Takestan Institute of Higher Education

Available online at **http://UCTjournals.com**

UCT Journal of Research in Science, Engineering and Technology

UCT. J. Resea. Scien. Engineer.Techno. (UJRSET)

Volume 3,Issue4 ***31-34 (2015)*** *ISSN 2382-977X*

Prediction and Diagnosis of Down Syndrome Disease by using the CHAID Algorithm

Sajad Saydali[1] and Hamid Parvin[2]

[1]Department of Computer Engineering, Yasuj Technical College, Technical & Vocational University, Yasuj, Iran
s.saydali@tvu.ac.ir
[2]Department of Computer Engineering, Assistant Professor, Qeshm International Branch, Islamic Azad University, Qeshm, Iran. parvin@iust.ac.ir

ABSTRACT

Today, development of technology and the use of modern medical equipment and update technology produce massive amounts of stored information in the medical database. Analysis and discovery of knowledge from medical database is difficult due to the high volume of data and is requires a newer technology that data mining technologies to achieve this important to help its powerful algorithms. Data mining techniques with extract knowledge and the unseen patterns from huge volumes of data and build models related to medical databases, designed decision support system that help decision-making to Medical. The main objective of this paper has been to examine how to apply data mining techniques to predict and diagnosis of Down syndrome disease based on the medical information of 200 patients referred to medical laboratories in the country by using data mining algorithm CHAID in Rapid Miner software. The results showed that the CHAID algorithm with 100% accuracy has ability diagnosis of Down syndrome disease.

Original Article:

Received 21 Sep. 2015
Accepted 19 Dec. 2015
Published 30 Dec. 2015

Keywords:
Data Mining, Database, Down Syndrome, Classification Techniques

Introduction

Modern medicine produce large amount of stored information in the medical databases. Need to science is essential to search in this data and knowledge discovery from databases. In fact, knowledge discovery from databases is the process of identify patterns and exist models in the data. Patterns and models that is valid, innovative, potentially useful and fully understood. Data mining is the stage of knowledge discovery process with the help of special algorithms of data mining and acceptable performance computing finds patterns or models in data. One of the fields that are requires use of these tools for analysis extensive data and predictive modeling with new computational methods are medical science. The purpose of data mining predictive techniques in clinical medicine is building a predictive model that helps doctors to improve methods of prevention, diagnosis and treatment programs [1]. Some of the chronic diseases such as diabetes, obesity and cardiovascular disease was the leading cause of death and disability in most countries [2] and can be predicted or detected with explored on previous similar patient data, compare and implementation common symptoms of the disease. Data mining tools can be help in the areas of predicting and diagnosing diseases, the effectiveness of treatment, identify side effects of medications and other medical science. Han and colleagues in 2008, by using data mining algorithms is identified diabetes in the patient's database [3]. Big Hwan Cho et al, in 2008 by using data mining is predicted existence neuropathy in patients with diabetes [4]. Kurosaki and colleagues in 2012, examined to predict cancer in patients with hepatitis c [5]. In the study of Silvera and colleagues in 2014, were investigated risk factors of lifestyle and dietary in patients with gastric cancer [6]. Tavakoli and colleagues in 2010 achieved results that show the use of data mining is much better than the hospital algorithm performance and the doctor's mental model [7]. Syndrome called to set of physical and mental symptoms of a specific condition and Down is the name of English physician that about 200 years before was discovered this set of symptoms. The cause of this phenomenon is a kind of disorder in the arrangement of chromosomes that occurs in embryonic stages and during cell division. Patients with Down syndrome in their body cells have 47 chromosomes instead of 46 chromosomes. Chromosomes are a very small portion of the cell, which genes collected in it and contain information that shapes our body. For example, skin color, eyes, hair, female or male is determined in this small section. The extra chromosome in the body of a person who have Down syndrome takes the impact in the process of the above formation and be causing changes of physical and mental. People with Down syndrome have differences with others. Some of these differences are related to the physical characteristics of individuals and the other part is related to intellectual characteristics, the syndrome occurs before birth.

Nowadays, using of screening methods to distinct high risk from low-risk cases. The purpose of this paper is to examine the role and scope of predictive data mining in medical science and propose an appropriate framework to the construction, operation and evaluation of data mining models in the prediction and diagnosis of Down syndrome. Then, in the second section, this paper is examining the data mining techniques in the field of predicting and diagnosis of disease. In the third section is expressed diagnosis of disease of Down syndrome by CHAID algorithm. The fourth section is contains the results and the fifth section is conclusions.

2. Data Mining Techniques

The main operations data mining divided into two categories of predictive and descriptive. Predictive tasks are used to predict future behavior. The order of prediction is use of multivariable or field in the database to predict future values or other unknown variables of interest. Descriptive tasks are determining data public properties. The purpose of describing is finding patterns in about the data that is interpreted to humans. Of techniques that for predictive methods can be used for classification, regression, slope detection. Classifieds and prediction is the process of identifying set of features and common models that describe and distinguish classes or data concepts [8]. Also, classifieds is the process of finding a model with a diagnosis categories or data concepts can predict of unknown other objects [9]. For example, the classification rules about a disease can be detected on the symptoms and characteristics of current known disease and to identify that disease in new disease used according to their disease symptoms. Medical diagnosis is an important application of the classification. Quentin Trautvetter et al., in 2002, have used the method of association rules and decision tree to extract knowledge from the medical database [10]. Anbananthen et al., in 2007, wielded neural networks and made decision tree of C4.5 algorithm for the diagnosis of diabetes [11]. In this study, Silvera and colleagues in 2004, was designed to evaluate the role of nutritional risk factors on the risk of cancer of the esophagus and stomach by using classification tree models [12]. Valera et al., in 2006, based on the classification tree model began to study predictors of colorectal cancer [13]. Shukla et al., in 2014, began to study techniques classification in prediction and diagnosis of diseases [14]. There are several methods that can be used for data mining of classification issues. One of the most widely used classification method is a methods based on decision tree. Decision tree is one of the most famous and oldest construction methods of classification model. The structure of the decision tree is a tree structure, like a flowchart. The highest node in the tree is the root node and leaf nodes indicate the class or class distribution. In classification algorithm based on decision trees provided the output knowledge in the form of a tree from different states of the characteristics values. Knowledge representation in the form of a tree has caused classification based on decision tree to be fully parsed. Decision trees, according to the decision rules are used to predict and classifieds. In cases that we want to express the results prediction in the form of classes, such as high-risk disease against low-risk disease or the existence or the lack of a disease, it will be a very efficient use of decision trees. Of the most commonly used algorithms of decision trees is CHAID. CHAID algorithm has ability to use all variables in the production of the prediction model. Generated tree by this algorithm is not necessarily binary.

3. Diagnosis of Down Syndrome Disease by using CHAID Algorithm

Down syndrome called to set of physical and mental symptoms of a specific condition. People with Down syndrome have a variety of symptoms, which the most important symbols is severe mental disabilities as well as heart problems, vision and hearing disorders, growth, appearance, and organs. In medical science to early diagnosis of Down syndrome, we use of screening tests. The most famous of these tests, are combinations tests of the first quarter and quadruple second quarter of Quad Rapple. first trimester combined screening test acts on the main variables, the two blood markers of Beta hormone- HCJ free (Free-BHCG), and Pep A article (PAPP-A), the result of blood tests and measuring the thickness of wrinkles the back of the fetal neck or simply NT, obtained by ultrasound for diagnosis of Down syndrome disease. In the first trimester combined screening tests, there are other variables. Table 1 shows the variables along with examples of records to patients. In this article, we attempt to put the first trimester combined screening test data sets by applying the CHAID algorithm, consider to the diagnosis of Down syndrome disease.

Table1. Variables and data sample to evaluate Down syndrome disease

FBHCG	MOM-FBHCG	PAPPA	MOM-PAPPA	NT	MOM-NT	Gestational		Age	Weight	Smoker	Diabetic	Twin	IVF	Down History
						Week	Day							
10.0	0.31	1690.0	0.51	0.70	0.47	12	4	25	55.0	No	No	No	No	No
24.2	0.81	3133.0	0.84	1.30	0.78	12	6	26	67.0	No	No	No	No	No
76.1	2.27	976.4	0.31	2.30	1.53	12	3	23	76.0	Yes	No	No	No	No
24.5	0.65	2708.0	1.06	0.90	0.65	11	6	28	58.0	No	No	No	No	No
41.3	1.38	11480.0	3.08	1.30	0.79	12	6	26	53.0	No	No	No	No	No
246.7	8.25	2061.0	0.55	2.20	1.37	12	6	19	83.0	No	No	No	No	No
48.5	1.35	1715.0	0.61	1.40	0.95	12	1	31	68.5	No	No	No	No	No
83.1	2.25	1192.0	0.44	2.10	1.46	12	0	32	87.0	No	Yes	No	No	No
28.1	0.90	2363.0	0.67	1.40	0.84	12	5	27	67.0	No	No	No	No	No
80.2	2.31	739.2	0.25	1.60	1.04	12	2	29	63.0	No	No	No	No	No

1.35	0.61	0.95
2.25	0.44	1.46
0.90	0.67	0.84
2.31	0.25	1.04

In conducted studies on the records, three features have the greatest impact for the diagnosis of Down syndrome disease and can be selected features are as input data mining models. Table 2 shows selected properties.

Table2. Effective fields in the diagnosis of Down syndrome disease after applying feature selection

MOM-FBHCG	MOM-PAPPA	MOM-NT
0.31	0.51	0.47
0.81	0.84	0.78
2.27	0.31	1.53
0.65	1.06	0.65
1.38	3.08	0.79
8.25	0.55	1.37

4. Conclusion

By applying the CHAID algorithm by Rapid Miner software, obtained decision tree and the results of the diagnosis of Down syndrome disease are shown in Figure 1 and Table 3.

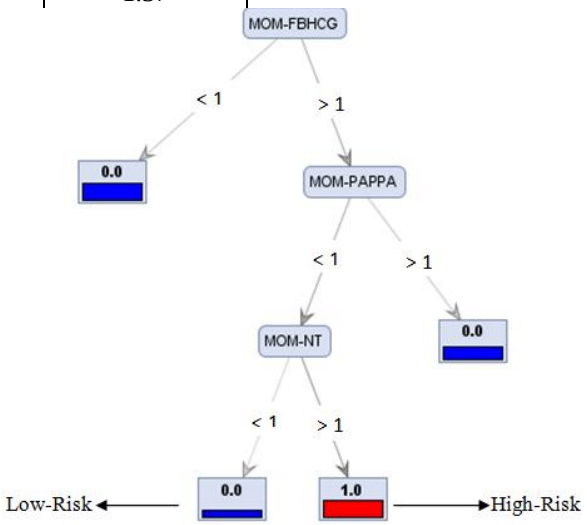

Figure1. Decision Tree obtained from CHAID algorithm

Code related to review decision tree obtained as follows:

if (MOM-FBHCG>1)and(MOM-PAPPA<1)and(MOM-NT>1)

Down_syndrom= High-Risk
else
Down_syndrom= Low-Risk

Table3. Results of the CHAID algorithm

Class	Class Precision	Class Recall
Low-Risk	100%	100%
High-Risk	100%	100%
Average- Precision= 100%		
Average- Recall= 100%		
Accuracy= 100%		

5. Conclusion

The medical field is rich in information, while, and is in need of knowledge discovery. Data mining techniques is a suitable solution to find the required knowledge from medical databases. In medical science, discovery and timely detection of disease can prevent from risk of many deadly diseases such as cancer, and causes lives saves. Using data mining and data modeling can identify patients with high-risk conditions. In fact, data mining by providing information to providers will help their care in identifying high-risk patients, so that, to improve the quality of their care and prevent their future problems and with design appropriate intervention, which resulted in the reduction of hospital admissions. In the conducted research showed that classification techniques plays the most widely used in order to predict and diagnose diseases in data mining. In this article, we discussed to the diagnosis of Down syndrome disease by using CHAID algorithm. The obtained results showed that the CHAID algorithm with 100% accuracy have ability to diagnosis of Down syndrome disease.

References

1. Bellazzi R, Zupan B. Predictive data mining in clinical medicine: Current issues and guidelines. International Journal of Medical Informatics, 2008; 77: 81–97.
2. Naren Ramakrishnan, David Hanauer, Benjamin J.Keller: Mining Electronic Health Records. IEEE Computer 43(10): 77-81, 2010.
3. Han J. Rodriguez J. C. & Beheshti M. Diabetes data analysis and prediction model discovery using rapid miner. In Future Generation Communication and Networking, 2008. FGCN'08. Second International Conference on. vol. 3, 2008; pp. 96-99. IEEE.
4. Cho B. H. Yu H. Kim K. W. Kim T. H. Kim I. Y. & Kim S. I. Application of irregular and unbalanced data to predict diabetic nephropathy using visualization and feature selection methods. Artificial Intelligence in Medicine, 42 no. 1, 2008; 37-53.
5. Kurosaki M, Hiramatsu N, Sakamoto M, Suzuki Y, Iwasaki M, Tamori A,et al. Data mining model using simple and readily available factors could identify patients at high risk for hepatocellular carcinoma in chronic hepatitis C. *J hepatol* 2012;56:602-8.
6. Navarro Silvera SA, Mayne ST, Gammon MD, Vaughan TL, Chow W-H, Dubin JA, et al. Diet and lifestyle factors and risk of subtypes of esophageal and gastric cancers: classification tree analysis. Ann Epidemiol. 2014 Jan; 24(1):50–7.
7. Tavakoli N, Jahanbakhsh M. Opportunities and Challenges of EHR Implementation in Isfahan [Project]. Isfahan: School of Informatics and Management, The university of Isfahan; 2010. p. 3. [In Persian].
8. Zhang D. and L. Zhou, Discovering Golden Nuggets: Data Mining in Financial Application, IEEE Transactions on Systems, Man and Cybernetics, Vol. 34(4), 2004 pp.513-522
9. Jiawei Han, Micheline Kamber, 2006, Data Mining Concepts & Techniques, Elsevier Inc.
10. Quentin-Trautvetter J. Devos P. Duhamel A. & Beuscart R. Assessing association rules and decision trees on analysis of diabetes data from the Diab Care program in France. Studies in health technology and informatics. 2002; 90, 557.
11. Anbananthen K. S. M. Sainarayanan G. Chekima A & Teo J. Artificial Neural Network Tree Approach in Data Mining. Malaysian Journal of Computer Science, 20 no. 1, 2007; 51.
12. Silvera SAN, Yale University. Dietary factors and risk of subtypes of esophageal and gastric cancer. Diss Abstr Int.
13. Valera VA, Walter BA, Yokoyama N, Koyama Y, Iiai T, Okamoto H, et al. Prognostic groups in colorectal carcinoma patients based on tumor cell proliferation and classification and regression tree (CART) survival analysis. Ann Surg Oncol. 2007 Jan; 14(1):34–40.
14. Shukla D.P, Shamsher Bahadur Patel, Ashish Kumar Sen. Literature Review in Health Informatics Using Data Mining Techniques. International Journal of Software and Hardware Research in Engineering, Volume 2, Issue 2, February 2014