Feature Selection for Microarray Data Analysis Using Mutual Information and Rough Set Theory

Wengang Zhou, Chunguang Zhou, Guixia Liu, Hong Zhu
College of Computer Science and Technology, Jilin University,
Changchun 130012, P. R. China
wgzhou@email.jlu.edu.cn

Abstract. Cancer classification is one major application of microarray data analysis. Due to the ultra high dimension of gene expression data, efficient feature selection methods are in great needs for selecting a small number of informative genes. In this paper, we propose a novel feature selection method based on mutual information and rough set (MIRS). First, we select some top-ranked features which have higher mutual information with the target class to predict. Then rough set theory is applied to remove the redundancy among these selected genes. Binary particle swarm optimization (BPSO) is first proposed for attribute reduction in rough set. Finally, the effectiveness of the proposed method is evaluated by the classification accuracy of SVM classifier. Experi-ment results show that MIRS is superior to some other classical feature selec-tion methods and can get higher prediction accuracy with small number of fea-tures. Generally, the results are highly promising.

1 Introduction

The development of microarray technology has made it easy to monitor the expression pattern of thousands of genes simultaneously and a huge amount of gene expression data has been produced during microarray experiments. These data has widely been applied to accurate prediction and diagnosis of cancer. Especially cancer classification [1] is an important issue because it can identify many genes relevant to cancer. The results reported in the literature have confirmed the effectiveness of mining cancer information from gene expression data. But microarray data often consists of small number of samples and large number of genes. The ultra high dimension of gene expression data makes it necessary to develop effective feature selection methods in order to reduce the computation cost and improve the classification accuracy.

Please use the following format when citing this chapter:
Zhou, Wengang, Zhou, Chunguang, Liu, Guixia, Wang, Yan, 2006, in IFIP International Federation for Information Processing, Volume 204, Artificial Intelligence Applications and Innovations, eds. Maglogiannis, I., Karpouzis, K., Bramer, M., (Boston: Springer), pp. 492–499

There are two general approaches to feature selection: filters [2] and wrappers [3]. In a filter method, features are selected based on the intrinsic characteristics which determine their relevance with the target classes. In wrapper type methods, the usefulness of a feature is directly judged by the estimated accuracy of a learning method and typically requires huge computational effort. Thus, it is difficult for wrappers to deal with large feature sets such as gene expression data. We mainly focus on the selection of a few tens features among several thousands by developing an efficient filter methods for cancer classification. When a small number of informative genes are selected, their biological relationship with the target disease can easily be identified.

Mutual information [4] has recently been proposed for feature selection. One common practice to use this method is to simply select the top-ranked genes with higher mutual information. But a deficiency of this simple ranking approach is that the features can be correlated among themselves. If gene g_i is ranked high for classification task, other genes highly correlated with gene g_i are also likely to be selected. This raises the issue of redundancy in feature set. Rough sets theory provides a feasible way to deal with redundancy [5]. An important concept is reduct in rough sets theory. Reduct is those minimal attribute sets of information system, which keep the same classify capability with original attribute set. The aim of reduction is to find out a minimum set of relevant attributes (features) that describe the dataset as well as all the original attributes do. Thus finding reduct can select the most relevant genes with the target class to predict and remove the redundancy among the selected features.

In this paper, we propose a novel feature selection method so called MIRS by integrating mutual information and rough set theory. First, mutual information is used to select some top-ranked genes which have higher mutual information from each data set. Then rough set theory is applied to remove the redundancy among these selected genes. Binary particle swarm optimization (BPSO) is first suggested as an attribute reduction algorithm for rough sets. Finally, the effectiveness of MIRS is evaluated by the classification accuracy of SVM classifiers. Experiment results show that the pro-posed method is superior to some other classical feature selection methods and can always get higher classification accuracy with fewer features.

2 Mutual Information for Feature Selection

In accordance with Shannon's information theory [6], the uncertainty of a random variable Y can be measured by the entropy $H(Y)$. For two variables X and Y, the conditional entropy $H(Y \mid X)$ measures the remaining uncertainty of Y when X is known. The mutual information (MI) $I(X;Y)$ measures the certainty about Y that is resolved by X. Apparently, the relation of $H(Y), H(Y \mid X)$ and $I(X;Y)$ is as follows:

$$I(X;Y) = H(Y) - H(Y \mid X) \tag{1}$$

The objective of training a classification model is to minimize the uncertainty about predictions on class labels Y for the known observations X. Thus, it is

equivalent to increase the MI $I(X;Y)$ as much as possible for training a classifier. The goal of feature selection process for classification is naturally to achieve the smallest subset of possible features which have higher values of $I(X;Y)$. With the entropy defined by Shannon, the prior entropy of Y is expressed as follows:

$$H(Y) = -\sum_{y \in Y} P(y) \log P(y) \tag{2}$$

where $P(y)$ represents the probability of Y. The conditional entropy $H(Y \mid X)$ is computed according to the following formula:

$$H(Y \mid X) = -\int P(x)(\sum_{y \in Y} P(y \mid x) \log P(y \mid x))dx \tag{3}$$

The mutual information MI between X and Y is presented formally as follows:

$$I(X;Y) = \sum_{y \in Y} \int P(y \mid x) \log \frac{P(y,x)}{P(y)P(x)} dx \tag{4}$$

The estimation of probability density is based on the distribution of mutual information [7] in a Bayesian framework by a second-order Dirichlet prior distribution. Beta approximation of the distribution is adopted in this paper. The top-ranked 700 genes and 500 genes with higher mutual information for the leukemia data set and the colon data set discussed in section 5 are selected respectively.

If too many genes are selected, there must be much noise retained in the data. On the other hand, if we select very few genes, some information contained in the data set for classification may be lost. The number of genes we select for the two data sets is determined with respect to the observation from classification experiments. Subsequently the redundancy among these selected genes will be removed in section 4.

3 Background on Rough Set Theory

Rough set theory is first introduced by Pawlak [8] in the 1980s as a mathematical tool to deal with uncertainty. In this section, we will introduce the principal concepts of rough sets theory related to our attribute reduction approach.

Information System: In rough sets theory, an information system S is denoted as $S = \{U, A, V, f\}$, where U is a finite set of instances $U = \{x_1, x_2, \cdots, x_n\}$. A is a finite set of attributes (features) and consists of condition attribute set C and decision attribute set D. $f : U \times A \rightarrow V$ is a function that $f(x_i, q) \in V_p$ for every $q \in A, x_i \in U$.

Indiscernibility Relation: Let $P \subseteq A, x_i, x_j \in U$, a binary relation IND called indiscernibility relation is defined as follows:

$$IND(P) = \{(x_i, x_j) \mid (x_i, x_j) \in U \times U, a \in P, f(x_i, a) = f(x_j, a)\} \tag{5}$$

Let $U / IND(P)$ denote the family of all equivalence classes of the relation $IND(P)$. For simplicity notation $U / IND(P)$ will be written as U / P.

Lower Approximation: Let $R \subseteq C$ and $X \subseteq U$, the R-lower approximation set of X is the set of all elements of U which can be certainly classified as elements of X according to knowledge R. It can be presented formally as follows:

$$\underline{R}X = \bigcup \{Y \in U / R : Y \subseteq X\} \tag{6}$$

Positive Region: The positive region of decision attribute set D with respect to R is the set of all objects from universe U that can be classified with certainty to classes of U / D employing attributes from R. It can be defined as follows:

$$POS_R(D) = \bigcup_{X \in U/D} \underline{R}X \tag{7}$$

4 BPSO for Attribute Reduction in Rough Set

Particle swarm optimization (PSO) is an evolutionary computation technique first introduced for use in real number space by Kennedy and Eberhart in 1995. It has been shown to be a powerful optimization method in many practical applications. In 1997, a binary version of particle swarm optimization (BPSO) is proposed and its performance has been tested on five benchmark functions [9]. But it has not been widely used and still need much further research.

Rough set can be used to find out all possible feature subsets. However, examining exhaustively all subsets of features for selecting the optimal one has been proved to be NP-hard [10]. Heuristic algorithms provide a new way to solve this NP-hard optimization problem. In this section, we suggest binary particle swarm optimization as an attribute reduction algorithm in rough set and apply it to find minimal reduct by removing the redundancy among the genes selected by mutual information.

4.1 Data Preprocessing

The values of gene expression level are continuous. But rough set can only handle discrete attribute value. Hence in order to use the attribute reduction algorithm, all the express level values of selected genes must be discretized firstly. The Entropy/MDL discretization algorithm of Rosetta [11] is used in our experiments. During the entire procedure of attribute discretization and attribute reduction, we combine the training samples with the testing samples together for each data set.

4.2 Population Initialization

Let n be the number of selected features (genes) by mutual information from the original data set. The velocity of i_{th} particle is initialized as a n-dimensional vector with the following form: $V_i = (v_{i1}, v_{i2}, \cdots, v_{in})$. Then the initial position of the i_{th}

particle $X_i = (x_{i1}, x_{i2}, \cdots, x_{in})$ can be computed according to Eq. (11). Where $v_{ij} \in [-6,6]$, $x_{ij} \in \{0,1\}$, $j \in \{1,2,\cdots,n\}$. x_{ij} is equal to 1 or 0 which indicates the corresponding feature is selected or not. We put all the selected features of the i_{th} particle into the attribute set P_i.

4.3 Fitness Evaluation

The goal of reduction is to use fewer features to achieve the same or better performance compared with that obtained using the complete feature set. Hence, individual evaluation contains the following two objectives: (1) Minimization of the feature numbers; (2) Maximization of the classification capability. We have to make some tradeoffs between the two objectives. In this paper, classification capability is always have higher priority. If two individuals have the same classification capability, the individual with fewer features will have higher fitness. A simple weighting method is adopted to define the fitness of particles as follows:

$$f(i) = w_1 \times fc(i) + w_2 \times (1 - fn(i)) \tag{8}$$

$$fc(i) = card(POS_{P_i}(D)) / card(U) \tag{9}$$

where w_1 and w_2 are weight coefficients, $fc(i)$ is the classification capability we can get by using the feature set P_i, $fn(i)$ is the number of features contained in P_i, $card(U)$ represents the cardinality of the set U.

4.4 Update Velocity and Position

Each particle represents a candidate solution with four state variables: v_i, x_i, p_i, p_g. These variables present the current velocity, current position, previous best position and current global best position of the i_{th} particle respectively. The velocity and position vector are updated according to the following equations:

$$V_{ij} = w \times V_{ij} + c_1 \times rand() \times (P_{ij} - X_{ij}) + c_2 \times rand() \times (P_{gj} - X_{ij}) \tag{10}$$

$$X_{ij} = \begin{cases} 0, if\ \rho \geq sig(V_{ij}) \\ 1, if\ \rho < sig(V_{ij}) \end{cases} \tag{11}$$

where c_1 and c_2 are known as acceleration coefficients, X_{ij} represents the j_{th} element of the n-dimensional vector X_i. *Rand* () produces a random number

between 0 and 1. ρ is a random number selected from the uniform distribution in [0, 1]. The function $sig(V_{ij})$ is a sigmoid limiting transformation.

5 Gene Expression Data Sets

There are several microarray data sets published from cancer gene expression studies. Two data sets of them are used to test the effectiveness of our proposed method. Because the benchmark data sets have been studied in many papers, we can compare the results of our method with others conveniently.

Leukemia data set [12] consists of 72 samples: 25 samples of acute myeloid leukemia (AML) and 47 samples of acute lymphoblastic leukemia (ALL). Each sample contains the expression levels of 7129 genes measured using high density oligonucleotide microarrays. In our experiments, 38 samples are used as training data and the remaining 34 samples are used as test data.

Colon data set consists of 62 samples of colon epithelial cells taken from cancer patients [13]. Each sample contains 2000 gene expression levels. 20 out of 62 samples are normal samples and the remaining are cancer samples. In our different experiments, 31 or 44 samples are used as training data and the remaining 31 or 18 samples are used as test data.

6 Experiment Results

In all the experiments, BPSO is run with a population size of 30 and it is terminated when the maximum generation of 600 is arrived. The five parameters in the Eq. (8) and Eq. (10) are set to as follows: $w = 1, c_1 = c_2 = 2, w_1 = 100, w_2 = 0.06$. All the algorithms are implemented in matlab 6.5. The features selected by BPSO are used for testing the classification accuracy by support vector machine (SVM) [14]. The classification accuracy is measured by the number of correct predictions made by the classifier over the test set.

In table 1, experiment results are displayed by using different feature selection methods (No indicates no feature selection). The classification accuracy is obtained by using linear SVM. For the colon data set, 44 samples are used as training data and 18 samples as testing data in this experiment. We can observe that our proposed feature selection method MIRS can always obtain higher classification accuracy with fewer features compared with using MI only. The effectiveness of MIRS is also verified by the remarkable improvement of classification accuracy compared with not using feature selection.

Table 1. The comparison results on feature number and classification accuracy

Dataset	Original No.	FS method	Selected No.	Accuracy
Leukemia	7129	No	7129	58.8%

Dataset	Original No.	FS method	Selected No.	Accuracy
	7129	MI	700	94.1%
	7129	MIRS	48	97.1%
Colon	2000	No	2000	50.0%
	2000	MI	500	66.7%
	2000	MIRS	32	94.4%

Table 2. Comparison of classification accuracy between MIRS and other methods

Data set	Feature selection	Linear SVM	Rbf SVM
Leukemia	MIRS	97.1%	97.1%
	PC	79.4%	79.4%
	CC	85.3%	85.3%
Colon	MIRS	80.7%	83.9%
	PC	64.5%	64.5%
	CC	64.5%	64.5%

We have also compared the performance of MIRS with some other classical feature selection techniques on the two real gene expression data sets. The comparison results are shown in table 2. The results of other techniques are extracted from a survey re-ported by Sung [15]. These feature selection techniques being compared include principal components (PC) and correlational coefficient (CC).

In this experiment, we use 31 samples as training data and the other 31 samples as testing data for the colon data set so that we can compare with the results of Sung directly. We try the following two kinds of support vector machines: (1) Linear SVM (no kernel); (2) Radial basis function SVM (RBF kernel). It is obvious that our proposed method is consistently better than the above methods in all the two data sets.

7 Conclusions

In this paper, we propose a novel feature selection method based on mutual informa-tion and rough set (MIRS). First, we select some top-ranked features which have higher mutual information with the target class to predict from two public available real gene expression data sets. Then rough set theory is applied to remove the redun-dancy among these selected genes. Binary particle swarm optimization (BPSO) is first proposed for attribute reduction in rough set. Finally, the effectiveness of the proposed method is evaluated by the classification accuracy of SVM classifier. Ex-periment results show that MIRS can always get higher prediction accuracy with small number of features compared with using MI only and is superior to some other classical feature selection methods. Generally, the results are highly promising.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant No. 60433020 and the Key Laboratory of Symbol Computation and Knowledge Engineering of the Ministry of Education.

References

1. Furey T., Cristianini N., Duffy N.: Support Vector Machine Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data. Bioinformatics, 16 (2000) 909-914
2. Model F., Adorjan P., Olek A., Piepenbrock C.: Feature Selection for DNA Methylation Based Cancer Classification. Bioinformatics, 17 (2001) 157-164
3. Kohavi R., John G.: Wrapper for Feature Subset Selection. Artificial Intelligence, 97 (1997) 273-324
4. Chow T., Huang D.: Estimating Optimal Feature Subsets Using Efficient Estimation of High-Dimensional Mutual Information. IEEE Transactions on Neural Networks, 16 (2005) 213-224
5. Zhong N., Dong J.Z.: Using Rough Sets with Heuristics for Feature Selection. Journal of Intelligent Information System, 16 (2001) 199-214
6. Cover T., Thomas J.: Elements of Information Theory. Wiley Series in Telecommunications, New York (1991)
7. Zaffalon M., Hutter M.: Robust Feature Selection by Mutual Information Distributions. Proceedings of the 14th International Conference on Uncertainty in Artificial Intelligence, (2002) 577-584
8. Pawlak Z.: Rough Sets. International Journal of Computer Information Science, 11 (1982) 341-356
9. Kennedy J., Eberhart R.C.: A Discrete Binary Version of the Particle Swarm Algorithm. Proceedings of the 1997 Conference on Systems, Man, and Cybernetics. Piscata-way NJ, IEEE Press (1997) 4104-4109
10. Skowron A., Rauszer C.: The Discernibility Matrices and Functions in Information Systems. Intelligent decision support: Handbook of applications and advances of rough set theory, 11 (1992) 331-362
11. Aleksander Øhrn: Institute of Mathematics, University of Warsaw, Poland. http://rosetta.lcb.uu.se/
12. Golub T.R., Slonim K.D., Tamayo P. et al.: Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science, 286 (1999) 531-537
13. Alon U., Barkai N. et al.: Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Cancer Tissues Probed by Oligonucleotide Arrays. PNAS, 96 (1999) 6745-6750
14. Vapnik V.: The Nature of Statistical Learning Theory. Springer-Verlag, Berlin Heidel-berg New York (1995)
15. Sung-Bae Cho, Hong-hee Won: Machine Learning in DNA Microarray Analysis for Cancer Classification. Proceedings of the First Asia-Pacific bioinformatics conference on Bioinformatics, 19 (2003) 189-198