Comparative Analysis of ID3 and C4.5 Decision Making Algorithms in B2B Marketing

C . Phani Ramesh[1], R . Swathi[2], G .V. Ramesh Babu[3], Prof . M . Padmavathamma[4]
1,3,4- Department of Computer Science, SVU College of CMCS, S.V.University, Tirupati.
2-Research Scholar, Department of CSE, S.V.U. College of Engineering, SV University.

***Abstract*— In order for B2B companies to grow internationally and at the same time not to risk a vast amount of time and money to set up sales offices, the role of Distributors are very interesting. But, for the last few years by examining the effectiveness of distributors in B2B marketing, it is no doubt that there are many mismatches between manufacturers and distributors. In order to avoid the complications from the Distributor side, the manufacturer plays a crucial role to select the efficient distributor among the available distributors. In this paper we extend our work with a comparative analysis of two algorithms ID3 and C4.5 including various parameters to make write choice to classify process patterns.**

Keywords: B2B Ecommerce, Decision Trees, ID3, C4.5.

I. INTRODUCTION

In recent years, the power of the Information Technology has drastically popularized the notion of Electronic Commerce, offers many benefits for businesses as well as regular daily life and made communication, collaboration, traveling extremely easy. Customers can access the digital business environments with a click of a mouse and participate in on-line business transactions more easily than by the use of traditional methods. This evolution is a prime phenomenon in the modern business world.

In B2B Electronic Commerce, businesses focus on selling to other businesses directly or through an intermediary. Several transactions worth huge amounts are carried out between companies through e-commerce channels, dealing in all kind of products and services and covers a broad spectrum of applications that enable businesses to form electronic relationships with their distributors, re-sellers, suppliers, customers, and other business partners. In addition, B2B e-commerce is believed to be by far the largest and successful form of e-commerce in terms of turnover and transactions made as it accounted for over 90% of all e-commerce transactions made in the last decade [1].

B2B e-Commerce, as one of the major business models brought about by the Internet technologies, has made a significant contribution to the e-marketers, and larger organizations are taking advantage of the vast array of suppliers/buyers via the B2B e-Marketplace.

An increasing number of Asian countries have adopted Internet economy that has bolstered their presence in the electronic marketplace. The region has witnessed increasing presence in the B2B marketplace due to the penetration and development of the Internet technology. Despite much interest from academics and business publications in western countries, a sharper focus on the B2B marketplace in Asia is timely and warranted for several reasons.

Here the primary focus on B2B research, the Internet is clearly constantly evolving as firms grow accustomed to doing business electronically. The Internet environment provides a more effective means by which businesses can be transacted electronically between trading partners, known as Business-To-Business Electronic Marketplace (B2B e-Marketplace).

The concept of B2B e-Marketplace provides a new dimension in facilitating marketers to work more effectively, particularly, when making critical marketing decisions. Benefiting from the Internet technologies, B2B e-Marketplace can serve both sellers and buyers as a new marketing channel to conduct or execute marketing functions such as sales and distribution.

Today more than ever before competitive advantage and profitability for manufacturers and distributors stem from the efficient management of their working relationships, sharing of information, transferring knowledge/expertise, and tapping into each other's capabilities Manufacturers and distributors want to deliver the high quality and right products to the customers' locations in the right quantities at the right time[2].

In the manufacturing world, there's a constant undercurrent of dissatisfaction with distributors. In the distribution world, there's a relentless litany of complaints about the manufacturers they represent.

In order to avoid the misunderstandings and also to overcome the heavy competition between the distributors, the Manufacturer requires appropriate decision, which will play a crucial role to select the efficient distributor among the available distributors. In this paper, we extend our work with a comparative analysis of two algorithms ID3 and C4.5 including various parameters to make write choice to classify process patterns.

II. RELATED WORK

Classification is one of the most frequently studied problems by Data Mining and machine learning (ML) researchers. It consists of predicting the value of a (categorical) attribute (the class) based on the values of other attributes (the predicting attributes). Since then, lots of efficient classification algorithms are being developed, some of them are, Statistical Classification, Decision Tree, Rule Induction, Fuzzy rule induction, neural networks [3].

Statistical classification is a procedure in which individual items are placed into groups based on the quantitative information of characteristics inherent in the items (referred to as variables, characters, etc.) and based on a training set of previously labeled items [4]. Some examples of statistical algorithms are linear discriminate analysis [5], least mean square quadratic [6], kernel and k nearest neighbors.

A decision tree is a set of conditions organized in a hierarchical structure [7]. It is a predictive model in which an instance is classified by following the path of satisfied conditions from the root of the tree until reaching a leaf, which will correspond to a class label. A decision tree can easily be converted to a set of classification rules.

Some of the most well-known decision tree algorithms are ID3 and C4.5. The ID3 algorithm (Quinlan86) is a decision tree building algorithm which determines the classification of objects by testing the values of their properties [8]. C4.5 is an improvement of ID3, making it able to handle real-valued attributes (ID3 uses categorical attributes) and missing attributes.

III. DECISION TREE

The decision tree is a structure that includes root node, branch and leaf node. Each internal node denotes a test on attribute, each branch denotes the outcome of test and each leaf node holds the class label. The topmost node in the tree is the root node. In a decision tree, each internal node splits the instance space into two or more sub-spaces according to a certain discrete function of the input attributes values [9]. In the simplest and most frequent case, each test considers a single attribute, such that the instance space is partitioned according to the attribute's value. In the case of numeric attributes, the condition refers to a range.

In a decision tree, each internal node splits the instance space into two or more sub-spaces according to a certain discrete function of the input attributes values. In the simplest and most frequent case, each test considers a single attribute, such that the instance space is partitioned according to the attribute's value. In the case of numeric attributes, the condition refers to a range.

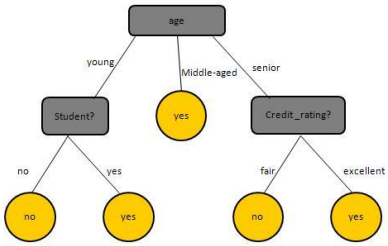


Fig:Decision tree representing computer buying.

A. ID3 Algorithm

ID3 is a simple decision tree learning algorithm developed by Ross Quinlan (1983) [10]. The basic idea of ID3 algorithm is to construct the decision tree by employing a top-down, greedy search through the given sets to test each attribute at every tree node. In order to select the attribute that is most useful for classifying a given sets, we introduce a metric---information gain.

The main ideas behind the ID3 algorithm are:

- Each non-leaf node of a decision tree corresponds to an input attribute, and each arc to a possible value of that attribute. A leaf node corresponds to the expected value of the output attribute when the input attributes are described by the path from the root node to that leaf node.
- In a "good" decision tree, each non-leaf node should correspond to the input attribute which is the *most informative* about the output attribute amongst all the input attributes not yet considered in the path from the root node to that node. This is because we would like to predict the output attribute using the smallest possible number of questions on average.
- *Entropy* is used to determine how informative a particular input attribute is about the output attribute for a subset of the training data. Entropy is a measure of uncertainty in communication systems introduced by Shannon (1948). It is fundamental in modern information theory.

B. C4.5 Algorithm

Several enhancements to the basic decision tree (ID3) algorithm have been proposed. C4.5,[11] a successor algorithm to ID3, proposes mechanism for 3 types of attribute test:

- The "standard" test on a discrete attribute, with one outcome and branch for each possible value of that attribute.

- A more complex test, based on a discrete attribute, in which the possible values are allocated to a variable number of groups with one outcome for each group rather than each value.

- If attribute A has continuous numeric values, a binary test with outcomes A_ Z and A>Z , based on comparing the value of A against a threshold value Z. Given *v* values of A, then *v*-1 possible splits are considered in determining Z, which are the midpoints between each pair of adjacent values.

The information gain measure is biased in that it tends to prefer attributes with many values. C4.5 proposes gain ratio, which considers the probability of each attribute value.

C4.5 is implemented recursively with this following sequence

1. Check if algorithm satisfies termination criteria
2. Computer information-theoretic criteria for all attributes
3. Choose best attribute according to the information-theoretic criteria
4. Create a decision node based on the best attribute in step 3
5. Induce (i.e. split) the dataset based on newly created decision node in step 4.
6. For all sub-dataset in step 5, call C4.5 algorithm to get a sub-tree (recursive call).
7. Attach the tree obtained in step 6 to the decision node in step 4
8. Return tree.

C. Algorithm C4.5

Input: an attribute-valued dataset D

1. Tree = { }
2. **if** D is "pure" OR other stopping criteria met **then**
3. terminate
4. **end if**
5. **for all** attributes a ∈ D **do**
6. computer information-theoretic criteria if we split on a
7. **end for**
8. a_{best} = Best attribute according to above computer criteria
9. Tree=Create a decision node that tests a best in the root
10. Dv=Induced sub-datasets from D based on a best
11. **for all** Dv **do**
12. Treey = C4.5 (Dv)
13. Attach Tree v to the corresponding branch of Tree
14. **end for**
15. **return** Tree

IV. METHODOLOGY

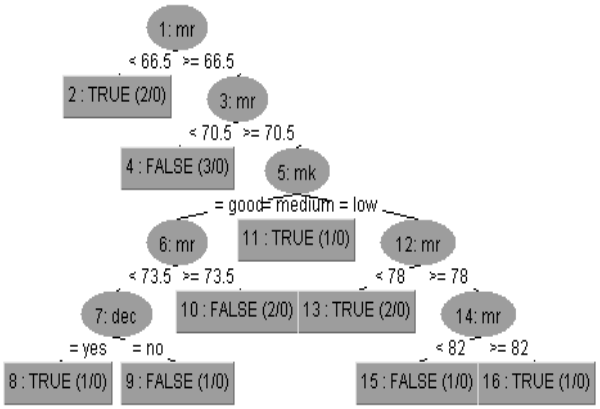
For this we have taken one B2B marketing Data set and constructed the System Model, the Manufacturer is responsible for the effective promotion of Business Products. This requires a thorough knowledge of the products being promoted, as well as the ability to solicit orders from outside Distributors, the manufacturer calls for bid from the Distributors. The Distributors will make registrations. There are n distributors (=d1,d2,d3,…..dn) and Manufacturer (M_B).The role of Manufacturer is to organize each auction, run and announces the bid result based on the attributes Forecast of Purchase (FP), Marketing Knowledge (Mk), Manufacturer Relationships (MR) and Advertising Support (As) and announces the selected distributor among the registered distributors using decision trees [12].

The channels between Manufacturer and Distributor are to secure and reliable. For selecting the best distributors, we proposed a secure decision making algorithm to select efficient distributor.

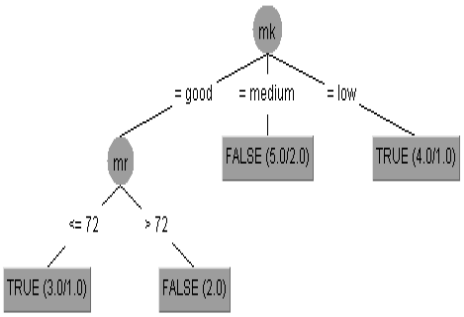
A. Symbolic Attribute Description

Attributes	Possible Values
Forecast of Purchase (Fp)	High, Low, Medium
Marketing Knowledge (Mk)	Good, Minimal, Low
Manufacturer Relationship (Mr)	Respectable, Imperfect
Advertising support(As)	Excellent, Satisfactory, poor
Decision	Yes, No

B. Distributor Decision Tree using ID3 Algorithm



C. Distributor Decision Tree using C4.5 Algorithm



V. Comparative Analysis Between ID3 and C4.5 Algorithm

Algorithm		
Size of Data Set	**ID3**	**C4.5**
14	35968KB	32481KB
24	35968KB	33688KB

A. Performance Parameters

1)Accuracy: The measurements of a quantity to that quantity's factual value to the degree of familiarity are known as accuracy.

Accuracy = No. of true positive + no. of true negatives
No. of true positive + false positive + false negative + true negative

Algorithm		
Size of Data Set	**ID3**	**C4.5**
14	94.155%	94.155%
24	78.472%	78.472%

Table 1: Accuracy Comparison between ID3 and C4.5 Algorithm in Percentage

2) Memory Used: How much amount of memory is used by a particular program to build an execute it successfully in different condition is known as memory used.

Algorithm		
Size of Data Set	**ID3**	**C4.5**
14	0 sec	0 sec
24	0 sec	0 sec

Table 2 : Memory Used comparison between ID3 and C4.5 algorithm in KB

3) Model Build Time: Extraction from dataset to data model is known as Model built time. It is depend upon the number of dataset used in training. And accuracy of program is depending upon the number of dataset, greater number of dataset there will be more accuracy.

Table 3: Models Build Time Comparison between ID3 and C4.5 Algorithm in Sec

4) Search Time: Search time is defined as after building model answering time of system is called search time.

Algorithm		
Size of Data Set	**ID3**	**C4.5**
14	0.0156 sec	0 sec
24	0 sec	0 sec

Table 4: Search Time Comparison between ID3 and C4.5 Algorithm in Sec

VI. Conclusions

In B2B Marketing, the Manufacturer plays a crucial role to select the efficient distributor among the available distributors, to avoid the misunderstandings and also to overcome the heavy competition between the distributors, the major responsibility of the manufacturers is to produce a product that can satisfy the expectations and desires of their customers and distributors. For this, there is a secure decision making mechanism is very much needed. By using the B2B marketing Data Set, we constructed the decision Tree to select the efficient distributors. In this paper, we performed comparative analysis between two classification algorithms ID3 and C4.5, and the result was satisfactory of C4.5. Experimental evaluation on real world data shows that C4.5 can learn to identify users simply by what commands they use and how often, and such an identification can be used to detect intrusions in a network computer system.

.

VII. REFERENCES

[1] Kamlesh K. Bajaj, Debjani Nag, "E-Commerce: The Cutting Edge of Business" Tata McGraw-Hill. p.10-14, 2005.
[2] C. Phani Ramesh, Prof. M. Padmavathamma,, "Threshold Secure B2B Model**",** IOSR Journal of Computer Engineering (IOSRJCE) ISSN: 2278-0661, ISBN: 2278-8727 Volume 5, Issue 6 (Sep-Oct. 2012), PP 11-14.
[3] H. Saini, D. Saini and N. Gupta, " E-Business system development: review on methods, design factors, techniques and tools with an extensive case study for secure online retail selling industry", International Journal of Science and Technology, Vol 2. No.5 (May 2009), pp.82-90.
[4] Lalanthika Vasudevan , S. E. Deepa Sukanya , N. Aarthi, "Privacy Preserving Data Mining Using Cryptographic Role Based Access Control Approach", Proceedings of the International MultiConference of Engineers and Computer Scientists 2008, Vol IIMECS 2008, Hong Kong, 19-21 March, 2008.
[5] Leonard A.Breslow and David W. Aha, "Simplified Decision Trees: A survey", The Knowledge Engineering Review, Vol.12:1, 1997, pp.1-40.
[6] Quinlan, J.R. "Induction of decision trees", Machine Learning", vol.1, No.1, 1986, pp. 81-106.
[7] C. Phani Ramesh, Prof. M. Padmavathamma, "Threshold Secure B2B Model", IOSR Journal of Computer Engineering, Vol.5, No.6, Sep-Oct, 2012, pp.11-14.
[8] Ueli Maurer, "Secure Multi-party computation made simple", Security in Communication Networks, Vol.2576, 2003, pp.14-28.
[9] Houston, F. and Gassenheimer, J.,"Marketing and exchange. Journal of Marketing," 51 (October), 3–18, 1987.
[10]Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, Clifford Stein - Introduction to Algorithms, 2nd edition - MIT Press and McGraw-Hill 2001.
[11] Veronica S. Moertini, "Towards the use of C4.5 Algorithm for Classifying Banking Dataset, INTEGRAL, Vol. 8 No. 2, October 2003.
[12] C. Phani Ramesh, Prof. M. Padmavathamma, " Threshold Secure B2B Model", International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 10, December- 2012