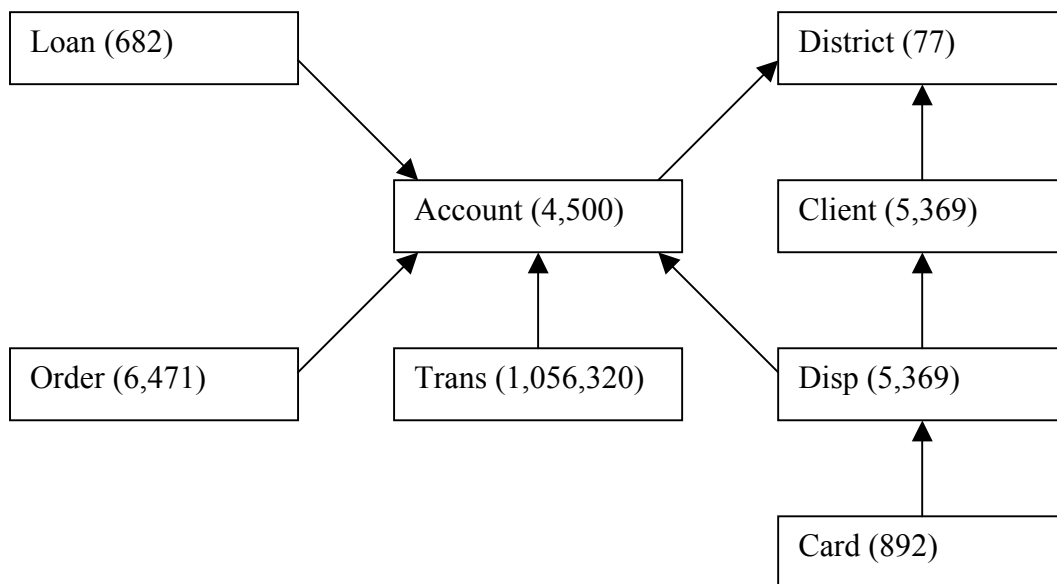


## A PKDD 1999 Challenge Financial Dataset Preparation

Mark-A. Krogel, Otto-von-Guericke-Universität, [krogel@iws.cs.uni-magdeburg.de](mailto:krogel@iws.cs.uni-magdeburg.de)

This is the description of a preparation of data made for finding predictive models for loan.status with the help of a RELAGGS variant, our system for propositionalization [1]. We hope that the pre-processed data might be useful for other applications as well.

The starting point is the dataset as provided by Petr Berka at <http://lisp.vse.cz/challenge>, where you can also find a description of the data. We find 8 tables depicted as rectangles with table names inside in the following figure, with arrows indicating foreign key relationships, always pointing from the table with the foreign key attribute to the one with the corresponding primary key attribute. Figures in brackets indicate numbers of records.

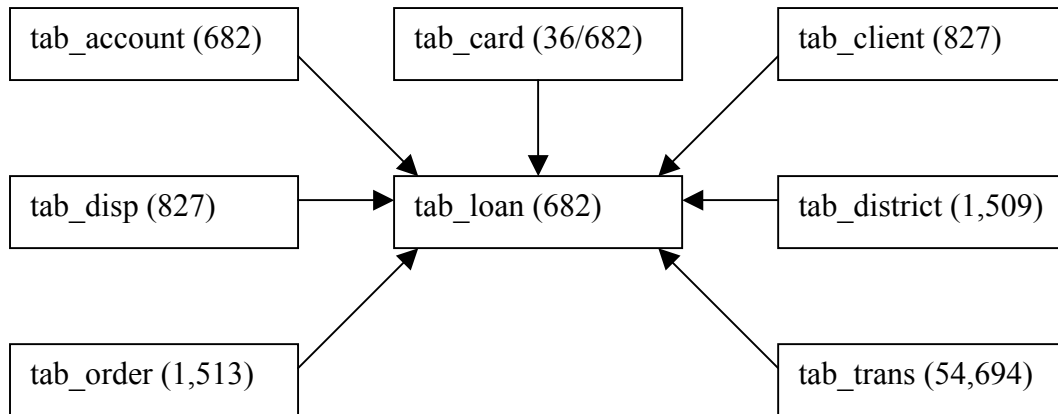


The following changes are applied to arrive at the pre-processed dataset:

- We reduce the dataset with respect to loans. For example, all transaction data on accounts without loans are not included.
- We reduce the dataset considering time constraints. For example, transaction data on a special account that were created only after the loan on that account was granted are not included. The task is predictive modeling, and so only those data should be used that were available in the moments of the decisions about loan granting. (However, we used all district information and order information – subject to discussion.)
- Dates are transformed to be of type date.
- Attribute birth\_number of table client is split into two attributes: birthday and sex.
- All tables (except loan, of course) are enriched with a foreign key attribute loan\_id.

We build SQL script files for the creation of the tables and their contents in MySQL databases. You can use the MySQL command source for the execution of the scripts, always create before inserts. (With the additional indexes, some work with the tables can be accelerated.) In principle, these files should work for other database servers as well.

Finally, the following database schema should result:



This looks like a star schema, however, it differs from the usual data warehouse facts-and-dimension-tables, eg. wrt. the direction of the arrows. The figure only shows foreign key relationships as caused by the freshly distributed attribute `loan_id`.

Note, that in `tab_card`, 36 records describing relevant cards are complemented by records for the loans on accounts without cards that are filled with NULL values in the corresponding fields. This was necessary for the current version of RELAGGS which works with inner joins.

Note also, that the figure for records in `tab_district` is reached by a UNION ALL of both accounts' districts and clients' districts.

The other record figures result from the reductions to loan specific information from the correct time intervals as described above.

Results obtained so far are the following. When A and C loans are combined into a positive class (loans without problems) and B and D into a negative class (loans with problems), we observe 11.1% of negative examples, which would be the default error rate for classification. Using WEKA <http://www.cs.waikato.ac.nz/ml/weka/> on RELAGGS results [2], especially J48 with 10-fold cross-validation, we achieve an average error rate of 7.8%. When the WEKA attribute selector based on information gain is applied before learning, this error rate even drops to 5.9%. The new variant of RELAGGS, which encompasses more aggregation functions, yields an error rate of 4.1%. This tool and another tool for the export of MySQL tables into ARFF files are available on request from [kroegel@iws.cs.uni-magdeburg.de](mailto:kroegel@iws.cs.uni-magdeburg.de).

## References

- [1] M.-A. Kroegel and S. Wrobel. Transformation-Based Learning Using Multirelational Aggregation. In: C. Rouveirol and M. Sebag (eds.) *Proceedings of the Eleventh International Conference on Inductive Logic Programming (ILP)*. LNAI 2157, Springer-Verlag, 2001.
- [2] M.-A. Kroegel and S. Wrobel. Feature Selection for Propositionalization. In: S. Lange, K. Satoh, and C. H. Smith (eds.) *Proceedings of the Fifth International Conference on Discovery Science (DS)*. LNCS 2534, Springer-Verlag, 2002.