Data Mining Methods for Linking Data Coming from Several Sources

Vicenç Torra[1], Josep Domingo-Ferrer[2] and Àngel Torres[2]

[1] Institut d'Investigació en Intel·ligència Artificial - CSIC,
Campus UAB s/n, E-08193 Bellaterra, Catalonia
e-mail vtorra@iiia.csic.es

[2] Dept. Computer Engineering and Maths (ETSE), Universitat Rovira i Virgili,
Av. Països Catalans 26, E-43007 Tarragona, Catalonia
e-mail {jdomingo,atorres}@etse.urv.es

Abstract. Statistical offices are faced with the problem of multiple-database data mining at least for two reasons. On one side, there is a trend to avoid direct collection of data from respondents and use instead administrative data sources to build statistical data; such administrative sources are typically diverses and scattered across several administration level. On the other side, intruders may attempt disclosure of confidential statistical data by using the same approach, *i.e.* by linking whatever databases they can obtain. This paper discusses issues related to multiple-database data mining, with a special focus on a method for linking records across databases which do not share any variables.

Keywords. Statistical disclosure control, Re-identification, Data mining, Artificial intelligence.

1 Introduction

Statistical offices are faced with the problem of multiple-database data mining at least for two reasons:

- On the good side, there is a trend to avoid direct collection of data from respondents and use instead administrative data sources to build statistical data; such administrative sources are typically diverses and scattered across several administration level. Linking administrative information held by municipalities with information held at higher administration levels can yield information that is more accurate and cheaper than the one that would be collected directly from respondents.

- On the bad side, statistical offices must realize that intruders may attempt disclosure of confidential statistical data by using exactly the same approach, *i.e.* by linking whatever databases they can obtain. This is the relevant side for statistical disclosure control (SDC).

This paper discusses issues related to multiple-database data mining, with a special focus on a method for linking records across databases which do not share any variables.

Section 2 is about general concepts of data mining and knowledge discovery in databases. Section 3 discusses the use of data mining in SDC, that is, how data mining can increase disclosure risk.

2 Data mining and knowledge discovery in databases

Several definitions are currently being used for both data mining and knowledge discovery in databases. While in some situations they are used as equivalent terms, data mining is often considered as one of the steps in the knowledge discovery process. Here, following Fayyad *et al.* (1996b), we use the latter approach, which is more suited for describing the relationships between this field and statistical disclosure control.

According to Fayyad *et al.* (1996a) and Frawley *et al.* (1991), knowledge discovery in databases (KDD) is defined as follows:

> *Knowledge discovery in databases* is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.

This process encompasses several steps. According to Kantardzic (2003), these are:

(i) Problem statement and hypothesis formulation

(ii) Data collection

(iii) Data pre-processing

(iv) Model estimation

(v) Model interpretation

We will focus on steps (iii) and (iv). Data pre-processing includes all those mechanisms used to improve data quality. Model estimation is also known as data mining, *i.e.* application of computational methods for building models from data. The next two sections describe data pre-processing and model estimation in greater detail. The last section reviews current trends in data mining and highlights their relation with Statistical Disclosure Control.

2.1 Data pre-processing

In realistic databases, data are not free from errors or inaccuracies, which can be due to accidental or intentional distortion. The pre-processing step is used to improve the quality of data. Several methods can be used to help in this task. Some of them are next recalled (see Kantardzic (2003)):

Simple transformations: Transformations that do not need major analysis of data and can be applied considering a single value at a time. These transformations include outlier detection and removal, scaling, data re-coding.

Cleansing and scrubbing: Tranformations of moderate complexity, *e.g.* involving name and address formatting.

Integration: This is applied to process data coming from various sources. Integration techniques are especially relevant for data mining in heterogeneous databases. Relevant tools include re-identification methods (in particular, record linkage algorithms) and tools for identifying attribute correspondences.

Aggregation and summarization: These transformations aim at reducing the number of records or variables in the database.

2.2 Model estimation

The main step of the knowledge discovery process is the actual construction of the model from the data. This is the data mining step. This step is defined in Fayyad *et al.* (1996a), page 9, as follows:

> The *data mining* component of the KDD process is mainly concerned with means by which patterns are extracted and enumerated from the data.

The patterns extracted by the data mining algorithm are supposed to constitute knowledge. In this setting, knowledge is understood as an interesting and certain enough pattern (Dzeroski, 2001). Of course, the terms *interesting* and *certain enough* are user- and application-dependent.

There is a large collection of data mining methods and tools available from the literature. A common classification, inherited from the machine learning field, is to divide data mining methods into two groups, one of them corresponding to supervised learning methods and the other to unsupervised learning methods. This classification is detailed below; data are considered as a flat table comprising variables and records.

Supervised learning: For one of the variables (modeled variable), a functional model is built that relates this variable with the rest of variables. Depending on the type of variable being modeled, two classes of methods are usually considered:

Classification: The variable for which the model is build is categorical. This categorical variable is called class. Descriptions are built to infer the class of a record given the values for the other variables.

Regression: This is similar to the classification problem, but the variable being modeled is continuous.

Unsupervised learning: Some knowledge about the variables in the database is extracted from the data. Unlike for supervised learning, there is no distinguished variable being modeled; instead, relationships between variables are of interest. Common unsupervised learning methods include clustering methods and association rules; the literature often includes in this group other (statistical) tools like principal components and dimensionality reduction methods.

Clustering: Clusters (groups) of similar objects are detected. Conceptual clustering is a subgroup of methods whose goal is to derive a symbolic representation from clusters.

Association rules: These specify tuples of values that appear very often in a database. They are commonly used in databases related to commercial transactions, *e.g.* to link purchases of product A with purchases of product B.

2.3 New trends in data mining

While data mining was in the past focused on the case of single flat files, currently there is a need for considering more complex data structures. In fact, two main situations arise, that correspond to two subfields:

Relational data mining In this situation, there is a single (relational) database consisting of multiple tables. Relational data mining looks for patterns that involve multiple relations in a relational database. It does so directly, without transforming the data into a single table first and then looking for patterns in such an engineered table. The relations in the database can be defined extensionally, as lists of tuples, or intensionally, as database view or sets or sets of rules. The latter allows relational data mining to take into account generally valid domain knowledge, referred to as background knowledge (Dzeroski, 2001).

Multi-database data mining In this case, there are different databases whose records must be linked before applying data mining techniques (Zhong, 2003). In this type of data mining the pre-processing step is especially critical, as the data must reach a quality level that allows records across databases to be linked.

3 Data mining in SDC

Several aspects of data mining are of interest in SDC. For example, most supervised learning methods and some unsupervised methods (*e.g.*, association rules) can be used to attack SDC because they allow relationships to be established between variables, which can lead to disclosure.

We will concentrate here on record linkage across databases. A current assumption in SDC is that record linkage can use variables shared across the databases. This assumption will be relaxed here and no set of shared variables will be assumed. What is needed in our approach is just a set of shared individuals or entities across the files —without such a set of shared individuals, record linkage does not make sense.

Example 1. A typical scenario where our relaxation is especially relevant is when data files with similar information (*e.g.* financial variables) are available for different time periods, (*e.g.* two consecutive years) which relate to nearly the same individuals (*e.g.* the companies of a certain region). In this case, even though variables are not the same (”2000 turnover” is not the same as ”2001 turnover”), re-identification via record linkage is possible.

Record linkage without shared variables is a subject of interest for both statistical disclosure control and data mining, because it highlights relationships between individuals that would otherwise remain implicit and undiscovered in the files to be linked.

Our approach to re-identification via record linkage without shared variables is rooted in the techniques of knowledge elicitation from groups of experts described in Torra and Cortés (1995) and Gaines and Shaw (1993). In these two references, a common conceptual structure is built from the information/opinion supplied by the group of experts, which should synthesize the information/opinions obtained from individual experts. In re-identification without shared variables, we assume that this common structure exists so that it makes sense to look for links between individuals in the different files. Note that, both in re-identification without shared variables and in knowledge elicitation from groups of experts, the initial information is similar: it consists of sets of records corresponding to roughly the same objects and evaluated according to a set of different variables in each file (in knowledge elicitation, the opinions of expert A on an object are different variables from the opinions of expert B on the same object).

Using the jargon in Gaines and Shaw (1993) for knowledge elicitation from groups, four cases can be distinguished depending on the coincidence or non-coincidence of variables and terminology (terminology is the domain of the variables, *i.e.* the terms used to evaluate the individuals):

Consensus Same variables and same terminology.

Correspondence Same variables but different terminology.

Contrast Different variables and different terminology.

Conflict Different variables and same terminology.

Classical record linkage falls into the case of consensus or correspondence, although in the latter case only small terminology differences are allowed (small inconsistencies among names, missing values and the like). However, based on the above classification, other types of record linkage are conceivable: correspondence when the degree of non-coincidence on the terminology is not limited to small variations of names (*e.g.* completely different terms, due, for example, to the use of different granularities), contrast and conflict.

We study in this paper the case of contrast, that is, record linkage when neither variables nor terminology are the same across the files to be linked. The only assumption of our approach is that a common structure underlies the files to be linked. In the context of Example 1, this assumption means that companies which are deemed similar according to some financial variables for the first year will also stay similar for the corresponding second year financial variables.

3.1 Re-identification without shared variables

As explained above, re-identification without common variables requires some assumptions, which are next summarized:

Hypothesis 1 *A large set of common individuals is shared by both files.*

Hypothesis 2 *Data in both files contain, implicitly, similar structural information. In other words, even though there are no common variables, there is substantial correlation between some variables in both files.*

Structural information of data files stands in our case for any organization of the data that allows explicit representation of the relationship between individuals. This structural information is obtained from the data files through manipulation of the data (*e.g.* using clustering techniques or any other data analysis or data mining technique). Comparison of the structural information implicit in both files is what allows two records that correspond to the same individual to be linked by the system.

Hypothesis 3 *Structural information can be expressed by means of partitions.*

In our approach, structural information is represented by means of partitions. Partitions obtained from data through clustering techniques make explicit the relation between individuals according to the variables that describe them. Common partitions in both files reflect the common structural information. We prefer partitions rather than other (more sophisticated) structures also obtainable with clustering methods, like dendrograms, because the former are more robust to changes in the data, as shown in Neumann and Norton (1986).

Although the main interest of our research is re-identification of individuals, the approach described below is not directly targeted to the re-identification of particular individuals. Instead, we try to re-identify groups of them. Due to this, we use the term of group-level re-identification; record-level re-identification is a particular case of group-level re-identification where one or more groups contain a single record. See Domingo-Ferrer and Torra (2003) for further details.

4 Conclusions

Data mining across different data sources has been discussed in this paper. Specifically, a method for linking records across databases which do not share any variable has been sketched. While such new data mining approaches can be cost-effective and useful to build statistical data from several administrative sources, they are rather a threat from the viewpoint of SDC. For that very reason, data protectors should use those methods if they with to obtain realistic estimates of disclosure risk.

Acknowledgments

This work is partially supported by the European Commission through “CASC” (IST-2000-25069) and by the Spanish MCyT and the FEDER fund through project “STREAMOBILE” (TIC-2001-0633-C03-01/02).

References

Dzeroski, S., (2001), Data Mining in a Nutshell, in S. Dzeroski, N. Lavrac, Relational Data Mining, Springer, 4-27.

Domingo-Ferrer, J., Torra, V., (2003), Disclosure risk assessment in statistical disclosure control via advanced record linkage, Statistics and Computing, to appear.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., (1996a), From Data Mining to Knowledge Discovery: An Overview, in U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy, Advances in Knowledge Discovery in Data Mining, MIT Press, 1-34.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R., (1996b), Advances in Knowledge Discovery in Data Mining, MIT Press, 1-34.

Frawley, W. J., Piatetsky-Shapiro, G., Matheus, C. J., (1991), Knowledge Discovery in Databases: An Overview. in G. Piatetsky-Shapiro, W. J. Frawley, Knowledge Discovery in Databases, Cambridge: AAAI/MIT Press, 1-27.

Gaines, B. R., Shaw, M. L. G. , (1993), Knowledge acquisition tools based on personal construct psychology, The Knowledge Engineering Review, 8, 49-85.

Kantardzic, M., (2003), Data Mining: Concepts, Models, Methods, and Algorithms, Wiley-Interscience.

Neumann, D. A., Norton, V. T. (Jr), (1986), Clustering and isolation in the consensus problem for partitions, Journal of Classification, 3 281-297.

Torra, V., Cortés, U., (1995), Towards an automatic consensus generator tool: EGAC, IEEE Transactions on Systems, Man and Cybernetics, 25 888-894.

Zhong, N., (2003), Mining Interesting Patterns in Multiple Data Sources, in V. Torra, Information Fusion in Data Mining, Springer, 61-77.