International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 6, Issue 1, No 01, 2018 ISSN: 2348-6600
Reference ID: IJCS-330 PAGE NO:2190-2195

Comparative Analysis of Various Data Mining Classification Algorithms Using R Software

R.Palanisamy[#1], Dr.S.S.Dhenakaran[*2]

[#]*M.Phil Research Scholar, Deparment of Computer Applications*
Alagappa University, Karaikudi
[1]unifypalani@gmail.com

[*]*Professor, Department of Computer Science*
Alagappa University
[2]ssdarvind@yahoo.com

***Abstract*— Data mining is the process of sorting the collection of stored data to identify/discover the patterns using some technical analysis. Evaluate the probability of future events data mining has more algorithms to segment and predict the data sets. Classification is one of the important mechanisms functioning by data mining and main goal of classification is to predict the target data accurately. For prediction numerous classification techniques have currently working, in this paper few of them are classified and compared with different data sets. Classification algorithms such as BayesNet, DT, J48, Logistic, Naïve Bayes, NBT, PART, RBFN are implemented and compared using R software with different test data.**

***Index Terms*— Data mining, R classification, Classification comparison, Prediction, Accuracy, Quality Metrics.**

I. INTRODUCTION

Data mining is the mechanism of identify the particular data sets from the large set of data with the help of methods includes database systems, statistics and machine learning. The characteristic of data mining is analysis of Knowledge Discovery Database process. Data mining is working with decision support system to perform accurate perdition of multiple groups in the data. Before mining of data, the initial preprocessing works should be done it includes data collection, data preparation and result interpretation. Data mining process is divided into four sections as data cleaning, data integration, data selection and data transformation.

Required classification techniques available in data mining which includes machine learning based approach, Neural Network, statistical procedure based approach. Some of the classification algorithms are explained briefly here.

A. K-Nearest Neighbors Algorithm

K-nearest neighbor is a pattern recognition algorithm specially used for regression and classification. The working principle of k-NN is the result objects should be classified from the majority choice of its neighbors. The output result is a class membership in the sense of k-NN as a classifier; k-NN is used as a regression the output is the property value for the object [1]. The properties of k-NN classifier are working with two aspects named as 1-nearest neighbor classifier and weighted nearest neighbor classifier. K-NN assigns weights based on the nearer objects using weighted nearest neighbor algorithm. Compare than the naive method, k-NN is easily tractable by computationally even for large data sets.

B. Naive Bayes Algorithm

Naïve Bayes classifier comes under the family of probabilistic classifier works with the help of bayes theorem. Naïve bayes classifiers are highly scalable [2]. Naïve Bayes algorithm works for both binary class and multiclass problems. Two kinds of probabilities are taken over here such as class probabilities and conditional probabilities. From the training dataset, the probability of each class is considered as a class probability, the class value is given from the conditional probabilities of each input.

C. ANN Algorithm

Artificial Neural Network is a collection of connection nodes of many artificial neurons; it is a computational structure or function of biological neural networks. ANN works for such areas are fault detection, speech recognition, product inspection, machine translation, social network filtering etc [3]. Neural network lead its process in three sectors are input layer, output layer and hidden layer. The activity of input layer is to collect raw inputs which are feed input network. Hidden unit represents to determine the

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 6, Issue 1, No 01, 2018 ISSN: 2348-6600
Reference ID: IJCS-330 PAGE NO:2190-2195

activity of each and every hidden unit. Output unit is entirely based on the hidden units and weight between hidden and output unit.

II. RELATED RESEARCH ON CLASSIFICATION

Classification technique is based on the inductive learning principle that analyzes and finds the patterns from the database. If the nature of an environment is dynamic, then the model must be adaptive i.e. it should be able to learn and map efficiently. Limère et al. (2004) presented a model for firm growth with decision tree induction principle [4]. It gives interesting results and fits the model to economic data like growth competence and resources, growth potential and growth ambitions. Hoi et al. (2006) developed a novel framework of learning the unified kernel machines for both labeled and unlabeled data. This framework includes semi supervised learning, supervised learning and active learning. Also, a spectral kernel is proposed, where it classifies the given labeled data and unlabeled data efficiently.

Xu et al. (2008) proposed a reproducing kernel Hilbert space framework for information theoretic learning [5]. The framework uses the symmetric nonnegative definite kernel function i.e. cross-information potential. Though this framework gives better result than the previous RKHS frameworks, still there is an issue to choose an appropriate kernel function for a particular domain. Shilton and Palaniswami (2008) defined a unified approach to support vector machines. This unified approach is formulated for binary classification and later on extended to one-class classification and regression.

Kumar et al. (2012) explored a binary classification framework for two stage multiple kernel learning [6]. The distinct advantage of this binary classification framework is that it is easier to leverage research in binary classification and to develop scalable and robust kernel based algorithms. Takeda et al. (2012) proposed a unified robust classification model that optimizes the existing classification models like SVM, minimax probability machine and fisher discriminant analysis. It provides several benefits like well-defined theoretical results, extends the existing techniques and clarifies relationships among existing models

Yee and Haykin (2013) viewed the pattern classification as an ill-posed problem [7], it is a prerequisite to develop a unified theoretical framework that classifies and solves the ill posed problems. Recent literature on classification framework has reported better results for binary class datasets alone. For multiclass datasets, there is a lack in accuracy and robustness. So, developing an efficient classification framework for multiclass datasets is still an open research problem.

III. DETAILS OF EXPERIMENTAL DATA

Data sets are collected from the open source of concern websites, which are listed as its name of data sets from banking customers, vehicle data sets, credit card data sets of various customers, diabetes data sets of individual patients. The details of data sets are described in the table below:

TABLE 1

DATA SET DESCRIPTION

Data Set	Attributes Count	Attributes	Data item count	Class
Bank	11	Age, Sex, Region, Income, Married, Children, Car, Save-Account, Current-Account, Mortgage, Pep	600	2
Car	6	Buying Capacity, Maintenance, Number Of Doors, Person Seating Capacity, Luggage Boot Space, Safety And Class	1728	4
Credit card	20	Checking Account, Month, Credit History, Purpose, Amount, Saving Account, Present Employment, Rate, Sex, Residence, Property, Age, Other Installments, Housing, Dependent, Telephone, Foreign Worker	1000	2
Diabetes	9	Number of times pregnant, plasma glucose concentration, blood pressure, triceps skin fold thickness, serum insulin, body mass index, Diabetes pedigree function, age	768	2

To classify theses collected data, nine distinct data mining classification algorithms are applied which are named as DT, BayesNet, Logistic, J48, NBT, Naive Bayes, PART, NBT, SMO, RBFN. R tool used here to test analytical performance

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 6, Issue 1, No 01, 2018 ISSN: 2348-6600
Reference ID: IJCS-330 PAGE NO:2190-2195

of these algorithms and produce exact results based on the data sets. These nine classification methods are tested using some parameters such as KS (Kappa Statistics), RMSE (Root Mean Squared Error, MAE (Mean Absolute Error). Kappa Statistics is used to classify and measure the data of observed accuracy with an expected accuracy. Root Mean Squared Error is used to predict the difference between observed and predicted values. Mean Absolute Error has provide average of the absolute errors, it is a measurement of difference between two continuous variables [8].

IV. RESULTS AND DISCUSSIONS

A. Classification on Bank dataset

Classification algorithms are tabulated in below table which is used for bank dataset measurement purpose using R tool. These measurements have compared with algorithms and also KS, MAE, RMSE. Among these algorithms, j48, BayesNet, SMO algorithms are having highest kappa statistics, MAE, RMSE respectively in bank dataset classification.

TABLE 2

CLASSIFICATION RESULT ON BANK DATASET

Algorithm	KS	MAE	RMSE
BayesNet	0.386	**0.397**	0.449
DT	0.612	0.299	0.36
J48	**0.819**	0.156	0.29
Logistic	0.452	0.361	0.43
Naive Bayes	0.37	0.377	0.44
NBT	0.771	0.177	0.319
PART	0.7	0.18	0.357
RBFN	0.459	0.359	0.432
SMO	0.406	0.292	**0.54**

Following Graphical Chart shows the representation of bank dataset classification using R software. Classification algorithms are compared and fluctuation highest results displayed. In this graphical chart different colors are displayed the each classification algorithm.

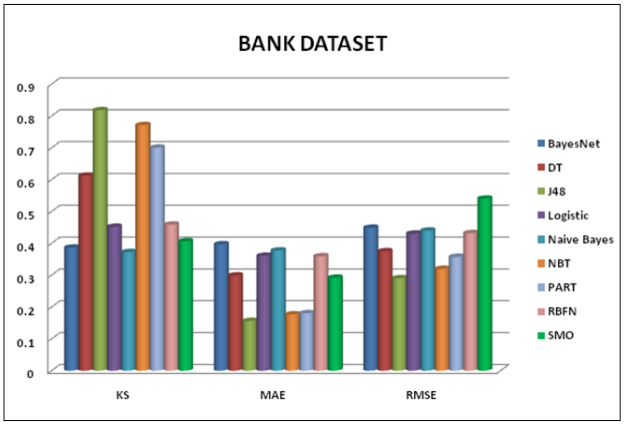

Fig. 1 Bank Data Set Classification Result

B. Classification on Car dataset

Classification algorithms are tabulated in below table which is used for car dataset measurement purpose using R tool. These measurements have compared with algorithms and also KS, MAE, RMSE. Among these algorithms, NBT, DT, SMO algorithms are having highest kappa statistics, MAE, RMSE respectively in bank dataset classification.

TABLE 3

CLASSIFICATION RESULT ON CAR DATASET

Algorithm	KS	MAE	RMSE
BayesNet	0.671	0.111	0.225
DT	0.799	**0.275**	0.322
J48	0.834	0.042	0.172
Logistic	0.85	0.043	0.152
Naive Bayes	0.667	0.114	0.224
NBT	**0.875**	0.068	0.157
PART	0.909	0.024	0.128
RBFN	0.874	0.068	0.157
SMO	0.865	0.256	**0.32**

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 6, Issue 1, No 01, 2018 ISSN: 2348-6600
Reference ID: IJCS-330 PAGE NO:2190-2195

Following Graphical Chart shows the representation of car dataset classification using R software. Classification algorithms are compared and fluctuation highest results displayed. In this graphical chart different colors are displayed the each classification algorithm.

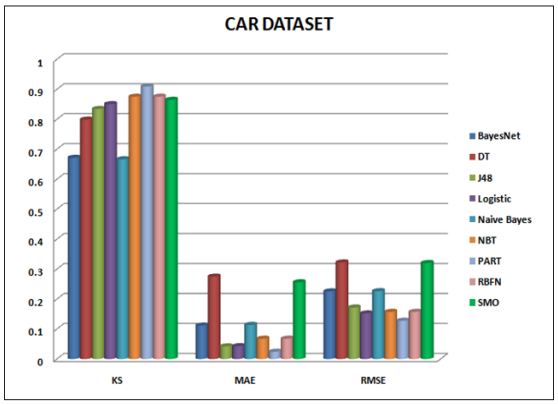

Fig 2. Car Data Set Classification Result

C. Classification on Credit card Dataset

Classification algorithms are tabulated in below table which is used for credit card dataset measurement purpose using R tool. These measurements have compared with algorithms and also KS, MAE, RMSE. Among these algorithms, NBT, DT, SMO algorithms are having highest kappa statistics, MAE, RMSE respectively in bank dataset classification.

TABLE 4

CLASSIFICATION RESULT ON CREDIT CARD DATASET

Algorithm	KS	MAE	RMSE
BayesNet	0.389	0.31	0.419
DT	0.203	**0.369**	0.432
J48	0.247	0.347	0.48
Logistic	0.375	0.31	0.409
Naive Bayes	0.381	0.294	0.42
NBT	**0.392**	0.31	0.422
PART	0.277	0.325	0.497
RBFN	0.334	0.339	0.42
SMO	0.365	0.249	0.499

Following Graphical Chart shows the representation of credit card dataset classification using R software. Classification algorithms are compared and fluctuation highest results displayed. In this graphical chart different colors are displayed the each classification algorithm.

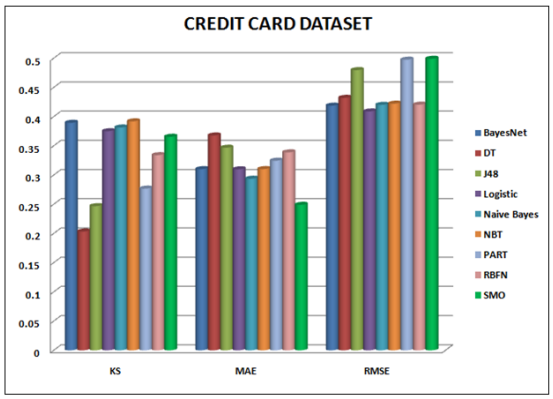

Fig 3. Credit Card Data Set Classification Result

D. Classification on Diabetes dataset

Classification algorithms are tabulated in below table which is used for Diabetes dataset measurement purpose using R tool. These measurements have compared with algorithms and also KS, MAE, RMSE. Among these algorithms, Logistic, RBFN, SMO algorithms are having highest kappa statistics, MAE, RMSE respectively in bank dataset classification.

TABLE 5

CLASSIFICATION RESULT ON DIABETES DATASET

Algorithm	KS	MAE	RMSE
BayesNet	0.429	0.299	0.421
DT	0.349	0.344	0.428
J48	0.416	0.316	0.446
Logistic	**0.473**	0.309	0.395
Naive Bayes	0.466	0.284	0.417
NBT	0.426	0.31	0.428
PART	0.439	0.31	0.415
RBFN	0.43	0.345	0.419
SMO	0.468	0.227	0.476

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 6, Issue 1, No 01, 2018 ISSN: 2348-6600
Reference ID: IJCS-330 PAGE NO:2190-2195

Following Graphical Chart shows the representation of credit Diabetes dataset classification using R software. Classification algorithms are compared and fluctuation highest results displayed. In this graphical chart different colors are displayed the each classification algorithm.

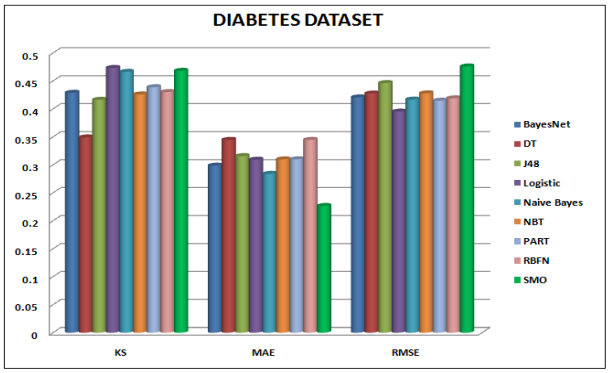

Fig 4. Diabetes Data Set Classification Result

V. COMPARATIVE ANALYSIS

The classification technique has compared with each other using R software. BayesNet, DT, J48, Logistic, Naive Bayes, NBT, PART, RBFN, SMO algorithms are taken for the experimental result with variety of datasets like bank, car, credit card, diabetes. Each classification algorithm provides various results based on the algorithm performance. Correctly classified index values are tabulated.

TABLE 6

CORRECTLY CLASSIFIED INDEX (CCI) COMPARISON

Algorithm	Bank	Car	Credit Card	Diabetes
BayesNet	71.12	86.76	**76.52**	75.42
DT	81.91	92.13	72.13	72.13
J48	**92.3**	93.3	71.5	74.9
Logistic	74.4	94.13	76.22	**78.24**
Naive Bayes	70.1	86.63	76.41	77.32
NBT	89.72	95.25	**76.52**	75.4
PART	86.19	**96.78**	71.24	76.29
RBFN	74.37	95.23	75.2	76.18
SMO	71.8	94.76	76.14	**78.36**

Following Graphical Chart shows the entire representation of Bank, Car and Credit card Diabetes dataset classification using R software. Classification algorithms are compared and fluctuation highest results displayed. In this graphical chart different colors are displayed the each classification algorithm.

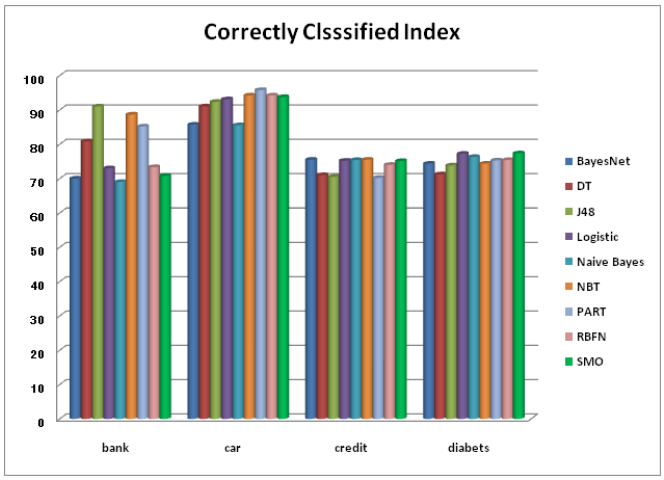

Fig 5. Correctly Classified Index (CCI) Value Comparison Chart.

VI. CONCLUSION

From the business point of view, classification has an important role to classify the data and predict exact results among large data sets. In this paper, SMO, RBFN, PART, Naive Bayes, NBT, J48,Logistic, BayesNet, DT classification algorithms are compared with four(bank, car, Credit card, Diabetes) data sets using R tool. BayesNet and NBT have highly corrected classified results for Credit card data set. For

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 6, Issue 1, No 01, 2018 ISSN: 2348-6600
Reference ID: IJCS-330 PAGE NO:2190-2195

Bank data set J48 has a highest score than the remaining algorithms. For car datasets PART has highest classified data, for credit card data set BayesNet and NBT working well for better classification. From this comparison, no one can provide the better classification for every data set. Every classification algorithms are needed to fulfill classification and prediction of different distinct data sets.

REFERENCES

[1] Available at: http://en.wikipedia.org/wiki/K-nearest_neighbors algorithm.
[2] Available at:https://en.wikipedia.org/wiki/Naive_Bayes_classifier
[3] Saravanan, Sasithra ”Review on Classification Based on Artificial Neural Networks” International journal Ambient Systems and Applications vol 2, no 4, 2014 p.p 11-18.
[4] Limère et al, “A classification model for firm growth on the basis of ambitions, external potential and resources by means of decision tree induction”, Working Papers2004027, University of Antwerp, Faculty of Applied Economics.
[5] Xu et al, “A Reproducing Kernel Hilbert Space Framework for Information-Theoretic Learning”, IEEE TRANSACTIONS ON SIGNAL PROCESSING, VOL. 56, NO. 12, DECEMBER 2008, 5891 – 5902.
[6] Kumar et al, “A Binary Classification Framework for Two-Stage Multiple Kernel Learning”, Appearing in Proceedings of the 29 th International Conference on Machine Learning, Edinburgh, Scotland, UK, 2012.
[7] Andrew Secker, Matthew N. Davies *et al*., “An Experimental Comparison of Classification Algorithms for the Hierarchical Prediction of Protein Function”, Expert Update (the BCS-SGAI) Magazine, **9**(3), 17-22, (2007).
[8] Available at: https://en.wikipedia.org/wiki/Mean_absolute_error.