Performance Evaluation of Decision Tree Classifiers on Medical Datasets

D.Lavanya
Research Scholar
Sri Padmavathi Mahila Visvavidyalayam
Tirupati - 2 , Andhra Pradesh

Dr. K.Usha Rani
Dept. of Computer Science
Sri Padmavathi Mahila Visvavidyalayam
Tirupati - 2 , Andhra Pradesh

ABSTRACT
In data mining, classification is one of the significant techniques with applications in fraud detection, Artificial intelligence, Medical Diagnosis and many other fields. Classification of objects based on their features into pre-defined categories is a widely studied problem. Decision trees are very much useful to diagnose a patient problem by the physicians. Decision tree classifiers are used extensively for diagnosis of breast tumour in ultrasonic images, ovarian cancer and heart sound diagnosis. In this paper, performance of decision tree induction classifiers on various medical data sets in terms of accuracy and time complexity are analysed.

Keywords— Data Mining, Classification, Decision Tree Induction, Medical Datasets.

1. INTRODUCTION
Classification is one of the fundamental tasks in data mining and has also been studied extensively in statistics, machine learning, neural networks and expert systems over decades [1,2]. The input for classification is a set of training records (training instances), where each record has several attributes. Attributes with discrete domains are referred to as Categorical, while those with continuous domains are referred to as numerical. There is one distinguished attribute called the class label. In general, given a database of records, each with a class label, a classifier generates a concise meaningful description for each class in terms of the attributes. The model is then used to predict class labels of unknown objects. Classification is also known as supervised learning, as the learning of the model is "supervised", that is, each training instance is labelled indicating its class. Classification has been successfully applied to a wide range of application areas, such as scientific experiments, medical diagnosis, weather prediction, credit approval, customer segmentation, target marketing and fraud detection [3,4]. Decision tree classifiers are used extensively for diagnosis of breast tumour in ultrasonic images, ovarian cancer, heart sound diagnosis and so on [5-10].

Data Mining with Decision trees plays a vital role in the field of medical diagnosis to diagnose the problem of a patient. In this paper, accuracy of various decision tree classifiers and their time complexity are compared on Medical Data sets. Decision tree classifiers are chosen as they [1]

- Provide human readable rules of classification
- Easy to interpret
- Construction of decision tree is fast
- Yields better accuracy

The rest of the paper is organized in three sections. In Section 2, the review of decision tree induction algorithms are presented. Related to medical data sets, the performance of most frequently used decision tree classifiers are compared and the results are presented in section 3and concluded in section 4.

2. DECISION TREE INDUCTION
Decision tree induction is a very popular and practical approach for pattern classification. Decision tree induction is the learning of decision trees from class-labelled training tuples.

A decision tree is a flow chart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node holds a class label.

The decision tree classifier has two phases [1]:

i) Growth phase or Build phase.
ii) Pruning phase.

The tree is built in the first phase by recursively splitting the training set based on local optimal criteria until all or most of the records belonging to each of the partitions bearing the same class label. The tree may overfit the data.

The pruning phase handles the problem of over fitting the data in the decision tree. The prune phase generalizes the tree by removing the noise and outliers. The accuracy of the classification increases in the pruning phase.

Pruning phase accesses only the fully grown tree. The growth phase requires multiple passes over the training data. The time needed for pruning the decision tree is very less compared to build the decision tree.

The table specified below represents the usage frequency of various decision tree algorithms [11].

Table 1-Frequency usage of decision tree algorithms

Algorithm	Usage frequency (%)
CLS	9
IDE	68
IDE3+	4.5
C4.5	54.55
C5.0	9
CART	40.9
Random Tree	4.5
Random Forest	9
SLIQ	27.27
PUBLIC	13.6
OCI	4.5
CLOUDS	4.5

By observing the above table the frequently used decision tree algorithms are ID3, C4.5 and CART. Hence, the experiments are conducted on the above three algorithms.

2.1 ID3 (Iterative Dichotomiser 3)

This is a decision tree algorithm introduced in 1986 by Quinlan Ross [12]. It is based on Hunts algorithm. The tree is constructed in two phases. The two phases are tree building and pruning

ID3 uses information gain measure to choose the splitting attribute. It only accepts categorical attributes in building a tree model. It does not give accurate result when there is noise. To remove the noise pre-processing technique has to be used.

To build decision tree, information gain is calculated for each and every attribute and select the attribute with the highest information gain to designate as a root node. Label the attribute as a root node and he possible values of the attribute are represented as arcs. Then all possible outcome instances are tested to check whether they are falling under the same class or not. If all the instances are falling under the same class, the node is represented with single class name, otherwise choose the splitting attribute to classify the instances.

Continuous attributes can be handled using the ID3 algorithm by discretizing or directly, by considering the values to find the best split point t by taking a threshold on the attribute values. Id3does not support pruning.

2.2 C4.5

This algorithm is an extension to ID3 developed by Quinlan Ross [13]. It is also based on Hunt's algorithm.C4.5 handles both categorical and continuous attributes to build a decision tree.In order to handle continuous attributes, C4.5 splits the attribute values into two partitions based on the selected threshold such that all the values above the threshold as one child and the remaining as another child. It also handles missing attribute values. C4.5 uses Gain Ratio as an attribute selection measure to build a decision tree. It removes the biasness of information gain when there are many outcome values of an attribute.

At first, calculate the gain ratio of each attribute. The root node will be the attribute whose gain ratio is maximum. C4.5 uses pessimistic pruning to remove unnecessary branches in the decision tree to improve the accuracy of classification.

2.3 CART

CART [14] stands for Classification And Regression Trees introduced by Breiman. It is also based on Hunt's algorithm. CART handles both categorical and continuous attributes to build a decision tree. It handles missing values.

CART uses Gini Index as an attribute selection measure to build a decision tree .Unlike ID3 and C4.5 algorithms, CART produces binary splits. Hence, it produces binary trees. Gini Index measure does not use probabilistic assumptions like ID3, C4.5. CART uses cost complexity pruning to remove the unreliable branches from the decision tree to improve the accuracy.

3. EXPERIMENTAL RESULTS

The Experimental data is collected from UCI Machine Learning Repository [15], which is publicly available. The results were analysed using Weka tool on the data using 10-fold cross validation to test the accuracy and time complexity of ID3, C4.5 and CART classifiers. The following table shows the characteristics of selected datasets related to medical domain.

Table 2- Data Set Characteristics

Data Set	No. of Attributes	No. of Classes	No. of Instances	Missing Values
Diabetes	8	2	768	No
Heart Stat log	13	2	270	No
Thyroid	28	6	9172	Yes
Breast Cancer	10	2	699	Yes
Arrhythmia	278	16	452	Yes

The above datasets contain both continuous and discrete attributes, where as ID3 algorithm cannot handle the continuous attributes. To evaluate the performance of the algorithms without any bias (for uniformity) discretization is done to convert continuous attributes in to categorical attributes. The type of discretization performed here is unsupervised discretization because supervised discretization produces complex search spaces. Some of the datasets contain missing values. Missing values cannot be handled by ID3 algorithm. So, pre-processing is done to replace the missing values with mean of the respective attributes. The Table 3 shows the accuracy of ID3, C4.5 and CART algorithms for classification applied on the above medical data sets using 10-fold cross validation is observed as follows:

Table 3-Classifiers Accuracy

Data Set	Accuracy (%)		
	ID3	C4.5	CART
Diabetes	57.5	73.8	75.1
Heart Statlog	61.4	76.6	78.5
Thyroid	65.60	67.92	69.16
Breast Cancer	90.41	94.56	94.84
Arrhythmia	42.69	64.38	70.57

The Table 4 shows the time complexity in seconds of various classifiers to build the model for training data.

Table 4-Execution Time to Build the Model

Data Set	Time(Secs)		
	ID3	C4.5	CART
Diabetes	0.03	0.08	0.36
Heart Statlog	0.01	0.06	0.13
Thyroid	1.41	4.34	47.83
Breast Cancer	0.01	0.09	0.44
Arrhythmia	0.38	1.47	5.69

To observe the performance of the classifiers on large data sets, only two data sets: Diabetes and Thyroid with increased size are considered for experiment. The performance in terms of accuracy and time complexity are presented in table 5 and Table 6.

Table 5-Enhanced Datasets - classifiers Accuracy

Data Set	Accuracy (%)		
	ID3	C4.5	CART
Diabetes	84.52	96.24	99.45
Thyroid	76.99	92.44	94.68

Table 6-Enhanced Datasets - Execution Time to Build the Model

Data Set	Time(Secs)		
	ID3	C4.5	CART
Diabetes	10.05	25.03	150.23
Thyroid	12.63	40.83	165.52

The classifiers accuracy on various data sets is represented in the form of a graph.

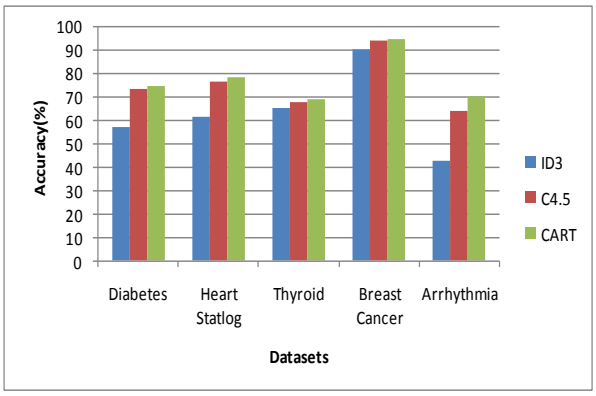

Fig 1: Comparison of Classifiers Accuracy

By observing the Experimental analysis, CART algorithm yields better accuracy compared to ID3 and **C4**.5 for both small and large data sets. The time complexities(in seconds) to build a decision tree model using ID3, C4.5 and CART classifiers on medical data sets are represented pictorially in figure2.

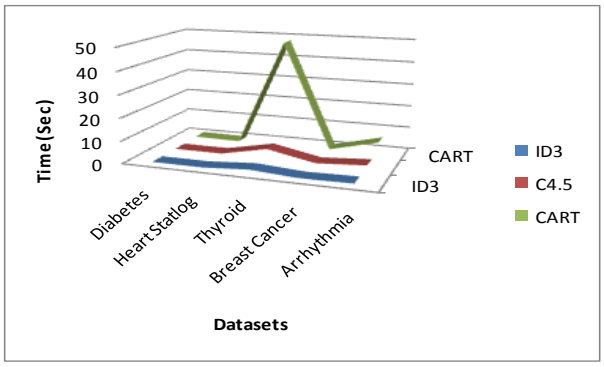

Fig 2: Time Complexities of Classifiers to Build the Model

The Figure.2 shows that time complexity of ID3 algorithm is less to build a model among the three classifiers.
Coming to the accuracy, CART algorithm produces better accuracy though the time complexity is high. Accuracy is more important for the classification of medical data. Hence, CART is the best algorithm for medical diagnosis.

4. CONCLUSIONS

Data Mining is gaining its popularity in almost all applications of real world. One of the data mining techniques i.e., classification is an interesting topic to the researchers as it is accurately and efficiently classifies the data for knowledge discovery. Decision trees are so popular because they produce human readable classification rules and easy to interpret than other classification methods.

Frequently used decision tree classifiers are studied and the experiments are conducted to find the best classifier for Medical Diagnosis. The experimental results show that CART is the best algorithm for classification of medical data. It is also observed that CART performs well for classification on medical data sets of increased size.

5. REFERENCES

[1] J. Han and M. Kamber, "Data Mining; Concepts and Techniques, Morgan Kaufmann Publishers", 2000.

[2] T. Mitchell, "Machine Learning", McGraw Hill, 1997.

[3] R. Brachman, T. Khabaza, W.Kloesgan, G.Piatetsky-Shapiro and E. Simoudis, "Mining Business Databases", Comm. ACM, Vol. 39, no. 11, pp. 42-48, 1996.

[4] U.M. Fayyad, G. Piatetsky-Shapiro and P. Smyth, "From Data Mining to knowledge Discovery in Databases", AI Magazine, vol 17, pp. 37-54, 1996.

[5] Antonia Vlahou, John O. Schorge, Betsy W.Gregory and Robert L. Coleman, "Diagnosis of Ovarian Cancer Using Decision Tree Classification of Mass Spectral Data", Journal of Biomedicine and Biotechnology • 2003:5 (2003) 308–314.

[6] Kuowj, Chang RF,Chen DR and Lee CC," Data Mining with decision trees for diagnosis of breast tumor in medical ultrasonic images" ,March 2001.

[7] H. Ren, "Clinical diagnosis of chest pain," Chinese Journal for Clinicians, vol. 36, 2008.

[8] My Chau Tu, Dongil Shin, Dongkyoo Shin, "A Comparative Study of Medical Data Classification Methods Based on Decision Tree and Bagging Algorithms", DASC '09 Proceedings of the 2009 Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing, IEEE Computer Society Washington, DC, USA ©2009.

[9] Sung Ho Ha and Seong Hyeon Joo, "A Hybrid Data Mining Method for the Medical Classification of Chest Pain", World Academy of Science, Engineering and Technology 70 2010.

[10] Matthew N.Anyanwu, Sajjan G.Shiva, "Comparative Analysis of Serial Decision Tree Classification Algorithms", International Journal of Computer Science and Security, volume 3.

[11] G Stasis, A.C. Loukis, E.N. Pavlopoulos, S.A. Koutsouris, D. "Using decision tree algorithms as a basis for a heart sound diagnosis decision support system", Information Technology Applications in Biomedicine, 2003. 4th International IEEE EMBS Special Topic Conference, April 2003.

[12] Quinlan, J.R, "Induction of decision trees". Journal of Machine Learning 1(1986) 81-106.

[13] J.R.Quinlan,"c4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, Inc, 1992.

[14] Breiman, Friedman, Olshen, and Stone. "Classification and Regression Trees", Wadsworth, 1984., Mezzovico, Switzerland.

[15] UC Irvine Machine Learning Repository, www.ics.uci.edu/~mlearn/MLRepository.html