Gene Selection Using Rough Set Theory[⋆]

Dingfang Li and Wen Zhang

School of Mathematics and Statistics, Wuhan University
Wuhan, 430072, China
dfli@whu.edu.cn, whu_zhangwen@whu.edu.cn

Abstract. The generic approach to cancer classification based on gene expression data is important for accurate cancer diagnosis, instead of using all genes in the dataset, we select a small gene subset out of thousands of genes for classification. Rough set theory is a tool for reducing redundancy in information systems, thus Application of Rough Set to gene selection is interesting. In this paper, a novel gene selection method called RMIMR is proposed for gene selection, which searches for the subset through maximum relevance and maximum positive interaction of genes. Compared with the classical methods based on statistics,information theory and regression, Our method leads to significantly improved classification in experiments on 4 gene expression datasets.

Keywords: Rough sets, gene selection, bioinformatics, classification learning.

1 Introduction

Recent studies on molecular level classification of cancer cells have produced remarkable results, strongly indicating the utility of gene expression data as diagnostic tools [1,2]. A major goal of the analysis of gene expression data is to identify the sets of genes that can serve, via expression profiling assays, as classification or diagnosis platforms [3]. The cancer classification procedure based on gene expression includes the following two steps. First of all, in order to decrease computational complexity and eliminate noisy genes, we have to choose a certain gene selection method [2,3]; secondly, in order to distinguish the tumor samples from the normal ones, we have to construct a fine-work classifier, which can analyze the gene expression data. Though the capability of a classifier is of great significance for the cancer classification, the gene selection method also plays an important role in improving the performance of classifiers [4,5,6,7,8]. Generally speaking, the goal of the gene selection is to select genes as few as possible while achieving better classification performance.

Since rough set theory (RS) was introduced by Pawlak in 1982, there has been great development in both theory and applications [9]. Rough set theory has been applied to feature selection for many years, and some achievements

[⋆] Supported by the National Key Research Program of China (No. 2003CSCA00200) and the National Key Lab Open Research Foundation (No. 2005C012).

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 778–785, 2006.

have been made [10,11,12]. In this paper, a novel gene selection method using rough set is proposed.

The organization of the rest is as follows. In section 2, a gene selection method named RMIMR is described in details; then evaluation experiments and discussion are presented in section 3; finally, conclusions are addressed in section 4.

2 Gene Selection Using Rough Set

Gene expression data can be represented by an matrix. The columns are MRNA/DNA samples, labelled $sample_1, sample_2, ..., sample_m$, rows represent genes, labelled $gene_1, gene_2, ..., gene_n$, where genes are more than samples. To handle the high-dimensional gene expression data, researchers have proposed different methods based on mutual information, statistical tests and regression [4,5,6,7,8]. Those approaches to gene selection fall into two types: filters and wrappers. In filter type, the characteristics in the gene selection are uncorrelated to that of the learning methods. Genes are selected based on the intrinsic characteristics [4], which determine their relevance or discriminant powers with regard to the targeted classes. In wrapper type methods, feature selection is "wrapped" around a learning method: the usefulness of a gene subset is directly judged by the estimated accuracy of the learning method [5]. The method proposed in this paper is of the filter type.

Recently, the rough set is applied to the analysis of genes, and it is usually used as a rule-based learning method[13]. Pawlak pointed out that one of the most important and fundamental roles of the rough sets philosophy is the need to discover redundancy and dependencies between features [9]. Although several methods using RS have been proposed for feature selection on common data sets, they can not be used for gene selection on gene expression data directly. The goal of attribute reduction in RS is reducing the attributes as well as maintaining the consistency of decision tables, which is defined as the power of classification[13]. According to Ron Kohavi's research, the best subset in feature selection may not be a reduct and even does not necessarily contains all core attributes, in fact the reduct may lead to the unfavorable performance when being used to train classifiers [12].

Then main contribution of this paper is that we define relevance of genes and interaction of genes using rough set, and propose the method call RMIMR (Rough Maximum Interaction-Maximum Relevance), which is verified to be effective and useful by analysis and experiments.

2.1 The Principle of Gene Selection

The goal of gene selection is to reduce the computational cost and noises so as to improve the classification accuracy. Therefor which gene should be reduced is the key issue. The common way is to reduce those genes that are irrelevant to the class variable. There have been many attempts to define what is an irrelevant or relevant gene. Dependency of attributes is an important concept in RS, which is used to denote the relativity degree between attributes and decision. In this paper, genes' relevance with respect to class variable is defined based on RS's dependency of attributes, then irrelevant genes can be reduced gradually or relevant genes can be selected.

Definition 1. *(the gene's relevance with respect to the class variable) Gene expression data contains n genes and a class variable D, the gene set is denoted by gene, gene* $= \{gene_1, gene_2, ..., gene_n\}$*, U is the universe of the data,* $|U|$ *denotes the cardinality of U. The relevance of* $gene_i$ *can be written as:*

$$relevance(gene_i) = \frac{|pos_{\{gene_i\}}(D)|}{|U|}, i = 1, 2, ..., n. \tag{1}$$

Definition 1 gives a way to evaluate the relevance of the gene for classification. One common practice of current filter type method is to simply select the top-ranked genes according to the relevance. That is, rank genes according to their relevance, then select the genes with high ranks into the gene subset. This method is simple to be realized, but sometimes it gives bad results [5]. It is frequently observed that simply combining a "very effective" gene with another "very effective" gene often does not form a better feature set. One reason is that these two genes may be highly interacted with each other. Some classifiers, such as Naive-Bayes, are sensitive to the interaction of genes. When the interaction is negative, performance of subset will decline rapidly. This raises the issue of "redundancy" of gene set. Besides the relevance, the interaction of genes must be considered. Based on the concept of dependency of attributes in RS, the interaction of genes can be defined as follows.

Definition 2. *(interaction of genes) Gene expression data contains n genes and a class variable D, the gene set is denoted by gene, gene* $= \{gene_1, gene_2, ...,$ $gene_n\}$*, U is the universe of the data,* $|U|$ *denotes the cardinality of U. Then the interaction of* $gene_i$ *and* $gene_j$ *is defined as:*

$$interaction(gene_i, gene_j) = \frac{|pos_{\{gene_i, gene_j\}}(D)|}{|U|} - \frac{|pos_{\{gene_i\}}(D)|}{|U|} - \frac{|pos_{\{gene_j\}}(D)|}{|U|}. \tag{2}$$

Where $\frac{|pos_{\{gene_i\}}(D)|}{|U|}$, $\frac{|pos_{\{gene_j\}}(D)|}{|U|}$ represents the relevance of $gene_i$ and $gene_j$ for classification, respectively, while $\frac{|pos_{\{gene_i, gene_j\}}(D)|}{|U|}$ represents the relevance of the gene combination. $interaction(gene_i, gene_j)$ reflects the interaction between $gene_i$ and $gene_j$, it can be illustrated as follows:

(1) If $interaction(gene_i, gene_j) > 0$, gene combination is better than the sum of isolated genes, combination has more relevance with class variable, there is positive interaction between $gene_i$ and $gene_j$;

(2) If $interaction(gene_i, gene_j) < 0$, gene combination is worse than the sum of isolated genes, combination has less relevance with class variable, there is negative interaction between $gene_i$ and $gene_j$;

(3) If $interaction(gene_i, gene_j) = 0$, gene combination is equivalent to the sum of isolated genes.

2.2 Gene Selection Using RS

In order to evaluate a gene subset with better generalization property, we should consider two basic rules: one is relevance of genes and the other is the interaction of genes. In the paper, a criterion called Maximum Interaction-Maximum Relevance is used to assess gene subset labelled *geneset*, which means that both relevance of genes and positive interaction of genes are both maximized. The criterion can be written as follows:

$$maxW\ W = \frac{1}{|geneset|} \sum_{gene_i \in\ geneset} relevance(gene_i). \tag{3}$$

$$maxV\ V = \frac{1}{|geneset|^2} \sum_{gene_i, gene_j \in\ geneset} interaction(gene_i, gene_j). \tag{4}$$

V is the average interaction of genes in subset, and W is average relevance of genes in subset, $|geneset|$ is the cardinality of gene subset labelled *geneset*. A well-performed gene subset has both maximum V and maximum W. Since value of interaction is between -1 and 1, for simplicity, we normalize it to [0,1]. Thus Eqs.4 can be amended as Eqs.5.

$$maxV\ V = \frac{1}{2 \times |geneset|^2} \sum_{gene_i, gene_j \in\ geneset} (interaction(gene_i, gene_j) + 1). \tag{5}$$

The maximum interaction-maximum relevance condition is to optimize Eqs.3 and Eqs.5 simultaneously, it can be denoted by Eqs.6.

$$\begin{cases} maxW\ W = \frac{1}{|geneset|} \sum\limits_{gene_i \in\ geneset} relevance(gene_i) \\ maxV\ V = \frac{1}{2 \times |geneset|^2} \sum\limits_{gene_i, gene_j \in\ geneset} (interaction(gene_i, gene_j) + 1). \end{cases} \tag{6}$$

The maximum interaction-maximum relevance gene subset is obtained by optimizing Eqs.6 simultaneously. Optimization of these two conditions requires combining them into a single criterion function. In this paper we treat the two conditions equally important, and consider the simple combined criteria:

$$max(W + V). \tag{7}$$

According to the combined criteria Eqs.7, a method named RMIMR (Rough Maximum Interaction-maximum Relevance) is proposed, it uses a simple heuristic algorithm to resolve the RMIMR optimization problem. The algorithm of RMIMR method is described as follows.

Algorithm 1. RMIMR

```
Data: Gene expression data contains n genes and a class variable, the
      gene set is denoted by $gene = \{gene_1, gene_2, ..., gene_n\}$
Result: Gene subset with s genes labelled $subset$
$subset \leftarrow \emptyset$;
for $i = 1$ to $n$ do
    $relevance(gene_i)$ is calculated according to Eqs.1;
end
for $i = 1$ to $n$ do
    if $relevance(gene_i)$ ranks highest then
        $subset \leftarrow subset + \{gene_i\}$;
        $gene \leftarrow gene - \{gene_i\}$;
        exit for;
    end
end
while s genes are selected do
    for $i = 1$ to $n$ do
        if $gene_i$ satisfys the Eqs.7 and has not been selected then
            $subset \leftarrow subset + \{gene_i\}$;
            $gene \leftarrow gene - \{gene_i\}$;
        end
    end
end
```

3 Experiments and Discussion

In order to evaluate the usefulness of the RMIMR approach, we carried out experiments on four gene expression datasets. The performance of the gene selection is evaluated by training SVM and Naive-bayes.

3.1 Data Sets and Discretization

Two-class datasets Leukemia and colon cancer are used for experiments [1,14], as well as other two multi-class datasets, Leukemia-3 and lung cancer [1,15], the details are listed in Table 1. The data are continuous, we discretize data beforehand. For each attribute, we assume that the mean of its data is μ, and the standard deviation is σ. Any data less than $\mu - \sigma/2$ are transformed to -1, any data between $\mu - \sigma/2$ and $\mu + \sigma/2$ are transformed to 0, any data greater than $\mu + \sigma/2$ are transformed to 1, then three intervals are obtained, meaning that genes are down-regulated, medium-regulated or up-regulated respectively.

3.2 Class Prediction Methods

SVM is a kernel-based learning method proposed by Vapnik, which has been extensively employed as a classification. The naive-Bayes method is a simple

Table 1. Datasets Used in Experiments

Dataset	Leukemia		Colon Cancer		Leukemia-3			Lung			
Gene	7129		2000		7129			1000			
Sample	72		62		72			197			
Name of class	ALL	AML	Tumor	Normal	T-cell	B-cell	AML	AD	NL	SQ	CO
Sample in class	47	25	40	22	9	38	25	131	17	21	20

approach to probabilistic induction that has been successfully applied in a number of machine learning applications

3.3 Results and Discussion

The experiments are carried out in two steps. First of all, a gene subset is selected using RMIMR; then the gene subset is used to train classifiers SVM and Naive-Bayes, and we assess classification performance using the "Leave-One-Out Cross Validation" (LOOCV). In order to demonstrate the advantages of RMIMR, we will compare our classification accuracy with the results presented in [7,8], in that paper information theory methods including MID, MIQ and statistical methods including BASELINE, TCD, TCQ are used, the compare result is plotted in Fig.1 to Fig.4.

In Fig.1 and Fig.2, we compare the RMIMR with TCD and TCQ on datasets Leukemia and colon, the classifier is SVM. In Fig.1, our gene subsets with 4, 6 or 10 genes respectively lead to LOOCV error of zero. In Fig.2, LOOCV error of RMIMR is less than those of other methods for each case, the advantage is obvious. In Fig.3 and Fig.4, we compare RMIMR against Baseline, MID and MIQ on datasets Leukemia and colon, the classifier is Naive-bayes. In fig.3, LOOCV errors of RMIMR is zero using subsets with 6, 9, 15 or 21 genes respectively. In Fig.4, though MIQ has higher accuracy than RMIMR when selecting 6 genes or 9 genes, average performance of RMIMR is better, and RMIMR has much less errors using 21 genes or 24 genes.

Fig.5 and Fig.6 display the results of RMIMR on multi-class datasets Leukemia-3 and Lung, Fig.5 shows the LOOCV error of Naive-bayes, while Fig.6 shows LOOCV error using SVM. In summary, the error rate of RMIRM

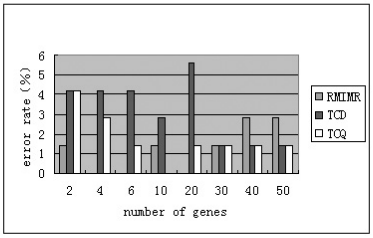

Fig. 1. The Classification Accuracy on Leukemia Data(SVM Classifier)

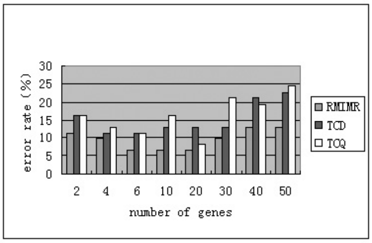

Fig. 2. The Classification Accuracy on Colon Data(SVM Classifier)

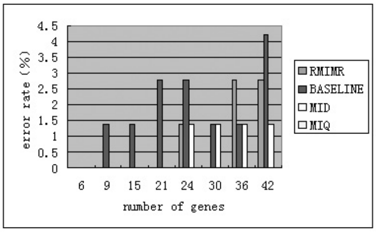

Fig. 3. The Classification Accuracy on Leukemia Data(Naive-bayes Classifier)

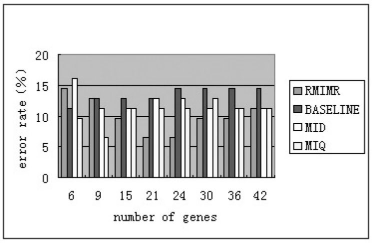

Fig. 4. The Classification Accuracy on Colon Data(Naive-bayes Classifier)

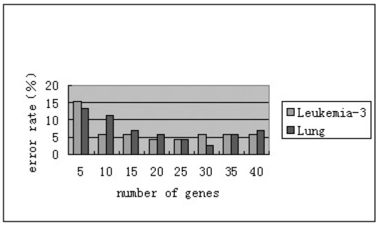

Fig. 5. The Classification Accuracy of RMIMR on Multi-calss Datasets(Naive-bayes classifier)

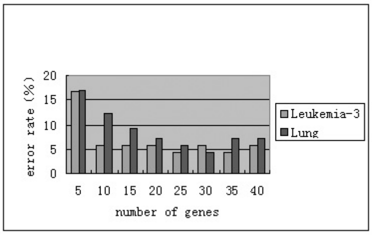

Fig. 6. The Classification Accuracy of RMIMR on Multi-class Datasets(SVM classifier)

on Leukemia-3 are below 6% with 10˜40 genes by training both Naive-bayes and SVM. Compared with the result obtained using partial least squares [8], in which LOOCV error is 4 when 69˜100 genes are selected , our method has fewer LOOCV errors with the small subset.

Experiment results suggest that RMIMR is an effective method for gene selection, it leads to significantly improved cancer diagnosis accuracy, finally it can results in a significant difference in a patient's chances for remission.

4 Conclusion

Because of the high dimension of expression data, selecting a small subset of genes out of the thousands of genes in Microarray is a crucial problem for accurate cancer classification. In this paper we investigated the problem of gene selection using RS, we proposed RMIMR method using RS. According to the analysis and experiments, we have found that our method lead to significantly improved classification accuracy, it is robust and generalized well to unseen data.

References

1. Golub T.R., Slonim D.K. and Tamayo, p., et al.: Classification of Cancer: Class discovery and Class Prediction by Gene Expression Monitoring. *Science*. **286** (1999) 315-333

2. Ben-Dor, A., Bruhm, L. and Friedman, N., et al: Tissue Classification with Gene Expression Profiles.*Computational Biology*, **7** (2000) 559-584
3. Jaeger, J., Sengupta, R., Ruzzo, W.L. : Improved gene selection for classification of microarrays. *Pacific Symposium on Biocomputing*, (2003) 53-64
4. Blum, A., Langley, P.: Selection of relevant features and examples in machine learning. *Artificial Intelligence.* **97** (1997) 245-271
5. Kohavi, R., John, G.H.: Wrappers for feature subset selection. *Artificial Intelligence.* **97** (1997) 273-324
6. Oh, I.S., Lee, J.S., Moon, B.R.: Hybrid genetic algorithms for feature selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* **26** (1982) 1424-1437
7. Ding, C., Peng, H.: Minimum redundancy feature selection from microarray gene expression data.IEEE Computer Society Bioinformatics Conference, (2003) 523-529
8. Nguyen, D.V., Rocke, D.M.: Multi-class cancer classification via partial least squares with gene expression profiles. *Bioinformatics.* **18** (2002) 1216-1226
9. Pawlak, Z.: Rough sets: present state and the future. *Foundations of Computing and Decision Sciences.* **18** (1993) 157-166
10. Mohamed Quafafou, Moussa Boussouf.: Generalized rough sets based feature selection.*Intelligent Data Analysis.* **4** (2000) 3-17
11. Han, J.C., Hu, X.H., Lin, T.Y.: Feature Subset Selection Based on Relative Dependency of Attributes[C]. Rough Sets and Current Trends in Computing: 4th International Conference, Uppsala, Sweden (2004) 176-185
12. Kohavi, R. and Frasca, B.: Useful feature subset and rough set reducts. Proceedings of the Third International Workshop on Rough Sets and Soft Computing, (1994) 310-317
13. Torgeir R.H., Bartosz, W., Andriy, K., Jerzy, T., Jan, K., Krzysztof, F.: Discovering regulatory binding-site modules using rule-based learning. *Genome Research.* **11** (2005) 855-865
14. Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., Levine, A. J.: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Natl. Acad. Sci.. **96** (1998) 6745-6750
15. Stefano, M., Pablo, T., Jill, M., and Todd, G.: A resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning Journal.* **52** (2003) 91-118.