Decision Tree Analysis on J48 Algorithm for Data Mining

Dr.G.Manikanta, MD.Adul Gafar, S.Govinda Raju

Professor[1],Associate Professor[2,3],

Dept. of CSE,

mail-id:gafar501@gmail.com, mail-id:skraj999@gmail.com

Anurag Engineering College,Anatagiri(V&M),Suryapet(Dt),Telangana-508206

Abstract— The Data Mining is a technique to drill database for giving meaning to the approachable data. It involves systematic analysis of large data sets. The classification is used to manage data, sometimes tree modelling of data helps to make predictions about new data. This research is focussed on J48 algorithm which is used to create Univariate Decision Trees. The research study also discuss about the idea of multivariate decision tree with process of classify instance by using more than one attribute at each internal node. The core concept behind the topic is to getdepth knowledge with new areas of research by explore more about data, information, knowledge, data mining techniques, and tools. All the results with experiment on Weka are finally examined.

Keywords— Data Mining; Classification Techniques; J48; Decision Trees; Univariate algorithm; Multivariate algorithm; Pruning

INTRODUCTION

Weka is open source software for data mining under the GNU General public license. This system is developed at the University of Waikato in New Zealand. "Weka" stands for the Waikato Environment for knowledge analysis. Weka is freely available at http://www.cs.waikato.ac.nz/ml/weka. The system is written using object oriented language java. Weka provides implementation of state-of-the-art data mining and machine learning algorithm. User can perform association, filtering, classification, clustering, visualization, regression etc. by using weka tool. Each and every organization is accession vast and amplifying amounts of data in different formats and different databases at different platforms. This data provides any meaningful information that can be used to know anything about any object. Information is nothing just data with some meaning or processed data. Information is than converted to knowledge to use with KDD.

Data Mining is a non trivial extraction of implicit, previously unknown, and imaginable useful information from data. Data mining finds important information hidden in large volumes of data. Data mining is the reasoning of data. It is the use of software techniques for finding patterns and consistency in sets of data [12]. Data Mining is an interdisciplinary field involving: Databases, Statistics, and Machine Learning. There are various techniques available for data mining as given below:-

A. Association Rule Learning: - This is also called market basket analysis or dependency modelling. It is used to discover relationship and association rules among variables.
B. Clustering: - This technique creates and discovers group of similar data items. This is also called unsupervised classification.
C. Classification: - This can classify data according to their classes i.e. put data in single group that belongs to a common class. This is also called supervised classification.
D. Regression: - It tries to find a function that model the data with least errors.
E. Summarization: - It provides easy to understand and analysis facility through visualization, reports etc [11].

It is possible to mine data with computer that automates this process. Various data mining tools are available in market some are:-

- Environment for DeveLoping KDD-Applications Supported by Index-Structures (ELKI)
- jHepWork
- Konstanz Information Miner (KNIME)
- Orange (software)
- RapidMiner
- Scriptella ETL — ETL (Extract-Transform-Load) and script execution tool
- Weka [11].

DECISION TREE

A decision tree is a decision support system that uses a tree-like graph decisions and their possible afterchance event results, resource costs, and utility. A Decision Tree, or a classification tree, is used to learn a classification function which concludes the value of a dependent attribute (variable) given the values of the independent (input attributes (variables). This verifies a problem known as supervised classification because the dependent attribute and the counting of classes (values) are given [4].

Decision trees are the most powerful approaches in knowledge discovery and data mining. It includes the technology of research large and complex bulk of data in order to discover useful patterns. This idea is very important because it enables modelling and knowledge extraction from the bulk of data available. All theoreticians and specialist are continually searching for techniques to make the process more efficient, cost-effective and accurate. Decision trees are highly effective tools in many areas such as data and text mining, information extraction, machine learning, and pattern recognition.

Decision tree offers many benefits to data mining, some are as follows:-
• It is easy to understand by the end user.
• It can handle a variety of input data: Nominal, Numeric and Textual
• Able to process erroneous datasets or missing values
• High performance with small number of efforts
•This can be implemented data mining packages over a variety of platforms [10].

A tree includes: - A root node, leaf nodes that represent any classes, internal nodes that represent test conditions (applied on attributes) as shown in figure 1.
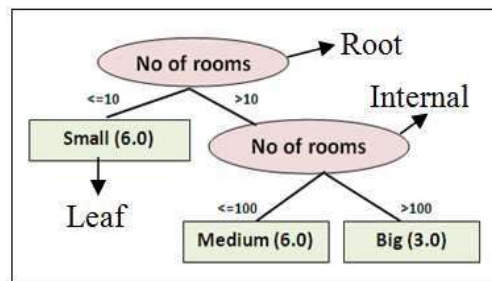

Figure 1 Decision tree

A. Tree Size

Basically, decision makers prefer a decision tree because it is not complex as well as easy to understand. Tree complexity has its effect on its accuracy. Usually the tree complexity can be measured by a metrics that contains: the total number of nodes, total number of leaves, depth of tree and number of attributes used in tree construction. Tree size should be relatively small that can be controlled by using a technique called pruning [13].

B. Rule Induction in trees

Decision tree induction is closely related with the rule induction. Each path that starts from the root of a decision tree and ends at one of its leave represents a rule. These rules can be generated very easily.

I.   DECISION TREE APPROACHES

There are two approaches for decision tree as:-
1) Univariate decision tree
In this technique, splitting is performed by using one attribute at internal nodes. For ex. X<2, y>=10 etc. There are many algorithms for creating such tree as ID3, c4.5 (j48 in weka) etc. This paper will illustrate that how to implement j48 algorithm and analysis its result.

This algorithm is an extension of ID3 algorithm and possibly creates a small tree. It uses a divide and conquers approach to growing decision trees that was leaded by Hunt and his co-workers (Hunt, Marin and Stone, 1966) [5].

A. Construction

Some basic steps are given below to construct tree:-
• First, check whether all cases belongs to same class, then the tree is a leaf and is labeled with that class.
• For each attribute, calculate the information and information gain.
• Find the best splitting attribute (depending upon current selection criterion) [4].

B. Counting information gain

"Entropy" is used in this process. Entropy is a measure of disorder of data. Entropy is measured in bits, nats or bans. This is also called measurement of uncertainty in any random variable. Just suppose that there is a fair coin, if single toss is performed on that coin than its entropy will be one bit. A series of two fair coins tosses will have entropy of two bits. Now if coin is not fair than there is uncertainty and this provides lower entropy rate.

Entropy for any P can be calculated as:-

$$Entropy(p) = -\sum_{j=1}^{n} \frac{|pj|}{|p|} \log \frac{|pj|}{|p|}$$

The conditional entropy is:-

$$Entropy(j|p) = \frac{|pj|}{|p|} \log \frac{|pj|}{|p|}$$

If base is 2 for logarithm than entropy measurement unit will be in bits, if base is 10 than unit is dits. Information Gain is used for measuring association between inputs and outputs. It is a state to state change in information entropy. Finally information gain can be calculated as:-

$$Gain(p,j) = Entropy(p - Entropy(j|p))$$

To get a small and efficient tree, splitting should be based on highest gain. Just suppose that there are 9 male (m) and 5 female (f) in a class instance. This instance in divided further into two different groups or instances on the bases of their calculated entropy and information gain. So 3m and 4f as left instance and 6m and 1f as right instance. Entropy and information gain can be measured just by putting values in formula as given below:-

Entropy_bef = -5/14*log (5/14) - 9/14*log (9/14)
Entropy_left = -3/7*log (3/7) -4/7*log (4/7)
Entropy_right = -6/7*log (6/7) -1/7*log (1/7)
Entropy_aft = 7/14*Entropy_left + 7/14*Entropy_right
Information_Gain = Entropy_bef - Entropy_aft

Figure 2 Evaluating entropy and information gain

C. Pruning

Pruning is very important technique to be used in tree creation because of outliers. It also addresses overfitting. Datasets may contain little subsets of instances that are not well defined. To classify them correctly, pruning can be used. There are two types of pruning:-
1.   Post pruning (performed after creation of tree)
2.   Online pruning (performed during creation of tree) [8].
Pruning algorithms
Separate and Conquer rule learning algorithm is basis to prune any tree. There are various rule learning schemes. All separate-and-conquer rule learning schemes are modifications of the same simple algorithm that starts with an empty set of rules and the full set of training instances. This simple Separate and Conquer algorithm is given below:-
Procedure SeparateAndConquer (D: instances)
•   R := empty set of rules
•   while not D is empty
•   r := best single rule for D
•   R := add r to R
•   remove those instances from D that are covered by r
•   return R
Some Separate and Conquer rule learning schemes are:-
•   Reduced-error pruning for rules
•   Incremental reduced-error pruning
•   Incremental tree-based pruning [3].

2) Multivariate decision tree
Multivariate DT is able to generalize well when dealing with attribute correlation and its result is also easy to humans. When working with Univariate DT's, they test single attribute more than once that may result in inefficient tree in some situations. Multivariate DT performs different tests with the data by using more than one attributes in test leaves. Test condition in these trees may be as x + y>10. This technique is a non linear combination of attributes at every test nodes [4]. Each test node will follow the form:-

$$\sum_{i=1}^{n+1} w_i y_i > 0$$

Where wi are real-valued coefficients, yi are attributes and n is total no of attributes in an instance. Figure 3 and 4 shows the difference between univariate and multivariate space partitioning and also represents that how multivariable test conditions are placed on internal nodes [4].
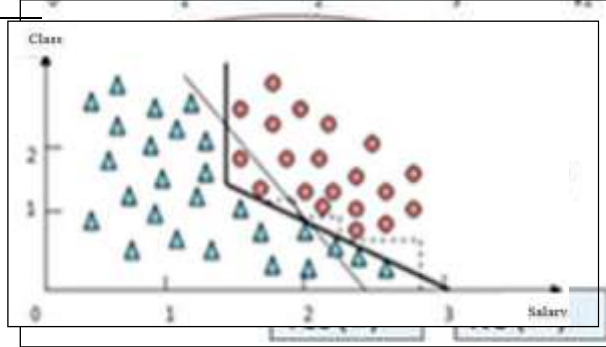
Figure 3 multivariate space partitioning
Figure 4 multivariate decision conditions

➢   Construction

Firstly we should have a set of training instances. This uses a top-down decision tree algorithm and merit selection criteria to choose the best splitting attribute to create a branch. Thus we have two partitions. Algorithm will apply same top-down analysis to make further more partitions. One of the stopping criteria is when all the attribute's values belong to a single class. There is only difference in splitting criteria if do comparison between multi and univariate tree construction. Multivariate DT uses LM (linear machine) [4].

➢   LM

A linear machine is a set of R linear discriminant functions that are used collectively to assign an instance to one of the R classes. Here p is an instance description that consists 1 and the n features that describe the instance. Discriminat function gi(p) for multivariate has the form as [4]:-

$$w_i^t p$$

Where w_i is a vector of n + 1 coefficients. The LM sates that instance p will belong to class i iff

$$(\forall j, i \neq j) gi(p) > gj(p)$$

w_i is the weight vector that have default value as w_i= 1,...N. Here, the absolute error correction rule, and the thermal perceptron are given below [4], [9].

➢   Absolute Error Correction rule

It is an approach to update the weight of the discriminat functions, which adjusts w_i, where i is the class to which the instance belongs, and wj, where j is the class to which the LM incorrectly assigns the instance. The correctly assignment is done by

$$w_i \leftarrow w_i + c_p$$

and

$$wj \leftarrow wj - c_p$$

Where

$$c = \left[\frac{(wj - wi)^t p}{2p^t p} \right]$$

It is a small real integer number such that its updates will classify the instance correctly [4], [9].

➢   Thermal Perceptron

This method is used for non linearly separable instances. It also adjusts wi and wj, and makes use of some constants as:-

$$c = \frac{B}{B+k}$$

and

$$k = \left[\frac{(wj - wi)^t p}{2p^t p} \right]$$

This process follows the algorithm as described below:
1.   assign value to B as :- B = 2;
2.   If LM is correct for all instances Or B < 0.001, RETURN
3.   Else, for each misclassified instance
a. Compute constant c & update w[i] and w[j]
b. Assign B ← aB - b with a = 0.99 and b = 0.0005
4.   Back to step 2 [4], [9].
 The main purpose behind this algorithm is to correct the weights-vector until all instances become correct, or in the worst case, a certain number of iterations are reached [4].

II. EXPERIMENT AND RESULTS

➢   Result of Univariate decision tree approach
Steps to create tree in weka
1 Create datasets in MS Excel, MS Access or any other & save in .CSV format.
2 Start the weka Explorer.
3 Open .CSV file & save in .ARFF format.
4 Click on classify tab & select J48 from choose button.
5 Select any appropriate test option.
6 Click on Start button & result will be displayed.

Relation: MYTABLE

No.	CATEGORY Nominal	RESULT Numeric	SALARY Numeric	CLASS Nominal
1	STUDENT	66.0		PASS
2	STUDENT	60.0		PASS
3	EMPLOYEE		35000.0	TEAC...
4	EMPLOYEE		35000.0	TEAC...
5	STUDENT	50.0		FAIL
6	STUDENT	44.0		FAIL
7	STUDENT	70.0		PASS
8	STUDENT	71.0		PASS
9	EMPLOYEE		22000.0	CLERK
10	EMPLOYEE		22000.0	CLERK
11	EMPLOYEE		50000.0	HOD
12	STUDENT	77.0		PASS
13	EMPLOYEE		50000.0	HOD
14	STUDENT	87.0		PASS
15	STUDENT	40.0		FAIL

Figure 5 Data for creating tree

Result is displayed in Classifier output window. Result can be viewed in separate window. To do this, right click on result list. To view tree in graphical form, click on "visualize tree" option in pop-up menu. A decision tree in shown in figure 6 that includes category, result, salary as root and internal nodes and classes for leaf nodes are pass, fail, HOD, teacher, & clerk.
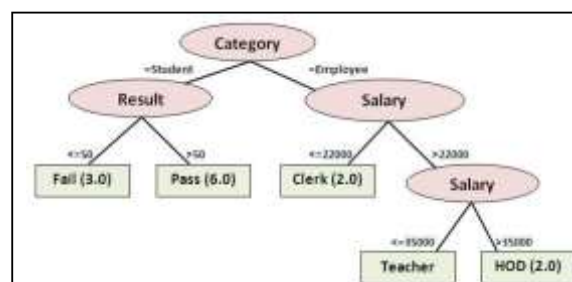
Figure 6 Decision tree visualization

Disadvantages of J48 algorithm

The run-time complexity of the algorithm matches to the tree depth, which cannot be greater than the number of attributes. Tree depth is linked to tree size, and thereby to the number of examples. So, the size of C4.5 trees increases linearly with the number of examples. C4.5 rules slow for large and noisy datasets Space complexity is very large as we have to store the values repeatedly in arrays [6].

➤   Result of Multivariate decision tree approach

Weka provides an algorithm, called M5P that is used to create classification and regression tree with a multivariate linear regression model where p stands for prime. This algorithm provides linear model as classes with some percent of approximated errors. Applying this algorithm on dataset is just same as applying J48 algorithm [1]. Result produced by M5P is displayed in figure 7 as shown below:-
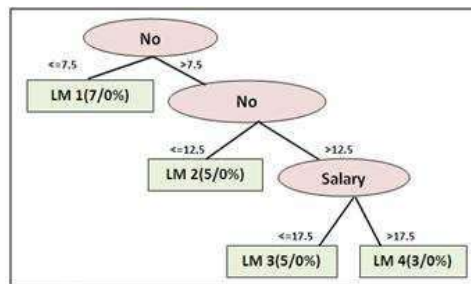

Figure 7 Result of M5P

CONCLUSION

This paper presents discussion about Decision Trees with the Univariate and the Multivariate approaches. Weka is used as data mining tool that provides various algorithms to be applied on data sets. The J48 algorithm is used to implement Univariate Decision Tree approach, while its results are discussed. The Multivariate approach is introduced as the Linear Machine approach that makes the use of the Absolute Error Correction and also the Thermal Perceptron Rules. Decision Tree is a popular technique for supervised classification, especially when the results are interpreted by human. Multivariate Decision Tree uses the concept of attributes correlation and provides the best way to perform conditional tests as compare to Univariate approach. The research study concludes that Multivariate approach is far better than Univariate approach while it allow us dealing with large amount of data.

REFERENCES

[1].   Dolado, J. J., D. . Rodríguez, and J. Riquelme. "A Two Stage Zone Regression Method for Global Characterization of a Project Database." (2007): 13. Web. 5 Apr. 2013.

[2].   Berzal, Fernando, Juan-Carlos Cubero, and Nicolas   Mar   Building multi-way decision trees with numerical attributes." 31. Web. 5 Apr. 2013.

[3].   Frank, Eibe. "Pruning Decision Trees and Lists." (2000): 218. Web. 5 Apr. 2013.

[4].   Korting, Thales S. "C4.5 algorithm and Multivariate Decision Trees." 5. Web. 2 Feb. 2013.[5].   Quinlan, J. R. "Improved Use of Continuous Attributes in C4.5." 14. Web. 11 Jan. 2013.

[6].   JUNEJA, DEEPTI, et al. "A novel approach to construct decision tree using quick C4.5 algorithm." Oriental Journal of Computer Science & Technology Vol. 3(2), 305-310 (2010) (2010): 6. Web. 18 Feb. 2013.

[7].   Ittner, Andreas, et al. "Non-Linear Decision Tree - NDT." In: Proceeding of 13th international conference on machine learning (ICML"96) 6. Web. 16 Mar. 2013.

[8].   Moertini, Veronica S. "TOWARDS THE USE OF C4.5 ALGORITHM FOR CLASSIFYING BANKING DATASET." Vol. 8 No. 2, October 2003 (2003): 12. Web. 24 Jan. 2013.

[9].   Utgoff, Paul E. "Linear Machine Decision Tree." (1991): 15. Web. 6 Feb. 2013.

[10].   Rokach, Lior, and Oded Maimon. "DECISION TREES." 28. Web. 1 Feb. 2013.

[11].   "Data Mining" from Wikipedia the free Encyclopedia. Web. <http://en.wikipedia.org/wiki/Data_mining>.

[12].   Term "INTRODUCTION OF DATA MINING","Data Mining: What is Data Mining ", source from http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm.

[13].   Rokach, Lior. "Data Mining with Decision Trees: Theory and Applications." 69 (2008): Web. 3 Feb. 2013.

[14].   Ga·sperin, Matej. "Case Study on the use of Data Minig Techniques in Food Science using Honey Samples." (February 2007): 18. Web. 8 May 2013.

[15].   Ozer, Patrick. "Data Mining Algorithms for Classification." (January 2008): 27. Web. 5 May 2013.

[16].   Gholap, Jay. "PERFORMANCE TUNING OF J48 ALGORITHM FOR PREDICTION OF SOIL FERTILITY." 5. Web. 2 May 2013.