Performance Analysis of Different Classification Methods in Data Mining for Diabetes Dataset Using WEKA Tool

Sujata Joshi
Assistant Professor, Dept. of CSE
Nitte Meenakshi Institute of Technology
Bangalore, India
e-mail:sujata_msrp@yahoo.com

S. R. Priyanka Shetty
Research scholar M tech, Dept. of CSE
Nitte Meenakshi Institute of Technology
Bangalore, India
e-mail:siddamshettypriya@gmail.com

Abstract— Data mining is the process of analyzing data based on different perspectives and summarizing it into useful information. Classification is one of the generally used techniques in medical data mining. The goal here is to discover new patterns to provide meaningful and useful information for the users. Recently data mining techniques are applied to healthcare datasets to explore suitable methods and techniques and to extract useful patterns. This paper includes implementation of different classification methods, measures, analysis and comparison pertaining to diabetes dataset. A detailed performance analysis and comparative study of these methods are done, which can be further used to choose the appropriate algorithm for future analysis for the given dataset.

Keywords- *Data mining, Classification algorithms, WEKA tool*

*****

I. Introduction

Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD)[1] is the computational process of discovering patterns in large data sets which involves methods using artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. It is an analytic process designed to explore data in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns.

Data mining includes classification as one of the fundamental task. Classification predicts categorical valued functions and prediction models predict continuous valued functions. The focus of this paper is to apply various classification algorithms such as Bayesian, Naïve, J48, RandomForest, Random tree, REP, k-NN, CART and Conjunctive rule learning on the selected dataset. A detailed performance analysis is done using different measures like sensitivity, specificity, accuracy, positive precision, negative precision and error rate.

Medical data mining has become prominent in the recent years since enormous amount of medical data is available which can be used to discover useful patterns. Among all diseases, diabetes is a chronic disease which is affecting almost 25% of the Indian population. Diabetes is the chronic disorder of the glucose-insulin metabolism characterized mainly by high blood glucose concentration in which the body cannot regulate the amount of sugar in blood; this condition is known as hyperglycemia which is associated with long term complications[1]. Diseases or chemicals that damage or destroy the pancreas can also cause diabetes. Examples include pancreatitis, pancreatic cancer, and hemochromatosis, a disorder in which excessive amounts of iron accumulate in the pancreas and other organs. Other specific types include diabetes due to genetic defects, drug induced diabetes etc.

There are mainly three types of diabetes mellitus. In Type1 diabetes, the pancreas fails to produce insulin and hence it requires patient to inject insulin or wear an insulin pump. Type2 diabetes results from insulin resistance, in which cells fail to use insulin properly. The third form, gestational diabetes occurs when pregnant women without a previous diagnosis of diabetes develop high blood glucose level. It mainly precedes development of type2 diabetes mellitus[2].

II. RELATED WORK

Classification has been successfully applied to a wide range of application areas, such as scientific experiments, medical diagnosis, weather prediction, credit approval, customer segmentation, target marketing and fraud detection [5,6]. Decision tree classifiers are used extensively for diagnosis of breast tumor in ultrasonic images, ovarian cancer, heart sound diagnosis and so on.

Arvind Sharma and P.C. Gupta discussed that data mining can contribute with important benefits to the blood bank sector. J48 algorithm and WEKA tool have been used for the complete research work. Classification rules performed well in the classification of blood donors, whose accuracy rate reached 89.9%[7].

As medical records systems become more standardized and commonplace, data quantity increases with much of it going unanalyzed. Taking into account the prevalence of diabetes among men and women the study is aimed at finding out the characteristics that determine the presence of diabetes and to track the maximum number of men and women suffering from diabetes with 249 population using WEKA tool[8].

III. CLASSIFICATION

Classification is a method used to extract models describing important data classes or to predict the future data. Classification is two step process: i) Learning or training step where data is analyzed by a classification algorithm. ii) Testing step where data is used for classification and to estimate the accuracy of the classification[4].

Classification Techniques:

A. Bayesian classification:

Bayesian classifiers are statistical classifiers. They can predict class membership probabilities, such as probability that a given tuple belongs to a particular class. It uses various searching algorithms and quality measures, based on bayes network classifier and provide data structure.

B. *Naïve Bayesian classification:*

Naïve Bayesian classification is also known as simple Bayesian classifier used to compare performance with decision tree and selected neural network classifiers. It assumes that the effect of an attribute value on a given class is independent of the values of the other attributes; this assumption is called as class conditional independence.

C. *Decision tree(J48*):

Decision tree is a flow chart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node holds a class label.

The decision tree classifier has two phases:

i) Growth phase or Build phase.

ii) Pruning phase.

The tree is built in the first phase by recursively splitting the training set based on local optimal criteria until all or most of the records belong to each partition.

The pruning phase handles the problem of over fitting the data in the decision tree. It removes the noise and outliers. The accuracy of the classification increases in this phase.

a) *J48 or C4.5:* This algorithm is based on Hunt's algorithm[9]. It handles both categorical and continuous attributes to build a decision tree. In order to handle continuous attributes, C4.5 splits the attribute values into two partitions based on the selected threshold such that all the values above the threshold as one child and the remaining as another child. It also handles missing attribute values. It uses Gain Ratio as an attribute selection measure to build a decision tree. C4.5 uses pessimistic pruning to remove unnecessary branches in the decision tree to improve the accuracy of classification.

D. *RandomForest:*

This was proposed to future enhance the scalability of decision tree induction. It adapts to the amount of main memory available and applies to any decision tree induction algorithm. It maintains an AVC set (attribute-value-class label) for each attribute, at each tree node, describing the training tuple at the node.

E. *Random tree:*

Fast decision tree learner. Randomly select an attribute at each node, fix some threshold and classify into tree. Builds a decision/regression tree using information gain/variance and prunes it using reduced-error pruning (with back fitting). Only sorts values for numeric attributes once. Missing values are dealt with by splitting the corresponding instances into pieces (i.e. as in C4.5).

F. *REP tree:*

Reduced error pruning (REP) class for constructing a tree that considers K randomly chosen attributes at each node. It performs only on numeric attribute and no pruning. Also has an option to allow estimation of class probabilities based on a hold-out set (back-fitting).

G. *CART:*

CART[10] stands for Classification And Regression Tree, based on Hunt's algorithm. CART handles both categorical and continuous attributes to build a decision tree. It handles missing values. CART uses Gini Index as an attribute selection measure to build a decision tree. CART produces binary splits. Hence, it produces binary trees. CART uses cost complexity pruning to remove the unreliable branches from the decision tree to improve the accuracy.

H. *k-NN:*

Nearest neighbor classifiers are based on learning by analogy. The training samples are described by n-dimensional numeric attributes. Each sample represents a point in an n-dimensional space. In this way, all of the training samples are stored in an n-dimensional pattern space. When given an unknown sample, a k-nearest neighbor classifier searches the pattern space for the k training samples that are closest to the unknown sample. These k training samples are the k "nearest neighbors" of the unknown sample.

"Closeness" is defined in terms of Euclidean distance, where the Euclidean distance between two points, X = (x11, x12, ….x1n) and Y = (y11, y12, …, y1n) is:

$$d(X,Y)=\sqrt{\sum_{i=1}^{n}(xi - yi)^2}$$

The unknown sample is assigned the most common class among its k nearest neighbors. When k = 1, the unknown sample is assigned the class of the training sample that is closest to it in pattern space. Classifier assigns equal weight to each attribute. It can also be used for prediction, i.e., to return a real-valued prediction for a given unknown sample.

I. *Conjunctive rule learning:*

Conjunctive Rule algorithm implements a single conjunctive rule learner that can predict for numeric and nominal class labels. It can be easily learnt by finding all commonalities shared by all positive examples.

A rule consists of antecedents "AND"ed together and the consequent (class value) for the classification or regression. In this case, the consequent is the distribution of the available classes (or mean for a numeric value) in the dataset. If the test instance is not covered by this rule, then it's predicted using the default class distributions/value of the data not covered by the rule in the training data.

This learner selects an antecedent by computing the Information Gain of each antecedent and prunes the generated rule using Reduced Error Pruning (REP) or simple pre-pruning based on the number of antecedents[11].

The limitation of this classifier: if a concept does not have a single set of necessary and sufficient conditions conjunctive learning fails.

IV. METHODOLOGY

A. *TOOL USED:*

WEKA Engine:

WEKA is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from our own Java code. WEKA contains tools for data preprocessing, classification, regression, clustering, association rules, and visualization. It is also well suited for developing new machine learning schemes. In our paper we used the WEKA as data mining engine, and made a bridge between the Diabetes Expert System interface and WEKA[12].

B. *Dataset*:

The dataset for the diabetes disease is acquired from UCI Pima Indian diabetes repository[13]. The diabetes dataset includes name of the attribute as well as the explanation of the attributes shown in table I. The dataset contains 768 record samples, each having 8 attributes and one class with two possibilities such as tested positive and tested negative.

TABLE I. DESCRIPTION OF ATTRIBUTES

S. No	**Attributes description**
1	Number of times pregnant
2	Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3	Diastolic blood pressure (mm Hg)
4	Triceps skin fold thickness (mm)
5	2-Hour serum insulin (mu U/ml)
6	Body mass index (kg/m)^2)
7	Diabetes pedigree function
8	Age (years)
9	Class variable (tested positive or tested negative)

V. RESULTS OF CLASSIFICATION ALGORITHMS USING WEKA TOOL

GainRatio calculation:

It is the modification of the information gain that reduces its bias. Gain ratio takes number and size of branches into account when choosing an attribute. It corrects the information gain by taking the intrinsic information (entropy of distribution) of a split into account. It evaluates the worth of an attribute by measuring the gain ratio with respect to the class.

Search Method: Attribute ranking.
Attribute Evaluator (supervised, Class (nominal): 9 class):
Gain Ratio feature evaluator
9 Class: tested_positive or tested_negative

TABLE II. RANKED ATTRIBUTES:

Attribute rank	Attribute nominal	Attribute
0.0986	2	Plas
0.0863	6	Mass
0.0726	8	Age
0.0515	1	Preg
0.0394	5	Insu
0.0226	7	Pedi
0.0224	4	Skin
0.0144	3	Pres

Selected attributes: 2,6,8,1,5,7,4,3 : 8

VI. RESULTS OF DIFFERENT CLASSIFICATION METHODS

A. *BayesNet:*

The results of diabetes dataset for bayes net classifier using WEKA tool is shown in table III.

TABLE III.

Correctly Classified Instances	601	78.25%
Incorrectly Classified Instances	167	21.74 %
Total number of instances	768	

B. *Naïve Bayes:*

The results of diabetes dataset for Naïve Bayes classifier using WEKA tool is shown in table IV.

TABLE IV.

Correctly Classified Instances	586	76.30%
Incorrectly Classified Instances	182	23.69%
Total number of instances	768	

C. *J48 pruned tree*

The results of diabetes dataset for J48 classifier using WEKA tool is shown in table V.

TABLE V.

Correctly Classified Instances	646	84.11%
Incorrectly Classified Instances	122	15.88%
Total number of instances	768	

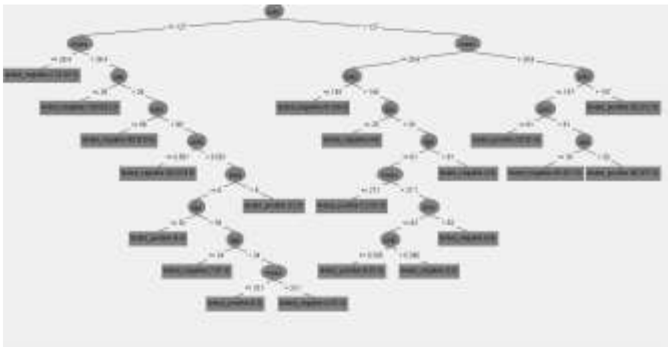

Figure 1. J48 decision tree.

D. *RandomForest*

The results of diabetes dataset for RandomForest classifier using WEKA tool is shown in table VI.

TABLE VI.

Correctly Classified Instances	756	98.43%
Incorrectly Classified Instances	12	1.56 %
Total number of instances	768	

E. *Random tree*

The results of diabetes dataset for bayes net classifier using WEKA tool is shown in table VII.

TABLE VII.

Correctly Classified Instances	768	100 %
Incorrectly Classified Instances	0	0 %
Total number of instances	768	

F. *REP tree*

The results of diabetes dataset for REP tree classifier using WEKA tool is shown in table VIII.

TABLE VIII.

Correctly Classified Instances	638	83.072%
Incorrectly Classified Instances	130	16.92 %
Total number of instances	768	

G. *CART*

The results of diabetes dataset for CART classifier using WEKA tool is shown in table IX.

TABLE IX.

Correctly Classified Instances	593	77.21%
Incorrectly Classified Instances	175	22.78 %
Total number of instances	768	

H. *k-NN*

The results of diabetes dataset for k-NN classifier using WEKA tool is shown in table X.

TABLE X.

Correctly Classified Instances	768	100 %
Incorrectly Classified Instances	0	0 %
Total number of instances	768	

I. *Conjunctive rule learner*

The results of diabetes dataset for conjunctive rule learner classifier using WEKA tool is shown in table XII.
(plas <= 127.5) => class = tested_negative

TABLE XI. CLASS DISTRIBUTIONS

Covered by the rule		Not covered by the rule	
tested_negative	tested_positive	tested_negative	tested_positive
0.810398	0.189602	0.372973	0.627027

TABLE XII.

Correctly Classified Instances	565	73.56 %
Incorrectly Classified Instances	203	26.43 %
Total number of instances	768	

VII. RESULTS

The results of applying different classification algorithms on diabetes dataset is shown in Table XIII, which consists of correctly classified and incorrectly classified instances. The accuracy of the different algorithms represented as shown in figure 2.

TABLE XIII. RESULTS OF DIFFERENT CLASSIFICATION ALGORITHMS.

Algorithms	Correctly classified	Incorrectly classified
Bayesian	78.25%	21.74%
Naïve bayes	76.30%	23.69%
J48	84.11%	15.88%
RandomForest	98.43%	1.565
Random tree	100%	0
REP	83.07%	16.92%
CART	77.21%	22.78%
k-NN	100%	0
Conjunctive rule learning	73.56%	26.43%

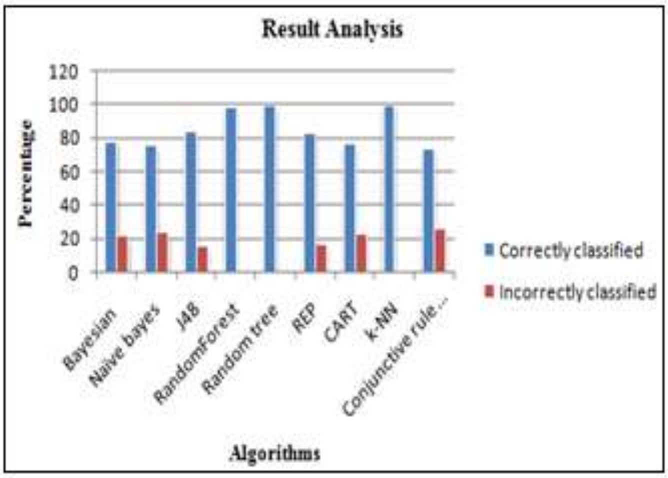

Figure 2. Result analysis of different classification algorithms.

A. *Evaluation Measures:*

Following are the evaluation measures for classification techniques:

a) Sensitivity = $\frac{TP}{TP+FN}$

b) Specificity = $\frac{TN}{FP+TN}$

c) Accuracy = $\frac{TP+FN}{TP+FP+TN+FN}$

d) Positive precision = $\frac{FP}{TP+FP}$

e) Negative precision =$\frac{FN}{TN+FN}$

f) Error rate = $\frac{FP+FN}{TP+FP+TN+FN}$

Where TP: True Positive FP: False Positive
TN: True Negative FN: False Negative

TABLE XIV. DIFFERENT MEASURE CALCULATION FOR DIFFERENT ALGORITHMS.

Sl.no	Classification Methods	Sensitivity	Specificity	Accuracy	Positive precision	Negative precision	Error rate
1	BayesNet	0.6902	0.832	0.3489	0.3122	0.1663	0.2174
2	Naïve Bayes	0.6156	0.842	0.3489	0.3237	0.1965	0.2369
3	J48	0.6641	0.936	0.3489	0.1523	0.1612	0.1588
4	RandomForest	0.9776	0.988	0.3489	0.0223	0.012	0.0156
5	Random tree	1	1	0.3489	0	0	0
6	REP tree	0.7798	0.858	0.3489	0.2535	0.1209	0.1692
7	CART	0.5597	0.886	0.3489	0.2753	0.2103	0.2278
8	k-NN	1	1	0.3489	0	0	0
9	Conjunctive rule	0.6492	0.782	0.3489	0.3851	0.1938	0.2643

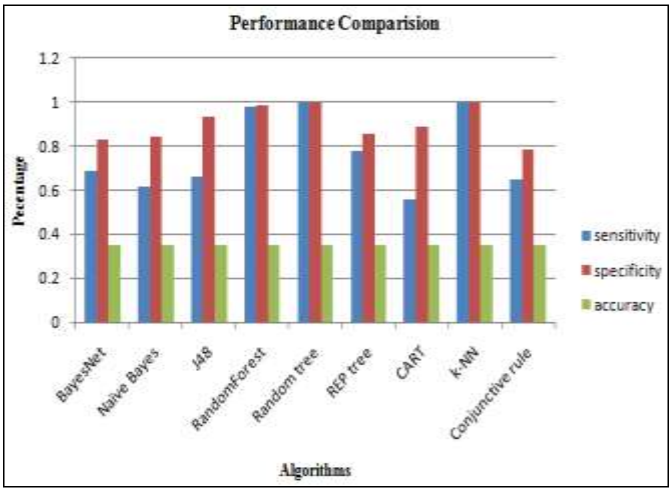

Figure 3. Performance comparison of sensitivity, specificity and accuracy measures of different classifiers.

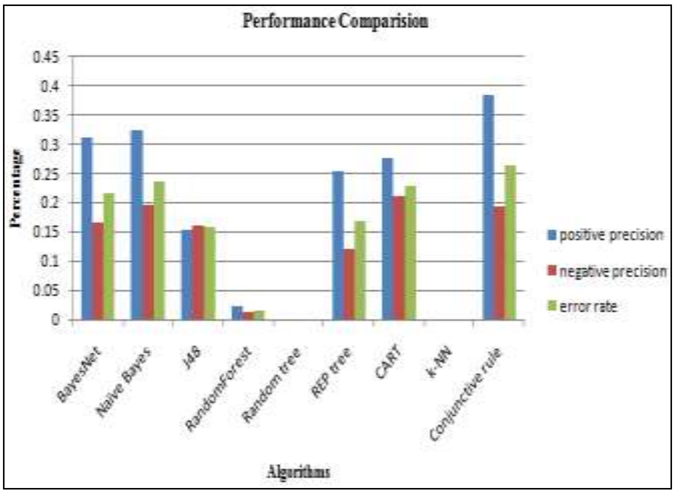

Figure 4. Performance comparison of positive precision, negative precision and error rate measures of different classifiers.

CONCLUSION

The data mining techniques and classification in particular is an interesting topic to the researchers as it accurately and efficiently classifies the data for knowledge discovery. In this paper the frequently used classification techniques namely Bayesian network, Naïve Bayes, J48, REP, RandomForest, Random tree, CART, KNN, Conjunctive rule learning are studied and the experiments are conducted on the diabetes dataset from UCI learning repository to find the best classifier for Diabetes Diagnosis. The performance indicators namely accuracy, specificity, sensitivity, precision, error rate are computed for the above dataset. The results show that the J48 algorithm works the best for the diabetes data set. It helps in identifying the state of the disease and also helps people to identify if they are diabetic, based on the attributes.

REFERENCES

[1] http://en.wikipedia.org/data-mining.

[2] http://www.emedicinehealth.com/diabetes. http://en.wikipedia.org/wiki/Diabetes_mellitus

[3] http://diabetes.co.in.

[4] Han, J., Kamber, M.: Data Mining; Concepts and Techniques, Morgan Kaufmann Publishers (2000).

[5] Klemetinen, M., Mannila, H., Ronkainen, P., Toivonen, H., and Verkamo, A. I (1994) "Finding interesting rules from large sets of discovered association rules," CIKM.

[6] Tsumoto S., (1997)"Automated Discovery of Plausible Rules Based on Rough Sets and Rough Inclusion," Proceedings of the Third Pacific-Asia Conference (PAKDD), Beijing, China, pp 210-219.

[7] Arvind Sharma and P.C. Gupta ―Predicting the Number of Blood Donors through their Age and Blood Group by using Data Mining Tool‖ International Journal of Communication and Computer Technologies Volume 01 – No.6, Issue: 02 September 2012.

[8] P.Yasodha, M. Kannan, "Analysis of a Population of Diabetic Patients Databases in WEKA Tool". International Journal of Scientific & Engineering Research Volume 2, Issue 5, May-2011 ISSN 2229-5518 Analysis of a Population of Diabetic Patients Databases in WEKA Tool

[9] J.R.Quinlan, "c4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, Inc, 1992.

[10] Breiman, Friedman, Olshen, and Stone. "Classification and Regression Trees", Wadsworth, 1984., Mezzovico, Switzerland.

[11] http://www.dbs.informatik.uni-muenchen.de/Lehre/KDD_Praktikum/WEKA/doc/WEKA/classifiers/rules/ConjunctiveRule.html.

[12] Svetlana S. Aksenova" Machine Learning with WEKA Explorer Tutorial for WEKA Version 3.4.3", School of Engineering and Computer Science.

[13] UCI Machine Learning Repository pima Indian diabetes dataset: http://archive.ics.uci.edu.ml/datasets.

[14] K. R. Lakshmi and S.Prem Kumar, "Utilization of Data Mining Techniques for Prediction of Diabetes Disease Survivability", International Journal of Scientific & Engineering Research, Volume 4, Issue 6, June-2013.

[15] K. Rajesh, V. Sangeetha, "Application of Data Mining Methods and Techniques for Diabetes Diagnosis", International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 3, September 2012.