PKDD99 Discovery Challenge - Financial Domain

Boris Levin, Abraham Meidan, Alex Cheskis, Ohad Gefen, Ilya Vorobyov
WizSoft
Abraham@wizsoft.com

In the following pages we will present the application of WizWhy (version 2.05) to the PKDD99 Discovery Challenge.

WizWhy is a data mining tool containing a proprietary Association Rules algorithm (i.e., an algorithm that reveals all the if-then rules in the data set, that meet pre-defined threshold parameters). On the basis of the discovered rules, WizWhy –
(1) Summarizes the data;
(2) Points out interesting phenomena in the data (= unexpected rules);
(3) Issues predictions for new cases.

A white paper describing WizWhy and a working demo can be downloaded from WizSoft's web: www.wizsoft.com

Proposed business objectives:

We limited ourselves to the following three questions:
(1) Which Accounts are unlikely to repay their loans? Whom should the bank watch carefully in order to minimize the bank losses from unpaid loans?
(2) Which Clients not having a credit card are likely to possess one? To whom to offer a credit card?
(3) Which Clients not using the service of payment orders are likely to use it? To whom to offer the service of payment orders?

The data mining effort:

To answer the first question (which Accounts are unlikely to repay their loans) we performed the following:
(1) We joined the Loan table with the Account and Demographic tables by an SQL query.
(2) We added the data per Account from the Transactions table.
(3) We divided the Loan table into two tables, one table contained the Accounts where the loan contracts had been concluded (status A or B) and the other table contained the Accounts having running contracts (status C or D).
(4) We checked and validated that the distribution of the values in each of the fields in one table was similar to the distribution in the other table. On the basis of this fact, we concluded the applicability of the patterns discovered in the first table on the second table.
(5) We assumed that the cost of a miss is 10 times higher than the cost of a false alarm.
(6) We ran WizWhy on the first table where the dependent variable was the Status field. WizWhy revealed the if-then rules explaining when Accounts did not repay the loans (status B).
(7) On the basis of these rules we issued predictions on the Accounts in the second tables ranking for each Accounts the likelihood of not paying the loan.

To answer the second question (to whom to offer a credit card) we performed the following:

(1) We joined the Client tables with the Disposition, Demographic and Credit tables by an SQL query.

(2) We ran WizWhy on the Client table where the dependent variable was the Type field. Since the number of records was small we distinguished between Clients having or not having a credit card and disregarded the type of card. WizWhy revealed the if-then rules explaining when Client have a credit card.

(3) On the basis of these rules we pointed out those Clients expected to have a credit card following the rules, but in fact did not have one. These clients are more likely to have a credit card than the others, and therefore the bank should offer them one.

To answer the third question (to whom to offer the service of payment orders) we performed the same method that we used in answering the previous question. Since we are limited to 6 pages, the results in regard to this question are not included in this report.

The discovered knowledge:

(1) Accounts unlikely to repay their loans:

Below is a list of Accounts having the highest probability of not repaying their loans. Since the report is restricted to 6 pages, the list contains the first 25 Accounts only. The Accounts are sorted by the expected probability of not repaying their loans.

STATUS	Probability of not repaying	Account ID	ACCOUNT_A
D	0.902	215	5267
C	0.897	240	5837
C	0.591	223	5385
C	0.59	227	5477
C	0.575	297	7614
C	0.555	189	4715
D	0.554	135	3189
D	0.548	127	3037
D	0.547	235	5700
C	0.546	222	5362
D	0.538	134	3166
C	0.53	195	4803
C	0.528	238	5742
D	0.526	151	3711
C	0.526	274	6950
D	0.525	315	8085
C	0.524	169	4260
D	0.521	26	808
C	0.516	171	4293
C	0.516	214	5263
D	0.516	77	2051
D	0.515	184	4618
C	0.513	199	4948
D	0.512	197	4858
C	0.506	320	8169

For each Account in the list, WizWhy can present the rules that entail the prediction.

Interestingly, some of these Accounts, although still repaying their loans, as can be seen from the Status field, are not expected to repay the loans completely.

The main field that explains the dependent variable is the total sum of the transactions in the Relation Transaction, where the value of the K_symbol field is SANC.UROK.

Example of a rule: When the total sum of sanction (see above) was between –638.00 and 31,491.00, and the value of the district_id field was between 6 and 31, *none* of the accounts repaid the loans. There were 9 accounts in this group, and the probability that this phenomenon was accidental was almost 0.

Applying the discovered knowledge:

In addition to watching the Accounts that unlikely to repay their loans, the bank can use the WizWhy predictor to check new accounts applying for a loan. Once the data of new accounts is entered, the WizWhy predictor will apply the rules and calculate the likelihood that the account will not repay the loan.

(2) Clients who are likely to have a credit card but in fact don't have one

Below is a list of the first 21 Clients who deviate from the discovered rules in regard to having a credit card. According to the rules each of these Clients should have a credit card, but in fact do not have one. Therefore, it makes sense for the bank to address these Clients and offer them a credit card. The Clients are sorted by the probability (according to the rules) of having a card. Note that the highest probability was 0.216. However, since we assumed that the cost of a miss is 10 times higher than the cost of a false alarm, it makes sense to address these Clients.

Probability of having a card	Client	Account
0.216	71	75
0.216	350	366
0.216	234	248
0.216	55	57
0.216	2545	2673
0.216	4430	10478
0.215	3876	4269
0.215	218	230
0.215	983	1029
0.215	4082	5794
0.215	2239	2356
0.215	4263	8316
0.215	85	90
0.214	395	414
0.214	22	22
0.214	266	281
0.214	368	386
0.214	3908	4375
0.214	3974	4681
0.214	3982	4768
0.214	274	289

Applying the discovered knowledge:

As mentioned, the bank can address those Clients who are likely to possess a credit card according to the rules, but in fact don't have one. The list above contains the first 21 Clients. The complete list contains many more.

(3) Clients likely to use the service of payment orders but in fact don't use it

We dealt with this issue exactly in the same way described above in regard to the second issue. However, as already mentioned, since this report is limited to 6 pages, we haven't presented the results here.