ISSN No. 0976-5697

Volume 5, No. 7, September-October 2014

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Performance Analysis On Decision Tree Algorithms

Miss Nalanda B Dudde
Student ME CSE,
Walchand Institute of Technology,
Solapur, India.

Dr Mrs S.S. Apte
HOD CSE Department,
Walchand Institute of Technology,
Solapur, India.

Abstract: In data mining, classification is one of the most widely use technique for medical diagnosis, fraud detection, artificial intelligence, sales and production, plant classification etc. Various algorithms have been implemented till now to build a decision tree which is a classification technique. Decision tree algorithms like ID3, C4.5, VFDT, RDT used widely because of its easy implementation and analysis is much simple. In this paper we are highlighting on performance evolution of these algorithms i.e. accuracy of each algorithm with the respective dataset used. And presenting the best one implemented so far.

Keywords: Classification, Decision trees, ID3, C4.5.

I. INTRODUCTION

An important technique in **machine learning is** decision trees, which are used extensively in **data mining**. It is used to turn out human-readable descriptions of trends within the underlying relationships of a dataset and might be used for classification and prediction tasks. A decision tree is predictive modeling technique used in classification, clustering and prediction tasks. It uses divide and conquer technique to split the problem search space into subsets [4]. It can be used to explain why a question is being asked. The decision tree assumes that questions are answered with a certain yes or no. In globe issues, the intuition of a personality's knowledgeable, or knowledgeable system package, is critical to see the possible finish node. Each end node represents a situation with known effective and efficient leadership styles. The technique has been used with success in many various areas, like diagnosis, plant classification, weather prediction, client selling methods etc. A decision tree is a representation of a decision procedure for determining the class of a given instance. Each node of the tree specifies either a class name or a specific test that partitions the space of instances at the node according to the possible outcomes of the test. Each subset of the partition corresponds to a classification sub problem for that subspace of the instances, which is solved by a sub tree [5][3]. Formally, one can define a decision tree to be either.

A leaf node (or answer node) represent successful/unsuccessful answer, and a non-leaf node (or decision node) that contains an attribute test with a branch to another decision tree for each possible value of the attribute [2]. An attribute check, with a branch to a different call tree for every doable price of the attribute, the positive and negative counts for every doable price of the attribute, and therefore the set of non-test attributes at the node, every with positive and negative counts for every doable price of the attribute.

Weather prediction is one of the most effective environmental constraints in our routine. We adjust ourselves with respect to weather condition from our dressing to strategic organizational planning activities.

There are 2 methods used for weather prediction one is empirical approach and another is dynamic approach [7].

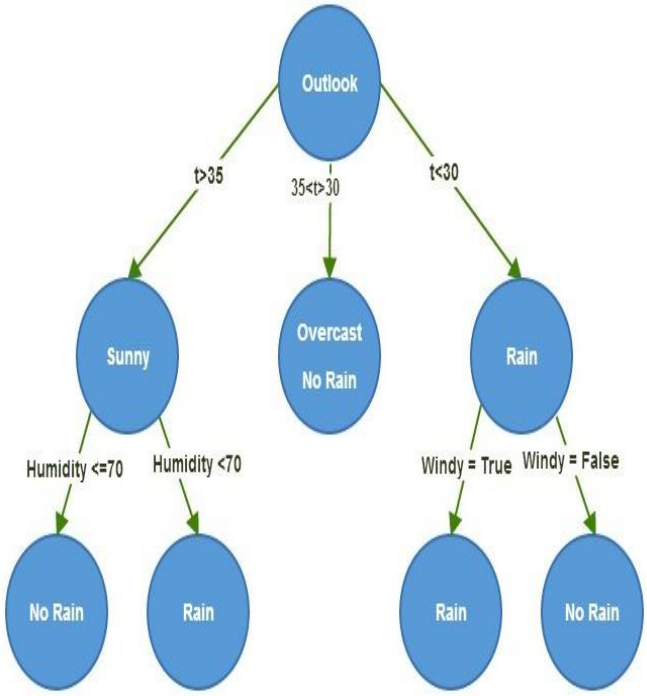

Figure1.1: Graphical Decision tree

The empirical approach is based upon the comparable cases (i.e., similar weather condition). The dynamical approach is based upon equations of the atmosphere and is commonly referred to as computer modeling.

II. LITERATURE REVIEW

The decision tree for numerical data [1] is proposed by Nandagaonkar S ,Attar V.Z , Sinha P.K. in paper Efficient Decision Tree Construction for Classifying Numerical Data which uses random strategy for building a decision tree. It

also uses heuristic method to compute the information gain for obtaining the split threshold of numerical attributes.

ID3 is a simple decision tree learning algorithm developed by Ross Quinlan (1983)[2]. The basic plan of ID3 formula is to construct the choice tree by using a top-down, greedy search through the given sets to check every attribute at each tree node.

C4.5 algorithm is an enhancement in the ID3 algorithm, it replaces the internal node with the leaf node hence, less error rate. It accepts continuous and categorical data. It has an enhanced method of tree pruning that reduces misclassification errors due to noise and too many details in the training data set [10]

An expert system for weather prediction based on animal behavior by Najat O. Alsaiari[6] it is an expert system for predicting short term weather based on the behavior of both birds and animals like sea- crab, which have been observed to have certain reactions before weather changes.

Weather prediction expert system approaches[7](Ceng-568 Survey) by Bulent Kiskac and Harun Yardimci it's a survey paper where Hybrid systems are promising for integration of current expert systems on large scale also many web weather report and forecast service systems enable implementations to fetch some weather information, hybrid system can reduce intensive forecast computations. Instead of computing some detailed information, here we can easily pick up knowledge about a satellite post-processing reports and comments

III. PERFORMANCE ANLYSIS

The table 4.1 shows the detailed analysis of prediction techniques which were adopted decision tree framework.

Table 1 Detailed analysis of various approaches

Name of Approach	Input Dataset	Input Dataset Type
ID3[8]	Diabetes	Static i.e. Non Continuous
C4.5[8]	Diabetes	Continuous and Discrete
VFDTc[9]	Weather	Only Streaming data

Though the ID3 algorithm worked well with the non-contiguous data and has the advantage that it generates a smaller depth decision tree, we would want to evaluate this algorithm with a larger and more complicated data set. Also, we might want to consider evaluating different types of decision trees along with clustering algorithms to determine if there is a better approach for the medical industry specifically for determination of the risk of heart disease. Lastly, there are vast amounts of data available and there are many ways it can be manipulated. So, using ID3 algorithms is an iterative process where processes are always being improved (such as when new attributes are added for considerations – for example, there may other factors for risk of heart disease such as weight or family history). For the medical industry, these decisions can determine if a patient is a high risk for heart disease along with making conclusions as to what insurance coverage a company should give a person based on this risk. ID3 algorithm depends entirely on the accuracy of the training data set for building its decision trees. The ID3 algorithm learns by supervision. It has to be shown what instances have what results. Due to this ID3 algorithm, cannot be successfully classify documents in the web. The data in the web is unpredictable, volatile and most of it lacks Meta data

In fact, since ID3 was first developed, there is now an improved version called C4.5.It has overcome all the drawback of ID3 and it handles missing attributes and continuous attributes i.e. it supports continuous data that means it does not need to entirely depend on accuracy of training data. But the C4.5 needs entire data to fit in memory. C4.5 is used in classification problems and it is the most used algorithm for building DT. It is suitable for real world problems as it deals with numeric attributes and missing values. The algorithm can be used for building smaller or larger, more accurate decision trees and the algorithm is quite time efficient. Compared to ID3, C4.5 performs by default a tree pruning process, which leads to smaller trees, more simple rules and more intuitive interpretations.

The VFDTc is the successful algorithm in classification algorithms for mining data streams and manage thousand of examples with similar performance to batch decision tree with enough examples. The VFDTc can work on numerical as well as discrete data and applies navie bayes classifier on tree leaves.The VFDTc is weak in antinoise capability and have the high memory cost, it is not suitable for data stream with high dimensional data and noisy data.

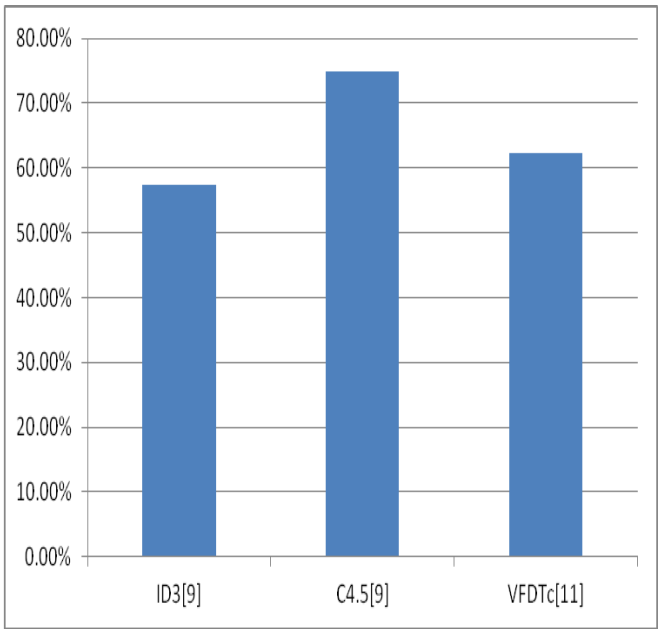

Figure 1: Accuracy Chart of various methodologies

IV. CONCLUSION AND FUTURE WORK

From the above analysis we can say that, the C4.5 and VFDTc are until now the best algorithms implemented using Decision Tree because they have come up with new behavioral model which is far better that ID3. VFDTc mainly used in fraud detection techniques and online transaction system.

Till now the prediction and forecasting of weather is done using weather maps, behavior of animals, satellite images, clouds positioning etc. We can propose a new model which can run on historical weather data that helps in classification and prediction. In the new model we can implement following things so, it would leads to better

accuracy in prediction as well as smaller trees for predicting the weather.

a. The input data may reside on web or it's a streaming data so our proposed approach takes less space.
b. It will not support windowing approach and backtracking so it can run in less amount of time as compare to ID3, C4.5 and VFDTc.
c. The weather data is a continuous data which changes timely, so the dataset is very huge for training and testing the tree. The maximum amount of data is directly proportional to the accuracy.
d. It can prune decision trees to stop them over fitting the training data.

V. REFERENCES

[1]. Nandagaonkar S., Attar V.Z. ; Sinha P.K. ," Efficient Decision Tree Construction for Classifying Numerical Data", Advances in Recent Technologies in Communication and Computing, 2009. , Page(s): 761 – 765

[2]. Quinlan J.R., "INDRODUCTION OF DECISION TREES" Machine learning. VOL 1986, 81-106

[3]. J. E. Gehrke, R. Ramakrishnan, and V. Ganti, "Rain-Forest - A framework for fast decision tree construction of large datasets," Data Mining and Knowledge Discovery, Vol. 4, No. 2/3, Jul. 2000, pp. 127-162.

[4]. J. Han and M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann Publish, 2001

[5]. Kalles D, Morris T. "Efficient incremental induction of decision tress", machine learning,1996,24(3); 231~242

[6]. Najat O. Alsaiari. "An expert system for weather prediction based on animal behavior".

[7]. Bulent Kiskac and Harun Yardimci. "Weather prediction expert system approaches"(Ceng-568 Survey)

[8]. D. Lavanya & Dr. K. Usha Rani, Sri padmavathi mahila visvavidyalayam, Tirupati , AP. "Performance Evaluation of decision tree classifiers on medical datasets"

[9]. Ricardo Rocha Projecto Mathematica Ensino Departamento de Mathematica 3810 Aveiro, Portugal. "Accurate Decision tree for mining high speed data streams"

[10]. Anuja Priyama, Abhijeeta, Rahul Guptaa, Anju Ratheeb and Saurabh Srivastava "Comparative Analysis of Decision Tree Classification Algorithms" International Journal of Current Engineering and Technology, Vol.3, No.2 (June 2013)

