Discovery of interesting rules and subgroups in a financial database

Irene Weber,
Institut für Informatik, Universität Stuttgart,
Breitwiesenstr. 20-22, 70565 Stuttgart, Germany

Abstract

The data mining task addressed here is discovery of interesting rules or interesting subgroups in the financial database. In particular, one aim is to find subgroups of clients where the fraction of credit card holders is especially large or small. The second aim is to find indicators for bad or good loans.

The rule space to be searched is declared manually, it involves symbolic, numerical, and derived attributes. The The interestingness of rules is determined by the statistical criterion, "implication intensity" combined with a minimum coverage requirement.

The data mining seems to have discovered interesting rules indicating good or bad loans that might help to improve the understanding of clients, while being less successful with the credit card holders problem.

1 Introduction

The data mining goal addressed here is to improve the understanding of groups of clients, in particular, to find characteristics of credit card holders (in order to identify, e.g., potential new card holders), and to find indicators for bad or good loans. In this application, the focus is not on learning classificators with high prediction accuracy, but rather on finding understandable and possibly general rules describing typical or indicative behaviours of clients. This data mining task is also known as interesting subgroup discovery. Rule or subgroup discovery involves an exhaustive search in the rule space. However, if a large number of rules is enumerated and evaluated, typically many rules describing significant associations are discovered, especially in large databases. Therefore, defining and/or choosing appropriate rule selection criteria for a given task is crucial for successful rule discovery. A specific problem of discovery tasks is that here, in contrast to classification and prediction tasks, a formal and generally accepted success criterion like predictive accuracy is lacking. For this reason, evaluation of the discovered rules necessarily is subjective and requires domain knowledge, and, ultimately, the judgement of the domain experts.

2 The data mining approach

The core of the data mining system applied here is a manual declaration of the rule space to be searched. This includes declaration of the attributes/features that are to occur in the condition part of rules, specification of a population of cases of interest, and specification of a target group belonging to the population. For example, when searching for interesting subgroups of credit card holders, the population may be declared to be the set of account owners, the target group consists of account owners holding a credit card, and condition attributes are, among others, age, sex, and home region of the account owners. A rule in the resulting rule space is $Region = prague, Age < 21 \rightarrow has\_card$. The user can restrict and control the search space by defining syntactical restrictions on rules, specifically, declare sets of literals that must or must not co-occur with a given a literal in the same rule (for details see [Web99]. The underlying rule language is a logical language. The rules are translated into Prolog clauses, and rules are evaluated by a Prolog interpreter. This approach offers great flexibility. It allows to process multi-relational databases and is able to incorporate Prolog predicates defining background knowledge.

The search algorithm conducts an exhaustive search through the search space in order to identify all acceptable rules. The search is organized in levels. In the first search level, rules with one condition literal are evaluated, in the second search level, rules with two condition literals are evaluated, and so on. Depending on the interestingness criterion, the algorithm is able to prune some rules from the search space, nevertheless, the deeper levels of the search often contain huge numbers of rules so that it is not feasible to investigate the rule space completely, and the search process has to be interrupted. This biases the algorithm towards general rules.

As interestingness criterion, the implication intensity suggested in [Fleu95] is used. The implication intensity is a statistical measure. For a given rule $A \rightarrow B$, the implication intensity computes the probability of observing the given or a smaller number of cases contracting the rule (cases for which the condition A is true, but the conclusion B is false, the so-called denial) if one assumes statistical independance of A and $\bar{B}$. The implication intensity is defined to be 1 minus that probability. If the implication intensity of a rule exceeds a user-defined threshold, the rule is accepted as potentially interesting. For the experiments conducted here, the implication intensity is combined with a minimum coverage requirement, that means, besides reaching the implication intensity threshold, a rule must also cover at least a minimum number of cases in order to be accepted. This avoids discovery of very specific rules, which I considered as less interesting in the data mining application at hand. If the search algorithm discovers an acceptable rule, it prunes th rule from from the search space, that is, its specialisations (rules with the same and additional condition literals) are excluded from the search.

I have also conducted some preliminary experiments with an alternative setting, namely discovery of the k best rules, and with an alternative interestingness criterion, distributional unusualness [Wro97]. Implication intensity was prefered, because it offers a statistical interpretation in terms of probability. The drawback of the "k best" setting was that it often produced k more or less specific variants of one rule as the k overall best.

3 Data mining efforts

3.1 Data preparation

Regarding the business goal "indicators for bad /good loans", we are interested in clients causing problems and losses with bad loans. Bad loans are loans still running or finished with problems, good loans are all loans finished or still running without problems. In total, there are 682 loans, 76 of which are bad.

For this task, the population of interest was declared to be the set of loans, the target group being either the group of bad loans or the good loans. This is sensible because a loan indirectly identifies a certain client, namely the owner of the account on which the loan is granted. I pre-computed a joined relation including selected attributes (as described below) from the loan, account, client, district, card and and transaction relation. The joined relation consisted of 682 tuples, one for each loan. The order relation was kept separate because the orders are not in zero-to-one or one-to-one relation to loans, but an arbitrary number of permanent orders might exist for each account.

In the card holder task, one is interested in clients. No client owns or disposes on more than one account, so that the disposition relation relates each client to exactly one account. All card holders are owners of an account, whereas nobody disposing on, but not owning an account holds a card. Since the dispositions of type owner uniquely identify a client, the set of all dispositions of type "owner" was declared to be the ground population, the target group is the set of dispositions occuring or not occuring in the card relation.

For this experiment, I pre-computed a joined relation including relevant attributes as above consisting of 4500 tuples, each tuple describing a disposition of type owner. Again, the order relation was kept separate.

3.2 Declaration of the search space

A first observation with regard to the available data is that the search space of potential rules is very large in comparison to the amount of available data. Although the transaction relation is contains more than 10^6 entries, all these entries describe not more than 682 loans or 4500 dispositions resp. in the task setting at hand. Among the vast wealth of attributes (up to, e.g., patterns in transactions as attributes of accounts),

	relation	attribute	explanation	values
1	loan	Status	good = {a, c}, bad = {b, d}	good, bad, a, b, c, d
2		Loan	a loan is or was granted for the account	yes, no
3	client	YoB*	year of birth of the account owner	$\geq$35, $\geq$42, $\geq$49,...,$\geq$77, <35, <42,...,<77
4		Sex	sex of the account owner	male, female
5	district	Region	home region of the account owner	prague, e/w/s/n/c_Bohemia, s/n_Moravia
6		Urb*	rate of urbanity of account owner's district	$\geq 4, \geq 5, \geq 6, \ldots \geq 10, \leq 9, \ldots \leq 4, \leq 3$
7		Avg_Sal*	avg. salary in account owner's district in thousands	$\leq$11, $\leq$10, $\leq$9, $\leq$8, $\geq$9, $\geq$10, $\geq$11, $\geq$12
8	account	Freq	frequency of issuance of statements	monthly,weekly, immediate
9	trans.	AvgInc*	avg. sum of monthly credits	<50000,<25000,>25000,>50000
10		VarInc*	standard deviation of monthly income	<40000, <20000,>20000,>40000
11		#Inc*	avg. number of monthly credits	$\geq 4, \geq 3, \geq 2, \leq 2, \leq 3, \leq 1$
12		AvgWD*	avg. sum of monthly withdrawals	<50000,<25000,>25000,>50000
13		VarWD*	standard deviation of monthly withdrawal	<40000,<20000,>20000,>40000
14		#WD*	avg. number of monthly witdrawals	$\geq$2, $\geq$4, $\geq$6, $\leq$6, $\leq$4, $\leq$2
15	card	Card	a credit card exists for the account	yes, no
16		Type		junior, classic, gold
17	order	Order	a perm. order exists for the account	yes, no
18		OType*	K_Symbol of order	household, leasing, loan, insurrance, unknown
19	disp.	Disp	a second client disposes on the account	yes, no
20	client	DSex	Sex of the second client	male,female

Table 1: Used attributes.

I have selected the attributes summarized in table 1. A large number of alternative attributes could be defined and investigated in further experiments.

Numerical attributes are discretized as follows. We define attribute values corresponding to interval boundaries. Combinations of several interval boundaries (where contradictive combinations are suppressed) define intervals of varying extensions (for details see [Web98] The number of interval boundaries is chosen quite arbitrarily, their position roughly realises "equi-distant" partitioning.

The asterisk marks * attributes of which several features can co-occur in a single rule. Besides the numerical attributes, this is attribute 18. As several permanent orders may exist on one account, it makes sense to ask for different combinations of types of permanent payments issued for an account. (However, in the experiments no such patterns were identified as significiant).

Attributes 7-12 describe properties of accounts and are derived from the transaction database by SQL statements as follows where *TYPE* is replaced by PRIJEM or VYDAJ:

```
SELECT ag.account_id, AVG(ag.in_sum), AVG(in_count), SQRT(VARIANCE(ag.in_sum)) FROM (
SELECT t2.account_id AS account_id, t2.monat, SUM(t2.amount) AS in_sum,
COUNT(t2.amount) AS in_count FROM trans t2
WHERE type = *TYPE* GROUP BY t2.account_id, t2.monat) AS ag  GROUP BY ag.account_id"
```

Regarding attributes 19 and 20, I found that no client owns or disposes on more than one account, but some accounts are operated by two clients. Thus, having a second client operating on it, as tested by attribute 19, possibly is an interesting feature of an account. Attribute 20 asks for the sex of the second disponent.

4 Results

4.1 Bad loans

After running some hours, the algorithm had discovered about 50 rules. Selected rules are shown below.

```
Minimum coverage:             0.015000  =  10 cases
ImpInt Threshold:             0.900000  => Index Threshold:  1.281550
Target class:                bad loans
Extension of target class:     76           non-target class: 606
Apriori fraction of target class: 0.111
1. VarInc > 20000.
   denial:    204  hits:     53  cover:    257  index: 1.612052   frac. target: 0.206
2. Disp = no, VarWD > 20000.
   denial:    159  hits:     44  cover:    203  index: 1.591772   frac. target: 0.217
   Disp = no, Card = no.
   denial:    332  hits:     71  cover:    403  index: 1.378772   frac. target: 0.176
3. VarInc < 40000, VarWD > 20000.
   denial:    151  hits:     40  cover:    191  index: 1.436619   frac. target: 0.209
4. AvgInc < 50000, VarWD > 20000.
   denial:    162  hits:     40  cover:    202  index: 1.305459   frac. target: 0.198
5. VarWD > 20000, Card = no.
   denial:    128  hits:     39  cover:    167  index: 1.673846   frac. target: 0.234
6. VarWD > 20000, Sex = female.
   denial:     94  hits:     29  cover:    123  index: 1.462862   frac. target: 0.236
7. Disp = no, #Inc >= 2, #Inc >= 3.
   denial:     60  hits:     22  cover:     82  index: 1.506826   frac. target: 0.268
8. AvgInc < 50000, #Inc >= 2, #Inc >= 3.
   denial:     59  hits:     20  cover:     79  index: 1.336361   frac. target: 0.253
9. AvgInc > 25000, Sex = female, Region = w_Bohemia.
   denial:     10  hits:      7  cover:     17  index: 1.313638   frac. target: 0.412
```

Rule 1 states that in the group of loan takers with medium or large variation in their monthly income (on that bank account), the fraction of bad loans is quite large (0.206% instead of 0.111%). Of the 76 bad loans, 53 (the "hits") belong to this group. A possible interpretion of that discovery is that varying income indicates low reliability of clients. Similarly, rule 2 indicates a high fraction of bad loans in the group of clients with medium or high variations in their monthly withdrawals that do not allow another person to dispose on their bank account. 44 of 76 bad loans belong to this group (Please note that the rules may and most probably do overlap.). Rule 3 indicates an increased fraction of bad loans for clients with medium or low variations in the average monthly income and medium or high variations in their average monthly withdrawals. Rules 4- 6 also show that the feature of medium or large variations in average monthly withdrawals in certain circumstances (no top monthly income (rule 4) or being female (rule 6)) indicates risky loans. Rules 7 and 8 may be interpreted such that getting one's monthly income distributed on 3 or 4 transactions in combination with certain other features hints on a risky loan. Rule 9 is surprising. It has only low coverage (17 cases) but the fraction of bad loans is very high here. Furthermore, it describes persons with medium or large monthly income whom one would expect to be good risks.

4.2 Good loans

A selection of discovered rules is shown below.

```
Minimum coverage:             0.150000 =  102 cases
ImpInt Threshold:             0.960000 => Index Threshold:   1.750690
Target class:                good loans
Extension of target class:     606         non-target class:  76
Apriori fraction of target class: 0.889
1. Disp = yes.
   denial:      0  hits:    145  cover:    145  index: 4.019746   frac. target: 1.000
```

```
   Card = yes.
   denial:      5  hits:    165  cover:    170  index: 3.203738   frac. target: 0.971
2. OType = household.
   denial:     20  hits:    421  cover:    441  index: 4.157294   frac. target: 0.955
3. VarInc < 40000, VarInc < 20000.
   denial:     23  hits:    402  cover:    425  index: 3.539816   frac. target: 0.946
4. VarWD < 40000, VarWD < 20000.
   denial:     32  hits:    405  cover:    437  index: 2.392807   frac. target: 0.927
5. #WD >= 2, #WD >= 4.
   denial:     22  hits:    351  cover:    373  index: 3.034818   frac. target: 0.941
6. OType = Unknown, Sex = male.
   denial:      5  hits:    109  cover:    114  index: 2.161417   frac. target: 0.956
7. AvgInc < 50000, AvgInc < 25000, OType = unknown.
   denial:      6  hits:    111  cover:    117  index: 1.949168   frac. target: 0.949
```

Rule 1 is very clear: all loans on accounts where a second person is allowed to dispose on are good loans. Rule 3 and 4 indicate that low variations in average monthly incomes and withdrawals resp. hint on a lower risk of bad loans. Rule 2 suggests permanent orders of type household as indicators of financial reliability, as do rules 6 and 7 with permanent orders of unknown type. Rule 6 seems especially interesting, because it describes persons with a low average income. Rule 5 indicates that distributing one's monthly withdrawals on more than 4 separate transactions indicates good loans. This rule is interesting when compared with rule 7 and 8 describing bad loans.

4.3 Credit card holders

With the given threshold settings, the search produced only few rules in the first four search levels. Rules 1- 4 indicate a high fraction of credit card holders among clients with medium or high average monthly income or withdrawals or medium or large variations of income or withdrawals. These rules as well as rule 5 (stating that card holders are rather frequent among clients born in or later than 1935) are probably not very surprising or novel. Rule 7 might give a hint on the personality of credit card holders. Rule 6 could be caused by the bank's policy of issuing credit cards. All these rules indicate potential new card holders, namely those contradicting a rule.

```
Minimum coverage:              0.080000  =  360 cases
ImpInt Threshold:              0.950000  => Index Threshold:  1.644850
Target class:                  has_card
Extension of target class:     892          non-target class: 3608
Apriori fraction of target class: 0.198
1. AvgInc > 25000.
   denial:    662  hits:    413  cover:   1075  index: 6.809351   frac. target: 0.384
2. AvgWD > 25000.
   denial:    519  hits:    346  cover:    865  index: 6.627571   frac. target: 0.400
3. VarWD > 20000.
   denial:    450  hits:    219  cover:    669  index: 3.730097   frac. target: 0.327
4. VarInc > 20000.
   denial:    507  hits:    200  cover:    707  index: 2.514070   frac. target: 0.283
5. YoB >= 35.
   denial:   2950  hits:    875  cover:   3825  index: 2.109115   frac. target: 0.229
6. Loan = good.
   denial:    441  hits:    165  cover:    606  index: 2.035933   frac. target: 0.272
7. Order = no.
   denial:    552  hits:    190  cover:    742  index: 1.759632   frac. target: 0.256
8. #Inc >= 2.
   denial:   2652  hits:    765  cover:   3417  index: 1.675036   frac. target: 0.224
```

4.4 Non-credit card holders

Although the threshold for implication intensity was rather strict, the algorithm produced a very large number of rules, many of which are more or less duals to rules discovered for credit card holders. A small selection is shown below. If novel, rules 1 - 4 are interesting because they may be operable, i.d., the bank could try to identify reasons for the low acceptance of credit cards in Moravia and improve its service accordingly.

```
Minimum coverage:              0.080000  =  360 cases
ImpInt Threshold:              0.990000  => Index Threshold:  2.326350
Target class:                   no_card
Extension of target class:     3608         non-target class: 892
Apriori fraction of target class: 0.802
   OType = household.
   denial:    587  hits:   2778  cover:   3365  index: 3.098259   frac. target: 0.826
   AvgInc < 50000, AvgInc < 25000.
   denial:    479  hits:   2946  cover:   3425  index: 7.672387   frac. target: 0.860
1. VarInc < 40000, Region = n_Moravia.
   denial:    120  hits:    630  cover:    750  index: 2.351097   frac. target: 0.840
2. AvgInc0 < 50000, Region = s_Moravia.
   denial:    121  hits:    643  cover:    764  index: 2.473701   frac. target: 0.842
3. YoB =< 76, Region = s_Moravia.
   denial:    110  hits:    593  cover:    703  index: 2.486323   frac. target: 0.844
4. YoB =< 76, Region = n_Moravia.
   denial:    111  hits:    595  cover:    706  index: 2.446771   frac. target: 0.843
```

5 Discussion and Conclusion

The main effort in this discovery task was to define useful attributes and to choose a good interestingness criterion. The discovery algorithm has discovered many rules, some which (in my opinion) show potential to help distinguish good (reliable) and bad clients, or indicate opportunities to improve the bank's services. In the experiments, only a selection of possibly useful attributes was investigated. Further experiments with other attributes might lead to additional interesting results. A final evaluation of the usefulness and interestingness of the discovered rules requires the knowledge of domain experts. A final evaluation of the usefulness and interestingness of the discovered rules requires the knowledge of domain experts.

References

[Fleu95] L. Fleury, C. Djeraba, J. Philippe, H. Briand. Contribution of the implication intensity in rules evaluations for Knowledge Discovery in Databases. ECML95 Workshop Notes Statistics, Machine Learning, and Knowledge Discovery in Databases , Heraklion, Greece, 1995.

[Web98] I. Weber. On pruning strategies for discovery of generalized and quantitative association rules. In L. Bing, W. Hsu, and W. Ke, editors, *Proc. of Knowledge Discovery and Data Mining Workshop, (Pricai'98)*, 1998.

[Web99] I. Weber. A declarative language bias for levelwise search of first-order regularities. In *Proc. ISMIS'99*, number 1609 in Lecture Notes in Artificial Intelligence. Springer-Verlag, 1999.

[Wro97] Stefan Wrobel. An algorithm for multi-relational discovery of subgroups. In J. Komorowsky and J. Zytkow, editors, *Proc. First European Symposium on Principles of Knowledge Discovery and Data Mining*, 1997.