# A Novel Decision Tree Algorithm for Numeric Datasets - C 4.5*Stat

Sudheep Elayidom.M
Associate Professor
School of Engineering
Cochin University,India
Email:sudheep@cusat.ac.in

Sumam Mary Idicula
Professor
Department of Computer Science
Cochin University,India
Email:sumam@cusat.ac.in

Joseph Alexander
Project officer
Nodal centre
Cochin University,India
Email:josephalexander@cusat.ac.in

## ABSTRACT

Decision trees are very powerful tools for classification in data mining tasks that involves different types of attributes. When coming to handling numeric data sets, usually they are converted first to categorical types and then classified using information gain concepts. Information gain is a very popular and useful concept which tells you, whether any benefit occurs after splitting with a given attribute as far as information content is concerned. But this process is computationally intensive for large data sets. Also popular decision tree algorithms like ID3 cannot handle numeric data sets. This paper proposes statistical variance as an alternative to information gain as well as statistical mean to split attributes in completely numerical data sets. The new algorithm has been proved to be competent with respect to its information gain counterpart C4.5 and competent with many existing decision tree algorithms against the standard UCI benchmarking datasets using the ANOVA test in statistics. The specific advantages of this proposed new algorithm are that it avoids the computational overhead of information gain computation for large data sets with many attributes, as well as it avoids the conversion to categorical data from huge numeric data sets which also is a time consuming task. So as a summary, huge numeric datasets can be directly submitted to this algorithm without any attribute mappings or information gain computations. It also blends the two closely related fields statistics and data mining.

**Keywords-** Statistical variance, Data Mining, Decision tree, Statistical mean, Accuracy

## 1. INTRODUCTION

Decision trees have proved to be very useful tool for the description, classification and generalization of data. Work on constructing decision trees from data exists in multiple disciplines such as statistics, pattern recognition, etc. There have been many variations for decision tree algorithms. The initial ID3 algorithm could not handle numeric data sets. Many later versions like C 4.5, C 5 etc handled numeric data sets. But not much literature is available where any alternative to information gain is given. [3] Surveys existing work on decision tree construction, attempting to identify the important issues involved and the current state of the art.

An authentic and classical literature for data mining techniques, is given in [1], where the decision tree algorithms using information gain has been thoroughly described. Article [6, 14] describes a way by which the classic decision tree algorithm C 4.5 works as well as its improvements, which is described in [7]. In [2] a way by which numeric attributes can be discretized for further data mining steps is given. Chi-square is a simple and general algorithm that uses the $\chi 2$ statistic to discretize numeric attributes repeatedly until some inconsistencies are found in the data, and achieves feature selection via discretization. [4] and [5] describes Naïve Bayes classifier as another classifier that processes numeric data sets efficiently and it assumes normal distribution for numeric attributes.

As shown in [8] there have been some efforts in the data mining community to incorporate statistical measures even before now, but it has been seen that they could not really make their place in algorithms of inducing decision trees. It has been seen that though C 4.5 has been the most popular decision tree implementation by Quinlan , a commercial version of it namely C5 has been released which includes many new features like new attribute types, pruning of misclassification costs etc. In C4.5, all errors are treated as equal, but in practical applications some classification errors are more serious than others. These concepts are described in [9] and [10]. But both C 4.5 and C 5.0 are based on information gain concepts. [11] and [12] publications by the same author which describes some practical application domains of decision trees.

This paper proposes variance as an alternative to information gain. Statistical variance is a concept which is straight forward to implement through programming than complex information gain concepts, which itself is a clear advantage as far as the processing of large data sets are considered.

RECENT SCIENCE Publishing Research

© RECENT SCIENCE PUBLICATIONS ARCHIVES | February 2013|$25.00 | 27702271 |
*This article is authorized for use only by Recent Science Journal Authors, Subscribers and Partnering Institutions*

## 2. PROBLEM STATEMENT

To design, implement and test a decision tree algorithm exclusively for numeric datasets where the various important concepts in statistics such as variance and mean can be used.

## 3. C 4.5 *STAT ALGORITHM

1. Let the set of training data be S. Put all of S in a single tree node.
2. If all instances in S are in same class, then stop
3. Split the next node by selecting an attribute A , for which there is minimum Statistical Variance
4. Put the split point as the Statistical Mean of the current subset of data.
5. Stop if either of the following conditions is met, otherwise continue with step 3:

(a) If this partition divides the data into subsets that belong to a single class and no other node needs splitting.

(b) If there are no remaining attributes on which the sample may be further divided.

In conventional decision tree algorithms like C4.5, the splitting will be done based on the maximum information gain concept. But here the statistical variance is used, which is defined as follows:

In general, the ***population variance*** of a *finite* population of size *N* is given by

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

Where μ is the population mean as given below:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Here the assumption is that, if a subset of the data is having low variance then there is a chance that they converge to a particular class in minimum number of iterations as there is minimum variation in the data for that attribute.

## 4. IMPLEMENTATION

For implementing the algorithm, the popular open source data mining package WEKA 3.4 has been used. There the program corresponding to new C4.5*stat algorithm has to be coded using JAVA, by integrating WEKA with Netbeans 7.0 IDE.

Many Weka core classes has been reused giving more accurate and systematic implementation. The advantage of using Netbeans IDE is that one can view the test results then and there itself in the IDE, just as you see the output in WEKA. All the existing decision tree algorithms have been tested with default parameter configurations in WEKA.

## 5. DETAILS OF DATASETS USED

UCI (University of California, Irvine) datasets are the standard benchmark datasets used in data mining research for testing new algorithms. Usually data mining research proceeds in different domains and researchers claim varying results across different domains. Hence in order to standardize performance evaluation, it is often required that scientists should prove the performances of the proposed algorithms over these standard datasets to see that their algorithms are having better performances other than the standard algorithms. These UCI datasets are a result of efforts from data mining researchers over many years for collecting datasets from various sources for their research purposes. Below is the description of the UCI datasets those are used for testing this algorithm. These bench marking datasets are freely downloadable from [15].

### 5.1 IRIS DATASET

The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant.

Predicted attribute: class of Iris plant.
Number of Instances: 150
Number of Attributes: 4, numeric

### 5.2 SEGMENT DATASET

The instances were drawn randomly from a database of 7 outdoor images. The images were hand segmented to create a classification for every pixel.

Predicted attribute:class of an image as one among 7 outdoor image classes. Number of Instances: 210
Number of Attributes: 19 continuous attributes

### 5.3 DIABATES DATASET

In particular, all patients here are females at least 21 years old of Pima Indian heritage.

Predicted attribute: class telling presence of diabetes
Number of Instances: 768Number of Attributes: 8 plus class to be predicted as one among positive or negative

### 5.4 LETTER Dataset

The objective is to identify each of a large number of black-and-white rectangular pixel displays as one of the 26 capital letters in the English alphabet.

Predicted attribute: Alphabet
Number of Instances: 20000
Number of Attributes: 16 pixel position details

#### 5.5 BREST CANCER Dataset

The objective is to classify patient data as malignant or not. Inputs are 9 cell related details. Predicted attribute: class telling presence of breast cancer
Number of Instances: 699
Number of Attributes: 9 cell specific details

## 5.6 Glass Dataset

The study of classification of types of glass was motivated by criminological investigation. At the scene of the crime, the glass left can be used as evidence, if it is correctly identified.

Predicted attribute: Type of glass
Number of Instances: 214
Number of Attributes: 10 glass specific details

## 5.7 Labor Dataset

It is used for the study of relationship between labor contract and the benefits offered by the employer. The class to be predicted is whether it is an acceptable contract or not. Predicted attribute: Type of labor contract
Number of Instances: 57
Number of Attributes: 16 types of labor conditions

## 6. TESTING

If one has to prove the worthiness of any new algorithm, it has to be tested against the standard UCI repositories which are the universally accepted benchmarking datasets. Cross validation using 4 folds has been used as a standard testing strategy. UCI repository has got different data sets out of which there are 7 purely numeric data sets that are supplied with Weka data mining suite are shown in the above section. They were used for testing the java implementation of the new C4.5*Stat algorithm. Many existing decision tree algorithms are used for comparison of accuracies. Also three popular algorithms for numeric data sets namely Naïve Bayes classifier, neural networks and C4.5 algorithm are also recorded in this paper along with this new algorithm, whose results are used for further theoretical analysis.

## 7. TEST RESULTS AND OBSERVATIONS

**TABLE 1: ACCURACIES OF VARIOUS DECISION TREE ALGORITHMS ON UCI DATA SETS**

| Dataset | ADTree | REPTree | RandomTree | C 4.5 * Stat |
|---|---|---|---|---|
| IRIS | NA | 96 | 94 | 95.3 |
| SEGMENT | NA | 94.81 | 89.13 | 94.07 |
| DIABATES | 73.17 | 73.56 | 68.09 | 74.47 |
| LETTER | NA | NA | 82.875 | 81.34 |
| BREAST-CANCER | 95.7 | 94.7 | 93.84 | 95.13 |
| GLASS | NA | 66.82 | 62.61 | 67.28 |
| LABOR | 82.45 | 68.42 | 85.96 | 82 |
| **MEAN** | **83.77** | **82.385** | **82.35786** | **84.23** |

Table.1 shows the accuracies of various decision tree algorithms over UCI data sets. In the table some cells are marked NA to designate that the algorithm does not support/high computing time such that accuracy values have not been used for tabulation. Finally table 2 shows the comparison of the accuracies of the 3 most important data mining classification algorithms namely C 4.5 ,neural networks and Naive Bayes algorithm for numeric data sets along with the new algorithm and that data is further used for detailed theoretical analysis. Table.2 also shows the summary of the test results. ANOVA testing is useful when the means of groups of data to be compared are two or more. A t-test can be used to compare two means.

A multiple t-test can also be used to compare more than two means. But the procedure may become more complicated as one has to first construct many pairs among many groups and then proceed with pair wise analysis. But in ANOVA, the procedure is much straight forward. In ANOVA, one starts with the assumption or the null hypothesis that the means are equal So in this work, the null hypothesis considered was that the mean of accuracies of all the algorithms namely C 4.5*stat, C 4.5, neural networks and Naive Bayes accuracies of all the algorithms namely C 4.5*stat, C 4.5, neural networks and Naive Bayes classifiers were comparable. Then after the test, the most important value to be observed is the p-value of the ANOVA test. According to ANOVA testing if this value is greater than 0.5, the null hypothesis can be accepted. But in this case it was 0.665. Hence the null hypothesis can be accepted and accuracies were concluded to be competent with each other. Neural networks showed a slightly higher accuracy, but lengthy and complex training time can be its limiting factor.

**Table 2: Accuracies of various classification algorithms on UCI Data set**

| Dataset | C4.5*stat | C 4.5 | Neural Network | Naïve Bayes |
|---|---|---|---|---|
| Iris | 95.3 | 95.3 | 95.3 | 94.66 |
| Segment | 94.07 | 96.17 | 95.18 | 77.03 |
| Diabetes | 74.47 | 73.3 | 75.78 | 75.78 |
| Letter | 81.34 | 86.59 | 82.56 | 63.99 |
| Breast-cancer | 95.13 | 95.56 | 96.28 | 95.99 |
| Glass | 67.28 | 66.82 | 67.28 | 45.32 |
| Labor | 82 | 73.68 | 89.47 | 92.98 |
| **Mean** | **84.23** | **83.92** | **85.97** | **77.96** |

## 8. ANALYSIS OF ALGORITHM

The time complexity of standard decision tree algorithm is O ($mn^2$), where m is the number of records and n is the number of attributes [13]. This is because there are total m records itself among all nodes in a particular level at a time and for computing information gain, it has to consider each of the n attributes. So in a particular level, the complexity is O (mn). In the worst case, there will be a split corresponding to each of the n attributes. So altogether it becomes like O ($mn^2$) in the worst case. But here as the numeric data are split based on statistical mean the number of levels in the worst case is $\log_2 m$. So the time complexity becomes O ($mn\log_2 m$). So as the numbers of attributes become very high, which is common in huge data sets like bioinformatics data, this algorithm will have an edge in terms of time.

## 9. CONCLUSIONS AND FUTURE SCOPE

The proposed new algorithm gives a new light to the data mining community in that, statistical measures like variance and means are good substitutes for conventional information theory based concepts. It has been also proved that this algorithm is competent as far as accuracy is concerned with respect to standard UCI data sets. Also this algorithm avoids the extra overhead of information gain computation.

Extending this work in developing fast decision tree algorithms for other data types also can be interesting. One of the future directions this work can be taken is testing with data sets of different domains other than standard UCI repository. Splitting with respect to other statistical parameters like mode, standard deviation etc can also be experimented. As it is well known among the data mining community, statistics and data mining are two closely related fields which give and take research contributions from domains of each other. Contributions of concepts like regression, Bayes classifier etc has been from statistics. In that light this new algorithm also can be considered as a contribution from statistics to the data mining community.

## REFERENCES

[1] I.J Witten, E. Frank. “Data Mining: Practical Machine Learning Tools and Techniques with Java Implementation,” Morgan Kaufmann, San Mateo, CA, 2000.

[2] Huan Liu, R. Setiono, “Chi2: feature selection and discretization of numeric attributes”, Proceedings of the seventh international conference on artificial intelligence, Herndon, USA, 1995, pp. 388-391.

[3] Sreerama K. Murthy, “Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey,” 1998, Springer Data Mining and Knowledge Discovery, 2(4), 345-389.

[4] Mark A. Hall, Geoffrey Holmes, “Benchmarking Attribute Selection Techniques for Discrete Class Data Mining”, 2003, IEEE TKDE, 15(6), 1437-1447.

[5] P. Langley, W. Iba, K. Thompson, An analysis of Bayesian classifiers," Proceedings of the Tenth National Conference on Artificial Intelligence, San Jose, CA, 1992, pp. 223-228, AAAI.

[6] J.R. Quinlan, “Induction of decision trees,” 1986, Machine Learning, 1(1), 81–106,

[7] J.R. Quinlan, “Bagging, boosting, and C4.5,” In Proceedings of 13th National Conference on Artificial Intelligence, pp. 725–730, 1996.

[8] Leo Breiman, “Bias, variance and arcing classifiers, Technical report 460, Statistics department, University of California at Berkley, 1996.

[9] Jeffrey P. Bradford, Clayton Kunz, Ron Kohavi, Cliff Brunk and Carla E. Brodley, “Pruning decision trees with misclassification costs”, 1998, Springer LNCS, 1398(1998), 131-136.

[10] R.E. Schapire, Yoav Freund, “A short introduction to boosting”, 1999, *Journal of Japanese Society for Artificial Intelligence*, 14(5), 771-780.

[11] Sudheep Elayidom.M, Sumam Mary idicula, Joseph Alexander, “Applying Data mining techniques for placement chance prediction”, Proceedings of international conference on advances in computing, control and telecommunication technologies at Trivandrum, India, 2009, pno:669-671.

[12] Sudheep Elayidom.M, Sumam Mary idicula, Joseph Alexander, “Analysis and implementation Of Data mining techniques using Naive-Bayes Classifier and Neural Networks", 2010, Global Journal of Computer Science and Technology, 10(10), 20-25.

[13] Zhang Harry, Su Jiang, “ A fast decision tree learning algorithm”, In Proceedings of the 21st national conference on Artificial intelligence, pp. 500-505, Boston, USA, 2006.

[14] U. Fayyad, R Uthurusamy, “From Data Mining to Knowledge Discovery in Databases,” AI Magazine, 17(3), pp. 37-54, 1996.

[15] http://archive.ics.uci.edu/ml/datasets.html

© RECENT SCIENCE PUBLICATIONS ARCHIVES | February 2013|$25.00 | 27702271 |
*This article is authorized for use only by Recent Science Journal Authors, Subscribers and Partnering Institutions*