



A HYBRID WEIGHTING ALGORITHM FOR FEATURE SELECTION METHODS

G. Baskar*, Dr.P.Ponmuthuramalingam
Ph.D. Research Scholar*, Associate Professor & Head
Department of Computer Science
Government Arts College (Autonomous)
Coimbatore, Tamil Nadu, INDIA

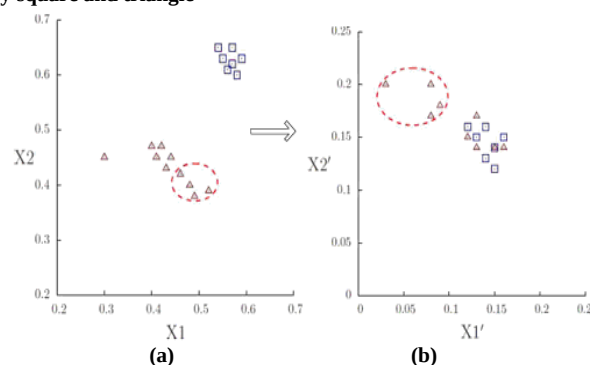
Abstract: Data mining is a system of searching huge amounts of data for patterns. It is a comparatively new perception which is straightly related to computer science. Gene expression microarray data are mostly used for different types of cancer studies, and the stability of gene selection is a popular topic in machine learning. Micro array data is a commonly used method in various cancer studies for candidate genes and the technology is powerful for biological exploration. In this paper, a solution which is based on (SVM-RFE) support vector machines recursive feature elimination and the versions of SVM-RFE, ReliefF is proposed under ensemble, sample weighting and hybrid weighting for micro array data set. A new framework is developed for improving the stability of feature selection algorithm and compares its performances.

Keywords: Micro Array, Colon, SVM-RFE, ReliefF, Hybrid weighting.

1. Introduction

Cancer diagnosis is an important problem in cancer genomics. Due to time, cost the experiment biological validation for high potential biomarker² candidates for validation is critical. For machine learning view point feature selection is a main problem for high dimensional data. Support vector machine (SVM) is an algorithm that attempts to find a linear separator (hyper-plane) between the data points of two classes in multidimensional space. SVMs are well suited to dealing with interactions among features and redundant features. A common problem with existing gene selection method is varying significantly with some variation of sample in the same dataset. Many major factors affect the stability of feature selection²; to overcome this we proposed a hybrid weighting frame work for improving the stability of feature selection methods. Stability has recently become a main interesting topic in both machine learning³ and in bioinformatics.

Figure1. Feature transformation based on hypothesis margin (a) is an original feature space and (b) is a Hypothesis margin feature space. According to the hypothesis margin in an original space is projected by each data in an original feature space. The class label of data is distinguished by square and triangle



$$x'_j = |x_j - x_j^M| - |x_j - x_j^H| \quad (1)$$

Where x'_j is the j th coordinate of x' in the new feature space and x_j , x_j^M , or x_j^H is the j th coordinate of x , x^H , or x^M in the original feature space R^d , respectively. Vector x' is called the margin vector of x , and R^d is called the margin vector feature space, where x^H and x^M are nearest hit and miss to x .

$$x'_j = \sum_{i=1}^m |x_j - x_j^{M_i}| - \sum_{i=1}^h |x_j - x_j^{H_i}| \quad (2)$$

Where x_j^{Ht} or x_j^{Mt} denotes the j th component of the i th neighbor to x with the same or opposite class label, respectively. m or h represents the total number of Misses or Hits ($m + h$ equals the total number of samples in the training set excluding x).

$$W(x) = 1/\sqrt{\text{dist}(x')} / \sum_{i=1}^n 1/\sqrt{\text{dist}(x')} \quad (3)$$

The distance measure by the above equations.

II. SVM-RFE Algorithm

Support Vector machine recursive feature elimination (SVM-RFE) for gene selection approach has been recently attracted many researchers and the method that used for weighting criterion of genes. SVM-RFE was introduced by Guyon et al.⁵. The performance of SVM-RFE become unstable at particular values of the gene filter, the features are eliminated, allowing to a principle related to discrimination function for support, and re-trained SVM at each step.

Algorithm Steps

Step 1 Train an SVM on the training set.

Step 2 Order features via the weights of the resulting classifier.

Step 3 Eliminate features with the lowest weight.

Step 4 Repeat the process with the training set limited to the remaining features.

III. Ensemble SVM-RFE

An Ensemble method is established to improve the stability of feature selection; the most popular bagging ensemble framework is compared with the sample weighting frame work. Then to aggregate the different ranking and the linear aggregation scheme is used to rank summing based on all bootstrapped¹.

IV. Sample Weighting Algorithm

The research is mainly focused on hypothesis margin, the feature selection is based upon the original data set, and each and every sample has been equally weighted. The margin based sample weighting is exploiting the weighting samples, in order to alleviate the effect of training data of feature selection results the margin vector feature space considers each class from one nearest neighbor. To overcome the noise and outliers in the training set of the transformed feature space the multiple nearest neighbors from each class is used. The main tasks are to exploit the discrepancy of margin vectors in this weight sample in the original space of feature selection methods the stability is improve under training data. We used lower weights to samples, the distance computing is involved by both feature space transformation and sample weighting, therefore it is effective and to give more efficient for micro array data.

Sample weighting algorithm flow

```

Calculate and store pair-wise distances among all margins
vectors  $x_i$ 
for  $i = 1$  to  $n$  do
For  $x_i$ , compute its weight according to Eq. (3)
end for
    
```

V. ReliefF Algorithm

A key idea of the original Relief algorithm is to estimate the quality of attributes allowing to how well their values discriminate between instances that are near to each other. For that purpose, given a randomly selected instance R_i , Relief searches for its two nearest neighbors: one of the same class, called nearest hit H , and the further from the different class, called nearest miss M . It updates the quality estimation $W[A]$ for every attributes A contingent on their values for R_i , M , and H . If instances R_i and H have different values of the attribute A then the attribute A separates two instances with the same class which is not desirable so decrease the quality estimation $W[A]$.

Input: For each training instance a vector of attribute value and the class value

Output: the vector W of estimations of the qualities of attributes

VI. Hybrid Weighting Algorithm

To improve the stability of feature selection we present a method based on the concept of boosting and our method evaluated the quality of feature on weighting criterion. The term fraction weighting is used in the margin along with sample weighting to increase the stability of feature selection algorithm, the combination of this hybrid weighting for feature selection is more stable and efficient than sample weighting over micro array data.

Margin Based Hybrid Weighting Algorithm

Input: data $D = \{x_i\}_{i=1}^n$

Output: Weight vector $W = \{w_1, \dots, w_n\}$ for all samples in D

Feature space transformation

for $i = 1$ to n do

```

    for  $j = 1$  to  $n$  do
    For  $x_i$ , compute  $x'_{i,j}$  according to equation
    end for
end for
Hybrid Weighting
Calculate and store pair-wise Euclidean distances among all
margin vectors  $x'_i$ 
    for  $i = 1$  to  $n$  do
    For  $x_i$ ,  $MF_v$  compute its weight according to equation
    end for

```

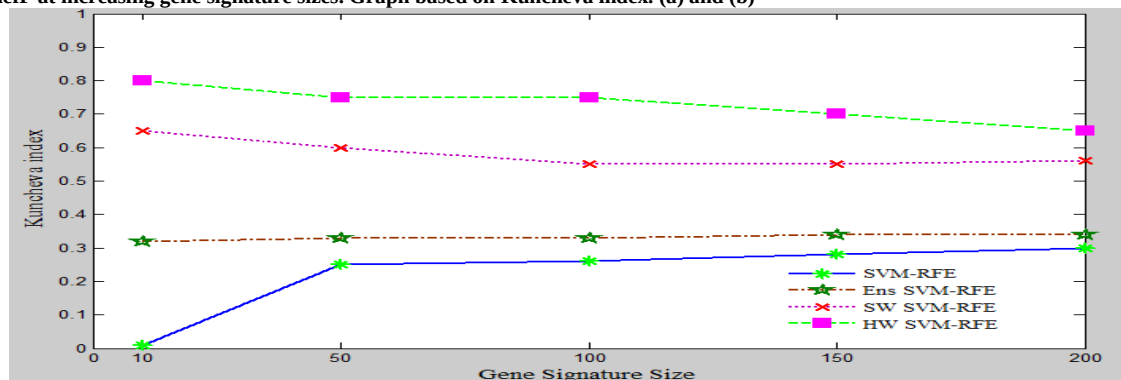
VII. Results

Colon tumor is a disease is found in the tissues of the colon in which cancerous growths (tumors). This dataset covers 62 samples. Among them, 40 tumors are biopsies and are (categorized as "negative") and 22 ordinary (categorized as "positive") biopsies are since healthy parts of the same patients colon. The number of total genes to be tested is 2000. The stability of the algorithm is based on kuncheva index and we can observe that the stability is consistently higher than SVM-RFE under measure. Although ReliefF shows relatively more stable result, Hybrid weighting still consistently improve its stability.

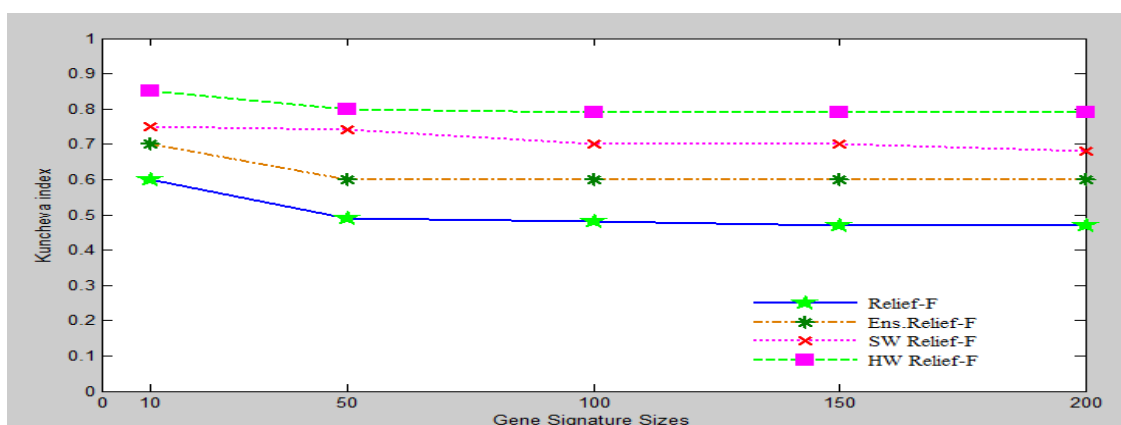
Table1. Classification Performance Measured by the AUC (Average Value $\pm$ Standard Deviation) of the Linear SVM for the Conventional, Ensemble, and Sample Weighting Versions of SVM-RFE and ReliefF at Increasing Gene Signature Sizes

Data	Selection Method	Gene Signature Size				
		10	50	100	150	200
Colon	SVM-RFE	76.4 $\pm$ 9.5	77.5 $\pm$ 8.2	79.2 $\pm$ 8.7	79.4 $\pm$ 8.5	80.1 $\pm$ 8.7
	Ens.SVM-RFE	80.3 $\pm$ 7.9	79.4 $\pm$ 9.0	78.6 $\pm$ 8.3	78.6 $\pm$ 9.1	79.4 $\pm$ 8.7
	SW SVM-RFE	79.5 $\pm$ 9.1	81.2 $\pm$ 8.4	78.4 $\pm$ 10.0	76.2 $\pm$ 10.0	76.2 $\pm$ 9.5
	HW SVM-RFE	78.4 $\pm$ 9.3	80.1 $\pm$ 9.3	78.3 $\pm$ 10.0	75.4 $\pm$ 10.0	75.5 $\pm$ 9.7
	Relief	78.8 $\pm$ 8.8	80.1 $\pm$ 8.8	78.5 $\pm$ 8.7	77.5 $\pm$ 8.9	76.1 $\pm$ 8.5
	Ens.ReliefF	78.9 $\pm$ 8.9	80.2 $\pm$ 9.9	79.1 $\pm$ 9.4	77.3 $\pm$ 9.6	76.1 $\pm$ 9.0
	SW ReliefF	78.3 $\pm$ 8.2	79.6 $\pm$ 9.4	78.1 $\pm$ 9.4	76.4 $\pm$ 10.0	75.4 $\pm$ 10.0
	HW ReliefF	78.1 $\pm$ 8.4	78.5 $\pm$ 9.3	77.5 $\pm$ 9.4	76.1 $\pm$ 10.0	74.5 $\pm$ 10.0

Figure 2. Stability of the conventional, Ensemble, Sample weighting, and Hybrid Weighting for the versions of SVM-RFE and ReliefF at increasing gene signature sizes. Graph based on Kuncheva index. (a) and (b)

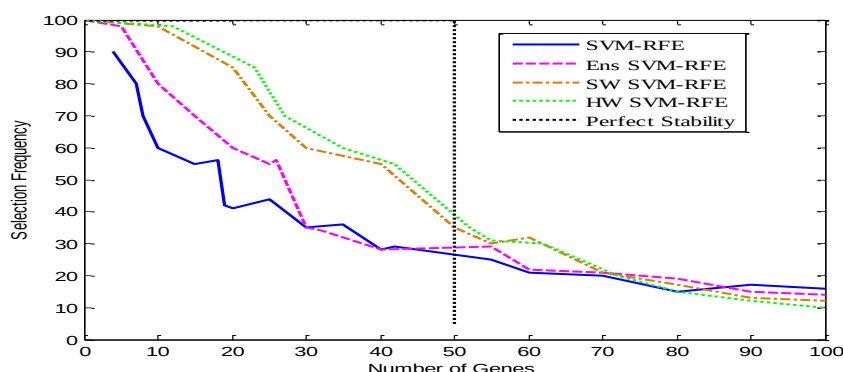


(a)



(b)

Figure 3. Selection frequency plots for colon dataset. Each plot shows how many genes occur in at least how many of the 100 gene signatures of size 50 selected by each version. Only the top 100 most frequently selected genes are included.



VIII. Conclusion and Future Work

This paper studies the stability of feature selection from gene expression microarray data sets. The first contribution of this paper is a general framework of hybrid weighting to improve the stability of existing feature selection methods. The framework weights each sample in training set according to its guidance and feature relevance for assessment, and then delivers the method according to the weighted training feature selection. The second contribution of this paper is the margin-based hybrid weighting algorithm established under the general framework. The algorithm assigns a weight to every sample giving to the outlying degree of its feature relevance to the local profile compared with other samples. Our experimental study based on gene expression data sets has exposed that the margin-based hybrid weighting algorithm is successfully improving the stability of illustrative SVM-RFE and ReliefF algorithms without losing their predictive performance. The results suggest that the general framework of hybrid weighting is a hopeful approach to improving the stability of feature selection methods for gene selection. In the future work, we can add some other weighting algorithm under the general frame work using single nucleoids polymorphism data.

References

- [1] Abeel.T, T. Helleputte, Y.V. Peer, P. Dupont, and Y. Saeys, "Robust Biomarker Identification for Cancer Diagnosis with Ensemble - Feature Selection Methods," *Bioinformatics*, vol. 26, no. 3, pp. 392- 398, 2010 [doi:10.1093/bioinformatics/btp630]
- [2] Boulesteix A.L, and M. Slawski, "Stability and Aggregation of Ranked Gene Lists," *Briefings in Bioinformatics*, vol. 10, no. 5, pp. 556-568, 2009 [doi: 10.1093/bib/bbp034]
- [3] Cortes.C and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, pp. [273-297, 1995. doi: 10.1023/A: 1022627411411]
- [4] Freund.Y and R.E. Schapire, "A Decision Theoretic Generalization of On-Line Learning and an Application to Boosting," *Computer Systems and Science*, vol. 55, no. 1, pp. 119-139, 1997.[doi:10.1007/3-540-59119-2166]
- [5] Guyon.I, J. Weston, S. Barnhill, and V. Vapnik, "Gene Selection for Cancer Classification Using Svm," *Machine Learning*, vol. 46, pp. 389-422, 2002.[doi: 10.1023/A:1012487302797]
- [6] Helleputte.T, and P. Dupont, "Partially Supervised Feature Selection with Regularized Linear Models," *Proc. 26th Int'l Conf. Machine Learning*, pp. 409-416, 2009. [doi:10.1145/1553374.1553427]
- [7] Kuncheva.L, "A Stability Index for Feature Selection," *Proc. 25thInt'l Multi-Conf.: Artificial Intelligence and Applications*, pp. 390-395, 2007. [doi:10.1109/34.574797]
- [8] Pepe M.S, R. Etzioni, Z. Feng, J.D. Potter, M.L. Thompson, M. Thornquist, M. Winget, and Y. Yasui, "Phases of Biomarker Development for Early Detection of Cancer," *J. Nat'l Cancer Inst.*, vol. 93, pp. 1054-1060, 2001. [doi:10.1093/jnci/96.18.1356]
- [9] I.H. Witten and E. Frank, "Data Mining Practical Machine Learning Tools and Techniques", Morgan Kaufmann Publishers, 2005. [doi: 10.1186/1475-925x-5-51]
- [10] J. Ye, J. Chen, R. Janardan, and S. Kumar, "Using Uncorrelated Discriminant Analysis for Tissue Classification with Gene - Expression Data," *IEEE/ACM Trans. Computational Biology and Bioinformatics*, vol. 1, no. 4, pp. 181-190, Oct.-Dec. 2004. [doi:10.1109/TCBB.2004.45]

Contributors



G.Baskar received his Master's degree in Information Technology in K.S.Rangasamy College of Technology, Tiruchengode, Tamil Nadu India in 2008 and M.Phil Degree in Computer Science from Bharathiar University, Coimbatore, Tamil Nadu, India in 2010, and He is currently working towards the PhD degree in Department of Computer Science, Government Arts College, Coimbatore, Tamil Nadu, INDIA in 2011. His area of interest includes Data mining, bioinformatics.



P.Ponmuthuramalingam received his Master Degree in Computer Science from Alagappa University, Karaikudi in 1988 and the Ph.D. in Computer Science from Bharathiar University, Coimbatore. He is working as Associate Professor and Head in Department of Computer Science, Government Arts College (Autonomous), Coimbatore. His research interest includes Text mining, Semantic Web, Network Security and Parallel Algorithm.