Predicting Data that People Refuse to Disclose

Bart Custers
eLaw – Center for Law and Digital Technologies
Leiden University
The Netherlands

Abstract: Data mining technologies are very good at predicting missing data. Datasets that are partially incomplete or incorrect can be completed or corrected by predicting characteristics. Completing or correcting data implies ascribing new characteristics to people. However, sometimes data are missing exactly because people refuse to disclose particular data, especially when these data are sensitive personal data. In general, people have a right to refuse disclosure of their personal data. This perspective on privacy, focusing on individual autonomy, is usually referred to as informational self-determination. When data mining technologies can easily deduce or predict missing data within slight margins of error, this undermines the right to informational self-determination. In fact, using predictions to fill in blanks may decrease the accuracy of the data, as predicted data may be less accurate than data provided by data subjects. As a result, paradoxically, people may be inclined to provide the personal data themselves.

Keywords: Big Data, informational self-determination, privacy, data mining

Introduction

Data mining technologies are useful tools for profiling, i.e., ascribing characteristics to individuals or groups of people.[1,2,3] Such profiles, in turn, may be useful for decision-making, selecting target groups, personalisation, etc.[4] Furthermore, most data mining technologies are very good at dealing with datasets that are incomplete or incorrect. Missing data generally do not constitute a problem when searching for patterns, as long as the total amount of missing data is not too large compared to the amount of data available. Hence, with the help of data mining predictions, the blanks (missing data) can be completed in datasets. Similarly, by predicting characteristics that are available, statements can be made about the accuracy of these data. Such predictions may show that the available characteristics are probably incorrect and can subsequently be corrected.
However, sometimes data is missing because people did not want to provide such data. Data subjects, i.e., the people the data in databases relate to, may have good reasons not to provide particular data. For instance, people may consider such information not to be someone else's business, they may consider disclosure as not improving their reputation, or they may fear disadvantageous judgments of others about themselves. Some information may not be considered appropriate for disclosure to anyone, but more often information may not be considered appropriate for disclosure to particular people or institutions. For instance, people may want to share medical information with their physician and their hospital, but not with their car insurance company, employer or supermarket. People may want to discuss their sexual preferences with friends, but not with their parents. Such a partitioning of social spheres is referred to as audience segregation.[5] In short, people may prefer that others who collect, process, analyse data have some blanks on them in their databases.

Informational Self-Determination

In fact, to some extent, people even have a right to refuse disclosure of their personal information. Everyone has a right to privacy, according to Article 12 of the Universal Declaration of Human Rights. What this right to privacy exactly means, is not entirely clear. When it comes to informational privacy (contrary to, for instance,

spatial privacy) a commonly used definition (particularly in the United States) is that of Westin, who refers to privacy in terms of control over information.[6] Privacy is a person's right to determine for himself when, how, and to what extent information about him is communicated to others. This definition is sometimes referred to as *informational self-determination* and has a strong focus on the autonomy of the individual.[7]
The traditional ways of collecting personal data are either directly, i.e., by asking data subjects for the data, or indirectly, for instance, by buying datasets or coupling databases. Predicting missing data is also a way of indirect data collection. This is rather new, as data mining technologies allow this way of indirect data collection on a large scale.

European Data Protection Legislation

Current European legislation protects collecting and processing personal data, but not the collecting and processing of anonymous data. For this reason, data controllers may prefer to process anonymous data, which allows profiling on an aggregate (group) level. Despite false negatives and false positives, such profiles may be sufficiently accurate for decision-making.[8] The characteristics may be valid for the group members even though they may not be valid for the individuals group members as such.[9] Predicting that people driving white cars cause less traffic accidents on average or predicting that people who refrain from eating peanut butter live longer on average may be (hypothetical) data mining results based on anonymous databases.
Ascribing an anonymous profile to a data subject (if John drives a white car, then he is likely to be a careful driver, or if Sue regularly eats peanut butter, then she is likely to live long), implies ascribing personal data to individuals. This process creates new personal data. These personal data, contrary to those data that a data subject voluntarily provided to a data controller, are much more difficult to get to know for a data subject (both the existence of such data and their contents). In fact, characteristics may be attributed to people that they did not know about themselves (such as life expectancies or credit default risks). People may be grouped with other individuals unknown to them.[10]
This process may seem harmless, but may be considered less harmless to the individuals involved when information is combined and used to predict or deduce, with slight margins of error, particular sensitive data. Furthermore, predicting or deducing missing values and subsequently ascribing them to individuals provides friction with informed consent from those individuals. In Europe, in many cases (though not always), data subjects have a right to consent to the use of their data. When people do not know the ways in which their personal data are processed, which characteristics are ascribed to them, and what are the consequences of this, it is very difficult for them to object.

Redlining

In the past, some financial institutions drew red lines on maps around entire neighbourhoods they deemed off-limits for loans. This practice, known as "redlining", is now strictly illegal.
Discrimination on the basis of race, sex, marital status, etc. is illegal, but because of group profiling, these characteristics may be linked to trivial characteristics, such as underwriting based on geographic location or credit history.

Discrimination Issues

Under discrimination laws, several characteristics are considered unacceptable for decision-making. For instance, ethnic background or gender should not be used to select job applicants. However, everyone knows that a trivial attribute like a name can often predict the ethnicity or gender of a person. The same may be true for attributes like profession (there are still very few female airline pilots or males working as an obstetrician) or zip code (some neighbourhoods are predominantly 'black' whereas others are predominantly 'white').
The use of data mining may further increase the possibilities of predicting sensitive characteristics. From a legal perspective, no employer looking for a new employee is allowed to ask for these characteristics and no job applicant has to provide them, but it is obvious that anti-discrimination legislation is extremely difficult to

enforce nevertheless. The point here is that hiding particular characteristics is not sufficient. In fact, research has shown that leaving out sensitive data like ethnic background and gender out of a database may still yield discriminatory data mining results.[11,12]

Solutions in Code

Are we running out of solutions? Using anonymous data is not really an option, as data may sooner or later be ascribed to individuals again.[13] In fact, when identifying characteristics, such as name, address, social security number, are missing, data mining technologies and database coupling may also be used to predict the missing characteristics. Deleting sensitive data from databases does not work either, as these sensitive characteristics may also be predicted. Prohibiting data mining at all (a radical measure) is not realistic with the enormous amounts of data we are facing in our information society, as it would imply less insight in and overview of the data available.

There are some (rather advanced) solutions, however.[14] These solutions require combining technological measures and legal measures. From a legal perspective it may be recommendable not to focus on (a priori) access limiting measures regarding input data, but rather focus on (a posteriori) accountability and transparency.[15] Instead of limiting access to data, which is increasingly hard to enforce in a world of automated and interlinked databases and information networks, rather the question how data can and may be used is stressed

But this requires also technological measures.[16] For instance, the architecture of data mining technologies can be adjusted ('solutions in code')[17] to create a value-sensitive design, that incorporates legal, ethical and social aspects in the early stages of development of these technologies.[18] This is exactly what privacy preserving data mining techniques aim at.[19] These may aim at protecting identity disclosure or attribute disclosure, but also at prevention or protection of the inferred data mining results. Similarly, discrimination-free data mining techniques have been developed, by integrating legal and ethical in data mining algorithms.[20]

Conclusion

Even though people may not want to disclose particular personal data to others, it may be possible to predict these data. Predictions can be based on other data available on these individuals and data available on the groups to whom they belong. Data mining technologies may be very useful to complete missing parts of the data in large datasets. However, when people explicitly refuse to disclose these data, predicting the missing characteristics may challenge their right to informational self-determination. Hence, their privacy and autonomy may be infringed.

The predictions of missing values usually contain margins of error. When the data mining results are used for decision-making, the decisions may contain errors as well. Since the predicted data may be less accurate than the data provided by individuals, people may be inclined to provide the (more correct) data themselves, to ensure (more) just decisions. This is a privacy paradox.[21]

Bibliography

1 Frawley, W.J., Piatetsky-Shapiro, G. and Matheus, C.J. (1993) Knowledge Discovery in Databases; an overview, In: *Knowledge Discovery in Databases,* G. Piatetsky-Shapiro and W.J. Frawley (eds.) Menlo Park, California: AAAI Press / The MIT Press.

2 Adriaans, P. and Zantinge, D. (1996) *Data mining*, Harlow, England: Addison Wesley Longman.

3 Fayyad, U-M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R. (1996) *Advances in knowledge discovery and data mining*. Menlo Park, California: AAAI Press / The MIT Press.

4 Even though a digital person is only a limited representation of a real person. Solove, D. (2004) *The Digital Person; Technology and Privacy in the Information Age*, New York: University Press.

5 Van den Berg, B. and Leenes, R. (2010) *Audience Segregation in Social Network Sites*, In: *Proceedings for SocialCom2010/PASSAT2010* (Second IEEE International Conference on Social Computing/Second IEEE International Conference on Privacy, Security, Risk and Trust). Minneapolis (Minnesota, USA): IEEE: 1111-1117.

6 Westin, A. (1967) *Privacy and Freedom*. London: Bodley Head.

7 Other common definitions of the right to privacy are the right to be let alone, see Warren and Brandeis (1890) and the right to respect for one's private and family life (Article 8 of the European Convention on Human Rights and Fundamental Freedoms).

8 Zarsky, T. Z. (2003) "Mine Your Own Business!": Making the Case for the Implications of the Data Mining of Personal Information in the Forum of Public Opinion. *Yale Journal of Law & Technology*, 5, pp. 56.

9 Vedder, A. H. (1999) KDD: The Challenge to Individualism, In: *Ethics and Information Technology*, Nr. 1, p. 275-281.

10 For more on risks of profiling, see Schermer, B.W. (2011) The Limits of privacy in automated profiling and data mining. *Computer Law & Security Review*, Volume 27, Issue 7, p. 45-52.

11 Verwer and Calders. (2010) Three Naive Bayes Approaches for Discrimination-Free Classification, in: *Data Mining*: *special issue with selected papers from ECML-PKDD 2010;* Springer

12 Pedreschi, D., Ruggieri, S., and Turini F. (2008) *Discrimination-aware Data Mining*. 14th ACM International Conference on Knowledge Discovery and Data Mining (KDD 2008): 560-568. ACM, August 2008.

13 Ohm, P. (2010) Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization, *UCLA Law Review*, Vol. 57, p. 1701-1765.

14 Custers B.H.M. (2010) *Data Mining with Discrimination Sensitive and Privacy Sensitive Attributes*. Proceedings of ISP 2010, International Conference on Information Security and Privacy, 12-14 July 2010, Orlando, Florida.

15 Weitzner, D.J., Abelson, H. et al. (2006) *Transparent Accountable Data Mining: New Strategies for Privacy Protection*. MIT Technical Report, Cambridge: MIT.

16 For more information, see also: http://wwwis.win.tue.nl/~tcalders/dadm/doku.php

17 Lessig, L. (2006) *Code Version 2.0*, New York: Basic Books.

18 Friedman, B., Kahn, P.H., Jr., and Borning, A. (2006). Value Sensitive Design and information systems. In: P. Zhang and D. Galletta (eds.) *Human-Computer Interaction in Management Information Systems: Foundations*, Armonk, New York; London, England: M.E. Sharpe, pp. 348–372.

19 Lindell, Y., and Pinkas, B. (2002) Privacy preserving data mining, *Journal of Cryptology*, no. 15, p. 177-206.

20 Calders, T., and Verwer, S. (2010) *Three Naive Bayes Approaches for Discrimination-Free Classification*. Special issue of ECML/PKDD.

21 Custers, B.H.M. (2004) *The Power of Knowledge* Tilburg: Wolf Legal Publishers, p. 157.