The Application of Data Mining in Health Care Informatics

***Hankun Hu[1], Fengjie Sun[2] and †Weidong Mao[2]**

[1]Zhongnan Hospital of Wuhan University, China
[2]School of Science & Technology, Georgia Gwinnett College, USA

*Presenting author: wb000400@whu.edu.cn
†Corresponding author: wmao@ggc.edu

Abstract

Health informatics is a multidisciplinary field that uses health information technology (HIT) to improve health care via any combination of higher quality, higher efficiency, and new opportunities. The availability of public health care information gives the opportunity for researchers to access and analyze the data. Data mining has been applied in health informatics in many area. Both environmental and genetic factors have roles in the development of some diseases. Complex diseases, such as Crohn’s disease or Type II diabetes, are caused by a combination of environmental factors and mutations in multiple genes. Patients who have been diagnosed with such diseases cannot easily be treated. However, many diseases can be avoided if people at high risk change their living style. How to identify their susceptibility to diseases before symptoms are found and help them make informed decisions about their health becomes an important topic in health informatics. The susceptibility to complex diseases can be predicted through the analysis of the genetic data. With the development of DNA microarray technique, it is possible to access the human genetic information related to specific diseases. This paper uses a combinatorial method to analyze the genetic case-control data for Crohn’s disease. A distance based cluster method has been applied to publicly available genotype data on Crohn’s disease for epidemiological study and achieved a high accurate result.

Keywords: Health Informatics, Data Mining, Susceptibility Prediction

Introduction

Health informatics (also called health care informatics, healthcare informatics, medical informatics, nursing informatics, clinical informatics, or biomedical informatics) is informatics in health care. It is a multidisciplinary field that uses health information technology (HIT) to improve health care via any combination of higher quality, higher efficiency (spurring lower cost and thus greater availability), and new opportunities. The disciplines involved include information science, computer science, social science, behavioral science, management science, and others. It deals with the resources, devices, and methods required to optimize the acquisition, storage, retrieval, and use of information in health and biomedicine. Health informatics tools include amongst others computers, clinical guidelines, formal medical terminologies, and information and communication systems [1] [2]. It is applied to the areas of nursing, clinical medicine, pharmacy, public health, occupational therapy, physical therapy, biomedical research, and alternative medicine[3]. The availability of public health care information gives the opportunity for researchers to access and analyze the data.

Data mining has been applied in health informatics in many areas and one of those important area is the disease control and prevention. Complex diseases, such as Crohn’s disease or Type II diabetes, are caused by a combination of environmental factors and mutations in multiple genes. Patients who have been diagnosed with such diseases cannot easily be treated. However, many diseases can be avoided if people at high risk change their living style. How to identify their susceptibility to diseases before symptoms are found and help them make informed decisions about their health becomes an important topic in health informatics. The susceptibility to complex diseases can be predicted through the analysis of the genetic data. With the development of DNA microarray technique, it is possible to access the human genetic information related to specific diseases. Assessing the association between DNA variants and disease has been used widely to identify regions of the genome and candidate genes that contribute to disease [4]. 99.9% of one individual's DNA sequences are identical to that of another person. Over 80% of this 0.1% difference will be Single Nucleotide Polymorphisms (SNP) and they promise to significantly advance our ability to understand and treat human disease. A SNP is a single base substitution of one nucleotide with another. Each individual has many single nucleotide polymorphisms that together create a unique DNA pattern for that person. It is important to study SNPs because they represent genetic differences among human beings. Genome-wide association studies require knowledge about common genetic variations and the ability to genotype a sufficiently comprehensive set of variants in a large patient sample [5]. High-throughput SNP genotyping technologies make massive genotype data, with a large number of individuals, publicly available. Accessibility of genetic data makes genome-wide association studies for complex diseases possible.

It's important to search for informative SNPs among a huge number of SNPs. These informative SNPs are assumed to be associated with genetic diseases. Tag SNPs generated by the multiple linear regression based method [6] are good informative SNPs, but they are reconstruction-oriented instead of disease-oriented. Although the combinatorial search method for finding disease-associated multi-SNP combinations has a better result, the exhaustive search is still very slow.

The distance-based algorithm and cluster analysis have been used to solve the classification problem. In this algorithm, each item that is mapped to the same class may be thought of as more similar to the other items in that class than it is to the items found in other classes. Therefore, similarity measures may be used to identify the "alikeness" of different items in the database [7]. In our algorithm, the similarity is measured by the distance between the item and some neighbor clusters whose class label are previously known. The algorithm can be applied in our case-control study to predict an individual's susceptibility to Crohn's disease by comparing its genetic data with that of other individuals to find the similarity.

In this paper, we first address the disease susceptibility prediction problem [8] [9] [10]. This problem is to assess accumulated information targeted to predicting genotype susceptibility to complex diseases with significantly high accuracy and statistical power. Next, we introduce the cluster-based distance algorithm and its application in disease susceptibility prediction problem. We will also introduce the case tagging algorithm which is used to reduce the size of data and improve prediction results. The proposed method is applied to a publicly available data for Crohn's disease [11].

Disease Susceptibility Prediction Problem

A SNP is a single base substitution of one nucleotide with another. Both substitutions have to be observed in the general population at a frequency greater than 1%. An example of a SNP is individual "A" has a sequence GAACCT, while individual "B" has sequence GAGCCT, the polymorphism is an A/G. Each individual has many single nucleotide polymorphisms that together create a unique DNA pattern for that person. Haplotype is the set of adjacent SNP's are present on alleles in a block pattern. The genotype is the descriptor of the genome which is the set of physical DNA molecules inherited from the organism's parents. A pair of haplotype consists a single genotype.

SNP's are bi-allelic and can be referred as 0 if it's a majority and 1, otherwise. If both haplotypes are the same allele, then the corresponding genotype is homogeneous, can be represented as 0 or 1. If the two haplotypes are different, then the genotype is represented as 2.

The case-control sample populations consist of N individuals which are represented in genotype with M SNPs. Each SNP attains one of the three values 0, 1, or 2. The sample G is an (0, 1, 2)-valued $N \ x \ M$ matrix, where each row corresponds to an individual, which is a sequence of 0, 1 and 2, each column corresponds to a SNP.

The disease susceptibility prediction problem can be formulated as follows:
Data sets have n genotypes. The input for a prediction algorithm includes:

(G1) Training genotype set G = (g_i | *i* = 1..n);
(G2) Disease status s(g_i) € {0,1}, indicating if g_i, is in case (1) or in control (0) , and
(G3) Testing genotype g_t without any disease status.

We will refer to the parts (G1-G2) of the input as the training set and to the part (G3) as the test case. The output of prediction algorithms is the disease status of the genotype g_t, i.e., s(g_t).

Cluster-Based Distance Algorithm

To find an individual's disease status, we compute its similarity with other individuals whose disease status are already known. In the distance-based algorithm, each item that is mapped to the same class may be thought of as more similar to the other items in that class than it is to the items found in other classes. Therefore, similarity measures may be used to identify the "alikeness" of different items in the database. The idea of similarity measures can be abstracted and applied to more general classification problems. The difficulty lies in how the similarity measures are defined and applied to the items.

In this algorithm. We determine the similarity among sequences by computing their hamming distance. The hamming distance between two strings (in our case, two genotypes, each represents an individual) of equal length is the number of positions for which the corresponding symbols are different. For example, the hamming distance between genotype 1 (01021011) and genotype 2 (01021011) is 0, but the hamming distance between genotype 3 (21021210) and genotype 4 (01021011) is 3.

For the training data set, we build graph-based clusters for each class. In other words, we generate N clusters in case class and N clusters in control class given the threshold N. First we generate the graph G_{case} and $G_{control}$ based on the hamming distance among individual genotypes in case class and in control class, respectively. The Kruskal's algorithm is used to find the minimum spanning tree (MST) for G_{case} and $G_{control}$. To generate *N* clusters for Gcase, we need to remove the largest *N-1* edges. As a result, the inter-cluster distance is maximized and the intra-cluster distance is minimized. For these *2xN* clusters, we choose the genotype that has the minimum distance with all other genotype in the same cluster as the centroid of the cluster.

K nearest neighbors (KNN) is used as the classification scheme based on the use of distance measures. The KNN technique assumes that the entire training set includes not only the data in the set but also the desired classification for each item. In effect, the training data become the model. When a classification is to be made for a new item, its distance to each item in the training set must be determined. Only the K closest entries in the training set are considered further. The new item is then placed in the class that contains the most items from this set of K closest items. In our case, the hamming distance of the testing genotype to each centroid in the training set (including both case and control set) will be computed, then we find out the K closest genotypes which have smaller hamming distance than others. From the set of K centroid, if most of them are coming from the case group, then the testing genotype will be classified as case, otherwise, it will be classified as control. Obviously, it will be better if k is an odd number. Figure 1 shows how to classify the testing item when N is 4 and K is 3.

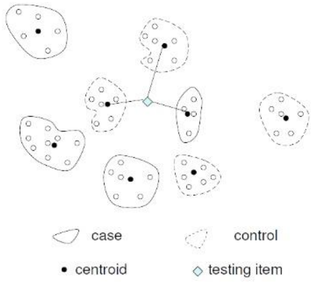

Figure 1. Classify the testing item, N=4, K=3

Results and Discussion

The genetic data is derived from the 616 kilo-base region of human Chromosome 5q31 that may contain a genetic variant responsible for Crohn's disease by genotyping 103 SNPs for 129 trios [11]. All offspring belong to the case population, while almost all parents belong to the control population. In the entire data, there are 144 case and 243 control individuals.

To measure the quality of prediction methods, we need to measure the deviation between the true disease status and the result of predicted susceptibility, which can be regarded as measurement error. We will present the basic measures used in epidemiology to quantify accuracy of our methods. The basic measures are sensitivity and specificity. Sensitivity is the proportion of persons who have the disease who are correctly identified as cases, while specificity is the proportion of people who do not have the disease who are correctly classified as controls. Sensitivity (accuracy in classification of case) and Specificity (accuracy in classification of controls) and Accuracy are calculated as follows in Eq. (1):

$$\begin{aligned} \text{Sensitivity} &= a\ (a + c) \\ \text{Specificity} &= d\ (b + d) \\ \text{Accuracy} &= (a + d)\ (a + b + c + d) \end{aligned} \tag{1}$$

a = True positive, people with the disease who test positive

b = False positive, people without the disease who test positive

c = False negative, people with the disease who test negative

d = True negative, people without the disease who test negative

Sensitivity is the ability to correctly detect a disease. Specificity is the ability to avoid calling normal as disease. Accuracy is the percent of the population that is correctly predicted.

We use K-fold cross validation method to measure the quality of the algorithm. In the K-fold cross validation, the data set is divided into K subsets, and the holdout method is repeated K times. Each time, one of the K subsets is used as the test set and the other K-1 subsets are put together to form a training set. In our experiment, we use 5-fold cross validation.

Table 1 is the experiment result. In this table, we compare the result when K is 1, 3, 5, 7 and the number of cluster N is 6, 10, 14, 18 and 22. The best result is as high as 100% for sensitivity, 100% for specificity, and 90% for accuracy, respectively.

Table 1. Experiment results

K	**Measures**	**Number of Clusters (N)**				
		6	**10**	**14**	**18**	**22**
1	Sensitivity	52	63	69	76	74
	Specificity	100	100	100	99	99
	Accuracy	76	82	85	88	87
3	Sensitivity	94	78	80	61	64
	Specificity	67	91	100	100	100
	Accuracy	81	85	90	81	82
5	Sensitivity	100	94	75	82	82
	Specificity	17	67	67	67	56
	Accuracy	59	81	71	75	69
7	Sensitivity	0	100	86	80	74
	Specificity	100	46	46	67	67
	Accuracy	50	73	66	74	71

Conclusions

In this paper, we discuss the potential of applying a cluster-based distance algorithm on how to predict the susceptibility for a complex disease, one of important problems in health care association studies. The proposed classification method based on cluster and distance is shown to have a high prediction rate without finding SNPs associated with the disease which may reduce the running time. The genetic factors associated with the disease have to be identified first. In our future work we are going to continue validation of the proposed method on various type of data.

References

[1] O'donoghue, John; Herbert, John (2012). "Data management within mHealth environments: Patient sensors, mobile devices, and databases". Journal of Data and Information Quality (JDIQ). 4 (1): 5.

[2] Mettler T, Raptis DA (2012). "What constitutes the field of health information systems? Fostering a systematic framework and research agenda". Health Informatics Journal. 18 (2): 147–56. doi:10.1177/1460458212452496. PMID 22733682.

[3] Popularity of usage of software in homeopathy is shown in example video of homeopathic repertorisation: Shilpa Bhouraskar, Working quick acute cases on Homeopathic Software (YouTube).

[4] Cardon, L.R., Bell, J.I.: Association Study Designs for Complex Diseases, Vol.2. Nature Reviews: Genetics (2001), 91-98.

[5] Hirschhorn, J.N.,Daly, M.J.: Genome-wide Association Studies for Common Diseases and Complex Diseases, Vol.6. Nature Reviews: Genetics (2005), 95-108.

[6] He, J. and Zelikovsky, A.: Tag SNP Selection Based on Multivariate Linear Regression, Proc. of International Conference on Computational Science (2006), LNCS 3992, 750-757.

[7] Margaret H.D., Data Mining - Introduction and advanced topics, prentice Hall, ISBN 0-13-088892-3.

[8] Mao, W., Brinza, D., Hundewale, N., Gremalschi, S. and Zelikovsky, A.: Genotype Susceptibility and Integrated Risk Factors for Complex Diseases, Proc. IEEE International Conference on Granular Computing (GRC 2006), pp. 754-757.

[9] Kimmel, G. and Shamir R.: A Block-Free Hidden Markov Model for Genotypes and Its Application to Disease Association. J. of Computational Biology (2005), Vol. 12, No. 10: 1243-1260.

[10] Listgarten, J., Damaraju, S., Poulin B., Cook, L., Dufour, J., Driga, A.,Mackey, J., Wishart, D., Greiner,R., and Zanke, B.: Predictive Models for Breast Cancer Susceptibility from Multiple Single Nucleotide Polymorphisms. Clinical Cancer Research (2004), Vol. 10, 2725-2737.

[11] Daly, M., Rioux, J., Schaffner, S., Hudson, T. and Lander, E.: High resolution haplotype structure in the human genome. Nature Genetics (2001) 29, 229-232.