Legal and Technical Issues of Privacy Preservation in Data Mining

Kirsten Wahlstrom
University of South Australia, Australia

John F. Roddick
Flinders University, Australia

Rick Sarre
University of South Australia, Australia

Vladimir Estivill-Castro
Griffith University, Australia

Denise de Vries
Flinders University, Australia

INTRODUCTION

To paraphrase Winograd (1992), we bring to our communities a tacit comprehension of right and wrong that makes social responsibility an intrinsic part of our culture. Our **ethics** are the moral principles we use to assert social responsibility and to perpetuate safe and just societies. Moreover, the introduction of new technologies can have a profound effect on our ethical principles. The emergence of very large databases, and the associated automated data analysis tools, present yet another set of ethical challenges to consider.

Socio-ethical issues have been identified as pertinent to data mining and there is a growing concern regarding the (ab)use of sensitive information (Clarke, 1999; Clifton et al., 2002; Clifton and Estivill-Castro, 2002; Gehrke, 2002). Estivill-Castro et al., discuss surveys regarding public opinion on personal privacy that show a raised level of concern about the use of private information (Estivill-Castro et al., 1999). There is some justification for this concern; a 2001 survey in InfoWeek found that over 20% of companies store customer data with information about medical profile and/or customer demographics with salary and credit information, and over 15% store information about customers' legal histories.

BACKGROUND

Data mining itself is not ethically problematic. The ethical and legal dilemmas arise when mining is executed over data of a personal nature. Perhaps the most immediately apparent of these is the invasion of privacy. Complete **privacy** is not an inherent part of any society because participation in a society necessitates communication and negotiation, which renders absolute privacy unattainable. Hence, individual members of a society develop an independent and unique perception of their own privacy. Privacy therefore exists within a society only because it exists as a perception of the society's members. This perception is crucial as it partly determines whether, and to what extent, a person's privacy has been violated.

An individual can maintain their **privacy** by limiting their accessibility to others. In some contexts, this is best achieved by restricting access to personal information. If a person considers the type and amount of information known about them to be inappropriate, then they perceive their privacy to be at risk. Thus, privacy can be violated when information concerning an individual is obtained, used, or disseminated, especially if this occurs without their knowledge or consent.

Huge volumes of detailed personal data are regularly collected and analysed by marketing applications (Fienberg, S. E. 2006; Berry and Linoff, 1997), in

which individuals may be unaware of the behind-the-scenes use of data, are now well documented (John, 1999). However, privacy advocates face opposition in their push for legislation restricting the secondary use of personal data, since analysing such data brings collective benefit in many contexts. DMKD has been instrumental in many scientific areas such as biological and climate-change research and is also being used in other domains where privacy issues are relegated in the light of perceptions of a common good. These include genome research (qv. (Tavani, 2004)), combating tax evasion and aiding in criminal investigations (Berry and Linoff, 1997) and in medicine (Roddick et al., 2003).

As **privacy** is a matter of individual perception, an infallible and universal solution to this dichotomy is infeasible. However, there are measures that can be undertaken to enhance privacy protection. Commonly, an individual must adopt a proactive and assertive attitude in order to maintain their privacy, usually having to initiate communication with the holders of their data to apply any restrictions they consider appropriate. For the most part, individuals are unaware of the extent of the personal information stored by governments and private corporations. It is only when things go wrong that individuals exercise their rights to obtain this information and seek to excise or correct it.

MAIN FOCUS

Data Accuracy

Mining applications involve vast amounts of data, which are likely to have originated from diverse, possibly external, sources. Thus the quality of the data cannot be assured. Moreover, although data pre-processing is undertaken before the execution of a mining application to improve data quality, people conduct transactions in an unpredictable manner, which can cause personal data to expire. When mining is executed over expired data inaccurate patterns are more likely to be revealed.

Likewise, there is a great likelihood of errors caused by mining over poor quality data. This increases the threat to the data subject and the costs associated with the identification and correction of the inaccuracies. The fact that data are often collected without a preconceived hypothesis shows that the data analysis used in DMKD are more likely to be **exploratory** (as opposed to the **confirmatory analysis** exemplified by many statistical techniques). This immediately implies that results from DMKD algorithms require further confirmation and/or validation. There is a serious danger of inaccuracies that cannot be attributed to the algorithms *per se*, but to their exploratory nature.

This has caused some debate amongst the DMKD community itself. Freitas (2000) has argued that mining association rules is a deterministic problem that is directly dependent on the input set of transactions and thus association rules are inappropriate for prediction, as would be the case of learning classifiers. However, most uses of association rule mining are for extrapolation to the future, rather than characterisation of the past.

The sharing of corporate data may be beneficial to organisations in a relationship but allowing full access to a database for mining may have detrimental results. The adequacy of traditional database security controls are suspect because of the nature of inference. Private and confidential information can be inferred from public information.

The following measures have thus been suggested to prevent unauthorised mining:

- **Limiting access to the data.** By preventing users from obtaining a sufficient amount of data, consequent mining may result in low confidence levels. This also includes query restriction, which attempts to detect when compromise is possible through the combination of queries (Miller and Seberry, 1989).
- **Anonymisation.** Any identifying attributes are removed from the source dataset. A variation on this can be a filter applied to the ruleset to suppress rules containing identifying attributes.
- Dynamic **Sampling**. Reducing the size of the available data set by selecting a different set of source tuples for each query.
- **Authority control** and cryptographic techniques. Such techniques effectively hide data from unauthorised access but allow inappropriate use by authorised (or naive) users (Pinkas, 2002).
- **Data perturbation**. Altering the data, by forcing aggregation or slightly altering data values, useful mining may be prevented while still enabling the planned use of the data. Agrawal and Srikant (2000) explored the feasibility of privacy-preservation by using techniques to perturb sensitive values in data.