Research Paper Engineering

Security Aspects in Knowledge Discovery and Data Mining

N.C Ramakrishna | IVth ECM, SNIST, Hyderabad

D. Rohit | IVth ECM, SNIST, Hyderabad

ABSTRACT

Recent developments in information technology have enabled collection and processing of vast amounts of personal data, such as criminal records, shopping habits, credit and medical history, and driving records. This information is undoubtedly very useful in many areas, including medical research, law enforcement and national security. However, there is an increasing public concern about the individuals' privacy. Privacy is commonly seen as the right of individuals to control information about them. The appearance of technology for Knowledge Discovery and Data Mining (KDDM) has revitalized concern about the following general privacy issues: secondary use of the personal information, handling misinformation, and granulated access to personal information. They demonstrate that existing privacy laws and policies are well behind the developments in technology, and no longer offer adequate protection.

KEYWORDS information Technology; IT, KDD, DM, Privacy,Threat

1. Introduction

Not surprisingly, data is treated today as one of the most important corporate assets of companies, governments and research institutions supporting fact-based decision making. It is possible to have fast access, to correlate information stored in independent and distant databases, to analyze and visualize data on-line and use data mining tools for automatic and semi-automatic exploration and pattern discovery. Knowledge Discovery and Data Mining (KDDM) is an umbrella term describing several activities and techniques for extracting information from data and suggesting patterns in very large databases. Marketing applications have adopted and expanded KDDM techniques.

Now, KDDM is moving to other domains where privacy issues are very delicate. Recent developments in information technology have enabled collection and processing of vast amounts of personal data, such as criminal records, shopping habits, credit and medical history, and driving records. This information is undoubtedly very useful in many areas, including medical research, law enforcement and national security. For the analysis of crime data, KDDM techniques have been applied by the FBI in the US as a part of the investigation of the Oklahoma City bombing, the Unabomber case, and many lower-profile crimes. Another example is the application of KDDM to analyzing medical data. Despite its benefits to social goals, KDDM applications inspire reservations. Individuals easily imagine the potential misuses from unauthorized tapping into financial transactions or medical records. There is an increasing public concern about the individuals' privacy. Surveys reveal growing concern about the use of personal information.

The data was so detailed that it generated strong public opposition and Lotus abandoned the project. However, this mostly affected small business, as large companies already had access and continued to use Lotus data sets [8]. At least 400 million credit records, 700 million annual drug records, 100 million medical records and 600 million personal records are sold yearly in the US by 200 super bureaus. Among the records sold are bank balances, rental histories, retail purchases, criminal records, unlisted phone numbers and recent phone calls. Combined, the information helps to develop data images of individuals that are sold to direct marketers, private individuals, investigators, and government organizations.

These data images are now the subject of analysis by automatic and semiautomatic Knowledge Discovery and Mining tools. We address revitalized general privacy issues and new threats to privacy by the application of KDDM. We distinguish them from threats to privacy or security resulting from the expansion of computer networks and on-line distributed information systems.

2. Privacy Threats

2.1 Use of the Personal Data

As we pointed out, recent surveys on privacy show a great concern about the use of personal data for purposes other than the one for which data has been collected. An extreme case occurred in 1989. Despite collecting over $16 million USD by selling the driver-license data from 19.5 million Californian residents, the Department of Motor Vehicles in California revised its data selling policy after Robert Brado used their services to obtain the address of actress Rebecca Schaeffer and later killed her in her apartment. While it is very

Unlikely that KDDM tools will reveal directly precise confidential data, the exploratory KDDM tools may correlate or disclose confidential, sensitive facts about individuals resulting in a significant reduction of possibilities. In fact, this is how they were applied in the investigation of the Unabomber case and other criminal investigations. They facilitated filtering large volumes of reports from informants so resources could be concentrated on much fewer promising leads and suspects. Thus, we would not expect that detailed personal addresses would be disclosed by a KDDM analysis.

A simple application of Link Analysis can correlate phone records and banking records to determine, with a certain degree of accuracy, if bank customers have a fax machine at home and how this impacts the likelihood of accepting offers on equity loans. Most individuals consider the use of information beyond its initial collection for secondary analysis a direct invasion of privacy, and perhaps even more if this reveals aspects like what a person has inside its home. Individuals understand that phone companies need to monitor length of phone calls for billing purposes, and that their bank must keep track of transactions in their accounts, but when this data is used for secondary purposes for which the individual neither has nor provided authorization, a serious issue in privacy appears.

2.2 Usage of Misinformation

Misinformation can cause serious and long-term damage, so individuals should be able challenge the correctness of data about themselves. For example, District Cablevision in Wash-

ington fired James Russell Wiggings, on the basis of information obtained from Equifax, Atlanta, about Wiggings' conviction for cocaine possession; the information was actually about James Ray Wiggings, and the case ended up in court.

This illustrates a serious issue in defining property of the data containing personal records. While individuals and legislators supporting the right of privacy favor the view that a person's data is the person property, data collectors favor the view that the data collector owns the data. This ethical issue has been illustrated by some cases of celebrities and other public figures that have been able to obtain rights on reproduction of their photographed image. However, for the average citizen, the horizon is not so promising.

2.3 Refined Access to Personal Data

The access to personal data should be on a need-to-know basis, and limited to relevant information only. For example, employers are obliged to perform a background check when hiring a worker but it is widely accepted that information about diet and exercise habits should not affect hiring decisions. There also seem to be two directions in this issue. Some advocate the removal of fields and even prohibiting their collection. Others support the release for very detailed data so research and analysis can advance. The new privacy laws in Germany illustrate the first approach. These regulations have dramatically reduced the number of variables in the census and the micro census. The second approach is illustrated by personal data placed on large on-line networked databases, like the Physician Computer Network in the US, with the intention to build and expand knowledge. Another example of this approach are more precise geo-referenced data sets in Geographical Information Systems and their databases that include cadastral data, aerial photography of individual properties and detailed features of private properties.

3. Novel Privacy Threats

We discuss new privacy threats posed by Knowledge Discovery and Data Mining (KDDM), which includes massive data collection, data warehouses, statistical analysis and deductive learning techniques. KDDM uses vast amounts of data to generate hypotheses and discover general patterns.

3.1 Guarding personal data from KDDM researchers

How planning decisions could be taken, if census data was not collected? How could epidemics be understood if medical records were not analyzed? Individuals benefit from data collection efforts in the process of building knowledge that guides society. The protection of privacy cannot simply be achieved by restricting data collection or restricting the use of computer and networking technology. Researchers feel that privacy regulations will enforce so many restrictions on data, that it would make the data useless.

3.2. Individuals from training Sets

The classification task in KDDM takes as input a set of cases and their classes (training set); the output is a classifier, that is, an operator that assigns classes to new, unclassified cases. For example, the cases may correspond to patients and classes to diagnoses.

- The first problem is how to provide the analyst with KDDM tools a training set. If such a set is provided from real data, then each record of a training set is a disclosure of the information of the individual corresponding to the record.

- The second problem is how to protect privacy if somebody has a classifier and a record that knows belongs to the training set that built the classifier, but does not have the class.

The KDDM classifiers are typically very accurate when applied to cases from the training set. Thus a classifier and knowledge that case A is in the training set reveals the class of case A. In this paper we argue that a classifier should be modified in such a way so as to have similar accuracy when applied to the cases from the training set, as when applied to the new cases.

When applied to KDDM, query restriction techniques may deny some particularly important information and obscure general patterns. A recent proposal has been based on this. The idea here is to supply a subset of the data so restricted that is not useful for the data miner. This has been criticized since, first, why would a miner will acquire or investigate data guaranteed not to have anything useful, second, the only way to guarantee that the data set is really contains no patterns is to find them all (which requires infinite computational time or to provide a very small set, and third, for this scheme to work, it is assumed that each miner will not cooperate with other miners (and in particular, that nobody gains access to more data).

4. Conclusion

KDDM revitalizes some issues and posses new threats to privacy. Some of these can be directly attributed to the fact that these powerful techniques may enable the correlation of separate data sets in other to significantly reduce the possible values of private information. Other can be more attributed to the interpretation, application and actions taken from the inferences obtain with the tools. While this raises concerns, there is a body of knowledge in the field of statistical databases that could potentially be extended and adapted to develop new techniques to balance the rights to privacy and the needs for knowledge and analysis of large volumes of information. Some of these new privacy protection methods are emerging as the application of KDD tools moves to more controversial datasets.

REFERENCES

[1] N.R. Adam and D.H. Jones. "Security of statistical | databases with an output perturbation technique". Journal | of Management and Information Systems, 6, (1): 101--110, | 1989. | [2] N.R. Adam and J.C. Wortmann. "Security-control methods | for statistical databases: A comparative study. ACM | Computing Surveys, 21(4): 515--556, 1989. | [3] M.J.A. Berry and G.Linoff. "Data Mining Techniques --- | for Marketing, Sales and Customer Support. John Wiley & | Sons, NY, USA, 1997. | [4] A. Berson and S.J. Smith. "Data Warehousing, Data | Mining, & OLAP". Series on Data Warehousing and Data | Management. McGraw-Hill, NY, USA, 1998. | [5] S. Bonorris. "Cautionary notes for the automated | Processing of data". IEEE Expert, 10(2): 53--54, April | 1995. | [6] C. Clifton and D. Marks. "Security and privacy | implications of data mining". In SIGMOD workshop on | Data Mining and Knowledge Discovery, Montreal, | Canada, June 1996. ACM. | [7] Private lives; public records --- in todays networked world | is personal information to easy to access?". | Computerworld, 31(6): 81-90, September 1997. | [8] M. J. Culnan. "``How did they get my name?' ': An | exploratory investigation of consumer attitudes towards | secondary information use". MIS Quarterly, 17: 341— | 361, 1993. |