Decision Tree Performance Analysis on Medical Data

Stenly R. Pungus
Faculty of Computer Science
Universitas Klabat
Manado, Indonesia

Debby E. Sondakh
Faculty of Computer Science
Universitas Klabat
Manado, Indonesia
Email: debby.sondakh [AT] unklab.ac.id

Abstract—**Healthcare database keeps large quantities of data about patients and their medical records. These data contains hidden patterns that can be extracted into 'valuable' information for medical professionals in diagnosing a disease. Data mining is a powerful tool for analyzing data from different dimensions. Classification, a technique in data mining, also has been widely used to recognize disease over symptoms. This paper present a research aims to compare and evaluate different approaches of decision tree classification algorithms for healthcare datasets. The algorithms considered here are Alternating Decision Tree, Best First Tree, J48, J48graft, Logistic Model Tree, Random Forest, and Random Tree. The algorithms were applied on five multivariate healthcare datasets. Five important performance indicators for data mining algorithms were tested on resulted classifiers, i.e. accuracy, precision, mean absolute error and root mean squared error rates, and classifier training time. Among the seven algorithms, this study concludes the best algorithm for the chosen datasets is J48. J48 provides classifier with high accuracy and precision values. It also takes few times to build the classifier.**

Keywords- Classification, Decision Tree, Healthcare Dataset

I. INTRODUCTION

Health information system's database stores mass of patients' medical record, which contains valuable information in the form of patterns. These patterns describe health data relations, and can be used for providing better diagnosis. Data mining has been widely used in many fields to analyze mass amount of data in order to find the 'hidden' patterns in the data, then produce valuable and useful knowledge. Data mining is the process of searching for valuable information or knowledge from the dataset in automatic or semi-automatic manner [2]. Automatic data mining, also called clustering or supervised learning, means the learning process is independent from predefined class label. Otherwise, semi-automatic data mining, also called classification or supervised learning, depends on predefined class label by an expert. Classification has become an important 'tool' used for extracting useful knowledge from medical database. It is adopted to identify a disease based on existing symptoms. This study aims to analyze the performance of decision tree algorithms on medical dataset, using datasets from University of California Irvine (UCI) repository [3]. Classification was conducted using Waikato Environment for Knowledge Analysis (WEKA) data mining software [4]. Algorithms' performances were evaluated using five parameters, i.e. accuracy, precision, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and classification model building time.

This paper has four sections. First is the introduction, explaining in general about data mining and its application in health, as well as the issues examined in this study, and related research as well. Section 2 elaborating the methodology used. Section 4 explains the classification results on the specified datasets using decision tree algorithms. The last section concludes the result and analysis.

A. *Decision Tree*

Classification is defined as the process of searching for a function or model that differentiates group of labeled training data. The model then will be applied in predicting other unlabeled data [1]. Model may be built using several techniques such as decision tree, classification rules, neural network, and regression analysis. Decision tree depicts a structural description of a set of data. Using this approach, classification model is built by decomposing the data into a hierarchical structure, based on the attribute values. Figure. 1 shows an example of a decision tree. It comprises of

a. *Internal nodes*; represents the tested attribute.
b. *Edge*; edge coming out from an internal node represents the conditions of one attribute values. It is the test result.
c. *Leaf* ; is the category or class of data.

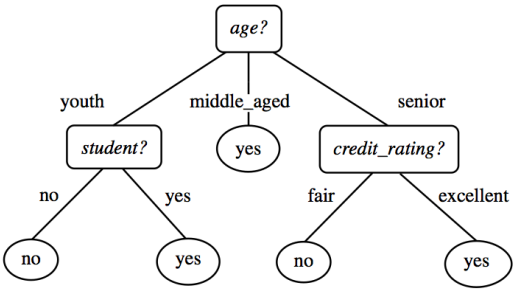

Figure 1. Decision Tree [1]

WEKA has 16 decision tree classifiers including Alternating Decision Tree (ADTree), Best First Tree (BFTree), Id3, J48, J48graft, Logistic Model Tree (LMT), NBTree, RandomForest (RF), Random Tree (RT), REPTree, and so on.

This study examined ADTree, BFTree, J48, J48graft, LMT, RF, and RT classifiers.

II. Related Research

A number of studies in evaluating classification techniques on medical datasets have been conducted. Akinola & Oyabugbe [5], Danjuma & Osofisan [6], Amin & Habib [7], Barnaghi, Sahzabi & Azuraliza [8], and Kumar & Sahoo [9] compared decision tree, Bayesian, and neural network on different datasets. The first three studies compared the J48 of decision tree, Naïve Bayes (NB) of Bayesian, and Multilayer Perceptron (MLP) of neural network, respectively on Ebola, Erythemato-squamous, and Hematological datasets taken from UCI repository, in terms of algorithms' accuracy and model building time. Result found that J48 is superior compare to the other two, and NB had the lowest performance [5,7]. J48's time taken to build the model was also the fastest [5]. On the other hand, Danjuma & Osofisan [6] discovered NB as the classifier with highest accuracy percentage.

Similar result was found by [9] when they investigated the performances of J48 decision tree with three Bayesian classifiers (Bayes Net, NB, and NB Updateable) and two neural network classifier (MLP and Voted Perceptron) on two datasets, i.e. Sick and Breast Cancer. The evaluated parameters were time and error rate. J48's got the smallest error rate, which means its accuracy is the higher. In terms of time, NB Updateable was the fastest. On the contrary, MLP is the slowest. Another comparison analysis by [8] also discovered J48 achieved the highest accuracy. Researcher compared J48 and LMT of decision tree, Bayes Net and NB of Bayesian, MLP and Radial Basis Function (RBF) of neural networks for classifying Liver Disorder data [8]. Similar to [5], this study aimed to find out whether classifier's performance is affected by training data size. Percentage split accuracy estimation method is applied. The results showed that classifiers' accuracy is fluctuated when the dataset's size increases. MLP, RBF, and J48 get the highest accuracy (79.41%) at 90/10.

Durairaj & Deepika [10] conducted a comparison accuracy and model building time between J48, NB, and lazy classifier lBk, applied to Leukimia Cancer dataset. All classifiers worked well in predicting leukemia cancer data. The lBk classifier is the fastest in build a model, but suffer in accuracy (82.35%) compare with NB and J48. NB builds the classification model in average of 0.16s with 91.17% of accuracy.

Gupta, Rawal, Narasimhan & Shiwani [11] compared another decision tree classifier, called J48graft, with Bayes Net, MLP, and JRip on Diabetes dataset. The highest percentage of accuracy, 81.33%, is the J48graft of decision tree.

III. Methodology

Figure. 2 depicts the methodology applied in this study. It comprises of four main steps, starting from data collection, followed by data preprocessing, data classification using WEKA tool, analyzing the classification results, and conclusion drawing.

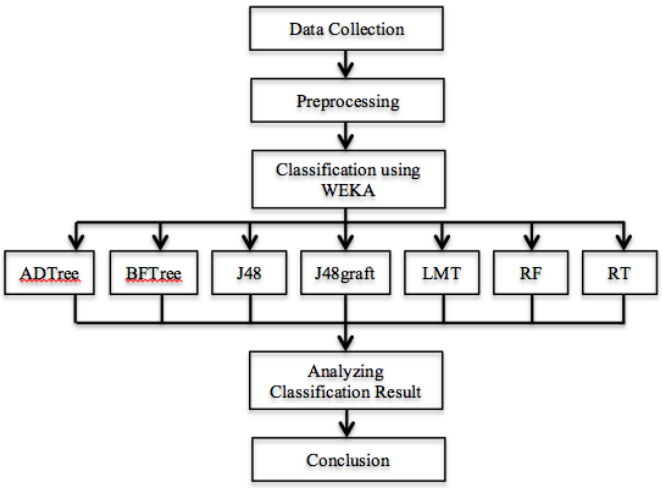

Figure 2. Methodology

At the first step, five medical dataset were collected from UCI repository [3], as listed in Table I.

TABLE I. Dataset Summary

Dataset	Number of Data	Number of Attributes
Echocardiogram	106	10
SPECT Heart	267	22
Chronic Kidney Disease	450	25
Mammographic Mass	961	6
Egg Eye State	14980	6

The next step is data preprocessing. All the datasets, but Chronic Kidney Disease, are availailabe in .txt format. Therefore, they have to be converted into format which is WEKA's format. The .txt dataset file was first converted into .csv using Ms.Excel. WEKA accept .csv file as well. Then, the .csv file was converted to .arff using WEKA.

IV. Result and Discussion

This section describes the analysis of decision tree classifiers resulting from classification process, using five parameters i.e. accuracy, precision value, time, error rates (Mean Absolute Error and Root Mean-Squared Error). Accuracy is percentage of data classifying correctly. Precision represents the ability of classifiers to put data as being under the correct category as opposed to all data in that category. It is defined as $P(\Phi(d_x, c_1) \quad T | \Phi(d_x, c_2) \quad T)$, conditional probability that a random object d is classified under c_1. MAE is measure the distance between the estimate and actual accuracy of each data. It is the total of absolute error divided by number of data on testing set that has the actual class labels. If the absolute error value were squared before it is averaged, then it resulting in the RMSE value. An ideal error rate has small MAE and RMSE values, in which the MAE must be smaller than RMSE.

Table II to VI show classification results of ADTree, BFTree, J48, J48graft, LMT, RF, and RT classifiers. Each table listed the five evaluated parameters of each dataset.

TABLE II. ECHOCARDIOGRAM DATASET RESULT CLASSIFICATION RESULT

ALGORITHMS	PARAMETER				
	ACCURACY	PRECISION	TIME	MAE	RMsE
ADTree	96.89%	0.965	0.02	0.307	0.312
BFTree	97.23%	0.97	0.3	0.221	0.278
J48	97.30%	0.974	0	0.0289	0.1157
J48graft	97.30%	0.974	0	0.0289	0.1157
LMT	95.95%	0.959	0.15	0.0366	0.124
RF	97.30%	0.973	0.013	0.0462	0.1249
RT	94.59%	0.946	0	0.0339	0.1763

TABLE III. SPECT DATASET RESULT CLASSIFICATION RESULT

ALGORITHMS	PARAMETER				
	ACCURACY	PRECISION	TIME	MAE	RMsE
ADTree	66.29%	0.659	0.03	0.4264	0.4647
BFTree	80.52%	0.778	0.33	0.275	0.3897
J48	80.90%	0.803	0.01	0.2422	0.3724
J48graft	70.41%	0.7	0.02	0.3745	0.4812
LMT	71.16%	0.71	0.49	0.3771	0.4544
RF	66.67%	0.661	0.02	0.374	0.4579
RT	66.29%	0.662	0	0.3567	0.5737

TABLE IV. CHRONIC KIDNEY RESULT CLASSIFICATION RESULT

ALGORITHMS	PARAMETER				
	ACCURACY	PRECISION	TIME	MAE	RMsE
ADTree	99.75%	0.998	0.023	0.0203	0.0539
BFT	97.00%	0.97	0.07	0.0397	0.1248
J48	99.00%	0.99	0.02	0.0225	0.0807
J48graft	98.75%	0.987	0.01	0.0244	0.0903
LMT	98.00%	0.981	0.84	0.0222	0.1068
RF	99.75%	0.998	0.017	0.037	0.0844
RT	95.50%	0.956	0	0.045	0.1677

TABLE V. MAMMOGRAPHIC MASS DATASET RESULT CLASSIFICATION RESULT

ALGORITHMS	PARAMETER				
	ACCURACY	PRECISION	TIME	MAE	RMsE
ADTree	82.83%	0.828	0.02	0.3195	0.3691

BFTree	81.99%	0.82	0.016	0.2511	0.371
J48	82.41%	0.824	0.03	0.2566	0.3631
J48graft	82.41%	0.824	0.01	0.2566	0.3631
LMT	83.66%	0.837	0.63	0.2359	0.3467
RF	78.04%	0.78	0.04	0.2487	0.401
RT	77.84%	0.778	0.01	0.2429	0.4429

TABLE VI. EGG EYE STATE DATASET CLASSIFICATION RESULT

ALGORITHMS	PARAMETER				
	ACCURACY	PRECISION	TIME	MAE	RMsE
ADTree	69.25%	0.691	1.6	0.4385	0.455
BFTree	84.38%	0.844	6.28	0.1857	0.3767
J48	84.50%	0.845	1.1	0.1691	0.3778
J48graft	84.75%	0.847	1.7	0.1669	0.3758
LMT	87.77%	0.878	279.99	0.1503	0.3128
RF	90.37%	0.906	1.18	0.1897	0.2758
RT	82.78%	0.828	0.13	0.1722	0.415

Comparison of accuracy percentage of the seven decision tree classifiers is presented at Figure. 3. RF classifier resulting models with the highest accuracy on three datasets (Echocardiogram, Chronic Kidney, and EEG Eye State), ADTree on Chronic Kidney, LMT on Mammographic Mass, and J48 on Echocardiogram and SPECT Heart. Classifiers performances are good with more than 80% average of accuracy, as follows: J48 **88.82%**, BFTree 88.22%, LMT 87.31%, J48graft 86.73%, RF 86.42%, RT 83.4%, and ADTree 83%.

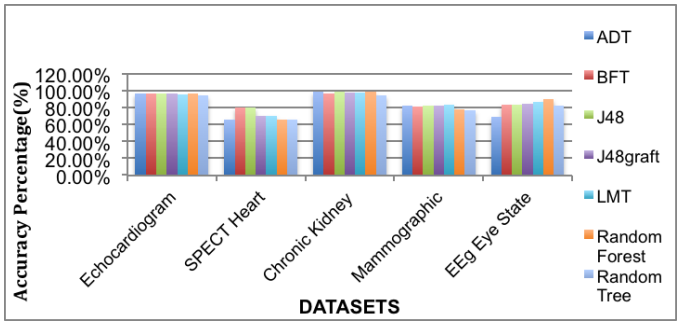

Figure 3. Accuracy

Similar results were found in precision values as shown in Figure. 4. RF classifier' produced a model with the highest precision values on Chronic Kidney 0.998 and EEG Eye State 0.906, ADTree on Chronic Kidney (0.998), LMT on Mammographic Mass (0.837), and J48 on two datasets Echocardiogram (0.974) and SPECT Heart (0.803). On average, J48 is the highest with **0.89** point, followed by BFTree 0.88, LMT and J48 graft 0.87, RF 0.86, RT and ADTree 0.83.

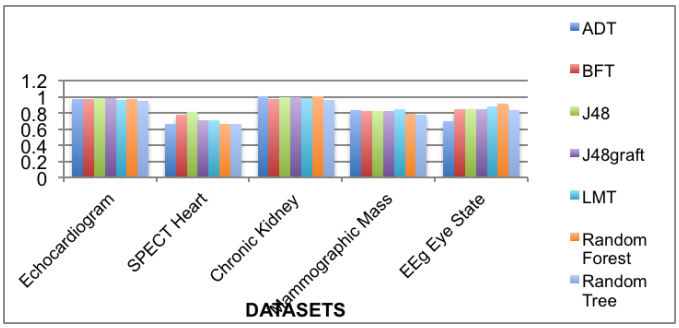

Figure 4. Precision

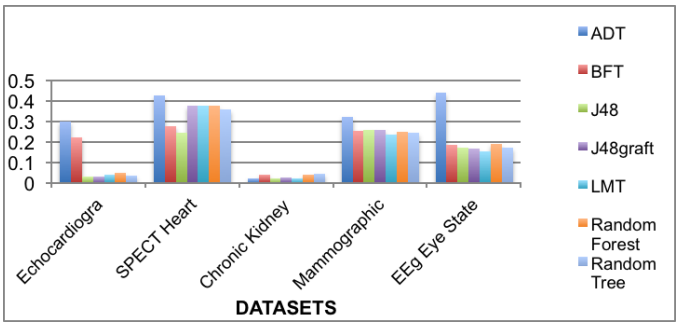

Figure 5. Error Rate – MAE

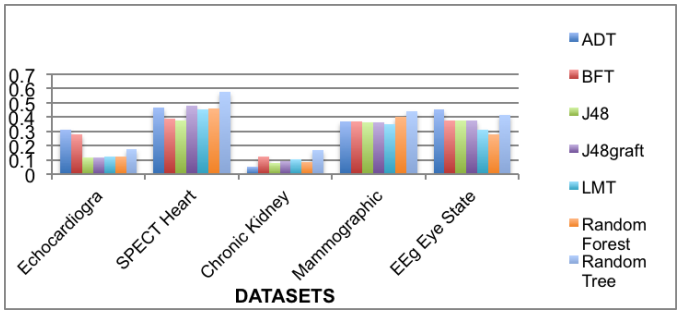

Figure 6. Error Rate - RMSE

Another parameter that is used to evaluate classifiers' performance is error rate. Figure. 5 and Figure. 6 present the MAE and RMSE of the resulting models. Low error rate means the model has high accuracy. J48 gives results with the lowest average MAE 0.14, while ADTree gives the highest average 0.3. As for RMSE, the J48 classifier's is the lowest with 0.14 and ADTree's is the highest with 0.36.

The last parameter evaluated to consider the best classifier among the seventh is time. It is shown as a graphical representation in Figure. 7. The graph in Figure. 7 represents the model building time of all classifiers. LMT requires longer time compare to the others. It spent 279.99 seconds to classify EEG Eye State, the biggest dataset (see Table VI). Classifying the medical datasets using LMT and BFT took long time. In more detail Figure. 8 illustrates ADTree, J48, J48graft, RF and RT time performance.

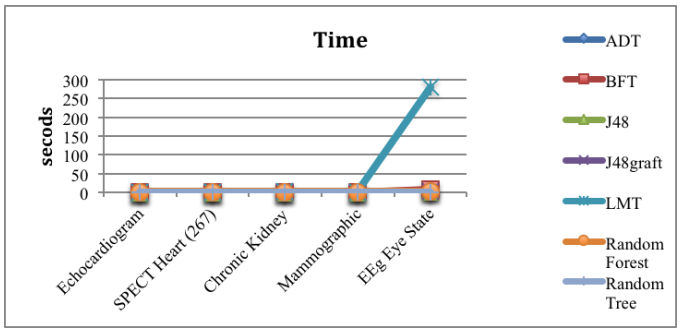

Figure 7. Model Building Time (a)

Overall, we can see that RT classifier is the fastest. RT requires average of 0.03 seconds, followed by J48 with average of 0.23 seconds, RF 0.25 seconds, J48graft 0.35 seconds, ADTree 0.33 seconds, BFTree 1.4 seconds, and LMT 56.42 seconds.

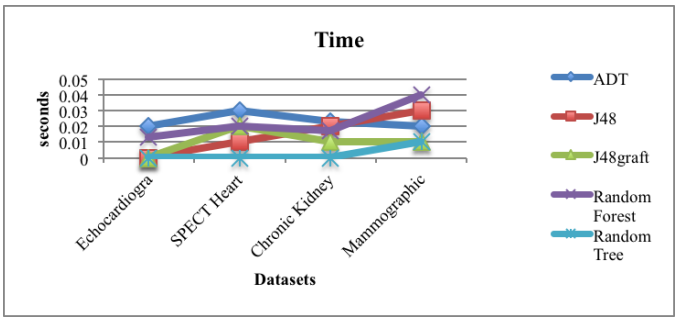

Figure 8. Model Building Time (b)

Table VII summarizes the results in terms of the best average accuracy, precision, error rates, and time. *Italic* format means the classifiers in the same columns rankings' are the same. For example, in column Precision, *LMT* and *J48graft* share the same ranking. From the results obtained after applying different classification algorithms on given datasets J48 showed the best accuracy compare to the other six classifiers. Otherwise, ADTree's results indicate that it is not good enough in classifying the given medical datasets.

TABLE VII. CLASSIFICATION RESULT SUMMARY

Ranking	Parameter				
	Accuracy	**Precision**	**MAE**	**RMSE**	**Time**
1	J48	J48	J48	J48	RT
2	BFTree	BFTree	LMT	*LMT*	J48
3	LMT	*LMT*	*J48graft*	*RF*	RF
4	J48graft	*J48graft*	*RT*	J48graft	J48graft
5	RF	RF	RF	BFTree	ADTree
6	RT	*RT*	BFTree	ADTree	BFTree
7	ADTree	*ADTree*	ADTree	RT	LMT

V. CONCLUSION

Classification has been conducted on five medical dataset, using seven decision tree algorithms in Weka, to measure and compare algorithms' performance in classifying health data. Analysis was carried out on five parameters, namely accuracy, precision, time taken to build the models, as well as the error rates. The result analysis then concluded as follows

1. J48 produces a more accurate classification model. Its performance is the highest compare to the other six algorithms, with an average accuracy of 88.82%, 0.89 precision value, and average error rate MAE and RMsE respectively 0.14 and 0.28. J48 requires an average of 0.23 seconds to build the classification model.
2. The classification results also discover that the J48 and LMT's model building time is directly proportional to the dataset's size. As for the other algorithms, the time fluctuated as the dataset increases.

REFERENCES

[1] J.Han & M. Kamber, *Data Mining Concepts and Techniques*, Academic Press,USA, 2001.

[2] I. H. Witten & Eibe Frank, *Data Mining Practical Machine Learning Tools and Techniques*, Edisi Kedua, Morgan Kaufmann Publishers, 2005.

[3] UCI. Availabel: https://archive.ics.uci.edu/ml/datasets.html

[4] WEKA. Available: http://www.cs.waikato.ac.nz/~ml/weka.

[5] S. O. Akinola and O. J. Oyabugbe, "*Accuracies and Training Time of Data Mining Clasification Algorithms: an Empirical Comparative Study*", *Journal of Software Engineering and Applications*, vol. 8, pp. 470-477, September 2015.

[6] K. Danjuma and A. Osofisan, "*Evaluation of Predictive Data Mining Algorithms in Erythemato-Squamous Disease Diagnosis*",

[7] M. N. Amin and M. A. Habib, "*Comparison of Different Classificaiton Techniques Using WEKA for Hematological Data*", *American Journal of Engineering Research*, vol. 4 (3), pp. 55-61, 2015.

[8] P. M. Barnaghi, V. A. Sahzabi, A. A. Bakar, "*A Comparative Study for Various Methods of Classification*", in *Proc. of Int. Conf. on Informatin and Computer Networks, Singapore, 2012.*

[9] Y. Kumar and G. Sahoo, "*Analysis of Bayes, Neural Network and Tree Classifier of Classification Technique in Data Mining using WEKA*", 2012.

[10] M. Durairaj and R. Deepika, "*Comparative Analysis of Classificatin Algorithms for the Prediction of Leukimia Cancer*", *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 5 (8), August 2015.

[11] N. Gupta, A. Rawal, V. L. Narasimhan, and S. Shiwani, "*Accuracy, Sensitivity and Specifity Measurement of Various Classificatin Techniques on Healthcare Data*", *IOSR Journal of Computer Engineering*, vol. 11 (5), pp. 70-73, May-June 2013.

[12] V. Mala and D. K. Lobiyal, "*Evaluation and Performance of Classification Methods for Medical Data Sets*", *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 5, issue 11, pp. 336-340, November 2015.

[13] S. Roy and A. Mohapatra, "*Performance Analysis of Machine Learning Techniques in Micro Array Data Classification*", *International Journal of Software and Web Sciences*, Vol. 4 (1), pp. 20-25, March-May 2013.