12

Major and Privacy Issues in Data Mining and Knowledge Discovery

Objectives:

- In this section, the major data mining issues, and OECD personal privacy guidelines, some privacy issues in knowledge discovery and data mining are enumerated.
- The scope of this section addresses major issues in data mining regarding mining methodology, user interaction, performance, and diverse data types.
- We discuss new privacy threats posed by knowledge discovery and data mining (KDDM).
- KDDM poses the following new challenges to privacy: stereotypes, guarding personal data from KDDM researchers, protecting privacy of individuals from training sets, and combination of patterns. We discuss the possible solutions and their impact on the quality of discovered patterns.
- KDDM technology includes massive data collection, data warehouses, statistical analysis and deductive learning techniques, and uses vast amounts of data to generate hypotheses and discover general patterns.
- KDDM is now moving to other domains where privacy issues are very delicate.
- The Organization for Economic Cooperation and Development (OECD) has provided probably the best known set of guidelines. This section focuses on the OECD guidelines since many nations have adopted then as statutory law.
- This section provides some insight into the problems concerning personal privacy and data faced by those who wish to employ knowledge discovery.

Abstract. In this section, the major data mining issues and OECD personal privacy guidelines, some privacy issues in knowledge discovery and data mining are enumerated. The major issues concerned with that of the data mining are individual privacy, which is a social one, issue, related to data integrity and technical issue whether to set up a relational database structure or a multidimensional one.

S. Sumathi and S.N. Sivanandam: *Major and Privacy Issues in Data Mining and Knowledge Discovery*, Studies in Computational Intelligence (SCI) **29**, 271–291 (2006)
www.springerlink.com

The scope of this section addresses major issues in data mining regarding mining methodology, user interaction, performance, and diverse data types.

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. The information is undoubtedly very useful in many areas, including medical research, law enforcement, and national security. However, there is an increasing public concern about 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 concerns about the following general privacy issues: secondary use of the personal information, handling misinformation, and granulated access to personal information. These issues demonstrate that existing privacy laws and policies are well behind the developments in technology, and no longer offer adequate protection. We discuss new privacy threats posed by KDDM. KDDM technology includes massive data collection, data warehouses, statistical analysis and deductive learning techniques, and uses vast amounts of data to generate hypotheses and discover general patterns. KDDM poses the following new challenges to privacy: stereotypes, guarding personal data from KDDM researchers, protecting privacy of individuals from training sets, and combination of patterns. We discuss the possible solutions and their impact on the quality of discovered patterns.

The Organization for Economic Cooperation and Development (OECD) has provided probably the best known set of guidelines. A number of countries have adopted these guidelines as statutory law, in whole or in part.

12.1 Major Issues in Data Mining

One of the key issues raised by data mining technology is not a business or technological one, but a social one. It is the issue of individual privacy. Data mining makes it possible to analyze routine business transactions and glean a significant amount of information about individuals' buying habits and preferences.

Another issue is that of data integrity. Clearly, data analysis can only be as good as the data that is being analyzed. A key implementation challenge is integrating conflicting or redundant data from different sources. For example, a bank may contain credit cards accounts on several different databases. The addresses (or even the names) of a single cardholder may be different in each. Software must translate data from one system to another and select the address most recently entered.

A hotly debated technical issue is whether it is better to set up a relational database structure or a multidimensional one. In a relational structure, data is stored in tables, permitting ad hoc queries. In a multidimensional structure, on the other hand, sets of cubes are arranged in arrays, with subsets created according to category. While multidimensional structures facilitate multidimensional data mining, relational structures thus far have performed better in client/server environments. And, with the explosion of the Internet, the world is becoming one big client/server environment.

Finally, there is the issue of cost. While system software costs have dropped dramatically within the past five years, data mining and data warehousing

tend to be self-reinforcing. The more powerful the data mining queries, the greater the utility of the information being gleaned from the data, and the greater the pressure to increase the amount of data being collected and maintained, which increases the pressure for faster, more powerful data mining queries. This increases pressure for larger, faster systems, which are more expensive.

The scope of this section addresses major issues in data mining regarding mining methodology, user interaction, performance, and diverse data types. These issues are introduced below:

Mining methodology and user interaction issues: These reflect the kinds of knowledge mined, the ability to mine knowledge at multiple granularities, the use of domain knowledge, ad hoc mining, and knowledge visualization.

- *Mining different kinds of knowledge in databases:* Since different users can be interested in different kinds of knowledge, data mining should cover a wide spectrum of data analysis and knowledge discovery tasks, including data characterization, discrimination, association, classification, clustering, trend and deviation analysis, and similarity analysis. These tasks may use the same database in different ways and require the development of numerous data mining techniques.
- *Interactive mining of knowledge at multiple levels of abstraction:* Since it is difficult to know exactly what can be discovered within a database, the data mining process should be *interactive.* For databases containing a huge amount of data, appropriate sampling techniques can first be applied to facilitate interactive data exploration. Interactive mining allows users to focus the search for patterns, providing and refining data mining requests based on returned results. Specifically, knowledge should be mined by drilling down, rolling up, and pivoting through the data space and knowledge space interactively, similar to what OLAP can do on data cubes. In this way, the user can interact with the data mining system to view data and discovered patterns at multiple granularities and from different angles.
- *Incorporation of background knowledge:* Background knowledge, or information regarding the domain under study, may be used to guide the discovery process and allow discovered patterns to be expressed in concise terms and at different levels of abstraction. Domain knowledge related to databases, such as integrity constraints and deduction rules, can help focus and speed up a data mining process, or judge the interestingness of discovered patterns.
- *Data mining query languages and ad hoc data mining:* Relational query languages (such as SQL) allow users to pose ad hoc queries for data retrieval. In a similar vein, high-level data mining query languages need to be developed to allow users to describe ad hoc data mining tasks by facilitating the specification of the relevant sets of data for analysis, the domain knowledge, the kinds of knowledge to be mined, and the conditions and constraints to be enforced on the discovered patterns. Such a language

should be integrated with a database or a data warehouse query language, and optimized for efficient and flexible data mining.

- *Presentation and visualization of data mining results:* Discovered knowledge should be expressed in high-level languages, visual representations, or other expressive forms so that the knowledge can be easily understood and directly usable by humans. This is especially crucial if the data mining system is to be interactive. This requires the system to adopt expressive knowledge representations techniques, such as trees, tables, rules, graphs, charts, crosstabs, matrices, or curves.
- *Handling noisy or incomplete data:* The data stored in a database may reflect noise, exceptional cases, or incomplete data objects. When mining data regularities, these objects may confuse the process, causing the knowledge model constructed to overfit the data. As a result, the accuracy of the discovered patterns can be poor. Data cleaning methods and data analysis methods that can handle noise are required, as well as outlier mining methods for the discovery and analysis of exceptional cases.
- *Pattern evaluation – the interestingness of problem:* A data mining system can uncover thousands of patterns. Many of the patterns discovered may be uninteresting to the given user, representing common knowledge or lacking novelty. Several challenges remain regarding the development of techniques to assess the interestingness of discovered patterns, particularly with regard to subjective measures that estimate the value of patterns with respect to a given user class, based on user beliefs or expectations. The use of interestingness measures to guide the discovery process and reduce the search space is another active area of research.

Performance issues: These include efficiency, scalability, and parallelization of data mining algorithms.

- *Efficiency and scalability of data mining algorithms:* To effectively extract information from a huge amount of data in databases, data mining algorithms must be efficient and scalable. In other words, the running time of a data mining algorithms must be predictable and acceptable in large databases. From a database perspective on knowledge discovery, efficiency and scalability are key issues in the implementation of data mining systems. Many of the issues discussed above under *mining methodology and user interaction* must also consider efficiency and scalability.
- *Parallel, distributed, and incremental mining algorithms:* The huge size of many databases, the wide distribution of data, and the computational complexity of some data mining methods are factors motivating the development of parallel and distributed data mining algorithms. Such algorithms divide the data into partitions, which are processed in parallel. The results from the partitions are then merged. Moreover, the high cost of some data mining processes promotes the need for incremental data mining algorithms that incorporate database updates without having to mine the entire data again "from scratch." Such algorithms perform knowledge

modification incrementally to amend and strengthen what was previously discovered.

Issues relating to the diversity of database types:

- *Handling of relational and complex types of data:* Since relational databases and data warehouses are widely used, the development of efficient and effective data mining systems for such data is important. However, other databases may contain complex data objects, hypertext and multimedia data, spatial data, temporal data, or transaction data. It is unrealistic to expect one system to mine all kinds of data, given the diversity of data types and different goals of data mining. Specific data mining systems should be constructed for mining specific kinds of data. Therefore, one may expect to have different data mining systems for different kinds of data.
- *Mining information from heterogeneous databases and global information systems:* Local- and wide-area computer networks (such as the Internet) connect many sources of data, forming huge, distributed, and heterogeneous databases. The discovery of knowledge from different sources of structured, semistructured, or unstructured data with diverse data semantics poses great challenges to data mining. Data mining may help disclose high-level data regularities in multiple heterogeneous databases that are unlikely to be discovered by simple query systems and may improve information exchange and interoperability in heterogeneous databases. Web mining, which uncovers interesting knowledge about Web contents, Web usage, and Web dynamics, becomes a very challenging and highly dynamic field in data mining.

The above issues are considered major requirements and challenges for the further evolution of data mining technology. Some of the challenges have been addressed in recent data mining research and development, *to a certain extent*, and are now considered *requirements*, while others are still at the research stage. The issues, however, continue to stimulate further investigation and improvement.

12.2 Privacy Issues in Knowledge Discovery and Data Mining

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, correlate information stored in independent and distant databases, analyze and visualize data on-line, and use data mining tools for automatic and semiautomatic 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. KDDM is now 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 an 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. Survey in the US reveal growing concern about the use of personal information. The newest Equifax/Harris Consumer Privacy Survey shows that over 70% of respondents are against unrestricted usage of their medical data for research purposes. As much as 78% believe that computer technology represents a threat to personal privacy and that the use of computers must be restricted sharply in the future, if privacy is to be preserved. Up to 76% believe they have lost control over their personal information and that their privacy is threatened. Time/CNN reveals that at least 93% of respondents believe companies selling personal data should be required to gain permission from individuals. In another study 96% of respondents believe that private information should never be used for another purposes without permission, and over 20% had personally experienced a privacy invasion. By contrast, in 1970 the same Equifax/Harris Survey found only 33% considered computer technology a threat to their privacy.

Massive amounts or personal data are available. In 1990, Lotus Development Corporation announced the release of a CD-ROM with the data on 100 million households in the US. The data was so detailed that it generated string public opposition and Lotus abandoned the project. However, this mostly affected small business, as large companies already had access and continued to use Lotus datasets. 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 superbureaus. 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 resold 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. In this section 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.

12.2.1 Revitalized Privacy Threats

Secondary Use of the Personal Information

Recent surveys have shown 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 precise confidential data directly, the exploratory KDDM tools may correlate or disclose confidential, sensitive facts about individuals resulting in a significant reduction of possibility. 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 form 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; however, with enough data about the patterns of behavior of young actresses in Los Angeles, the potential addresses may be reduced to a few possibilities making it feasible to visit them all. A simple application of link analysis can correlate phone 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 for secondary analysis a direct invasion of privacy, and perhaps even more if this reveals aspects like what does a person have 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 consider unauthorized secondary use of their data to be a violation of privacy.

Handling Misinformation

Misinformation can cause serious and long-term damage, so individuals should be able to challenge the correctness of data about themselves. For example, District Cablevision in Washington 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's property, data collectors favor the view that the data collector owns the data. These ethical issues have been illustrated by some cases of celebrities and other public figures who have been able to obtain rights on reproduction of their photographed image. However, for the average citizen the

horizon is not so promising. The existing market of personal data postulates that the gathering institution owns the data. However, the attitude of data collectors and marketers toward privacy is more moderate than 20 years ago when marketers believed that there was "too much privacy already."

The reason for this change, apart from the fact that privacy is under much bigger threat now, is probably the fear of massive public opposition and losing the trust of customers. Many data owners acknowledge that there is a "Big Brother" aspect in the exploitation of personal datasets, and data collectors now take some measures to preserve the customers trust. Others imply that the "sinister purpose" of data mining is the "product of junk science and journalistic excess," but nevertheless believe that "marketers should take a proactive stance and work to diffuse the issue before it becomes a major problem,"

Granulated Access to Personal Information

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 seem to be two directions in this issue. Some advocate the removal of variables 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 microcensus. 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 datasets in geographical information systems and their databases that include cadastral data, aerial photography, and detailed features of private properties.

Scholars form diverse backgrounds in history, sociology, business, and political science have concluded that the existing privacy laws are far behind the developments in information technology and do not protect privacy well even for the issues mentioned above. The OECD [Organization for Economic Cooperation and Development] Principles on Data Collection addresses some of the concerns raised by the issues in this section. However, only 24 countries have adopted them, thus far with varying degrees. Twelve nations have adopted all OECD's principles in statutory law. Australia, Canada, New Zealand, and the US do not offer protection to personal data handled by private corporations. In Australia, the Privacy Act 1988 predates the growing on-line purchasing and other massive networked data-collection mechanisms. Australia's Privacy Commissioner, Moira Scollay, has taken steps to simplify privacy regulations and provide a single national framework for data-matching systems such as FLY-BUYS cards. However, so far, she has only proposed for discussion a draft of principles for fair handling of personal information.

12.2.2 New Privacy Threats

Stereotypes

General patterns may be used for guessing confidential properties. Also, they may lead to stereotypes and prejudices. If the patterns are based on properties such as race, gender, or nationality, this issue can be very sensitive and controversial. Examples are debates over studies about intelligence across different races. The issue raises debate because KDDM tools may allow the application of different commercial standards based on race or ethnic group. Banks may use KDDM tools to find a different pattern of behavior between two racial or ethnic groups and then deny credit or apply a different policy based on the attribute. An example of different commercialization based on ethnic group occurs in marketing of petrol in Canada because it is well know that French-Canadians are more likely to buy premium petrol than Anglo-Canadians. While these examples seem harmless, in the current situation where more and more business consider data analysis as necessary for ensuring efficiency and competitiveness, wide application of KDDM patterns can easily spread other kinds of stereotypes. For example, data analysis has made it possible for car insurance companies to adjust their premiums based on the age of the driver. It has allowed health insurance companies also adjust their premiums based on some patterns of behavior of the policyholder (for example, the practicing of sports like scuba driving or hand gliding). Stereotypes may have serious implications into the type of data mining technology applied to some task. For example, an inductive learner may be used to create a classifier out of previous credit card applications. If the learner is of the type that can explain why it is denying or granting some credit level, the corporation using it may detect that it is using controversial attributes (i.e., gender, nationality) to perform the credit ranking. However, if the tools represents knowledge in implicit forms that are hard to interpret (a la artificial neural networks), it may be possible that credit assignment is been mainly based on controversial attributes without the awareness of the credit-card company and the application. Many legal issues may be raised if the applicant is to discover this.

Another problem is that KDDM tools use parametric and nonparametric methods from the fields of statistics and machine learning. Their limitations are difficult to explain to a lay person. However, it may be easy to claim that the pattern was derived by a sophisticated computer program, and thus, accepted as truth about individuals. In particular, the fact that they are obtaining rules of thumb and generalization that work in most cases (and are about groups) but are false of each individual is one of the hardest concepts to grasp. This is also true for statistics. One toss of a fair coin will end up in heads or in tail and not in between, but such in between is the "expected value" of such a single event. Similarly, statistics allows us to predict accurately values about populations, like how many smokers will develop cancer this year, but it will not guarantee than a Mr. X will develop cancer, and despite him being a smoker.

Guarding Personal Data From KDDM Researchers

The protection of privacy cannot simply be achieved by restricting data collection or restricting the use of computer and networking technology. How could planning decisions 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 guide's society. Researchers feel that privacy regulations may enforce so many restrictions on data, that it would make the data useless. On the other hand, researchers who apply KDDM tools to datasets containing personal information should not be given access to individual data. But such a restricted access can make KDDM tasks very difficult or even impossible. Thus an appropriate balance between a need for privacy and a need for knowledge discovery should be found.

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 paints and classes to diagnoses. There are two security problems regarding the training set. The first is how to provide the analyst with a training set. If such a set is provided from real data, then each record of a training set is a discloser of the information of the individual corresponding to the record. If the training set contains artificial or perturbed data, then the quality of a classifier may be seriously affected. The second problem is how to protect privacy if somebody has a classifier and a record that is known to belong to the training set that built the classifier. The KDDM classifiers are typically very accurate when applied to case 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 section, we argue that a classifier should be modified in such a way so as to have similar accuracy when applied to the case from the training set, as when applied to the new cases.

Combination of Patterns

Combining two or more general patterns may lead to a disclosure of individual information, either with certainty, or with a high probability. Also, knowledge of totals and other similar facts about the training data may be correlated to facilitate compromising individual values. For example consider a dataset where there are 10 people: 2 females and 8 males; there are 8 cases of disease A, and none of the females has disease A. if it is known that Mr. X's information is part of the data, it is possible to infer that Mr. X has disease A. The problem here is how to protect individual values while preserving values of parameters that are calculated by statistical inference. These parameters usually described

a probabilistic model and their accurate inference hardly reveals anything about individual records. Thus, perturbation to the data that preserve the estimators of statistical inference technique could potentially be sufficient for the problem.

How to hide what can be inferred by all possible methods? This implies in a sense that we discover first everything that is to be discovered about a dataset in order to be sure that we conceal it. Or in other words, to understand how to protect values in individual records, we first must understand very well the capabilities of KDDM tools and technology. We have clearly illustrated that tools for knowledge discovery pose new threat to privacy. However, researchers in KDDM seem to have contradictory views on privacy issues. Some believe that KDDM is not a threat to privacy, since the derived knowledge is only about and from groups. Clearly, this is only true if the identified group is large, but otherwise, information about an individual is partially disclosed. Also, performing inferences about groups does not prevent the creation of stereotypes. Others clearly support our views saying that KDDM deals mainly with huge amounts of microdata. Some fear there would be different academic standards: "statutory limitations that vary from country to country ... suggests that the practice ... varies from country to country." Europe has adopted the OECD directives and investigators across all fields of scholarly research now require "the subject's written consent or data may not be processed." Others think that subject approval may not be sufficient for data miners to refer or disclose incidentally discovered patterns or relationships. That is, how can one ask approval from an individual about something yet unknown?

12.2.3 Possible Solutions

We discuss the possible solutions to the privacy problems imposed by KDDM and their impact on the quality of discovered patterns. The issue of how data may create or reinforce stereotypes is more an issue of the application of the technology than a problem of the technology. It is similar to considering the impact of and biased reporting or political propaganda on large audiences as a disadvantage of TV. This is not a disadvantage of the technology, but a very complex issue of the social structures that allow the production and broadcasting of materials, later distributed through the powerful technology. Thus, the issue of created or reinforced stereotypes with KDDM tools falls in the sociological, anthropological, and legal domain. The interpretation of patterns from KDDM is as difficult as the interpretation of statistics. For example, the times/distance/etc. achieved by female athletes vs. male athletes in Olympic competitions constitute data that statistically distinguishes two groups. When does finding two groups justify creating different policies is not an issue of the technology that discovered the two clusters. In what follows, we concentrate on technological solutions that could alleviate some of the risk to privacy associated with the other issues listed in previous section. Those

threats to privacy listed in previous section are not totally new and there is some understanding of them from pervious attempts to ensure privacy in the facts if the need to perform data analysis. In particular, the field of statistical database has done much progress in this direction and we believe it can provide much insight into new issues like those listed previously.

The field of statistical database has developed methods to guard against the disclosure of individual data while satisfying request or aggregate statistical information. In particular, the experience form this field indicates that removing identifiers such as names, addresses, telephone numbers, and social security numbers is a minimum requirement for privacy but it is enough. Reidentification based on remaining fields may still be possible and removing identifiers is considered the weakest approach, and should never be used on its own. In fact, the information models used in KDDM and statistical database is quite similar. The main objectives of a security control mechanism in a statistical database are to provide statistical users with a sufficient amount of high-quality statistics (statistics quality is measured by the consistency, bias, and precision) and at the same time, to prevent disclosure of confidential individual information. Similar to the context of statistical databases, the objective of a security control mechanism for KDDM is to prevent the disclosure of confidential individual values. However, unlike in statistical databases, another objective is not to minimize the distortion of all statistics, but rather to preserve general patterns; in order words, the objective is to preserve those statistics based on attributes that have a high impact on the patterns and rules to be discovered.

Various techniques have been proposed for security protection in statistical database, but none of them is both effective and efficient. All methods tradeoff privacy of individual values for statistics and/or pattern distortion. The methods can be classified into two main categories, "query restriction" and "noise addition." *Query restriction* techniques provide exact answers to some queries and reject others that may lead to a database compromise. Hence, the statistical quality of released information is high, but these techniques are typically overly restrictive. They are inadequate against skilled users, who have previous knowledge about information stored in the database, and/or require high initial implementation effort. When applied to KDDM, query restriction techniques may deny some particularly important information and obscure general patterns. The idea here is to supply a subset of the data so restricted that is not useful for the data miner. This has been much criticized. First, why would a miner acquire or investigate data guaranteed not to have anything useful? Second, the only way to guarantee that the dataset contains no patterns is to find them all (which require infinite computational time) or to provide a very small set. Third, for this scheme to work, it is assumed that each miner will not co-operate with other miners (and in particular, that nobody gains access to more data). Since the objective of query restriction is not to maximize the number of answerable queries, but rather to answer all, or most of, the important queries partitioning is a potential approach. This requires

the development of methods for grouping record on the basis of their values of important attributes. These attributes may be identified using rough set theory. *Noise addition* techniques prevent compromise by introducing an error either to data or to result of queries. These techniques are robust (resistant to users' supplementary knowledge) and provide answers to all queries, but at the expense of allowing partial disclosure and/or providing information with less statistical quality. O'Leary has suggested that noise addition by itself could provide protection from all those KDDM techniques that are very sensitive to noise. In particular, probability distribution data perturbation techniques referred to as *data swapping*, seems suitable for privacy protection in KDDM. Data swapping refers to interchanging the values in the records of the database in such a way that low-order statistics are preserved. Statistics that employ exactly k attributes are called k-order statistics. A database D is said to be d-transformable if there exists a database D', but has no records in common with D, but has the same $k-$order counts as D, for $k \in \{0, \ldots, d\}$. However, finding a general data swap is thought to be an intractable problem.

Recent investigations have shown that it is possible to transform the data by data swapping, so the rules produced by decision trees (a ubiquitous KDDM technique) on the new set corresponds to finding a-transformation of a given database, but relaxing the condition that D and D' have no records in common. The exchange of values of the confidential attribute consists of randomly shuffling within heterogeneous leaves of the decision tree. Thus all the statistics, which do not involve this attribute, are preserved. Similarly, the statistics that involve the confidential attribute and whose query sets are defined by nodes of a decision tree will also be preserved. All the perturbed statistics are those whose query sets could be obtained by the repeating process of splitting heterogeneous leaves, until homogeneous leaves are reached. Since the heterogeneous leaves have a vast majority of records belonging to a single class and no straightforward way for further splitting, we can argue that the most seriously disturbed statistics will be those that involve a small number of records. Furthermore, we can balance the statistical precision against the security level by choosing to perform the swapping in the internal nodes, rather than in the leaves of the decision tree: the closer to the root, the higher the security but lower the precision.

12.3 Some Privacy Issues in Knowledge Discovery: The OECD Personal Privacy Guidelines

Several countries have generated principles to protect individuals from the potential invasion of privacy that data collection and retrieval poses. The Organization for Economic Cooperation and Development (OECD) has provided probably the best known set of guidelines. A number of countries have adopted these guidelines as statutory law, in whole or in part. The OECD has specific guidelines pertaining to data privacy that directly affect those

performing knowledge discovery generally, and those who use the so-called "personal data" in particular. In this section we address such questions as

- What are the implications of the existing privacy guidelines, especially those of the OECD, for knowledge discovery?
- What are the limitations of these guidelines?
- How do the restrictions on knowledge discovery about individuals affect knowledge discovery on groups?
- How do legal systems influence knowledge discovery?

The answers we pose to these and other related issues will be helpful in generating a larger dialog on this important subject. There has been relatively little investigation into the privacy and security issues relevant to knowledge discovery, in particular and intelligent systems in general. Developers have proposed and used intrusion-detection systems as the basis of security systems designed to protect privacy. Typically, intrusion-detection systems determine if a user is an intruder to a legitimate user, generally by way of various internal systems profiles. Earlier studies of security issues in intelligent systems included issues of privacy and the security of systems knowledge. There has been some concern about knowledge discovery as a different kind of threat to data base security as well.

12.3.1 Risks Privacy and the Principles of Data Protection

The classic definition of the invasion of privacy refers to the "abuse or disclosure of intimate personal data." Recently, there has been an effort to expand this definition to include other issues, such as the protection of general privacy and protection from the unauthorized use of one's "personal" data taken from computer databases.

Increasingly, companies and organizations are using computer-based systems to capture personal data. Although this method typically increases both efficiency and productivity, there are a number of risks to individual privacy. In particular, those risks include the following:

- the data can be used for some purpose other than that for which it was collected;
- the data can be inaccurate, incomplete, or irrelevant;
- there are risks of unauthorized access to personal information;
- individual databases can be linked, increasing the range of information about individuals;
- The increased ability to construct individual profiles from multiple sources may affect "decisions concerning the individual's qualifications, credit eligibility, health, insurance consumption patterns social security, employment and so on."

As both the amount of information and number of users on the Internet grows, these risks become increasingly likely to manifest themselves. This

is particularly true for joining previously disparate databases. Hence, many feel that additional guidelines and statutory-based controls are necessary to prevent the invasion of personal privacy. These concerns have led organizations to generate guidelines to mitigate these privacy risks, including the OECD and the council of Europe. This section focuses on the OECD guidelines since many nations have adopted then as statutory law.

OECD principles of data collection. The following are the OECD principles of data protection:

1. *Collection limitation:* Data should be obtained lawfully and fairly, while some very sensitive data should not be held at all.
2. *Data quality:* Data should be relevant to the stated purposes, accurate, complete, and up-to-data; proper precautions should be taken to ensure this accuracy.
3. *Purpose specification:* The purpose for which data will be used should be identified, and the data should be destroyed if it no longer serves the given purpose.
4. *Use limitation:* Use of data for purposes other than specified is forbidden, except with the consent of the data subject or by authority of the law.
5. *Security safeguards:* Agencies should establish procedures to guard against loss, corruption, destruction, or misuse of data.
6. *Openness:* It must be possible to acquire information about the collection, storage, and use of personal data.
7. *Individual Participation:* The data subject has a right to access and challenge the data related to him or her.
8. *Accountability:* A data controller should be accountable for complying with measures giving effect to all these principles.

The OECD principles arose to help nations cope with the shipment of data outside the country of origin. They attempt to ensure that when data is transported across country borders the data subjects enjoy the same level of privacy as in the original country. Thus far 24 countries have adopted the OECD guidelines to varying degrees, including Australia, Austria, Belgium, Canada, Denmark, Finland, France, Germany, Greece, Iceland, Ireland, Italy, Japan, Luxembourg, the Netherlands, New Zealand, Norway, Portugal, Spain, Sweden, Switzerland, Turkey, the UK, and the US. Not all countries employ the OECD guidelines as statutory law, and not all countries have adopted all eight guidelines. Instead, the level of participation – the number of guidelines adopted – varies from country to country.

Twelve nations have adopted all eight of the principles in statutory law; Japan adopted seven of the principles (excluding #7), and UK has adopted six of the principles (excluding #7 and #8), as statutory law. Alternatively, Australia, Canada, New Zealand, and the US do not offer protection to personal data handled by private corporations. However, those for four countries have similar statutory constraints on personal data held in the public sector.

Scope of application: personal data. The OECD developed the primary protective guidelines for personal data. Consequently, the KDD community must determine what kinds of data fall under the heading of personal. According to Neisingh and de Houwer, personal data is data gathered by corporations and government, including financial, educational, economic, social, political, medical, criminal, welfare, business, and insurance data. As a result, it is easy to see that these principles affect many different kinds of datasets.

12.3.2 The OECD Guidelines and Knowledge Discovery

Now, we discuss the impact, implications, and limitations of the guidelines for knowledge discovery.

Collection limitation. This principle states that, "some very sensitive data should not be held at all," thereby limiting the scope of knowledge discovery from data. If the data is "very sensitive," knowledge discovery researchers should probably not have access to the data, as it could lead to repercussions. Such sensitive data is likely to include information about religious beliefs, race, national origin, and other issues.

However, it is not clear what it means for data to be sensitive. What may be deemed sensitive depends on the context and country in which the authorities develop the legislation. What is sensitive in one country may not be sensitive in another, suggesting that knowledge discovery could differ from country to country. Accordingly, such cultural differences could from the international differences in computer science practices.

Data quality. Knowledge discovery may influence the data quality principle. For example, knowledge discovery may lead to questions about additional categories of information, such as derived data. The data quality principle suggests that researchers differentiate derived data from original data and not include it in the database, since its accuracy could not be assured. Over time the data on which derived data is based may change, thereby changing the derived data as well. As a result, researchers should not store this data, as it could be outdated. If the derived data is kept, researchers should treat it with the same concerns as the original data.

Also, this principle's recommendation that proper precaution be taken suggests that there be quality standards in knowledge discovery. However, since the discipline is still evolving, it may be premature to talk about generating standards.

Purpose specification limitations. This principle indicates that databases be used only for the declared purposes. Goals for the use of data should be generated, and the data should be used to accomplish those goals exclusively. Any other uses would require the consent of the data subject. Consequently, it is critical that if a database is planned for knowledge discovery, then the use of knowledge discovery is specified.

In addition, if knowledge discovery is only done on databases for which knowledge discovery has been declared, then only those databases generated since the gathering of purpose information began are available for this activity. Accordingly, legacy and existing databases are probably outside the scope of knowledge discovery. Users may have to declare the specific knowledge discovery tasks when gathering the data, instead of declaring anticipated knowledge discovery for some general purpose.

The purpose principle is critical for knowledge discovery using multiple databases. If the data was gathered for use in a single database, analysis across multiple databases generally would violate the purpose principle. This could limit knowledge discovery using individual personal data to particular databases.

This principle threatens knowledge discovery's potential to expand on its on discoveries as well. Feedback can play a very important role in knowledge discovery tasks. As the system generates more knowledge, that knowledge can form the basis of the search for additional knowledge. Therefore, if the principle limits the knowledge discovery task to the first-level findings specified in the original purpose, it limits the power of knowledge discovery significantly.

Another possible limitation is the required level of detail in the statement of purpose. It is possible, in an extreme case, that authorities would require researchers to elicit each specific knowledge discovery activity, not just the fact that knowledge discovery would be done.

Use limitation. This principle is closely related to the purpose specification principle, as it specifies that if data is to be used for some purpose other than the originally specified purpose, the data subject must provide consent. By extension, the data subject will need to provide consent when his or her personal data is to be used for knowledge discovery. The purpose specification principle requires users to identify the original use of the information, and the use limitation principle constraints data use to the original purpose. Both principles require data subject consent if changes in the use of the data occur.

The use limitation principle has a direct impact on performing knowledge discovery from related databases. Generally, expanding the analysis of knowledge discovery from one database to multiple databases would require data subject consent, since the interaction of multiple, previously unconnected databases would suggest alternative uses beyond the original scope.

Acquiring data subject consent may be very difficult, in part because most data subject would have difficulty understanding the technology of knowledge discovery. Further, it is unclear what level of detail data subjects would need. For example, would the awareness that knowledge discovery was being done be sufficient or would users need to explain the individual task level?

Security safeguards. This principle calls for establishing safeguards against the misuse of data. In some cases, knowledge discovery may qualify as a misuse of data, especially if unauthorized users perform knowledge discovery, or if knowledge discovery occurs without gathering consent. As a result,

authorization procedures for knowledge discovery must be established. The limitations associated with the statement of purpose also influence the security safeguard principle. A particular concern is how to secure a database from knowledge discovery without eliminating access to virtually all users.

Openness. Taken to one extreme, the openness principle suggests that data subjects should be able to acquire information about the uses of knowledge discovery and the specific knowledge discovered about them. Requiring analysis to inform individuals about particular derived data could limit the general use of knowledge discovery and thereby inhibit its use. If knowledge discovery does not lead to inferences about individual data subjects, there would not necessarily be an openness issue.

However, since it is virtually impossible to deter users of a database from performing knowledge discovery, it will be equally difficult to knowledge discovery, it will be equally difficult to know for certain whether knowledge discovery is being done using information about a particular data subject. Thus, the individual participation and accountability principles play a critical role in controlling inappropriate knowledge discovery.

Individual participation. This principle suggests that data subjects should be able to challenge knowledge discoveries related to them. These discoveries might pertain to the individual only or to the individual's relationship to specific groups. The knowledge discovered may directly influence how the users perceive and treat the data subject, possibly adversely affecting that person's available options.

In light of the right to challenge knowledge discoveries, it is critical to document the development of conclusions. Substantiating the quality of different knowledge discovery approaches and algorithms will become increasingly important. The development and use of standards will help mitigate the challenges to knowledge discovery findings.

Accountability. This principle calls for a data controller who is accountable for user compliance with the OECD measures. Thus, a knowledgeable data controller should authorize and be responsible for the adherence of knowledge discovery activity to the OECD measures. In addition, the data controller should inform data subjects of the use and findings from knowledge discovery.

However, due to the decentralization of databases and the difficulty of controlling knowledge and the difficulty of controlling knowledge discovery activity by those who have access to databases, data controllers will have great difficulty monitoring knowledge discovery effectively. Accordingly, it will be important for the data controller to inform database users and maintenance personnel about the policies regarding knowledge discovery activities, including the consequences of inappropriate use.

12.3.3 Knowledge Discovery about Groups

This section has thus far focused primarily on privacy issues associated with individual personal data. The OECD guidelines do not refer explicitly to

knowledge discovery about groups. As a result, unless the knowledge discovered directly affects the individual personal data, there would be no general application of the guidelines. Instead alternative legislation or guidelines could be used to guide knowledge discovery about groups. For example, in the US it is illegal to discriminate against certain groups based on sex, race, religion, or national origin. Knowledge discovery about groups, then, could comply with these laws in its use of knowledge pertaining to these categories.

Further, the OECD guidelines suggest that individuals have the right to control the use of data about themselves, even in apparently innocuous knowledge discovery about groups. As a result, individuals could request that they not be included in the generation of knowledge about groups of which they may be a member.

One drawback of these individual privacy constraints is that they could interfere with important knowledge discoveries. For example, certain diseases seem to strike some groups and not others. As a result, information relating to groups could be the key to the discovery of cures, or other important kinds of knowledge.

12.3.4 Legal Systems and other Guidelines

The OECD guidelines form one basis of analysis. This could also be extended to investigate alternative sets of guidelines and statutory laws. The Council of Europe issued a similar set of guidelines for the European community that included the eight OECD principles and some additional constraints relating to the so-called transborder data flows. As alternative legal structures develop, researchers could analyze them for their impact on knowledge discovery.

Legal systems offer bases for the interpretation of different terms and situations in knowledge discovery as well. Many states, for the purposes of protecting litigants from undue invasions of privacy by adverse parties, have statutes defining personal or consumer information. For example, the California Code of Civil Procedure, section 1985.3, provides detailed definitions about personal records; they are the "original or any copy of books, documents, or other writings pertaining to a consumer and which are maintained by any 'witness' which is a physician, chiropractor. . . "

In the specific case of litigation, there are laws regarding the disclosure of information. For example, the California Code of Civil Procedure, section 1985.3, deals with "Sub-poena for production of personal records," while section 1985.4 summarizes the law regarding "production of consumer records maintained by state or local agency."

Further, in many cases different levels of government regulate certain industries, to a certain extent. Such industries include insurance, law, accounting, and medicine. As a result, these industries are likely to have regulations on limitations of disclosure of information. In other cases, the industries are self-regulated.

The limitations to the use of knowledge discovery that we have discussed are not limited to the new methods of knowledge discovery developed by the artificial intelligence community. Rather, they apply to all methods used to generate knowledge, including more traditional statistical and database approaches. The OECD guidelines limit the knowledge that can be obtained using any process, including direct examination, classic database updates, queries, or statistical methods.

This section provides some insight into the problems concerning personal privacy and data faced by those who wish to employ knowledge discovery. When it comes to personal data, there often are statutory limitations that vary from country to country. As a result, it suggests that the practice of computer science and artificial intelligence varies from country to country, based on different cultural and legal differences. However, it is clear that there are some general principles of data collection and maintenance that a number of countries adhere to. Those principles influence what data can be used in knowledge discovery and how users process and maintain discovered data.

Many of the above limitations are a result of the OECD personal privacy legislation predating knowledge discovery's widespread use in the artificial intelligence community. The OECD guidelines could not anticipate or address many important issues regarding knowledge discovery, and thus several principles are too general or unenforceable.

12.4 Summary

The mentioned above issues are considered major requirements and challenges for the further evolution of data mining technology. Some of the challenges have been addressed in recent data mining research and development, *to a certain extent*, and are now considered *requirements*, while others are still at the research stage. The issues, however, continue to stimulate further investigation and improvement.

Knowledge discovery and data mining revitalizes some issues and posses new threats to privacy. Some of these can be directly attributed to the fact that this powerful technique may enable the correlation of separate datasets in other to significantly reduce the possible values of private information. Others can be attributed more 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.

The OECD guidelines could not anticipate or address many important issues regarding knowledge discovery, and thus several principles are too general or unenforceable.

12.5 Review Questions

1. What is mining methodology and state some of the user interaction issues?
2. Discuss in detail on performance issues and issues relating to the diversity of database types.
3. Explain about privacy issues on data mining.
4. Give details on the Organization for Economic Cooperation and Development(OECD) personal privacy guidelines.
5. What are the OECD principles of data collection in the mining technology?