ISSN:0975-9646

ID3 and Its Applications in Generation of Decision Trees across Various Domains- Survey

L.Surya Prasanthi[#1], R.Kiran Kumar[#2]

[#1]*Research Scholar, Department of Computer Science, Krishna University, Machilipatnam, India.*
[#2]*Department of Computer Science, Krishna University, Machilipatnam, India.*

***Abstract*— Data Mining is widening its scope by using new algorithms applicability in several domains which assist us in gaining more knowledge and further in decision making. Decision Tree is one such important technique which builds a tree structure by incrementally breaking down the datasets in smaller subsets. Decision Trees can be implemented by using popular algorithms such as ID3, C4.5 and CART etc.The present study considers ID3 algorithm to build a decision tree. Entropy and Information Gain are used by ID3 to construct a decision tree. This paper presents a survey on the application of ID3 in various fields such as Medicine, Health, Education, Computer Forensics, Web attacks, Food database.**

***Keywords*— Data Mining, Decision Trees, ID3, Entropy, Information Gain.**

I. INTRODUCTION

DATA MINING

Data mining is relatively a new concept emerged in 90's as a new approach to data analysis and knowledge discovery. Data mining has originated from statistics and machine learning as an interdisciplinary field. The degree of information is growing in bulk and people's ability of using information technology to collect and produce data is significantly enhancing. In such a huge volume of data, discovering useful knowledge and improving the effectiveness of information utilization are the challenges to be addressed. It was under this background, Data Mining [1] evolved.

Many researchers and analysts gave many definitions to data mining. It is the process of investigating data from different outlook and summarizing it into useful information that can be used in many fields for analysis. In simple, data mining is picturizing, classifying a dump of confused data into a meaningful and understandable data.

Data mining have several classes of tasks. Classification is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam". One of the Classification techniques is Decision Tree which generates a tree like mode.

DECISION TREE —

In today's commercial era, every organization is competitively working to be the top most in the market. Data Analysis and Useful Information are needed for the organizations to grow in the right direction.

Decision making is the most important step which makes to reach the standards. Decision makers need to analyze the existing data and know the conditions that need to be implemented via effective techniques. Decision Trees is one such effective technique. The use of Decision tree in data mining is to predict the data from the existing one that is similar to classification and segmentation.

Usually, decision trees are divided into two categories:

1) Classification trees- that classify objects into one of the predefined categories (e.g., letters during text recognition)
2) Regression trees- that predict an actual value (e.g., the profitable price of new consumer goods).

A decision tree is grouping of Mathematical, logical and Computational methods to create a model from a dataset that predicts the value of target variable based on several input variables. Decision tree induction is one of the most employed methods to extract knowledge from data, since the representation of knowledge is very intuitive and easily understandable by humans [9]. The most successful approach for inducing decision trees, the greedy top-down approach, had been continuously improved by researchers over the years. A decision tree as the name implies is a directed tree which represents rules with a node called ROOT that has no incoming edges. A node that represents outside is known as test node and remaining are considered as leaves or decision nodes.

A decision tree is a classifier in the form of tree expressed as a recursive partition of the instance space. A decision tree is a flow chart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node holds a class label. [17] Each node in a decision tree can be either a leaf node or a decision node where a leaf node indicates the value of target attribute and decision node specifies tests to be carried out on attributes & values with an outcome of one branch and a sub tree.

In construction of a decision tree each internal node splits the dataset into two or more subspaces based upon the input attribute values. [2] To keep it simple each test considers a single attribute so that the input dataset is segmented based upon values. Similarly each leaf node is assigned to one class that represents appropriate target value.

To construct an effective decision tree with less complexity, a good decision maker prefers only less

complex decision tree. The complexity of a decision tree increases as number of nodes increases, number of leaves increases, attributes and depth of the tree. A disadvantage of this model is that there will always be information loss, because a decision tree selects one specific attribute for dividing at each stage with a single starting point. A decision tree has only single starting point hence it can present only one outcome. Therefore, decision trees are suited for data sets where there is one clear attribute to start with. Small errors in a training data set can also lead to very complex decision trees.

A decision tree can be built using many algorithms. Among them Iterative Dichotomiser (ID3) tree is most used method which is discussed in the next section.

ID3 (Iterative Dichotomiser)

To construct a decision tree, there are many algorithms like ID3, C4.5, CART, etc., ID3 constructs decision tree by employing a top-down, greedy search through the given sets of training data to test each attribute at every node. ID3 algorithm select the attribute to be splitted based on two metrics.

1) Entropy Metric: It measures the amount of information in an attribute. Entropy is calculated for all the remaining attributes. Split occurs at the attribute that has smallest entropy.

2) Information Gain: It is a statistical property which measures how well a given attribute separates training examples into targeted classes. The one with the highest information (information being the most useful for classification) is selected based on entropy.

Algorithm:

Step 1: Initially calculate classification entropy.

Step 2: Select the attributes and for each attribute, calculate information gain

Step 3: Highest information gain attributes are figured out.

Step 4: Remove node attribute, for future calculation.

Repeat steps 2-4 until all attribute have been used.

ID3 (E, T_A, A)

E- Examples are the training examples.

T_A- Target Attribute is the attribute whose value is to be predicted by the tree. [5]

A- Attributes are the list of attributes which may be tested by the learned decision tree.

- Initially create a root node.
- Return the single-node tree Root, as + If all examples are positive,
- Return the single-node tree Root, as - If all examples are negative,
- If number of predicting attributes is empty, then Return the single node tree Root, with label = most common value of the target attribute in the examples
- Otherwise Begin
  - A ─ The Attribute that best classifies examples
  - Decision Tree attribute for Root ─ A
  - For each positive value, v_i , of A, *f*
  - Add a new tree branch to Root, corresponding to the test A = v_i
  - If v_i = empty
    - Then add a leaf node with label as most common target value in examples.
    - Else add the sub tree ID3 (E (v_i), T_A, A)
- End
- Return Root

Because of this potential, ID3 can be applied in several domains to generate decision tress.

APPLICATIONS OF ID3

The decision tree tells what customers want the most. Nishant mathur and Sumit Kumar et al.,[6] used Havrda and Charvat entropy instead of Shannon Entropy to find the information of different properties to identify the node of decision tree. This modified entropy has reduced the size of tree as well as decreased the rules, which will help to understand customer. They concluded that, for lower value of alpha (α) = 0.25, tree is small and less complex as compared to the use of Shannon entropy. As value of α increase the size of tree may also vary and complexity may increase.

1) ID3 Application on Food Database

Ashvini Kale, Nisha Auti[7] in their research explained about the use of ID3 algorithm in implementation of Automatic menu planning for children as recommended by dietary management system. This research was carried out using Indian food database as many of the Indian children are affecting to mal nutrition due to mothers ignorance on nutrition facts. Approximately 30% of the new born children are having problems of low weight and hence easily susceptible to diseases. Vitamins and mineral deficiencies also affect children's survival and development. Anemia affects 74% of children below the age of three, more than 90% of adolescent girls and 50% of women. The proposed method of food suggestion for children is based on the factors such as food preferences, availability of food, medical information, disease information, personal information, activity level of a child, for Indian food database. The important task in implementation is to recommend the particular food item from the food database based on certain attributes such as likeliness, availability, its nutritional contents and decision of child. This result helps to select the food from the database such that the deficiency will not occur in near future and proper diet plan will be given to the child. Based on the outcome, one can suggest efficient food to be assigned to child. The ID3 algorithm is used to take proper decision among available foods. Related to the proposed research work it concludes that diet plan can be built using ID3 algorithm, as its error rate is zero.

2) Web Attack Detection Using ID3

In a research by Garcia, YH., Monroy, R., Quintana, M.,[8], explained how ID3 can be used in detection of web attacks. As today's smart technologies enable every operation to perform online, transactions has increased a lot. At the same time,the attacks on these online sites have also increased. Hence many organizations prefer an Intrusion Detection System. The main problem of existing

IDS is that they cannot detect mimicry attacks and new attacks as this problem prevails in well known IDSs, like snort. To solve this issue, IDS researchers have turned their attention to machine learning techniques, including classification rules and neural networks.

In their experiments unlike other IDS, ID3 was able to classify even unseen Web application queries as an attack. To classify this, they used a training set of 400 web application attacks queries from three vulnerabilities, and also gathered 462 web application non attacked queries. They used various window sizes (5, 8, 10, 12, 15) among which size 10 gave best decision tree which precisely captured the examples. After building a decision tree, inputs were assigned to ID3 and they indicate both the false alarm rate and the missing alarm rate, considering two sets of attacks. The results obtained prove that IDS is a competitive alternative for detecting Web application attack queries by using ID3.

3) Application of ID3 in Diabetes

Diabetes is one among the challenging diseases that the human race is finding difficulties which is prevailing since years. Many of the analysts around the globe have been working on diabetes and making aware of the signs and effects of the disease on the various organs of the body.

Daveedu Raju Adidela et al.,[10] in their research described some rules in prediction of diabetes. According to the survey, causes for Diabetes are of two types, Insulin Dependent and Non-Insulin Dependent diabetes. According to WHO, around 300 million people would be affected by this decease by the year 2025 around the world. It is observed that the most of the people affected by non-insulin dependent diabetes and it is genetic in nature up to 95%. There are several risk factors that influence the second type diabetes with hyper tension, obesity, over consuming of alcohol and fat diet, desk-bound nature, aging and many more. There are other specific factors that imply the cause of diabetes such as in take of certain medicines, infected pancreas or damaged pancreas which fails to produce insulin properly and gestational diabetes (caused during pregnancy). The information gain is calculated for all the attributes and selects an attribute which has maximum information gain for the root node. Fuzzy ID3 algorithm an extension of ID3 is used to obtain decision tree for individual clusters. The decision tree of each cluster produce's a set of adaptation rules from which diabetic effected patient is identified. The eight attributes of the individual person is given to the system and by utilizing these adaptation rules, the system will predict the person as tested_positve or tested_negitive of diabetes. [10]

The authors finally concluded that this process can be much more helpful and needs some improvement by taking some more attributes in to consideration like Eye vision, food habits of family, etc., More over this process can be applied on many other diseases and reduce the number of disease affected people. This model can also be implemented in other areas like weather forecasting, business transactions, agricultural fields, etc.

4) Efficiency of ID3 in monitoring Heart Attack:

D.Senthil Kumar [11] in his research took three diseases Heart attack, Hepatitis and Diabetes and used decision tree algorithms like ID3, C4.5, CART and found the complexity of ID3 as 65% in identifying the diseases. This can be much more improved by conducting a regular diagnosis of various other diseases dataset. The heart disease dataset of 473 patients is used in this experiment and has 76 attributes, 14 of which are linear valued and are relevant. The hepatitis disease dataset has 20 attributes, and there are 281 instances and 2 classes. The diabetic dataset of 768 patients with 9 attributes. Features like age, sex, chest pain, etc., depending on the disease are considered and 4 performance measures like Precision, Recall, ROC, F-Measure are used and a confusion matrix is constructed to classify the data with the above measures.

5) ID3 in Identifying Cancer:

Priyadharsini.C, Dr. Antony Selvadoss Thanamani et. al.[12] have analyzed an important process of identifying cancer in early stages using ID3 algorithm. As we know cancer has become a dreadful disease and affecting many people's health, this analysis helped to identify the early stages of cancer. For this analysis, they used Multi-dimensional Array model with modified ID3 algorithm. The modified ID3 algorithm compares the current database with the previous dataset and identifies the results as positive or negative. In case a patient is affected with that disease, this algorithm shows the infection percentage [12]. Here ID3 is used to split training examples in to target classes, the one which gives highest classification is selected and used.

6) Application of ID3 in Computer Forensics

Data analysis is the most crucial part in computer crime forensics system. The result of data analysis has a direct impact on the validity and credibility of the evidence. In the prototype system of computer crime forensics, the general practice is making use of ID3 algorithm directly, but in this way it does not effectively mine a reasonable model because of the versatility of ID3 algorithm and the uniqueness of forensic data. According to characteristics of computer crime forensics data, this paper [13] puts some improvements of ID3 algorithm in terms of information gain to make it more suitable for computer crime forensics field data, and experiments show that the improved algorithm is effective. As the diversity of attacks, in the extraction of features and attributes of behaviors, if we still choose the largest value of information gain as the property of division to construct division tree, it will generate very much redundant information, and even result in error message when march between the input event and the rule base.

This work has been explained very clearly by Yuesheng Tan, Zhansheng Qi, Jingyu Wang. [13]. Initially they took Snort an Intrusion detection tool which is used in TCP/IP to identify wide variety of network Traffic. Their work was carried out in Windows XP environment which has much vulnerability and hence suffers lot of attacks. They got a huge data of 400 pieces which is enough for data analysis.

Entire work has been performed by comparing alarm levels and number of attacks. From a list of 15 attributes, 9 were Positive and 6 were Negative, whose information gain is calculated and recorded. After experts experience on the data classification of alarm level and number of attacks, they conclude that Alarm level is not sufficient to classify the attacks. After looking into the table of number of attacks, they choosed it as best classification to divide the sample set into two classifications. In their conclusion they observed that time complexity and space complexity of ID3 has been improved than earlier algorithm and the decision tree constructed by improvised algorithm has less branches and nodes than earlier algorithms.

7) Application of ID3 in Knowledge Acquisition for Tolerances Design:

In a research on Knowledge Acquisition by Xinyu Shao, Guojun Zhang, Peigen Li, and Yubao Chen [14], ID3 algorithm has been improved using previous knowledge. Tolerance Design is the total amount by which a given dimension may vary, or the difference between the limits. Tolerance engineering affects areas like Product design, Quality Control and Manufacturing. Knowledge processing can be used to aid engineering design. Knowledge processing technology is utilized in Intelligence and can be incorporated in existing CAD systems. Prior to implementation of ID3 some premises should be checked, they are:

i) Tolerance Description
ii) Function satisfied
iii) Parting line related
iv) Mold design (Cool well designed and Gate well designed)
v) Machine capability.
vi) Design

The above attributes are checked thoroughly and later on Information gain is calculated.

Knowledge acquisition is a time consuming process as it needs much discussion and verification as sometimes knowledge may be delude which causes loss of resources and inefficiency in designing. This research mainly concentrated on this and implemented ID3 algorithm as there is no need to assign properties to attributes, hence we can skip the task of building hierarchy. The results obtained by ID3 itself creates a hierarchy. This method can be adopted in real engineering design areas where large amounts of data knowledge and experience is scattered.

8) Application of ID3 to reduce Cost sensitive Decision Tree

Cost sensitive learning is one of the challenging concepts in data mining. Many decision tree researchers got attracted to this challenge and tried to get a solution to design a cost efficient decision tree. *Fan Min and William Zhu et al.,[15]* worked on this problem and prepared a model for competition strategy to cost sensitive Decision tree. In their research, they used ID3 algorithm and advanced its work using λ values for λ-ID3, where λ-ID3 coincides with ID3. In their work initially they developed a population of decision trees using ID3 and EG2 using information gain and test cost.

It works as follows:

Step 1: A number of decision trees are produced using different algorithms and/or different parameter values.

Step 2: These decision trees are post-pruned.

Step 3: The decision tree with the least cost on the training set is selected for classification.

Each time they used a mushroom dataset in their work they used 60% of the training set and remaining as test set. This strategy is simple because it does not change the structure of any candidate decision tree. It is efficient because not too many candidate decision trees are needed. It is effective in selecting a good decision tree for classification.

8) Use of ID3 for Breast Tumor Diagnosis

Decision tree classifiers are used extensively for diagnosis of breast tumor in ultrasonic images, ovarian cancer and heart sound diagnosis. D.Lavanya, Dr. K.Usha Rani [17] in their research on various decision tree algorithms showed that ID3 is used majorly. They compared the time complexities for ID3, CART, and C4.5 for different diseases and concluded that time complexity of ID3 algorithm is less to build a model among the three classifiers but Accuracy is very less compared to CART, which further needs to be improved.

ID3 in multi array model algorithm is explained as follows:

$E = D1 \times D2 \times \ldots \times Dn$ be finite-dimensional vector n, where Dj is a finite set of discrete symbols, E elements e = is the sample, vjDj, j = 1, 2, ..., n. PE is the positive sample set, NE is the anti-sample set, and the number of samples which are p and n depiction to the regulations of information theory.

The proposed sample data used by ID3 has certain requirements, which are:

Attribute-value description, Predefined classes, an example's attributes, discrete classes, sufficient examples. The proposed model is simple to understand and interpret, requires little data preparation, Able to handle both numerical and categorical data, possible to validate a model using statistical tests, Robust, Performs well with large datasets.

9) ID3 Algorithm for Predicting Heater Outlet Temperature:

ID3 algorithm, used in the field of machine leaning, is applied to structure identification of premises of a fuzzy model. This method was applied to a system to predict heater outlet temperature. Good results were obtained and the system has been operated in the required accuracy for a year. ID3 algorithm is used to select the effective variables in premises of fuzzy model and compute their boundary values. [18]

Many kinds of machine learning methods have been reported. We choose “ID3" algorithm because:

(1) ID3 can classify many cases with information content.
(2) ID3 can generate IF-Then type rules automatically.
(3) ID3 can select effective variables in premises in order.

They applied the fuzzy modeling to match with operators' procedure. About 60 cases of past two years are examined. "Feed rate" is selected as the most effective variables in premise of rules. And other variables were found to be not important operators' experience. Information content is used to define variables in premise of fuzzy model.

10) ID3 for Preserving Decision Tree:

Chris Clifton et al., had performed a research on ID3 algorithm for preserving decision tree over vertically partitioned data when there are more than two parties involved in a work with an arbitrary number of parties where only one party has the class attribute. To perform this and prove, they evaluated security of algorithm under the basic framework of Secure Multiparty Computation. The proof that was given by them was assuming semi honest adversaries due to space constraints. All algorithms can't provide complete privacy like constituent algorithms, in full algorithm the leaked information can be pretended by knowing the distribution counts at each node, so overall privacy is maintained.

11) Application of ID3 in Educational Field:

The main aim of any educational institution is to give a quality education to their wards, by knowing their interest and needs which should have better classification techniques. In an analysis by Brijrsh Kumar Baradwaj et al.[19], took a student dataset in a college of 50 students consisting data of attendance, previous semester performance, assignments, and class tests and hence predicting the fail ratio and trying to improve it in further years by generating a rule set [19]. By this task, they extracted a knowledge that describes students' performance in end examinations which help the dropout students who need special care and hence institution gets an idea of what else can be implemented to improve those students.

Like the above results, many other researchers have also worked on ID3 algorithm on various databases and concluded many results in the field of education for Placement analysis of fourth year students by classifying their overall performances and also to identify the first year student's dropout classification.

CONCLUSION

This paper focused on ID3 Decision Tree algorithm for classification. The present study reviewed Robust Decision tree algorithm ID3 and its applications in wide range spectrum of domains such as Health, medical, Education, Engineering etc. Across all the domains, the performance of ID3 has resulted in good performance. However, splitting criterion and pruning can be further improved to achieve higher accuracy and generalization. A minute increase in performance and generalization will yield better results and analysis, particularly in Health care domain. Hence our future work focuses on developing a simplified decision tree algorithmic model by using a novel splitting criterion and a pruning technique, with the objective of increasing accuracy and generalization.

REFERENCES

1. Chen Jin, Luo De-lin, Mu Fen-xiang, "*An Improved ID3 Decision Tree Algorithm*" Proceedings of 2009 4th International Conference on Computer Science & Education 978-1-4244-3521-0/09/$25.00 ©2009 IEEE
2. K. Gra¸bczewski, "Meta-Learning in Decision Tree Induction", 11Studies in Computational Intelligence 498, DOI: 10.1007/978-3-319-00960-5_2, © Springer International Publishing Switzerland 2014.
3. Lior Rokach and Oded Maimon "*Top-Down Induction of Decision Trees Classifiers*" – A Survey IEEE TRANSACTION ON SYSTEMS, MAN AND CYBERNETICS: PART C, VOL. 1, NO. 11, NOVEMBER 2002
4. MichalWozniak "*A hybrid decision tree training method using data streams*" Knowl Inf Syst (2011) 29:335–347 DOI 10.1007/s10115-010-0345-5 Published online: 5 October 2010 published with open access at www.springer.com
5. Anand Bahety. "*Extension and Evaluation of ID3 – Decision Tree Algorithm*". University of Maryland, College Park.
6. Nishant Mathur, Sumit Kumar, Santosh Kumar, and Rajni Jindal "*The Base Strategy for ID3 Algorithm of Data Mining Using Havrda and Charvat Entropy Based on Decision Tree*" International Journal of Information and Electronics Engineering, Vol. 2, No. 2, March 2012
7. Ashvini Kale, Nisha Auti "*Automated Menu Planning Algorithm for Children: Food Recommendation by Dietary Management System using ID3 for Indian Food Database*" Procedia Computer Science 50 (2015) 197 – 202, Published by **Elsevier**.
8. Garcia, YH., Monroy, R., Quintana, M., "*Web Attack Detection Using ID3*" 2006, in IFIP International Federation for Information Processing, Volume 218, Professional Practice in Artificial Intelligence, eds. J. Debenham, (Boston: **Springer**), pp. 323-332.
9. Rodrigo C. Barros, Marcio P. Basgalupp, Alex A. Freitas, "*Towards the Automatic Design of Decision Tree Induction Algorithms*" GECCO'11, July 12–16, 2011, Dublin, Ireland. Copyright 2011 **ACM** 978-1-4503-0690-4/11/07
10. Daveedu Raju Adidela, Lavanya Devi. G, Jaya Suma. G, Appa Rao Allam "*Application Of Fuzzy Id3 To Predict Diabetes*" International Journal of Advanced Computer and Mathematical Sciences ISSN 2230-9624. Vol 3, Issue 4, 2012, pp541-545
11. D.Senthil Kumar, G.Sathyadevi and S.Sivanesh "*Decision Support System for Medical Diagnosis Using Data Mining*" IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 3, No. 1, May 2011 ISSN (Online): 1694-0814 www.IJCSI.org
12. Priyadharsini.C, Dr. Antony Selvadoss Thanamani "*Prediction of Missing Values in Blood Cancer & Occurrence of Cancer Using Improved Id3 Algorithm*" International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 2, Issue 8, August 2014.
13. Yuesheng Tan, Zhansheng Qi, Jingyu Wang *"Applications of ID3 Algorithms in Computer Crime Forensics"* 978-1-61284-774-0/11/$26.00 ©2011 **IEEE**.
14. Xinyu Shao, Guojun Zhang, Peigen Li, and Yubao Chen, "*Application of ID3 in* Knowledge Acquisition for Tolerance Design" Journal of Materials Processing Technology 117 (2001) 66-74. Published by **Elsevier** Science B.V.
15. Fan Min and William Zhu "*A Competition Strategy to Cost-Sensitive Decision Trees*" T. Li et al. (Eds.): RSKT 2012, LNAI 7414, pp. 359–368, 2012. **Springer**-Verlag Berlin Heidelberg 2012.
16. Shikha Chourasia "Survey paper on improved methods of ID3 Decision Tree Classification" International Journal of Scientific and Research Publications, Volume 3, Issue 12, December 2013 1 ISSN 2250-3153.
17. D.Lavanya, Dr. K.Usha Rani *"Performance Evaluation of Decision Tree Classifiers on Medical Datasets"* International Journal of Computer Applications (0975 – 8887) Volume 26– No.4, July 2011
18. Tetsuji TAN1 and Woto sakoda "*Fuzzy modeling by ID3 algorithm ad its application to prediction of Heater outlet Temperature*" 0-7803-M36-2 192 $3.00 8 1992 **IEEE.**
19. Brijesh Kumar Baradwaj, Saurabh Pal "Mining Educational Data to Analyze Students' Performance" IJACSA Vol. 2, NO.6, 2011.