J. Appl. Environ. Biol. Sci., 5(11S)365-369, 2015

ISSN: 2090-4274
Journal of Applied Environmental and Biological Sciences
www.textroad.com

The Use of Data Mining Classification Techniques to Predict and Diagnose of Diseases

Sajjad Saydali[1*], Dr. Hamid Parvin[2]

[1]Department of Computer Engineering, Master of Computer Engineering, Qeshm International Branch, Islamic Azad University, Qeshm, Iran.
[2]Department of Computer Engineering, Assistant Professor, Qeshm International Branch, Islamic Azad University, Qeshm, Iran.

Received: May 14, 2015
Accepted: August 27, 2015

ABSTRACT

Nowadays biological progresses and technology development and use of new technology and modern medical equipment produce massive amounts of information stored in medicine databases. Analysis and knowledge discovery is difficult from the medical database, due to the large volume of information and is needed to newer technology that data mining technologies has achieved to this with the help of its powerful algorithms. Data mining techniques extracted the hidden patterns among the data and with build the model of databases design the decision support system that in the field of decision will help to doctors. In this article, which is a review article, its purpose is to discuss on the application of data mining and classification techniques in predicting and diagnose diseases.
KEYWORDS: Database, Data Mining Techniques, Classification Techniques, Prediction and Diagnosis of Diseases

1. INTRODUCTION

Today, with advances in science and technology has been provided the possibility to save important data with a large volume and is necessary requiring to science to search in this data and knowledge discovery from databases. In fact, knowledge discovery from databases is a process of pattern recognition and in the data existing models. Patterns and models that is valid, exquisite, potentially useful and completely understandable. Data mining is the process of knowledge discovery process with the help of special algorithms of data mining and with computing acceptable performance find patterns or models in data. One of the areas, which require the use of these tools for large-scale data analysis and predictive modeling with new computational methods, is medical science. The goal of predictive of data mining techniques in clinical medicine is building a predictive model, which will help to doctors to improve the prevention methods, diagnosis and your treatment programs [1]. Data mining tools can help in the areas of predicting and diagnosing of diseases, the effectiveness of treatment, and identification adverse effects of drugs to medical science. This article shows that predictive of data mining provides essential tools for researchers and clinicians to improve in the prevention of diseases, diagnostic methods and treatment programs. Then, in the second part of this article, we discussed to the review, the role and scope of predictive of data mining in medical science to predict and diagnosing of diseases. Classification of data mining techniques in the field of prediction and diagnosis of disease expressed in third part. The fourth part is includes conclusion.

2. The use of data mining to predict and diagnosing of diseases

Modern medicine produces massive amounts of information stored on medical databases. Data mining can have numerous and varied applications in the medical field, which most important of them are to predict and diagnose of diseases. Some of the chronic diseases, such as diabetes, obesity and cardiovascular disease has been the major cause of death and disability in most countries [2], and can be predicted or detected by exploring on previous similar patient data, compare and implementation of the common signs of disease. Jianko Han et al. in 2008, by using the decision tree algorithm ID3, detected the presence of diabetes in the patient's database [3]. Bik Hwan Cho et al. in 2008, by using the category of support vector machine has predicted the existence of neuropathy in diabetic patients [4]. Kurosaki et al. in 2012 began to predict cancer in patients suffering from c hepatitis. They looked at characteristics to predict cancer, such as age, platelet count, albumin, and aminotransferase, and began to identify patients who have a high chance of developing cancer [5]. In the study of Silvera and colleagues in 2014, were examined lifestyle and dietary risk factors in patients with gastric cancer. In patients with squamous cell carcinoma of the esophagus, smoking is as the primary risk factor and after that, Esophageal Reflux, income, race, non-citrus

[*] **Corresponding Author:** Sajjad Saydali, Department of Computer Engineering, Master of Computer Engineering, Qeshm International Branch, Islamic Azad University, Qeshm, Iran.

fruits and energy intake have been proposed as a risk factor [6]. Tavakoli et al. in 2010, achieved results that show the use of data mining is much better than the performance of the hospital algorithm and doctors mental model [7].

3. Data mining techniques

The main operations of data mining divided into two categories, predictive and descriptive. Predictive tasks used to predict their future behavior. The order of predict, which use several variables or fields in a database to predict future values or unknown of other interest variables. Descriptions tasks define general properties of the data. The goal of description is to find patterns in the data, which is interpreted to humans. Data mining methods is shown in Figure1.

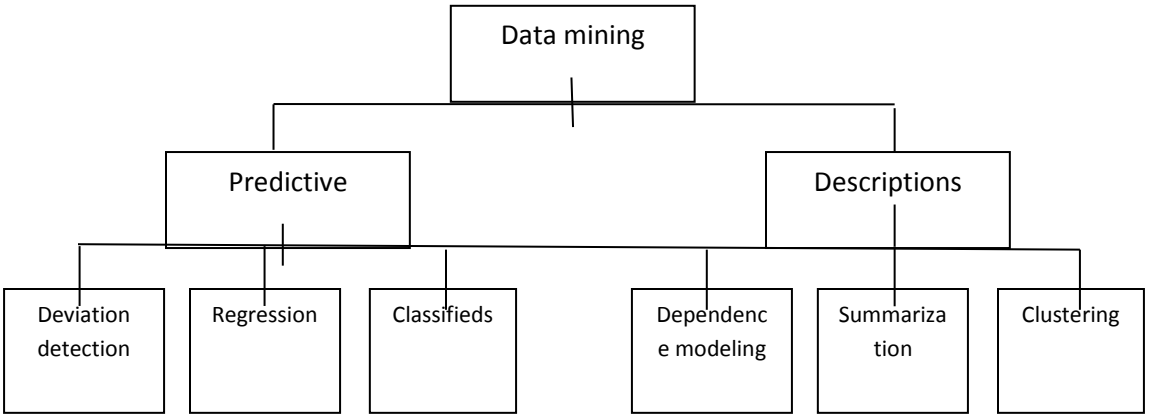

Figure1. Data mining methods

In the following, we will examine the classification techniques and applications of classification techniques in predicting and diagnose of diseases.

3. 1. Classification

Classification and prediction are the process of identifying a set of features and common models, which describe and distinguish classes or concepts of data [8]. Classification is the process of finding a model to identify categories or concepts of data can predict the unknown category of other object [9]. For example, the classification rules about a disease can be discovered from the signs and characteristics of current known disease and to identify the disease in new disease used according to their disease symptoms. In the classification of existing data divided into two parts of the training data set and test data sets. Using the training data set made the model and from the test data set used for validation and calculation accuracy of the model. Each record is includes a set of features.

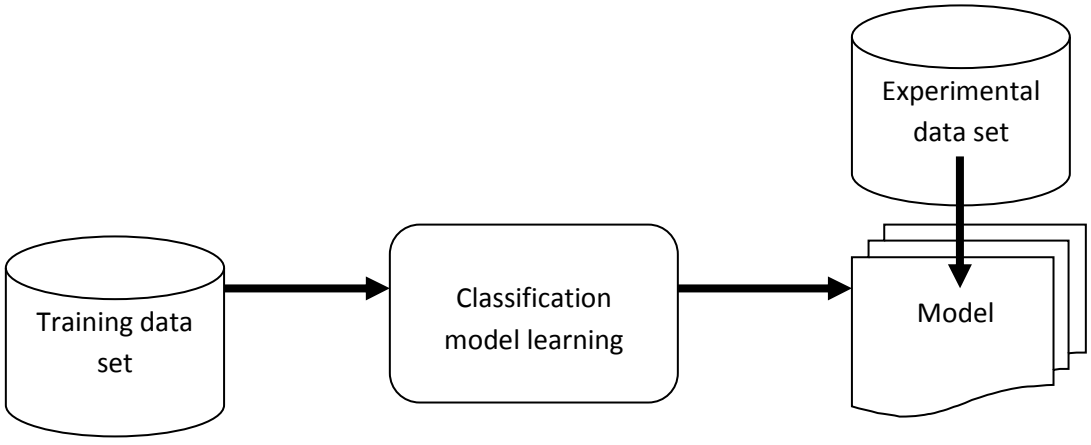

Figure2. The process of classification operations method

One of the features called a class feature. In the training stage, the training data set given to one of the classification algorithm, up based on values of other features for the class feature values built model. Form of built

model depends on the type of learning algorithm. For example, if the learning algorithm is a decision tree algorithm, the built model will be a decision tree. In any case, according to used learning algorithm in the training stage, the model is built. After building the model at the evaluation stage, the built model accuracy with the help of experimental data sets, which was built the model during the training stage, has not seen this data set will be evaluated. A set of test data is not used in the teaching and build the model. Figure 2 above shows the classification operations process.

3. 2. The use of classification to predict and diagnosing of diseases

Data mining techniques, as well as play an important role in finding patterns and knowledge extraction in order to contribute to the effective detection of diseases and provide better services and medical care [10]. Used algorithms in data mining techniques trying to find and present the closest model to the characteristics of the desired data. Quentin-Trautvetter has used the method of association rules and decision tree to extract knowledge from the medical database [11]. Anbananthen et al., in 2007, have used the neural network and made decision tree from the C4. 5 algorithms for the diagnosis of diabetes [12]. In the study of Silvera and colleagues in 2004, was conducted to evaluate the role of nutritional risk factors on the risk of cancer of the esophagus and stomach by using of the tree classification model [13]. Valera et al., in 2006, based on the classification tree model began to study predictors of colorectal cancer [14]. Shukla et al., in 2014, began to explore data mining techniques in predicting and diagnose of diseases as shown in Table1 [15].

Table1. Data mining techniques to predict and diagnose of diseases [15]

Use	Technique
Find the diseases common items in medical databases	Appriori and FPGrowth
Classification of medical data	Genetic Algorithm
Extraction patterns, identify trends	Neural Networks
Find common patterns	Association Rule Mining
Classification	Bayesian Ying Yang (BYY)
Decision Support	Decision Tree Algorithms Such as ID3, C4. 5, C5, and CART .
To improve the accuracy of classification	Outlier Prediction Technique
The analysis of medical images	Fuzzy cluster analysis
Classification of diseases	Classification Algorithm
Modeling and analysis of medical data	Bayesian Network algorithm
Improve the accuracy of classification	Naïve Bayesian
The exact classification from the medical data	Combined use of Kmeans , SOM and Naïve Bayes
Medical diagnosis	Time Series Technique
Classification of medical data	Combination of SVM, ANN and ID3
Clustering and classification of medical database	Clustering and classification
Classification of diseases	SVM
Classification of drugs and health effects	Fuzzy Cognitive Maps
Classification of diseases	k-NN

Medical diagnosis is an important application of the classification. In classification techniques, at first in training stage on the training data set built models and in the performance evaluation stage and will determine the accuracy of the model. Suppose, we have a record set related to the Down syndrome disease that each record is related to a patient. From each patient, we have a number of features including MOM-FBHCG, MOM-PAPPA, MOM-NT, age, smoker, diabetes and history of Down syndrome disease and also know what is the kind of disease risk of each patient, ie, have a class feature to the name of the risk type of Down syndrome disease, which is two values of high-risk and low-risk. Each record has a class feature values, meant that one of the high risk and low risk. For example, suppose the problem is includes thousands of records, one hundred patients with high risk and nine hundred healthy person with low risk. The purpose of building a model is to high and low risk class. So that, if entered into a new patient that model to identify the new patient belongs to which of the two classes. However, when with new patient given the patient characteristics to the model recognize the class models associated with the patient. Obviously, the diagnosis is based on classes that model in training stage is faced with them. So, there will not be new class detection in classification application. In Tables 2 and 3 show the various stages of the process of classification to predict and diagnosis of Down syndrome disease. As can be seen in this process, there is a learning algorithm, which is based on the training data set builds a model. This model applied on experimental data set and in the next step will be calculated its accuracy.

Table2. Learning data set for model training

Record	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6	Feature 7	Class
1	1. 18	1. 40	0. 85	33	No	No	No	**Low risk**
2	2. 27	0. 31	1. 53	23	No	No	No	**High risk**
3	0. 43	0. 74	0. 89	25	No	No	No	**Low risk**
4	0. 91	1. 30	0. 90	30	No	Yes	No	**Low risk**
5	2. 15	0. 92	1. 12	40	Yes	No	No	**High risk**
6	0. 93	0. 91	0. 77	21	No	No	No	**Low risk**
7	1. 42	1. 78	0. 69	25	No	No	No	**Low risk**
8	2. 25	0. 44	1. 46	32	No	No	No	**High risk**
9	3. 27	4. 30	1. 22	26	Yes	No	No	**Low risk**
10	0. 61	1. 20	0. 96	22	No	No	No	**Low risk**

Table3. Experimental data set for model actions

Record	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6	Feature 7	Class
1	0. 90	2. 23	0. 89	26	No	Yes	No	**?**
2	2. 30	0. 31	1. 50	23	No	No	No	**?**
3	1. 28	0. 71	0. 85	24	No	No	No	**?**
4	3. 19	0. 50	1. 08	32	Yes	No	No	**?**
5	1. 66	0. 55	0. 34	29	No	No	No	**?**

Each of the experimental data set records compared with the training data record set and receive answers their class. For example, if we have a new disease with the following characteristics the patient characteristics given to the model, the model determines the class related to that patient. Due to the characteristics of this patient is on the record No. 2, the result of its class is equal to "High risk".

Record	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6	Feature 7	Class
1	2. 35	0. 40	1. 47	24	No	No	No	**High risk**

4. Conclusions

The medical field is rich in information, while 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 be prevented from getting to many life-threatening diseases, such as cancer, and lead to saves people lives. By using of data mining and data modeling can identify the patients with high-risk conditions. In fact, data mining by providing information to care providers helping them in identifying high-risk patients, so that to improve the quality of their care and avoid from problems of their future and to design appropriate interventions, which resulted in the reduction of hospital admissions. Research conducted shows that classification techniques in data mining plays the most widely used in order to predict and diagnose of diseases. In classification, based on available data set related to patient built a model, then be evaluation the new patient information based on that model and is recognized the class of patient.

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. Harleen Kaur, Siri Krishan Wasan and Vasudha Bhatnagar, THE IMPACT OF DATA MINING TECHNIQUES ON MEDICAL DIAGNOSTICS, Data Science Journal, Volume 5, pp119-126, 19 October 2006
11. Quentin-Trautvetter J. Devos P. Duhamel A. & Beuscart R. Assessing association rules and decision trees on analysis of diabetes data from the DiabCare program in France. Studies in health technology and informatics. 2002; 90, 557.
12. 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.
13. Silvera SAN, Yale University. Dietary factors and risk of subtypes of esophageal and gastric cancer. Diss Abstr Int.
14. 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.
15. 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