International Journal of Management, IT & Engineering
Vol. 10 Issue 01, January 2020
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com Double-
Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gate as well as in Cabell's Directories of Publishing Opportunities, U.S.A

A study of Fuzzy Based Approach for Securing Information in Data mining

Shrikant Zade (IEEE member: 92529270)
Research Scholars,CSE Deptt.
Mewar University,Rajasthan (India)

Dr.Chandikaditya Kumawat
Professor, CSE, MEWAR,Uni.
Rajasthan

Dr.Pradeep Chouksey,
Asso. Prof., CSE Deptt,
LNCT,Bhopal

Abstract:

Extracting previously unknown patterns from huge volume of data is the main objective of any data mining technique. In recent days there are tremendous growth in data collection due to the computerization. The patterns revealed by data mining algorithm can be used in various domains like Image Analysis, Marketing and weather forecasting, etc. As a side effect of the mining algorithm some sensitive information is also revealed. There is a need to preserve the privacy of individuals which can be achieved by using privacy preserving data mining. In this paper we study a new approach to preserve sensitive information using fuzzy logic.

Keywords: Privacy preserving, k-means, fuzzy data set, classification

1. Introduction

In the present era, use of information technology has resulted in the generation of data at a phenomenal rate in all areas including business, manufacturing, and healthcare. This explosive growth in stored data has generated an urgent need for new techniques that can transform the vast amount of data into useful information and knowledgebase. Data mining is, perhaps, the most suitable technique to satisfy this need.

Typically, data are stored in a table, and each record (row) corresponds to one individual. Each record has a number of attributes, which can be divided into the following three categories: 1) Attributes that clearly identify individuals. These are known as explicit identifiers and include, e.g., Social Security Number. 2) Attributes whose values when taken together can potentially identify an individual. These are known as quasi-identifiers, and may include, e.g., Postal code, Birth-date, mobile number and Gender. 3) Attributes that are considered sensitive, such as Disease and Salary. When releasing such data, it is necessary to prevent the sensitive information

of the individuals from being disclosed. Two types of information disclosure have been identified in the literature [1, 2]: identity disclosure and attribute disclosure. Identity disclosure occurs when an individual is linked to a particular record in the released table. Attribute disclosure occurs when new information about some individuals is revealed, i.e., the released data make it possible to infer the characteristics of an individual more accurately than it would be possible before the data release. Identity disclosure often leads to attribute disclosure. Once there is identity disclosure, an individual is re-identified and the corresponding sensitive values are revealed.

We have studied an equivalence class of an anonymized table to be a set of records that have the same values for the quasi-identifiers. To effectively limit disclosure, we need to measure the disclosure risk of an anonymized table. The introduced k-anonymity as the property that each record is indistinguishable with at least k-1 other records with respect to the quasi-identifier [3, 4]. In other words, k-anonymity requires that each equivalence class contains at least k records. While k-anonymity protects against identity disclosure, it is insufficient to prevent attribute disclosure. [5]

In such cases, one of the solutions that has been suggested is to perturb or map the data [2]. While such mapping should not enable the receiving party to derive the original values from the mapped values, they should also be able to retain the same association with other attributes as the original data. Such challenges also frequently arise when a data owner decides to offload the data mining activity to a third party such as a cloud operator. Here, the data owner needs to provide data to the cloud operator with a risk of losing privacy over the sensitive data. So, it is best for the data owner to transform the sensitive fields and provide the modified data to the cloud operator. However, there should be a straight forward and computationally simple means for the data owner to map the association rules derived from the modified data to those that apply to original data.

The primary goal of privacy preserving is to hide the sensitive data before it gets published. For example, a hospital may release patient's records to enable the researchers to study the characteristics of various diseases. The raw data contains some sensitive information of individuals, which are not published to protect individual privacy. However, using some other published attributes and some external data we can retrieve the personal identities. Table 1

shows a sample data published by a hospital after hiding sensitive attributes. (Ex. Patients name).

Table 1. Raw-Data Attribute

ID				
	Age	Sex	Zip Code	Disease
1	22	F	613001	Fever
2	33	M	613002	Fever
3	44	M	613003	Headache
4	55	F	613004	Cough

Table 2. Voter Registration List

ID				
	Name	Age	Sex	Zip Code
1	Asha	22	F	613001
2	Sachin	33	M	613002
3	Dhoni	44	M	613003
4	Lata	55	F	613004

Table 2 shows a sample voter's registration list. If an opponent has access to this table he can easily identify the information about all the patients by comparing the two tables using the attributes like (zip-code, age, sex).These types of attributes are called as Quasi identifier attributes. The rest of the paper is organized as follows:

Section 2 describes the various methods that can be used for privacy preserving in data mining. Section 3 provides an insight on the conventional K-means algorithm. Section 4 explains about the fuzzy based membership function approach and how it can be used for privacy preserving. Section 5 shows the conclusion and scope of fuzzy based for privacy preserving in data mining.

2. Conventional k -means clustering for PPDM: Let r be the number of parties, each having different attributes for the same set of entities. n is the number of the common entities. The parties wish to cluster their joint data using the k-means algorithm. Let k be the number of clusters required. The final result of the k-means clustering algorithm is the value/position of the means of the k clusters, with each side only knowing the means corresponding to their own attributes, and the final assignment of entities to clusters.

Let each cluster mean be represented as μi, i

= 1, . . . , k. Let μij represent the projection of the mean of cluster i on party j. Thus, the final result for party j is

• the final value/position of μij , i = 1 . . . k

• cluster assignments: clusti for all points (i 1, . . . , n)

The k-means algorithm also requires an initial assignment (approximation) for the values/positions of the k means.

Their technique uses classic k-means clustering done over multiple subsamples of the data, followed by clustering the results to get the initial points. For simplicity, we assume that the k means are selected arbitrarily. Thus, for i = 1 . . . k, every party selects its share μ0 ij of any given mean. This value is local to each party and is unknown to the other parties. The basic algorithm directly follows the standard k-means algorithm. The approximations to the true means are iteratively refined until the improvement in one iteration is below a threshold. At each iteration, every point is assigned to the proper cluster, i.e., we securely find the cluster with the minimum distance for each point Once these mappings are known, the local components of each cluster mean can be computed locally.

3. Various methods for privacy preserving in data mining:

3.1 Random Perturbation methods:

One of the earliest work in the privacy-preserving data mining is by Agrawal and Srikant [2]. Here, the authors investigate how different privacy preserving perturbations affect the association rules. In particular, they investigate how the original data distribution can be

estimated based on the perturbed data. They have considered three types of perturbation: Value class membership where every data value corresponding to a class (e.g., bin in a histogram) is assigned the same value. They also look into value distortion where data is perturbed by a random number. They determined that such random perturbation with Gaussian distribution is most effective [6]. They observe that random perturbations do not necessarily preserve data privacy. Clifton et al summarize different techniques for privacy preserving [7]. Here, some high-level techniques are suggested for solving the problem [8].

We studied three recent papers [9,10,11]. Here, the authors propose a fuzzy-based mapping to map sensitive numerical data to another domain for privacy preservation. Similarly, they propose a decision-tree based approach for mapping categorical data values. In this paper, we concentrate on the numerical data and fuzzy-based mapping.

Let us take a brief look at their method. If the sensitive field values are in the range min-max, then this range is divided into *k* fuzzy sets. Each data value is mapped into one of these five fuzzy sets. In order to distinguish one value of the set from another, a numerical affinity (they refer to it as intensity) term is used. In other words, numerical data is mapped into different fuzzy sets, yet retaining their value in terms of an affinity value.

3.2 Data-driven approaches:

The 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 [12, 13] (either ordinal or nominal), time series, locations (for location privacy), access logs, search logs [14], graphs [7, 12] (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 tradeoff between information loss and disclosure risk.

In order to give a more formal definition of these measures, let X represent the original data, β the data protection mechanism, and X' the perturbated data.

Naturally, $X' = \beta(X)$.

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 β is the analysis, the information loss IL corresponds to:

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

If X is a numerical database, β 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 re-identification. In this setting, 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.

4. Fuzzy sets-based approaches in privacy

There are several methods for data protection based on fuzzy sets. These methods are for data protection based on fuzzy techniques, measures for information loss based on fuzzy techniques, and also re-identification methods based on fuzzy approaches.

4.1. Data protection methods based on fuzzy techniques:

Using the notation given above, a data protection method is a function F that applied to X returns the file X'. In the literature there are different families of functions F 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.

1. *Improved Representation through* n-*ary quantifiers and semi-fuzzy quantifiers* An n-ary *fuzzy quantifier* Q on a base set E ≠ Ø assigns to each choice of fuzzy subsets X1,X2.X3,…, Xn of E a gradual result Q (X1…Xn) [0; 1]. Fuzzy quantifiers constitute an expressive class of operators because they introduce a second order construct for fuzzy sets.[17]

However, they are often hard to define because the familiar concept of cardinality of crisp sets is not applicable to the fuzzy sets that form the arguments of a fuzzy quantifier. It is necessary to introduce a simplified representation, which must be still powerful enough to embed all quantifiers in the sense of TGQ. An n-ary *semi-fuzzy quantifier* on a base set E is a mapping which to each choice of *crisp* subsets Y1; : : : ; Y n of E assigns a gradual result Q(Y1; : : : ; Yn) [0; 1]. Because semi- fuzzy quantifiers must be defined for crisp input only, they are much easier to define than fuzzy quantifiers. In particular, the usual crisp cardinality is applicable to their arguments and can hence be used to provide an interpretation for semi-fuzzy quantifiers[18].

2. A *quantifier fuzzification mechanism:*
(QFM) F assigns to each semi-fuzzy quantifier Q a fuzzy quantifier F(Q) of the same arity and on the same base set. These are applicable both to crisp and fuzzy arguments. QFMs are useful because the concepts of TGQ can be easily adapted to the case of semi- fuzzy quantifiers and fuzzy quantifiers. We can then require that a certain property of a quantifier be preserved when applying the QFM, and that F be compatible with certain constructions on (semi-) fuzzy quantifiers. This can be likened to the well-known mathematical concept of a homomorphism (structure-preserving mapping).

3. We require *compatibility* with concepts of TGQ; an adequate QFM should preserve all properties of linguistic relevance.

International Journal of Management, IT & Engineering
Vol. 10 Issue 01, January 2020
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com Double-
Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at:
Ulrich's Periodicals Directory ©, U.S.A., Open J-Gate as well as in Cabell's Directories of Publishing Opportunities, U.S.A

4. We should find *models of the axioms*, i.e.

'reasonable' choices of F, and characterize interesting classes of such models in terms of distinguished properties;

4.2. 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 β = *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 [16]. 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 intuitionist fuzzy partitions. 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].

4.3. 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 [15] 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. From a formal point of view, given a certain distance *d* with parameter *p*, we have that the disclosure risk is defined by

$$D\ (Y/X')R|\{y\ x^*(y)|y\ Y|/|Y|$$

Where

$$x^*(y) \ \arg\min dp\ (x, y)$$

$$xX'$$

Naturally, when $dp = AM$, where AM stands for the arithmetic mean, the results of this formulation are equivalent to the ones obtained with above equation. Using this definition, we can use any other distance. 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 $X0$, and a distance function d defined in terms of a parameter p. This p is the best parameter that an intruder might have, and, thus, an upper bound of the disclosure risk. [19,20]

5. Conclusion: In fact, the use of a model for re-identification 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 equi-probable 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.

[1] Qian Wang, Xiangling Shi," (a, d) Diversity: Privacy Protection Based on l-Diversity" Software Engineering, 2009 pp.367-372.

[2] Agrawal, R. and Srikant, R, "Privacy-preserving data mining", In Proc. SIGMOD00, 2000, pp. 439-450.

[3] P. Samarati, "Protecting respondents' identities in microdata release," in Transactions on Knowledge and Data Engineering, pp. 1010 – 1027, 2001.

[4] L. Sweeney, "*k*-anonymity: a model for protecting privacy", International Journal on Uncertainty, Fuzziness and Knowledge based Systems, 2002, pp. 557-570.

[5] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam," *l*-diversity: Privacy beyond *k*-anonymity". In *Proc. 22nd Intnl. Conf. Data Engg. (ICDE)*, page 24, 2006.

[6] H. Kargupta, S. Datta, Q. Wang, K. Sivakumar, Random data perturbation techniques and privacy preserving data mining, Knowledge and Information Systems 7 (4) (2003) 387–414.

[7] C. Clifton, M. Kantarcioglu, J. Vaidya, X. Lin, M. Zhu, Tools for privacy preserving distributed data mining, ACM SIGKDD Explorations 4 (2) (2002) 28–34.

[8] M.D. Singh, P.R. Krishna, and A. Saxena, "A Privcaypreserving Jaccard Similarity function for Mining Encrypted Data," *IEEE TENCON 2009 Proc*., 2009, pp. 1-4.

[9] V. Vallikumari, S.S. Rao, K.V.S.N. Raju, K.V. Ramana, and B.V.S. Avadhani, "Fuzzy-based approach for privacy preserving publication of data," *Intl. J. Computer Science And Network Security*, Vol. 8, No. 1, 2008, pp. 115-121.

[10] E. Poovammal and M. Ponnavaikko, "An Improved Method for Privacy Preserving Data Mining," *2009 IEEE IACC, 2009*, pp. 1453-1458.

[11] E. Poovammal and M. Ponnavaikko, "Preserving Micro data Release: categorical and Numerical Data," *2009 IEEE SETIT*, March 2009, pages 5.

[12] 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.

[13] Sweeney, L. (2002) *k*-anonymity: a model for protecting privacy, Int. J. of Unc., Fuzz. and Knowledge Based Systems 10:5 557-570.

[14] 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

[15] Winkler, W. E. (2004) Re-identification methods for masked microdata, PSD 2004, Lecture

Notes in Computer Science 3050 216-230.

[16] 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

[17] G. Bordogna and G. Pasi. "A fuzzy information retrieval system handling users", preferences on document sections. In Dubois et al. [3].

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

[19] Torra, V., Navarro-Arribas, G., Abril, D. (2011) Supervised Learning for record linkage through weighted means and OWA Operators, Control and Cybernetics, in press.

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