Data Mining Decision Trees in Economy

Badulescu, Laviniu-Aurelian and Nicula, Adrian

University of Craiova, Faculty of Automation, Computers and Electronics, University of Oradea, Faculty of Economics

25 May 2007

DATA MINING DECISION TREES IN ECONOMY

Ph.D. s. Laviniu Aurelian Bădulescu, University of Craiova, Faculty of Automation, Computers and Electronics, Software Engineering Department, laviniu_aurelian_badulescu@yahoo.com
Ph.D. s. Adrian Nicula, University of Oradea, Faculty of Economics, Str. Universităţii nr. 1, Oradea, anicula@uoradea.ro

Data Mining represents the extraction previously unknown, and potentially useful information from data. Using Data Mining Decision Trees techniques our investigation tries to illustrate how to extract meaningful socio-economical knowledge from large data sets. Our tests find 5 attributes selection measures that perform more accurate then the best performance of the 17 algorithms presented in literature.

Data Mining, Decision Trees, classification error rate

1. Introduction

In today's knowledge-driven economy, Data Mining (DM) is an essential tool in the pursuit of enhanced productivity, reduced uncertainty, delighted customers, mitigated risk, maximized returns, refined processes and optimally allocated resources.[8] DM is defined as the process of discovering patterns in data. The process must be automatic or semiautomatic. The patterns discovered must be meaningful in that they lead to some advantages, usually an economic advantage. The data is always present in substantial quantities.[16] Machine learning provides the technical basis of DM. The main ideas behind DM are often completely opposite to mainstream statistics.[6]

DM ultimately provides a framework for dealing with uncertainty. As organizations and the global economy become more complex, sources of uncertainty become more plentiful. To make decisions more confident, new and sophisticated approaches are required. [8]

The tasks of DM like the explanation and the prediction of economic phenomena, to forecast the future or to discover the hidden laws that underlie the unknown dynamics, are the ultimate goals of economics.[2] DM techniques help to decide what tasks, activities and transactions are more economical and beneficial to use at the time.[11] DM has become more viable economically with the advent of cheap computing power based on UNIX, Linux, and Windows operating systems. [15]

Using Decision Trees (DT) techniques our investigation tries to illustrate how to extract meaningful socio-economical knowledge from large data sets derived from the Current Populations Survey. However, the extraction of useful knowledge from such large data sets is a very demanding task that requires the use of sophisticated techniques. The data sets need to be filtered and preprocessed to eliminate irrelevant attributes and incompleteness before building classifiers and predictors to efficiently extract information. [10]

DT are often used in credit scoring problems in order to describe and classify good or bad clients of a bank on the basis of socioeconomic indicators (*e.g.*, age, working

conditions, family status, etc.) and financial conditions (*e.g.*, income, savings, payment methods, etc.). Conditional interactions describing the client profile can be detected looking at the paths along the tree, when going from the top to the terminal nodes. Each internal node of the tree is assigned a partition of the predictor space and each terminal node is assigned a label class/value of the response. As a result, each tree path, characterized by a sequence of predictor interactions, can be viewed as a production rule yielding to a specific label class/value. The set of rules produced constitutes the predictive learning of the response class/value of new objects, where only measurements of the predictors are known. As an example, a new client of a bank is classified as a good client or a bad one by dropping it down the tree according to the set of splits of a tree path, until a terminal node labeled by a specific response-class is reached. [14]

Before starting the experiment, we need to specify the knowledge we want to extract, because the knowledge specificity determines what kind of mining algorithm to be chosen. In our investigation, we want to learn in which income class a person should be in real world. That is a categorized problem, therefore we decide to use DT, the one of the basic techniques for data classification, to represent the knowledge that would be mined.

Classification is a form of data analysis, and it can be used to extract models describing important data class or make future predictions. Through this mining experiment, we build a DT in order to get some classification rules and use them to predict what amount of credit line should be given to a new applicant.

We chose an attribute named "class <=50K, >50K" as class target attribute since we want to learn, based on census data, the proper classification of income (income<=$50K/yr or income>$50K/yr) for every person. These two groups can be targeted for special treatments, treatments too costly to apply to the people base as a whole. Based on income level the manager of a company could identify customers who might be attracted to different services the enterprise provides, one they are not currently enjoying, to target them for special offers that promote this service. In today's highly competitive, customer-centered, service-oriented economy, data is the raw material that fuels business growth—if only it can be mined [16].

Our investigation can be used for a demographic and socioeconomic segmentation. Demographic and socioeconomic segmentation is based on a wide range of factors including age, sex, family size, income, education, social class and ethnic origins. So it is helpful in indicating the profile of people who buy a company's products or services [12].

2. Performance tests

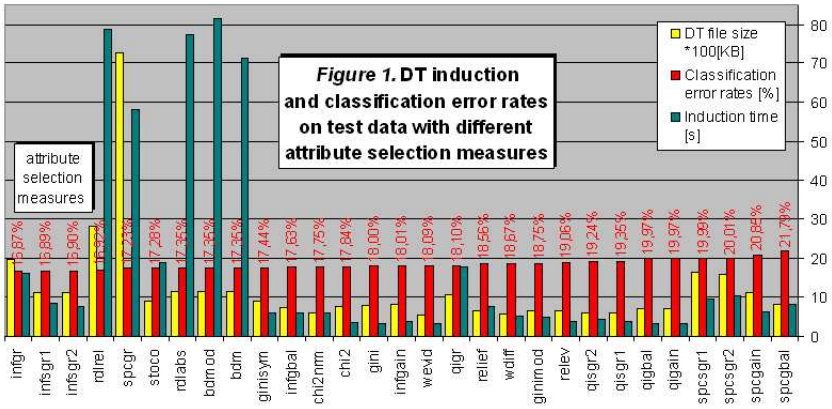
2.1. Decision Tree Induction

For the performance tests we use software developed by C. Borgelt [4]. At first, DT was induced on the 32561 test records of the *Adult Database*. *Adult Database* [5] was donated by Ron Kohavi[7] and has 48842 instances (train=32561, test=16281) and 15 attributes: *age*, *workclass*, *fnlwgt*, *education*, *education-num*, *marital-status*, *occupation*, *relationship*, *race*, *sex*, *capital-gain*, *capital-loss*, *hours-per-week*, *native-country*, *class* (target attribute). Missing values are confined to attributes *workclass*, *occupation* and *native-country*. There are 6 duplicates or conflicting instances. For the label ">50K" the probability is 23.93% and

for the label '<=50K' it is 76.07%. Extraction was done by Barry Becker from the 1994 Census database. Prediction task is to determine whether a person makes over 50K a year. *Adult Database* was used in many others publications [9].

There has been used 29 attribute selection measures on which the splitting of a node of the DT has to be realized. They are found in the literature, some of them being used in the induction of some very well-known DT. Attribute selection measures [3, 4] used for induction, pruning and execution of DT are: information gain (*infgain*), balanced information gain (*infgbal*), information gain ratio (*infgr*), symmetric information gain ratio 1 (*infsgr1*), symmetric information gain ratio 2 (*infsgr2*), quadratic information gain (*qigain*), balanced quadratic information gain (*qigbal*), quadratic information gain ratio (*qigr*), symmetric quadratic information gain ratio 1 (*qisgr1*), symmetric quadratic information gain ratio 2 (*qisgr2*), Gini index (*gini*), symmetric Gini index (*ginisym*), modified Gini index (*ginimod*), relief measure (*relief*), sum of weighted differences (*wdiff*), χ^2 (*chi2*), normalized χ^2 (*chi2nrm*), weight of evidence (*wevid*), relevance (*relev*), Bayesian-Dirichlet/K2 metric (*bdm*), modified Bayesian-Dirichlet/K2 metric (*bdmod*), reduction of description length - relative frequency (*rdlrel*), reduction of description length - absolute frequency (*rdlabs*), stochastic complexity (*stoco*), specificity gain (*spcgain*), balanced specificity gain (*spcgbal*), specificity gain ratio (*spcgr*), symmetric specificity gain ratio 1 (*spcsgr1*), symmetric specificity gain ratio 2 (*spcsgr2*).

Performances regarding the *size of the file which contains DT* and the *time needed to induce DT* have been noticed. As it can be seen on the Figure 1, excepting 5 attribute selection measures (*rdlrel*, *spcgr*, *rdlabs*, *bdmod*, *bdm*) that needed long periods of time for inducing DT, the other 24 measures had small values at this performance. Related to the file size containing DT, only one measure (*spcgr*) has a very large value in comparison to the other 28 measures.



DT induced at this step, has been executed on the 16281 test samples of the *Adult Database*. The most important performance for the classification of the different DT, the *classification accuracy on the test data*, data completely unknown at the training of DT, has been distinguished. This performance is expressed by *classification error rate* on the test

data and is represented next to the performance of the *file size containing DT* and the amount of *time* of the DT induction in the Figure 1 chart. In this chart, the performances are sorted in the ascending order of the classification error rates values on the test data. It can be noticed that the highest performance for the error rate on the test data is obtained by the *infgr* measure.

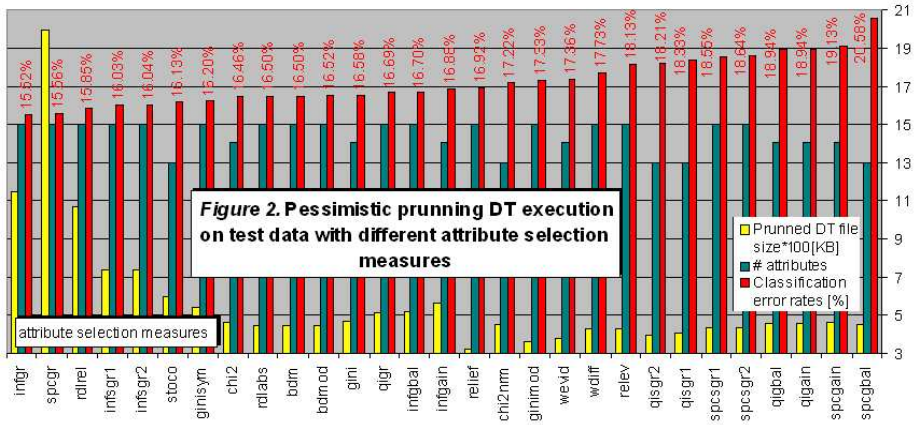


Figure 2. **Pessimistic prunning DT execution on test data with different attribute selection measures**

2.2. Decision Tree pruning with pessimistic pruning method

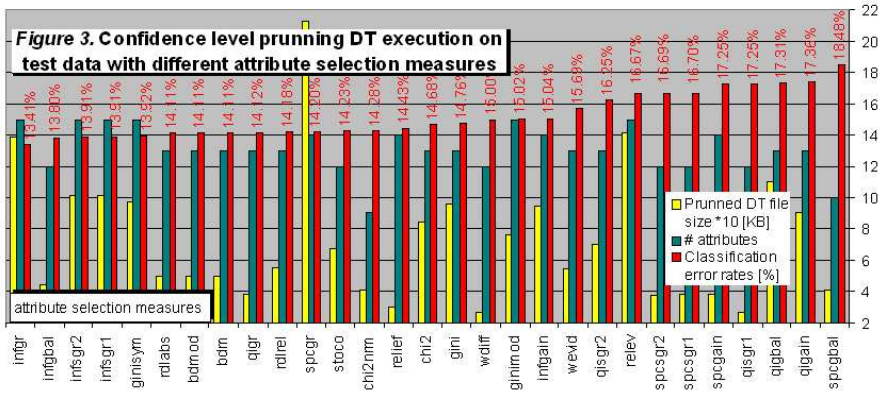
DT induced at the previous step was pruned by using the pessimistic pruning method. The performances that were brought in sight at this pruning were *number of attributes used by DT* and *pruned DT file size*. We can see that a number of 5 measures (*chi2nrm, qisgr1, qisgr2, spcgbal, stoco*) needs only 13 attributes in the construction of pruned DT, a number of 7 measures (*chi2*, *gini*, *infgain*, *qigain*, *qigbal*, *spcgain*, *wevid*) needs 13 attributes in the construction of pruned DT, and the other 17 measures use all the 15 attributes in the construction of pruned DT. DT pruned at this step with pessimistic pruning method, was executed on the 16281 test data of the *Adult Database*. The most important performance for the classification of the different DT, the accuracy of classification on the test data, which are completely unknown at the DT training, is represented along with the performance *number of attributes used by DT* and *pruned DT file size*, in the Figure 2 chart. In this chart, the performances are sorted in ascending order by values of the classification error rates. Correlation coefficient between pruned DT file size and classification error rates, *-0.503*, suggests a negative correlation. The greater the pruned DT file sizes, the smaller the error rate value.

We can notice that the best performance at the classification error rate is obtained by the same *infgr* measure. For pruned DT the accuracy of the classification error rate is better than for unpruned DT.

2.3. Decision Tree pruning with confidence level pruning

DT induced at first step was pruned using confidence level pruning. The performances taken into consideration at this pruning were the *number of attributes used by DT* and *pruned DT files size*. It's noticeable that a (*chi2nrm*) measure needs only 9 attributes to build pruned DT, a (*spcgbal*) measure needs 10 attributes to build pruned DT, a number of 6 measures (*infgbal*, *qisgr1*, *spcsgr1*, *spcsgr2*, *stoco*, *wdiff*) needs 12 attributes to build pruned DT, a number of 11 measures (*bdm*, *bdmod*, *chi2*, *gini*, *qigain*, *qigbal*, *qigr*, *qisgr2*, *rdlabs*, *rdlrel*, *wevid*) needs 13 attributes to build pruned DT, a number of 4 measures (*infgain*, *relief*, *spcgain*, *spcgr*) needs 14 attributes to build pruned DT, and the other 6 measures use all the 15 attributes to build the pruned DT. It is noticeable that the number of the necessary attributes of the pruned DT for the classification has decreased unlike the pessimistic pruning.

Pruned DT at this step with confidence level pruning was executed on the 16281 test samples of the *Adult Database*. The accuracy of the classification on the test data is expressed in the classification error rate and is represented along with the performance of the *number of attributes used by DT* and *pruned DT file size* in the Figure 3 chart. In this chart the performances are sorted in the ascending order of the values of the classification error rates on the test data. We can notice that the best performance of the classification error rate on the test data is obtained by the same *infgr* measure. The accuracy of the classification is better than for the unpruned DT and for the DT pruned with pessimistic pruning method.



3. Conclusions

From documentation of *Adult Database*[1] we find that the following algorithms, with the classification error rates specified in square brackets: *FSS Naïve Bayes* [14.05%], *NBTree* [14.10%], *C4.5-auto* [14.46%], *IDTM (Decision table)* [14.46%], *HOODG* [14.82%], *C4.5 rules* [14.94%], *OC1* [15.04%], *C4.5* [15.54%], *Voted ID3 (0.6)* [15.64%], *CN2* [16.00%], *Naive-Bayes* [16.12%], *Voted ID3 (0.8)* [16.47%], *T2* [16.84%], *1R* [19.54%], *Nearest-neighbor (3)* [20.35%], *Nearest-neighbor (1)* [21.42%], *Pebls* [Crashed],

were run on *Adult* test data, all after removal of unknowns and using the original train/test split. The best performance of classification accuracy on test data is performed by *FSS Naïve Bayes algorithm* with value of 14.05% for classification error rate. Our tests find 5 attributes selection measures that outperform the best performance of the 17 algorithms presented above. Thus, for confidence level pruning DT our tests were showed that *infgr* measure obtain an error rate of 13.41%, *infgbal* an error rate of 13.80%, *infsgr1* and *infsgr2* an error rate of 13.91%, and *ginisym* an error rate of 13.92%.

Our task was to use DM tools, like DT algorithms, to address what facts are and how they affect the income of a person. From the view of business values, this investigation has commercial benefits in real business world today to attract more and more potential and valuable customers, enlarge market shares in the industry, and minimize the risks for the financial companies [13].

BIBLIOGRAFY

[1] *adult.names*, ftp://ftp.ics.uci.edu/pub/machine-learning-databases/adult/.

[2] Alves A., Camacho R., Oliveira E., "Inductive Logic Programming for Data Mining in Economics", in Proc. of the 2nd International Workshop on Data Mining and Adaptive Modeling Methods for Economics and Management, Pisa, September 2004.

[3] Borgelt C., "A decision tree plug-in for DataEngine", in Proc. European Congress on Intelligent Techniques and Soft Computing *(EUFIT)*, vol. 2, 1998, pp. 1299-1303.

[4] Borgelt C., http://fuzzy.cs.uni-magdeburg.de/~borgelt/dtree.html.

[5] ftp://ftp.ics.uci.edu/pub/machine-learning-databases/adult/.

[6] Jessen H.C., Paliouras G., "Data Mining in Economics, Finance, and Marketing", in Lecture Notes in Computer Science, Vol. 2049/2001, Springer Berlin/Heidelberg, 2001, p. 295.

[7] Kohavi R., "Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid", in Proc. of the 2nd International Conf. on Knowledge Discovery and Data Mining, 1996, pp. 202-207.

[8] Kudyba S.(ed.), "Managing Data Mining, Advice from Experts", IT Solutions Series, Idea Group, USA, 2004, pp. VII-VIII.

[9] Larose D.T., "Data Mining Methods and Models", John Wiley & Sons, Hoboken, New Jersey, 2006, pp. 18-25.

[10] Lazăr A., "Knowledge Discovery for Large Data Sets", Youngstown State University, 2003, http://www.cis.ysu.edu/~alazar/pdf/2003ResearchProposal.pdf.

[11] Nayak R., "Data Mining and Mobile Business Data", in Khosrow-Pour M.(ed.) Encyclopedia of information science and technology, vol. II, Idea Group, 2005, p. 700.

[12] Payne A., "Handbook of CRM: Achieving Excellence in Customer Management", Elsevier Butterworth-Heinemann, Great Britain, 2005, p. 67.

[13] Peng J., Du P., "Classification with Different Models on Adult Income", 2002, http://citeseer.ist.psu.edu/cache/papers/cs/27570/http:zSzzSzwww.cas.mcmaster.cazSz~cs4tf3zSzprojectzSzreport_he.pdf/classification-with-different-models.pdf

[14] Siciliano R., Conversano C., "Decision tree induction", in Wang, J.(ed.), Encyclopedia of data warehousing and mining, Idea Group, USA, 2006, p. 353.

[15] Thomasian A., "Active disks for Data Mining", in Wang, J.(ed.) Encyclopedia of Data Warehousing and Mining, Idea Group, USA, 2006, p. 6.

[16] Witten I.H., Frank E., "Data mining: practical machine learning tools and techniques", 2nd ed., Elsevier, Morgan Kaufmann, USA, 2005, p. 5.