Data Warehousing with Graphical Models

Zdeněk Kouba
The Gerstner Laboratory for Intelligent Decision Making and Control
Faculty of Electrical Engineering, Czech Technical University
Technická 2, 166 27 Prague 6
Czech Republic
kouba@labe.felk.cvut.cz

Kamil Matoušek
The Gerstner Laboratory for Intelligent Decision Making and Control
Faculty of Electrical Engineering, Czech Technical University
Technická 2, 166 27 Prague 6
Czech Republic
matousek@labe.felk.cvut.cz

Abstract – Statistical data analysis can profit from data warehousing technology. Data warehouse is an ideal means for storing huge multivariate frequency tables. Both the facts and dimensions can be considered as random variables. Then the methods of statistical data analysis can help in determining associations among those facts and dimensions during the on-line data analysis (OLAP) process. Thus methods of statistical data analysis can be used for on-line analysis of data stored in the data warehouse. On the other hand statistical data analysis techniques can take part in the data warehouse building.

I. INTRODUCTION

Data warehousing represents a powerful knowledge based approach for gathering large amounts of data and for receiving meaningful results based on historical data analyses [11]. The requested information and/or knowledge is acquired in a reasonably short time period. Data warehouse on-line analysis engine OLAP enables many kinds of users or external applications (such as expert or knowledge based systems) to access the required chunks of data with respective data security. It enables in an easy way to define and manipulate with data views – selecting the most interesting information. To achieve this, just the interesting dimensions are selected and all the fact fields are recalculated using chosen aggregation functions.

Statistical data analysis, on the other hand, provides well-formed theory and methodology, which effectively handles data analysis process regardless the implementation and the data extent and is well-known for yielding correct and reliable results.

Discrete multivariate data analysis, as a field, is focused on analysis of multivariate frequency (contingency) tables. It can profit from data warehousing technology in sense of large multivariate frequency tables storage. Both users and the analytical process itself need fast evaluation of marginal counts (marginal frequencies) of respective random variables. Data warehouse and its OLAP engine represent ideal means for achieving this task. Both the facts and dimensions of a data warehouse can be considered as random variables from statistical data analysis point of view.

Then the methods of statistical data analysis can help in determining associations among those facts and dimensions during the on-line data analysis process performed by OLAP. Thus methods of statistical data analysis can be helpful, if used for on-line analysis of data stored in the data warehouse.

II. COLLAPSIBILITY

Two basic OLAP operations are roll-up and drill-down change the granularity level of data being analyzed. Changing the granularity level can bring the user to incorrect interpretation of data analysis results.

Let us illustrate possible problems in data analyzing on various data granularity levels on following example [6] based on data published in [9].

Let us have the three-dimensional data cube introduced in Table I. The data cube represents number of sentences in 4863 murder cases in Florida in years 1973-79. The three dimensions of the data cube represent:

- Color of the victim's skin { black, white }
- Color of the murderer's skin { black, white }
- The sentence { execution, other }

<table>
<thead>
<tr>
<th>Victim's skin colour</th>
<th>Execution</th>
<th>Other</th>
<th>Murderer's skin colour</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black</td>
<td>11</td>
<td>2309</td>
<td>Black</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>111</td>
<td>White</td>
</tr>
<tr>
<td>White</td>
<td>48</td>
<td>238</td>
<td>Black</td>
</tr>
<tr>
<td></td>
<td>72</td>
<td>2074</td>
<td>White</td>
</tr>
</tbody>
</table>

Table I: Three-Dimensional Data Cube

291
TABLE III
DATA CUBE AFTER ROLL-UP

<table>
<thead>
<tr>
<th>Murderer's skin colour</th>
<th>Execution</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black</td>
<td>59</td>
<td>2547</td>
</tr>
<tr>
<td>White</td>
<td>72</td>
<td>2185</td>
</tr>
</tbody>
</table>

After roll-up over Victim’s skin colour we get two-dimensional data cube introduced in Table II.

When analyzing the rolled-up data cube (Table II) the conclusion can be made that relatively more white murderers (3.2%) have been executed than black ones (2.3%). However when analyzing the original (Table I) data cube it is clear that both in black victim sub-cube and white victim sub-cube the percentage of black murderers having been executed is much higher than the percentage of white murderers.

This example should warn OLAP developers. Users tend to use the analytical tools in a carefree way. OLAP operations should therefore provide explanatory capabilities as much as possible.

Statistical multivariate data analysis provides suitable methods and tools. The strange situation in the above example is a consequence of the fact, that the three-dimensional table was not collapsible. The reason is that the structure of mutual associations among all three dimensions of the three-dimensional data cube has been corrupted by rolling-up the cube. For example examining of all two-dimensional sub-cubes of a complex data cube may be very misleading if any of the data cube dimensions are interrelated. However, the dimensionality of a data cube may be reduced by well founded collapsing the data cube over all totally independent variables. Collapsibility is explained in details in [1].

III. DATA WAREHOUSING AND STATISTICAL DATA ANALYSIS

The structure of the dependence among factors (or, alternatively in the case of a data warehouse, dimensions) may be described by so-called log-linear models. For example, for a three-factor case, the log-linear model looks like this:

\[
\log p(A,B,C) = \theta^0 + \theta^A + \theta^B + \theta^C + \theta^{AB} + \theta^{AC} + \theta^{BC} + \theta^{ABC} .
\]  

Parameter \( \theta^0 \) is always nonzero; the other parameters \( \theta^S \) (\( S \) denotes the set of factors) express how the probability \( p(A,B,C) \) is affected by the mutual interaction of factors.

The presence and/or absence of some \( \theta \) parameters in (1) defines the particular dependency structure. The term „model“ denotes the dependency structure in the following text.

Thus the case when factors \( A, B, C \) are mutually independent will be represented by a model having all \( \theta^S \), \( \theta^{BC}, \theta^{AC}, \) and \( \theta^{ABC} \) equal to zero vectors, i.e. the following log-linear model will represent the particular dependency structure:

\[
\log p(A,B,C) = \theta^0 + \theta^A + \theta^B + \theta^C .
\]  

Another well-known dependency structure called conditional independence can be recognised easily having corresponding log-linear model. Let \( A \) be conditionally independent on \( B \) given \( C \) (i.e. \( A \perp B | C \)). Then \( \theta^{AB} \) and \( \theta^{BC} \) will be equal to zero vectors and the corresponding log-linear model will look like:

\[
\log p(A,B,C) = \theta^0 + \theta^A + \theta^B + \theta^C + \theta^{BC} + \theta^{AC} .
\]  

This easy interpretation of log-linear models is the main reason of their popularity.

It can be tested by statistical means whether or not this particular model fits given data sufficiently. One of crucial aims of statistical multivariate data analysis is to find the simplest model (or set of them) which interpret the data cube sufficiently well.

Having such a model we can make decisions concerning the collapsibility. Further we can use it for sophisticated forgetting of non actual data with minimal information lost as was suggested in [5]. Having the dependency structure at disposal we can carry out simulations and predictions based on data stored in the data warehouse.

Statistical multivariate data analysis provides us with theory and methodology of log-linear models and their special subclasses - graphical and decomposable models [6]. The problem is that there exist no analytical method deducing the dependency structure (expressed by means of a log-linear model) from the contents of a data cube. The only possibility is to make a hypothesis about the model and consequently to test it. Of course, the number of possible log-linear models for a data cube with higher number of dimensions may be extremely huge and testing all of them will not be computationally feasible. However, for a subclass of log-linear models known as graphical models Havránek published an algorithm [3] making chance to reduce dramatically the number of necessary tests. This algorithm was further improved by Havránek and Edwards [4].

IV. SUMMARY

This paper shows that both statistical data analysis and data warehousing technology can profit from each other. Data warehouse appears to be a suitable knowledge-based technology for storing large multivariate contingency tables and fast evaluation of required marginal counts. Random variables can be represented by the facts as well as dimensions. By examples of collapsibility and data forgetting, the paper illustrates, how the statistical data analysis methods can affect development of OLAP engine and multidimensional data models.

On the other hand, methods of statistical data analysis can be used for on-line analysis of data stored in the data warehouse, determining specific associations among dimensions and facts, data prediction and interpretation [7].
V. REFERENCES


