Applying the Fisher score to identify Alzheimer's disease-related genes

J. Yang[1,3], Y.L. Liu[2], C.S. Feng[1] and G.Q. Zhu[1]

[1]College of Computer Science, Sichuan Normal University, Chengdu, China
[2]College of Mathematics and Software, Sichuan Normal University, Chengdu, China
[3]Visual Computing and Virtual Reality Key Laboratory of Sichuan Province, Chengdu, China

Corresponding author: J. Yang
E-mail: yanghuajun@163.com

Genet. Mol. Res. 15 (2): gmr.15028798
Received May 16, 2016
Accepted June 3, 2016
Published June 24, 2016
DOI http://dx.doi.org/10.4238/gmr.15028798

ABSTRACT. Biologists and scientists can use the data from Alzheimer's disease (AD) gene expression microarrays to mine AD disease-related genes. Because of disadvantages such as small sample sizes, high dimensionality, and a high level of noise, it is difficult to obtain accurate and meaningful biological information from gene expression profiles. In this paper, we present a novel approach for utilizing AD microarray data to identify the morbigenous genes. The Fisher score, a classical feature selection method, is utilized to evaluate the importance of each gene. Genes with a large between-classes variance and small within-class variance are selected as candidate morbigenous genes. The results using an AD dataset show that the proposed approach is effective for gene selection. Satisfactory accuracy can be achieved by using only a small number of selected genes.

Key words: Alzheimer's disease; Fisher score; Feature selection; Gene microarray

INTRODUCTION

Alzheimer's disease (AD) (Liang et al., 2008) is a form of dementia. This neurodegenerative brain disease usually develops at an age over 65 years. Because it is irreversible, the disease causes extensive damage to the patient's health.

Although many studies on AD have been carried out, the cause and mechanism for the progression of AD are not well understood. A full understanding of the potential molecular mechanisms would provide key information to enable the successful treatment of AD. In particular, identifying genes that are involved in AD could be beneficial to both explaining the causes of the disease and designing treatments.

Recently, advances in gene microarray technology have enabled biologists to measure the expression level values of many genes simultaneously in one experiment, which provides an opportunity for machine learning methods to be used to extract valuable biological information from these large datasets (Schena et al., 1995). By analyzing these high-throughput microarray gene databases, researchers are aiming to gain deep insights into the causes, processes, and biological mechanisms of human diseases.

Using AD microarray data, researchers have developed various methods for exploring the genes associated with the disease. Clustering analysis technology (Pang et al., 2010; Guttula et al., 2012; Hu et al., 2013; Yang et al., 2013) has often been employed to cluster the genes by their expression level and to identify co-expressed genes in the same group. A special local clustering algorithm (Pang et al., 2010) was used to cluster the gene data and identify the so-called isolated points genes with significantly different expression values. Gene order computing (Hu et al., 2013) was utilized to generate higher quality gene clustering patterns than most other clustering methods. Fuzzy cluster analysis (Yang et al., 2013) was used to analyze gene data by grouping gene sequences together that were expected to have close relationships each other and similar functions and characters. A hierarchical cluster analysis (Guttula et al., 2012) was performed to group genes based on their expression pattern. Independent component analysis was employed to reveal meaningful biological patterns in AD gene expression data (Wei et al., 2009).

In this paper, we present a novel approach for utilizing AD microarray data to identify the morbigenous genes. Different from the methods described above, we classify the data without using clustering approaches.

We evaluate the ability of a gene to distinguish normal individuals from AD patients by its expression value. The genes with a strong ability to distinguish samples between the two classes are regarded as the key genes for the disease. The problem of identifying morbigenous genes is transformed to a problem of feature selection, which involves seeking the best feature subset to differentiate samples from the different classes.

The rationality of this approach is that the number of samples in gene microarray data is usually far less than the number of genes (Dougherty, 2001), and many genes are irrelevant to the disease. Gene selection mainly has two merits (Peng et al., 2005; Saeys et al., 2007). First, it can reduce dramatically the number of genes used in classifying the disease and overcome the problem of the "curse of dimensionality". Second, the selected genes are likely to be biologically relevant to AD and can be further explored. Those explorations may help to better understand the essential molecular mechanisms associated with AD. For these reasons, microarray data have been widely used for classifying diseases, and as a result, many approaches have been proposed to classify cancer using microarray data (Veer et al., 1981; Zirvi et al., 1989; Shipp et al., 2002; Rifkin et al., 2003).

The Fisher criterion was used to assess the capacity of genes to classify samples in this paper. The results from the experiment using AD gene data show that the selected feature genes were able to accurately predict the type of the testing samples.

MATERIAL AND METHODS

Gene selection by Fisher score

The Fisher score (FS) is a supervised feature selection technique. In this section, we briefly review the principle of the Fisher criterion for feature selection.

The generic problem of supervised feature selection is as follows. Given a sample set $\{(x_i, y_i)\}$, where $i \in 1..n$, $x_i \in R^d$, $y_i \in \{1,2,...,C\}$, and y_i denotes the class label of the sample x_i; n is the number of samples; and d is the dimension of the features, that is, the number of genes, the aim is to construct a subset with m features that contains the most discriminative information. We use $X = [x_1, x_2, ..., x_n] \in R^{d \times n}$ to represent the sample data matrix, where x_j denotes the *j*th sample.

The core idea of Fisher discrimination analysis is to construct a subset of features such that in the data space spanned by the features in the subset, the distances between samples from different classes are as large as possible, while the distances between samples from the same class are as small as possible. In particular, when m features are selected, the original data matrix $X \in R^{d \times n}$ will be represented by $Z \in R^{m \times n}$. Then, the FS is computed as follows,

$$f(Z) = tr(A_b) / tr(A_w) \quad \text{(Equation 1)}$$

where $tr(\ )$ denotes the trace of a matrix; A_b is the between-class scatter matrix; and A_w is the within-class scatter matrix, which is defined as

$$A_b = \frac{1}{n} \sum_{k=1}^{c} n_k \left(m_k - m\right)\left(m_k - m\right)^T \quad \text{(Equation 2)}$$

where m_k and n_k are the mean and sample number of the k-th class, respectively, in the reduced data space and *m* is the mean vector of all samples.

The number of candidate subsets is C_d^m, so the optimal feature subset selection problem can be solved by combination optimization, but this is very time-consuming and challenging. To reduce the difficulty, a heuristic strategy is often used to calculate a score for each feature independently using some criterion f. Specifically, let u_k^j and σ_k^j be the mean and deviation of samples from the *k*-th class, corresponding to the *j*-th feature. Let u^j and σ^j denote the mean and deviation of the whole samples corresponding to the *j*-th feature. Then, the FS of the *j*-th feature is calculated as follows,

$$f(j) = \frac{\sum_{k=1}^{C} n_k \left(u_k^j - u^j\right)^2}{\sum_{k=1}^{C} n_k (\sigma_k^j)^2} \quad \text{(Equation 3)}$$

After obtaining the FS of each feature, we select the features with first-m large scores to construct the feature subset.

The details of feature selection are described in Algorithm 1.

Algorithm 1: feature selection by FS

Input: the samples $X \in R^{d \times n}$ and their class label $y \in N^n$; the expected feature number m

Output: the selected feature subset $S \in N^m$

Begin

For each feature in the feature space

Evaluate the corresponding FS by Equation 3 and record the score in a score array.

End

Sort the score by descending order.

Select the top-m features with a high score and place their feature index into the set S.

End

RESULTS

Dataset and procedure

We used the proposed feature selection approach for gene selection to classify or predict the status of individuals, i.e., normal or AD disease.

We used a dataset from GEO Datasets deposited by Blalock et al. that featured hippocampus gene expression from control and AD samples (Blalock et al., 2004). The hippocampal specimens were obtained from the Brain Bank of the Alzheimer's Disease Research Center at the University of Kentucky. The human Gene Chips (HG-U133A) of Affymetrix and Microarray Suite 5 were used in the microarray data collection. The procedures for total RNA isolation, labeling, and microarray construction were described previously (Blalock et al., 2003, 2004).

There are 31 samples, and each sample contains 22,283 gene expressions. Four cases of control, incipient, moderate, and severe data are provided in the original data. Here, we perform the classification experiment for the control and moderate classes, so the 9 control samples and 8 moderate samples are selected to form the dataset X. Then, the FSs are evaluated in X, and the features (the key genes) are selected. With these gene expressions, we use some classifier to classify the testing sample and calculate the recognition accuracy. Because the number of samples in gene microarray data is generally far less than the number of genes, we construct the testing set using the leave-one-out method. Specifically, each time, we select one sample from X as a testing sample and the remainder as the training samples. The procedure is repeated until every sample in X has been selected as a testing sample and classified. Then, we compute the recognition accuracy by Equation 4.

$$ac = (nc \,/\, n) \times 100\% \qquad \text{(Equation 4)}$$

where *nc* is the number of correct classifications and *n* is the number of testing samples. Because the main goal is to measure the identification ability of genes for AD, simple classifiers such as the nearest neighbor classifier (NNC) or nearest class mean classifier (NMC) are employed in

the experiment. The detailed procedure for classifying a testing sample by NNC is described in Algorithm 2.

Algorithm 2: classifying a testing sample by NNC

Input: training samples and their label $\{x_i, y_i\}$; the testing sample t ; selected feature set $S \in N^m$

Output: the predicted class label $c(t)$ of the testing sample.

Begin

By S , retain the m gene expressions for every x_i and t .

For every training sample x_i

Compute the distance between x_i and t by some distance measure.

Store the distance in the distance array.

End

Sort the distance array by ascending order.

Set $c(t)$ as the label of the sample at the first position in the distance array.

End

Experiment results

First, we calculate the FS of genes and sort their scores by descending order. The sorted scores are shown in Figure 1.

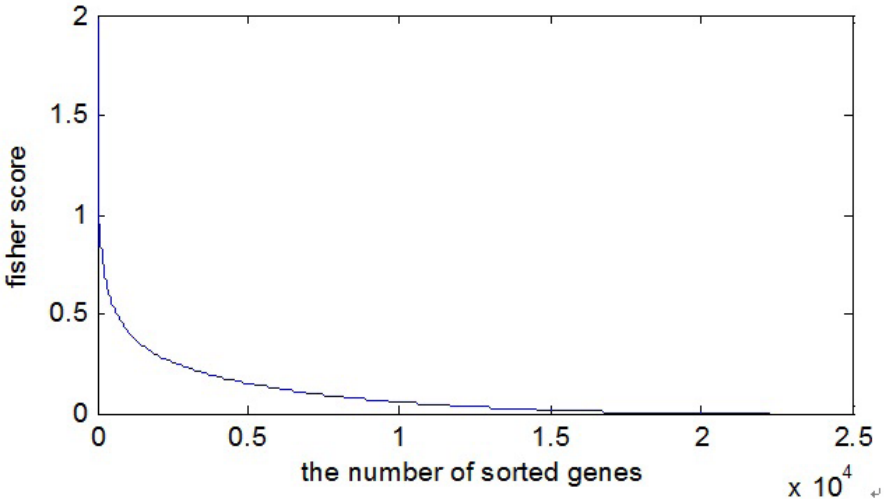

Figure 1. Fisher score of genes.

From Figure 1, we can see most of the genes' scores are close to zero. That means those genes are irrelevant to the classification or the disease. Then, we use the genes with high scores as features to classify or recognize the testing sample. The results are shown in Table 1. Here, the distance measure is the Euclidean distance, following Eq. (5). In Table 1, we also present the calculated recognition accuracies based on the feature genes obtained by the normalized mutual information (NMI) approach proposed previously (Peng et al., 2005; Liu et al., 2005).

$$d(x_i,t)=\sqrt{(x_i-t)^T(x_i-t)} \quad \text{(Equation 5)}$$

From Table 1, we can see that when all genes are selected as features, the accuracy of NNC is 52.94%, and the accuracy of NNC is 76.47% if only one gene is selected as a feature. With more than one gene, the accuracy can reach 100%. These results indicate that many genes are irrelevant to the disease and that a relatively inferior accuracy will be obtained if those genes are selected. Moreover, we find that few genes are informative for classification. When using only 10 genes, the NNC can reach 100% accuracy. The result from NMC is similar. The first gene selected by the mutual information approach is closely related to the class label, and the accuracy of NMI+NNC can reach 94% using only the first gene. When more genes are considered as features, the FS method is superior to the mutual information method. When 10 genes are selected as features, the accuracy of FS+NNC is 6% higher than that of NMI+NNC.

Table 1. The recognition accuracies of using different genes by NNC and NMC.

The number of genes	Accuracy (%)			
	FS+NNC	FS+NMC	NMI+NNC	NMI+NMC
1	76.47	82.35	94	94
10	100	100	94	94
100	100	100	100	100
200	100	100	100	100
500	100	100	88	100
5000	94.12	88.24	88	88.24
22283	52.94	64.71	52.94	64.71

To visualize the most informative genes that are selected, the expression values of each sample are drawn in Figures 2, 3, and 4.

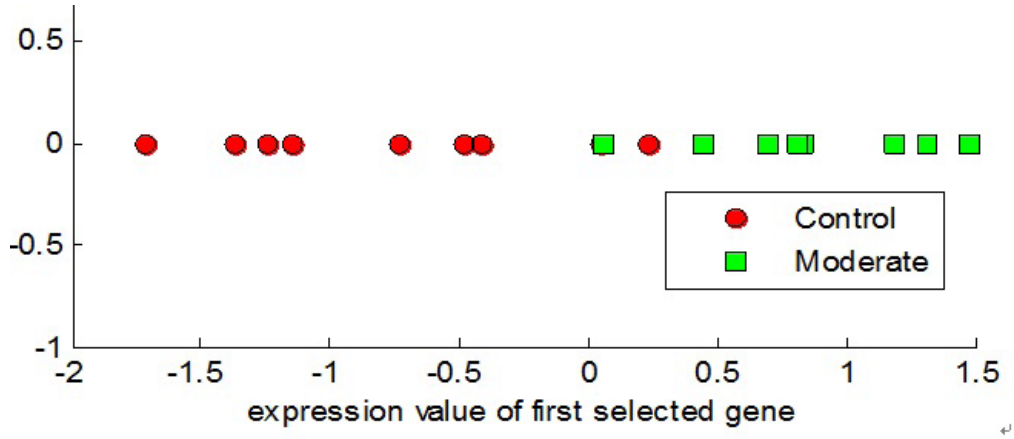

Figure 2. The first selected gene expression values of samples with two classes.

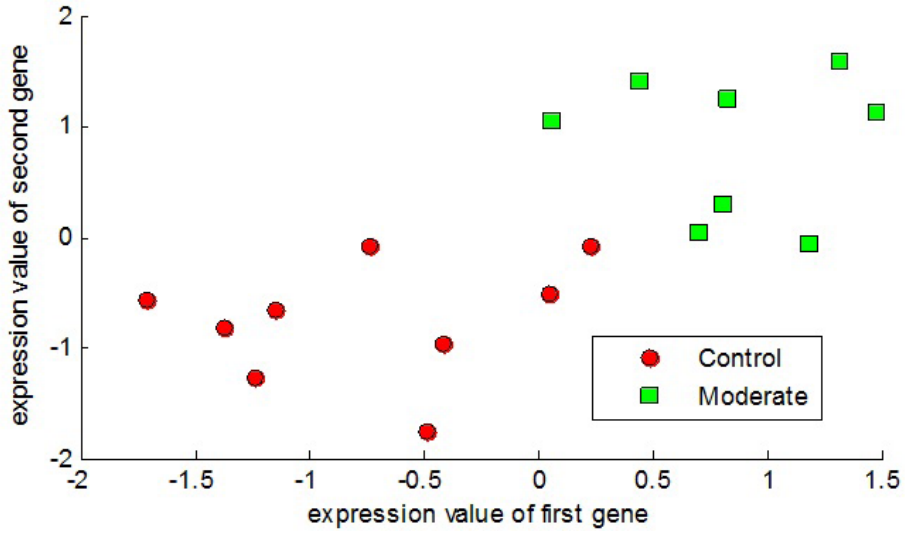

Figure 3. A scatter diagram of samples from two classes (Control and Moderate) with the first two genes.

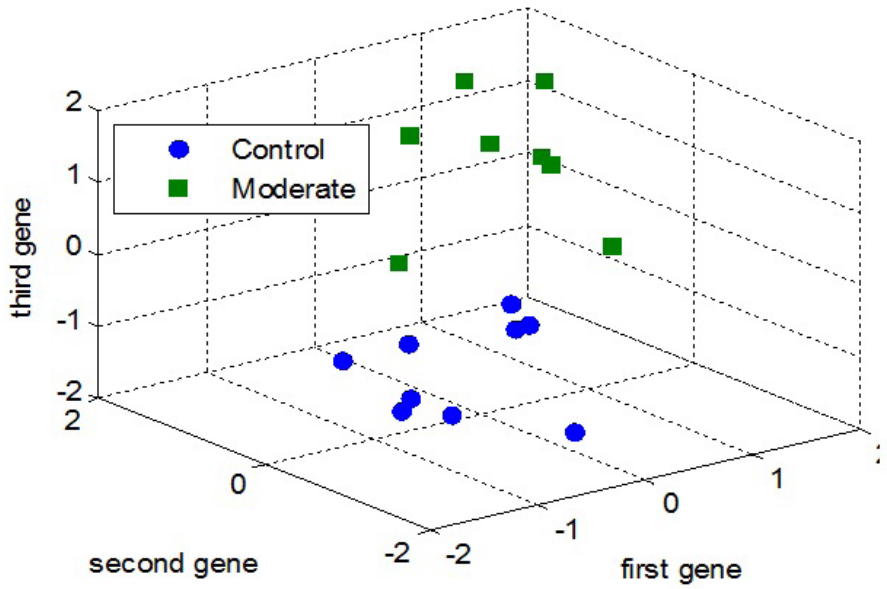

Figure 4. A scatter diagram of samples from two classes with the first three genes.

Table 2 lists the names of the 10 selected genes with the highest scores. From the viewpoint of disease recognition, these genes are related to AD. Researchers can further analyze the function of these genes and identify the relationship between these genes and AD from a biology point of view.

Table 2. Selected genes by the proposed approach.

Ranking number	Number of gene	Name of gen
1	16245	216333_x_at
2	13488	213567_at
3	20939	221035_s_at
4	12048	212122_at
5	7997	207941_s_at
6	12370	212445_s_at
7	862	200793_s_at
8	21104	221201_s_at
9	14855	214940_s_at
10	12518	212593_s_at

DISCUSSION

Here, we provide a novel approach to select genes related to AD. The genes with high FSs are informative for classifying patients and are selected as feature genes. The experiment results show that the expression values of these genes can distinguish samples between the different classes. Therefore, these genes are very likely associated with AD and can be further analyzed by researchers.

Conflicts of interest

The authors declare no conflicts of interest.

ACKNOWLEDGMENTS

Research supported by the National Nature Science Foundation of China (#61373163), the National Science and Technology Support Program (#2014BAH11F02, #2012BAH76F01, and #2014BAH11F01), the Scientific Research Fund of Sichuan Provincial Education Department (#15ZA0039), the Project of Visual Computing, Virtual Reality Key Laboratory of Sichuan Province (#PJ2012001), and the scientific research project of Sichuan Normal University (#14yb02).

REFERENCES

Blalock EM, Chen KC, Sharrow K, Herman JP, et al. (2003). Gene microarrays in hippocampal aging: statistical profiling identifies novel processes correlated with cognitive impairment. *J. Neurosci.* 23: 3807-3819.

Blalock EM, Geddes JW, Chen KC, Porter NM, et al. (2004). Incipient Alzheimer's disease: microarray correlation analyses reveal major transcriptional and tumor suppressor responses. *Proc. Natl. Acad. Sci. USA* 101: 2173-2178. http://dx.doi.org/10.1073/pnas.0308512100

Dougherty ER (2001). Small sample issues for microarray-based classification. *Comp. Funct. Genomics* 2: 28-34. http://dx.doi.org/10.1002/cfg.62

Guttula SV, Allam A and Gumpeny RS (2012). Analyzing microarray data of Alzheimer's using cluster analysis to identify the biomarker genes. *Int. J. Alzheimers Dis.* 2012: 649456. http://dx.doi.org/10.1155/2012/649456

Peng H, Long F and Ding C (2005). Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* 27: 1226-1238. http://dx.doi.org/10.1109/TPAMI.2005.159

Hu B, Jiang G, Pang C, Wang S, et al. (2013). Assessment of gene order computing methods for Alzheimer's disease. *BMC Med. Genomics* 6 (Suppl 1): S8. http://dx.doi.org/10.1186/1755-8794-6-S1-S8

Liang WS, Reiman EM, Valla J, Dunckley T, et al. (2008). Alzheimer's disease is associated with reduced expression of

energy metabolism genes in posterior cingulate neurons. *Proc. Natl. Acad. Sci. USA* 105: 4441-4446. http://dx.doi.org/10.1073/pnas.0709259105

Liu X, Krishnan A and Mondry A (2005). An entropy-based gene selection method for cancer classification using microarray data. *BMC Bioinformatics* 6: 76. http://dx.doi.org/10.1186/1471-2105-6-76

Pang CY, Hu W, Hu BQ, Shi Y, et al. (2010). A special local clustering algorithm for identifying the genes associated with Alzheimer's disease. *IEEE Trans. Nanobioscience* 9: 44-50. http://dx.doi.org/10.1109/TNB.2009.2037745

Rifkin R, Mukherjee S, Tamayo P, Ramaswamy S, et al. (2003). An analytical method for multiclass molecular cancer classification. *SIAM Rev.* 45: 706-723. http://dx.doi.org/10.1137/S0036144502411986

Saeys Y, Inza I and Larrañaga P (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics* 23: 2507-2517. http://dx.doi.org/10.1093/bioinformatics/btm344

Schena M, Shalon D, Davis RW and Brown PO (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270: 467-470. http://dx.doi.org/10.1126/science.270.5235.467

Shipp MA, Ross KN, Tamayo P, Weng AP, et al. (2002). Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat. Med.* 8: 68-74. http://dx.doi.org/10.1038/nm0102-68

Veer LJ, Hongyue D, Mark J, Yudong DH, et al. (1981). Expression profiling predicts outcome in breast cancer. *Biochem. Pharmacol.* 30: 1855-1856.

Wei K, Mou X, Liu Q, Chen Z, et al. (2009). Independent component analysis of Alzheimer's DNA microarray gene expression data. *Mol. Neurodegener.* 4: 1-14.

Yang J, Si J, Gu X and Shi O (2013). Fuzzy cluster analysis of Alzheimer's disease-related gene sequences. *Engineering* 5: 530-533. http://dx.doi.org/10.4236/eng.2013.510B109

Zirvi KA, Dasmahapatra KS, Atabek U and Lyons MA (1989). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Brain Res.* 501: 205-214.