Effects of Data Anonymization on the Data Mining Results

Ines Buratović, Mario Miličević and Krunoslav Žubrinić
Department of Electrical Engineering and Computing - University of Dubrovnik
Ćira Carića 4, 20000 Dubrovnik, Croatia
E-mail: ines.buratovic@stud.unidu.hr, mario.milicevic@unidu.hr, krunoslav.zubrinic@unidu.hr

Abstract - This article examines the possibility of publication of students' data, such as secondary school success, state graduation exam scores and success during their first year of university study for analyses. In order to discover data patterns and relationships using data mining techniques, the data must be released in the form of original tuples, instead of pre-aggregated statistics. These records contain sensitive and even confidential personal information, which implies significant privacy concerns regarding the disclosure of such data. Removing explicit identifiers prior to data release cannot guarantee anonymity, since the datasets still contain information that can be used for linking the released records with publicly available collections that include students' identities. One of the privacy preserving techniques proposed in the literature is the *k*-anonymization. The process of anonymizing a data set usually involves generalizing data records and, consequently, it incurs loss of relevant information. In the primary research undertaken in the University of Dubrovnik's students' database the effect of anonymization has been measured by comparing the results of mining the original data set with the results of mining the altered data set to determine if it is possible to use anonymized data for research purposes.

I. INTRODUCTION

Privacy, as defined in [1], is the right of a person to determine which personal information about himself/ herself may be communicated to others. In the terms of data publishing, privacy is the right of a person or an entity to be secure from unauthorized disclosure of sensitive information. Sensitive information could be contained in an electronic repository, or can be derived as aggregate or complex information from data stored in an electronic repository.

Privacy preservation has become an important issue in many data mining applications. Traditionally, the data are published in the form of representative statistics, or pre-aggregated parts that others might be interested in. Data released in these forms lack flexibility as it cannot be used for data mining purposes. In order to discover data patterns and relationships using data mining techniques, the data must be released in the form of microdata, i.e., data in the original form of individual tuples. Obviously the release of microdata offers significant advantages in terms of information availability. However, the publication of microdata raises privacy concerns when published records contain sensitive or confidential information.

Croatian legal system settles data privacy with the *Law on the protection of data confidentiality* and the *Law on the protection of personal information*. Trade secret, as described in the *Law on the protection of data confidentiality* [2], is data defined as trade secret by law, other regulation or general act of a company, institution or other legal entities, which represent the manufacturing secret, the results of research or design work whose disclosure to an unauthorized person could have harmful effects on the person's economic interests. The *Law on protection of the personal information* [3] regulates the protection of individuals' personal data as well as supervision over collecting, processing and use of personal information in the Republic of Croatia. The purpose of the protection of personal information is the protection of private life and other basic human rights and freedoms during collecting, processing and using personal data.

According to the Croatian legislature [3], personal data are any information which refer to an identified person, or a person who can be identified directly or indirectly, based on one or more specific traits based on his or her physical, psychological, mental, economic, cultural or social identity. Personal information contained in the databases must be adequately protected from deliberate or accidental misuse, destroying, loss or unauthorized access or tampering. Persons with authority to manage databases containing personal information, as well as the database users, must take all possible technical, human resources and organizational actions in order to protect the databases from the described unwanted events, as well as from any other misuse. It is also necessary for the keepers of personal information to know and understand their obligations in terms of protecting confidentiality of this information.

Privacy preserving data mining [4] has been proposed as a paradigm of exercising data mining while protecting the privacy of individuals. To protect the privacy of the respondents to which the data refer, released records are usually "sanitized" by removing all explicit identifiers such as names, personal identification numbers, addresses, and phone numbers. Although apparently anonymous, the de-identified data may contain other data that often combine uniquely and can be linked to publicly available information to re-identify individuals. To avoid such linking attacks, while preserving the integrity of the released data, Samarati and Sweeney have proposed the concept of *k*-anonymity [5, 6].

A dataset is said to provide *k*-anonymity when the contained data do not allow the recipient to associate the released information to a set of individuals smaller than *k*, meaning that each record is indistinguishable from at least $k - 1$ other records within the dataset. Since it is highly impractical to make assumptions on which data are known to a potential attacker and can be used to identify respondents, *k*-anonymity requires that, in the released dataset, the respondents be indistinguishable (within a given set) with respect to the set of attributes, called quasi-identifier, that can be exploited for linking [7]. In other words, k-anonymity requires that if a combination of values of quasi-identifying attributes appears in the dataset, then it appears with at least *k* occurrences [8].

II. PROBLEM DEFINITION

Students enrolling in programs at the University of Dubrovnik come from different surroundings, and have diverse backgrounds and abilities. They have finished secondary education in various secondary schools, while having significantly diverse success. At the same time, in every generation at the University, there are several students who are unable to keep pace with the lectures and cannot achieve the learning outcomes of the curriculum.

Experience shows that factors like background knowledge, success in secondary school and social background, among others can influence academic success or failure. Discovering the patterns and regularities between these factors and academic success can be used to identify potentially unsuccessful future students already upon enrollment. These students could then be assisted and supported to approach university studies in a way that will increase their chances of success (e.g. additional tutoring, mentorship).

Dataset containing information about students' background and academic success should be disclosed for the analyses to the various departments at the University, and perhaps even outside the institution. To enable performing detailed analyses of the students' data, the data ought to be released in the form of original tuples. Meanwhile, the original dataset contains students' personal and confidential data; therefore its disclosure raises significant privacy concerns, and the dataset should be anonymized prior to the disclosure.

The given dataset consists of records for the students who enrolled in the University for the first time in the 2010/2011 academic year. It contains information on 255 students. Each student is described by 21 attributes.

In order to describe students' backgrounds and secondary school performances, several attributes are used in the dataset. For instance, their place of residence, finished secondary school, profession, secondary school GPA (*Grade Point Average*) and scores on state graduation exams (SGE) in general education subjects that student was taught during secondary school, i.e. the Croatian language, mathematics and a foreign language. For the description of student's academic success, attributes, such as first year GPA, earned credit points (*European Credit Transfer System* - ECTS) and number of passed courses in the first year, can be used.

Even without explicit identifiers such as student's name and student identification number, there are still attributes in the dataset that can be used in combination to identify certain students. E.g. place of residence, secondary school and scores on the state graduation exams combined can distinguish certain students, making them recognizable in the set.

III. EXPERIMENTAL RESULTS

A. Data anonymization

Two main techniques that have been proposed for enforcing *k*-anonymity on a dataset are generalization and suppression, both preserving the truthfulness of the data. The application of *k*-anonymity algorithms produces more general datasets that provide protection of the identities. While applying those algorithms it is important to minimize loss of precision and completeness, in order to keep data mining results as accurate as possible.

Ensuring *k*-anonymity of the students' data in the described dataset is done using the Samarati's algorithm. That algorithm uses generalization and tuple suppression over quasi-identifiers to find a *k*-minimal generalization that suppresses tuples. Data owner sets a maximum number of tuples that can be suppressed, and the algorithm computes a generalization that satisfies *k*-anonymity within that constraint. The proposal considers the application of generalization at the attribute (column) level and suppression at the tuple (row) level [7].

However, if the data contains a large number of attributes which may be considered quasi-identifiers, sometimes it becomes difficult to anonymize the data without an unacceptably high amount of information loss [9].

Among the attributes in given student dataset, the following were chosen as quasi-identifiers:

- place of residence (postal code),
- secondary school,
- age at enrollment,
- profession,
- secondary school GPA and
- scores on the state graduation exams.

Generalization hierarchies are determined for each of the quasi-identifiers. Generalization at the attribute level ensures that all values of an attribute belong to the same domain. However, as a result of the generalization process, the domain of an attribute can change. Note that, since the domain of an attribute can change and since generalized values can be used in place of more specific ones, it is important that all the domains in a generalization hierarchy be compatible [6].

Figure 1 shows an example on generalization hierarchy for postal code attribute. Postal codes can be generalized by dropping, at each generalization step, the least significant (rightmost) digit. Since majority of students at the University of Dubrovnik comes from Dubrovnik and Dalmatia area (postal codes starting with

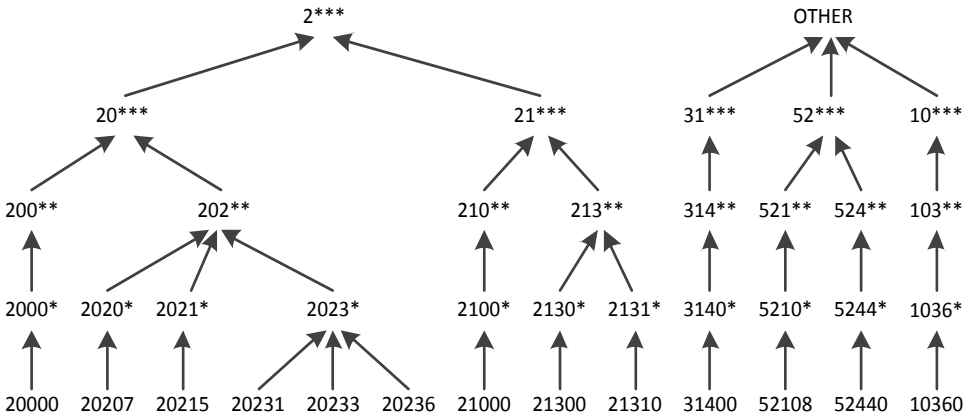

Figure 1. An example on generalization hierarchy for postal code

digit "2"), to decrease the probability of suppressing the minority who comes from different cities, the last step of postal code generalization consists of grouping the postal codes into two groups: "postal codes starting with digit 2" and "others".

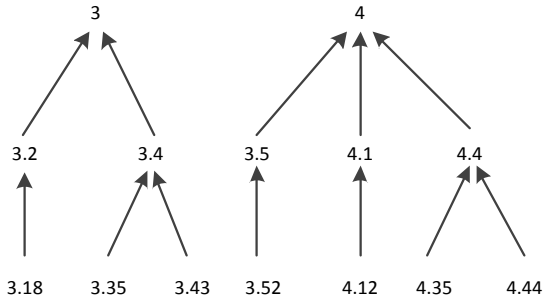

Figure 2. An example on generalization hierarchy for secondary school GPA

For the example of generalization hierarchy for secondary school GPA consider Figure 2. At each generalization step, the GPA is rounded to a certain number of decimal places, while that number being smaller at each step.

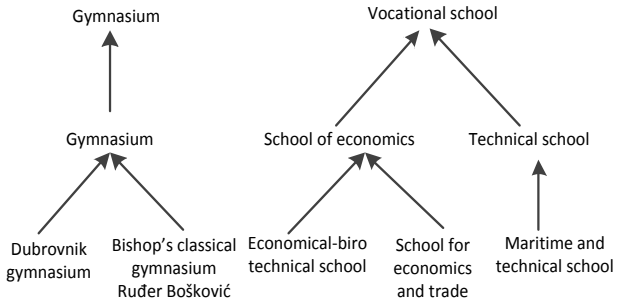

Figure 3. An example on generalization hierarchy for secondary school

Secondary schools, as shown in Figure 3, can be generalized by grouping specific schools into more general ones at each step.

Anonymity of the data is provided using Samarati's algorithm exploiting both the generalization and suppression with the level of anonymity *k*=2 and the maximum number of tuples that can be suppressed set to 10%. The goal of 2-anonymization is to alter the data in the dataset in a way that every combination of values of quasi-identifiers can be indistinctly matched to at least 2 students. Since the dataset is relatively small and level of anonymity *k*=2 ensures basic anonymity, that level is selected for the experiment.

The attributes are generalized each one step at a time, until the 2-anonymity of the data is satisfied, or there are less than 10% outliers that can be suppressed left. When the second condition is satisfied, the outliers are suppressed to avoid the overgeneralization of the significant attributes.

In order to keep the utility of the data, while still preserving students' privacy it is important to choose the order of the generalization of the attributes carefully. It is assumed that in determining the outcome of data mining, less significant attributes can be generalized more than more significant attributes. Based on that presumption, the quasi-identifiers are generalized in the reverse order of their significance. The result of that is that the most significant attributes are less generalized then those of lower importance.

The order of the attributes to be generalized is determined using *ReliefF* method [10]. *ReliefF* algorithm is a general and successful attribute relevancy estimator. It is able to detect conditional dependencies between attributes and provides a unified view on the attribute relevancy estimation in regression and classification.

In this case the *ReliefF* evaluation algorithm gives all attributes an average merit score and an average rank according to the importance of attributes in the classification:

0.1011765 2 Department
0.092549 1 Undergr_Study
0.0729412 18 Enroll_Date
0.05435 11 Sec_School_GPA
0.0478431 8 Profession
0.0321569 4 Sec_School
....

TABLE 1. THE NUMBERS OF DISTINCT VALUES THROUGH THE PROCESS OF ANONYMIZATION

Attribute	Distinct values				
	Original dataset	*After 1st step of generalization*	*After 2nd step of generalization*	*After 3rd step of generalization*	*After 4th step of generalization*
Postal code	52	32	19	11	3
Secondary school	35	7	3	-	-
Age at enrollment	9	3	2	-	-
Profession	13	10	5	-	-
Secondary school GPA	158	27	4	-	-
SGE results - Croatian language	114	79	12	4	2
SGE results - Foreign language	190	89	13	4	2
SGE results - Mathematics	54	48	12	4	2

-0.0074952 15 State_Grad_FL_Result
-0.0094118 10 Sel_Rank
-0.0109804 12 State_Grad_FL_Level
-0.0243137 7 Gender

Table 1 shows hierarchical levels for attributes that are finally chosen and how many possible values there are for each attribute through the process of anonymization. After performing the generalization, there are still 22 instances left that don't satisfy the given level of anonymity. Those instances are suppressed.

The anonymity of the dataset is enforced through the combination of SQL statements and scripts written in C programming language. Development of such data anonymization software is planned for the future work.

B. *Data mining*

The patterns and regularities in the student dataset are discovered using data mining techniques. The data are analyzed with different data mining methods using 10-fold cross-validation. For the purpose of this research, students' success is evaluated by the number of earned ECTS credits, where the workload for an academic year totals to 60 ECTS credits. Credits are analyzed as continuous and discrete class, but since the goal is to identify students who will not successfully meet all requirements specified with the program of studies even two-class classification gives satisfactory results. The tool used for data mining research is WEKA (*Waikato Environment for Knowledge Analysis*) [11].

The best results are achieved using the classical *Naive Bayes* method [11] based on Bayes' rule which says that for a hypothesis H and evidence E that bears on that hypothesis, the conditional probability of H given E can be calculated as

$$Pr[H \mid E] = (Pr[E \mid H]Pr[H])/Pr[E], \quad (1)$$

where Pr[A] denotes the probability of event A and Pr[A|B] denotes the probability of A conditional on

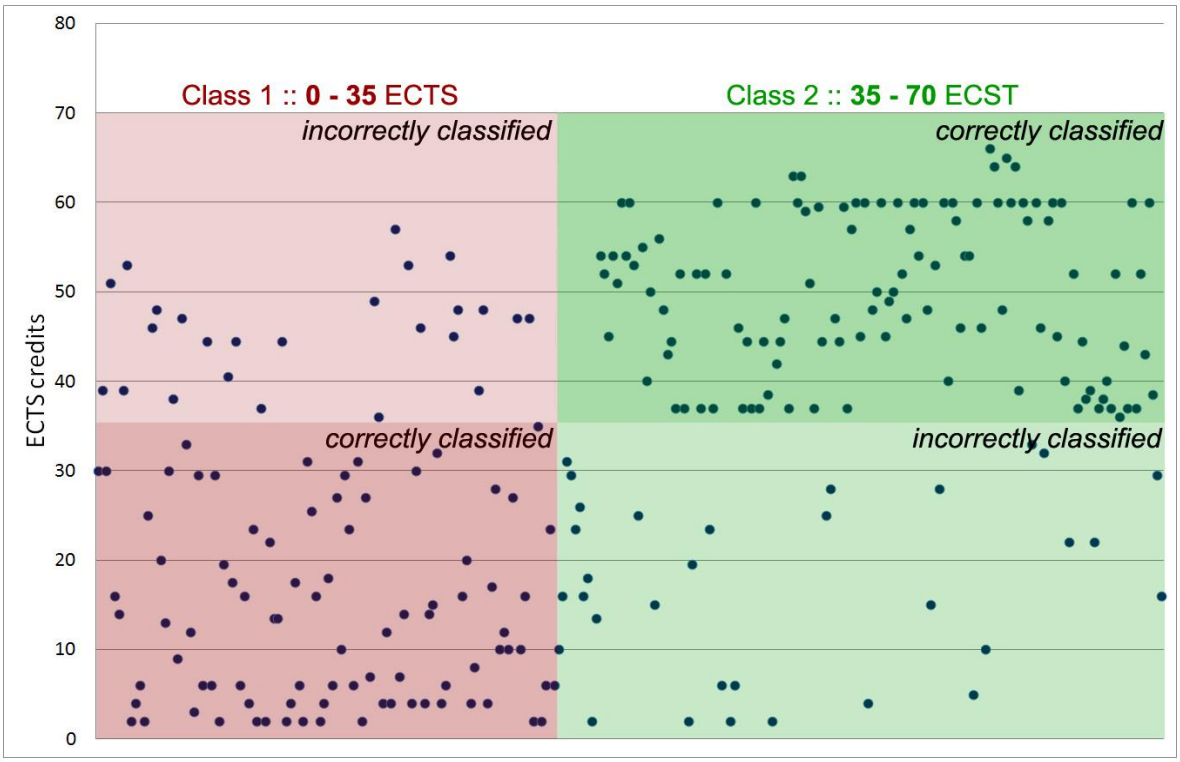

Figure 4. Two-class classification results

another event B.

This method goes by the name of *Naive Bayes* because it assumes that attribute values are independent of each other given the class. The assumption that attributes are independent (given the class) in real life surely is a simplistic one, but *Naive Bayes* works very effectively when tested on actual datasets.

First, the two-class classification is performed with the original learning dataset (with no anonymization applied) and 77.647% of instances were correctly classified. Figure 4 shows the results of two-class classification.

It is important to mention that confusion matrix shows that among 22.353% incorrectly classified instances there are only 9.8% of students who will not achieve satisfactory results (false positives).

In the second step, the effects of anonymization (i.e. generalization and suppression) were analyzed. It was necessary to verify that it is still possible to achieve satisfactory performance of the classification process. Experiments have shown that the classification accuracy remained almost unchanged (76.394%). Good results are partly a consequence of the fact that the process of anonymization of the dataset takes into account the impact of individual attributes on the results of the classification. That result has proven our presumption on role of significance of quasi-identifiers in the generalization process.

IV. CONCLUSION

Release of the personal data in the form of microdata for the analyses poses a threat to the privacy of respondents' whose information is included in the datasets. In order to protect the respondents' privacy and still be able to perform analyses of their information, the data in the dataset must be de-identified.

This article provides an example of combining generalization and suppression using described algorithm to achieve *k*-anonymity within a dataset containing personal information of students at the University of Dubrovnik. It demonstrated that carefully planned and implemented data anonymization that takes into account the impact of the individual attributes on the result of data mining, can at the same time preserve privacy of the students and keep the results of data mining almost intact, making it possible to release anonymized data for research purposes.

REFERENCES

[1] E. Bertino, D. Lin, and W. Jiang, „A Survey of quantification of privacy preserving data mining algorithms" in Privacy-Preserving Data Mining, vol. 34, Springer US, 2008, pp. 183– 205

[2] Law on the protection of data confidentiality, Official Gazette of the Republic of Croatia, 108/1996, in Croatian

[3] Law on the protection of personal information, Official Gazette of the Republic of Croatia, 103/2003, in Croatian

[4] R. Agrawal and R. Srikant. „Privacy-preserving data mining". In ACM-SIGMOD Conference on Management of Data, pp. 439–450, 2000.

[5] P. Samarati, L. Sweeney, „Generalizing data to provide anonymity when disclosing information" in Proceedings. of the 17th ACM SIGMOD-SIGACT-SIGART Symposium on the Principles of Database Systems, 1998

[6] L. Sweeney, „k-anonymity: a model for protecting privacy". International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570.

[7] P. Samarati, „Protecting Respondents' Identities in Microdata Release", IEEE Transactions on Knowledge and Data Engineering, Vol 13 (6), 2001, pp. 1010–1027.

[8] V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati, „k-Anonymous Data Mining: A Survey", in Privacy-Preserving Data Mining, Vol. 34 (2008), pp. 105-136.

[9] C. C. Aggarwal. "On k-anonymity and the curse of dimensionality". In Proc. of the 31th VLDB Conference, Trondheim, Norway, September 2005.

[10] M. Robnik-Sikonja, I. Kononenko, „Theoretical and Empirical Analysis of ReliefF and RreliefF" in Machine Learning Journal, Vol. 53, 2003, pp. 23-69

[11] I. H. Witten, E. Frank, M. A Hall. "Data mining: practical machine learning tools and techniques", third edition, Morgan Kaufmann, San Francisco, 2011.