A Hybrid Data Perturbation Approach To Preserve Privacy

Thanveer Jahan, Dr.G.Narsimha, Dr.C.V.Guru Rao

Abstract— The challenging approach in data mining applications is about balanced privacy and data quality, a pair of contradictive factors. Various techniques were focused on privacy preserving scheme using data perturbation. In this paper a sensitive attribute is analyzed using a hybrid data transformation approach which is a combination of fuzzy logic and non-negative factorization method. The proposed method is used to protect sensitive information of a confidential attributes without losing accuracy in results. The accuracy is measured using data mining techniques such as classification and clustering. Different classifiers and K-means clustering method are used and compared on the dataset.

Index Terms— Privacy, Data Perturbation, Fuzzy Logic, Non-negative Factorization, Classification, Clustering.

———————————— ◆ ————————————

1 INTRODUCTION

The enormous increase of sharing individual's data has lead to privacy concern during data publishing. Privacy preserving data mining is widely used under these privacy constraints. Many approaches have been adopted in preserving privacy such as data perturbation or data distortion. The major task of data perturbation is balancing the privacy and quality of data, which are considered as pair of contradictive factors. Data perturbation is classified into a twofold approach, probability distribution and value distortion [3]. The probability distribution approaches the data with another sample from the same distribution or by the distribution itself. The value distortion approach perturbs data elements or attributes directly by either additive noise [7], multiplicative noise or some other randomization procedures,. There are different methods in data perturbation can be categorized as noise additive, matrix decomposition methods such as singular valued decomposition ,sparsified singular valued decomposition, Non negative matrix factorization method and fuzzy logic[2].

Fuzzy logic is used to solve imprecise problem with better solutions [4]. To decrease complexity of a system we need to understand a system well. Fuzzy sets are the extension of generic set theory and have a different approach to preserve privacy. The main characteristics of fuzzy sets, contrasting with crisp set are the progressive transition from one set to another. This natural characteristic of fuzzy logic provides automatic mechanisms to deal with imprecision and uncertainty, which are inherent to real world knowledge. The assessment of data set can be done using fuzzy membership in fuzzy sets [5].A fuzzy set is a pair (A, μ_A) where A is a set and $\mu_A : A \rightarrow [0, 1]$. For all x ϵ A, $\mu_A(x)$ is called the grade of membership of x. Fuzzy sets model the linguistic variables of a given domain, both in terms of shape(Triangular, Trapezoidal, S-shaped, e.t.c) and partitioning of the attributes domains [6].Each linguistic term can be represented as a fuzzy set having its own membership function[8]. Fuzzy membership functions: S-shaped membership function is represented as

$$f(x;a,b)=\left\{\begin{array}{cc} 0, & x\le a \\ 2\left(\dfrac{x-a}{b-a}\right)^2, & a\le x\le \dfrac{a+b}{2} \\ 1-2\left(\dfrac{x-b}{b-a}\right)^2, & \dfrac{a+b}{2}\le x\le b \\ 1, & x\ge b \end{array}\right\}$$

Z-shaped membership function is represented as:

$$f(x;a,b)=\left\{\begin{array}{cc} 1, & x\le a \\ 1-2\left(\dfrac{x-a}{b-a}\right)^2, & a\le x\le \dfrac{a+b}{2} \\ 2\left(\dfrac{x-b}{b-a}\right)^2, & \dfrac{a+b}{2}\le x\le b \\ 0, & x\ge b \end{array}\right\}$$

Triangular membership function is represented as:

$$f(x;a,b,c)=\left\{\begin{array}{cc} 0, & x\le a \\ \dfrac{x-a}{b-a}, & a\le x\le b \\ \dfrac{c-x}{c-b}, & b\le x\le c \\ 0, & c\le x \end{array}\right\}$$

- *Thanveer Jahan is currently pursuingPh.D in computer science and engineering in JNTU,Hyd,India , PH-+919550666795. E-mail :thanvijahan@mail.com*
- *Dr.G.Narasimha ,Associate Professor,Jntu,Jagityal,Karimnagar.*
- *Dr.C.V.Guru Rao,Principal ,S.R Engg College,Warangal.*

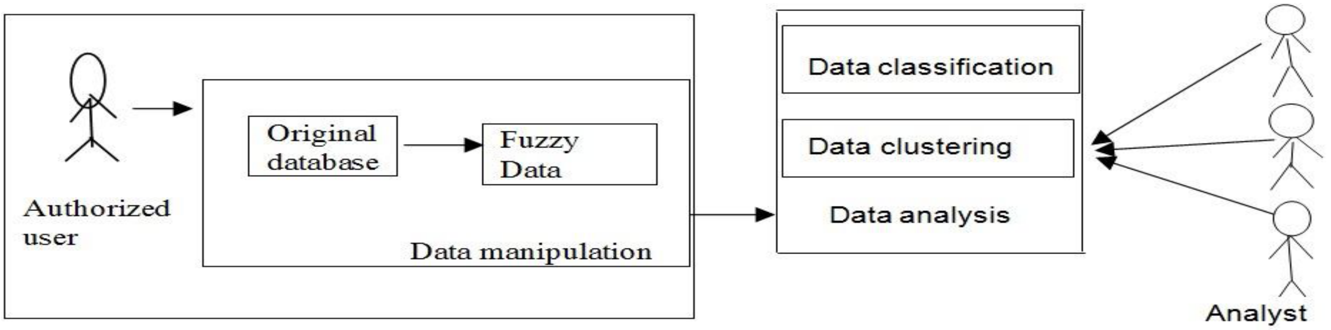


Figure 1

Non –Negative matrix factorization method:It is a recent method used in data analysis tasks to find suitable and useful representation of task, which makes latent structure reducing the dimensionality of the data .The computational method for a non-negative data matrix is represented as V.
NMF is a linear, non-negative approximate data representation. It assumes data that our data consists of T measurements of N non-negative scalar variables, denoting (N-dimensional) measurement vectors V^t (t=1……T). It finds an approximate factorization V≈WH into non-negative factors W and H.
NMF can control in approximating high dimensional data in a lower dimensional space**.**

2 Proposed Work

In this we have used a hybrid method fuzzy logic and nonnegative matrix factorization method on the data set to preserve privacy [9]. The proposed method for data perturbation is shown in the figure 1.The original data having numerical attributes is perturbed with fuzzy data using the fuzzy membership function [1]. The dimensionality of fuzzy data is reduced using Non-Negative matrix factorization method. These hybrid transformations are applied on original data and analyzed by data mining techniques.

2.1 Algorithm for Hybrid Data Transformation:

1. Data owner owns an original data (D).
2. Data having confidential numerical attribute is perturbed using(S,T,Z membership functions).
3. An S-membership function is taken as Fuzzy data (D′).
4. Fuzzy data (D) is then used by a Non-Negative Factorization matrix Transformation to reduce the dimensionality (D″).

$$[W\ H] = (D'_{P*Q}, K) \text{ where } k=\min(P,Q),$$

$$D'' = W*H$$

5. The hybrid Transformed data (D″) is published to the data analyst for analysis.
6. Analysis uses Data mining techniques: classification and clustering are applied on original data (D), fuzzy data(D′), hybrid transformed data(D″).

3 Experimental Results

We have downloaded data set from UCI machine Learning Repository Fertility datasets. These data sets are perturbed before publishing to preserve privacy. Initially an authorized owner owns the original data and does not release unless it is distorted. The experiments are carried on MATLAB and Tanagra tool for data analysis.

The original data having confidential attributes are perturbed using fuzzy membership functions. Fuzzy membership functions (S, Z, T) are applied on original datasets. Comparisons are done on the datasets. Fuzzy data (D′) used by membership functions(S and Z) accuracy is compared with the original data set (D) and found to be same as shown in Figure 2.

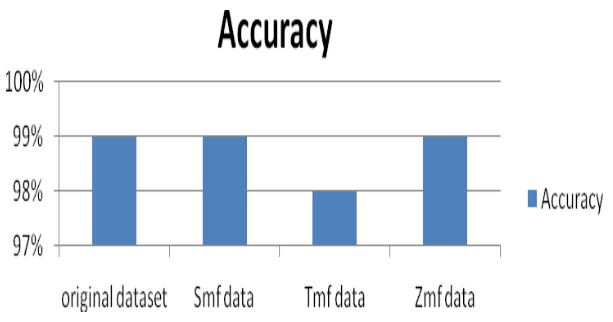
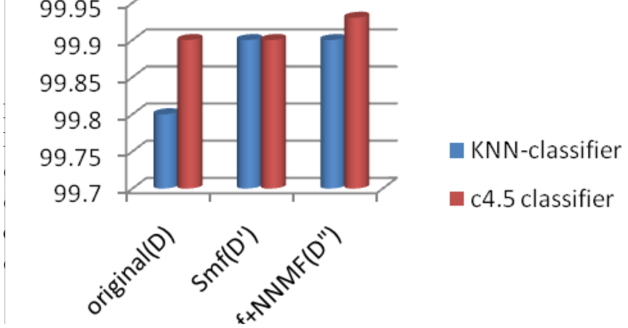


Figure 2

Fuzzy data (D′) is transformed using Non-Negative matrix Factorization method to reduce the dimensionality of data. The hybrid transformed data (D″) is analyzed by data mining techniques. Accuracy is measured using different classifiers on original data ,Fuzzy data (D)and hybrid transformed data(D″).Classifiers are KNN and c4.5 are used .Hence proved that the KNN classifier is classifying better than c4.5 on the hybrid transformed data(D″) than Fuzzy data and original



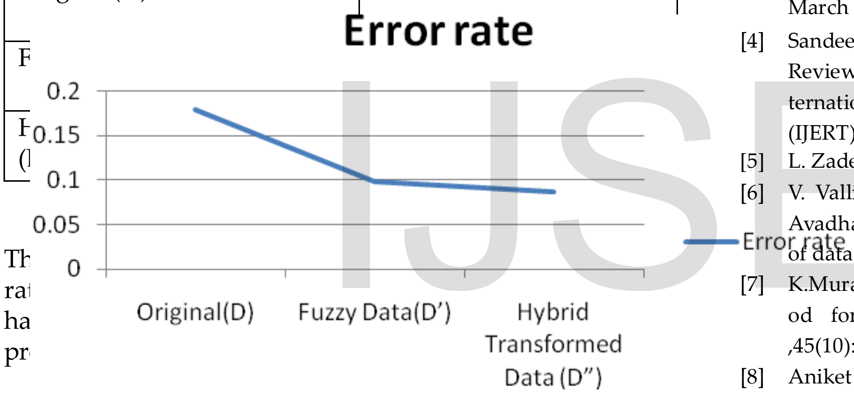
ginal
).The
mis-

$$M_{Ei} = 1/N \times \sum_{j=1}^{k} (|Cluster_j(T_i)| - |Cluster_j(T_1)|)$$

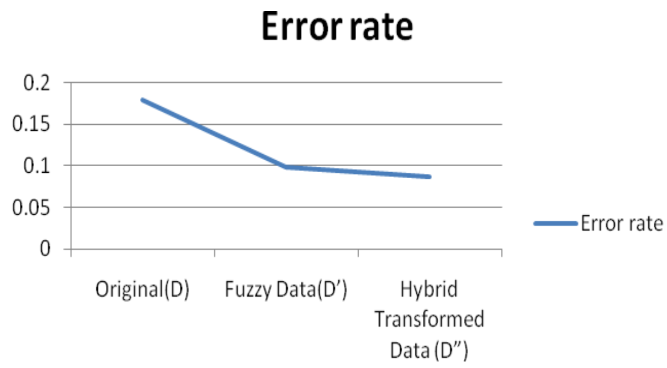
The formula N means the number of points in the original dataset .K is the number of clustering (|Cluster (T)|) means the number of data points of the i^{th} cluster of the original data set ($|Cluster_i(T^i)|$) means the number of data points of the i^{th} cluster of the transformed data set[10]. The Experimental results are shown in the Table 1.

Table I
MISCLASSIFICATION ERROR RATES

Datasets	Error rate
Original(D)	0.18
F	
H (	



Th
ra
ha
pr



4 Conclusion

The major concern of preserving privacy is the accurate data. In this paper the experiments have proved that hybrid transformations are protecting confidential information present in the data sets .The hybrid transformed data is thus preserving privacy during publishing. The accuracy of the original data and transformed are measured by classification and clustering. Further work will be proceeded by using a multivariate datasets and apply data perturbation on multi columns of the datasets.

References

[1] M.Naga Lakshmi and K.Sandhya Rani "Privacy Preserving Clustering Based on Fuzzy Data Transformation Methods", in International Journal of Advanced Research in Computer Science and Software Engineering ISSN: 2277 128X, Volume 3, Issue 8, August 2013,Pg No 1027-1033

[2] S.R. M. Oliveira, O.R. Zaiyane (2003), "Privacy Preserving Clustering by Data Transformation", in proceedings of 18th Brazilian Conference on Databases.

[3] Samir Patel, Gargi Shah, Aniket Patel, Techniques of Data Perturbation for privacy Preserving Data Mining" , International Journal of Advent Research in Computer & Electronics (IJARCE) Vol.1, No.2, March 2014 ,Pg No 5-10.

[4] Sandeep Kumar Singh, Mr.Ganesh Wayal, Mr.Niresh Sharma "A Review: Data Mining with Fuzzy Association Rule Mining" in International Journal of Engineering Research & Technology (IJERT)Vol. 1 Issue 5, July – 2012,ISSN: 2278-0181.

[5] L. Zadeh, "Fuzzy sets'. Inf. Control. 8, 338–353 (1965).

[6] V. Vallikumari, S. Srinivasa Rao, KVSVN. Raju, KV. Ramana, BVS. Avadhani, "Fuzzy based approach for privacy preserving publication of data. Int. J. Comput. Sci. Netw. Secur. 8(1),(2008).

[7] K.Muralidhar,R.Sarathi, "A General additive data perturbation method for data base security" ,journal of Management Science ,45(10):1399-1415,2002.

[8] Aniket Patel, HirvaDivecha "A Study of Data Perturbation Techniques For Privacy Preserving Data Mining" ,in International Journal of software & hardware research in Engineering ISSN No:2347-4890 vol 2 ,Issue 2 ,Feb2014,Pg No 42-46.

[9] S.R. M. Oliveira, O.R. Zaiyane (2003), "Privacy Preserving Clustering by Data Transformation", in proceedings of 18th Brazilian Conference on Databases.

[10] B. Karthikeyan, G. Manikandan, V. Vaithiyanathan," A fuzzy based approach for privacy preserving clustering", J. Theor. Appl. Inf. Technol. 32(2), 118–122 (2011).