Fuzzy methods for database protection

Vicenç Torra[1] Daniel Abril[1] Guillermo Navarro-Arribas[2]

[1]IIIA, Artificial Intelligence Research Institute
CSIC, Spanish Council for Scientific Research
Campus Universitat Autònoma de Barcelona
08193 Bellaterra (Catalonia, Spain)
[2]DEIC, Dept. of Information and Communications Engineering
UAB, Universitat Autònoma de Barcelona
Campus Universitat Autònoma de Barcelona
08193 Bellaterra (Catalonia, Spain)

Abstract

Data privacy has become an important topic of research. Ubiquitous databases and the eclosion of web technology eases the access to information. This information can be related to individuals, and, thus, sensitive information about users can be easily accessed by interested parties. Data privacy focuses on tools and methods to protect the privacy of the respondents and data owners.

In the last years, a large number of methods have been developed for data privacy. Some of them are based on fuzzy sets and systems. In this position paper we present a review of some of our results in this area. In particular, we focus on the use of fuzzy sets for data protection, for measuring information loss and for measuring disclosure risk. The techniques used in this field and reviewed in this paper range from fuzzy clustering to fuzzy integrals.

Keywords: Data privacy, fuzzy clustering, fuzzy measures, fuzzy integrals.

1. Introduction

Data protection is a field of research that is receiving much attention. The amount of data and information available in databases is now increased by the information in web pages and social networks. Some of this information is sensitive, and the aggregation of these data might lead to detailed profiles of people.

To avoid the disclosure of sensitive information, data protection methods have been developed. The fields that study these methods are Statistical Disclosure Control (SDC) [22] and Privacy Preserving Data Mining (PPDM) [2]. The differences between these two fields root in their origin. SDC has its origin in the National Statistical Offices and the need of publishing the data obtained from census and questionnaires for researchers or policy makers. PPDM has its origin in the data mining community, and methods are more oriented to companies that need to share the data either with other companies or with researchers.

At present there exists a large number of techniques devoted to data protection. They can be roughly classified according to different dimensions. Most methods have been developed so that data owners apply them, and the differences are about owner's knowledge about the computation intended by a third party. Then, in this scenario, and following [16], we classify data protection methods into computation-driven, data-driven, and result-driven approaches.

- Computation-driven methods are those in which the data owner knows which computation has to be applied to the data by a third party. Then, protection methods are developed so that the third party obtains the expected results with the protected data. It is usual here to use cryptographic protocols and, then, the results of the third party are exactly the same that would be obtained from the original data but without any disclosure of sensitive information.
- Data-driven methods are those in which the data owner does not know which analysis will be performed by the third party. It is even possible that the data owner does not know who the third party is. This is the case of data published in the web by a National Statistical Office (NSO) as any researcher can use the data. Data-driven methods often use perturbation approaches. These approaches are applied to the data and then used by the third parties in their analysis. Then, the results obtained by the third party are only an approximation of the results that would be obtained with the original data.
- Result-driven methods are those in which privacy concerns to the results of the data mining process. This would be the case of applying algorithms for mining association rules. We want that the application of these algorithms does not return rules that lead to disclosure.

We have applied some techniques based on fuzzy sets in the setting of data protection. More specially, in the case of data-driven methods. In this

paper, we will discuss in Section 2 the typical scenario of data-driven or perturbative methods, then in Section 3 we will describe some of our contributions in this area in which fuzzy sets play a role. The paper finishes with some conclusions and lines for future research.

2. Data-driven approaches

As stated in the introduction, data-driven approaches corresponds to methods that are applied when the data owner does not know the type of analysis to be applied to the data. In this case, protection is driven by the data. That is, methods are selected according to the type of data available. Different methods exist according to the data available. The literature presents methods for e.g. databases with numerical or categorical [13, 14] (either ordinal or nominal), time series, locations (for location privacy), access logs, search logs [8], graphs [7, 11] (for online social networks).

All data-driven methods follow a similar strategy. They modify the data introducing some kind of perturbation. This perturbation is expected to be enough to ensure protection of the sensitive information and at the same time low enough so that the data utility is not lost.

In order to evaluate these methods, information loss measures (utility measures) and disclosure risk measures have been developed. Then, a good data protection method is one that achieves a good trade-off between information loss and disclosure risk.

In order to give a more formal definition of these measures, let X represent the original data, ρ the data protection mechanism, and X' the perturbated data. Naturally, $X' = \rho(X)$.

Then, information loss measures are defined in terms of the divergence, for a particular set of analyses, between the results of an analysis on the original data X and the same analysis on the protected data X'. So, if α is the analysis, the information loss IL corresponds to:

$$IL(X, X') = divergence(\alpha(X), \alpha(X')).$$

If X is a numerical database, α can be e.g. the mean on the variables or a clustering algorithm.

There exist different definitions for disclosure risk measures. We follow a computational approach based on reidentification [23]. In our setting, we assume that when the protected data set X' is published, an intruder tries to link her data with the published file X'. Then, if correct links are established between intruder's information and the published data file X', disclosure takes place. For this attack, we presume that the information the intruder has corresponds to a subset of the original data set X. Let us consider the following example.

Example 1 *Let X be data from a hospital that is expected to be transferred to a data miner so that she can establish some relationships between illness, residence, age and profession. This data is protected using a certain protection method leading to X'. Then, let us assume that the data miner, acting as an intruder, wants to* attack *the protected data X' based on some acquaintances she has. If she knows some tuples* (person | (residence, age, profession))*, she might be able to infer the illness for these acquaitances. That is, the intruder, has a subset Y of X annotated with the name of the people.*

In a more formal way, we have X, X' and the file of the attacker Y which is a subset of X. Y will be annotated with the identifiers (the names or social security numbers) to make disclosure possible.

Under this scenario, the proportion of correct links between Y and X' is a measure of risk.

In order to link files Y and X' we can use record linkage algorithms. In its simplest form, record linkage algorithms link each record of the intruder y in Y with the nearest record x in the protected file X' using the Euclidean distance. If the nearest record, say x^*, corresponds to the protected version of x then we have a confidentiality breach. That is, let

$$x^*(y) = \arg\min_{x \in X'} ||x - y||, \tag{1}$$

then the following measure of disclosure risk (DR) is considered:

$$DR(Y, X') = \frac{|\{y = x^*(y) | y \in Y\}|}{|Y|}. \tag{2}$$

In this definition $|\cdot|$ denotes the cardinality of the set.

3. Fuzzy sets-based approaches in privacy

We have developed several methods for data protection based on fuzzy sets. We review them in this section. There are methods for data protection based on fuzzy techniques, measures for information loss based on fuzzy techniques, and also reidentification methods based on fuzzy approaches. We will review other contributions in this section.

3.1. Data protection methods based on fuzzy techniques

Using the notation given above, a data protection method is a function ρ that applied to X returns the file X'. In the literature there are different families of functions ρ for this purpose. The main three classes are perturbative methods, non-perturbative methods and methods for synthetic data generation. Perturbative methods modify the original data introducing some noise (some kind of error is introduced to the records), the non-perturbative methods modify the original data changing the level of detail but there is no erroneous data (e.g., change of the granularity), and the synthetic data is based

on constructing models of the data and then replacing the original data by the one generated with the models.

We have contributed in perturbative methods and on synthetic data generators using techniques inspired on fuzzy sets.

Among perturbative approaches, microaggregation is a very effective method. This method proceeds with the construction of small clusters of data, and then replaces each record by the cluster representative. If all clusters are required to be of a certain minimum size, privacy is ensured. If all clusters are required to be small, the perturbation of the original data is small (i.e., information loss is small and data utility is kept maximal). We have contributed with methods based on fuzzy clustering. Fuzzy c-means permits us to solve one of the inconveniences of applying standard crisp clustering in data privacy: an intruder can easily guess which is the centroid assigned to a particular record. This information can be easily used by the intruder to attack the database. This problem as well as a possible solution is explained in detail in [19].

In addition, we have developed approaches that take into account the existing constraints on the variables of the data. We understand constraints here as in data editing [21]. That is, the data in files have to satisfy some established constraints. For example, in a data file we have that the variable corresponding to retail price is defined in terms of net price plus taxes, or we might have ages in a data file and these ages should be positive and less than 130. Then, when a file is protected, the protected dataset should satisfy also these constraints. If not, the protected database will be inconsistent with the schema of the database. Data editing [6, 21] is the field in official statistics that focuses on the construction of a database satisfying the constraints.

Microaggregation is a data protection mechanism that is specially well suited when data edit constraints are present. We have developed an approach for this problem based on fuzzy c-means (see [15] for details). Our variation of fuzzy c-means is such that the resulting centroids are required to satisfy the constraints on the variables even when these constraints are not satisfied in the original data.

We have also contributed to data protection methods based on synthetic data generators as e.g. [12, 3]. Our contribution was based on fuzzy c-regression models. Our approach permits the user to have some flexibility that is not available for other methods. Fuzzy c-regression models use a parameter c that corresponds to the number of clusters (i.e., regression models). Then, the larger the parameter c, the more clusters we have, and the more regression models we have. This parameter permits the user to find a better trade-off between information loss and disclosure risk. That is, large c has low information loss but high risk, and small c has large information loss and low risk. Then, an appropriate c permits to have a good balance between information loss and disclosure risk. This flexibility is not present in other approaches based on synthetic data generators. Details on this approach can be found in [4].

3.2. Information loss measures based on fuzzy techniques

An important issue once methods are built is the evaluation of its information loss. Strictly speaking, information loss depends on the data uses. That is, if a third party wants to apply a regression model (i.e, following the notation above $\alpha = regression$), then the information loss should be measured in terms of the divergence in the regression.

We have studied information loss when the user wants to apply clustering to the data and, more specifically, to the case of applying fuzzy clustering to the data. We have considered several methods for fuzzy clustering. Among them, α being the Fuzzy c-means, noise clustering, possibilistic c-means and fuzzy possibilistic c-means. We have studied information loss for some standard methods for data protection as e.g. microaggregation but also for some synthetic data generators [10] as IPSO [3] and our method based on fuzzy c-regression models [5]. We have also compared [9] the measures of information loss obtained with our approach and the ones obtained with some generic information loss measures (e.g., measures based on standard statistics as the mean, variances, covariances and correlations of the database).

The comparison of fuzzy clusters is not an easy task. Two problems arise, one is about the comparison of two fuzzy partitions. Another problem is that fuzzy clustering methods are typically implemented with methods that only ensure the convergence to a local optima. We have addressed the two problems. In addition, the need to compare fuzzy clusters taking into account the uncertainty of the fuzzy clustering methods have lead us to the definition of interval-valued or intuitionistic fuzzy partitions [18, 20]. That is, fuzzy partitions in which the membership value of an element to a cluster is an interval instead of a number in $[0, 1]$.

3.3. Disclosure risk measures based on fuzzy techniques

Fuzzy techniques have also been applied to measure disclosure risk. As stated above, one of the approaches for computing disclosure risk [23] is to count the number of records of an intruder that can be linked to the protected file. The most standard approach for linking the two files is to use an Euclidean distance to measure the dissimilarity between pairs of records. Nevertheless, other distances can be used. The literature includes results using e.g. Mahalanobis distance and also kernel-

based distance. From a formal point of view, given a certain distance d with parameter p, we have that the disclosure risk is defined by

$$DR(Y, X') = \frac{|\{y = x^*(y) | y \in Y\}|}{|Y|},$$

where

$$x^*(y) = \arg \min_{x \in X'} d_p(x, y).$$

Naturally, when $d_p = AM$, where AM stands for the arithmetic mean, the results of this formulation are equivalent to the ones obtained with Equation 1 in the expression DR in Equation 2. Using this definition, we can use any other distance.

We have used distances based on weighted means, OWA operators and also Choquet integrals. In the case of a weighted mean and the OWA operator, the parameter p is a weighting vector; in the case of the Choquet integral the parameter p is a fuzzy measure.

The consideration of a parameterized aggregation operator for computing the distance permits us to consider the corresponding optimization problem. That is, we can study which is the parameter p which maximizes the disclosure risk. Then, given a pair of files Y and X', and a distance function d defined in terms of a parameter p, we can determine the p that maximizes the number of reidentifications. This p is the best parameter that an intruder might have, and, thus, an upper bound of the disclosure risk.

We have applied this optimization approach to the weighted mean, the OWA [17] and the Choquet integral [1]. The optimization problem is a quadratic problem with linear constraints. Due to this, the (global) optimal solution can be found, although it might require a large computation power.

We have seen that the weighted mean gives better results than the OWA, and that the Choquet integral leads to better results than the weighted mean. We have also seen that for most of the cases in which all variables are protected with the same data protection method, the use of a Choquet integral only improves in a minor way the results of the weighted mean and the OWA operator. In contrast, in the case of data files in which different data protections are applied to different variables, the use of a Choquet integral is useful as the number of correct links is significantly increased with respect to the results of the other methods.

In fact, the use of a model for reidentification based on a Choquet integral permits us to elicit a fuzzy measure from the data. This fuzzy measure represents the relationships between the variables in the dataset. When the variables are all protected using the same method, we have that the measure learnt from the dataset leads to results similar to a Choquet integral with an equiprobable probability distribution on the weights. That is, in this case the fuzzy measure represent independent variables. In contrast, when different methods are applied to different variables, the learnt fuzzy measure represents the relationships between the variables: the measure shows that some variables are protected together and others not.

4. Conclusion and future work

In this paper we have reviewed the application of fuzzy approaches to data privacy. We have shown that methods based on fuzzy sets can be developed and used in several different problems. In particular, we have shown that they can be either used when defining a data protection method or also in the evaluation of the methods. With respect to the evaluation of methods, fuzzy approaches can be used either for measuring information loss or for assessing disclosure risk.

Most of the results described above are for files with numerical data, and all of them have been applied only to files with this type of data. Nevertheless, for some of the methods, other types of data could be used in a similar way. For example, data protection methods based on fuzzy c-means could be easily extended to non-numerical data (e.g., categorical or time-series). Information loss measures based on fuzzy clustering could also be applied to non-numerical data. Finally, the results on reidentification could also be applied to other types of data.

Acknowledgements

Partial support by the Spanish MEC (projects ARES – CONSOLIDER INGENIO 2010 CSD2007-00004 –, eAEGIS – TSI2007-65406-C03-02 –, and N-KHRONOUS – TIN2010-15764) is acknowledged.

References

[1] Abril, D., Navarro-Arribas, G., Torra, V. (2010) Choquet Integral for Record Linkage, manuscript.

[2] Aggarwal, C. C., Yu, P. S. (eds.) (2007) Privacy-Preserving Data Mining, Springer.

[3] Burridge, J. (2003) Information preserving statistical obfuscation, Statistics and Computing, 13:321–327.

[4] Cano, I., Torra, V. (2009) Generation of Synthetic Data by means of fuzzy c-Regression, Proc. FUZZ-IEEE 2009 (ISBN: 978-1-4244-3597-5), DVD-ROM, Jeju Island, Korea, August, 2009, 1145-1150.

[5] Cano, I., Ladra, S., Torra, V. (2010) Evaluation of Information Loss for Privacy Preserving Data Mining through comparison of Fuzzy Partitions, Proc. FUZZ-IEEE 2010 / WCCI (ISBN in press), Barcelona, 2010.

[6] Granquist, L. (1997) The new view on editing, Int. Statistical Review 65:3 381-387.

[7] Hay, M., Miklau, G., Jensen, D. (2008) Anonymizing Social Networks, Proc. VLDB 2008.
[8] Jones, R., Kumar, R., Pang, B., Tomkins, A. (2007) "I know what you did last summer": query logs and user privacy, Proc. CIKM pp. 909-914
[9] Ladra, S., Torra, V. (2008) On the comparison of generic information loss measures and cluster-specific ones, Int. J. of Unc., Fuzz. and Knowledge Based Systems 16 (Supplement-1) 107-120.
[10] Ladra, S., Torra, V. (2010) Information loss for synthetic data through fuzzy clustering, Int. J. of Unc., Fuzz. and Knowledge Based Systems 18:1 25-37
[11] Liu, K., Terzi, E. (2008) Towards identity anonymization on graphs, Proc. SIGMOD 2008.
[12] Muralidhar, K., Sarathy, R. (2008) Generating Sufficiency-based Non-Synthetic Perturbed Data, Transactions on Data Privacy 1:1 17 - 33
[13] Sweeney, L. (2002) Achieving k-anonymity privacy protection using generalization and suppression, Int. J. of Unc., Fuzz. and Knowledge Based Systems 10:5 571-588.
[14] Sweeney, L. (2002) k-anonymity: a model for protecting privacy, Int. J. of Unc., Fuzz. and Knowledge Based Systems 10:5 557-570.
[15] Torra, V. (2009) On the Definition of Linear Constrained Fuzzy c-Means, Proc. of the EUROFUSE 2009 (ISBN: 978-84-9769-242-7), Pamplona, Spain, September 2009, 61-66.
[16] Torra, V. (2010) Privacy in Data Mining, in O. Maimon, L. Rokach (Eds.), Data Mining and Knowledge Discovery Handbook, 2nd Edition (invited Chapter).
[17] Torra, V., Navarro-Arribas, G., Abril, D. (2011) Supervised Learning for record linkage through weighted means and OWA Operators, Control and Cybernetics, in press.
[18] Torra, V., Min, J.-H. (2010) I-Fuzzy Partitions for Representing Clustering Uncertainties, Proc. ICAISC 2010, Lecture Notes in Artificial Intelligence 6113 (2010) 240-247.
[19] Torra, V., Miyamoto, S. (2004) Evaluating fuzzy clustering algorithms for microdata protection, Lecture Notes in Computer Science 3050 (2004), 175-186.
[20] Torra, V., Miyamoto, S. (2011) A definition for I-fuzzy partitions, Soft Computing, in press.
[21] De Waal, T. (2008) An overview of statistical data editing. Statistics Netherlands.
[22] Willenborg, L., de Waal, T. (2001) *Elements of Statistical Disclosure Control*, Lecture Notes in Statistics, Springer-Verlag.
[23] Winkler, W. E. (2004) Re-identification methods for masked microdata, PSD 2004, Lecture Notes in Computer Science 3050 216-230.