Examining Classification Techniques in Data Mining for PIMA Indian Diabetes Dataset

S.Janani[1] D. RamyaChitra[2]
[1]Research Scholar [2]Assistant Professor
[1,2]Department of Computer Science and Engineering
[1,2]Bharathiar University, Coimbatore

Abstract— Classification techniques have been widely used in the medical field for accurate classification than an individual classifier. This paper presents computational intelligence techniques for Diabetes Patient Classification. This paper evaluates the selected 5 classification algorithms (Naïve Bayes, Multilayer Perception, Decision Table, J48 and FT) for the classification of diabetes patient datasets. The aim of this paper is to investigate the performance of different classification techniques. In this paper we are analyzing the performance of 5 classification algorithms. We use the PIMA Indian diabetes datasets for calculating the performance of classification algorithms by using the training set parameter. And finally a comparative analysis based on the performance factors such as the Classification Accuracy, Error Rate and execution time is performed on all the algorithms.

Keywords: Data Mining, Classification, PIMA Indian Diabetes Dataset

I. INTRODUCTION

Knowledge Discovery in Databases creates the context for developing the tools needed to control the flood of data facing organizations that depend on ever-growing databases of healthcare information. Knowledge Discovery has the preprocessing, Data mining and Post processing phases. KDD is the iterative or cyclic process that involves sequence of steps of processes and data mining is the core component of the KDD process. Data mining is a collection of techniques for efficient automated discovery of previously unknown, valid, novel, useful and understandable patterns in large databases. These patterns must be actionable so that may be used in an enterprise's decision making [1].

The Diabetes is a metabolic disease which reveals itself through hyperglycemia. It appears as a result of insulin deficiency or inefficacy, and chronically progresses. Being a hormone secreted by the pancreas, insulin provides the glucose in blood to be used by tissues. Without insulin, tissues cannot effectually use the nutrients taken in and the glucose in blood increases. This metabolic malfunction leads to modifications in many organs and diabetic individuals have an important place in the health system. The diabetes will probably continue to be the most important disease and death reason with its worldwide increasing number in the future, as well (Harrison İç Hastalıkları, 2004). If the insulin hormone is totally missing, this diabetes is called "Type 1 Diabetes (the diabetes dependent on insulin)" (The Society of Endocrinology and Metabolism of Turkey, 2012). Generally, it is seen in children or patients at young ages. If there is the insulin hormone in the body, but its amount is low or there is resistance to insulin in the tissues, this diabetes is called "Type 2 Diabetes". Type 2 Diabetes is the most prevalent metabolism disease in the adult society.

Type 2 diabetes is a serious global health problem in most countries. Type 2 Diabetes has developed with aging population, increasing urbanization, diet changes, decreasing physical activities, unhealthy life style, behavior patterns and fast cultural and social changes (Pickup, 2003). In this study, many classification algorithms have been implemented on PIMA Diabetes data set by UCI and the performance of this algorithm has been analyzed by the data mining tool WEKA.In this study, 5 different data mining algorithms have been used to classify the PIMA diabetes Data set. The remaining section of the paper is structured as follow Section 2 describes the literature review, Section 3 describes the methodology for the PIMA diabetes dataset and Section 4 describes our experimental result. And finally Section 5 gives the Conclusion and Future work.

II. LITERATURE REVIEW

A Research Paper given by Sudajai Lowanichchai, SaisuneeJabjone, Tidanut Puthasimma, Informatic Program Faculty of Science and Technology Nakhon Ratchsima Rajabhat University it proposed the application Information technology of knowledge-based DSS for analysis diabetes of elder using decision tree. [2].

In another Research paper presented by Yang Guo ,GuohuaBai, Yan Hu School of computing Blekinge Institute of Technology Karlskrona, Sweden, The discovery of knowledge from medical databases is important in order to make effective medical diagnosis. The dataset used was the Pima Indian diabetes dataset. Preprocessing was used to improve the quality of data. Classifier was applied to the modified dataset to construct the Naïve Bayes model. Finally weka was used to do simulation, and the accuracy of the resulting model was 72.3%. [3].

Literature Review on Diabetes, by National Public health: Women tend to be hardest hit by diabetes with 9.6million women having diabetes. This represents 8.8% of the adult population of women 18 years of age and older in 2003 and a two fold increase from 1995 (4.7%).. By 2050, the projected numbers of all persons with diabetes will have increased from 17 million to 29 million. [4]

In the proceeding of Aljaruallah A.A in International Conference on Innovation Technology gives the detailed information about the Type II Diabetes used classification algorithms available in the tool Weka. It explains the concept with a suitable example.[5]

Hongjun Lu,et al(2000)., build an efficient scalable classifiers in the form of decision tables by exploring capabilities of modern relational database management systems. They implemented the unique features of the approach that include its high training speed, linear scalability and simplicity [6].

Kenneth J McGarry, et al., compared the Knowledge Extraction from Radial Basis Function Networks and Multilayer Perceptions. RBF networks are localist types of learning technique Local learning systems generally contain elements that are responsive to only a limited section of the input space [7].

Gaganjot kaur et.al(2014) proposed the new algorithm improved J48 classification algorithm which is used in prediction of diabetes, it shows that the classification algorithm J48 is computed and the performance is improved by a new improved J48 classification algorithm.[8]

Ashok kumar et.al(2013) tells how the classification algorithm is applicable for diabetes dataset and how the performance measure will validate the classification. Finally BayesNet algorithm is suggested for the classification of diabetes dataset.[9]

III. Methodology

Using the three classification algorithms here found the best algorithm for the PIMA Diabetes dataset. The flow diagram for the comparative analysis is shown in Fig 1.

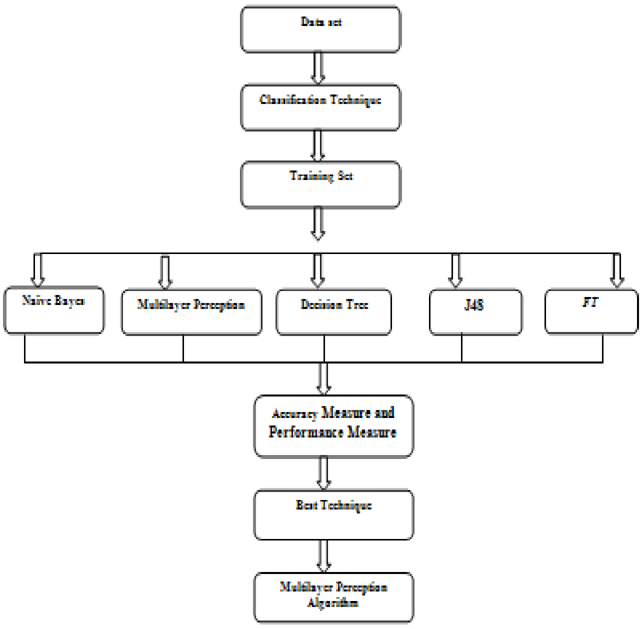

Fig. 1: Process Flow Diagram for Comparative Analysis

A. Data set:

The diabetes data set has been taken from the web site of UCI (UC-Irvine archive of machine learning datasets (UCI Machine Learning Repository, 2012)). The data comprise statistically important information about diabetic patients. The diabetes data set consists of 99 patients and 9 attributes. There are no missing data in the data set. Detailed portray of the data set is seen in Table 1.

S.No	Name of the Attribute	Definition of the Attribute
1.	Number of pregnancy	Numerical values
2.	Plasma glucose concentration	Glucose concentration in the 2nd hour in oral glucose tolerance test
3	Diastolic blood pressure	mm Hg
4	Triceps derma thickness (mm)	Triceps derma thickness
5	Serum insulin in the 2nd hour	Insulin (mu U/ml)
6	Body mass index	(Kg weight/(m height)^2)
7	Diabetes family history	Whether there is the diabetes or not in the family
8	Age	Age
9	Class	1 – test result for the diabetes positive (yes) 0 - test result for the diabetes negative (no)

Table 1:The Diabetes Data Set

B. Classification:

In the Data mining, the classification technique can be used to predict group membership for data instances. The classification is similar to the clustering technique, and in that it also sectors the customer records into distinct sector called classes. In order to predict the outcome of the datasets, the algorithm processes a training set containing a set of attributes and the respective outcome, usually called target or prediction attribute. In this paper we have analysed the classification algorithms to predict which algorithm is suitable for the PIMA Diabetes data set. In the classification the comparative of five algorithms shows which one fitted effectively for the PIMA diabetes dataset.

C. Algorithms:

1) Naive Bayes:

Bayes classifiers are statistical classifiers. Bayes predicts the membership probabilities of the data, that is, their probability about belonging to a specific category. Bayes classifier is based on the Bayes theorem explained below:

A sample in a data set is composed of the input values X = {x1,x2,xm}. If it is pretended that the total number of the categories is m, the probability calculations are done with Equation 1 for the sample whose category is to be determined.

2) Multilayer Perception:

MLP (Multilayer Perception) is an artificial neural network model which is mostly used and learns best [7]. It is known that Multilayer Perception algorithm has a very strong function in classifying prediction problems [7]. The purpose of this model is to minimize the difference between the target result (output) of the network and the attained result. This model is expressed as propagation algorithm since it makes the mistake spreading it over the network or as back-propagation since it uses back-propagation learning algorithm in the process of that the multilayer perception is getting trained.

3) Decision Table:

Decision Table algorithm classifier summarizes the dataset with a decision table which contains the same number of attributes as the original dataset. Then, a new data item is assigned a category by finding the line in the decision table that matches the non-class values of the data item. By eliminating attributes that contribute little or nothing to a test model of the dataset, the algorithm reduces the likelihood of over-fitting and creates a smaller and condensed decision table [6].

4) J48:

Decision Tree Algorithm is to find out the way the attributes-vector behaves for a number of instances. Also on the bases of the training instances the classes for the newly generated instances are being found. J48 is an extension of ID3. The additional features of J48 are accounting for missing values, decision trees pruning, continuous attribute value ranges, derivation of rules, etc. In other algorithms the classification is performed recursively till every single leaf is pure, that is the classification of the data should be as perfect as possible [8].

5) FT

FT combines a standard univariate DT, such as C4.5, with linear functions of the attributes by means of linear regressions. While a univariate DT uses simple value tests on single attributes in a node, FT can use linear combinations of different attributes in a node or in a leaf. In the constructive phase a function is built and mapped to new attributes. A model is built using the constructor function. This is done using only the examples that fall at this node. Later, the model is mapped to new attributes.[10]

IV. Experimental Results

In this paper the experimental measures are measured by using the performance factors such as the classification accuracy, performance measures and error rate to determine the best algorithm for the PIMA Diabetes data set. The accuracy measure, performance measure and the error rate measure for the chosen classification algorithms is shown in Table 2 – Table 4. And its Comparison result is shown in Fig 2 – Fig 4.

Algorithm	Correctly Classified(% Value)	Incorrectly Classified(%Value)
Naïve Bayes	70.7071	29.2929
Multilayer Perception	95.9596	4.0404
Decision Table	74.7475	25.2525
J48	72.7273	27.2727
Ft	76.7677	23.2323

Table 2:Accuracy Measure for classification Algorithm

From the experimental results (table 2), it is inferred that for the training set parameter using PIMA Diabetes data set, the Multilayer Perception algorithm gives the more correctly classified instances compared to other algorithm. The Comparison of accuracy measure for classification algorithm is shown in Fig2.

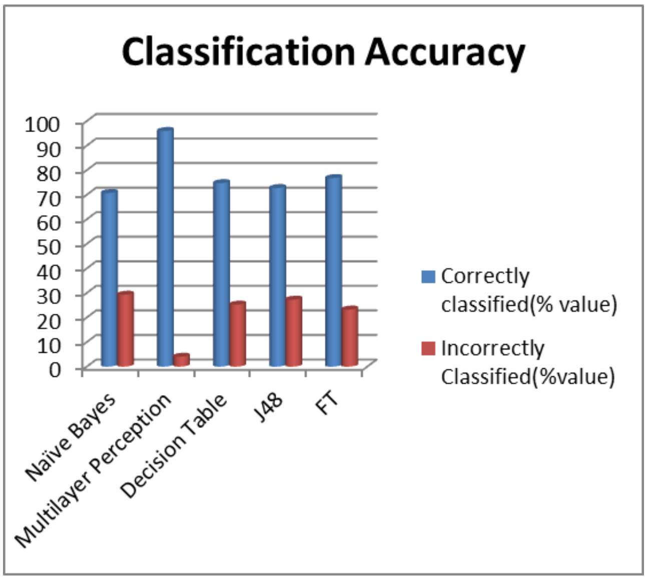

Fig. 2: Comparison of accuracy measure for classification algorithm

ALGORITHM	TP	FP	PRECISION	RECALL	F-MEASURE	ROC AREA	KAPPA STATISTIC
Naïve Bayes	0.707	0.37	0.7	0.707	0.702	0.709	0.3476
Multilayer Perception	0.96	0.047	0.96	0.96	0.96	0.822	0.9127
Decision Table	0.747	0.299	0.746	0.747	0.747	0.655	0.4511
J48	0.727	0.465	0.779	0.727	0.677	0.727	0.3077
FT	0.768	0.383	0.794	0.768	0.741	0.572	0.434

Table 3:Performance Measure for Classification algorithm

From the experimental results (Table 3), it is inferred that for the training set parameter by using PIMA Diabetes dataset for Multilayer Perception algorithm, the performance measures like the TP Rate, Precision, F-Measure, Kappa values and ROC increases and the FP rate decreases. The comparison of performance measures for classification algorithm is shown in Fig3.

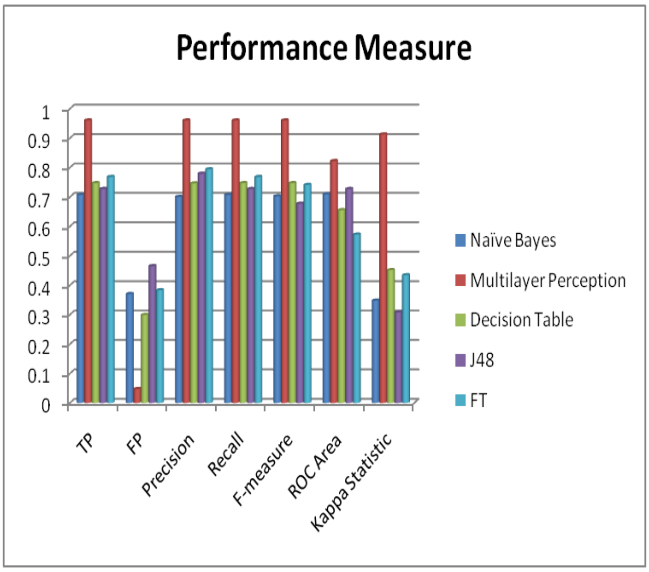

Fig. 3: Comparison of performance measure for classification algorithm

ALGORITHM	MA	RMSE	RAE(%)	RRSE(%)
Naïve Bayes	0.3016	0.4612	65.0655	95.8626
Multilayer Perception	0.1097	0.2174	23.6628	45.1983
Decision Table	0.3437	0.4095	74.1516	85.1191
J48	0.3884	0.4407	83.7952	91.6111
FT	0.2323	0.482	50.1187	100.1966

Table 4: Error rate measure for classification algorithm

From the experimental results (Table 4), it is inferred that for the training set parameter by using PIMA Diabetes dataset for Multilayer Perception algorithm, the error measures like MA, RMSE, RAE and RRSE decreases. The comparison of Error rate measures for classification algorithm is shown in Fig4

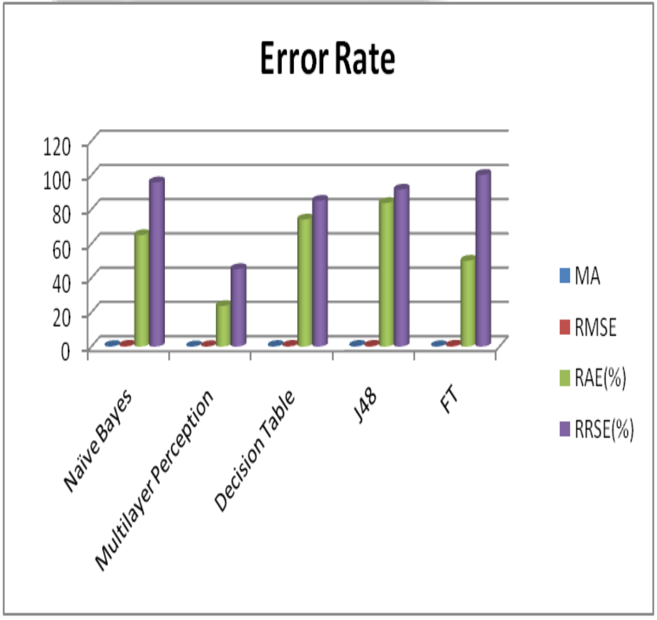

Fig. 4: Comparison of error rate measure for classification algorithm

The analysis results shows from table2 - table 4 stated clearly that the Multilayer Perception Algorithm performs better in classifying the PIMA Diabetes data set. The Algorithm which classifies the data set 95.95%correctly and also it has the least Error rate for all the error measure values. And it reaches the maximum Kappa value 0.9127. The analyses of the classification algorithm shows that Multilayer Perception algorithm is performing well for the PIMA Diabetes Data set.

V. Conclusion

The Algorithm with highest accuracy value and the lowest Error rate is declared as the best algorithm. This paper presents the performance of 5 different classification algorithms like Naïve Bayes, Multilayer Perception, Decision Table, J48 and FT on the PIMA diabetes dataset which is downloaded from the UCI repository. The paper shows the analysis of those algorithm and comparison finally says clearly that the Multilayer Perception algorithm performs well than other algorithm for this dataset. Hence it is proved that Multilayer Perception Algorithm performs better for the PIMA Diabetes Dataset. In future an approach that will be used for hybrid model construction of community health services. These classification algorithms can be implemented for other dominant diseases prediction and classification. An another scope is to seeing whether by applying new algorithms will made any improvements over techniques which are used in this paper in future.

Reference

[1] Fayyad, U, Data Mining and Knowledge Discovery in Databases: Implications for scientific databases, Proc. of the 9th Int. Conf. on Scientific and Statistical Database Management, Olympia, Washington, USA, 2-11, 1997.

[2] SudajaiLowanichchai, SaisuneeJabjone, TidanutPuthasimma, ”Knowledge-based DSS for an Analysis Diabetes of Elder using Decision Tree”

[3] Yang Guo , GuohuaBai , Yan Hu School of computing Blekinge Institute of Technology Karlskrona, Sweden, “Using Bayes Network for Prediction of Type-2 Diabetes”

[4] Beckles GLA, Thompson-Reid PE, editors. Diabetes and Women’s Health Across the Life Stages: A Public Health Perspective. Atlanta: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion,

[5] AlJarullah, AA., “Decision Tree Discovery for the Diagnosis of Type II Diabetes”, International Conference on Innovations in Information Technology 2011

[6] Hongjun Lu and Hongyan Liu,” Decision Tables: Scalable Classification Exploring RDBMS Capabilities”,Proceedings of the 26th International Conference on Very Large Databases, Cairo, Egypt, 2000.

[7] Kenneth J.McGarry,Stefan wermter and John MacIntyre,” Knowledge Extraction from Radial Basis Function Networks and Multi_layer Perceptrons”, Neural Networks, 1999. IJCNN '99. International Joint Conference on (Volume:4) ISSN :1098-757DOI: 10.1109/IJCNN.1999.833464 Publisher:IEEE

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

[9] Dr.D. Ashok kumar et.al “Performance Evaluation of Classification Data mining Techniques in Diabetes” IJCSIT Vol 6(2)2015

[10] TrilokChand Sharma et.al “Weka Approach for Comparitive Study of Classification Algorithm”IJARCCE Vol2,Issue 4 2013