Contribution to the „Discovery Challenge" at the 3rd European Conference on Principles and Practice of Knowledge Discovery in Databases, PKDD '99, September 15-18, 1999, Prague, Czech Republic

Visual, Interactive Data Mining with InfoZoom – the Financial Data Set

Michael Spenke, Christian Beilken
GMD – German National Research Center for Information Technology
FIT – Institute for Applied Information Technology, http://www.gmd.de/fit
Schloss Birlinghoven, D-53754 Sankt Augustin
Michael.Spenke@GMD.de, http://fit.gmd.de/hci/pages/michael.spenke.html

Abstract

This paper describes the application of the data analysis tool *InfoZoom* to the Financial Data Set for the PKKD'99 Discovery Challenge. No automatic method for data mining is used. Instead, InfoZoom enables the user to interactively explore different visualizations of the data. In this way, the user gets a feeling of the data, detects interesting knowledge, and gains a deep understanding of the data set.

Introduction

InfoZoom has been developed at *GMD – the German National Research Center for Information Technology*. It is now marketed by the GMD spin-off company *humanIT*.

InfoZoom displays database relations as tables with attributes as rows and objects as columns. In Figure 1 each column corresponds to a bank account with a granted loan. The attributes are hierarchically ordered like files in a directory.

The menu left of each attribute name shows the possible values and their frequency. Selecting a value from the menu restricts the table to the objects with this value. Clicking on the arrow outline right of an attribute sorts the table by this attribute.

A table can be compressed by clicking the second of the three mode buttons in the upper left corner of the table (above „682 of 682 objects").

In **Compressed Mode**, the column width is reduced until all the objects

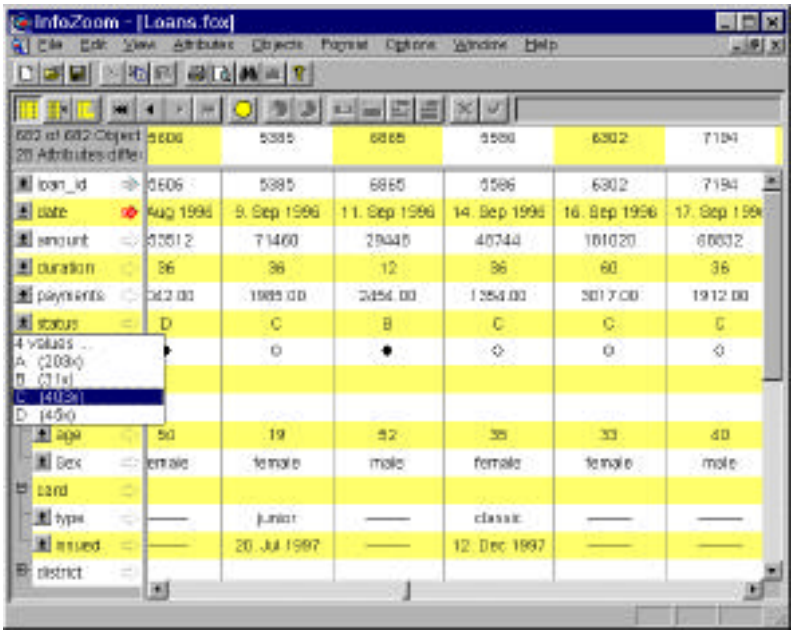

Figure 1: Wide Table Mode

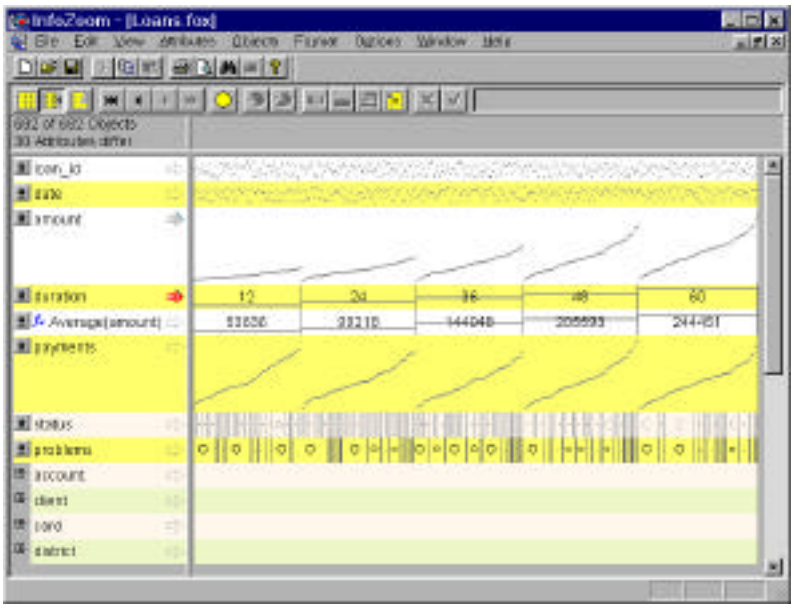

Figure 2: Compressed Table Mode

fit on the screen. In Figure 2 the column width is about one pixel. In large tables the column width will be even smaller. Some techniques make the table readable in spite of this compression. The most important is that neighbouring cells with identical values are combined into one larger cell. Because the table is sorted by the attribute *duration*, all its values are readable. The width of each cell indicates the number of subsequent objects with this value. If a cell is to small to display a numeric value, a short horizontal line still indicates its relative height. In this way, it can be seen that loans with a short *duration* typically also have a small *amount*.

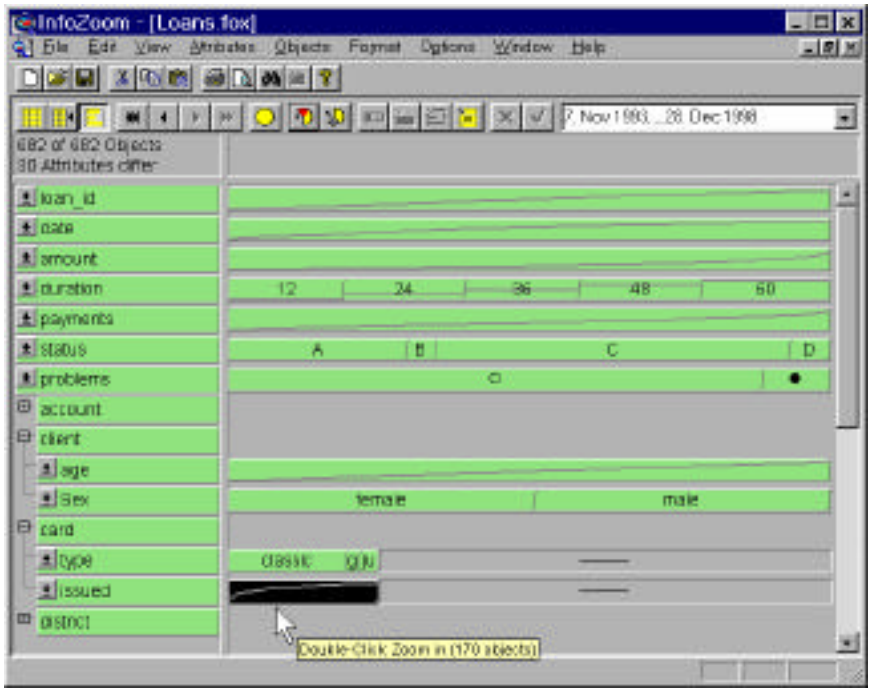

Figure 3: Overview Mode

Instead of selecting a value from the menu, an attribute can also be restricted by selecting and double-clicking a value or value-range directly in the table. In a short animation, the clicked cells grow while the others shrink. This looks like **zooming** into the table.

Like the formula-cells in a spreadsheet program, derived attributes can be defined which are automatically updated by InfoZoom when necessary. For example, in Figure 2 the average *amount* for each *duration* will be recomputed after each zoom operation.

In **Overview Mode** the values of each attribute are sorted independently. It is important to understand that this is not a table but something like a bar chart displaying the value distributions for the attributes.

The Financial Data Set

The Financial Data Set consists of 8 relations describing bank accounts, clients, credit cards, permanent orders, transactions, and loans. The semicolon-separated tables could be directly imported into InfoZoom. Next, the correct date formats had to be defined, so that InfoZoom correctly interprets and sorts the values.

We have concentrated on the question if clients have had any problems with paying back granted loans and if these cases correlate with other information about the accounts and the clients. For each loan there is an attribute *status* indicating whether the loan contract is finished and whether there are or have been any problems.

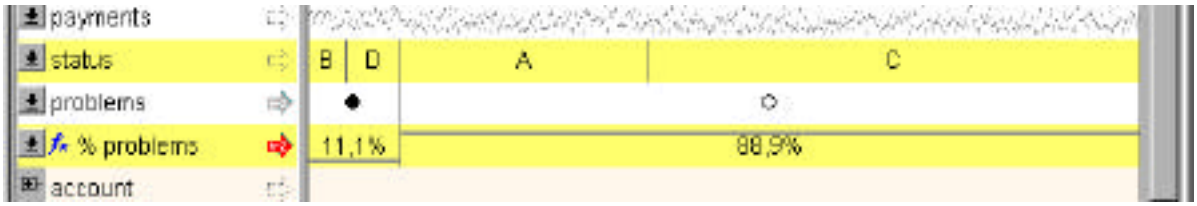

Figure 4: Definition of the target attribute *problems*

We have defined a new attribute *problems* which is true for *status* values *B* and *D*, which indicate that the client is in debt. This is our **target attribute**. We have also defined a derived attribute which always shows the

percentage of problematic loans. The goal of the analysis is to find correlations between *problems* and the other attributes.

Data Mining Techniques

We begin our analysis with a table which has exactly one column for each account. It only contains the 682 accounts with a loan. For 76 of them problems have been reported so far. As the derived attribute *% problems* shows, this is equivalent to 11.1% of the accounts.

The easiest technique for the detection of correlations between the attribute *problems* and other attributes is performed in the *Overview Mode* where the value distributions are shown: We zoom-in on the loans with problems and watch how the other value distributions change. This is done by double-clicking on the filled circle in the row of the attribute *problems*.

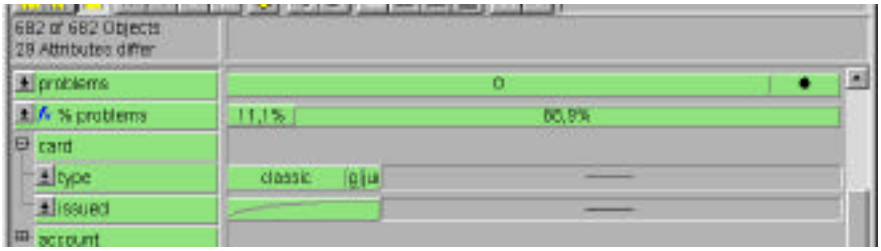

Figure 5: A credit card has been issued for about a quarter of the accounts.

This starts a short animation where the field with the filled circle is growing and the field with the empty circle becomes smaller. Finally, only the filled circle remains, indicating that only loans with problems are displayed now.

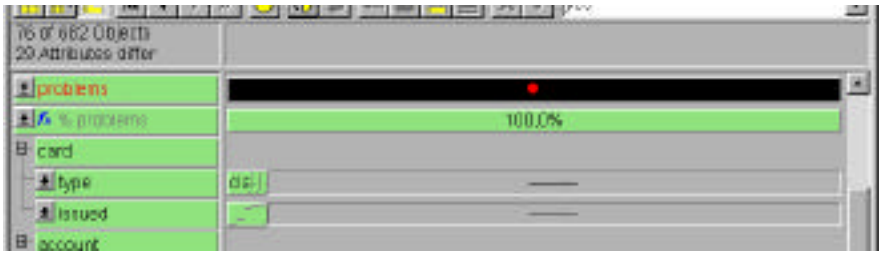

Figure 6: Only a few credit cards where issued for accounts with problems.

Simultaneously during the animation those fields shrink which indicate that credit cards (*classic*, *gold*, and *junior*) where issued for some accounts. Obviously, there is a strong correlation between *problems* and *card type*. The underlying causal dependency can however only be determined by a domain expert: Maybe certain criteria have to be met before a credit card is issued. Maybe the credit card is withdrawn as soon as problems occur...

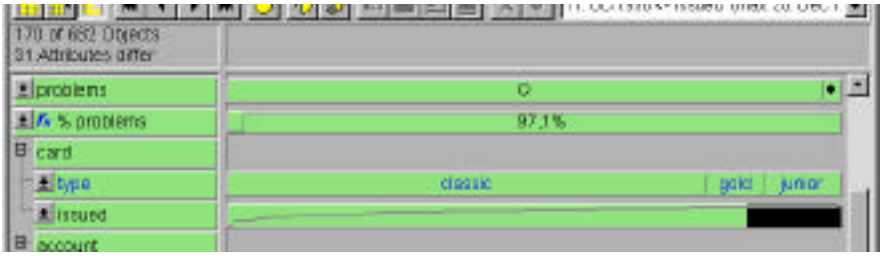

Figure 7: Accounts with credit cards rarely have loan problems.

Vice versa, the same correlation can be observed if we zoom-in on the credit cards by selecting and double-clicking the fields *gold*, *classic*, and *junior*. The fraction of accounts with problems drops to 2.9%. If we further zoom-in on the cards issued lately (selected in Figure 7), the problematic loans completely disappear.

With the same techniques many correlations between *problems* and other attributes can be easily detected (see below). The animation makes it possible to look for growing or shrinking fields in multiple attributes at the same time. Significant movements immediately catch the eye of the user.

For example, we can select and zoom-in on the right part of the curve for the attribute *amount* in Figure 3 and watch how the filled circle for *problems* grows. If we use the opposite technique of double-clicking the filled circle and watching the *amount* attribute, we can observe that the curve is distorted, but it is hard to decide whether the amounts increase or decrease. In this case it is a good idea to introduce a derived attribute which divides the amounts into some coarse classes:

After zooming-in on the problematic loans, the cells of the higher value classes grow:

Another technique builds on **consecutive sorts** of the compressed table. In Figure 8 the table of all accounts with loans has been joined with the *disposition* relation. As a consequence, for some accounts there is more than one column in the table because there are additional users besides the owner. The attributes have been sorted in the following order: *age*, *Sex*, *% problems*, *type*, *#clients*.

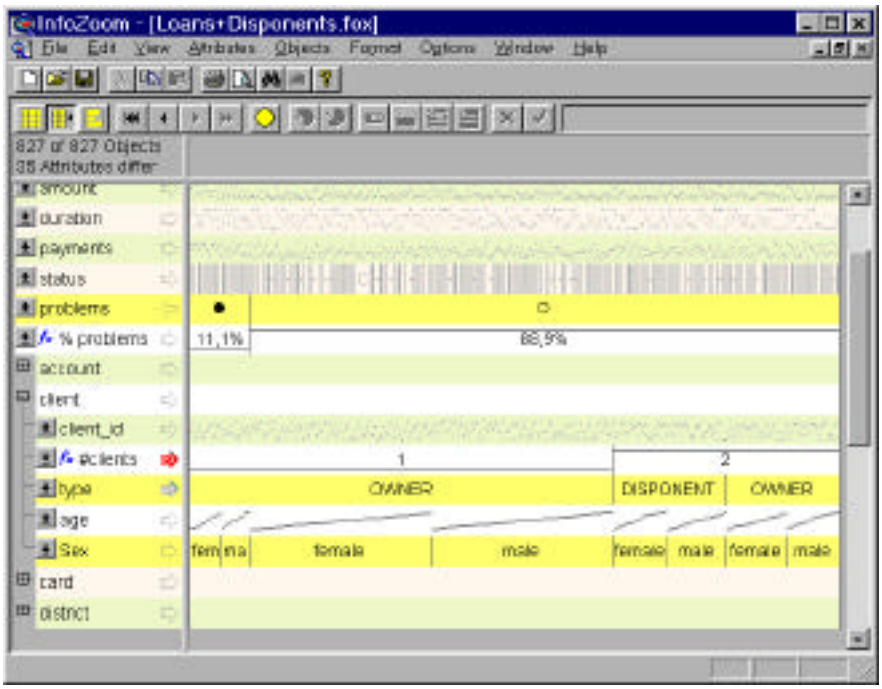

Figure 8: Accounts with two users never had problems!

The derived attribute *#clients* shows that there are never more than two users. Roughly a quarter of the accounts have two users. Surprisingly, these accounts **never** have had any problems with paying back loans! This strange observation is confirmed when we zoom-in on the problematic cases again: The *DISPONENT* field completely disappears and all remaining accounts have only a single user, the owner itself.

Derived attributes computing **percentages** can be used to search for correlations very efficiently. In Figure 9 an attribute has been defined which computes the percentage of loans currently displayed in the table for each district. As long as the complete table is shown, the result is *100%* for each district. In Figure 9 we have already zoomed-in on the problematic loans.

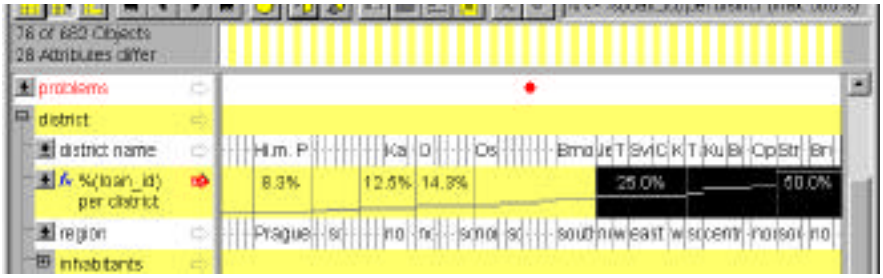

Figure 9: Percentage of problematic loans per district

For some districts 50% of the loans are still displayed even though we have already zoomed-in on the 11.1% cases were there are problems. This means that every second loan in these districts is problematic (instead of the average of 11.1%). In order to see the names of these districts, one can simply double-click the 50% field and zoom into these districts.

In order to search for correlations with the **demographic data** available for each district, we first select the values from 25.0% to 50.0% as in Figure 9. Next we sort by one of the demographic attributes by clicking at the arrow after the attribute name. In Figure 10 the attribute *average salary* was chosen.

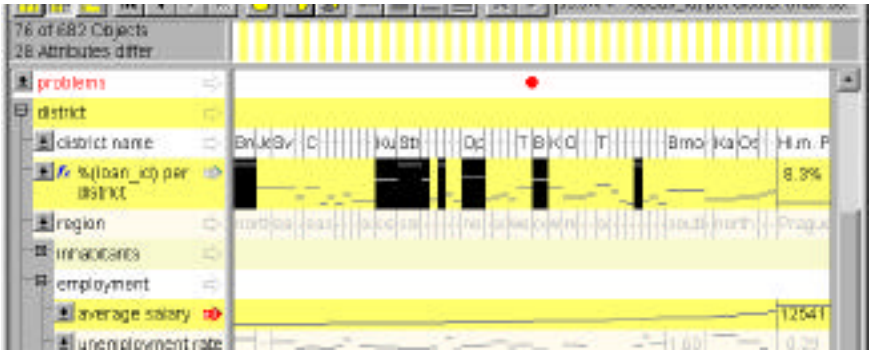

Figure 10: Watching the distribution of selected cells after a re-sort

The sorting action does not clear the selection of the percent values. It is, however, broken into smaller pieces scattered over the row. The districts with a low average salary are now displayed on the left side. The selected high percentages concentrate a bit on the left side. Therefore, we can conclude that the danger of loan problems is somewhat higher in districts with a low average salary. However, this is only a weak correlation. Maybe surprisingly, no strong correlations with any of the demographic attributes were found.

Provided that there is sufficient main memory, InfoZoom can handle very large tables. For example, the table with 1,056,320 transactions could be analyzed. In Figure 11 it is shown that only 4.1% of the accounts with at least one transaction concerning the bank *WX* had problems with a loan. Here a table of the 191,556 transactions for accounts with a loan was used.

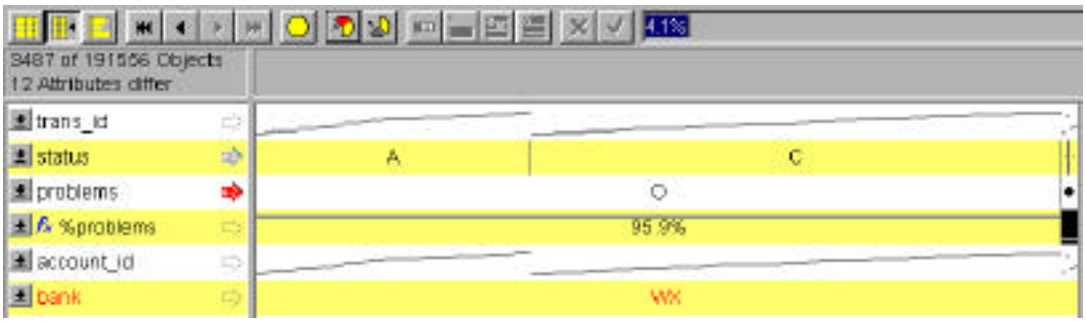

Figure 11: Transactions concerning the bank *WX*

Findings

The attribute *problems* was used as the target attribute. It describes whether there have been any problems with a loan given to a customer. In 11.1% of the loans there have already been problems. The following correlations between *problems* and other attributes were found using the techniques described above:

- There are less problems with **newer** loans. However, this may be because the newer loans are still running and may cause problems in the future.
- If the **amount** of a loan is higher, problems are more probable and vice versa.
- The **monthly payments** are exactly computed by the formula *amount / duration*. Therefore, the *payments* attribute is redundant.
- Contracts over **12 month** cause slightly less problems than longer contracts.
- Accounts with a **weekly issuance** of statements cause somewhat more problems.
- It does not make a difference whether the owner of the account is male, female, young, or old.
- Owner of **credit cards** have caused problems in only 2.9% of the contracts.
- In the **districts** *Bruntal*, *Domazlice*, *Sokolov*, and *Strakonice* 50% of the loans are problematic.
- In the **region** *North Bohemia* only 1.6% of the loans have caused problems.
- There is practically no correlation with the **demographic data** like *unemployment rate*, or *average salary*.
- If there are **permanent orders** of type *household*, the probability of problems is only 4.5%. If we compute the sum of these orders per account and look at the half of the accounts with a high sum, the probability is even only 2.7%.
- If among the **transactions** for an account there are collections from other banks, the danger to get problems with a loan is only 6.2%. If we compute the sum of all these transactions for each account, and consider only those accounts with a sum higher than the average sum, the danger even drops to 3.8%.
- If at least one transaction of an account was a **sanction interest** (because of a negative balance) there have been problems with the loan in 90%.
- If the **average balance** of an account is high, problems are rare.
- If there is at least one transaction regarding the **bank** *WX*, the percentage of problems drops to 4.1. If it is a credit, there have been no problems at all.

Finally, the most surprising result is the following:

- Accounts where there is a **second authorized user** besides the owner have caused no problems at all.

We do not draw any conclusions about the underlying causal dependencies leading to these results. This can only be judged by domain experts.

Conclusion

We have shown the interactive techniques for visual data mining supported by InfoZoom. The goal of our approach is not a completely automatic algorithm that searches for interesting results. Instead, InfoZoom enables the user to interactively explore the data set and to get a feeling of the contained information. It introduces a unique visualization of the whole data set on a single screen. The user can easily view the data set from many different views. Zooming into the data is simply performed by selecting parts of the displayed data. Moreover, one can quickly perform powerful queries and immediately see the results. Derived attributes can be defined like in a spreadsheet program and are automatically updated when necessary.

We are convinced that using InfoZoom is simple enough to be used by domain experts in order to understand their data and to detect the hidden knowledge.

References

[1] Spenke, M.; Beilken, Chr.; Berlage, Th., *The Interactive Table for Product Comparison and Selection*, Proceedings of the UIST 96 Ninth Annual Symposium on User Interface Software and Technology, Seattle, November 6 - 8, 1996. ACM 1996, pp. 41 – 50.

[2] http://fit.gmd.de/hci/projects/focus/focus.html – The InfoZoom home page, [1] is available online.

[3] http://www.humanIT.de – Download a free test version of InfoZoom.