Gene Selection with Rough Sets for Cancer Classification

Lijun Sun
Dept. of Comp. Sci. & Tech.
Tongji University, Shanghai,
P.R. China
sunlj1028@yahoo.com.cn

Duoqian Miao
Dept. of Comp. Sci. & Tech.
Tongji University, shanghai,
P.R. China
miaoduoqian@163.com

Hongyun Zhang
Dept. of Comp. Sci. & Tech.
Tongji University, Shanghai,
P.R. China
zhanghy586@sina.com

Abstract

A new method combining correlation based clustering and rough sets attribute reduction together for gene selection from gene expression data is proposed. Correlation based clustering is used as a filter to eliminate the redundant attributes, then the minimal reduct of the filtered attribute set is reduced by rough sets . Three different classification algorithms are employed to evaluate the performance of this novel method. High classification accuracies achieved on two public gene expression data sets show that this method is successful for selecting high discriminative genes for classification task. The experimental results indicate that rough sets based method has the potential to become a useful tool in bioinformatics.

Keywords: gene selection, correlation, rough sets , reduction, cancer classification

1. Introduction

The emergence of cDNA microarray technologies makes it possible to record the expression levels of thousands of genes simultaneously. Generally, different cells or a cell under different conditions yield different microarray results, thus comparisons of gene expression data derived from microarray results between normal and cancer cells can provide the important information of cancer diagnosis and treatment. In practice, clustering and classification algorithms are widely adopted to analyze gene expression data[1][2][3][4][5][11][14][15][16], in this paper, we focus on cancer classification using gene expression data, which is a hot topic in recent years and has received general attention by many biological and medical researchers. A reliable and precise classification of tumors based on gene expression data may lead to a more complete understanding of molecular variations among tumors, and hence, to better diagnosis and treatment strategies.

Microarray experiments usually generate large datasets with expression values for thousands of genes but not more than a few dozens of samples, thus very accurate classification of tissue samples in such high dimensional problems is difficult. Among a large amount of genes encoded in the microarray data, only a very small fraction of them are informative for a certain task [1][12][15][16]. How to select the most useful features (genes) for cancer classification is becoming a very challenging task.

Previous work on feature selection from gene expression data can be generally classified as filter and wrapper approaches. (1)Filter type methods are essentially data pre-processing or data filtering methods. Features are selected based on the intrinsic characteristics, which determine their relevance or discriminate powers with regard to the targeted classes. Simple methods based on mutual information [5],

Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007)
0-7695-2874-0/07 $25.00 © 2007 IEEE

statistical tests (t-test, F-test) have been shown to be effective [1][6]. They also have the virtue of being easily and very efficiently computed. In filters, the characteristics in the feature selection are uncorrelated to that of the learning methods, therefore they have better generalization property. (2)In wrapper type methods, feature selection is "wrapped" around a learning method: the usefulness of a feature is directly judged by the estimated accuracy of the learning method. One can often obtain a set with a very small number of non-redundant features, which gives high accuracy, because the characteristics of the features match well with the characteristics of the learning method. GSVM-RFE is reported can find multiple compact cancer-related gene subsets on each of which high leave-one-out validation accuracy can be achieved[7].

The theory of rough sets [8] is a major mathematical tool for managing uncertainty that arises from granularity in the domain of discourse—that is, from the indiscernibility between objects in a set. Rough sets have been applied mainly in mining tasks like classification, clustering and feature selection. A quick search of biological literatures shows that rough sets are still seldom used in bioinformatics. A major obstacle for using rough sets to deal with gene expression data may be the large scale of gene expression data and the comparatively slow computational speed of rough sets algorithms. In this paper, we introduce a combinational method using correlation based clustering and rough sets attribute reduction for feature selection. This paper is organized as follows. The next section gives the background of rough sets. Then, our method is detailed in Section 3. And in Section 4, experimental results are listed. The discussions of these results are given. Finally, the conclusion is drawn in Section 5.

2. Rough sets theory

In rough sets theory, a decision table is denoted by $T = (U, A, C, D)$, where U is universe of discourse, A is a set of primitive features, $C, D \subset A$ are two subsets of features that are called condition and decision features, respectively, where $C \cap D = \Phi$. Rows of the decision table correspond to objects, and columns correspond to attributes [8].

Definition 1 Let $a \in A$, $P \subseteq A$. A binary relation *IND(P)*, called the indiscernibility relation, is defined as the following:

$$IND(P) = \{(x, y) \in U \times U \mid \forall a \in P, a(x) = a(y)\}$$

Let $U / IND(P)$ denotes the family of all equivalence classes of the relation $IND(P)$, $U / IND(P)$ is also a definable partition of the universe induced by P.

Definition 2 An attribute $c \in C$ is a core attribute if

$$Card(U / IND(C - \{c\})) \neq Card(/ IND(C - \{c\} \cup D))$$

Definition 3 An attribute $c \in C$ is a superfluous attribute if

$$Card(U / IND(C - \{c\})) = Card(U / IND(C - \{c\} \cup D))$$

Definition 4 The subset of attributes $R \subseteq C$ is a reduct of attribute C if

$$Card(U / IND(R \cup D)) = Card(U / IND(C \cup D))$$

And $\forall Q \subset R$

$$Card(U / IND(Q \cup D)) \neq Card(U / IND(C \cup D))$$

The goal of rough sets based feature selection is to find a minimal subset R of all features C which has the same discriminate power with the original condition attribute set C, thus R is used instead of C for classification task. In rough sets theory R is called a reduct of C. Reducts obtained in a decision table usually is more than one, generally the reduct with the fewest attributes is optimal. Obtaining all reducts or minimal reducts of a decision table is a NP-hard problem, thus heuristic knowledge deriving from the dependency relationship between condition attributes and decision attributes in a decision table is mainly utilized to assist the attribute reduction. Many methods have been proposed to search for the minimal attribute reducts, which are classified into several categories: 1) positive region [8]; 2) frequency function [9]; 3) information entropy [10]; etc.

3. Rough sets based gene selection method

Our learning problem is to select high discriminate genes for cancer classification from gene expression data. We may formalize this problem as a decision system $T = (U, A, C, D)$, where universe $U = \{x_1, x_2, \ldots\ldots, x_m\}$ is a set of tumors. The conditional attributes set $C = \{g_1, g_2, \ldots\ldots, g_n\}$ contains each gene, the decision attribute $D = \{d\}$ corresponds to class label of each sample. Each attribute $g_i \in C$ is represented by a vector $g_i = \{x_{1,i}, x_{2,i}, \ldots\ldots, x_{m,i}\}$, $i=1,2,\ldots\ldots,n$, where $x_{k,i}$ is the expression level of gene i at sample k, $k=1,2,\ldots\ldots m$.

In thousands of genes many are highly correlated, this "redundancy" will increase the computational cost and at the same time decrease the accuracy of classification. Thus correlation based clustering is applied to decrease the dimensions of gene space as the first step. Correlation coefficient between two genes is defined as

$$d(g_i, g_j) = \frac{\operatorname{cov}(g_i, g_j)}{\operatorname{var}(g_i) * \operatorname{var}(g_j)}$$

$$1 \le i, j \le n, i \ne j$$

where var(·) responds to standard deviation and cov(·) is covariance. Generally, if $d(g_i,g_j)$ is greater then 0.8, g_i and g_j are considered as remarkably linear correlated. In our experiments, $\varepsilon = 0.8$ is used as the threshold. Genes with correlation coefficient greater than the threshold are grouped into one cluster, and each cluster is represented by the gene with minimal information entropy because the higher attribute entropy means the more expected information is needed using the attribute to classify the samples. Given the partition by D, $U/IND(D)$, of U, the entropy based on the partition by $a \in C, U / IND(a)$, of U, is given by

$$E(a) = -\frac{1}{|U|} \sum_{X \in U/IND(D)} \sum_{Y \in U/IND(a)} |X \cap Y| \log_2 \frac{|X \cap Y|}{|Y|}$$

The above operation can be seen as a filter of the original attribute set; reduct is then constructed from the filtered attribute set by adding attributes using information entropy as the heuristic information. The attribute with lowest information entropy will be selected increasingly until reduct is founded. As the next step superfluous attributes are deleted from the reduct to get a minimal reduct. The algorithm is formulated as the following:

1. Initialization

Calculate entropy of each gene $g \in C$, denoted by $E(g)$

2. Correlation based clustering

a) $C1 \leftarrow \phi$

b) For each gene $g_i \in C$, if $d(g_i,g_j) \ge threshold$, then assign gene g_j into $cluster_i$, where $1 \le i, j \le Card(C)$, $i \ne j$, and at the same time delete g_j from C, $C \leftarrow C - \{g_j\}$

c) For each cluster $cluster_i$, select g_1 which satisfied with $E(g_1) = \min\limits_{g \in cluster_i} E(g)$, and assign g_1 to C_1, $C1 = C1 \cup \{g_1\}$

3. Searching for attribute Reduct

a) $RED(C1) \leftarrow \phi$

b) While

$Card(U / IND(C1 \cup D)) \ne Card(U / IND(RED(C1) \cup D))$

if $E(g_1) = \min\limits_{g \in C1 - RED(C1)} E(g)$ then

$$RED(C1) = RED(C1) \cup \{g_1)$$

4. Delete superfluous attributes

$\forall g \in RED(C1)$

if $Card\,(U / IND\,(RED\,(C1) \cup D))$
$= Card\,(U / IND\,(RED\,(C1) - \{g\} \cup D))$

then $RED\,(C1) = RED\,(C1) - \{g\}$

5. Output the minimal reduct $RED(C1)$

Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007)
0-7695-2874-0/07 $25.00 © 2007 IEEE

4. Experimental results

Two well known gene expression data sets: the colon cancer data set and leukemia data set, which are the same data sets used in many publications for gene selection and cancer classification[1][3][12][13], are used to evaluate the performance of our method.

The colon data set consists of 62 samples and 2000 genes, the samples are composed of 40 tumor biopsies collected from tumors and 22 normal biopsies collected from the healthy part of the colons of the same patient, each sample has been preclassified into one of the two classes: 40 normal and 22 cancer. The leukemia data consists of 72 samples and 7129 genes, including 25 AML type of leukemia and 47 ALL type of leukemia, the samples are taken from 63 bone marrow samples and 9 peripheral blood samples.

First, a simple method introduced in [12] is used to discretize the domain of each attribute because rough sets methods require discretization input. Any data larger than $\mu+\sigma/2$ were transformed to state 1; any data between $\mu+\sigma/2$ and $\mu-\sigma/2$ were transformed to state 0; any data smaller than $\mu-\sigma/2$ were transformed to state -1. where σ is standard deviation, μ is mean of a gene. These three states correspond to the over-expression, baseline, and under-expression. Then our method is employed to searching for informative genes for classification. After clustering with correlation coefficient threshold ε =0.8, 1227 genes are left in colon data set and 4991 genes are left in leukemia data sets. As the next step, the filtered data sets undergo rough sets attribute reduct, only 6 genes and 4 genes are left in the reduct of colon data set and leukemia data set respectively. The obtained genes are listed in table1 and table2.

Table 1. **Informative genes found in colon data set**

gene	*Description*
X63629	H.sapiens mRNA for p cadherin.
J05032	Human aspartyl-tRNA synthetase alpha-2 subunit mRNA
H08393	COLLAGEN ALPHA 2(XI) CHAIN (Homo sapiens)
U32519	Human GAP SH3 binding protein mRNA, complete cds.
M76378	Human cysteine-rich protein (CRP) gene, exons 5 and 6.
U09564	Human serine kinase mRNA, complete cds.

Table 2. **Informative genes found in leukemia data set**

gene	*Description*
M84526_at	DF D component of complement (adipsin)
M89957_at	IGB Immunoglobulin-associated beta (B29)
M11722_at	Terminal transferase mRNA
J05243_at	SPTAN1 Spectrin, alpha, non-erythrocytic 1 (alpha-fodrin)

Three different classification algorithms: KNN, C5.0 and Naive Bayes are employed to evaluate classification power of the obtained genes, and LOOCV(leave-one-out cross validation), which is a widely used process for gene classification, is employed to evaluate the performance of classification process. With LOOCV, each object in data set will in turn be the test set, and the left are training set. Thus each sample of the data set will be predicted once by classifier trained with the left samples. All iterations are then averaged to obtain an unbiased number of performance estimates. A summary of the experimental results is shown in table3 and table4, experimental results on entire data sets are also listed.

We acquired 79.0%, 82.25%, 90.32% classification accuracy on colon data set and 93.1%, 94.44%, 94.44% classification accuracy on leukemia data with KNN, Naive Bayes and C5.0 algorithms respectively, which are all compared or partially suprior to

Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007)
0-7695-2874-0/07 $25.00 © 2007 IEEE

Table 3. **Experimental results of LOOCV on colon data**

classifier	classification accuracy for each class		overall classification accuracy	classification accuracy on entire data set
	Tumor	Normal		
KNN	82.94%	72.6%	79.0%	79.0%
Naive Bayes	72.5%	90.90%	82.25%	35.5%
C5.0	95.0%	81.81%	90.32%	82.3%

Table 4. **Experimental results of LOOCV on leukemia data**

classifier	classification accuracy for each class		overall classification accuracy	classification accuracy on entire data set
	ALL	AML		
KNN	95.71%	88.18%	93.1%	82.4%
Naive Bayes	97.87%	88.0%	94.44%	41.2%
C5.0	97.87%	88.0%	94.44%	91.2%

[1][3][12][13].

The above results indicate that our method has successfully achieved its objectives: automatic gene selection for predicting the class of new object. The classification accuracy of leukemia data set is higher than colon data set, the reason maybe because the scale of leukemia data set is larger. In theory, the more information there is about the problem, the more likely for rough sets method to finding informative features. If possible we can get more accurately predicted result through constructing as large data set as possible to train a rough set model.

5. Conclusion

In this paper, a successful gene selection method based on rough sets theory is presented. Correlation based clustering is done first as a preprocessing to eliminate redundancy in gene expression data set, then the minimal reduct of the filtered attribute sets is reduced by rough sets. Two well known public datasets are used to test the performance of this novel method, high prediction accuracies have been achieved through LOOCV, this suggests our method can select informative genes for cancer classification and rough sets approach holds a high potential to become a useful tool in bioinformatics.

Acknowledgement

This research is supported by National Nature Science Foundation of P.R. China (No.60475019).

References

[1] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander, "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring", Science, Vol. 286, No.5439, Oct. 1999, pp. 531–537.

[2] J. Li, H. Liu, J.R. Downing, A.E. Yeoh, and L. Wong, "Simple Rules Underlying Gene Expression Profiles of More Than Six Subtypes of Acute Lymphoblastic Leukemia (ALL) Patients", Bioinformatics, Vol. 19, No.1, 2003, pp. 71–78.

[3] A. Au, K.C.C. Chan, A.K.C. Wong, and Y. Wang, "Attribute Clustering for Grouping, Selection, and Classification of Gene Expression Data", IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol.2, No.2, June 2005, pp. 83-101.

[4] F.D. Smet, NLMM. Pochet, K. Engelen, T.V. Gorp, P.V. Hummelen, K. Marchal, F. Amant, D. Timmerman, B.D. Moor, and I. Vergote, "Predicting the Clinical Behavior of Ovarian Cancer from Gene Expression Profiles", International Journal of Gynecological Cancer, Vol.16, No.s1, Feb.2006, pp.147–151.

[5] Y. Wang, I.V. Tetko, M.A. Hall, E. Frank, A. Facius, K.F.X. Mayer, and H.W. Mewes, "Gene Selection from

Microarray Data for Cancer Classification—A Machine Learning Approach", Computational Biology and Chemistry ,Vol.29, No.1, Feb 2005, pp.37–46.

[6] C. Ding, "Analysis of Gene Expression Profiles: Class Discovery and Leaf Ordering", Proceedings of 6th Annual Conference on Research in Computational Molecular Biology, ACM Press, New York, 2002, pp.127-136.

[7] T. S. Furey, N. Cristianini, N. Duffy, D.W. Bednarski, M. Schummer, and D. Haussler, "Support Vector Machine Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data", Bioinformatics, Vol.16, No.10, 2000, pp. 906–914.

[8] Z. Pawlak, "Rough Set- Theoretical Aspects of Reasoning about Data", Kluwer Academic Publishers, Dorderecht, Boston, London, 1991.

[9] J. Wang, J. Waog, "Reduction Algorithms Based on Discernibly Matrix: The Ordered Attributes Method", Journal of Computer Science And Technology, Vo1.16, No.6, 2002, pp.489-504.

[10] D.Q. Miao, G.R. Hu, "A Heuristic Algorithm for Reduction of Knowledge", Journal of Computer Research and Development. Vol.36, No.6, 1999, pp.681-684.

[11] J. J. Valdes, A.J. Barton, **"**Gene Discovery in Leukemia Revisited: A Computational Intelligence Perspective", Proceedings of the 17th International Conference on Industrial & Engineering Applications of Artificial International Conference & Expert Systems, Springer Verlag, 2004, pp.118-127.

[12] C. Ding, H.C. Peng, "Minimum Redundancy Feature Selection from Microarray Gene Expression Data", Journal of Bioinformatics and Computational Biology, Vol.3, No.2 Apr. 2003, pp.185-205.

[13] A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and Z. Yakhini, "Tissue Classification with Gene Expression Profiles", Proceedings of the 4th Annual International Conference on Computational Molecular Biology (RECOMB), Universal Academy Press, Tokyo, Apr. 2000, pp.54-64.

[14] V.S. Tseng, C.P. Kao, "Efficiently Mining Gene Expression Data via a Novel Parameterless Clustering Method", IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol.2, No.4 , 2005, pp.355-365.

[15] L.P. Wang, C. Feng, and X. Xie, "Accurate Cancer Classification Using Expressions of Very Few Genes", IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol.4, No.1, 2007, pp.40-53.

[16] S.Mitra, Y.Hayashi, "Bioinformatics with Soft Computing", IEEE Transactions on Systems, Man and Cybernetics-Part C: Applications and Reviews, Vol. 36, No.5, 2006, pp.616-635.