International Journal of Computational Intelligence Research
ISSN 0973-1873 Volume 13, Number 8 (2017), pp. 1999-2004

http://www.ripublication.com

Micro Array Gene Expression Classification Using Support Vector Machine Ensembles

C.Kanimozhi*

Assistant Professor, Department of Information Technology,
University College of Engineering, B.I.T Campus,
Anna University, Trichirappalli.

Dr.A.Valarmathi

Assistant Professor, Department of Computer Applications,
University College of Engineering, B.I.T Campus,
Anna University, Trichirappalli.

Abstract

Micro array gene expression data is a high dimensional data which contains expression levels for thousands of genes with relatively small number of samples. The treatment failure in the most common childhood malignancy of Acute Lymphoblastic Leukemia (ALL) are better expressed through the expression levels of genes. This paper aims to develop a classification algorithm by analyzing the expression levels of ALL dataset for solving the multiclass problem. An Ensemble of Support Vector Machines is employed for better accuracy in the classification.

Keywords: Micro array gene expression; Ensemble; Support Vector Machines; Classification.

1. INTRODUCTION

Gene expression microarray data is a form of high-throughput genomics data providing relative measurements of mRNA levels for thousands of genes in a biological sample. The analysis of large expression data sets is becoming a challenge in cancer classification as they are characterized by small number of samples with thousands of expression levels for genes. The fourth most common childhood malignancy is Acute Lymphoblastic Leukemia (ALL). Leukemia relapse represents

the outgrowth of a clonal cell population not completely eliminated by the treatment. The treatment failure in childhood Acute Lymphoblastic Leukemia ie., re-growth of cancerous cells can be termed as early relapse (< 30 months) and late relapse (>30 months). About 15-20% of the children are prone to treatment failure and 80-85% is a success, which is a no relapse.

The aforementioned conditions are expressed in genes and hence micro arrays are employed for the classification of relapse as early, late and no relapse. As unsupervised methods indirectly identifies the classes through clusters, there is a possibility of misclassification. Support Vector machines (SVM) are the most powerful supervised learning algorithm reported in the literature that are used for classification tasks. The objective of training a Support Vector Machine is to find the optimal hyperplane that separates the data of different classes. The effectiveness of an SVM depends upon the kernel function, parameters of the kernel and the soft margin parameter C.

2. RELATED WORKS

2.1. SVMs

The Support Vector Machine (SVM) is a state-of-the-art classification method introduced in 1992 by Boser, Guyon, and Vapnik[1]. The SVMs based on empirical risk minimization are stable classifiers. When the training data are not linearly separable in the input space, SVMs can use kernel functions to project the training data to a feature space of a higher dimension, in which the separation becomes easier. In recent years, SVMs have been employed successfully in the classification of micro array gene expression data, with their advantages to solve difficulties such as small-size samples and high dimension. But training a SVM during the training phase involves solving a quadratic optimization problem and requires high computational cost of $O(|T|^2)$ [2], where T is the number of samples or training instances. In addition, SVMs with a Gaussian Radial Basis Function (RBF) exhibits better classification for multi class problems [3] represented as

$$K(Xi, Xj)=e^{-\gamma|Xi-Xj|2} \qquad (1)$$

For this reason, a Gaussian RBF with parameter δ is chosen, and the best combination of C and δ is selected by a grid search. Besides the excellent classification performances by the SVMs with only limited number of training samples, many variations to the SVMs are suggested to improve the classification performance such as semisupervised SVMs [4],[5], active learning with SVMs [6],[7], applying feature extraction or feature selection [8],[9].

2.2. Ensembles

Although the above mentioned works promises very good results, it is still possible to improve the classification performance by means of ensemble of classifiers. Ensembles are a group of base classifiers where the prediction is the combination of all the base classifiers in some order to reduce the generalization error[11]. An ensemble can be constructed using different supervised learners as base classifiers or from the same supervised learner. When the same base learner is used, diversity must

be introduced by means of manipulation of the training samples, parameters to the kernel function and features of the instances.

3. PROPOSED WORKS

3.1. Dataset

The pediatric acute leukemia dataset was obtained from the Gene Expression Omnibus(GEO) at the National Center for Biotechnology Information(NCBI), a public repository for Gene expression data. These datasets have been preprocessed for removal of missing values and other discrepancies using appropriate tool before actual processing. This dataset consists of 197 samples with 3 classes namely early relapse, no relapse and late relapse.

3.2. Methodology

An Ensemble model is built by employing M Support Vector Machines as base classifiers. The Radial Basis Function(RBF) can be chosen as the kernel function as it supports multi class classification. Micro array gene expression data is divided into M disjoint training sets. Each base classifier is trained and tested on the disjoint training sets. The collective classification accuracy is the average classification of all the base classifiers.

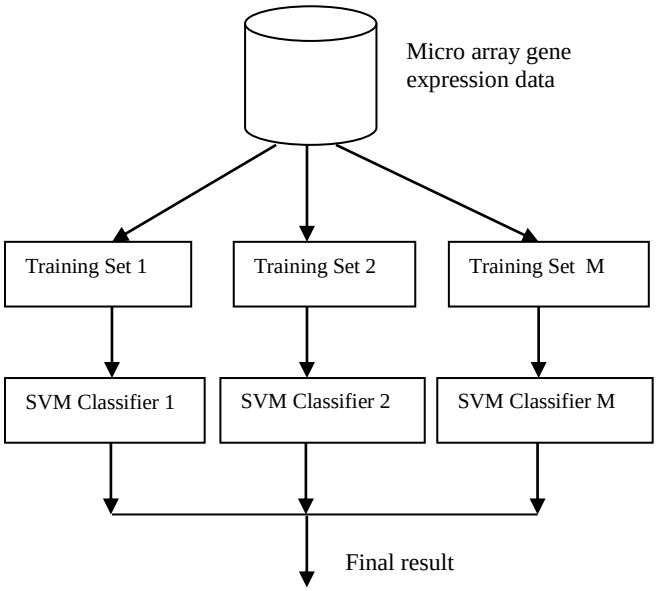

Fig 1. Schematic Diagram of proposed approach

4. EXPERIMENTS AND RESULTS

All experiments were performed on a computer having the following features Intel(R) Core(TM) i7-4702MQ CPU running @2.20 GHz and 4GB RAM on Windows 8.1 Pro Operating System.

4.1. Experimental Settings

SVM Ensemble classifiers are evaluated using micro array gene expression dataset for Pediatric acute lymphoblastic leukemia(ALL) . The dataset is validated by means of 10-fold cross validation technique. Thus the dataset is divided into 10 disjoint subsets and used for training/testing. For training SVM, the freely available library LibSVM [12] is used through the interface e1071 in R. A Gaussian RBF Kernel is used as a base classifier for the proposed approach. An ensemble model is created using three SVM base classifiers. The values for C and δ is chosen by a grid search to introduce diversity among the base classifiers.

4.2. Experimental Results

The labeled instances of the dataset is divided into training and test set in the ratio of 2:1. A single SVM classifier is build with hyper parameters value set as C=0.25 and δ=0.0676. The classifier is trained with the training dataset and tested with test dataset. The single Support Vector Machine is observed to achieve 83.09% classification accuracy. The second model is build as an Ensemble, employing K-Nearest Neighbor (KNN), Random Forest (RF) and Support Vector Machine as base classifiers. This model is subjected to bootstrap aggregation technique in which each instance of the data set possess the probability of getting selected during each iteration of the training process. A classification accuracy of 84.79% is attained through this ensemble. The third model is also build as an Ensemble employing C5.0, Gradient Boosting Method (GBM), K-Nearest Neighbor and Support Vector Machine as base classifiers. Boosting technique is employed in the model to reduce bias and the classification accuracy is 83.26%.

The proposed model is build as an Ensemble with M SVM base classifiers. Ensemble diversity is ensured by means of dividing the entire dataset into M disjoint training sets. Each base classifier is trained and tested using the disjoint data sets. As bagging and boosting techniques applied to ensembles of SVM classifiers does not improve the performance of an Ensemble, simple averaging of the base classifiers is considered as the classification accuracy of the Ensemble.

All the above mentioned models are trained and tested for 10 runs and an average accuracy value of the model is considered for comparison with all other models. The table below mentions the techniques used for construction of a model, the training algorithms used as base classifiers and the accuracy obtained from each model.

Table I. Comparison of the classification accuracy by different classification techniques.

Techniques	**Base Classifiers**	**Accuracy (%)**
Single SVM	-	83.09
Ensemble-Bagging	KNN,RF and SVM	84.79
Ensemble-boosting	C5.0, GBM, KNN and SVM	83.26
SVM Ensemble	SVM	**88.06**

The classification performance of SVM ensembles is 88.06% and it outperforms the classification accuracy when compared to single SVM and the other ensembles using bagging and boosting techniques as shown in Table I. The classification accuracy can better be depicted by the following figure using different techniques employed for building a model.

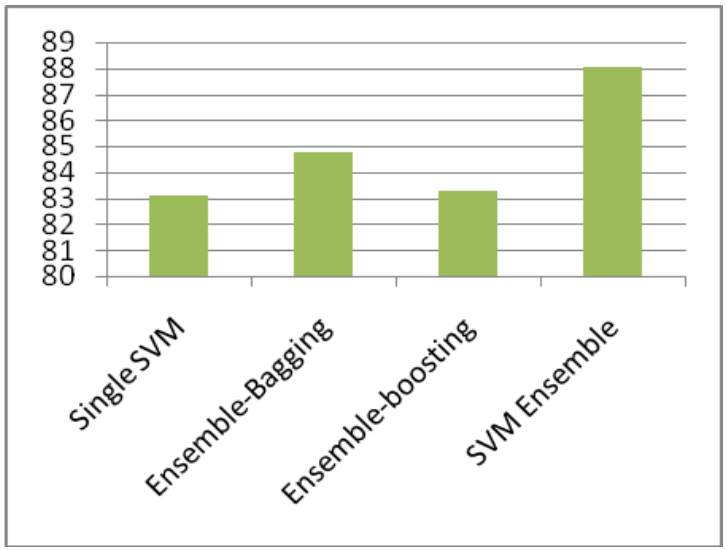

Fig 2. Classification accuracy using different techniques

CONCLUSION

In this paper, Support Vector Machine ensembles have been proposed by introducing diversity in the base classifiers by training them using disjoint training sets from micro array gene expression data for pediatric Acute Lymphoblastic Leukemia(ALL). The proposed Support Vector Machine ensemble is employed to classify the treatment failure in ALL data set as early relapse, late relapse and no relapse which is interestingly a multiclass problem. When compared with the Single SVM and the ensembles constructed using different base classifiers, the proposed Support Vector Machine ensembles trained with disjoint sets of data gives better classification performance. As the micro array gene expression data contains thousands of expression levels of genes, the future work may be set in to reduce the feature set of genes which are highly relevant for the disease.

REFERENCES

[1] Vapnik VN. The Nature of Statistical Learning Theory. New York: Springer-Verlag 1995.

[2] O.Chapelle, “Training a support vector machine in the primal,” Neural Comput., vol. 19, no. 5, pp. 1155–1178, 2007.

[3] F.Melgani and L.Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," IEEE Trans. Geosci. Remote Sens., vol. 42, no. 8, pp. 1778–1790, Aug. 2004.

[4] G. Camps-Valls, T. V. Bandos-Marsheva, and D. Zhou, "Semi-supervised graph-based hyperspectral image classification," IEEE Trans. Geosci.Remote Sens., vol. 45, no. 10, pp. 2044–3054, Oct. 2007.

[5] M. Chi and L. Bruzzone, "Semisupervised classification of hyperspectral images by SVMs ptimized in the primal," IEEE Trans. Geosci. Remote Sens., vol. 45, no. 6, pp. 1870–1880, Jun. 2007.

[6] D. Tuia, F. Ratle, F. Pacifici, M. F. Kanevski, and W. J. Emery, "Active learning methods for remote sensing image classification," IEEE Trans. Geosci. Remote Sens., vol. 47, no. 7, pp. 2218–2232, Jul. 2009.

[7] W. Di and M. M. Crawford, "Active learning via multi-view and local proximity co-regularization for hyperspectral image classification," IEEE J. Sel. Top. Signal Process., vol. 5, no. 3, pp. 618–628, Jun. 2011.

[8] M. Pal and G. M. Foody, "Feature selection for classification of hyperspectral data by SVM," IEEE Trans. Geosci. Remote Sens., vol. 48, no. 5,pp. 2297–2307, May 2010.

[9] W. Liao, A. Pizurica, P. Scheunders, W. Philips, and Y. Pi, "Semisupervised local discriminant analysis for feature extraction in hyperspectral images," IEEE Trans. Geosci. Remote Sens., vol. 51, no. 1, pp. 184–198,Jan. 2013.

[8] L. Xu, A. Krzyzak, and C. Y. Suen, "Methods of combining multiple classifiers and their applications to handwriting recognition," IEEETrans.Syst., Man Cybern., vol. 22, no. 3, pp. 418–435, May/Jun. 1992.

[9] R. Ranawana and V. Palade, "Multi-classifier systems: Review and a road map for developers," Int. J. Hybrid Intel. Syst.,vol.3,no.1,pp.1–41, Jan. 2006.

[10] A. Merentitis, C. Debes, and R. Heremans, "Ensemble learning in hyperspectral image classification: Toward selecting a favorable bias-variance tradeoff," IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.,vol.7, no. 4, pp. 1089–1102, Apr. 2014.

[11] L. I. Kuncheva, Combining Pattern Classifiers: Methods and Algorithms.New York, NY, USA: Wiley-Interscience, 2004.

[12] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for Support Vector Machines," ACM T.Intel. Syst. Tec., vol. 2, no. 3, pp.27:1–27:27, 2011.