Privacy Preserving in Data Mining – Experimental Research on SMEs Data

O. Grljevic*, Z. Bosnjak* and R. Mekovec**
* University of Novi Sad, Faculty of Economics Subotica, Subotica, Serbia
** University of Zagreb, Faculty of Organization and Informatics, Varaždin, Croatia
oliverag@ef.uns.ac.rs, bzita@ef.uns.ac.rs, renata.mekovec@foi.hr

***Abstract*—Analysis of data on individuals and business sensitive data as well as revealing the results of such analysis without disclosing confidential and sensitive information is a very important issue. Many techniques for preserving privacy of data are currently being used. This paper is addressing some of the basic techniques: randomization, k-anonymity, distributed privacy preserving and application effectiveness downgrading. Most of the techniques should be applied in the phase of data collection or their preprocessing, which can lead to different results (better or worse) of data mining than would be obtained on original data. For this reason, data analysts should be encouraged to quantify the ratio between privacy preserved in data with application of each technique and the loss of data or quality of outputs. This paper illustrates the application of certain techniques for preserving privacy on experimental dataset, and reveals the effects that their use has on the results.**

I. INTRODUCTION

Rapid development of hardware and technology resulted in increased ability of data collecting and storage, including private data and data on individual habits and preferences. Consequently, questions of unauthorized access to data, data misuse and privacy violation arose, and applications of data mining methods and techniques were seen as threats to privacy preserving.

These trends led to many researches on privacy-preserving data mining, resulting in diversified privacy-preserving techniques that combine data mining with cryptography and different data hiding/masking methods, with the aim of transforming the data in a way that no one can identify the respective entity the data belong to based on values stored in a database.

Individuals' privacy concerns usually refer to his anxiety regarding who can collect (and how) information about him, which information are being collected and how collected information will be used (who can access them and for what purpose). On the other side, with proliferation of new technologies (like Internet, mobile and sensor technologies) individual's privacy concerns are expanded on various privacy threats that can occur in this digital environment. According to Accenture, [1] individuals consider that companies are obligated to protect customer personal data, and also reveal information about their practice regarding the collection, processing, dissemination or storage of customers' data. Considering that in today's global and networked environment data are stolen or lost on daily basses, data protection is becoming more difficult for many companies to address. Companies that are data holders are left with the responsibility of processing customer personal data in a way not compromising privacy, confidentiality or their own interests. DataLossDB in [3] reports on irregular data usage, pointing out that sensitive personal data are at risk and can be used for various purposes. These findings emphasize the need for changes in the field of data protection.

One important European regulation regarding privacy-protection is the EU Data Protection Directive (Directive 95/46/EC) which establishes common principles for data protection in Europe, [4]. Those principles are integrated in privacy regulations and laws of EU Member States, and were also used as a base for the development of many international privacy regulations. Section VIII of this Directive determines the purpose and means of personal data processing regarding its confidentiality and security. It is stated that appropriate measures should be used to protect data *'against accidental or unlawful destruction or accidental loss, alteration, unauthorized disclosure or access, in particular where the processing involves the transmission of data over the network, and against all other unlawful forms of processing'*. Evidently, those stated obligations clearly define responsibilities of parties included in the data processing. Nevertheless, during the implementation of those principles in practice many problems arise.

Discussing ways of data privacy protection also incorporates privacy-preserving in data mining process. This article investigates the effects of different data privacy-preserving techniques on the results of data mining. Section II gives an overview of the most frequently used techniques from the viewpoint of data mining. Section III describes some experimental results on privacy-preserving of data on small and medium sized enterprises (SMEs) in Vojvodina. The analyses were performed in Weka[1] tool. Section IV summarizes conclusions and future research directions.

II. OVERVIEW OF PRIVACY-PRESERVING METHODS IN DATA MINING

The majority of privacy-preserving methods used in a data mining process actually transform a source data prior to the application of data mining method or technique. In other words, privacy-preserving techniques are implemented in a time of data gathering or in a data preprocessing phase. The essence of these techniques is

[1] Waikato Environment for Knowledge Analysis, [11].

978-1-4577-1974-5/11/$26.00 ©2011 IEEE

granularity reduction, which results not only in increased data protection, but also in decreased effectiveness. This unpreferred effect on data mining algorithms comes from a data loss in the transformed dataset compared to the original one.

As it is stated in [2], the most often applied privacy-preserving methods are randomization, k-anonymity model, l-diversity model, distributed privacy-preserving, and output privacy-preserving (so called downgrading application effectiveness).

Randomization Models introduce some noise into data in order to hide all identification attributes in the dataset, [2, 6, 7]. The amount of noise has to be sufficient to disable identification of original data values, while aggregated distributions can be made in order to conduct data analysis. Besides the randomization, methods of multiplicative perturbation can be used, where random data projection or random data rotation techniques perturb the original data. The above mentioned methods represent the two most favored randomization approaches. Closely connected with them, and also frequently used with k-anonymity method, that is also described in the sequel, is data swapping. Data swapping preserves data privacy by interchanging the values of different records.

Since in some datasets the indirect identification of an entity is possible even if the fields for identification are removed, the *k-anonymity model* has been developed. Namely, the combination of pseudo-identifiers, such as age, ZIP code or similar attributes, can reveal the identity of a person, or at least narrow the potential candidates. The k-anonymity model reduces the data granularity by generalization and data suppression techniques. Generalization transforms the values of an attribute to a given interval (like transforming for e.g. a date of birth to age) or groups the categorical values into different value sets. Data suppression completely disregards the attribute.

With k-anonymity an original dataset can be transformed where it is difficult for an intruder to determine the identity of the individuals in that dataset. The transformed dataset has the property that each record is similar to at least another k-1 other records (regarding the identifying variables), and is indistinguishable from them, [9].

L-diversity model is an extension of k-anonymity model. It is developed in order to overcome the drawback of the latter to effectively prevent conclusion of delicate attribute values. This is achieved by distribution of quasi-identifier values within a group [8].

Distributed privacy-preserving is applicable in cases when more entities participate with their own datasets in cumulative dataset which they individually access and conduct unique analyzes, [2, 5, 10]. For example, distributed privacy-preserving is applied when competitive companies wish to derive aggregate results (such as demand/sales trends in the industry) and conduct some data analysis on the common dataset, but at the same time keep hidden the subtle and specific business data from the competition. In such a situation, companies usually do not share the whole dataset, but just some horizontal or vertical dataset partitions.

Horizontal data partitioning involves distributing data records with the same attributes among entities, while vertical data partitioning involves distribution of attributes among entities, which means there may be different attributes or views of the same dataset by different entities.

Downgrading application effectiveness refers to those data mining models whose outputs often reveal more of important information about the business policy and activities, or sensitive individuals' data, then the input attributes do. Thus, output can be used to draw conclusions about the data on which the model is built. Typical examples are classification results and associative rules that can reveal important information about marketing or business strategies of companies. For downgrading effectiveness of association rules, the technique of association rule hiding is most often utilized. This technique implies application of distortion (inputs in certain transactions are modified to different values) or blocking (input value is incomplete – replaced with question mark, and the truthfulness of the data is preserved).

Each of the mentioned privacy-preserving methods has its advantages, but also some drawbacks. Randomization is, for example, a very simple method applicable also on data gathering points, thanks to the independency of added noise on different tendencies in data, but it is difficult to hide outliers, and in case of publicly available data, the identity of a data owner can easily be identified.

K-anonymity is also an easy way of privacy-preserving, with lots of algorithms implementing the model. However, despite the large ability of the model to make the data owner undistinguishable from other entities, i.e. hide the identification attributes, the inefficiency to prevent concluding the sensitive attributes is its disadvantage. This is the reason why the l-diversity model was developed, assuring diversity of sensitive attributes. Namely, the problem of the k-anonymity originates from the fact that data are grouped into records that could be related with no more than k respondents from the group under consideration. However, the sensitive attribute within a data group can have a same value, enabling easy identification of data owners, especially if there exists a background domain knowledge or quasi-identifier that can reduce the number of possibilities.

The advantage of distributed privacy-preserving data mining is the derivation of useful statistics on the aggregated data set, while preserving the privacy of individual data sets it comprises of. The method efficacy is highly dependent on the degree of confidence in other data analysis participants.

The usefulness of preserving privacy of the results of certain data mining applications (for instance, association rules, classification, etc.) reflects in the prevention of the usage of their results to draw conclusions about the dataset and companies' business policy/strategy. Key shortcomings are reflected in the fact that many rules that do not contain sensitive data may be lost in the process, while spurious association rules can be derived that are of little interest to the user, so-called ghost rules.

III. Case Study

Weka tool offers series of filters that can be used as a certain technique for privacy preserving in data mining. Such filters, that are supposed to be applied in the preprocessing phase of a data mining project, are: randomization, resampling, discretization, adding noise, adding values, adding expression, etc. Generally, they are

```
  J48 pruned tree
------------------

ConditionOfProductsOnMarket <= 0
|   No_Limitations_for_More_Inovations = FALSE
|   |   More_Inovations_Limited_by_Insufficient_Stimulants = TRUE
|   |   |   RatesAsPrimarProblemForUsingMoreCredits = TRUE: no investments (3.0)
|   |   |   RatesAsPrimarProblemForUsingMoreCredits = FALSE
|   |   |   |   Pitanje33Inovacije <= 0: no answer (2.0)
|   |   |   |   Pitanje33Inovacije > 0: no investments (2.0/1.0)
|   | More_Inovations_Limited_by_Insufficient_Stimulants = FALSE: no answer (70.0/3.0)
|   No_Limitations_for_More_Inovations = TRUE
|   |   InovationsInTechnology = FALSE: no answer (2.0/1.0)
|   |   InovationsInTechnology = TRUE: foreign (2.0)
ConditionOfProductsOnMarket > 0
|   InovationsInTechnology = FALSE
|   |   LeasingAsExternalFinantialSource = FALSE: no investments (323.0/161.0)
|   |   LeasingAsExternalFinantialSource = TRUE: foreign (20.0/14.0)
|   InovationsInTechnology = TRUE: foreign (112.0/73.0)

Number of Leaves  :         9

Size of the tree :  17
```

Figure 1. J4.8 tree

more focused on preserving the data privacy from the end users that benefit from data mining results, than from data analysts.

This paper presents an application of different filters in Weka and their implications on data mining results. The research was conducted on a dataset on small and medium sized enterprises in province of Vojvodina. For illustration purposes a data subset was derived from a larger set of data using Select Attribute option that evaluates and measures the influence of input attributes on the output or class attribute. The dataset contains the following data on SMEs: Work experience (in years) of a director of the SME; innovations as part of business policy of SME; innovations in technology in the last two years; limiting factors for increase of innovations in the enterprise: insufficient stimulants, non exiting limitations; source of financing business activities (working and investment capital); leasing as an external financial source; rates as the greatest problem in usage of credits; the condition of goods or services on the market; quality as a competitive factor of SME; number of employees; average age of employees; deficit of skilled workers. The output attribute was set to the type of investments the SMEs conducted in the last five years, in terms of introducing new technologies or similar capital investments.

The J4.8 classifier was applied on the presented dataset in order to classify instances according to their investment activities: no investments, no answer on the given question, foreign investments, domestic investments or both foreign and domestic investments. The percentage split option was used that divides the presented dataset into training and test sets giving an opportunity to evaluate the effects of classification. The J4.8 classifier applied to the original data values resulted in 56.59% correctly classified instances. Figure 1 shows the J4.8 pruned tree. The number of instances that have reached the branch correctly is placed prior the slash, while the number of incorrectly classified instances is stated after the slash.

In the sequel of the paper the results of application of different filters in Weka on the same dataset are presented.

The *Randomize filter* in Weka randomly shuffles the order of instances passed through it. Whenever a new set of instances is passed in, the random number generator is reset with the given seed value. In presented research, the default value of 42 for randomSeed was used, as well as values 54, 15, and 100. These values influence the randomization of instances and reflect on the results. After randomization the J4.8 classifier generated the results given in Table I.

As can be seen from the above example, the classification results decreased from not very satisfactory ones in the original classification to almost unacceptable ones after the application of the J4.8 filter.

A similar filtering technique in Weka is the *Resample filter* that produces a random subsample of a dataset using either sampling with replacement or without replacement. Opposite from the previous example, after each application of the Resample filter, the generated result was more satisfying. There were 60.44% correctly classified instances in the first run of the filter. With the second run there were 62.64%, then 77.47%, 82.42%, going to 95%. Unfortunately, the size of the classification tree grew simultaneously, and ended at the size of 193 with 162 leaves.

AddNoise filter is an instance filter that changes a percentage of a given attribute values. The attribute must be nominal and the missing values can be treated as values itself. The percentage of introduced noise to data was varied, which then influenced the result of the J4.8 classifier. As figure 2 illustrates, adding noise to data changed the structure of data.

The greater the percentage of introduced noise, greater is the black surface on the graph, representing added noise to data. The data represents investments in technology in the last two years. The results after application of this filter are given in Table II. Again, it can be seen that they are unsatisfactory.

TABLE I
Classification results on randomized instances

Random seed value	Correctly Classified Instances	Incorrectly Classified Instances
15	51.0989 %	48.9011 %
15	45.0549 %	54.9451 %
54	49.4505 %	50.5495 %
54	47.2527 %	52.7473 %
54	52.1978 %	47.8022 %
100	40.6593 %	59.3407 %

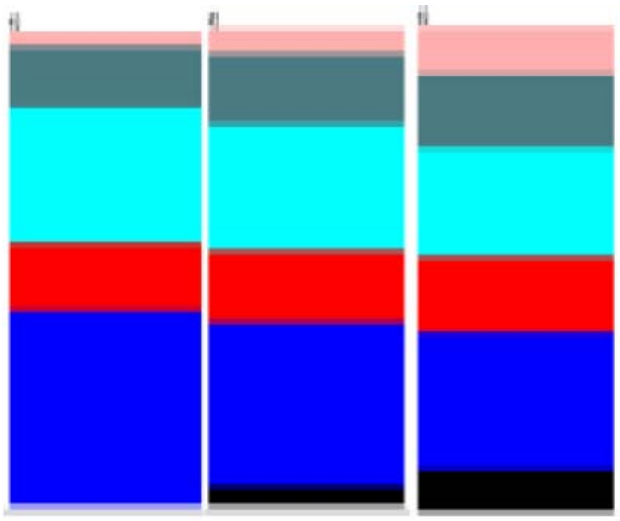

Figure 2. Noise added to data

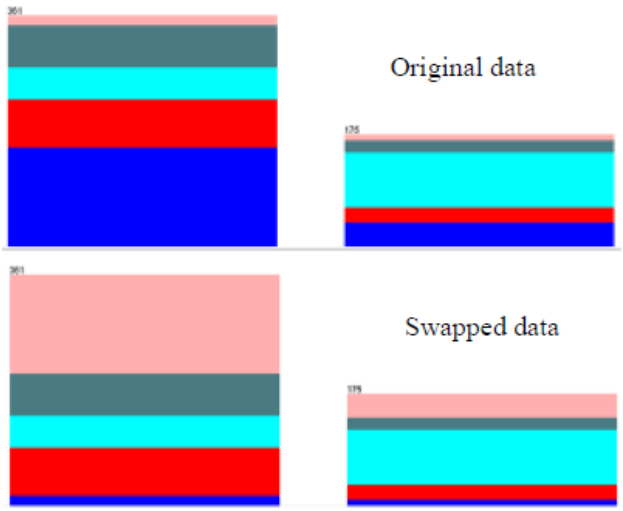

Figure 3. Data structure prior and after swapping

The *SwapValue filter* swaps two values of a nominal attribute. A user can specify which attribute(s) will be processed by this filter. After swapping the values of the attribute Interest Rate, the number of correctly classified instances using J4.8 classifier decreased to 28.30%. Figure 3 illustrates how swapping values changed the structure of data on Interest Rate as a limiting factor for SMEs for broader utilization of credit funds. The first column represents SMEs that stated this factor is a limiting one (value TRUE) and the second column represents enterprises that assessed this factor to be false.

Although the filters in Weka tool do not provide an absolute privacy protection, users can benefit from their correct and combined utilization. As it is emphasized in the paper, privacy-preserving is of utmost importance in today's business environment, unfortunately from the point of application of data mining methods and techniques it imposes certain difficulties. To determine real benefit utilization of different techniques for privacy preserving brings, it is necessary to measure the achieved degree of privacy preserved in data and the amount of data and information that is lost in this process. In addition, presented paper illustrates on the practical examples that application of different filters in Weka, whose functionality is similar to the techniques for privacy-preserving described in the first part of this paper, led to quite poor results.

TABLE II.

Classification results after adding noise to data

Random seed value	Correctly Classified Instances	Incorrectly Classified Instances
10	40.11%	59.89%
10	43.43%	56.57%
20	35.26%	64.74%
20	32.94%	67.06%
35	14.81%	85.19%
35	30.82%	69.18%
5	28.66%	71.34%
5	31.45%	68.55%

IV. CONCLUSION

Data privacy protection is extremely important in contemporary business environment. The privacy-preserving is an issue of concern for individuals as well as for business companies. Companies able to protect and secure customer data will gain better relationship with their customers, customers' satisfaction, trust and loyalty.

But empirical research shows that this issue is also a source of many difficulties, proportional to its importance. Furthermore, available privacy-preserving methods and techniques often fail to give satisfactory results and thus jeopardize the data mining process. In this paper experiments in Weka tool are described that speak in favor of the above statement. It can also be concluded that generalization is hardly possible, so each data mining task brings a unique privacy-preserving challenge to data analysts.

Further research directions should include measuring the degree of privacy preserved during data mining against data or information loss, and possible improvements in achieved performance by combining diverse privacy-preserving techniques.

REFERENCES

[1] Accenture, How global organizations approach the challenge of protecting personal data, including a comparison with Dutch organizations, 2010, <http://www.accenture.com/NR/rdonlyres/836A71D2-4E7E-4E42-A778-4D8A4059 6296/0/Accenture_Data_privacy_ reportLD.pdf>, (pristupano 20.6.2011.).

[2] Charu C. Aggarwal, Philip S. Yu, "Privacy-Preserving Data Mining: Models and Algorithms", Kluwer Academic Publishers Boston/Dordrecht/London.

[3] DataLossDB open security foundation, Data Loss Statistics <http://datalossdb.org/statistics>, (pristupano 20.6.2011.).

[4] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, Official Journal of the European Comunities No L 281/31, 1995, <http://ec.europa.eu/justice_home/fsj/privacy/docs/95-46-ce/dir1995-46_part1_en.pdf>, (pristupano 28.01.2010.).

[5] Sweeney, L. "Achieving k-anonymity privacy protection using generalization and suppression", International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 571-588.

[6] Waikato Environment for Knowledge Analysis, available at http://www.cs.waikato.ac.nz/ml/weka/.

[7] Machanavajjhala A., Kifer D., Gehrke J., Venkitasubramaniam M., "ℓ-Diversity: Privacy Beyond k-Anonymity", ACM Journal Transactions on Knowledge Discovery from Data (TKDD), Vol. I, Issue 1, March 2007, Pages 1–47.

[8] Du W., Zhan Z., "Using randomized response techniques for privacy-preserving data mining", In Proceedings of The 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 24-27 2003.

[9] Kargupta H., Datta S., Wang Q., Sivakumar K., "Random Data Perturbation Techniques and Privacy Preserving in Data Mining", Journal Knowledge and Information Systems, Volume 7, Issue 4, Springer-Verlag New York, May 2005.

[10] Doganay M.C., Pedersen T., Saygin Y., Savas E., Levi A., "Distributed privacy preserving k-means clustering with additive secret sharing", Published in: Proceeding PAIS '08 Proceedings of the 2008 international workshop on Privacy and anonymity in information society, ISBN: 978-1-59593-965-4, 2008.

[11] Urabe S., Wong J., Kodama E., Takata T., "A high collusion-resistant approach to distributed privacy-preserving data mining", Published in Proceeding PDCN'07 Proceedings of the 25th conference on Proceedings of the 25th IASTED International Multi-Conference: parallel and distributed computing and networks, 2007.