



AN EFFICIENT UNIFORM GENE SELECTION FOR MICRO ARRAY DATA VIA HYBRID WEIGHTING

¹G. Baskar, ²Dr.P.Ponmuthuramalingam

¹ PhD Research Scholar, ² Associate Professor & Head

Department of Computer Science
Government Arts College (Autonomous)
Coimbatore, Tamil Nadu, INDIA

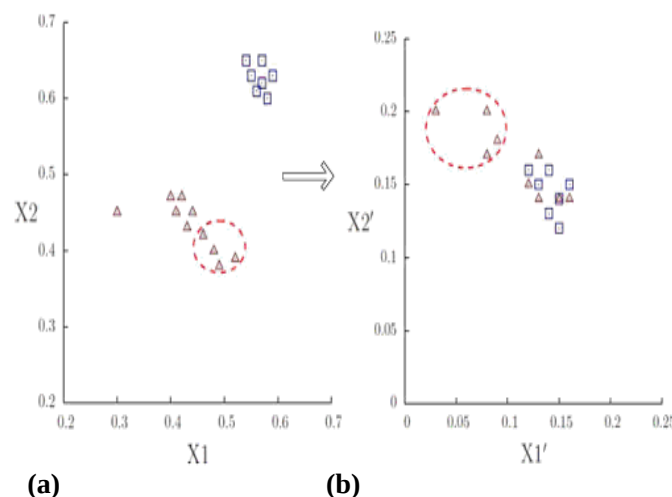
Abstract: Feature Selection is a common standard method used for gene expression micro array data, and an important phase in calculating a selection method is the stability of the certain gene below same data set with different samples. The high performance applications of machine learning algorithms have been enhanced by recent developments. In this paper, a solution which is based on (SVM-RFE) support vector machines recursive feature elimination and the versions of SVM-RFE is proposed under ensemble, sample weighting and hybrid weighting for micro array data set. A new framework is introduced, and we developed the framework for improving the stability of feature selection algorithm and compare its efficiency and effectiveness with representative method.

Keywords: With Feature Selection, Classification, Micro Array, SVM-RFE, Hybrid weighting.

1. Introduction

The identification of cancer identification is a main problem in cancer genomics, due to time consuming and cost it is important to select candidates for validation. Gene expression microarray data are mostly used for different types cancer studies, and the stability of gene selection is a popular topic in machine learning. Some feature selection results are affected by selecting gene, data distribution, and sample sizes. In this paper a framework has been developed, we proposed a hybrid weighting to improve the stability of selection methods.

Figure 1. Feature space 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 a original space is projected by each data in a 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.

The rest of the paper is organized as follows: Section 2. SVM-RFE, Section 3. Ensemble SVM-RFE, Section 4. Sample weighting Algorithm, Section 5. Hybrid weighting Algorithm, Section 6. Results, Section 7 Conclusion and Future work

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[7]. The performance of SVM-RFE become unstable at particular values of the gene filter, the features are eliminated, allowing to a principle related to their support to the discrimination function, and the SVM is re-trained 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 bootstrapped [1].

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 to improve the stability of feature selection methods 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. 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.

VI. Result over Feature Selection Algorithm

The micro array data set here used is a colon(cancer) with 2000 features, 62 samples and the performance is evaluated using Kuncheva index, this index considers over lapping gene between two gene subsets but not

involving non overlapping, and highly correlated genes corresponding to molecular changes. To address this issue we used a measure called percentage of overlapping genes-related nPOGR

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

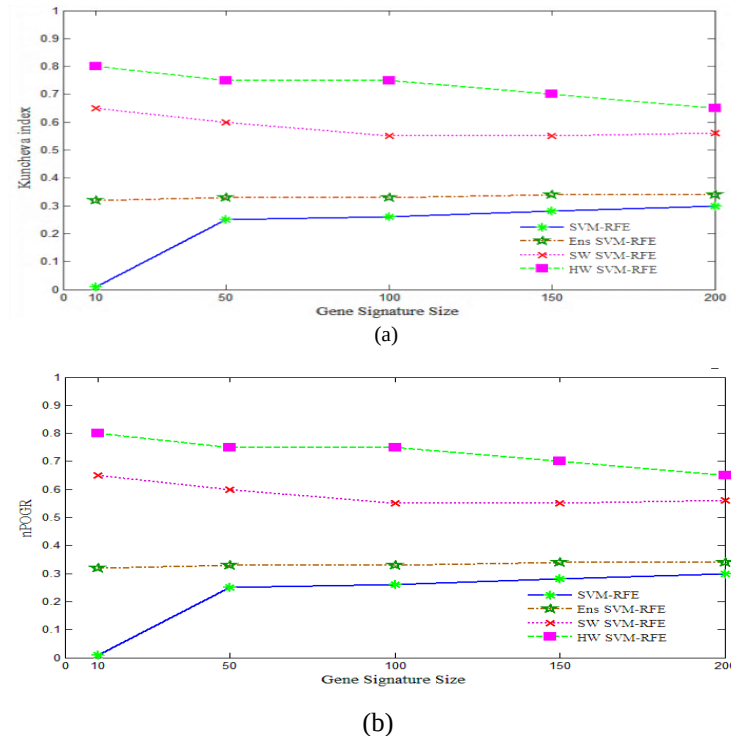


Table 1. The number of Genes Selected Frequencies across 100 Signature of Size.

Data	Selection Method	Frequency Intervals		
		[1,100]	[50,100]	[85,100]
Colon	SVM-RFE	642	16	1
	Ens.SVM-RFE	512	25	7
	SW SVM-RFE	350	36	16
	HW SVM-RFE	280	39	20

VII. Conclusion and Future Work

The research is based on hypothesis margin on feature selection algorithm on gene expression micro array data, Here we proposed the versions of SVM-RFE under ensemble, sample weighting and hybrid weighting. A new framework is introduced that is hybrid weighting, and the method is compared, the results is show in above graph. The SVM-RFE version using hybrid weighting improving the stability of feature selection when compared with sample weighting and Ensemble, and the results are effectiveness with representatives' method. In future we plan to use some other feature selection algorithm.

VIII. 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:10.1007/3-540-59119-2_166]
- [5] Guyon.I, J. Weston, S. Barnhill, and V. Vapnik, "Gene Selection for Cancer Classification Using Support Vector Machines," *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]

Authors profiles



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.Ponnuthuramalingam received his Masters 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



International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0063
ISSN (Online): 2279-0071

International Journal of Software and Web Sciences (IJSWS)

www.iasir.net