Effectiveness of Feature Selection and Classification Techniques for Gene Expression Data Analysis

Abdallah N. ElSheikh[a] Tamer M. Jarada[a] Mohamad Nagi[b] Ghada Naji[d] Panagiotis Karampelas[g]
Omer Şair Peter Peng[a] Keivan Kianmehr[e] Tansel Özyer[f] Mick Ridley[b] Jon Rokne[a] Reda Alhajj[a,c,g]

[a]Dept of Computer Science
University of Calgary
Calgary, Alberta, Canada

[b]School of Computing
University of Bradford
Bradford, UK

[c]Dept of Computer Science
Global University
Beirut, Lebanon

[d]Dept of Biology
Lebanese University
Tripoli, Lebanon

[e]Dept of Computer Engineering
University of Western Ontario
London, Ontario, Canada

[f]Dept of Computer Engineering
TOBB University
Ankara, Turkey

[g]Dept of Information Technology
Hellenic American University
Manchester, NH, USA

***Abstract*—Gene expression data is characterized by high dimensionality and small number of samples. Reducing the dimensionality is essential for effective analysis of the samples for efficient knowledge discovery. Actually, there is a tradeoff between feature selection and maintaining acceptable accuracy. The target is to find the reduction level or compact set of features which once used for knowledge discovery will lead to acceptable accuracy. Realizing the importance of dimensionality reduction for gene expression data, this paper presents novel framework which integrates dimensionality reduction with classification for gene expression data analysis. In other words, we present techniques for feature selection and demonstrate their effectiveness once coupled with data mining techniques for knowledge discovery. We concentrate on four feature selection techniques, namely chi-square, consistency subset, clustering-based and community-based. The effectiveness of the feature reduction techniques is demonstrated by coupling them with classification techniques, namely associative classification, support vector machines (SVM) and naive Bayesian classifier. The reported test results are encouraging; they demonstrate the applicability and effectiveness of the proposed framework.**

***Index Terms*—Feature selection, classification, associative classifier, naive Bayesian classifier, SVM classifier, gene expression data, chi-square, social networks, clustering, consistency subsets**

I. Introduction

The human body like all other organisms is composed of cells. A cell is the smallest building block of the body. Though tiny and invisible by eye, it is a huge factory that contains the deoxyribonucleic acid (DNA) which encodes genes that lead to proteins. In other words, DNA is the hereditary material in humans and almost all other organisms. Nearly every cell in the body has the same DNA. DNA consists of four different nucleotides, namely tyrosine (T), adenine (A), cytosine (C), and guanine (G). It exists in the form of a double helix where G pairs with C, and A pairs with T. Long chains of double-stranded DNA are called chromosomes. Within chromosomes is where an organism's genetic material is stored [16]. The genetic information that is stored in DNA allows molecules to combine to form functioning, living cells and organisms. This information is then responsible for defining traits and characteristics in living organisms, such as height, eye color, sex, etc. The sections of DNA that determine these traits are called genes and are located on the chromosomes. At its basic level, a gene codes for the synthesis of a protein via ribonucleic acid (RNA) molecule.

The expression level of genes is one of the important aspects that guide researchers in their effort to reveal various phenomena. Fortunately, the development in technology produced the microarray which has been successfully used to simultaneously study the expression levels of thousands of genes. Accordingly, a huge amount of gene expression data has been produced. Gene expression data analysis is an important research area that has attracted the attention of a large number of research groups around the world, e.g., [4], [5], [6], [8], [15], [26], [30], [31], [32], [34]. The target is to identify biomarkers for various diseases. A biomarker is a molecule (mainly gene or protein) that exists in the cell and does not function normally. Experimental and computational approaches have been applied in the process. However, pure wet-lab experiments are generally not feasible because of the high dimensionality which is characterized by the large number of genes. Hence, computational approaches [12], [11] are more effective as preprocessing step for experimental methods. In other words, computational approaches are used to reduce the number of genes that could be used to classify samples into two groups, namely infected and normal [28], [17], [18], [24].

In this paper, we present an integrated framework for gene expression data analysis. We describe four techniques for feature selection and then study how they positively affect classification techniques when applied to gene expression data analysis. The techniques employed for feature reduction are chi-square, consistency subset, clustering, and community discovery by employing the social network model.

Chi-square, denoted X^2, test is a very commonly used method for feature reduction. It is a preprocessor method where each feature is assessed independently of all other features. Each feature is evaluated based on its ability to predict the classification. In other words, chi-square test evaluates features individually by measuring their chi-squared statistic with respect to the classes, and hence it is criticized for ignoring the interaction between features which may better guide the reduction process. Consistency subset is accepted as a wrapper method. It is a forward feature selection method that investigates the different subsets of features. There are two approaches that this method can take, forward selection and backward elimination. Forward selection methods start with an empty set of features and continue to add new features until the addition of new features no longer helps the relevance of the subset. In other words, feature subset evaluation is done to look for combinations of features whose values divide the data into subsets containing a strong single class majority. The search is in favor of small feature subsets with high class consistency. Clustering based feature selection distributes the features into groups based on their expression levels in samples. Then representatives are selected from each cluster to constitute the actual reduced set of features. The number of the representative features per cluster depends on the size and homogeneity of the cluster. Finally, a two-mode social network is build between genes and samples based on the expression levels of genes in samples. Then the social network is folded to produce a one-mode social network of genes. The social network of genes is then analyzed to find the communities of genes. From each community, the most influential gene is selected as the representative of the community in the reduced set of features. As a result, we get four different reduced sets of features, one from each of the four methods employed in the framework.

We used the outcome from the feature reduction process to compare the four methods in order to decide on the most effective for microarray data analysis [29], [25]. For this purpose, we build three classifiers, namely associative classifier, naive Bayesian [20], and SVM. The test results on different datasets favor both the clustering and the social network model based approaches as effective and efficient methods for feature reduction.

The rest of the paper is organized as follows. Feature reduction by clustering and social network analysis are covered in Section II and Section III, respectively. The employed classification approaches are described in Section IV. The test results are reported in Section V. Section VI is conclusions and future work.

II. Feature Reduction by Clustering

Clustering is known as unsupervised learning technique very suitable to categorize features into groups such that the similarity within a group is maximized and the similarity across the groups is minimized. The basic input that all clustering algorithms require is the mean for computing the similarity measure, which is mostly a distance function. In addition to the distance function, some other parameters may be required by clustering algorithms which are classified into categories according to how they proceed to produce the final outcome.

In this work, we applied k-means clustering with validity analysis. K-means is one of the most commonly used clustering algorithms. It requires specifying the number of clusters as input; then a centroid is determined for each cluster and the objects are iteratively distributed into clusters until a stable solution is obtained. Because it is hard to find the number of clusters in advance, we decided on running k-means by ranging the number of clusters between 2 and 50. Then, we applied cluster validity indexes on the outcome to select the most appropriate solution, which is the solution favored by the majority of the validity indexes [21].

The solution returned as the most appropriate for the given data is further analyzed to decide on the number of features to represent each cluster. For compact and homogeneous clusters, the feature closest to the centroid is selected as representative; compactness and homogeneity are decided based on the average variance of each cluster compared to the overall average variance of the clusters. Clusters which get their own average variance above the overall average variance are considered homogeneous and will have only one representative per cluster. However, clusters which do not satisfy the aforementioned property of variance are decomposed further into subclusters such that each subcluster satisfies the average variance property. The decomposition into subclusters is recursively done until all the clusters satisfy the average variance property. As a result, there will be n clusters leading to n features as the reduced set of features. Each of the n features is selected as the closest to the centroid of the cluster in which it is located. Of course n varies depending on a number of factors, including the size of the original feature set and how diverse are the features, i.e., the homogeneity of the clustering result and how many levels of decomposition are needed for achieving homogeneous set of clusters.

III. Feature Reduction by Social Network Analysis

Social network is a powerful model which could be effectively used to tackle a number of problems, including feature reduction as demonstrated in this paper. A social network models a given problem by identifying two constructs, actors and links. Actors may all form one category leading to one-mode network. It is also possible to have actors from two (may be more) categories leading to two (or higher) mode network. In the setting of the problem tackled in this paper, there are two sets of actors, namely samples and genes. A link between two actors reflects a kind of relationship. While links connect actors within the same group in one mode networks, in two (or higher order)-mode social networks a link connects two actors if they are related and they belong to two different groups. Researchers have used bipartite graphs to tackle other problems in bioinformatics, e.g., [2], [3].

Any n-mode ($n \geq 2$) social network could be folded into m-mode social networks, where $1 \geq m < k$. For the model described in this paper, a two mode network is folded to produce a one-mode network of genes which are the actual features in the dimensionality reduction problem addressed in this paper.

A social network is represented as a graph which is manipulated by considering the corresponding adjacency matrix. It is a two dimensional matrix where (for the problem addressed in this paper) a row corresponds to a feature or gene and a column corresponds to a sample. An entry (i, j) reflects the expression level of gene i in sample j. In other words, a link between gene i and sample j indicates that gene i is expressed in sample j.

Because our target is to reduce the number of genes that could be used to classify samples, the two mode network is folded into one mode network where the actors are only genes and the link between two genes reflect the number of samples in which the two genes are co-expressed. The folding process is done by multiplying the adjacency matrix by its transpose to produce a new matrix where rows and columns are all genes. The one-mode social network of genes is processed further to find communities of genes. A community of genes includes a set of genes which are more connected to each other than to other genes outside the group. We eliminate edges (links) from the graph by considering the betweenness centrality. For each edge, its betweenness centrality is determined as the number of shortest paths that pass through the edge. The higher the betweenness centrality the more becomes the edge a candidate to be removed. Edges are removed from a social network based on two criteria, they should have high betweenness centrality and their removal should lead to more communities in the network. Our strategy for the network of genes is to satisfy either of the following two constraints, the one that could be achieved first: (1) to have the number of communities equal to the number of clusters produced by the method presented in Section II; (2) not to remove any edge whose betweenness centrality is below the average betweenness centrality of all the edges in the network. This will lead to reasonable number of communities where each community is somehow homogeneous. We select for each community a representative which has the highest average of two centrality measures, namely degree centrality and eigen-value centrality. Interestingly, the conducted experiments reported close to 90% overlap (at least in functionality) between the reduced features produced by the clustering and the social networks based methods.

IV. CLASSIFICATION

In general, pattern recognition methods can be divided into two categories: unsupervised and supervised. While clustering is unsupervised, classification is supervised. A supervised method is a technique that one uses to develop a predictor or classification rule using a learning set with known classes. The predictor is subsequently used to classify unknown objects. Given a set of objects (samples in our case), the target is to construct a classifier model by proceeding as follows. First, it is important to know the class for each of the samples in hand. This information will allow us to use part of the samples for building the classifier model and the rest for testing the model. The former part is called the training set and the latter part constitute the test set. The training set should well represent all the cases to be covered by the classifier in order to produce a classifier with good accuracy. The accuracy is determined as the percentage of the correctly classified instances from the test set. In other words, classification [12], [11] centers around exploring through data objects (training set) to find a set of rules which determine the class of each object according to its attributes. The discovered rules are later used to build a classifier to predict the class or missing attribute value of unseen objects whose class might not be known.

The classification problem can be formalized as follow: Given a matrix of n rows and m columns (individuals), which correspond to k classes. Individuals' annotation can be written as:

$$C_{11}, C_{12}, ..., C_{1r}, C_{21}, C_{22}, ..., C_{2d}, ..., C_{kf},$$

where r, d, and f are the number of individuals in classes 1, 2, and k, respectively, and $r + d + ... + f = m$. The goal is to construct a classifier model which can predict the class of a new individual which was not considered while building the model. Support vector machines (SVMs), associative classification, k nearest neighbors, Bayesian networks, decision trees, and neural networks are well established techniques for classification.

SVM represents a particular instance of a large class of learning algorithms known as kernel machines; it is a powerful supervised learning algorithm for classification. Two class SVM projects data into higher dimensional space where the two classes are linearly separable. It finds in the space of the data points a hyperplane that separates the two classes of the data, and maximizes the width of a separating band between the data points and the hyperplane. The support vectors are defined as the ones nearest to this margin; only the support vectors define the model and need to be stored.

The objective of SVM is to select the optimal hyperplane which can separate the two classes as there may exist many hyperplane that can separate the two sets of points. The optimal hyperplane is defined as the hyperplane which can separate the classes with largest margin. A hyperplane equation can be determined by two parameters: w and b, where w is a weight vector perpendicular to the hyperplane, and b is a bias that moves the hyperplane parallel to itself. The equation of a hyperplane can be written as:

$$\vec{w}^T.\vec{x_i} + b = 0 \tag{1}$$

Recall that a classification task usually involves training and testing data which consist of data instances, and each instance in the training set has several features and a target value (class label). The goal of SVM is to produce a model which can predict the target value of data instances in the testing set which just have the features. Given a training set of instance-label pairs $(x_i, y_i), i = 1, ..., l$, where $X_i \in R^n$ and $y \in \{1, -1\}^l$, the SVM requires solving the following optimization problem:

$$min_{w,b,\epsilon}\frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{l}\epsilon_i \qquad (2)$$

subject to

$$y_i(w \bullet x_i + b) \geq 1 - \epsilon_i, \epsilon_i > 0, i = 1, ..., l \qquad (3)$$

where x_i is an input vector, l is the number of instances in the training set, C is a cost factor for misclassification, and ϵ is a slack variable for misclassification points.

There are many fundamental advantages of SVM compared to other methods. First, SVM produces a unique solution because it is basically a linear method and does not have such a pitfall as multiple local minima. Second, SVM is inherently able to deal with very large amount of dissimilar information. Third, the discriminant function is characterized by only a comparatively small subset of the entire training data set, thus making the computations noticeable faster. Recent comparison study among all existing classification methods have shown the outperforming nature of SVM [23].

A Bayesian network [22] is a graphical probabilistic model that consists of a directed acyclic graph and a set of conditional probability tables. The nodes in the network represent features or variables and links encode the conditional independence between the variables. The probability distribution is unconditional for a node without any parents. If a node has one or more parents, the probability distribution is a conditional distribution, where the probability of each node value depends on the values of the parents. This requires the probability distribution for each node be defined by a probability table and by considering its parent nodes.

The learning process in a Bayesian network consists of two stages. First the network structure is built (structure learning) and then probability distribution estimations are calculated in form of probability tables. Structure learning often has high computational complexity as the number of possible structures is huge. To solve the computational complexity, heuristic and approximate learning algorithms have been proposed [7], [10]. There are many combinations of structure learning and search technique that can be used to create Bayesian networks.

Associative classification is a simple classifier model which derives a set of association rules where the consequent of each rule is the class variable. In other words, given a set of objects (reduced set of genes in our study), which are the relevant characteristics of the samples, and a class label per sample. The association rules mining process proceeds as follows to discover associative classification rules. First, each sample is considered as a transaction and genes together with the class label constitute the set of items. Then, association rules are derived by finding first frequent itemsets that contain the class variable as an item. From each frequent itemset only one rule is produced, which is the rule with the class variable as the consequent. Only rules which have high confidence and are interesting are maintain as classification rules.

V. Experiments

We have running a set of experiments to demonstrate the applicability and effectiveness of the proposed framework. For this purpose, we have used different existing software tools for the feature reduction and the classification tasks. We used three cancer datasets. The environment, the datasets and the results are described next in this section.

A. Testing Environment

ORA was used to realize the social network based feature selection method, and WEKA was used for the other three feature selection methods. For the consistency subset feature selection method, we used the greedy stepwise search algorithm with forward selection with default settings. The output from WEKA consists of a minimal subset of features which can best represent the complete original dataset. For the chi-square dimensionality reduction method, we used the ranker search method (also with the default settings). Here Weka does not return a reduced set of genes. It rather produces a ranked list showing the importance of each feature based on its chi-square value. Then, the reduced set of features is extracted from the ranked list. Explicitly, We decided on the final reduced set of features after analyzing the effect of using different sizes of the reduced feature set (starting with the top ranked feature as set by itself) and by considering the rank of each feature. Finally, the k-means clustering of WEKA was used to find clustering outcome when the number of clusters is in the range 2 to 50. Further, we applied in the process five of our implemented clustering validity indexes, namely Dunn, Davies-Bouldin, Silhouette, Jaccard and Rand. The majority voting was used to find the best number of clusters.

To classify our reduced feature sets two tools were used, namely MATLAB and WEKA. MATLAB was used to run the SVM classifier and for the naive Bayes classifier. On the other hand, WEKA was used to realize the associative classifier. For each of the classification algorithms the feature reduced training set was used to train the classifier. Then the accuracy was determined based on the test data.

B. Datasets

Three data sets have been used in the experiments conducted in this study. The essential information related to the data sets is summarized below:

1) **Leukemia (AML/ALL) [9]:** The ALL/AML data set resulted from affymetrix microarray with 6817 genes. The data has 73 ALL/AML samples, 38 (27AML/11 ALL) samples for training and 35 (23AML/ 12 ALL) for testing.
2) **Colon:** The Colon data set contains 62 samples collected from cancer patients. Among them 40 tumor biopsies and 22 normal biopsies are from healthy parts of the colons of the same patients. Two thousands out of around 6500 genes were selected based on the confidence in the measured expression levels. The Colon data set was downloaded from the University of Texas, Human Genetics Center http://www.sph.uth.tmc.edu/hgc/

default.aspx?id=2775. Samples were split as: 15 normal samples were used for training and 7 for testing; 23 tumor samples were used for training and the other 17 were used for testing.

3) **Prostate [27]:** The expression profile of 12,600 genes were derived from 136 samples: 102 training (52 prostate vs 50 normal) vs 34 testing (25 prostate vs 9 normal).

C. Gene Selection

On each of the three datasets, we applied the the four dimensionality reduction methods to extract genes that have the discrimination potential to be used as biomarkers. For the clustering-based method, genes closer to the centroid are selected from each data set. For the social network based method, genes which are the most influential within their groups are selected. Finally, the top genes extracted from each data set after employing the social network based method are shown in Table I, where genes are ranked from top to bottom based on their influence within their communities.

TABLE I
TOP 11 GENES EXTRACTED FROM THE DATA USING THE SOCIAL NETWORK BASED METHOD

Leukemia	Colon	Prostate
CFD	*IL8*	*PHF16*
CD33	*CSRP1*	*VGLL4*
CST3	*CKS2*	*GRSF1*
MYB	*DARS*	*ZNF148*
CSTA	*CSRP1*	*LAMP2*
CEBPD	*DES*	*HNRNPM*
ELA2	*FBL*	*CCR2*
CXCL8	*HNRNPA1*	*CALM2*
LEPR	*GUCA2B*	*MEGF9*
SPTAN1	*CLNS1A*	*CYTSA*
PPBP	*CXCL2*	*HTATIP2*

By analyzing the genes reported in Table I, we noticed that there are no common genes which can be used as "general" cancer biomarkers. This may indicate that different cancer types have totally different signatures.

D. Classification Accuracy

Classifying samples is a challenging computational task as the number of significant genes (sample features) is small. In this analysis, we used SVM, naive Bayesian network and associative classifier. The accuracy of classifiers is based on 10-fold cross validation, i.e., partitioning the data into 10 groups, train the model with 9 groups and test the model with the tenth group. The process is iteratively applied 10 times by considering one of the 10 groups as the test set in each run. The final result is computed as the average of the results reported by the ten individual runs.

The classifiers have been employed for the three data sets used in the experiments. The final accuracy results are reported in Table II, Table III and Table IV for SVM, naive Bayesian and the associative classifier, respectively. These classifiers were built based on the outcome from the social network based feature reduction method. The outcomes from the testing using the other feature reduction approaches have been left out for space limitation. They will be reported in a future publication that extends the existing work.

As we can notice from the results reported in Table II, Table III and Table IV, the accuracy is different for each data set. This of course depends on the selected features and their distribution. This is not surprising and demonstrates the fact that the effectiveness of the outcome from the feature reduction method depends on how the method fits the characteristics of the analyzed dataset. This is supported well by having the reduced feature sets returned by the two methods (clustering based and social network based) almost overlap if not in the genes, at least in the functionality of the genes reported in the reduced features set. The same justification and analysis is valid for the classifiers which have been constructed based on the reported reduced feature set for each of the tested datasets. We noticed that different classifiers reported better accuracy for different datasets. That is, each of the three classifiers used in the framework favors certain dataset(s).

TABLE II
SVM CLASSIFICATION ACCURACY FOR THE THREE CANCER DATASETS

	Leukemia	Colon	Prostate
Accuracy	*96.71%*	*77.05%*	*90.05%*
Cross Validation	*96.89%*	*95.28%*	*90.45%*

TABLE III
NAIVE BAYES CLASSIFICATION ACCURACY FOR THE THREE CANCER DATASETS

	Leukemia	Colon	Prostate
Accuracy	*94.95%*	*80.09%*	*90.75%*
Cross Validation	*95.76%*	*87.16%*	*92.10%*

TABLE IV
ASSOCIATIVE CLASSIFIER CLASSIFICATION ACCURACY FOR THE THREE CANCER DATASETS

	Leukemia	Colon	Prostate
Accuracy	*97.15%*	*72.80%*	*91.00%*
Cross Validation	*97.80%*	*81.30%*	*93.20%*

VI. CONCLUSIONS

Gene expression data analysis is one of the main research areas that have attracted the attention of a large number of research group who are all trying to identify biomarkers for different diseases. The main challenge has always been the high dimensionality of the data in the sense that there are large number of molecules to consider as biomarkers and there are small number of samples to utilize for building models that could differentiate between the infected and uninfected samples. Pure wet-lab based analysis is hard to achieve if at all feasible to the high cost and tremendous effort required to undertake the experiments. Fortunately computational techniques have provided attractive methods to help in the analysis. This brings up the beauty of computation whether discrete or applied, including statistical, machine learning and data mining techniques. However, applied computational techniques do suffer and do no produce satisfactory results when the dataset to be analyzed is characterized by high dimensionality. Here comes the role of dimensionality reduction approaches which are capable of finding a reduced set of feature that could be used to discriminate between infected and uninfected samples. Four dimensionality reduction features have been utilized in this study. The utilized features have different characteristics and capabilities. Some of them like chi-square evaluate the features individually, while the other three (consistency subset, clustering based and social network based) consider the features collectively and hence produce more robust results. The conducted testing demonstrates that features are related in some sense and considering them in isolation could lead to information loss and hence negatively affects the final outcome. Our study demonstrated further that both clustering based and social network based approaches work fine and produce good discriminative set of features.

Not having common genes between the set of genes reported as discriminating features for classification is another interesting phenomenon to comment on. While the reported genes have close functionalities when they are different, the reason for having different genes is the fact that two of the feature reduction methods (clustering and social network) selected representative features as closest to centroid or most influential in the community, respectively. Checking the other genes close to the centroids or next influential genes in communities might lead to almost total overlap in the reported set of genes. This has been left as future work. It will be investigated by building a fuzzy model where the degree of membership in a community is determined by the degree of influence and the degree of membership in a cluster is determined based on the distance to the centroid. As future work, we will also investigate further the characteristics of each dataset to decide on general guidelines regarding the most appropriate feature reduction and classifier model for each dataset. For instance, SVM reported consistently high accuracy for all the three datasets utilized in the testing. Naive Bayesian classifier also reported high accuracy for all the tested data. But for the colon data, the accuracy reported by naive Bayesian classifier is better, and the associative classifier reported best accuracy for the leukemia data. Having all the three classifiers reported high acceptable accuracy for all the datasets is encouraging to use any of the three classifiers in practice. On the other hand, attaining the highest accuracy is always favored especially in sensitive cases like the gene expression data analyzed in this study.

References

[1] M. Alshalalfa and R. Alhajj, “Motif Location Prediction by Divide and Conquer,” *Proceedings of the International Conference on Bioinformatics Research and Development,* pp.102-113, July 2008.

[2] J. Ballesteros and K.G. Palczewski, “protein-coupled receptor drug discovery: implications from the crystal structure of rhodopsin,” *Current Opinion in Drug Discovery and Development,* 4(5):561-74, 2001.

[3] K. Bleakley and Y. Yamanishi,“Supervised prediction of drug-target interactions using bipartite local models,” *Bioinformatics,* 25(18):2397, 2009.

[4] S. Bicciato, M Pandin., G. Didon and C. Di Bello, “Pattern identification and classification in gene expression data using an autoassociative neural network model.” *Biotechnology and Bioengineering,* 81:594-606, 2002.

[5] R. Bijlani, et al., “Prediction of biologically significant components from microarray data: independently consistent expression discriminator(iced),” *Bioinformatics,* 19:62-70, 2003.

[6] F. Chu and L. Wang, “Cancer classification with microarray data using support vector machines.” *Bioinformatics*, 176:167-189, 2005.

[7] G. Cooper and E. Herskovits, “A Bayesian method for the induction of probabilistic networks from data”. *Machine Learning,* 9, 309-347, 1992.

[8] T. S. Furey, et al., “Support vector machine classification and validation of cancer tissue samples using microarray expression data,” *Bioinformatics,* 16:906-914, 2000.

[9] T. R. Golub, et al., “Molecular Classification of Cancer: class discovery and class prediction by gene expression monitoring,” *Science*, 286, 531-537, 1999.

[10] D. Heckerman, D. Geiger and D. M. Chickering, “Learning Bayesian networks: The combi- nation of knowledge and statistical data”. *Machine Learning,* 20, 197-243, 1995.

[11] R. J. Henery, *Machine Learning, Neural and Statisical Classification.* http://www.maths.leeds.ac.uk/~charles/statlog/whole.pdf. 1994.

[12] M. Kantardzic, *Data Mining - Comcepts, Models, Methods, and Algorithms.* Picastaway: Wiley-Interscience, 2003.

[13] K. J. Kechris, E. van Zwet, P. J. Bickel, and M. B. Eisen. “A Boosting Approach for Motif Modeling using ChIP-chip Data,” *Bioinformatics,* 21(11):2636-2643, 2005.

[14] M.J. Keiser, et al., “Predicting new molecular targets for known drugs,” *Nature,* 462(7270):175-181, 2009.

[15] O. Kent and J. Mendell, “A small piece in the cancer puzzle:microRNAs as tumor suppressors and oncogenes,” *Oncogene,* 25:6188-6196, 2006.

[16] D. E. Krane and M. L. Raymer. *Fundamental Concepts of Bioinformatics,* San Francisco: Pearson Education, 2003.

[17] B. Krishnapuram, L. Carin and A. Hartemink, Gene expression analysis: Joint feature selection and classifier design. 2000.

[18] J. Li and L. Wong, “Identifying good diagnosis gene group from gene expression profile using the concept of emerging patterns,” *Bioinformatics*, 18:725-734, 2002.

[19] E. Moler, M. Chow, and I. Mian, “Analysis of molecular profile data using generative and discriminative methods,” *Physiol genomics,* 4:109-126, 2000.

[20] A. Moore, Lecture on Bayesian Networks. http://www.autonlab.org/tutorials/bayesnet09.pdf. 2001.

[21] Özyer T. and Alhajj R., “Achieving natural clustering by validating results of iterative evolutionary clustering approach,” *Proceedings of IEEE International Conference on Intelligent Systems,* pp.488-493, 2006.

[22] J. Pearl, “Bayesian Networks: A Model of Self-Activated Memory for Evidential Reasoning”. *Proceedings of the Conference of the Cognitive Science Society,* University of California, Irvine, CA, pp. 329-334, August 15-17, 1985.

[23] M. Pirooznia, J. Yang, M. Yang and Y. Deng, “A comparative study of different machine learning methods on microarray gene expression data,” *BMC Genomics*, 2007.

[24] A. Qabaja, M. Alshalalfa, R. Alhajj and J. Rokne, “Multiagent Approach for Identifying Cancer Biomarkers,” *Proceeding of IEEE International Conference on Bioinformatics and Biomedicine,* Nov 2009.

[25] J. Quackenbush, “Microarray data normalization and transformation.” *Nature Genetics Supplement,* December 2002: 496-501.

[26] S. R. Setlur, et al., “Integrative Microarray Analysis of Pathways Dysregulated in Metastatic Prostate Cancer,” *Cancer Research,* 67:10296, 2007.

[27] D. Singh, P. G. Febbo, et al., “Gene expression correlates of clinical prostate cancer behavior,” *Cancer Cell*, 1, 203209, 2002.

[28] R. Soumya , P. D. Sutphin, J. T. Chang and R. B. Altman. “Basic Microarry Analysis: Grouping and Feature Reduction.” *TRENDS in Biotechnology,* pp.189-193, 2001.

[29] D. Stekel, *Microarray Bioinformatics.* New York: Cambridge University Press, 2003.

[30] R. Varshavsky, et al., “Novel unsupervised feature filtering of biological data,” *Bioinformatics*, 22:507-513, 2006.

[31] J. Wang, et al., “Tumor classification and marker gene prediction by feature selection and fuzzy c-means clustering using microarray data,” *BMC Bioinformatics*, 4:60-70, 2003.

[32] L. Wang, F. Chu, and W. Xie, Accurate cancer classification using expressions of very few genes.” *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* 4:40-53, 2007.

[33] Z. Xia, L.-Y. Wu, X. Zhou and S.T.C. Wong, ”Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces,” BMC Systems Biology, 4(Suppl 2):S6, 2010.

[34] X. Zhang and H. Ke, “ALL/AML cancer classification by gene expression data using SVM and CSVM,” *Genomics informatics,* 11:237-239, 2000.