INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 5 ISSUE 1 (2019) PAGES 4096 - 4104
RECEIVED : 15.12.2018 PUBLISHED : 06.01.2019
Jan 06, 2019

COMPARATIVE ANALYSIS OF CLASSIFICATION TECHNIQUES FOR DIABETES DIAGNOSIS

P.Janani[1], N. Nirmala Devi [2]
[1]Assistant Professor, [2]Assistant Professor
Department of Computer Science, Auxilium College, Katpadi, Vellore, India.

ABSTRACT

Data mining is the process of analyzing data from different perspectives and summarizing it into an u seful information.In this paper we proposing an different classification algorithm to identify the accurac y on diabetic data sets.we are proposing an efficient two level for classifying data. Initial phase we use tra ining data for analyzing the optimality of dataset then new dataset is formed as optimal training dataset n ow we apply our classification mechanism on new diabetic datasets. The data mining methods and techni ques will be explored to identify suitable methods and techniques for efficient classification on diabetic da ta set and in mining it in useful patterns.

Keywords: Datamining, Diabetic dataset, classification, Naïve Bayes classification, Random forest

1. INTRODUCTION

The aim of this study focuses to investigate the performance and accuracy on different classification m ethods using weka.A problem that occurs in the bioinformatics or the medical science is to reach the corr ect diagonosis of certain important medical information.For ultimate diagnosis generally many tests are do ne involving clustering and classification of the data.All of this testing procedures are necessary to reach t he ultimate diagnosis.In the other hand too many tests will be complicated and the process is difficulty in obtaining the end results.This kind of difficulties were resolved by using the various classification method s. Diabetes mellitus or simply diabetes is a set of related diseases in which the body cannot regulate the a mount of sugar in blood level.In every age group this is common.It charges plenty of money and its growi ng quickly.This is also known as metabolic disease or heredity diseases in which the person will have the high blood sugar level either the body will not produce enough insulin or the beta cells in body will not re spond to the insulin that is produced in the pancreas.This is classified into the three types based on the sy mptoms polyuria,polydipsia,polyphagia.The diabetic person has risk and leads to other disease such as bl ood vessel damage,blindness,heart diseases,nerve damage,and kidney diseases.Diabetics also classified as two types such as type 1 insulin diabetes and non insulin dependent,diabetes is a disease in which the bl ood glucose increases which is due to the defects of secretion of insulin,or its action or both.

Diabetes is a prolonged medical disease. In diabetes the cells of person produce insufficient amount of in sulin or defective insulin may insulin or may unable use insulin properly and efficiently that further leads to hyperglycemia and type-2 diabetes.In the diabetes federation has claimed that presently 246 million pe ople are suffering for diabetes worldwide and its number is expected to increase up to 380 million 2025.T here are substantial amount of research has been done on the medical data with various algorithms such as bayes,J48 graft,c4.5,conjuctive rule learner.

INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 5 ISSUE 1 (2019) PAGES 4096 - 4104
RECEIVED : 15.12.2018 PUBLISHED : 06.01.2019
Jan 06, 2019

2. RELATED WORK

A good amount of data mining techniques applied in the medical diagnosis.RESHEDUR M. RAHM AN,FARHANA AFROZ [1] proposed a techniques with comparison different classification algorithms a nd different tools such as Weka,Tangara and MATLAB to find the accuracy of single diabetes data sets w ith different tools and different classification algorithm.It achieves the 79.19%using Machine learning,78. 98% using Weka,and 81.33% using J48 in weka.In tangara 83.85%,100%,90.63 using same ML,NB and J 48.

DR.R.S.KAMATH[2] identifies exploration of mining algorithm in diabetic patients database.Here the cl assification techniques are applied to classify the data and the data of the diabetes patient is evaluated wi th 10 fold cross validation and results comparison is done with that validation.It includes classification te chniques such as Naïve bayes,k-star,One R,Simple cart.which gives the accuracy such as 77.80%,71.17% ,75.76%,76.02.

VINCENT LABATUT,HOCINE CHEERIR[3] proposes performance measures for different classifiers.It includes ground truth index(GTI)for getting the classification accuracy.The comparison done with the the oretical point of view.The classification accuracy is done with different measures jaccard's coefficients.

MUKESH KUMAR,DR.RAJAN VOHRA,ANUSHAL ARORA[4]for prediction of diabetes using bayes classification.The data set are collected from the hospital and classification is used for analysis.It uses the trees J48 and bayes.net and gives the accuracy of 76.75% and 75.30 %using the single datamining tool W EKA.

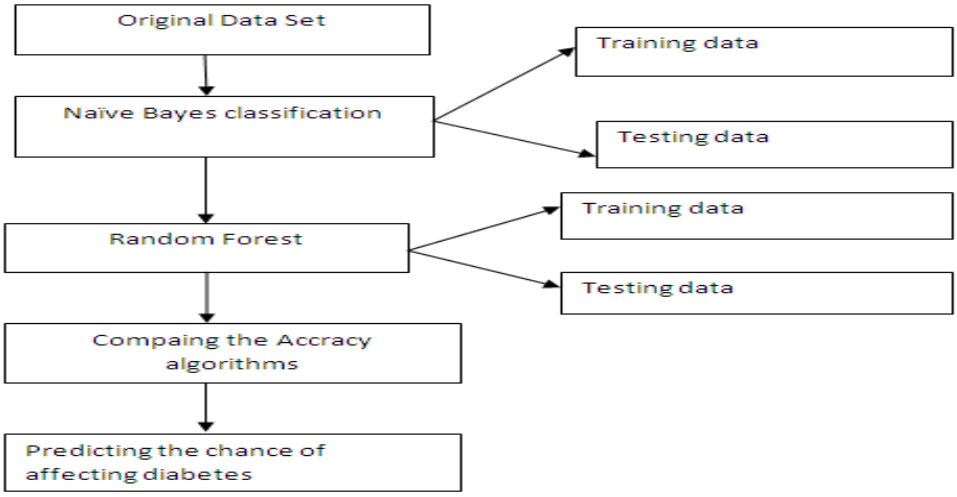
G.KEERTHANA,DR.V.SRIVIDHYA[5]Identifies the performance with both by using classification and clustering algorithm.It includes the various classifications such as Bayes.Net,Naïve bayes,One R,cluster ing algorithm to achieve the accuracy of the dataset.

NADAV DAVID MAROOM,LIOR ROKACH,ARMIN SHMILOVICI[6]for improving the confusion m atrix values in classification.It demonstrate the benefits of methods by applying it to error correcting code using Adaboost with orthogonal array code matrix.

3. RESEARCH FRAMEWORK

Our methodology adopted for the implementation of research problem which begins with the data collection. The training data set used for data mining is the pima Indian diabetes databases from t he UCI machine learning repository.As per the requirements the dataset is converted in the required form. comparitive study of algorithms is carried out to select the efficient one. Our objective of the classi fication is to assign the class to find previously unseen records as accurately as possible.The motive is to find a classification model based on the class attributes and to find the accuracy of the model. The given data set is divided with split percentage as training set and test set. The training set is to build a model an d test set is used to validate the accuracy when the new data set arrives.

INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 5 ISSUE 1 (2019) PAGES 4096 - 4104
RECEIVED : 15.12.2018 PUBLISHED : 06.01.2019
Jan 06, 2019



Architecture diagram of proposed work

4. DATA SET DESCRIPTION

The characteristics of the data set used in the research,the detailed description of the data sets are avail able in UCI repository[8].Our objective of the data set is to diagnosis of diabetes data set based on the per sonal details of the patients such as Age,Blood pressure,Body MassIndex(BMI),Insulin,plasma it is useful to decide where the Pima Indian Diabetes Data(PIMA),belongs to the class tested positive or tested negat ive. The data set available publicly from UCI machine learning respository. The problem posed here is to predict where the person would tested as positive or tested negative. This is referred as two class problem were 1 being interpreted as tested positive and 2 being interpreted as tested negative. 500 belongs t o the class 1 and 268 belongs to the class 2.

TABLE1:DATASET

Dataset	No.of example	Input attributes	No.of classes	Total no.of attributes	Noisy attributes
PIMA	768	7	2	8	NO

The purpose of the study is to investigate the relationship between the diabetes and diagonostic results a nd a list of variables that represents the measurements and medical attributes that includes the 768 tuples with the input attributes of 7 and classes is 2 and total no of attributes as 8.The attributes that used for the classification is given below:

INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 5 ISSUE 1 (2019) PAGES 4096 - 4104
RECEIVED : 15.12.2018 PUBLISHED : 06.01.2019
Jan 06, 2019

Plasma glucose concentration

Blood pressure(mm Hg)

Triceps skin fold thickness(mm)

Insulin (mm U/ml)

Body mass index(weight in kg/(height in m)^2)

Diabetes pedigree function

Age(years)

Class variable (positive 1 or negative 2)

5. NAÏVE BAYES

The Bayesian classification represents a supervised learning as well as the statistical method for classificat ion.Assumes an underlying probabilistic model and allows us to capture uncertainity about the model in a principled way of determining probabilities of the outcomes.

It can solve diagnostic and predictive problems. This classification is named by the Thomas Bayes(17 02-1761) who proposed the bayes theorem. Bayesian classification provides a useful perspectives for u nderstanding and evaluating many learning algorithms.

It calculates explicit probabilities for hypothesis and its robust to noise in input data. Bayesian classification is based on bayes theorem.Simple bayesian classification known as the naïve bayesian class ifier to be comparable in performance with the decision tree and neural network classifiers.Bayesian classi fiers have also exhibited high accuracy and speed when applied to the large datasets.Bayes theorem is usef ul in that it provides a way of calculating the posterior probability,p(H|X)from p(H),P(X),and p(X|H),Bay es theorem.

Where p(X|H)is posterior probability of X on H,P(X) is the prior probability of X.P(H)which is independe nt on X.

RANDOM FOREST

Random Forest is the supervised machine learning algorithm. Random Forest has tremendous techniqu e potential of becoming of popular technique for future classifiers. Random forest algorithm is one of the best among the classification algorithm for classifying the large sets of data. It is also a combinati on of tree predictor where each tree depands on the values of a random vector sampled independen tly with same distribution for all trees in the forest. Introducing a right kind of randomness make them accurate classifiers. The algorithm was developed by Leo Breiman and Ad ele cutler.Random forest grows many classification trees.

INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 5 ISSUE 1 (2019) PAGES 4096 - 4104
RECEIVED : 15.12.2018 PUBLISHED : 06.01.2019
Jan 06, 2019

1)If the number of cases in the training set is N,sample N cases at random but with the replacem ent from the original data .This sample will be the training set for growing the trees.

2)If there are M input variables a number mM is specified such that at each mode,m variables are selected at random out of the M and the best split on these m is used to split the node.The value of the m is held constant during the forest growing.

3)Each tree is grown to the largest extent possible.

IMPLEMENTATION USING WEKA

The aim of data mining is to make sense of large amounts of mostly unsupervised data,in some domain.cl assification maps into predefined groups.It is often referred to as supervised learning as the classes are det ermined prior to examining the data.Two classes tested positive and tested negative are defined based on the data attribute value during the analysis of diabetes dataset.

6. PERFORMANCE METRICS

We measure the performance of the classifiers with respect to different metrics like accuracy, precis ion,recall,F-measures,ROC curve and gamma statistics along with the confusion metrics. The TP is defin ed as the True positive which it is calculated by using the formula

TRUE POSITIVE

TP rate =diagonal element of confusion metrics/sum of relavant rows

Which refers the positive tuples that were correctly labeled by the classifier.

FALSE POSITIVE

False positive is calculated by using the formula

FP rate=non_diagonal element of the confusion metrics/sum of relavant row

Which refers the negative tuples that were incorrectly labeled by the classifier

TRUE NEGATIVE

True negative is calculated by using the formula

TN=Diagonal element of the confusion metrics/sum of relavant column.

True negative refers the negative tuples correctly labeled by classifier.

FALSE NEGATIVE

False negative is calculated by using the formula

INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 5 ISSUE 1 (2019) PAGES 4096 - 4104
RECEIVED : 15.12.2018 PUBLISHED : 06.01.2019
Jan 06, 2019

FN=Nondiagonal element of the confusion metrics/sum of relavant column.

False negative refers positive tuples that were incorrectly labeled by classifier.

PRECISION

Precision in weka is defined as fraction of retrieved element that are relavant to find the precision

Precision of class A=diagonal element/sum of relavant column.

Precision of class B=non diagonal element/sum of relavant column.

RECALL

Recall is defined fraction of the elements that are relavant to the query that are successf ully retrieved.

Recall =tp/tp+fn

ACCURACY

Accuracy is determined by using the formula

F_MEASURES

A measure that combines precision and recall is the harmonic mean of precision and recall the F-measure is defined by:.

CONFUSION MATRIX

Evaluation the confusion matrix at each iteration enables making decision regarding the next one agains t all classifier that should be added to the current code to demonstrate the benefits of the method by apply ing it to error correcting code orthogonal arrays as the basic code matrix.

TABLE 1:

TRUE POSITIVE	FALSE POSITIVE
FALSE NEGATIVE	TRUE NEGATIVE

POSITIVE NEGATIVE

For two class matrix the true positive,true negative,false positive,false negative will be in the order whi ch is given in the table 1.

COMPARISION OF NAÏVE BAYES AND RANDOM FORESTALGORITHM IN WEKA

Our comparison is based on the time taken,accuracy,and confusion matrix.

TABLE 2:

Timetaken	Naïve bayes	RandomForest
Train	0.02 seconds	0.14 seconds
Test	0.02 seconds	0.11 seconds

TABLE 3:

	Correctly Classified Instances	Incorrectly Classified Instances
Naïve Bayes Train	76%	23%
Naïve Bayes Test	75%	25%
Random Forest Train	97%	2%
Random Forest Test	97%	2%

TABLE 4:

Confusion Matrix:Naïve Bayes

a b <-- classified as

156 112 | a = tested_positive

70 430 | b = tested_negative

Confusion Matrix:Naïve Bayes Test

a b <-- classified as

86 59 | a = tested_positive

37 202 | b = tested_negative

INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 5 ISSUE 1 (2019) PAGES 4096 - 4104
RECEIVED : 15.12.2018 PUBLISHED : 06.01.2019
Jan 06, 2019

Confusion Matrix:Random Forest Train

a b <-- classified as

266 2 | a = tested_positive

15 485 | b = tested_negative

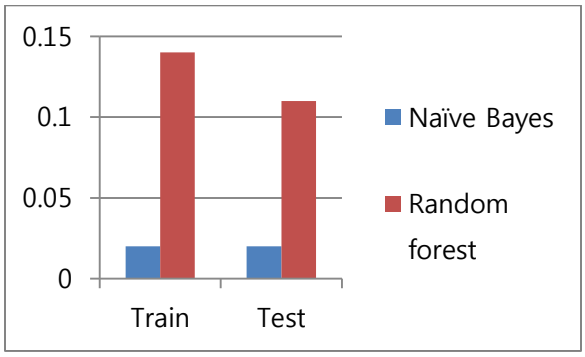
Confusion Matrix:Random Forest Test

a b <-- classified as

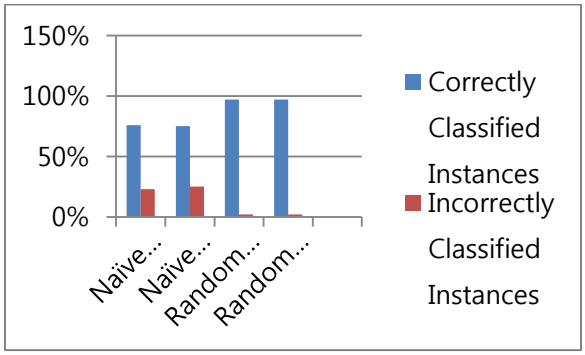
143 2 | a = tested_positive

9 230 | b = tested_negative

Graph: TABLE 2



GRAPH: TABLE 3



INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 5 ISSUE 1 (2019) PAGES 4096 - 4104
RECEIVED : 15.12.2018 PUBLISHED : 06.01.2019
Jan 06, 2019

CONCLUSION

In this research data mining technique applied to classify Diabetes data and predict the patient has chances of being affected by diabetes or not.Different types of classification algorithm applied to the single Pima Indian Diabetes Dataset and the above the results obtained tabulated the table. The research can be extended by applying association mining. This work extends to utilize the implementation of different medical dataset.

REFERENCES

RashedurM.Rahman.FarhanaAfroz, "comparison of various classification techniques using different data mining tools for diabetes diagnosis",Journal of Software Engineering and Applications,2013.

Dr.R.S.Kamath,"Weka Approach for Exploration Mining in Diabetic Patients Database",Chatrapati Shahu Institute of Business Education and Research Kolhapur,India.2013

Vincent Labatut,Hocine cherifi",Evaluation of Performance Measures For Classifiers Comparison",UniversityofBurgundy,Turkey,2013

Mukesh kumari,Dr.Rajan vohra,Anushal arora,"Prediction of Diabetes Using Bayesian Network",P.D.M college of Engineering,Bhadurgarh,International Journel of Computer Science and Information Technologies 2014.

G.K.Keerthana,Dr.V.Srividhya,"Performance Enhancement of Classifiers Using Integration of Clustering and Classification Techniques,International Journal of Computer Science and Engineering, 2012

Nadav David Marom,Lior Rokach,Armin Shmilovici,"Using the confusion Matrix For Improving Ensemble Classifier,2010 IEEE 26th convention of Electrical and Electronics Engineers in Isreal.

.