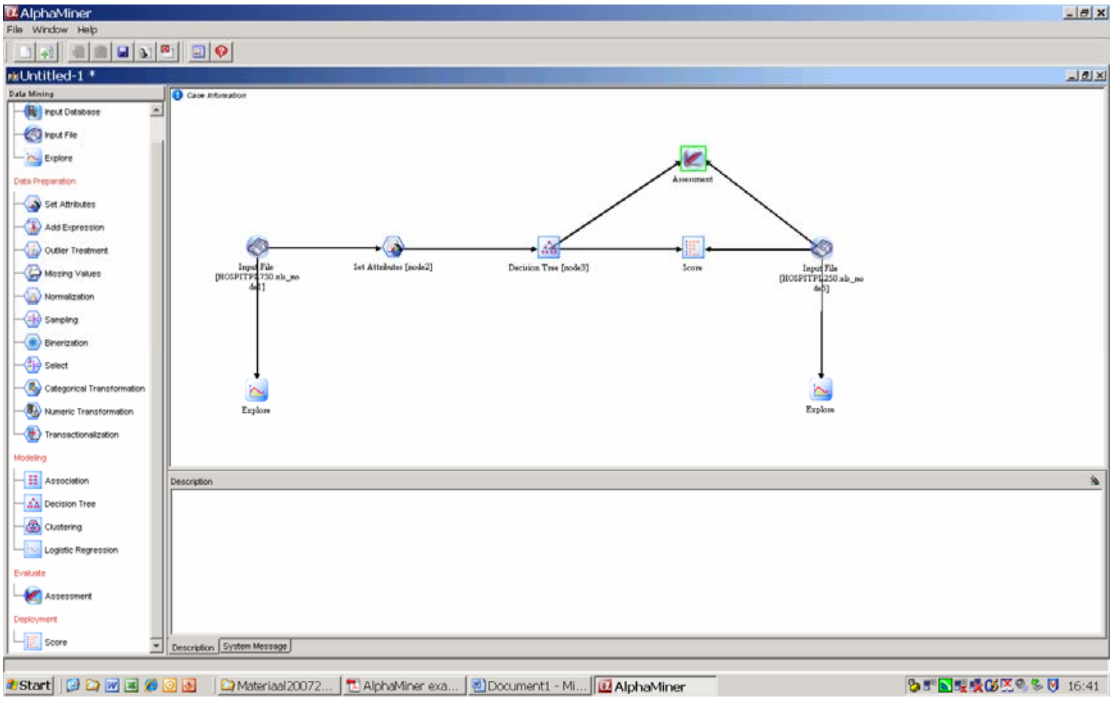
Available material:
Alpha Miner program
UserGuideAlpha Miner
Example Session Alpha Miner
HOSPITPL750.xls
HOSPITPL250.xls
INCOME.XLS

Install the Alpha Miner Program and use the Example Session document and the User Guide to become a little bit familiar with the Decision Tree Miner.

Assignment 1.1 Hospital data

The files HOSPITPL750.xls and HOSPITPL250.xls contain information about 750 and 250 patients (diagnoses (D1, D5), gender, and age, insurance, and hospital time (one, two, 3_5, 6_9, 10). Use the 750 patient in combination with decision tree to build a decision tree and classification rules. Use the 250 patients to test the classification performance of your model.



Answer the following questions:

During building the decision-tree uses only the default settings!

1. How many rules are there?
2. What is the classification performance (Precision) on the learning material?
3. What is the classification performance (Precision) on the test material?
4. What is in your opinion the best rule (give arguments)

In a second learning session we try to get a larger tree and more rules. This is not so simple, because an automatic pruning algorithm is used to prune the tree. But let we try. Open the decision tree and in the tree setting change the confidence factor from 0.25 to 1 (the confidence factor is used for pruning (smaller values incur more pruning)).

5. How many rules are there?
6. What is the classification performance (Precision) on the learning material?
7. What is the classification performance (Precision) on the test material?

In a third learning session we try to get a smaller tree and fewer rules. Open the decision tree and in the tree setting change the confidence factor to 0.05 (the confidence factor is used for pruning (smaller values incur more pruning)).

8. Are there important differences between this model and the previous one?
9. Verify that in the Decision-tree-view, you can find information about the predicted classification of each of the 750 patients in the learning material and in Score-view the classification of the learning material.

Assignment 1.2

Below, a description of CRM mining assignment is given. Beside this description the following material/knowledge is needed:

- The AlphaMiner software. The software is available via the AlphaMiner web side http://www.eti.hku.hk/alphaminer/ following the instruction on this web side to install the software on your machine.
- User Guide alpha-Miner and example session alpha-Miner.
- The dataset INCOME.XLS

Introduction

A bank has developed a new financial product. The new product is only attractive for households with a totally yearly income of €50.000 or more. To promote the new financial product, the bank will get in contact with potential buyers of the product. De costs to get in contact with potential buyers are €50 (information material via direct mailing + contact via the telephone). If the product is really bought by a consumer the profit is about €3010 (minus €50). The bank has on his disposal a database with about 1 million personal descriptions of there consumers. Beside address and telephone number, the following information is available:

Age: continuous
Workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
Education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
Marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
Occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
Relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
Race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
Sex: Female, Male.
Capital-gain: continuous
Capital-loss: continuous
Hours-per-week: continuous

Information about the yearly income is unfortunately not in the database. Meanwhile, the mentioned promotion campaign is started and about 20.000 families are already contacted. Only 24.75% of them have an income from more then €50.000 en only 11.1% of the families with an income above the €50.000 will really buy the new financial product (only 2.7% of the contacted families). The bank considers using a knowledge discovering technique on the 20.000 already contacted consumers to select a subset of the remaining families that will be contacted by direct mailing and a telephone call. From the 20.000 contacted families only about 540 really bought the new product. For that reason the bank is reserved to use the 20.000 records to predict directly, which families will buy the product (which families to contact). As a first step they will try to use knowledge discovery techniques to select families with an income above the €50.000 boundary.

The assignment:
Use the database with the 20.000 already contacted families for witch we know if they have a total salary above or bellow the 50K boundary (about 25%). It's clear that this hands-on tasks are pretty open. Use this database in combination with AlphaMiner software to develop classification model. Normally we will use 10 fold cross validation to search for the optimal parameters settings and to have an idea of the classification quality of the model. However, performing such experiments is time consuming. Therefore, in this assignment[1], you can use the default parameter settings of the alpha miner; beside the minimal number of leaf nodes (change this value from 2 to 50).
An important practical aspect of this classification problem is the following. If we inaccurate select a family whit a salary below the 50K boundary we will lose €50, however if we miss a family with a salary above the 50K, we miss a potential profit of €3010-€50=€2960. However, AlphaMiner, as many other practical knowledge discovery tools, gives information about the reliability of the inducted rules. Without performing 10-fold-CV experiments, in this assignment, it is allowed to use the (dangerous) hypotheses that the reliability information of the mined rules is correct. This gives the possibility to search for a boundary in the rule set to get optimal (estimated) profit. Or in other words: if the decision rule predict only 45% >50K incomes, it is possible profitable to direct mail this group of people.

Compare the estimated profit if we will mail all clients, with the estimated profit if we use the mined classification model to select for mailing. Explain how the mined classification model is used to reach this profit. Use about one A4 page (with a possible appendix) to report your results. Name your document TonWeijters01.doc (please use your one name!) and post it to the study web in the proper directory. Deadline: February 11 2008, 8:30

Success,
Ton Weijters

[1] In the Alpha Miner we have a combination of tree building and pruning based on statistical techniques. Moreover, we have a relative large learning data set (20.000 cases) and a splitting stop criteria of 50 cases. For this reason, the dangerous for over fitting in this assignment is not very high. However, methodological it is absolute wrong!