"Dimitrie Cantemir" Christian University
Knowledge Horizons - Economics
Volume 8, No.4, pp. 44–47
P-ISSN: 2069-0932, E-ISSN: 2066-1061
© 2017 Pro Universitaria
www.orizonturi.ucdc.ro

TECHNIQUES AND ALGORITHMS USED FOR KNOWLEDGE EXTRACTION FROM LARGE VOLUMES OF DATA

Ana-Maria Ramona STANCU[1], Mihaela MOCANU[2]
[1] Academy of Economic Studies, Bucharest, Romania , ana_maria_ramona@yahoo.com,
[2] "Dimitrie Cantemir" Christian University , rmocanu99@yahoo.fr

Abstract Large volumes of data have raised the problem of their use from the exploitation up to the result, and the Data Mining technology uses complex search methods that aim to identify some patterns and clusters of data, some trends in the consumers' behavior that can be used to anticipate their future behavior. Methods for knowledge extraction from data represent classes of problems that are subject to different solving algorithms. Of all algorithms, the current paper is dealing with decision trees and we will present a classifying application on which we will study the decision trees.

Key words:
algorithm, tree, model, rules, techniques
JEL Codes:
C8 - Data Collection and Data Estimation Methodology; Computer Programs (C82 - Methodology for Collecting, Estimating, and Organizing Macroeconomic Data; Data Analysis)

1) INTRODUCTION

The existence of large data volumes has raised the problem of reorienting their use from the retrospective exploitation process to the prospective one. Data relating to prior periods can be processed with the help of Data Mining technology, and a model is created based on them. The development of Data Mining techniques can be explained by the accumulation of data volumes that have developed over the years.

Data Mining techniques can be applied upward and downward. For the downward approach we have considered the assumptions made by different methods and means, while for the upward approach we have taken into account the knowledge extraction from available data. The larger the data volume the results obtained rely on, the more relevant they are. The data can be exploited by various techniques to obtain information, such as: genetic algorithms, group analysis, decision trees, neural networks.

2) DECISION TREES

Decision (or classification) trees have a structure used for dividing large collections of items into smaller sets by applying sequences of simple rules of decision. This technique builds up the tree for shaping the classification process. After its construction, it is applied to each tuple (item) of the database and to the classification results. The decision tree is a classifier under the form of a tree structure in which there is a leaf node and a decision node. The leaf node (or node events) indicates the target attribute value (and is represented by a square), and the decision node specifies some tests to be performed on a single attribute (and is represented by a circle).

The decision trees are part of the most important Data Mining classifying methods [1] and assist in the decision making process. [2]

The decision trees can be built up from right to left and from top to bottom by running several tests and trying to reach the best sequence in order to predict the scope. Each test creates branches leading to different stages until the test ends up with a leaf node. Their structure consists of internal nodes which indicate a

test on an attribute made up of branches that represent the result of the test and of leaf nodes representing the class labels.

We can say that the decision trees are analysis schemes that can help the decision makers by designing the possible outcomes.

a. Rules extraction from the decision trees

A system that is based on decision trees adopts the *top-down* type strategy, a strategy that searches only an area of the searching space guaranteeing the finding of a simple tree, even if it is not the simplest. The Decision trees (D.T.) can be used both as descriptive models and predictive ones. The prediction can be achieved by covering the trees. This can be done, more often than not, by sets of rules generating an output file containing the specified values. In order to extract the rules from a decision tree it is necessary to follow several steps, and the most important ones can remind that: the node leaf is included in the prediction, class, knowledge is presented in the form of rules, and the rules are created for the path of node leaf and the root node and are easily perceived by people.

b. Classification by decison trees induction

The most widely used and known algorithms that implement the decision trees (D.T.) are: CART, CHAID, ID3/C4,5 and C5,0 and SPRINT. The algorithm *CART* (Classification and Regression Trees) was introduced by Breiman in 1984 and it is based on classifications and regressions, the main construction is based on the binary division of attributes, and the *CHAID* algorithm (Chisquared Automatic Interaction Detection) uses the "chi-square" tests for segmentation, and the number of branches varies from two to the number of types of predictions. Another algorithm is the *ID3/C4.5* algorithm and the *C5.0* one.

The *ID3* algorithm was introduced by Ross Quinlan in 1986 and it is used to generate a decision tree out of a set of data based on Hunt's algorithm. These algorithms produce trees that have multiple branches for a single node combining the decision trees into a single classifier using the knowledge gained for division, the *SPRINT* algorithm is used in large data sets, and the division is based on a single attribute value.

The most widely used algorithm of those listed above is the algorithm ID3 / C4.5 that adopts the crossing to be done from top to bottom searching only a part of the search space. The algorithm *C4.5* is an extension of algorithm *ID3*, extending the classification domain from the enumeration type attributes to some of numeric type.

The following algorithm is called *ID3* and it is a forerunner of *C4.5*.

Input: The set of algorithms used in entailing the algorithm A is a set of elements whose attributes have discrete values. The set of candidate attributes is called the *list-of-attributes*.
Output: a decision tree

The algorithm ID3: *Generates the decision tree*
Step1: creates node N;
Step2: if the elements of set A belong to the same class C, then
Stage3 returns N as a leaf with class C label
Stage4 if A is void then
Stage5 returns N as a leaf labelled with the class that contains the most elements from M (the majority)
Stage6 selects the testing attribute A_i out of the set A, the attribute that devides the data most efficiently
Stage7 labels the node N with the attribute A_i
Stage8 for every known value of the attribute A_i // the X set is partitioned
Stage 9 creates a new branch from node N for condition A_i=a
Stage 10 be S_i the set of those elements from set A for which A_i=a
Stage 11 the S_i= Θ , then
Stage 12 attaches a leaf labelled with the most common value from X
Stage 13 thus attaches the node returned by *Generates the decision tree (S_i, A-{A_{ij}})*

Be S a set of data. Supposing we want to classify this set according to the attribute X with card (X) = m; it results that there will be m distinct classes C_i, with i=$\overline{1,m}$. Be s_i the number of elements from S that belong to the class C_i. Then the information required to classify a drive set is:

$$I(s_1, s_2, \ldots s_3) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

Formula 1 - Shannon's formula for entropy [3]
where p_i = the probability that an element may belong to C_i , $p_i=\frac{s_i}{s}$

Be attribute A with card (A) = v A = {a1, a2, ... a_v }. The attribute A partitions S in v {S1, S2, ... Sv}, where S_j contains elements y∈ S, A(y) = a_i. If the attribute A were to be selected as the test attribute (that is the best for the classification that divides the data most efficiently), then these subsets correspond to the

arcs starting from the node that contains the set S. Be s_{ij} = card ($C_i \cap S_j$).

Entropy, that is the information expected, can be obtained by partitioning the set S by the attribute A, is given by:

$$E(A)=\sum_{j=1}^{v}\frac{s_{1j}+...+s_{mj}}{s}I(s_{1j},....s_{mj}).$$

Formula 2 –Entropy calculation [4]

The lower the entropy is, the more homogeneous the partitioned sets will be. The information obtained by dividing the tree according to node A is $G(A)=I(s_1, s_2, \ldots s_m)-E(A)$.

The algorithm computes the information obtained by each attribute. The attribute with the most information gathered by following the tree is chosen as the testing attribute for the set of S entailed. A node is created and it is labelled with the chosen attribute, and for each value of that attribute arcs are created, the set entailed is partitioned according to that attribute.

c. The advantages of the decision trees

In 1997, Kamber M., L. Winstone, W. Gong, S. Cheng, and J. Han declared that the decision trees require restrictions on the data studied [5]. As stated by Bounsaythip C. and Rinta-Runsala in 2001, the main advantages are that are easy to understand and produce effective models. They also argue that they are applicable to real problems, for example, in commercial matters.

3) The building of the decision tree based on classification techniques

The application we developed on the classifications in Oracle Data Mining is called **w_classification_15**, an application performed using the *V_DM_Questionnaire table* from which we chose the fields: Fees and Interest from the Data Analysis which we grouped according to the Time_Expected and from the Building Class we chose the target: *Image_Name* and according to *Code_Client*.

From the connection made, we selected the model DT (Decision Trees) and then we looked at the results that occurred.

DT Model - a pattern that involves choosing a smaller set of data on which it will be built, and the remaining data set will be used to test the newly obtained model and hence to check the accuracy and trust coefficients associated to the predictive values.

Within a decision tree, at each level a type *Then If* instruction is displayed describing a rule and, as an extra layer is being added to the tree, the instruction *If – Then* appears.

From *Figure 5.3* we can notice that the company's image (pe_ansamblu_imaginea) that occurs in nodes: 0 (with a confidence of 34.92%), 1 (with a confidence of 40.63%), 2 (with a confidence of 62.98%) and 3 (with a confidence of 70.89%) represents a major impact on the people who wish to use a banking service.

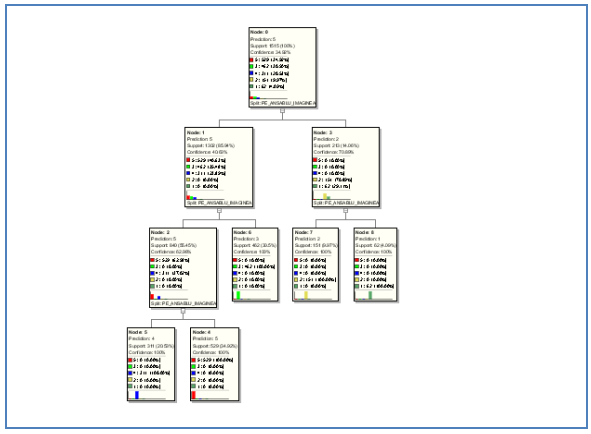

Figure 1 – The decision tree display

That is,

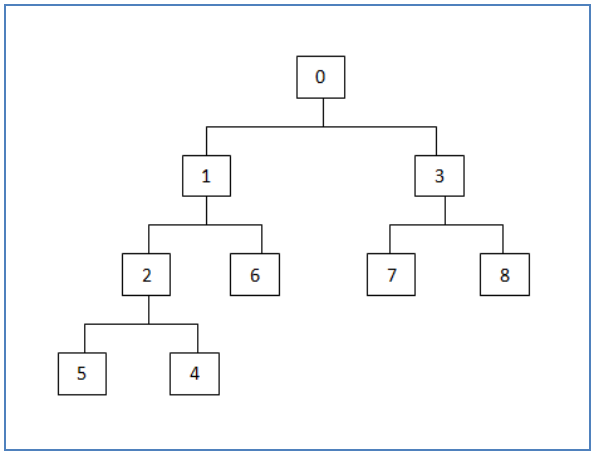

Figure 2 – Re-writing the decision tree

where:

Node	Prediction	Support	Confidence	Division
0	5	1515 (100 %)	34,92 %	Overall image
1	5	1302 (85,94 %)	40,63 %	Overall image
2	5	840 (55,45 %)	62,98 %	Overall image
3	2	213(14,06 %)	70,89 %	Overall image
4	5	529 (34,92 %)	100 %	-
5	4	311 (20,53 %)	100 %	-
6	3	462 (30,5 %)	100 %	-
7	2	151 (9,97 %)	100 %	-
8	1	62 (4,09 %)	100 %	-

Table 1 – The results following the application of the decision tree

CONCLUSIONS

For the information extraction from the data warehouses we used the Data Mining instrumental, which allows decision making to be more accurate. The Data Mining techniques used may be helpful in telecommunications, financial and banking institutions and can be used as models in achieving their market trends and the new models that are to be launched. By means of the application performed we wanted to find out if the bank's image influences people when they make use of a bank's services. According to the outcome we obtained after applying the decision tree, we noticed that the bank's image is an important factor when appealing to a bank's services.

BIBLIOGRAPHY

[47] Berry M.J., Linoff G.S., *Data Mining Techniques: For marketing, sales, and customer support*, John Wiley & Sons, Inc., (1997)

[48] Quinlan J.R., *Decision Trees and Decision-Making*, IEEE Transactions on Systems, Man and Cybernetics, vol. 20, no. 2, (1990), pp. 339-346

[333] Gh. M. Panaitescu, Transmiterea si codarea informatiei [Information Transmission and Coding], Course notes, Oil and Gas University, Ploiesti, Department of Automation, Computers and Electronics, 2015

[444] Gh. M. Panaitescu, *Transmiterea si codarea informatiei [Information Transmission and Coding]*, Course notes, Oil and Gas University, Ploiesti, Department of Automation, Computers and Electronics, 2015

[55] Kamber M., Winstone L., Gong W., Cheng S., and Han J., *Generalization and Decision Tree Induction: Efficient Classification in Data Mining*, Proceedings of 1997 Int'l Workshop on Research Issues on Data Engineering (RIDE'97), Birmingham, England, April 1997, pg. 111-120