SINGLE DECISION TREE CLASSIFIERS' ACCURACY ON MEDICAL DATA

Md. Rajib Hasan[1], Nur Azzah Abu Bakar[2], Fadzilah Siraj[2], Mohd Shamrie Sainin[2], Shariful Hasan[3]

[1]*Universiti Utara Malaysia, Malaysia, rajib@live.com.my*
[2]*Universiti Utara Malaysia, {nurazzah,fad173,shamrie}@uum.edu.my*
[3]*Universiti Putra Malaysia, Malaysia, shariful_hsn@yahoo.com*

ABSTRACT. Decision tree is one of the classification techniques for classifying sequential decision problems such as those in medical domain. This paper discusses an evaluation study on different single decision tree classifiers. There are various single decision tree classifiers which have been extensively applied in medical decision making; each of these classifies the data with different accuracy rate. Since accuracy is crucial in medical decision making, it is important to identify a classifier with the best accuracy. The study examines the performance of fourteen single decision tree classifiers on three medical data sets, i.e. Wisconsin's breast cancer data sets, Pima Indian diabetes data sets and hepatitis data sets. All classifiers were trained and tested using WEKA and cross validation. The results revealed that classifiers such as FT, LMT, NB tree, Random Forest and Random Tree are the five best single classifiers as they constantly provide better accuracy in their classifications.

Keywords: decision tree classifier, machine learning algorithm, decision tree evaluation

INTRODUCTION

Decision tree is a classification scheme which generates a tree and a set of rules from a given dataset. The tree represents the model of different classes to which the data belong. Hans and Kamber (2006) describe decision tree as a structure consisting of nodes and branches. Each internal node denotes a test on an attribute while the leaf nodes represent the classes or class distributions, and the branch represents an outcome of the test. Decision tree is used to divide a large collection of records into successively smaller sets of records by applying a sequence of simple decision rules. According to Chandra (2011), decision tree predicts the value of a dependent attribute given the values of the independent attributes. Its structures are organized as classification schemes such that it can facilitate decision making in sequential decision problems, and it is a form of multiple variable (or multiple effect) analyses, including prediction, explanation, description, or classification of an outcome or target (Witten & Frank, 2005).

Decision trees have been widely used both to represent and to conduct decision processes (Lopez-Vallverdu, Riano & Bohada, 2012). In medical decision making, the decision makers often face problems with sequential decision problem involving decisions that lead to different outcomes. Ishwaran and Rao (2009) emphasize that when decision process involves many

sequential decisions, the decision problem becomes difficult to visualize and implement. Thus, decision trees are indispensable graphical tools in such settings as it allows for intuitive understanding about the problem and can aid in decision making. Medical decision making are made for various purposes including screening, diagnosing, pruning and drug and therapy prescription (Fauci et al., 2009). Over the years, multiple computer-based structures have been proposed to formalize these decision processes. They range from statistical approaches such as Bayesian Networks (Arsene, Dumitrache & Mihu, 2011; Lucas, van der Gaag & Abu-Hanna, 2004; Velikova, de Carvalho & Lucas, 2007) or probabilistic models (Husmeier, Dybowski & Roberts, 2004) to symbolic approaches such as decision trees (Chapman & Sonnenberg, 2003), decision tables or decision rules (Yeh, Cheng & Chen, 2011). Among them, decision trees have been particularly successful and widely used both to represent and to conduct decision processes. Medical decision trees can be provided by experts or induced from medical databases. In medical domain, decision tree is often referred to as classification tree in which the outcome is a classification label such as the disease status of a patient.

DECISION TREE CLASSIFIERS

Single decision tree classifier (DTC) is one of the versatile classifiers in data mining. DTCs are very helpful in classifying medical data, which is important in decision making process for medical practitioners (Lavanya & Rani, 2012).

This study evaluates and compares fourteen single DTC algorithms in successive testing and cross validation. The DTCs include the Alternating Decision Tree (AD Tree), Best-first tree (BF Tree), Classification and Regression Tree (CART), Decision Stump, Functional Trees (FT), Logical Analysis of Data (LAD Tree), Logistic Model Trees (LMT), Naïve Bias Tree (NB Tree) and Reduced Error Pruning Tree (REP Tree), J48, J48 Graft, Hoeffding, Random and Random Forest. This comparison is used as a basis to identify the base classifiers to build an ensemble model, however, discussion on ensemble model is not included in this paper. Figure 1 summarizes the 14 DTCs.

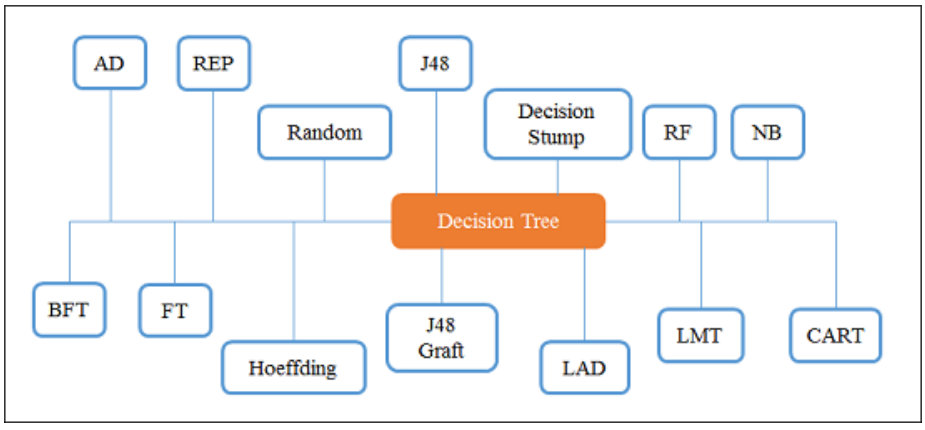

Figure 1. Single decision tree classifiers

Previous studies on DTCs revealed various performance results (Huanhuan Chen, Yao & Tino, 2011; Parvin, MirnabiBaboli, & Alinejad-Rokny, 2015; Tao, 2014; Parvin, Ghatei, Alinejad-Rokny, & Minaei, 2011). AD trees, for example, is only pay off if the data contains many thousands of instances and yield no benefit on small datasets. The studies by Witten et al. (2011) and Mohamed, Salleh and Omar (2012) found that REP tree can reduce misclassifi-

cation of data for the better accuracy. Subramanian, Srinivasan, & Ramasamy (2012) conducted the study on classification algorithms for Network Intrusion Systems and found that Random trees can be generated efficiently and the combination of large sets of random trees generally leads to accurate models. J48 and REP tree are both competitive in producing better classification results in Subramanian, Srinivasan and Ramasamy (2012) and Mohamed et al. (2012). The average difference performance of J48 and REP tree is 7.10% for accuracy and 6.28% for complexity of tree structure (Mohamed et al., 2012). Zhao and Zhang (2007) found that J48 is the optimal choice considering both accuracy and speed for finding active object while Decision Stump is only better in terms of speed.

DATASET

The study uses three different datasets which were obtained from UCI Machine Learning repository. They are Wisconsin's Breast Cancer (WBC) dataset, Pima Indian diabetes dataset and hepatitis dataset. These datasets consist of actual real life data and are relatively large with different size of missing values, i.e. 16 for WBC dataset and 167 for hepatitis dataset. Pima Indian diabetes dataset however, does not contain any missing value data. A built-in function in WEKA, i.e. the Replace Missing Values function that serves as an unsupervised filter is used in this study to automatically replace all the missing values at preprocessing phase.

Table 1 summarizes the properties of each dataset.

Table 1. Datasets

	Wisconsin's Breast Cancer dataset	**Pima Indian diabetes dataset**	**Hepatitis dataset**
Data Set Characteristics	Multivariate	Multivariate	Multivariate
Attribute Characteristics	Integer	Integer, Real	Categorical, Integer, Real
Associated Tasks	Classification		Classification
Classification		N/A	
Number of Instances	699	768	155
Number of Attributes	10	8	19
Missing Values	16	No	167
Area	Life	Life	Life

RESULT

All the 14 DTCs were trained and tested using WEKA and the testing results are as shown in Table 2. LMT classifier produced the highest accuracy for Wisconsin's breast cancer data, i.e. 74.23%, followed by NB Tree and Random Forest (71.13%) and Random Tree (70.10%). For Pima Indian diabetes dataset, FT produced the best accuracy which is 80.84%, followed by Hoeffding tree (80.08%), LMT (79.31%), and NB tree (78.16%). For hepatitis dataset,

Hoeffding tree and NB tree showed an equal performance with 81.13% accuracy whereas LMT, Random forest and Random tree performed better with 84.91%, 86.79% and 83.02% respectively.

Table 2. Testing performance for each single DTC

Algorithm (Training) (Testing percentile 66)		Wisconsin's Brest Cancer Attribute: 699 Instance: 10	Pima Indian-Diabetes Attribute: 768 Instance: 9	Hepatitis Attribute: 155 Instance: 20	Average
AD Tree	Testing	67.0103%	75.8621 %	75.4717 %	72.78%
BF Tree	Testing	65.9794 %	76.6284 %	73.5849 %	72.06%
Decision Stump	Testing	65.9794%	77.3946 %	73.5849 %	72.32%
FT	Testing	68.0412 %	80.8429 %	79.2453 %	76.04%
Hoeffding Tree	Testing	65.9794 %	80.0766 %	81.1321 %	75.73%
J48	Testing	68.0412 %	76.2452 %	79.2453 %	74.51%
J48 graft	Testing	64.9485 %	76.6284 %	79.2453 %	73.61%
LAD Tree	Testing	69.0722 %	74.3295 %	75.4717 %	72.96%
LMT	Testing	74.2268 %	79.3103 %	84.9057 %	79.48%
NB Tree	Testing	71.134 %	78.1609 %	81.1321 %	76.81%
Random Forest	Testing	71.134 %	75.8621 %	86.7925 %	77.93%
Random Tree	Testing	70.1031 %	72.4138 %	83.0189 %	75.18%
REP Tree	Testing	65.9794 %	75.4789 %	77.3585 %	72.94%
Simple CART	Testing	65.9794%	76.2452 %	71.6981 %	71.31%

It is learnt from Table 2 that FT, LMT, NB tree, Random Forest and Random tree consistently perform better than the other classifier algorithms on each single dataset. In line with this, the average performance on three datasets also shows that they are the five best DTC algorithms where LMT score the highest average performance with 79.48%. Random Forest is the second best with 77.93%, followed by NB tree (76.81%), FT (76.04%), and random tree (75.18%). Figure 2 summarizes the results of testing these five algorithms on single and all datasets.

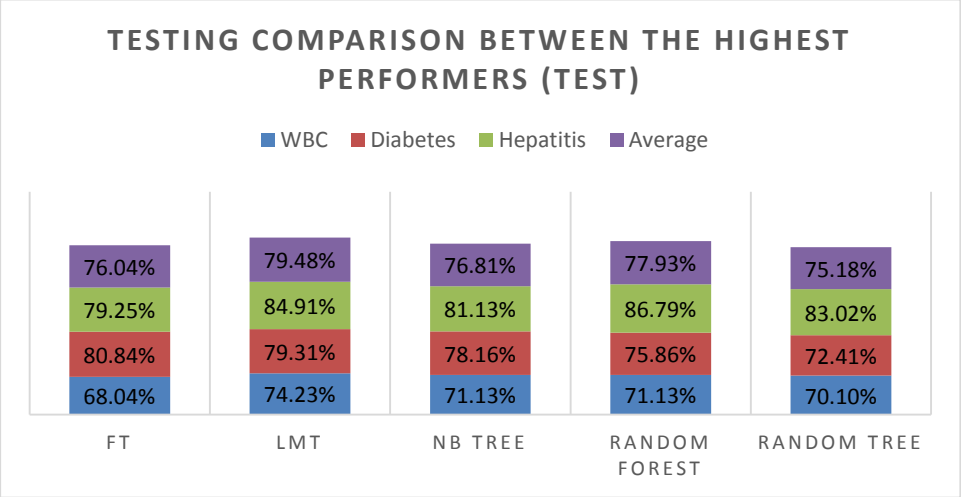

Figure 2. Comparison between the six best DTCs (Testing)

Cross Validation

The 10 folds cross validation was used on the five classifiers and revealed that FT classifier scores the highest accuracy for WBC data, i.e. 96.99%. For Pima Indian diabetes data, LMT performed better with the accuracy of 77.47% meanwhile for hepatitis data, Random Forest performed better at 83.87%. Table 3 summarizes this results.

Table 3. Cross validation result (10 folds)

Algorithm (Training) (Testing percentile 66)	Wisconsin's Brest Cancer Attribute: 699 Instance: 10	Diabetes Attribute: 768 Instance: 9	Hepatitis Attribute: 155 Instance: 20	Average
FT	96.9957 %	77.3438 %	81.2903 %	85.2099 %
LMT	95.9943 %	77.474 %	83.2258 %	85.5647 %
NB tree	95.8512 %	73.5677 %	82.5806 %	83.9998 %
Random forest	95.1359 %	74.349 %	83.871 %	84.4519 %
Random tree	94.5637 %	68.099 %	76.7742 %	79.8123 %

CONCLUSION

The study reported the accuracy for each of the fourteen DTC based on the standard accuracy measurement in terms of testing. Cross validation was applied on the selected base classifiers, i.e. FT, LMT, NB tree, Random forest and Random tree. As discussed above, the accuracy varied between different single classifiers.

REFERENCES

Arsene, O., Dumitrache, I., & Mihu, I. (2011). Medicine expert system dynamic Bayesian network and ontology based. *Expert Systems with Applications*, 38, 15253–15261.

Chandra, B. (2011). Heterogeneous Node Split Measure for Decision Tree Construction, 872–877.

Chapman, G. B., & Sonnenberg, F. A. (2003). Decision making in health care: Theory, psychology and applications. In G. B. Chapman & F. A. Sonnenberg (Eds.). *Cambridge series on judgement and decision making*. Cambridge: Cambridge University Press.

Fauci, A. S., Braunwald, E., Kasper, D. L., Hauser, S. L., Longo, D. L., Jameson, J. L., & Al., E. (2009). *Featuring the Complete Contents of Harrison's Principles of Internal Medicine* (17th ed.). McGraw Hill. Harrison's Online.

Han, J., & Kamber, M. (2006). *Data Mining: Concepts and Techniques* (2nd ed.). Morgan Kaufmann.

Husmeier, D., Dybowski, R., & Roberts, S. (2004). *Probabilistic Modelling in Bioinformatics and Medical Informatics*. Springer.

Huanhuan Chen, Yao, X., & Tino, P. (2011). Ensemble Learning through Diversity Management: Theory, Algorithms, and Applications. *The 2011 International Joint Conference on Neural Networks*, 1(1), 1–6. Retrieved from http://www.cs.bham.ac.uk/~hxc/tutorial/

Ishwaran, H., & Rao, J. S. (2009). Decision Tree: Introduction. In M. Kattan (Ed.). *Encyclopedia of medical decision making*, 323-328, California: Sage Inc.

López-Vallverdú, J. A., Riaño, D., & Bohada, J. A. (2012). Improving medical decision trees by combining relevant health-care criteria. *Expert Systems with Applications*, 39(14), 11782–11791. doi:10.1016/j.eswa.2012.04.073.

Lucas, P., van der Gaag, L., & Abu-Hanna. (2004). Bayesian networks in biomedicine and health-care. *Artificial Intelligence in Medicine*, 30(3), 201–214.

Lavanya, D. & Rani, K. U. (2011). Performance evaluation of decision tree classifiers on medical datasets. *International Journal of Computer Applications*, 26(4), 2–5.

Mohamed, W. N. H. W., Salleh, M. N. M., & Omar, A. H. (2012). A comparative study of Reduced Error Pruning method in decision tree algorithms. *Proceedings of 2012 IEEE International Conference on Control System, Computing and Engineering*, 1(1), 392–397. doi:10.1109/ICCSCE.2012.6487177.

Parvin Hamid, MirnabiBaboli, M., & Alinejad-Rokny, H. (2015). Proposing a classifier ensemble framework based on classifier selection and decision tree. *Engineering Applications of Artificial Intelligence*, 37(1), 34–42. doi:10.1016/j.engappai.2014.08.005

Parvin, H., Ghatei, S., Alinejad-Rokny, H., & Minaei, B. (2011). An innovative combination of particle swarm optimization, learning automaton and great deluge algorithms for dynamic environments. *International Journal of Physical Sciences*, 6(22), 5121–5127.

Subramanian, S., Srinivasan, V. B., & Ramasamy, C. (2012). Study on classification algorithms for network intrusion systems. *Journal of Communication and Computer*, 9, 1242–1246.

Tao, Y. (2014). Computational verb decision trees. *International Journal of Computational Cognition*, 5(3), 57–62.

Velikova, M., de Carvalho Ferreira, N., & Lucas, P. (2007). Bayesian network decomposition for modeling breast cancer detection. *Artificial Intelligence in Medicine*, AIME 2007, 4594, 346–350.

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools & Techniques* (2nd ed.). San Francisco: Diane Cerra.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining Practical Machine Learning Tools and Techniques*. Burlington, USA: Elsevier.

Yeh, D., Cheng, C., & Chen, Y. (2011). A predictive model for cerebrovascular disease using data mining. *Expert Systems with Applications*, 37(7), 8970–8977.

Zhao, Y., & Zhang, Y. (2007). Comparison of decision tree methods for finding active objects. *Advances of Space Research*, 1(1), 1–10.