DATA MINING METHODS WITH TREES

E + M

Marta Žambochová

1. Introduction

The contemporary world is characterized by the explosion of an enormous volume of data deposited into databases. Sharp competition contributes to the development of new processes in plotting data. Data processing from extensive databases and data warehouses has many forms. The conventional approaches to analysing data through reports and sheets are mostly based on Structured Query Languages (SQL) and On-Line Analytical Processing (OLAP). These techniques make it possible to keep an overview of the company's present situation. But they can't answer the question: Which data from the database are fundamental to commercial assimilation? Data mining deals with searching for hidden information and hidden dependencies among data.

Data often can't be analysed using standard statistical methods because they contain missing figures or are in qualitative units, and also because some databases are greatly widespread. Each organization must be able to extract important information from an extensive database. These were the main reasons why data mining was started.

Data mining has recently received much attention. It combines computer science and statistics and creates processes for the nontrivial extraction of implicit, previously unknown, potentially useful information. Data mining tools use many different statistical and non-statistical methods. Data mining hasn't replaced contemporary methods for exploring large quantities of data, but rather, appropriately complements them. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions.

Data mining is necessary for preparing the exploitation campaign and creating the sales strategy of supermarket chains, banking houses and insurance companies for their offers and so on in many countries (headed by the US). Questions of data mining are dealt with in [7] and [5].

Some goals and tasks of data mining are:

- Classification,
- Prediction,
- Estimation of values of autonomous variable,
- Segmentation (clustering),
- Analysis of relations,
- Prediction in time series,
- Detection of deviation.

Techniques used in data mining are:

- Decision trees,
- Statistical methods,
- The nearest neighbor method,
- Neural networks,
- Genetic algorithms,
- Rule induction,
- Data visualization,

2. Trees

Tree structures are used in many areas, such as computer science (data structure), biology (classification), psychology (decision theory) and social science (organization structure).

The tree is a connected acyclic graph consisting of nodes and branches (joints of nodes). One node may be chosen as the main node. It is called the „root" and is drawn at the top (if the tree is drawn from the top downward) or on the left (if the tree is drawn from left to right). Two or more nodes can follow it (and are connected to it). But some nodes don't have any offspring. These nodes are called „leaves".

3. Decision Trees

A very extensive group of trees includes decision trees. These are trees in which each internal node denotes a test on an attribute value, each branch represents an outcome of the test, and each tree leaf represents a category or

a real number (a value of a dependent variable). Decision trees are described in [1], [2], [3], [4] and [8].

Decision trees are constructed through an iterative process. They are formed recursively by dividing the space of predictor values. The first node, the root, contains all the space of values. The algorithm searches for an acceptable rule for partitioning the space. It assigns a new node to each part. It repeats this procedure until the stopping rule is fulfilled. Sometimes we can optimize the final tree; we prune it.

Decision trees differ along several dimensions:

- Splitting criterions
- Stopping rule
- Kinds of conditions of branch operation
  - Multivariate (testing on several features of the input at once)
  - Univariate (testing on only one of the features)
- The style of branch operation
  - All of the tests have two outcomes, i.e., all of the internal nodes branch into two child nodes (binary tree)
  - Some of the tests have more than two outcomes, i.e., some of the internal nodes branch into more than two offspring nodes
- Type of final tree (the describing of leaf nodes)
  - Classification trees (all leaf nodes contain assignment of the class)
  - Regression trees (all leaf nodes contain assigment of the real value - the estimate of the dependent variable‘s value)

4. ID3 Algorithm

One of the algorithms for constructing decision trees is the Iterative Dichotomiser 3 (ID3) algorithm. It uses the top-down induction method of decision trees. It is intended for smaller decision trees, but there is an extended version of the base algorithm ID3 called algorithm C4.5. (it is described in [6])

The algorithm starts with the root node. In each step, a leaf node with an inhomogeneous sample set is selected and a test for this node is chosen. This test will develop new branches with new leaves.

The attribute that minimizes the average entropy is chosen for the test. The entropy is a measure characterizing the homogeneity of sets of examples and is defined by the formula:

$$E(b) = \sum_{c} -(\frac{n_{bc}}{n_b}) \log_2 (\frac{n_{bc}}{n_b}) \qquad (1)$$

where n_b signifies the number of instances in branch b, n_{cb} signifies the number of instances in branch b of class c.

If all the instances on the branch are positive, then the probability that an instance on a branch b is positive is 1. If all the instances on the branch are negative, then the probability that an instance on a branch b is positive is 0.

The average entropy is defined by the formula:

$$E = \sum_{b} (\frac{n_b}{n_t}).E(b) \qquad (2)$$

where n_t means the total number of instances in all branches.

5. Illustration of the Construction of the Decision Tree

We want to forecast a description of clients who will have some interest in our offer. We have data containing client information. This is our training data set out of which we will construct the decision tree. (Tab. 1)

First we must find the best attribute with the best test for splitting our training data set. We will compute the entropies of all the attributes and choose the attribute with the smallest average entropy.

Computation of the average entropy of the attribute „educational attainment“:

- elementary education
  - one client hasn‘t any interest
- secondary education
  - two clients have interest
  - two clients haven‘t any interest
- university education
  - three clients have interest

$$E_1 = \frac{1}{8} \cdot (-\log_2 \frac{1}{1}) + \frac{4}{8} \cdot (-\frac{2}{4} \cdot \log_2 \frac{2}{4} - \frac{2}{4} \cdot \log_2 \frac{2}{4}) + \frac{3}{8} (-\log_2 \frac{1}{1}) = 0{,}5$$

Computation of the average entropy of the attribute „marital status“:

- single
  - one client has interest
  - two clients haven't any interest
- married
  - two clients have interest
  - one client hasn't any interest
- divorced
  - two clients have interest

$$E_2 = \frac{3}{8}\cdot\left(-\frac{1}{3}\log_2\frac{1}{3}-\frac{2}{3}\cdot\log_2\frac{2}{3}\right)+\frac{3}{8}\cdot\left(-\frac{2}{3}\cdot\log_2\frac{2}{3}-\frac{2}{3}\cdot\log_2\frac{1}{3}\right)+\frac{2}{8}\left(-\log_2\frac{2}{2}\right)=0{,}69$$

Computation of the average entropy of the attribute „habitation“:

- apartment rentals
  - two clients have interest
  - one client hasn't any interest
- cooperative apartment
  - three clients have interest
- self-contained home
  - two clients haven't any interest

$$E_3 = \frac{3}{8}\cdot\left(-\frac{2}{3}\cdot\log_2\frac{2}{3}-\frac{1}{3}\log_2\frac{1}{3}\right)+\frac{3}{8}\cdot\left(-\frac{3}{3}\cdot\log_2\frac{3}{3}\right)+\frac{2}{8}\left(-\log_2\frac{2}{2}\right)=0{,}69$$

Computation of the average entropy of the attribute „car“:

- yes
  - five clients have a car
- no
  - three clients haven't a car

$$E_4 = \frac{5}{8}\cdot\left(-\frac{2}{5}\cdot\log_2\frac{2}{5}-\frac{3}{5}\log_2\frac{3}{5}\right)+\frac{3}{8}\cdot\left(-\frac{3}{3}\cdot\log_2\frac{3}{3}\right)=0{,}606844$$

Tab. 1: Illustration Data

Name	Educational attainment	Marital status	Habitation	Car	Interest in offer (dependent attribute)
Novák	university education (UE)	divorced	apartment rentals (AR)	yes	yes
Nový	secondary education (SE)	married	cooperative apartment (CA)	no	yes
Novotný	elementary education (EE)	single	apartment rentals	yes	no
Moudrý	secondary education	single	self-contained home (SCH)	yes	no
Šikovný	university education	single	apartment rentals	yes	yes
Pokorný	secondary education	divorced	cooperative apartment	no	yes
Blecha	university education	married	cooperative apartment	no	yes
Průcha	secondary education	married	self-contained home	yes	no

Source: [9]

The minimal value of the entropy is seen to be E1= 0,5 for the attribute „educational attainment“. We choose this attribute for the test and splitting. We are now continuing with the analysis /examination of the next test.

Computation of the average entropy of the attribute „marital status“:

$$E_5 = \frac{1}{1}(-\frac{1}{1}\cdot\log_2\frac{1}{1}) + \frac{1}{4}(-\frac{1}{1}\cdot\log_2\frac{1}{1}) + \frac{2}{4}(-\frac{1}{2}\cdot\log_2\frac{1}{2} - \frac{1}{2}\cdot\log_2\frac{1}{2}) + \frac{1}{4}(-\frac{1}{1}\log_2\frac{1}{1}) +$$
$$+ \frac{1}{3}\cdot(-\frac{1}{1}\cdot\log_2\frac{1}{1}) + \frac{1}{3}\cdot(-\frac{1}{1}\cdot\log_2\frac{1}{1}) + \frac{1}{3}\cdot(-\frac{1}{1}\cdot\log_2\frac{1}{1}) = 0{,}5$$

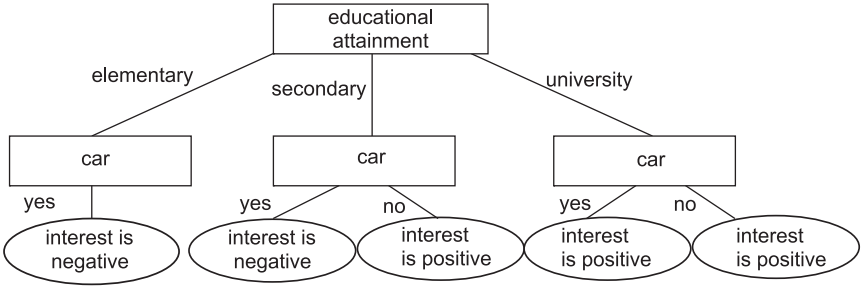
Computation of the average entropy of the attribute „habitation“:

$$E_6 = \frac{1}{1}(-\frac{1}{1}\cdot\log_2\frac{1}{1}) + \frac{2}{4}(-\frac{1}{2}\cdot\log_2\frac{1}{2} - \frac{1}{2}\cdot\log_2\frac{1}{2}) + \frac{2}{4}(-\frac{1}{2}\cdot\log_2\frac{1}{2} - \frac{1}{2}\cdot\log_2\frac{1}{2}) + \frac{2}{3}(-\frac{2}{2}\log_2\frac{2}{2}) +$$
$$+ \frac{1}{3}\cdot(-\frac{1}{1}\cdot\log_2\frac{1}{1}) = 1{,}0$$

Computation of the average entropy of the attribute „car“:

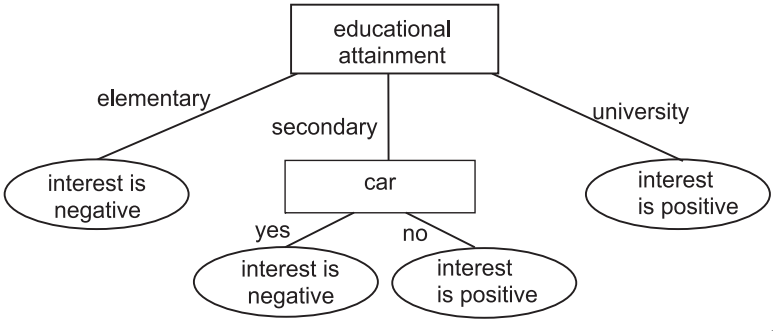
$$E_7 = \frac{1}{1}(-\frac{1}{1}\cdot\log_2\frac{1}{1}) + \frac{2}{4}(-\frac{2}{2}\cdot\log_2\frac{2}{2}) + \frac{2}{4}(-\frac{2}{2}\log_2\frac{2}{2}) + \frac{2}{3}\cdot(-\frac{2}{2}\cdot\log_2\frac{2}{2}) + \frac{1}{3}\cdot(-\frac{1}{1}\cdot\log_2\frac{1}{1}) = 0{,}0$$

Fig. 1: The Completed Decision Tree



Source: [9]

Fig. 2: The Final Tree



Source: [9]

The minimal value of the entropy is E7= 0.0 for attribute „car“. We choose this attribute for the test and splitting. This test is the closing test because the value of the entropy is zero.

We can sketch the completed decision tree in the figure 1.

We can prune this tree now. The final tree is delineated in figure 2.

This decision tree is indicative of the sorts of potentially interested persons. We can see that interest is positive among people with university education or people with secondary education but without a car.

6. Conclusion

The decision tree method is increasing in popularity for both classification and prediction. It can also be used for cluster analysis and time series in some situations. The main advantages of this method are its simplicity, non-parametric nature, robustness, and the ability to process both quantitative and qualitative variables. Decision trees can be easily converted to classification rules that can be expressed in common language.

References:

[1] ANTOCH, J. *Klasifikace a regresní stromy.* Sborník ROBUST 88 (http://www.statspol.cz/robust/index.htm)

[2] BERIKOV, V., LITVINENKO, A. Methods for Statistical Data Analysis with Decision Trees. (http://www.math.nsc.ru/AP/datamine/eng/decisiontree.htm)

[3] BREIMAN, L., FRIADMAN, J., OLSHEN, R., STONE, C. *Classification and Regression Trees.* California: The Wadsworth Statistics/Probability Series, Wadsworth International Group. Belmont. CA, 1984

[4] KEPRTA, A. *Nebinární klasifikační stromy.* Sborník ROBUST 94 (http://www.statspol.cz/robust/index.htm)

[5] KLÍMEK, P. *Aplikovaná statistika pro ekonomy.* 1.vyd. Zlín: Univerzita Tomáše Bati, 2003. ISBN 80-7318-148-74.

[6] QUINLAN, J.R. C4.5: *Programs for Machine Learning (Morgan Kaufmann Series in Machina Learning).* 1993. ISBN 1-55860-238-0.

[7] RUD, O.P. *Data mining - Praktický průvodce dolováním dat pro efektivní prodej, cílený marketing a podporu zákazníků (CRM).* 1. vyd. Praha: Computer Press, 2001. ISBN 80-7226-577-6.

[8] SAVICKÝ, P., KLASCHKA, J., ANTOCH, J. Optimální klasifikační stromy. *Sborník ROBUST 2000*, 1. vyd. Praha: Jednota českých matematiků a fyziků, 2001. ISBN 80-7015-792-5.

[9] ŽAMBOCHOVÁ, M. Použití stromů ve statistice. *Sborník Ekonomika, regiony a jejich výhledy,* FSE-Univerzita J.E. Purkyně Ústí nad Labem, 2006. ISBN 80-7044-795-8.

RNDr. Marta Žambochová
J.E. Purkyně University, Ústí nad Labem
Faculty of Social and Economic Studies
Department of Mathematics and Statistics
zambochova@fse.ujep.cz

Doručeno redakci: 25. 6. 2007
Recenzováno: 7. 10. 2007
Schváleno k publikováni: 14. 1. 2008

ABSTRACT

DATA MINING METHODS WITH TREES

Marta Žambochová

Present world is characterized by ever growing volume of data collected and saved into databases. Data often can't be analysed by using standard statistical methods because they contain many missing figures or are in qualitative units, and because some databases are in very wide usage. Each organization must be able to extract important information from an extensive database. These were the main reasons why data mining was initiated.

Tree structures are used in many diverse areas. Tree structures are frequently used in statistical data analysis, particularly in data mining.

This paper describes decision trees, their data structure and their implementation in statistical data analysis. Decision trees offer a non-algebraic method for partitioning data. Using decision trees is attractive because they offer visualization, simplicity of interpretation and high accuracy. We can utilize them to solve various classificatory and predictive exercises. They are a perfect instrument to help managers in the decision-making processes.

The decision trees are also used to form different groups of clients in order to prepare special offers and campaigns. Their potential lies in the ability to predict potential debtors on which may be decided whether to give or reject a loan or insurance to a particular costumer. The decision trees are also used to predict the potency for a new product designed for targeted customer, detect an insurance fraud, or foretell the number of people, who want to attend the competition and so on.

There are quite a few algorithms, which have been described and are being used to form decision trees. The following two are among the basic ones: algorithm ID3 and its improved version C4.5. The author is J. R. Quinlan. The first one is very illustrative and it is really important in order to acquire the basic understanding in decision trees problematic. The article contains an example of this ID3 algorithm application.

Key Words: *data mining, decision tree, ID3 algorithm*

JEL Classification: *C19, C44, C63*