On the complementarity of OLAP and rich associations mining

David Chudán, Vojtěch Svátek

Department of Knowledge Engineering, University of Economics, Prague

Motivation

- OLAP analysis and data mining are the most common techniques of (large) data analysis, but they are mostly used separately.
- Isn´t it possible to use these two techniques complementary (on one analytical question at the same time) to get better results?

Technologies - OLAP

- Objective is to gain insight into the meaning contained in databases
- Based on OLAP cubes
- The basic operations in OLAP analysis are:
  - Drill-down/up
  - Slice/dice
  - Pivot (rotate)

Technologies – rich association rules mining represented by GUHA method

- Original Czech data exploration method
- Overtakes the market basket analysis and also the a-priori algorithm in terms of representational power
- Rich history, many implementations, the most recent in LISp-Miner DM software

Technologies - differences

- Level of data granularity
  - In OLAP analysis, the fields are displayed in terms of aggregation, the internal structure is not altered during the analysis
  - GUHA relies on preprocessing of numerical data into discrete categories
- Degree of automation
  - Manual browsing vs. generation of many hypotheses

The task

- The dataset contains data about the fictional bank and it´s clients
- The objective of the task is to identify subgroups of clients with many occurrences of bad loan´s quality
- In the task, the following attributes were used: *district*, *amount*, *duration*, and *status (A, B, C, D)*.

Sample course of analysis – BI view

		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					

Sample course of analysis

Associations mining with high data granularity

The formulation of the task:

District (subset 1-1) & Duration (subset 1 - 1) & Amount (subset 1-1) $=>_{0,9;5}$ *Quality (Bad)*

=> Implication *quantifier* *

0.9 - confidence

10 - support

*Petr Hájek, Tomáš Havránek
Mechanizing Hypothesis Formation
Mathematical Foundations for a
General Theory Originally
published: Springer-Verlag Berlin
Heidelberg New York in 1978
ISBN 3-540-08738-9

Sample course of analysis

Results of the task in LISp-Miner DM software:

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

Sample course of analysis

- Question: *Is it really dangerous for us (as a bank management) to provide a loan to a client from Bruntal or is it only bound to a certain amount of money or to certain duration?*
- Answer can be found out by two ways, by changing threshold values in DM task, or in the OLAP analysis

Sample course of analysis

OLAP analysis, filtered analyzer report (only Bruntal district included) in Pentaho BI suite.

Good loan quality

Bad loan quality

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%

Sample course of analysis

Associations mining with lower data granularity

The formulation of the task:

Region(subset 1-1) & *Duration(subset 1 - 1)* & *Amount(subset 1-1)* $=>_{0,6;20}$ *Quality (Bad)*

Result of the task: 1 rule

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

Sample course of analysis

- Question: *Is it really dangerous for us to provide a loan to clients from Karlovarsky region?*

No, clients from Cheb district have no problem with paying their loans.

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

Conclusion

Both analysis methods, OLAP and rich association rules mining, are useful to be used as complements – this can prevent premature and inacurrate conclusions.

Future work

- More substantial case study and 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