International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org, editorijettcs@gmail.com
Volume 9, Issue 1, January - February 2020 ISSN 2278-6856

Comparative Analysis of Classification Techniques in Data Mining

V.Saranya[1] and A.Vigneswari[2]

[1]Assistant Professor, Department of Computer science,
Karur Velalar College of Arts and Science for Women, Karur

[2]Associate Professor, Department of Computer science,
Karur Velalar College of Arts and Science for Women, Karur

Abstract: *Nowadays, an enormous library of Data Mining techniques has been extended to carry out a stacks of trouble in fields such as medical imaging, sales, business administration, marketing and traffic analysis, manufacturing process astronomy and etc. Currently, Data Mining had a major force on the information industry, due to the broad availability of immense datasets. Classification technique is one kind of generally applied process of data mining in healthcare. Classification is frequently used in marketing, surveillance, fraud detection and scientific discovery. This paper compared the a few classification algorithm gives the best result. The researchers applied a variety of classification algorithms such as K-Nearest Neighbour classifiers, decision Tree, Bayesian Network, Support Vector Machine and Artificial Neural Networks. This paper presents the comparative analysis on different classification algorithms such as Naive Bayes, IBK, Decision Tree and J48. The experimental result shows that the Naïve Bayes classification algorithm gives high classification accuracy than the rest of the algorithms. These algorithms are evaluated by precision, f-measures, recall, TP Rate and FP Rate.*

Keywords: Data mining, Classification, Decision Tree, and Bayesian Network.

1. INTRODUCTION

Data mining is an influential novel technology improved and so fast grown. It is a technology used with enormous potential to help business and companies target on the most significant information of the data. Data mining can be a method of extracting useful pattern or information and relationships within enormous amounts data. The term data processing also referred as “Knowledge mining from data”. The universal goal of the data mining method is to extract information from information set and associated it into a comprehensive construction for future use. Data mining be a multidisciplinary field of research that combines database technology, machine learning, artificial intelligence, knowledge engineering, statistical research, object-oriented methods, Information retrieval, highly performance computing and data visualization of the recent technology [1].

Data Classification is accomplished of processing a wider variety of data than regression and is upward in popularity. Classification techniques in data mining are able to processing a t amount of data. It can predict categorical class labels and classifies data based on training set and class labels and that can be applied for classifying recently obtainable data. Subsequently it can be outlined as a predictable part of data mining and is gaining more popularity. Classification techniques applied in many applications such as machine earning, statistics, database system and artificial intelligence [2].

2. REVIEW OF LITERATURE

Alexander Statnikov et al [3] had compared the classification algorithms namely Random Forest and Support Vector Machine(SVM) algorithm for 22 cancer diagnostic and prognostic datasets. An experiment showed that the SVM classification algorithm gives better performance than Random Forest algorithm. **Hem Jyotsna Parashar et al [4]** proposed a new method for classification using Decision Tree algorithm that proposed method creates decision tree and extract rules for classification. It improves the quality and classify data more accurately. It gives better result than ID3 and C4.5. **Sathya Devi et al [5]** presented to evaluate the performances in terms of classification accuracy of AD Tree, NB Tree and AdaBoost and LogitBoost algorithms using various accuracy measures like FP Tree, TP rate, Recall, Precision and F-Measure. An experimental result showed that the highest accuracy is found in AD Tree 92.6% and 91% accuracy is found in NB Tree, 86% accuracy in LogitBoost algorithm and 83.33% accuracy is found in AdaBoost algorithm. From this classification results, the performance of AD Tree is better than the other algorithm. **Bendi Venkata Ramana et al [6]** evaluated the selected classification techniques for the classification of some liver patient datasets. The classification algorithms are Naive Bayes classifier, C4.5, Back Propagation, Neural Network and Support Vector Machines. These algorithms are evaluated based on four criteria like accuracy, precision, sensitivity and specificity. From the experimental results, KNN, Back Propagation and SVM are produce better results than the Naive Byaes classifier and C4.5. **Sweety Maniar et al [7]** analyzed the performance of various algorithms based on the classification accuracy. The classification algorithms namely Classification, Regression Tree (CART), Neural Network, Naive Bayes (NB), Decision Tree (DT) K-Nearest Neighbor are evaluated. By comparing these classification algorithms, Neural Network results showed better performance among the other methods and it gives the best classification accuracy. **Kalaiselvi et al [8]** analyzed the performance of the various classification algorithms such as Bagging, Dagging, Decorate, Multiclassifier and MultiboostAB are compared. From the experimental results, the Robot Navigation datasets are used and the classification time and accuracy is calculated by 10-fold validation method. Bagging algorithm is the best classification algorithm to finding the accuracy than other algorithms. **Anita Ganpanti et al [9]** evaluated the performance of three kinds of Meta classification algorithms nalely END, Bagging and Dagging. Comparing these three algorithms are based on the performance factors namely classification accuracy and error rate. From the experimental results, it observed that END algorithm performs better than other algorithms. **R.Porkodi [10]** compared the classification algorithms are aive Bayes, K-Nearest Neighbor, CN2, Support Vector Machine (SVM) and Random Forest. For Lung Cancer dataset, Random Forest algorithm outperforms better than the remaining algorithms. In the outset, the classification algorithms of K-Nearest Neighbor, CN2, Naïve Bayes and Random Forest gives better performance and the Support Vector Machine algorithm obtained poor result for this data set.

Payal pathway et al [11] had compared the classification algorithms like MultiLayer Perception (MLP), Aggregating One Dependence Estimators (AODE) and Naïve Bayes classifier a three datsets namely PTP 1B INHIBITORS, Selective Inhibitors and Drug dataset. The classification algorithms like AODE and Naïve Bayes taken minimum time for classification. From the experimental results, the MLP algorithm is more accurate than other algorithms based on the performance factor such as precision and recall. **Sudhamathy et al [12]** presented the three decision tree classification methods like rpart, ctree and randomforest are compared. This comparative analysis based on performance measure precision. From this, the random forest gives better performance for cancer datasets. **Pardeep Kumar et al [13]** had compared the different classification algorithms namely one neural network (Back Propagation), Three Decision tree algorithms (CHAID, QUEST and C4.5), one statistical (Logistic regression) and one support vector machine (AdaboostM1-SVM and LibSVM) with and without boosting and one clustering algorithm (k-means). These classification algorithms are performed with four datasets with various repositories such as Mushroom, Vote, Nursery and Credit. Compared these algorithms based on performance factors such as predictive accuracy, error rate, training time, classification index and comprehensibility. Finally concluded that Genetic algorithm is the first preference when predictive accuracy and comprehensibility are the selection criterion and decision tree(C 5.0) is the first preference when training time is a selection criterion. SVM is the first preference in terms of training time and predictive accuracy. **Kishansingh Rajput et al [14]** examined about comparison of various classification techniques in data mining such as Support Vector Machine, Decision Tree, K-Nearest Neighbor, Apriori algorithm. These algorithms compared based on their performance factors like kappa statistics, time taken, correctly and incorrectly classified instances, root mean squared error, and so on. After comparison concluded that all four of them have their own advantages and disadvantages and they can best be applied in various situations. **Jianhua Shao et al [15]** proposed a CARM based method for finding range-based rules from numerical data in order to build classification and characterization models. The experimental results showed that the CARM based method outperformed than other rule mining methods such as C4.5 and RIPPLE. **Meraj Nabi et al [16]** evaluated performance of Naive Bayes, Logistics Regression, Decision Tree and Random Forest classification algorithm using Pima Indian Diabetes dataset. The Logistic Regression produces higher accuracy than the Naive Bayes, J48, Decision Tree and Random Forest. The accuracy measurements compared in terms of MAE(Mean Absolute Error), RMAE(Root Mean Absolute Error) and confusion matrices. **Rafael Jiménez et al [17]** compared the classical machine learning techniques such as Decision Tree, Artificial Neural Network and modern statistical techniques such as K-Nearest Neighbor, Naive Bayes used in drug related context and to examine the frequent reasons for high school students use drugs and the reason differ from the type of substance used. The result showed that all the analyzed techniques alcohol and tobacco contains the lower percentage of correct classifications concerned the better predictors. The classification technique analyzed substance use risk and predictive factors. The friends use and pleasant activities are the main motives in order to distinguish between adolescent substance users and not users.

3 CLASSIFICATION ALGORITHMS

A Classification Algorithm is a procedure for choosing an assumption from a collection of alternatives that best fits a set of observations. The classification algorithm contains five types of algorithms such as Decision Tree, Bayesian Network, K-Nearest neighbour, Support Vector Machine and Artificial Neural Networks.

3.1 Decision Tree

Decision Tree is the one of the classification algorithm in data mining that builds classification in the form of tree structure. The result of a tree contains two types of nodes like decision nodes and leaf nodes. Decision tree helps to implement complex decision into easy process and the complex decision is subdivided into simple decision. A decision tree contains methods namely ID3, CART, C4.5, C5.0 and J48 [18].

3.2 Bayesian Network

A Bayesian Network is a graphical form for probability interaction among a set of variables characteristics. The approach applied in Naïve Bayes classifier is very simple. With the facility of little amount of training data, it is possible to classify the instances. The naive bays classifier is statistical algorithm providing astonishingly higher results [19].

3.3 K Nearest Neighbour

KNN algorithm is based on similarity measure and used to store all accessible cases that used to identify the unknown data point based on the nearest neighbor. It is an easy to implement classification technique and training is very fast. KNN is particularly well suited for multimodal classes.

3.4 Support Vector Machine

Support Vector Machines has gained prominence in the field of machine learning and pattern classification. It uses a nonlinear mapping to change training data into higher dimension. An SVM with a small number of support vectors can have high-quality simplification even when the dimensionality of the data is high.

3.5 Artificial Neural Network

Artificial Neural Network (ANN) is organization composed of a number of interconnected units. ANN does not form one network, but a various relations of networks. It includes three types of layer such as input layer, output layer and hidden layer [20].

4 METHODOLOGY

The main objective of the study is to find the best decision tree based classification algorithms from five algorithms namely ID3, C4.5, C5.0, PART and Bagging CART. The classification algorithms are validated based on the performance measures such as precision, recall, f-measures, accuracy and kappa statistic.

4.1 Dataset

This study uses four datasets namely Iris, Contact lenses, Balance scale and Pima which are collected from UCI Repository. The instances and attributes of the two datasets are listed in Table 1.

Table 1: Dataset Description

Measure/ Attributes	Datasets	
	Iris	Ecoli
Instances	150	336
Attributes	5	9

The Table 2 describes the performance measures for Iris dataset for different classification algorithms like Naïve Bayes (NB), IBK (K-Nearest Neighbor), Decision Tree (DT) and J48. From this Naïve Bayes and J48 classifier produces best accuracy measure compared to all other classifier algorithms. The next highest performance measure is IBK achieved 95% and the lowest measure is Decision Tree classification algorithm.

Table 2: Performance Measures for Iris Dataset

Algorithms	Precision (%)	Recall (%)	F-Measures (%)	TP Rate (%)	FP Rate (%)
NB	96	96	96	96	2
IBK	95	95	95	95	2
DT	92	92	92	92	3
J48	96	96	96	96	2

The Figure 1 represented the Naïve Bayes and J48 classification algorithm gives more accuracy than the rest of algorithms for Iris dataset.

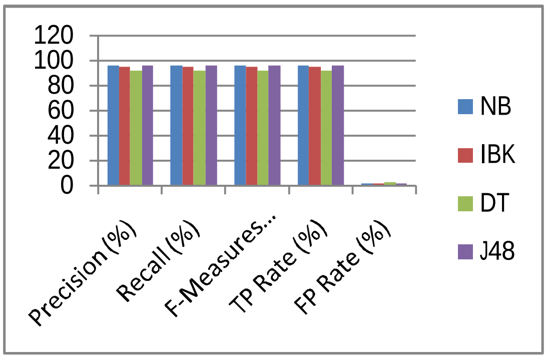

Figure 1 Performance measures for Iris dataset

The Table 3 describes the correctly and incorrectly classified instances for the Iris dataset for various classification algorithms like Naïve Bayes, IBK, Decision Tree and J48 algorithm. In Naïve Bayes algorithm, 144 instances are correctly classified and 6 instances are incorrectly classified. In IBK algorithm, 143 instances are correctly classified and 7 instances are incorrectly classified. In Decision Tree algorithm, 139 instances are correctly classified and 11 instances are incorrectly classified. In J48 algorithm, 144 instances are correctly classified and 6 instances are incorrectly classified.

Table 3: Accuracy Measures for Classification Algorithms

Algorithms	Correctly classified Instances	Incorrectly Classified Instances
NB	144	6
IBK	148	7
DT	139	11
J48	144	6

From Table 3, the Naïve Bayes and J48 algorithms produces the best accuracy measures for the Iris dataset. Figure 2 depicts the details of Table 3 in bar chart.

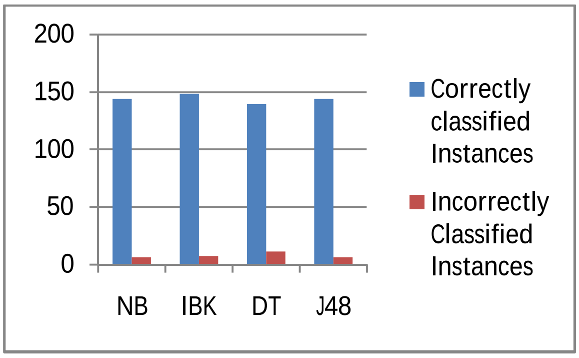

Figure 2 Accuracy measures for Iris dataset

The Table 4 describes the performance measures for Ecoli dataset for different classification algorithms like Naïve Bayes (NB), IBK (K-Nearest Neighbor), Decision Tree (DT) and J48. From this Naïve Bayes classifier produces best accuracy measure compared to all other classifier algorithms. The next highest performance measure is J48 achieved 84% and the lowest measure is Decision Tree classification algorithm.

Table 4 Performance Measures For Iris Dataset

Algorithms	Precision (%)	Recall (%)	F-Measures (%)	TP Rate (%)	FP Rate (%)
NB	86	86	86	86	3
IBK	80	80	80	80	5
DT	76	76	76	76	8
J48	84	84	84	84	4

The Figure 3 represented the Naïve Bayes classification algorithm gives more accuracy than the rest of algorithms for Iris dataset.

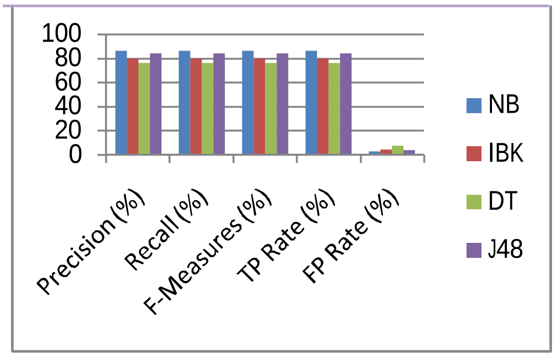

Figure 3 Performance measures for Ecoli dataset

The Table 5 describes the correctly and incorrectly classified instances for the Ecoli dataset for various classification algorithms like Naïve Bayes, IBK, Decision Tree and J48 algorithm. In Naïve Bayes algorithm, 286 instances are correctly classified and 50 instances are incorrectly classified. In IBK algorithm, 270 instances are correctly classified and 66 instances are incorrectly classified. In Decision Tree algorithm, 258 instances are correctly classified and 78 instances are incorrectly classified. In J48 algorithm, 283 instances are correctly classified and 53 instances are incorrectly classified.

Table 5 Accuracy Measures for Classification Algorithms

Algorithms	Correctly classified Instances	Incorrectly Classified Instances
NB	286	50
IBK	270	66
DT	258	78
J48	283	53

From Table 5, the Naïve Bayes algorithm produces the best accuracy measures for the Ecoli dataset. Figure 4 depicts the details of Table 5 in bar chart.

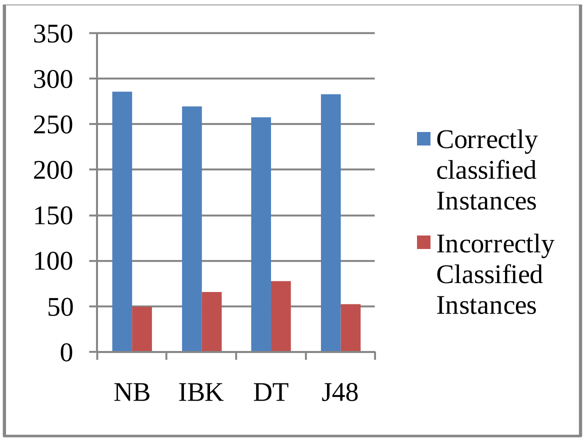

Figure 4 Accuracy measures for Ecoli dataset

5 CONCLUSION

This paper deals among plenty of classification techniques employed in data mining. Classification of Data Mining contains quite a lot of applications such as customer segmentation, credit analysis, bio-medical, marketing, business modeling and drug response modeling. In Data Mining, Classification technique contains various algorithms namely Decision Tree, Naïve Bayes, K-Nearest Neighbor, Support Vector Machine and Artificial Neural Networks. From this analysis, the classification technique generates more precise and accurate system results. This paper presents improve the accuracy and performance of the Classifier. The comparison table shows that how classification algorithms performed in different datasets and identified which one is gives the best accuracy among the different classifier. The result shows that the Naïve Bayes and J48 classification algorithm suitable for the Iris dataset. And then an experimental result demonstrates that the Naïve Bayes classification algorithm suitable for the Ecoli dataset. Because the Naïve Bayes classifier produces best accuracy measure compared to all other classification algorithms such as IBK, Decision Tree (DT) and J48 algorithm. In future work can be extended to include a number of other classification algorithms for different types of dataset.

References

[1] LiangZhao, Deng-Feng Chen, Sheng-Jun Xu and Jun Lu, “The Research of Data Mining Classification Algorithm that Based on SJEP”, International Journal of Database Theory and Application, Vol.8, No.2, pp. 223-234, 2015.

[2] Delveen Luqman Abd AL-Nabi, Shereen Shukri Ahmed, “Survey on Classification Algorithms for Data Mining:(Comparison and Evaluation)”, Computer Engineering and Intelligent Systems, Vol.4, No.8, 2013.

[3] Alexander Statnikov, Lily Wang and Constantin F Aliferis, “A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification”, BMC Bioinformatics, pp. 1-10, 2008.

[4] Hem Jyotsana Parashar, Singh Vijendra, and Nisha Vasudeva, "An Efficient Classification Approach for Data Mining", International Journal of Machine Learning and Computing, Vol. 2, No. 4, pp. 446-448, 2012.

[5] T.Sathya Devi, Dr.K.Meenakshi Sundaram, " A Comparative Analysis Of Meta And Tree Classification Algorithms Using Weka", International Research Journal of Engineering and Technology (IRJET), Vol.3 No.11, pp. 77-83, 2016.

[6] Bendi Venkata Ramana, Prof. M.Surendra Prasad Babu, Prof. N. B. Venkateswarlu, "A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis", International Journal of Database Management Systems(IJDMS), Vol.3, No.2, pp. 101-114, 2011.

[7] Sweety Maniar, Jagdish S. Shah, "Survey and Comparison of Classification Algorithm for Medical Image", International Journal of Engineering And Computer Science, Vol.5, No.8, pp.17679-17684, 2016.

[8] P.Kalaiselvi, Dr.C.Nalini, "A Comparative Study of Meta Classifier Algorithms on Multiple Datasets", International Journal of Advanced Research in Computer Science and Software Engineering, Vol.3, No.3, pp. 654-659, 2013.

[9] Anita Ganpati, "A Performance Comparison Of End, Bagging and Dagging Meta Classification Algorithms", Proceedings of Academics World 24th International Conference, 2016.

[10] R. Porkodi, "A Study on Performance Analysis of Data Mining Classification Algorithms over Lung Cancer Dataset", IJRIT International Journal of Research in Information Technology, Vol. 2, No.3, pp. 49-58, 2014.

[11] Payal Pahwa, Manju Papreja, Renu Miglani, "Performance Analysis of Classification Algorithms", International Journal of Computer Science and Mobile Computing, Vol.3, No.4, pp.50-58, 2014.

[12] Sudhamathy G, Thilagu M, Padmavathi G., "Comparative Analysis of R Package Classifiers Using Breast Cancer Dataset", International Journal of Engineering and Technology (IJET), Vol.8, No.5, pp.2127-2136, 2016.

[13] Pardeep Kumar, Nitin and Vivek Kumar Sehgal and Durg Singh Chauhan, "A Benchmark To Select Data Mining Based Classification Algorithms For Business Intelligence And Decision Support Systems", International Journal of Data Mining & Knowledge Management Process (IJDKP), Vol.2, No.5, pp.25-42, 2012.

[14] Lakshmanaprabu S. K, Shankar K, Ashish Khanna, Deepak Gupta, "Effective Features to Classify Big Data Using Social Internet of Things", IEEE Access, Vol.6, 2018.

[15] Jianhua Shao, Achilleas Tziatzios, "Mining range associations for classification and characterization", Data & Knowledge Engineering, 2018.

[16] N. Chandra Sekhar Reddy, K. Sai Prasad and A. Mounika, "Classification Algorithms on Data mining: A Study", International Journal of Computational Intelligence Research, Vol.13, No.8, pp. 2135-2142, 2017.

[17] Rafael Jiménez, Joella Anupol, Berta Cajal, Elena Gervilla, "Data mining techniques for drug use research", Addictive Behaviors Reports, 2018.

[18] Himani Sharma, Sunil Kumar, "A Survey on Decision Tree Algorithms of Classification in Data Mining", International Journal of Science and Research (IJSR), Vol.5 No.4, pp. 2094-2097, 2016.

[19] Disha A. Katariya, Prof. U.R.Gandhi, "Analysis the Data Mining Classification Algorithm", International Journal for Research in Applied Science & Engineering Technology (IJRASET), Vol.6 No.7, pp. 463-466, 2018.

[20] Arun K Pujari, "Data Mining Techniques", second edition, ISBN: 978-81-7371-6720, 2010.

[21] Sandeep Kumar Budhani, C. K. Jha, Amir Ahmad, "Comparative Study of Meta Classification Algorithm: Bagging, AdaboostM1 and Stacking with Concept Drift based Synthetic Dataset Hyperplane1 and Hyperplane2", International Journal of Engineering Science and Computing, 2018.

[22] Mrs.Yogita Bhapkar, "Comparative analysis of classification based data mining algorithms for credit risk analysis", International Journal of Engineering & Scientific Research Vol. 6, No.2, pp. 1-9, 2018.

AUTHOR

V.Saranya received the M.Sc. and M.Phil. degrees in Computer Science from Bharathiar University in 2017 and 2019 respectively. She is currently an Assistant Professor, Computer Science department in Karur Velalar College of Arts and Science for Women, Karur. Her research interest includes Data Mining, Artificial Intelligence, Network Security and Machine Learning.

A.Vigneswari received the M.Sc. and M.Phil. degrees in Computer Science from Bharathidasan University in 2006 and 2008 respectively. And also she received MCA degree from Periyar University in 2009. She working as an Associate Professor, Computer Science department in Karur Velalar College of Arts and Science for Women, Karur. Her research interest includes Machine Learning, Computer Networks, Artificial Intelligence, and Data Mining.