ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 2, February 2014

A Successive Feature Selection Algorithm for Gene Ranking

T.Revathi[1], Dr.P.Sumathi[2]

Doctoral Research Scholar, Manonmaniam Sundaranar University,Tirunelveli[1]

Assistant Professor, PG & Research Department of Computer Science, Govt.Arts College, Coimbatore[2]

Abstract: Identification and classification of cancer for the gene is most vital. The importance of the each gene is to be found by the gene raking measurement. Modified Successive Feature Selection is used for gene ranking in this paper. Then the Support Vector Machine classifier is trained with that dataset. Genes are collected from the dataset. Many of the feature selection algorithms produced fault for their ranked gene performance. To prevent this, proposed method produces the better accuracy by producing a feature selection algorithm in gene expression data analysis of sample classifications. That the proposed method selects the gene and divides the genes into subset, from the features, gene ranks are selected. From the Lymphoma and Leukemia dataset genes are selected. The proposed method shows promising classification accuracy for the entire test data sets.

Keywords: Successive Feature Selection, Modified Successive Feature Selection, Support Vector Machine, Lymphoma dataset, Leukemia dataset.

I. INTRODUCTION

CANCER research is one of the major research areas in the medical field. Accurate prediction of different tumor types has great value in providing better treatment and toxicity minimization on the patient. Previously, cancer classification has always been morphological and clinical based. These conventional cancer classification methods are reported to have several limitations (Azuaje 2000) in their diagnostic ability. It has been suggested that specification of therapies according to tumor types differentiated by pathogenetic patterns may maximize the efficacy of the patients (Alizadeh 2000). Also, the existing tumor classes have been found to be heterogeneous and comprises of diseases that are molecularly distinct and follow different clinical courses.

Microarray technology has the potential to address many interesting questions in genetics by revealing patterns of expression for genes and classifying samples (such as tumor samples) based on such patterns. However, basic questions about microarray data persist without satisfactory answers. The simplest microarray experiment studies the variation in gene expression across the categories of a single factor, such as tissue types, strains of mice, or drug treatments. The purpose of such an experiment is to identify differences in gene expression among the varieties.

Increasing availability of a variety of gene-related biological data sources and ranging from microarray expression data from protein to protein interaction data then the promising approach is to use bioinformatics methods that can analyze this data and rank genes based on potential relevance to a disease and such methods can being valuable in helping to prioritize genes for further biological study (ShivaniAgarwal and ShiladityaSengupta 2009). In recent time the problem of ranking objects has gained considerable attention in machine learning and data mining and the ranking problems arise in a variety of domains ranging from document retrieval to collaborative filtering and a variety of new learning methods have been developed that directly optimize ranking performance.
Support Vector machine-one-against- all (SVM-OAA) and Linear Discriminate Analysis (LDA) is used as a classifier for performance evaluation. Datasets is randomly divided into two parts, one for training and another part for testing and gene ranking that is ANOVA P-Values can be computed using one-way ANOVA. Top genes were selected from the ranked data and gene combination has been performed.

The classifier is trained using all possible gene combinations and the classifier is validated using 5 fold or 10 fold cross validation methods. The best gene combination can be selected from the result of accuracy. Compared with the previous result obtained by ELM (Zhang et al 2007) SVM OAA attains best accuracy with the use of very few genes than LDA. The same classifier is used on Leukemia and Liver datasets for both the gene selection and classification that improves the strength of the model.

Some of the widely used Gene ranking techniques are T-Score, ANOVA, etc. But those techniques will sometimes wrongly predict the rank when large database is used. To overcome this modification of Successive Feature Selection algorithm is used.

II. RELATED WORK

Microarray technology evolved from Southern blotting, where fragmented DNA is attached to a substrate and then probed with a known DNA sequence. The first reported use of this approach was the analysis of 378 arrayed lysed bacterial colonies each harboring a different sequence which were assayed in multiple replicas for expression of the genes in multiple normal and tumor tissue. This was expanded to analysis of more than 4000 human sequences with computer driven scanning and image processing for quantitative analysis of the sequences in human colonic

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 2, February 2014

tumors and normal tissue and then to comparison of colonic tissues at different genetic risk. The use of a collection of distinct DNAs in arrays for expression profiling was also described in 1987, and the arrayed DNAs were used to identify genes whose expression is modulated by interferon. These early gene arrays were made by spotting cDNAs onto filter paper with a pin-spotting device. The use of miniaturized microarrays for gene expression profiling was first reported in 1995, and a complete eukaryotic genome (Saccharomyces cerevisiae) on a microarray was published in 1997 (Maskos 1992).
(Wang and Gotoh 2009) presented a method for cancer classification using a single gene with the use of micro array gene expression profiling. The gene selection has been made by the use high class-discrimination capability according to their depended degree by the classes. The classifier is developed the foundation of the rules generated by the selection of single genes. The method called rough sets based soft computing could be used for cancer classification with a single gene. Data set such as leukemia, lung cancer and prostate cancer from the website: http://datam.i2r.a-star.edu.sg/data/krbd/. Before do gene selection and classification the data are preprocessed. In the single genetic method the prediction procedure and result are easily understood because this model is based on the rules evaluated with the help of single genes. This model is simple and effective as well as achieved better classification accuracy in all of this data set than multi-gene models.

Machine learning is a bough of Artificial Intelligence (AI) that uses a variety of statistical, probabilistic and optimization systems that permits computers to "learn" from past examples and to detect hard-to-discern patterns from large, noisy or complex data sets. Therefore, machine learning is frequently used in cancer diagnosis and detection. In the research work by (Osareh et al 2010), SVM, K-nearest neighbors and probabilistic neural networks classifiers are jointed with signal-to-noise ratio feature ranking, sequential forward selection-based feature selection and principal component analysis feature extraction to distinguish between the benign and malignant tumors of breast. The overall accuracy for breast cancer diagnosis achieved equal to 98.80% and 96.33% in order that using SVM classifier models against two widely used breast cancer benchmark datasets.

(Rao et al 2007) worked on ANNs and statistical techniques to identify prostate cancers and classify them using metrics call values. (Ziaei et al 2006) presented a system for lymphoma cancer classification where genes were ranked based on their signal to noise (S/N) ratios. They used PCA for more dimensionality reduction. Selected genes were applied to Perceptron neural network for classification. Their study was based on 40 patients and 4026 genes.

III. METHODOLOGY

A. Successive Feature Selection

Successive Feature Selection SFS procedure (SFS) a set of $x \leq 10$ features 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. There could be more than one best subset of features in a given level. A feature is dropped in level 1 that gives four different subsets of features. The best set in level 1 is$\{x_1, x_2, x_4\}$ which is selected for level 2. In a similar way a feature is dropped from the best set of features of level 1 into level 2, which gives three different subsets of features. The best sets in level 2 are $\{x_2, x_4\}$ and $\{x_1, x_2\}$supposing that their classification accuracies are the same and are higher than those of other subsets and the best set in level 3 is$\{x_2\}$.

This process is terminated when all the features are ranked. Two ranked sets are obtained in SFS: namely$R_1 = \{x_2, x_4, x_1, x_3\}andR_2 = \{x_2, x_1, x_4, x_3\}$,

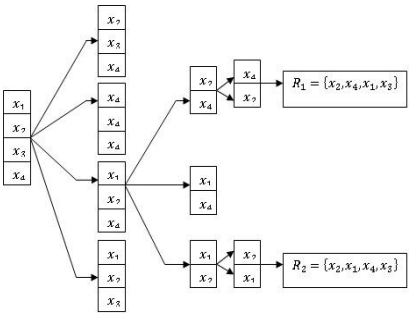

Figure 1 Successive Feature Selection

B.Modified Successive Feature Selection

In the SFS two ranked SFS are obtained, which indicate that $x_2$is the top-ranked feature and that x3 is the bottomrankedor least important feature. Want to select the three top-ranked features, then 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 $F_k = F_1 \cup F_2 = \{x_1, x_2, x_4\}$

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.

The Modified successive feature algorithms is given as

Step1: Find the set of features from $F_1 = \{x_2, x_4, x_1\}$ and$F_2 = \{x_2, x_1, x_4\}$.

Step2: Top three genes are selected by intersection$F_1$and F_2, that is $F_k = F_1 \cup F_2 = \{x_1, x_2, x_4\}$

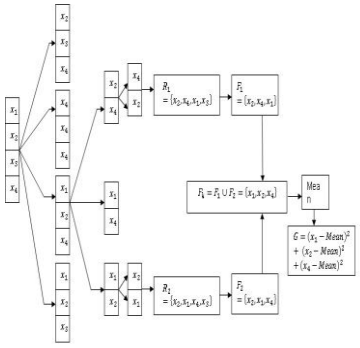

Figure 2 Modification of Successive Feature Selection

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 2, February 2014

Step3: Then the gene rankings are finding out by using three features

Step4: Gene Ranking =

$$G = (x_1 - Mean)^2 + (x_2 - Mean)^2 + (x_4 - Mean)^2$$

Step5: The output of Value G gives the gene ranking. This Modified successive feature selection algorithm provides and investigates the importance gene.

IV. SUPPORT VECTOR MACHIN (SVM)

To test the idea of using the weights of a classifier to produce a feature ranking, we useda state-of-the-art classification technique: Support Vector Machines (SVMs) (Boser et al 1992;Vapnik, 1998). SVMs have recently been intensively studied and benchmarked against a variety of techniques (Guyon 1999). They are presently one of the best-known classification techniques with computational advantages over their contenders(Cristianini 1999) Support vector machine is a common technique in the field of machine learning. The fundamental theory of SVM regression is to map nonlinearly the original data x into a high-dimensional feature space and to provide solution a linear regression problem in this feature space. Let $\{(x_i, y_i)\}_i^m \in X \times \{(\pm 1)\}$, Where x_i represents the input vector, y_i indicates the equivalent output value and m represents the total number of data patterns, the SVM regression function is:

$$f(x) = w \cdot x + b \quad (1)$$

x denotesthe high-dimensionalfeature space, $w$represents the weight vector and $b$indicates the bias term. The coefficients w and b are calculated by reducing the following regularized risk function:

$$R(c) = \frac{1}{2}\|w\|^2 + C\frac{1}{m}\sum_{i=1}^{m} L_\varepsilon(y_i, f(x_i)) \quad (2)$$

Where Crepresents a cost function determining theempirical risk.$\frac{1}{2}\|w\|^2$is the regularization term.$L_\varepsilon(y_i, f(x_i))$is called the e-insensitive loss function,which is given as:

$$L_\varepsilon(y_i, f(x_i)) = \begin{cases} |y_i - f(x_i)| - \varepsilon & |y_i - f(x_i)| \geq \varepsilon \\ 0 & |y_i - f(x_i)| < \varepsilon \end{cases} \quad (3)$$

In (3.3), the e-insensitive loss function equals zero when the error of forecasting value is lower than ε, otherwise the loss equals value ahead of ε. Two positive slack variables ξ and ξ* are established to represent the distance from real values to the equivalent boundary values of the ε-tube. Then, R(C) is changed into the following constrained form:

$$\min \phi(w, \xi, \xi^*) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}(\xi + \xi^*)$$
$$s.t \begin{cases} y_i - w \cdot x - b \leq \varepsilon + \xi \;\; \xi \geq 0 \\ w \cdot x + b - y_i \leq \varepsilon + \xi^* \; \xi^* \geq 0 \end{cases} \quad (4)$$

This constrained optimization difficulty is solved usingthe following Lagrangian form:

$$\max H(\partial, \partial^*) = \sum_{i=1}^{m} y_i(\partial_i - \partial_i^*) - \varepsilon\sum_{i=1}^{m}(\partial_i + \partial_i^*)$$
$$-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\partial_i - \partial_i^*)(\partial_j - \partial_j^*)K(x_i, x_j) \quad (5)$$
$$s.t \sum_{i=1}^{m}(\partial_i - \partial_i^*) = 0 \;\; \partial_i, \partial_i^m \in [0, C]$$

Where $\partial_i$and ∂_i^m are the so-called Lagrangianmultipliers. By the Lagrange multipliers $\partial_i$and ∂_i^mcomputed, an optimal preferred weight vector isacquired, that is:

$$w^* = \sum_{i=1}^{m}(\partial_i - \partial_i^*)K(x_i, x) \quad (6)$$

Therefore, the regression function is:

$$f(x) = \sum_{i=1}^{m}(\partial_i - \partial_i^*)K(x_i, x) + b \quad (7)$$

In accordance with the Karush–Kuhn–Tucker's (KKT)conditions of solving quadratic programming difficulty,the equivalent data points of $\partial_i - \partial_i^* \neq 0$are supportvectors, which are engaged in determining the decisionfunction. SVM built by using Radial Basis Function (RBF)has excellent nonlinear forecasting performance andless free parameters require determination. In (7),

$$K(x_i, x) = \exp(-\|x_i - x\|^2/\sigma^2),$$

C, σ and εparameters are provided by the user, the determination of theparameters plays a significant role in the performance ofSVM(Wun et al 2003).

One of the important features of SVM is that this transformation does not require to be implemented to determine the separating hyperplane in the possibly veryhigh dimensional feature space, a kernel representation can be exploited for determining the separating hyperplane, in which the solution evaluated at the support vectors is written as a weighted sum of the values of certain kernel function.

V. EXPERIMENTAL RESULTS

The innovative of the paper purpose to find the ranking gene with accurate cancer classifications for this SVM classification is selected, it is a sufficiently good classifiers. The proposed methodology was applied to the publicly available cancer datasets namely Lymphoma and Leukemia cancer dataset and the experimented using MATLAB.

(i) 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 leukaemia (CLL). The entire data set contain 4026 genes. In this data set, a small part of data is missing.

(ii) 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.

TABLE 1
DATASET USED IN THE EXPERIMENT

Dataset	Class	Number of Gene	Training samples	Test samples
Lymphoma dataset	3	4026	44	21
Leukemia dataset	2	7129	40	19

Table 1 gives about the dataset. There are two datasets. Lymphoma dataset containing 4026 number of genes. This

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 2, February 2014

is a 3-class classification problem. There are 44 samples for training and 21 samples for testing. Then the Leukemia dataset contains 7129 number of genes and it is a 2 class classification problem and it has 40 samples for training and 19 samples for testing.

TABLE 2
COMPARISON BETWEEN THE METHODS ON THE DATASET

	Successive Feature Selection with SVM		Modified Successive Feature Selection with SVM	
	Time (sec)	Error rate	Time (sec)	Error rate
Lymphoma Dataset	11	0.254	10	0.089
Leukemia Dataset	12	0.266	8	0.095

Table 2 shows the comparison of proposed method with Existing method. It gives the accuracy for the SFS with SVM classification and MSFS with SVM classification used by Lymphoma and Leukemia dataset.

The Table 3 gives the performance analysis for the SFS with SVM and MSFS with SVM used by Lymphoma and Leukemia dataset. That the MSFS with SVM produces the error rate of 0.089 and take the classification time of 10 seconds for the Lymphoma dataset. For the Leukemia dataset classification time of 8 seconds with the error rate of 0.095.

For the existing method of SFS with SVM produce the error rate of 0.25432 for the Lymphoma dataset and 0.26672 for the Leukemia dataset

TABLE 3
PERFORMANCE EVOLUTION

Dataset	Methods	Number of Selected Genes	Classification Accuracy (%)
Lymphoma dataset	SFS with SVM	150	94
	MSFS with SVM	150	97
Leukemia dataset	SFS with SVM	140	93
	MSFS with SVM	140	95

A. Execution Time

The CPU execution time is the execution time taken to complete the process and can be used as a measure to measure efficiency and scalability of the algorithm while using a large dataset

B. Accuracy

Because of the very small sample sizes, we took special care in evaluating the statistical significance of the results. In particular, we address:

1. How accurately the test performance predicts the true classifier performance (measured on an infinitely large test set).
2. With what confidence we can assert that one classifier is better than another when its test performance is better than the other.

Figure 3 shows the comparison of execution time in sec for the SFS with SVM classification and MSFS with SVM classification used by the Lymphoma dataset and Leukemia dataset. By the comparison clearly noticed that the MSFS with SVM classification produce the better result in the reduced time.

Figure 4 shows the comparison of accuracy in percentage for the proposed method of MSFS with SVM classification and the Existing method of SFS with SVM, from the above clearly noticed that the proposed method provides better results by their accuracy in percentage.

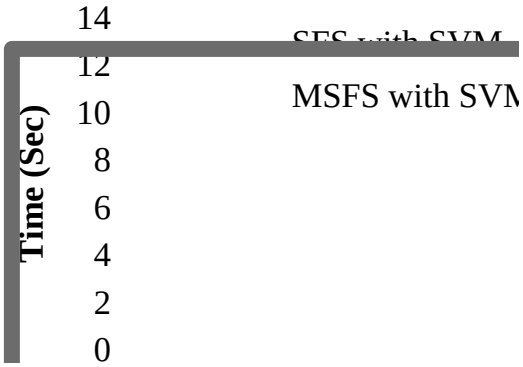

Figure 3 Execution time comparisons

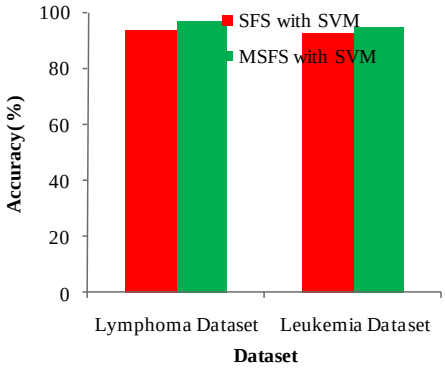

Figure 4 Performance Measurement

VI. CONCLUSION

Cancer is one of the important characteristic in the bio-medicine field. Accurate calculation of several tumor kinds has greater value in offering treatment and toxicity on the patients. Early cancer categorization is generally depends on morphological and clinical examination. These methods used before for cancer classification techniques stated to have many disadvantages in their diagnosis also. Technique of gene ranking is proposed here. Then the classifier is trained with that dataset. The classification of gene for identifying the cancer is been obtained. In this paper proposed a method of Modified Successive Feature

Selection for the gene ranking used by the classification of SVM. That the genes are divided into feature and it is divided into subsets. The proposed algorithm analyzes this occurrence and provides a way to investigate important genes. It is observed that the algorithm finds a small gene subset that provides high classification accuracy on several gene expression data sets. From the results clearly observed that the proposed method of SFS with SVM classification provides the better results.

REFERENCES

[1] Azuaje A. "Interpretation of genome expression patterns: computational challenges and opportunities", IEEE Engineering in Medicine and Biology. (2000).
[2] Alizadeh A. "Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling", Nature. 403:503–511, (2000).
[3] ShivaniAgarwal and ShiladityaSengupta, "Ranking Genes by Relevance to a Disease",Massachusetts Institute of Technology, Cambridge (2009).
[4] Zhang, Huang, G.B., Sundararajan, N. And Saratchandran, P., "Multicategory Classification Using An Extreme Learning Machine For Microarray Gene Expression Cancer Diagnosis", IEEE/ACM Transactions On Computational Biology And Bioinformatics, Vol. 4, No.3, Pp. 485 – 495, 2007
[5] Maskos U, Southern EM. Oligonucleotide hybridizations on glass supports: a novel linker for oligonucleotide synthesis and hybridization properties of oligonucleotides synthesised in situ. Nucleic Acids Res, 20 (7), 1992, 1679–84.
[6] Wang X And Gotoh O, "Cancer Classification Using Single Genes", Genome Informatics, Vol. 23, Pp.179-188, 2009.
[7] Osareh, A.; Shadgar, B.; "Machine learning techniques to diagnose breast cancer", 5th International Symposium on Health Informatics and Bioinformatics (HIBIT), pages 114 – 120, 2010.
[8] Wun-Hwa Chen, Sheng-Hsun Hsu and Hwang-Pin Shen, "Application of SVM and ANN for intrusion detection", Computers & Operations Research, Vol. 32, Pp. 2617-2634, 2003.
[9] K.V. G. Rao, P. P. Chand, M.V.R. Murthy, "A neural Network Approach in Medical Decision Systems" Journal of Theoretical and Applied Information Technology, vol. 3 No. 4, 2007
[10] L. Ziaei, A. R. Mehri, M. Salehi, "Application of Artificial Neural Networks in Cancer Classification and Diagnosis Prediction of a Subtype of Lymphoma Based on Gene Expression Profile," Journal of Research in Medical Sciences, vol. 11, No. 1; Jan. & Feb. 2006.
[11] Boser, B., Guyon, I., & Vapnik, V. (1992). An training algorithm for optimal margin classifiers. InProceedings of the Fifth Annual Workshop on Computational Learning Theory (pp. 144–152). Pittsburgh: ACM.
[12] Vapnik, V. N. (1998). Statistical learning theory. Wiley Interscience.
[13] Guyon, I. (1999). SVM Application Survey: http://www.clopinet.com/SVM.applications.html.
[14] Cristianini, N. &Shawe-Taylor, J. (1999).An introduction to support vector machines. Cambridge, MA: Cambridge University Press
[15] The lymphoma data set (http://llmpp.nih.gov/lymphoma).
[16] http://www.cs.purdue.edu/commugrate/data/Leukemia/