Selection of Informative Gene on Autism Using Statistical and Machine Learning Methods

***Nur Amalina Rupawon*[1], *Zuraini Ali Shah**[2]**

Department of Software Engineering, Faculty of Computing, Universiti Teknologi Malaysia, 81310 Johor Bahru, Johor, Malaysia

[1]*nurama92@gmail.com*, [2]*aszuraini@utm.my*

Abstract

Disorder of neurodevelopmental in autism might be caused by the mutations in multiple general transcriptions. This may risks to a disease in sporadic cases of autism that probably associated with the alterations of the regulation in expression of genes at the global levels. The selection of informative gene on autism is one of the ways to test the premises which will produce an evaluation about the distribution of the gene expression level. The most major subject of this research that requires for the selection of gene on autism is about finding the significant gene. Thus, this thesis presents the application of different statistical and machine learning methods for the recognition of the most significant genes that related to autism. Three different statistical methods are used for the statistical analysis using Fisher's discriminant analysis, two sample t-test and feature correlation with class. The machine learning approaches involves the application of genetic algorithm hybrid with three different classifiers such K- nearest Neighbor (KNN), Support Vector Machine (SVM) and Linear Discriminant Analysis (LDA). Results from every method is examined through the profiles of gene expression analysis and followed by the comparative studies about the performance of the gene selection based on the different approached methods.

Keywords: Gene Selection, Gene Expression of Autism, Statistical Methods, Machine Learning

1.0 Introduction

Autism is a severe neurodevelopment disorder that related to high degree of heritability. People with autism usually characterize such as having social and communication deficits instead of ritualistic and monotonous behavior. Autism can be divided into two that consist of minority and majority cases. The identification of small number of genetic mutations is accounts to minority cases meanwhile sporadic are counterpart (Alter *et al*., 2011). The causes that lead to autism can be both environmental factors and genetic susceptibility (Stephan, 2008).

On the other hand, gene expression is the process which genetic instructions are used to synthesize gene products. Protein is the usual product that been synthesized performing the essential functions as enzymes, hormones and receptors. All living cells' functions and adaptability are controlled by the highly regulated mechanism of gene expression. There are many methods that can be used to study and quantify the regulation of gene expression such as in DNA microarray. DNA microarray is one of the methods that been used widely in field related to the molecular biology. Practically, DNA microarray is a sophisticated method that useful for analyzing the expression of genes in specific cells under specific condition at a given time (Rinaldis, 2007). This method also consists of several down to it. This is because the method produces huge information of dataset as well as has high dimensionality (Ammu *et al*., 2013; Latkowski *et al*., 2014). In addition, the fact that the numbers of parameters in samples commonly are very small as compared to the number of samples (Osowski, 2014) used for classifier training may lead the classifier to be over fitting.

Other than that the other problems that bring the difficulties for gene selection in DNA microarray is that data stored in medical databases are normally noisy instead of having a large variance for some of the gene sequences (Wang *et al*., 2010). Thus, such problems is requires to be overcome with the purpose of discovering biomarkers that related to a specific gene. This is because biomarkers discovery will allow the production of an observation which can provide more understanding and deeper finding of the relationship of such markers with a disease (Wang *et al*., 2010). One of the ways to achieve the goal of learning the knowledge about the markers is through the selection of the small numbers of significant genes. The reason is that it can be consider as one of the way that can be associated to the process of tracing the disease (Latkowski *et al*., 2014).

A part from that, one of the steps to alleviate such problems is through the employment of the feature selection (FS) methods. Feature selection is the process of selecting a subset of relevant features that useful in a construction of a model. It is useful when a data is assumed to have irrelevant or redundant features. Irrelevant features can be define as an invaluable information that related to a context meanwhile redundant features provide repetitive or no more information than the currently selected features. Feature selection methods also known as variable selection or attribute selection. The variables of an original representation are not altered when using these methods instead it is simply selecting a subset of them. (Saeys *et al*., 2007). Thus, these methods are able to offer the advantages of interpretability by a domain expert due to its preservation towards the original semantics of a variable.

This research is aiming on selecting small and significant genes in gene expression of autism through the application of different feature selection. Three objectives are set to satisfied the aim of this research by; (i) implement statistical method using Fisher's discriminant analysis, two sample t-test and feature correlation with class to select small number of genes expression of autism (ii) apply statistical and machine learning methods to select the informative gene in gene expression of autism (iii) measure the performance of the selected gene based on different approached methods.

2.0 Gene Selection Methods

2.1 Autism Dataset

This research will present the problem of gene selection applied to gene expression of autism dataset. The dataset applied to the research is GDS4431 that can be retrieved and downloaded from publicly available database at GEO (NCBI) repository (http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS4431). The database can be divided into two classes that consist of children with autism (n=82) and normal (healthy) children (n=64). It is represented as autism and control respectively in the database. In addition, the type of dataset is expression profiling by array with count of 146 samples with total number of 54613 genes. Total RNA was extracted with Affymetrix Human U133 Plus 2.0 39 Expression Array (NCBO base, 2011).

2.2 First Stage Selection

Data with 54613 numbers of genes are reduced into 250 genes by log transformation using application of GEO2R tool. The reduced genes used as input data for the first stage of selection using statistical methods. Three methods are used in this stage consist of fisher discriminant analysis (FDA), two sample t-test (TT) and feature correlation with class (COR) which are able to provide higher statistical independence. In Fisher test, discriminative ability is evaluated based on the value of its discrimination measure. A higher value of this measure represents a good discrimination. Parameters are set to represent the mean values, standard deviations of the class 1 and class 2 respectively. The test that can be represented in the following form (Osowski, 2013):

$$S(f) = \frac{|c_{1-}\, c_2|}{\sigma_1 + \sigma_2} \tag{3.1}$$

The c_1 and c_2 define as mean values of class 1 and class 2 respectively while σ_1 and $\sigma_2$are standard deviation of the appropriate classes. $S(f)$ is used to represent the discriminative ability of the features. Specifically, higher value of $S(f)$ considers significant discrimination for gene recognition from both classes.

Two samples t-test is used to determine either means of two different populations is same. It returns *h,* which is equal 1 or 0. The value of 1 indicates a rejection of the null hypothesis at the 5% significance level otherwise it indicates a failure to reject the null hypothesis at the same significance level. The function returns the *p*-value of the test. The lower value of p indicates that the compared populations are significantly different. The statistical test is formulated as

$$t = \frac{x_1 - x_2}{\sqrt{\frac{{\sigma_1}^2}{n} + \frac{{\sigma_2}^2}{m}}} \tag{3.2}$$

The formula is described as x_1, x_2, ${\sigma_1}^2$, ${\sigma_2}^2$ and n and m that represent the class 1 and class 2, variance class 1 and variance class 2 and samples sizes of both classes respectively . TT has been implemented in MATLAB as *ttest2* function (Matlab User Manual, 2012).

In feature correlation with class method, the correlation of gene f with a uniform distribution of two classes is examined by the formula describes below (Osowski, 2013):

$$S(f) = \frac{\left(m_1(f) - m(f)\right)^2 + \left(m_2(f) - m(f)\right)^2}{2\sigma^2(f)} \tag{3.3}$$

where m, m_1 and m_2 is mean value of feature for all data, feature of class 1 and feature of class 2 respectively. Meanwhile, $\sigma^2(f)$ represents the variance of the feature. The large value of $S(f)$ represents a good discriminative ability of feature in the two classes' recognition. Each statistical method ranks and selects the 100 best of genes with the purpose of identifying informative gene related to the autism.

3.0 Second Stage Selection

Meanwhile, machine learning is applied to optimize the selection of the best genes selected in the previous stage. At this stage of selection, Genetic Algorithm (GA) works as the genetic selection to the subset of data and three different types of classifiers are applied to evaluate the accuracy and performance of the selected genes in classification. The classifiers used for the classification in this research consist of K-nearest Neighbor (KNN), Support vector Machine (SVM) and Linear Discriminant Analysis (LDA).Using GA approach the gene is coded in a binary way where value one (1) work as input signal to the classifiers that represent significant gene related to autism. Meanwhile negative one (-1) means exclusion of the particular gene as input signal. This method of GA is reproduces by selecting parent and performs crossover with mutation assigns to the bit which representing the children. The process occurs randomly in generated chromosomes. A fitness function is determined and evaluated to select chromosomes from the current population forming a new population. The new population is used in the next iteration of the algorithm.

Briefly, the gene selection processes that involves in all stages can be represented as in the figure below:

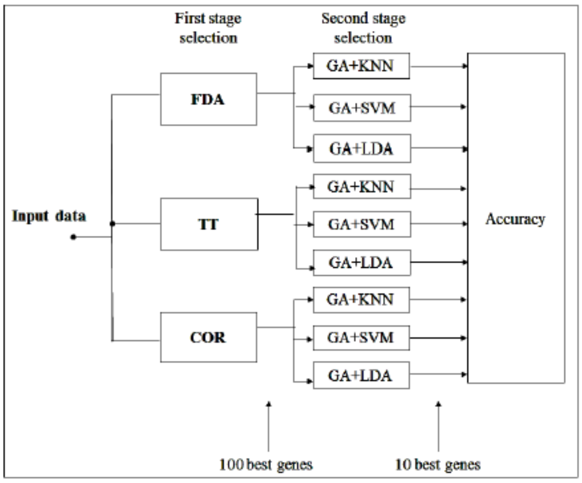

Figure 6 General scheme of selecting gene of autism

4.0 Result and Discussion

At first stage of selection, the statistical methods are applied into dataset to get the order of gene. Gene is sorted based on the significant values and as the result, 100 best of genes are selected by each individual method. The implementation of each methods produce different set of gene during selection. A table is provided to display the percentage of identical genes among the 100 best genes selected by every method.

Table 1 Overlapping percentage of selected 100 best genes by different statistical methods

Methods	**FDA**	**TT**	**COR**
FDA	100	98	98
TT	98	100	100
COR	98	100	100

Based on the table, it shows that the three methods are able to select set with high number of identical genes. The overlapping result between method TT and COR cover 100% of similarity. Meanwhile, the result shown by method FDA against TT and COR are 98% of similarity respectively. In comparison, the overlapping percentages produced by TT and COR in this paper are similar to previous paper Latkowski *et al.* (2014) which is 100%. However, there is slightly different overlapping percentage produced by FDA in this research. The result shows that FDA produced 8% more of identical genes than in Latkowski *et al.* (2014). The overlapping percentage is crucial to increase the quality of selection. The quality of selection can be assessed by analyzing the expression profiles of selected genes. The quality of selected genes is indicated by the difference of means value between two classes. The comparison of the quality of selection for significant and insignificant gene is presented in the table below.

Table 2 Difference of mean value in two case of gene selection.

Case of selection	The most significant gene		The least significant gene	
Class	Autism	Control	Autism	Control
Mean value	60.94	53.74	9.13	9.16
Standard deviation	9.50	7.00	1.13	1.49
Difference of mean value	7.20		0.03	

Based on the table it shows that the mean between the class of autism and control are significantly differs in the case of selection involved the most significant gene. Compare to the case of selection related to the least significant gene, the mean between the two classes are barely noticeable.

The 100 best genes selected by each statistical method are used as input attributes in the second stage of selection. In this stage, the genetic algorithm is combined with KNN, SVM and LDA classifiers and set with two different parameters; Parameter 1 and Parameter 2. The parameter consists of crossover probability and mutation rate are set as 0.8 and 0.02 and 0.3 and 0.5 in Parameter 1 and Parameter 2 respectively. Parameter 1 is set based on Latkowski *et al. (*2015) that applied GA combined with SVM classifier in their study and Parameter 2 is our own new set parameter. The optimal number of genes require to be selected by the algorithm is set as 10. All selected genes created common input to classifiers and 10-fold cross-validation process. The cross-validation process produce the percentage of the mean error for every classification involved. Table below shows the result of the accuracy produced by different application of classifiers in two different parameters.

Table 3 Result of accuracy based on different parameters and statistical methods input data

	Parameter 1			Parameter 2		
	FDA	**TT**	**COR**	**FDA**	**TT**	**COR**
Classifiers						
KNN	81.27	81.58	80.41	81.54	81.47	80.72
SVM	82.81	81.51	81.58	82.81	81.51	81.58
LDA	89.15	88.84	88.74	88.70	88.94	88.60

Based on the table, the differences of accuracy produced in comparison to both parameters are insignificantly displayed. However, in comparison of different classifiers it shows that LDA classifier able to produce the highest accuracy for both parameters using input data from all the statistical methods of this research. The result outperforms the best result of 77.05% error reported by the paper Latkowski *et al*. (2015) for the autism data base. It shown that methods used in this research are able to produce better result than Latkowski *et al*. (2014) and Latkowski *et al*. (2015) for appropriate statistical and machine learning methods. In biological aspect, among the 10 informative genes selected by the algorithm the commonly selected genes by each method such as: HIST1H2BG (histone cluster 1, H2bj), RAB8A (RAB8A, member RAS oncogene family) and FGF1 (fibroblast growth factor 1 (acidic)). Gene of HIST1H2BG shows the highest repetition of gene identified by different classifier using different input data which is 66.67% of selection frequency.

5.0 Conclusion

This paper focus on the selection of gene related with the gene microarray of autism. Several methods for data mining has been examined and applied into this study. The main purposed is to produce small number of subset that can be significantly associated to the autism. The selection of the important genes involved two stage of selection. First stage implements the statistical methods to identify the 100 best of genes. The next stage used the machine learning approaches to select the top 10 genes that related to autism by using optimization algorithm and classification. This research implements and adapting method used in the previous researches by Latkowski *et al*. (2014) and Latkowski *et al*.(2015) with some modification and improvement. This research proposed a simpler way to select the informative genes of autism in gene microarray data by imply two stage of selection. Though this research use simpler approaches but the result produced shows its ability of producing better performances comparing with the more complicated process proposed by the previous research. In addition, with the application of many classifiers it displays the effectiveness of different methods towards dataset of autism. Thus, this open up the opportunity for better analysis related to autism. Instead, for future planning and for the sake of improving result gained by this research we would like to suggest additional filtering during pre-processing. Multiple filtering seems to be a better way of improving the performance measure of this dataset due to the fact that this microarray data of autism has a very large variance and dimensionality. Lastly, different dataset of autism also can be applied into this approaches to prove the effectiveness of this methods used in the research for different cases.

References

Alter M. D., Kharkar R., Ramsey K. E., Craig D. W., Melmed R. D., Grebe T., Curtis-Bay R., Ober-reynolds S., Kirwan J., Jones J., Blake- Turner J., Hen R., Stephan D.and Stephan, D. A. (2011). Autism and increased paternal age related changes in global levels of gene expression regulation. PloS one, 6(2), e16715.

Ammu, P. K., and Preeja, V. (2013). Review on Feature Selection Techniques of DNA Microarray Data. International Journal of Computer Applications (0975–8887) vol, 39-44.

http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS4431

Latkowski T., and Osowski S. (2014). Feature selection methods in application to gene expression: autism data.

Latkowski T., and Osowski S. (2015). Developing Gene Classifier System for Autism Recognition. In Advances in Computational Intelligence (pp. 3-14). Springer International Publishing.

Osowski S. (2013). Methods and tools in data mining. BTC, Warsaw.

Rinaldis D E., and Lahm A. (Eds.). (2007). DNA microarrays: current applications. Horizon Scientific Press.

Saeys Y., Inza I., and Larrañaga P. (2007). A review of feature selection techniques in bioinformatics. bioinformatics, 23(19), 2507-2517.

Stephan D. A. (2008). Unravelling autism. The American Journal of Human Genetics, 82(1), 7-9.

Wang X., and Gotoh O. (2010). A robust gene selection method for microarray-based cancer classification. Cancer informatics, 9, 15.