Comparative Study Between ID3 Algorithm and Improved ID3 Algorithm in Datamining

O Yamini[1], G V Ramesh Babu[2]

[1]*Research Scholar, Department of Computer Science, S V University, Tirupati, A P*

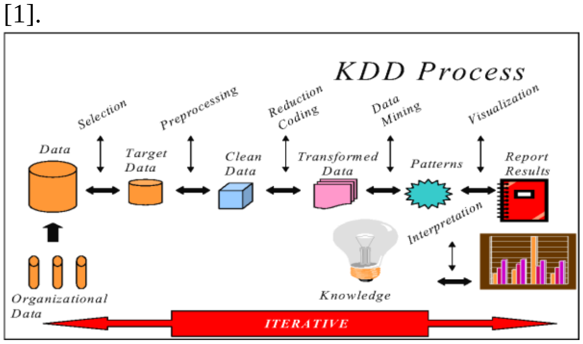
[2]*Research Supervisor, Department of Computer Science, S V University, Tirupati, A P*

***Abstract*—Data mining includes several important technologies such as classification, clustering, regression, etc. Categorical mining technology, among data mining technologies, is becoming the most mature and active research direction enabling successful applications. Categorical mining can be applied to uncover useful information from large amounts of data stored in a large number of fields such as hospitals, inventory, banking, etc. Decision tree methods, neural network methods and statistics exploratory methods. A large number of decision tree algorithms such as Iterative Dichotomizer 3, C4.5 and Classification and Regression Tree (CART) are used in different fields. Most of the decision tree methods are developed from the ID3 method. Based on the disadvantages of the disadvantages of the ID3 algorithm, the improved ID3 algorithm is developed using Taylor's attitude. In this article, we will compare ID3 Algorithm and Advanced ID3 Algorithm to see how effective the Advanced ID3 Algorithm is.**

***Index Terms*—Data mining, ID3 classification algorithm, Improved ID3 algorithm.**

I.DATA MINING

Data mining is a step in the knowledge discovery process in a database that includes the application of data discovery and analysis algorithms that, within acceptable limits of computational efficiency, generate a table that lists specific patterns on the data [1].



1.1 Data mining functions

Data mining functions or tasks can be used to specify the types of patterns or knowledge to be discovered during data mining. Some of the main data mining functions are association, clustering, classification, outlier analysis, regression and prediction, etc. [2]

a) Clustering:

Clustering is used to partition or segment data objects (or observations) into subsets known as groups or clusters. Objects that are close to each other are positioned in the same group. Like classification, clustering classifies data objects that are similar but unlike classifiers, unknown class label (i.e. unsupervised learning) [4]. Cluster analysis is one of the most popular techniques used not only in data mining but also in other fields such as statistics, image segmentation, pattern recognition, object recognition, information retrieval, bioinformatics, etc [5].

A large collection of clustering algorithms has been proposed by many researchers [6, 4, 7] in the past two decades. Some popular clustering algorithms are presented in Table. Clustering algorithms based on probability model, fuzzy set, expectation maximization, correlation using PCA, graph have also been proposed by some researchers. proposed research.

b) Classification

A classification algorithm is used to predict the data classes [4]. To date, a large collection of classification algorithms (or classifiers) have been proposed by researchers [4, 8].

c) Outlier analysis

The outlier's area unit typically discarded by most of knowledge mining strategies as noise or exceptions. Sometimes, outliers might have additional data as compared to alternative information objects. So outlier analysis is very important for a few application areas like intrusion detection, fraud detection, anomaly detection, etc. [9].

Several data processing techniques typically use bunch to detect the outliers as a noise. The outlier detection strategies is classified as classification-based strategies, applied mathematics strategies, clustering-based strategies, supervised, semi-supervised and unsupervised strategies, deviation-based strategies and proximity-based strategies [4]

d) Regression

Regression predicts the worth of attribute supported regression technique(s) over time. The long run values of variables area unit foretold with the assistance of historical statistic plot [4]. Analytic thinking (also known as as evolution analysis) discovers fascinating patterns within the evolution history of the objects. Identification of patterns in associate degree object's evolution and matching of the objects' dynamical trends area unit the 2 major aspects of analytic thinking [10]. Trends of the objects, whose behavior evolves over time, is delineated victimization analytic thinking and regression models. Analytic thinking exposes time-varying trends of the information objects at intervals the dataset. The association analysis can even be used for evolution analysis [11].

e) Prediction

Regression analysis is accustomed model the link between one or additional freelance or predictor variables and a dependent or response variable (which is continuous-valued). within the context of knowledge mining, the predictor variables area unit the attributes of interest describing the tuple (i.e., creating up the attribute vector). In general, the values of the predictor variables area unit glorious. (Techniques exist for handling cases wherever such values could also be missing.) The response variable is what we would like to predict—it is what we have a tendency to spoken in Section half-dozen.1 because the foretold attribute. Given a tuple delineated by predictor variables, we would like to predict the associated worth of the response variable.

II. CLASSIFICATION BY DECISION TREES

A decision tree could be a non-parametric supervised learning formula, that is used for each classification and regression tasks. It's a gradable, tree structure that consists of a root node, branches, internal nodes and leaf nodes. As you'll be able to see from the subsequent diagram.

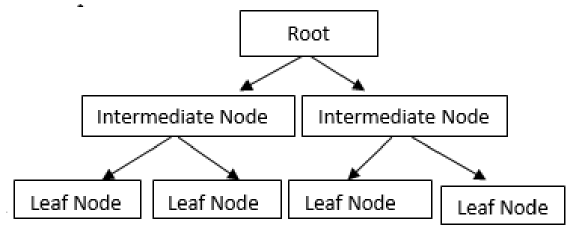


Fig: Decision Tree

Types of decision trees

Hunt's algorithm, developed in the 1960s to model human learning in psychology, is the basis of many popular decision tree algorithms, such as the following:

A) ID3: Ross Quinlan is credited with developing ID3, which stands for "Iterative Dichotomiser 3".This algorithm exploitsentropy and gain information as metrics to evaluate candidate splits. You can find some of Quinlan's research on this algorithm from 1986.

B) C4.5: This algorithm is considered a later version of ID3, also developed by Quinlan. It can use gain or gain rate information to evaluate split points in the decision tree.

C) CART: The term CART is an acronym for "classification and regression tree" and was introduced by Leo Breiman. This algorithm typically uses Gini impurities to determine the ideal property for separation. The Gini impurity measures how often a randomly selected attribute is misclassified. When evaluating using Gini impurities, a lower value would be more ideal. The following table shows the comparison of parameters between different decision tree algorithms. These algorithms are among the most influential data mining algorithms in the research community [4].

III. ID3 ALGORITHM

Iterative Dichotomiser 3 is an algorithm used to generate decision trees. This algorithm is based on Occam's razor: it prefers small decision trees over large decision trees[12]. However, it does not always produce the smallest tree, and is therefore a

heuristic. ID3 (Example, Target_Attribute, Attribute) Create a root node for the tree.
• If all samples are positive, return the tree to a node.
• Root, labeled = +. If all examples are negative, return a one-node tree
• Root, with label = . If the number of predictor attributes is empty, then return
• the root of a single-node tree, with label = the most current value of the target attribute. Otherwise begin
• A= Attribute best classified examples.
• Decision tree attribute for root = A
• For each possible value, vi
• from A, Add a new branch under the root, corresponds to the test A = vi .
Let Examples (vi), which is a subset of examples with the value vi for A.
If Example (vi) is empty, then below this new branch add a leaf node labeled = most common target value in the examples.
Otherwise, below this branch, add subtree ID3 (• Example vi), Target_Attribute, Attribute{A}) End
• Return Root

ID3 search in attributes of training instances and extract the attribute that best separates the given examples. If the attribute ranks the training sets perfectly, ID3 stops; otherwise it works recursively on m (where m = number of possible values of an attribute) partitioned subset to obtain their "best" attribute. The algorithm uses greedy research, i.e. it chooses the best attribute and never goes back to reconsider previous choices. The central goal of the decision tree growth algorithm is to select the attribute with a heterogeneous class distribution because the algorithm uses the concept of entropy.

From the above Algorithm let us discuss about a few crucial terms such as Entropy and Information Gain.

Formula for Entropy

$$Ent(D) = -\sum_{d=1}^{|k|} P_d \log_2 P_\mathrm{d} \qquad \text{--(1)}$$

Formula for Information Gain

$$Gain(D,a) = Ent(D) - \sum_{v=1}^{|V|} \frac{|D^v|}{|D|}, Ent(D^V) \quad \text{---(2)}$$

where D={(x_1, y_1), (x_2, y_2), _ _ __, (x_m, y_m)} the training sample set |D| represents the number of training samples. A = {a_1, a_2, --- , a_d} denotes the attribute set of |D|, d = {1, 2, ---, |k|}. p_k(k = 1, 2, --- , |D|) stands for the probability training set S belongs to class C_i.

In a homogenous data set, p_k is 1, and $\log_2(p_k)$ is zero. Hence, the entropy of a homogenous data set is zero. Assuming that there are V different values (V = { a_1, a_2, --- , a_V} in an attribute a_i, D_i represents sample subsets of every value, and |D_i| represents the number of current samples. A transport mode dataset is taken as an example to compute the ID3 algorithm.

Attributes				Classes
Gender	Car Ownership	Travel cost	Income level	Transportation
Male	0	Moderate	Low	Bus
Male	1	Moderate	Medium	Bus
Female	1	Moderate	Medium	Train
Female	0	Moderate	Low	Bus
Male	1	Moderate	Medium	Bus
Male	0	Standard	Medium	Train
Female	1	Standard	Medium	Train
Female	1	Expensive	High	Car
Male	2	Expensive	Medium	Car
Female	2	Expensive	High	Car

Table : Data set for Transportation mode

Calculating Highest Information Gain

(i) Calculating Entropy for the classes:

The record set for classes are : [Bus, Bus, Train, Bus, Bus, Train, Train, Car, Car, Car]

Entropy: 1.571

(ii) Calculating Entropy for the attributes

<u>Gender</u>

Record Set: [Male, Male, Female, Female, Male, Male, Female, Female, Male, Female].

Entropy: Male: 1.522

Entropy: Female: 1.371

Information Gain: 0.12

<u>Car Ownership</u>

Record set: [0, 1, 1, 0, 1, 0, 1, 1, 2, 2]

Entropy calculated for distinct data of particular column

Information Gain: 0.534

<u>Travel cost</u>

Record set: [Moderate, Moderate, Moderate, Moderate, Moderate, Standard, Standard, Expensive, Expensive, Expensive]

Information Gain: 1.21

<u>Income level</u>

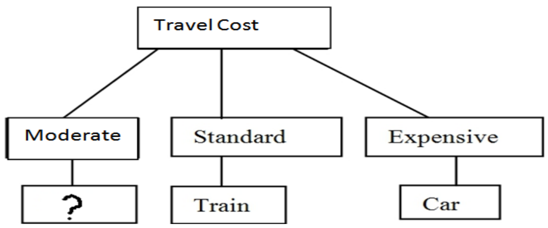
Calculating gain for record set:[Low, Medium, Medium, Low, Medium, Medium, Medium, High, Medium, High]

Information Gain: 0.695

(iii) Generating an Information Gain Table

Attributes	Information Gain
Gender	0.125
Car Ownership	0.534
Travel cost	1.21
Income level	0.695

Fig : Initial Decision tree primarily based on ID3



Same procedure is followed for each and every column and Information is represented as:

Attributes	Information Gain
Gender	0.322
Car Ownership	0.171
Income level	0.171

Table: Generation information gain after removing travel cost column

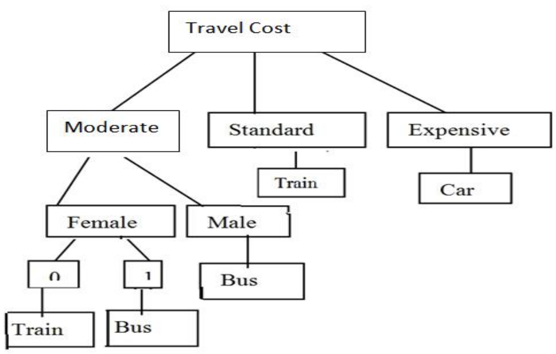


Fig : Transportation decision tree

Benefits

- Educational information is used to create easy to understand prediction rules.
- It builds the fastest tree as well as the fastest tree.
- ID3 searches the entire dataset to create the entire tree.
- It exposes the leaf nodes for this reason allowing to truncate the test information and reduce the number of checks.
- ID3 computation time is characteristic of linearity due to feature diversity and large number of nodes

Defects

- For a small sample, the information may be over-fitted or over-classified.
- It's up to you, only one feature to be tested on the site, this will take a lot of time.
- Classification of contiguous records can also be computationally expensive, as multiple trees must be generated to see where the continuity breaks.
- The disadvantage of ID3 is that when given multiple input value types, it is too sensitive to the possibility of multiple value types.

IV. IMPROVED ID3 ALGORITHM

According to the differentiation concept in advanced mathematics, which means that of Taylor formulation will regulate superior capabilities. The Taylor method is an increased form at any factor, and the Maclaurin components is a function that can be expanded into Taylor's collection at point 0[14]. The difficulty in computation of the knowledge entropy regarding the ID3 rule maybe reduced supported an approximation formulation of Maclaurin method this is useful to create a choice tree in a brief period of time [13].

The Taylor series is given by means of:

$$f(x)$$

$$= f(x_0)+f'(x_0)(x-x_0)+\frac{f''(x_0)}{2}(x-x_0)^2+......+\frac{f^{(n)}(x_0)}{n!}(x-x_0)^n \quad \text{------(3)}$$

When x=0, then the above equation may be changed in the following form. For easy calculation, the very last equation implemented here is written as:

$$f(x) = f(0) + f^{(1)}(0) + \frac{f^2(0)}{2!}x^2 + \cdots + \frac{f^n(0)}{n!}x^n \quad (4)$$

Let us assume that there are n counter examples and p high quality examples in pattern set D, the information entropy of D can be written as:

$$Ent(\mathrm{D}) = -\frac{\mathrm{p}}{p+n}\log_2\frac{\mathrm{p}}{p+n} - \frac{\mathrm{n}}{p+n}\log_2\frac{\mathrm{n}}{p+n} \quad \text{----- (5)}$$

Assuming that there are V extraordinary values protected in the characteristic a_i of D, and each value consists of n_i counter examples and p_i tremendous examples, the facts advantage of the attribute a_i can be written as:

$$Gain(D, a_i) = Ent(D) - \sum_{i=1}^{v} \frac{p_i+n_i}{p+n}\ Ent(D^v) \quad (6)$$

Where

$$Ent(D^v) = \frac{-p_i}{p_i+n_i}\log_2\frac{p_i}{p_i+n_i} - \frac{n_i}{p_i+n_i}\log_2\frac{n_i}{p_i+n_i} \quad \text{-- (7)}$$

For the simplification purpose,

$$Ent(D) = \frac{2pn}{p+n}$$

$Gain(D, a_i) == \frac{2pn}{p+n} - \sum_{i=1}^{v} \frac{2P_i n_i}{p_i+n_i}$

Hence, Equation (6) can be used to calculate the data advantage of every characteristic and the attribute that has the maximal records advantage may be selected because the node of a decision tree. The new information advantage expression in Equation (6) that only includes handiest the Arithmetic Operators greatly reduces the difficulty in computation and increases the facts managing potential. No logarithmic values are used right here. Calculating Information Gain for equal dataset reduces the time and complexity. Let us calculate Entropy and Information Gain for the above dataset that's utilized in ID3 Algorithm using Improved ID3Algorithm. Calculating Entropy for lessons of above statistics set.

Calculating Entropy for lessons of above statistics set

Entropy Ent(D)=2pn/(p+n)= 7.2

Calculating Information Gain for attributes of above dataset

Information Gain

$$\text{Gain}(D,a_i) = 2pn/(p+n) - \sum_{i=1}^{v} \frac{2P_i n_i}{p_i+n_i}$$

Information Gain for Gender:

$2pn/(p+n) - \sum_{i=1}^{2} \frac{2P_i n_i}{p_i+n_i} == 4.4$

Information Gain for Car Ownership:

$2pn/(p+n) - \sum_{i=1}^{3} \frac{2P_i n_i}{p_i+n_i} == 5.6$

Information Gain for Travel Cost:

$2pn/(p+n) - \sum_{i=1}^{3} \frac{2P_i n_i}{p_i+n_i} == 7.2$

Information Gain for Income Level:

$2pn/(p+n) - \sum_{i=1}^{3} \frac{2P_i n_i}{p_i+n_i} == 5.2$

Calculating Information Gain for attributes of above dataset after removing column 2 which is having highest Information Gain.

Entropy $Ent(D) = \frac{2pn}{p+n} = 1.6$

Information Gain

$Gain(D, a_i) == \frac{2pn}{p+n} - \sum_{i=1}^{v} \frac{2P_i n_i}{p_i+n_i}$

Information Gain for Gender:

$\frac{2pn}{p+n} - \sum_{i=1}^{2} \frac{2P_i n_i}{p_i+n_i} == 0.6$

Information Gain for Car Ownership:

$\frac{2pn}{p+n} - \sum_{i=1}^{3} \frac{2P_i n_i}{p_i+n_i} == 0.3$

Information Gain for Income Level:

$\frac{2pn}{p+n} - \sum_{i=1}^{3} \frac{2P_i n_i}{p_i+n_i} == 0.3$

Advantages

- Taylors collection widely used in Calculus, Computational series, Numerical Methods, Inequalities, Local maxima and minima.
- Basically, Taylor's set provides an approximation of possibilities equal to a polynomial.
- Very useful for derivative formulas and can be used to obtain theoretical error limits.
- The power series can be reversed to produce the inverse property.

V. RESULTS AND GRAPHS

Algorithm	Time
ID3	0.1683
Improved ID3	0.0364

Table: ID3 vs Improved ID3 comparison time

The resulting graph is fully plotted based on the ID3 algorithm runtime in addition to the Improved ID3algorithm using MATlab. It is clear that the Improved ID3Algorithm runs in much less time than the ID3 Algorithm. The complexity is also reduced with the proposed new algorithm.

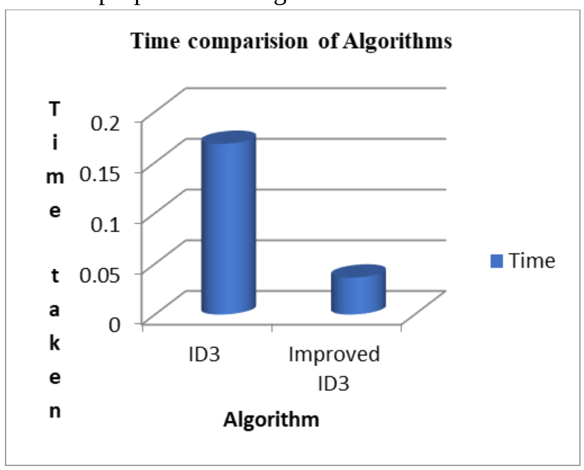


Fig: Comparison between ID3 and IID3 algorithms

The time taken to run properties is also shown below compared to Improved ID3and ID3. The bar chart comparisons for all attributes are

Algorithm	ID3	IID3
ENTR0PY	0.0602	0.0126
IGG	0.0654	0.0253
COIG	0.0712	0.0268
TCIG	0.0775	0.0271
ILIG	0.0839	0.0273

Tab : Entropy and gain calculation time for all attributes

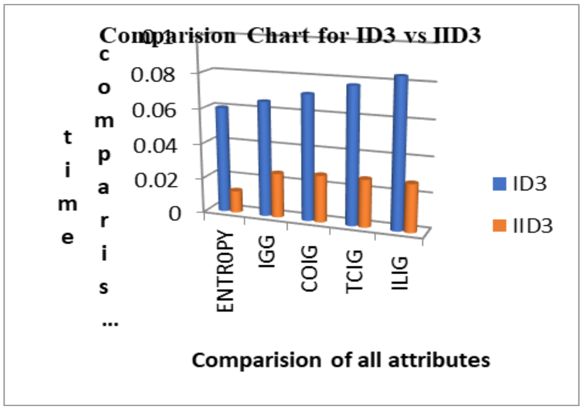


Fig: Time comparison chart for attributes of ID3 and IID3 Algorithm

VI CONCLUSION

This paper shows that here it is complex to calculate the entropy and attributes and using ID3 Algorithm whereas it is simple to calculate arithmetic expressions using Improved ID3 Algorithms. So that it is easy to construct a decision tree using this process within a short period of time. The time taken between ID3 and Improved ID3 algorithms along with entropy and its attributes using MATLab. Comparison chart has been shown in above along with graphs between ID3 and Improved ID3 Algorithms.

REFERENCE

[1] Fayadd U, Piatesky-Shapiro G, Smyth P (1996) From data mining to knowledge discovery in databases. AAAI Press/The MIT Press, Massachusetts Institute of Technology. ISBN 0–26256097–6.

[2] Chen M, Han J, Yu PS (1996) Data mining: an overview from a database perspective. IEEE Trans Knowl Data Eng 8(6):866–883.

[3] *Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth,* From Data Mining to Knowledge Discovery in Databases, American Association for Artificial Intelligence. Magazine Volume 17 Number 3 (1996), 0738-4602.

[4] Han J, Kamber M, Pei J (2012) Data mining concepts and techniques, 3rd edn. Elsevier, Netherlands

[5] Jain AK, Murty MN, Flynn PJ (1999) Data clustering: a review. ACM Comput. Surv 31(3):1–60

[6] Fan Chin-Yuan, Fan Pei-Shu, Chan Te-Yi, Chang Shu-Hao (2012) Using hybrid data mining and machine learning clustering analysis to predict the turnover rate for technology professionals. Expert Syst Appl 39:8844–8851

[7] Tan KC, Teoh EJ, Yua Q, Goh KC (2009) A hybrid evolutionary algorithm for attribute selection in data mining. Expert Systems Appl 36(4):8616–8630

[8] Liao TW, Triantaphyllou E (2007) Recent advances in data mining of enterprise data: algorithms and applications. World Scientific Publishing, Singapore, pp 111–145.

[9] Mabroukeh NR, Ezeife CI (2010) A taxono Hea Z, Xua X, Huangb JZ, Denga S (2004) Mining class outliers: concepts, algorithms and applications in CRM. Expert Syst Appl 27(4):681e97.

[10] Chen YL, Weng CH (2009) Mining fuzzy association rules from questionnaire data. Knowl Based Syst 22(1):46–56.

[11] Tan KC, Teoh EJ, Yua Q, Goh KC (2009) A hybrid evolutionary algorithm for attribute selection in data mining. Expert Syst Appl 36(4):8616–8630.

[12] http://www.roselladb .com

[13] S. Alamelu Mangai, B. Ravi Sankar, K. Alagarsamy , Taylor Series Prediction of Time Series Data with Error Propagated by Artificial Neural Network, International Journal of Computer Applications (0975 – 8887), Volume 89 – No.1, March 2014.

[14] Qiang Yang, Haining Henry Zhang and Hui Zhang, ―Taylor Series Prediction: A Cache Replacement Policy based on Second-order Trend Analysis, Hawaii, International Conference on System Sciences 2001; 5:5023, ISBN 0-7695-0981-9.