimum dimension that can be achieved without acceptable loss of precision? Or, which components of KPCA should be selected to represent the data set in fewer dimensions?

This study proposes a wrapper method for choosing the best value of $d < m$. The concept of this wrapper method is to use a nearest-neighbor (NN) classifier to evaluate the classification performance on different low-dimensionality representations of the data to choose the most appropriate value of d.

Due to the small number of observations in the training data set and to obtain a result that can generalize well, c-fold cross-validation technique is used to determine the classification accuracy of the classifier. It is usually called k cross–validation, but c is used here to differentiate it from the parameter k of the NN classifier. The accuracy is evaluated on several lower-dimensionality representations of the data to find the value of d that best describes the data.

Step 1.3: feature weighting. Feature weighting[23] is a technique used to estimate the relative influence of individual features with respect to the classification performance. When successfully weighted, high-impact features receive a high-value weight, whereas a low weight is assigned to low-impact features. The output of this step is a weight vector that is stored as a list of weights, labeled the "Weight list" in Figure 2, to be used in the distance measurement formula. Feature weighting is needed for instance-based learning algorithms such as NN. Giving weights to the features based on their quality and usefulness has the potential to lead to accurate distance measurement. Two hypotheses are proposed and tested in this study to address this issue.

Hypothesis 1: Eigenvalues can be used as weights for features. The first hypothesis for feature weighting is based on the eigenvalues derived from KPCA. We observed that the dimensionality reduction is achieved by discarding features with a low eigenvalue and retaining only those features with a high eigenvalue. One idea is to use these eigenvalues in the kNN as a vector weight and then use it in the Euclidean distance formula.

Hypothesis 2: Genetic algorithm can be used to seek the weights for features. Genetic algorithm (GA)[24] is considered a general purpose search process for optimization problems. Because all optimization algorithms have an objective function, we have designed a fitness function as called in the GA, for optimization of classification performance by searching for the best features (genes) encoding weights for the similarity measurement. The goal of the GA is to minimize the classification error of the training data set. A wrapper feature-weighting method based on GA is used to propose a weight-learning GA. The concept of this wrapper method is to use a GA to seek the best weights of features with the kNN classifier based on the GA fitness function (Fig. 3). The fitness function is computed by subtracting the accuracy from the number one. C-fold cross-validation technique is also used here to determine the classification accuracy of the classifier. The accuracy is computed by averaging the accuracies of the c-folds for the kNN classifier using the generated weights in its Euclidean distance measure.

Step 1.4: oversampling. Many gene expression data sets associated with rare diseases have the imbalanced-classes problem. That is, at least one of the classes constitutes only a very small minority of the data. For such problems, the effect is on practical classification, whereby the interest usually leans toward correct classification of the minor class. Generally, most classification techniques assume that training samples are evenly distributed among different categories. However, in practical applications, data sets often exist in an unbalanced

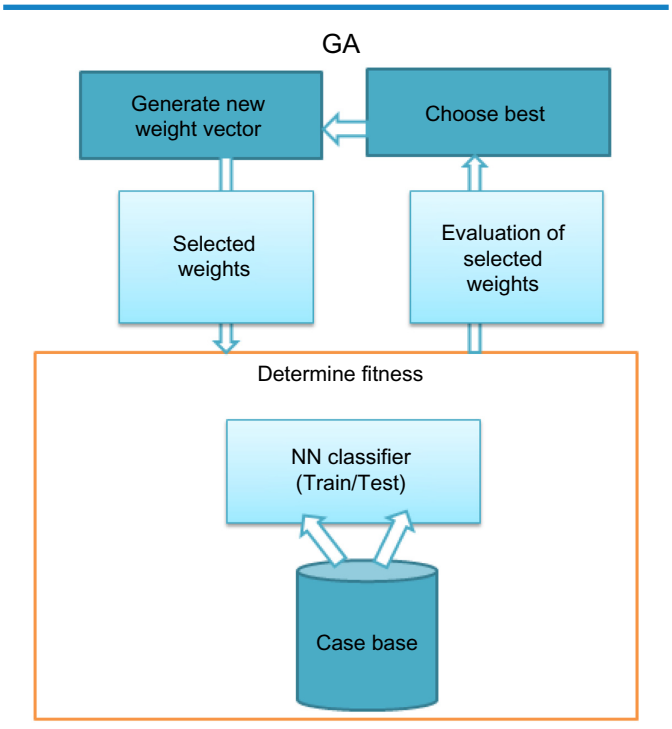

Figure 3. Weight-learning GA and kNN.

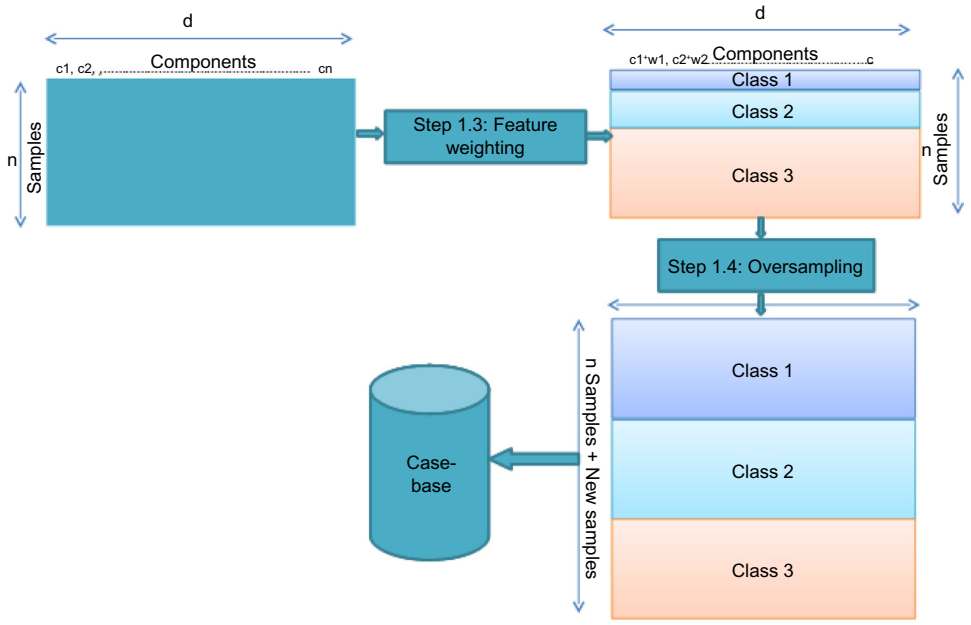

Figure 4. Modification of the training data set in the feature-weighting and sampling techniques section.

form. In addition, gene expression data sets often have a low number of samples. With these types of data sets, a poor classification performance is often achieved and can result in trivial classifiers that completely ignore the minority class.

One approach to dealing with unbalanced data sets is to use oversampling techniques to increase the number of samples in the gene expression data set. In this study, we consider the use of a well-acknowledged oversampling method to balance the training set before the learning phase, which is the synthetic minority oversampling technique (SMOTE) methodology.[25] SMOTE, a widely used technique,[26–28] is applied in this framework to add new, artificial minority examples by interpolating between original minority-class examples. Figure 4 shows the feature-weighting and oversampling steps in the case-based retrieval framework.

Module 2: retrieval. The second module of the framework is related to the query samples (Fig. 5). A new query sample comes in the high-dimensional space. Irrelevant features are filtered based on the determined relevant features obtained from the feature selection step of Module 1. The next step is to transform the new sample into a lower-dimensional space by projecting the filtered features onto the dimensionality reduction vector obtained from the dimensionality reduction step of Module 1. Once the new sample is passed through the preprocessing steps of the test samples, it is presented to the *k*NN classifier to retrieve similar previous cases from the case base using the feature's weight obtained from the feature-weighting step of Module 1.

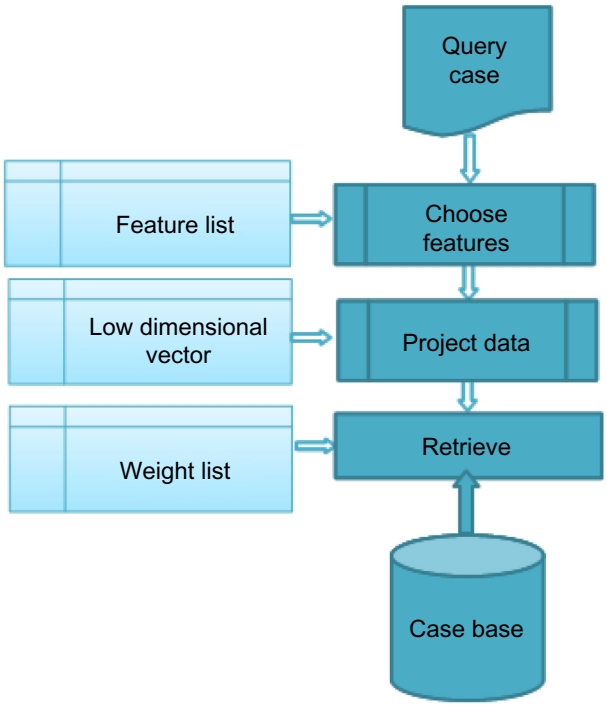

Figure 5. Preprocessing the test sample in the case-based retrieval framework and retrieving similar cases.

Results and Discussion

Several experiments are performed in each step in the case-based retrieval framework to demonstrate the validity of the proposed framework and to evaluate the framework using different data sets.

Data sets. The main experiments are performed on a childhood leukemia gene expression data set that has been collected from The Children's Hospital at Westmead. This data set is also available in the public domain and can be explored through the Oncogenomics Section of the Paediatric Oncology

Branch at the NCI, National Institutes of Health, USA (http://pob.abcc.ncifcrf.gov/cgi-bin/JK). The entire childhood leukemia gene expression data set is composed of 110 patients with expression values for 32,678 probes. The patients of this data set are classified into three categories based on their risk of relapse. A stratified random sampling is applied on the gene expression data set, and it is divided into training and test data sets. The training and test data sets are composed of 70 and 40 patients, respectively. The distribution of patients in each data set is shown in Table 1.

In addition to the childhood leukemia data set, we have chosen three other publicly available microarray data sets: NCI-60, Colon cancer, and Prostate cancer data sets. The NCI-60 cancer cell line data set is a well-studied publicly available microarray benchmark collected by Ross et al[29] and was produced using Affymetrix HG-U133A chips. The data set consists of 60 samples that are classified into eight categories. Each sample is measured over 5,244 gene expression values.

Experiments on the childhood leukemia data set. The first step of Module 1 is applied on the training data set for selecting the relevant features. The feature selection algorithm BIRF[18] is involved in this step. Detailed explanation about the algorithm and experiments can be found in the report by Anaissi et al.[18] This step produced a feature list of 107 genes, which are selected as the most significant biomarkers.

As mentioned above, not all steps of the framework must be involved in processing the data before measuring the similarity between patients. Consequently, we present the test data set to the *k*NN algorithm to evaluate the performance in terms of retrieval of similar cases from the case base before continuing with the remaining steps of the framework.

In general, the *k* parameter of the *k*NN classifier plays an important role in classification, especially when the distribution of classes in the training set is uneven. Accordingly, and based on the number and the distribution of patients in the training childhood leukemia data set, it is readily perceived that the value of *k* affects the performance of the NN classifier, as shown in Table 2.

A cross-validation procedure is used here to determine the optimal value of *k*. The best average accuracy result is achieved at $k = 5$ (Table 2). The average accuracy is calculated based on the average of the sensitivity and specificity of each class because the traditional accuracy does not reflect the actual classification result in the case of an imbalanced data set. All the reported results in the following experiments are based on these measurements.

Table 1. The number of patients in the training and test data sets.

	HIGH RISK	MEDIUM RISK	STANDARD RISK	TOTAL
Training dataset	6	53	11	70
Test dataset	5	25	10	40

The Colon cancer data set is a publicly available microarray data set that was obtained with an Affymetrix oligonucleotide microarray.[30] The Colon data set contains 62 samples, with each sample containing the expression values for 2,000 genes. Each sample indicates whether or not it came from a tumor biopsy. This data set has been used in many different research papers, eg, Ben-Dor et al,[31] Brazma and Vilo,[32] and Getz et al.[33] The Prostate cancer data set also has been used in the experiments. The data set contains 52 prostate tumor samples and 50 nontumor prostate samples, with around 12,600 genes.

Table 2. Comparison of the performance of *k*NN on the childhood leukemia test data set for different values of *k*.

K	AVERAGE ACCURACY
3	0.71
4	0.69
5	**0.73**
6	0.66
7	0.62
8	0.58
9	0.58
10	0.52

*The average accuracy is calculated based on the average of the sensitivity and specificity of each class.

As can be seen from Table 2, the accuracy does not improve as *k* further increases. This result can be justified by looking at the nature of the data set and how the three classes are distributed. It can be noticed that a new high-risk patient is hard to classify correctly if *k* is large. All the results reported in the following experiments on the childhood leukemia data set were obtained based on the value $k = 5$. The initial classification performance of the 5NN classifier on the test data set is presented in the confusion matrices in Tables 3 and 4. These show that the classification results of the data set processed only by the first step (feature section) were poor, with most patients predicted incorrectly, especially the high-risk ones. These results suggest that further steps of the framework are required to enhance the performance of retrieving similar cases from the case base. Therefore, the dimensionality reduction step is applied to the training data set.

Kernel principal component analysis. An eight-fold cross-validation is applied to the training childhood leukemia gene expression data set. Many applications use ten-fold cross-validation, but because there are not many samples in our data set, and to have a reasonable number of training and test samples (especially for high-risk patients), the value of *d* (the target dimensionality) is determined using eight-fold

Table 3. Results of classification performance tests on the childhood leukemia test data set.

	PREDICTED HIGH	PREDICTED MEDIUM	PREDICTED STANDARD
Actual High	2	2	1
Actual Medium	1	22	2
Actual Standard	1	3	6

Table 4. Statistics by class for the confusion matrix of the test data set presented in Table 3.

	HIGH	MEDIUM	STANDARD
Sensitivity	0.40	0.88	0.60
Specificity	0.94	0.66	0.90

cross-validation. Each sample from the eight folds is taken from the training data for validation, and the remaining data are processed by KPCA and reduced into different values of *d*. After each reduction, the test sample is projected into the space of the reduced data and classification accuracy is evaluated. This process is repeated eight times, and the eight results are averaged for each value of *d* so that we have a single estimate at each value of *d* for the eight validations. Classification accuracy of the obtained reduced data set is evaluated for each value of *d*, which is determined based on the classification performance. Figure 6 shows the classification results for different values of *d*. As can be seen, the best average value is realized at $d = 50$. Moreover, the accuracies of the eight folds at the value 50 consistently have the highest values, as shown in Figure 6. The variance accounted for the chosen *d* represents 95% of the total variance.

An important reduction is achieved in the dimensionality because the data set will be in a low-dimensional space and distance measurements can be applied on the data set to compute the similarity between the patients in the case base and the new patients. The classification performance of the 5NN classifier on the test set after processing the data set with the second step of the framework is presented in the confusion matrices in Tables 5 and 6.

As can be seen from Table 6, there are significant improvements in the classification results for each class. The average

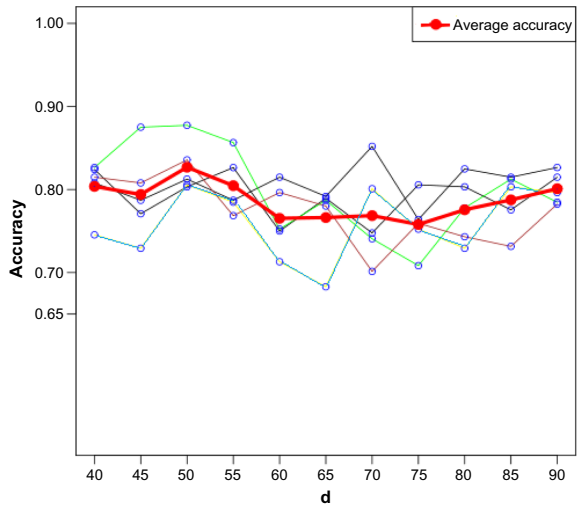

Figure 6. Accuracy according to the different dimensionality reduction processes using eight-fold cross-validation on the test childhood leukemia data set.

Table 5. Results of the classification performance on the childhood leukemia test data set after reducing the dimensionality.

	PREDICTED HIGH	PREDICTED MEDIUM	PREDICTED STANDARD
Actual High	3	1	1
Actual Medium	0	23	2
Actual Standard	2	1	7

accuracy increased from 0.73 to 0.82. The sensitivity of the minor class (high-risk patients) also increased from 0.4 to 0.6. This result indicates that the minority class becomes recognized better by the NN classifier. However, further improvements may be achieved if we use a weighted-feature-based similarity and process the training data set using the third step of the framework, ie, by using the feature-weighting step.

Feature weighting. The aim of the initial experiments performed in this section was to decide which of the hypotheses presented in the Methods section should be used for feature weighting. Experiments for Hypothesis 1 are first conducted and for that, eight-fold cross-validation is applied on the training data set to compute the accuracies of the unweighted 5NN classifier. Eight accuracies are obtained from this experiment for the eight test folds. The same procedure is applied on the weighted 5NN with the eigenvalues, and eight accuracies are computed for the same eight test folds. A paired *t*-test was conducted to determine whether differences in accuracy between the eigenvalue approach and the unweighted approach are significant. The *P*-values of the *t*-test are also used to judge the degree of the performance improvement. The paired *t*-test generates a *P*-value of 0.0383, which indicates that the eigenvalues-weighted 5NN and unweighted 5NN processes do not have the same accuracy. The results of a one-tailed *t*-test indicate that the eigenvalue approach gives little accuracy improvements over the unweighted 5NN because the average balanced accuracy increases from 0.82 to 0.86.

With respect to the Hypothesis 2, the training data set is randomly partitioned into eight subsamples. Each fold is taken from the training data set for validation, and the remaining data are processed by a GA. For each set of feature weights generated by GA, the validation fold is presented to the 5NN classifier and the classification accuracy is evaluated by calculating the average balanced accuracy of the obtained confusion matrix. This process is repeated eight times, and the eight

Table 6. Statistics by class for the confusion matrix of the test data set presented in Table 5.

	HIGH	MEDIUM	STANDARD
Sensitivity	0.60	0.92	0.70
Specificity	0.94	0.86	0.90

Table 7. The parameters for the genetic algorithm for this task.

NO. OF VARIABLES	BEQ	LOWER	UPPER	POPULATION SIZE	NO. OF GENERATIONS	OTHER PARAMETERS
50	1	zeros(1,50)	ones(1,50)	100	50	Default

results from each fold are averaged for each set of weights so that we have a single estimate for each of the eight folds.

The methodology is implemented using the Matlab Global Optimization Toolbox. The parameters for the GA for this task are shown in Table 7.

The program runs for 50 generations with a population size of 100 individuals. During each round of iteration, if the obtained population forms a solution with a better fitness value, the populations will converge to the relevant weights of the feature. Figure 7 shows the running process of the GA by plotting the fitness value of each generation. The program stops at the 50th iteration, and the best individual is then selected as an encoding weight for the feature.

A paired *t*-test was also conducted to determine whether accuracy differences between the GA approach and the unweighted approach are significant. The *P*-value of 3.53×10^{-8} on the *t*-test indicates a significant performance improvement in accuracy. The results of the one-tailed *t*-test indicate that the GA approach outperforms the unweighted 5NN. The average balanced accuracy for the eight-fold cross-validation increased from 0.82 to 0.93. Tables 8 and 9 show the classification performance of the test data set applied on the GA-weighted 5NN classifier.

Hypothesis 2 is supported in this study because it leads to a significant enhancement in the classification performance as shown in Tables 8 and 9. The GA wrapper method provided good feature weights that substantially improve the performance of the NN classifier. These results outlined in Tables 8 and 9 indicate that assigning different weights to features in the domain of gene expression data sets improves the classification accuracy of the NN algorithm.

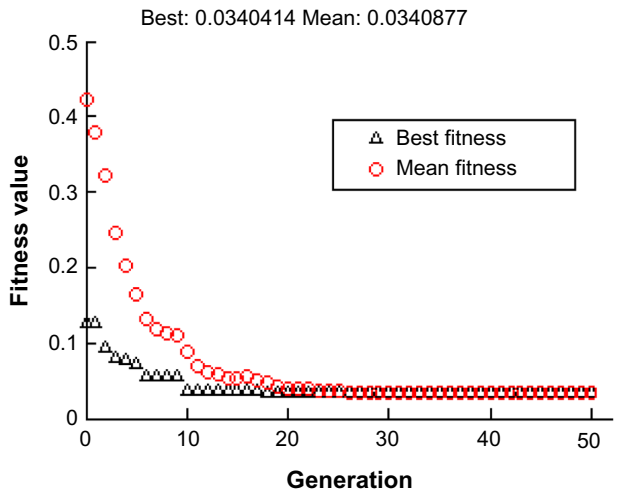

Figure 7. Fitness value for each generation.

Table 8. Results of the classification performance on the childhood leukemia test data set applied on the weighted 5NN classifier.

	PREDICTED HIGH	PREDICTED MEDIUM	PREDICTED STANDARD
Actual High	4	0	1
Actual Medium	0	25	0
Actual Standard	0	1	9

The NN is a white-box classifier that allows us to look at the classification outputs in detail. Analysis of the obtained results reveal that the classification probability for some patients is not high enough because some patients have two classifications with the same probabilities, but they were classified correctly based on the score voting of their similarity to the neighboring samples. Classification probabilities for each patient are calculated and presented in Table 10. The probabilities are computed for each patient in the test data set based on the five retrieved patients. For example, if the five retrieved patients for a high-risk patient in the test data set are two medium, one standard, and two high risk, then the classification probabilities for this patient are 0.4, 0.2, and 0.4, respectively. As can be seen from this table, some patients – and especially those in the minority class – are hardly classified with the actual category. These results indicate that the imbalanced-classes problem affects the classification performance of the process for the minority class. Therefore, oversampling may be required to enhance the classification performance in the case of the minority class.

Oversampling. SMOTE[25] is used in this study to oversample the minority class by introducing synthetic samples. Minority classes are oversampled at 30%, 50%, 100%, 200%, 300%, and 400%. The best accuracy is achieved at 100%, as shown in Table 11. Oversampling the training data set consistently provides an improvement in classification of test data. Moreover, it provides a more stable classifier for the imbalanced classes. The confusion matrix and probabilities

Table 9. Statistics by class for the confusion matrix of the test data set presented in Table 8.

	STANDARD	MEDIUM	HIGH
Sensitivity	0.90	1.00	0.80
Specificity	0.96	0.93	1.00

Table 10. Results of classification probability on the childhood leukemia test data set.

ACTUAL	PREDICTED HIGH	PREDICTED MEDIUM	PREDICTED STANDARD
1 High	0.4	0.2	0.4
2 High	0.4	0.4	0.2
3 High	0.4	0.4	0.2
4 High	0.8	0.2	0
5 High	0.2	0	0.8
6 Medium	0	1	0
7 Medium	0	1	0
8 Medium	0	0.6	0.4
9 Medium	0	0.8	0.2
10 Medium	0	1	0
11 Medium	0	0.6	0.4
12 Medium	0	1	0
13 Medium	0	0.6	0.4
14 Medium	0	0.6	0.4
15 Medium	0	0.6	0.4
16 Medium	0	0.8	0.2
17 Medium	0	1	0
18 Medium	0	0.8	0.2
19 Medium	0	1	0
20 Medium	0	0.8	0.2
21 Medium	0	1	0
22 Medium	0	0.6	0.4
23 Medium	0	1	0
24 Medium	0	1	0
25 Medium	0	1	0
26 Medium	0	0.6	0.4
27 Medium	0	0.6	0.4
28 Medium	0	0.8	0.2
29 Medium	0	0.8	0.2
30 Medium	0	0.8	0.2
31 Standard	0	0.2	0.8
32 Standard	0	0	1
33 Standard	0	0	1
34 Standard	0	0	1
35 Standard	0.2	0.2	0.6
36 Standard	0.2	0.2	0.6
37 Standard	0.2	0.2	0.6
38 Standard	0.2	0.2	0.6
39 Standard	0.2	0.4	0.4
40 Standard	0.2	0.4	0.4

of the 5NN classifier are presented in Tables 12 and 13, respectively.

For instance, SMOTE does not show major improvements in the classification performance. However, the classification probability becomes higher than before when the weighted NN classifier is used without oversampling the training data set (Table 13).

Table 11. Classification performance on the childhood leukemia training data set after 100% oversampling.

	PREDICTED HIGH	PREDICTED MEDIUM	PREDICTED STANDARD
Actual High	18	0	0
Actual Medium	0	53	0
Actual Standard	0	0	33

Experiments on three public microarray data sets. One of the most important aspects of any experiment is validating the framework on other data sets. Validation is achieved by applying the proposed case-based retrieval framework on three publicly available microarray data sets. We have to keep in mind that this framework is proposed for very complex gene expression data sets. This means that some gene expression data sets may not need to be processed by all steps of the framework. For example, Golub's leukemia data set[16] is considered a very simple data set and does not need to go through all the steps of the framework. According to our experiments, the feature selection algorithm BIRF with an NN classifier was enough to accurately (zero error rate) retrieve similar cases from the case base without applying any dimensionality-reducing or feature-weighting algorithms.

The case-based retrieval framework is validated on the NCI-60 data set. Initially, we have applied the feature selection step, followed by the NN classifier to see whether it could be enough to retrieve similar cases from the case base. The average accuracy of 0.68 resulting from the eight-fold cross-validation indicates that further steps are required to achieve better accuracy. Consequently, the dimensionality reduction step is applied on the training data set, and a *t*-test is used to determine whether accuracy differences between the feature selection/NN classifier (FS/NN) approach and the FS/dimensionality reduction (DR)/NN are significant or not. The *P*-value (3.789×10^{-8}) of the paired *t*-test indicates a significant performance improvement in the average classification accuracy. The average accuracy of the eight-fold cross-validation is calculated after processing the data set with the dimensionality reduction step. The resulting value

Table 12. Results of classification performance on the childhood leukemia test data set after 100% oversampling of the training data set.

	PREDICTED HIGH	PREDICTED MEDIUM	PREDICTED STANDARD
Actual High	4	0	1
Actual Medium	0	25	0
Actual Standard	0	0	10

Table 13. Results of classification probability on the childhood leukemia test data set after application of SMOTE.

ACTUAL	PREDICTED HIGH	PREDICTED MEDIUM	PREDICTED STANDARD
1 High	0.8	0.2	0
2 High	1	0	0
3 High	1	0	0
4 High	0.8	0.2	0
5 High	0.4	0	0.6
6 Medium	0	1	0
7 Medium	0	1	0
8 Medium	0	0.8	0.2
9 Medium	0	0.8	0.2
10 Medium	0	1	0
11 Medium	0	0.6	0.4
12 Medium	0	1	0
13 Medium	0	0.8	0.2
14 Medium	0	1	0
15 Medium	0	0.6	0.4
16 Medium	0	0.8	0.2
17 Medium	0	1	0
18 Medium	0	0.8	0.2
19 Medium	0	1	0
20 Medium	0	0.8	0.2
21 Medium	0	1	0
22 Medium	0	0.6	0.4
23 Medium	0	1	0
24 Medium	0	1	0
25 Medium	0	1	0
26 Medium	0	0.6	0.4
27 Medium	0	0.6	0.4
28 Medium	0	1	0
29 Medium	0	1	0
30 Medium	0	0.8	0.2
31 Standard	0	0.2	0.8
32 Standard	0	0	1
33 Standard	0	0	1
34 Standard	0	0	1
35 Standard	0.2	0	0.8
36 Standard	0	0.2	0.8
37 Standard	0.1	0.1	0.8
38 Standard	0.2	0.2	0.6
39 Standard	0.2	0.2	0.6
40 Standard	0.4	0	0.6

of 0.88 indicates a substantial performance improvement of the NN classifier by involving the dimensionality reduction step. Further improvements can be achieved if we apply the feature-weighting step and use a weighted feature-based similarity. Similar to the results on the childhood leukemia data set, the results of the *t*-test applied on the outcomes of the weighted NN classifier indicate that assigning different weights to features improves the classification accuracy of the NN algorithm (Table 14).

Table 14. Average balanced accuracy results of the three public microarray data sets processed by the case-based retrieval framework.

DATASETS	FS/NN	FS/DR/NN	FS/DR/FW/NN
NCI	0.68	0.88	0.95
Colon	0.86	0.93	–
Prostate	0.90	0.98	–

The framework is also applied on the Colon data set, which is processed by two steps of the framework, namely, feature selection and dimensionality reduction, before presenting the data set to the NN classifier. An average balanced accuracy of 0.93 is achieved without applying feature weighting. With reference to the Prostate cancer data set, an average balanced accuracy of 0.98 is achieved with feature selection and dimensionality reduction.

Table 14 represents the average balanced accuracies of the three data sets obtained at each step of the framework, and it mainly shows the effect of the dimensionality reduction step on the NN classifier.

Conclusion

A case-based retrieval framework is proposed in this article for gene expression similarity measurements. The framework initially applies the feature selection algorithm BIRF to select the features relevant to a specific trait of interest, followed by a dimensionality reduction algorithm KPCA. KPCA reduces the dimensionality of the childhood leukemia data set by projecting the data set to a lower dimensional space for better calculation of distance measurements. A weight-learning GA is proposed for feature weighting in the NN classifier. The weighted NN classifier has been successfully applied and it enhances the classification performance. The results show that the weight-learning GA improved the unweighted NN algorithm. Introducing weights to the features in the NN algorithm leads to improvement in the classification performance. SMOTE approach also provides an improvement in the classification of imbalanced class data sets.

The ultimate goal of this study is to apply this case-based retrieval framework to a CBR system so that we can have a clinical tool to help predict the risk of relapse for childhood leukemia sufferers by comparing them to previous patients based on their gene expression measurements.

Acknowledgments

The authors thank The Children's Hospital at Westmead for providing the childhood leukemia data set and for giving the biological point of view for the results. DRC and PJK are

grateful for funding from the Cancer Institute NSW, which contributed to data analysis in this study.

Author Contributions

Conceived the framework and designed the experiments: AA, PJK. Worked on the CBR section: MG. Worked on the literature review: AB. Performed experiments and analyzed the data: AA. Worked on the biological analysis of the results: DRC. All authors read and approved the paper.

REFERENCES

1. Watson I. Case-based reasoning is a methodology not a technology. *Knowl Based Syst*. 1999;12(5):303–8.
2. Smyth B, Keane M. Adaptation-guided retrieval: questioning the similarity assumption in reasoning. *Artif Intell*. 1998;102:249–93.
3. Marling C, Whitehouse P. Case-based reasoning in the care of Alzheimer's disease patients. *Case-Based Reason Res Dev*. 2001;1:702–15.
4. Lieber J, Bresson B. Case-based reasoning for breast cancer treatment decision helping. *Adv Case-Based Reason*. 2000;22:1–10.
5. Díaz F, Fdez-Riverola F, Corchado JM. GENE-CBR: a case-based reasoning tool for cancer dignosis using microarray data sets. *Comput Intell*. 2006; 22(3–4):254–68.
6. Leake D. Case-based reasoning. *Knowl Eng Rev*. 1994;9(01):61–4.
7. Ma S, Li J, Liu D. The case retrieval strategy based on hierarchical clustering. *Second Pacific-Asia Conference on Web Mining and Web-based Application*. New York NY: IEEE; 2009:81–5.
8. Wess S, Althoff K, Derwand G. Using k-d trees to improve the retrieval step in case-based reasoning. *Topics in Case-Based Reasoning*. London, UK: Springer-Verlag; 1994:167–81.
9. Bellman RE. *Adaptive Control Processes: A Guided Tour*. Princeton, NJ: Princeton University Press; 1961:4.
10. Beyer K, Goldstein J, Ramakrishnan R, Shaft U. *When is "Nearest Neighbor" Meaningful? In Database Theory – ICDT*'99. Berlin, Heidelberg: Springer; 1999:217–35.
11. Baldi P, Hatfield G. *DNA Microarrays and Gene Expression: from Experiments to Data Analysis and Modeling*. Cambridge: Cambridge University Press; 2002.
12. Salamó M, López-Sánchez M. Rough set based approaches to feature selection for case-based reasoning classifiers. *Pattern Recognit Lett*. 2011;32(2):280–92.
13. Arshadi N, Jurisica I. Maintaining case-based reasoning systems: a machine learning approach. *Adv Case-Based Reason*. 2004;1:439–520.
14. Park CS, Han I. A case-based reasoning with the feature weights derived by analytic hierarchy process for bankruptcy prediction. *Expert Syst Appl*. 2002;23(3):255–64.
15. Huang ML, Hung YH, Lee WM, Li R, Wang TH. Usage of case-based reasoning, neural network and adaptive neuro-fuzzy inference system classification techniques in breast cancer dataset classification diagnosis. *J Med Syst*. 2012;36(2):407–14.
16. Golub T, Slonim D, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*. 1999;286(5439):531–7.
17. Gordon G, Jensen R, Hsiao L, et al. Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer Res*. 2002;62(17):4963.
18. Anaissi A, Kennedy PJ, Goyal M, Catchpoole DR. A balanced iterative random forest for gene selection from microarray data. *BMC Bioinformatics*. 2013;14:1–10.
19. Duan K, Rajapakse J, Wang H, Azuaje F. Multiple SVM-RFE for gene selection in cancer classification with expression data. *IEEE Trans Nanobioscience*. 2005;4(3):228–34.
20. Inza I, Sierra B, Blanco R, Larrañaga P. Gene selection by sequential search wrapper approaches in microarray cancer class prediction. *J Intell Fuzzy Syst*. 2002;12:25–33.
21. Bellman R. Dynamic programming and Lagrange multipliers. *Proc Natl Acad Sci U S A*. 1956;42(10):767.
22. Schölkopf B, Smola A, Müller K. Kernel principal component analysis. *Artif Neural Networks – ICANN*. 1997;97:583–8.
23. Wettschereck D, Aha DW, Mohri T. A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms. *Artif Intell Rev*. 1997;11(l–5):273–314.
24. Holland JH. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. MIT Press: Cambridge; 1992.
25. Chawla N, Hall L, Kegelmeyer W. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res*. 2002;16:321–57.
26. Sáez JA, Luengo J, Stefanowski J, Herrera F. Managing borderline and noisy examples in imbalanced classification by combining SMOTE with ensemble filtering. In: Corchado E, Lozano JA, Quintián H, Yin H, eds. *Intelligent Data Engineering and Automated Learning-IDEAL* 2014. Berlin, Heidelberg: Springer; 2014:61–8.
27. Verbiest N, Ramentol E, Cornells C, Herrera F. Preprocessing noisy imbalanced datasets using SMOTE enhanced with fuzzy rough prototype selection. *Appl Soft Comput*. 2014;22:511–7.
28. Zhou B, Yang C, Guo H, Hu J. A quasi-linear SVM combined with assembled SMOTE for imbalanced data classification. *Neural Networks (IJCNN), The* 2013 *International Joint Conference on*. New York NY: IEEE; 2013:1–7.
29. Ross DT, Scherf U, Eisen MB, et al. Systematic variation in gene expression patterns in human cancer cell lines. *Nat Genet*. 2000;24(3):227–35.
30. Alon U, Barkai N, Notterman D, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci U S A*. 1999;96(12):6745–50.
31. Ben-Dor A, Bruhn L, Friedman N, Nachman I, Schummer M, Yakhini Z. Tissue classification with gene expression profiles. *J Comput Biol*. 2000;7(3–4):559–83.
32. Brazma A, Vilo J. Gene expression data analysis. *FEBS Lett*. 2000;480:17–24.
33. Getz G, Levine E, Domany E. Coupled two-way clustering analysis of gene microarray data. *Proc Natl Acad Sci U S A*. 2000;97(22):12079.