A Survey on Data Mining Techniques in the Medicative Field

Chinky Gera
M.Tech(CSE) Student
Department Of Computer Science & Engineering
RIMT Institute of Engineering and Technology

Kirti Joshi
Assistant Professor
Department Of Computer Science & Engineering
RIMT Institute of Engineering and Technology

ABSTRACT

Data mining is the process of releasing concealed information from a large set of database and it can help researchers gain both narrative and deep insights of exceptional understanding of large biomedical datasets. Data mining can exhibit new biomedical and healthcare knowledge for clinical decision making. Medical assessment is very important but complicated problem that should be performed efficiently and accurately. The goal of this paper is to discuss the research contributions of data mining to solve the complex problem of Medical diagnosis prediction. This paper also reviews the various techniques along with their pros and cons. Among various data mining techniques, evaluation of classification is widely adopted for supporting medical diagnostic decisions.

General Terms

Data Mining, Classification, Medical.

Keywords

Classification, Decision Tree, K means Clustering, Naïve Bayes, WEKA.

1. INTRODUCTION

Data mining is one of the provoking and significant area of research. Data mining is implicit and non-trivial task of identifying the viable, novel, inherently efficient and perspicuous patterns of data. Figure 1 represents the data mining as part of KDD process [15]. The hidden relationships and trends are not precisely distinct from reviewing the data. Data mining is a multi-level process involves extracting the data by retrieving and assembling them, data mining algorithms, evaluate the results and capture them. [10] Data Mining is also revealed as necessary process where bright methods are used to extract the data patterns by passing through miscellaneous data mining processes.

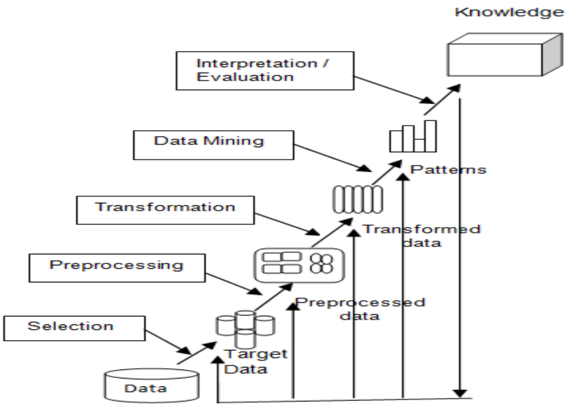

Fig 1: Data Mining Processes [15]

The cause of data mining effort is usually either to design a descriptive model or a predictive model. Descriptive mining functions specify the common properties of the data in the database. Predictive mining functions perform the reasoning on the present data in order to form predictions. From a general view, there is a robust consent between both researchers and executives about the standard that all data mining techniques must meet. Classification and Prediction are two designs of data analysis that can be used to extract models illustrating the essential data classification or to predict the future data mode. Such analysis can assist to provide with a better knowledge of the data tremendously. [4]

2. DATA MINING TECHNIQUES IN MEDICAL SCIENTIFIC DISCIPLINE

In medical data mining, some techniques like Multilayer Perceptron, J48 Decision Tree, KNN, K-Clustering, Random Forest has been widely used by authors in their research work. Table 1 enlists the pros and cons data mining techniques. We have briefly discussed these techniques:

2.1 Neural Network

Classification is one of the most dynamic research and significant area of neural networks. This paper summarizes vital growth in neural network classification analysis. In medical field, the neural network manipulates the predictive decision making by the described set of rules. Neural network provide powerful mechanism to help the physicians to review, model and make sense of complex clinical data across medical applications. [6] [9]

2.2 Decision Tree

There are some standard tree algorithms are implemented: ID3 and C4.5 (called version J48). J48 is mostly used as compared to ID3 as J48 yields better results in any context. In the WEKA data mining tool, J48 is an enhanced version of C4.5 algorithm. The decision tree generated by C4.5 is used for classification process. [14] [1]

2.3 K-Nearest Neighbor

K-Nearest Neighbor is a simple and powerful statistical unsupervised clustering approach. KNN can work with little information or no prior information of data distribution. KNN is also known by various names: Lazy learning, memory based reasoning, example based reasoning, instance based learning, case based reasoning.

2.4 K-means Clustering

It is a simplest, quantitative and iterative method used for aggregating enormous sets of data.. An algorithm used to organize the objects based on features into number of clusters. Its main purpose is to define k centers, one for every cluster. These centers should be placed by a deceptive means as different location needs different results. [2]

2.5 Naïve Bayes

Naïve Bayes is one of the oldest traditional classification algorithm. Naïve Bayes is extremely attractive because of its simplicity, grace and robustness. It is a statistical and supervised learning method which can resolve difficulty involving both categorical and continuous estimated attributes.

Table 1: Prons and Cons of Classification Techniques

Methods	Advantages	Disadvantages
Neural network	1. error prone 2. robust in noise environment 3. very efficient	1.high complexity model with long duration of training 2. local minima 3. over fitting
Decision Tree	1.easy to understand 2.easily incorporated into real time system 3.produce a set of rules that are transparent	1.Small variation in data can lead to different decision trees. 2.Does not work very well on a small training set
K-NN	1. easy to use and understand 2. fast training required 3. high presentation ability 4. no optimization	1. computation slow 2. sensitive to representation 3.slow testing 4. huge area for storage
K-means Clustering	1. Fast 2. Easy to understand 3. Fairly efficient	1. Algorithm fails for non linear dataset 2. Unable to handle turbulent data and outliers
Naïve Bayes	1. easy to perform 2.Good outcomes in most cases 3. often used as punching bag for elegant algorithms	Loss of validity

3. DISCUSSION OF PAPERS

In this literature Survey we present an overview of the current research being carried out using the data mining techniques for the diagnosis and prognosis of various diseases. The following algorithms have been identified: Decision Trees, Artificial neural networks and Naïve Bayes. Analysis shows that it is very difficult to name a single data mining algorithm as the most suitable for the diagnosis and/or prognosis of diseases. At times some algorithms perform well than others. The Table 2 below shows the survey made in various techniques used in data mining for the medical diagnosis.

Table 2: Literature review

Author	Year	Knowledge Type	Knowledge resource	DM techniques/ applications
Jyoti Soni et.al [13]	2011	Heart Disease Prediction	Decision tree outperforms and sometimes Bayesian classification ’s having similar accuracy as Decision tree.	Classification: Clustering Bayesian classification, Neural Networks Decision Tree, KNN
Samar Al-Qarzaie et.al [8]	2011	Breast Cancer Disease	WEKA tool is used to give 93.4675% accuracy in testing set and in the training set it yields 96.8% accuracy	Classification: Decision Tree
Arvind Sharma et.al [11]	2012	Blood Donors	By using WEKA tool, J48 decision tree acquires 89.99% accuracy	Classification: J48 Decision Tree
Shweta Kharya [5]	2012	breast cancer diagnosis and prognosis	Decision tree is a best predictor with 93.62% accuracy	Classification: Neural Network, Association, Naive.Bayes, C4.5 decision tree algorithm
Dr. Bushra M. Hussan [3]	2012	Prediction of medical data by K means Clustering	On changing the instances it shows 97% of accuracy.	Classification: K-means, Clustering
M. Durairaj et.al [2]	2013	Applications in healthcare sector	WEKA tool gives 97.77% accuracy for cancer prediction and about 70% accuracy for success rate of IVF treatment.	Classification: Artificial Neural Network

Aarti Sharma et.al [12]	2014	Applications of Data Mining	Decision tree shows good accuracy for processing raw data into information and find patterns	Classification: association, rule clustering, prediction and Evaluation pattern
Mehak Naib et.al [7]	2014	Predict Primary Tumors using multiclass classifier approach	Multiclass classifier gives good accuracy by using WEKA tool than binary classifier	Classification: Random Forest

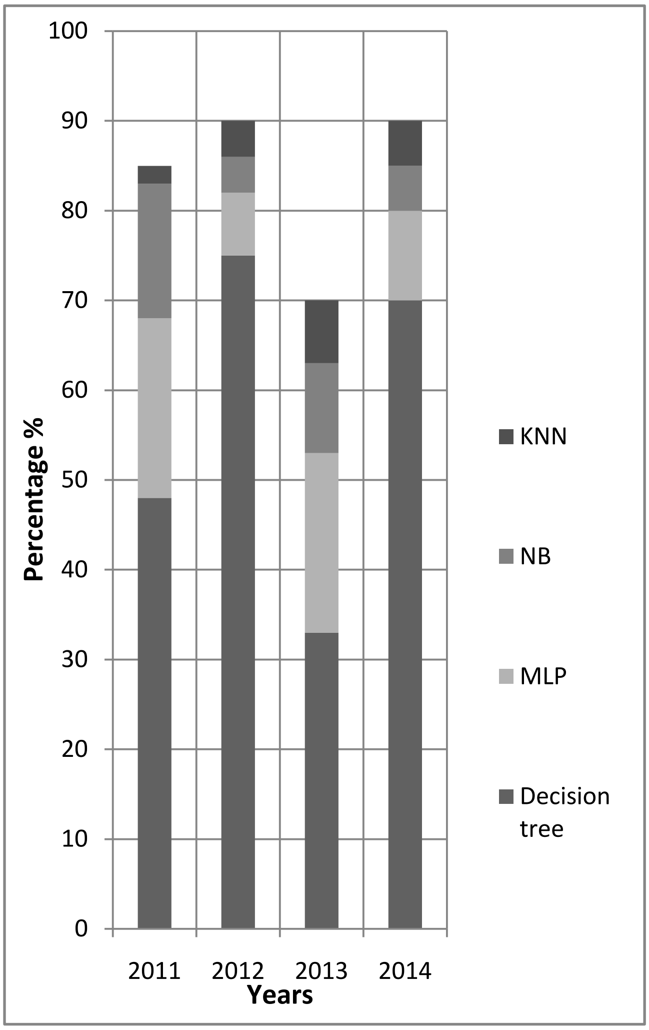

Figure 1: Survey of Algorithms used in Medical datasets

Different algorithms like Decision tree, K- Nearest Neighbor (KNN), Naïve Bayes (NB), Multilayer perceptron (neural network) are used in figure 1 presents that which algorithm is used most in the Literature review in medical field. The outcome reveals that decision tree is the best among other algorithms which shows higher accuracy as compared to other algorithms.

4. CONCLUSION AND FUTURE SCOPE

The main purpose of this survey was to discover the most typical data mining algorithms. The ideas of future work include the evaluation of chosen algorithms on the basis of chosen medical dataset. Other algorithms can be applied on built-in dataset and the algorithm which gives best result will be applied on the test dataset. The experiments would be conducted on the selected medical records which design the analysis even more accurate. The good idea is taking also other algorithms to the experiments and compares their performance in medical field. This would evolve a new class and assist in scheming Medical Decision Support Systems by the selection of the most acceptable algorithms. Other methods which are not included in this survey on comparison basis and can discover the best one by assessing the advantages and limitations of the prevailing one.

5. ACKNOWLEDGEMENT

This paper using an opportunity to convey the gratitude to parents, experts and friends for their aspiring guidance and helpful encouraging criticism.

6. REFERENCES

[1] Chaurasia V. et al. Mining Approaches to Detect Heart Diseases, International Journal of Advanced Computer Science and Information Technology, Vol. 2, No. 4, ISSN: 2296-1739, 56-66.

[2] Durairaj M. et al. 2013. “Data Mining Applications in Healthcare Sector: A Study”, International Journal of Scientific and Technology Research, Vol. 2, ISSN 2277-8616.

[3] Hussan D. 2012. “Data Mining based Prediction of Medical data using K-means algorithm”, Basrah Journal of Science(A), Vol. 30(1), 46-56.

[4] Jain N. et al. 2013. “Data Mining Techniques: A Survey Paper”, International Journal of Research in Engineering and Technology, Volume: 02, Issue: 11, eISSN: 2319-1163 | pISSN: 2321-7308.

[5] Kharya S. 2012. “Using Data Mining Techniques for Diagnosis and Prognosis of Cancer Disease”, International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol. 2, No. 2.

[6] Mittal P. et al. 2013. “Study and Analysis of Predictive Data Mining Approaches for Clinical Dataset”, International Journal of Computer Applications, Volume 63, No. 3.

[7] Naib M. et al. 2014. “Predicting Primary Tumors using Multiclass Classifier Approach of Data Mining”, International Journal of Computer Applications (0975 – 8887), Volume 96, No. 8.

[8] Qarzaie S. et al. “Using the Data Mining Techniques for Breast Cancer Early Prediction”.

[9] Rani K. 2011. “Analysis of Heart Dksiseases Dataset Using Neural Network Approach”, International Journal of Data Mining and Knowledge Management Process (IJDKP), Vol. 1, No. 5.

[10] Saurkar A. et al. 2014. “A Review Paper on various Data Mining Techniques”, International Journal of Advance Research in Computer Science and Software Engineering, Volume 4, Issue 4, ISSN: 2277 128X, pp. 98-101.

[11] Sharma A. et al. 2012. "Predicting the Number of Blood Donors through their Age and Blood Group by using Data Mining Tool", International Journal of Communications and Computer Technologies, Volume 01, No. 6, ISSN Number: 2278-9723.

[12] Sharma A. et al. 2014. "Applications of Data Mining - A Survey Paper", International Journal of Computer Science and Information Technologies, Vol. 5(2), ISSN: 0975-9646, 2023 - 2025.

[13] Soni J. et al. 2011. "Predictive Data Mining Diagnosis: An Overview of Heart Disease Prediction", International Journal of Computer Applications (0975-8887), Vol. 17, No. 8.

[14] Upadhyay N. et al. 2014. "A Survey on the Classification Techniques in Educational Data Mining", International Journal of Computer Applications Technology and Research, Vol. 3, ISSN 2319-8656, pp. 725-728.

[15] Gennaro Costagliola et.al 2009. "Monitoring Online Tests through Data Visualization", IEEE Transactions on Knowledge & Data Engineering, Issue No.06 – June, vol.21, pp: 773-784.