A Hybrid Approach for Gene Selection and Classification using Support Vector Machine

Jaison Bennet[1], Chilambuchelvan Ganaprakasam[1], and Nirmal Kumar[2]
[1]Department of Computer Science and Engineering, RMK Engineering College, India
[2]Software Developers, Wipro Technologies, India

Abstract: *Deoxyribo Nucleic Acid (DNA) microarray technology allows us to generate thousands of gene expression in a single chip. Analyzing gene expression data plays vital role in understanding diseases and discovering medicines. Classification of cancer based on gene expression data is a promising research area in the field of bioinformatics and data mining. All genes do not contribute for efficient classification of samples. Hence, a robust feature selection method is required to identify the relevant genes which help in the classification of samples effectively. Most of the existing feature selection methods are computationally expensive. Redundancy in gene expression data leads to poor classification accuracy and also acts bad on multi class classification. This paper proposes an ensemble feature selection technique which is a combination of Recursive Feature Elimination (RFE) and Based Bayes error Filter (BBF) for gene selection and Support Vector Machine (SVM) algorithm for classification. The proposed ensemble gene selection method yields comparable performance on classification when compared to existing classifiers and provides a new insight in feature selection.*

Keywords: *BBF, classification, microarray, RFE, SVM.*

Received February 14, 2013; accepted August 12, 2014; published online August 9, 2015

1. Introduction

Deoxyribo Nucleic Acid (DNA) acts as a template for making copies of itself and also as a blueprint for a molecule called Ribo Nucleic Acid (RNA). The process of transcribing a gene's DNA sequence into RNA is called gene expression. A gene's expression level indicates the approximate number of copies of that gene's RNA produced in a cell and it is correlated with the amount of the corresponding proteins made [9]. Microarray is the technology for measuring the expression levels of tens of thousands of genes in parallel in a single chip. Each chip is about 2cm by 2cm and microarrays contain up to 6000 spots. Different Microarray technologies include Serial Analysis of Gene Expression (SAGE), nylon membrane, and illumina bead array. Thus, microarrays offer an efficient method of gathering data that can be used to determine the expression pattern of thousands of genes. High dimensionality of gene expression data is a big challenge in most classification problems. Large number of features (genes) against small sample size and redundancy in expressed data are the main two reasons which lead to poor classification accuracy [17]. Subsequently dimension reduction is essential to classification. Support Vector Machine (SVM) is a supervised computer learning technique used for data classification. SVM's have been performing well in evaluating microarray expression data [10]. It performs classification on data by placing an optimal hyper plane which maximizes the functional margin [3, 11, 13].

In the past several decades, to increase the accuracy of classification minimizing the number of features against sample size was usually in practice. Many methods are available for feature selection to remove noisy genes from data set and to improve classification performance. Some of the methods are chi-squared, information gain, gain ratio, relief**,** Support Vector Machine-Recursive Feature Elimination (SVM-RFE). After feature selection, there are several classifiers for classifying class labels. The classifiers are meant to get trained on feature selection data and then tested on an independent test dataset to evaluate the accuracy of it [12]. Some of the classifiers in practice are nearest neighbor, logistic model tree, bayes network, artificial neural networks, SVM etc.

Among various gene selection methods, in this paper we propose hybrid feature selection technique which is a combination of SVM-RFE and Based Bayes Filter (BBF) for gene selection and sequential minimal optimization algorithm for training SVM classification method. The experiments have been conducted on publicly available leukemia data set. The results bring new insights on feature selection and achieve better accuracy than existing classification methods.

The rest of the paper is organized as follows. Related work is being discussed in section 2. In section 3 we describe various existing gene selection methods. We present the proposed hybrid gene selection technique in sections 4 and 5 gives details on experiments being conducted on leukemia dataset.

Section 6 briefs results and discussion. In section 7 we conclude with some scope for future work.

2. Related Works

Jianchen *et al.* [5] presents feature selection technique which plays a vital role in gene classification. This paper proposes SVM-RFE for feature selection in multi class classification. Here, class interval is used as the evaluation criterion and it eliminates features in a recursive manner. Chaos particle swarm optimization algorithm is used for feature selection. The results are being validated with the publicly available data sets from UCI repository.

Yuchun *et al.* [18] carried feature selection by SVM-RFE in two stages to avoid instability. The first stage is a pre-filtering process, which specifically eliminate irrelevant, redundant and noisy genes while keeping informative genes. It involves multiple iterations to generate gene subsets which clearly avoid redundancy. In stage two, all gene subsets are combined together and eliminate one gene at each step. This final gene subset promises effective result on classification. Linear SVM is used for classification. Publically available datasets such as ALL/AML, colon cancer and lymphoma are used for performance evaluation. This feature selection method is compared with correlation ranking method in terms on accuracy and area under ROC.

Kai-Bo *et al.* [8] proposes a new feature selection method that uses a backward elimination procedure similar to that implemented in SVM-RFE. This approach unlike SVM-RFE method at each step, computes the feature ranking score from a statistical analysis of weight vectors of multiple linear SVMs trained on subsamples of the original training data. The proposed method was tested on four gene expression datasets for cancer classification. Results show that the proposed feature selection method yields better gene subsets than the original SVM-RFE and improves the classification accuracy.

Ji-Gang *et al.* [4] proposed a new method, Based BBF, to select relevant genes and remove redundant genes in classification analysis of microarray data. The effectiveness and accuracy of this method is demonstrated through analysis of five publicly available microarray datasets such as colon cancer, DLBCL, leukemia, prostate and lymphoma dataset. To assess the performance they use Leave One Out Cross Validation (LOOCV). This provides realistic assessment of classifiers which generalize well to new data. This feature selection technique combined with SVM yields better accuracy.

3. Gene Selection Methods

3.1. Information Gain

In this method, rank each feature according to some univariate metric and only the highest ranking features are used while the remaining low ranking features are eliminated. Hence, only genes with high ranking are used for classification. However, gene ranking based on univariate methods has some drawbacks and to mention one is that the genes selected are most probably redundant. Highly ranked genes may carry similar discriminative information towards the defined class. Elimination of one high ranked gene may not cause any degradation of classification accuracy [7, 14]. Therefore it does not perform well for similar class labels.

3.2. Genetic Algorithm

In this method, a search is conducted in the space of genes, evaluating the goodness of each gene subset by the estimation of the accuracy percentage of the specific classifier to be used, training the classifier only with the found genes. It is claimed that this approach obtains better predictive accuracy estimates than the previous approach. A common drawback in this method is that they have a higher risk of over fitting than filter techniques and are very computationally intensive. In contrast, it incorporate the interaction between genes selection and classification model, which make them unique compared to existing ones [18].

3.3. Recursive Feature Elimination

In RFE method [3], nested subsets of features are selected in a sequential backward elimination manner, which starts with all the feature variables and removes one feature variable at a time. At each step, the coefficients of the weight vector w of a linear SVM are used to compute the feature ranking score. In the SVM-RFE, the gene being removed should change the objective function *j* least Equation 1.

$$j = \frac{\|\omega\|^2}{2} \quad (1)$$

3.4. Based Bayes Error Filter

BBF [4], for gene selection is implemented in two steps. First the relevant candidate genes are selected by a criterion function and second the criterion controlling the upper bound of the Bayes error is applied to the relevant candidate genes in order to remove the redundant genes. This method can effectively perform gene selection with reasonably low classification error rates and a small number of selected genes. This not only obtains a small subset of informative genes for classification analyses, but also provides a balance between selected gene set size and classification accuracy.

The probability of classification error of any classifier is lower bounded by the Bayes error [5]. Therefore, attention has focused on approximations

and bounds for the Bayes error. One of these bound estimations for the Bayes error is provided by the Bhattacharyya distance. The Bhattacharyya distance, dB, [2, 16] can be a seperability measure between two classes and also can give lower and upper bounds of the Bayes error.

The Bhattacharyya distance (d_B) is given by:

$$d_B = \frac{1}{8}(M_2 - M_1)^T \left[\frac{\Sigma_1 + \Sigma_2}{2}\right]^{-1} (M_2 - M_1) + \frac{1}{2} ln \frac{\left|(\Sigma_1 + \Sigma_2)/2\right|}{|\Sigma_1|^{1/2} |\Sigma_2|^{1/2}} \quad (2)$$

Herem M_2, M_1 are the mean vectors of class and Σ_1, Σ_2 are the covariance matrixes of class. The first term of Equation 2 gives the class separability due to the difference between class means, and the second term gives the class separability due to the difference between class covariance matrices.

4. Hybrid Gene Selection Technique

This new ensemble approach is the combination of SVM-RFE and BBF. SVM-RFE yields good performance on classification but lacks on poor seperability in redundant class labels. BBF avoids redundant class labels in selection. Both combined achieves comparable performance.

The Figure 1 shows the schematic view of overall process carried. The feature selection from dataset is performed with SVM-RFE and BBF. After selection it is undergone with classifier for training. Finally evaluation carried with testing data.

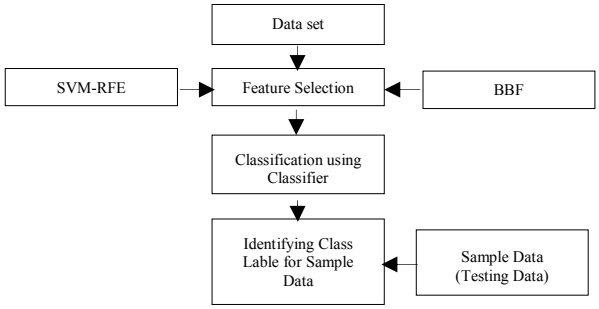

Figure 1. Schematic view of proposed system.

4.1. Feature Selection Method

The recursive elimination procedure of SVM-RFE [3] is implemented as follows:

1. Start: ranked set R=[ ]; picked feature subset S=[1, …, d].
2. Repeat until all features in subset gets ranked:
   a. Train the features with SVM from set S as input variables.
   b. Calculate the weight vector for each feature.
   c. Calculate the ranking score for features in set S.
   d. Identify the feature with the smallest ranking score.
   e. Update.
   f. Eliminate smallest ranking feature.
3. Result: Ranked feature set R[ ].

After ranking genes by SVM-RFE we eliminate redundancy by applying BBF. First the relevant candidate genes are selected by a criterion function and second the criterion controlling the upper bound of the Bayes error is applied to the relevant candidate genes in order to remove the redundant genes. This method can effectively perform gene selection with reasonably low classification error rates and a small number of selected genes. This not only obtains a small subset of informative genes for classification analysis, but also provides a balance between selected gene set size and classification accuracy.

4.2. Classification with SVM

SVM [3] is a data mining technique which classifies data in an intelligent manner. The SVM learns itself by separating data with a plane on a given training data and regression rules from data. SVM was first outlined by Vapnik *et al.* [15, 19] from statistical learning methods in the 1960 for classifying the data. SVM classifies data in large data sets by identifying a linear or non-linear separating surface in the input space of a data set. The separating surface depends only on the subset of the original data known as a set of support vectors. A SVM classifies data by placing one or more planes on data such that it achieves good classification results. A good result on separation is achieved by plane that has the largest distance to the nearest data points of any class, called functional margin. If this functional margin is large, then the generalization error of the classifier will be small and vice versa.

5. Experiment

In this section we apply SVM-RFE and BBF for gene selection and SVM for classification on the publicly available Leukemia dataset.

This dataset, provided by Golub *et al.* [1] contains the expression levels of 7,129 genes for 27 patients of Acute Lymphoblastic Leukemia (ALL) and 11 patients of Acute Myeloid Leukemia (AML). After data preprocessing, 3,051 genes remain in the dataset. The source of the 3,051 gene expression measurements is publicly available at the website: http://ligarto.org/rdiaz/Papers/rfVS/.

6. Results and Discussion

The above data set is utilized in the described process of feature selection and undergone classification using SVM. After proper ranking by SVM-RFE which reduces computational complexity, the resultant is given to BBF which eliminates redundancy. This method shows less classification error because the Bayes error depends only on the gene space and not on the classifier. Hence, it is possible to find out an

optimal gene set for a given classification problem, rendering the minimum classification error.

Figure 2 shows classification accuracy of 95.833% on given dataset with 72 instances in that it classifies 69 instances correctly 3 instances incorrectly.

```
Time taken to build model: 1.92 seconds

   Stratified cross-validation


Correctly Classified Instances          69               95.8333 %
Incorrectly Classified Instances         3                4.1667 %
Kappa statistic                          0.9072
Mean absolute error                      0.0417
Root mean squared error                  0.2041
Relative absolute error                  9.1472 %
Root relative squared error             42.7913 %
Total Number of Instances               72

   Detailed Accuracy By Class

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.979     0.08       0.958     0.979     0.968      0.949    ALL
                 0.92      0.021      0.958     0.92      0.939      0.949    AML
Weighted Avg.    0.958     0.06       0.958     0.958     0.958      0.949

   Confusion Matrix

  a  b      classified as
 46  1 |  a = ALL
  2 23 |  b = AML
```

Figure 2. Classification results before feature selection.

The same dataset is applied to hybrid feature selection SVM-RFE and BBF and instances incorrectly classified are reduced to 2. Accuracy ends with 97.22 % as shown in Figure 3 and shows better performance than single feature selection algorithms and the existing ensemble feature selection algorithms.

```
Time taken to build model: 4.7 seconds

   Stratified cross-validation


Correctly Classified Instances          70               97.2222 %
Incorrectly Classified Instances         2                2.7778 %
Kappa statistic                          0.9376
Mean absolute error                      0.0278
Root mean squared error                  0.1667
Relative absolute error                  6.0981 %
Root relative squared error             34.9389 %
Total Number of Instances               72

   Detailed Accuracy By Class

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0.08       0.959     1         0.979      0.983    ALL
                 0.92      0          1         0.92      0.958      0.983    AML
Weighted Avg.    0.972     0.052      0.973     0.972     0.972      0.983

   Confusion Matrix

  a  b      classified as
 47  0 |  a = ALL
  2 23 |  b = AML
```

Figure 3. Classification results after hybrid feature selection.

False Positives (FP rate) refer to negative samples that are classified as positive and True Positives (TP rate) refer to the correct classifications of positive examples. In our proposed method before feature selection the FP rate was two and false negative is one. The accuracy is 95.2 with ALL/AML (46/23). After this feature selection method, false negative falls to zero thus correctly classify entire ALL and accuracy reaches 97.5 with ALL/AML (47/23). Table 1 gives the summary of classification results.

Table 1. Classification accuracy before and after feature selection.

Feature Selection	No. of Instances	Correctly Classified	Incorrectly Classified	Classification Accuracy
Before	72	69	3	95.83
After	72	70	2	97.22

Precision is a measure of the accuracy provided that a specific class has been predicted. It is defined by $Precision= TP/(TP+FP)$, where TP and FP are the numbers of TP and FP predictions for the considered class. The precision for ALL and AML is 0.958 before preprocessing and after preprocessing precision for ALL remains same but for AML it is improved to 1 by reducing false negative to zero.

Recall is a measure of the ability of a prediction model to select instances of a certain class from a data set and corresponds to the TP rate. It is defined by the formula.

$$Recall=TP/(TP+FP)$$

Where TP and false negative are the numbers of TP and false negative predictions for the considered class. Here, Recall for ALL and AML before preprocessing is 0.979 and 0.92 respectively. On applying this method all ALL instances are correctly classified and hence, recall reaches one.

Table 2. Performance Comparison of Gene selection methods on different classifiers.

	Classifier		
Feature Selection Method	**JJ48**	**BBNet**	**SSVM**
SVM-RFE and BBF	991.33	994.4	997.2
Symmetrical Uncertainty	889.1	992.8	995.8
Information Gain	990.33	993.8	995.8
Relief	891.9	993.1	995.8

Table 2 shows the performance comparison of our Feature selection methods with SVM and other existing methods. It is very obvious that SVM-RFE and BBF with SVM achieves higher classification accuracy than the existing ensemble methods.

7. Conclusions

Classification of cancer based on gene expression data is a promising research area in the field of data mining. In this paper, hybrid gene selection technique which combines SVM-RFE and BBF has been proposed for gene selection. Based on the experimental results on leukemia dataset it is found that the performance of SVM-RFE and BBF combined with SVM for classification was superior to the previous related works in terms of gene selection and classification. SVM-RFE ranks the genes and BBF is applied to remove redundancy on top ranked genes. Moreover, several gene selection methods against different classifiers were compared. This approach can play a vital role in accurate cancer classification thus, eliminating the morphological and clinical means of diagnosis. There was a limitation in terms of time complexity which is left as a direction for future research. In future work, it can also be extended to perform classification on multi class labels.

References

[1] Golub T., Slonim D., Tamayo P., Huard C., Gaasenbeek M., Mesirov J., Coller H., Loh M., Downing J., Caligiuri M., Bloomfield C., and Lander E., "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," *Science*, vol. 286, pp. 531-537, 1999.

[2] Goudail F., Refregier P., and Delyon G., "Bhattacharyya Distance as a Contrast Parameter for Statistical Processing of Noisy Optical Images," *Journal of the Optical Society of America A: Optics, Image Science, and Vision*, vol. 21, no. 7, pp. 1231-1240, 2004.

[3] Guyon J., Weston S., Barnhill S., and Vapnik V., "Gene Selection for Cancer Classification using Support Vector Machines," available at: http://axon.cs.byu.edu/Dan/778/papers/Feature%20Selection/guyon*.pdf, last visited 2002.

[4] Ji-Gang Z. and Hong-Wen D., "Gene Selection for Classification of Microarray Data based on the Bayes Error," *BMC Bioinformatics*, vol. 8, no. 1, pp. 370-383, 2007.

[5] Jian L., Jin-Mao W., Tian Y., and Hai-Wei Z., "Feature Selection based on Bayes Minimum Error Probability," *in Proceedings of the 9th International Conference Fuzzy Systems and Knowledge Discovery*, Sichuan, pp. 706-710 , 2012.

[6] Jianchen W., Ganlin S., Xiusheng D., and Wen B., "Improved SVM-RFE Feature Selection Method for Multi-SVM Classifier," *in Proceedings of International Conference on Electrical and Control Engineering*, Yichang, pp. 1592-1595, 2011.

[7] Jung-Hsien C. and Shing-Hua H., "A Combination of Rough-Based Feature Selection and RBF Neural Network for Classification Using Gene Expression Data," *IEEE Transactions on Nano Bioscience*, vol. 7, no. 1, pp. 91-99, 2008.

[8] Kai-Bo D., Jagath R., and Haiying W., "Multiple SVM-RFE for Gene Selection in Cancer Classification with Expression Data," *IEEE Transactions on Nano Bioscience*, vol. 4, no. 3, pp. 228-234, 2005.

[9] Lee Y. and Han J., "Cancer Classification Using Gene Expression Data," *Information Systems*, vol. 28, pp. 243-268, 2003.

[10] Lee Y. and Lee C., "Classification of Multiple Cancer Types by Multi-Category Support Vector Machines using Gene Expression Data," *Bioinformatics*, vol. 19, no. 9, pp. 1132-1139, 2003.

[11] Luo L., Huang D., Ye L., Zhou Q., Shao G., and Peng H., "Improving the Computational Efficiency of Recursive Cluster Elimination for Gene Selection," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 8, no. 1, pp. 122-129, 2011.

[12] Meena K., Subramaniam K., and Gomathy M., "Gender Classification in Speech Recognition using Fuzzy Logic and Neural Network," *the International Arab Journal of Information Technology*, vol. 10, no. 5, pp. 477-484, 2013.

[13] Statnikov A., Wang L., and Aliferis C., "A Comprehensive Comparison of Random Forests and Support Vector Machines for Microarray based Cancer Classification," *BMC Bioinformatics*, available at: http://www.biomedcentral.com/content/pdf/1471-2105-9-319.pdf, last visited 2008.

[14] Sung-Bae C. and Hong-Hee W., "Machine Learning in DNA Microarray Analysis for Cancer Classification," *in Proceedings of the 1st Asia-Pacific Bioinformatics Conference on Bioinformatics*, pp. 69-78, 2003.

[15] Vapnik V., *Statistical Learning Theory*, Wiley, 1998.

[16] Xuan G., Zhu X., Chai P., Zhang Z., Shi Y., and Fu D., "Feature Selection based on the Bhattacharyya Distance," *in Proceedings of the 18th International Conference on Pattern Recognition*, Hong Kong , pp. 957-960, 2006.

[17] Yu L. and Liu H., "Redundancy based Feature Selection for Microarray Data," *in Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, USA, pp. 737-742, 2004.

[18] Yuchun T., Yan-Qing Z., and Zhen H., "Development of Two-Stage SVM-RFE Gene Selection Strategy for Microarray Expression Data Analysis," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 4, no. 3, pp. 365-381, 2007.

[19] Zhang H., Ho T., and Kawasaki S., "Wrapper Feature Extraction for Time Series Classification Using Singular Value Decomposition," *International Journal of Knowledge and Systems Science*, vol. 3, pp. 53-60, 2006.

Jaison Bennet received his Ms degree in computer science and engineering from Anna University, India in 2007. Currently, he is working towards the PhD degree in computer science and engineering and working as Assistant professor in RMK Engineering College, Chennai. He has published 6 papers in international journals and conferences His research interests include data mining, bioinformatics and knowledge discovery. He is a life member in IAENG, IACSIT and ISTE.

Chilambuchelvan Gnanaprakasam received his PhD from Anna University, Chennai in 2008. Currently, he is working as Professor in the Department of Computer Science and Engineering, RMK Engineering College, Chennai, India. He is in the teaching profession for the past 22 years and his areas of interest are embedded system, VLSI design, soft computing and bio medical engineering. He has published 30 papers in International Journals and Conferences. He has produced three PhD and currently guiding twelve PhD scholars. He is a Life member in ISTE, IETE, ISOI, BMESI and SSI.

Nirmal kumar completed his postgraduate degree in computer science and engineering from RMK Engineering College, affiliated to Anna University, Chennai, India. He is working as Software developer in Wipro technologies, India. His areas of interest include data mining, artificial intelligence, machine learning and application development.