Selecting Informative Genes from Microarray Dataset Using Fuzzy Relational Clustering

Soudeh Kasiri-Bidhendi and Saeed Shiry Ghidary

Department of Computer Engineering, Amirkabir University of Technology,
Tehran, Iran
{kasiri,shiry}@aut.ac.ir

Abstract. Selecting informative genes from microarray experiments is one of the most important data analysis steps for deciphering biological information imbedded in such experiments. This paper presents a novel approach for selecting informative genes in two steps. First, fuzzy relational clustering is used to cluster co-expressed genes and select genes that express differently in distinct sample conditions. Second, Support Vector Machine Recursive Feature Elimination (SVM-RFE) method is applied to rank genes. The proposed method is tested on cancer datasets for cancer classification. The results show that the proposed feature selection method selects better subset of genes than the original SVM-RFE does and improves the classification accuracy.

Keywords: Gene selection, microarray data, redundancy reduction.

1 Introduction

The DNA microarray technology is emerging recently in the field of computational biology [1]. Classification based on microarray data faces with many challenges. The main challenge is the overwhelming number of genes compared to the number of available training samples, and many genes are not relevant to the distinction of samples. Gene selection is a process that selects a small subset of genes from the full set, prior to data classification [1]. Gene selection problem can be broadly divided into two categories: Gene ranking and gene-subset evaluation. Gene ranking involves a criterion function for measuring the discriminative power of individual genes. Some examples of criterion function are: TNoM score [2] and Park score [3]. This ranking is simple but it has three drawbacks: (i) there is not any prior knowledge to determine the size of the subset. (ii) Selected genes may be redundant. (iii) Ranking considers only the individual gene discriminative ability and the combined effect of genes is ignored [4].

Combining two low ranked genes may obtain higher discriminative information than combining two high ranked genes. There are attempts to minimize the redundancy [5] by measuring pairwise gene correlation within the selected set. Gene-subset evaluation methods have subsequently been proposed to overcome those drawbacks. The most informative subset must be found by performing greedy search to evaluate all possible gene combinations [6] or stochastic search [7] is often an alternative choice for obtaining a sub-optimal solution. In [8], linear SVMs are used in a backward

H. Sarbazi-Azad et al. (Eds.): CSICC 2008, CCIS 6, pp. 782–785, 2008.

elimination procedure for gene selection, and the selection procedure is referred to as SVM recursive feature elimination (SVM-RFE). Compared with other feature selection methods, SVM-RFE is a scalable, efficient wrappers method.

In this paper fuzzy relational clustering (FRC) is used for redundancy reduction and eliminating genes with similar expressions. Then SVM-RFE is used to rank genes and select the most informative genes. This method is tested on cancer classification tasks based on gene expression data. The remainder of the paper is organized as follows: In section II, proposed method is described. In Section III, numerical experiments on publicly available colon and Leukemia cancer datasets are reported. Finally, conclusion remarks are presented in Section IV.

2 Proposed Method

The first part of proposed method, fuzzy relational clustering, is used to eliminate redundancy of gene expression. Only one gene from each cluster would be selected as informative gene. FRC uses cosine distance, and clusters genes having similar expressions in samples. Then SVM-RFE is used as ranking criterion to select informative genes.

2.1 Fuzzy Relational Clustering

Traditional fuzzy relational clustering (FRC) can be summarized as follows [9]:

- Determine the set of samples to be clustered. Let $X = \{x_1, x_2, ..., x_m\}$ be a set of data where x_i is a $\ltimes N$ vector with real values.
- Establish the fuzzy similarity matrix: to construct the fuzzy similarity matrix R, the first step is to calculate the similarity indices $r_{ij} = R(x_i, x_j)$ of x_i and x_j where r_{ij} can be any arbitrary similarity function.
- Transform fuzzy similarity matrix, R, into a fuzzy equivalence matrix $\mathbf{R}^*$=matrix $(r_{ij}^*)_{i,j=1..m}$: A fuzzy similarity matrix of size $m \times m$ should be composed by itself at most $m-1$ times to be converted to a fuzzy equivalence matrix.
- Calculate λ-cut matrix of. The λ-cut matrix $\mathbf{R}^*$ can be defined as follows:

$$r_{ij}^{\lambda} = \begin{cases} 1 & r_{ij}^* \geq \lambda \\ 0 & r_{ij}^* < \lambda \end{cases} \tag{1}$$

Similar rows of matrix R_{λ}^* form a cluster. With an equivalent matrix and different thresholds, different clustering results will be obtained. However, in the case of microarray with high dimensions, fuzzy equivalence matrix calculation is computationally too complex to be considered practically. This problem can be solved by finding connected components in an undirected graph [10].

2.2 SVM Recursive Feature Elimination

SVM-RFE feature selection method was proposed in [8] to conduct gene selection for cancer classification. 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 of a linear SVM are used to compute the feature ranking score.

3 Experiments

The proposed method is evaluated on Colon and Leukemia cancer dataset. The expression data in raw format are available at [11] and [12] respectively. Classification accuracy of applying proposed and SVM-RFE methods on datasets is shown in Table 1. Fig. 3 shows the test errors of linear SVM classifiers on gene subsets selected respectively by SVM-RFE and proposed method. As shown in table 1 and fig. 3 using fuzzy relational clustering for omitting redundancy of co-expressed genes leads to better performance with fewer genes. The fuzzy relational clustering does not need a pre-determined number of clusters. In particular, by choosing different values for λ, different number of clusters and different cluster shapes may be acquired. Last but not least, it's worth noting that clusters resulted from applying other methods have hyper-spherical shapes which may not be the case in many problems while clusters resulted from applying Fuzzy relational clustering do not have restriction. Therefore, genes that express differently in distinct sample conditions are selected for ranking with SVM-RFE and it improves the classification accuracy with fewer genes.

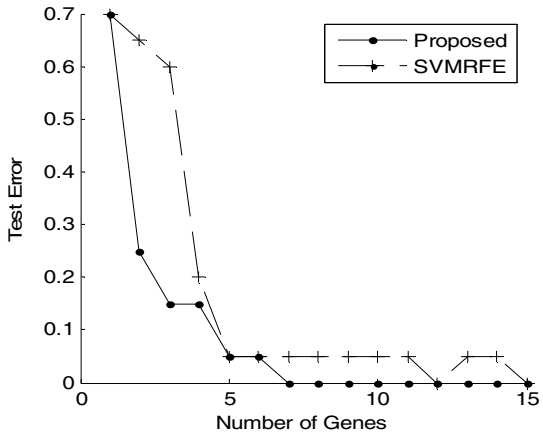

Fig. 1. Performance of feature subsets selected by SVM-RFE and Proposed Method on Colon dataset

Table 1. Results of experiments

Method	Dataset	Number of Genes	Train Accuracy	Test Accuracy
SVM-RFE	Colon	7	100%	95%
Proposed	Colon	7	100%	100%
SVM-RFE	Leukemia	4	100%	97%
Proposed	Leukemia	4	100%	100%

4 Conclusion

Gene selection plays an important role in analysis of microarray datasets. In this paper, fuzzy relational clustering is used to cluster co-expressed genes and select genes expressed differently in distinct sample conditions. Then Support Vector Machine Recursive Feature Elimination method is applied to rank genes. We conclude that the proposed method can select better gene subsets than SVM-RFE and improve the cancer classification accuracy.

References

1. Manfred, N., Laiwan, C.: Informative Gene Discovery for Cancer Classification from Microarray Expression Data. In: IEEE Workshop on Machine Learning for Signal Processing, pp. 393–398 (2005)
2. Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M., Yakhini, Z.: Tissue Classification with Gene Expression Profiles. Journal of Computational Biology 7, 559–584 (2000)
3. Park, P.J., Pagano, M., Bonetti, M.: A Nonparametric Scoring Algorithm for Identifying Informative Genes from Microarray Data. In: Proceedings of the Pacific Symposium on Biocomputing, pp. 52–63 (2001)
4. Guyon, I., Elisseeff, A.: An Introduction to Variable and Feature Selection. Journal of Machine Learning Research (3), 1157–1182 (2003)
5. Wang, M., Wu, P., Xia, S.: Improving Performance of Gene Selection by Unsupervised Learning. In: Proceedings of Neural Networks and Signal Processing, vol. 1, pp. 45–48 (2003)
6. Inza, I., Sierra, B., Blanco, R.: Gene Selection by Sequential Search Wrapper Approaches in Microarray Cancer Class Prediction. Journal of Intelligent and Fuzzy Systems 12, 25–34 (2002)
7. Deutsch, J.M.: Evolutionary Algorithms for Finding Optimal Gene Sets in Microarray Prediction. Bioinformatics 19, 45–52 (2003)
8. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection, for cancer classification using support vector machines. Machine Learning 46, 389–422 (2002)
9. Bojadziev, G., Bojadziev, M.: Fuzzy Sets, Fuzzy Logic, Applications. World Scientific, New Jersey (1995)
10. Dong, Y., Zhuang, Y., Chen, K., Taib, X.: A hierarchical clustering algorithm based on fuzzy graph connectedness. Fuzzy Sets and Systems 157, 1760–1774 (2006)
11. `http://www.broad.mit.edu/cancer`
12. `http://microarray.princeton.edu/oncology`