IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 12, December 2016

A Survey on Data Mining Approaches to Diabetes Disease Diagnosis and Prognosis

B. Senthil Kumar [1], Dr. R. Gunavathi[2]

Assistant Professor, Department of Computer Science, Sree Narayana Guru College, K. G. Chavadi, Coimbatore, Tamilnadu, India[1]

Associate Professor, Department of MCA, Sree Saraswathi Thyagaraja College, Pollachi, Tamil Nadu, India[2]

Abstract: Data Mining plays an important role in the field of healthcare, because disease diagnosis and analysis have huge size of data. These circumstances create huge number of data handling issues, and that to be handled effectively. The health dataset's are uncertain and dynamic in nature and it is very tedious to maintain and to manipulate. To overcome the above issues, several studies introduced numerous Machine learning approaches for various disease diagnosis and prognosis. This paper a different data mining and machine learning techniques used in diabetes are analyzed and compared. The task of disease diagnosis and prognosis is a part of classification and prediction. The recent and popular data mining techniques used in clinical data includes Bayesian, Random forest algorithms, Artificial Neural network, SVM and Decision Tree etc. This paper gives the problems and findings about those techniques with various factors.

Keywords: Data Mining, Clinical data, Diabetes Mellitus, classification, Regression.

I. INTRODUCTION

Diabetes mellitus is simply referred as diabetes which is a chronic metabolic disorder has now become widespread and extensive grown in India. It is caused by insufficient secretion or impaired insulin action. Diabetes mellitus consist of different group of disorders characterized by hyperglycemia, which is an abnormality in high blood glucose level and altered metabolism of carbohydrates, lipids and proteins [1]., characterized by hyperglycaemia, insufficient insulin and insulin resistance or both, leading to altered metabolism of carbohydrates, protein, lipids and an increased risk of vascular complications, defect in reactive oxygen species scavenging enzymes and high oxidative stress induced damage to pancreatic beta cells. Various types of reported diabetes mellitus can be classified into two categories named as type 1 diabetes and Type2 diabetes. The Type 1 diabetes is insulin dependent diabetes mellitus (IDDM), where the body does not produce any insulin and it mostly diagnosed in children and young adults [2]. There are several unique symptoms can help to diagnose Type 1 diabetes such as increased thirst, frequent urination, hunger, fatigue and blurred vision. The Type 1 diabetes Treatment aims at maintaining normal blood sugar levels through regular monitoring, insulin therapy, diet and exercise.

Type 2 is noninsulin dependent diabetes mellitus (NIDDM), where the body does not produce enough insulin. This type of diabetes can be diagnosed in elderly people. The type 2 diabetes can handle with oral hypoglycemic agents (sulphonylureas, biguanides etc,). As the disease progresses, the symptoms become more severe and potentially dangerous. Diabetes mellitus lead to various complications like visual impairment, stroke, cardio vascular disease, leg amputation and renal failure if diagnosis is not done in the right time. Analyzing and diagnosing the diabetes mellitus using data mining approach become very popular and essential due to high dimensional nature.

Data mining is the process of extracting valuable knowledge from huge dataset and it has played an important role in health care domain. Data mining techniques would be a valuable asset for diabetes researchers because it can expose hidden knowledge from a huge amount of diabetes related data [3]. Data mining is the method of retrieving useful information and patterns from massive dataset. Especially in health care the dataset size are huge and dynamic in nature and hard to predict based on statistics. In the context of clinical field, many data mining techniques are proposed. The most popular and well used data mining techniques in the field of clinical are: statistics based analysis, prediction and classification. With the enormous amount of health care data collected from various sources, it is gradually more important that the technique is necessary to be developed with more powerful features for analysis, interpretation and decision process. This process refers to the nontrivial extraction of unknown/hidden or new and potentially useful information from data in databases.

The fig 1.0 shows the basic steps involved in the knowledge mining process with clinical data. Clinical Data Sources**:** the clinical data sources are collected from different sources, such as sensors, web repositories and manual synthetic datasets.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 12, December 2016

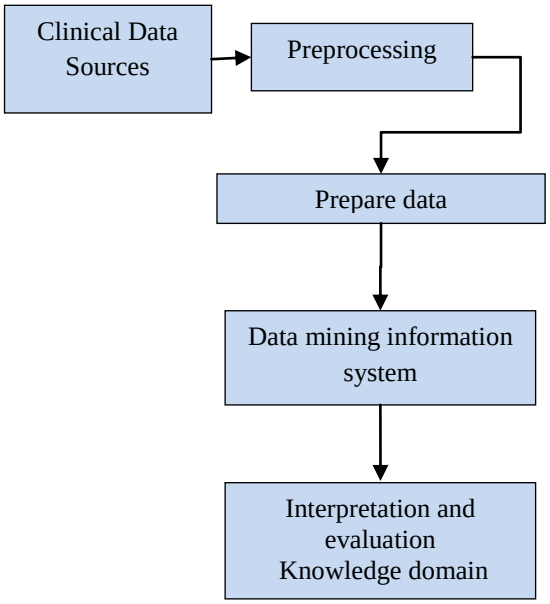

Fig 1.0. Steps in the Evolution of Clinical Data Mining

After the successful collection, the system performs the preprocessing step, which includes the data cleaning, data integration and so on. Data cleaning is also called as data cleansing process, this finds and eliminates the noisy, irrelevant and missing details form the collected data. Data integration is the next stage of cleansing; this collects data from multiple different data sources are combined in a common source format. Clinical Data mining is the extraction of unknown and hidden predictive information from hefty and dynamic databases. It is considered as a powerful and new technology with great potential to focus on the most important information in the clinical data sources shown in fig 1.0. This paper surveys about the various techniques and tools available in data mining to manage clinical activities. Data mining can significantly help diabetes research and ultimately improve the quality of health care for diabetes patients. In this paper, we surveyed different latest techniques and tools used in data mining to handle diabetes data. Finally we provide the summary and findings related to the work.

II. LITERATURE STUDY

Data mining has been participated an essential role in the intelligent healthcare systems which is stated in paper [4] [5][6]. The associations of disease and the real causes of the disease and the effects of symptoms that are impulsively seen in patients can be assessed by the users via data mining techniques. In the application of health care systems, Massive datasets can be applied as the input data to the system to find the association between features. The effects of associations have not been assessed adequately in the literature. This have been explored the relationships of hidden knowledge placed among the large medical databases and it has been searched relevant features by means of finding frequent items using candidate generation.

A. Heart disease:

Learning of the risk factors associated with diabetes helps health care professionals to identify patients at high risk of having diabetes disease. Statistical analysis and data mining techniques [7] assists healthcare professionals in the diagnosis of heart oriented diseases. Such analysis has identified the disease of the heart and blood vessels, using statistical values, and this comprises cerebrovascular disease known as stroke, coronary heart disease also known as heart diseases, raised blood pressure [hypertension], heart failure, rheumatic heart disease, peripheral artery disease and congenital heart disease.

In [8] authors presented an effective approach for the prediction of heart disease risk levels from the heart disease dataset with clustering technique. Initially the heart disease dataset is clustered using the K-means clustering algorithm and it will extract the features and data relevant to heart disease from the dataset. It allows the dataset to be portioned into k segments. The proposed approach mines the frequent patterns subsequently from the extracted data related to heart disease. The authors used MAFIA a maximal frequent Item set algorithm, which is a machine learning (ML) algorithms trained with selected significant prototype [9]. It basically predicts the heart disease and its risk factor of it. Additionally some practice from [10] resolves the prediction accuracy oriented problems. The approach utilizes the ID3 algorithm [11] for training process and applied in the vast amount of disease dataset. The consequences of the study showed that the designed prediction system is capable of predicting the heart disease effectively. But the prediction of diabetes is slightly different from the above. A study on the prediction of heart disease risk levels from the heart disease database with the use of bayes algorithms [12] carried out in the literature. This exploits the basic data mining classification techniques with 11 important features and multiple instances; the approach is effective due to the bagging method, which is an iterative process. From the results of [13] bagging technique is accurate and capable than the J48 and Bayesian classification algorithms for heart disease prediction.

B. Diabetes:

The predictive models were used in different clinical datasets, and it is also used in the diabetes data's, in such techniques, the risk scores are calculated to estimate the risk of diabetes, so there is a need of diabetes index process. The need of diabetes index has been recognized in [14], this conducted a survey regarding the diabetes risk factors. They found that most indexes were additive in nature and none of the surveyed indices have taken interactions among the risk factors into account.

In Paper [15] used association rule mining to systematically explore associations of different features and attribute associated with disease. The generated association rules do not establish a diabetes index since the study does not designate a particular outcome of interest and they do not assess or predict the risk of diabetes in patients dataset, but they discovered some significant

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 12, December 2016

associations between diagnosis codes. In specific, the author in [16] used FP-Growth and Apriori algorithms for diabetic prognosis. But the rules for disease forecast suffer from huge iterations and size. Some authors used split and merge algorithms with quantitative association rules for diabetic and its co-morbid condition prognosis. But only few researches succeeded in this research area.

Random forest mechanisms are used in the clinical assessment. It is a tree based algorithm and it is effective for huge datasets. The algorithm proposed in [17] is a combination of tree predictors so each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them. Random forests are an effective tool in prediction and classification techniques. And the main advantage of this algorithm is it reduces the over fitting problem. Injecting the right kind of randomness makes them accurate classifiers and regressions. In addition, the framework in terms of strength of the individual predictors and their correlations gives insight into the ability of the random forest to predict.. For a while, the conventional thinking was that forests could not compete with arcing type algorithms in terms of accuracy. The consequences scatter this certainty but lead to several problems. So the changes with the boosting and arcing algorithms have the ability to reduce bias as well as variance. Using a random selection of features to segment, each node yields error rates that compare favorably to Adaboost [18] but it more robust with respect to noise.

In the diabetes diagnosis process, Machine learning approaches are used. There are tremendous researches on health have been developed with the help of data mining techniques, we gave a deep analysis about those techniques and problems in it.

Chang-Shing Lee and Mei-Hui Wang proposed A Fuzzy Expert System for Diabetes Decision Support Application in [19] which is a five-layer fuzzy ontology that includes different fuzzy layers to describe the knowledge with uncertainty. The layers such as a fuzzy knowledge layer, fuzzy group relation layer, fuzzy group domain layer, fuzzy personal relation layer, and fuzzy personal domain layer. The author developed semantic decision making process in diabetic disease diagnosis. Even though the techniques are effective it has certain limitations such as the application was tested with a single dataset, the adaptation of the technique should be evaluated. The fuzzification approach is only applied in the fuzzy expert system is still more important rather than the ontology model. The approach suffers from the accuracy in disease diagnosis.

SonuKumari and Archana Singh proposed [20] an intelligent and effective methodology for the automated detection of Diabetes Mellitus using Neural Network. There are numerous methods have been proposed to diagnose diabetes mellitus, the given algorithm consists of three parts which are Network Designing, Training, and back propagation error handling.

Fayssal Beloufa and Chikh proposed modified Artificial Bee Colony (ABC) [21] for diabetes disease diagnosis. In this work, the authors used a blended crossover operator (BLX- α) of genetic algorithm. This enhanced the diversity of ABC without compromising with the solution quality. The modified ABC has been used as an evolutionary algorithm to create an optimal fuzzy classifier. The study tunes the optimal rules and membership functions with high accuracy and reliability. And a fuzzy classifier built with the modified ABC technique. This classifier can be used for the effective decision support system with few active fuzzy rules. This has provided a successful diagnosis tool for diabetes data. The author compared the MABC performance with the existing ABC and proved the efficiency. However, the technique needs more training samples to achieve the accuracy and effectiveness.

SVM (Support Vector Machine) is a supervised learning method for data classification. It is a classification and regression prediction tool which maximizes the prediction accuracy. And the main advantage of using this is it avoids the over-fitting problem. Based on the SVM clinical data have been classified and diagnosed effectively.

Wenxin Zhu and Ping Zhong [22] investigated the importance and incorporation of hidden information in one-class SVM, which is the extension of SVM technique. In this paper, the authors utilized the advantages provided by the SVM+ and focus on the one-class classification problem. Because the training samples are rather limited and insufficient, so in the system the hidden data's are identified from the hidden data. Authors prove the classification efficiency by embedding the additional information into the corresponding optimization problem and they derived an m-SVM style SVM+ framework for one-class classification.

The new one-class SVM model yields better generalization performance. But, there are some downsides of the techniques, which is more complex and tough than one-class m-SVM. It needs more tuning parameters and deep study is necessary. And as like prior methods, it cannot be performed using statistical analysis. The group attribution selection process needs much attention.

Giveki, Davar, et al introduced a novel automatic approach to diagnose Diabetes disease based on Feature Weighted Support Vector Machines (FW-SVMs) and Modified Cuckoo Search (MCS) [23]. The model proposed in this paper consists of three stages: Initially, PCA is applied to select an optimal subset of features out of set of all the features. Then the Mutual Information is deployed to construct the FW-SVM by weighting different features based on their degree of weight. At last, the parameter selection plays a vital role in classification accuracy of SVMs; the MCS is applied to select the best parameter values. The method by the authors achieved 93.58% accuracy on UCI dataset. This method is suited for the limited number of features, where the method only considered four features from diabetes dataset.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 12, December 2016

C. Summary:

Based on the above different techniques and algorithms several applications are created. Some applications automatically assign the appropriate meal and 'moment of measurement' to incomplete glycaemia data. The above mentioned machine learning techniques were studied, and every algorithm is designed with the intension of improving accuracy. Several fusion based approaches were used to increase the classification efficiency in different clinical datasets. However, the algorithms and techniques could cover a single factor of improvement than all. The followings are the summary and finding from the literature.

Table 1.0 accuracy comparison table

Models	Accuracy (%)
Decision Tree (DT)	85.090
Artificial Neural Network (ANN)	84.532
Naïve Bayes (NB)	81.010
Bagging	85.333
Boosting	84.098
Random Forest (RF)	85.558
SVM	87.6
SVM+	94.2
fuzzy	93.8
Artificial Bee Colony	91.9
PCA	90.6
Genetic	96.5
FW-SVM	93.5
MABC (modified ABC)	96.5

The table 1.0 shows the percentage of the accuracy achieved in every algorithm, this shows the genetic and modified ABC algorithms have higher accuracy value than others.

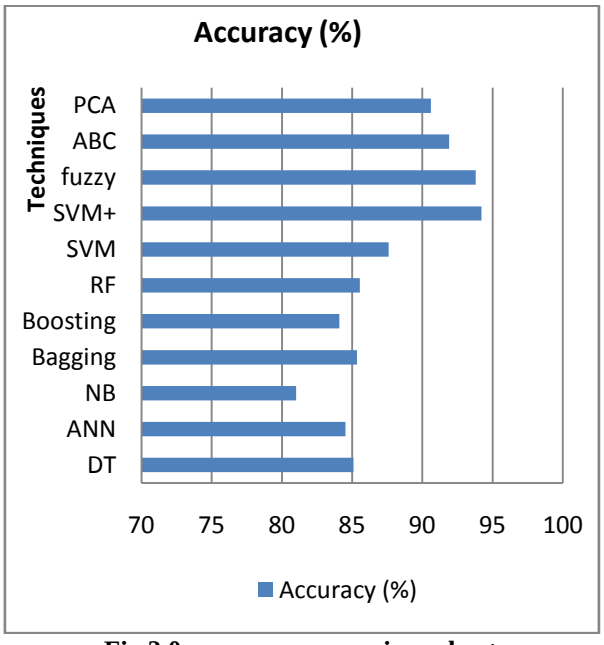

Fig 2.0 accuracy comparison chart

The existing techniques and tools for diabetic disease diagnosis and prognosis are based on the supervised techniques, which need more and accurate training samples. The training samples should be collected and included with the test data types, which means the class distribution should be considered. As said, the classification accuracy is based on the accurate training samples. However, some techniques proposed to find hidden factors in the dataset that is not fully studied. So the effective classifier needs complete training samples. Few studies used semi-supervised classification and un-supervised learning process to reduce the issues of the data collection. This can be reducing the issues of supervised learning problem, but the accuracy and class imbalance problem arises. The major task of classifier is it should reduce the iterations and over-fitting problems. The tree based algorithm may suffer by such issues. In this study we compared several algorithms and techniques, which handles clinical datasets. While comparing those techniques there are several problems found and recognized with set of negative features.

The first problem in effective classification is Class imbalance problem, which needs complete and effective data collection. The class imbalance need appropriate class distribution process, and thus increases the classification accuracy. The training samples and test samples should be equally distributed, however the distribution is not always be perfect. So the proposed system should consider the class imbalance problem in the uncertain data environment.

Another problem in clinical data mining is the Dimensionality problem, which can be effectively overcome by advanced feature extraction techniques.

In general, clinical datasets are dynamic and uncertain in nature, for example the clinical datasets are unique in values and the trend changes the training samples every time. The Diagnosis and prognosis of all type of diabetes with uncertain dataset should be handled effectively for accurate and effective results.

When comparing the above algorithms, certain algorithms are supervised and some of them are semi-supervised nature. So the active and self learning framework should be performed to reduce the complications in data collection. Along with the above problems, there are some general issues such as iteration, time reduction, improving accuracy and reducing false alarms. So, in this survey we summarized the problems of the existing research and planned to cover almost all the problems in the future with effective algorithms.

III. CONCLUSION

The clinical activity analysis plays an important role in current trend. Detection and analysis of clinical activity is the most important issue in real time scenario, because the lack of training samples and sufficient data's make these processes much complicated. This clinical data analysis can be performed by effective data mining techniques and approaches. There are several different methods to

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 12, December 2016

diagnosis and prognosis diabetes mellitus. This survey presents a various techniques of the data mining approach to solve the diabetes disease diagnosis problem. From the analysis we discover several problems and finds in clinical datasets handling process.

REFERENCES

[1] Assal, J. P., and L. Groop. "Definition, diagnosis and classification of diabetes mellitus and its complications." World Health Organization (1999): 1-65.
[2] National Diabetes Data Group. "Classification and diagnosis of diabetes mellitus and other categories of glucose intolerance." Diabetes 28.12 (1979): 1039-1057.
[3] Koh, Hian Chye, and Gerald Tan. "Data mining applications in healthcare."Journal of healthcare information management 19.2 (2011): 65.
[4] SA, S. "Intelligent heart disease prediction system using data mining techniques." International Journal of Healthcare & Biomedical Research 1 (2013): 94-101.
[5] Tomar, Divya, and Sonali Agarwal. "A survey on Data Mining approaches for Healthcare." International Journal of Bio-Science and Bio-Technology 5.5 (2013): 241-266.
[6] Milovic, Boris, and Milan Milovic. "Prediction and decision making in health care using data mining." Kuwait Chapter of the Arabian Journal of Business and Management Review 1.12 (2012): 126.
[7] Yoo, Illhoi, et al. "Data mining in healthcare and biomedicine: a survey of the literature." Journal of medical systems 36.4 (2012): 2431-2448.
[8] Chaurasia, Vikas, and Saurabh Pal. "Early prediction of heart diseases using data mining techniques." Carib. j. SciTech 1 (2013): 208-217.
[9] Methaila, Aditya, et al. "Early Heart Disease Prediction Using Data Mining Techniques." Computer Science & Information Technology (CS & IT),© CS & IT-CSCP (2014).
[10] Kavousi, Maryam, et al. "Evaluation of newer risk markers for coronary heart disease risk classification: a cohort study." Annals of Internal Medicine 156.6 (2012): 438-444.
[11] Ranganatha, S., et al. "Medical data mining and analysis for heart disease dataset using classification techniques." Research & Technology in the Coming Decades (CRT 2013), National Conference on Challenges in. IET, 2013.
[12] Songthung, Phattharat, and Kunwadee Sripanidkulchai. "Improving type 2 diabetes mellitus risk prediction using classification." Computer Science and Software Engineering (JCSSE), 2016 13th International Joint Conference on. IEEE, 2016.
[13] Perveen, Sajida, et al. "Performance Analysis of Data Mining Classification Techniques to Predict Diabetes." Procedia Computer Science 82 (2016): 115-121.
[14] Juliyet, L. Cynthiya, and Mr K. Mohamed Amanullah. "The Surveillance on Diabetes Diagnosis Using Data Mining Techniques."
[15] Patil, B. M., R. C. Joshi, and Durga Toshniwal. "Association rule for classification of type-2 diabetic patients." Machine Learning and Computing (ICMLC), 2010 Second International Conference on. IEEE, 2010.
[16] Sankaranarayanan, S. "Diabetic Prognosis through Data Mining Methods and Techniques." Intelligent Computing Applications (ICICA), 2014 International Conference on. IEEE, 2014.
[17] Butwall, Mani, and Shraddha Kumar. "A Data Mining Approach for the Diagnosis of Diabetes Mellitus using Random Forest Classifier." International Journal of Computer Applications 120.8 (2015).
[18] Vijayan, V. Veena, and C. Anjali. "Prediction and diagnosis of diabetes mellitus—A machine learning approach." Intelligent Computational Systems (RAICS), 2015 IEEE Recent Advances in. IEEE, 2015.
[19] Lee, Chang-Shing, and Mei-Hui Wang. "A fuzzy expert system for diabetes decision support application." IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 41.1 (2011): 139-153.
[20] Kumari, Sonu, and Archana Singh. "A data mining approach for the diagnosis of diabetes mellitus." Intelligent Systems and Control (ISCO), 2013 7th International Conference on. IEEE, 2013.
[21] Beloufa, Fayssal, and M. A. Chikh. "Design of fuzzy classifier for diabetes disease using Modified Artificial Bee Colony algorithm." Computer methods and programs in biomedicine 112.1 (2013): 92-103.
[22] Zhu, Wenxin, and Ping Zhong. "A new one-class SVM based on hidden information." Knowledge-Based Systems 60 (2014): 35-43.
[23] Giveki, Davar, et al. "Automatic detection of diabetes diagnosis using feature weighted support vector machines based on mutual information and modified cuckoo search." arXiv preprint arXiv:1201.2173 (2012).