Section A-Research paper
ISSN 2063-5346

BROAD PATTERNS OF GENE EXPRESSION REVEALED BY CLUSTERING ANALYSIS OF BIOLOGICAL AND CLINICAL DATA

Dr.T.SHANMUGVADIVU

Assistant Professor in Computer Science, Arulmigu Palaniandavar Arts College For Women, Palani.

DOI: 10.48047/ecb/2023.12.si4.1754

ABSTRACT: Due to a large number of genes and a small sample size, gene expression microarray data poses a severe challenge to the accurate classification of diseases or phenotypes. Gene selection is a frequently used technique in preprocessing microarray data for successful classification of diseases or phenotypes. Widely used gene selection methods are mainly focused on filter approaches. They have been proved to be efficient and effective. However, the researcher finds that some genes discarded by many existing methods are helpful for classification at certain conditions and cannot be removed blindly. With more and more biological information generated, the most pressing task of bioinformatics has been to analyze and interpret various types of data, including nucleotide and amino acid sequences, protein structures, gene expression profiling and so on. The researchers have applied the data mining techniques of feature generation, feature selection, and feature integration with learning algorithms to tackle the problems of disease phenotype classification and patient survival prediction from gene expression profiles, and the problems of functional site prediction from DNA sequences.

Keywords: Feature generation, Feature selection, Gene expression, Gene Classification

1. INTRODUCTION

The rapid advances in microarray technology enable biologists to measure the expression levels of thousands or ten thousands of genes simultaneously. The initial information from microarray experiments goes through various data processing steps including image processing, quality control and normalization. The resulting data set is called gene expression microarray data set, which is a two dimensional array with thousands of columns (genes) but a small number (often less than one hundred) of rows (samples). Such data set poses a very severe challenge to sample classification and usually results in the known problem of "curse of dimensionality" and over-fitting of the training data for traditional sample classification technologies. Therefore, selecting a small number of discriminative genes from thousands of genes is essential for successful sample classification. In recent years, feature selection has been extensively applied to gene selection for sample classification. Feature selection is a process that selects a subset of original features by reducing the number of features, removing irrelevant, redundant, or noisy data. It has been

proved to be effective in improving classification accuracy, speeding up a classification algorithm and enhancing result comprehensibility. With regard to how to evaluate the goodness (quality) of a subset of features, the feature selection methods fall into two broad categories: the filter approach and the wrapper approach. In the filter approach, a good feature set is selected as a result of pre-processing based on properties of the data itself and independent of the classification algorithm. The wrapper approach requires one predetermined mining algorithm in feature selection and uses its performance to evaluate and determine which features are to be selected. It tends to give superior performance as it finds features better suited to the predetermined classification algorithm, but it is more computationally expensive than the filter approach. For this reason, the filter model is widely used in gene selection for microarray data. However, the high dimensionality and small sample size of microarray data also poses severe challenges to filter approaches in terms of effectiveness. Some of the recent research efforts have been focused on these challenges whereas the researchers find that selecting some genes from the ones that are discarded by these filter methods can lead to even higher classification accuracy. In this work aim to develop a novel hybrid solution for gene selection in sample classification of microarray data which can select discriminative genes and improve classification accuracy more effectively.

2. MOTIVATION

The aim of data mining is to automatically or semi-automatically discover hidden knowledge, unexpected patterns and new rules from data. There are a variety of technologies involved in the process of data mining, such as statistical analysis, modeling techniques and database technology. During the last ten years, data mining is undergoing very fast development both techniques and applications. Its typical applications include market segmentation, customer profiling, fraud detection, (electricity) loading forecasting, and credit risk analysis and so on. In the current post-genome age, understanding floods of data in molecular biology brings great opportunities and big challenges to data mining researchers. Successful stories from this new application will greatly benefit both computer science and biology communities. Good feature subsets contain features highly correlated with the class, yet uncorrelated with each other. This principle makes use of the properties of the data itself to evaluate the goodness of features and is independent of the classification algorithm. It is simple and efficient for high dimensional data.

3. RELATED WORK

In this section, the researcher summarizes the earlier studies related to this work. Feng Chu et al. [1] described their research in "Applications of Support Vector Machines to Cancer Classification with microarray data". Microarrays, also known as gene chips or DNA chips, provide a convenient way of obtaining gene expression levels for a large number of genes simultaneously. Each spot on a microarray chip contains the clone of a gene from a tissue sample. Some mRNA samples are labeled with two different kinds of dyes, for example, Cy5 (red) and Cy3 (blue). After mRNA interacts with the genes, i.e. hybridization, the color of each spot on the chip will change. The resulted image reflects the characteristics of the tissue at the molecular level. Microarrays can thus be used to help classify and predict

different types of cancer. Traditional methods for diagnosis of cancer are mainly based on the morphological appearances of the cancers; however, sometimes it is extremely difficult to find clear distinctions between some types of cancer according to their appearances. Hence the microarray technology stands to provide a more quantitative means for cancer diagnosis. For example, gene expression data have been used to obtain good results in the classifications of lymphoma, leukemia, breast cancer and liver cancer.

Brown et al. introduced a new method of functionally classifying genes using gene expression data from DNA microarray hybridization experiments. The method is based on the theory of support vector machines (SVMs). Brown et al. described SVMs that use different similarity metrics including a simple dot product of gene expression vectors, polynomial versions of the dot product, and a radial basis function. Compared to the other SVM similarity metrics, the radial basis function SVM appears to provide superior performance in identifying sets of genes with a common function using expression data. In addition, SVM performance is compared to four standard machine learning algorithms. SVMs have many features that make them attractive for gene expression analysis, including their flexibility in choosing a similarity function, sparseness of solution when dealing with large data sets, the ability to handle large feature spaces, and the ability to identify outliers.

Each data point produced by a DNA microarray hybridization experiment represents the ratio of expression levels of a particular gene under two different experimental conditions. An experiment starts with microarray construction, in which several thousand DNA samples are fixed to a glass slide, each at a known position in the array. Each sequence corresponds to a single gene within the organism under investigation. Messenger RNA samples are then collected from a population of cells subjected to various experimental conditions. These samples are converted to cDNA via reverse transcription and are labeled with one of two different fluorescent dyes in the process. A single experiment consists of hybridizing the microarray with two differently labeled cDNA samples collected at different times. Generally, one of the samples is from the reference or background state of the cell, while the other sample represents a special condition set up by the experimenter, for example, heat shock. The level of expression of a particular gene is roughly proportional to the amount of cDNA that hybridizes with the DNA affixed to the slide. By measuring the ratio of each of the two dyes present at the position of each DNA sequence on the slide using laser scanning technology, the relative levels of gene expression for any pair of conditions can be measured. The result, from an experiment with n DNA samples on a single chip, is a series of n expression-level ratios. Typically, the numerator of each ratio is the expression level of the gene in the condition of interest to the experimenter, while the denominator is the expression level of the gene in the reference state of the cell. The data from a series of m such experiments may be represented as a gene expression matrix, in which each of the n rows consists of an m-element expression vector for a single gene. In these experiments the number of experiments m is 79 and the number of genes n is 2467. Brown et al.[2] define Xi to be the logarithm of the ratio of gene X's expression level in experiment i to X's expression level in the reference state. This log ratio is positive if the gene is induced (turned up) with

respect to the background and negative if it is repressed (turned down).

Bald et al. [3] described their research in "A Bayesian Framework for the Analysis of Microarray Expression Data": Regularized t-Test and Statistical Inferences of Gene Changes". DNA microarrays are now capable of providing genome-wide patterns of gene expression across many different conditions. The first level of analysis of these patterns requires determining whether observed differences in expression are significant or not. Current methods are unsatisfactory due to the lack of a systematic framework that can accommodate noise, variability, and low replication often typical of microarray data. Bald et al. [3] developed a Bayesian probabilistic frame work for microarray data analysis. At the simplest level, Bald [3] model log-expression values by independent normal distributions, parameterized by corresponding means and variances with hierarchical prior distributions. Bald [3] derived point estimates for both parameters and hyper parameters, and regularized expressions for the variance of each gene by combining the empirical variance with a local background variance associated with neighboring genes. An additional hyper parameter, inversely related to the number of empirical observations, determines the strength of the background variance. Simulations show that these point estimates, combined with a t-test, provide a systematic inference approach that compares favorably with simple t-test or fold methods, and partly compensate for the lack of replication. The approach is implemented in software called Cyber-T accessible through a Web interface at www.genomics.uci.edu/software.html. The code is available as Open Source and is written in the freely available statistical language R.

4. FEATURE SELECTION TECHNIQUE FOR DATA MINING

A known problem in classification (in general machine learning) is to find ways to reduce the dimensionality of the feature space to overcome the risk of over-fitting. Data over-fitting happens when the number of features is large ("curse of dimensionality") and the number of training samples is comparatively small. In such a situation, a decision function can perform very well on classifying training data, but does poorly on test samples. Feature selection is concerned with the issue of distinguishing signal from noise in data analysis.

Feature selection techniques can be categorized according to a number of criteria. One popular categorization is based on whether the target classification algorithm will be used during the process of feature evaluation. A feature selection method, that makes an independent assessment based only on the general characteristics of the data is named "filter" while, on the other hand, if a method evaluates features based on accuracy estimates provided by certain learning algorithm which will ultimately be employed for classification, named as "wrapper". With wrapper methods, the performance of a feature subset is measured in terms of the learning algorithm's classification performance using just those features. The classification performance is estimated using the normal procedure of cross validation, or the bootstrap estimator. Thus, the entire feature selection process is rather computation-intensive. For example, if each evaluation involves a 10-fold cross validation, the classification procedure will be executed 10 times. For this reason, wrappers do not scale well to data sets

containing many features. Besides, wrappers have to be re-run when switching from one classification algorithm to another. In contrast to wrapper methods, filters operate independently of any learning algorithm and the features selected can be applied to any learning algorithm at the classification stage.

Filters have been proven to be much faster than wrappers and hence, can be applied to data sets with many features [45]. Since the biological data sets discussed in the later chapters of this thesis often contain a huge number of features (e.g. gene expression profiles), the researcher concentrates on filter methods.

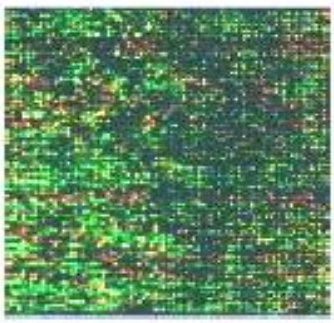

Figure-1: An illuminated microarray (enlarged). A typical dimension of such an array is about 1 inch or less, the spot diameter is of the order of 0.1 mm, for some microarray types can be even smaller.

Figure-1 shows an illuminated enlarged microarray. Another taxonomy of feature selection techniques is to separate algorithms evaluating the worth or merit of subset features from those of individual features. Most of the feature selection methods introduced in evaluate how well an individual feature contributes to the separation of samples in different classes and produce a simple feature ranking. However, there is also one method in this chapter, the correlation-based feature selection that assesses and selects a subset of features. The researchers also present a new feature selection algorithm, ERCOF, which first evaluates features individually and then forms the final representative feature set by considering the correlations between the features. There are some other dimensions to categorize feature selection methods. For example, some algorithms can handle regression problem, that is, the class label is numeric rather than a discrete valued variable; and some algorithms evaluate and rank features independently from class, i.e. unsupervised feature selection. The restrict their study to the data sets with discrete class label since this is the case of the biological problems analyzed of this thesis, though some algorithms presented can be applied to numeric class label as well.

The frequency of a k-gram pattern is used as the value of this feature. For example,

- UP-X (DOWN-X), which counts the number of times the letter X appears in the up-stream (down-stream) part of a functional site in its nucleotide acid or amino acid sequence.
- UP-XY (DOWN-XY), which counts the number of times the two letters XY appear as a substring in the up-stream (down-stream) part of a functional site in its nucleotide acid or amino acid sequence.

where X and Y range over the 4 nucleotide acid letters or the standard 20 amino acid letters and the special stop codon symbol.

5. ENTROPY-BASED RANK SUM TEST AND CORRELATION FILTERING (ERCOF)

In this strategy, combine the above presented methods of entropy measure and Wilcoxon rank sum test, as well as Pearson correlation coefficient test together to form a three-phase feature selection process. The researcher name this combined feature selection process ERCOF - Entropy-based Rank sum test and Correlation Filtering.

ERCOF - Entropy-based Rank sum test and Correlation Filtering Methodology

Step-1: k=1
Step-2: Rank all features in group F on class entropy in an ascending order, $f_1, f_2, \ldots f_l$.
Step-3: Let $S_k = \{f_1\}$ and remove f_1 from F.
Step-4: For each f_i (i>1)
Calculate Pearson correlation coefficient $r(f_1,f_i)$;
If $r(f_1,f_i) > r$, then
Add f_i into S_k and remove it from F;
Step-5: k=k+1 and go to Step-2 until F=Φ

Using ERCOF in gene expression data analysis where there are often more than thousands of features, expects to identify a subset of sharply discriminating features with little redundancy. The entropy measure is effective for identifying discriminating features. After narrowing down by the Wilcoxon rank some test, the remaining features become sharply discriminating. Then, with the correlation examination, some highly correlated features are removed to reduce redundancy. CFS selects only one feature if the class entropy of this feature is zero. However, Pearson correlation coefficient also has a shortcoming - the calculation of correlation is dependent on the real values of features which are sensitive to some data transformation operations. Therefore, other algorithms are being implemented to group correlated features.

6. RESULTS AND DISCUSSION

The feature selection techniques reviewed in the preceding sections have been used as key steps in the handling of high-dimensional biomedical data. For example, their use is prevalent in the analysis of microarray gene expression data. Besides, they have been also used in the prediction of molecular bioactivity in drug design, and more recently, in the analysis of the context of recognition of functional site in DNA sequences. One issue should be addressed here is the so-called “multiple comparisons problem” which happens when selects features by choosing a statistical confidence level (like standard 5% or 1%) for t-test,

$\aleph^2$ -test, and other statistical measures. The description of the problem is: when performing m multiple independent significance tests, each at the α level, the probability of making at least one Type I error (rejecting the null hypothesis inappropriately) is $1-(1-\alpha)^m$. For example, suppose we consider m=200 features and perform independent statistic tests to each of them at the standard α=5% level, then the probability of getting at least one significant result is $1-0.95^{200}$=0:99996. So, when we get a significant feature among the tests. In fact, under this setting, we would still expect to observe approximately 10 (200*0.05) "significant" features, even when there were actually no features that can distinguish the two classes. Obviously, the problem becomes serious when the total number of considered features is large, which is the case in some biological data such as gene expression profiling. Table-1 shows a sample data set.

Data Set	Program	TP	FN	Specificity	Precision	CC
CDs	Erpin	880	102	89.6 %	84.3 %	0.483
	Polyadq	862	120	87.8 %	82.0 %	0.459
	Proposed	887	95	90.3 %	85.4 %	0.497
Introns	Erpin	741	241	75.5 %	69.5 %	0.320
	Polyadq	718	264	73.1 %	67.5 %	0.293
	Proposed	775	207	78.9 %	72.8 %	0.363
Simple shuffling	Erpin	888	94	90.4 %	85.4 %	0.494
	Polyadq	826	156	84.1 %	77.8 %	0.415
	Proposed	942	60	95.9 %	93.3 %	0.570
Markov 1st order	Erpin	772	210	78.6 %	72.3 %	0.354
	Polyadq	733	249	74.6 %	68.7 %	0.309
	Proposed	775	217	78.9 %	71.9 %	0.361

Table-1: Sample Data Set

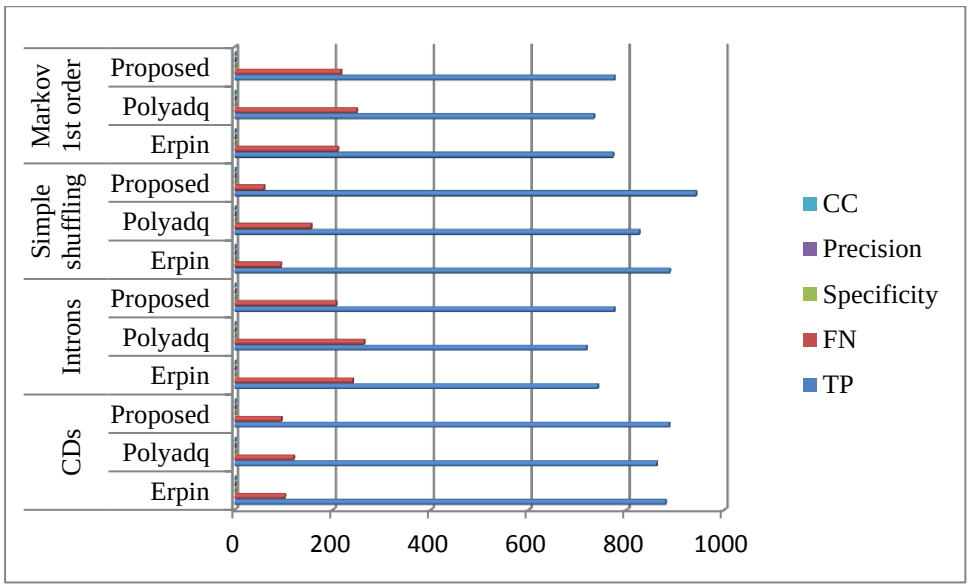

Figure-2: Performance Comparison

Figure-2 shows the validation results by different programs on different sequences not containing PASes, coding sequences (CDS), introns, and two types of randomized UTR sequences (simple shuffling and 1st order Markov simulation). TN is the number of true negatives. FP is the number of false positives. CC is correlation is

Co-efficient and CC= $\frac{(TP+TN-FP+FN)}{\sqrt{(TP+FP)*(TP+FN)*(TN+FP)*(\mathrm{TN}+\mathrm{FN})}}$.

Calculations of Precision and CC use TP and FN features ranking by their entropy values (the less the entropy value is, the more important the feature is). Some of these top features can be interpreted by those reported motifs, for example, it clearly visualizes both USE and DSE as characterized by G/U rich segments since UP-TGT, UP-T, DOWN-TGT, DOWN-T, UP-TG and UP-TT are among top features showed in Table-2.

Rank	1	2	3	4	5	6	7	8	9	10
Feature	UP-TGT	DOWN-A	UP-T	UP -AG	DOWN –TGT	DOWN-T	UP -TG	UP -TT	DOWN -AA	UP -A

Table-2: The top 10 features selected by entropy-based feature selection method for PAS classification and prediction in human DNA sequences.

6.1 Prediction of PAS in mRNA sequences

When the researcher applies the model to 312 true PASes that were extracted from mRNA sequences, the results obtained are not good - only around 20% of them can be predicted correctly. Besides, the program Erpin performs even worse on these PASes - with prediction accuracy at only 13%. These poor results may indicate that the good features used in the model for PAS prediction in DNA sequences are not efficient for mRNA. Therefore, the researcher decides to build another model for mRNA sequences without poly(A) tails. This model is also expected to provide a new way for predicting the mRNA cleavage site/poly(A) addition site.

7. CONCLUSION

This paper focuses on how to effectively apply data mining technologies to biological and clinical data. Some problems arising from gene expression profiling and DNA sequence data are studied in depth using data mining techniques of feature generation, feature selection, and feature integration with learning algorithms. In order to identify genes associated with disease phenotype classification or patient survival prediction from gene expression data, a new feature selection strategy, ERCOF (Entropy based Rank sum test and Correlation Filtering), is worked out by combining entropy measure, Wilcoxon rank sum test and Pearson correlation coefficient test. ERCOF conducts a three-phase feature filtering aiming to find a subset of sharply discriminating genes with little redundancy. In the first phase, it selects genes using an entropy-based method that generally keeps only 10% of the

features. In the second phase, a non-parametric statistics called the Wilcoxon rank sum test is applied to the features kept by the first phase to further filter out some genes and divide the remaining ones into two groups - one group consisting of genes that are highly expressed in one type of samples (such as cancer) while another group consisting of genes that are highly expressed in another type of samples (such as non-cancer). In the third phase, correlated genes in each group are determined by Pearson correlation coefficient test and only some representatives of them are chosen to form the final set of selected genes.

REFERENCES

[1]. Feng Chu and Lipo Wang. Applications of Support Vector Machines to Cancer Classification with Microarray data. International Journal of Neural Systems, vol. 15, no. 6 (2005) 475–484.

[2]. Michael P. S. Brown,William Noble Grundy, David Lin, Nello Cristianini, and Charles Sugnet. Support Vector Machine Classification of Microarray Gene Expression Data. UCSC-CRL-99-09.

[3]. Pierre Baldi, Anthony D. Long. A Bayesian Framework for the Analysis of Microarray Expression Data: Regularized t-Test and Statistical Inferences of Gene Changes. , Bioinformatics 19 (2003), pp. 1–11.

[4]. M. Hall. Correlation-based feature selection for discrete and numeric class machine learning. In Proceedings of the 17th International Conference on Machine Learning, pages 359-366, 2000.

[5]. T. R. Golub et al. Molecular classifications of cancer: Class discovery and class prediction by gene expression monitoring. Science, 286(5439):531–537, 1999.

[6]. M. Robnik-Sikonja and I. Kononenko. Theoretical and empirical analysis of Relief and ReliefF. MachineLearning, 53:23-69, 2003.

[7]. Li-Juan Zhang and Zhou-Jun Li, Gene Selection for classifying microarray data using grey relational analysis. Proceedings of Discovery Science'2006, Lecture Notes in Computer Science Vol.4265: 378-382, 2006.

[8]. Wai-Ho Au, Keith C.C.Chan, Andrew K.C.Wong, and Yang Wang, Attribute Clustering for Grouping, Selection, and classification of gene expression data, IEEE/ACM transactions on computational biology and bioinformatics Vol 2, No.2, April-June 2005.

[9]. Li-Juan Zhang, Zhou-Jun Li, Huo-Wang Chen and Jian Wen, Minimum Redundancy Gene Selection based on Grey Relational analysis. In Workshops Proceedings of

ICDM'2006 pages 120-124, IEEE Computer Society, 2006.

[10]. Blaise Hanczar et.al. Improving Classification of Microarray Data using Prototype-based Feature Selection, SIGKDD Explorations Volume 5, Issue 2, pages23-30, 2003.

[11]. C. Ding and H. Peng. Minimum redundancy feature selection from microarray gene expression data. In Proceedings of the Computational Systems Bioinformatics Conference, pages 523-529, 2003.

[12]. Y. Wu and A. Zhang. Feature selection for classifying high-dimensional numerical data. In IEEE Conference on Computer Vision and Pattern Recognition 2004, volume 2, pages 251–258, 2004.

[13]. L. Yu, H. Liu, Efficient feature selection via analysis of relevance and redundancy, J. Mach. Learning Res. 5 (2004) 1205–1224.

[14]. L. Yu and H. Liu. Redundancy based feature selection for microarray data. In Proceedings of the Tenth ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pages 737–742,2004.

[15]. Li-Juan Zhang, Zhou-Jun Li and Huo-Wang Chen, An Effective Gene Selection Method Based on Relevance Analysis and Discernibility Matrix. Accepted by PAKDD 07.

[16]. David A. Bell and Hui Wang, A Formalism for Relevance and Its Application in Feature Subset Selection, Machine Learning, 41,175-195, 2000.

[17]. R. Kohavi, G. John, Wrappers for feature subset selection, Artif. Intell. 1-2 (1997) 273-324.

[18]. A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma and I. S. Lossos et al., Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling, Nature 403 (2000) 503–511.

[19]. T. Golub, D. K. Slonim, P. Tamayo, C. Huard and M. Gaasenbeek et al., Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science 286 (1999) 531–536.

[20]. X. J. Ma, R. Salunga, J. T. Tuggle, J. Gaudet and E. Enright et al., Gene expression profiles of human breast cancer progression, in Proc. Natl. Acad. Sci. USA, Vol. 100 (2003), pp. 5974–5979.

[21]. X. Chen, S. T. Cheung, S. So, S. T. Fan and C. Barry, Gene expression patterns in human liver cancers, Molecular Biology of Cell 13 (2002) 1929–1939.