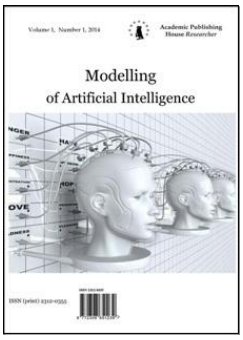
Published in the Russian Federation
Modeling of Artificial Intelligence
Has been issued since 2014.
ISSN: 2312-0355
Vol. 6, Is. 2, pp. 171-182, 2015

DOI: 10.13187/mai.2015.6.171
www.ejournal11.com



UDC 004

Data-Mining Techniques to Classify Microarray Gene Expression Data Using Gene Selection by SVD and Information Gain

[1] Halit Vural
[2] Abdülhamit Subaşı

[1] International Burch University, Bosnia and Herzegovina
Francuske revolucije bb. Ilidza, Sarajevo 71000
PhD (Information Technologies)
E-mail: hvural@ibu.edu.ba
[2] International Burch University, Bosnia and Herzegovina
Francuske revolucije bb. Ilidza, Sarajevo 71000
Doctor of Information Technologies, professor
E-mail: asubasi@ibu.edu.ba

Abstract

Microarray data analysis can provide valuable information for cancer prediction and diagnosis. One of the challenges for microarray applications is to select an appropriate number of the most significant genes for data analysis. Besides, it is hard to accomplish a satisfactory classification results by using data mining techniques because of the dimensionality problem and the over-fitting problem. For this reason, it is desirable to select informative genes in order to improve classification accuracy of data mining algorithm. In this study, Singular Value Decomposition (SVD) is used to select informative genes and reduce the redundant information. Furthermore, information gain is used to determine useful features of data to get better classification performance. In the last step, the classification technique is applied to the selected features. We conducted some experimental work on subset of features from available datasets. The experimental results show that the proposed feature selection and dimension reduction gives better classification performance in terms of the area under the receiver operating characteristic curve (AUC) and the prediction accuracy.

Keywords: data mining techniques; gene selection; information gain; microarray gene expression; singular value decomposition (SVD).

Introduction

Cancer is one of the major causes of death in all countries. Many different cancer types have been diagnosed in various organs and tissues. Since it is related with genetic abnormalities in the cell, DNA microarrays, which allow the simultaneous measurement of expression levels of genes, have been used to characterize gene-expression profiles of tumor cells. Thus, these measurements allow capturing anomalies in the cell. Microarray technology also allows a standardized, clinical assessment of oncological diagnosis and prognosis. However, those studies are specific to limited cancer types, and their results have limited use due to inadequate validation in large patient age

group. Beside its challenging data retrieval process, microarray techniques have been used as a promising tool to improve cancer diagnosis and treatment in recent decades. On the other hand, data produced from gene expressions contain high level of noise and the intense number of genes relative to the number of available samples. For this reason, it shows a great challenge for classification and statistical techniques with microarray data. In this study, we applied some methodology to resolve these kinds of problems. After then, data mining techniques have been used to classify gene expression data of cancer and normal tissues.

Discovering genes commonly regulated in cancer may have an important implication in understanding the common biological mechanism of cancer. Many researchers analyzed the global gene-expression profiles of various cancer types over the past years. They described many gene-expression signatures that are associated with cancer progression, prognosis and response to therapy [1]. The tumor type-specific signatures from these studies show little convergence in gene structures. Recent research in molecular oncology have provided few useful molecular markers of tumors, due to limitations with sample availability, acquisition, integrity, preparation, and identification [2]. And yet, cancer remains a challenge to catch an essential, common transcriptional feature of neoplastic transformation and progression because of its heterogeneous characteristic.

There are some studies that focus on decision-making support system to help doctors and clinicians in their diagnosis and prognosis process [2]. Those kinds of systems usually lie in three phases: feature selection, modeling and knowledge discovery [3]. There are many data mining techniques that have been used to model the gene expression data in this manner. In last decades, supervised learning methods attracted more attention from many researchers. Among them the artificial neural network (ANN) became mostly studied supervised learning method deployed in medical research [4]. Other studies focus on Bayesian supervised learning methods, decision trees [5] and support vector machines (SVM) [6]. Some researchers tried multiple classifier systems to improve the performance of single approach [6]. In the process of clarifying the disease problem, the insignificant or redundant genes should be eliminated in microarray datasets. Otherwise, classification performance becomes defective. So, it is necessary to select informative genes to get the best performance from classification of microarray data. For this reason, reducing the number of features using some technique to get relevant informative genes is needed. We studied the effect of this process to the classification performance. Also we propose a model that increases the classification performance of genes according to cancer types. This method differs from other studies in combining dimension reduction and feature selection techniques. We conducted some experimental work on some subset of features from existing microarray datasets. As the performances of classifier are compared, it is clearly seen that the feature selection and dimension reduction from a high-dimensional dataset slightly increases the performance of classification on gene expression data of cancer and normal tissues by selecting highly relevant and informative genes.

In this study, we used a feature selection method called Singular Value Decomposition (SVD) to select informative genes and reduce the redundant information. Then information gain (IG) is used to determine useful features of data to get better classification performance. In the last step, the classification technique is applied to the selected features. The block diagram of proposed gene selection methodology is shown in Figure 1. The proposed dimension reduction (SVD), feature selection approach (IG), and data mining tools will be introduced in Section 0. In Section 0, we represent the experimental outcomes and discussion on 6 common microarray datasets, which becomes a proof to the proposed approach. Our approach improves the average classification accuracy rates as well as it reduces the variation of classification performance. Section 0 concludes this paper.

Figure 1: Block diagram of proposed gene selection system

Materials and methods

We applied the method to broadly used public microarray datasets which are AML-ALL, Leukemia-1, Colon, SRBCT, Lung Cancer and DLBCL. Those cancer types are common in human diseases and these datasets are publicly available at (http://www.gems-system.org) [7]. These datasets are used commonly in the researches to make comparisons with other problem solving models.

Leukaemia (AML-ALL). The leukaemia dataset was taken from a collection of leukaemia patient samples reported by Golub et al.[8]. This dataset is well-known as a benchmark for microarray studies. It comprises measurements corresponding to acute lymphoblast leukaemia (ALL) and acute myeloid leukaemia (AML) samples from peripheral blood and bone marrow. The dataset consisted of 73 samples: 25 samples of AML, and 48 samples of ALL. Each sample was taken from bone marrow and were analysed using Affymetrix microarrays containing 7,129 genes. This dataset is divided into two subsets as train and test by [8]. We used those datasets as two parts as other researchers do. The training data consists of 38 samples (27 ALL and 11 AML), and the test data consists of 34 samples (20 ALL and 14 AML).

Leukaemia-1. There are three types of classes in Leukaemia-1 cancer dataset: acute lymphoblastic leukaemia (ALL) B-cell, ALL T-cell, and acute myeloid leukaemia (AML). There are 72 samples totally which contains 5,327 genes [8].

Colon Tumor. The samples in colon dataset are measurements of colon adencarcinoma specimens snap-frozen in liquid nitrogen within 20 minutes of removal from patients [1]. The microarray dataset consists of 22 normal and 40 tumor tissue samples. In this dataset, each sample contains 2,000 genes. Dataset is publicly available at (http://genomics-pubs.princeton.edu/oncology/affydata/ index.html).

SRBCT (Small Round Blue Cell Tumors). This dataset includes 4 different small round blue cell tumors: Ewing family tumor (EWS), rhabdomysarcoma (RMS),, neuroblastoma (NB) and Burkitt lymphoma (BL). The training set and testing set contain 63 and 20 samples, respectively. The cDNA microarrays consist of 2,308 genes. Small round blue cell tumors (SRBCT) of childhood are diagnosed by using 96 perfectly selected features for predicting the test data classes [9].

Lung Cancer. This dataset contains the gene expression information on 203 lung tissue samples and 12,600 features [10]. The samples are categorized into five diagnostic classes according to histological diagnose. Those classes include four different lung tumors: adenocarcinomas (AD), small-cell lung carcinomas (SMCL), squamous cell carcinomas (SQ) and carcinoids (COID) and normal lung tissue (NL).

DLBCL (The diffuse large B-cell lymphoma and follicular lymphomas). This data set contains 58 samples from DLBCL patients and 19 samples from follicular lymphoma. The gene expression profiles were analysed using Affymetrix human 6,817 oligonucleotide arrays [11].

Singular Value Decomposition (SVD). Singular value decomposition (SVD) is a data-driven mathematical framework that can be used to reduce dimension for the gene expression data [12]. SVD is a common technique for analysis of multivariate data, like principle component analysis (PCA). There could be thousands of measurements of genes in a single microarray dataset. When the experiments includes less than ten assays or more than hundreds, finding singular values has the ability to extract small signals from noisy gene expression data [12]. Using SVD, we easily reduced the number of features in gene expression data.

We explain mathematical definition of the SVD here to clarify the process of data. We consider definition in Van Der Heijden et al. [13] and let H denote an N×M matrix of real-valued microarray data which can be composed into the product:

$$H = U\sum\sum^{T} \quad (1)$$

where U is an orthonormal N×R matrix, $\sum$ is a diagonal R×R matrix, and V is an orthonormal M×R matrix. R is the rank of the matrix H, where without loss of generality and $R \leq M \wedge R \leq N$. The matrix $V=[v_0 \ldots v_{M-1}]$ contains the (unit) eigenvectors v_i of H^TH. Leaving the proofs to Van der Heijden et al. [13] , the correlated eigenvalues are all places on the diagonal of the matrix S where :

$$\sum = S^{1/2} \quad (2)$$

We calculate the square roots of all (diagonal) elements in S to obtain $\sum$ matrix. Without the loss of generalization level it is assumed that they are sorted in descending order as

$$S_{i,j} \geq S_{i+1,j+1}. \quad (3)$$

As stated in [21], the i^{th} row of H forms the transcriptional response of the i^{th} gene g_i, the j^{th} column of H forms the expression profile of j^{th} assay a_j . We refer to the columns of U, the left singular vectors $\{u_k\}$ as eigen-assays and the rows of V^T, the right singular vectors $\{v_k\}$ as eigen-genes. Eigen-assays and eigen-genes form orthonormal basis for genome-wide array expression profiles and gene transcriptional response respectively. SVD is then the linear transformation of the expression data from the Ngenes×Msamples space to the reduced M-eigenassays×M-eigenassays as below:

$$H=USV^T \quad (4)$$

The elements of S other than the diagonal are zero, and the diagonal elements are called the singular values. Thus, $S=diag(s_1, \ldots , s_M)$ and $s_k>0$ for $1 \leq k \leq r$, and $s_i=0$ for $(r+1) \leq k \leq M$ where s_i indicates the relative importance of the j^{th} eigen-gene and eigen-assay in the explanation of data. The i^{th} eigen-gene is represented only in the related j^{th} eigen-assay with the related eigen-expression level s_i. Hence, the expression of each eigenvector (eigen-gene or eigen-assay) is decoupled from that of all other eigenvector. The decorrelation of the eigenvectors shows that the eigen-gene denotes independent gene expression pattern across all arrays and the related eigen-assay denotes independent sample states across all genes [12]. Thus, SVD is proper for stability analysis of the ratio between the largest and smallest singular values. The sensitivity of the inverse to noise in microarray data can easily be measured by this ratio [13].

Feature Selection. The researchers regard feature selection as an important issue in classification. Processing thousand genes in one dataset makes significant differences in classifying microarray data. Most of the existing classifiers need a feature selection scheme at their design and it is very important to select a good feature selection method. Otherwise, performance can be seriously degraded. Gene expression data includes a large number of gene expression values in a very small sample size. Moreover, the high-correlation between many genes leads redundancy [6].

Feature selection aims to find out a powerful subset within a database to reduce the number of features presented to the modeling process [6]. Feature selection is the major bottleneck of machine learning and data mining [14]. In this study, we tried to improve the performance of the classification process by using some feature selection schemes.

Table 1 gives the selected genes used in the experiments. These genes are gathered after dimension reduction with SVD and feature selection with Information Gain (IG) methods. These selected subsets are used to get high performance from the experiments. The effectiveness of using SVD and IG together for feature selection is discussed in Section 0.

Table 1: Selected features of microarray datasets in this study

Name of gene-expression dataset	Number of features in the dataset	Number of Samples	Number of classes	Reduced number of features with SVD	Number of selected features by InfoGainAttributeEval
(ALL-AML)	7,129	73	2	38; 35	3
Leukemia-1	5,327	72	3	72	6
Colon	2,000	62	2	62	3
SRBCT	2,308	83	4	83	8
Lung Cancer	12,600	203	5	203	9
DLBCL	6,817	77	2	77	7

Attribute Subset Evaluators. Information Gain (IG) is a statistical property that measures how well a given attribute separates the training sample according to its target classification. IG is used to select proper attributes among given ones of the training set. In order to determine information gain precisely, entropy value is used commonly. Entropy is defined as the impurity of an arbitrary collection of given samples [15]. Entropy can be calculated for the target attribute which takes n different values, then the entropy of collection A can be written as:

$$Entropy(A) = \sum_{k=1}^{n}(-p_k \log_2 p_k) \quad (5)$$

Here, p_k is the proportion of A that belongs to class k.
Relative to a collection of samples A, the information gain G(A, a) of an attribute a is:

$$Gain(A, a) \equiv Entropy(A) - \sum_{v \in Values\ (a)} \frac{|A_v|}{|A|} Entropy(A) \quad (6)$$

The expected entropy described in Eq. (6) is the sum of the entropies of each subset A_v. Here, Gain (A,a) is the information providing the target value, given the value of some other attribute a [15]. There may be some redundant or irrelevant attributes that causes accuracy to decrease [16]. We used InfoGainAttributeEval and Ranker evaluation tools of WEKA [17] for gene selection that is relevant to clinical outcomes.

InfoGainAttributeEval tool, which evaluates feature based on information gain, can be used to select relevant genes of clinical outcomes. It uses MDL-based discretization method to discretize numeric features [16].

Our second evaluation tool is Ranker. It ranks individual features according to their evaluation. Single-feature evaluating methods are used with the Ranker search method to build a ranking list from which Ranker discards a given number [16]. Table 1 lists the selected attributes by InfoGainAttributeEval and Table 2 lists their Ranker results.

Table 2: Selected features with ranker values from cancer datasets

Dataset Name	Number of selected features	Selected Features and their given importance values by Ranker Method
Leukemia (AML-ALL) train set	3	F3:0.377;F2:0.266;F6:0.264
Leukemia-1	6	F3:0.716; F10:0.409; F6:0.285; F4:0.246; F5:0.225; F1:0.211

Colon Tumor	3	F3:0.297; F2:0.197; F6:0.186
SRBCT	8	F4:0.536; F7:0.488; F6:0.447; F8:0.329; F3:0.306; F2:0.263; F10:0.245; F16:0.217; F13:0.177
Lung Cancer	9	F3:0.725; F4:0.431; F2:0.42; F7:0.382; F1:0.315; F5:0.216; F8:0.16; F6:0.107; F24:0.101
DLBCL	7	F7:0.326; F2:0.176; F16:0.167; F12:0.167; F1:0.167; F8:0.142; F3:0.142

Data Mining Techniques. Different data mining techniques were used for microarray classification and most researchers applied the Support Vector Machine (SVM) for classifying the microarray dataset and they obtained very good results [6]. We additively used Artificial neural networks (ANNs) and decision tree algorithms to compare the results with SVM. As a common approach in classification studies, ANNs are used for comparison to other methods.

Artificial Neural Networks (ANNs). The neural network was first recognized as a useful for nonlinear statistical modeling [18]. Artificial Neural Networks is is developed as a model of brain. A neuron represents a decision mechanism which is similar to human brain. The neurons fire when the value of the signal passes a predefined threshold. Multi layer perceptron (MLP) neural networks consist of units arranged in layers [19]. Each layer consists of nodes in the fully connected network. Each MLP is consists of a minimum of three layers composed of an input layer, one or more hidden layers and an output layer. The input layer distributes the inputs to subsequent layers and input nodes have might have different activation functions. Each hidden unit node and each output node require thresholds related with them in addition to the weights. The hidden unit nodes must have nonlinear activation functions and the outputs might have different activation functions. As a result, each signal feeding into a node in a succeeding layer has the original input multiplied by a weight with a threshold added and then is delivered through an activation function that may be linear or nonlinear (hidden units)[19].

Support Vector Machines (SVMs). Vapnik [20] developed a powerful data mining technique called the support vector machine (SVM) which is based on the statistical learning theory and it shows a high classification performance based on the principle of structural risk minimization [20]. It minimizes the total empirical risk and the bound of risk confidence. The preference of an SVM classifier over traditional classifiers like multi-layer perceptron (MLP) networks is its better global minimization and generalization ability [21].

Typically, a conventional classifier can be defined to separate the positive samples from negative ones. Suppose we have data points in the training set that are vectors of n members. We are supposed to develop a classifier that finds a hyper plane which separates negative and positive members. Real-time problems are usually not ideal for separation and they involve noisy data. That means a hyper plane that separates all positive members from negative ones cannot be defined easily. In that case, SVM maps a given training set into a possibly high-dimensional feature space to determine a hyper plane that separates the members. There may be many high-dimensional candidate planes available for that job. SVM selects the most proper hyper plane among those candidates. In that attempt, SVM can select that maintains a maximum margin in the given training set. That hyper plane will lead to maximal generalization ability for the classification of unseen data. A kernel function can be defined in the algorithm to find a separating hyper plane [21].

$$x_i \in R^d, \{x_i, y_i\}, i=1,\ldots,N, \{y_i | y_i \in \{-1,1\}\} \quad (7)$$

It is assumed that the training set is linearly separable after being mapped into a high-dimensional data space by a non-linear function φ(x), the training data satisfy

$$\begin{aligned} w^T\phi(x_i) + b \geq 1 for y_i = 1 \\ w^T\phi(x_i) + b \leq -1 for y_i = -1 \end{aligned} \quad (8)$$

These two formulas define two parallel hyperplanes and the distance between them is $2/\|w\|$ which is called margin.

$$y_i(w^T(x_i) + b) - 1 \geq 0, \forall i \tag{9}$$

Larger margin allows better classification performance.

$$\begin{gathered} minimize: \zeta(w, b, \zeta_i) = \frac{1}{2} w^T w + C \sum_{i=1}^{N} \zeta_i \\ \text{subject to: } y_i[w^T(x_i) + b] \geq 1 - \zeta_i, for i = 1, \ldots, N \\ \zeta_i \geq 0 for i = 1, \ldots, N \end{gathered} \tag{10}$$

The classifier uses training set to maximize the classification margin and minimize the total error terms of training samples.

$$\begin{gathered} y_i[w^T(x_i) + b] \geq 1 - \zeta_i \\ \zeta_i \geq 0 \end{gathered} \tag{11}$$

If the input space is transformed into a high-dimensional, non-linear space, by using a kernel function independent from feature space. Kernel function K(x, x_i) can be represented as:

$$y(x) = sign(w^T \phi(x) + b) = sign\left(\sum_{i=1}^{N} \alpha_i \, y_i K(x, x_i) + b\right) \tag{12}$$

where α is a Lagrangian variable.

Random Forest

Random forests combine tree predictors such that each tree depends on the values of a random vector. Those vectors are sampled independently and with the same distribution for all trees in the forest [22]. Each variable (feature) in the forest has an importance measure which is an internal estimation of the decrease in the classifier's overall accuracy. So, that particular variable can be considered to have more classification ability if it has a larger importance measurement. Random Forest classifier is an extended version of classification tree by an integration of the bagging idea [22]. That is because it constructs many classification trees using various assisting samples of a fixed-size matrix from the original data. When a new input vector is classified, it computes every tree in the forest for that vector. Each computation gives a classification result to be compared with other results. The algorithm chooses the most accurate classification among others (over all the trees in the forest) [22].

Results and Discussion

We evaluate the performance of the proposed feature extraction (SVD and IG) approach for different classifier based on 6 well-known gene expression datasets, namely Leukemia-AML-ALL, Leukemia-1, Colon Tumor, Lung Cancer, SRCBT and DLCBL. Table 1 shows the 6 microarray datasets with their properties. Dimension reduction and feature selection (SVD and IG) algorithms combined with SVM, ANN, and Random Forest classifiers on the 6 datasets as shown in Figure 1. In present study, K-fold cross validation [16] is used to evaluate the performance of the classifiers. The cross-validation accuracy (CVA) is the average of the k individual accuracy measures

$$CVA = \frac{1}{k} \sum_{j=1}^{k} A_i \tag{13}$$

where k (10 in this case) is the number of folds used, and A_i is the accuracy measure of each fold, i = 1, . . . , k.

Furthermore, three different classifiers are compared against each other on the basis of average accuracy, true negative rate (specificity), true positive rate (sensitivity), and the value of area under the ROC curve (AUC). ROC (Receiver operating characteristic) curve evaluates the classifier's ability of discrimination.The true accuracy can be calculated by using:

$$ACC = (TP + TN) / (P+N) \quad (14)$$

where TP is true positives, TN is true negatives, FP is false positives, and FN is false negatives. Another measure that evaluates the results is F measure:

$$F\text{-measure} = 2TP / (2TP+FP+FN) \quad (15)$$

AUC gives an overall accuracy measure that is independent of any particular treshold. That measure can be used as the index of performance [23]. Likewise AUC identifies sensitivity and specifity values.

Experimental Results

In this study, we used feature reduction to decrease the number of features and eliminate the noise in the data (Figure 1). After reducing the number of features, we selected highly relevant genes which increase the classification accuracy. Although selecting informative genes is the most important issue in classification, dimensionality of the data does not allow good classification accuracy because of the noise and irrelevant data. Another problem is that there are not enough samples according to the number of features. It is not possible to get high classification results with high-dimensional features and small number of instances. Hence, at first step of the method, we applied SVD feature reduction to get less number of features which is close to the number of instances. As a second step, to get high accuracies for classification techniques, we applied information gain to determine more informative genes among the reduced number of features. We used Info Gain Attribute Eval attribute evaluation tool of WEKA to select informative features. With those small numbers of relevant features to clinical outcomes, we can expect more accurate classification results. Numbers of those features are listed in Table 1 as 3, 3, 6, 3, 8, 9, and 7 for Leukaemia (AML-ALL) train and test sets, Leukemia-1, Colon Tumor, SRBCT, Lung Cancer, and DLBCL, respectively. Selected features are listed in Table 1 and 2. Table 2 lists the ranking values of selected features as well. Those ranking values are recorded by the Ranker method in WEKA. In the last step of the experiments, we applied classification methods to the selected features. Those steps are summarized in Figure 1.

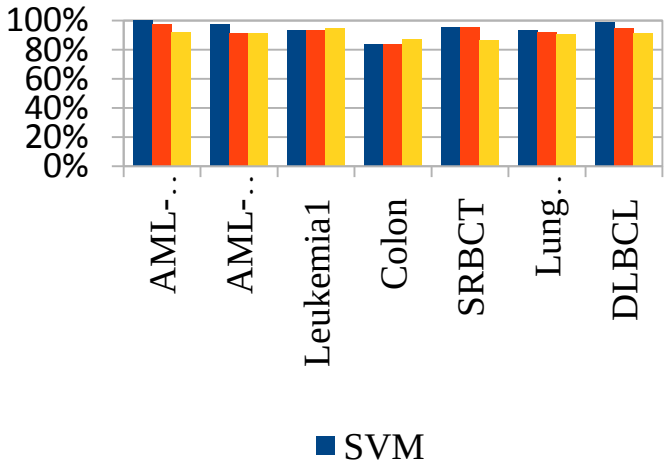


Figure 2: Graphical representation of evaluation performance of classifiers for different microarray dataset

We divided the whole microarray data into training and test sets and applied 10-fold cross validation as in [24] to calculate the performance of each model. In the process, the model is trained using nine data parts of the collection and the remaining part is utilized in testing [25]. These techniques have been extensively used in previous studies and all have shown high accuracy in classification of microarray data.

We applied dimension reduction and feature selection by using SVD and IG methods and we recorded classification accuracies to compare the performances in Table 3 and Figure 2. It can be seen easily that the performance of our gene selection is better than results of [8] and comparable to the results of [26]. It can be seen from Table 3 and Figure 2 that SVD+IG performs better with SVM than other classifiers in microarray data analysis, which shows the success of the proposed approach.

Experimental results obtained in this study demonstrate that features extracted using SVD and IG improved the classification accuracy of classifiers. Similarly, the related AUC and F-measure are also calculated in the same way (Table 4 and 5). We tried to choose the best parameter values for SVM classifier to maximize the prediction. We applied the same technique to find the best parameters for all other classifiers in the study.

Table 3: Classification algorithms and their accuracies(%) for cancer datasets

Classifier	**AML-ALL (train)**	**AML-ALL (test)**	**Leukemia-1**	**Colon**	**SRBCT**	**Lung Cancer**	**DLBCL**
SVM	100	97.14	93.06	83.87	95.18	93.60	98.70
ANN	97.37	91.43	93.06	83.87	95.18	92.12	94.81
Random Forest	92.11	91.43	94.44	87.10	86.75	90.64	90.91

The experimental results are shown in Table 3, 4, and 5. In Figure 2, those results are compared using bar chart. The classification performances of methods compared with each other are shown in Table 3. We used Leukemia (AML-ALL) dataset as in two parts as in [8]. First part is training set with 38 samples, and the second is testing set with 35 samples. To compare the methods for train and test sets, SVM gives highest accuracies for both training and testing sets as 100% and 97.14%, respectively. AUC values with SVM are as 1.00 and 0.971. The following method is ANN which gives 97.37% and 91.43% accuracies where AUC values are as 0.974 and 0.914. Random Forest gives accuracies as 92.11% and 91.43% for training and testing sets respectively. AUC values for Random Forest are 0.917 and 0.915. As we compare F-Measures for the algorithms accordingly, SVM gives 1.00 and 0.964 for training and testing sets respectively. ANN gives similar measures as 1.00 and 0.929. Random Forest follows ANN with the measures 0.941 and 0.983.

Random Forest shows the highest accuracy as 94.44% for Leukemia-1 beside other classification tree algorithms. SVM and ANN follow Random Forest with 93.06% accuracies. The AUC values for Leukemia-1 are 0.928, 0.932, and 0.943 for SVM, ANN, and Random Forest, respectively. The F-Measures are 0.928, 0.949, and 0.966 accordingly.

By the nature of Colon tumor data, dataset gives lower results for all of our classification algorithms. We approve that our results are similar with other studies. For instance, Rojas-Galeano et al. [27] gets 88% with their kernel-based algorithm. Futschik et al. [2] reported that Colon dataset gives lower accuracy compared to other datasets. They had 90% accuracy by their method. Xiong et al. [28] reported 87% for their test set in Colon tumor dataset. Our result for Colon data is 83.87% for SVM and ANN, and 87.10% for Random Forest classifier. F-measures are 0.84, 0.839, and 0.871, respectively.

As shown in Table 3-5 and Figure 2, total accuracy, AUC and F-measures achieved with the SVM classifier were equal to 95.18%, 0.952 and 0.974 respectively for SRCBT. These results were better than those achieved by the Random Forest classifier. ANN classifier gives same accuracy and AUC value where F-Measure is 0.989. Indeed, the total accuracy, AUC, and F-measure values are equal to 86.75%, 0.866, and 0.973 respectively for the Random Forest classifier. From these comparisons, it is found that SVM has achieved a better classification performance (higher classification accuracy rate, AUC and F-measure) than the other algorithms in designing a classification system for all 6 microarray data sets. The performance result of classifiers verified for microarray data classification with two special considerations: feature extraction and selection of the classifier. The most important attributes which are derived from the microarray data are dependent upon the feature selection and dimension

reduction methods used. The selected attributes, which are most outstanding for microarray data classification, should be used as the inputs of the classifier.

Table 4: Classification algorithms and their AUC for cancer datasets

Classifier	AML-ALL (train)	AML-ALL (test)	Leukemia-1	Colon	SRBCT	Lung Cancer	DLBCL
SVM	1.00	0.971	0.928	0.84	0.952	0.936	0.987
ANN	0.974	0.914	0.932	0.839	0.952	0.922	0.949
Random Forest	0.917	0.915	0.943	0.871	0.866	0.898	0.908

Table 5: Classification algorithms and their F-Measure for cancer datasets

Classifier	AML-ALL (train)	AML-ALL (test)	Leukemia-1	Colon	SRBCT	Lung Cancer	DLBCL
SVM	1.00	0.964	0.928	0.834	0.952	0.936	0.987
ANN	0.974	0.914	0.932	0.839	0.952	0.922	0.949
Random Forest	0.917	0.915	0.943	0.871	0.866	0.898	0.908

Lung Cancer dataset gives 93.60% with SVM and 92.12% with ANN. Random Forest is the highest among other classification algorithms with 90.64%. AUC values for Lung Cancer datasets are 0.936, 0.922, and 0.898 for SVM, ANN, and Random Forest, respectively. F-Measures for ANN, Random Forest, and SVM are respectively as 0.979, 0.967, 0.955.

Out of 58 samples in the DLBCL dataset, with10-fold cross-validation, Table 3 shows that our model achieved competitive classification accuracies with fewer genes. We can easily see that the method is consistent over different classifiers, i.e., with the classifiers ANN, SVM, and Random Forest; our model all achieved significantly good classification accuracies. Table 3 shows the prediction performance of the selected genes and the power of gene sets involded in our study. We should note here that we used regular classifiers and we didn't apply any heuristic approach to achieve these performances. DLBCL dataset gives high accuracies and it has the highest average among all other datasets. For instance, SVM gives the highest accuracy as 98.7%.

When we compare the results overall, we can see that SVM and ANN classifiers give high performance compared to other classifier algorithms. Besides we saw that our algorithm works under different validations, and there is no significant difference. Also Random Forest (RF) is applied by considering its robustness and it gives competitive high accuracies on all datasets including Colon dataset, which is 87.1%. Colon datasets is low in average for all classifier algorithms where Leukaemia (AML-ALL) is well classified by all classifiers. Accordingly, comparing the results by Figure 2 we can see that SVM gives higher accuracies for all datasets except Leukemia-1. Random Forest is more accurate than other algorithms. SVM and ANN classifiers give same accuracies which are higher than RF for Leukemia-1 and SRBCT.

The results show that our method achieved high accuracies on all datasets we used. And the method catches good results with applied classifiers. It is possible to say that selecting one or more of these classifiers as a proper method to apply gene mining on other datasets.

Discussion

Great dimensionality is a critical problem in analyzing gene expression data. The noise should be filtered out after normalization and set of biomarker genes which are related to cancer should be

selected to get good results in classification process. Furthermore, the main target of gene expression data analysis is to determine biologically relevant genes. Besides, over-fitting is another problem which can be solved by selecting informative genes. Rapaport [29] and Chuang [30] proposed a method that integrates biological priori information in the gene selection process. In this study, we used SVD to decompose microarray data, and we identified the eigen-genes corresponding to select the only genes which play important roles in determining biological effect to cancer. Therefore, we could identify genes in terms of the strength of association with clinical outcomes. We experimentally examined 6 public datasets. The DLBCL, SRBCT, Leukaemia (AML-ALL) test and train sets, Lung Cancer, and Leukemia-1 microarray data using our approach give good results where Colon gave lower results which can be approved as compatible. We then list the results of these datasets here to show our method is good enough and indeed useful and powerful tool for gene selection and classification when diagnosing cancer. Our identification of gene set makes it possible to get highly related to the distributions of cancer and normal samples. SVM and ANN algorithms give high accuracies with all datasets when Random Forest gives lower results. It can be said that the proposed gene selection with SVD and IG works better with SVM and ANN algorithms.

Another issue is discussing Colon tumor data with the given method. Colon dataset gives lower accuracies for all classification algorithms than other datasets. It seems that the method should be alternated to select more relevant features for Colon tumor. Heuristic approach can be useful for this aim. Since the default settings for the classification methods and Ranker, it is possible to get higher accuracies with new settings and different selections.

Conclusion

In this study, a dimension reduction and feature selection scheme is proposed to discover knowledge from the gene expression data. A method is used to reduce features from thousands to tens at first step when second step is the feature-selection part. The experimental results illustrate that the classification with relevant features achieve promising performance on the testing set using a few features than using the whole dataset. We examined 6 public microarray data using our approach and we obtained good results to demonstrate the study here that our method is useful and powerful tool for gene selection and classification when diagnosing cancer. In this approach, we tried to select correct informative marker genes and improved the final classification performance by integrating SVD and IG into our gene selection method. Thus, we tried to define an original pathway information with a better combination of gene expression data. We can easily say here, the results from our study show that the system is effective for gene selection and it is useful in improving classification performance.

We can also say that feature-selection significantly improves the performance of classifiers. The proposed method reduces the number of features, then selects the relevant ones for classification. The proposed method differs from single feature-ranking methods. Actually a more precise methodology can be found to reduce features for better results and also can be designed a better feature selection method to find relevant genes to that cancer type. We can conclude that the features selected by the search methods such as Rank search with the evaluator Information Gain Attribute Eval yields better results. Furthermore, applying more than one feature selection technique (Incremental feature selection) is essential for the effective performance. The results show that the approach described in this study can be used as an effective gene selection tool and improved microarray data classification method for cancer.

References:

1. Alon U., Barkai N., Notterman D.A., Gish K., Ybarra S., et al. (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, p. 6745.

2. Futschik M.E., Reeve A., Kasabov N. (2003) Evolving connectionist systems for knowledge discovery from gene expression data of cancer tissue., Artificial Intelligence in Medicine. 28, pp. 165–189.

3. Zhu Z., Ong Y.S., Dash M. (2007) Wrapper-Filter Feature Selection Algorithm Using a Memetic Framework, IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics). 37, pp. 70–76. doi:10.1109/TSMCB.2006.883267.

4. Xu Y., Selaru F.M., Yin J., Zou T.T., Shustova V., et al. (2002) Artificial neural networks and gene filtering distinguish between global gene expression profiles of Barrett's esophagus and esophageal cancer, Cancer Research. 62, p. 3493.

5. Sevon P., Toivonen H., Ollikainen V. (2006) TreeDT: Tree Pattern Mining for Gene Mapping, IEEE/ACM Transactions on Computational Biology and Bioinformatics. 3, pp. 174–185. doi:10.1109/TCBB.2006.28.
6. Chen Z., Li J., Wei L. (2007) A multiple kernel support vector machine scheme for feature selection and rule extraction from gene expression data of cancer tissue, Artificial Intelligence in Medicine. 41, pp. 161–175.
7. Statnikov A., Aliferis C.F., Tsamardinos I., Hardin D., Levy S. (2005) A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis, Bioinformatics. 21, pp. 631–643.
8. Golub T.R., Slonim D.K., Tamayo P., Huard C., Gaasenbeek M., et al. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science. 286, p. 531.
9. Khan J., Wei J.S., Ringnér M., Saal L.H., Ladanyi M., et al. (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, Nat. Med. 7, pp. 673–679. doi:10.1038/89044.
10. Bhattacharjee A. (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses Proceedings of the National Academy of Sciences. 98, pp. 13790–13795. doi:10.1073/pnas.191502998.
11. Shipp M.A., Ross K.N., Tamayo P., Weng A.P., Kutok J.L., et al. (2002) Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning, Nat. Med. 8, pp. 68–74. doi:10.1038/nm0102-68.
12. Wall M.E., Rechtsteiner A., Rocha L.M. (2003) Singular Value Decomposition and Principal Component Analysis, http://arxiv.org/abs/physics/0208101.
13. Van Der Heijden F., Duin R.P., De Ridder D., Tax D.M.J. (2004) Classification, parameter estimation and state estimation, Wiley Online Library, http://onlinelibrary.wiley.com /doi/ 10.1002/0470090154.fmatter/summary (accessed March 6, 2013).
14. Guyon I., Elisseeff A. (2003) An introduction to variable and feature selection, The Journal of Machine Learning Research. 3, pp. 1157–1182.
15. Mitchell T.M. Machine Learning, McGraw-Hill, New York, 1997.
16. Witten I.H., Frank E., Hall M.A. (2011) Data mining practical machine learning tools and techniques, third edition., Morgan Kaufmann Publishers, Burlington, Mass.
17. Weka 3 - Data Mining with Open Source Machine Learning Software in Java, (n.d.),. http://www.cs.waikato.ac.nz/ml/weka/index.html (accessed December 11, 2011).
18. Du K.-L., Swamy M.N.S. (2006) Neural networks in a softcomputing framework, Springer, London.
19. Delashmit W.H., Manry M.T. (2005) Recent developments in multilayer perceptron neural networks, in: Proceedings of the Seventh Annual Memphis Area Engineering and Science Conference, MAESC.
20. Vapnik V. Statistical learning theory., in: 1998.
21. Chow T.W.S., Cho S.-Y. (2007) Neural networks and computing : learning algorithms and applications, Imperial College Press, Distributed by World Scientific, London, Singapore, Hackensack, NJ.
22. Breiman L. Random forests, Machine Learning. 45 (2001), pp. 5–32.
23. Muller K.-R., Mika S., Ratsch G., Tsuda K., Scholkopf B. (2001) An introduction to kernel-based learning algorithms, IEEE Transactions on Neural Networks. 12, pp. 181–201. doi:10.1109/72.914517.
24. Salzberg S.L. (1997) On comparing classifiers: Pitfalls to avoid and a recommended approach, Data Mining and Knowledge Discovery. 1, pp. 317–328.
25. Han J., Kamber M. (2006) Data mining: concepts and techniques, Elsevier, Morgan Kaufmann, Amsterdam, Boston, San Francisco, CA.
26. Han B., Li L., Chen Y., Zhu L., Dai Q. (2011) A two step method to identify clinical outcome relevant genes with microarray data, Journal of Biomedical Informatics. 44, pp. 229–238. doi:10.1016/j.jbi.2010.11.007.
27. Rojas-Galeano S., Hsieh E., Agranoff D., Krishna S., Fernandez-Reyes D. (2008) Estimation of Relevant Variables on High-Dimensional Biological Patterns Using Iterated Weighted Kernel Functions, PLoS ONE. 3, p. e1806. doi:10.1371/journal.pone.0001806.
28. Xiong M., Jin L., Li W., Boerwinkle E., et al. (2000) Computational methods for gene expression-based tumor classification, Biotechniques. 29, pp. 1264–1271.
29. Rapaport F., Zinovyev A., Dutreix M., Barillot E., Vert J.P. (2007) Classification of microarray data using gene networks, BMC Bioinformatics. 8, p. 35.
30. Chuang H.-Y., Lee E., Liu Y.-T., Lee D., Ideker T. (2007) Network-based classification of breast cancer metastasis, Molecular Systems Biology. 3, doi:10.1038/msb4100180.