COPING WITH DISCOVERY CHALLENGE BY GUHA

David Coufal, Martin Holeňa, Anna Sochorová
Institute of Computer Science of the Academy of Sciences of the Czech Republic
Pod Vodárenskou věží 2, 182 07 Praha 8, Czech Republic

Introduction

This paper describes a solution of discovery challenge for financial data set, announced as part of 3rd European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'99), held in Prague, September 15-18, 1999. The structure of the paper is following: section 1 gives short introduction to GUHA method that forms theoretical base of our approach to challenge solution. Relevant references on literature with full description of the method are included. Section 2 describes preprocessing of original data set for data mining stage of knowledge discovery, section 3 presents discovered knowledge and section 4 concludes the paper. It has to be strongly emphasized that to full understanding of the paper, reader has to be familiar with the challenge task definition.

1. Principles of GUHA method

Basic ideas of GUHA (General Unary Hypotheses Automaton) method were given in [1] already in 1966. Starting notion of the method is an object. Object has properties expressed by variables given on this object. For example object can be a man with properties given by variables sex, age, color of eyes, etc. In order to make reasonable knowledge discovery we need to have set of the objects of the same kind which differ in values of variables defined on them.

The aim of GUHA method is to generate hypotheses on relations among properties of the objects which are in some sense interesting. This generation is processed systematically; the machine generates in some sense all possible hypotheses and collects the interesting ones. The hypothesis is generally composed from two parts; from an antecedent and a succedent. The antecedent and succedent are tied together by so called generalized quantifier, which describes the relation between them. The antecedents and succedents are propositions on the object in sense of the classical propositional logic, so they are true or false on particular object. These propositions can be simple or compound similarly as in propositional logic. Compound propositions are usually composed by conjunction connective. Formulation of these propositions is enabled through original variable categorization. Given an antecedent and a succedent, frequencies of four logically possible combinations can be computed and expressed in compressed form as so-called four-fold table (ff-table). General ff-table looks like this:

ff-table	succedent	non(succedent)
antecedent	*a*	*b*
non(antecedent)	*c*	*d*

Where *a* is the number of the objets satisfying antecedent and succedent, *b* is the number of the objects satisfying antecedent but not satisfying succedent, etc.

A generalized quantifier is a decision procedure assigning 1 or 0 to each ff-table. If the value is 1 then we accept hypothesis with this ff-table, if it is 0 then we do not accept it. The basic generalized quantifier defined and used in GUHA is given by Fisher exact test known from mathematical statistics. In this case decision for accepting is made on the base of a statistical test on some level alpha. For each hypothesis, value of Fisher statistic given by values *a, b, c,* and *d* of ff-table is computed. Its value, simply said, describes measure of association between the antecedent and succedent. More precisely, value of Fisher statistic refers to level on which we can accept hypothesis (in statistical sense) that conditional probability of the succedent on condition of the antecedent P(S/A) is greater than single probability of the succedent P(S). In other words that the presence of antecedent increase probability of the presence of the succedent. Accepting or not accepting of tested hypothesis is given by value of Fisher statistic in comparison with value of alpha. Note that Fisher statistic have following symmetry property, the level for accepting P(S/A)>P(S) is the same that the level for accepting P(A/S)>P(A). Further property of Fisher statistic is negation symmetry property: Fisher(A,S)=Fisher(non(A),non(S)). Several other quantifiers are used in GUHA.

Notice that above presented level alpha of the statistical test is local significance level justified for a test of single hypothesis, saying nothing about the simultaneous validity of the whole set of hypotheses. This limitation is very common in data mining [2]. GUHA method does not suffer from this disadvantage since the statistical component of the underlying theory entails the possibility to make use of methods for multiple hypotheses testing and to obtain sets of simultaneously valid hypotheses. Existing methods for multiple hypotheses testing differ with respect to how the validity of some initial assumption is ensured. In our computations, the following three methods have been used: Bonferroni method, Holm method and Simes method [3, 4, 5].

Mathematical exact basis for GUHA method can be found in book [6].

The theoretical principles of GUHA method are implemented in GUHA+- software package [7], which enables effective realization of the process of generating hypotheses and their testing and thus exploring of knowledge about given data. For a previously implementation see [8].

2. Data preprocessing

As stated in the task definition, bank data are given by eight tables. Each table consists of several variables characterizing a property that can be associated with an account. Hence the object of the investigation is the account with its properties. The account has static and dynamic characteristic. Variables of static characteristic are given by tables account, client, disposition, permanent order, loan, credit card and demographic data, variables of dynamic characteristic are given by table transaction.

In the preliminary investigation of the original tables we have explored the fact that in data each client is associated only with one account and there are at most two clients, which can manipulate with one account. If it is so then one from the clients has owner right to the account, second has user right to the account and they live in the same district. If there is only one client who can manipulate with the account then he has owner right. Other exploration. There is at most one card issued to one account. There are cases of different district address of the bank where account is held and district address of the owner of this account.

Within the preprocessing stage, new variables were computed on base of original data and each variable was categorized into several categories union of them covering the whole range of the variable. Following lists contain all variables used in data mining stage by GUHA +- software. Variables are given together with parent variable(s) from original tables on base of them they were computed. Parent variable(s) is(are) given in brackets together with letter coinciding with first letter of the name of the original table. First list consists of variables computed on base of variables from tables describing static characteristic of the account:

- **acc_year** [date | a]; **acc_freq** [frequency | a]; **acc_dist%** [district_id | a]
- **owner_sex** [birth_number | c]; **owner_age** [birth_number | c]; **owner_dist%** [district_id | c]
- **user** [type | d]
- **order_sum** [amount | p]; **order_other** [amount, k_symbol | p];
  order_insurance [amount, k_symbol | p] **order_household** [amount, k_symbol | p];
  order_leasing [amount, k_symbol | p]; **order_loan** [amount, k_symbol | p]
- **loan_status** [status | l]; **loan_amount** [amount | l]; **loan_duration** [duration | l];
  loan_pay [payments | l]; **loan_year** [date | l]
- **card_type** [type | c]; **card_year** [issued | c]

Most of the names of the variables are self-explanatory. In case of the **acc_dist%**, this shortcut represents 15 variables with names **acc_dist_id**, **acc_dist_region**, **acc_dist_a4**, … , **acc_dist_a16** given by original table demographic data **- dist_id** [A1 | d]; **dist_region** [A3 | d]; **dist_a4** [A4 | d]; **dist_a5** [A5 | d]; … ; **dist_a16** [A16 | d]. Similarly for **owner_dist%** that represents variables **owner_dist_id**, **owner_dist_region**, **owner_dist_a4**, … , **owner_dist_a16**. Variable **user** marks if here is another client who can manipulate with the account besides the owner. Variable **order_sum** represents total money sum of all permanent orders issued on the account. It means that 17+17+1+6+5+2=48 variables describing static characteristic of account were defined.

Variables describing dynamic characteristic of the account are given in the table transaction. Preprocessing of data from this table was done in two phases. First phase yielded computing new variables described below.

- **amount_sign** [type, amount | t]: variable's value is signed cash flow given by transaction,
  amount_sign = (-1)·amount, if type="VYDEJ" else **amount_sign** = amount
- **volume** [type, amount | t]: variable's value gives turnover on account given by transaction,
  thus **volume** = abs(**amount_sign**)
- **balance** [balance | t]: variable's value is the same as value of original variable balance given in table transaction
- **withdrawal_card** [operation, amount | t]: **withdrawal_card** = amount if operation = "VYBER KARTOU" else **withdrawal_card** = 0
- **credit_cash, collection, withdrawal_cash, remittance**: variables were computed in the similar manner as variable **withdrawal_card**

- **insurance** [k_symbol, amount | t]: **insurance** = amount if k_symbol = "POJISTNE" else **insurance** = 0
- **statement, credited_interest, sanction_interest, household, pension, loan**: variables were computed in the similar manner as variable **insurance**

In second phase, values of these new variables except the variable **balance** were summed within one month. On base of variable **balance** value of new variable **Ibalance** for each month was computed using the following equation:

$$\mathbf{Ibalance} = \frac{1}{30} \sum_i balance_i \cdot duration\_of\_the\_balance_i\_in\_days$$

.

Where *i* denotes particular month's day if some transaction was realized in this day. Meaning of the variable **Ibalance** is to characterize of monthly cash flow on account in terms of interest given to client by bank. If there is interest rate *k*% p.a. given by bank, then monthly interest is given by **Ibalance**·(1+0.01*k*)/12. In the end, monthly values were averaged through number of month. For variable **amount_sign, volume** and **Ibalance** minimum and maximum were found too. Final list of variables used in data mining stage characterizing dynamics of the account is:

- **AvgM_amount_sign, MinM_amount_sign, MaxM_amount_sign**
- **AvgM_volume, MinM_volume, MaxM_volume**
- **AvgM_Ibalance, MinM_Ibalance, Max_Ibalance**
- **AvgM_withdrawal_card, AvgM_credit_cash, AvgM_collection, AvgM_withdrawal_cash, AvgM_remittance**
- **AvgM_insurance, AvgM_statement, AvgM_credited_interest, AvgM_sanction_interest, AvgM_household, AvgM_pension, AvgM_loan**
- **AvgM_transaction_#**

Variable **AvgM_transaction_#** refers to averaged monthly number of realized transactions on the account.

3+3+3+5+6+1=21 variables describing dynamic of the account were defined. Hence total number of variables describing the account used in data minig stage was 69.

As it was stated in beginning of this paragraph each variable was categorized into several categories at a medium into 5 for each variable. Because of limitation of size of the paper, not all categories are described here, only categories relevant to results of data mining stage, i.e. that appear in hypotheses explored by data mining.

Discovered knowledge

Because we are not bank's experts, our definition of good or bad client comes out from natural criterions. We have this basic point of view on this definition; if a loan for client is granted then client is good if there were/are no problems with loan payments, and clearly client is bad when there were/are problems with loan payment. In our discovery work we concentrated on exploring appropriate characterizations of this notion of good or bad client. The consequence is that we aimed only on data of clients with granted loan.

We assume that eminent interest of every bank is to reveal in advance if client asking for the loan is good or bad according to our definition. This prediction has to be based on information the bank knows in time of asking for the loan. That is why we carried out our exploration from data (namely data from table transaction) that were older than date when the loan was granted.

In first phase we restricted ourselves on single antecedents hypotheses and exploration was divided into two subcases: subcase 1a - characterization of the good client and subcase 1b characterization of the bad client. Before we will give the explored hypotheses we have to explain categories **loan_status**__ok and **loan_status**__bad appearing in succedent of these hypotheses. Category **loan_status**__ok is satisfied if loan status given in original table loan was marked as A (contract finished no problems) or C (running contact, OK so far), else it is not satisfied. Category **loan_status**__bad is only satisfied if loan was marked as B (contract finished, loan not paid) or D (running contract, client in debt).

subcase 1a - good client: by proceeding preprocessed data by GUHA+- software we have explored hypotheses summarized in table 1. We have used Fisher quantifier with the significance level alpha=0.05 and restricted ourselves only on hypotheses supported at least by 15 objects (*a* value in ff-table). Explored hypotheses are valid globally in sense of Simes method on significance level 0.05. In table presented statistic prob is given by fraction a/(a+b) values of ff-table, and characterizes hypotheses in sense of a implication

Table 1

#	antecedent	Succedent	ff-table		Fisher	Prob
1	**AvgM_sanction_interest**__no	**loan_status**__ok	603	50	6.12144e-024	0.9234
			3	26		
2	**order_household**__yes	**loan_status**__ok	421	20	5.0375e-013	0.9546
			185	56		
3	**AvgM_remittance**__yes	**loan_status**__ok	375	16	9.08928e-012	0.9591
			231	60		
4	**user**__yes	**loan_status**__ok	145	0	3.75844e-009	1.0000
			461	76		
5	**MinM_lbalance**__positive	**loan_status**__ok	606	70	1.59916e-006	0.8965
			0	6		
6	**AvgM_amount_sign**__positive	**loan_status**__ok	606	70	1.59916e-006	0.8965
			0	6		
7	**loan_year**__**98**	**loan_status**__ok	154	4	1.11229e-005	0.9747
			452	72		
8	**card_type**__card_yes	**loan_status**__ok	165	5	1.38617e-005	0.9706
			441	71		
9	**AvgM_lbalance**__>40000	**loan_status**__ok	296	19	4.947e-005	0.9397
			310	57		
10	**MinM_amount_sign**__-20000-0	**loan_status**__ok	256	16	0.000193676	0.9412
			350	60		
11	**loan_pay**__to2000	**loan_status**__ok	125	4	0.000330908	0.9690
			4	72		
12	**AvgM_household**__yes	**loan_status**__ok	377	32	0.000652951	0.9218
			229	44		
13	**MaxM_amount_sign**__20000-40000	**loan_status**__ok	294	22	0.000819922	0.9304
			312	54		
14	**AvgM_withdrawal_cash**__yes	**loan_status**__ok	606	73	0.00133557	0.8925
			0	3		
15	**AvgM_statement**__yes	**loan_status**__ok	558	61	0.00201398	0.9014
			48	15		
16	**acc_year**__**97**	**loan_status**__ok	117	5	0.00264987	0.9590
			489	71		
17	**AvgM_collection**__yes	**loan_status**__ok	197	13	0.00332071	0.9381
			409	63		
18	**order_sum**__>10000	**loan_status**__ok	241	18	0.00388401	0.9305
			365	58		
19	**owner_dist_region**__**5**	**loan_status**__ok	61	1	0.0043812	0.9839
			545	75		
20	**acc_dist_region**__**5**	**loan_status**__ok	60	1	0.00491192	0.9837
			546	75		

Hypotheses are sorted increasingly according to value of Fisher statistic. Hypotheses can be divided into several groups according to properties of the account they are regarding.

One group of hypotheses regards to absence or presence of some kind of operations on the account – hypotheses #1, #3, #12, #14, #15 and #17. Hypothesis #1 says that there is strong association between absence of sanction interest payment and good loan payments. Similarly for hypotheses #3, #12, #14, #15 and #17 good loan payment is associated with presence of any transaction of the type given by antecedents of these hypotheses.

Hypotheses #2 and #18 are connected with the permanent orders issued on the account - #2 associates good loan payment with presence of household permanent order. Hypothesis #18 says that there are no problems with payment if total money sum of permanent orders is greater than 10000 – solvent clients.

Hypotheses #5, #6, #9, #10 and #13 characterize cash flow on the account and can be interpreted in following manner, loan payment is without problems if #5: “balance” on the account is always positive – client was not in debt; #6: average monthly increment on account is positive – so amount of the money on the account is increasing; #9: similarly as #5, average “balance” on the account is high; #10: maybe surprising hypothesis can be explained as saying that there was investment with money from the account, but this investment was not too big because there was also category **MinM_amount_sign**__<-20000 defined; #13: maximal increase on the account within one month was between 20000 and 40000.

Possibly interpretation of hypotheses #7 and #16 is that usually if there are problems with loan payments then these problems are not in the beginning of the duration of the loan. Thus if it is year 1999 now, loans granted in year 1998 and 1997 can be without problems from this reason.

In sense of the implication strong hypothesis #4 says that if there is another client who can manipulate with account then (in data) loan payment is always without problem. Hypothesis #8 characterizes clients with modern approach to goods or services payment – they use credit card. Hypothesis #11 says that there are no problems if there are small loan payments – granted amount of money is small. In the end, hypotheses #19 and #20 give association between client district address and good loan payment. Reason for simultaneous presence of hypotheses #19 and #20 is that the most of client's district addresses are the same as the district addresses of the bank where the account is held.

subcase 1b - bad client: because of symmetry negation property of Fisher quantifier, characterization of account with **loan_status**__bad property is tightly connected with hypotheses from table 1. This characterization can be formulated as, if the account do not satisfy some property given by antecedents from table 1 there is increasing possibility of bad loan payments. On base of table 1 we can derive table 2 of the hypotheses with the antecedents given by the negations of antecedents of hypotheses from table 1.

Table 2

#	Antecedent	succedent	ff-table		Fisher	Prob
1	**AvgM_sanction_interest**__yes	**loan_status**__bad	26	3	6.12144e-024	0.8966
			50	603		
2	**order_household**__no	**loan_status**__bad	56	185	5.0375e-013	0.2324
			20	421		
3	**AvgM_remittance**__no	**loan_status**__bad	60	231	9.08928e-012	0.2062
			16	375		

In second phase of our exploration we aimed on hypotheses with compound antecedents of the length 2. We used Fisher quantifier with significance level alpha=0.001, again with restriction on hypotheses supported at least by 15 objects; Bonfferoni method gave 300 (Simes 758) simultaneously valid hypotheses. So we applied other restriction to choose only 100% implication hypotheses.

Table 3

#	Antecedent	succedent	ff-table		Fisher	prob
1	**order_other**_no **order_household**_yes	**loan_status**__ok	208	0	1.30383e-013	1
			398	76		
2	**order_other**_no **AvgM_remittance**_yes	**loan_status**__ok	179	0	1.75133e-011	1
			427	76		
3	**acc_dist_a7**__>3 **order_household**__>5000	**loan_status**__ok	139	0	9.3298e-009	1
			467	76		
4	**owner_dist_a13**__>3 **AvgM_sanction_interest**__yes	**loan_status**__bad	21	0	6.27134e-022	1
			55	606		
5	**owner_dist_a7**__to6 **AvgM_sanction_interest**__yes	**loan_status**__bad	20	0	7.41362e-021	1
			56	606		
6	**acc_dist_a7**__to6 **AvgM_sanction_interest**__yes	**loan_status**__bad	20	0	7.41362e-021	1
			56	606		
7	**AvgM_credited_interest**__to200 **AvgM_sanction_interest**__yes	**loan_status**__bad	19	0	8.62321e-020	1
			57	606		
8	**owner_dist_a6**__to27 **AvgM_sanction_interest**__yes	**loan_status**__bad	17	0	1.1127e-017	1
			59	606		
9	**owner_dist_a8**__exactly1 **AvgM_sanction_nterest**__yes	**loan_status**__bad	15	0	1.35051e-015	1
			61	606		
10	**owner_dist_a12**__>3 **AvgM_sanction_interest**__yes	**loan_status**__bad	15	0	1.35051e-015	1
			61	606		

These hypotheses are 100% implications as it can be seen from the four fold tables. In antecedents of hypotheses #1, #2 and #3 there are combinations of other properties of the accounts that (in data) gives 100% warranty that loan will be pay without problems. Category **order_other**_no says that there is no other kind of permanent order on the account besides for insurance, household, leasing or loan payment. Category **order_ household__**>5000 clearly denotes clients with household permanent order payment greater than 5000. Category **acc_dist_a7__**>3 says that bank address is in the district with more than 3 municipalities with number of inhabitants from 2000 to 10000.

Form of antecedents for **loan_status__**bad succedent reveals that critical property for bad loan payment is presence of sanction interest on the account, but even in this case (see ff-table of hypothesis #1 in table 2) the loan could be pay ok. However, if there is sanction interest on account together with a property from set given by first members of antecedents of hypotheses in table 3, there were always problems with loan payment. Explanation of categories: **owner_dist_a13__**>3 – unemployment rate in year 1996 was greater than 3% in district where client lives; **owner_dist_a7__**to6 - client address is in district with maximally 6 municipalities with number of inhabitants from 2000 to 10000; **acc_dist_a7__**to6 – see preceding category; **AvgM_credited_interest__**to200 – credited interest to account is maximally 200, because there was defined also category **AvgM_credited_interest__**>200, category **AvgM_credited_interest__**to200 denotes account with small credited interest; **owner_dist_a6__**to27 **-** address of the client is in district with maximally 27 municipalities with number of inhabitants from 500 to 2000; **owner_dist_a8__**exactly1 **-** address of the client is in the district where exactly one municipality with number of inhabitants greater than 10000 is; **owner_dist_a12__**>3 – client lives in district where unemployment rate in year 1995 was greater than 3%.

4. Conclusion

We have presented 50 hypotheses discovered by GUHA; each can be interpreted as saying: *the antecedent increases the probability of the succedent.* Besides, hypotheses from tables 1 and 3 give highly true implications and would admire the interpretation as saying: *the probability of the succedent conditioned by the antecedent is high.*

According to the theory of global interpretation of multiple hypotheses testing (see [6] chapter 8) the global significance of our results was considered. From this point of view the results as a whole can be interpreted as sufficiently reliable knowledge on the universe from which the data form a random sample.

References

[1] Chytil M., Hájek P., Havel I. *The GUHA method of automated hypotheses generation*, Computing, 293-308, 1966

[2] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors. Advances in Knowledge Discovery and Data Mining, AAAI Press, Menlo Park, 1996

[3] Y. Hochberg, and A.C. Tamhane. *Multiple Comparison Procedures*. John Wiley and Sons, New York, 1987.

[4] E. Samuel-Cahn. *Is the Simes improved Bonferroni procedure conservative?* Biometrika, 83:928–933, 1996.

[5] P.H. Westfall. *Multiple Testing of General Contrasts Using Logical Constraints and Correlations*. Journal of the American Statistical Association, 92:299–306, 1997.

[6] Hájek P., Havránek T. Mechanizing Hypothesis Formation. *Mathematical Foundations for a General Theory.* Springer Verlag, Berlin – Heidelberg – New York, 1978

[7] Honzíková Z. *GUHA +- User’s guide*, manual for GUHA+- software package, 1999

[8] Hájek P., Sochorová A., Zvárová J. *GUHA for personal computers*. Computational Statistics and Data Analysis, 19, 149-153, 1995