Review of Decision tree data mining algorithms : ID 3 and C4.5

Conference Paper · July 2015

CITATIONS
4

READS
1,927

1 author:

Rajeev Kumar Bedi
Beant College of Engineering and Technology, Gurdaspur
47 PUBLICATIONS **100** CITATIONS

SEE PROFILE

Proceedings of International Conference on Information Technology and Computer Science
July 11-12, 2015, ISBN:9788193137307

REVIEW OF DECISION TREE DATA MINING ALGORITHMS: ID3 AND C4.5

DavinderKaur
Mtech Student, CSEDept ,BCET
Gurdaspur, India

Rajeev Bedi
Assistant Professor, BCET
Gurdaspur, India

Dr. Sunil Kumar Gupta
Associate Professor, BCET,
Gurdaspur, India

ABSTRACT

Data mining is a process of identification of useful information from large amount of random data. It is used to discover meaningful pattern and rules from data. Classification, clustering, association rules are data mining techniques. Classification is a process of assigning entities to analready defined class by examining the features. Decision tree is a classification technique in which a model is created that anticipates the value of target variable depends on input values. ID3 and C4.5 are commonly used decision tree algorithms. These algorithms are based on Hunt's algorithm. Goal of this study is to provide review of these decision tree algorithms. At first we present concept of Data Mining, Classification and Decision Tree. Then we present ID3 and C4.5 algorithms and we will make comparison of these two algorithms.

KEYWORDS- *Data mining, classification, decision tree*

I. INTRODUCTION

Data mining is a process of extraction useful information from large amount of data. It is used to discover meaningful pattern and rules from data. Data mining is a part of wider process called knowledge discovery [4]. The steps of knowledge discovery are

- Selection
- Processing
- Transformation
- Data mining
- Interpretation/Evaluation

Data mining uses two types of approaches i.e supervised learning or unsupervised learning.

A. Classification

Classification is the process of assigning newly presented entities to already defined class by examining the features of entities. Classification is to make decision from unseen cases by building examples of past decisions [2]. There are two steps in classification process.

- In first step, model is built from training data in which value of class label is known. Classification algorithms are used to create model from training data sets.
- In second step, accuracy of model is checked by test data and if correctness of model is satisfactory then the model is used to classify data with unknown class label.

Among classification algorithm, decision tree algorithms is usually used because it is easy to follow and economical to implement.

B. Decision Trees

Decision tree is a classification technique. It is a tree like structure where internal node contains splits and splitting attributes. It represents test on an attribute. Arcs between internal node and its child contain consequences of test. Each leaf node is associated with a class label. Decision tree is constructed from training set. Then this decision tree is used to classify the tuples with unknown class label [2].

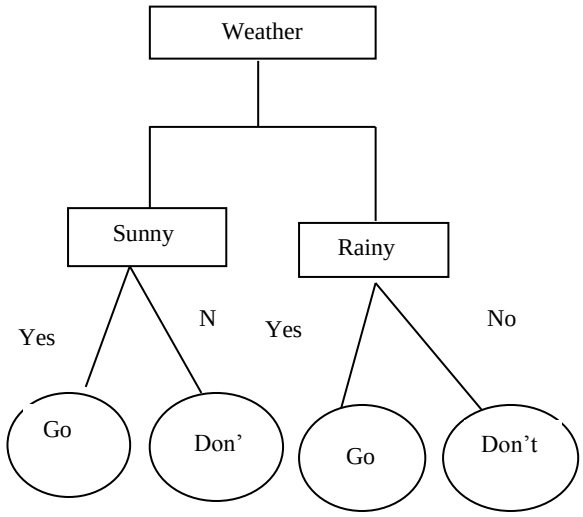


Fig. 1 Decision Tree showing whether to go for trip or not depending on weather

I. DECISION TREE ALGORITHMS

Decision tree learning methods are most commonly used in data mining. The goal is create a model to predict value of target variable based on input values. Training dataset is used to create tree and test dataset is used to test accuracy of the decision tree. Each leaf node representsthe target attribute's value depend on input variables represented by path by path from root to leaf node. First, an attribute that splits data efficiently is selected as root node in order to create small tree. The attribute with higher information is selected as splitting attribute[4].

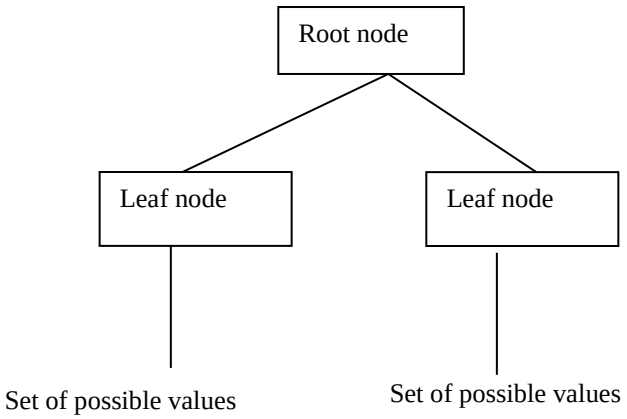


Fig. 2 Decision tree induction

Decision tree algorithm involves three steps:

1. For a given dataset S, select an attribute as target class to split tuples in partitions.
2. Determine a splitting criterion to generate a partition in which all tuples belong to a single class. Choose best split to create a node.
3. Iteratively repeat above steps until complete tree is grown or any stopping criterion is fulfilled.

A. *ID3 (Iterative Dichotomiser 3)*

ID3 algorithm is presented by J.R. Quinlan, 1986.ID3 uses Information gain as splitting criterion. Topmost decision node is the best predictor, it is called root node. The attribute with highest Information Gain is selected as split attribute. Information gain is used to create tree from training instances. This tree is used toclassify test data. When information gain approaches to zero or all instances belong to single target then growing of tree stops. [1].

It grows tree classifiers in three steps:

1. Selection of target attribute and calculation of entropy of attributes.
2. Select attribute with highest information gain measure
3. Create node containing that attribute. Iteratively apply these steps to new tree branches and stop growing tree after checking of stop criterion.

The ID3 decision makes use of two concepts when creating a tree from top-down [1]:

1. Entropy
2. Information Gain (as referred to as just gain) Using these two concepts, the nodes to be created and the attributes to split on can be determined.

Entropy

Entropy is degree of randomness of data. It is used to calculate homogeneity of data attribute. If entropy is zero then sample is totally homogeneous and ifis one then sample is completely uncertain.

Information Gain

Information gain is decrease in entropy. Attribute with highest information gain isselected as best splitting criterion attribute

$$ET(X, S) = \sum_{j=1}^{k} \frac{|Sj|}{|S|}.ET\ (Sj)$$

$$IG(X, S) = E(S) - E(X, S)$$

B. C4.5

C4.5 algorithm is enhancement to ID3.C4.5 can handle continuous input attribute..It follows three steps during tree growth [3]:

1. Splitting of categorical attribute is same to ID3 algorithm. Continuous attributes always generate binary splits.

2. Attribute with highest gain ratio is selected.

3. Iteratively apply these steps to new tree branches and stop growing tree after checking of stop criterion. Information gain bias the attribute with more number of values. C4.5 used a new selection criterion which is Gain ratio which is less biased.

The Gain ratio measure is a selection criterion which is used less biased towards selecting attributes with more number of values [3].

$$GR(X, S)= \frac{IG(X,S)}{SI(X,S)}$$

$$\mathbf{SI}(X, S)=-\sum_{j=1}^{k} \frac{|Sj|}{|S|} \log \frac{|Sj|}{|S|}$$

Advantages: C4.5 made improvements to ID3 [10]:

1. It can handle both discrete and numerical attributes.
2. It can handle missing value attribute.
3. It can avoid over fitting of decision tree by providing the facility of pre and post pruning.

II. IMPLEMENTATION OF ID3 AND C4.5

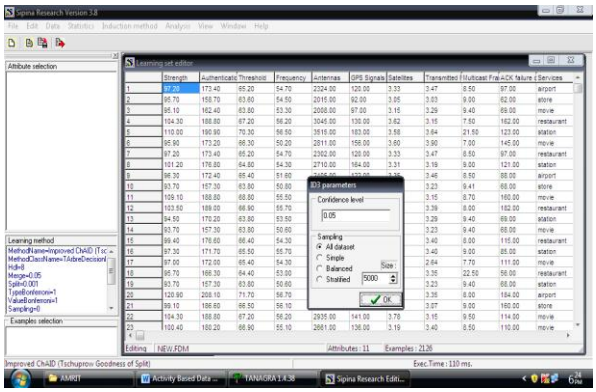


Fig. 3Import Data Set

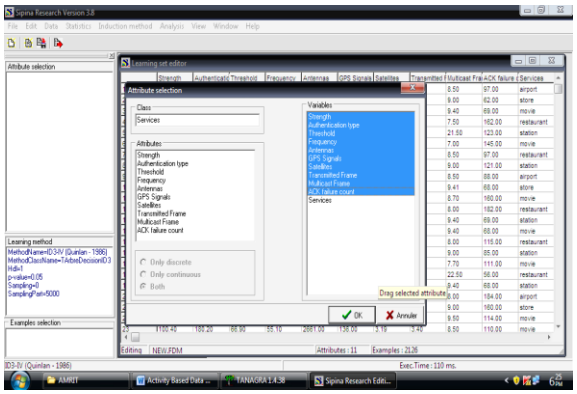
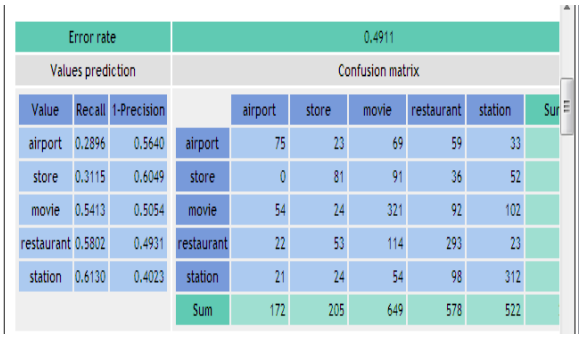


Fig. 4Define Class Variables

Fig. 5 Generating Tree based on class attribute services



Error rate			0,4911					
Values prediction			Confusion matrix					
Value	Recall	1-Precision		airport	store	movie	restaurant	station
airport	0.2896	0.5640	airport	75	23	69	59	33
store	0.3115	0.6049	store	0	81	91	36	52
movie	0.5413	0.5054	movie	54	24	321	92	102
restaurant	0.5802	0.4931	restaurant	22	53	114	293	23
station	0.6130	0.4023	station	21	24	54	98	312
			Sum	172	205	649	578	522

Fig. 6Calculate Error Rate

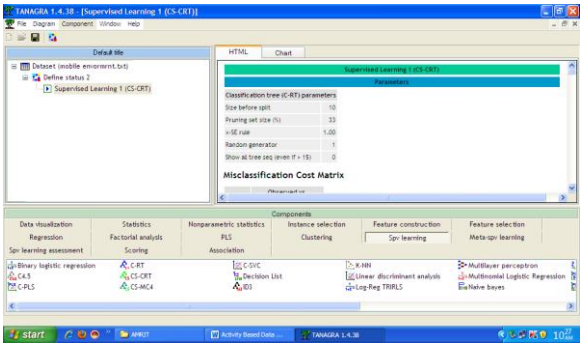


Fig. 7Parameters of C4.5 Algorithm

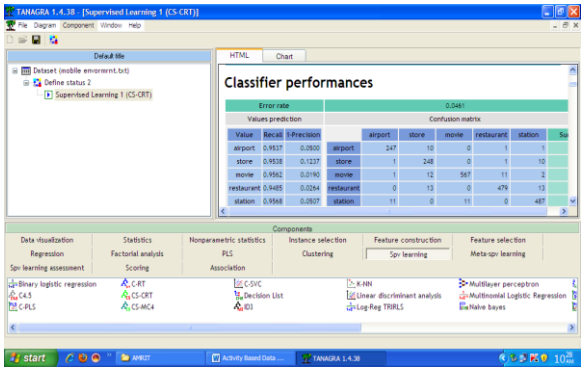


Fig. 8 Calculate error rate

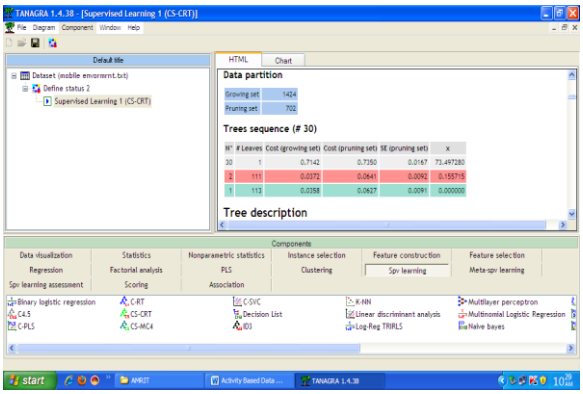


Fig. 9 Calculate no of nodes

TABLE I

RESULT:

	ID3	C4.5
Error rate	0.40	0.04
nodes	61	217
leaves	31	109
Execution time	94ms	125ms

IV. CONCLUSION

In this Research paper, we presented classification technique decision tree. We presented decision tree algorithm ID3 and C4.5.We focused on key elements of construction of decision tree. We did comparison of ID3 AND C4.5 algorithms. It is concluded that C4.5 is more accurate and consume less execution time to mine data with minimum error rate 0.04. C4.5 is a best algorithm for mining a data set.

REFERENCES

[1] Fong, P.K. and Weber-Jhanke, J.H (2012), "Privacy Preserving Decision Tree Learning using Unrealized Data Sets",*IEEE Transactions on knowledge and Data Engineering*, Vol.24,No.2, February 2012, pp. 353-364

[2] Kabra, R.R. and Bichkar, R.S. (2011),"Performance Prediction of Engineering Students using Decision Tree", International Journal of Computer Applications, Vol.36, No.11, December 2011, pp. 8-12.

[3] Karaolis, M.A. &Moutiris, J.A (2010), "*Assessment of the Risk Factors of Coronary Heart Events Based on Data Mining with Decision Trees",* IEEE Transactions on Information Technology in Biomedicine, Vol.14, No.3, May 2010, pp. 559-566.

[4] Kesavraj, G. and Sukumaran, S. (2013), "A Study on Classification Technique in Data Mining", 4th ICCNT-2013.

[5] Sautikar, A.V., Bhujada, V., Bhagat, P.&Khaparde, A.(2014)," A Review paper on Various Data Mining Techniques", International Journal of Advanced Research in Computer Science and Software Engineering, Vol.4,Issue 4, April 2014, pp. 98-101.

[6] Li, L. & Zhang, X. (2010), "Study of Data Mining Algorithm based on Decision Tree", 2010 International Conference on Computer Design and Applications (ICCDA 2010), Vol.1, pp. 155-158.

[7] Yi-Yang, G. and Man-ping, R. (2009), "Data Mining and Analysis of Our Agriculture based on the Decision Tree", ISECS International Colloquium on Computing, Communication, Control and management, 2009, pp. 134-138.

[8] Zhang, X.F. and Fan, L.(2013)," *A Decision Tree Approach for Traffic accident Analysis of Saskatchewan Highways*", 26th IEEE Canadian Conference of Electrical and Computer Engineering(CCECE) 2013.

[9] Zhang, T., Fulk, G.D. & Tang, W.(2013),"Using Decision Tree to Measure Activities in People with stroke", *35th Annual International Conference of the IEEE EMBS, July 13, pp.6337-6340.*

[10] Suknovic, .M, Delibasic, B., Jovanovic, M., Vukecevic, M., Obradovic, Z.(2011),"Reusable components in decision tree induction algorithm",Comp Stat Februaury 2011.