Performance Comparison of Data Mining Techniques

T.Praveena
Ph.D Research Scholar, Department of Computer Science,
Nandha Arts and Science College, Erode, Tamil Nadu, India.
Email: praveenasurya14@gmail.com
Dr.S.Prasath
Research Supervisor & Assistant Professor, Department of Computer Science,
Nandha Arts and Science College, Erode, Tamil Nadu, India.
Email:softprasaths@gmail.com

Abstract-

In recent years the healthcare industry has generated large amounts of data. The value based treatment in hospitals and digitization of world likes to have the computerized data rather than hard copy form. Diabetes is one of the common and rapidly increasing diseases in the world. It is a major health problem in most of the countries. Diabetes is a condition in which your body is unable to produce the required amount of insulin needed to regulate the amount of sugar in the body. This leads to various diseases including heart disease, kidney disease, blindness, nerve damage and blood vessels damage. Hence, there is a requirement of a model that can be developed easily providing reliable, faster and cost effective methods to provide information of the probability of a patient to have diabetes. In this paper dealt with an attempt is made a comparative study on diabetes using classification and clustering.

Keywords- Data Mining, Big Data, Health Care, Diabetes, Clustering, Classification.

1. INTRODUCTION

Digital data have a vital role in the computerized world. Now we are living in data world. Everywhere we are seeing only data. The storage of data occupies more space, since the data usage is increasing every day. The important thing is 'data storing' and 'data processing'. The data in computing are represented as structured format in the olden days. It is stored as tabular format. Now a days so much data is not natively in structured format. The problems start right away during data acquisition, when the massive data requires us to make decisions, about what data to retain and what to remove and how to store what we save reliably with the correct metadata. Dealing with these Big Data is a highly challenging issue for the data analysts.

Data Mining

Data mining is an extraction of hidden predictive information from large database. Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. These tools can include statistical models, mathematical algorithms, and machine learning methods.

Consequently, data mining consists of more than collecting and managing data; it also includes analysis and prediction [1]. With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important to develop powerful means for analysis and perhaps interpretation of such data and for the extraction of interesting knowledge that could help in decision-making.

Limitation of Data Mining:

i. Data Mining are primarily data or personnel-related rather than technology-related.
ii. Data Mining difficult to handle the large amount of data.
iii. It is difficult to handle the dynamic data.
iv. Big data can handle enormous amount data and also it capable of handle the dynamic data .So that Big data is better than data mining to handle such kind of data and easily solve the problems.

Big Data

Big data is a buzzword that is used to describe the large amount of data either structured or unstructured data format. Exactly, if the data which is beyond to the storage capacity & which is beyond to the processing power, that data we are calling 'BIG DATA'. Big data is so large and is difficult to process using the old database and software techniques. The health care data includes Electronic Health Reports (EHR) of patients data, clinical reports, doctor's prescription, diagnostic reports, medical images, pharmacy information, health insurance related data, data from Social Medias and medicinal journals [13]. All these information collectively forms Big Data in health care. By employing the analysis of big data will produce the predicted results for understanding the trends to improve the health care and life time expectancy, proper treatment at early stages at low cost.

The analytics associated with big data is described by four characteristics: volume, velocity, variety and veracity [8]. The accumulation of health-related data continuously, resulting in an incredible volume of data; Velocity is accessing those data in real-time at a rapid speed; Variety includes diabetic glucose measurements, blood pressure readings, and various EHRs; Whereas veracity assumes the simultaneous scaling up in performance of the architectures and platforms, algorithms and tools to match the need of big data Having data larger it needs different Approaches, Techniques, Architectures, and Tools.

Healthcare

Big data and its related technologies have improved healthcare enormously from understanding the Origins of diseases, Better diagnoses, Helping patients to monitor their own conditions. Healthcare organizations can improve their quality of service by analyzing the effectiveness of a treatment and also the efficiency of the healthcare delivery process. Since information is in the digital form, healthcare providers can use some available tools and technologies to analyze that information and generate valuable insights. Full view for every patient is created by electronic health records, scanned documents, medical images, notes from physicians, information about environment.

Diabetes

Diabetes is a disease that arises when the insulin production in the body is insufficient or the body is unable to use the produced insulin in an appropriate manner, as an outcome, this leads to high blood glucose. The body cells break down the food into glucose and this glucose needs to be transported to all the cells of the body. The insulin is the hormone that directs the glucose that is produced by breaking down the food into the body cells. Any modification in the creation of insulin leads to a rise in the blood sugar levels and this can lead to harm to the tissues and failure of the organs. Generally a person is considered to be suffering from diabetes, when blood sugar levels are above normal (4.4 to 6.1 mmol/L). There are three main types of diabetes, viz. Type 1, Type 2 and Gestational.

Types of Diabetes

The three main types of diabetes are described below:

1. Type 1 – There are only around 10% of diabetes patients have this form of diabetes. But, there has been a rise in the number of cases of this type in the world. The disease manifest as an autoimmune disease happening at a very young age of below twenty years. It called as juvenile-onset diabetes. In this type of diabetes, the pancreatic cells that produce insulin have been demolished by the defence system of the body. Injections of insulin along with regular blood tests and dietary limits have to be followed by patients suffering from Type 1 diabetes.
2. Type 2 – Almost 90% of the diabetes patients are affected by this type 2 diabetes. It called as, the adult-onset diabetes or the non-insulin dependent diabetes. In this situation the various organs of the body become insulin resistant, and this raises the demand for insulin. At this point, pancreas doesn't make the required amount of insulin [4].To keep this type of diabetes away, the patients have to follow a strict diet, routine exercises and keep track of the blood glucose. Obesity, overweight, physically inactive can lead to type 2 diabetes. Also with ageing, the risk of emerging diabetes is measured to be more. Most of the Type 2 diabetes patients are in border line diabetes or the Pre-Diabetes, a situation where the blood glucose levels are upper than normal but not as high as a diabetic patient.
3. Gestational diabetes – It tends to occur in pregnant women due to the high sugar levels as the pancreases don't produce enough amount of insulin. Taking no treatment can lead to difficulties during child birth [4]. Monitoring the diet and taking insulin can control this form of diabetes. All these types of diabetes are serious and It needs treatment.

Symptoms, Diagnosis and Treatment

The common symptoms of a person suffering from diabetes are:

a. Polyuria (frequent urination)
b. Polyphagia (excessive hunger)
c. Polydipsia (excessive thirst)
d. Weight gain or strange weight loss.
e. Healing of wounds is not quick, blurred vision, fatigue, itchy skin, etc

Urine test and blood tests are conducted to detect diabetes by checking for excess body glucose. The commonly conducted tests for determining whether a person has diabetes or not are

i. A1C Test
ii. Fasting Plasma Glucose (FPG) Test
iii. Oral Glucose Tolerance Test (OGTT).

Though both Type 1 and Type 2 diabetes cannot be cured they can be controlled and treated by special diets, regular exercise and insulin injections. The complications of the disease include neuropathy, foot amputations, glaucoma, cataracts, increased risk of kidney diseases and heart attack and stroke and many more.

2. RELATED WORKS

This section is to provide the general overview of related works in the field of diabetes. In particular those works are related and focused on the clustering and classification techniques and its various methods.

To conduct a systematic review ofthe applications of machine learning, data mining techniques and its tools in the field of diabetes [5].The predictive analysis algorithm in Hadoop/Map Reduce environment to predict the diabetes type's prevalent, complications associated with it and the type of treatment to be provided [12].Diabetes mellitus and its spread over the country particularly in Tamil Nadu, theresearch region. It already developing medical intelligence using clinical big data and proposes aforecasting and prediction system for Diabetes mellitus[10].To find the solutions to diagnose the disease by analyzing the patterns found in the data through classification analysis by employing. To propose a quicker and more efficienttechnique of diagnosing the disease, leading to timely treatment of the patients[1].

A hybrid algorithm of Modified-Particle Swarm Optimization and Least Squares Error.Support Vector Machine is proposed for the classification of type II DM patients. LS-SVM algorithm is used for classification by finding optimal hyper-plane which separates various classes.Since LS-SVM is so sensitive to the changes of its parameter values, Modified-PSO algorithm is used as an optimization technique for LS-SVM parameters [8].The diabetes data with a total instance of 768 and 9 attributes (8 for input and 1 for output) will be used to test and justify the differences between the classification methods. Subsequently, the classification technique that has the potential to significantly improve the common or conventional methods will be suggested for use in large scale data, bioinformatics or

other general applications [11].The modified J48 classifier is used to increase the accuracy rate of the data mining procedure for the diabetes [4].

To classify the risk of diabetes mellitus. Four well known classification models that are Decision Tree, Artificial Neural Networks, Logistic Regression and Naive Bayes were first examined.Then, findings suggest that the best performance of disease risk classification is Random Forest algorithm is best for the diabetes [7].The adaboost and bagging ensemble techniques using J48 (c4.5) decision tree as a base learner along with standalone data mining technique J48 to classify patients with diabetes mellitus using diabetes risk factors. The performance of adaboost ensemble method is better than bagging as well as standalone J48 decision tree [11].

A survey also highlights applications, challenges and future issues of Data Mining in healthcare. Recommendation regarding the suitable choice of available Data Mining technique [3].The Classification of diabetic's data set and the K-means algorithm to categorical domains. Before classify the data set preprocessing of data set is done to remove the noise in the data set. This algorithm is also used to improve the classification rate and cluster the data set using two attributes namely plasma and pregnancy attribute [6].

Three data mining algorithms, namely Self-Organizing Map (SOM), C4.5 and RandomForest, are applied on adult population data from Ministry of National Guard Health Affairs (MNGHA), Saudi Arabia to predict diabetic patients using 18 risk factors. RandomForest achieved the best performance compared to other data mining classifiers [13].To concentrate upon predictive analysis of diabetes diagnose using artificial neural network as a data mining technique. The Pima Indian diabetes database was obtained from UCI server and used for analysis [2].

In the literature review mainly focused on the analysis of diabetes. Clustering and classification techniques are used to identify the diabetes. This system provides an efficient way to cure and care the patients with better outcomes like affordability and availability.

3. Clustering

This pattern partitions the records in database into diverse gatherings. In the same gathering, the gatherings have the comparative properties and the distinctions ought to make as bigger as could be expected under the circumstances and in the same gathering, the distinctions ought to be as littler as would be prudent. There is no predefined class in this gathering it goes under the unsupervised learning. Techniques included in bunch examination are partioning systems, various leveled routines, thickness Based strategies, network based techniques, model-based routines, grouping high-dimensional information, requirement based bunching and Outlier investigation.

i. K-means Clustering
ii. Hierarchical clustering
iii.Density based clustering

4. Classification

Classification divides data samples into target classes. The classification technique predicts the target class for each data points. For example, patient can be classified as "high risk" or "low risk" patient on the basis of their disease pattern using data classification approach. It is a supervised learning approach having known class categories. Binary and multilevel are the two methods of classification. In binary classification, only two possible classes such as, "high" or "low" risk patient may be considered while the multiclass approach has more than two targets for example, "high", "medium" and "low" risk patient.Figure 1 shows the various classification techniques used to identify the diabetes in easier way.

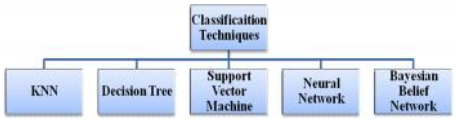

Fig. 3.1 Different Classification Techniques

The research work revealed that there is no single best algorithm which yields better result for dataset. Classification techniques are also used for predicting the treatment cost of healthcare services which is increases with rapid growth every year and is becoming a main concern for everyone [7]. Classification tree approach to predict the cost of healthcare [8] by using the dataset of 3 years collected from the insurance companies to perform the experiment. The first two year data was used to train the classifier and last one year data was used for comparing the predicted results of classifier. Finally there are various classification technique are used to identify the diabetes prediction.

5. Data Mining Tools

The data mining tools on which the integrated clustering-classification technique has been implemented.

5.1 WEKA tool

WEKA is Waikato Environment for Knowledge Analysis, data mining/machine learning tool developed by Department of Computer Science, University of Waikato, New Zealand. It is a collection of open source of many data mining and machine learning algorithms, including pre-processing on data, classification, regression, clustering, association rule extraction and feature selection which supports .arff (attribute relation file format) file format.

5.2 Tanagra

Tanagra was written an aid to education and research on data mining by Ricco Rakotomalala. The entire user operation of Tanagra is based on the stream diagram paradigm. Under the stream diagram paradigm, a user builds a graph specifying the data sources and operations on the data. Paths through the graph can describe the flow of data through manipulations and analysis. Tanagra simplifies this paradigm by restricting the graph to be a tree with only one parent to each node and the other one for data source of an each operation.

5.3 Orange

Orange is a component-based data mining and machine learning software suite, featuring a visual programming front-end for explorative data analysis, visualization, Python bindings and libraries for scripting. It includes set of components for data preprocessing, feature scoring and filtering, modeling, model evaluation and exploration techniques. It is implemented in C++ and Python.

6. Experiments and Results

The dataset "pima Indian Diabetes" are consider with the use of K-means, Hierarchical and Density based clustering technique and the different classification algorithms available on data mining tools. The Pima Indian diabetes data sets available on UCI machine learning repository.

The experiment is performed on the dataset results in Table 6.1 shows the accuracy measure of K-means clustering technique for different classifiers used. SVM provides the highest accuracy in the range of 66-71%, followed by Naïve Bayes with accuracy in the range of 64-66% and KNN with accuracy ranging between 62-68%.

Table 6.1 Accuracy for K-means clustering

Classifier	Weka	Tanagra	Orange
NB	66.18 %	64.87%	65.38%
KNN	68.71 %	62.11%	62.90%
SVM	71.13 %	66.45%	66.17%

The experiment is performed on the dataset results in Table 6.2 shows the accuracy measure of K-means clustering technique for different classifiers used. SVM provides the highest accuracy in the range of 64-68%, followed by Naïve Bayes with accuracy in the range of 61-67% and KNN with accuracy ranging between 60-66%.

Table 6.2 Accuracy for Hierarchical clustering

Classifier	Weka	Tanagra	Orange
NB	61.20 %	62.14%	67.14%
KNN	60.84 %	65.74%	66.28%
SVM	64.45 %	66.12%	68.17%

The experiment is performed on the dataset results in Table 6.3 shows the accuracy measure of K-means clustering technique for different classifiers used. SVM provides the highest accuracy in the range of 63-68%, followed by Naïve Bayes with accuracy in the range of 60-62% and KNN with accuracy ranging between 60-62%.

Table 6.3 Accuracy for Density based clustering

Classifier	Weka	Tanagra	Orange
NB	60.12 %	61.14%	62.78%
KNN	61.27 %	62.45%	63.54%
SVM	68.23 %	63.18%	64.27%

6. CONCLUSION

This paper dealt with the survey of automatic diagnosis of diabetes is an important real-world medical problem. Detection of diabetes in its early stages is the key for treatment. Clustering, classification techniques and various methods are used to model actual diagnosis of diabetes for local and systematic treatment, along with presenting related work in the field. The performance of the techniques was investigated for the diabetes diagnosis problem. This research work also shows the importance of the diabetes approach for the performance of classification and clustering techniques it shows better result for the patients. In future it is planned to gather the information from different locales over the world and make a more precise and general prescient model for diabetes conclusion. Future study will likewise focus on gathering information from a later time period and discover new potential prognostic elements to be incorporated. The work can be extended and improved for the automation of diabetes analysis.

REFERENCES

[1] Aiswarya Iyer, S.Jeyalatha and Ronak Sumbaly, “Diagnosis of Diabetes Using Classification Mining Techniques”, International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015.

[2] Divya Tomar and Sonali Agarwal , “A survey on Data Mining approaches for Healthcare” ,International Journal of Bio-Science and Bio-Technology Vol.5, No.5 (2013)

[3] Gaganjot Kaur,Amit Chhabra, “Improved J48 Classification Algorithm for the Prediction of Diabetes”, International Journal of Computer Applications (0975 – 8887) Volume 98, No.22, July 2014.

[4] Ioannis Kavakiotis, Olga Tsav, Athanasios Salifoglou, Nicos Maglaveras,Ioannis Vlahavas, Ioanna Chouvarda, “Machine Learning and Data Mining Methods in Diabetes Research”,Computational and Structural Biotechnology, Journal 15 (2017) 104–116.

[5] M. Kothainayaki, P. Thangaraj, “Clustering and Classifying Diabetic Data Sets Using K-Means Algorithm”, Journal of Applied Information Science, Volume 1 Issue 1 June 2013

[6] Nongyao Nai-aruna, Rungruttikarn Moungmai, “Comparison of Classifiers for the Risk of Diabetes Prediction”, 7th International Conference on Advances in Information Technology Procedia Computer Science 69 (2015) 132 – 142.

[7] Omar S Soliman, Eman AboElhamd, “Classification of Diabetes Mellitus using Modified Particle Swarm Optimization and Least Squares Support Vector Machine”, International Journal of Computer Trends and Technology (IJCTT) – volume 8 number 1– Feb 2014.

[8] R. Shantha Mary Joshitta L. Arockiam, "A Predictive Model to Forecast and Pre-TreatDiabetes Mellitus using Clinical Big Data in Cloud", International Journal Of Applied Engineering Research , January 2015.

[9] Rashedur M.Rahman, Farhana Afroz, "Comparison of Various Classification Techniques Using Different Data Mining Tools for Diabetes Diagnosis", Journal of Software Engineering and Applications, 2013, 6, 85-97.

[10] Sajida Perveen, Muhammad Shahbaz, Aziz Guergachib, Karim Keshavjee, "Performance Analysis of Data Mining Classification Techniques to Predict Diabetes", Symposium on Data Mining Applications, SDMA2016, 30 March 2016, Riyadh, Saudi Arabia.

[11] Saravana kumar N, Eswari T, Sampath & Lavanya , "Predictive Methodology for Diabetic Data Analysis in Big Data", 2nd International Symposium on Big Data and Cloud Computing (ISBCC'15).

[12] Tahani Daghistani, Riyad Alshammari "Diagnosis of Diabetes by Applying Data Mining Classification Techniques", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 7, 2016