Impact Factor Value: 4.029 ISSN: 2349-7084

IJCERT International Journal of Computer Engineering In Research Trends

Volume 4, Issue 12, December - 2017, pp. 556-561 www.ijcert.org

Comparative Performance Analysis of Different Data Mining Techniques and Tools Using in Diabetic Disease

[1]Sarangam Kodati, [2]Dr. R P. Singh

[1]Research Scholar, Department of Computer Science and Engineering, Sri Satya Sai University of Technology and Medical Science, Sehore, Bhopal, Madhya Pradesh, India

[2.]Professor, Department of Computer Science and Engineering, Sri Satya Sai University of Technology and Medical Science, Sehore, Bhopal, Madhya Pradesh, India

---

Abstract: - Data mining means to the process of collecting, searching through, and analyzing a significant amount of data in a database. The most essential and popular data mining methods are classification, association, clustering, prediction or sequential patterns. In health concern businesses, data mining plays a vital role in the early prediction of diseases toughness. This paper explores the early prediction diabetic diabetes using various data mining methods and data mining tools. The dataset has taken 768 instances from PIMA Indian Dataset by determining the accuracy of the data mining techniques in prediction. The analysis proves that Modified J48 Classifier provides the highest comparative durability accuracy than other techniques.

Keywords: Data mining Techniques, Data mining Tools, Diabetic disease, Performance Accuracy.

---

1. Introduction

The process of turning the low-level data into high-level knowledge. Hence, Knowledge discovery in databases refers to imitation on the nontrivial extraction over implicit, previously weird and potentially useful data from information of databases. While data mining and Knowledge discovery in databases are often dealt with hence equal words however of real data mining is an important step in the Knowledge discovery in databases process. Indicates data mining namely like a step between an iterative knowledge discovery processes.

- Data cleaning: also known as much data cleaning that is a phase in which noise data and irrelevant data are eliminated from the collection.
- Data integration: at this stage, more than one data sources, of heterogeneous, may be combined of a familiar source.
- Data selection: at it step, the data relevant to the analysis is decided regarding and retrieved from the data collection.
- Data transformation: also known as like data consolidation, it is a phase into which the selected data is transformed into forms excellent for the mining procedure.
- Data mining: it is the crucial step into which clever methods are utilized according to extract patterns Potentially useful.
- Pattern evaluation: this step, strictly interesting patterns representing knowledge are identified based on given measures.
- Knowledge representation is the final phase of which the discovered knowledge is visually represented by the user. In this quadrant noticing techniques are old in conformity with assist users to understand and interpret the data mining results.

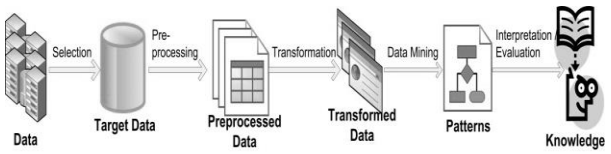

Fig 1: Knowledge discovery in databases (KDD) Process

Data mining is a key role of the intelligent health domain [1]. There are several software's, and tools have been used after diagnose and classify health information's based on the attributes. The large volume databases are included in this process as like input. This process resulted in data collection complication. The

followings are the basic information about diabetes then its basic reasons and symptoms.

Diabetes risk Prediction Model can support medical experts and practitioners in predicting risk status primarily based on the clinical data records. In the biomedical field, data mining and its methods play a fundamental role in prediction and analyzing special type regarding health issues. The healthcare industry gives large quantities of healthcare data and that need to stand mined to ascertain hidden data for valuable decision selection. Determining hidden patterns and relationships can also often dead tough and unreliable. The health report is categorized and predicted if they have the symptoms of Diabetes risk and the usage of risk factors because of disease [2]. It is indispensable in conformity with finding the best match algorithm that has greater accuracy, speed and memory utilization of prediction in the case of Diabetes.

Data mining methods have proved because of early prediction of disease with higher accuracy to save human life and reduce the treatment cost. This paper explores some Data mining techniques specific as much Navie Bayes, MLP, Bayesian Network, C4.5, Amalgam KNN, ANFIS, PLS-LDA, Homegenity-Based, ANN, Modified J48, etc. are analyzed according to predict the diabetes disease. Veena 2014 combined Amalgam KNN then ANFIS in conformity with improving the accuracy of prediction. In that, K-means and KNN are combined according to overcome the computational complexity respecting a large number of the dataset. And the training set is verified with fuzzy systems and neural networks in imitation of produce a better result. Sapna 2012 implemented genetic algorithm with data mining techniques after test the patients affected by way of access to diabetes based totally on the fitness value and the precision chromosome value.

Gaganjot Kaur 2014 proposed a new approach because predicting diabetes the usage of WEKA and MATLAB because generating J48 classifiers together with improved existing J48 algorithm. Murat Koklu 2013 formed a selection support system the use of data mining and artificial intelligence classification algorithms namely Multilayer Perceptron, Navie Bayes classification and J48 according to to diagnosing illness. To achieve excellent part of predicting the onset of diabetes, Manaswini Pradhan 2011 suggested or experimented ANN based classification mannequin and Genetic algorithm because of feature selection. Hence, this paper usually focused on Data mining methods and analyzed its accuracy including various tools.

2. Diabetic Disease

Diabetes is a disease that takes place when the insulin production in the body is inadequate, or the body is unable to utilize the produced insulin among a proper manner. As a result, this leads to high blood glucose. The body cells break down the food between glucose and this glucose needs in imitation of stay transported to every the cell of the body. The insulin is the hormone that directs the glucose so is produced using breaking down the food into the body cells. Any alternate in the production of insulin leads to increase between the blood sugar levels, and this can lead to damage by the tissues and failure on the organs. A person is considered by being suffering from diabetes when blood sugar levels above normal (4.4 to 6.1 mmol/L).

3. Diabetes is Categorized

There are three main three types of diabetes Type 1, Type 2 and Gestational.

- Type 1 Diabetes: It is a chronic condition amongst which the pancreas produces little or no insulin. This type of Diabetes results from the pancreas's failure according in imitation of produce enough insulin. This necessitates the individual in imitation over insert insulin or carries an insulin pump. This form was once earlier referred to conformity with as "insulin-dependent diabetes mellitus." The cause of type 1 diabetes is unknown[3].
- Type 2 Diabetes begins with insulin resistance, a situation within which cells fail according to respond according to insulin properly. As the disease progresses a lack of insulin might also develop. This form was earlier referred according to consequently "noninsulin-dependent diabetes mellitus" or "adult-onset diabetes." The primary cause is excessive body weight and now not enough exercise.

- Gestational diabetes is the 0.33 main form or takes place then pregnant female without previous data as regards diabetes develop high blood-sugar levels.

As a consequence regarding the human, our bodies malfunction according to generate insulin, and necessitates the individual between imitation of insert insulin and raise an insulin pump. This category was once previously indicated as much permanency "Insulin-Dependent Diabetes Mellitus." The second category about DM is recognized namely "Type II DM" therefore a consequence as regards insulin confrontation, a situation of any cells are ineffective of accordance with exploit insulin appropriately, occasionally merged collectively with an absolute insulin insufficiency. This category additionally called "Non-Insulin Dependent Diabetes Mellitus" yet "adult-onset diabetes." At last, "gestational diabetes" takes place when conceived women without an earlier.

4. Diabetes Symptoms, Diagnosis and Treatment

- The common symptoms of a person who has diabetes are:

• Polyuria (frequent urination)
• Polyphagia (excessive hunger)
• Polydipsia (excessive thirst)
• Weight gain or strange weight loss
• Healing of wounds is not quick, blurred vision, fatigue, itchy skin, etc. The 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:

• A1C Test
• Fasting Plasma Glucose (FPG) Test
• Oral Glucose Tolerance Test (OGTT).

Though both Type 1 and Type 2 diabetes can't be cured, they may be controlled and treated with the aid of 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 much more. The earlier diagnosis of diabetes, the risk of the complications can be dodged. Hence a faster technique of predicting the disease has been presented in this paper.

5. Application of Data Mining Techniques in Diabetes

Medical data can be trained using data mining methods in imitation of foretelling diabetes. For this, the dataset has to be preprocessed to remove noisy and fill the missing values. Pima Indian Diabetes Dataset used to be taken by evaluating data mining Classification. The dataset comprises 9 attributes and 768 instances. The following table1 shows the description of the characteristics. Data mining methods perform keep applied by the effective factors such as BMI, DPF, age, and skin according to predict diabetes. Insulin and GTT measure is used for testing diabetes. Pregnancy and BP are also considered as testing factors. The above attributes may be categorized and cluster using various methods such as Navie Bayes, J48, PLS-LDA, Support Vector Machine, BLR, MLP, K-Nearest neighbors, Bayesian Network. With the help concerning the above attributes type 1, type 2 diabetes and gestational diabetes can be diagnosed. Obesity, age factor, and family history are the main cause of type 2 diabetes. The class volatile one indicates the diabetic test is positive and 0 indicates the test is negative. Tangara, WEKA and MATLAB tools help after functional data mining task together with whole machine learning algorithms. Data excavation supervised instruction algorithms are used in imitation of categorization task. DM method can predict the hidden patterns from the previous history. Classification is the usually used method within medical data mining. The predictive accuracy of the classifier is estimated. The application of data mining method can minimize the number of exams required for detecting disease.

Table 1: Attributes of Diabetes Dataset

S NO.	Attribute	Description
1	Plasma	Plasma glucose concentration a 2 hours in an oral glucose tolerance test
2	Pressure	Diastolic blood pressure(mmHg)
3	Skin	Triceps skin fold thickness(mm)
4	Insulin	2-Hour serum insulin (mu U/ml)
5	Pregnancy	Number of times pregnant
6	Mass	Body Mass Index(BMI)
7	Pedigree	Diabetes Pedigree function
8	Age	Age(in years)
9	Class	Class variable(0 or 1)

6. Various Data Mining Techniques Used to Predict Diabetes

The diabetic patients suffer from various diseases, and also that affects some parts of vile organs. If the treatments are not taken to control the disease, that leads the patient to death. Hence, effective measures have after being taken by predicting the disease at the earliest and control. In it paper, some data mining methods are analyzed in imitation of diagnosing diabetes mellitus with the best methods using a range of tools. As per the data are given in table 2, Gaganjot compared the accuracy and error rate concerning various data mining algorithms such as much Naive Bayes, MLP, Random forest, Random Tree and Modified J48.The result provides 99.87% accuracy in Modified J48 Classifier. Radha experimented C4.5, Support Vector Machine, K-Nearest Niebuhr, PNN or BLR classification methods to classify the patients with and besides diabetes. The resulting present shows to that amount C4.5 decision tree algorithm provides 86% concerning accuracy. Mohtaram Bayesian network technique predicts the diabetic patients. The Bayesian Network trained with the given data set and provided 90.4% accuracy in prediction. The genetic algorithm creates the best solution to foretell the illnesses including 80.5% concerning accuracy. The computing device learning algorithm Multilayer Perceptron fed with training dataset and trained in imitation of classifying the feature vectors. The result obtained

within MLP is 97.61%. This paper deals with different data mining methods concerning the overall performance concerning the system according to predict diabetes.

7. Different Types of Data mining Techniques and Data Mining Tools

Table 2: Different types of data mining techniques and data mining tools.

Author Name & Year	Data Mining Techniques	Data Mining Supporting Tools	Best DM Techniques	Performance Accuracy
Gaganjot Kaur and Amit Chhabra, 2014 [4]	Navie Bayes, MLP, Random Tree, REP tree, RAD, Random Forest, J48, Modified J48 Classifier	WEKA, MATLAB	Modified J48 Classifier	99.87%
P.Radha and Dr. B. Srinivasan, 2014 [5]	C4.5, SVM, k-NN,PNN,BLR	Tanagara	C4.5	86.5%
Mohtaram Mohammadi, Mitra Hosseini, Hamid Tabatabaee, 2014 [6]	Bayesian Network, Decision Tree	MATLAB	Bayesian Network	90.4%
Sudesh Rao and N. Arun Kumar, 2014 [7]	Genetic algorithm	Clementine	Genetic algorithm with fuzzy logic	80.5%
Veena Vijayan. V and Aswathy Ravi kumar, 2014 [8]	EM, KNN, K-means, amalgam KNN and ANFIS algorithm	Sharper Light	Amalgam KNN and ANFIS	80.%
Arwa Al-Rofiyee, Maram Al-Nowiser, Nasebih Al-Mufadi, 2013 [9]	MLP	WEKA	MLP	97.61%
K.R. Lakshmi and S. Prem kumar, 2013 [10]	C4.5, SVM, k-NN, PNN, BLR, MLR, PLS-DA, PLS-LDA, k-means & Apriori	Tanagara	PLS-LDA	76.78%
Murat Koklu and Yavuz Unal, 2013 [11]	Multilayer Perceptron, J48 and Navie Bayes Classifier	WEKA	Navie Bayes Classifier	76.3%
Rupa Bagdi, Prof. Pramod Patil, [12]	ID3 , C4.5 Decision Tree	ID3 , C4.5 Decision Tree	C4.5 Decision Tree	74%
Ashwin kumar.U.M and Dr.Ananda kumar.K.R, 2012 [13]	Decision Tree and Incremental learning	WEKA	C4.5	68%
S.Sapna , Dr. Tamilarasi and M. Pravin Kumar,[14]	Genetic Algorithm	MATLAB	Generic Genetic Algorithm	80%
Manaswini Pradhan and Dr. Ranjit Kumar Sahu, 2011 [15]	Artificial Neural Network, Genetic algorithm	Tanagara	Artificial Neural Network	73.43%
Muhammad Waqar Aslam and Asoke Kumar Nandi, 2010 [16]	Genetic Programming	GP Lab tool Box	Genetic Programming	78.5%
Huy Nguyen Anh Pham and Evangelos Triantaphyllou, 2008 [17]	Homogeneity-Based algorithm	Rapid Miner	Homogeneity-Based algorithm	80.1%

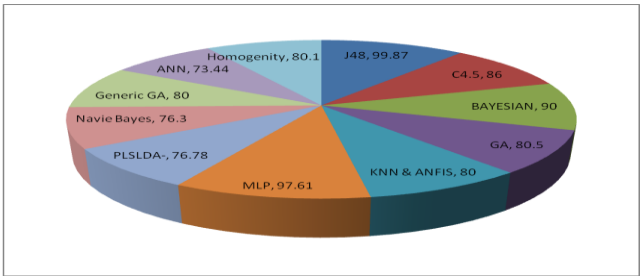

Fig 2: Performance accuracy comparative of various data mining techniques

8. Results

The results obtained from the given dataset categorized between two classes, i.e., patients with diabetes and without diabetes using some data mining techniques. The precision to predict the diabetes disease using different methods is shown in graphical representation within the fig 2. Based on the results demonstrated, Modified J48 classifier provides the highest accuracy 99.87% for predicting the diseases. The performance concerning the algorithm is calculated using the equation for Total Accuracy and Random Accuracy. Here, True positive and True Negative, False positive and False Negative parameters are taken to consider the equation. Radha compared classification methods and found the C4.5 decision tree algorithm gives better accuracy 86% into the prediction. Arwa Al-Rofiyee et al. used machine learning algorithm multilayer perceptron to predict the disease with 97.61% performance analysis accuracy.

9. Conclusion

In the medical field accuracy in prediction of the diseases is the most important factor rather than the execution time. In the analysis of data mining techniques and tools Modified J48 Classifier gives 99.87% of highest accuracy using WEKA & MATLAB tool. Since the diabetes is a chronic disease, it has to be prevented before it affects people. In future diabetes can be prevented using gene analysis and previous history of the diabetic

References

1. Han, j. And M. Kamber, Data Mining Concepts, and Techniques. 2006: Morgan Kaufmann Publishers. 2. Lee, I.-N., S.-C. Liao, and M. Embrechts, Data mining techniques applied to medical information. Med. inform, 2000.

2. Obenshain, M.K., Application of Data Mining Techniques to Healthcare Data. Infection Control and Hospital Epidemiology, 2004.

3. Sandhya, J., et al., Classification of Neurodegenerative Disorders Based on Major Risk Factors Employing Machine Learning Techniques. International Journal of Engineering and Technology, 2010. Vol.2, No.4.

4. Gaganjot Kaur, Amit Chhabra, "Improved J48 Classification Algorithm for the prediction of Diabetes", International Journal of Computer Applications(0975-8887) vol.98 No.22, July 2014.

5 P. Radha, Dr. B. Srinivasan, " Predicting Diabetes by consequencing the various Data mining Classification Techniques", International Journal of Innovative Science, Engineering & Technology, vol. 1 Issue 6, August 2014, pp. 334-339

6. Mohtaram Mohammadi, Mitra Hosseini, Hamid Tabatabaee, "Using Bayesian Network for the prediction and Diagnosis of Diabetes" , MAGNT Research Report, vol.2(5), pp.892-902.

7. Sudesh Rao, V. Arun Kumar, "Applying Data mining Technique to predict the diabetes of our future generations", ISRASE eXplore digital library, 2014.

8. Veena vijayan, Aswathy Ravikumar, " Study of Data mining algorithms for prediction and diagnosis of Diabetes Mellitus", International Journal of Computer Applications (0975-8887) vol. 95-No.17, June 2014

9. Arwa Al-Rofiyee, Maram Al-Nowiser, Nasebih Al-Mufad, Dr. Mohammed Abdullah AL-Hagery, "Using Prediction Methods in Data mining for Diabetes Diagnosis"http://www.psu.edu

10. K.R Lakshmi, S.Premkumar, " Utilization of Data mining Techniques for prediction of Diabetes Disease survivability", International Journal of Scientific & Engineering Research, vol.4 Issue 6, June 2013.

11. Murat Koklu and Yauz Unal, " Analysis of a International population of Diabetic patients Databases with Classifiers", International Journal of Medical,Health,Pharmaceutical and Biomedical Engineering", vol.7 No.8, 2013.

12. Rupa Bagdi, Prof. Pramod Patil," Diagnosis of Diabetes Using OLAP and Data Mining Integration", International Journal of Computer Science & Communication Networks,Vol 2(3), pp. 314-322.

13. Ashwinkumar.U.M and Dr. Anandakumar K.R, "Predicting Early Detection of cardiac and Diabetes symptoms using Data mining techniques", International conference on computer Design and Engineering, vol.49, 2012.

14. S. Sapna, Dr. A. Tamilarasi and M. Pravin Kumar, "Implementation of Genetic Algorithm in predicting Diabetes", International Journal of computer science, vol.9 Issue 1, No.3, January 2012.

15. Manaswini pradhan, Dr. Ranjit kumar sahu, " predict the onset of diabetes disease using Artificial Neural Network", " International Journal of Computer Science & Emerging Technologies, vol.2 Issue 2, April 2011.

16. Muhammad Waqar Aslam and Asoke Kumar Nandi, "Detection of Diabetes using Genetic Programming", European Signal Processing Conference (EUSIPCO-2010), ISSN 2076-1465.

17. Huy Nguyen Anh Pham and Evangelos Triantaphyllou, " Prediction of Diabetes by Employing New Data
mining approach which balances Fitting and Generalization Springer 2008.

