Feature Selection of DNA Microarray Data

Mohammed Liakat Ali

Course: 60-520
Fall 2005
University of Windsor
E-mail: ali1p@uwindsor.ca
December 2, 2005

Abstract. The feature selection is a vital process for data classification. To find a minimum optimal feature set is a NP-complete problem. Various approximation algorithms are proposed for feature selection.
Various feature selection methods are discussed. Deployment of the methods are also discussed.
We have compared experimental results of a redundancy based methods with our gene ranking methods. We have found that redundancy based feature selection methods can select compact gene subsets. We also found that only gene ranking based methods also perform well.

1 Introduction

In pattern recognition, feature selection is done before doing classification of the data. As per Horn and Martinez (1994)[5], finding an optimal minimum feature set for classification is a NP-complete problem. Various approximation algorithms to select minimum number of features have been proposed in the literature. Both filter and wrapper methods are used extensively. Using the selected minimum feature set will improve significantly the performance of DNA Microarray data classifiers.

Minimum feature set may infer that a group of 20 genes behave differently, say, from cancer and healthy patients. That would lead to devising efficient drugs and/or treatment for controlling the behavior of those genes.

The analysis of DNA Microarray data is quite important nowadays and this tendency will prevail in the near future by helping to enhance health care, environmental studies, and analysis of biological systems.

Before delving into the details, let us have a closer look to the terms like microarray data, representation of objects, classifiers, feature selection and feature extraction, and optimal feature set.

1.1 Microarray Data

Microarray technology is one of the most promising tools available to life science researchers. The cDNA arrays and the Affymatrix technologies are used to produce DNA chip or DNA microarray. The final result of microarray experiment is a set of numbers representing expression level of DNA segments. These DNA segments are called genes.

A gene encodes specific instructions to produce a specific product typically a protein by a cell. The amount of protein produced by a gene is called expression level of the gene.

As per the central dogma of molecular biology, the flow of genetic information is from DNA to RNA to proteins. In the cell nucleus, the information in the DNA is copied into RNA. The process is called transcription. RNA copy of the genetic information act as a messenger. RNA transport the information to the cell cytoplasm. In the cell cytoplasm, RNA is translated into proteins. The proteins are the products of the genes.

Genes are only a portion of the entire amount of DNA in a cell. Genes are delimited by a specific sequence of bases. This sequences are called signal. They controls the production of protein. They turn on and off the production of protein through protein and DNA interactions.

Every cell of individual organism will have same DNA. Thus they carry same information. But not all genes are expressed in the same way in all cells. The different pattern of gene activations control the production of protein in different cells.

1.2 Representation of Objects

On its own right, representation of objects by features is an important topics. Objects are represented by their characteristic features. But objects has to be represented with due consideration of their domain and purpose at hand. Thus we can not choose same features for all the purposes.

In microarray data for cell type classification, features are gene expression. For human genome there are 30,000-40,000 genes. It is having a very high dimension. High dimension data has a well known problem in pattern recognition called 'curse of dimensionality'.

Three main reasons to keep dimensionality low are measurement cost, classification accuracy, and to identify and monitor the target disease or function types.

Thus it is very important to represent an object with features with high discriminating ability for classification.

1.3 Classifiers

A classifier will assign the object to a category i.e., class. It will use features of an object and a discriminant function to make the decision. We can divide classifiers as linear and non-linear.

Domain independent theory of classification is based on the abstraction provided by the features of the input data.

1.4 Feature Selection vs. Feature Extraction

In feature selection we try to find the best subset of the input feature set. Feature selection methods are based on individual feature ranking as per a criterion function, minimum redundancy among the features, maximum mutual information among the features and the class level etc.

In feature extraction we create new features based on transformation or combination of the original feature set. Feature extraction methods are Principle Component Analysis (PCA), Linear Discriminant Analysis (LDA) etc.

1.5 Optimal Feature Set for Classification

To find optimal feature subset we have to evaluate objective function for

$$\sum_{m=0}^{d} \binom{d}{m}$$

subsets. Clearly it has an exponential complexity, $O(2^d)$. As mensioned in the beginning finding an optimal minimum feature set for classification is a NP-complete problem.

2 Deployment of Feature Selection Methods

Feature selection methods can be grouped as embedded, filter, and wrapper based on their relation with the induction algorithm for classification.

2.1 Embedded method

They will be a part of induction algorithms. Blum, and Langley (1997)[1] stated that methods for inducing logical descriptions provide the clearest example of feature selection methods embedded within a basic induction algorithm. In the context of microarray data classification, very few implementations use the method now-a-days.

2.2 Filter method

They are separate processes from the induction algorithms. The filter are preprocessing methods. They uses general characteristics of the data set to select features. Considering the high dimension of microarray data the filter methods are being frequently used by various implementations to bring the dimension of the data to a manegable value before using other methods.

2.3 Wrapper method

They are also separate processes from induction algorithm but they use an induction algorithm as a subroutine. They evaluate alternative subsets by an induction algorithm.

3 Feature Selection Methods

We can divide feature selection methods as Optimal Selection Methods and Suboptimal Selection Methods based on the optimal solution of the problem.

3.1 Optimal Selection Methods[6]

As per Jain, Duin, and Mao (2000), The following methods are guaranteed to find optimal minimum feature subset.

Exhaustive Search The method will evaluate all possible subsets consisting of m features of total d features i.e.,

$$\sum_{m=0}^{d} \binom{d}{m}$$

subsets to find the best subset. It is an exponential problem.

Branch and Bound Search In this method only a fraction of all possible feature subsets will be evaluated. Criterion function must satisfy the monotonicity property i.e., $J(x_1, \ldots, x_i) \leq J(x_1, \ldots, x_i, x_{i+1})$

3.2 Sub Optimal Selection Methods[6]

Again as per Jain, Duin, and Mao (2000), There is no guaranty that the following methods will find an optimal minimum feature subset.

Best individual Features Here all the d features are evaluated individually using an scalar criterion function. Then select m best features. Clearly a sub optimal method because as per Cover (1974)[2] the best two independent measurements are not the two best. Complexity of the method is O(d).

Sequential Forward Selection (SFS) At the beginning it will select the best feature using a scalar criterion function. Then it will add one feature at a time. The new feature along with already selected features will maximize the criterion function, J(.). It is a greedy algorithm. It cannot retract. Once a feature is selected there is no way to discard the feature. Complexity of the method is O(d).

Sequential Backward Selection (SBS) At the beginning it will select all d features. Then it will delete one feature at a time and select the subset which maximize the criterion function, J(.). It is also a greedy algorithm, It also cannot retract. Once a feature is discarded there is no way to add the feature. Complexity of the method is O(d). But require more computation than SFS.

Plus l take away r Selection The method will at first add l features by forward selection, then discard r features by backward selection. There is a need to decide optimal l and r. The method has no subset nesting problems like SFS and SBS.

Sequential Forward Floating Search (SFFS) The method is a generalized plus l take away r algorithm. The value of l and r are determined automatically. Performance is close to optimal with an affordable computational cost.

Sequential Backward Floating Search (SBFS) It is also a generalized plus l take away r algorithm like SFFS. The value of l and r are also determined automatically. Close to optimal solution as SFFS. More efficient than SFFS for m closer to d than to 1.

4 Class Separability Measures[8]

We shall consider now two examples of measuring the discrimination effectiveness using mainly feature vectors. For multiple features, correlation among features usually influence classification capability.

4.1 Divergence

As per Bayes rule, given two classes $\omega 1$ and $\omega 2$ and a feature vector x, we select $\omega 1$ if P(ω 1| x) > P(ω 2|x). Hence ratio $\frac{P(\omega 1|x)}{P(\omega 2|x)}$ has discriminating capability. For given P($\omega 1$) and P($\omega 2$) same information resides in $D_{12}(x) = \ln \frac{P(x|\omega 1)}{P(x|\omega 2)}$. For completely overlapping classes $D_{12}(x) = 0$.

Since x takes different values, it is natural to consider mean value over class $\omega 1$

$$D_{12} = \int_{-\infty}^{+\infty} p(x \mid \omega 1) \ln \frac{p(x \mid \omega 1)}{p(x \mid \omega 2)} dx$$

Similarly for $\omega 2$

$$D_{21} = \int_{-\infty}^{+\infty} p(x \mid \omega 2) \ln \frac{p(x \mid \omega 2)}{p(x \mid \omega 1)} dx$$

The sum is $d_{12} = D_{12} + D_{21}$.

The multi-class divergence can be computer by summing up divergence of every class pairs.

4.2 Scatter Matrices

Computation of Divergence is not easy for non Gaussian distribution. On the other hand scatter matrices can be computer very easily. Within class scatter matrix is defined as

$$Sw = \sum_{i=1}^{M} P_i S_i$$

where S_i is the covariance matrix for class ω_i

$$S_i = E[(x - \mu_i)(x - \mu_i)^T]$$

Between class scatter matrix is defined as

$$S_b = \sum_{i=1}^{M} P_i(\mu_i - \mu_0)(\mu_i - \mu_0)^T$$

where μ_0 is the total mean

$$\mu_0 = \sum_{i}^{M} P_i \mu_i$$

Total Mixture scatter matrix is defined as

$$S_m = E[(x - u_0)(x - u_0)^T]$$

We can calculate that $S_m = Sw + S_b$

The following criterion functions can be defined among others

$$J_1 = \frac{trace\{S_m\}}{trace\{S_w\}}$$

$$J_2 = \frac{\mid S_m \mid}{\mid S_w \mid} = \mid S_w^{-1} S_m \mid$$

$$J_3 = trace\{S_w^{-1} S_m\}$$

For equally probable two classes problem $\mid Sw \mid$ is proportional to $\sigma_1^2 + \sigma_2^2$ and $\mid S_b \mid$ is proportional to $(\mu_1 - \mu_2)^2$. The Fisher's discriminant ratio is defined as

$$FDR = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}$$

5 Review of Minimum Redundancy feature selection methods

5.1 Paper 1: Ding and Peng (2003)[3]

Filter method is used. The algorithm is a SFS. The first feature was selected using $maxV_1, V_1 = \frac{1}{|S|} \sum_{i \in S} I(h, i)$ for all genes in the set S.

Suppose the algorithm already selected m features for the set X. The additional features will be selected from the set $Y = S - X$. The following two conditions will be optimized simultaneously

$$max_{i \in Y} I(h, i)$$

$$min_{i \in Y} \frac{1}{\mid X \mid} \sum_{j \in X} I(i, j)$$

Mutual information, I of two variable x and y is defined as

$$I(x, y) = \sum_{i,j} p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i) p(y_j)}$$

The main contribution of the paper was claimed by the authors was to point out the importance of minimum redundancy in gene selection.

5.2 Paper 2: Yu and Liu (2004)[9]

Here also a filter method is used.

Algorithm is:

Relevance analysis

1. Order features based on decreasing ISU values

Redundancy analysis

2. Initialize F_i with the first feature in the list
3. Find and remove all features for which F_i forms an approximate redundant cover
4. Set F_i as the next remaining feature in the list and repeat step 3 until the end of the list

The implementation combines SFS with elimination technique. The entropy of a variable X is defined as

$$H(X) = -\sum_i P(x_i)\log_2(P(x_i))$$

The entropy of X after observing values of another variable Y is defined as

$$H(X \mid Y) = -\sum_j P(y_j)\sum_i P(x_i \mid y_j)\log_2(P(x_i \mid y_j))$$

The amount by which the entropy of X decreases reflects additional information about X provided by Y, is called Information Gain

$$IG(X \mid Y) = H(X) - H(X \mid Y)$$

Symmetrical uncertainty is defined as

$$SU(X,Y) = 2\big[\frac{IG(X \mid Y)}{H(X)+H(Y)}\big]$$

Individual C-correlation (ISU_i): The correlation between any feature F_i and the class C is called Individual C-correlation, (ISU_i)

Combined C-correlation (CSU_i): The correlation between any feature F_i and $F_j (i \neq j)$ and the class C is called combined C-correlation, $CSU_{i,j}$

Approximate redundant cover: For two features F_i and F_j, F_i formed an approximate redundant cover for F_j iff $ISU_i \geq ISU_j$ and $ISU_i \geq CSU_{i,j}$

6 Comparison with our Experimental Results

To investigate the problem of feature selection we implement a filter method. We used FDR[4] as criterion function. Initial gene selection was based on gene ranking. Then Fisher and Loog-Duin[7] Discriminant techniques are applied to transform the feature space. Then linear and quadratic classifier are used. 10-fold cross validation was applied.

We used Leukemia, Lung cancer, and Breast cancer data from UCI repository.

Dataset		RBF			Our Experiment					
							FQ	LDQ	FL	LDL
Name	Total Genes	Samples	Selected Genes	Acc	Samples	Selected Genes	Acc	Acc	Acc	Acc
Leukemia	7129	72	4	87.50	72	80	98.75	59.23	98.75	95.00
Lung cancer	12533	181	6	98.34	197	367	67.12	49.89	77.32	73.60
Breast cancer	24481	97	67	79.38	97	273	78.63	68.72	78.63	74.70

Table 1. Comparison of gene selection results.

Where, RBF = Redundancy Based Filter Algorithm used in paper 2[9]
FQ = Fisher Discrminant Analysis followed by Quadratic Classifier
FL = Fisher Discrminant Analysis followed by Linear Classifier
LDQ = loog-Duin Discrminant Analysis followed by Quadratic Classifier
LDL = Loog-Duin Discrminant Analysis followed by Linear Classifier

From the table 1 we can observed that RBF selected very compact gene sets for all the cases. FQ and FL out perform LDQ and LDL in all 3 datasets. RBF out perform all methods in 1 dataset by big margin. FQ and FL jointly out perform others in 1 dataset also in big margin. RBF, FQ, and FL have comparable result in 1 dataset.

7 Conclusions

We can conclude that minimum redundancy methods select very compact gene sets. It can help to identify and monitor the target disease or function types.

From our experience, we know that on average the performance of LDQ is better than FQ because Fisher discrminant analysis is linear in nature. Here we selected genes by FDR ranking. Due this performance of FQ and FL may get enhancement.

From the result we can also conclude that gene selection by only ranking has some merits.

The seminar was on feature selection of DNA microarray data for classification. We have tried to explain the technical aspects in easy to understand format.

References

1. A. Blum and P. Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97(1-2):245–271, 1997.
2. T.M. Cover. he Best Two Independent Measurements Are Not the Two Best. *IEEE Trans. Systems, Man, and Cybernetics*, 4:116–117, 1974.
3. C. Ding and H. C. Peng. Minimum Redundancy Feature Selection from Microarray Gene Expression Data. *Proc. Second IEEE Computational Systems Bioinformatics Conf.*, pages 523–528, 2003.
4. R. Duda, P. Hart, and D. Stork. *Pattern Classification*. John Wiley and Sons, Inc., New York, NY, 2nd edition, 2000.
5. K. S. V. Horn and T. Martinez. The Minimum Set Problem. *Neural Networks*, 7(3):491–494, 1994.

6. Duin R. P. W. Jain, A. K. and J. Mao. Statistical Pattern Recognition: A review. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 22(1), 2000.
7. M. Loog and P.W. Duin. Linear Dimensionality Reduction via a Heteroscedastic Extension of LDA: The Chernoff Criterion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(6):732–739, 2004.
8. S. Theodoridis and K. Koutroumbas. *Pattern Recognition*. Elsevier Academic Press, second edition, 2003.
9. L. Yu and H. Liu. Redundency Based Feature Selection for Microarray Data. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 737 – 742, 2004.