Computers in Biology and Medicine 56 (2015) 82–88

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Computerized system for recognition of autism on the basis of gene expression microarray data

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

[a] Military University of Technology, Institute of Electronic Systems, Warsaw, Kaliskiego 2, Poland
[b] Warsaw University of Technology, Institute of the Theory of Electrical Engineering, Measurement and Information Systems, Warsaw, Koszykowa 75, Poland

ARTICLE INFO

Article history:
Received 13 June 2014
Accepted 2 November 2014

Keywords:
Gene expression microarray
Gene selection
Ensemble of classifiers
SVM
Random forest

ABSTRACT

The aim of this paper is to provide a means to recognize a case of autism using gene expression microarrays. The crucial task is to discover the most important genes which are strictly associated with autism. The paper presents an application of different methods of gene selection, to select the most representative input attributes for an ensemble of classifiers. The set of classifiers is responsible for distinguishing autism data from the reference class. Simultaneous application of a few gene selection methods enables analysis of the ill-conditioned gene expression matrix from different points of view. The results of selection combined with a genetic algorithm and SVM classifier have shown increased accuracy of autism recognition. Early recognition of autism is extremely important for treatment of children and increases the probability of their recovery and return to normal social communication. The results of this research can find practical application in early recognition of autism on the basis of gene expression microarray analysis.

1. Introduction

Autism spectrum disorders (ASD) are pervasive neurodevelopmental disorders that affect different aspects of human functions [1]. There are many studies [2–4] oriented at biomarker identification for ASD focused on genetic variants. An important research direction is studying the gene expression microarray in search for gene expression signatures which are informative with respect to identification of ASD.

Gene microarray technology is a technique for detecting alterations in the expression of thousands of genes simultaneously between different biological conditions [5]. The analysis of the expression levels allows to detect altered gene expression of particular genes in a considered disease when compared to healthy controls. The most relevant genes strictly associated with the mechanism of disease formation allow to predict the potential danger of being affected by such a disease.

The most important problem in this analysis is a small number of observations (usually in the range of hundreds) related to a very large number of gene expressions, usually tens of thousands. This considerable imbalance of the number of genes and observations makes the selection an ill-conditioned problem. Moreover, the data stored in medical databases related to autism are typically noisy and some gene sequences have large variance [6].

Recent progress in data mining techniques, which started from the pioneering work of Golub team [7], has laid solid foundations for discovering the genes which are best associated with a particular disease. Actual approaches to the task of gene selection include different clustering methods [8], application of neural networks and Support Vector Machines [9,10], statistical tests [11], linear regression methods applying forward and backward selection [12,13], fuzzy logic based algorithms [14], rough set theory [15], various statistical methods [8,16], as well as a combination of many selection methods [10,17].

The progress in this field depends on the type of investigated illness. ASDs belong to the most difficult cases because of large variation of gene expression levels among individuals [3,6,18]. The recent approach presented in the paper [3] is based on the analysis of data belonging to phenotypic subgroups. However, the actually reported accuracy of recognition between the reference class and the combined autistic group is still only 81.8% [3]. This rate is not satisfactory from the practical point of view.

This paper proposes a different approach to the problem. We apply many gene validation and selection methods cooperating with the support vector machine (SVM) classifiers. They form the so-called ensemble of classifiers integrated into the final system by a random forest of decision trees [19]. Applying different techniques

* Corresponding author at: Military University of Technology, Institute of Electronic Systems, Warsaw University of Technology, Institute of the Theory of Electrical Engineering, Measurement and Information Systems, Warsaw, Koszykowa 75, Poland. Tel.: +48 22 234 7235.
E-mail addresses: tlatkowski@wat.edu.pl (T. Latkowski), sto@iem.pw.edu.pl (S. Osowski).

http://dx.doi.org/10.1016/j.compbiomed.2014.11.004
0010-4825/

of gene selection allows looking on the selection problem from different points of view. After fusing their results into a single outcome the probability of proper recognition of classes is increased. We will demonstrate that our approach is able to increase significantly the autism recognition accuracy.

The most important contribution of the paper is developing the fusion system of the results of many selection/classification approaches to the final decision of the increased accuracy. The presented approach is in contrast to the majority of papers, where different methods have been tried, but only one (the best one) was used as the final solution. The results of numerical experiments performed on the NCBI autism database [20] have confirmed the superiority of the proposed approach.

2. Materials

The numerical experiments in autism recognition have been performed on a publicly available database, downloaded from the GEO (NCBI) repository [20]. The number of observations in this dataset equals 146 and the number of genes—54,613. The database consists of two classes. The first one is related to children with autism and the number of such instances is $n=82$. The second (control) group is composed of healthy children and the number of such instances is $n=64$.

All subjects in the base are male. Probands and controls were all recruited from the Phoenix area. Blood draws for all subjects were done between the spring and summer of 2004. Total RNA was extracted for microarray experiments with Affymetrix Human U133 Plus 2.0 39 Expression Arrays. Children with autism were diagnosed by a medical professional (developmental pediatrician, psychologist or child psychiatrist) according to the DSM-IV criteria and the diagnosis was confirmed on the basis of ADOS and ADI-R criteria [21]. In an attempt to obtain a homogenous population of children with autism, non-classic forms of autism were excluded, including autism with regression and Asperger's syndrome, a higher functioning form of autism where individuals have language skills within normal range. In addition, each subject had a normal high-resolution chromosome analysis, and a negative Fragile X DNA test.

IQ scores of autistic and control children were not done for this study, but all individuals did demonstrate a language impairment as part of the diagnostic criteria. For additional analysis, paternal ages were available for 78 children with autism and 57 controls. The group of children with autism was significantly younger than the control group (autism: mean—5.5 years SD—2.1; control: mean—7.9 SD—2.2). Paternal age was similar between the groups. The study population was primarily Caucasian and there were no group level differences in ethnicity. The box plots of the age of children and their parents are presented in Fig. 1.

Our main task is to build a computer program that would be able to separate the autistic group from the control group with the highest accuracy on the basis of expression levels only. The labels of the testing group of subjects were not known for the program. We solved the problem in two phases. In the first one we applied a few methods of selecting small but optimal subsets of genes with good class discriminative abilities. In the next phase these genes were used as the input attributes to the SVM classifiers forming an ensemble. The final decision on the ensemble was worked out by an additional random forest classifier.

3. Genetic approach to gene selection

In the task consisting in assessment of the class discrimination ability of genes we treated each gene as the diagnostic feature and applied well-known feature validation methods to solve the selection problem. The following methods were applied: 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. A short description of them is provided in the Appendix. The operation principle of these methods relies on different foundations which allows to see the selection problem from different points of view. As a result of application of these methods we get 8 different sets of genes ordered according to their class discrimination ability, from the most to the least discriminative. However, at this stage none of the methods indicates their optimal size.

The important task was to find the optimal population size of genes that will guarantee the best performance in the classification stage. To solve this problem we performed the following calculations using only a limited number of the most significant genes, selected at the first stage of validation. This task was done using the genetic algorithm [22,23] cooperating with the Support Vector Machine of Gaussian kernel [24].

Only 100 best genes were used at this stage of processing. We limited this number to optimize the performance and increase the speed of the genetic algorithm. It is known that the classifier of good generalization ability should be supplied by the limited number of input attributes (usually no more than few tens). Therefore, 100 best genes from each of the selection procedure should be enough to provide vast choice for genetic operations. The application of genetic algorithm was repeated separately for all 8 sets of genes chosen in the course of individual selection procedures.

In this solution we used a binary code representation of an individual gene. The value 1 means the inclusion of the gene while zero indicates the lack of this gene in the input vector $\mathbf{x}$ to the classifier. In all experiments an elitist strategy of passing two fittest population members to the next generation was used. This guarantees that the fitness is never declined from one generation to the next, which is a desirable property in our application. The algorithm created crossover children by combining pairs of parents in the current population using the roulette rule. The crossover probability applied in the solution was 0.8. The mutation of children was created by randomly changing the genes of individual parents. The assumed mutation rate was 0.03.

Each chromosome in the genetic algorithm is associated with the input vector $\mathbf{x}$ applied to the SVM classifier. The value 1 of the particular element of the chromosome vector means real inclusion of the gene and zero—no such gene in the actual vector. Two data sets were involved in the GA based training: the learning set and the validation set extracted from the learning one (20% of the learning data). The classifier is trained on the learning data and then tested on the validation data set. The testing error on the validation data forms the basis for the definition of the fitness function. Fitness is defined as the error function taken with a minus sign. The genetic algorithm maximizes the value of the fitness function (equivalent to the minimization of the error function) by performing subsequent operations of selection of parents, crossover among parents and finally mutation.

The described process is repeated until a termination condition is reached. The applied terminating conditions are as follows: a solution is found that satisfies minimum criteria, fixed number of generations is reached, allocated computation time is reached, the highest ranking solution's fitness is reached or a plateau such that successive iterations no longer produce better results is reached. In our approach fitness function is directly defined on the basis of a 10-fold cross validation error of the SVM classification system with the Gaussian kernel.

Genetic algorithm presented a very effective way of finding the best set of the most significant features. Application of the GA to feature selection was found superior to other methods. GA performs two tasks simultaneously. It selects the most important features in the predefined set and at the same time determines