# Analysis of Gene Expression Microarray Dataset for Feature Selection

G. Baskar
PhD Research Scholar
Department of Computer Science
Government Arts College (Autonomous)
Coimbatore, Tamil Nadu, INDIA

P. Ponmuthuramalingam
Associate Professor& Head
Department of Computer Science
Government Arts College (Autonomous)
Coimbatore, Tamil Nadu, INDIA

## ABSTRACT
Microarray is a powerful technology for biological exploration which enables to simultaneously measure the level of activity of thousands genes in various cancer study .clustering is important data mining technique to extract useful information from various high dimensional datasets. A wide range of clustering algorithm is available and still in an open area of research k-Means algorithm is one of the basic and most simple partitioning clustering technique is given by Mac Queen in 1967. In this paper a sample weighting and efficient margin based sample weighting algorithm to improve the stability of feature selection. We proposed a weighted k-means to improve the cluster stability and presented an experimental evaluation of the proposed method, the experiment of microarray dataset show the feature selection algorithm such as SVM-RFE are more stable in gene selection.

## Keywords
Feature selection, Classification, Clustering, Gene expression microarray.

## 1. INTRODUCTION
The identification and validation of molecular biomarkers for cancer diagnosis, is an important problem in cancer genomics. Gene expression microarray data [1] are widely used for identifying candidate genes in various cancer studies. selection of candidate genes in this context can be regarded as a problem of feature selection, Many feature selection methods have been adopted for gene selection from microarray data, and have shown good classification performance of the selected genes [2], [3]. The stability issue of feature selection has recently become a topic of strong interest in both the machine learning and the bioinformatics communities. Data mining is the process of discovering useful information (i.e. patterns) Underlying the data. Powerful techniques are needed to extract patterns from large data because traditional statistical tools are not efficient enough anymore, Clustering is an important data mining technique that puts together

Similar objects into a collection in which the objects exhibit certain degree of similarities.

In this paper we focus on the stability of feature selections method under sample weighting and efficient margin based sample weighting algorithm with weighting k-means algorithm using colon and leukemia dataset. The rest of the paper is organized as follows: In section 2 svm-rfe baseline algorithm, In section 3 Weighting k-means algorithm, In section 4 stability measures, In section 5 margin based sample weighting, In section 6 results and We conclude the result of the algorithm and future work in section 7.

## 2. SVM-RFE BASELINE ALGORITHM
Although much simpler feature selection methods are Available [4], SVM-RFE is chosen as a baseline because it is known to provide state-of-the-art classification performance and widely used in microarray data. Recently, the original algorithm has been improved by several studies such as bootstrapped SVM-RFE, two-stage SVM-RFE, and SVM-RFE combined with MRMR filter [5]. SVMRFE is intrinsically a multivariate feature selection method in the sense that it considers feature interaction while evaluating the relevance of features.

The main process of SVM-RFE is to recursively eliminate features of low weights, using SVM to determine feature weights. Starting from the full set of features, at each iteration, the algorithm trains a linear SVM classifier based on the remaining set of features, ranks features according to the squared values of feature weights in the optimal hyper plane, and eliminates one or more features with the lowest weights. This recursive feature elimination (RFE) process stops until all features have been removed or a desired number of features are reached. We implemented of SVM-RFE on Weka's [6] . As suggested by the authors of SVM-RFE, 10 percent of the remaining features are eliminated at each iteration to speed up the RFE process is chosen as another representative algorithm for margin-based feature selection. It is a simple and efficient feature weighting algorithm which considers all features together in evaluating the relevance of features. The main idea is to weight features according to how well their values distinguish between samples that are similar to each other. Specifically, for a two-class problem, the weight for each feature is determined.

Feature weighting algorithm; it produces feature weights which does not explicitly construct the margin vector for each sample but takes an average of the margins over all samples. Our sample weighting algorithm produces sample weights by explicitly projecting each sample to its margin vector in the margin vector feature space and a successive sample weighting procedure in the margin vector feature space. Our sample weighting algorithm can be used as a pre-processing step for any feature selection algorithms which can be extended to incorporate sample weights.

## 3. WEIGHTING K-MEANS ALGORIHM
Weighted k-means attempts to decompose a set of objects into a set of disjoint clusters, taking into consideration the fact that the numerical attributes of objects in the set often do not come from independent identical normal distribution. Weighted k-means algorithms are iterative and use hill-climbing to find an

optimal solution (clustering), and thus usually converge to a local minimum.

## 4. STABILITY MEASURES

We take a similarity-based approach where the stability of a feature selection method is measured by the average over all pair wise similarity comparisons among all feature subsets (gene signatures) obtained by the same method from different sub samplings of a data set the stability of a feature selection method depends on the specific choice of the similarity measure. Simple measures such as the percentage of overlap or Jaccard index can be applied. These measures tend to produce higher values for larger subsets due to the increased bias of selecting overlapping features by chance. To correct this bias, Kuncheva suggested the use of the Kuncheva index, The Kuncheva index only considers overlapping genes between two gene subsets, without taking into account non overlapping but highly correlated genes which may correspond to coordinated molecular changes. To address this issue, Zhang et al. proposed a measure called percentage of overlapping genes-related, POGR.

We take a similarity-based approach where the stability of a feature selection method is measured by the average over all pair wise similarity comparisons among all feature subsets (gene signatures) obtained by the same method from different sub samplings of a data set the stability of a feature selection method depends on the specific choice of the similarity measure. Simple measures such as the percentage of overlap or Jaccard index can be applied. These measures tend to produce higher values for larger subsets due to the increased bias of selecting overlapping features by chance. To correct this bias, Kuncheva suggested the use of the Kuncheva index, The Kuncheva index only considers overlapping genes between two gene subsets, without taking into account non overlapping but highly correlated genes which may correspond to coordinated molecular changes. To address this issue, Zhang et al. proposed a measure called percentage of overlapping genes-related, POGR.

## 5. MARGIN BASED SAMPLE WEIGHTING

In a recent study, Han and Yu proposed a theoretical framework about stable feature selection which defines the stability of feature selection from a sample variance perspective and shows that the stability of feature selection under training data variations can be improved by variance reduction techniques. The sample weighting framework proposed in this study is motivated by importance sampling, one of the commonly used variance reduction techniques.

The theory of importance sampling suggests that in order to reduce the variance of a Monte Carlo estimator (e.g., the estimate of feature relevance by a feature weighting algorithm based on a training set), instead of performing i.i.d. sampling, we should increase the number of samples taken from regions which contribute more to the quantity of interest and decrease the number of samples taken from other regions. When given only the empirical distribution in a training set, although we cannot redo the sampling process, we can simulate the effect of importance sampling by increasing the weights of samples taken from more important regions and decreasing the weights of those from other regions. Therefore, the problem of variance reduction for feature selection boils down to finding an empirical solution to estimating the importance of samples with respect to feature evaluation and weighting samples.

## 6. RESULTS

**Table 1: Result of the algorithm**

| Data | Selection method | Frequency interval | | |
|---|---|---|---|---|
| | | [1, 100] | [50,100] | [85,100] |
| Colon | SVM-RFE | 642 | 18 | 1 |
| | Weighting K-Means | 463 | 20 | 8 |
| | Sw SVM-RFE | 350 | 36 | 16 |
| leukemia | SVM-RFE | 688 | 18 | 4 |
| | Weighting K-Means | 582 | 24 | 11 |
| | Sw SVM-RFE | 469 | 28 | 14 |

The performance of the SW SVM-REF had good result than weighting k-means algorithm over colon and leukemia.

We used two different cancer datasets to make a study, The Leukemia data set is a collection of gene expression measurements from 72 leukemia (composed of 62 bone marrow and 10 peripheral bloods) samples reported by Golub. It contains an initial initial training set composed of 47 samples of acute lymphoblastic leukemia (ALL) and 25 samples of acute myeloblastic leukemia (AML). The Colon dataset is a collection of gene expression measurements from 62 Colon biopsy samples reported by Alon. It contains 22 normal and 40 Colon cancer samples. The Colon dataset consists of 2000 genes.

## 7. CONCLUSIONS AND FUTURE WORK

The result suggests that sample weighting algorithm improving the stability of feature selection method for gene expression microarray dataset. In future the sample weighting algorithm can be improve using fuzzy logic and other variant of k-means.

## 8. REFERENCES

[1] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander, "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene

Expression Monitoring," Science, vol. 286, pp. 531-537, 1999.

[2] T. Li, C. Zhang, and M. Ogihara, "A Comparative Study of Feature Selection and Multiclass Classification Methods for Tissue Classification Based on Gene Expression," Bioinformatics, vol. 20, pp. 2429-2437, 2004.

[3] Y. Saeys, I. Inza, and P. Larranaga, "A Review of Feature Selection Techniques in Bioinformatics," Bioinformatics, vol. 23, no. 19, pp. 2507-2517, 2007.

[4] H. Liu, J. Li, and L. Wong, "A Comparative Study on Feature Selection and Classification Methods Using

Gene Expression Profiles and Proteomic Patterns," Genome Informatics, vol. 13, pp. 51-60, 2002.

[5] P.A. Mundra and J.C. Rajapakse, "SVM-RFE with MRMR Filter for Gene Selection," IEEE Trans. NanoBioscience, vol. 9, no. 1, pp. 31- 37, Mar. 2010

[6] I.H. Witten and E. Frank, Data Mining - Practical Machine Learning Tools and Techniques. Morgan Kaufmann Publishers, 2005.

[7] B.Y. Rubinstein, Simulation and the Monte Carlo Method. John Wiley & Sons, 1981.

[8] Y. Tang, Y.Q. Zhang, and Z. Huang, "Development Two-Stage SVM-RFE Gene Selection Strategy for Microarray Expression Data Analysis," IEEE/ACM Trans. Computational Biology and Bioinformatics, vol. 4, no. 3, pp. 365-381, July 2007.

[9] Pawan Lingras, Chad West. Interval set Clustering of Web users with Rough k-Means, submitted to the Journal of Intelligent Information System in 2002.

[10] Yeung K.Y, Haynor D.R, Ruzzo W.L. Validating clustering for gene expression data. Bioinformatics. 2001.