Volume 5, Issue 9, September 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Comparing Analysis of Decision Tree Algorithms Based on Healthcare

N. Saravanan
Research Scholar
Department Computer Science
Periyar University, Salem, Tamil Nadu, India

Dr. V. Gayathri
Professor, Department of Computer Science
Gonzaga College of Arts and Science for Women
Krishnagiri, Tamil Nadu, India

Abstract: ***This study used data mining modelling techniques to examine the healthcare classification. The healthcare system which recommended patients based on choice of various healthcares to be offered. Here in this paper we compare the five classification algorithm to choose the best classification algorithm for Healthcare system. These five classification algorithms are Nitre, J48, LMT, RondomForest & Simple Cart, and Classification Algorithm. We compare these five Healthcare using open source data mining tool Weka & present the result. We found that LMTree classification algorithm works better for this Healthcare System than other five classification algorithms.***

Keywords: NBTree, J48, LMT, RondomForest and SimplreCart Classification Algorithm, Weka

I. INTRODUCTION

The healthcare system which recommended patients based on choice of various healthcares to be offered E.g. If the patients is interested in healthcare system would like LMT classification is better like Advanced System. Here we use open source for data collection & Weka to check the results. A framework for healthcare System is explained in.

II. LITERATURE REVIEW

In research [1], they conducted experimental comparison of ADTree, Simple CART, J48, ZeroR and Naïve Bays on Moodle database of a college. The experimental results show that all ensemble methods outperform ADTree. The experimental results also show that all five methods benefit from data preprocessing, including gene selection and discretization, in classification accuracy. In addition to comparing the average accuracies of ten-fold cross validation tests on seven data sets, they used two statistical tests to validate findings.

In the paper [2], they presented an analysis of the in customer membership card model two algorithms for C5 and CART will be used and comparative using data mining techniques. Applying data mining classification algorithm in The customer membership card classification model can help to understand Children number and income levels factors are affecting on the card ranks. Therefore, we are able to these two attributes as the main standards to recommend card to a new customer. Moreover, we can refer other attributes to help judge which cart to recommend to future customers, they have investigated two data mining techniques: The C5.0 and CART Classification two algorithms. The achieved prediction performances are comparable to existing techniques. They found out that C5.0 algorithm has a much better performance than the other techniques.

In paper [3], three existing decision tree algorithms ID3, C4.5 and CART have been applied on the educational data for predicting the student's performance in examination. They efficiency of various decision tree algorithms can be analyzed based on their accuracy. C4.5 is the best algorithm among all the three because it provides better accuracy and efficiency then other algorithms

In the paper [4], they conducted experiment in the WEKA environment by using four algorithms namely ID3, J48, NBTree and NEDTA and Alternating Decision Tree on banking dataset were compared in terms of classification accuracy. According to their simulation results, the NEDTA classifier outperforms the ID3, J48 and NBTree in terms of classification accuracy.

In paper [5], they presented a large-scale empirical comparison between ten supervised learning methods: SVMs, neural nets, logistic regression, naive bayes, memory-based learning, random forests, decision trees, bagged trees, boosted trees, and boosted stumps. They also examined the effect that calibrating the models via Platt Scaling and Isotonic Regression has on their performance.

III. CLASSIFICATION ALGORITHMS

Classification is a data mining task that maps the data into predefined groups and classes. It is also called as supervised learning. It consists of two steps. First step is the model construction which consists of set of predetermined classes. Each tuple sample is assumed to belong to a predefined class. The set of tuple used for model construction is training set.

The model is represented as classification rules, decision trees, or mathematical formulae. Second step is model usage which is used for classifying future or unknown objects. The known label of test sample is compared with the classified result from the model. Accuracy rate is the percentage of test set samples that are correctly classified by the model. Test set is independent of training set, otherwise over-fitting will occur [6]. Here we consider the brief introduction of each classification algorithm.

A. LMT Classification Algorithm

Logistic Model Trees use logistic regression functions at the leaves. This method can deal with missing values, binary and multi-class variables, numeric and nominal attributes. It generates small and accurate trees. It uses CART pruning technique. It does not require any tuning parameters. It is often more accurate than C4.4 decision trees and standalone logistic regression [7]. LMT produces a single tree containing binary splits on numeric attributes, multiway splits on nominal ones and logistic regression models at the leaves. It also ensures that only relevant attributes are included in the latter.

B. J48 Classification Algorithm

J48 is a tree based learning approach. It is developed by Ross Quinlan which is based on iterative dichtomiser (ID3) algorithm. J48 uses divide-and =-conquer algorithm to split a root node into a subset of two partitions till leaf node (target node) occur in tree. Given a set T of total instances the following steps are used to construct the tree structure.

Step1: if all the instances in T belong to the same group class or T is having fewer instances, than the tree is leaf labeled with the most class in T.

Step2: If step 1 does not occur than select a test based on a single attribute with at least two or greater possible outcomes. Then consider this test as a root node of the tree with one branch of
each outcome of the test, partition T into corresponding T1, T2,T3…, according to the result for each respective cases, and the same may be applied in recursive way to each sub node. [8].

Step3: Information gain and default gain ratio are ranked using tow heuristic criteria by algorithm J48

C. Simple Cart Classification Algorithm

Classification And Regression Trees it is developed by Breiman, Friedman, Olshen, Stone in early 80's.Introduced tree-based modelling into the statistical mainstream and Rigorous approach involving cross-validation to select the optimal tree. Our philosophy in data analysis is to look at the data from a number of different viewpoints. Tree structured regression offers an interesting alternative for looking at regression type problems. It has sometimes given clues to data structure not apparent from a linear regression analysis. Like any tool, its greatest benefit lies in its intelligent and sensible application

D. NBTree Classification Algorithms

NBTree: NBTree (Naive Bayesian tree) consists of [9] naïve Bayesian classification and decision tree learning. The naive Bayesian tree learner, NBTree (Kohavi 1996). An NBTree classification sorts the example to a leaf and then assigns a class label by applying a naïve bayes on that leaf. NBTree frequently achieves higher accuracy than either a naive Bayesian classifier or a decision tree learner.

The steps of NBTree algorithm are:

(a) At each leaf node of a tree, a naive bayes is applied.
(b) By using naive bayes for each leaf node, the instances are classified.
(c) As the tree grows, for each leaf a naive bayes is constructed.
(d) This process repeated until no example is left.

E. RandomForest

Random forests are an ensemble method used for classification. The methodology includes construction of decision trees of the given training data and matching the test data with these. Random forests are used to rank the importance of variables in a classification problem. To measure the importance of a variable in a data setDn= {(Xi, Yi)} i=1n we fit a random forest to the data. During the fitting process the error for each data point is calculated and averaged over the forest. To measure the importance of the i-th feature after training, the values of the i-the feature are permuted among the training data and the error is again computed on this data set. The importance score for the i-the feature is computed by averaging the difference in error before and after the permutation for all the trees. Normalization of the score is done by the standard deviation of these differences [10].

Table 1: Result Using Different Classification Algorithm

Classification Algorithm → **Health Care ↓**		**NB Tree**	**J48**	**LMT**	**Random Forest**	**Simple Cart**
Breast Cancer	Correctly Classify Instance	203	216	215	192	198
	Incorrectly Classify Instance	83	70	71	92	88
Contact_Lenses	Correctly Classify Instance	17	20	17	19	19

	Incorrectly Classify Instance	7	4	7	5	5
Diabetes	Correctly Classify Instance	565	567	595	568	577
	Incorrectly Classify Instance	203	201	173	200	191
Iris	Correctly Classify Instance	142	144	141	141	143
	Incorrectly Classify Instance	8	6	9	9	7
Iris_2D	Correctly Classify Instance	142	144	142	140	142
	Incorrectly Classify Instance	8	6	8	10	8
Overall Result	Total	1378	1378	1378	1376	1378
	Total Correctly Classify Instance	1069	1091	1110	1060	1079
	%	77.58	79.17	80.55	77.03	78.30
	Total Incorrectly Classify Instance	309	287	268	316	299
	%	22.42	20.83	19.45	22.97	21.70

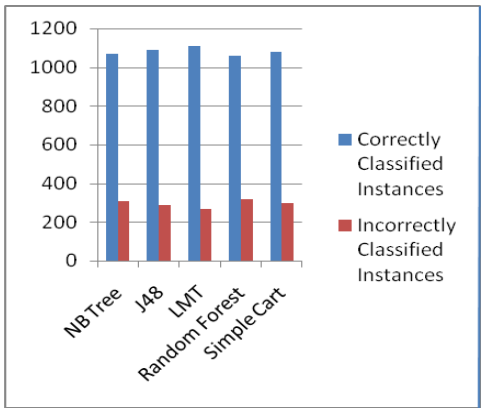

Figure 1: This graph shows accuracy comparison of NBTree, J48, LMT, Random Forest and Simple Cart Algorithms.

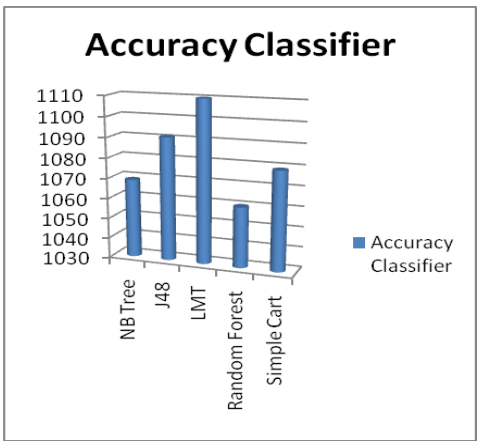

Figure 2: This graph shows accuracy Classifier of NBTree, J48, LMT, Random Forest and Simple Cart Algorithms.

IV. WEKA

WEKA is a data mining system developed by the University of Waikato in New Zealand that implements data mining algorithms using the JAVA language. WEKA is a state of- the-art facility for developing machine learning (ML) techniques and their application to real-world data mining problems. It is a collection of machine learning algorithms for data mining tasks. The algorithms are applied directly to a dataset. WEKA implements algorithms for data preprocessing, classification, regression, clustering and association rules; It also includes visualization tools. The new machine learning schemes can also be developed with this package. WEKA is open source software issued.

The data file normally used by Weka is in ARFF file format, which consists of special tags to indicate different things in the data file (foremost: attribute names, attribute types, and attribute values and the data). The main interface in Weka is the Explorer. It has a set of panels, each of which can be used to perform a certain task. Once a dataset has been loaded, one of the other panels in the Explorer can be used to perform further analysis [11].

V. EXPERIMENTAL RESULT

Here we are considering the sample data extracted from open source and we use Weka. In this data consider five medicine oriented such as Breast Cancer, Contact lenses, Diabetes, Iris, Iris_2D

In table1, we are considering only those from table 1 for which the classification algorithm classifies highest percentage From table 1, we can observe that LMT has highest percentage of correctly classified instance & lowest

percentage of incorrectly classified instances. Random Forest classification algorithm has lowest percentage of correctly classified instances & highest percentage of incorrectly classified instances. J48 has the 79.17% & 20.83% percentage for correctly & incorrectly classified instances. NBTree & Simple Cart classification algorithm has 77.85% & 78.30% correctly classified instances respectively & 22.42%, and 21.70%, and 12.88% incorrectly classified instances.

Ascending order of classification algorithm considering the classification accuracy into account is LMT, J48, Simple Cart, NBTree and Random Forest, so we consider the LMT as classification algorithm for healthcare System as classification accuracy for LMT is highest. Figure1, This graph shows accuracy comparison of NBTree, J48, LMT, Random Forest and Simple Cart Algorithms. Figure2, This graph shows accuracy Classifier of NBTree, J48, LMT, Random Forest and Simple Cart Algorithms.

VI. CONCLUSION

Here in this paper we compare the five classification algorithm to choose the best classification algorithm for recommending the healthcare to various medicines but suitable algorithm patients to choices. These five classification algorithms, we consider for comparison, are NBTree, J48, LMT, RondomForest & Simple Cart, and Classification Algorithm. We use the open source data mining tool Weka to check the result. We found that LMT classification algorithm works better for this Healthcare System as incorrectly classified instance for this algorithms are less than other five classification algorithms.

VII. FUTURE WORK

Future works include the combination of other data mining algorithms to recommend the Healthcare to the patients from the data obtained from the open source.

REFERENCES

[1] Comparative Study of Classification Algorithms, International Journal of Information Technology, Knowledge Management, July-December2012, Volume5, No.2 PP.239-243.

[2] Customer Card Classification based on C5.0 & CART Algorithms, international Journal of Engineering Research and Applications, Vol.2, Issue 4, July-August 2012, PP.164-167.

[3] Survey on Decision Tree Classification algorithms for the Evaluation of Student Performance, International Journal of Computers & Technology, Volume 4 No.2, March-April, 2013, ISSN 2277-3061.

[4] Classification of data using New Enhanced Decision Tree Algorithm (NEDTA), International Association of Scientific Innovation and Research (IASIR),ISSN(Print):2279-0047.

[5] Rich Caruana Alexandru Niculescu-Mizil, "An Empirical Comparison of Supervised Learning Algorithms".

[6] Sunita B. Aher and Lobo L.M.R.J. "Data Mining in Educational System using WEKA". IJCA Proceedings on International Conference on Emerging Technology Trends (ICETT) (3), 20-25, 2011. Published by Foundation of Computer Science, New York, USA (ISBN: 978-93-80864- 71-13).

[7] Niels Landwehr, Mark Hall, and Eibe Frank: Logistic Model Trees. In Machine Learning 59 (1-2) 161-205(2005)

[8] Performance Analysis of Classification Tree Learning Algorithms, International Journal of Computer Applications (0975-8887), Volume 55-No.6, October 2012.

[9] Yumin Zhao, Zhendong Niu_ and Xueping Peng, "Research on Data Mining Technologies for Complicated Attributes Relationship in Digital Library Collections","Applied Mathematics & Information Sciences, An International Journal", Appl. Math. Inf. Sci. 8, No. 3, 1173-1178 (2014)

[10] http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#prox

[11] Dr. Sudhir B. Jagtap, Dr. Kodge B. G. "Census Data Mining and Data Analysis using WEKA", (ICETSTM – 2013) International Conference in "Emerging Trends in Science, Technology and Management-2013, Singapore.

