Data Mining and Knowledge Discovery, 11, 117–119, 2005

DOI: 10.1007/s10618-005-0009-3

Privacy in Data Mining

Published online: 17 August 2005

The widespread computerization and, especially, the booming use of Internet have enabled in the last years an unprecedented level of automated data collection. Parallel to this, data mining has emerged as an important discipline providing powerful tools for data analysis. Beyond the positive consequences of higher information accuracy, a negative point is an Orwellian feeling of dwindling privacy for individual persons (or companies, for that matter).

Privacy in administrative, statistical and other databases is about finding tradeoffs between the societal *right to know* and the individual *right to private life*. Thus, the passive subject of privacy is the individual citizen or, in business data collection, the individual company.

Several disciplines have been active subjects in studying privacy:

Statistics: Most national statistical laws contain commitments to respondents' privacy, which are essential to encourage citizens' response;
Philosophy: The ethics of information society regards privacy as a value to be defended;
Computer science: Privacy is regarded as part of data security.

In recent times, data mining has also endeavored to become compatible with privacy. The seminal paper along this line was probably (Agrawal and Srikant, 2000), which introduced the notion of privacy-preserving data mining (PPDM). The so-to-speak culture of PPDM stems from the computer science privacy tradition commenced in the 1970s by Schlörer and Denning (Schlörer, 1975; Denning et al., 1979), among others. However, data mining actually draws on both computer science and statistics. So, PPDM might actually benefit from looking at the statistical privacy tradition dating as far back as 1974, when Dalenius published (Dalenius, 1974). It is interesting to note that, while the computer science literature on privacy is extremely scarce in the period 1985–1999, the statistical privacy literature continued alive through the 1980s and substantially increased from 1996 on (see Willenborg and DeWaal (2001) for a review of statistical disclosure control).

Thus, it makes a lot of sense to bring together computer science and statistics to foster the development of PPDM. Such is the goal of the present special issue. We next describe the four papers in it.

The first paper (Bertino et al., 2005) is a computer science-oriented contribution to PPDM. The authors describe an evaluation framework for estimating and comparing different kinds of PPDM algorithms. Their goal is to determine which PPDM techniques are best to protect sensitive information; to that end, quality and privacy measures must be defined.

The proposed framework is then applied to evaluate a specific set of algorithms aimed at association rule hiding. The paper makes a connection with previous evaluation initiatives undertaken by the statistical community.

The second paper (Fienberg and Slavkovic, 2005) analyzes the problem of confidentiality in categorical statistical databases when association rules are to be preserved. Interestingly enough, this is a paper written by statisticians which tackles a data mining topic usually dealt with by computer scientists (association rule release). The authors treat association rules as conditional tables and apply the large existing body of knowledge on tabular data confidentiality. They consider the inferences that an intruder can make on confidential categorical data using the information released on one or several association rules.

Statistical information loss caused by PPDM is the topic of the third paper (Mateo-Sanz et al., 2005). Any algorithm for privacy of individual data should try to reach an optimum tradeoff between data quality (low information loss) and privacy protection (low disclosure risk). While risk measures are naturally bounded between 0 and 1, information loss measures often are not, which makes it awkward to combine both types of measures to achieve an optimal tradeoff. The paper proposes to use probabilities to define [0, 1] bounded information loss measures for any statistic of interest. The resulting probabilistic measures of data quality are complementary to those described in previous literature and in the Bertino et al. paper.

Finally, the fourth paper (Domingo-Ferrer and Torra, 2005) is about disclosure risk in PPDM. More specifically, the paper deals with *k*-anonymity, which is a useful concept to manage the conflict between data quality and individual privacy. *k*-anonymity had been previously defined in the literature as a situation in which, for each combination of values of quasi-identifiers in a database, at least *k* records exist sharing that combination, so that individual re-identification is prevented. The practical problem is that the methods proposed in the literature to achieve *k*-anonymity are ill-suited for continuous (numerical) data. The authors propose a unified approach to *k*-anonymity based on microaggregation which suits any kind of data (categorical and continuous).

We hope and wish that initiatives like this special issue can encourage joint work between computer science and statistics to the greater progress of privacy-preserving data mining.

Josep Domingo-Ferrer
Department of Computer Engineering and Maths
Rovira i Virgili University of Tarragona
Av. Països Catalans 26
E-43007, Tarragona, Catalonia
josep.domingo@urv.net

Vicenç Torra
Institut d'Investigació en Intel·ligència Artificial-CSIC
Campus UAB
E-08193, Bellaterra, Catalonia
vtorra@iiia.csic.es

References

Agrawal, R. and Srikant, R. 2000. Privacy preserving data mining. In Proceedings of the ACM SIGMOD, ACM, pp. 439–450.

Bertino, E., Fovino, I.N., and Provenza, L.P. 2005. A framework for evaluating privacy preserving data mining algorithms. Data Mining and Knowledge Discovery. This issue.

Dalenius, T. 1974. The invasion of privacy problem and statistics production: An overview. Statistik Tidskrift, 12:213–225.

Denning, D.E., Denning, P.J., and Schwartz, M.D. 1979. The tracker: A threat to statistical database security. ACM Transactions on Database Systems, 4(1):76–96.

Domingo-Ferrer, J. and Torra, V. 2005. Ordinal, continuous and heterogenerous k-anonymity through microaggregation. Data Mining and Knowledge Discovery. This issue.

Fienberg, S.E. and Slavkovic, A.B. 2005. Preserving the confidentiality of categorical statistical data bases when releasing information for association rules. Data Mining and Knowledge Discovery. This issue.

Mateo-Sanz, J.M., Domingo-Ferrer, J., and Sebé, F. 2005. Probabilistic information loss measures in confidentiality protection of continuous microdata. Data Mining and Knowledge Discovery. This issue.

Schlörer, J. 1975. Identification and retrieval of personal records from a statistical data bank. Methods Inform. Med., 14(1):7–13.

Willenborg, L. and DeWaal, T. 2001. Elements of Statistical Disclosure Control. New York:Springer-Verlag.