*International Journal of Engineering & Technology*, 7 (2.24) (2018) 235-238

International Journal of Engineering & Technology

*Website: www.sciencepubco.com/index.php/IJET*

***Research paper***

# Privacy Protection and Perfect Classification Nature of C4.5 Algorithm

**K. Chokkanathan*, S. Koteeswaran**

[1]*Research Scholar,* [2]*Associate Prof, Department of Computer Science and Engineering, Vel Tech Dr.RR & Dr.SR Technical University*

## Abstract

C4.5 algorithm is developed by Ross Quinlan which is the extension of ID3 algorithm used for generating a decision trees.Since the tree generated by C4.5 can be used for classification, so it's also referred to as statistical classifier.Even though the Random Decision Tree is used to avoid the information leakage there are some problems and issues related to privacy maintenance.When we try to instantiate more instances for one class it leads to ambiguity at the same time creating new classes more and more will increase the complexity in RDT. These problems can be resolved by using our C4.5 algorithm.We can have any number of nodes in a network, each node can create its own tree or class and each class can initiate many number of instances for a disseminated classification consuming secure amount or threshold homomorphic encryption. The main objective of this paper is to discuss the ideal nature of the C4.5 algorithm and how they support this algorithm to be utilized in various datamining process.

***Keywords:*** *Decision Tree, C4.5, ID3, Random Decision Tree, classification.*

## 1. Introduction

When data is moving in a public network environment, there are possibilities for getting modified by the intruders and personal information may be misused in a variety of means. In networking applications it is easy to distribute data across multi parties and gathering on a large volume for distributing and sharing information. The most widely used algorithm is Decision tree algorithm for distributing and classifying the decision making data packets in the network environment[1].One of the datamining tasks is Classification, which is can be done effectively by C4.5 in a distributed environment. This will be useful for producing classifier that have good prediction accuracy without compromising privacy even for small data sets.

### Algorithms for Classification

Most of the research works are carried out by the concept of classification from datamining. So many techniques are involved in datamining, among them classification is approach is implemented by many researchers for their research work. Classification is the most appropriate technique for networking field, where this technique is used for classifying real time and non-real time packet during the packet transformation from source to destination. When the data from various applications are handled by this technique, from huge volume of data set it will identify and classify valid, invalid, real-time, non-real time packets accordingly. Among other techniques classification is an effective method to classify the network based packets in an efficient manner. There are various approaches under classification such as Decision Tree Induction, Neural Networks, and Classification by Lazy Learners, K-Nearest Neighbor Classifiers, Rule-Based Classification and Bayesian Classification.Under decision tree induction, C4.5 is playing vital role in network packet traffic classification.

### Decision Tree algorithm (C4.5)

A Predominant approach for the classification problems is Decision Tree approach. A model will be constructed to demonstrate the classification process. In general a decision tree consists of nodes, branches and leaves to represent variables, conditions, and outcomes respectively. Basically it's a recursive process for constructing a tree structure for a particular scenario. Decisions are made for various stages.
First choose a unique attribute and fix it as root and make possible branches from the root node, this has to be repeated up to visiting all the attributes and derive as much as possible branches. Tree structure will be completed once after finding all the training data at the end node after successful classification. The following figure-1 shows the model diagram for the Decision tree structure.
Different types are derived from decision tree such as ID3, C4.5 and Random Forest.ID3 is used to generate the decision tree for different types of datasets; it builds the tree from the top down, with no backtracking. C4.5 is a widely-used classification method in different fields that is descended from an earlier method ID3.Among decision tree algorithms C4.5 is most widely used and accepts attributes in both formats like category as well as continuous for constructing the decision tree. Using Gain or Gain-Ratio approach it will produce optimally splitted attributes. The main advantage of decision tree approach is knowledge or data can be extracted and represented in the form of classification rules.For each and every rule a unique path will be identified from root to each leaf node.

**Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.**

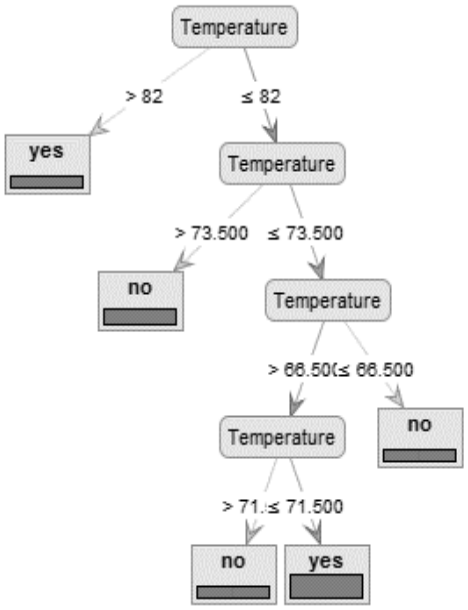


**Fig.1:** Decision Tree example for Temperature measurement

## Classification and Prediction

In datamining concept Decision tree is an efficient and effective algorithm for classification and prediction. Initially it compares the attributes of objects and find the optimal values and fix them as root node, later based on the selected root nodes or attributes it determines the leaf node of the tree branch. One major advantage is it does not need more detailed background about the attributes during the learning process, it will find them from the training examples and their expressions which will become the conclusion of the model. Important characteristics of Decision tree are, 1. It has very simple structure and easy to understand. 2. When we have large amount data then it's the best option for selecting the Decision tree. 3. Always produces high accuracy. In general C4.5 is the improved version of ID3 and has some added advantages compare to ID3, such as, 1. It increases the processing speed while increasing the value of attributes. 2. It uses information gain ratio to select the attributes. 3. It can easily handle the missing values during training process. By using trimming techniques it uses to avoid unevenness of the tree structure. And it has the ability of cross validation of the K iterations. These are some of the synaptic advantages of C4.5 for choosing using it in datamining fields.

## Capabilities of C4.5

In C4.5 a training dataset which consists of several general attributes and class attributes can be used to for classification to predict the class attributes of a new transaction. Constructing a decision tree is playing vital role in decision making tree .In a decision tree each leaf node is used to represent a classification result and each non-leaf nodes are used to represent the testing attributes also called general attributes [2].

When packets are transmitted in the network they are accumulated continuously, the problem is whether they are analyzed effectively in a proper manner or not. So we want to analyze those using data mining technology. Most of the cases decision tree based C4.5 algorithm is used for packet analysis to help the classifier which will gain insight knowledge about transmitted packets and strategies to obtain the maximum benefit [3].Even though the C4.5 algorithm is successor of ID3, it has more advantages over ID3.When number of nodes getting increase obviously we have to divide into sub classes, but when we split in ID3, it causes high error classification and will take more time for processing the subsets.C4.5 is recommended to overcome these concerns using gain ratio to select the records associative only to the attributes. Even though the C5.0 is descendant of C4.5, we cannot use it for two reasons first, data set has limitations, a data set used by a researchers may not be applied by other researchers, second; datamining tool Weka simulation tool will support C5.0.Because of these reasons it's advisable to use C4.5 in data classification or network traffic classification [6].

## Preserving Nature of C4.5

When there is a huge volume of data sets, it's very difficult to search the similar datasets associated to the particular environment. Classification is the process of generalizing the data packets from the known and huge volume of training data structure. Particularly when the data or information need to be shared or exchanged among group of nodes is a point to get loophole for data security [4].Without the permission of owner the data should not be used for secondary purpose. So the principle priority and security must be provided to the data during the transmission.C4.5 is an extension of ID3 which will provide the optimum solution and uses the information gain rate to find the properties of huge volume of information in a tree structure [5].In a knowledge discovery environment it's more important to preserve the privacy of data. It's one of the most popular methods from supervised machine learning family used for network traffic classification. If follow divide and conquer approach to solve the problems by learning from given data set of independent instances.

## Network traffic classification

When there are so many applications using different types of networks, the emerging concept called network traffic classification is playing vital role with different techniques. One of the supervised machine learning approaches called Decision Tree, which is widely used in network traffic classification. There are many techniques available in Decision Tree approach, among them C4.5 will provide the highest classification accuracy than others [6].

Network traffic identification and classification are playing vital role in network management and security areas where different applications using different services and consuming more network resources. This classification can be used for detection of intruders, to detect the malicious applications which can cause damages to our data or information and to recognize the DoS attack. The traditional approaches such as port-based and payload based identification and classification are becoming more difficult with new coming applications, because they use dynamic port numbers, masquerading techniques and encryption to avoid detection. So the concept of Machine learning approach has been introduced to overcome the issues in traditional approaches by considering distinctive characteristics of flow statistics. In general the classification will be done as given in the following flow diagram-figure-2.

For network traffic classification various parameters are used such as port number, arrival time, type of protocol, packet length and etc. to categorize into number of classes. Because of increasing applications and huge of size of data traffic several challenges are faced by the network engineers [7].Some of the traffic classification techniques based on this ML are impacting the dataset size and feature selection by experiments. These machine learning algorithms are providing increased classification accuracy with less information. For example C4.5, NB-Tree, and Random Forest are producing high classification accuracy [8].Even though they have more computational complexity they produce more precise classification results.

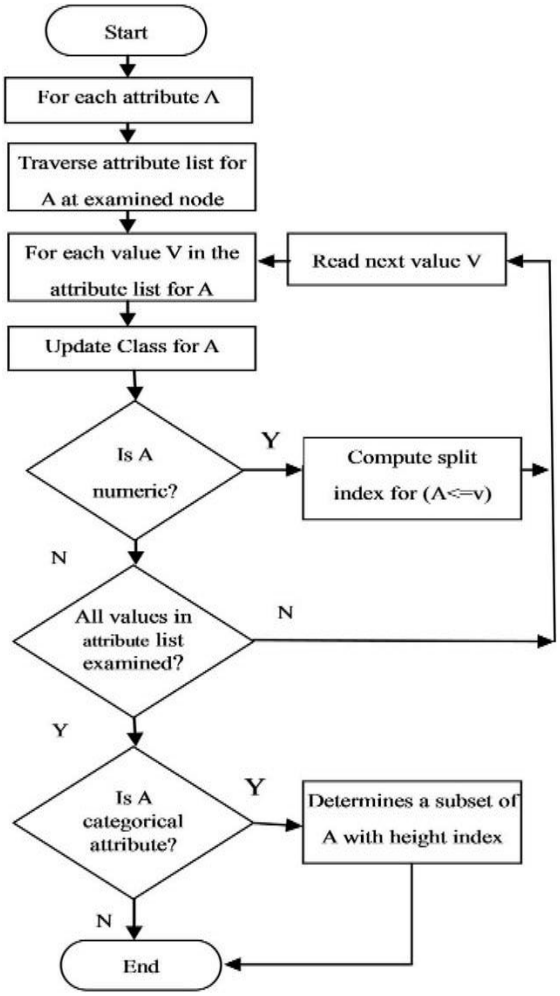


**Fig.1:** Classification flow chart in C4.5

## Pruning in C4.5

Whenever we generate a decision tree for training data set it will over-fit for sample data in most of the scenario. Those trees will work for the trained data nut the performance may not be opt for the unseen sample data sets. So to avoid the prediction error and reduce error rate we need to implement the concept of pruning. Wherever we have less possibility of classifying the instances there we need a techniques which will provide the facility to reduce the size of decision tree. Such a powerful and efficient technique is known as Pruning. So we can reduce the noisy or erroneous data by reducing and removing the necessary part from the decision tree and we can reduce the complexity of the final classification process also. The pruning algorithm is based on a pessimistic estimate of the error rate associated with a set of N cases, E of which do not belong to the most frequent class. Instead of E/N,C4.5 determines the upper limit of the binomial probability when E events have been observed in N trials, using a user-specified confidence whose default value is 0.25 [9].In general the pruning process starts at leaves and end at root..Forasub-tree,C4.5adds the estimated errors of the branches and compares this to the estimated error if the sub-tree is replaced by a leaf.

## Discretization

In datamining analysis process the most significant algorithm is Decision tree algorithm which can be used for classification, description and generalization of data. There are so many disciplines such as signal processing, pattern recognition, decision theory, statistics, machine learning and artificial neural networks are working based on decision tree algorithms for making best decisions. It's acting like an unsupervised filter in constructing accurate trees.C4.5 algorithm is used for constructing decision tree in discretization. It is expressed in two different phases: in the first phase, all the attributes are discretized instead of dealing with numerical values. The second phase is about evaluating the performance by constructing the decision tree [10].Some of the clustering and classification algorithms will not be comfortable in handling numeric scales, they will deal with nominal attributes only. So they must be discredited into a small number of distinct range of values.

Although most decision tree and decision rule learners can handle numeric attributes, some implementations work much more slowly when numeric attributes are present because they repeatedly sort the attribute values [11].

**Table 1:** Sample Discretization Process Chart for Bank data

| Attributes | Range | Code |
|---|---|---|
| AGE | 23 – 30.6 | A |
| | 30.6– 38.5 | B |
| | 38.5– 45.6 | C |
| | 45.6– 53.4 | C |
| | 53.4– 61.6 | E |
| BALANCE | 2000-5000 | A |
| | 5000-10000 | B |
| | 10000-50000 | C |
| | 50000-100000 | D |
| | 100000-200000 | E |
| DAY | 4-10 | A |
| | 10-20 | B |
| | 20-40 | C |
| | 40-50 | D |
| | 50-60 | E |

This discretization is very simple and easy to implement.It requires only the number of intervals and the number of points to be included in any given interval [12].The above table-1 shows the sample values for the Discretization process of bank data.The main objective of this discretization is number of values must be reduced and group them into a specific number of intervals.

## Tree Generation

From the labeled trained class tuples algorithm will learn the decision tree. Basically the decision tree has flow-chart like a tree structure. Top most node is called root. Each and every internal non-leaf nodes are called test condition on attributes. Each leaf node is the outcome of the condition on attributes.Suppose a customer wants to buy a computer then starting from quotation collection and up to purchase of computer we can generate as a tree structure. Overall tree will give guidance that where he purchased, what configuration and how much he paid etc. Sometimes the tree generation will be done with only binary tree, where each internal node will have exactly two branch nodes Whereas others can generate non-binary trees. Greedy or non-backtracking techniques is used in C4.5, which is following the top-down approach. This approach starts with a training set of tuples and their associated class labels. Training data set is continuously splitted into smaller subsets for building the tree. The following steps are used for constructing a decision tree [13].
The steps are,

1. Select the data set input
2. Choose the classifier
3. Calculate entropy, gain, gain ratio of attributes
4. Processing the data set
5. Algorithm will generate Tree structure

## C4.5 for Distributed Environment

Protecting or masking sensitive information is the extension of datamining concept to protect the privacy of data and provide the high level security in the public environment. This concern about how to provide security for the personal data where there is a possibilities of misusage for various purpose. In order to

overcome this problem various techniques have been introduced in datamining domain to preserve and protect the data. Nowadays the advanced database technologies are facilitating the multi parties to collect and share / distribute the huge volume of data across the networks. Most of the business decisions are taken accurately with the help of the underlying techniques such as association rule mining and decision tree learning etc. which are part of distributed datamining concept. Even though there are so many business organizations are ready to collaborate with one another due to some legal constraints and competitions they were unable to disclose their private information to other party during the data mining process.

However the growing technologies are gaining knowledge from the huge volume of data, there is a possibility of violation of individual privacy data. So the privacy preserving data mining comes to an action and addressing the privacy issues and becoming a challenging research topic in datamining domain. Datamining tasks such as association rule, classification, clustering are the basic for preserving privacy of data. Data will be controlled from central part and subset will be distributed horizontally to all parties involving in the transaction. Each and every subset contains all the attributes and every involving party has the same number of transaction records.

Another issues is to construct decision tree in a distributed environment. First as an experimental step we have to generate the decision tree for a small dataset and check for privacy and accuracy. In C4.5 there is a provision of getting more accuracy in a distributed environment because of its improved time complexity and support of multiparty environment [14]. The advantage is for every party only one tree will be constructed and time will be reduced. During the construction of tree it applies pruning technique to improve the accuracy of output.

## 2. Conclusion

The main objective of this paper is to highlight the idle nature of the C4.5 algorithm and how it is supporting and protecting the privacy of data in a distributed network environment. It's one of the most widely used classical approach in classification algorithms. It provides the best performance in constructing decision tree for various issues and mining rules from the data sets. When more and more emerging applications are introduced with similar properties, the overall accuracy will get more impact. In addition to the better ML approach we need to find the best approach to handle the sample data set from various applications and integration of method to achieve the more accurate traffic classification. That accuracy can be provided by C4.5 algorithm.With less execution time C4.5 will be able to provide the maximum accuracy in implementation.It's one of the best algorithms for mining the data set. This will improve the performance in terms of time saving and increased efficiency too. Not only improving the growing speed of the tree but also better information of rules can be generated.

## References


[1] Privacy Preserving Distributed Classification Using C4.5 Decision Tree

[2] Optimization of C4.5 Decision Tree Algorithm for Data Mining Application

[3] Application of Decision Tree Based on C4.5 in Analysis of Coal Logistics Customer

[4] Quinlan, J.R., "C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, 1993.

[5] Agrawal R, Srikant R,"Privacy Preserving Data Mining", Association for Computing Machinery Special Interest Group on Management of Data (ACM SIGMOD) Conference, 2000.

[6] Alhamza Munther et.al "Network Traffic Classification - A Comparative Study of TwoCommon Decision Tree Methods: C4.5 and Random Forest" 2nd International Conference on Electronic Design (ICED), August 19-21, 2014, Penang, Malaysia.

[7] J. Park, H.-R. Tyan, and C.-C. Kuo, "Internet traffic classification for scalable qos provision," in Multimedia and Expo, 2006 IEEE International Conference on, 2006, pp. 1221-1224.

[8] Li Jun et.al "Internet Traffic Classification Using Machine Learning" CHINACOM '07. Second International Conference on 22-24 Aug. 2007, Page(s)-239 – 243.

[9] Arbres de décision, Ingénierie des connaissances (Master 2 ISC).

[10] Ihsan A. Kareem , Mehdi G. Duaimi "Improved Accuracy for Decision Tree Algorithm Based on Unsupervised Discretization", International Journal of Computer Science and Mobile Computing, Vol.3 Issue.6, June- 2014, pg. 176-183

[11] IanH. Witten, Eibe Frank, & Mark A. Hall, "Data Mining Practical Machine Learning Tools and Techniques", Third Edition, Morgan Kaufmann, 2011.

[12] KrzysztofJ.Cios,WitoldPedrycz,RomanW.Swiniarski,&LukaszA .Kurgan,"DataMining AKnowledge Discovery Approach", Springer Science Business Media, LLC, 2007.

[13] Gaurav L. Agrawal, Prof. Hitesh Gupta, "Optimization of C4.5 Decision Tree Algorithm for Data Mining Application" Volume 3, Issue 3, March 2013.

[14] S.MerlinJ.Jesu, Vedha Nayhi, "Privacy Preserving Distributed Classification Using C4.5 Decision Tree", International Journal of Innovative Research in Computer andCommunication Engineering, Vol.3, Special Issue 3, April-2015.

[15] S.V. Manikanthan , T. Padmapriya "An enhanced distributed evolved node-b architecture in 5G tele-communications network" International Journal of Engineering & Technology (UAE), Vol 7 Issues No (2.8) (2018) 248-254.March2018

[16] S.V.Manikanthan and T.Padmapriya "Recent Trends In M2m Communications In 4g Networks And Evolution Towards 5g", International Journal of Pure and Applied Mathematics, ISSN NO:1314-3395, Vol-115, Issue -8, Sep 2017.

[17] S.V. Manikanthan, T. Padmapriya, Relay Based Architecture For Energy Perceptive For Mobile Adhoc Networks, Advances and Applications in Mathematical Sciences, Volume 17, Issue 1, November 2017, Pages 165-179