Effect of Training Set Size in Decision Tree Construction by using GATree and J48 Algorithm

Namdeo Venkatrao Kalyankar

***Abstract*—The Genetic Algorithm based approach provided by GATree has been used to evolve optimal decision trees for classification problems in data mining applications. The standard databases for data mining available on the Internet have been used in this work. The generated decision trees have been compared with those obtained J48 algorithm. The effects of various parameters and training set size on the classification accuracy and tree size have been analyzed.**

***Index Terms*— Genetic Algorithm, Data Mining, Decision Tree, Weka.**

I. Introduction

A typical database user retrieves data from databases using an interface to standard technology such as Structured Query Language (SQL). A data mining system takes this process a step further, allowing users to discover new knowledge from the data. Data mining, from a computer scientist's point of view, is an interdisciplinary field. Data handling techniques such as neural networks, genetic algorithms, regression, statistical analysis, machine learning and cluster analysis are prevalent in the literature on data mining [3][5].

Genetic algorithms are search algorithms based on the mechanics of natural selection and natural genetics. They are proven to provide robust search in complex spaces. These algorithms are computationally simple yet powerful in their search for improvement. They are now finding more widespread applications in business, scientific, and engineering circles [6].

Genetic Algorithm is a population based search technique. It uses operators such as selection, crossover and mutation. It is an iterative method where the search continues until desired criteria are satisfied.

1. A decision tree is a predictive modeling technique used in classification, clustering and prediction tasks. Decision trees use a '*divide and conquer'* technique to split the problem search space into subsets. The strategy to construct trees in decision tree algorithm is based on following three points: Criteria for selecting a variable to split
2. Criteria to find the split point/points of a selected variable
3. Criteria to decide when to stop the growing process of the tree.

Manuscript received March 06, 2018; revised March 30, 2018.

Namdeo Venkatrao Kalyankar, Vice-Chancellor, Gondwana University, Gadchiroli, Maharashtra, India phone: +91-7132-222221; fax: +91-7132-223105 (e-mail: drkalyankarnv@yahoo.com).

If the target variable or response variable or class variable is a nominal/categorical variable then it is called as *classification tree*. If the target variable is continuous then the tree is called as *regression tree*[1,2][4][7].

In this paper we use Genetic Algorithm to evolve optimal decision trees for classification problems in data mining. The emphasis will be on evolving decision trees which gives higher classification accuracy and are small in size.

A large number of data sets are available on the Internet for data mining. One such source is UCI Repository of machine learning datasets [10]. The data sets used in this work include *Iris*, *Students*, *Votes*, *Contact Lenses* and *Labor*. The *Iris* and *Contact Lenses* data sets are selected from the UCI Repository; whereas, the other data sets are form the WEKA's data library [11]. These data sets often include noisy records. As a result, there are limitations on achievable efficiency of classification.

II. Data sets and tools

The data is important for any organization. It can be any type such as numerical, character, etc. The UCI Repository, WEKA provides several data sets that can be used for the research in decision trees. I can also use data generation software to generate the data [10].

A. Weka Software

WEKA is a comprehensive tool for machine learning and data mining. Weka was developed at the University of Waikato in New Zealand. "Weka" stands for the Waikato Environment for Knowledge Analysis [11]. Weka provides implementations of state-of-the-art learning algorithms that you can apply to your dataset. It also includes a variety of tools for transforming datasets, like the algorithms for discretization. We can preprocess a dataset, feed it into a learning scheme, and analyze the resulting classifier and its performance.

Suppose we have some data and we want to build a decision tree from it. A common situation is for the data to be stored in a spreadsheet or database. However, Weka expects it to be in ARFF format because it is necessary to have type information about each attribute, which cannot be automatically deduced from the attribute values. An ARFF (Attribute-Relation File Format) file is an ASCII text file that describes a list of instance sharing a set of attributes. ARFF files were developed for Weka machine learning Software.

Weka contains tools for classification, regression, clustering, association rules, visualization, and data pre-

ISBN: 978-988-14047-9-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

processing. Weka is open source software under the GNU GPL. It is easily extensible, which allows researchers to contribute new learning algorithms to Weka, keeping it up-to-date with the latest developments in the field. As a result, Weka has become very popular with academic and industrial researchers, and is also widely used for teaching purposes. Figure 2.1 shows the classification of the tree in WEKA.

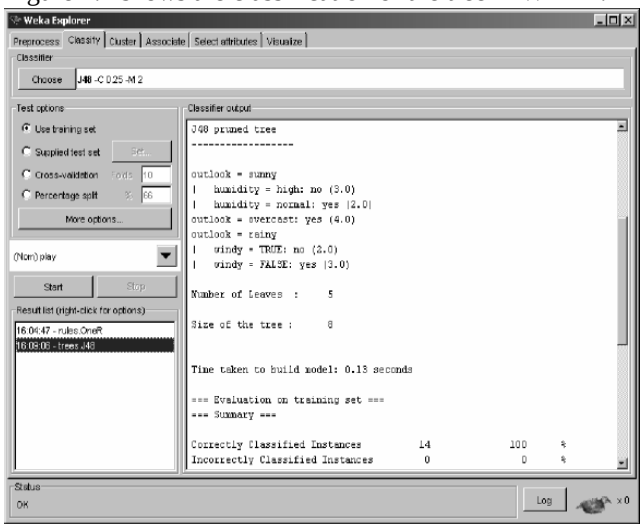

Fig 2.1 Classification of the tree in WEKA

B. *GATree Software*

GATree is a decision tree builder that is based on Genetic Algorithms (GAs). The idea behind it is rather simple but powerful. Instead of using statistic metrics that are biased towards specific trees we use a more flexible, global metric of tree quality that try to optimize accuracy and size [8,9]. GATree offers some unique features not to be found in any other tree inducers while at the same time it can produce better results for many difficult problems. Figure 2.2 shows the screen shot from the program's interfaces that uncover its basic functionality.

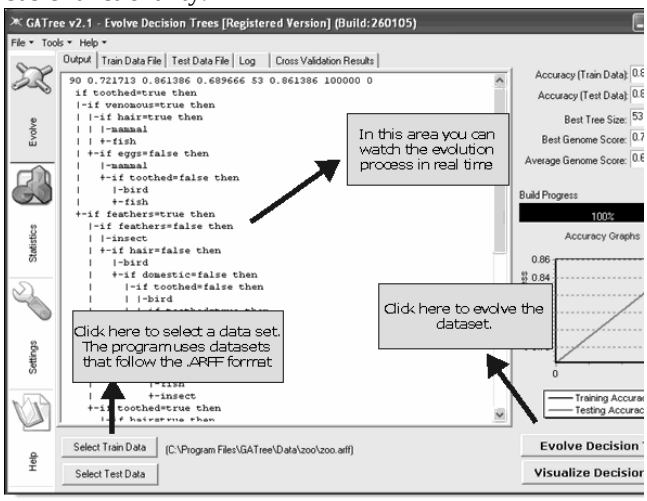

Fig 2.2 The main evolution tab of GATree

A few unique characteristics of GATree are as follows:

a) The user can control the characteristics of the output (More Accurate vs Smaller Trees)
b) There is no upper bound on its results given that we can provide the system with unlimited processing power and time
c) The system evolves complete solutions to the problem. We can stop evolution whenever the results are satisfactory
d) The system evolves a set of possible solutions (e.g, decision trees) that closely match the input data. This gives us alternative hypotheses for the same data

III. Experimental Results and Analysis

The WEKA and GATree software are used to construct decision trees on various data sets provided in the UCI Repository and WEKA database. To understand the effectiveness of GA in the decision tree construction process, the decision trees constructed using J48 algorithm provided in WEKA software are compared with the trees obtained using GA-based approach provided in GATree software. The comparison is with respect to the classification accuracy and the tree size.

Experiments are also conducted to understand the effect of training set size on the quality of constructed trees. The experiments conducted include the training as well as cross-validation.

A. *Experiment using WEKA*

Experiments are conducted to study the effect of training set size on the quality of generated trees. The decision tree is constructed using training sets that comprise of limited number of samples namely 5%, 15%, 30%, 50%, and 75%. The result of each experiment depends on the seed used for random number generator in the process of selection of training subset. Hence, to get a correct estimation of classification accuracy, each experiment was conducted five times with seed values from 1 to 5 and the average of the classification accuracies in these runs is determined. Other options in J48 algorithm in "More Options …" window were set to default values. In particular, the "Preserve order for % split" was unselected.

Table 3.1 and figure 3.1 shows the Iris dataset with splitting on different percentages.

Table 3.1 The classification results of J48 (C4.5) algorithm for Iris dataset.

Dataset	Seed	Classification Accuracy				
		5% Split	15% Split	30% Split	50% Split	75% Split
IRIS	1	44.05	94.53	95.23	94.66	94.73
	2	67.83	80.46	93.33	94.66	94.73
	3	55.24	97.65	95.23	94.66	94.73
	4	63.63	93.75	93.33	94.66	100.00
	5	30.76	93.75	93.33	90.66	92.10
Average		52.30	92.03	94.09	93.86	95.26

Such like this I also uses the datasets of Students, Votes, Contact Lenses and Labor for making the decision trees. Finally the graph i.e. figure 3.2 is generated which represents the accuracy and split of the datasets.

ISBN: 978-988-14047-9-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

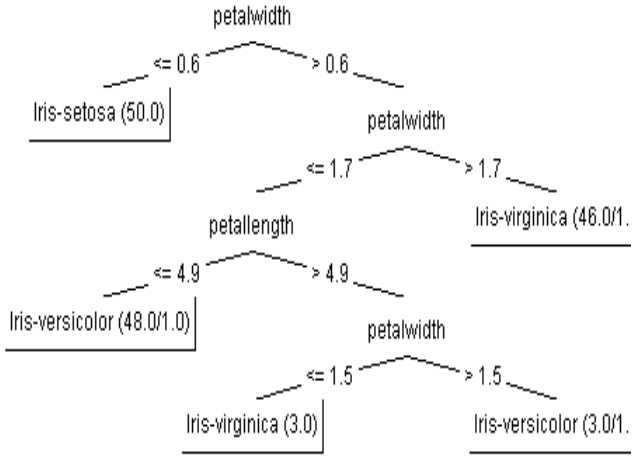

Fig 3.1 The decision tree generated by J48 algorithm for Iris dataset.

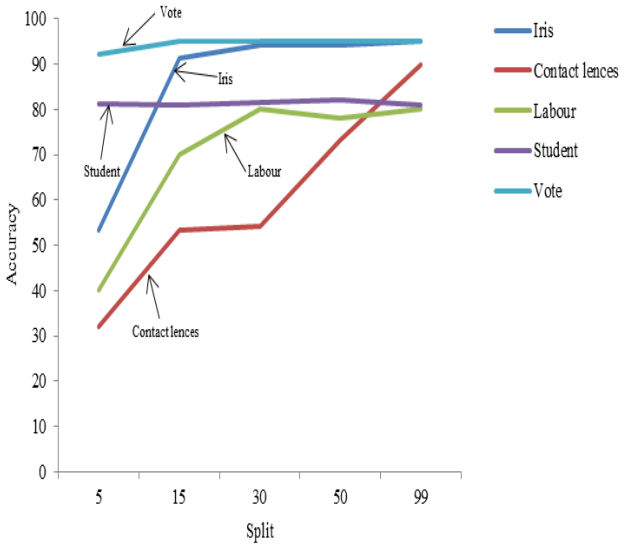

Fig 3.2 Graphical results for J48 algorithm on various datasets.

B. *Experiment using GATree*

The GATree is GA implementation for construction of decision trees. As GA is a robust search and optimization technique, I expect to get improvements over the results obtained using the J48 (C4.5) algorithm in WEKA software. Hence, the second experiment is conducted in which GATree is used to perform cross-validation for various datasets. The J48 algorithm from WEKA software is also used to get the cross-validation results on these datasets for the purpose of comparison.

The parameter for accuracy Vs. size preference of GATree was set to 50% in this experiment and all other parameters were set to their default values. The results are summarized in Table 3.2 along with the results of J48 algorithm for five datasets used earlier and graphically represented in figure 3.3. It can be observed that the GATree results are more accurate (at the cost of larger tree size) except for the Students dataset where the GATree result is slightly inferior. Also observe that the GATree has obtained a tree having higher accuracy and less size in case of Contact Lenses dataset. These results clearly demonstrate the superiority of the Genetic Algorithmic approach for the evolution of decision trees.

Table 3.2 Comparison results of J48 (C4.5) algorithm with GATree

SN	DataSet	J48 algorithm		GATree	
		Accuracy	Size	Accuracy	Size
1	Iris	94.8	11	97.04	13
2	Contact-Lenses	94.5	45	95.49	9
3	Labor	88.4	13	94.23	21
4	Students	87.3	19	86.64	37
5	Vote	95.6	7	96.94	39

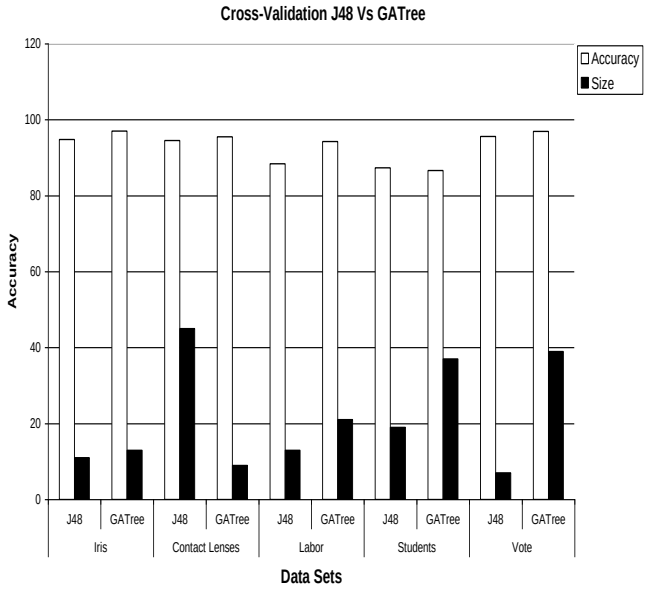

Fig 3.3 Graphical representation of comparison of cross-validation results of J48 algorithm and GATree on various datasets

IV. Conclusions

Although the classification accuracy depends on the number of instances and the distribution of the samples in various classes, it is not necessary to use the entire datasets for the construction of decision trees. For small datasets, 15% to 30% of the instances also give good classification accuracy. For large datasets, I obtain a very good accuracy even if we use as few as 5% instances for tree construction. This conclusion is particularly important as the databases available with the companies may have huge number of instances and databases having several million records are not very uncommon.

The Genetic Algorithmic approach has given better results compared to the J48 algorithm. This clearly demonstrates the capabilities of GA-based approach and we should use GA-based implementations to obtain decision trees that give better classification accuracy or the trees that are smaller in size. Further, GA gives multiple solutions with varying classification accuracies and tree sizes and we can select the one as per our needs.

References

[1] Bohanec, M., and Bratko,L., 1994, 'Trading Accuracy for Simplicity in Decision Trees', Machine Learning, Vol. 3, pp. 223-250.

ISBN: 978-988-14047-9-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

[2] Esposito, F., Malerba, D., Semerao, G., 1997, 'A Comparative Analysis of Methods for Pruning Decision Trees', IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 19, pp. 475-490.
[3] Gehrke, J.E., Ramakrishnan, R., and Ganti, V., 2000, 'RainForest- A Framework for Fast Decision Tree Construction of Large Datasets', Data Mining and Knowledge Discovery, Vol. no. 4, pp. 127-162.
[4] Kusiak, A., Kern, J.A., Kernstine, K.H., Tseng, T.L., 2000, 'Autonomous decision-making: a data mining approach', IEEE Transactions on Information Technology in Biomedicine, Vol. 4, pp. 274 – 284.
[5] Olaru, C., Wehenkel, L., 1999, 'Data mining', IEEE Tran. on Computer Applications in Power, Vol. 12, pp. 45-63.
[6] Papagellis, A., Kalles, D., 'GATree: Genetically Evolved Decision Trees', Computer Technology Institute, Partras, Greecce.
[7] Quinlan, J .R., 1986, 'Induction of Decision Trees', Machine Learning, Vol. no. 1, pp. 81-106.
[8] Wall, M., 1996, 'GAlib: A C++ Library of Genetic Algorithm Components', M.I.T. Vol. no. 4.
[9] Website of GATree software: http://www.gatree.com
[10] Website of UCI Machine learning repository: http://archive.ics.uci.edu/ml/
[11] Website of WEKA Software: http://www.cs.waikato.ac.nz

ISBN: 978-988-14047-9-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)