

Entity Level Data Integration by Statistical Methods

Hans-J. Lenz

Free University, Berlin
hjlenz@wiwiss.fu-berlin.de

Abstract

In most cases unique identifiers are required to join data from different databases. If global unique keys are absent or corrupted the supplement of data extracted from different sources becomes difficult. The main question is: Does a given record is related to an entity which is identical to an entity corresponding to another record, or not? This leads to a classification problem with at least two classes: identical and not identical.

Classifying pairs of records needs a three-step procedure. The first step is to define suitable common properties (attributes) of data for all different sources. Secondly, to allow comparisons the values of the records are transformed to this common properties. Finally, the classification is performed on an almost finite subset, the range of an appropriate comparison function.

Different classification techniques can be applied like Association Rules, Classification Trees, Neural networks or Record Linkage techniques. The unknown parameters of the classification rules are computed by sampling and supervised learning. Unbiased error rates can be estimated for instance by cross validation. Special attention must be paid to control the computing complexity of the identification process. The approach will be illustrated for data from two library databases and from the planned German administrative record census, which will become a substitute of a regular census.

References

- [1] Agrawal, R., Imielinski, T. and Swami, A. N. (1993): Mining association rules between sets of items in large databases. ACM SIGMOD Int. Conf. on Management of Data, Washington, DC, pp. 207216.
- [2] Alvey, W. and Jamerson, B. (Eds.) (1997): Record Linkage Techniques 1997. Int. Workshop and Exposition. Fed. Committee on Stat. Methodology, Off. of Management and Budget, Washington, DC.
- [3] Breiman, L., Friedman, J., Olshen R., and Stone C. (1984): Classification and regression trees. Chapman and Hall.

- [4] Fellegi, I. P. and Sunter, A. B. (1969): A theory of record linkage. Journal of the American Statistical Association, 64, 1183-1210. Wirtschaftswissenschaft der FU Berlin.
- [5] Neiling, M. and Lenz, H.-J. (2000): Data integration by means of object identification in information systems. 8th Eur. Conf. on Information Systems, Vienna, Austria, July 2000.