A Fuzzy Approach for Privacy Preserving in Data Mining

M.Sridhar
Associate Professor
R.V.R & J.C College of Engineering
Guntur, India.

B.Raveendra Babu, PhD.
Professor
VNR VJIET
Hyderabad, India.

ABSTRACT

Advances in hardware technology have increased storage and recording capabilities regarding individual's personal data. Privacy preserving of data has to ensure that individual data publishing will refrain from disclosing sensitive data. Data is anonymized to address the data misuse concerns. Recent techniques have highlighted data mining in ways to ensure privacy. Most 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 there were many problems with this approach and remedies like l-diversity and t-closeness were proposed to overcome them. 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, and also better accuracy of mining results were achieved.

General Terms

Privacy Preserving Data Mining (PPDM)

Keywords

Privacy Preserving Data Mining (PPDM), K-Anonymity, l-Diversity, Fuzzy Logic, Adult Dataset

1. INTRODUCTION

Advancement in hardware technology has led to vast storage capability giving rise to recording of personal data about individuals in various fields. In turn, this paved misuse of personal data for different actions. Data mining is now viewed as a threat to privacy of data. This augments the concern about the privacy of the underlying data. To preserve privacy, a number of techniques have been proposed for modifying or transforming the data. The techniques for performing privacy-preserving data mining (PPDM) are done through cryptography, data mining and information hiding [1]. Apart from these communities PPDM is an independent field with broader perspectives.

With the advent of global net database sharing is a common task. In data publishing tasks for statistical use of a patient data and voters list the identifying attributes are combined and suppressed (Quasi Identifier [QI] attributes) and sensitive information are revealed (sensitive attributes) [2, 3]. Transformation on the data is done to perform the PPDM that in turn reduce the representation granularity rather than privacy. This granularity reduction results in loss of effectiveness in data management/ mining algorithms. This is the cause of information loss and privacy.

The main objectives in PPDM are,

Privacy-Preserving Data Publishing: Here, techniques include methods such as randomization, k-anonymity, l-diversity, t-closeness & fuzzy logic for privacy [4]. The other issues are perturbed data usage in conjunction with classical data mining methods (association rule mining). The utility-based methods are used to determine the privacy preservation of data.

To preserve privacy changes Data Mining Applications results: The association rule or classification rule mining can compromise the data privacy.

Auditing Query: Either modifying / restricting the query results are performed.

Cryptographic Methods for Distributed Privacy: A variety of cryptographic protocols are used to communicate among the data distributed across multiple sites. And a secure function computation is done devoid of revealing sensitive information.

Theoretical Challenges in High Dimensionality: In case of real data sets (high dimensional) the PP process is hard. The optimal k-anonymization is NP-hard.

To ensure data privacy following techniques are commonly used:

Randomization method: In this, noise is added to the data that in turn masks the records attribute values. The added noise is huge so there is loss in record value. To derive aggregate distributions from the perturbed records some techniques are needed. Two kinds of perturbation are possible with the randomization method:

Additive & Multiplicative Perturbation: In additive perturbation, randomized noise is added to the data records. The overall data distributions can be recovered from the randomized records. Data mining and management algorithms re designed to work with these data distributions. In multiplicative perturbation, the random projection or random rotation techniques are used in order to perturb the records [5].

K-Anonymity: The data is transformed to equivalence classes where each class have a set of K- records that differs from others [6]. To reduce the granularity representation of the pseudo-identifiers techniques Viz., generalization & suppression are used. The attribute values are generalized to a range so as to reduce the granularity (for example, date of birth generalized as year of birth). It reduces identification risk. The value of the attribute is removed completely to reduce the identification risk with public records (suppression). The k-anonymity is a good technique because of its simplicity in definition and also many algorithms are available to process the anonymization [7, 8, 9]. But it faces many attacks where background knowledge is known to the attacker. Some kinds of such attacks are as follows:

Homogeneity Attack: Though the data is k-anonymized, the sensitive attribute values are same for group of k- tuples and can be identified [10].

Background Knowledge Attack: Here quasi-identifier attributes with the sensitive attribute narrows down available values of the sensitive field.

k-anonymity is better preventer in record identification rather in preventing inference of the sensitive values of the attributes of that record. Therefore, the technique of l-diversity was proposed which not only maintains the minimum group size of k, but also focuses on maintaining the diversity of the sensitive attributes.

l-diversity: In social networks, vertices are partitioned into equivalence groups in all the vertices groups. It yields privacy although the data publisher has no knowledge possessed by the adversary [11]. It insists all records that share the similar values of quasi identifiers to have l- diverse values for their sensitive attributes. It's too prone to adversary attacks but it ensures a low breach probability [12, 13]. Anatomy is the other l-diversity method. It does not violate the l-diversity property but it confirms that a prompt individual is involved in the data. t-closeness is the other method that possess table-wise Sensitive Attribute values distribution and it is repetitive among every anonymised group [14].

The numerous problems with K-anonymity is identified in the literature and have proposed techniques to counter them or avoid them. l-diversity and t-closeness are such techniques to name a few. 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 anonymization is achieved and, it is evaluated for classification accuracy using data mining algorithms. The paper is organized as follows: Section 2 reviews some related works in the literature, section 3 details the materials and methods, section 4 gives the results and section 5 concludes the paper.

2. RELATED WORKS

Shang, et al., [15] proposed a novel scheme for selective content distribution encoded as documents, preserving user privacy based on an efficient and novel group key management scheme. The proposed approach is based on access control policies that specify which user can access either documents or sub-documents. On this basis, a broadcast document is divided into multiple subdocuments. Each subdocument is encrypted with a different key. Conforming to modern attribute-based access control, policies are specifically against user identity attributes. But this approach preserves privacy such that users get access to specific documents, or subdocument, based on policies without needing to provide information about identity attributes to the publisher. Under this approach, the document publisher does not learn identity values of users, and also does not know what policy conditions are verified by users which in turn prevents inferences about identity attributes values being prevented. Also, the proposed key management scheme on which the broadcasting approach is based is efficient as decryption keys need not be sent to users together with the encrypted document. Users can reconstruct keys to decrypt the authorized document portions of a document based on subscription information from the document publisher. Another advantage is that the scheme efficiently handles user's new and revoked subscriptions.

Wang, et al., [16] proposed a new model, Unique Distinct l-SR diversity based on private information sensitivity. Also, two performance measures were presented: Entropy Metric and Variance Metric, as to how sensitive information could be inferred from an equivalence class. Unique Distinct l-SR diversity was achieved through implementation of l-SR diversity algorithm. The latter was tested on one benchmark and three synthetic data sets, and compared with other l-diversity algorithms. The results revealed that the proposed algorithm functioned better on minimizing sensitive information inference reaching comparable generalization data quality in contrast to other data publishing algorithms.

Kumari, et al., [17] suggested a holistic approach to achieve maximum privacy without information loss and minimum overheads. Studies showed that l-diversity and t-closeness techniques increased computational effort to infeasible levels, while increasing privacy. A few techniques account for maximum information loss when achieving privacy. The proposed method addresses this problem using fuzzy set approach which is a total paradigm shift and a new way of looking at data publishing privacy problem. This method allows personalized privacy preservation being useful for both numerical and categorical attributes and only necessary tuples are transformed.

Bayardo, et al., [18] proposed and evaluated k-anonymization optimization algorithm for powerful de-identification procedure. A k-anonymized dataset record is indistinguishable from others. Simple restrictions of optimized k-anonymity are NP-hard, resulting in major computational challenges. A new approach exploring possible anonymizations taming problem combinatorics is presented. Data management strategies are developed to reduce reliance on expensive operations like sorting. Real census data experiments revealed the proposed algorithm could locate optimal k-anonymizations under two representative cost measures and a wide range of k. It also revealed that the algorithm could provide good anonymizations under circumstances where input data/input parameters preclude locating an optimal solution within reasonable time.

3. MATERIALS AND METHODS

3.1 Adult Dataset

UCI Machine Learning Repository [19] provides the 'Adult' dataset used for evaluation. It contains 48842 instances, including categorical and integer attributes from 1994 Census information. It has about 32,000 rows with 4 numerical columns, the column and ranges including age {17 – 90}, fnlwgt {10000 – 1500000}, hrsweek {1 – 100} and edunum {1 – 16}. The age column and native country are anonymized using k-anonymization. Table 1 shows the original attributes of the Adult dataset.

Table 1: Attributes of the Adult dataset

Age	Native-country	Class
39	United-States	<=50K
50	United-States	<=50K
38	United-States	<=50K
53	United-States	<=50K
28	Cuba	<=50K
37	United-States	<=50K
49	Jamaica	<=50K
52	United-States	>50K
31	United-States	>50K
42	United-States	>50K

3.2 Fuzzy Logic

All the above techniques amplify computational effort to almost infeasible levels, though they boost up privacy. Some techniques results in information loss even its privacy achievement is great. So, maximum privacy achievement devoid of information loss and minimum overheads is required. To overcome this, the data privacy is achieved using fuzzy set. Fuzzy logic useful in both the cases numerical and definite attributes. Fuzziness is a means to symbolize improbable, prospect and approximation. These sets are an annex of classical set theory and are applied in fuzzy logic. In classical set theory, the associated elements in relation to a set is assessed in binary terms with respect to a crisp condition (i.e., an element belongs to / does not belongs) to the set. In contradiction, fuzzy set theory allows the slow assessment of the associated elements in relation to a set. This is defined by a membership function: $\mu \rightarrow [0, 1]$.

Following Figures 1 to 3 shows the membership functions used to approximate the data to achieve privacy. Figure 1 gives the values for the input variable being attribute.

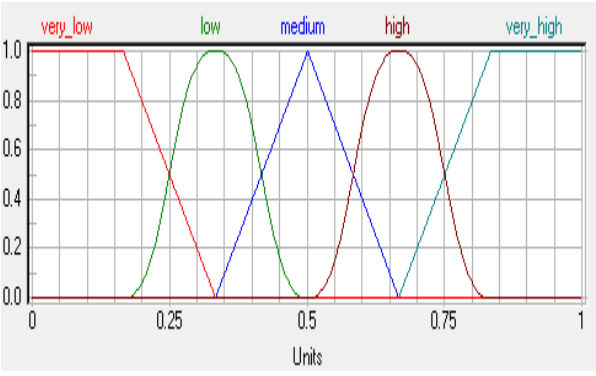

Figure 1: Membership function for Input variable: Attribute

Figure 2 gives the class count values, class count being the number of classes of the same type for the given input attribute.

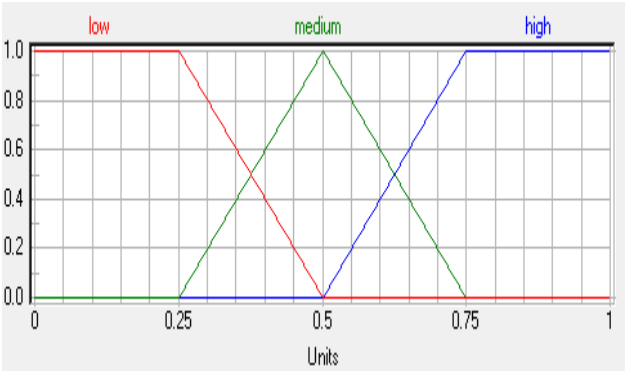

Figure 2: Membership function for class count

Figure 3 shows the new membership function of the new attribute created**.**

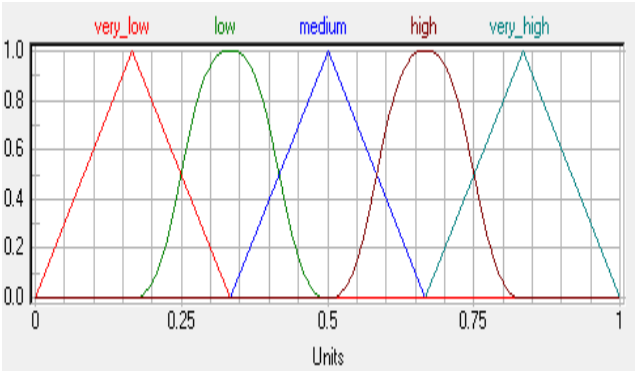

Figure 3: Membership function of the new attribute created

The rule blocks are based on the control strategy of a fuzzy logic system. Each rule block contains rules restricted to the same context. A context is obtained by the same input and output variables of the rules. The 'if' part of the rules describes the context and the 'then' part describes the response of the fuzzy system. The degree of support (DoS) is used to weigh each rule according to its importance.

The processing of the rules starts with calculating the 'if' part. The operator type of the rule block determines which method is used. The operator types MIN-MAX is used. Some of the rules generated are shown in Table 2.

Table 2: Rule Generated

IF		THEN	
Attr	**ClassCount**	**DoS**	**NewAttr**
very_low	low	0.88	very_low
very_low	low	0.94	Low
very_low	low	0.06	Medium
very_low	low	0.05	High
low	high	0.09	Low
low	high	0.91	Medium
low	high	0.27	High
low	high	0.40	very_high
medium	low	0.86	very_low

IF		THEN	
medium	low	0.99	Low
medium	low	0.56	Medium
medium	low	0.80	High
high	low	0.69	very_low
high	low	0.98	Low
high	low	0.84	Medium
high	low	0.64	High
high	low	0.72	very_high
very_high	low	0.33	very_low
very_high	low	0.32	Low
very_high	low	0.87	Medium
very_high	low	0.59	High

4. RESULTS

The classification accuracy for the original attributes without anonymization and after fuzzy anonymization is evaluated from k nearest neighbor, J48 and Bagging. The dataset is classified using 10 fold cross validation of the original and fuzzy anonymized dataset. The classification accuracy obtained is tabulated in Table 3 and is shown in Figure 4. It is seen from the figures that the classification accuracy in fact increases by 0.3% to 0.9% on Fuzzy anonymization of the dataset. All the possible results have been discussed above.

Table 3: Classification Accuracy

Technique used	Classification Accuracy
kNN without anonymization	79.32%
J48 without anonymization	85.32%
Bagging without anonymization	85.01%
kNN with fuzzy anonymization	79.56%
J48 with fuzzy anonymization	86.08%
Bagging with fuzzy anonymization	85.99%

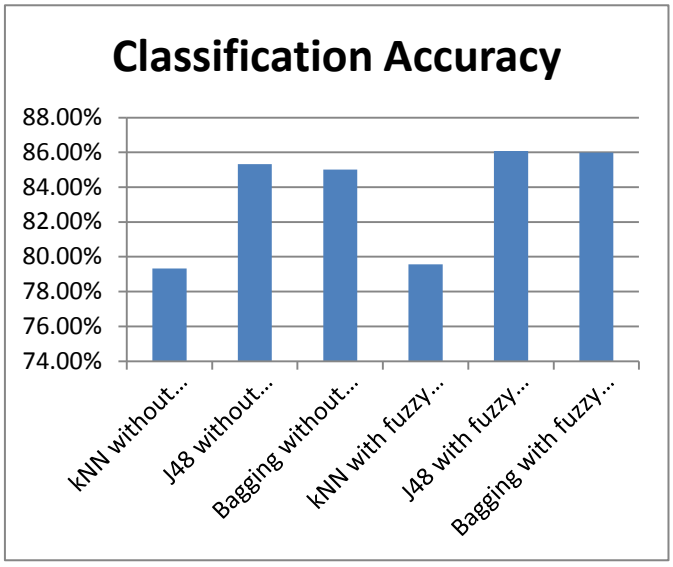

Figure 4: Classification Accuracy obtained

5. CONCLUSION

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. It is seen that the classification accuracy increases by 0.3% to 0.9% on Fuzzy anonymization of the dataset.

6. ACKNOWLEDGMENTS

Profusely thank our most generous managements for allowing to make use of the infrastructure on the campus and the support extended for the fulfilment of the research paper.

7. REFERENCES

[1] Agrawal R., Srikant R. Privacy-Preserving Data Mining. Proceedings of the ACM SIGMOD Conference, 2000.

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

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

[4] Martin, D., Kifer, D., Machanavajjhala, A., Gehrke, J., And Halpern, J. 2006. Worst-case background knowledge in privacy. Tech. rep., Cornell University.

[5] Muralidhar, K., Batrah, D., & Kirs, P. J. (1995). Accessibility, security, and accuracy in statistical databases: The case for the multiplicative fixed data perturbation approach. Management Science, 41(9), 1549-1564.

[6] P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report, CMU, SRI, 1998.

[7] Bayardo, R. J. And Agrawal, R. 2005. Data privacy through optimal k-anonymization. In Proceedings of the International Conference on Data Engineering (ICDE'05).

[8] Lefevre, K., Dewitt, D., And Ramakrishnan, R. 2005. Incognito: Efficient fulldomain k-anonymity. In SIGMOD.

[9] Zhong, S., Yang, Z., And Wright, R. N. 2005. Privacy-enhancing k-anonymization of customer data. In Proceedings of the International Conference on Principles of Data Systems (PODS).

[10] Ohrn, A. And Ohno-Machado, L. 1999. Using boolean reasoning to anonymize databases. A. I. Medicine 15, 3, 235–254.

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

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

[13] 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.

[14] 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.

[15] N. Shang, F. Paci, M. Nabeel, and E. Bertino. A privacy-preserving approach to policy-based content dissemination. Technical Report 2009-14, Purdue University Center for Education and Research in Information Assurance and Security (CERIAS), 2009.

[16] Wang, Y., Cui, Y., Geng, L., and Liu, H. A newperspective of privacy protection: Unique distinct l-SRdiversity. In Proceedings of PST. 2010, 110-117.

[17] V. Valli Kumari, S.Srinivasa Rao, KVSVN Raju, KV Ramana and BVS Avadhani, Fuzzy based approach for privacy preserving publication of data, IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.1, January 2008, pp:115-121.

[18] Bayardo R. J., Agrawal R.: Data Privacy through Optimal k-Anonymization. Proceedings of the ICDE Conference, pp. 217–228, 2005.

[19] Frank, A. & Asuncion, A. (2010). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.