Anonymization techniques for Knowledge Discovery in Databases

Willi Klösgen
German National Research Center for Information Technology (GMD)
D-53757 Sankt Augustin, Germany
kloesgen@gmd.de

Abstract

KDD deals with the ready data, available in all scientific and applied domains. However, some domains with comprehensive and conclusive data have severe data security problems. To exclude the reidentification risk of individual cases, e.g. persons or companies, the access to these data is rigidly restricted, and often KDD applications are not allowed at all. In this paper, we discuss data privacy issues based on our experience with some applications of the discovery system Explora and other data analysis approaches. At first, some examples of applications are presented referring to a simple classification organized according to two dimensions important for the privacy discussion. Then we treat the reidentification risk and discuss anonymization methods to overcome these problems. Aggregation and synthetization methods are discussed in more detail. There is a tradeoff between the reduction of the reidentification risk and the preservation of the statistical content of data. We analyse for some main KDD patterns, how far the statistical content of anonymized data is still sufficient. In principle, KDD needs aggregate events. Since the event space of a dataset is very large, a static precomputation of all possible events is often not viable. We propose an architectural solution of a modular KDD system including a separate data server handling also data security requirements and ensuring that only dynamically aggregated data leave the server and can be analysed by the discovery modules of the KDD system. Finally, some other data privacy aspects are addressed.

Introduction

Knowledge discovery in databases (KDD) is the search for patterns that exist in databases, but are hidden among the volumes of data. KDD aims at partially automating analytical processes and is based on large scale search, statistical methods mainly of exploratory analysis, and advanced data management. A database can be regarded as a sample drawn from a joint distribution of its variables. Patterns typically based on unusual marginal distribution characteristics of some variables can supply valuable knowledge, if this sample is sufficiently representative for the domain.

Often these sample databases consist of micro data, i.e. data on individual cases like persons, companies, transactions. Therefore, privacy and security issues have to be observed for KDD. The investigation of these issues within the KDD context is still limited. O'Leary discusses intrusion-detection systems (1992), KDD as special threat to database security (1991), and the impact of privacy guidelines on KDD (1995).

The security risks and protection regulations for micro data are relevant also for KDD. This first problem area relates to security of input data and the questions, whether an analyst is allowed to access a special micro dataset and use KDD methods to analyse it.

These problems are given mainly for public institutions and especially for data gathered when executing administrative processes, and then, as a secondary application, shall be exploited for KDD purposes. A well known example for the security and privacy controversy about micro data is the public discussion on censuses. Some other examples of very sensitive data are tax data gathered when processing tax returns, or intelligent traffic systems that automatically capture data on individual cars and their routes to bill driving on highways. For planning purposes, the exploration of these data would be highly valuable (e.g. to reduce and guide individual traffic or to plan efficient traffic systems), but an individual car driver must be protected from the reidentification risk and the derivation of any kind of personal mobility profile.

The knowledge discovered by KDD techniques is usually expressed as a set of harmonized statements on groups of cases. Therefore, KDD derives findings on aggregates. In principle, if micro data cannot be accessed, KDD techniques could use also aggregate input data, which should be available for an aggregation level as fine as possible. The finest aggregation level excluding for the most part the risk of reidentifying an individual case relies on aggregates of three cases.

Various patterns (Klösgen & Żytkow 1995) are used in KDD systems to determine when an aggregate (e.g., a subgroup of persons) is interesting. Although these

are aggregate rules or patterns, and KDD is not intended to identify single cases, there arise problems of discriminating groups. All members of a group are often identified with the group and the discovered group behaviour is attached to all members when the discovered patterns are subsequently processed. If a group of persons with a high risk of being ill is identified by a KDD application, a personell manager may hesitate to employ an individual member of this group. Even if the group may be to some extent heterogeneous as for this risk, some kind of collective behaviour is assigned to the group. The second privacy problem area relates therefore to the output of KDD methods.

In our applications of the KDD system Explora (Klösgen 1995a, 1995b, 1992), we were mainly confronted with the first privacy problem, data security. Because data security problems resulting from the application of information technology are treated for a long time, solutions and regulations are available also for KDD. Even important however is the problem of group discrimination. Although KDD can be regarded as only one possible technique of data analysis and many more other techniques supplying results on groups are being applied for a long while, the discrimination discussion has been vigorously arisen only recently and attributed especially to KDD applications.

In the context of data privacy issues, two simple binary dimensions are important to classify KDD applications. Much more severe regulations are used for applications run in a public environment like government, public administrations, or public institutes (e.g., Research Institutes, Statistical Offices) than in private institutions. This may be explained by the fact that public applications are critically observed and any collection and analysis of micro data by public institutions is mistrusted. On the other side, this leads to a very cautious treatment of data security issues by public institutions trying to avoid any public controversy.

The OECD guidelines (see below) for data collection and data protection have been adopted by 24 countries. But not all countries offer protection to personal data handled by private corporations. In Australia, Canada, New Zealand and the United States, only data held in the public sector are regulated. This also indicates the importance of this dimension.

Another important distiction is, whether data were especially collected for data analysis and KDD applications (primary applications) or were produced when executing administrative processes or business transactions, and as a secondary application are then used for KDD purposes. Especially this second group of applications must be treated very sensitively.

Privacy relevant application groups

Public primary applications

A typical example are data collected by National Statistical Institutes. A controversial public discussion on data privacy and population censuses has lead in some countries to their abolition. In Germany, the number of variables collected in the census was fundamentally reduced and the access and analysis of censuses is very restrictively regulated by special laws. An analyst within an Institute may use census data for discovery processes supporting well defined analysis problems related to the scope of the Institute. An external analyst however is not allowed to access census data on the micro level, because Institutes are very restrictive on their release, even for of research applications.

This is also the case with our KDD application exploiting "Micro Census" (a 1 percent sample of the German population questioned yearly). Discovery processes in this voluminuous data set (800.000 persons, 200 variables) aim at the identification of risk profiles for unemployment and of changes in education and health status of population groups. External KDD applications must rely on aggregate data (see below); within the Statistical Institute, the microdata are used.

Public secondary applications

Data compiled for an administrative process such as tax return or granting of public transfers (help for housing, education, children) are analysed to support planning of legislation (Klösgen 1994). Special laws regulate the availability and analysis of these data for secondary purposes, where the restrictions in the tax field (because of the guaranteed secrecy of tax data) are more severe than in the transfer field when a citizen is claiming for a subsidy and must agree on the exploitation of the data. For the tax domain, synthetical data (see below) are used in internal and external applications. The transfer domain can exploit micro data for internal applications within government.

Private primary applications

A comparably easy situation concerning the availability and analysis of data is given in the case when private institutions collect data to be exploited by data analysis methods including KDD. Explora applications of this type relate to the analysis of data collected by institutions of market research and opinion poll, e.g. a survey on the financial behaviour of people as clients of banks and insurance companies. This data and the corresponding analysis tools are freely marketed by the institutations based on the given permission of the questioned persons. The clients of the institutes exploit the data, but are not allowed to release them to others.

Private secondary applications

A sensitive and often only loosely regulated application group contains e.g. medical data collected during hospital treatment or client and transaction data stored by banks on financial transactions. The legal foundation of these applications is often based on very general permissions. You usually have to sign a contract if you open a bank account agreeing that data on transactions are stored and used for all purposes connected to

the management of the account. To these purposes implicitely belong also planning and especially discovery. Usually the client or patient has no choice, he must simply accept this clause of the contract.

Reidentification risk

One main risk of micro data is the reidentification of cases. Persons or firms may be willing to provide their data for a special purpose to a (governmental) office, but it must be prevented that an unauthorized intruder gets known of the data. A company surely will not agree that sensitive data is accessed by a competitor and a person will disclose data on his health status to a doctor, but possibly not to his employer.

National legislations regulating the access to statistical microdata contain mostly the general condition that microdata can only be released to third analysts if the assignment of data to really existing individual cases is not possible. Paass and Wauschkuhn (1985) describe the results of a project performed at GMD in cooperation with National Statistical Institutes and governmental organizations with the aim to develop operational criteria for the release of micro data and techniques to determine the reidentification risk of the cases included in data. They proved, that for most large statistical databases, e.g. Microcencus, Income and Consumption Sample (Germany), and Statistics of Income File, Current Population Survey (U.S.A.), the reidentification risk is high, when only some simple anonymization techniques are applied. Generally, the reidentification risk is sufficiently low, if the number of attributes and the value domain of the attributes are small (less than 10 to 15 attributes), considering already the sample property and the inevitable data error and noise. If an intruder can apply an own, comprehensive database holding more than 10 to 15 attributes overlapping with the attributes of a micro database, the reidentification risk is very high.

In socio-economic research, a lot of methods were developed to exclude the reidentification risk that an intruder can identify a case in a micro dataset (e.g., Feige & Watts 1972, Spruill 1983). Since an intruder can possibly apply additional knowledge on the goal case, simple anonymization techniques (omitting identification number, name and address) are not sufficient to exclude this risk. Anonymization techniques generate aggregate or synthetical data to reduce the reidentification risk under preservation of the statistical content (data quality principle).

Input data for KDD

We first propose a technical solution to solve the data security problem when applying a KDD system.This solution relies on a strictly modular architecture of a KDD system. The data access is managed and controlled by a separate data server handling also data security requirements and ensuring that ony sufficiently aggregated data leave the server and can be analysed by the discovery modules. Then we discuss two main anonymization techniques and treat the tradeoff between anonymization and statistical content.

Data access in KDD systems

Protection methods to ensure data security could be incorporated into a KDD system. The versatility of Explora is based on a general search approach which allows to embed various pattern types. This partition results in four subsystems which are highly independent and communicate only via some narrow interfaces. A verification method associated to a pattern type evaluates a pattern instance (see (Klösgen 1995a) for evaluation functions). These verification methods rely on a defined interface to access data.

A small set of aggregated data structures (e.g. histograms for dependent and independent variables and their conjunctions) is supported by this interface. A data management subsystem is responsible for the efficient generation of the aggregate structures (Klösgen 1995b). Explora relies on a dynamic in-core management of data limiting the number of cases that can be exploited in the system (about 50,000 records). Other approaches (Holsheimer et al. 1995a) rely on parallel data management techniques and database servers to realize a very efficient data management subsystem, allowing also very large data bases to be exploited.

The search components are independent of the data management subsystem and exploit only the results of the verification method. Search is also modularly implemented with independent subsystems, e.g. for the management of the structure of the seach space (partial ordering). Another architecture of KDD system is described by Holsheimer et al. (1995b).

The data management system can be extended by protection modules ensuring security. These modules must guarantee that only aggregates are generated on the server compatible with the given security requirements of an application. Micro data reside on a database server which is run under a special management and access software ensuring micro data security.

This dynamic computation of aggregates (as and when needed by the discovery modules) is often necessary, because the given data are sparse in the event space, i.e. the number of events is much higher than the number of records in the database. One could compile only those events that are defined by a limited number of conjunctions and are represented in the database. But also these events would often contain only one or two cases, so that they have to be further aggregated (for data security reasons) by omitting some conjunctive term in their definition. Therefore, a complete event space of a special order n (with n conjunctive terms in the event definitions) can often not be constructed under the restriction that each event contains at least, e.g., 3 cases. Sometimes however, a complete event space can be statically precomputed and KDD can run on this event space. Especially, this

is possible, when the given analysis problem involves ony a limited number of not too fine grained attributes, and no continuous attributes are necessary for the analysis or can be reduced to discrete attributes with a domain consisting of a finite number of intervals.

For practical applications, advanced solutions based on dynamic accesses via a separate data server ensuring data protection have not yet been implemented, and the problem is treated by generating aggregate or synthetical data on which KDD systems are operated. These solutions are discussed in the following.

Aggregate data

One anonymization technique is based on combining several cases to aggregates. A simple technique combines a group of at least three to five similar cases *(triple aggregates)* by averaging over the groups. The triples can be seen as artificial cases. In detail, many variants of this method are discussed in literature.

For instance, a number of discrete attributes is selected and records are hierarchically ordered according to their values and those of one continuous attribute. These should be "important" attributes, if such a qualification is possible for the application. Groups are built according to the values of the discrete attributes. Within the groups, the triple groups are built according to the sorted values of the continuous attribute. A refined variant allows the selection of hierarchically lower grouping attributes to be different for different values of higher grouping attributes. The values of the continuous and of the remaining attributes are averaged in the triples. For discrete attributes, indicator attributes are constructed for each of their values counting the average number of occurences of a value in the (triple) group. For the remaining attributes, the triples may contain averages of very different values, so that a lot of information is lost and noise is introduced. This leads to a reduced statistical content of the triples. More elaborate methods use special attributes (principal components) representing the multivariate structure of the sample to hold this reduction as small as possible. Also clustering can be applied to construct small homogeneous groups.

Paas and Wauschkuhn (1985) have proposed a statistical model for analyzing the statistical content of the derived dataset anonymized by such aggregation techniques. The reduction of the statistical content caused by an anonymization method that is still tolerable depends on the kind of analysis that will be done on the data. Different requirements must be satisfied, e.g. for the generation of simple summary descriptions, conclusive statistical tasks, or explorative analyses. For explorative analyses, a higher degree of reduction will be tolerable than for conclusive studies.

KDD is an exploratory approach testing a large number of hypotheses. To alleviate the problems inherent in this approach, a high level of significance is applied to filter potentially interesting hypotheses (Klösgen 1992). In a refinement phase, the potentially interesting hypotheses are then elaborated to generate an "optimal" set of harmonized hypotheses.

A simple approach to determine the reduction of the statistical content due to anonymization techniques compares the discovery results derived on the original and anonymized datasets. Ideally, the results should be equal. Some first experiments we executed on a test dataset suggest that the discovery results are very similar, and that triple aggregation does not reduce the statistical content too much for KDD purposes.

To study these problems in more detail, a lot of experiments have to be run, considering different patterns, various search, evaluation, and refinement strategies, and also different types of attributes according to their relevance for the discovery problem. Surely, the attributes used to construct the aggregates will not be critical, but the remaining attributes which have been aggregated in the triples may be critical.

This relates to a similar problem. How robust are the discovery methods considering the sample error? To judge the anonymization error, it should be compared to the sample error. If the anonymization error does not exceed considerably the sample error, anonymization is not critical for KDD applications.

For instance, the probabilistic rule and mean patterns offered in Explora require for a potentially interesting subgroup of cases, that the share of cases of a target group (discrete dependent variable) or the mean of a continuous dependent variable deviates from an apriori value (the share or mean in the total population) by at least $5s$, where s is an estimation of the standard deviation of the share or the mean (Klösgen 1992). We regard the given data set as a sample of a total population and consider the random error of an index x (share or mean) as the distribution of the difference $x_1 - x_2$ between two randomly selected samples. For share and mean, the standard deviation for these distribution of differences can be easily computed. The statistical evaluations underlying the rule and mean pattern ensure, that if a hypothesis is evaluated as potentially interesting based on one sample, it is (stochastically) also potentially interesting for another sample.

The anonymization error of an index x is the difference $x_o - x_a$ in original micro data and anonymized data. If the distribution of the anonymization error is stochastically smaller than the distribution of the random error, anonymization is not critical for the KDD-pattern based on this index. The empirical distribution of anonymization errors can be estimated by the distribution of the errors computed for one anonymized dataset. The distribution for similar indices built for a large number of subgroups of cases is studied and compared with the distribution of the random error.

Empirical studies (Paass & Wauschkuhn 1985) show that for more than the half of all tested subgroups, the anonymization error is smaller than the random error. But for a lot of subgroups, the indices in the

anonymized and original micro dataset deviate significantly. Especially the remaining continuous attributes that were not used in constructing the triples and had to be averaged are critical, but also in some cases the remaining discrete variables. Therefore, subgroups defined by these variables have to be treated cautiously. On the other side, one must consider that for large microsets, the significance values used to evaluate a subgroup are much higer than the requested multiple (e.g. $5s$). Therefore, anonymization based on triple aggregates does not destroy the statistical content too much, if KDD is run as an exploratory approach in large datasets. Measuring the interestingness of subgroups in a KDD system (Klösgen 1995a) should in this case weight also the importance of attributes based on their role in the anonymization process.

Another approach is based on performing KDD in an event space. An event space is given by a projection of the database tuples to the cross product of the (possibly coarsened) value domains of selected variables (which are relevant for an analysis problem). One of our KDD applications running on an event space is the external analysis of micro census data. The selected variables include regions, industries, jobs and their extensive hierarchical classifications. The event space can be seen as a super table containing many cells. Each cell (with the number of occurrences above a threshold) can also be seen as an artificial case with a weight corresponding to the occurrences.

Another type of aggregate applications of Explora was performed for election research, where the aggregates correspond to given election districts. KDD runs on these districts, where personal election results are averaged over districts and further socio economic data are aggregated and associated to districts.

Aggregation techniques solve also some performance problems or capacity limits of systems. While future systems relying on parallel techniques allow to exploit very large databases, e.g. census data may exceed the limits of existing systems. Instead of using 800,000 cases, an aggregate application of the micro census in Explora relies on 50,000 events.

Synthetical data

For secondary public applications, security constraints are so severe, that micro analyses may not be allowed even internally. Tax returns and tax legislation is such an example. Here we use synthetical data to render KDD possible. A synthetical dataset is an artificial sample of cases derived from various available tabulations and other aggregations. Often by the combination of various aggregate sources, a database can be constructed offering more information for being analysed than it is possible by aggregate analyzing techniques.

By order of several German ministries, we have constructed a synthetical micro database for income tax planning (Gyarfas 1990). This database is built by using available tax statistics which are officially gathered based on the individual tax returns. These statistics are regulated by law in detail. Other statistical processing of individual tax returns is not allowed.

Interpolation is a first technique used in this context. For a continuous variable (e.g. income), a discrete distribution is given by histograms, i.e. a set of intervals (income classes) and for each interval, the number of cases in the interval and the average income. This discrete distribution shall be refined, assuming a smooth (unknown) distribution of the continuous variable. The unsteadyness of histograms can be attributed to the given discretization by intervals. A finer discretization (based on smaller intervals) can be derived by spline interpolation. It is assumed that the smooth distribution can be represented in each interval by a polynomial of degree three. The constraints (continuous and smooth distribution) allow to construct and solve a system of linear equations to determine the coefficients of the polynomials. Alternatively, distributions which are characterized by a small number of parameters can be estimated using the given tabulation (e.g. lognormal distribution). Compared to interpolation, this leads however to a worse fit.

Based on given marginal distributions of some variables including cross tabulations and other aggregates like correlations and regressions, an artificial micro dataset is generated which is consistent with the given aggregates. This can be treated as a combinatorical optimization problem, and simulated annealing is applied to generate the artificial micro dataset. Based on the interpolation results, an initial weighted sample is given (e.g. with values of the interpolated income variable). These cases receive values for the other variables by the optimization procedure, so that the given aggregates are approximated.

The tasks of generating this synthetical dataset and the identification of interesting findings during discovery can be regarded as inverse tasks. A micro dataset can be seen as a sample drawn from a joint distribution of the variables. Some partial information on this common distribution is given by the available marginals, the remaining information on the common distribution is inferred by the generation procedure. This generation is based on information theoretic approaches minimizing the information gain. The joint distribution is generated which maximizes entropy under the given constraints. On the other side, KDD evaluations should not infer this additional distributional information not available in the given marginal information as interesting. This is ensured, since e.g. for a finite discrete distribution, entropy is maximal for an equal distribution, which is not interesting for a KDD pattern. Implicitely, most KDD patterns are evaluated the more interesting the more unequal a corresponding distribution is (Klösgen 1995a).

Therefore, KDD methods of course can not find additional knowledge in synthetical data which is not al-

ready contained in the given aggregate information. The profit of synthetical micro data lies in the uniform framework of a simple data structure (database in form of a large table) analysed by KDD methods. Also other techniques (e.g. simulation models) may be easily applied to micro data to infer additional variables which are then analysed by KDD techniques.

Examples for the derivation of additional variables in the tax application are given by the variables that can be calculated by applying known regulations. Tax regulations define, how much income tax a person has to pay given his/her relevant socio-economic variables.

Disvovered knowledge, data security and data privacy

The results of KDD applications are aggregate findings on some groups of cases. Here the question arises whether these results must be hold confidentially or may be published or released to persons who are not allowed to access the input micro data. The transmission of results usually is regulated for an individual application. E.g., for KDD analyses of census data, the results can be published, if the reidentification risk is excluded. This is ensured given that the groups are large, i.e. contain at least a fixed number of cases.

Another problem may arise with some discovery patterns. If the input is a complete population (not only a sample) and exact rules (with 100 percent certainty) are discovered, the values of the rule conclusion can be exactly contributed to all members of the group, which may contradict to the non-reidentification requirement. A finding "all persons owning more than 3 appartment houses and living in Berlin pay no income tax", resulting from a discovery process in tax data, would surely contradict the tax secrecy.

If some groups must be excluded by national laws, the corresponding sensitive variables like religion, beliefs, race should be deleted in input data and not be usable for KDD. Like any other tools, KDD systems may be used involving great responsibility or misused. A mature state of awareness within KDD community on discriminative, manipulative and other irresponsible applications is however necessary to be developed. OECD has listed eight principles of data protection (O'Leary 1995; we discuss some KDD specific impact of the principles in this number of IEEE Expert).

Conclusion

There are two privacy problems of KDD, the input and the output problem. Regulations determine whether an analyst may access a special micro dataset and use KDD. This is usually done on a higher level, e.g. data analysis for planning purposes is allowed to a limited user group. If data analyses are allowed for pre-existing databases, also KDD can be applied. Access regulations for micro data are most restrictively handled by public applications, especially for secondary public applications relying on data gathered for an administrative process. In these cases, some methods to exclude the reidentification risk of a micro dataset and preserving the statistical content of data as far as possible can be used to allow KDD to be applied on a modified dataset. Some aggregation and synthetization methods were summarized.

The output problem refers to the results of KDD applications. Which findings may be discovered, published and used for which following purposes? KDD ethics must be developed outlawing e.g. discrimination, manipulation, or watching of groups. Since ethics alone cannot exclude these applications, penal regulations may be needed. However, this is not a specific KDD problem, but concerns all kinds of data analyses.

References

Gyarfas, G. 1990. *Ein Simulationsmodell der Einkommensbesteuerung auf der Grundlage synthetischer Mikrodaten.* München/Wien: R. Oldenbourg Verlag.

Holsheimer, M., Kersten, M., and Siebes, A. 1995a. Data Surveyor: Searching the nuggets in parallel. In *Knowledge Discovery in Databases II,* eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Cambridge, MA: MIT Press.

Holsheimer, M., Klösgen, W., Mannila, H., and Siebes, A. 1995b. A Datamining Architecture. To appear.

Feige, E. L., and Watts, H.W. 1972. An Investigation on the Consequences of Partial Aggregation of Microeconomic Data. *Econometrica,* Vol. 40, No. 2.

Klösgen, W. 1992. Problems for Knowledge Discovery in Databases and their Treatment in the Statistics Interpreter EXPLORA. *International Journal for Intelligent Systems,* vol 7(7), 649-673.

Klösgen, W. 1995a. Explora: A Multipattern and Multistrategy Discovery Assistant. In *Knowledge Discovery in Databases II.*

Klösgen, W. 1995b. Efficient Discovery of Interesting Statements in Databases. *The Journal of Intelligent Information Systems,* Vol. 4, No 1.

Klösgen, W., and Żytkow, J. 1995. Knowledge Discovery in Databases Terminology. In *Knowledge Discovery in Databases II.*

O'Leary, D. 1991. Knowledge Discovery as a Threat to Database Security. In *Knowledge Discovery in Databases,* eds. G. Piatetsky-Shapiro, and W. Frawley, Cambridge, MA: MIT Press.

O'Leary, D. 1992. Intrusion Detection Systems. *The Journal of Information Systems,* Vol. 6, No 1, 63-74.

O'Leary, D. 1995. Some Privacy Issues in Knowledge Discovery: OECD Personal Privacy Guidelines. *IEEE Expert,* April 1995.

Paass, G., and Wauschkuhn, U. 1985. *Datenzugang, Datenschutz und Anonymisierung.* München/ Wien: R. Oldenbourg Verlag.

Spruill, N. 1983. Testing Confidentiality of Masked Business Microdata. Working Paper, PRI 83-07.09, The Public Research Institute, Alexandria, Virginia.