ORIGINAL ARTICLE

THE JOURNAL OF NEUROBEHAVIORAL SCIENCES
NÖRODAVRANIŞ BİLİMLERİ DERGİSİ

Year : 2016
Volume : 3
Issue Number : 3
Doi Number : 10.5455/JNBS.1471026038

Article history:
Received 12 August 2016
Received in revised form 10 September 2016
Accepted 26 September 2016

CLASSIFICATION OF SCHIZOPHRENIA AND BIPOLAR DISORDER BY USING MACHING LEARNING ALGORITHMS

BİPOLAR BOZUKLUK VE ŞİZOFRENİ HASTALIĞININ MAKİNE ÖĞRENMESİ ALGORİTMALARI KULLANILARAK SINIFLANDIRILMASI

Cemil Can Saylan[1], Kaan Yilancioglu[2*]

Abstract

Data mining based investigations of disease mediating factors and related potential diagnostic biomarkers using genomic information obtained from gene expression analysis tools become very informative and useful. In the present study, public Gene Expression Omnibus (GEO) genome wide expression dataset (ID: GSE12654) consisting of schizophrenia, bipolar disorders patients besides normal groups were analyzed by using different classification algorithms including kNN, naïve bayes and decision tree. A set of most differentially expressed genetic features (p<0.05) were used for creating the classifiers which can predict disease states in test set with high accuracy. Data mining tools are suggested to be applicable for developing genome-based diagnostic biomarkers.

Keywords: Data Mining, kNN, Naive Bayes, Decision Tree

Özet

Gen ifadesi çalışmaları sonucunda elde edilen genomik bilgiler, data madenciliği temelli çalışmalarda, hastalık oluşturucu faktörlerin ve bu hastalıklar ile ilişkili potansiyel teşhis biomarkörlerinin bulunması açısından oldukça kullanışlı ve bilgi vericidir. Bu çalışmada, Gene Expression Omnibus (GEO) veri bankasından alınmış, tüm genom ekspresyon verisi (ID: GSE12654) kulllanılmıştır. Veri normal grupların yanısıra, bipolar ve şizofreni hastalarının gen ifadesi bilgisini içermektedir. kNN, naïve bayes ve decision tree bilgisayarlı öğrenme algoritmaları kullanılarak veri analizi gerçekleştirilmiştir. Gruplar arasında istatistiksel olarak anlamlı bir şekilde (p<0.05) farklı eksprese olmuş bir grup gen kullanılarak klasifikasyon yapılmıştır ve gruplar yüksek doğruluk oranında tahmin edilmiştir. Genom tabanlı teşhiş biyomarkörlerinin bulunması açısından, veri madenciliği tekniklerinin yararlı ve uygulanabilir olduğu görülmektedir.

Anahtar Kelimeler: *Veri Madenciliği, kNN, Naive Bayes, Decision Tree*

[1]Department of Molecular Biology and Genetics, Faculty of Engineering and Natural Sciences, Üsküdar University, İstanbul, Turkey.
[2]Department of Bioengineering, Faculty of Engineering and Natural Sciences, Üsküdar University, İstanbul, Turkey.
[*]Corresponding author: Kaan Yilancioglu, PhD; Department of Bioengineering, Faculty of Engineering and Natural Sciences, Üsküdar University, İstanbul, Turkey. E-mail: kaan.yilancioglu@uskudar.edu.tr

1. Introduction

Schizophrenia is one of the most concerned chronic and severe brain disorders, which interferes managing of emotions, speech and thinking processes. Diagnosis of schizophrenia cannot be made easily, although psychiatrists or other licensed mental health professionals diagnose it through interviews and symptoms which reflect biologically heterogeneous characteristics.

Bipolar disorder is identified with extreme mood swings from depression to mania and manic depressive disorder. Causation of bipolar disorder has not been understood entirely, however genetic and environmental factors are thought to have some roles. Diagnosis of the disease mostly relies on clinical examinations.

In order to predict individual disease risk, diagnostic classification of brain disorders might be used by interpreting brain visualization and genetic variation analysis results. This approach has been gained importance for finding potential diagnostic and prognostic biomarkers. Modified statistical methods, which are combined by genomics, have been held in more research activities since the strong genetic associations were demonstrated for various diseases (Orru, Pettersson-Yeo, Marquand, Sartori, & Mechelli, 2012).

In this decade, Microarray and Next Generation Sequencing (NGS) platforms have come into prominence in analyzing genomic data with machine learning algorithms, which have been suggested to be successfully utilized in training classifiers to decode genetic profiles of interest from genomic data (Lu & Han, 2003).

Presently, genomic information is used for classification of patient with brain disorders. In a previous study, it was demonstrated that SVMs can classify both bipolar and schizophrenia from normal subjects with high accuracy by using gene expression data (Struyf, Dobrin, & Page, 2008). Beside this, Genome-Wide Association Studies (GWAS) are also used to classify on bipolar disorder(Emamian, Hall, Birnbaum, Karayiorgou, & Gogos, 2004). It is believed that reciprocal action of genetic predisposition and environmental factors play a role as developmental effects play in bipolar disorder and schizophrenia.

In this article, we present an examination of publicly available microarray gene expression data by using machine learning methods to classify schizophrenic and bipolar disorder individuals respectively.

2. Data and Analysıs methods

2.1. Data

Microarray expression data set (ID: GSE12654) is available in GEO database (Iwamoto, Kakiuchi, Bundo, Ikeda, & Kato, 2004). The dataset was divided into three groups as schizophrenia, bipolar disorders patients and control groups. The samples correspond to 3 bipolar and 4 controls, 4 schizophrenia and 4 controls for testing, 8 bipolar disorder and 11 controls and 9 schizophrenia and 11 controls for training. For each subject, demographic and clinical information were described in the original paper (Iwamoto et al., 2004). The gene expression data was obtained using Affymetrix Human Genome U95 Version 2 oligonucleotide arrays containing 54,676 probe sets (Affymetrix, Santa Clara, CA). Probe level data was normalized using the robust multi-array average (RMA) method (Wu & Irizarry, 2005). The data set includes the RMA value of each probe set as a numerical feature. All computational analyses were done by using R (v3.2.2) [R core team, 2013]. For microarray data preparation "LIMMA" package was used (Ritchie et al., 2015).

2.2. K Nearest Neirghbor Classification Algorithm (K-NN)

The k-Nearest Neighbor (k-NN) is one of the most commonly used non-parameter algorithms, which can be used for predicting test samples according to training model, which finds nearest neighbors to the test samples(Geva & Sitte, 1991). The classification method has been applied by using RapidMiner 7.0 data mining software [RapidMiner 7.0, 2006].

2.3. Decision Tree Algorithm

Decision tree algorithm is a commonly used method in data mining studies (Holzinger, 2015).It has been used as tree-shaped model, which has more limpid representations of results compared with other classification methods. The aim is to create a model that classifies the target attribute based on input variables of training set.

2.4. Naive Bayesalgorithm

Naive Bayes Algorithm classifies samples based on Bayesian rule (Domingos & Pazzani, 1997). Its name comes from the strong (naïve) statistical independence assumptions. Beside this, it often works remarkably well in practice. Naïve Bayes assumes nominal features, which means that numerical features must be discretized prior to running Naive Bayes. Naive Bayesian model can build with uncomplicated repetitive parameter estimation which makes it especially useful for very large datasets. The Naive Bayesian classifier is widely used because it often outperforms more complicated classification methods.

3. Result and Discussion

3.1. Expression Analysis

In schizophrenia, 437 differentially expressed probes 99-(p-value<0.05) were extracted among 54,676parallelly controlled with GEO2R(Sean & Meltzer, 2007). Top 20 probes were manually evaluated and LRRFIP1 and ABCA2 genes were found related with schizophrenia disorder according to previous literature. LRRFIP1 is described as a transcriptional repressor, which may regulate expression of the TNF gene (Suriano et al., 2005) which is implicated in synaptic formation and scaling, long-term potentiation, and neurogenesis (Bilbo & Schwarz, 2009), and was found to be associated with schizophrenia (Morar et al., 2007).ABCA2 was enriched for brain-critical exons which are highly expressed in human brain under strict purifying selection. The significant enrichment within 'Brain-Critical Exons' implicates the pathogenic potential of the hub gene, ABCA2,also found in co-expression network of prenatal frontal cortex in schizophrenia disorder(Wang et al., 2015).

JNBS 2016 Published by Üsküdar University www.jnbs.org

For bipolar disease, 325 differentially expressed probes (p-value<0.05) were extracted. We manually evaluated top 20 differentially expressed probes. NCAM1 and ARHGAP4 were found to be associated with bipolar disease. NCAM1 is expressed in both neurons and glial cells with functions in cellular migration, cell recognition, synaptic plasticity (Kiss & Muller, 2001) and central nervous system development(Ronn, Hartz, & Bock, 1998). In bipolar disease patients, secreted isoform of NCAM1 is increased in the hippocampus whereas the concentration of a proteolytic cleavage isoform of NCAM (cNCAM) was not changed in the brain of patients with bipolar disease patients but increased in schizophrenia(Vawter, Howard, Hyde, Kleinman, & Freed, 1999; Wood et al., 1998).

Some of the proposed genetic markers involvingSEC24C, PGLYRP1, ARHGAP4, RPL22, SLC6A11, and SYK together were described as the "switchboards" were proposed as targets for drug development for bipolar disease(Lee et al., 2011). These genes were also found in our differentially expressed probes list.

3.2. K-NN Classification and Classifier Performance

For classification, top 325 most differentially expressed genes ($p < 0.05$) between control (n=11) and bipolar (n=8), and 437 ($p < 0.05$) most differentially expressed genes between schizophrenia (n=9) and control (n=11) were used to train the model. For validation, same gene set was used in the test data from 4 controls and 3 bipolar patients, and 4 controls and 4 schizophrenia patients. According to results, prediction accuracy was found on testing data as ~86%. Results were shown in Table 1, Table 2 and Table 3 individually for each group.

Table 1: k-NN classification of Bipolar disorder patients. Table shows confusion matrix of classifier on testing data for bipolar disorder patients.

	True Control	True Bipolar	Class Precision
Pred. Control	4	1	80.00%
Pred. Bipolar	0	2	100.00%
Class Recall	100.00%	66.67%	Accuracy: 85.71%

Table 2: k-NN classification of Schizophrenia patients. Table shows confusion matrix of classifier on testing data for schizophrenia patients.

	True Control	True Schizophrenia	Class Precision
Pred. Control	4	1	80.00%
Pred. Bipolar	0	2	100.00%
Class Recall	100.00%	66.67%	Accuracy: 85.71%

3.3. Decision Tree Classification and Classifier Performance

In order to apply decision tree classification method on bipolar disorder, same dataset has been trained as in k-NN classification method. According to the tree shown in Figure 1, it is shown that NCAM1 (41289_at) and ARPGAP4 (39649_at) genes were shown to be important in classifying the disorder. Decision tree model classifies bipolar and control patients with high accuracy on testing data as ~86% shown in Table 3. In addition, k-NN and decision tree model results were found to be similar.

For schizophrenia, GUCY2C (34450_at) and C2orf72 (39394_at) genes were shown to classify the disorder successfully by using decision tree demonstrated in Figure 2. GUCY2C gene is a paralog gene to GUCY1A2 that is previously shown to be associated with schizophrenia [23]. Decision tree model classifies schizophrenia and control patients at accuracy on testing data as 62.5% shown in Table 4.

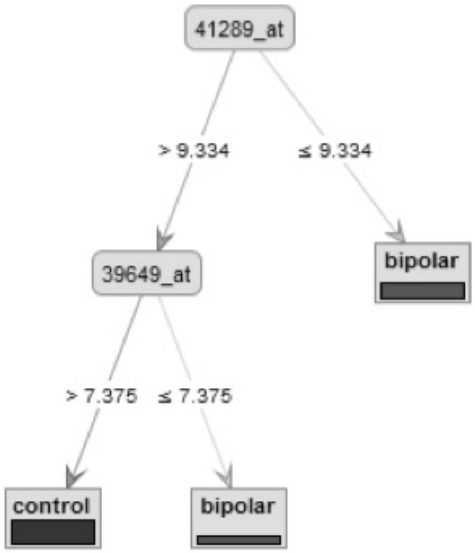

Figure 1: Decision tree classification of bipolar disorder. 41289_at and 39649_at probes represent to NCAM1 and ARPGAP4 genes respectively. Bipolar patients are determined as ≤9.334 expression level of NCAM1 and ≤7.375 expression level of ARPGAP4. Control is determined as>7.375 expression level of ARPGAP4.

Table 3: k-NN classification of Bipolar disorder patients. Table shows confusion matrix of classifier on testing data for bipolar disorder patients.

	True Control	True Bipolar	Class Precision
Pred. Control	4	1	80.00%
Pred. Bipolar	0	2	100.00%
Class Recall	100.00%	66.67%	Accuracy: 85.71%

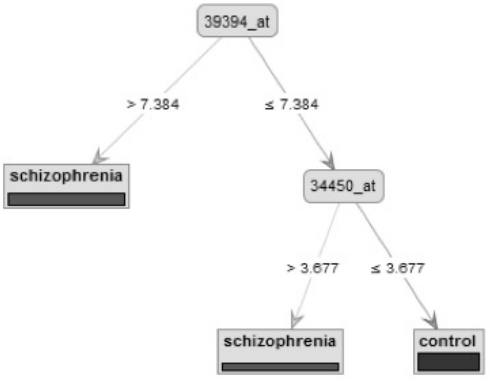

Figure 2: Decision tree classification of schizophrenia and control samples. 39394_at and 34450_at represent to GUCY2C and C2orf72 genes respectively. Schizophrenia patients are determined as >7.384 expression level of GUCY2C and >3.677 expression level of C2orf72. Control is determined ≤3.677 expression level of ARPGAP4.

Table 4: Decision tree classification of Schizophrenia. Table shows confusion matrix of classifier on testing data for Schizophrenia patients.

	True Control	True Schizophrenia	Class Precision
Pred. Control	3	2	60.00%
Pred. Schizophrenia	1	2	66.67%
Class Recall	75.00%	50.00%	Accuracy: 62.50%

3.4. Naive Bayes Classification and Classifier Performance

Naive Bayes classification was applied to the same probe set which was used in previous analyses. Test group includes4 controls with 3 bipolar patients, and 4 controls with 4 schizophrenia patients. Accordingly, results were shown on Table 5 and Table 6. k-NN and Naive Bayes classification methods demonstrated similar accuracy on bipolar and schizophrenia classifications.

Table 5: Naive Bayes classification of Bipolar disorder. Table shows confusion matrix of classifier on testing data for bipolar patients.

	True Control	True Bipolar	Class Precision
Pred. Control	3	0	100.00%
Pred. Bipolar	1	3	75.00%
Class Recall	75.00%	100.00%	Accuracy: 87.71%

Table 6: Naive Bayes classification of Schizophrenia. Table shows confusion matrix of classifier on testing data for Schizophrenia patients.

	True Control	True Schizophrenia	Class Precision
Pred. Control	3	0	100.00%
Pred. Schizophrenia	1	4	80.00%
Class Recall	75.00%	100.00%	Accuracy: 87.50%

4. Conclusions

Differentially expressed genes used as classifying features might be useful for revealing important genes and gene families associated with schizophrenia and bipolar disease. More importantly the classifier method might be applicable for developing effective Microarray-based diagnostic tests.

References

Bilbo, S. D., & Schwarz, J. M. (2009). Early-life programming of later-life brain and behavior: a critical role for the immune system. Frontiers in Behavioral Neuroscience, 3. doi: ARTN 1410.3389/neuro.08.014.2009

Domingos, P., & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning, 29(2-3), 103-130. doi: Doi 10.1023/A:1007413511361

Emamian, E. S., Hall, D., Birnbaum, M. J., Karayiorgou, M., & Gogos, J. A. (2004). Convergent evidence for impaired AKT1-GSK3beta signaling in schizophrenia. Nat Genet, 36(2), 131-137. doi: 10.1038/ng1296

Geva, S., & Sitte, J. (1991). Adaptive nearest neighbor pattern classification. IEEE Trans Neural Netw, 2(2), 318-322. doi: 10.1109/72.80344

Hofmann, M., Klinkenberg, R. (2013). "RapidMiner: Data Mining Use Cases and Business Analytics Applications (Chapman & Hall/CRC Data Mining and Knowledge Discovery Series)," CRC Press, October 25.

Holzinger, A. (2015). Data Mining with Decision Trees: Theory and Applications. Online Information Review, 39(3), 437-438. doi: 10.1108/Oir-04-2015-0121

Iwamoto, K., Kakiuchi, C., Bundo, M., Ikeda, K., & Kato, T. (2004). Molecular characterization of bipolar disorder by comparing gene expression profiles of postmortem brains of major mental disorders. Mol Psychiatry, 9(4), 406-416. doi: 10.1038/sj.mp.4001437

Kiss, J. Z., & Muller, D. (2001). Contribution of the neural cell adhesion molecule to neuronal and synaptic plasticity. Reviews in the Neurosciences, 12(4), 297-310.

Lee, S. A., Tsao, T. T. H., Yang, K. C., Lin, H., Kuo, Y. L., Hsu, C. H., Kao, C. Y. (2011). Construction and analysis of the protein-protein interaction networks for schizophrenia, bipolar disorder, and major depression. Bmc Bioinformatics, 12. doi: Artn S2010.1186/1471-2105-12-S13-S20

Lu, Y., & Han, J. W. (2003). Cancer classification using gene expression data. Information Systems, 28(4), 243-268. doi: Pii S0306-4379(02)00072-8

Morar, B., Schwab, S. G., Albus, M., Maier, W., Lerer, B., & Wildenauer, D. B. (2007). Evaluation of association of SNPs in the TNF alpha gene region with schizophrenia. American Journal of Medical Genetics Part B-Neuropsychiatric Genetics, 144B(3), 318-324. doi: 10.1002/ajmg.b.30451

Orru, G., Pettersson-Yeo, W., Marquand, A. F., Sartori, G., & Mechelli, A. (2012). Using Support Vector Machine to identify imaging biomarkers of neurological and psychiatric disease: A critical review. Neuroscience and Biobehavioral Reviews, 36(4), 1140-1152. doi: 10.1016/j.neubiorev.2012.01.004

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y. F., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research, 43(7). doi: ARTN e4710.1093/nar/gkv007

Ronn, L. C., Hartz, B. P., & Bock, E. (1998). The neural cell adhesion molecule (NCAM) in development and plasticity of the nervous system. Exp Gerontol, 33(7-8), 853-864.

Sean, D., & Meltzer, P. S. (2007). GEOquery: a bridge between the gene expression omnibus (GEO) and BioConductor. Bioinformatics, 23(14), 1846-1847. doi: 10.1093/bioinformatics/btm254

Struyf, J., Dobrin, S., & Page, D. (2008). Combining gene expression, demographic and clinical data in modeling disease: a case study of bipolar disorder and schizophrenia. Bmc Genomics, 9. doi: Artn 53110.1186/1471-2164-9-531

Suriano, A. R., Sanford, A. N., Kim, N., Oh, M., Kennedy, S., Henderson, M. J.,Sullivan, K. E. (2005). GCF2/LRRFIP1 represses tumor necrosis factor alpha expression. Molecular and Cellular Biology, 25(20), 9073-9081. doi: 10.1128/Mcb.25.20.9073-9081.2005

Vawter, M. P., Howard, A. L., Hyde, T. M., Kleinman, J. E., & Freed, W. J. (1999). Alterations of hippocampal secreted N-CAM in bipolar disorder and synaptophysin in schizophrenia. Mol Psychiatry, 4(5), 467-475. doi: DOI 10.1038/sj.mp.4000547

Wang, Q., Li, M. X., Yang, Z. X., Hu, X., Wu, H. M., Ni, P. Y., Li, T. (2015). Increased co-expression of genes harboring the damaging de novo mutations in Chinese schizophrenic patients during prenatal development. Scientific Reports, 5. doi: Artn 1820910.1038/Srep18209

Wood, G. K., Tomasiewicz, H., Rutishauser, U., Magnuson, T., Quirion, R., Rochford, J., & Srivastava, L. K. (1998). NCAM-180 knockout mice display increased lateral ventricle size and reduced prepulse inhibition of startle. Neuroreport, 9(3), 461-466. doi: Doi 10.1097/00001756-199802160-00019

Wu, Z., & Irizarry, R. A. (2005). Stochastic models inspired by hybridization theory for short oligonucleotide arrays. J Comput Biol, 12(6), 882-893. doi: 10.1089/cmb.2005.12.882

JNBS 2016 Published by Üsküdar University www.jnbs.org