Evaluating SVM Algorithms for Bioinformatics Gene Expression Analysis

Heena Farooq Bhat
Department of Computer Science, University of Kashmir, Srinagar, India
heenafarooq14@gmail.com

***Abstract*— Support Vector Machines SVMs are trendy and dominant in learning systems because of providing good generalization properties, attending high dimensional data, their ability to classify input patterns with minimized structural classification risk and finding the optimal separating hyper-plane between two classes in feature space. Recent work in bioinformatics has seen an increasing use of SVM algorithms due to their benefits in dealing with high dimensional data, small sample size and compound data structures. The main aim of this paper is to provide a review of the most widely used SVM algorithms in bioinformatics namely gene expression based on the objectives using DNA microarrays classified into into three groups namely gene finding, class discovery and class prediction. These algorithms are then applied on cancer datasets: Leukemia and Lymphoma to produce better accuracy results.**

***Keywords*-** Support Vector Machines, Bioinformatics, Gene Expression, Class Discovery, Class Prediction, Recursive Feature Elimination.

I. INTRODUCTION

Support Vector Machines (SVMs) are supervised learning methods used for classification and regression. It involves analyzing a given set of labeled data so as to predict the labels of unlabelled future data. The purpose of SVM is to separate the data points by computing a hyperplane or a decision function [1]. The criterion used by SVMs is based on margin maximization between the two data classes. The margin is the distance between the hyper planes bounding each class. The separating hyperplane has to be determined in such a way that the margin between positive class and a negative class is maximized to produce good generalization ability.

The Support Vector Machine (SVM) [2] is a supervised learning algorithm which is useful for recognizing restrained patterns in composite datasets. This algorithm has been applied in various domains which includes text categorization, image-recognition, hand-written digit recognition [3]. Although, SVM's are based on statistical learning theory and have the aim of determining the location of decision boundaries that produce the optimal separation of classes, with the two-class pattern recognition problem, in which the classes are linearly separable, the SVM selects from among the infinite number of linear decision boundaries, the one that minimizes the generalization error. Thus, the selected decision boundary will be one that leaves the greatest margin between the two classes, where margin is defined as the sum of the distances to the hyper-plane from the closest points of the two classes [4]. SVMs have been shown to provide a better generalization performance than traditional techniques such as neural networks [5].

The paper is organized as follows: Section I (current section) provides the introduction of SVM classification. Section II gives a brief idea of bioinformatics. Section III reviews and classifies the SVM algorithms in bioinformatics that has been used for gene expression. The experimental results of various gene expression algorithms presented in this paper are evaluated in section IV. Section V concludes the paper.

The working procedure of SVM classification consists of two stages i.e. training phase and testing phase. Training phase involves the training set as input, minimize the decision function and obtain the prediction model as output. Testing phase involves prediction model and validation set as input, apply prediction model to the validation set and obtain an accuracy of prediction. Given N elements in input, and two disjointed classes in output, the basic SVM takes the input elements, elaborates them (learns by them, actually), and, for each of them, predicts if it belongs to the first class or the second one. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other. The Figure 1 below shows the flowchart of SVM classification algorithm.

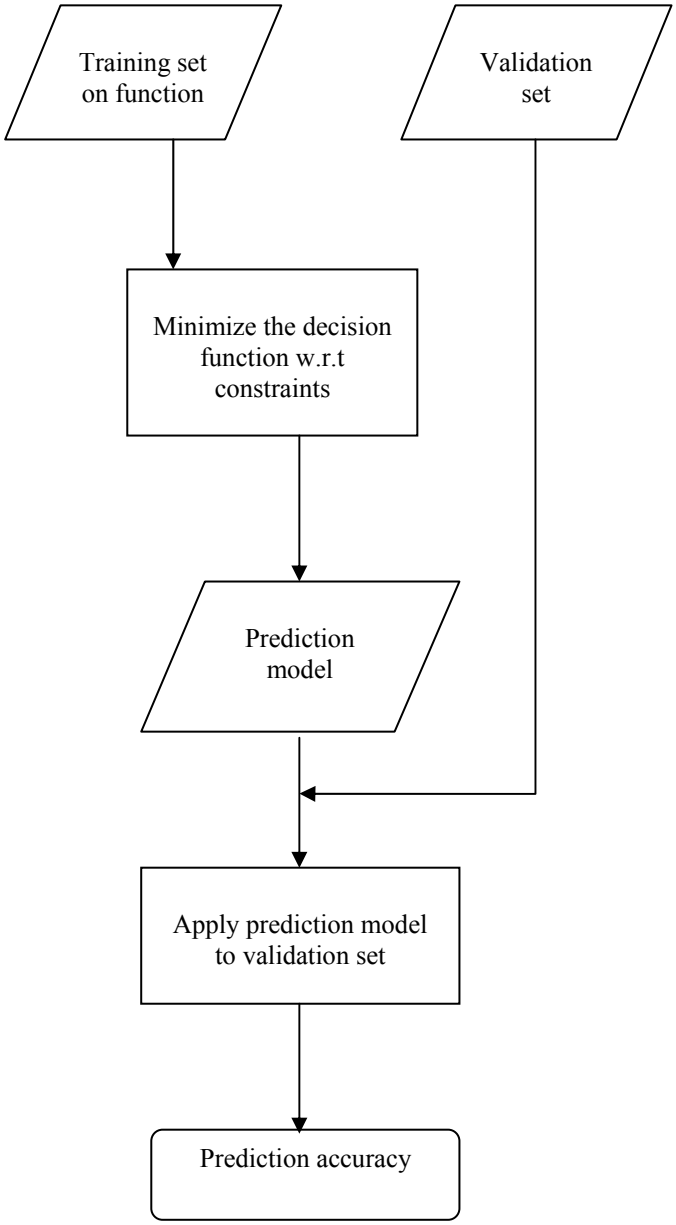

Figure 1. Flowchart of SVM classification algorithm

The support vector machine is originally introduced for binary SVM classification problems and has an excellent ability to solve these problems. For binary classification problems, the main aim of SVMs is to separate the data in some optimal methods. One of the key processing steps in the development of SVM algorithms is to employ an optimizer to solve the quadratic programming problem. Typically, the conventional SVMs have used an optimizer based on quadratic programming (QP) or linear programming (LP) methods to solve the optimization problem [6]. Due to its immense size, the QP problem that arises from SVMs cannot be easily solved via standard QP techniques.

A number of papers discuss binary SVM for various applications. Binary SVMs have been extended to solve multi-class classification tasks. Various methods have been proposed to construct a multi-class classifier by combining binary classifiers. SVMs algorithms have been applied to many domains successfully, it includes the bioinformatics area.

II. BIOINFORMATICS

Bioinformatics can be defined as the application of computer technology to the management of biological information. Bioinformatics is the science of storing, extracting, organizing, analyzing, interpreting and utilizing information from biological sequences and molecules. It has been mainly fueled by advances in DNA sequencing and mapping techniques. Over the past few decades rapid developments in genomic and other molecular research technologies and developments in information technologies have combined to produce a tremendous amount of information related to molecular biology. The primary goal of bioinformatics is to increase the understanding of biological processes. Applications of SVMs to bioinformatics include gene finding, protein function domain detection, function motif detection, protein function inference, disease diagnosis, protein and gene interaction network reconstruction, data cleansing, and protein sub-cellular location prediction. For example, microarray technologies are used to predict a patient's outcome. On the basis of patients' genotypic microarray data, their survival time and risk of tumor metastasis or recurrence can be estimated. Machine learning can be used for peptide identification through mass spectroscopy. Correlation among fragment ions in a tandem mass spectrum is crucial in reducing stochastic mismatches for peptide identification by database searching. An efficient scoring algorithm that considers the correlative information in a tunable and comprehensive manner is highly desirable. This paper reviews SVM algorithms in Bioinformatics area of gene expression.

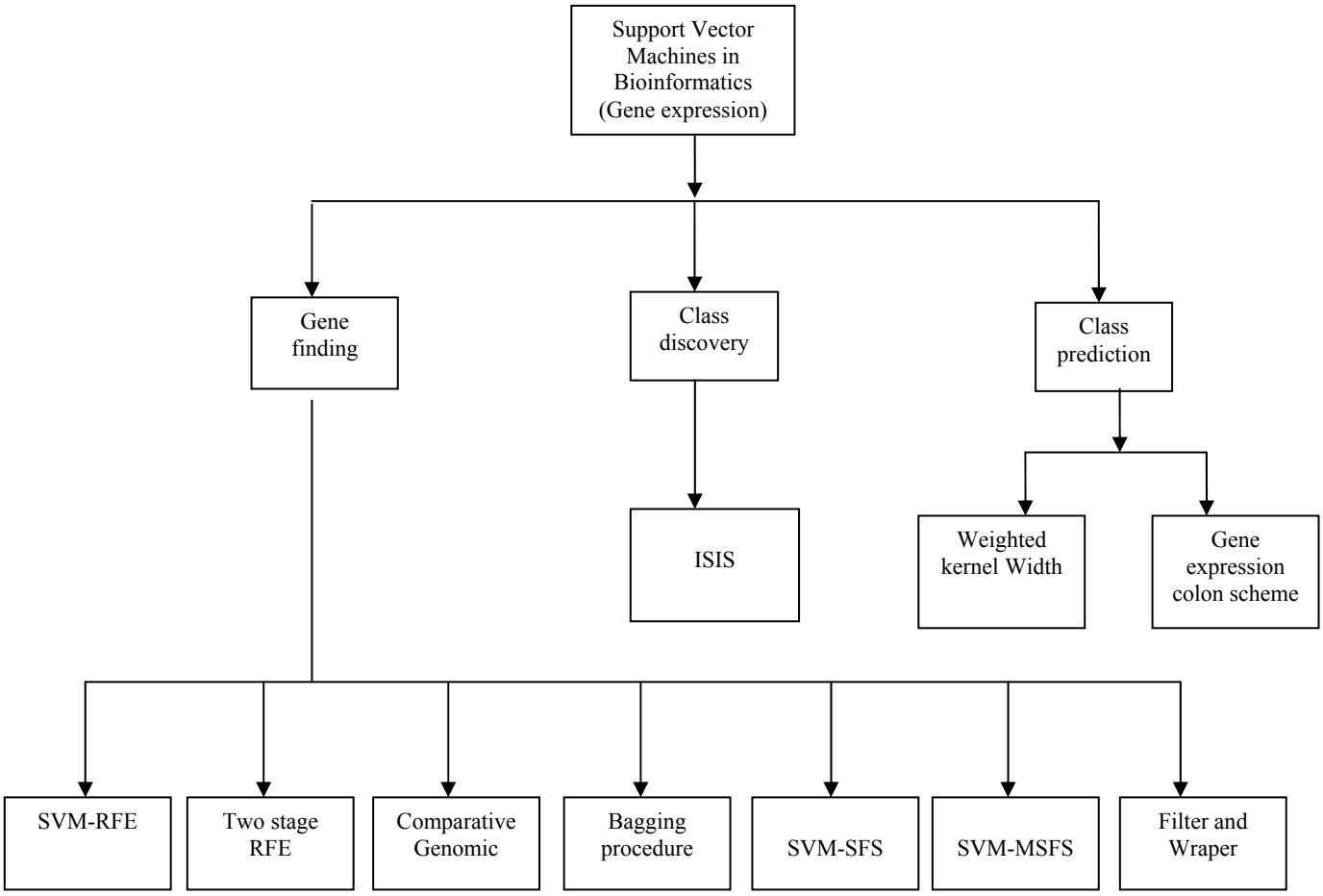

Figure 2. Classification of Support Vector Machines in Bioinformatics

III. SUPPORT VECTOR MACHINES IN BIOINFORMATICS

In the recent years, SVMs have been applied in many bioinformatics domains [7] including recognition of translation start sites [8], protein remote homology detection [9][10][11], protein fold recognition [12], microarray gene expression analysis [13][14][15][16][17], functional classification of promoter regions [18], prediction of protein–protein interactions [19] and peptide identification from mass spectrometry data [20]. SVM and its variants have been successfully applied in many domains, for example in two-class classification of microarray data [14] [22]. Bioinformatics data sets usually contain measurements for thousands of genes, which prove problematic for many traditional methods, while SVM are well suited to obtain classification models with such high dimensional data. In this paper, the main focus will remain on the gene expression analysis for cancer classification for classifying the unknown tissue samples.

Support vector machine (SVM) is one of the state-of-art kernel based machine learning techniques and has been widely used for the classification of microarray gene expression data [24]. The gene expression data has proved to be very useful for tissue classification and prediction [21]. The high dimensionality and relatively few examples characterized in gene expression data also suggest that Support Vector Machines (SVMs), a novel machine learning technique, is more dominant in classifying the tissues and in removing the non-informative genes than other methods such as Fisher linear discriminant and decision trees [13]. Since, the microarray data suffers from the curse of dimensionality, the small number of samples, and the level of irrelevant and noise genes [23]. These make the classification task of a test sample a very challenging problem. As a result, it is important to eliminate those irrelevant genes and identify the informative genes. Hence it concludes that the objectives using DNA microarrays can be classified into three major groups [25]: i) Gene finding, ii) Class discovery, iii) Class prediction as shown in Figure 2 above.

A. Gene Finding Algorithms

The main approach here is the feature selection. Its objective is to reduce the dimensionality of microarray dataset by selecting the most informative genes. This review mainly focuses on the gene expression data. But SVMs have been applied to many other biological data such as protein and DNA sequence. Various authors investigated a number of techniques for gene finding. Some of them are reviewed as below.

I) SVM-Recursive Feature Elimination (RFE)

This is used for removing redundant genes. [14] states that in tissue classification, feature selection methods matter more than the classification methods. RFE is basically a weight based saliency analysis on the basis of which weights connected to important features attain large absolute values and weights connected to unimportant features attain small absolute values. This technique detects weights with small values by evaluating the magnitude of weights and then removes the features emanating these small weights.

It assumes that a smaller "filter-out" factor in the SVM-RFE, which results in a smaller number of gene features eliminated in each recursion, should lead to extraction of a better gene subset [26]. As this is highly sensitive to the "filter-out" factor, simulations have shown that this assumption is not always correct and that the SVM-RFE is an unstable algorithm. To select a set of key gene features for reliable prediction of cancer types or subtypes and other applications, a new two-stage SVM-RFE algorithm has been developed [26].

Duan et al [27] proposed a feature selection method for multiclass classification. The proposed method selects features in backward elimination and computes feature ranking scores at each step from analysis of weight vectors of multiple two-class linear Support Vector Machine classifiers from one-versus-one decomposition of a multi-class classification problem.

II) Two Stage SVM-RFE

It involves the processing of sample classification to be carried out in two stages as:

Stage one: At this stage, SVM-RFE is designed to effectively eliminate most of the irrelevant, redundant and noisy genes while keeping information loss small.

Stage two: After the elimination of redundant genes, a fine selection for the final gene subset is then performed at the second stage.

The two-stage SVM-RFE overcomes the instability problem of the SVM-RFE to achieve better algorithm utility.

A three-stage of gene selection algorithm for microarray data was proposed by [28]. The proposed approach combines information gain (IG), Significance Analysis for Microarrays (SAM), mRMR (Minimum Redundancy Maximum Relevance) and Support Vector Machine Recursive Feature Elimination (SVM-RFE). In the first stage, intersection part of feature sets is identified by applying the (SAM–IG). While, the second minimizes the redundancy with the help of mRMR method, which facilitates the selection of effectual gene subset from intersection part that recommended from the first stage. In the third stage, (SVM-RFE) is applied to choose the most discriminating genes.

Similarly, Furey et al [15] analyses the classification of the tissue samples and an investigation of the data for mis-labeled tissue results using support vector machines. It also uses the neighborhood analysis to select important genes. In the mean time, [16][29][30] proposed the feature scaling methods for gene selection. Campbell et al [31] also proposed the automatic relevance determination for gene selection.

III) SVM-Comparative Genomic Hybridization (CGH)

Another method for feature selection was developed by [32] and is called Comparative Genomic Hybridization (CGH) which is one of the important mapping techniques for cancerous cells. For SVM-based classification, kernel used is substantially better than the standard kernel for SVM. Our approach of greedily selecting features with the maximum influence on an objective function results in significantly better classification and feature selection.

After the successful method of CGH for data classification, another technique has been applied to gene expression data. In this approach, SVM begins with a set of genes (first set) that have a common function: for example, genes coding for ribosomal proteins or genes coding for components of the proteasome. In addition, a separate set of genes (second set) that are known not to be members of the functional class is specified. These two sets of genes are combined to form a set of training examples in which the genes are labeled positively if they are in the functional class and are labeled negatively if they are known not to be in the functional class [34]. SVM is generally considered as the best "off-the-shelf" classifier.

IV) SVM-Bagging Procedure

One of the simplest ways to use SVM in the group structure is to apply bagging procedure with the base classifier of SVM given by [33]. The following work is done by various authors by applying the bagging procedure:

Guan et al [33] applied the bagging procedure for constructing a collection of SVMs for gene function prediction. The ensemble of SVMs consistently outperformed the single SVM classifier.

An extension of above work has been done later by [34] based on the feature perturbation for microarray classification. The study examined an SVM trained using a set of selected genes by Fisher criterion, an SVM trained using the feature set obtained by Neighborhood Preserving Embedding (NPE), a set of SVMs trained using a set of orthogonal wavelet coefficients of different wavelet mothers and a set of SVMs trained using texture descriptors extracted from the microarray, considering it as an image. The positive results obtained provide further affirmation that ensembles of classifiers obtain more reliable results.

As stated above [25], the feature selection algorithms aims to reduce the dimensionality of dataset by choosing useful genes but many of these algorithms produced fault for their ranked gene performance. Hence, [36] proposed a method by producing a feature selection algorithm in gene expression data analysis of sample classifications to produce the better accuracy results. This method selects the gene and divides the genes into subset, from the features, gene ranks are selected.

Mao et al [37] proposed two different constructed multiclass classifiers with gene selection which are fuzzy support vector machine (FSVM) with gene selection and binary decision classification tree based on SVM with gene selection.

V) SVM-Successive Feature Selection (SFS)

Revathi et al [36] proposed Successive Feature Selection SFS procedure (SFS) a set of features which is processed one at a time that the value of x is taken due to memory constraints and it is experimentally found that the suitable values of x is equal to or lower than 10. The output is the rank of features. In the successive level that the feature is dropped once at a time and a subset of features is obtained. That the classification accuracy using classifiers evaluated, and the best subset of features is processed to the next level. This process is terminated when all the features are ranked. Two ranked sets are obtained in SFS: namely $R_1 = \{x_2, x_4, x_1, x_2\}$ and $R_2 = \{x_2, x_1, x_4, x_2\}$.

VI) SVM-Modified Successive Feature Selection (MSFS)

In the SFS two ranked SFS are obtained, which indicate that x_2 is the top-ranked feature and that x_3 is the bottom ranked or least important feature. To select the three top-ranked features, the result will be $F_1 = \{x_2, x_4, x_1\}$ and $F_2 = \{x_2, x_1, x_4\}$. If the order of features is not important, then instead of two sets, F1 and F2, selected a common top 3 ranked features from the set $F_k = F_1 \cup F_2 = \{x_1, x_2, x_4\}$ [36]. Then the Gene ranking are find out by Mean and Standard Deviation. That the Mean of the common top 3 ranked features to the Standard deviation for the common top 3 ranked features. Then the Gene ranking are find out by the maximum value of this.

VII) SVM- Filter Method (FM) and Wrapper Method (WM)

The development of feature selection has two major directions i.e. filters and wrappers. Some gene selection methods do not assume any specific distribution model on the gene expression data and they are referred to as model-free gene selection methods or usually called Filter method (FM). The filters work fast and are efficient in selecting features but do not always produce satisfactory results when performing on wide feature sets. While other gene selection methods assuming certain models are referred to as model-based gene selection methods or may called Wrapper method (WM) [38].

The filter work was done by Deisy et al [39]. They used the analysis of symmetrical uncertainty with information gain. By calculating the difference between the entropy of the whole class and the features, features with less information can easily be identified.

Backstrom et al [40] presented an internal wrapper feature selection method for cascade correlation. The internal wrapper feature selection method selects features while hidden units are being added to the growing cascade correlation network architecture.

In other hand, some researcher applied Filter method and Wrapper method, called Hybrid Method. Moreover, hybrid gene selection methods search for an optimal subset of features is built into the classifier construction, and can be seen as a search in the combined space of feature subsets and hypotheses [41].

Another approach came into existence which is the extension of previously mentioned work. Here, a novel approach to combine feature (gene) selection and transductive support vector machine (TSVM) was proposed by [42]. It demonstrates that the potential gene markers could be identified and the TSVMs improved the prediction accuracy as compared to the standard inductive SVMs (ISVMs). The selected genes of the microarray data were then exploited to design the TSVM.

Saberkari et al [43] uses selective independent component analysis (SICA) for decreasing the dimension of microarray data. Using this selective algorithm, the instability problem occurred in the case of employing conventional independent component analysis (ICA) methods.

Chakraborty et al [44] proposed an effective classification technique that uses Naïve Bayes classifier, k-NN and SVM. The dimensionality reduction of the gene expression dataset is performed by using statistical approaches. From the dimensionality reduced data, the important genes are identified and also features are extracted. The well-trained classifier is used for the classification of micro array gene expression dataset.

B. Class Discovery Algorithms

This is the second objective where the approach is clustering. Its main aim is to determine new disease. Class discovery differs from the gene finding in a way that it does not involve any predefined classes. It involves grouping together specimens that are based on the similarity of their expression profiles with regard to the genes represented on the array [45]. Cluster analysis or clustering is often used for class discovery.

The objective of clustering expression profiles of tumors is to determine new disease (cancer) classifications. Although cancer classification has improved over the past 30 years, there has been no general approach for identifying new cancer classes (class discovery) or for assigning tumors to known classes (class prediction) [46]. Class discovery methods could also be used to search for fundamental mechanisms that cut across distinct types of cancers. For example, one might combine different cancers (for example, breast tumors and prostate tumors) into a single data set, eliminate those genes that correlate strongly with tissue type, and then cluster the samples based on the remaining genes. Discovery of a new class is usually achieved by an unsupervised machine learning method with the help of a clustering technique such as hierarchical clustering, k-means clustering and self organizing maps [47] [48]. The various methods of class discovery are as:

I) SVM-ISIS Method

This class discovery problem is approached by a method called ISIS (for "identifying splits with clear separation"). A method was proposed by [49] where the objective is to discover biologically relevant structures in the gene expression profiles of different tissue samples in an unsupervised fashion. This method searches for binary partitions in the set of samples that show clear separation. Mathematically, each class distinction is characterized according to the size of margin achieved by a support vector machine (SVM) separating the two classes.

It consists of two steps: First, based on the classification method Diagonal Linear Discriminant Analysis a score function called DLD is proposed to quantify the degree of separability of a given binary class distinction of the set of samples. This score function is defined on the graph of all bipartitions of the set of samples. Secondly, all the bipartitions are declared representing the local maxima in the graph that is for which the score does not increase if the class label of a single sample is changed. Since, this search over all bipartitions is not feasible. Hence, a fast heuristic to find candidate partitions as starting points for a search of local maxima have been proposed. This result in a variation of the original ISIS algorithm called SVM-ISIS. It detects the known tumor subtypes in an unsupervised fashion.

Ibrahim et al [50] adapted two semi-supervised machine learning approaches, namely self-learning (where miRNA and gene based classifiers are enhanced independently) and co-training (where both miRNA and gene expression profiles are used simultaneously)to enhance the quality of cancer sample classification.

Chakraborty et al [51] presents a combination of kernelized fuzzy rough set (KFRS) and semi-supervised support vector machine (S^3VM) for predicting cancer biomarkers from one miRNA and three gene expression data sets.

C. Class Prediction Algorithms

Here the approach is classification. Its objective is to classify the unknown samples whether they are cancerous or normal. It is basically categorized into two methods namely statistical and supervised machine learning methods [48]. In supervised method we need to train the classifier before we start in classifying process. Supervised methods are usually more effective in cancer classification researches. They are used for cancer prediction in a way that a classifier is trained with a part of the samples in the cancer microarray dataset. Then, the trained classifier is used to predict the samples in the rest of the dataset to evaluate the effectiveness of the classifier [52]. It is indicated that supervised methods are better than unsupervised methods.

The support vector machines are used in the quality control of DNA sequencing data. As a result, the classification of quality of the entire DNA sequencing data will automatically be made as high or low quality [53]. A new method is devised to fulfill the quality screening of DNA chromatograms which is a composition of feature extraction and support vector machines.

Multiclass classification tasks are very common in biological world. However, SVMs are originally developed for binary classification and hindered to solve multi-classification directly. Usual approach has been the resolution of the multi-class problem into a series of binary ones [54]. Most of the studies were confined towards binary gene selection problems and only a very few considered multi class gene selection and classification. This is so because multiclass gene selection and classification is significantly harder than the binary problems.

Since, cancer diagnosis is successfully implemented by the classification method, SVM and is one of the most important emerging clinical applications of gene expression microarray technology. A computer system for powerful and reliable cancer diagnostic model creation based on microarray data needs to be developed. To keep a realistic perspective on clinical applications the main focus might be on multiclass diagnosis. To equip the system with the optimum combination of classifier, gene selection and cross-validation methods, we performed a systematic and comprehensive evaluation of several major algorithms for multiclass classification, several gene selection methods, multiple ensemble classifier methods and two cross-validation designs [56].

The Support Vector Machine (SVM) is one of the classification methods successfully applied to the cancer diagnosis problems. The Multiclass SVM is a recently proposed extension of the binary SVM and is applied to multiclass cancer diagnosis problems. Comparable classification accuracy and its flexibility render the Multiclass SVM a viable alternative to other classification methods [57]. The various methods are:

I) SVM-Weighted Kernel Width Method

Apart from cancer classification, Yuvaraj et al [58] devised a new method for tumor classification using gene expression data. In the proposed method, genes are selected first using Nonnegative Matrix Factorization (NMF). In order to improve the performance of classification, Symmetry NMF (SymNMF) was used in this approach. Then, features are extracted from the selected genes by virtue SymNMF. At last, an efficient machine learning approach was used to classify the tumor samples using the extracted features. For a better classification, Support Vector Machine with Weighted Kernel Width (WSVM) was used in this classification approach.

A general framework is proposed by [59] for prediction of predefined tumor classes using gene expression profiles from microarray experiments. The framework consists of 1) evaluating the appropriateness of class prediction for the given data set, 2) selecting the prediction method, 3) performing cross-validated class prediction, and 4) assessing the significance of prediction results by permutation testing. This framework is designed to reduce the occurrence of spurious findings, a legitimate concern for high-dimensional microarray data.

Multiclass support vector machines (MC-SVMs) are the most effective classifiers in performing accurate cancer diagnosis from gene expression data. The one-Against-All method of MC-SVM techniques by [60] [61] were found to be the best methods in this domain. Gene selection techniques can significantly improve the classification performance of both MC-SVMs [56].

Support Vector Machines-One against All (SVM- OAA) SVMs are the most modern method applied to classify gene expression data, which works by separating space into two regions by a straight line or hyper plane in higher dimensions [63]. SVMs were formulated for binary classification (2 classes) but cannot naturally extend to more than two classes. SVMs are able to find the optimal hyper plane that minimizes the boundaries between patterns. SVMs are power tools used widely to classify gene expression data [62].

To develop multiclass classification models with optimal parameters and features, [64] performed a systematic evaluation of machine learning algorithm and four feature selection methods using three-fold cross validation and a grid search. We obtained an accuracy of 100%, relative classifier information (RCI) of 1.0, and a kappa index of 1.0 by applying the model of support vector machine namely one-versus-rest SVM (OVR) which selected only four features to categorize the 12 groups, resulting in a time-saving and cost-effective strategy for diagnosing neuromuscular diseases.

One key element in understanding the molecular machinery of the cell is to understand the meaning, or function, of each protein encoded in the genome. One of the most powerful such homology detection methods is the SVM-Fisher method [9]. The pair-wise SVM method uses a pair-wise sequence similarity algorithm such as Smith-Waterman in place of the Hidden Markov Method (HMM) in the SVM-Fisher method.

II) SVM-Gene Expression Based Colon Classification (GECC)

Rathore et al [65] proposed a novel gene expressions based colon classification scheme (GECC) that exploits the variations in gene expressions for classifying colon gene samples into normal and malignant classes. A majority voting based ensemble of support vector machine (SVM) has been proposed to classify the given gene based samples. Previously, individual SVM models have been used for colon classification. However, their performance is limited. In this research study, we propose an SVM-ensemble based new approach for gene based classification of colon, wherein the individual SVM models are constructed through the learning of different SVM kernels, like, linear, polynomial, radial basis function (RBF), and sigmoid. The predicted results of individual models are combined through majority voting. In this way, the combined decision space becomes more discriminative.

However, [54] proposed a new voting approach to the assignment of class label for a test observation after pair-wise training of SVM classifiers. The approach investigates the correlations between "scores" - the real valued vector produced for observations by a set of binary SVM classifiers. These score vectors are then combined using a majority voting mechanism to assign the class membership for the test observations. The performance of the algorithm is evaluated on various gene expression profiles, and two typical multi-class SVM algorithms, namely the max-wins voting by [66] and pair-wise coupling by [22], are compared with the proposed method. The experimental results on synthetics data and microarray data show the effectiveness of the proposed method and that the new multi-class SVM is superior to max-wins and pair-wise coupling in terms of the classification of multiple-labeled microarray.

In order to analyze non-coding RNAs (ncRNAs) by kernel methods including support vector machines, [67] developed a new technique based on directed acyclic graphs (DAGs) derived from base-pairing probability matrices of RNA sequences that significantly increases the computation speed of stem kernels. Also, profile-profile stem kernels for multiple alignments of RNA sequences which utilize base-pairing probability matrices for multiple alignments instead of those for individual sequences were proposed. Stem kernels can be utilized as a reliable similarity measure of structural RNAs, and can be used in various kernel-based applications.

Support vector machines (SVMs) [2] have been shown superior performance in the analysis of microarray gene expression data than other classification algorithms such as decision trees, Parzen windows and Fisher's linear discrimination [16][68].

IV. EXPERIMENTAL RESULTS

The purpose of this paper is to evaluate the existing algorithms of gene expression on some datasets to produce sufficiently good classifiers. The existing algorithms of gene expression were applied to the available cancer datasets namely Leukemia and Lymphoma dataset and experimented using Matlab.

A. Leukemia dataset

The leukemia data set contains expression levels of 7129 genes taken over 72 samples. Labels indicate which of two variants of leukemia is present in the sample. This dataset is of the same type as the colon cancer dataset and can therefore be used for the same kind of experiments.

B. Lymphoma dataset

Lymphoma data set contain 42 samples derived from diffuse large B-cell lymphoma (DLBCL) and 9 samples from follicular lymphoma (FL) after that 11 samples from chronic lymphocytic leukemia (CLL). The entire data set contain 4026 genes. In this data set, a small part of data is missing.

Table I. performance evaluation

Dataset	Algorithms	No. of Genes	Accuracy in %age
Leukemia dataset	SVM-RFE	140	92
	SVM-CGH	140	91
	SVM-SFS	140	93
	SVM-MSFS	140	95
Lymphoma dataset	SVM-RFE	150	91
	SVM-CGH	150	92
	SVM-SFS	150	94
	SVM-MSFS	150	97

The Table 1 gives the performance analysis for the four SVM gene expression (feature selection) algorithms SVM-RFE, SVM-CGH, SVM-SFS and SVM-MSFS used by Leukemia and Lymphoma dataset. It has been concluded that the SVM-MSFS provides good classification accuracy results for Leukemia and Lymphoma datasets as compared to other SVM algorithms.

V. CONCLUSION

The gene expression data has become an important approach used for cancer classification. The main aim of gene expression dataset is to train known tissue samples in order to classify unknown samples. It has been concluded that the main objectives of microarrays can be classified into three groups: gene finding, class discovery and class prediction. A number of algorithms have been used for cancer classification which comes under the above mentioned three groups. This paper reviews and analyzes some of those algorithms and provides a way to investigate important genes. The algorithms for gene selection were then implemented on two datasets: Leukemia and Lymphoma. It has been observed that Modified Successive Feature Selection (MSFS) with SVM classification provides the better results.

REFERENCES

[1] Heena Farooq Bhat, MAW, (2013) Modified One-Against-All Algorithm Based on Support Vector Machine. International Journal of Advanced Research in Computer Science and Engineering (IJARCSSE).
[2] Vapnik, V. N. (1998). Statistical Learning Theory. Wiley.
[3] Cristianini,N. and Shawe-Taylor,J. (2000) An Introduction to Support Vector Machines. Cambridge University Press, MA.
[4] V.N. Vapnik, "The Nature of Statistical Learning Theory," Springer-Verlag, New York, 1999.
[5] Kreßel, U.H.G.: Pairwise Classification and Support Vector Machines. In: Advances in Kernel Methods, pp. 255–268. MIT Press, Cambridge (1999).
[6] Watanachaturaporn, P., Arora, M. K., & Varshney, P. K. (2004, May). Evaluation of factors affecting support vector machines for hyperspectral classification. In the American Society for Photogrammetry & Remote Sensing (ASPRS) 2004 Annual Conference, Denver, CO.
[7] Noble, W.S. (2004) Support vector machine applications in computational biology. Kernel Methods in Computational Biology, MIT Press, Cambridge, MA.
[8] Zien,A., Rätch,G., Mika,S., Schölkopf,B., Lengauer,T. and Müller,K.-R. (2000) Engineering support vector machine kernels that recognize translation initiation sites. Bioinformatics, 16, 799–807.
[9] Jaakkola,T., Diekhans,M. and Haussler,D. (1999) Using the Fisher kernel method to detect remote protein homologies. Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology, AAAI Press, Menlo Park, CA, pp. 149–158.
[10] Liao,L. and Noble,W.S. (2002) Combining pairwise sequence similarity and support vector machines for remote protein homology detection. Proceedings of the Sixth Annual International Conference on Computational Molecular Biology, pp. 225–232.
[11] Leslie,C., Eskin,E., Weston,J. and Noble,W.S. (2003) Mismatch string kernels for SVM protein classification. In Becker,S., Thrun,S. and Obermayer,K. (eds), Advances in Neural Information Processing Systems. MIT Press, Cambridge, MA.
[12] Ding,C. and Dubchak,I. (2001) Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics, 17, 349–358.
[13] Brown,M.P.S., Grundy,W.N., Lin,D., Cristianini,N., Sugnet,C., Furey,T.S., Ares,J.M. and Haussler,D (2000) Knowledge-based analysis of microarray gene expression data using support vector machines. Proc. Natl Acad. Sci., USA, 97, 262–267.
[14] Guyon,I., Weston,J., Barnhill,S. and Vapnik,V. (2001) Gene selection for cancer classification using support vector machines. Machine Learning, 46, 389–422.
[15] Furey,T.S., Cristianini,N., Duffy,N., Bednarski,D.W., Schummer,M. and Haussler,D. (2001) Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics, 16, 906–914.
[16] Mukherjee,S., Tamayo,P., Slonim,D., Verri,A., Golub,T., Mesirov,J. and Poggio,T. (1999) Support vector machine classification of microarray data. Technical Report AI Memo 1677, Massachusetts Institute of Technology.
[17] Vert,J.-P. and Kanehisa,M. (2003) Graph-driven features extraction from microarray data using diffusion kernels and kernel CCA. In Becker,S., Thrun,S. and Obermayer,K. (eds), Advances in Neural Information Processing Systems 15. MIT Press, Cambridge, MA.
[18] Pavlidis,P., Furey,T.S., Liberto,M., Haussler,D. and Grundy,W.N. (2001) Promoter region-based classification of genes. In Altman,R.B., Dunker,A.K., Hunter,L., Lauderdale,K., and Klein,T.E., (eds), Pacific Symposium of Biocomputing, Singapore, pp. 151–163. World Scientific.
[19] Bock,J.R. and Gough,D.A. (2001) Predicting protein–protein interactions from primary structure. Bioinformatics, 17, 455–460.

[20] Anderson,D.C., Li,W., Payan,D.G. and Noble,W.S. (2003) A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: support vector machine classification of peptide MS/MS spectra and SEQUEST scores. J. Proteome Res., 2, 137–146.

[21] Slonim DK, Tamayo P, Mesirov JP, Golub TR, Lander ES. Class prediction and discovery using gene expression data. Proceedings 4th Annual International Conference on Computational Molecular Biology (RECOMB), Tokyo, Japan, 2000; 263–272.

[22] J. Zhu, S. Rosset, T. Hastie, R. Tibshirani. 1-norm support vector machines. Advances in Neural Information Processing Systems 16(1): 49-56, 2004.

[23] Gupta, B., & Mishra, R. B. (2013). Neuro-psychiatric Disease Prediction Using support Vector Machine. DEMENTIA, 200(71), 129.

[24] Li, T., Zhang, C., & Ogihara, M. (2004). A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. Bioinformatics, 20(15), 2429-2437.

[25] Alshamlan, H. M., Badr, G. H., & Alohali, Y. (2013). A Study of Cancer Microarray Gene Expression Profile: Objectives and Approaches. In Proceedings of the World Congress on Engineering (Vol. 2).

[26] Yuchun Tang; Yan-Qing Zhang; Zhen Huang, "Development of Two-Stage SVM-RFE Gene Selection Strategy for Microarray Expression Data Analysis," Computational Biology and Bioinformatics, IEEE/ACM Transactions on , vol.4, no.3, pp.365,381, July-Sept. 2007

[27] Duan, K. B., Rajapakse, J. C., & Nguyen, M. N. (2007). One-versus-one and one-versus-all multiclass SVM-RFE for gene selection in cancer classification. In Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics (pp. 47-56). Springer Berlin Heidelberg.

[28] Abou-Taleb, A. S., Mohamed, A. A., Mohamed, O. A., & Abedelhalim, A. H. Hybridizing Filters and Wrapper Approaches for Improving the Classification Accuracy of Microarray Dataset 2013.

[29] Chapelle O, Vapnik V, Bousquet O, Mukherjee S. Choosing kernel parameters for support vector machines. AT&T Labs technical report, 2000.

[30] Weston J, Muckerjee S, Chapelle O, Pontil M, Poggio T, Vapnik V. Feature selection for SVMs. Advances in Neural Information Processing Systems 13. In: Solla SA, Leen TK, Muller KR (eds) MIT Press, 2001.

[31] Campbell C, Li Y, Tipping M. An efficient feature selection algorithm for classification of gene expression data. Neural Information Processing Systems: Natural and Synthetic, British Columbia, Canada, 2001, pp 123–134.

[32] Liu, J., Ranka, S., & Kahveci, T. (2008). Classification and feature selection algorithms for multi-class CGH data. Bioinformatics, 24(13), i86-i95.

[33] Guan Y, Myers C, Hess D, Barutcuoglu Z, Caudy A, Troyanskaya O. Predicting gene function in a hierarchical context with an ensemble of classifiers. Genome Biology. 2008; 9:S3.

[34] Chicco, D. (2012). Support Vector Machines in Bioinformatics: a Survey.

[35] Nanni, L., Brahnam, S., & Lumini, A. (2012). Combining multiple approaches for gene microarray classification. Bioinformatics, 28(8), 1151-1157.

[36] Revathi, T., & Sumathi, P. A Successive Feature Selection Algorithm for Gene Ranking (2014).

[37] Mao, Y., Zhou, X., Pi, D., Sun, Y., & Wong, S. T. (2005). Multiclass cancer classification by using fuzzy support vector machine and binary decision tree with gene selection. BioMed Research International, 2005(2), 160-171.

[38] Y. Kun, C. Zhipeng, L. Jianzhong, and L. Guohui, “A stable gene selection in microarray data analysis,” BMC Bioinformatics, vol. 7, no. 1, pp. 1–16, 2006.

[39] Deisy, C., Subbulakshmi, B., Baskar, S., & Ramaraj, N. (2007). Efficient dimensionality reduction approaches for feature selection, International Conference on Computational Intelligence and Multimedia Applications. India: Sivakasi (pp.121–127).

[40] Backstrom, L., & Caruana, R. (2006). C2FS: An algorithm for feature selection in cascade neural networks, IEEE International Joint Conference on Neural Networks. Canada: Vancouver, BC, pp. 4748–4753

[41] S. Yvan, I. aki, and L. Pedro, “A review of feature selection techniques in bioinformatics,” Bioinformatics, vol. 23, no. 19, pp. 2507–2517, Sep. 2007.

[42] Maulik, U.; Mukhopadhyay, A.; Chakraborty, D., "Gene-Expression-Based Cancer Subtypes Prediction Through Feature Selection and Transductive SVM," Biomedical Engineering, IEEE Transactions on , vol.60, no.4, pp.1111,1117, April 2013

[43] Saberkari, H., Shamsi, M., Joroughi, M., Golabi, F., & Sedaaghi, M. H. (2014). Cancer Classification in Microarray Data using a Hybrid Selective Independent Component Analysis and υ-Support Vector Machine Algorithm. Journal of Medical Signals and Sensors, 4(4), 291–298.

[44] www.ijsrp.org/research-paper-1114/ijsrp p3507.pdf (2014)

[45] A. L. Tarca, R. Romero, and S. Draghici, “Analysis of microarray experiments of gene expression profiling,” American journal of obstetrics and gynecology, vol. 195, no. 2, pp. 373–388, 2006.

[46] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., & Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. science, 286(5439), 531-537.

[47] G. Sheng-Bo, L. M. R., and T.-M. Lok, “Gene selection based on mutual information for the classification of multi-class cancer,” in Proceedings of the 2006 international conference on Computational Intelligence and Bioinformatics - Volume Part III, ser. ICIC’06. Springer-Verlag, 2006, pp. 454–463.

[48] Y. T. Young, “Efficient multi-class cancer diagnosis algorithm, using a global similarity pattern,” Comput. Stat. Data Anal.,vol. 53, no. 3, pp. 756–765, Jan. 2009. [Online]. Available: http://dx.doi.org/10.1016/j.csda.2008.08.028

[49] www.markowetzlab.org/docs/markowetz@gfkl2002.pdf

[50] Ibrahim, R.; Yousri, N.A.; Ismail, M.A.; El-Makky, N.M., "miRNA and gene expression based cancer classification using self-learning and co-training approaches," Bioinformatics and Biomedicine (BIBM), 2013 IEEE International Conference on , vol., no., pp.495,498, 18-21 Dec. 2013

[51] Chakraborty, D.; Maulik, U., "Identifying Cancer Biomarkers From Microarray Data Using Feature Selection and Semisupervised Learning," Translational Engineering in Health and Medicine, IEEE Journal of , vol.2, no., pp.1,11, 2014

[52] H. Jorng-Tzong, W. Li-Cheng, L. Baw-Juine, K. Jun-Li, K. Wen- Horng, and Z. Jin-Jian, “An expert system to classify microarray gene expression data using gene selection by decision tree,” Expert Syst. Appl., vol. 36, no. 5, pp. 9072–9081, Jul. 2009.

[53] Öz, E., & Kaya, H. (2013). Support vector machines for quality control of DNA sequencing. Journal of Inequalities and Applications, 2013(1), 1-9.

[54] www.qub.ac.uk/researchcentres/EPIC/Research/IntelligentSystems/AdvancedLearningAlgorithmforMicroarraryDataAnalysis/

[55] J. Friedman. (1996) Another Approach to Polychotomous Classification. Dept. Statist., Stanford Univ.,Stanford, CA. [Online]. Available: http:// wwwstat. stanford.edu/reports/friedman/ply.ps.Z.

[56] 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(5), 631-643.

[57] Lee, Y., & Lee, C. K. (2003). Classification of multiple cancer types by multicategory support vector machines using gene expression data. Bioinformatics, 19(9), 1132-1139.

[58] Yuvaraj, N.; Vivekanandan, P., "An efficient SVM based tumor classification with symmetry Non-negative Matrix Factorization using gene expression data," Information Communication and Embedded Systems (ICICES), 2013 International Conference on , vol., no., pp.761,768, 21-22 Feb. 2013
[59] Radmacher, M. D., McShane, L. M., & Simon, R. (2002). A paradigm for class prediction using gene expression profiles. Journal of Computational Biology,9(3), 505-511.
[60] Crammer, K. and Singer, Y. (2000). On the learnability and design of output codes for multiclass problems. In Computational Learning Theory, pages 35–46.
[61] Weston,J. and Watkins,C. (1999) Support vector machines for multi-class pattern recognition. In Proceedings of the Seventh European Symposium On Artificial Neural Networks (ESANN 99), Bruges, April 21–23
[62] Yoonkyung Lee, Cheo Koo Lee .2003. Classification of multiple cancer types by Multicategory support vector machines using gene expression data, Vol. 19 no. 9, Bioinformatics.
[63] Mallika, R., & Saravanan, V. (2010). An svm based classification method for cancer data using minimum microarray gene expressions. World Academy Of Science, Engineering And Technology, 62.
[64] Soo Beom Choi; Jee Soo Park; Jai Won Chung; Tae Keun Yoo; Deok Won Kim, "Multicategory classification of 11 neuromuscular diseases based on microarray data using support vector machine," Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE , vol., no., pp.3460,3463, 26-30 Aug. 2014
[65] Rathore, S.; Hussain, M.; Khan, A., "GECC: Gene Expression Based Ensemble Classification of Colon Samples," Computational Biology and Bioinformatics, IEEE/ACM Transactions on , vol.11, no.6, pp.1131,1145, Nov.-Dec. 1 2014.
[66] J. Friedman. (1996) Another Approach to Polychotomous Classification. Dept. Statist., Stanford Univ.,Stanford, CA. [Online]. Available: http:// wwwstat. stanford.edu/reports/friedman/ply.ps.Z.
[67] Sato, K., Mituyama, T., Asai, K., & Sakakibara, Y. (2008). Directed acyclic graph kernels for structural RNA analysis. BMC bioinformatics, 9(1), 318.
[68] M. Brown, W. Grundy, D. Lin, N. Cristianini, C. Sugnet, J. Manuel Ares, and D. Haussler. “Support vector machine classification of microarray gene expression data”, Technical Report UCSC-CRL-99-09, Department of Computer Science, University of California, Santa Cruz, 1999.