e-Health – For Continuity of Care
C. Lovis et al. (Eds.)

doi:10.3233/978-1-61499-432-9-511

Predicting the Disease of Alzheimer With SNP Biomarkers and Clinical Data Using Data Mining Classification Approach: Decision Tree

Onur ERDOĞAN [a,1] and Yeşim AYDIN SON [a,]
[a] *Middle East Technical University, Ankara, Turkey*

Abstract. Single Nucleotide Polymorphisms (SNPs) are the most common genomic variations where only a single nucleotide differs between individuals. Individual SNPs and SNP profiles associated with diseases can be utilized as biological markers. But there is a need to determine the SNP subsets and patients' clinical data which is informative for the diagnosis. Data mining approaches have the highest potential for extracting the knowledge from genomic datasets and selecting the representative SNPs as well as most effective and informative clinical features for the clinical diagnosis of the diseases. In this study, we have applied one of the widely used data mining classification methodology: "decision tree" for associating the SNP biomarkers and significant clinical data with the Alzheimer's disease (AD), which is the most common form of "dementia". Different tree construction parameters have been compared for the optimization, and the most accurate tree for predicting the AD is presented.

Keywords. data mining, single nucleotide polymorphism, integrating genotype and phenotype data, decision tree, alzheimers disease

Introduction

A challenging aspect of the post-genome biology of human is to understand the biological effects of inherited variations in DNA structure between individuals. The most common genetic variations are represented as single DNA building block alterations that they have received much attention recently in terms of the detection of a particular disease [1]. SNPs are single nucleotide alterations found in every 300 to 1000 nucleotides in genomic DNA and cause personal differences in phenotypes as well as underlying reason of a particular disease [2]. SNPs can be used as genomic markers revealing individuals susceptibility to certain disease to produce new approaches for treatment applications and to take prohibitive precaution. SNP association studies are widely done to determine possible relations between genetic variations and diseases.

In the last decades, developments of data mining methods have become a promising approach in bioinformatics in order to solve biological problems [3], [4].

[1] Corresponding author: *Onur ERDOĞAN, Middle East Technical University, Informatics Institute, Department of Health Informatics, PhD Student, Ankara, Turkey.* onur.erdogan@tubitak.gov.tr

Drawing conclusions from the high amount of data requires intelligent computational analysis [5]. The main aim of data mining is the classification of the data, and clustering of the data based on the similarities or dissimilarities. It has been recently shown that the data mining applications are extremely useful in order to make decisions over high dimensional data such as whole human genome data that consists of around 3.4 billion nucleotide pairs. Outputs of mining methods in such genetic studies have revealed interesting findings inheritable tendency to contract specific diseases [6].

Alzheimer's disease (AD) is a complex and genetic disorder and the most common cause of dementia. Early onset is the rare form and symptoms start before the age of 65 and caused by mutations in the genes such as APP (amyloid precursor protein), PSEN-1 and PSEN-2 (presenilin genes) [7]. On the other hand late onset is the common form of AD presents a complex genetic inheritance, where the differential diagnosis is critical. Few number of variants in genes have now been identified [7] such as APOE (apolipoprotein E) in chromosome 19, but the genetic factors underlying the late onset Alzheimer's risk are still being investigated.

In this study we have investigated the AD associated genotypes and identified the representative SNP subsets and clinical features through data mining. The performance of the genomic AD model build based on integrated genotyping and clinical data showed its potential as a predictive and diagnostic tool.

1. Methods

1.1. Dataset

The Alzheimer's disease genotyping and phenotyping data is obtained from GENADA [8] study through the dbGAP database. With authorized access, 1718 study participants' individual level data is available through GENADA study. Genotyping and clinical data is collected from eligible individuals who have late onset with mild to moderate level of Alzheimer's disease.

Clinical data included in this study that may be associated to the Alzheimer's disease are cholesterol (mmol/l), hemoglobin (g/l), HBA1C_PCT, HDL cholesterol (mmol/l), LDL cholesterol (mmol/l), triglycerides (mmol/l) and amount of white blood cells. Additional data in medical records are also used such as age onset of first symptoms and body mass index at the moment of the first diagnose [8]. Genotyping data included 410907 SNPs, for all case and controls. SNPs, which have the minor allele frequency greater than 1% is included in the association studies order to reveal the genomic variations associated with AD susceptibility [9]. Interrogating the DNA sequences of people, SNP alleles are included [8].

1.2. Preprocessing

Data preprocessing that usually takes 60% of efforts is the most important step in knowledge extraction process [10]. At first we have eliminated 108 samples whose has missing data in their lab findings. Furthermore, the inconsistent data is corrected. Onset ages of 14 people in the case group were not entered in the dataset; hence missing values in age attribute are filled with the mean age of the cases. All the genotyping data was available, hence there was no need for imputation. After the elimination of the

missing values over clinical findings, 1480 samples were available for the decision model construction.

GENADA study contains different repositories in terms of genotype data, phenotype data, clinical findings and laboratory results. Combining the tables from different sources requires structured query language. We used SQL to transfer all related data tables into one table and to create relationship between keys.

Dimension reduction is beneficial for both computational efficiency and enhancement of model accuracy since biological data contains large amount of variables. First, the p-value associations of 410969 SNP are revealed by PLINK analysis, then SNPs that present statistical significant association with AD are selected. A second selection step for biological relevance of the SNPs is done by METU-SNP's AHP based prioritization algorithm [11]. After prioritization, SNPs with the AHP score above 0.40 are selected for further analysis. There were 958 biologically and statistically significant SNPs identified in total.

1.3. Constructing the Model

Two different decision trees were constructed in order to see how much genotyping data and clinical findings contribute to the prediction of Alzheimer's disease. The first tree is constructed using only representative SNPs found by AHP scoring. The second tree is constructed using both clinical findings and genotyping data. The C4.5 decision tree algorithm with gain ratio which is the best single feature selection criterion is used since it can handle not only categorical but also numeric attributes [12]. For investigation of accuracy of C4.5 prediction model, k-fold cross validation was applied. The process of data classification that models our implementation contains data preprocessing that takes approximately 60% of efforts, data modelling using training data and model validation using test set. The final step is the visualization and knowledge extraction.

Tree construction depends on entropy reduction, but in each division, over-fitting can lead the poor accuracy. In response to this problem, pruning strategy is used. The minimal gain ratio is identified as 0,01 for splitting the dataset for both tree constructions. 11 fold-cross validation is chosen for the error prediction and model accuracy. We also applied C4.5 algorithm to dataset with its 958 independent variables considering the selected genotype data, 8 clinical and 2 demographic features [6] in order to predict whether a person is case or control. Finally, IF_THEN rules are extracted from decision trees and we compared results whether addition of clinical attributes contribute to the prediction of the complex disease or not.

2. Results

Even though combinations of a variety of genetic and other factors are suspected in complex diseases, the genetic factors underlying the Alzheimer's disease are still unknown. In this study we have attempted to construct a decision tree for the differential diagnosis of late onset Alzheimer's disease based on patients' genotyping and clinical data.

The first decision tree which is constructed only with representative SNP biomarkers included 38 significant SNPs, selected among 958 that were associated with the disease. With these 38 SNPs, decision tree generated 26 rules for the

diagnoses of the late onset AD. The genetic variations included in the first decision tree were located on genes such as ABCC4, ANGPT2, ANGPT2, ARHGAP26, ATG5, C9orf3, DBT, DDO, DISC1, ENPP6, FGD4, FMNL2, FOXO3, GABBR2, GSN, KCNN3, KIF26B, LIPH, MAML3, NBN, PDZD8, PLCB1, PTPRM, SEMA3C, SEMA3C, SEMA5A, SLC35A3, SNW1, SYN3, TLL2, TPO, TRHDE, C9orf3, CAMKK2, DOK1, HMGA1, PIKFYVE, STK39. The most significant genetic variations are SNP_A4213932, SNP_A2146889, SNP_A2258450 and SNP_A1849082 when we consider the top levels of the decision tree [6]. When we have integrated the relevant clinical data with SNP data in the second decision tree, the clinical features such as HBA1C_PCT, Body Mass İndex, Hemoglobin, WBC, Trig, Cholesterol and HDL are included. Only 27 SNPs out of 958 are chosen by attribute selection criterion. According to decision tree build by integrating genotyping and clinical data, the most significant genetic and clinical variations were the SNP_A4213932, SNP_A2146889, age and cholesterol (mmol), placed at the first two level of the tree [6]. The performances of classifiers for both decision trees after 11 fold cross validation method are summarized in Table 1.

Table 1. Performances comparison of three model

	Only Genotype		Only Clinical Findings&Phenotype		Genotype and Clinical Findings&Phenotype	
	Value	Confidence Interval	Value	Confidence Interval	Value	Confidence Interval
Accuracy	56,08% *(std.dev:1,96)*	[54,76%-57,40%]	53,31% *(std.dev:3,65)*	[50,86%-55,76%]	55,07% *(std.dev:2,49)*	[53,40%-56,74%]
Sensitivity (True Positive Rate)	57,53%	-	42,20%	-	61,96%	-
Specificity (True Negative Rate)	54,62%	-	65,54%	-	55,07%	-

3. Discussion

Alzheimer ’s disease (AD) is the neuro-degeneration of the brain tissue beyond the effects of normal aging process. Today the only effective diagnostic method for diagnosis of the Alzheimer patients is the confirmation of the lesions in brain tissue during autopsy. AD is a complex and chronic disease and inherited in 80% of the cases. Various genetic changes including the APOE polymorphisms can only explain the 25% of the genetic background of the disease [13], [14]. Studies on large populations of AD cases and controls continue to identify the rest of the genomic variations associated with AD.

Here we have identified 958 biologically and statistically significant SNPs associated with late onset AD, which are later used to construct a decision tree model for the differential diagnosis with the accuracy rate of classification of 56,08%. Even though the performance of the model is not too satisfying and should be improved in further studies, this is a promising demonstration of how genome wide association studies can benefit from data mining approaches for the interpretation of the SNP variations in disease models. In this study [6], not only genotyping data but also clinical information is used. The aim of the clinical information is to show how it

contributes the prediction of disease. Surprisingly, integration of clinical information didn't improve the accuracy rate of the model. Hence, only using genotype data, Alzheimer's disease can be predicted for the new data samples [6]. Considering the difficulties of diagnostic methods, implementation of decision tree using genotypic information of individuals can support distinguishing the Alzheimer's disease patients from dementia [6].

With the guidance of this study, data mining classification methods can be implemented to higher dimensional genome databases in order to extract novel and important patterns.

References

[1] Krishnan VG, Westhead DR. A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function. Bioinformatics [Internet]. 2003 Nov 20 [cited 2014 Jan 29];19(17):2199–209. Available from: http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btg297

[2] Human Genetic Variation [Internet]. National Human Genome Research Institute; 2007 [cited 2012 Sep 11]. Available from: http://www.sciencemag.org/content/318/5858/1842.short

[3] Pirooznia M, Seifuddin F, Judy J, Mahon P. Data mining approaches for genome-wide association of mood disorders. Psychiatr Genet [Internet]. 2012 [cited 2014 Feb 5];22(2):55–61. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3306768/

[4] Bar-or A, Schuster A, Wolff R, Keren D. Decision Tree Induction in High Dimensional, Hierarchically Distributed Databases. Proceedings of SDM'05 Newport Beach. California; 2005.

[5] Raza K. Application of Data Mining in Bioinformatics. Indian J Comput Sci Eng. 2009;1(2):114–8.

[6] Erdoğan O. Predicting the Disease Alzheimer (AD) With SNP Biomarkers and Clinical Data Based Decision Support System Using Data Mining Classification Approaches [Internet]. METU; 2012. p. 153. Available from: http://etd.lib.metu.edu.tr/upload/12614832/index.pdf

[7] Farlow JL, Foroud T. The genetics of dementia. Semin Neurol [Internet]. 2013 Sep;33(4):417–22. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24425039

[8] Fornazzari L, Gauthier S, St. George-Hyslop P, Wherrett J, Keren R, Feldman H, et al. Multi-Site Collaborative Study for Genotype-Phenotype Associations in Alzheimer's disease [Internet]. 2002 [cited 2014 Feb 5]. p. 1–38. Available from: http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/GetPdf.cgi?id=phd002049.1

[9] Rapley R, Harbron S, editors. Overview of Microarrays in Genomic Analysis. Molecular Analysis and Genome Discovery. West Sussex, England: John Wiles & Sons, LTD; 2004.

[10] Cabena P, Hadjinian P, Stadler R, Verhees J, Zanasi A. Data Mining: From Concept to Implementation. New Jersey; 1998.

[11] Üskünkar G. An Integrative Approach To Structured SNP Prioritization and Representative SNP Selection For Genome-wide Association Studies:Algorithms and Systems [Internet]. Middle East Technical University; 2008 [cited 2012 Sep 5]. p. 150. Available from: http://etd.lib.metu.edu.tr/upload/12610009/index.pdf

[12] Kohavi R, Quinlan R. Decision Tree Discovery. Citeseer [Internet]. 1999 [cited 2012 Sep 5];3. Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.4.5353&rep=rep1&type=pdf

[13] Roses a D, Lutz MW, Amrine-Madsen H, Saunders a M, Crenshaw DG, Sundseth SS, et al. A TOMM40 Variable-length Polymorphism Predicts the Age of Late-onset Alzheimer's disease. Pharmacogenomics J [Internet]. Nature Publishing Group; 2010 Oct [cited 2012 Jul 27];10(5):375–84. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2946560&tool=pmcentrez&rendertype=abstract

[14] Bertram L, Lange C, Mullin K, Parkinson M, Hsiao M, Hogan MF, et al. Genome-wide Association Analysis Reveals Putative Alzheimer's Disease Susceptibility Loci in Addition to APOE. Am J Hum Genet [Internet]. 2008 Nov [cited 2012 Jul 13];83(5):623–32. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2668052&tool=pmcentrez&rendertype=abstract