An Efficient Gene Selection Algorithm Based on Tolerance Rough Set Theory

Na Jiao[1,⋆] and Duoqian Miao[2]

[1] Department of Computer Science and Technology, Tongji University, Shanghai 201804, P.R. China
zdx.jn@163.com
[2] Key Laboratory of Embedded System and Service Computing, Ministry of Education of China, Tongji University, Shanghai 201804, P.R. China

Abstract. Gene selection, a key procedure of the discriminant analysis of microarray data, is to select the most informative genes from the whole gene set. Rough set theory is a mathematical tool for further reducing redundancy. One limitation of rough set theory is the lack of effective methods for processing real-valued data. However, most of gene expression data sets are continuous. Discretization methods can result in information loss. This paper investigates an approach combining feature ranking together with feature selection based on tolerance rough set theory. Compared with gene selection algorithm based on rough set theory, the proposed method is more effective for selecting high discriminative genes in cancer classification task.

Keywords: Microarray data, gene selection, feature ranking, tolerance rough set theory, cancer classification.

1 Introduction

DNA microarray is a technology to measure the expression levels of thousands of genes, which is quite suitable for comparing the gene expression levels in tissues under different conditions, such as healthy versus diseased.

Discriminant analysis of microarray data has been widely studied to assist diagnosis. Because lots of genes in the original gene set are irrelevant or even redundant for specific discriminant problem, gene selection is usually introduced to preprocess the original gene set for further analysis.

There are two basic categories of feature selection algorithms, namely filter and wrapper models. Filter methods select feature subsets independently of any learning algorithm and rely on various measures of the general characteristics of the training data. Some statistical tests (t-test, F-test) have been shown to be effective. The idea of these methods is that features are ranked and the top ones or those that satisfy a certain criterion are selected. Wrapper methods use the predictive accuracy of a predetermined learning algorithm to determine the goodness of the selected subsets and are computationally expensive.

⋆ Corresponding author.

H. Sakai et al. (Eds.): RSFDGrC 2009, LNAI 5908, pp. 176–183, 2009.

Features using existing feature selection such as filter and wrapper have redundancy because genes have similar scores in similar pathways. Rough set theory can be used to eliminate such redundancy. Rough set theory [1-6], proposed by Pawlak in 1982, is widely applied in many fields of data mining such as classification and feature selection. However, traditional rough set theory-based methods are restricted to the requirement that all data must be discrete. Existing methods [7] are to discretize the data sets and replace original data values with crisp values. This is often inadequate, as degrees of objects to the descretized values are not considered. Discretization ignores their discrimination. This may cause information loss. A better choice to solve the problem may be the use of tolerance rough set theory.

This paper presents a gene selection method based on tolerance rough set theory. By using tolerance relations, the strict requirement of complete equivalence can be relaxed, and a more flexible approach to subset selection can be developed. The proposed method is comprised two steps. In step 1, we rank all genes with the t-test and select the most promising genes. In step 2, we apply tolerance rough set theory-based method to the selected genes in step 1. The experimental results demonstrate that the proposed algorithm is more effective than gene selection approach based on rough set theory for achieving good classification performance.

2 Preliminaries

2.1 Rough Set Theory

There is a classificatory feature in gene expression data sets. We can formalize the gene expression data set into a decision system.

Definition 1. Decision table.

A decision table is defined as $T = \langle U, C \cup D, V, f \rangle$, where U is a non-empty finite set of objects; C is a set of all condition features (also called conditional attributes) and D is a set of decision features (also called decision attributes); $V = \bigcup_{a \in C \cup D} V_a$, V_a is a set of feature values of feature a; and $f : U \times (C \cup D) \rightarrow V$ is an information function for every $x \in U$, $a \in C \cup D$.

For any $B \subseteq C \cup D$, an equivalence (indiscernibility) relation induced by B on U is defined as Definition 2.

Definition 2. Equivalence relation.

$$IND(B) = \{(x, y) \in U \times U | \forall b \in B, b(x) = b(y)\}. \tag{1}$$

The family of all equivalence classes of $IND(B)$, i.e., the partition induced by B, is given in Definition 3.

Definition 3. Partition.

$$U/IND(B) = \{[x]_B | x \in U\}, \tag{2}$$

where $[x]_B$ is the equivalence class containing x. All the elements in $[x]_B$ are equivalent (indiscernible) with respect to B. Equivalence classes are elementary sets in rough set theory.

For any $X \subseteq U$ and $B \subseteq C$, X could be approximated by the lower and upper approximations.

Definition 4. Lower approximation and upper approximation.

$$\underline{B}X = \{x | [x]_B \subseteq X\}, \tag{3}$$

$$\overline{B}X = \{x | [x]_B \cap X \neq \emptyset\}. \tag{4}$$

Let $B \subseteq C$, the positive region of the partition $U/IND(D)$ with respect to B is defined as Definition 5.

Definition 5. Positive region.

$$POS_B(D) = \cup_{X \in U/IND(D)} \underline{B}X, \tag{5}$$

and it is the set of all samples that can be certainly classified as belonging to blocks of $U/IND(D)$ using B.

By employing the definition of the positive region it is possible to calculate the rough set degree of dependency of a set of features D on B.

Definition 6. Degree of dependency of feature.

$$\gamma_B(D) = |POS_B(D)| / |U|. \tag{6}$$

2.2 T-Test

Feature subset selection is an important step to narrowing down the feature number prior to data mining. We assume that there are two classes of samples in a gene expression data set.

Definition 7. T-test.

The t-value for gene a is expressed by:

$$t(a) = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}, \tag{7}$$

where μ_i and σ_i are the mean and the standard deviation of the expression levels of gene a for i =1, 2. When there are multiple classes of samples, the t-value is typically computed for one class versus all the other classes. The top genes ranked by t-value can be selected for data mining. Feature set so obtained has certain redundancy because genes in similar pathways probably all have very similar score. If several pathways involved in perturbation but one has main influence it is possible to describe this pathway with fewer genes, therefore feature selection based on rough set theory is used to minimize the feature set.

2.3 Gene Selection Algorithm Based on Rough Set Theory

Gene selection algorithm based on rough set theory for gene expression data is composed of t-test and feature selection based on rough set theory. T-test

is helpful for reducing dimensionality. The algorithm without the t-test preprocessing will get worse performance. After feature ranking, top ranked n genes are selected to form the feature set. The values of all continuous features are discretized. Rough set theory-based feature selection method starts with the full set and consecutively deletes one feature at a time until we obtain a reduction.

Algorithm 1. Gene selection algorithm based on rough set theory (GSRS)

(1) Calculate t-value of each gene, select top ranked n genes to form the feature set C
(2) Discretize the feature set C
(3) Set $P = C$
(4) **do**
(5) **for each** $a \in P$
(6) **if** $\gamma_{P-\{a\}}(D) == \gamma_C(D)$
(7) $P = P - \{a\}$
(8) **until** $\gamma_{(P-\{a\})}(D) < \gamma_{(C)}(D)$
(9) **return** P

The loop continues to evaluate in the above manner by deleting conditional features, until the dependency value of the current reduct is less than that of the dataset.

3 Gene Selection Algorithm Based on Tolerance Rough Set Theory

3.1 Similarity Measures

In this approach, suitable similarity measure, given in [2,3], is described in Definition 8.

Definition 8. Similarity measure.

$$S_a(x, y) = 1 - \frac{|a(x) - a(y)|}{|a_{\max} - a_{\min}|}, \tag{8}$$

where $a \in C \cup D$, and $a_{\max}$ and $a_{\min}$ denote the maximum and minimum values respectively for feature a. When considering more than one feature, the defined similarities must be combined to provide a measure of the overall similarity of objects. For a subset of features, B, the overall similarity measure is defined as Definition 9.

Definition 9. Overall similarity measure.

$$(x, y) \in S_{B,\tau} iff \frac{\sum_{a \in B} S_a(x, y)}{|B|} \geq \tau, \tag{9}$$

where τ is a global similarity threshold; it determines the required level of similarity for inclusion within tolerance classes. This framework allows for the specific

case of traditional rough set theory by defining a suitable similarity measure and threshold (τ =1). From this, for any $B \subseteq C \cup D$, $0 < \tau \leq 1$, the so-called tolerance classes that are generated by a given similarity relation for an object are defined as Definition 10.

Definition 10. Similarity relation.

$$S_{B,\tau}(x) = \{y \in U | (x,y) \in S_{B,\tau}\}. \tag{10}$$

For any $X \subseteq U$, $B \subseteq C$ and $0 < \tau \leq 1$, lower and upper approximations are then defined in a similar way to traditional rough set theory.

Definition 11. Modified lower approximation and upper approximation.

$$\underline{B_\tau} X = \{x | S_{B,\tau}(x) \subseteq X\}, \tag{11}$$

$$\overline{B_\tau} X = \{x | S_{B,\tau}(x) \cap X \neq \emptyset\}. \tag{12}$$

The tuple $\left\langle \underline{B_\tau} X, \overline{B_\tau} X \right\rangle$ is called a tolerance-based rough set. Based this, the positive region and the dependency function can be defined as follows.

Let $B \subseteq C$ and $0 < \tau \leq 1$, the positive region is defined as Definition 12.

Definition 12. Modified positive region.

$$POS_{B,\tau}(D) = \cup_{X \in U/S_{D,\tau}} \underline{B_\tau} X. \tag{13}$$

For $B \subseteq C$ and $0 < \tau \leq 1$, the tolerance rough set degree of dependency is given in Definition 13.

Definition 13. Modified degree of dependency of feature.

$$\gamma_{B,\tau}(D) = |POS_{B,\tau}(D)| / |U|. \tag{14}$$

From these definitions, a feature selection method can be formulated that uses the tolerance-based degree of dependency, $\gamma_{B,\tau}(D)$, to gauge the significance of feature subsets.

3.2 Tolerance Rough Set Theory-Based Gene Selection Method

Gene selection algorithm based on tolerance rough set theory for gene expression data combines feature ranking together with feature selection based on tolerance rough set theory. Similarly, t-test can eliminate such redundant genes. T-test is used to feature ranking as the first step and select top ranked n genes to form the feature set. Tolerance rough set theory-based feature selection method can judge every feature and delete the features that are superfluous.

Algorithm 2. Gene selection algorithm based on tolerance rough set theory (GSTRS)

(1) Calculate t-value of each gene, select top ranked n genes to form the feature set C
(2) Set $P = C$

(3) **do**
(4) **for each** $a \in P$
(5) **if** $\gamma_{P-\{a\},\tau}(D) == \gamma_{C,\tau}(D)$
(6) $P = P - \{a\}$
(7) **until** $\gamma_{P-\{a\},\tau}(D) < \gamma_{C,\tau}(D)$
(8) **return** P

The stopping criteria is automatically defined through the use of the dependency measure when the deletion of further features does not result in a decrease in dependency.

3.3 A Simple Example

To illustrate the operation of feature selection algorithm based on tolerance rough set theory, it is applied to a simple example dataset in Table 1, which contains three real-valued conditional features and a crisp-valued decision feature. Set τ =0.8. $C = \{a, b, c\}$. $D = \{d\}$.

Table 1. Example dataset

Objects	a	b	c	d
1	0.3	0.4	0.2	R
2	0.3	1	0.6	A
3	0.4	0.3	0.4	R
4	0.9	0.4	0.7	R
5	0.9	0.7	0.7	A
6	1	0.4	0.7	A

The following tolerance classes are generated:

$U/S_{D,\tau} = \{\{1,3,4\},\{2,5,6\}\}$,
$U/S_{C,\tau} = \{\{1\},\{2\},\{3\},\{5\},\{4,6\}\}$,
$U/S_{C-\{a\},\tau} = U/S_{\{b,c\},\tau} = \{\{1\},\{2\},\{3\},\{5\},\{4,6\}\}$,
$U/S_{C-\{b\},\tau} = U/S_{\{a,c\},\tau} = \{\{1\},\{2\},\{3\},\{5\},\{4,6\}\}$,
$U/S_{C-\{c\},\tau} = U/S_{\{a,b\},\tau} = \{\{1,3\},\{4,6\},\{2\},\{5\}\}$,
$U/S_{C-\{a,b\},\tau} = U/S_{\{c\},\tau} = \{\{1\},\{2\},\{3\},\{4,5,6\}\}$,
$U/S_{C-\{a,c\},\tau} = U/S_{\{b\},\tau} = \{\{1,3,4,6\},\{2\},\{5\}\}$,
$U/S_{C-\{b,c\},\tau} = U/S_{\{a\},\tau} = \{\{1,2,3\},\{4,5,6\}\}$.

Considering feature set , the lower approximations of the decision classes are calculated as follows:

$\underline{C_\tau}\{1,3,4\} = \underline{\{a,b,c\}_\tau}\{1,3,4\} = \{x|S_{\{a,b,c\},\tau}(x) \subseteq \{1,3,4\}\} = \{1,3\}$,
$\underline{C_\tau}\{2,5,6\} = \underline{\{a,b,c\}_\tau}\{2,5,6\} = \{x|S_{\{a,b,c\},\tau}(x) \subseteq \{2,5,6\}\} = \{2,5\}$.

Hence, the positive region can be constructed:

$$POS_{C,\tau}(D) = \cup_{X \in U/S_{D,\tau}} \underline{C_\tau} X = \underline{C_\tau}\{1,3,4\} \cup \underline{C_\tau}\{2,5,6\} = \{1,2,3,5\}.$$

The resulting degree of dependency is:

$$\gamma_{C,\tau}(D) = \frac{|POS_{C,\tau}(D)|}{|U|} = \frac{|\{1,2,3,5\}|}{|\{1,2,3,4,5,6\}|} = \frac{4}{6}.$$

For feature set $C - \{a\}$, the corresponding dependency degree is:

$$\gamma_{C-\{a\},\tau}(D) = \frac{\left|POS_{C-\{a\},\tau}(D)\right|}{|U|} = \frac{|\{1,2,3,5\}|}{|\{1,2,3,4,5,6\}|} = \frac{4}{6},$$
$$\gamma_{C-\{a\},\tau}(D) = \gamma_{\{b,c\},\tau}(D) = \gamma_{C,\tau}(D) = \frac{4}{6}.$$

Feature a is deleted from feature set C. Similarly, the dependency degree of feature set $\{b,c\} - \{b\}$ is:

$$\gamma_{\{b,c\}-\{b\},\tau}(D) = \frac{\left|POS_{\{b,c\}-\{b\},\tau}(D)\right|}{|U|} = \frac{|\{1,2,3\}|}{|\{1,2,3,4,5,6\}|} = \frac{3}{6},$$
$$\gamma_{\{b,c\}-\{b\},\tau}(D) = \frac{3}{6} < \gamma_{C,\tau}(D) = \frac{4}{6}.$$

Therefore, the algorithm terminates and outputs the reduct $\{b,c\}$.

4 Experiments

To evaluate the performance of the proposed algorithm, we applied it to two benchmark gene expression data sets: Lymphoma data set (http://llmpp.nih.gov /lymphoma) and Liver cancer data set (http://genome-www.stanford.edu/hcc/). The Lymphoma data set is a collection of 96 samples. There are 42 B-cell and 54 Other type samples having 4026 genes. The Liver cancer data set is a collection of gene expression measurements from 156 samples and 1648 genes. There are 82 cases of HCCs and 74 cases of nontumor livers.

GSRS and GSTRS are run on the two data sets. Firstly, t-test is employed as a filter on Lymphoma and Liver cancer. The top ranked 50 largest t-test values genes are selected. When there are missing values in data sets, these values are filled with mean values for continuous features and majority values for nominal features [8]. As two data sets are real-valued, for GSRS algorithm, discretization of every feature of the two data sets is Equal Frequency per Interval [7]. For GSTRS algorithm, set τ =0.9. The reduction results are listed in Table 2.

Two factors need to be considered for comparing GSRS and GSTRS. One is the number of selected genes. From Table 2, we can find that the number of selected genes by GSRS is equal to the number of selected genes by GSTRS.

The other considered factor is classification accuracy of the selected genes of two data sets. Two classifiers, C5.0 and KNN, are respectively adopted. As there are a relatively small number of samples, leave-one-out accuracy is adopted. The results are shown in Table 3.

Table 2. Reduction results

Data sets	Genes	Samples	GSRS	GSTRS
Lymphoma	4026	96	7	7
Liver cancer	1648	156	6	6

Table 3. The classification accuracy of two data sets

Data sets	Lymphoma		Liver cancer	
	GSRS	GSTRS	GSRS	GSTRS
KNN	93.5%	94.8%	89.6%	92.5%
C5.0	95.2%	97.4%	91.3%	94.3%

Experimental results show the selected genes by GSTRS have higher classification accuracy than the selected genes by GSRS when we take KNN classifier. While C5.0 classifier is adopted, the classification accuracy of selected genes by GSTRS is highest of all. The reason may be that GSTRS can retain the information hidden in the data.

5 Conclusions

In this paper, we address gene selection of tolerance rough set theory. By constructing an example, we show how the technique works. This paper extends the research of traditional rough set theory and establishes one direction for seeking an efficient algorithm for gene expression data. Our method is applied to the gene selection of cancer classification. Experimental results show its validity.

Acknowledgments. This paper is supported by The National Natural Science Foundation of P.R.China (no. 60475019, 60775036) and The Research Fund for the Doctoral Program of Higher Education (no. 20060247039).

References

1. Pawlak, Z.: Rough Sets. International Journal of Information Computer Science 11(5), 341–356 (1982)
2. Jensen, R., Shen, Q.: Tolerance-Based and Fuzzy-Rough Feature Selection. In: Proceedings of the 16th International Conference on Fuzzy Systems (FUZZ- IEEE 2007), pp. 877–882 (2007)
3. Parthalin, N.M., Shen, Q.: Exploring The Boundary Region of Tolerance Rough Sets for Feature Selection. Pattern Recognition 42, 655–667 (2009)
4. Miao, D.Q., Wang, J.: Information-Based Algorithm for Reduction of Knowledge. In: IEEE International Conference on Intelligent Processing Systems, pp. 1155–1158 (1997)
5. Wang, G.Y.: Rough Set Theory and Knowledge Acquisition. Xi'an Jiaotong University Press (2001) (in Chinese)
6. Li, D.F., Zhang, W.: Gene Selection Using Rough Set Theory. In: Wang, G.-Y., Peters, J.F., Skowron, A., Yao, Y. (eds.) RSKT 2006. LNCS (LNAI), vol. 4062, pp. 778–785. Springer, Heidelberg (2006)
7. Grzymala-Busse, J.W.: Discretization of Numerical Attributes. In: Klsgen, W., Zytkow, J. (eds.) Handbook of Data Mining and Knowledge Discovery, pp. 218–225. Oxford University Press, Oxford (2002)
8. Grzymala-Busse, J.W., Grzymala-Busse, W.J.: Handling Missing Attribute Values. In: Maimon, O., Rokach, L. (eds.) Handbook of Data Mining and Knowledge Discovery, pp. 37–57 (2005)