International Journal of Scientific Engineering and Research (IJSER)
www.ijser.in
ISSN (Online): 2347-3878, Impact Factor (2014): 3.05

A Verified Technique for Colon Cancer Analysis with Minimum Number of Features

Mohammed A. El-Shrkawey[1], Ben Bella S. Tawfik[2]

[1, 2] Suez Canal University, Faculty of Computers & Informatics, Information System Department, Ismailia 41522, Egypt

Abstract: ***Gene expression data is characterized by high dimensionality and small number of samples. Many researches work in data reduction, in other words selecting the most influence features (features selection). This work differs in verifying each step of selection; also, it reaches smaller number of features with high discrimination. Reducing data dimensionality lead to effective analysis of gene features. Actually, there is a tradeoff between feature selection and acceptable accuracy. The target is to find the compact set of features used for knowledge discovery and acceptable accuracy. So, we present a novel framework which integrates dimensionality reduction with classification for gene expression data analysis. In order to achieve our objective, we will use Oligonucleotide arrays. It provides a broad picture of the cell state by monitoring the expression level of thousands of genes at the same time. The developed techniques make to extract useful information from the resulting data sets. Gene expression is analyzed using 40 tumor and 22 normal colon tissue samples with 2000 human genes. The first phase of preprocessing, the introduced data is arranged and normalized. The second phase performs the features reduction in two steps. First step implements the features reduction from 2000 to 602 using t-test (lowest p-value). Second step, the reduction is implemented using sequential forward correlation which comes with only three gene features. With these only three genes a quadratic classification is done to test the features significance. The result of these classification attempt more than 96% of success.***

Keywords: Features selection, QDA classifier, gene expression, t-test, and p-value

1. Introduction

Cells are the smallest building blocks of the human body as all other organisms. In spite of cell is very minute and undetectable by eye, it is a huge industrial unit. Each cell contains the deoxyribonucleic acid (DNA) which the hereditary material in humans and almost all other organisms [1]. Every cell in the body has the same DNA that consists of Long chains of double stranded called chromosomes. Each chromosome carries thousands of genes that store the information responsible for defining traits and characteristics in living organisms. At its basic level, a gene codes for the creation of a protein via ribonucleic acid (RNA) molecule [2]. However, the investigation of the genes various phenomena are challenged by two basic problems [3], [4], [5]. The first of them is the exploiting efforts of exploration the huge amount of genes and their features. The second is caused by a biomarker (mainly gene or protein) that exists in the cell and does not function normally. Many Experimental approaches have been applied to overcome these problems. But, pure wet-lab experiments are generally not feasible due to the high dimensionality considered by the large number of genes [6]. So, multiples computational approaches are used to reduce the number of genes that could be used to classify samples into two groups, namely infected and normal [7], [8], [9].Therefore, reducing the number of features (dimensionality) is very important aspect in statistical of huge amount of gene expression data. So, Gene expression data analysis is an important research area that has attracted the attention of a of research groups as in[10], [11], [12]. Reducing features can also save storage and computation time and increase comprehensibility.

There are two main approaches to reducing features: feature selection and feature transformation [13]. Feature selection algorithms select a subset of features from the original feature set; feature transformation methods transform data from the original high-dimensional feature space to a new space with reduced dimensionality.

In this paper, we presents novel framework to integrate dimensionality reduction and classifying the gene expression data analysis. The rest of the paper is organized as follows. Section II, Data arrangement and preprocessing. Section III, Phase One, Feature reduction by estimating p-value (using t-test). Section IV, Phase Two, features reduction by removing the redundancy and selecting the most relevant features. Section V, analysis of the results is introduced. Section VI, Conclusion. Finally, section VII, References.

2. Preprocessing

Gene expression data (2000 genes for 62 samples) is obtained from the microarray experiments of Colon tissue samples of Alon et al. [14]. The total number of observations is 62. Hence, the data is arranged into two groups of observations 40 – 'Normal', and 22 – 'Cancer'. The data is divided into two equal groups Training and Testing. For each observation we got 2000 measures of gene concentration. The measures are normalized to get values from 0:1 each.

3. Implemented phases

3.1 Phase One: Features Reduction using p-value

Usually, Filters are used as a pre-processing step due to their simplicity and fast processing. A widely-used filter method for bioinformatics data is to apply a univariate criterion separately on each feature. This filter method assumes that there is no interaction between features. In this work, t-test is applied on each feature and compare p-value (or the absolute values of t-statistics) for each feature as a measure of how effective it is at separating groups. In order to get a general idea of how well-separated the two groups are performed by each feature. Figure 1, illustrates the number of features with

its p-values. There are about 15% of features having p-values close to zero and over 30% of features having p-values smaller than 0.05. This means, there are more than 602 features among the original 5000 features that have strong discrimination power. One can sort these features according to their p-values (or the absolute values of the t-statistic) to select some features from the sorted list. However, it is usually difficult to decide how many features are needed unless one has some domain knowledge or the maximum number of features. Generally, these features should be considered in advance based on outside constraints. The first ten features with the minimum p-value are shown in table 1.

Figure 2 shows number of used features versus the percentage of success. With only seven features more than 96% of success can be attempted.

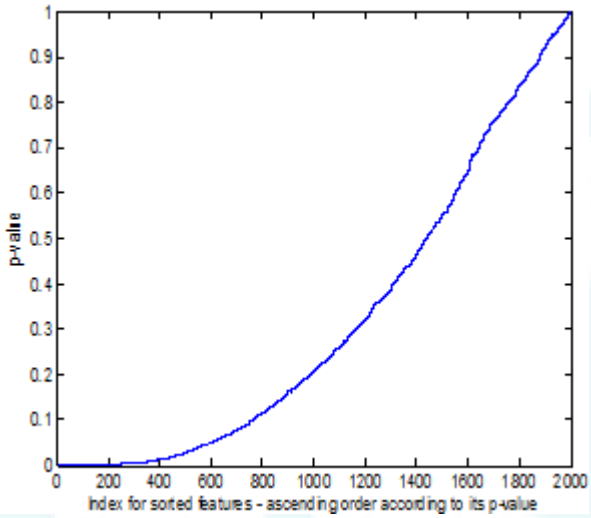

Figure 1: p-value for the 2000 features (sorted).

Table 1: The first 10 sorted p-value with its Gene (feature) number

p-value ($\times 10^{-6}$)	Gene number
0.0000	493
0.0001	377
0.0005	249
0.0009	1635
0.0127	1423
0.0159	625
0.0865	245
0.0951	1771
0.1552	765
0.1572	1772

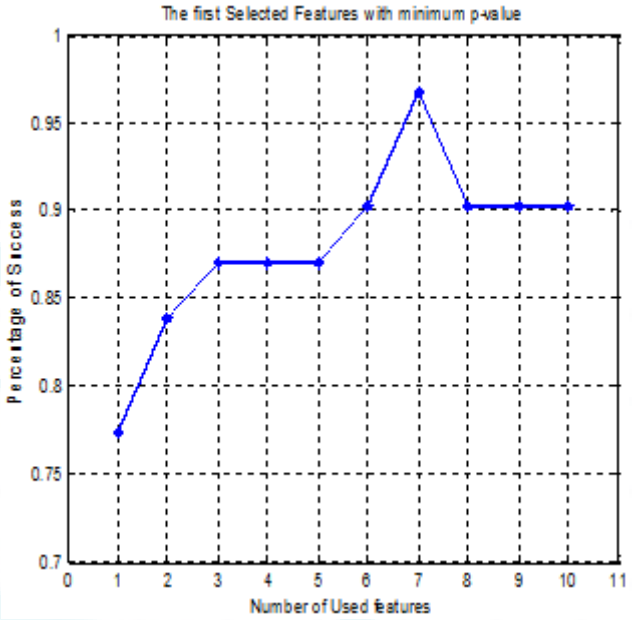

Figure 2: Number of used features versus the percentage of success

3.2 Phase Two: Features Reduction -removing redundancy

Strong relevance of a feature indicates that the feature is always necessary for an optimal subset. It cannot be removed without affecting the original conditional class distribution. On other hand, weak relevance suggests that the feature is not always necessary. But, it may become necessary for an optimal subset at certain conditions. Irrelevance indicates that the feature is not necessary at all. So, an optimal subset should include all strongly relevant features, a subset of weakly relevant features and none of irrelevant features. However, it is not given in the definitions which of weakly relevant features should be selected and which of them should be removed. Therefore, it is necessary to define feature redundancy among relevant features [15].Notions of feature redundancy are normally in terms of feature correlation. It is widely accepted that two features are redundant to each other if their values are completely correlated. In reality, it may not be so straightforward to determine feature redundancy when a feature is correlated (perhaps partially) with a set of features. We now formally define feature redundancy in order to devise an approach to explicitly identify and eliminate redundant features.

In summary, our method approximates relevance and redundancy analysis by selecting all predominant features and removing the rest features. It uses both C– and F-correlations to determine feature redundancy. Also, it combines sequential forward selection with elimination so that it not only circumvents full pair-wise F-correlation analysis but also achieves higher efficiency than pure sequential forward selection or backward elimination. The implementation of this phase is shown in figure3.We start with 602 features. But, after removing the redundancy we reached only three features. Figure 3 shows that the percentage of success is almost the same with using two or three features. Being relevant is not necessarily the same as being useful in this sense. Some researchers have argued that the presence of all relevant features may not be necessary

and may actually reduce performance. This is particularly the case for features that represent the same factor, or are correlated and so are redundant. Variables that are independently and identically distributed are not truly redundant. We use forward sequential feature selection in order to find important features. More specifically, since the typical goal of classification is to increase the percentage of success, the feature selection procedure performs a sequential search. The training set is used to select the features and to fit the Quadratic Discriminant Analysis QDA model, and the test set is used to evaluate the performance of the finally selected feature.

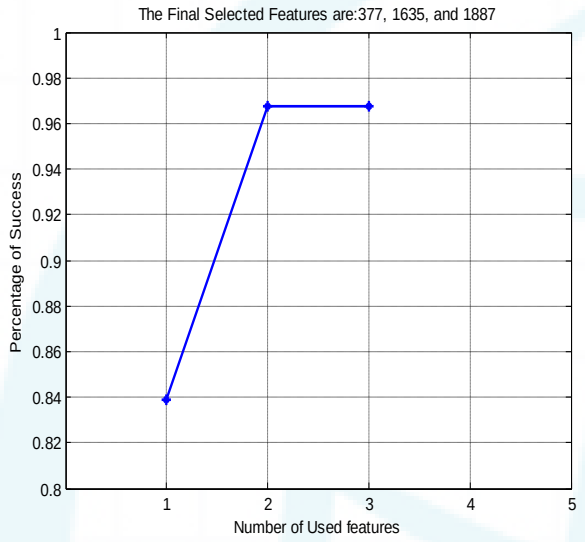

Figure 3: The final used features (three features) versus the percentage of success

After removing the redundancy only three features can reach more than 96% of success.

4. Conclusions

This work differs from other techniques of features selection in many items. Namely, it shows the importance of data preprocessing. Also, the model output reaches a smaller number of features. It shows also how the selected features attempt high classification high-dimensional data. More specifically, it shows how to perform sequential feature selection, which is one of the most popular feature selection algorithms. It also shows how to use the Quadratic Discriminant Analysis to verify the selection in each step. The selection is done in two phases staring from 2000 features space to 602 features to only three features. 96% of success is reached with these features.

References

[1] Z. Xia, L.-Y. Wu, X. Zhou and S.T.C. Wong, "Semi-supervised drug protein interaction prediction from heterogeneous biological spaces," BMC Systems Biology, 4(Suppl 2):S6, 2010.

[2] D. E. Krane and M. L. Raymer. Fundamental Concepts of Bioinformatics, San Francisco: Pearson Education, 2003.

[3] M.J. Keiser, et al., "Predicting new molecular targets for known drugs," Nature, 462(7270):175-181, 2009. R. Bijlani, et al., "Prediction of biologically significant components from microarray data: independently consistent expression discriminator(iced)," Bioinformatics, 19:62-70, 2003

[4] F. Chu and L. Wang, "Cancer classification with microarray data using support vector machines." Bioinformatics, 176:167-189, 2005.

[5] Qabaja, M. Alshalalfa, R. Alhajj and J. Rokne, "Multiagent Approach for Identifying Cancer Biomarkers," Proceeding of IEEE International Conference on Bioinformatics and Biomedicine, Nov 2009.

[6] A. Qabaja, M. Alshalalfa, R. Alhajj and J. Rokne, "Multiagent Approach for Identifying Cancer Biomarkers," Proceeding of IEEE International Conference on Bioinformatics and Biomedicine, Nov 2009.

[7] L. Wang, F. Chu, and W. Xie, Accurate cancer classification using expressions of very few genes." IEEE/ACM Transactions on Computational Biology and Bioinformatics, 4:40-53, 2007.

[8] M. Pirooznia, J. Yang, M. Yang and Y. Deng, "A comparative study of different machine learning methods on microarray gene expression data," BMC Genomics, 2007.

[9] R. Varshavsky, et al., "Novel unsupervised feature filtering of biological data," Bioinformatics, 22:507-513, 2006.

[10] L. Wang, F. Chu, and W. Xie, Accurate cancer classification using expressions of very few genes." IEEE/ACM Transactions on Computational Biology and Bioinformatics, 4:40-53, 2007.

[11] S. R. Setlur, et al., "Integrative Microarray Analysis of Pathways Dysregulated in Metastatic Prostate Cancer," Cancer Research, 67:10296, 2007.

[12] Md. MonirulKabir, Md. Shahjahan, Kazuyuki Murase , A New Local Search based Hybrid Genetic Algorithm for Feature Selection, Neurocomputing,Vol.74, Issue 17 ,2011, pp.2914-2928

[13] IffatA.Gheyas, Leslies. Smith, Feature subset selection in large dimensionality domains, Pattern Recognition, 2010, pp 5 – 13

[14] http://microarray.princeton.edu/oncology/affydata/index.html

[15] Lei Yu, and HuanLiu , 'Efficient Feature Selection via Analysis of Relevance and Redundancy', Journal of Machine Learning Research 5 (2004) 1205–1224.