Estimation of Association Summarization Techniques Performance in Prediction of Diabetes Mellitus

J.Omana
Part Time Research Scholar, Anna University, India.

Dr.M.Moorthi
Professor, Dept. of Electronics and Communication Engineering, Prathyusha Engineering College, India.

Abstract – Diabetes is a menacing disease which is evolving as a challenge for human life irrespective of age groups. In order to reduce severe complications and mortality rate, prediction of diabetes in earlier stage has become mandatory. So the health care institutions all over the world are concerned in predicting and monitoring the risk factors by using different association rule techniques. To surpass the influential demerits such as data redundancy, less data space coverage which were encountered so far, an extension method of incorporating diabetes risk factors to produce an optimal summary is carried out. In this paper, various summarization techniques are studied and a comparative evaluation on the performance of each algorithm is performed to suggest an algorithm with high accuracy.

Index Terms – Association rule summarization techniques, Data redundancy, Risk factors, Data space coverage.

1. INTRODUCTION

Diabetes Mellitus is a disease which is caused due to lack of insulin production from pancreas or lack of cell response to the produced insulin. If diabetes is not treated earlier, it may lead to severe complications. According to recent report by Centres for disease control out of 30.3 million people, 23.1 million people have been diagnosed with either Type1 or Type2 diabetes. Diagnosing Diabetes manually is a complicated task. Hence, data mining techniques are used to predict prevalence of the disease.

Data Mining plays an important role in processing the huge amount of dataset and extract valuable information from them. This information is converted into useful knowledge which should be understandable in nature. Data mining include different techniques such as Clustering, Classification, Prediction, Association, Sequential patterns etc. Clustering is the process of grouping the data with similar features. Hence the data within a cluster is similar and dissimilar to the data in another cluster. Classification is used to analyze a new set of data and predict the group to which the data is belonging. The purpose of prediction is to identify the relationship between dependent and independent variables. Sequential pattern analysis is used to uncover similar patterns that existing in a transaction over a period. Those patterns are useful in business decision making. The Association rule mining generates .rules from the frequently occurring factors in a transaction. Association Rules generated from diabetes risk factor also provide justifications, which may serve as a guide for diabetes care .Data mining is used in various domain such as healthcare, Bioinformatics, Finance, Business in order to improve the performance in future, reduce the cost, enhance the efficiency and accuracy.

2. BACKGROUND KNOWLEDGE

2.1 KDD Process

Knowledge Discovery from Database in shortly known as KDD. The main objective of KDD process is to explore useful knowledge from large databases and predict the interesting patterns among them. KDD is an iterative process in which the following steps are repeated until an interesting, understandable pattern is obtained.

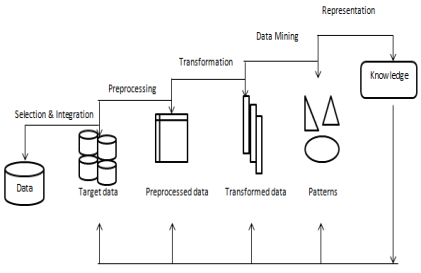

Fig 2.1.1 Steps involved in KDD

The steps involved in KDD process are as follows,

Initially develop an understanding on the domain and create/select a dataset. Later follow the steps given below,

- Data Pre-processing and Cleansing- the noise data, missing data and outliers are handled in this step.

- Data Integration- Various data sources are combined.
- Data Selection- Gleaning the data from database, that are relevant to analysis task
- Data transformation- Obtained data are converted into appropriate forms suitable for carrying out aggregation or summery operations.
- Data mining- Involves extracting data patterns by applying various methods.
- Pattern evolution- Involves identifying pattern representing knowledge from interesting measures.
- Knowledge presentation- Knowledge extracted are visualised by users using this knowledge representation technique.

2.2 Machine Learning

Machine learning is a kind of artificial intelligence, which enables software to predict the accurate outcomes based on the relations learnt from previous datasets. Machine learning used to generate algorithm to receive inputs and perform statistical analysis on the input to predict the output. ML algorithms are categorized into three, namely supervised learning, unsupervised learning and reinforcement learning.

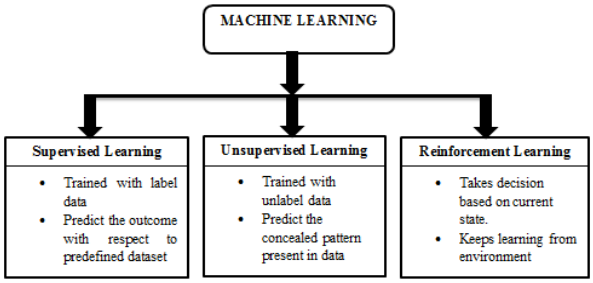

Fig 2.2.1 Classification of machine learning

The supervised learning algorithm is carried out on labelled dataset. In supervised learning, an algorithm is trained with predefined inputs. In most of the cases it is unable to figure out the function that always make the correct predictions .Hence the computer containing the input is also fed with valid output and from which the system should be able to learn the patterns. Based on the relationships that is learned between the target output and input values, the output for new dataset is predicted. In Unsupervised Learning, the computer is actually trained with unlabelled data. In this learning there is no any predefined input or output relationships. Here the algorithm use different techniques to analyse the data and discovers the interesting patterns between the data. This algorithm is useful, when the experts don't know what to be found from the given dataset. In Reinforcement Learning, the agent learns from the interactions with environment, inorder to take actions and to maximize the reward. Learning from environment is done in a iterative fashion. The agent must understand the current state, to make valid actions

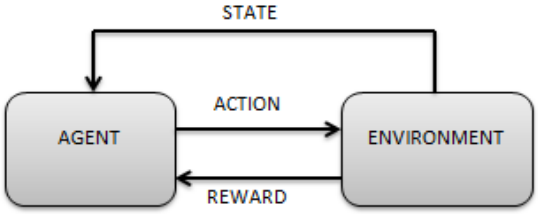

Fig 2.2.2 Reinforcement Learning

2.3 Association Rule Mining

Association rule mining is defined as, to find the frequent patterns and correlations among the relational database, transactional database and other data storage repositories. Association rule mining was initially introduced in the market as basket analysis tool. This technique is majorly concentrated in analysing unsupervised data. This plays a vital role in biology and bio informatics to reveal of a gene information data, as a result from a huge amount of raw data. Association rule performance depends on the SUPPORT, CONFIDENCE, MINIMUM SUPPORT THRESH HOLD and MINIMUM CONFIDENCE THRESHHOLD. The association rules are applied in the database transaction. Each data in the transactions are known as the item. By applying some rules the frequent pattern is revealed out and basket analysis is one of the examples for the frequent pattern occurrence in the association rule mining.

2.4 Classification Technique

Classification is the process of analysing the given dataset and identifies the predefined group or classes to which the analysed data is related to. Some of the classification algorithms are Decision trees, linear classifiers, Quadratic classifiers, Kernel estimation, Support vector machines, Neural networks, learning vector quantization. The calculation of Id3 is done in a data set starts with root. The values in the data set are classified based on the Entropy value. The data belongs to same class comes under the one class. When there is no more data of same class it will make as a separate class. ANN (Artificial Neural Networks) is more or less similar to biological neural systems. Because of its adaptive nature it process on a large volume of data input and is able for machine learning.

3. DIABETES MELLITUS

Diabetes Mellitus (DM) can be defined as a metabolic disorder which is due to high blood glucose level in the body for a prolonged period. This disease is mainly caused when there is an irregular segregation of insulin by the body and when there is no proper response of the human body cell

towards the segregated insulin. Symptoms of diabetes are Polyuria, Polyphagia, and Polydipsia in common. Untreated diabetes may lead to severe complications ranging from cardio vascular disease, macro vascular disease to death

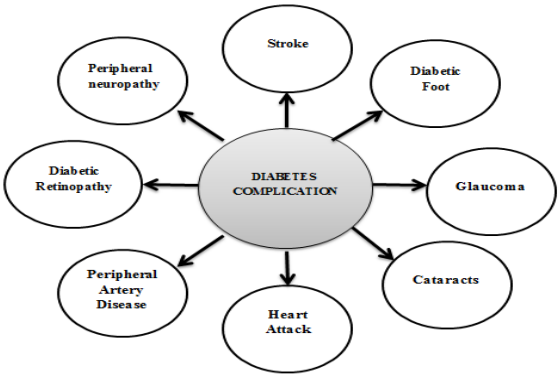

Fig 3.1 Consequences of diabetes

In the year of 2012, 2.2 million life losses were recorded which the root cause was diagnosed as the high blood sugar level. In the year of 2014 the adults who crossed age of 18 had diabetes. In 2015 the death rate caused by the diabetes is 1.6 million. The WHO (World Health Organization) extends its service to monitor, control and prevent occurring in medium income countries to reduce the loss of life. WHO provides "Global Report on Diabetes" which presents as a guideline for the government civil society and private sector. It also includes the burden and complications of diabetes and its side effects. It suggests people to follow the healthy diet and physical exercise to reduce the effect of this harmful disease.

3.1 Types of Diabetes

There exist few types of diabetes. They are Type 1DM, Type 2DM, Gestational diabetes.

Type 1 Diabetes:

This is caused due to lack of insulin production from pancreas. The insulin production is reduced by depleting Beta cell. This type of diabetes is also referred as IDDM(Insulin dependent diabetes mellitus). It mostly occur in younger age groups

The symptoms are constant hunger, vision changes, fatigue, vomiting etc. This type is treated with insulin injection, which balances the lack of insulin in the human body. The type 1 diabetes cannot be prevented in earlier stages, but they are controlled with the use of insulin injections.

Type 2 Diabetes:

Type 2 diabetes mellitus is due to lack of cell response to the produced insulin. They are also referred as NIDDM (non-Insulin dependent diabetes mellitus).The type 2 diabetes occurs in adult age group specifically greater than 35. Type 2 diabetes symptoms are similar to type 1.

These symptoms spread very slowly and may also be absent in some cases, hence this type is considered as menacing when compared to other types. Th type 2 diabetes is treated with proper diet, medications, weight loss surgery and insulin injection in severe cases.

Gestational Diabetes:

Gestational diabetes occurs during maternity period of women. Usually it will be naturally cured or it may result in type 2 diabetes after pregnancy. Gestational diabetes requires a careful medical monitoring to cure it completely. If gestational diabetes is not treated in earlier stages it may affect either the mother or babies health.

3.2 Comparison of Diabetes Types

FEATURES	TYPE 1	TYPE 2
Age Of Onset	Diagnosed during childhood(age<20)	Diagnosed in adults(age>30)
Body weight	Thin or Normal	Obese
Ketoacidosis	Common	Rare
Auto antibodies	Present	Absent
Symptom Growth	Develop rapidly	Develop slowly and may be subtle or absent.
Defects	Beta cells are destroyed.	Insulin resistance; Other defects
Treatment	Insulin Injections or Insulin Pump devices are used	Tablets and proper Diet is followed. Insulin Injections are also used in severe cases.
Prevalence	Less	More

4. COMPARISON OF ALG0RITHM

4.1 Adaboost:

[2]The objective of this paper is to classify the diabetes patients in three different age groups using the risk factors identified. The algorithms involved in this process are Adaboost and Bagging which uses J48 Decision tree as base learner. In this paper, chi-square test is been carried out to determine the presence of diabetes, the result produced was adult had more possibilities to develop diabetes when compared with other age groups. In order to find out the efficiency of classifiers the dataset from CPCSSN is divided into three cohorts and AROC curve test is performed..Finally, the performance of Adaboost is found to be efficient, while dealing with small classes of data.

4.2 Bagging:

[2]In this paper the efficiency of each ensemble techniques namely Adaboost, Bagging that is applied along with J48 decision tree as base learner is determined. According to bagging, each bootstrap replicates contain 63.2% of original data .Hence by processing repeatedly, the result obtained from weak learners is composed with strong learners to improve the accuracy. Once the diabetes presence is predicted, the efficiency of ensemble techniques with J48 decision tree is predicted using AROC curve test. By doing so, the Bagging algorithm Proves to be more efficient while processing large datasets with 0.98% of Perception capability of classifiers.

4.3 K-Nearest Neighbors:

[3]The objective of this paper is to group the patients with Diabetes Mellitus using Classification algorithm namely J48 Decision Tree, Support Vector Machine, KNN, Random Forest and to predict the algorithm with better performance in terms of Accuracy, Sensitivity, Specificity.KNN algorithm is used for classifying the available data based on the vote rendered by the neighbours and Euclidean function is used here to measure the distance. It was suggested that how large is the k-value, the accuracy will be that much greater. In order to estimate the performance of algorithms based on the above mentioned criteriaan Confusion Matrix is constructed.

Accuracy=(TP+TN)/(TP+TN+FP+FN)

Sensitivity=TP/(TP+TN)

Specificity=TN/(TN+FP)

The above mentioned equations are used in this paper to calculate the accuracy, sensitivity, specificity from the constructed matrix. Based on the determined criteria the performance of algorithm is predicted before pre-processing, the result produced was J48 decision tree is optimal with 73.82% accuracy, 59.7% sensitivity, 81.4% specificity. After PreProcessing, once again the Performance comparison is done, in which KNN at K=1 is found to be optimal with 100% of accuracy, sensitivity and specificity.

4.4 Random Forest:

[3]In this paper the performance of Classification algorithms are compared in two different stages namely before PreProcessing and after PreProcessing. Random Forest algorithm uses bagging approach to achieve better prediction while analysing the algorithms. Performance before cleaning the data, Random forest has shown only 71.74% of accuracy, 53.81% of sensitivity, 80.4% of specificity. The performance of Random Forest has reached 100% after cleaning the data. From this paper, the importance of PreProcessing in identifying the efficient algorithm is clear.

4.5 Decision Stump:

[4]In this paper Adaboost algorithm is used for predicting the prevalence of diabetes. The process involved in this paper are global and local dataset collection, training global data with Adaboost under different base learners namely Support vector machine, Decision Tree, Naïve Bayes and Decision Stump. And then local dataset is validated on the classifiers mentioned above. Finally the performance comparison of classifiers in terms of accuracy, error rate, sensitivity and specificity is done.

Decision stump is a kind of decision tree that contains only one level under each root. The performance of classifiers without Adaboost algorithm is estimated. In which SVM shows the highest accuracy rate 79.6%, but the decision stump had the lowest accuracy rate 74.4% while comparing with other classifiers. After including Adaboost algorithm with classifiers, the accuracy rate of decision stump 80.72% has become the highest than all other classifiers and even the error rate of decision stump 19.27% is the least one. The performance of SVM remains the same even after including boosting algorithm, but the decision stump is proves to be efficient is predicting in the prevalence of diabetes while working as a base classifier for Adaboost algorithm.

4.7 Apriori:

[8] In this paper Apriori algorithm is used for predicting diabetes medications. The process carried out in this paper is collection of raw data, pre-processing and the data is converted into binary matrix form. This representation is subjected to two divisions with minimum support of distinct values (30% & 50%).

While analysing with minimum support of 30%, the result produces all symptoms and all common medications are preferred. While analysing with minimum support of 50%, the result produced was Type 1 Diabetes symptoms ,based on that appropriate insulin medications are suggested.

5. RESULT

As a result of this survey it was estimated that each algorithm has its own efficiency, irrespective of parameters used. From this survey, efficiency of each individual algorithm is estimated in which it was identified that the KNN algorithm is capable of producing 100% of accuracy, sensitivity and specificity, once the data been used is pre-processed. Similarly the random Forest algorithm can also reach 100% in accuracy, sensitivity, specificity with pre-processed data. By this the importance of PreProcessing in improving the efficiency of algorithm is identified. Decision Stump is a kind of Decision tree algorithm which can improve its performance up to 80.72% with the help of boosting algorithms. If decision tree grows larger, the complexity will also increases, whereas decision stump has an advantage(i.e) it has only one level below it. The Apriori algorithm efficiency varies based on the support count value used. The boosting algorithm bagging is able to perform well with larger datasets with higher efficiency.

NAME OF THE ALGORITHM	ACCURACY (%)	SENSITIVITY (%)	SPECIFICITY (%)
KNN(before PreProcessing)	70.18	52.98	79.4
KNN(after PreProcessing)	100	100	100
Random Forest(before PreProcessing)	71.74	53.81	80.4
Random Forest(after PreProcessing)	99.5	99.8	99.6
Decision stump(with boosting algorithm)	80.72	88.4	64.5
Bagging	98	62	74.2

Table 5.1: Comparison of Association Summarization Techniques

Based on the comparison of association rule summarization techniques performance, the KNN and Random Forest algorithm processing with pre-processed data is identified to be efficient in prediction of risk factor of diabetes mellitus.

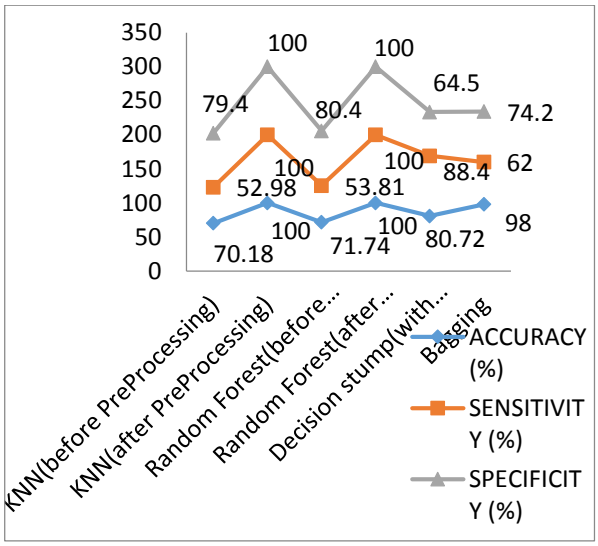

Fig 5.1 Depiction of comparison

6. CONCLUSIONS

In this paper a study on various existing data mining techniques are performed. The study includes the details about performance of algorithms in different scenarios. As per the study, it is understood that bagging algorithm can be used for reducing variances and it is capable of handling large dataset. The Adaboost algorithm is efficient in classifying smaller groups. Decision Stump show higher accuracy in prediction when boosted with Adaboost algorithm. KNN and Random Forest both the algorithm renders 100% accuracy, while processing with a pre-processed data. This study can be used as a guideline for future enhancement of these techniques.

REFERENCES

[1] J Omana, S Monika, B Deepika, "Survey on Efficiency of Association Rule Mining Techniques" International Journal of Computer Science and Mobile Computing, Volume 6 Issue 4, April 2017, Pages 5-8

[2] Omana .J, Sujithra .S, Vishali .S, Yuvashree .K. Data Mining Techniques in Prediction of Risk Factors of Diabetes Mellitus, International Journal of Advance Research, Ideas and Innovations in Technology, Volume 4 Issue 1, Pages 228-232

[3] IoannisKavakiotis ,Olga Tsave , AthanasiosSalifoglou ,NicosMaglaveras , Ioannis Vlahavas ,IoannaChouvarda.MACHINE LEARNING AND DATA MINING METHODS IN DIABETES RESEARCH.ELSEVIER 2017.

[4] SajidaPerveen ,Muhammadshahbaz ,Aziz guergachi, KarimKeshavjee .PERFORMANCE ANALYSIS OF DATA MINING CLASSIFICATION TECHNIQUES TO PREDICT DIABETES.ELSEVEIR 2016.

[5] J.Pradeep Kandhasamy,S.Balamurali.PERFORMANCE ANALYSIS OF CLASSIFIER MODELS TO PREDICT DIABETES MELLITUS.ELSEVIER 2015.

[6] V. VeeranVijayan, C.Anjali. PREDICTION AND DIAGNOSIS OF DIABETES MELLITUS-A MACHINE LEARNING APPROACH.IEEE Recent Advances in Intelligent Computational Systems Dec 2015.

[7] Gyorgy J. Simon, Pedro J. Caraballo, Terry M. Therneau, Steven s. Cha, M.Regina Castro, Peter W. Li. EXTENDING ASSOCIATION

RULE SUMMARIZATION TECHNIQUES TO ASSESS RISK OF DIBETES MELLITUS. IEEE Transactions On Knowledge And Data Engineering Jan 2015

[8] KonstantiaZarkogianni, EleniLitsa ,KonstantinosMitisis, Po-Yen Wu, Chanchala D. Kaddi, Chi-Wen Cheng, D. Wang,Konstantina S. Nikita. A REVIEW OF EMERGING TECHNOLOGIES FOR THE MANAGEMENT OF DIABETES MELLITUS. IEEE Transaction On Biomedical Engineering Dec 2015.

[9] R.Priya , R.Roshma . PREDICTION OF CO-MORBID CONDITIONS ASOCIATED WITH DIABETES USING SPLIT AND MERGE ALGORITHM. International Journal of Innovative Research in Computer and Communication Engineering Jul 7 2015.

[10] DhirajPandey, Santosh Kumar. PREDICTION SYSTEM TO SUPPORT MEDICAL INFORMATION SYSTEM USING DATA MINING APPROACH. International Journal of Engineering Research and Applications May-Jun 2012.

[11] J Omana, S Dhanalakshmi, VM Divyalakshm,"Categorization of Drugs Using SVM Classification" International Journal of Computer Science Trends and Technology (IJCST) – Volume 5 Issue 2, Mar – Apr 2017

[12] I Mohan, "Knowledge Discovery using Big Data" Journal of Current Computer Science and Technology Volume 5 Issue 05 April 2015

[13] I Mohan, Ajith Kumar, Ajith Kumar, S Bhuvanesh "Relevance Feature Discovery for Text Mining Using Feature Clustering" International Journal of Scientific Research in Computer Science, Engineering and Information Technology Volume 2 Issue 2 Pages 661-665 April 2017