A BRIEF STUDY OF DATA MINING TECHNIQUES TO STUDY PATTERNS IN MEDICAL SCIENCES

Amanpreet Kaur[1], Priyanka[2]

[1,2] *Computer Science and Engineering Swami Vivekanand Institute of Engineering and Technology*

---***---

Abstract - *In the current era of technology, evolution of medical sciences becomes an active field of research as people have more curiosity towards their health. Different techniques of data mining are used to mine the information from various data patterns. Prior, PC was used to manufacture an information based clinical result which utilizes learning from therapeutic specialists and moves this information into PC calculations physically. This process takes lot of time and gives subjective results as this information only depends on medical professional only. To overcome these type of problems various techniques of machine learning are used to extract important medical patterns from the raw data. In this paper, we have critically analyzed various data mining techniques to gather informative patterns from data sets in medical sciences.*

Key Words: **data mining, machine learning, data sets, informative patterns, raw data**

1.INTRODUCTION

Computational health informatics is an emerging research topic which involving various sciences such as biomedical, medical, nursing, information technology, computer science, and statistics [1]. To extract hidden patterns and relationships from large databases, data mining merges statistical analysis, machine learning and database technology. In several areas of medical services, including prediction of the effectiveness of surgical procedures, medical tests, medication, and the discovery of relationships among clinical and diagnosis data, data mining techniques have been applied [2]. In medical science, doctor's facilities introduced different data frameworks with a lot of information to manage medical insurance and patient information but unfortunately, data are not mined to discover hidden information for effective decision [2][3]. Clinical test outcomes are regularly made on the basis of doctors' perception and experience rather than on the knowledge enrich data masked in the database and sometimes this procedure prompts inadvertent predispositions, doctor's expertise may not be capable to diagnose it accurately which affects the disease diagnosis system [2] [3].

Data warehousing and Data Mining offers a comprehensive support for gathering, analyzing and presenting medical data. Clinical decisions are often made using the doctor's prescription and experience in his or her field rather than the knowledge base which is rich in data hidden in the database [4]. Usually such practices results in errors, wrong advice to the patients in case if the doctors are in fatigue and stress, unwanted biases and also leads to the extravagant medical price which directly affects the quality of services provided to patients.[4]

Clinical decision support integrated with computer generated patient records could enhance patient safety, diminish medical errors, improve victim outcome and reduce unfavorable practice variation [5].

2. DATA WAREHOUSING AND DATA MINING

The term data warehouse was introduced by W. H. Immon as: "A Data Warehouse is a subject-oriented, time-variant, integrated, and non-volatile collection of data in support of management's decision making process" [3]. The aim of Data warehouses is to collect and archive important operational data and pure data. Data Warehouses are used to predict similar trends instances in the data and is used for decision support and performing the analysis on historical data and legacy data. One of the major tasks in data warehousing is to perform data cleansing on the data collected from various input sources. Schemas in data warehousing tends to be in multidimensional form, which involves one or a few some large fact tables and some amount of smaller dimension tables.

The term data mining generally refers to the process of automatically examining large databases to extract useful and hidden patterns. The term data mining generally refers to the process of automatically examining large databases to extract useful and hidden patterns. The same way knowledge discovery in broad area artificial intelligence which is also known as statistical analysis or machine learning, the concept of data mining evolves to discover hidden rules and patterns from the huge amount of data. The term Data mining is different from machine learning and statistics in the way that it manages large volumes of data, which is primarily stored on disk. That is, data mining deals with "knowledge discovery in databases."

The extraction of non trivial, implied, hidden and actionable knowledge from large volume of datasets is performed by data mining. Discovery of the useful knowledge that can improve proficiency of processes cannot be handled manually.

Data mining usually implements two types of strategies:

- Supervised Learning Strategy
- Unsupervised Learning Strategy

In a supervised learning method, a training set is already available which is used to learn parameters. Classification algorithm uses supervised learning method approach. Each of these data mining techniques uses a different approach depending upon the purpose of modeling objective [10]. There are usually two common modeling objectives viz. Classification and Prediction. Classification model predicts the categorical data that is in discrete and unordered form whereas prediction model predicts the continuous valued data [10]. In unsupervised learning method, no training set is available to learn the parameters. Clustering algorithm uses unsupervised learning method approach. There are various clustering algorithms like K-mean clustering algorithm, K-mediod algorithm, DBSCAN and OPTICS, hidden markov algorithm. Unsupervised learning provides the capability to learn more larger and complex models. In unsupervised learning strategy, the learning can be preceded in hierarchical fashion from the observations resulting into ever more deeper and abstract levels of representation.

3. LITERATURE REVIEW

Candice MacDougall, Jennifer Percival and Carolyn McGregor (2009), *"Integrating Health Information Technology into Clinical Guidelines"* brings the use of research based evidence into practice so as to develop clinical guidelines into practice. This paper provides a review of current research on the integration of Health Information Technology (HIT) into clinical guidelines so as to achieve more accurate results [8].

K.Srinivas, B.Kavihta Rani and Dr. A.Govrdhan (2010), *"Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks"* uses data mining application techniques in the Healthcare and the prediction of Heart Attacks. The author has deeply examined the use of data mining techniques in classification such as Rule based technique, Decision Tree technique, Naïve Bayes Classification technique and Artificial Neural Network technique for the extraction of huge amount of patterns from the abundant data which is not mined so as to discover the hidden information from the data. The author implemented the data preprocessing tasks and decision making one dependency augmented Naïve Bayes classifier (ODANB) and Naïve credal classifier 2 (NCC2) and then the ODANB is equated with the existing methods that improve the Naïve Bayes with the Naïve Bayes itself. It can predict whether the patient is suffering from a heart disease or not using several medical parameters such as patient's age, sex, blood pressure and blood sugar etc. In the rule based technique the rules were piled in the database in the form of IF-THEN rules that is the antecedent part resulting in the conclusion part. Decision Tree includes Classification and Regression Tree (CART), Iterative Dichotomized 3 (ID3) and C4.5. The results when compared show that the ODANB is better than the other methods likes Naïve Bayes. The author also concluded that there are several problems and constraints of using different algorithms of data mining. The hidden patterns can be extracted regarding the prediction of heart attack from data warehouses [11].

Mai Shouman, Tim Turner, Rob Stocker (2012), "*Using Data Mining Techniques in Heart Disease Diagnosis and Treatment",* determines the gap resulting from the previous theories of research on diagnosis and treatment of heart disease and introduces a model to systematically close those occurring gaps to discover if we apply techniques of data mining to the heart disease treatment data then it can provide a reliable performance than it is achieved in diagnosing heart disease. In this paper author has used hybrid data mining techniques which includes naïve density, bagging algorithm and support vector machine. In this literature survey, the different datasets which are being used in the previous year papers using different-different techniques are being defined and their accuracy is measured so as to determine the various algorithms of data mining used in the diagnosis of heart disease. The author uses single type algorithm like Naïve Bayes, Decision Tree, Bagging algorithm and hybrid type algorithm like Fuzzy-AIRS-K-nearest neighbor, Neural network ensembles and then determine the accuracy. The results show that the hybrid type approach is having more accuracy than the single type as the maximum accuracy achieved by using single data mining technique is 84.14% by naïve bayes while the accuracy achieved by using hybrid data mining technique is 89.01% by neural network ensemble. This paper results in the output that heart disease can be predicted with higher accuracy with hybrid data mining techniques [10].

Bata Sundar V, T Devi and N Saravanan (2012), "*Development of data Clustering Algorithm for Predicting Heart"* finds out the accuracy of the result by using the K-means Clustering Algorithmic Technique for the prediction and diagnosis of Heart disease. It uses two datasets – real and artificial datasets. The real dataset is the dataset taken from the real life patients of hospitals and patients of laboratory tests whereas the artificial dataset is the dataset taken from the UCI machine learning Databases, 2004.The author in this research mainly focuses on the prediction of Heart disease using K-means Clustering in context of Data Mining. The author first pre-processes the dataset which includes the various steps like eviction of duplicate records, finding out the missing values from the dataset, removing the outliers and noise and normalizing the various values which are used to portray the information in the databases. Then the preprocessed heart disease data is taken and clustered using the K-means algorithm. 13 attributes were taken in the dataset and the performance and analysis of the various algorithms like Decision Trees, Naïve Bayes , Neural network and K-means is done. Each algorithm is compared with another algorithm in terms of its accuracy and time taken to predict .According to this research been taken place the highest accuracy is of the K-means with 66.00% and the time taken is 8 second. The second highest accuracy is 39.96% with the shortest time taken that is of 4sec by the neural network. The lowest accuracy is of Decision Tree with 24.73% with time taken of 10 seconds [12].

Nirmala Devi M., Appavu alias Balamarugan. S, Swathi U.V (2013), *"An Amalgam ANN to predict Diabetes Mellitus"* presents the development of an amalgam model for classifying Pima Indian diabetic database (PIDD). This amalgam model combines K-means with K-Nearest Neighbour (KNN). They compare the results of simple KNN with cascaded K-means and KNN for the same k-values. It uses the following algorithms: KNN Classifier, K-means partitioning and Amalgam KNN. The dataset is taken from UCI Machine learning data repository for diabetes mellitus that is PIDD. This dataset is from Indian Pima Diabetes Datasets. The results are then compared by measuring the statistical measures such as accuracy, sensitivity and specificity and calculated using WEKA tool. For k=5, K-means and KNN has accuracy of 97% while the simple KNN has the accuracy of 73.17% and Amalgam KNN has accuracy of 97.4%. For k=3, Amalgam KNN has accuracy of 96.87% and simple KNN has accuracy of 72.65%.The author concluded that performance of the algorithm increases if the value of K increases [9].

Gunasekar Thangarasu, Assoc. Prof. Dr. P.D.D. Dominic (2014), *"Prediction of Hidden Knowledge from Clinical Database using Data mining Techniques"* predicts the diabetic disease from clinical database by using Neural Network algorithm. This research was being conducted with the various objectives like it identifies the various complications that cause diabetes from clinical databases through Fuzzy logic Techniques. It develops a Hybrid Genetic Algorithm that computes the best fitness value which is used for evaluating the prediction accuracy of diabetes from clinical databases. It also identifies the type of diabetes the patient is suffering from through data clustering algorithms from the clinical database. Then, it will analyse the performance of projected algorithms. It uses hybrid system model to identify diabetes mellitus, its types and complications which uses certain algorithms like Neural-network algorithm, Fuzzy logic techniques, Hybrid Genetic Algorithm and Data clustering algorithms. The dataset is collected via questionnaire distribution with participants. All the dataset was organized and analysed using a computer program SPSS 20. This model was successfully implemented with input as symptoms that may appear in an individual during the early stages of diabetes and also based on the physical condition of the individual. This research study avoids the patients from undertaking certain blood tests, checking the diastolic and systolic blood pressure etc. Thereby creating a user friendly interface and environment for the patient's without any requirement of a doctor or hospital staff [13].

ShravanKumar Uppin and M A Anusuya (2014), "*Expert System Design to Predict Heart and Diabetes Diseases*" designs an expert system that predicts the heart disease and diabetes disease. The author uses reduced number of attributes and then uses data mining technique in which he applied C4.5 classification algorithm so that there is more accuracy and less run time. The author also applies decision tree algorithm for the prediction of heart disease and diabetes and foretells that whether disease is present or not. According to the author the existing method takes 0.05 sec whereas the proposed method took around 0.025 sec and the accuracy is also increased from 84.35% to 85.96% [5].

4. METHODOLOGY AND PROPOSED WORK

The methodology provides an understanding of implementation of novel algorithm. The methodology includes the following steps:

1. Collection of data: Dataset were mainly collected from UCI repository and from various hospitals of Hemodialysis disease. Patient's data were collected which contains 8 attributes.

2. Preprocessing and Filtering: In preprocessing step, it selects an attribute for selecting a subset of attribute so that it can provide good predicted capability. It also contains the conversion of data types like numeric to nominal or vice versa. It handles all the missing values and remove them. If an attribute contains more than 5% missing values, then the records should not be deleted and it is advised to put the values where the data is missing using some suitable methods and helps in feature selection and class label

3. Classification: Classification is a technique for machine learning by which it is used to predict the grouping membership of different data instances. It will perform the task by which it will generalize the well-known structure so as to apply it on new data. Here Random forest classifier has been used for quality measurement of dataset will be consider on the basis of percentage of correctly classified instances. For validation phase we use 10-fold cross validation method. Random forest classifier helps in identifying the characteristics of patient with Diabetes diseases.

The proposed work includes:

1. To apply pre-processing and filtering on the Diabetes disease raw data and convert into formatted dataset.

2. To apply Weighted Hierarchical clustering and Balance Random forest on Diabetes patient's dataset,

- Using Weighted Hierarchical algorithm by updating the Euclidean distance formula for clustering on the formatted patient's data.
- Balanced Random forest on clustered data for the classification from the clusters of patient's data and analyze the predictions.

3. To compare and analyze the results of proposed technique with the existing techniques on the basis of following parameters:

a) Accuracy Rate(AR)
b) Precision
c) Recall
d) F-Measure

5. CONCLUSION

In this work, we have discussed about previous data processing techniques to retrieve information and current development in the research of medical sciences. Further we have elaborated terminologies and techniques of learning in data mining and data warehousing. Literature survey is discussed to study about previous researches in this field of technology. This work automatically clears that current information gathering by using various data mining techniques are far better than the manual system.

REFERENCES

[1] Emrana Kabir Hashi, Md. Shahid Uz Zaman and Md. Rokibul Hasan, "An Expert Clinical Decision Support System to Predict Disease Using Classification Techniques", International Conference on Electrical, Computer and Communication Engineering (ECCE), pp. 396-400, February 16-18, 2017.

[2] P. K. Anooj, "Clinical decision support system: Risk level prediction of
heart disease using weighted fuzzy rules," J. of King Saud Uni. Comput. and Inform. Sci., ELSEVIER, Vol. 24, pp. 27-40, 2012.

[3] Purushottam, K.Saxena and R. Sharma, "Efficient Heart Diseasem Prediction System," Proced. Comput. Sci., ELSEVIER, Vol. 85, pp. 962
– 969, 2016.

[4] Parthiban Latha and Subramanian R. (2008) , "Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm" International Journal of Biological and Life Sciences 3:3 2008, Pondicherry.

[5] Uppin ShravanKumar and M A Anusuya (2014), "Expert System design to predict Heart and Diabetes Diseases", International Journal of Scientific Engineering and Technology Vol. 03,Mysore, India.

[6] Fayyad, U, "Data Mining and Knowledge Discovery in Databases: Implications from scientific databases", Proc. of the 9th Int. on the Scientific and Statistical Database Management, Olympia, Washington, USA, 2-11, 1997.

[7] Palaniappan Sellappan and Awang Rafiah (2008), "Intelligent Heart Disease Prediction System Using Data Mining Techniques ", International Journal of Computer Science and Network Security, VOL. 8 No. 8, August 2008, Selangor, Malaysia.

[8] Mac Dougall Candice, Percival Jennifer and Mc Gregor Carolyu (2009), "Integrating Health Information Technology into Clinical Guidelines", Annual International Conference of the IEEE, EMBS Minneapolis, Minnesota, USA, September 2-6, 2009.

[9] M Nirmala Devi , Balamurugan.S Appavu alias, U.V Swathi (2013), "An amalgam KNN to predict Diabetes Mellitus", 2013 IEEE International Conference on Emerging Trends in Computing, Communication and Nanotechnology, Madurai, Tamil Nadu, India.

[10] Shouman Mai, Tumer Tim, Stocker Rob (2012), "Using Data Mining Techniques in Heart Disease Diagnosis and Treatment", International Conference on Electronics, Communications and Computers 2012, IEEE, Northcott Drive, Canberra.

[11] Srinivas K, Kavihta Rani B. and Dr. Govrdhan A. (2010), "Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks", International Journal on Computer Science and engineering Vol. 02,No. 02, 2010 ,250-255, Jagtial, Karimnagar.

[12] Sundar V Bata and Tevi T, Saravanan N (2012), "Development of a Data Clustering Algorithm for Predicting Heart", International Journal of Computer Applications(0975-888) Volume 48- No. 7, June 2012, Coimbatore, India.

[13] Thangarasu Gunasekar and Assoc. Prof. Dr. Dominic P.D.D. (2014), "Prediction of Hidden Knowledge from Clinical Database using Data mining Techniques", 2014 IEEE 978-1-4799-0059-6, Tronoh Perak, Malaysia.