Fuzzy based approach for privacy preserving publication of data

V. Valli Kumari, S.Srinivasa Rao, KVSVN Raju, KV Ramana and BVS Avadhani

Andhra University, Visakhapatnam, India

Summary

Data privacy is the most acclaimed problem when publishing individual data. It ensures individual data publishing without disclosing sensitive data. The much popular approach, is K-Anonymity, where data is transformed to equivalence classes, each class having a set of K- records that are indistinguishable from each other. But several authors have pointed out numerous problems with K-anonymity and have proposed techniques to counter them or avoid them. l-diversity and t-closeness are such techniques to name a few. Our study has shown that all these techniques increase computational effort to practically infeasible levels, though they increase privacy. A few techniques account for too much of information loss, while achieving privacy. In this paper, we propose a novel, holistic approach for achieving maximum privacy with no information loss and minimum overheads (as only the necessary tuples are transformed). We address the data privacy problem using fuzzy set approach, a total paradigm shift and a new perspective of looking at privacy problem in data publishing. Our practically feasible method in addition, allows personalized privacy preservation, and is useful for both numerical and categorical attributes.

Key words:

Privacy preserving, data privacy, fuzzy information, anonymity

1. Introduction

Data base sharing has become a common phenomenon with the advent of technology and global networking. Assume a hospital publishes patient information for statistical use, by suppressing the identifying attributes, if any. As voter registration lists are publicly accessible, a careful correlation of the attributes in the two lists would reveal sensitive information about an individual [11,14, 19]. For example, disease, which he did not wish to reveal, might get revealed. The attributes that help in revealing information when combined with other attributes are called as Quasi Identifier [QI] attributes. The attributes that hold private information about an individual and should not be disclosed are called as sensitive attributes.

K-anonymity[7] is one widely discussed approach for achieving data privacy. In K-anonymized data, privacy is achieved through generalization and suppression. Suppression of directly identifiable attributes, like name, SSN is done by not publishing them. Then the data set shown in table 1 is divided into equivalence classes. Each equivalence class has a distinct tuple occurring k-times, which is called generalization. Thus, generalization means replacing a tuple with a more generalized tuple, which is indistinguishable from several other tuples in the equivalence class as in table 2. This is also called as anonymization[1]. But several problems are identified with K-anonymity [2, 17].

A K- anonymous table may allow an adversary to derive the sensitive information of an individual with 100% confidence. There is considerable information loss from the data. And it does not take into account personalized anonymity requirements[20].

Table 1: Microdata

Age	Gender	Zip code	Disease	Name
52	Male	530001	Diabetes	A
31	Female	530209	Malaria	B
40	Female	536702	Typhoid	C
28	Male	538796	Typhoid	D
60	Male	537777	Typhoid	E
20	Male	534444	Viral fever	F

Table 2: K-Anonymity property satisfied (K=2)

Age	Gender	Zip code	Disease	Name
50-60	Male	53****	Diabetes	A
30-40	Female	53****	Malaria	B
30-40	Female	53****	typhoid	C
20-30	Male	53****	Typhoid	D
50-60	Male	53****	Typhoid	E
20-30	Male	53****	Viral fever	F

Manuscript received January 5, 2008
Manuscript revised January 20, 2008

K-Anonymity allows an attacker to discover the value of sensitive attributes, when there is little diversity in the sensitive attributes [6, 21]. To counter this, another scheme called l-diversity was proposed. *l-diversity* provides privacy even when the data publisher does not know what kind of knowledge is possessed by the adversary. It ensures that, all tuples that share the same values of quasi identifiers should have l- diverse values for their sensitive attributes. Even l-diversity is prone to attacks by an adversary, as it guarantees a low breach probability [16,23]. Anatomy [24] is another l-diversity specific method. Though it does not violate the l-diversity property, it confirms that a particular individual is included in the data. *t-closeness* is another scheme, which recommends table-wise distribution of Sensitive Attribute values to be repeated within each anonymised group [8].

Personalised privacy preservation is another method which allows each sensitive attribute in a record in the table to have a privacy constraint[10, 16]. However, the computational effort is too high as generalization has to be done again on the sensitive attribute column. Personalized privacy preservation uses a tree based approach, for personalized privacy. Greedy algorithm is used and hence is not optimal, so does not achieve minimal loss. p-sensitive k-anonymity is almost similar to l-diversity[5].

Extended p-sensitive K-anonymity is a scheme that extends p-sensitive k-anonymity property that which is similar to the personalized privacy method, where in the protection is offered at different levels in the taxonomy for the sensitive attribute [4]. Another scheme in [18] assumes hierarchy in each QI attribute, and that all partitions in a general domain should be at the same level of hierarchy.

Contributions of the paper: Our study has shown that all these techniques increase computational effort, though they increase privacy. A few techniques account for too much of information loss, while achieving privacy. We address the data privacy problem using fuzzy set approach, a total paradigm shift and a new perspective of looking at privacy problem in data publishing. The domain generalization based solution completely disassociates the sensitive values with the identifying attributes. Our practically feasible method in addition, allows personalized privacy preservation, and is useful for both numerical and categorical attributes.

Outline of the paper Section 2 consists of a brief overview of fuzzy sets. Section 3 contains a description of the fuzzy based privacy preserving model. Section 4 discusses the experimental results and the Informativeness metric along with the distinguishability metric. Section 5 concludes the paper with a discussion on what needs to be done further.

2.Background

2.1 Fuzzy sets overview

Fuzziness[3, 9] is a way to represent uncertainty, possibility and approximation. Fuzzy sets are an extension of classical set theory and are used in fuzzy logic. In classical set theory the membership of elements in relation to a set is assessed in binary terms according to a crisp condition- an element either belongs to or does not belong to the set. By contrast, fuzzy set theory permits the gradual assessment of the membership of elements in relation to a set; this is described with the aid of a membership function:

$$\mu \rightarrow [0, 1]$$

The domain of the membership function, which is the domain of concern and from which elements of the set are drawn, is called the '*universe of discourse'*. For example, the Universe of discourse of the fuzzy set 'High Income' can be the positive real line $[0, \infty)$.

The notion central to fuzzy systems is that truth values (in fuzzy logic) or membership values (in fuzzy sets) are indicated by a value on the range [0.0, 1.0], with 0.0 representing absolute false and 1.0 representing absolute truth. For example, let us take the statement:
"Jane is old."

If Jane's age was 75, we might assign the statement the truth value of 0.80. The statement could be translated into set terminology as "Jane is a member of the set of old people." This statement would be rendered symbolically with fuzzy sets as:

$$\mu_{\text{OLD}}(\text{Jane}) = 0.80$$

Where μ is the membership function, operating in this case on the fuzzy set of old people, which returns a value between 0.0 and 1.0. The modifiers of fuzzy values are called Hedges. To transform the statement, "Jane is old" to "Jane is very old", the hedge "very" is usually defined as follows:

$$\mu_{\text{"very"}} A(x) = \mu\, A(x) \mathbin{\char`^} 2.$$

For example, If μ_{OLD} (Jane) =0.8 then μ_{VERYOLD} (Jane) =0.64. Every input value is associated with a linguistic variable.

A linguistic variable represents a concept that is measurable in some way either objectively or subjectively, like temperature or age. Linguistic variables are characteristics of an object or situation. For each linguistic variable it should be assigned a set of linguistic terms (values) that subjectively describe the variable.

Most of the times, linguistic terms are words that describe the magnitude of the linguistic variable, as “hot” and “large”, or how far they are from a goal value as in “exact” or “far”. Each linguistic term is fuzzy set and has its own membership function. It is expected that for a linguistic variable to be useful the union of the support of the linguistic terms cover its entire domain.

3. The privacy-preserving model

Our proposed privacy preserving model primarily has two objectives: preserving privacy while revealing useful information for i) numerical attributes, and ii) categorical (non-numerical) attributes. and to find a generalized table T^*, such that it includes all the attributes of T and an individual tuple from T is not identifiable in T*.

Table 3. Microdata of employee table

Name	Age	Gender	Zipcode	Income
Arun	52	M	12000	10000
Keller	60	M	18000	23000
Mani	81	M	19000	20000
Joe	42	M	22000	58000
Syam	19	M	24000	85000
Rama	21	F	58000	94000

Table 4. Microdata of patient table

Name	Age	Gender	Zipcode	disease
Arun	52	M	12000	Gastric ulcer
Keller	60	M	18000	Pneumonia
Mani	81	M	19000	bronchitis
Joe	42	M	22000	Pneumonia
Syam	19	M	24000	Pneumonia
Rama	21	F	58000	Flu

Let T be a relation holding information about a set of individuals each associated with a tuple t. The attributes in T are classified as: $\forall t \in T$.

1. *Identifier Attributes (A^i)*

These attributes uniquely identify the individual associated with the tuple, as anonymisation requires that the data be disassociated with the identifiers. One specific example is the name attribute.

2. *Sensitive attributes (A^s)*

These attributes should not be disclosed to the public or may be disclosed after disassociating its value with an individual’s other information. A few examples are Income and Disease, as shown in Tables 3 and 4.

3. *Quasi Identifier(A^{qi})*

These values may be published, but it so happens that with a combination of these attributes an individual may get identified. For instance, age and zipcode might disclose the identity.

Thus, to summarise, Name is a member of A^i, sensitive attributes are Income and disease and are members of A^s, members of A^{qi} are Age, Zipcode and Gender. We give a mathematical formalization for these,

A^i={Name}
A^s={Income, Disease}
A^{qi}={Age, zipcode, gender}

When compared to the existing research, we claim that privacy can be achieved in both the cases whether the sensitive attribute is categorical or numerical .

3.1 Privacy and disclosure levels

To reduce the computational effort and increased control of the owner on the his data, an attribute PL (Privacy level) is introduced into the table. The value set to PL tells whether the data is to be released or not , in other words whether, the sensitive attribute needs to be transformed or not. If the user prefers to release the data, he can decide on the level of disclosure by setting the parameter DL (Disclosure level). The value set to DL decides whether the data has to be partially released or can be released in full. The DL is valid only for categorical attributes. Both PL and DL are boolean attributes.

Privacy level (PL)

The user may be given a chance to select his level of privacy by setting the PL to true (t)/false(f). Setting PL to true means, the user does not want to disclose the data at all. So whole of the data in the row pertaining to the user is suppressed. But if the selection is false, the user is willing to give the data. This increases personalized privacy and also, reduces the computational effort.

Disclosure Level (DL)

A DL attribute is attached to each categorical attribute in the table, which allows the user to choose graded personalized privacy.

3.2 Fuzzy based Privacy preserving for numerical attributes

Assume, the data in table 3 is to be published and that the user specified sensitive attribute is Income. Then, the following procedure is followed to transform the table in to a publishable form. In the table, A^s= {income}. As income is a sensitive attribute and is numerical, Rule1 is applied for transforming its values. L is the linguistic term set and $\{l_1, l_2,..l_n\}$ are the linguistic values, A_i^s is the linguistic variable for the attribute A^s and 'n' is the number of linguistic values, 'i' refers to the numerical attributes of T which are sensitive.

Rule 1: If L=$\{l_1, l_2,..l_n\}$,

then $\forall t \in T \quad \forall i \; A_i^s,\; F(A_i^s) \quad \to$ L

Suppose the linguistic term set for the variable income L(A^s=income) is: *{High, Medium, Low}* with membership functions defined as below. The minimum and maximum values of income according to the business organization are *min* and *max* respectively and *a_1, a_2, a_3* are the midpoints of each fuzzy set and *k* is the number of fuzzy sets. The *k* fuzzy sets will have ranges of :

$\{min\text{-}a_2\}, \{a_1\text{-}a_3\}, \{a_{(i-1)}\text{-}a_{(i+1)}\},\ldots,\{a_{(k-1)}\text{-}max\}$.

For fuzzy set with midpoint a_1, the membership function is given by

$$\begin{array}{lll} f_1(x) & = 1.0 & \text{if } x=min \\ & = (x- a_2) / (min\text{-}a_2) & \text{if } x< a_2 \\ & = 0 & \text{if } x>= a_2 \end{array}$$

For the fuzzy set with midpoint a_i, 2<=i<=k-1, the membership function is given by

$$\begin{array}{lll} f_i(x) & = 0 & \text{if } x<=a_{(i-1)} \\ & = (x\text{-}a_{i-1})/(a_i\text{-}a_{i-1}) & \text{if } a_{(i-1)}<x<a_i \\ & = 1.0 & \text{if } x= a_i \\ & = (a_{i+1}\text{-}x)/(a_{i+1}\text{-}\ a_i) & \text{if } a_i<x< a_{i+1} \\ & = 0 & \text{if } x>=a_{(i+1)} \end{array}$$

For fuzzy set with midpoint a_k, the membership function is given by

$$\begin{array}{lll} f_k(x) & =0 & \text{if } x<=a_{(k-1)} \\ & = (x\text{-}\ a_{(k-1)}) / (max\text{-}\ a_{(k-1)}) & \text{if } x>a_{(k-1)} \\ & =1.0 & \text{if } x=max \end{array}$$

For k=3, f_1, f_2, f_3 are functions of low, medium and high respectively. The transformed income attribute values of table 3 after applying the above transformations along with the values of weight (f1, f2, f3) are as given in table 5. This helps the end user of the data to make out the distinction between two attribute values, even though they are mapped to the same linguistic term. For instance , in table 5, both 10000 and 23000 are mapped to low. The relativeness (informativeness) is still maintained by the weight. The weight associated tells that low associated with 10000 is still lower than the low associated with 23000. The data in publishable form will have weight associated with every transformed value as in table 5. However, the Income attribute values are not published.

Table 5. Transformed values

Income	10000	23000	58000	85000	94000
Weight	1.0	0.71	0.73	0.66	0.86
changed to	low	low	Medium	high	high

3.3 Taxonomy based privacy preserving transformation for categorical attributes

For categorical attributes like disease, the following taxonomy tree is taken.

0.0 Disease
1.0 Respiratory Problem
1.1 Flu
1.2 Pneumonia
1.3 Bronchitis
2.0 Digestive problem
2.1 Gastric Ulcer
2.2 Dyspepsia
2.3 Gastritis

The representation for this taxonomy tree is as shown in figure 1.

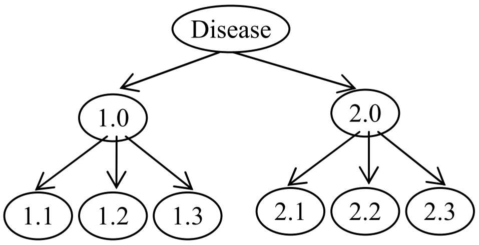

Figure 1. Taxonomy tree for disease

In the Taxonomy tree, pertaining to a specific attribute, we associate the sensitivity level with each such attribute. The user has choice of defining the sensitivity level. For instance for Disease, we have DL ={t, f} If the selected level is true, then the ancestor of the attribute value is returned as a response to query on the that tuple. If the selected level is false, the attribute value itself is returned in response to any query on that tuple.

For each row we have PL as the privacy level, as fixed by the user. The possibilities are t/f. The user may set this value to t if he wants the data to be revealed and f when he does not want to reveal the data. For each sensitive attribute we assign a Disclosure level (DL) specifier. If the user doesn't mind revealing the data, the DL is set to t, if he wants to disclose the data but disassociate the data with himself, the DL is set to f.

Table 6. Transformed values for categorical attribute

PL	Disease	DL	Transformed to
t	Gastric Ulcer	t	Gastric Ulcer
f	Pneumonia	t	Not published
f	Pneumonia	f	Not published
t	Gastric Ulcer	f	Digestive problem

Other attributes: Other attributes like Zip Code and Gender, a taxonomy tree can be constructed.

Thus, T^* contains only those tuples of T for which the owner has set PL to t. These selected tuples contain transformed sensitive categorical and numerical attribute values according to the preferences set by the owner of the data.

4. Experimental Results and metrics

The experiments were carried out on Adult data set from UCI Machine learning repository [22]. The tables were modified by adding attributes disease and income. The taxonomy trees were constructed based on information obtained from literature. It was seen that as fuzzy transformation is just about mapping a given value to a term in the fuzzy set, it took affordable time delay for mapping

4.1 Informativeness metric

The gains that can be had are more information in the data published when compared to the existing methods. Informativeness may be defined as the extent of information or knowledge that can be extracted from the published data. In earlier works, the data was either perturbated with noise or was generalized. When noise is added, informativeness is almost zero, Though goal of privacy will be achieved, there is no way in which the user can use the data. When data is generalized, for instance if age variable takes values, 30,32,34,36,38, and 60, the generalized term would be [30-60]. All k-anonymity based works use this kind of generalisation. The information in the set [30-60] is that the person with age 30 and age 60 are both members of the same set. Further, it can be seen that while five of them are in thirties, only one is in sixty, and still sixty is the member of the set. But in the fuzzy based privacy preservation, when the above set values are mapped to fuzzy set low, it can be seen that when 60 is transformed, it is associated with lesser membership value in the low set and relatively the others are mapped to the same set with higher membership values and the difference exists between different values. It is these membership values that preserve information and informativeness of the proposed method is high when compared to perturbation methods and other k-anonymity related methods.

4.2 Distinguishability

Though each tuples' mapped value is distinct from the others, the proposed method still offers complete privacy as none of the attributes is associated with an identifier.

5. Conclusions and future work

A practically feasible approach for achieving maximum privacy with more information and minimum overheads (as only the necessary tuples are transformed) is proposed. Though dimensionality reduction [12] is proposed in earlier work, that is not be necessary in our work. The data privacy problem is addressed using fuzzy set approach, a total paradigm shift and a new perspective of looking at privacy problem in data publishing. The domain generalization [13] based solution completely disassociates the sensitive values with the identifying attributes. Our practically feasible domain generalisation method in addition, allows personalized privacy preservation, and is useful for both numerical and categorical attributes. Furthermore, we would like to extend our experimental work to optimize the time spent on processing the tuples, and enforcing the privacy constraints at the time of data retrieval.

6. References

[1] Benjamin C.M. Fung, Ke Wang, and Philip S. Yu. "Anonymizing classification data for privacy preservation". In IEEE Transactions on Knowledge and Data Engineering, vol-19, No. 5, 2007.

[2] M. Ercan Nergiz and C. Clifton. "Thoughts on k-Anonymization". In Proc of 22nd Intl. Conf. on Data Engineering (ICDEW06), IEEE Computer Society, 2005.

[3] K.Sridevi, KVSVN Raju, V.Valli Kumari and S.Srinivasa Rao. "Privacy Preserving in Clustering by Categorizing Attributes using Privacy and Non Privacy Disclosure Sets", WORLDCOMP'07,The 2007 Intl. Conf. on Data Mining, pp301-307,June 2007,LasVegas,USA.

[4] Alina Campan and Trajan Marius Truta. "Extended p-Sensitive k-Anonymity". Studia University, Vol 1, 2006.

[5] Trajan Marius Truta, Bindu Vinay. "Privacy Protection: P-Sensitive K-Anonymity Property", Proceedings of the Workshop on Privacy Data Management, In 22nd IEEE Intl. Conf. of Data Engineering (ICDE), Atlanta, Georgia, 2006.

[6] Machanavajjhala A., Gehrke J., Kifer D. , "l-diversity: privacy beyond k-anonymity". Proceedings of the 22nd IEEE Intl. Conf. on Data Engineering, 2006.

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

[8] Ninghui Li , Tiancheng Li and Suresh.V. " t-Closeness: Privacy beyond k-anonymity and l-diversity". ICDE 2007,23rd IEEE Intl. Conf. on Data Engineering, 2007

[9] L. A Zadeh, "Fuzzy sets", Information and control, vol.8, pp.338-353, 1965.

[10] Vijay S. lyengar. "Transforming Data to Satisfy Privacy Constraints". In proc. ACM SIGKDD '02 Edmonton, Alberta, Canada, 2002

[11] L. Sweeney. "Datafly: A system for providing anonymity in medical Data". In Intl. Conf. on Database Security, pages 356–381, 1998.

[12] Kristen LeFevre, David J. DeWitt and Raghu Ramakrishnan. "Multidimensional K-Anonymity". Dept. of Computer Sciences Technical Report 1521, 2005

[13] Kristen LeFevre, David J. DeWitt and Raghu Ramakrishnan. "Incognito: Efficient Full Domain K-Anonymity". In proceedings of ACM SIGMOD'05, USA, 2005.

[14] B.C.M. Fung, K. Wang, and P.S. Yu, "Top-Down Specialization for Information and Privacy Preservation," Proc. 21st Intl. Conf. Data Engineering (ICDE '05), pp. 205-216, Apr. 2005.

[15] G. Xexeo, Belchio, A.D Rocha, A. R., "Evaluating Software Quality Requirement using Fuzzy Theory," Proceedings of ISAS 96, Orlando, July 1996.

[16] X. Xiao and Y. Tao, "Personalized Privacy Preservation," Proceedings of ACM SIGMOD International Conference. Management of Data, June 2006.

[17] K. Wang, P. Yu, and S. Chakraborty, "Bottom-Up Generalization: A Data Mining Solution to Privacy Protection," Proc. Fourth IEEE Intl. Conf. on Data Mining (ICDM '04), Nov. 2004.

[18] P. Samarati, "Protecting Respondents' Identities in Microdata Release," IEEE Transactions on Knowledge Engineering., vol. 13, no. 6, pp. 1010-1027, 2001.

[19] L. Sweeney. "K-anonymity: A model for protecting privacy". Intl. Journal on Uncertainty, Fuzziness, and Knowledge-based Systems, 10(5):557{570), 2002.

[20] I. Dinur and K. Nissim. "Revealing information while preserving privacy". In PODS, pages 202–210, 2003.

[21] S. Zhong, Z. Yang, and R. N. Wright. "Privacy-enhancing k-anonymization of customer data". In PODS, 2005

[22] D.J. Newman, S. Hettich, C.L. Blake, and C.J. Merz, "UCI Repository of Machine Learning Databases, Available at www .ics. uci. edu/~ mlearn/MLRepository.html, University of California, Irvine, 1998.

[23] K. Wang, B.C.M. Fung, and P.S. Yu, "Handicapping Attacker's Confidence: An Alternative to k-Anonymization," *Knowledge* and Information Systems: J. (KAIS), *2006.*

[24] Xiao, X., Tao, Y Anatomy: Simple and Effective privacy preservation. In Proceedings of the 32nd Very Large Data Bases conference (VLDB), pp. 139-150, Seoul, Korea, September 12-15, 2006.

Dr. V. Valli Kumari holds a PhD degree in Computer Science and Systems Engineering from Andhra University Engineering College, Visakhapatnam and is presently working as Professor in the same department. Her research interests include Security and privacy issues in Data Engineering, Network Security and E-Commerce. She is a member of IEEE and ACM and is a fellow of IETE

Mr. S. Srinivasa Rao holds an MTech in Computer Science and Systems Engineering from Andhra University and is currently working as Associate Professor in MVGR Engineering College, Vizianagaram. His research interests include Data Mining.

Dr. KVSVN Raju holds a PhD degree in Computer Science from IIT, Kharagpur, and is presently working as Professor in the department of Computer Science and Systems Engineering at Andhra University Engineering College, Visakhapatnam. His research interests include Software Engineering Data Engineering and Data Security.

Mr. KV Ramana holds an M.Tech degree in Computer Science and Systems Engineering from Andhra University, Visakhapatnam and is presently working as Assistant Professor in the same department. His research interests include privacy issues in Data Engineering and Web Technologies.

Dr.BVS.Avadhani holds a PhD in Applied Mathematics from Andhra University and is currently working in department of Computer Science and Systems Engineering Andhra University His research interests include mobile computing, optimization and Data Engineering.