# An Improved Decision Tree Classification Using ID3 Algorithm Using Data Mining

**G.Priyadarshini,M.Sc.,M.Phil**
[1, 2] Dept of Computer Applications
[1, 2] KG COLLEGE OF ARTS AND SCIENCE, COIMBATORE-35

***Abstract***- *Data mining is the process of finding the previously unknown and potentially interesting patterns and relation in database. Decision tree learning algorithm has been successfully used in expert systems in finding the knowledge. The main work is to performed in these systems is using inductive methods to the given values of attributes of an unknown object to determine appropriate classification according to decision tree rules This paper suggests several procedures and methods for building decision tree, such as ID3, C4.5, and CART. Good choice for decision making tree methods . Decision tree learning method is also one of the methods that are used for classification or diagnosis. Decision tree learning method is used in Medical science for diagnosis purpose. This paper suggests that decision tree construction with ID3 algorithm for Diabetic patient database. For this database I have choose Iterative Dichotomizer algorithm. This algorithm based on the homogenous mixture Entropy, Information Gain for the best split.*

***Keywords***- Data mining, decision tree, ID3(Iterative Dichotomizer),CART(Classification and Regression tree ) algorithm, C4.5 algorithm.

## I. INTRODUCTION

Data mining is the process of discovering meaningful new patterns, correlation, and trends by sifting through large amount of data stored inrepositories using pattern recognition technologies as well as statistical and mathematical techniques. The main objective of Data mining is to discovered knowledge for the purpose of explaining current behavior, predicting future outcomes or providing support for business decision.

Data mining involves many different algorithms to accomplish different tasks. Data mining algorithms can be characterized as consisting of three parts:-

Model: The Task of the algorithm is to fit a model to the data.

Preference: - Some basic criteria must be used to fit one model over the another model.

Search: - All algorithms the required some of techniques to search the data.

Data mining is the step in the knowledge discovery in database process (KDD) .the structures that are the outcome of the data mining process must meet certain condition so that these can be considered as knowledge. These conditions are validity, understandability, utility, novelty, interestingness.

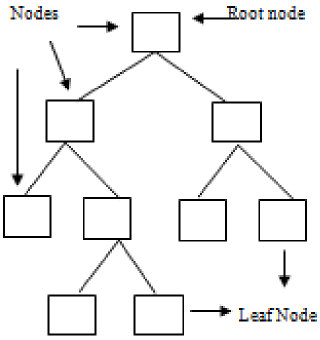
Researcher identify two fundamental goals of data mining : prediction and description there are several data mining techniques some of these are association, classification, sequential patterns and clustering.

Data mining is the process of discovering more knowledge, such as patterns, changes, associations, significant structures and irregular, from large amounts of data stored in databases or data warehouses or other information repositories [1]. It has been widely used in recent years due to the availability of huge amounts of data in electronic form, and there is a need for turning such data into useful information and knowledge for large applications. These applications are found infields such as Business management, Artificial Intelligence, Learning through machine, Analysis in market, Statistics and Database Systems and Decision Support.

## II.DECISION TREE

Decision tree learning method is one of the methods that are used for classification. As for many other machine learning methods, the learning in decision tree is done by using a data set of already classified instances to build a decision tree which will later used as classifier. The set of instances used to “train” the decision tree is called the training set.

Decision tree learning has main advantages. In that one is of the advantages is that it gives a graphical representation of the classifiers which makes it easier to understand. Decision tree is same as tree structure. The top most nodes in the tree is the root node. Each node in the tree specifies a test on some attribute and each branch descending from the node corresponds to one of the possible values of the attributes. Except for the terminal nodes that represent class. Moving down the tree branch corresponding to the values of the attribute in the given example. This processes repeated for sub tree rooted at the current node. There are several procedures and methods for building decision tree, such as *Iterative Dichotomize ,Classification, Regression tree* algorithm and C4.5 algorithm. My concept is *Iterative Dichotomize* Algorithm. This concept is based on based on the Entropy, mutual gain.



An Decision Tree

Advantages of Decision Tree Classifications

- Decision Trees able to generate understandable rules,
- They are able to handle both numerical and the categorical attributes,
- clear indication is provided in field of which fields are most important for prediction or classification

*A. CART Algorithm*

CART (Classification and Regression Tree) is one of the popular methods of building decision trees in the machine learning community; CART builds a binary decision tree by splitting the record at each separate node, according to a single attribute of a function. CART uses the gain index for determining the best split. At the every record of the training set has been assigned to some leaf of the full decision tree, At the end of the growing process. CART is nonparametric. Therefore this method does not require specification of any functional form. CART does not require variables to be selected in advance. CART algorithm will itself identify the most significant variables and eliminate non-significant ones. To test this property, one can include insignificant (random) variable and compare the new tree with tree, built on initial dataset. Both trees should be grown using the same parameters (splitting rule and N minparameter). We can see that the final tree 5.1, build on new dataset of three variables, is the identical to tree 3.2,built on two-dimensional data's

*B.C4.5 Algorithm*

In building a decision tree, we can make with training sets that have records with unknown attribute values by evaluating the gain, or the ratio, for an attribute by considering only those records where those attribute values are available. We can classify records that have unknown attribute values by estimating the probability of the various possible results. Unlike CART, which generates a binary decision tree, Variable branches per node are produced tree by C4.5. When a discrete variable is chosen as the splitting attribute in C4.5, there will be one branch for each value of the attribute. A decision tree model consists of a set of rules for dividing a large heterogeneous population into smaller, more groups are homogeneous with respect to a particular target variable. A decision tree may be painstakingly constructed by hand in the manner of Linnaeus and the generations of taxonomists that followed him, or it may be grown automatically by applying any one of several decision tree algorithms to a model set comprised of pre-classified data. The target variable is usually categorical and the decision tree model is used either to calculate the probability that a given record belongs to each of the categories, or to classify the record by assigning it to the most likely class. Decision trees can also be used to estimate the value of a continuous variable, although there are other techniques more suitable to that task

*C. ID3 (Iterative Dichotomizer) Algorithm*

Quinlan introduced the ID3 Algorithm, Iterative Dichotomizer 3, for constructing decision tree from the data. The most important features of ID3 algorithm is its capability to break down a complex decision tree into a collection of simpler decision tree. Dataset is used to generate a decision tree. ID3 is the precursor to the C4.5 algorithm, and is mainly used in the machine learning and natural language processing domains.

- Every attribute can provide most of one condition on a path given.
- The training data can be created from understandable prediction rule.
- Whole data set is searched to create a tree.
- One current hypothesis is maintained.
- No backtracking: this can't be changed, once an attribute is selected,.
- Attribute are selection by computing information gain on the full training set.

- A top down search through the given sets to test each attribute at every tree node by starting ID3 algorithm builds a decision tree by starting.
- ID3 does not guarantee an optimal solution, it can get stuck in local optimums.
- By selecting the best attribute to split the dataset on each iteration is used by a greedy approach. One changes that can be made on the algorithm can be to use backtracking during the search for the optimal decision tree.
- ID3 can over fit to the training data, to protect over fitting, smaller decision trees should be preferred over larger ones. This algorithm but it does not always produce the smallest possible tree, will produces small trees,.
- Using on continuous data ID3 is harder. If the values of any given attribute is continuous, then the attributes are many more places to split the data on this attribute, and searching for the best value to split by can be time consuming. The ID3 algorithm is a classification algorithm based on Information Entropy, that all examples are mapped its basic idea to different categories according to different values of the condition attribute set; its core is to determine the best classification attribute form condition attribute. The algorithm chooses information gain as attribute selection criteria; usually the attribute that has the highest information gain is selected as the splitting attribute of current node, To make information entropy that the divided subset seed smallest According to the different values of the attribute, branches can be established, and the process above is recursively called on International Journal of Data Mining & Knowledge Management Process (IJDKP) each branch to create other nodes and branches until all the samples in a branch belong to the same category. The splitting attributes, the concepts of Entropy and Information Gain are used to select.

*B. Main steps in ID3 Algorithm are*

For each attribute in the database, computes its Entropy.

The current node is the attributes (A) with highest information gain;

For every values of the attribute A builds a sub tree;

If A= value one then generate subtree1
If A= value two then generate subtree2

For each sub tree, repeat this process from the first step;
When there is no attributes in left the process stops.

*1) Entropy:*

Advanced networks in wireless using 4G technology Putting together a decision tree is all a matter of choosing which attribute to test at each node in the tree. We shall a measure a define called information gain which will be used to decide which attribute to test at each node. Information gain is itself evaluated using a measure called entropy, which the case of a bi nary decision problem is the first define and then define for the general case. Entropy measures the impurity of set of training objects. For a collection S, entropy is given as

Entropy (S) =

$$\sum_{i=1}^{c} - \text{pi log2 pi}$$

For a collection S having +ve and -ve example

Entropy (S) = -$P_{+}$i log$_2$ $P_{+}$i – p_i log2 p_ i

Where P+ is the positive of proportion examples
Where P- is the proportion of negative examples
Where S is a set, consisting of S data sample, pi is the portion of s belonging to the class I Notice that entropy is 0 when all members of S belong to the same class.

Entropy is 1 when the collection contains an equal number of positive and negative.

If the collection contains unequal number of positive and negative examples, the entropy is between zero and one.

*2) Information Gain:*

Each attributes for Information Gain is based on the computed entropy , and reduction in entropy is expected in states .The information Gain of an attribute A relative to a set of objects S is defined as

Gain(S, A) = Entropy(S) - $\sum_{v\in \text{Values(A)}}$ |S $_v$ | / S E (S $_v$)

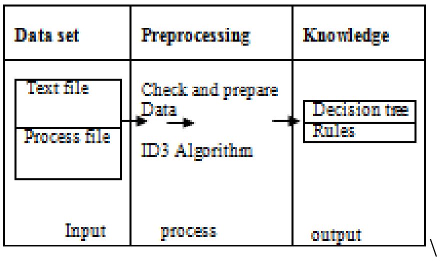
Where values (A) is the set of all possible values for attribute A, S $_v$ is the subset of S for which attribute A has value v.

The attribute having the highest information gain is to be preferred as root node. Information gain is precisely the measure used by ID3 to select the best attribute at each step in growing the decision tree.

## III. SYSTEM DESCRIPTION

For small application of data mining in medicine the DM_ ID3 system was build

TABLE I SYSTEM DESCRIPTION



| Data set | Preprocessing | Knowledge |
|---|---|---|
| Text file<br>Process file<br>Input | Check and prepare Data<br>ID3 Algorithm<br>process | Decision tree<br>Rules<br>output |

\

## IV.CONCLUSION

This system provides various algorithms for constructing decision tree. Such as ID3, C4.5, and CART. For Diabetes mellitus prediction system ID3 algorithm is better choice. Because of Entropy and information gain calculation for best splitting .So I suggest this algorithm for Diabetes Mellitus prediction system. Privacy preservation in data mining activities is of significant importance for many applications. However, the privacy preserving process sometimes reduces the utility of training datasets, which causes inaccurate data mining results. Privacy preservation approaches focus on different areas of a data mining process, and data mining methods also vary.

This focuses on privacy protection of the training samples applied for decision tree data mining. The privacy preserving approach discussed in this thesis would not function,if there is leakage is all training datasets, because reconstruction algorithm is the dataset the generic. Therefore, research is further required to eliminate this limitation. Future research should also explore means to reduce the storage requirement associated with the derived dataset complementation approach. This thesis relies on theoretical proofs with limited real tests, so testing with lab samples should be the next step to gain solid ground on real-life application.

## REFERENCES

[1] Almuallim H., An Efficient Algorithm for Optimal Pruning of Decision Trees.Artificial Intelligence 83(2): 347-362, 1996.

[2] Jaiwe Han, Michelin kamber, *Data mining: Concept and technique.*

[3] M.James.*classification algorithms*.Johnwiley& sons

[4] J.R Quinlan.*inducation of decision trees*, centre for Advanced computing sciences, New Wales institute Of Technology

[5] Tom M. Mitchell, (1997). *Machine Learning*, Singapore, McGraw-Hill.

[6] Paul E. Utgoff and Carla E. Brodley, (1990). 'An Incremental Method for Finding Multivariate Splits for Decision Trees', *Machine Learning:Proceedings of the Seventh International Conference,* (pp.58). Palo Alto, CA: Morgan KaufmannHttp://

[7] www.google.com