Fuzzy Based Approach for Privacy Preserving in Data Mining

Shrikant Zade[#1], Dr. Pradeep Chouksey[*2],Dr.R.S.Thakur[$3]
[#]*Research Scholars: CSE Department, Mewar University, Chhitorgarh, India*
**Associate Professor: CSE Department, TIT-College, Bhopal, India*
[$]*Associate professor, MCA Deptt, MANIT, Bhopal*

Abstract— *Data privacy is the major issue in privacy preserving. It confirms that data of individual publish without disclosing sensitive data of that person. The most popular scheme, is k-anonymity, where data is transformed into equivalence classes, each class having a set of k- records that are indistinguishable from each other. But several authors have pointed out number of issues with k-anonymity and have proposed techniques to counter them or avoid them. The l-diversity and t-closeness are another technique for the same. We studied all possible technique with computational efforts, though they increase privacy, some techniques have too much of information loss, while achieving privacy.*

In this paper, we propose a novel, holistic approach to achieve maximum privacy with no information loss and minimum overheads. We address the data privacy problem using fuzzy inference system approach by using Gaussian membership transform function, a total paradigm shift and a new perspective of looking at privacy problem in privacy preserving data mining. Our approach is for both numerical and categorical attribute.

Keywords— ***Privacy preserving, data privacy, fuzzy inference system, Gaussian membership function***

I. INTRODUCTION

With the advancement of digitization, data sharing over the internet is a common practice. Let us consider a hospital publishes patient information for data mining activities, by suppressing the identifying field, if any. As election voter data are publicly accessible, if a careful correlation takes place of the attributes in the two dataset, it find out sensitive information about any individual [8,9]. For example, disease, which he did not wish to disclose, might get be disclosed. The attributes that help in revealing information when combined with attribute of other data sets are called as Quasi Identifier [QI] attributes. The attributes that hold private data of an individual and should not be disclosed are called as sensitive attributes.

K-anonymity[7] is one widely known scheme for achieving privacy in data publication. In k-anonymized data, privacy is achieved through generalization and suppression. Suppression of directly identifiable attributes, like name, aadhar card number is done by not giving for publish. 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 is a process in which we replace a tuple with a more generalized tuple, which is identical from several other tuples in the equivalence set as in table 2. This is also called as anonymization [1]. But several problems are identified with k-anonymity [2, 10].

Table I: Microdata

Age	Gender	Postal Code	Disease	Name
20	M	440011	Viral Fever	A
60	M	447777	Typhoid	B
28	M	448796	Typhoid	C
40	F	447602	Typhoid	D
31	F	440209	Malaria	E
52	M	440001	Diabetes	F

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

Age	Gender	Postal Code	Disease	Name
20-30	M	44****	Viral Fever	A
60-70	M	44****	Typhoid	B
20-30	M	44****	Typhoid	C
40-50	F	44****	Typhoid	D
30-50	F	44****	Malaria	E
50-60	M	44****	Diabetes	F

A k- anonymous data may permit an adversary to find out sensitive information of an person with 100% confidence. There may be some information loss from the data. And it does not take into account personalized anonymity requirements[10].

K-anonymity data set allows an adversary to find out the value of sensitive attributes, when there are some diversity in the sensitive attributes [6, 11]. To overcome this, another approach called as l-diversity was proposed. *l-diversity* provides privacy even when the data publisher does not know what kind of knowledge is possessed by the attacker. It ensure that, all tuples that share the common 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 little violate probability [12,13]. Anatomy [14] is another l-diversity specific method. Though it does not violate the l-diversity property, it confirm that a particular person 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 approach which allows each sensitive attribute in a data set to have a privacy constraint [10, 15]. However, the computational effort is too high as compared to the generalization approach. Personalized privacy preservation uses a decision tree based approach. Greedy algorithm is used for the computational and it is not optimal, so does not achieve minimal loss.

P-sensitive k-anonymity is nearly similar to l-diversity [5]. Extended p-sensitive k-anonymity is a scheme that extends p-sensitive k-anonymity characteristic that which is similar to the personalized privacy method, where in the privacy is offered at different hierarchical levels in the classification for the sensitive attribute [4]. Another scheme in [16] assumes hierarchy in each QI attribute, and that all partition in a general domain should be at the same level of hierarchy.

Contributions of the paper: We studied all PPDM techniques with increase computational effort, though they increase privacy. Some of them techniques account for too much of information loss, while achieving privacy. We address the data privacy problem using Fuzzy Inference System a new approach, a total paradigm shift and a new perspective of looking at privacy problem in data publishing. The domain simplification based solution completely disassociates the sensitive values with the identifying attributes. Our practically feasible solution, allows personalized privacy preservation, and is useful for both numerical and categorical attributes.

Outline of the paper: Section 2 gives a brief overview of fuzzy logic. Section 3 contains a description of the fuzzy inference system 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 future scope.

II. BACKGROUND

A. Fuzzy sets overview

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

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

The domain of the membership function, which is the domain of concern and from which elements of the set are strained, 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 idea central to fuzzy systems is that fact 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 complete false and 1.0 representing complete truth.

B. Fuzzy based Privacy preserving for numerical attributes:

Assume, the data in table 1 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 $L=\{l_1,l_2,l_3...l_n\}$ are the linguistic values, is the linguistic variable for the attribute and 'n' is the number of linguistic values, '*i*' refers to the numerical attributes of *T* which are sensitive.

Rule 1: If $L= \{ l_1,l_2,l_3....l_n\}$,

(2)

Then, , F(

(3)

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-a2}*, $\{a_1\text{-}a_3\}$, $\{a_{(i-1)}\text{-}a_{(i+1)}\}, \ldots, \{a_{(k-1)}\text{-}max\}$.

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

$$f_1(x) = \\ = (x - a_2)/(min - a_2) \text{ if } x < a_2$$

(4)

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

$$f_i(x) = 0 \text{ if } x \leq a_{(i-1)} \\ = (x - a_{(i-1)})/(a_i - a_{(i-1)}) \text{ if } a_{(i-1)} < x < x$$

=

(5)

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

$$f_k(x) = 0 \text{ if } x \leq a_{(k-1)} \\ = (x - a_{(k-1)})/max - a_{(k-1)}) \text{ if } x > a_{(k-1)}$$

(6)

III. BACKGROUND OF GAUSSIAN FUZZY SETS,FUSSY INFERENCE SYSTEM, AND THE SUFFICIENT CONDITION:

A. Gaussian Fuzzy sets:

A Gaussian MF (as in Fig 1) can be represents as

$$\mu_G(x; c, \qquad (7)$$

where *c* is centre of fuzzy set, and parameterizes the width of the fuzzy set. A of a fuzzy set is a crisp set that contain all the elements of the universe set X that have a membership grade equal to or greater than ,where Based on the of a fuzzy set, the width of fuzzy set, is ac or cb, which can be determine by

$$W_\alpha = ac = cb = \sqrt[2]{}$$

(8)

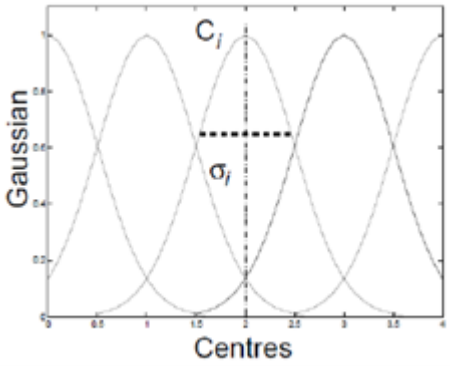

Fig 1: Fuzzy membership function

B. Fuzzy Inference System

The fuzzy production rules for an n-input FIS model, where n>0, can be represented as follows.

$R^{j_1,j_2,\dots j_n}$:

If

$$\left(x_1 \text{ is } A_1^{j_1}\right) AND \left(x_2 \text{ is } A_2^{j_2}\right) \dots AND \left(x_n \text{is } A_n^{j_n}\right), THEN \left(y \text{ is } B^{j_1,j_2,\dots j_n}\right) 1 \le j_i \le M_i.$$

(9)

The AND operator in the rule antecedent part is the production function. For the domain, its MFs are

$$\mu_i^1(x_i), \mu_i^2(x_i.$$

The upper and lower limits for the universe of discourse of respectively. The output is obtained by using the weighted average of a representative value, with respect to its compatibility grade, as in

$$y = \frac{\sum_{j_n=1}^{j_n=M_n} \dots \sum_{j_2=2}^{j_2=M_2} \sum_{j_1=1}^{j_1=M_1} \left(\mu_1^{j_1}(x_1) x \left(\mu_2^{j_2}(x_2) x \dots x \left(\mu_n^{j_n}(x_n) \right) \right) \right)}{\sum_{j_n=1}^{j_n=M_n} \dots \sum_{j_2=2}^{j_2=M_2} \sum_{j_1=1}^{j_1=M_1} \left(\mu_1^{j_1}(x_1) x \left(\mu_2^{j_2}(x_2) x \dots x \left(\mu_n^{j_n}(x_n) \right) \right) \right)}$$

(10)

where is the representative value of, i.e.

$$b^{j_1,j_2,\dots j_n} =$$

This value represents the overall location of the MF. It can be obtained via fuzzification, or be represented by the point whereby the membership value is 1.

C. Scientific Condition:

We first derivative of FIS model, as in (3) returns the weighted series. The sufficient condition assume that all the component in the weighted series are always Very Low, Low, Medium, High, Very High,

IV. EXPERIMENTAL RESULTS:

The experiments were carried out on Cardiology data set downloaded from UCI Machine learning repository [17]. The taxonomy trees were constructed based on information is 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. Clustering can be takes place on original, fuzzified and defuzzified data by k-mean algorithm. 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 generalization.

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.

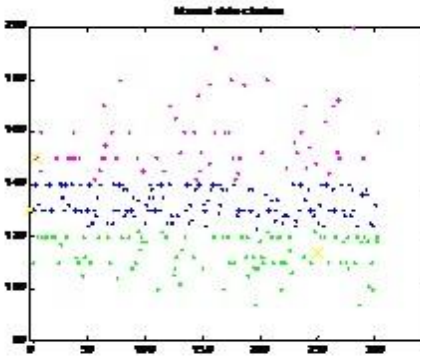

Fig 2. Clustering with original data set

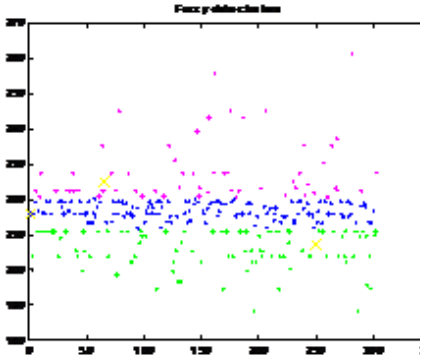

Fig 3. clustering with fuzzy data set

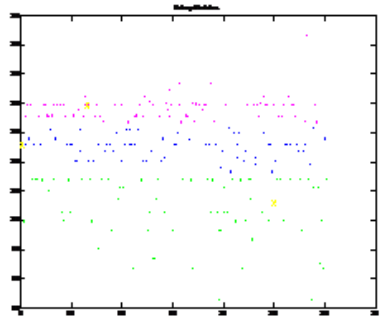

Fig 4. clustering with defuzzified data

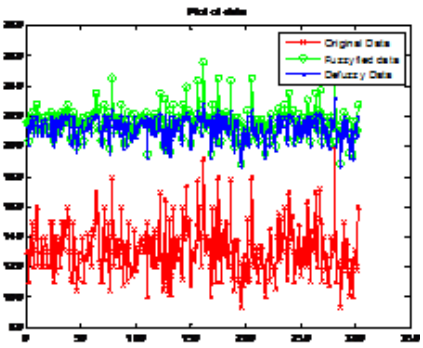

Fig 5. data plot

Fig 4: Data plot

V. CONCLUSION AND FUTURE SCOPE

Most data anonymization techniques are taken from various fields like data mining, cryptography and information hiding. K-Anonymity is a popular approach where data is transformed to equivalence classes and each class has a set of k- records indistinguishable from each other. But it amplifies computational effort to infeasible levels, though they boost up privacy. Some techniques results in information loss even its privacy achievement is great. This paper addresses the problem of Privacy Preserving in Data Mining by transforming the attributes to fuzzy attributes. Due to fuzzification, exact value cannot be predicted thus maintaining individual privacy. The dataset is classified using 10 fold cross validation of the original and fuzzy anonymized dataset. The experimental results demonstrate the effectiveness of the fuzzy anonymization.In future, we will implement this on cloud computing.

References

[1] G. Eason, B. Noble, and I.N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529-551, April 1955. (*references*)

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.

[3] I.S. Jacobs and C.P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350.

[4] K. Elissa, “Title of paper if known,” unpublished.

[5] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740-741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

[8] Wu Y.-H., Chiang C.-M., Chen A. L. P.: Hiding Sensitive Association Rules with Limited Side Effects. *IEEE Transactions on Knowledge and Data Engineering*, 19(1), 2007.

[9] Atallah, M., Elmagarmid, A., Ibrahim, M., Bertino, E., Verykios, V.: Disclosure limitation of sensitive rules, *Workshop on Knowledge and Data Engineering Exchange*, 1999.

[10] Ciriani V., De Capitiani di Vimercati S., Foresti S., Samarati P.:
k-Anonymity. *Security in Decentralized Data Management*, ed. Jajodia S., Yu T., Springer, 2006.

[11] Gedik B., Liu L.: A customizable *k*-anonymity model for protecting location privacy, *ICDCS Conference*, 2005.

[12] Li N., Li T., Venkatasubramanian S: *t*-Closeness: Privacy beyond *k*-anonymity and *l*-diversity. *ICDE Conference*, 2007.

[13] Machanavajjhala A., Gehrke J., Kifer D., and Venkitasubramaniam M.: *l* Diversity: Privacy Beyond k-Anonymity. *ICDE*, 2006.
[14] Xiao X., Tao Y. Anatomy: Simple and Effective Privacy Preservation. *VLDB Conference*, pp. 139-150, 2006.
[15] Xiao X., Tao Y.: *m*-Invariance: Towards Privacy-preserving Republication of Dynamic Data Sets. *SIGMOD Conference*, 2007.
[16] Atallah, M., Elmagarmid, A., Ibrahim, M., Bertino, E., Verykios, V.: Disclosure limitation of sensitive rules, *Workshop on Knowledge and Data Engineering Exchange*, 1999.
[17] https://archive.ics.uci.edu/ml/datasets.html