International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 1 | ISSN : 2456-3307

Classification Algorithms in Data Mining : A Survey

C. Parimala*[1], R. Porkodi[2]
*[1]PG Scholar, Department of Computer Science, Bharathiar University, Coimbatore, Tamilnadu, India
[2]Assistant Professor, Department of Computer Science, Bharathiar University, Coimbatore, Tamilnadu, India

ABSTRACT

Classification is a data mining task that assigns items in a collection to target categories or classes. The scope of classification is to accurately predict the target class for each case in the data. In the hypothesis build training procedure, a classification algorithm find relationships between the worth of the predictors and the values of the goal. Different classification algorithms use dissimilar techniques for finding relationships. These relationships are summarized in a model, which container afterward be apply to a different data set in which the class assignments are unknown. Classification has many applications in customer segmentation, business modeling, marketing, credit analysis, bio medical and drug responsemodeling. This paper presents the study and analysis of five classification algorithms manually Bayesian network, j48, logistic model tree, random tree and rep tree for liver disorders dataset and the performance of these algorithms are compared using the various performance metrics such as Precision, Recalland F measure in which random tree algorithm gives 100% accuracy. The experimental result shows that random tree provides high accuracy than the Bayesian algorithm, j48, logistic model tree and rep tree.

Keywords : Classification Algorithm, Bayesian Net, J48, LMT, Random Tree, REP Tree.

I. INTRODUCTION

Data mining is the pattern of examining large pre-existing database in order to generate new information.The information obtained from data mining is hopefully some new and useful.The overall target of the data mining process is to extract information from a data set and modified it into an understandable structure for further use. The raw analysis step, it involves database and data management features, data preprocessing,model and assumption considerations, interestingness metrics, complicated considerations, post processing of observed structures, visualization and online updating. Data mining is the analysis step of theKDD(knowledge discovery in database) process [1]. The authentic data mining task is the semiautomatic or automatic analysis of large quantities of data to extract formerly undiscovered, interesting patterns such as groups of data records, unusual records and dependencies. Data processing is based on complex algorithms that allow for the segmentation of data to identify patterns and trends, detect anomalies, and predict the probability of various situational outcomes.

Fresh trends are Distributed Data Mining, Multimedia Data Mining, Spatial and Geographic DataMining, Ubiquitous Data Mining, Time Series and Sequence Data Mining, Application Exploration, scalable and interactive data mining methods, integration of data mining with database systems, data warehouse systems and web database systems, Visual data mining, New methods for mining complex types of data, Biological data mining, Data mining and software engineering, Web mining, Real

time data mining and Privacy protection and information security in data mining [2]. Data mining process applications are life sciences, customer relationship management, web applications, manufacturing, competitive benefit, intelligence, retail, finance, banking, computer, network, security, monitoring, surveillance, teaching support, climate modeling, astronomy financial data analysis, retail industry, telecommunication industry, biological data analysis, other scientific applications and intrusion discovery. Data mining advantages are Marketing, Retail, finance, banking, manufacturing and governments. Data mining disadvantages are privacy issues, security issues, Misuse of information and inaccurate information. Classification is the process of automatically creating a model of classes from a set of records that contain class labels.

The section 1 discuss about the introduction of data mining and the classification. Section 2 gives the brief explain literature survey.Section 3 explains methodology and used in classification algorithms. Section 4 discussion results are explained. Section 5 and concludes this analysis work.

II. II.LITERATURE SURVEY

Yun Wan et.al(2015) [3] proposed an Ensemble Sentiment Classification System of Twitter Data for Airline Services Analysis by customer feedback.The proposed system provides ensemble sentiment classification strategy was appliedbased on Majority Vote principle of multiple classificationmethods, including Naive Bayes, SVM, Bayesian Network, C4.5, Decision Tree and Random Forest algorithms.When comparing the two different type of datasets BayesianNetwork is the better accuracy when comparing other algorithms.

Hazem M. El-Bakry et.al(2016) [4] discussedan effective hybrid system for breast cancer classification.The proposed system combines K-means clustering algorithm, fuzzy rough feature selection (FRFS), and discernibility nearest neighbor classifier.Itis proved that the proposed model outperforms with accuracy up to 98.9%.

Anita kumar (2015) [5] proposed data mining classification techniques applied for cancerPerpetuation using cancer data set. Classification techniques such as CART, Random Forest, LMT, and Naive Bayesian are used.Random forest gives better accuracy than other algorithms.

Hakizimana Leopordet.al (2016) [6]proposed survey and analysis for existing techniques on both classification and regression models techniques that have been applied for diseases outbreak prediction in datasets.Classification algorithm are decision tree, svm, naive Bayes.Comparison of various diseases are heart,breast cancer,lung cancer on Naïve Bayes Technique.Naive Bayes gives high accuracy value when comparing to other algorithms.

Nitya Upadhyay et.al [7] proposed the prediction of student's behaviour and academic performance of the education system to improve the result in examination. Classification algorithms are decision tree,C4.5,naïve Bayes,id3,j48.Decision tree gives high accuracy when comparing other algorithms.

IFabien Lotteet.al (2007) [8]proposed classification algorithms used to design Brain Computer Interface (BCI) systems based on ElectroEncephaloGraphy (EEG).The proposed system provides a commonly employed algorithms and describe their criticalproperties SVM and observed that algorithm gives better result for synchronous BCI.

Patel Pinky Set.al(2015) [9] proposed survey of email classification algorithms in Data Mining. The proposed system provides Classification is a data mining technique based on machine learning.The classification algorithms are naïve Bayes, support vector machine, neural network and k-nearest neighbor have been used. Each technique has got its pros and cons.

Arvind Kumaret.al (2015) [10] proposed a Survey on Hoeffding Tree Stream Data Classification Algorithms. The proposed system provides a survey of machine learning techniquesdiscussed.Classification algorithms are hoeffding tree, streaming randomforest and concept adapting very fast decision tree. Hoeffdingtree are better than batch trees in terms of learning time required.

Vandana Korde et.al (2012) [11] proposed a text classification survey based on classification algorithms. The proposed system provides methodologies to organize and extract pattern and knowledge from the documents. The classification algorithms are Bayesian classifier, Decision Tree, K-nearest neighbour (KNN), SupportVector Machines(SVMs), Neural Networks, Rocchio's.SVM produces highest accuracy result than other algorithms.

Divya Jainet.al (2016) [12] proposed Utilization of Data Mining Classification Approach for Disease Prediction.The proposed system provides a comparative study of numerous classification approaches used for predictive analysis of several diseases. Classification algorithms are KNN, Naïve Bayes, support vector machines,decision trees. Naive Bayes and C4.5 algorithms are the highest accuracy and least complexity than other algorithms.

III. METHODOLOGY

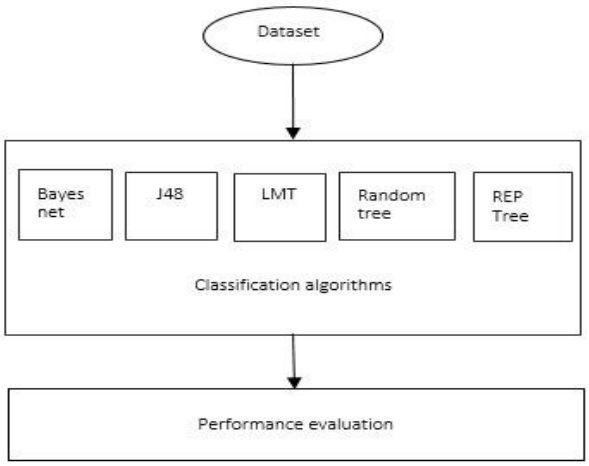

Figure 1. Methodology ofthe proposed research work

The figure 1 shows the methodology of the proposed work that consists of two phases namely classification phase and performance evaluation phase. The liver disorders dataset is chosen as an experimental dataset. The classification phase uses five classification algorithms namely Bayes network, j48,logistic model tree,random treeand rep tree. These algorithms are analyzed and validated using the different performance evaluation metrics.

1. Classification Algorithms

Classification is the data mining task that allocates all record in the data set to one of the few predefined classes. Data set is divided into training and test sets. Training set has a known class labels while the test set labels are unknown. Some of the most popular and common are adapted and presented here in based on their capabilities simplicity and robustness. Classification is likewise characterized as the task of target function learning for mapping each attribute set to its corresponding, class label. There are numerous classification algorithms such as Bayes network,j48,logistic model tree,random tree and rep tree [13].

1.1BayesNetwork:

Bayesian network is also called belief networks, is a graphical model for probability relationships among a set of variables features, This Bayesian network consist of two components. First component is mainly a directed acyclic graph (DAG) in which the nodes in the graph are called the random variables and the edges between the nodes represents the probabilistic dependencies among the corresponding random variables. Second component is a set of parameters that describe the conditional probability of each variable taken its parents. A Bayesian network describes a system by specifying relationships of conditional dependence between its variables.The conditional dependences are

represented by a directed acyclic graph, in which, each node [14].

1.2 J48:

J48 is an open source Java implementation of the C4.5 algorithm.C4.5 is a program that creates a decision tree based on a set of labeled input data. It is a simple C4.5 decision tree for classification, it creates a binary tree. The decision tree approach is the most useful in the classification problem. With this technique, a tree is built to model the classification process. Once the tree is built, it is functional to each tuple in the database and results in classification for that tuple though the tree is built. It ignores the missing values. The basic idea of J48 is to divide the data into range based on the attribute values for that item that are found in the training Sample. It allows classification either in the form of decision tree or rules created based on the test set provided [15].

1.3 LogisticModelTree:

It is a classification model with an associated training algorithm that unify logistic regression (LR) and decision tree learning [16]. It is also called a logic model, which is used to model dichotomous outcome of variables. Logistic model trees are based on the earlier notion of a model tree: a decision tree that has linear regression models at its leaves to provide a piecewise linear regression model.

1.4 Random Tree:

It is a supervised classifier, it is an ensemble learning algorithm that generates many individual learners. It employs a bagging idea to produce a random set of data for constructing a decision tree, random tree is a collection of tree predictors that is called forest [17].

1.5 Rep Tree:

REP means reduced-error pruning. REP Tree is a quickly decision tree learner which builds a decision and regression tree using information gain as thesplitting criterion, and prunes it using reduced errorsnipping [18].

IV. RESULTS AND DISCUSSION

A. Dataset Description

The Table 1 shows the liver disorders data set with 345 instances and the following attributes are mcv, alkphos, sgpt, sqot, gammagt, drinks and selection has been used for analysis liver disorders data due to its proficiency of disease [19].

Table 1. Liver disorders dataset

ATTRIBUTE	DESCRIPTION
MCV	Mean corpuscular volume
ALKPHOS	Alkaline phosphotase
SGPT	Alamine aminotransferase
SQOT	Aspartate aminotransferase
GAMMAGT	Gamma-glutamyl transpeptidase
DRINKS	Number of half-print equivalents of alcoholic beverages drunk per day
SELECTOR	Selector field used to split data into two sets

The Table 2 shows that the five classification models, generated with the selected data mining algorithms, are compared by using the following evaluation measures: % of correctly and incorrectly classified instances and kappa statistic. These are well known measures for evaluation of data mining models for classification. Kappa statistic is a measure of the degree of non-random agreement between observer and measurement of the same categorical variable. When comparing the time taken for different algorithms the Bayes net and random tree takes the least computing time. The figure 2 shows the correctly and incorrectly classified instances from the result of the five classification algorithms and observed that random tree algorithms gives better classification result than the other algorithms.

Table 2. Results of classifiers

Evaluation criteria	Classification Algorithms				
	Bayes net	J48	LMT	Random tree	Rep tree
Correctly classified instances	218	292	261	345	256
	64%	85%	76%	100%	74%
Incorrectly classified instances	127	53	84	0	89
	36%	15%	24%	0%	26%
Kappa statistics	0.24	0.67	0.48	1	0.43
Time taken sec	0.02	0.03	0.53	0.02	0.03
Accuracy	64%	85%	76%	100%	74%

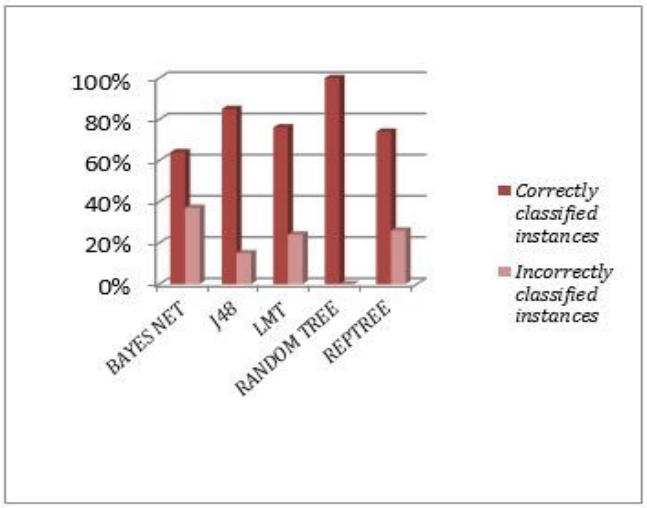

Figure 2. Correctlyand Incorrectly Classified Algorithms

Table 3. Accuracy Results

Classifier	Precision	Recall	Fmeasure
BAYES NET	0.63	0.63	0.63
J48	0.85	0.84	0.84
LMT	0.75	0.75	0.75
RANDOM TREE	1	1	1
REO TREE	0.76	0.74	0.72

The five classification algorithms have been validated using the five important metrics such as precision, recall and F-measure and these validation results are shown in Table 3 and same is depicted in fig.3. The result shows that random tree classifier gives better classification accuracy as 1.0 than the other four algorithms.

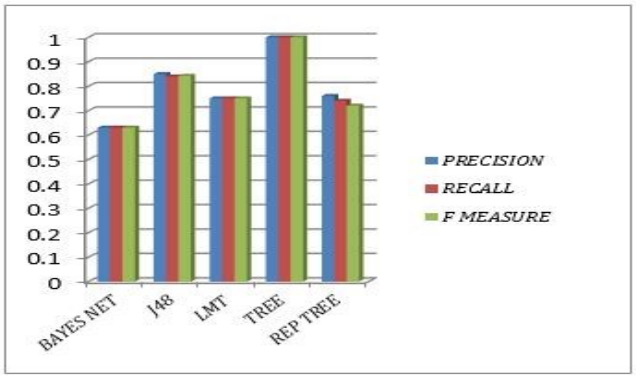

Figure 3. Performance Evaluation

V. CONCLUSION

This paper conducted an extensive study on five classification algorithms and the experimental result shows that the random tree classifier gives better accuracy 100%, which takes 0.02 seconds for training. The second best algorithm is the j48 which gives accuracy 85% and taken 0.03 seconds than the random tree algorithm. The third best algorithm is logistic model tree which gives accuracy 76% and takes 0.53 seconds. The fourth best algorithm is rep tree which gives accuracy 74% and takes 0.03 seconds for training. Finally Bayes network gives the least accuracy 64% and takes 0.02 seconds for training the instances.

VI. REFERENCES

[1]. Fayyad, Ussama; Piattetsky-Shapiro,Gregory; Smyth, Padhraic (1996). "From Data Mining to Knowledge Discovery in Databases" (PDF). Retrieve December 2008.

[2]. https://graduatedegrees.online.njit.edu/resources/mscs/mscs-articles/current-trends-in-data-mining/

[3]. Yun Wan, Dr. QigangGao," An Ensemble Sentiment Classification System of Twitter Data for Airline Services Analysis",2015 IEEE

15th International Conference on Data Mining Workshops.

[4]. Ibrahim M. El-Hasnony, Hazem M. El-Bakry, Ahmed A. Saleh, ," Classification of Breast Cancer Using Softcomputing Techniques", International Journal of Electronics and Information Engineering, Vol.4, No.1, PP.45-54, Mar. 2016.

[5]. Anita kumar,"A Study on Cancer Perpetuation Using the Classification Algorithms", International Journal of Recent Research in Mathematics Computer Science and Information Technology Vol. 2, Issue 1, pp: (96-99), Month: April 2015 – September 2015, Available at: www.paperpublications.org

[6]. HakizimanaLeopord, Dr. Wilson KiprutoCheruiyot, Dr. Stephen Kimani," A Survey and Analysis on Classification and Regression Data Mining Techniques for Diseases Outbreak Prediction in Datasets", The International Journal Of Engineering And Science (IJES) || Volume || 5 || Issue || 9 || Pages || PP -01-11 || 2016 || ISSN (e): 2319 – 1813 ISSN (p): 2319 – 1805.

[7]. NityaUpadhyay, VinodiniKatiyar," A Survey on the Classification Techniques In Educational Data Mining", International Journal of Computer Applications Technology and ResearchVolume 3– Issue 11, 725 - 728, 2014, ISSN: 2319–8656.

[8]. Fabien Lotte, Marco Congedo, Anatole Lecuyer, FabriceLamarche, Bruno Arnaldi, "A review of classification algorithms for EEG-based Brain computer interfaces", Journal of Neural Engineering, IOP Publishing, 2007, 4, pp.24. <inria-00134950>.

[9]. Patel Pinky S. Devendra V. Thakor," A Survey of Email Classification Algorithms in Data Mining", International Journal of Engineering Technology, Management and Applied Sciences www.ijetmas.com January 2015, Volume 3 Issue 1, ISSN 2349-4476.

[10]. Arvind Kumar, ParminderKaur andPratibha Sharma,"A Survey on Hoeffding Tree Stream Data ClassificationAlgorithms", CPUH-Research Journal: 2015, 1(2), 28-32ISSN (Online): 2455-6076 http://www.cpuh.in/academics/academic_journals.php

[11]. VandanaKorde and C NamrataMahender," Text classification and classifiers:A survey", International Journal of Artificial Intelligence & Applications (IJAIA), Vol.3, No.2, March 2012.

[12]. Divya Jain, VijendraSingh ," Utilization of Data Mining Classification Approach for Disease Prediction: A Survey", I.J. Education and Management Engineering, 2016, 6, 45-52 Published Online November 2016 in MECS (http://www.mecs-press.net) DOI:10.5815/ijeme.2016.06.05.

[13]. L. Tao, F. Sun, and S. Yang, A fast and robust sparse approach for hyper spectral data classification using a few labelled samples," IEEE Transactionson Geoscience and Remote Sensing, vol. 50, no. 6, pp. 2287-2302, 2012.

[14]. Delveen Luqman Abd AL-Nabi1, Shereen Shukri Ahmed2, Survey on Classification Algorithms for Data Mining:(Comparison and Evaluation),Computer Engineering and Intelligent SystemsISSN 2222-1719 (Paper) , ISSN 2222-2863 (Online)Vol.4, No.8, 2013.

[15]. R. Sivanesan1, K. Devika Rani Dhivya2 , "A Review on Diabetes Mellitus diagnoses using classification on Pima Indian Diabetes Data Set", International Journal of Advance Research in Computer Science and Management Studies Research Article / Survey Paper / Case Study, Volume 5, Issue 1, January 2017, Available online at: www.ijarcsms.com

[16]. Wikipedia (2017). Logistic model tree. Available:https://en.wikipedia.org/wiki/Logistic model tree.

[17]. Jiawei Han and Micheline Kamber Data Mining:Concepts and Techniques, second edition.

[18]. Ghosh, S. R. and Waheed, S. (2017). Analysis of classification algorithms for liver disease

diagnosis. Journal of Science, Technology and Environment Informatics, 05(01), 361-370. https://doi.org/10.18801/jstei.050117.38 .

[19]. C.L. Blake, D.J. Newman, S. Hettich and C.J. Merz. (2012) UCI machine learning repository databases. [Online]. Available: http://mlr.cs.umass.edu/ml/machine-learning-databases/0022

[20]. Cheng, Hong, et al. "Discriminative frequent pattern analysis for effective classification." Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on. IEEE, 2007.

[21]. M. El-Hasnony, H. M. El Bakry, A. A. Saleh, "Comparative study among data reduction techniques over classification accuracy," International Journal of Computer Applications, vol. 122, no. 2, pp. 8,15, 2015.

[22]. John C. Bailar, Thomas A. Louis, Philip W. Lavori, Marcia Polansky, "A Classification for Biomedical Research Reports," N Engl J Med, Vol. 311, No. 23 pp. 1482-1487, in the year 2010.

[23]. Ada, Rajneet Kaur. "Using Some Data Mining Techniques to Predict the Survival Year of Lung Cancer Patient." (2013).

[24]. M. A. Nishara Banu, B. Gomathy ,"Disease Forecasting System Using Data Mining Methods", 2014 International Conference on Intelligent Computing Applications.