Expert Systems with Applications 42 (2015) 864–872

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Data mining for feature selection in gene expression autism data

Tomasz Latkowski [a,1], Stanislaw Osowski [a,b,*]

[a] Military University of Technology, Faculty of Electronics, Kaliskiego 2, 00-908 Warsaw, Poland
[b] Warsaw University of Technology, Faculty of Electrical Engineering, Koszykowa 75, Warsaw, Poland

ARTICLE INFO

Article history:
Available online 6 September 2014

Keywords:
Gene expression microarrays
Feature selection
Clustering
Classification
Autism

ABSTRACT

The paper presents application of data mining methods for recognizing the most significant genes and gene sequences (treated as features) stored in a dataset of gene expression microarray. The investigations are performed for autism data. Few chosen methods of feature selection have been applied and their results integrated in the final outcome. In this way we find the contents of small set of the most important genes associated with autism. They have been applied in the classification procedure aimed on recognition of autism from reference group members. The results of numerical experiments concerning selection of the most important genes and classification of the cases on the basis of the selected genes will be discussed. The main contribution of the paper is in developing the fusion system of the results of many selection approaches into the final set, most closely associated with autism. We have also proposed special procedure of estimating the number of highest rank genes used in classification procedure.

© 2014 Elsevier Ltd. All rights reserved.

1. Introduction

Gene microarray technology is a sophisticated technique used in molecular biology for detecting alterations in the expression of thousands of genes simultaneously between different biological conditions (De Rinaldis, 2007). The analysis of the expression levels allows to detect altered gene expression of particular genes in a given disease when compared to healthy controls. From the practical point of view biologists need to identify only a small number of the most significant genes that can be used as biomarkers in the disease tracing. The most relevant genes increase our understanding of the mechanism of disease formation and allow to predict the potential danger of being affected by such disease.

The main problem in this analysis is a limited number of observations related to very large number of gene expressions. Number of observations is usually in the range of hundreds and number of genes tens of thousands. Because of the large imbalance of the number of genes and observations (patients) the selection is an ill conditioned problem. Moreover, data stored in medical databases are typically noisy and some gene sequences have large variance (Alter et al., 2011). It makes the gene selection in DNA microarrays even more difficult task.

Progress in bioengineering and data mining, which has been observed in recent years, has created the solid foundations for discovering the genes which are the best associated with the particular disease. Data analysis of microarrays is widely examined and introduced in literature starting with pioneering Golub investigation in 1999 (Golub et al., 1999).

Actual approaches performing this task include different clustering methods (Eisen, Spellman, & Brown, 1998), application of neural networks and Support Vector Machines (Alonso-González & Moro-Sancho, 2012; Guyon, Weston, Barnhill, & Vapnik, 2002; Wilinski & Osowski, 2012), statistical tests (Baldi & Long, 2001), linear regression methods applying forward and backward selection (Huang & Pan, 2003), fuzzy expert system based algorithms (Kumar, Victoire, Renukadevi, & Devaraj, 2012; Woolf & Wang, 2000), rough set theory (Wang & Gotoh, 2010), use of chaotic binary particle swarm optimization (Chuang, Yang, Wu, & Yang 2011), application of ReliefF method combined with different classifiers (Alonso-González & Moro-Sancho, 2012), various statistical methods (Golub et al., 1999; Mitsubayashi, Aso, Nagashima, & Okada, 2008), as well as fusion of many selection methods (Wilinski & Osowski, 2012; Yang, 2011). Most of the papers studied particular methods and then selected the best one as the most appropriate for the gene selection task, neglecting the others.

Although the progress in this field is high, there is still a need for better understanding and improvement of the research, especially in the medical area not well covered in recent research. To such examples belong autism data (Alter et al., 2011; Esteban & Wall, 2011; Hu & Yinglei, 2013). These data belong to the most

---

[*] Corresponding author at: Warsaw University of Technology, Faculty of Electrical Engineering, Koszykowa 75, Warsaw, Poland. Tel.: +48 22 234 7235; fax: +48 22 234 5642.

E-mail addresses: tlatkowski@wat.edu.pl (T. Latkowski), sto@iem.pw.edu.pl (S. Osowski).

[1] Tel.: +48 22 234 7235.

http://dx.doi.org/10.1016/j.eswa.2014.08.043
0957-4174/© 2014 Elsevier Ltd. All rights reserved.

demanding, because of very large variability of gene expression values representing the same class of data (Alter et al., 2011). The large variance in the distribution of gene expression levels is associated with many types of symptomatic profiles of autism represented in the base. Therefore, the application of standard methods, which serve very well in recognition of other cases, for example different types of cancer, does not lead to the acceptable results for autism.

Autism is a neurodevelopmental disorder that impairs the normal development of emotional interactions and other forms of social communication (Yang & Gill, 2007). Genetic approaches to autism study aim to identify risk variance at specific genes and in this way to find association of their expression level with the disease. There is a general idea that alterations at the level of gene expression might be important sign in mediating the risk for autism.

This paper is devoted to the task of selection of the genes and gene sequences which are the most closely associated with the disease. The selected genes of the particular expression levels form the most characteristic pattern for the autism. Applying a classifier to such chosen data, should lead to the improved accuracy of the recognition between autism and reference (healthy) cases. These two tasks (gene selection and classification problem) will be considered in the paper.

In the numerical experiments we will analyze different gene ranking methods. It is known that different selection algorithms may provide differing results for the same datasets (Wilinski & Osowski, 2012). The results of individual selection methods will be fused and lead to the final set of genes. The application of several methods gives opportunity to look on the selection problem from different points of view. After fusing their results the probability of proper selection of the most important genes is increased. The results of numerical experiments concerning selection of the most important genes in autism as well as classification of cases on the basis of the selected genes will be discussed.

The other contribution of the paper is developing the fusion system of the results of many selection approaches into the final set, most closely associated with the disease. This is in contrast to the majority of papers, where different methods have been tried, but only one (the best) was treated as the final solution. We have also proposed special procedure of estimating the number of higher rank genes using the self-organization procedure. In the task of classification we have implemented the ensemble of classifiers integrated into the final system, which is responsible for recognition of autism from the reference cases. The trained classifier system may then be used to predict the autism or non-autism class of the newly acquired data.

2. Applied feature selection methods

Feature selection is the most important operation in processing the data stored in gene microarrays. The application of feature selection methods allows to identify a small number of important genes that can be used as biomarkers of the appropriate disease. In this paper some chosen feature selection methods will be examined and integrated into the final system. Using the set of methods instead of single one will increase the probability of finding the globally optimal set of genes which are the best associated with the particular disease.

The paper will apply the following methods: Fisher discriminant analysis, ReliefF algorithm, two sample t-test, Kolmogorov–Smirnov test, Kruskal–Wallis test, stepwise regression method, feature correlation with a class and SVM recursive feature elimination. These methods rely their operation principle on different foundations and thank to this allow to look on the selection problem from different points of view.

2.1. Fisher discriminant analysis

In Fisher discriminant analysis the greatest weight is assigned to feature which is characterized by a large difference of the mean values in two studied classes and a small value of standard deviations within each class. The two class discrimination measure of the feature f is defined in the form (Duda, Hart, & Stork, 2003; Guyon & Elisseeff, 2003):

$$S_{12}(f) = \frac{|c_1 - c_2|}{\sigma_1 + \sigma_2} \qquad (1)$$

where c_1 and c_2 represent the mean values for classes 1 and 2, respectively, while σ_1 and σ_2 are the appropriate standard deviations. A large value of $S_{12}(f)$ indicates good class discriminative ability of the feature.

2.2. ReliefF algorithm

The ReliefF algorithm ranks the features according to its the highest correlation with the observed class while taking into account the distances between opposite classes (Robnik-Sikonja & Kononenko, 2003). The main idea of the ReliefF algorithm is to estimate the quality of the features according to how well their values distinguish between observations that are near to each other. ReliefF selects randomly an instance R_i of observation and then searches for k of its nearest neighbors from the same class, called nearest hits H_j and also k nearest neighbors from each of the different classes, called nearest misses $M_j(C)$. It updates the quality estimation $W(A)$ for all attributes A depending on their values for R_i, hits H_j and misses $M_j(C)$. If instances R_i and H_j have different values of the attribute A then the attribute A separates two instances with the same class which is not desirable. So the quality estimation $W(A)$ is decreased. If instances R_i and M_j have different values of the attribute A then this attribute separates two instances of different class values which is desirable. So the quality estimation $W(A)$ is increased. The algorithm averages the contribution of all hits and misses. The detailed description of the procedure can be found in Robnik-Sikonja and Kononenko (2003).

2.3. Two-sample t-test

The next used selection method is a two-sample Student t-test. One explicit assumption of t-test is that each of two compared populations of genes (autism and controls) should follow a normal distribution. Checking the condition of normality distribution of genes in our data base we found that in about 80% cases it was fulfilled. The null hypothesis of t-test is that data in the class 1 and 2 are independent random samples of normal distributions with equal means and equal but unknown variances against the alternative hypothesis that the means are not equal. The test statistic is formulated in the form

$$t = \frac{c_1 - c_2}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \qquad (2)$$

where and n and m represent the sample sizes of both classes (Sprent & Smeeton, 2007).

Two sample t-test is implemented in MATLAB as *ttest2* function (Matlab user manual – statistics toolbox, 2013). The test result returns h, which is equal 1 or 0. The value of 1 indicates a rejection of the null hypothesis at the 5% significance level, while $h = 0$ indicates a failure to reject the null hypothesis at the same significance level. The function returns also the p-value of the test. Low value of p indicates that the compared populations are significantly different.

دریافت فوری ← متن کامل مقاله

ISIArticles
مرجع مقالات تخصصی ایران

- ✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
- ✓ امکان دانلود نسخه ترجمه شده مقالات
- ✓ پذیرش سفارش ترجمه تخصصی
- ✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
- ✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
- ✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
- ✓ دانلود فوری مقاله پس از پرداخت آنلاین
- ✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات