On the complementarity of OLAP and rich associations mining

David Chudán, Vojtěch Svátek
Faculty of Informatics and Statistics, University of Economics, Prague
Nám W. Churchilla 4, 130 67 Prague, Czech Republic
e-mail: xchud01@vse.cz, svatek@vse.cz

ABSTRACT

The paper presents a comparison of possibilities and proposal for complementary use of OLAP and the rich variant of association rule mining based on the GUHA method. The rationale is to determine the point when it is useful for the analyst to proceed from OLAP to descriptive data mining, as well as the point of return from data mining results to OLAP in order to see them in a broader view.

1 INTRODUCTION

In large datasets (or warehouse environments) it is possible to use many types of analysis. OLAP analysis and data mining (DM) are the most frequent. They are based on completely different techniques and algorithms and often applied on different kinds of analytical problems; yet they are often part of a single business intelligence (BI) solution, see Fig. 1. With a certain simplification we can say that the regular use of OLAP analysis can solve analytical questions like "what happened last year with the profit in the Northern region", while DM is trying to answer the question "why it happened". The task of (especially, the predictive type of) DM is very often complementary to OLAP analysis, when OLAP only pinpoints some problem while DM can provide an insight into the problem and estimate its reason, which may eventually lead to finding a solution to the problem.

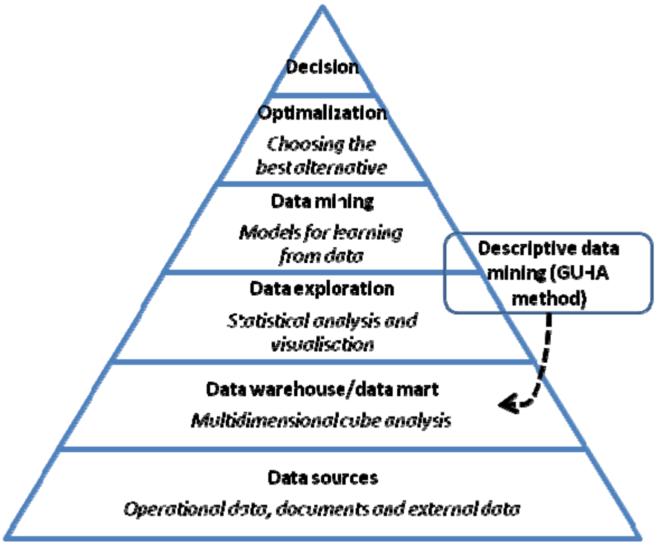

Decision
Optimalization
Choosing the best alternative
Data mining
Models for learning from data
Data exploration
Statistical analysis and visualisation
Data warehouse/data mart
Multidimensional cube analysis
Data sources
Operational data, documents and external data
Descriptive data mining (GUHA method)

Figure 1 *– The main components of BI [1] with the loop from data mining back to OLAP analysis*

In this paper we however advocate a somewhat different scenario, suitable for descriptive DM such as the GUHA method detailed in Section 3. Descriptive DM represents a transitional analysis type, which exhibits higher representational power than statistical analysis while preserving its exploratory nature (in contrast to predictive DM model building). The crucial novelty of the scenario can be outlined as follows. Let us imagine that we found some interesting rules using DM and we would like to examine them in a broader perspective. Is it possible and helpful at this point to switch back to OLAP analysis and see the surroundings of the rules (the loop in *Figure 1*)?
Sections 2, 3 and 4 briefly present and verbally compare OLAP and the DM method, Section 5 explains the data set and task, section 6 shows the course of complementary analysis and finally, Section 7 wraps up the paper.

2 OLAP (MULTIDIMENSIONAL) ANALYSIS

The objective of OLAP (multidimensional) analysis is to gain insight into the meaning contained in databases [2]. OLAP analysis is based on the OLAP cube, a data structure that overcomes several limitations of two-dimensional relational databases. (Despite the word "cube", there is no limitation of three dimensions. The number of dimensions can be tens and depends on the used BI platform.)
The basic operations in OLAP analysis are [2]:

- **Drill-down/up** – Analytical technique that enables to navigate among levels of data granularity from the most summarized to the most detailed.
- **Slice/dice –** A slice is a subset of a multi-dimensional array corresponding to a single value for one or more members of the dimensions not in the subset. Slicing enables to see different slices of information presented in the OLAP cube. Dice means to slice a data cube in more than one dimension.
- **Pivot (rotate) –** Pivoting means the changing of the dimensional orientation of the data view.

The examples of OLAP used throughout this paper were provided using the Pentaho BI suite. It is an open source suite that integrates ETL (extract, transform, load), dashboard, reporting, workflow and data mining capabilities. Due to the Open Source Business Intelligence and Reporting survey [3], open source solutions are more preferable than commercial ones, and Pentaho is the open source leader with market share of 30%.

3 RICH ASSOCIATION RULES MINING

Association rules are one of the data mining techniques (sixth most used DM method[1]) used to discover interesting relations between variables in large datasets. Although the idea is much older, it has been widely popularized in 90s with market basket analysis where the aim of is to determine which items in a supermarket are bought together [4].

In the context of this paper we however do not consider traditional associations but rich associations, in particular those produced by the LISp-Miner data mining software [5] with its underlying GUHA method [6], the original Czech method of exploration analysis. GUHA provably overtakes the market basket analysis [4] and also the a-priori algorithm [7] in terms of representational power, while its recent implementations guarantee high throughput. The rich representational features of GUHA, such as (runtime-generated) value groupings in rule literals, make it particularly suitable for alignment with OLAP analysis.[2]

4 MAJOR DIFFERENCES

Both described methods, OLAP analysis and rich association rules mining, can be used to analyze the same kind of tabular data. The main difference lies in the level of model granularity and the degree of automation. The GUHA attributes are not a priori labeled for a role in the analysis (antecedent / consequent) as the dimensions / measures in OLAP. In OLAP analysis the (numerical) *measure* fields are displayed in terms of aggregations such as sums or averages; the internal structure of *dimensions* is not altered during the analysis. GUHA, in contrast, relies on pre-processing of numerical data into *discrete* categories, and these can be further amalgamated at *runtime*.

Furthermore, OLAP analysis is carried out *manually*, as the analyst "browses" through data and inspects the view of data s/he currently needs. GUHA, on the other hand, generates many individual hypotheses in one shot [8], even if some of them are very similar to others.

5 DATASET AND TASK

The dataset used in this demonstrative case study is an updated version of the Financial Data Set first introduced in the PKDD'99 Discovery Challenge [9]. The Financial Dataset consists of 8 tables describing the operations of bank customers. For this task only one table is used, the Loans table, which contains the following columns: *loan_id, birth number, district, salary, amount, payments, duration* and *status.*

The task is to identify, between 6181 customers available in the dataset, subgroups with high occurrences of bad loans. From the available columns, for better clarity of the analysis and inserted figures only four attributes are used: *district* (the location where the client lives), *amount* (overall amount of the loan), *duration* (the time period for which the loan is granted) and *status***.** Status is the key indicator of the bad loans and potentially unreliable clients. Status has four possible values, A, B, C and D [9].

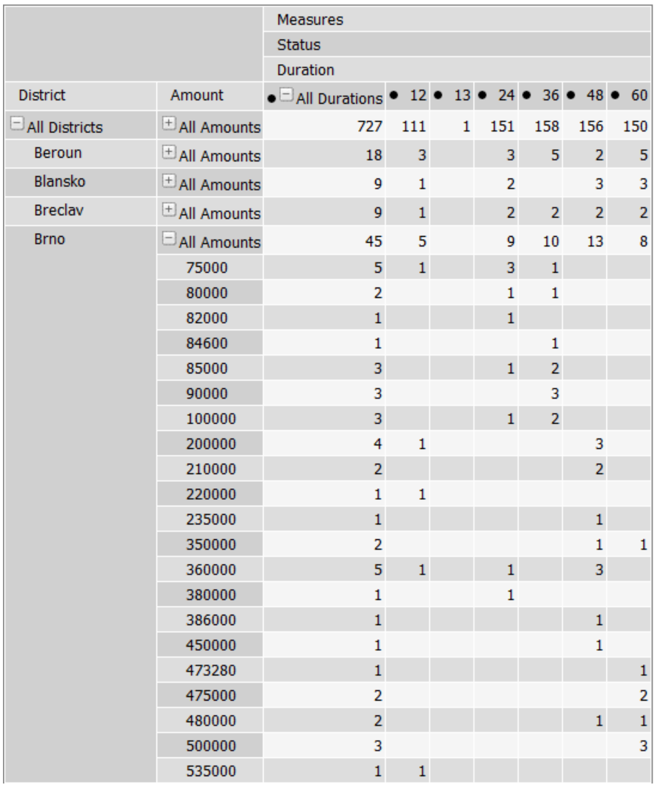
- **A** stands for finished contracts with no problem.
- **B** stands for finished contracts with loan not payed.
- **C** stands for running contracts that are OK so far
- **D** stands for running contracts, where the client is in debt and it is probable than s/he will have problems with paying the loan

The task considered here is to describe groups of clients that tend to belong to groups B and D.

6 SAMPLE COURSE OF ANALYSIS

6.1 OLAP analysis

OLAP analysis is easy to use, as it is very similar to contingency tables known from spreadsheet processors. We choose individual columns from the dataset, thus creating the dimensions of the cube, the data element that categorizes each item into non-overlapping regions. Let the dimensions in this analysis be *district*, *amount* and *duration*. The measure can generally be a variety of key performance indicators in business environment, e.g. days, amount of money or more complex ratios as 'cost per person per week'. In our analysis let the measure simply be the count of status, more precisely the count of status that indicates bad loan quality (status B and D). Fig. 2 shows the absolute numbers of aggregated status B and D drilled through all dimensions (for the example of Brno district).



		Measures						
		Status						
		Duration						
District	Amount	● ⊟ All Durations	● 12	● 13	● 24	● 36	● 48	● 60
⊟ All Districts	⊞ All Amounts	727	111	1	151	158	156	150
Beroun	⊞ All Amounts	18	3		3	5	2	5
Blansko	⊞ All Amounts	9	1		2		3	3
Breclav	⊞ All Amounts	9	1		2	2	2	2
Brno	⊟ All Amounts	45	5		9	10	13	8
	75000	5	1		3	1		
	80000	2			1	1		
	82000	1			1			
	84600	1				1		
	85000	3			1	2		
	90000	3				3		
	100000	3			1	2		
	200000	4	1				3	
	210000	2					2	
	220000	1	1					
	235000	1					1	
	350000	2					1	1
	360000	5	1		1		3	
	380000	1			1			
	386000	1					1	
	450000	1					1	
	473280	1						1
	475000	2						2
	480000	2					1	1
	500000	3						3
	535000	1	1					

Figure 2 *- Analysis view in Pentaho BI suite*

[1] http://www.kdnuggets.com/polls/2007/data_mining_methods.htm

[2] We however do not show these features here for simplicity.

As we can see, three dimensions are on the border of clarity, and with further dimensions the table would be rather confusing.
Based on this view the analyst can conclude that no matter of the amount of the loan, only few clients from Brno region have problem with paying short-term loans (for one year). But finding such rules manually is very time consuming, as the analyst has to drill down and roll up through the individual districts and individual amounts.
For more dimensions this task becomes impossible, hence more complex relationships are beyond the reach of OLAP.

6.2 Associations mining with high data granularity

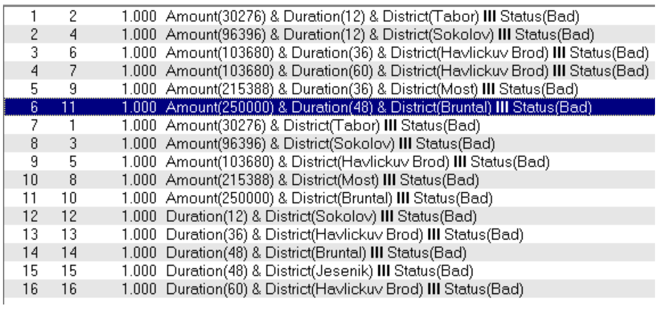
In the DM process and important task is to prepare the data for analysis via data pre-processing. The task of data pre-processing usually consists in grouping data into subsets with common characteristics, and in 'binning' numerical values into intervals. In the first DM task shown here only basic pre-processing was done, which consists in creating a group 'good loan quality' (A and C) and 'bad loan quality' (B and D) for the *Status* attribute.

The task is formulated as a template rule with three elements in antecedent, such that every antecedent attribute can be valued by a subset of its domain of size exactly (i.e. minimum as well as maximum) 1:

District(subset 1-1) & *Duration(subset 1 - 1)* &
& *Amount(subset 1-1)* => *Quality (Bad)*

The interest measures (confidence and support) thresholds were set as follows: *minConf=0.9* and *support=10.* Detailed explanation of the DM setting in LISp-Miner is out of the scope of this paper, details can be found at [10].

The DM tool finds 16 rules in data, all of which have *Conf=1* (see Fig. 3).

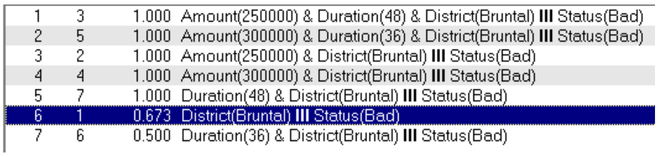


1	2	1.000	Amount(30276) & Duration(12) & District(Tabor) **III** Status(Bad)
2	4	1.000	Amount(96396) & Duration(12) & District(Sokolov) **III** Status(Bad)
3	6	1.000	Amount(103680) & Duration(36) & District(Havlickuv Brod) **III** Status(Bad)
4	7	1.000	Amount(103680) & Duration(60) & District(Havlickuv Brod) **III** Status(Bad)
5	9	1.000	Amount(215388) & Duration(36) & District(Most) **III** Status(Bad)
6	11	1.000	Amount(250000) & Duration(48) & District(Bruntal) **III** Status(Bad)
7	1	1.000	Amount(30276) & District(Tabor) **III** Status(Bad)
8	3	1.000	Amount(96396) & District(Sokolov) **III** Status(Bad)
9	5	1.000	Amount(103680) & District(Havlickuv Brod) **III** Status(Bad)
10	8	1.000	Amount(215388) & District(Most) **III** Status(Bad)
11	10	1.000	Amount(250000) & District(Bruntal) **III** Status(Bad)
12	12	1.000	Duration(12) & District(Sokolov) **III** Status(Bad)
13	13	1.000	Duration(36) & District(Havlickuv Brod) **III** Status(Bad)
14	14	1.000	Duration(48) & District(Bruntal) **III** Status(Bad)
15	15	1.000	Duration(48) & District(Jesenik) **III** Status(Bad)
16	16	1.000	Duration(60) & District(Havlickuv Brod) **III** Status(Bad)

Figure 3 *–The results in LISp-Miner system*

These rules describe 16 groups of clients with bad loans quality. From the list of rules we can easily identify problematical districts (Sokolov, Havlickuv Brod, Bruntal etc.) and problematical amounts of loan. However, the bank analyst can, for example, ask the question (referring to the highlighted rule in Fig. 3): *Is it really dangerous for us to provide a loan to a client from Bruntal or is it only bound to a certain amount of money or to certain duration?*

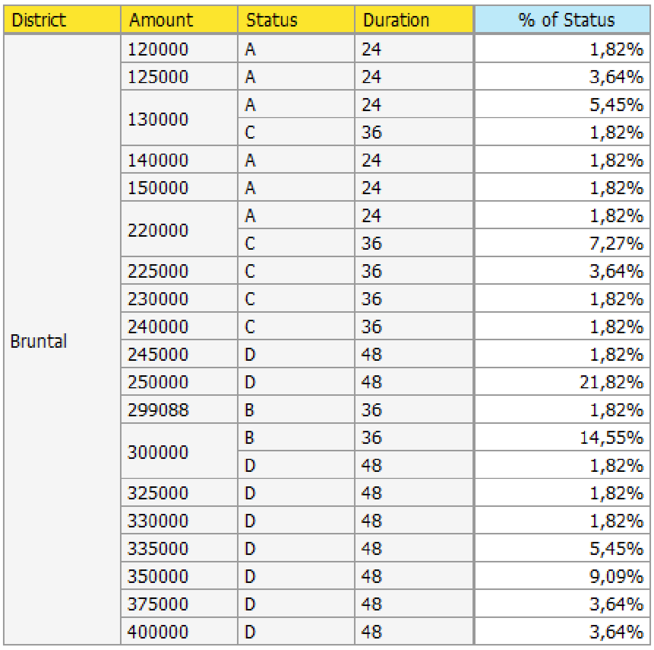
We can resolve the question in several ways. One option is to change the confidence and support thresholds in the DM task itself. Another option is to exclude all districts except Bruntal (cf. Fig. 4) in the task (but even in this case it is necessary to change the thresholds, to *minConf=0.5* and *support=7*). Now we can see that the probability that a client from Bruntal will have problems with paying his/her loan is 67,3% (see the highlighted rule in Fig. 4).



1	3	1.000	Amount(250000) & Duration(48) & District(Bruntal) **III** Status(Bad)
2	5	1.000	Amount(300000) & Duration(36) & District(Bruntal) **III** Status(Bad)
3	2	1.000	Amount(250000) & District(Bruntal) **III** Status(Bad)
4	4	1.000	Amount(300000) & District(Bruntal) **III** Status(Bad)
5	7	1.000	Duration(48) & District(Bruntal) **III** Status(Bad)
6	1	0.673	District(Bruntal) **III** Status(Bad)
7	6	0.500	Duration(36) & District(Bruntal) **III** Status(Bad)

Figure 4 *- Results of data mining task, all disctricts except Bruntal excluded*

However, another way is to go back to OLAP analysis, drill down to district Bruntal, and display all variants of *amount* and *duration.* As we can see in Fig. 6, [3] the bad status in not related to lower amounts of the loan. So, these clients, based on this dataset, are credible for the loan.



District	Amount	Status	Duration	% of Status
Bruntal	120000	A	24	1,82%
	125000	A	24	3,64%
	130000	A	24	5,45%
		C	36	1,82%
	140000	A	24	1,82%
	150000	A	24	1,82%
	220000	A	24	1,82%
		C	36	7,27%
	225000	C	36	3,64%
	230000	C	36	1,82%
	240000	C	36	1,82%
	245000	D	48	1,82%
	250000	D	48	21,82%
	299088	B	36	1,82%
	300000	B	36	14,55%
		D	48	1,82%
	325000	D	48	1,82%
	330000	D	48	1,82%
	335000	D	48	5,45%
	350000	D	48	9,09%
	375000	D	48	3,64%
	400000	D	48	3,64%

Figure 5 *– Filtered analyzer report, only district Bruntal included*

6.3 Associations mining with lowered data granularity

In the second DM task more preprocessing was done. The granularity of the field *district* (77 values) was decreased by grouping them into *regions* [11] (14 values) and the *amount* was divided into four intervals, starting from (0:100000> and ending by 300000+.

[3] This is a different view on data in Pentaho BI Suite, the „Analyzer report", which is designed to create reports from data, which enables more options than the standard analysis view.

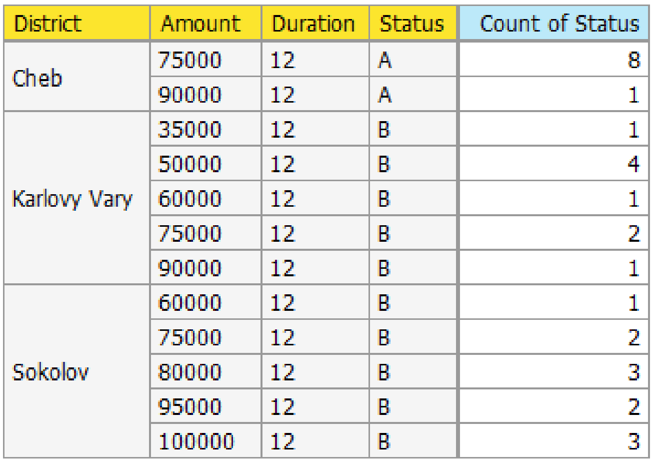
The task is formulated analogously:

Region(subset 1-1 & Duration(subset 1 - 1) &
& Amount(subset 1-1) => Quality (Bad)

The interest measures were set as *minConf=0.6* and *support=20.* With this setting, one rule was found:

Region(Karlovarsky) & Amount(0;100000> & Duration(12) => Quality(Bad) with Conf=0.69

This rule is potentially very interesting for the bank, because it tells that clients from the Karlovarsky region (which contains three districts: Cheb, Sokolov and Karlovy Vary), with relatively low loan amount (up to 100000) and duration of the loan for one year, are unreliable to pay their commitments, with probability of nearly 70%. As in the previous task we can now return to OLAP and check the data there (see Fig. 6).



District	Amount	Duration	Status	Count of Status
Cheb	75000	12	A	8
	90000	12	A	1
Karlovy Vary	35000	12	B	1
	50000	12	B	4
	60000	12	B	1
	75000	12	B	2
	90000	12	B	1
Sokolov	60000	12	B	1
	75000	12	B	2
	80000	12	B	3
	95000	12	B	2
	100000	12	B	3

***Figure 6** – Filtered analyzer report, districts from Karlovarsky region with duration of 12 month included*

Now it is clear that association rules mining could lead the bank management to incorrect conclusions. The rate of unreliable clients in this region is very high (status B) and, depending on the bank's risk management policy, they could deny all loans from the whole region. However, as we can see in Fig. 7, the Cheb district has no bad loans for the duration 12 months at all, so this decision could eliminate many reliable clients.

7 CONCLUSIONS

This paper shows that using both analysis methods, OLAP and rich association rules mining, on the same dataset as complements can be useful and may prevent premature, inaccurate conclusions. OLAP analysis is very time-consuming and with an increasing number of attributes it becomes impossible to manually discover all interesting relationships between data. On the other hand, association rules mining is (with appropriate knowledge) fast and gives us many clear and accurate results. But it is usually at the high data granularity level, so it is possible to lose the complex view to data and arrive to premature conclusions.

The imminent future work consists in a more substantial case study and in construction of a formal apparatus for relating OLAP and GUHA hypotheses. In longer term, implementation of a software facility allowing for system interoperation and user interface integration is envisaged.

8 ACKNOWLEDGEMENTS

The research is partially supported by the CSF project no. 201/08/0802.

References

[1] C. Vercellis: Business intelligence: data mining and optimization for decision making. pp. 10. John Wiley and Sons, 2009. ISBN: 0470511389

[2] OLAP and OLAP server definition. The OLAP Coouncil. 1995. Available at: *http://www.olapcouncil.org/research/glossaryly.htm*

[3] J. C. Diaz: Adoption and Usage Survey: Open Source and Business Intelligence and Reporting, 2009, BeyeNetwork

[4] R. Agrawal; T. Imielinski; A. Swami: Mining Association Rules Between Sets of Items in Large Databases", SIGMOD Conference 1993: pp. 207-216

[5] M. Šimůnek: Systém LISp-Miner — akademický systém pro dobývání znalostí z databází, Historie vývoje a popis ovládání. skripta VŠE, Praha, Oeconomica, 2010, 106 stran, ISBN: 978-80-245-1699-8.

[6] P. Hájek; T. Havránek: Mechanising Hypothesis Formation – Mathematical Foundations for a General Theory. Berlin – Heidelberg – New York, Springer-Verlag, 1978, 396 pp.

[7] R. Agrawal; R. Srikant: Fast Algorithms for Mining Association Rules. Proc. 20th Int. Conf. Very Large Data Bases. 1994.

[8] T. Kliegr; D. Chudán; A. Hazucha; J. Rauch: SEWEBAR-CMS: A System for Postprocessing Association Rule Models. Alexandria 21.10.2010 – 23.10.2010. In: *RuleML-2010 Challenge.* Washington : CEUR-WS, 2010, s. 1–8. ISSN 1613-0073.

[9] http://lisp.vse.cz/pkdd99/

[10] J. Rauch; M. Šimůnek: An Alternative Approach to Mining Association Rules. In:Lin T Y, Ohsuga S, Liau C J, and Tsumoto S (eds): Foundations of Data Mining and Knowledge Discovery. Springer-Verlag, 2005, pp. 219-239.

[11] http://obce.sweb.cz/