Efficient Gene Selection with Rough Sets from Gene Expression Data

Lijun Sun[1], Duoqian Miao[2], and Hongyun Zhang[3]

[1] Department of Computer Science and Technology,
Tongji University, Shanghai, 201804, P.R. China
Sunlj1028@yahoo.com.cn

[2] Department of Computer Science and Technology,
Tongji University, Shanghai, 201804, P.R. China
Miaoduoqian@163.com

[3] Department of Computer Science and Technology,
Tongji University, Shanghai, 201804, P.R. China
Zhanghongyun583@sina.com

Abstract. The main challenge of gene selection from gene expression dataset is to reduce the redundant genes without affecting discernibility between objects. A pipelined approach combining feature ranking together with rough sets attribute reduction for gene selection is proposed. Feature ranking is used to narrow down the gene space as the first step, top ranked genes are selected; the minimal reduct is induced by rough sets to eliminate the redundant attributes. An exploration of this approach on Leukemia gene expression data is conducted and good results are obtained with no preprocessing to the data. The experiment results show that this approach is successful for selecting high discriminative genes for cancer classification task.

Keywords: Gene selection, Feature ranking, Rough sets, Attributes reduction.

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 tumor cells can provide the important information for tumor classification [1]. 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.

Gene expression data set has very unique characteristics that are very different from all the previous data used for classification. Most publicly available gene expression data usually has the following properties:

- high dimensionality: Up to tens of thousands of genes,

G. Wang et al. (Eds.): RSKT 2008, LNAI 5009, pp. 164–171, 2008.

- very small data set size: Not more than a few dozens of samples,
- most genes are not related to tumor classification.

With such a huge attribute space, it is almost certain that very accurate classification of tissue samples is difficult and among a large amount of genes, only a very small fraction of them are informative for classification task [1] [2] [3] [4] [5] [12] [14] [15], thus performing gene selection prior to classification makes help to narrowing down the attribute number and improving classification accuracy and time-complexity of classification algorithms. More importantly, with the ”noise” from the large number of irrelevant genes removed, the biological information hidden within will be less obstructed; this would assist in drug discovery and early tumor discovery. How to select the most useful genes for cancer classification is becoming a very challenging task.

A good number of algorithms have been developed for this purpose [1] [2] [3] [5] [11] [12] [14] [15]; feature-ranking approach is most widely used. In this approach, each feature/attribute is measured for correlation with the class according to some measuring criteria. The features/attributes are ranked and the top ones or those that satisfy a certain criterion are selected. The main characteristic of feature ranking is that it is based on individual feature correlation with respect to class separately. Simple method such as statistical tests (t-test, F-test) has been shown to be effective [1] [6]. This kind of approach also has the virtue of being easily and very efficiently computed.

Feature sets so obtained have certain redundancy because genes in similar pathways probably all have very similar scores and therefore no additional information gain, rough sets attribute reduction can be used to eliminate such redundancy and minimize the feature sets. The theory of rough sets [7] , as 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, has been applied mainly in data mining tasks like classification, clustering and feature selection. Recent years, rough sets theory has been used in gene selection task by some researchers. Evolutionary rough feature selection is employed on three gene expression datasets in [19], not more than 10 genes are selected on each data set while high classification accuracies are obtained; In [20], with the positive region based reduct algorithm, more than 90% of redundant genes are eliminated.

In this paper, we introduce a pipelined method using feature ranking and rough sets attribute reduction for gene 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 conclusions are drawn in Section 5.

2 Rough Sets Based Feature Selection

In rough sets theory, a decision table is denoted by $T = \{U, A)$, where $A = C \cup D$,C is called condition attribute sets,$D = \{d\}$ is decision feature, and U

is universe of discourse. Rows of the decision table correspond to objects, and columns correspond to attributes [7].

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

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

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. Indispensable and Dispensable Attribute. An attribute $c \in C$ is an indispensable attribute if

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

An attribute $c \in C$ is a dispensable attribute if

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

Definition 3. Reduct. The subset of attributes $R \subseteq C$ is a reduct of attribute set C if

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

And $\forall Q \subset R$

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

Definition 4. Core. The set of all indispensable features in C is

$$CORE(C) = \cap RED(C)$$

where $RED(C)$ is the set of all reducts of C with respect to D

The reduct represent the minimal set of non-redundant features that are capable of discerning objects in a decision table. An optimal feature subset selection based on the rough set theory can be viewed as finding such a reduct R,$R \subseteq C$ with the best classifying properties.R,instead of C , will be used in a rule discovery algorithm. It is obvious that all of indispensable features in core cannot be deleted from C without losing the accuracy of a decision table; the feature(s) in core must be the member of feature subsets. Therefore, the problem of feature subset selection will become how to select the features from dispensable features for forming the best reduct with core. Obtaining all 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 attribute reducts, which are classified into several categories: 1) positive region [7]; 2) frequency function [8]; 3) information entropy [9]; 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, where universe $U = \{x_1, x_2, ..., x_m\}$ is a set of tumors, the conditional attributes set $C = \{g_1, g_2, ..., g_n\}$ contains each gene; the decision attribute d corresponds to class label of each sample. Each attribute $g_i \in C$ is represented by a vector $g_i = \{x_{1,i}, x_{2,i}, ..., x_{m,i}\}$, $i = 1, 2, ..., n$, where $x_{k,i}$ is the expression level of gene i at sample k, $k = 1, 2, ..., m$.

To select genes, t-test is widely used in the literature [1] [6]. Assuming that there are two classes of samples in a gene expression data set, the t-value for gene g is given by:

$$t = \frac{\mu_1 - \mu_2}{\sqrt{\sigma^2/n_1 + \sigma^2/n_2}} \tag{1}$$

Where μ_i and σ_i are the mean and the standard deviation of the expression levels of a gene g for class i respectively, and n_i is the number of samples in class i 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 then be selected for data mining. Feature sets so obtained have certain redundancy because genes in similar pathways probably all have very similar score and therefore no additional information gain. If several pathways involved in perturbation but one has main influence it is possible to describe this pathway with fewer genes, therefore Rough Sets attribute reduction is used to minimize the feature sets.

Reduct is constructed from core because it represents the set of indispensable features, thus all attributes in core must be in the reduct, then we adding attributes using information entropy as the heuristic information until a reduct is find. The attribute with lowest information entropy will be selected first 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 $c \in C$, $U/IND(c)$, of U, is given by

$$E(c) = -\frac{1}{|U|} \sum_{X \in U/IND(D)} \sum_{Y \in U/IND(c)} |X \cap Y| \log_2 \frac{|X \cap Y|}{|Y|} \tag{2}$$

We can formulate our method as the following steps:

1. Calculate t-value of each gene, select the top ranked n genes to form the attribute set C.
2. Calculate core attribute sets of C using Discernibly Matrix [8],denoted by $CORE(C)$.
3. Calculate the reduct of C using information entropy as the heuristic information. Let $RED(C) \leftarrow CORE(C)$, while $Card(U/IND(C \cup D)) \neq Card(U/IND(RED(C) \cup D))$, we calculate information entropy of each gene $g \in C - RED(C)$, denoted by $E(g)$, if $E(g_1) = \min\limits_{g \in C - RED(C)} E(g)$ then we assign g_1 to $RED(C)$. Repeat the above operation until we find a reduct of C.

4 Experimental Results

A well known gene expression data sets, leukemia data set of Golub et al. (1999), which is the same data sets used in many publications for gene selection and cancer classification [1] [3] [5] [11] [12] [19] [20], is used to evaluate the performance of our method. The acute leukemia dataset (http://www.genome.wi.mit.edu/MPR) consists of 38 samples including 27 cases of acute lymphoblastic leukemia (ALL) and 11 cases of acute myeloid leukemia (AML). The gene expression measurements were taken from high-density oligonucleotide microarrays containing 7129 genes. An independent test set of 20 ALL and 14 AML samples also exists.

First t-test is employed as a filter on the training set; the top ranked 50 genes are selected. Then entropy based discretization introduced in [16] is used to discretize the domain of each attribute because rough sets methods require discrete input. Entropy based attribute reduction algorithm is employed on the data set to find a minimal reduction. As the result, X95735 is the only gene to be selected in the reduction. A box plot of X95735 expression levels in the training set is presented in Fig. 1. This figure clearly indicates that the expression levels of X95735 can be used to distinguish ALL from AML in the training set.

Two rules are induced by Rough sets: if the expression level of X95735 $\geqslant$ 938 then the sample is classified as AML; If the expression level of X95735 <938 then the sample is classified as ALL. With the simples rules induced by Rough sets, 31 test samples are correctly classified; there are only three mistakes, one for AML, and two for ALL.

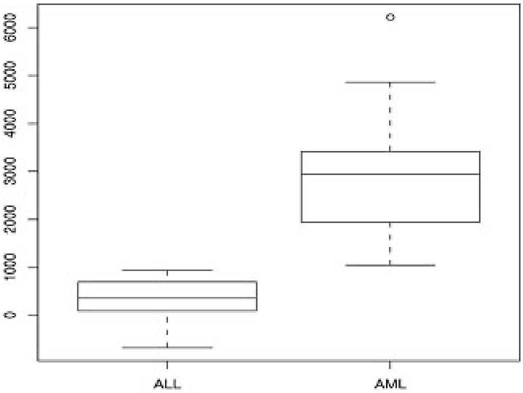

Fig. 1. Expression Levels of X95735 in Training Set

It is interesting that X95735 is also selected by many other methods. It is reported in [5] that X95735 is the only gene identified by J48 pruned tree and the emerging patterns algorithm, and X95735 is also selected by voting machine [1], SVM [10], Deb's NSGA-algorithm [21] and Cho's work [22]. An approach using clustering in combination with Rough Sets and neural networks has been investigated in [11], X95735 is repeated selected, and the classification accuracy is 91.2% on test data set. For comparison, the feature selection and classification results obtained by our method and some results in previous publishers are shown in table 1.

Table 1. The Comparison of Feature Selection and Classification Results

Method	Number of features	Classification Results
Rough sets	1	31
J48	1	31
Emerging Patterns	1	31
SVM	7	34
NSGA-II	3	34

The results obtained by us suggest that the expression level of X95735 plays an important role in distinguishing two types of acute leukemia. Role of X95735 in discerning between two types of acute leukemia samples is also verified by biological researchers [17] [18].

5 Conclusions

Gene expression data set usually has thousands of genes while a few dozens of samples, among a large amount of genes, only a very small fraction of them are informative for classification task. In order to achieve good classification performance, and obtain more useful insight about the biological related issues in cancer classification, gene selection should be well explored to reduce the noise and avoid overfitting of classification algorithm.

In this paper, a successful gene selection method based on rough sets theory is presented. Filter kind of method is done first as a preprocessing to select top ranked genes; the minimal reduct of the filtered attribute sets is induced by rough sets. Acute leukemia gene expression dataset is used to test the performance of this novel method; only one gene X95735 is selected, and high prediction accuracies have been achieved on the test data set. Gene X95735 is also selected by many other methods, and has been verified by biological researchers to play an important role in distinguish two different types of acute leukemia, AML and ALL.

Acknowledgements

This research support by The National Natural Science Foundation of China (Granted No. 60475019, No. 60775036) and The Research Fund for the Doctoral Program of Higher Education (Granted No. 20060247039).

References

1. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.S.: Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science 286, 531–537 (1999)
2. Wang, L.P., Feng, C., Xie, X.: Accurate Cancer Classification Using Expressions of Very Few Genes. IEEE/ACM Transactions on Computational Biology and Bioinformatics 4, 40–53 (2007)
3. Au, A., Chan, K.C.C., Wong, A.K.C., Wang, Y.: Attribute Clustering for Grouping, Selection, and Classification of Gene Expression Data. IEEE/ACM Transactions on Computational Biology and Bioinformatics 2, 83–101 (2005)
4. Smet, F.D., Pochet, N.L.M.M., Engelen, K., Gorp, T.V., Hummelen, P.V., Marchal, K., Amant, F., Timmerman, D., Moor, B.D., Vergote, I.: Predicting the Clinical Behavior of Ovarian Cancer from Gene Expression Profiles. International Journal of Gynecological Cancer 16, 147–151 (2006)
5. Wang, Y., Tetko, I.V., Hall, M.A., Frank, E., Facius, A., Mayer, K.F.X., Mewes, H.W.: Gene Selection from Microarray Data for Cancer Classification-A Machine Learning Approach. Computational Biology and Chemistry 29, 37–46 (2005)
6. Ding, C.: Analysis of Gene Expression Profiles: Class Discovery and Leaf Ordering. In: 6th Annual Conference on Research in Computational Molecular Biology, pp. 127–136. ACM Press, New York (2002)
7. Pawlak, Z.: Rough Set- Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, Dorderecht (1991)
8. Wang, J., Waog, J.: Reduction Algorithms Based on Discernibly Matrix: The Ordered Attributes Method. Journal of Computer Science And Technology 16, 489–504 (2002)
9. Miao, D.Q., Hu, G.R.: A Heuristic Algorithm for Reduction of Knowledge. Journal of Computer Research and Development 36, 681–684 (1999)
10. Furey, T.S., Cristianini, N., Duffy, N., Bednarski, D.W., Schummer, M., Haussler, D.: Support Vector Machine Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data. Bioinformatics 16, 906–914 (2000)
11. Valdes, J.J., Barton, A.J.: Gene Discovery in Leukemia Revisited: A Computational Intelligence Perspective. In: Orchard, B., Yang, C., Ali, M. (eds.) IEA/AIE 2004. LNCS (LNAI), vol. 3029, pp. 118–127. Springer, Heidelberg (2004)
12. Ding, C., Peng, H.C.: Minimum Redundancy Feature Selection from Microarray Gene Expression Data. Journal of Bioinformatics and Computational Biology 3, 185–205 (2003)
13. Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M., Yakhini, Z.: Tissue Classification with Gene Expression Profiles. In: 4th Annual International Conference on Computational Molecular Biology (RECOMB), pp. 54–64. Universal Academy Press, Tokyo (2000)
14. Tseng, V.S., Kao, C.P.: Efficiently Mining Gene Expression Data via a Novel Parameterless Clustering Method. IEEE/ACM Transactions on Computational Biology and Bioinformatics 2, 355–365 (2005)
15. Mitra, S., Hayashi, Y.: Bioinformatics with Soft Computing. IEEE Transactions on Systems, Man and Cybernetics-Part C: Applications and Reviews 36, 616–635 (2006)

16. Fayyad, U.M., Irani, K.B.: Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning. In: Proceedings of the 13th International Joint Conference of Artificial Intelligence, pp. 1022–1027. Morgan Kaufmann, Chambery, France (1993)
17. Van, D.G.E., Leccia, M., Dekker, S., Jalbert, N., Amodeo, D., Byers, H.: Role of Zyxin in Differential Cell Spreading and Proliferation of Melanoma Cells and Melanocytes. J. Invest. Dermatol. 118, 246–254 (2002)
18. Yagi, T., Morimoto, A., Eguchi, M., Hibi, S., Sako, M., Ishii, E., Mizutani, S., Imashuku, S., Ohki, M., Ichikawa, H.: Identification of a Gene Expression Signature Associated with Pediatric AML Prognosis. Blood 102, 1849–1856 (2003)
19. Banerjee, M., Mitra, S., Banka, H.: Evolutinary-Rough Feature Selection in Gene Expression Data. IEEE Transaction on Systems, Man, and Cybernetics, Part C: Application and Reviews 37, 622–632 (2007)
20. Momin, B.F., Mitra, S., Datta Gupta, R.: Reduct Generation and Classification of Gene Expression Data. In: Proceeding of First International Conference on Hybrid Information Technology (ICHICT 2006), pp. 699–708. IEEE Press, New York (2006)
21. Deb, K., Reddy, A.R.: Reliable Classification of Two Class Cancer Data Using Evolutionary Algorithms. BioSystems 72, 111–129 (2003)
22. Cho, S.B., Ryu, J.: Classification Gene Expression Data of Cancer Using Classifier Ensemble with Mutually Exclusive Features. In: Proceedings of the IEEE, Special Issue on Bioinformatics Part-I: Advances and Challenges, pp. 1744–1753. IEEE Press, New York (2002)