Construction of a Decision Tree with the Use of Association Rules: A Review Paper

Sneha D. Dhole
Department of Technology,
Shivaji University, Kolhapur, Maharashtra, India

Rashmi J. Deshmukh
Department of Technology,
Shivaji University, Kolhapur, Maharashtra, India

Abstract- Now a days as we all know that Internet has grown widely all over the world and there exist 80% of population which uses Internet in their day- to- day life to make their work done easily. People share most of their data on Internet so there exist huge amount of entity related data, sometimes this data may be confidential. If we have no particular way to store data on internet this makes very difficult to access data from internet and it will take more time for retrieval. For example in our academics if we have all the subjects as a one book as a syllabus for whole course then it will be very difficult for us to learn and understand. But if we get each subject as separate book then it will be easy for study. Similarly data on internet needs to be classified in such a way that it will reduce the data retrieval time. Doing classification manually is more complex and takes more time. There exist many methodologies for classification of data automatically such as decision tree algorithm, rule induction model, Naive Bayes Classifier and so many. In this review we are going to use decision tree algorithm with association rules.

Keywords – Association Rules, Decision Tree Algorithm, Rule Induction Model and Naive Bayes Classifier

I. INTRODUCTION

We all live in a B2C (Business to Customer) world. Almost everyone produces large amount of data over the internet. And this is increasing rapidly as Internet is growing vastly.Classification of data is the most important thing to make it more efficient. Manual classification is complex and time consuming. There exist many methodologies for classification of data automatically such as decision tree algorithm, rule induction model, Naive Bayes Classifier and so many**.**
In this review paper we are using decision tree algorithm with the use of association rules.Firstly we use the association rule mining to get rules and then we use it to build decision tree. Finally, we apply the decision tree to classification of information.

1.1 Association Rule Mining

Association rules determine relation between variables of large database. Most commonly association rules are used to analyze and predict customer behaviour. Consider an example "if customer buys bread he is 80% likely to buy butter also". These activities are used in marketing such as product promotion and product pricing. Association rule is based on two popular measurements *Support* and *Confidence.*Relating to considered example *Support* is a probability that contains bothBread and Butter. Whereas *Confidence* denotes the probability that a transaction containing Bread also contains Butter. For example In a Super Market there are total 100 transactions Out of which 20 contains *Bread* So, 20/100*100=20% which is nothing but *Support*. And out of 20 transactions 9 transactions contain *Butter* So, 9/20*100=45% which is nothing but *Confidence*.

There are various algorithms are available for generation of association rules such as Apriori algorithm, Elcat algorithm and F.P Growth algorithm. In this review we are going to use Apriori algorithm.

1.2 Decision Tree

A decision tree also known as a classifier in the form of tree. Decision tree contains two types of nodes one is decision node and other is leaf node. A decision node defines a choice and tests. Leaf node indicates final classification. Consider an example of bank to approve a loan or not. Bank does so many tests before approving the loan. There are so many tests and choices and done. Below shows a decision tree for the same problem.

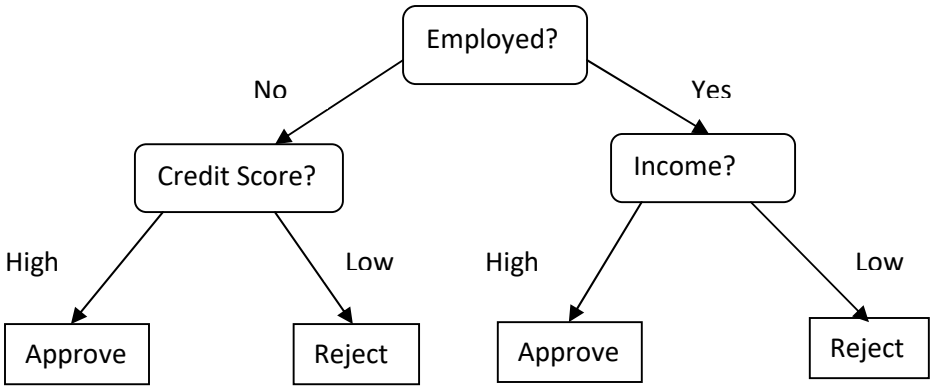

Figure 1. An example of decision tree

The first test made here is “Is the applicant is employed?” if no then next tests made is “What is his/her credit score?” if its high then approve the loan else reject the application. On the other side if applicant is employed then next test is to check for his income if income is high then approve the loan else reject the application.

II. REVIEW WORK

R. Agrawal, T. Imielinski, and A. Swami [3]They have presented an efficient algorithm that generates all significant association rules between items in the database. The algorithm incorporates buffer management and novel estimation and pruning techniques. Also present results of applying this algorithm to sales data obtained from a large retailing company, which shows the effectiveness of the algorithm.

D. M. Blei [5] Latent Dirichlet allocation (LDA), a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modelled as a finite mixture over an underlying set of topics. Each topic is, in turn, modelled as an infinite mixture over an underlying set of topic probabilities. In the context of text modelling, the topic probabilities provide an explicit representation of a document. They have presented efficient approximate inference techniques based on vibration methods and an EM algorithm for empirical Bayes parameter estimation

J. Han, J. Pei, and Y. Yin [7] Here they have proposed a novel frequent pattern tree structure, which is an extended prefix tree structure for storing compressed, crucial information about frequent patterns and develop an efficient FP tree based mining method, FP growth, for mining the complete set of frequent patterns by pattern fragment growth.

FabrizioSebastiani [1] A general inductive process automatically builds a classifier by learning, from a set of pre-classified documents, the characteristics of the categories. The advantages of this approach over the knowledge

engineering approach (consisting in the manual definition of a classifier by domain experts) are a very good effectiveness, considerable savings in terms of expert labor power, and straightforward portability to different domains. This survey discusses the main approaches to text categorization that fall within the machine learning paradigm.

III.PROPOSED ALGORITHM

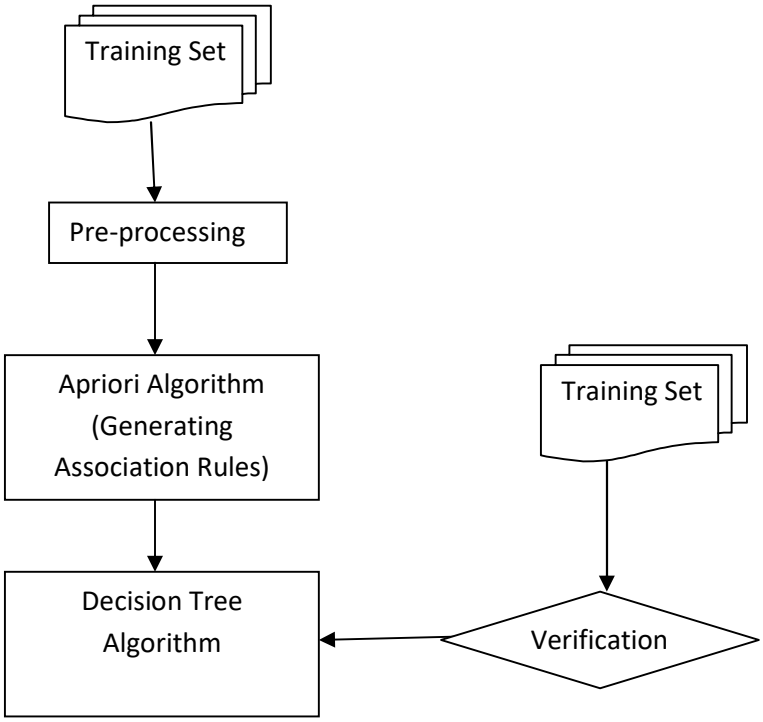

Figure 2. An idea of proposed model

3.1 Generating Association Rules:

We can't directly give raw data to association rule mining. We have to remove all the unrelated data such as "a","an","the" and change form of some phrases from document. All the category labels are needed to be predefined in the form of$C = c_1, c_2, \dots, c_3$

Algorithm ARM (T)

Input - A set of documents with the form $D(t_1, t_2, \dots, t_n, c_i)$A minimum support threshold and minimum confidence threshold;

Output - A set of association rules with the form $R(t_1, t_2, \dots, t_n, c_i)$

Method:

(1) $C_1 \leftarrow init - pass($
(2) Find the one frequent set that satisfies min(Supp) and min(Conf);
(3) $for(k = 2; FK - 1 \neq \emptyset; k++)$
(4) Find the k frequent set that satisfy the min(Supp) and min(Conf);
(5) $endf$

(6) $returnR \leftarrow D_k$

In the algorithm, basically associates each item in I with every category label, Line 2 determines whether the candidate 1-ruleitems are frequent. In Line 3, we generate k-condition , then we generate all the frequent rule items by making multiple passes over data, then at last we determine which is actually frequent(4) and generate the final rule set R(6).

3.2 Construction of Decision Tree:

Once we have association rules generated further we need to construct a decision tree. With the rule $\{t_i\} \rightarrow \{c_j\}$, if we get t_i in one document then we can make t_i a decision nodeand make it left child labelled with c_i, and for every t_i, if we have the rule ,we name it effective rule. And then we start with the most frequent category to find the decision tree node. If we cannot find any effective rule, we compute the entropy of each attribute and find the most proper attribute as the decision node.

Algorithm Decision Tree (D,A,R,T)

Input-A set of documents in the form of $D(c_i, t_1, t_2, \dots, t_n)$

A set of association rule in the form of $R(t_1, t_2, \dots, t_n) \rightarrow C(c_i)$

Output - A decision tree classifier

Method:

(1) D contains only training examples of the same class
(2) thmake T a leaf node labeled with class
(3) $elseif A \neq \emptyset th$ make T a leaf node labeled with
(4) $else$ (find effective rules in R) make a decision node on
(5) elcompute D's entropy and the best attribute
(6) Make T a decision node on
(7) end
(8) end
(9) end

IV.CONCLUSION

In this paper we have presented a new decision tree model which uses association rules for building decision tree, this combination performs better than traditional decision tree construction algorithms. An Apriori algorithm is used for generation of association rules. It uses a "bottom up approach", where frequent subsets are extended one item at a time. On other side decision tree requires less effort from users for data preparation and the advantage of decision tree is nonlinear relationships between parameters do not affect the performance.

REFERENCES

[1] FABRIZIO SEBASTIANI, "Machine Learning in Automated Text Classification", ACM Computing Surveys , F. 2009.
[2] D. Barbará, C. Domeniconi, N. Kang, "Mining Relevant Text from Unlabelled Documents", Proceedings of theThird IEEE International Conference on Data Mining, pp. 489 – 492, 2000.
[3] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In Proc.1993 ACM -
[4] SIGMOD Int. Conf. Management of Data, pp 207–216, Washington, D.C., May 1993.
[5] O. R. Za¨ıane and M.-L. Antonie. Classify- ing text documents by associating terms with text categories. In Thirteenth Australasian Database Conference (ADC'02), pages 215–222, Melbourne, Australia, January 2002.
[6] D. M. Blei, et al., "Latent Dirichlet allocation," Journal of Machine Learning Research, vol. 3, pp. 993-1022, 2003.

[7] T. Joachims. Text categorization with support vector machine learning with many relevant features. In 10th European Conferenceon Machine Learning (ECML-98), pp. 137–142,1998.
[8] J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. In ACM-SIGMOD, Dallas, 2000.
[9] H. Gao, Y. Fu and J.P. Li, "Classification of Sensitive Web DOCUMENTS", "Apperceiving Computing and IntelligenceAnalysis, 2008. ICACIA 2008. International Conference", pp. 295-298 , Dec. 2008.
[10] Zaki, Mohammed J; SPADE: An Efficient Algorithm for Mining Frequent Sequences, Machine Learning Journal, 42, pp. 31–60 (2001).
[11] Witten, Frank, Hall: Data mining practical machine learning tools and techniques, 3rd edition
[12] Hájek, Petr; Feglar, Tomas; Rauch, Jan; and Coufal, David; *The GUHA method, data preprocessing and mining*, Database Support for Data Mining Applications, Springer, 2004.