A MODEL FOR PRESERVING PRIVACY OF SENSITIVE DATA

Ms Shalini Lamba #1 , Dr S. Qamar Abbas #2
# 1 Research Scholar, Shri Venkateshwara University,Meerut, U.P., India
#2 Visiting Guide, Shri Venkateshwara University,Meerut, U.P., India

***Abstract*— Advancement in the field of Information technology has resulted in tremendous growth in data collection and extraction of unknown patters from this huge pool of data which has become the primary objective of any data mining algorithm. Besides being so effective, some sensitive information is also revealed by mining algorithm. The extracted knowledge is highly confidential and it needs refinement before giving to data mining researchers and the public in order to concentrate on privacy concerns. The data mining techniques have been developed by the researchers to be applied on data bases without violating the privacy of individuals. Over the last decade, many techniques for privacy preserving data mining have come up. In this paper we propose a new approach to preserve sensitive information using fuzzy logic. Clustering is done on the original data set, and then we add noise to the numeric data using a fuzzy membership function that results in distorted data. Set of Clusters generated using the fuzzified data is also equivalent to the original cluster as well as privacy is also achieved. It is also proved that the processing time of the data is considerably reduced when compared to the other methods that are being used for this purpose.**

***Index Terms*— K-Means, S-Shaped fuzzy Membership Function, Privacy, Clustering** *(key words)*

I. INTRODUCTION.

Data mining is the process used to analyze large quantities of data and gather useful information from them. It extracts the hidden information from large heterogeneous databases in many different dimensions and finally summarizes it into categories and relations of data [1] In order to learn a system in detailed manner, we should be able to decrease the system complexity and increase our understanding about the system. For any application, if the information available is imprecise then fuzzy reasoning provides a better solution [13].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. Patient's name).

Table 1 – Patient Medical Data

ID	Attributes			
	Age	*Sex*	*Pin code*	*Disease*
1	22	M	789001	Dengu
2	33	F	789002	Dengu
3	44	F	789003	SwineFlu
4	55	M	789004	Malaria

Table 2 - Patients Registration List

ID	Attributes		
	Age	*Sex*	*Pin code*
	Name	*Age*	*Sex*
1	Anup	22	M
2	Rima	33	F
3	Asha	44	F
4	Sagar	55	M

Table 2 shows a sample patient'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 (Pin-code, age, sex). One can clearly identify who is having Dengu, Swine flu Malaria etc which is sensitive attribute. This Microdata is a valuable source of information for the research and allocation of public funds, trend analysis and medical research. Our work is publishing this data without revealing sensitive information about them as shown in Table 3.

Table 3- A 2-ANONYMOUS TABLE

ID	Attributes			
	Non Sensitive			Sensitive
	Age	*Sex*	*Pin code*	*Disease*

1	2*	*	7890*	Dengu
2	3*	F	7890*	Dengu
3	4*	F	7890*	SwineFlu
4	5*	*	7890*	Malaria

This idea of using fuzzy logic is applied to preserve the individual information while revealing the details in public. This paper mainly focuses on converting the sensitive data into modified data by using S – shaped fuzzy membership function. K means clustering algorithm is applied on the modified data and it is found that the relativity of the data is also maintained. There are a number of methods used for preserving the privacy of the data while clustering. Some of the methods are use of cryptographic algorithms, noise addition, and data swapping. All of these methods introduce a bit of complexity in the algorithm and increase the processing time. Our main aim is to reduce this processing time and at the same time provide an optimum solution to the problem of privacy preserving. For this purpose we are using the concept of fuzzy approach.

S – shaped fuzzy membership function is given by,

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

Where x is the value of the sensitive attribute, a & b is the minimum and maximum value in the sensitive attribute list.

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 explains the proposed system based around the fuzzy based membership function approach and how it can be used for privacy preserving. Section 4 shows the proposed method experimental result and comparison with K means algorithm.

II. LITERATURE SURVEY

In recent year's lot of research work has been carried out to preserve data privacy before releasing the data for various research purposes which adopts various techniques like Data Auditing, Data Modification, Cryptographic methods and k-anonymity.

A. Modification-Based Techniques

A number of techniques have been developed for a quantity of data mining techniques like classification, association rule discovery and clustering, based on the hypothesis that discerning data modification or sanitization is an NP-Hard problem, and for this basis, alteration can be used to address the complexity issues.

- Swapping values between records (e.g. [10])
- Replacing the original database by a sample from the same distribution (e.g. [18] [21] [23])
- Adding noise to the values in the database (e.g. [22] [11])
- Adding noise to the results of a query (e.g. [20])
- Sampling the result of a query (e.g. [25]).

B. Cryptography-based techniques

In Cryptographic methods [2] data is encrypted using protocols like secured multiparty computation (SMC). It is a study of mathematical techniques, related to aspects of information security such as confidentiality, data integrity, entity authentication, and data origin authentication is shaping the way that information is safely and securely transmitted over the Internet. Sensitive information is quit large, Such as Credit Card Information, Social Security Numbers, Private correspondence, Military statement, Bank account information.

C. Refurbishing-based techniques

Refurbishing-based techniques are techniques where the original circulation of the data is reconstructed from the randomized data.

Algorithm for Refurbishing

Step 1 : Creating randomized data replica by data perturbation of entity data records.

Step 2 : Recreate distributions, not values in individual records.

Step 3 : By means of using the reconstructed distributions, fabricate the original data.

For refurbishing the predicament of reconstructing original distribution from a given distribution can be viewed in the broad-spectrum framework of inverse problems [24]. In [12], it was publicized that for smooth sufficient distributions (e.g. slowly varying time signals), it is possible to fully recuperate original distribution from Proceedings of the World Congress on Engineering 2008 Vol I non-overlapping, adjoining partial sums. Such partial sums of true values are not available to us. We cannot formulate priori assumptions about the original distribution; we merely be acquainted with the distribution used in randomizing values of an attribute. There is rich query optimization literature on estimating attribute distributions from partial information [17]. In the OLAP literature, there is work on approximating queries on sub-cubes from higher-level aggregations (e.g. [19]). However, these works did not have to cope with information that has been intentionally distorted.

Kargupta et al. used a data refurbishing approach to derive private information from a disguised data set. Namely, a new data set X¤ is reconstructed from the disguised data using certain algorithms, and the difference between X¤ and the actual original data set X indicates how much private information can be disclosed. The further apart X¤ is from X, the higher level of the privacy preservation is achieved. Therefore, the difference between X¤ and X can be used as the measure to quantify how much privacy is preserved.

In Noise addition methods [3] we add some random noise (number) to numerical attributes. This random number is usually drawn from a normal distribution with a small standard deviation and with zero mean. Noise is added in a controlled way so as to maintain means, variances and co-variances of the attributes of a data set. However, noise addition to categorical attributes is not as straightforward as the noise addition to numerical attributes, due to the absence of natural ordering in categorical values.

Data swapping [4] [5] interchange the attribute values between different records. Similar attribute values are interchanged with higher probability. The unique feature of this approach is all original values are kept back within the data set and only the positions are swapped.

In Aggregation [6] [7] instead of individual values the records are replaced by a group representative. Aggregation refers to both combining a few attribute values into one, or grouping a few records together and replacing them with a group representative.

In [15] the clustering operation is performed after applying 2-dimensional transformations to the data. A different approach for privacy preservation in data mining is given in [16]. This introduces the concept of fuzzy sets which is just an extension to the generic set theory. By using fuzzy sets we can perform a gradual assessment of the data set given to us and this is done by using a fuzzy membership function. Each linguistic term can be represented as a fuzzy set having its own membership function.

Fuzzy c-Means (FCM) can be used for clustering (Fig. 1). But any element in the set may have membership in more than one category [14].

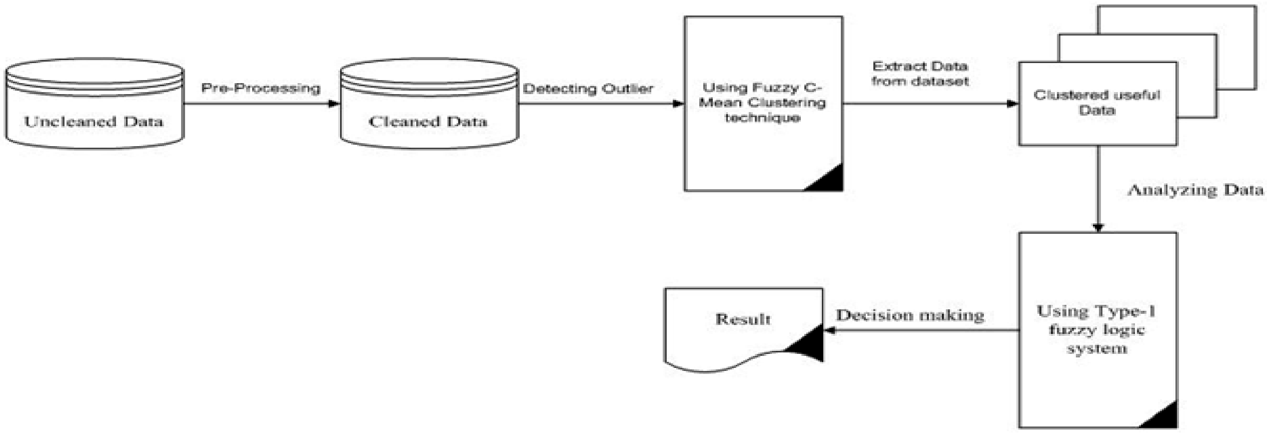

Fig.1 For Refurbishing

III. PROPOSED SYSTEM

A. Objective of the Problem

The sensitive attribute can be selected from the numerical data and it can be modified by any data modification technique. After modification, the modified data can be released to data mining researchers or any agency or firm. If they can apply data mining techniques such as clustering, classification, etc for data analysis, the modified table does not affect the result. In this work, we have applied k-means clustering algorithm to the modified data and verified the result. The steps involved in this work are,

1. Sensitive Attribute Selection.
2. Data Transformation perturbative masking technique for modifying the sensitive attribute.
3. Applying k-means algorithm for original and fuzzified data.
4. Compare the processing time of the original and fuzzified data.

B. Proposed Algorithm

Algorithm for Data Transformation

1. Consider a database D consists of T tuples. D={t1,t2,…tn}. Each tuple in T consists of set of attributes T={A1,A2,…Ap} where Ai €T and Ti € D.

2. Identify the sensitive or confidential numeric attribute AR

3. Identified sensitive attribute values are modified by using S – shaped fuzzy membership function and the fuzzified data is sent back to the user.

4. The received data is grouped into different clusters Using K – means algorithm.The objective of the Clustering algorithms is to group the similar data together depending upon the characteristics they possess.[1]K-means clustering clusters the similar data with the help of the mean value and squared error criterion.[1]Apply the k-means clustering for both original and modified data set to get the clusters.

C. Proposed system architecture

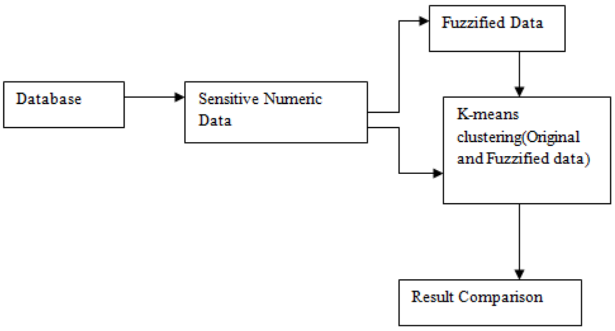

Fig.2 Proposed system architecture

IV. RESULT ANALYSIS

For our experimental purpose we have used data sets like census details, air distance for our experiment. For a given original data, the equivalent fuzzified data is given in Table 4. Table 5 shows the result obtained by performing the clustering on the original data and the result of the clustering performed on the fuzzy data. It is evident that in both the cases, resultant clusters contain the same set of elements. This shows that, the use of fuzzy conversion does not hamper the relativity of the cluster data.

Table 4 - Example Data Set

Original Data	2	4	10	12	3	20	30	11	25
Fuzzified Data	0	0.0102	0.1632	0.2551	0.2551	0.7449	1	0.2066	0.9362

Say for the original data set {2,4,10,12,3,20,30,11,25}-we calculate the mean which is 13 and thus form two clusters, one of which has elements less than 13 and the other with elements greater than 13 as shown in Table 5.

We now calculate the fuzzified data using S – shaped fuzzy membership function where x is the original data set {2,4,10,12,3,20,30,11,25} and a is 2 and b is 30.

Case 1 when x=2 then the fuzzified data will be 0 as (x=2)<=(a=2)

Case 2 when x=4 then the fuzzified data will be 2((4-2)/(30-2))2=2(2/28)2=1/98=0.0102

Case 3 when x=10 then the fuzzified data will be 2((10-2)/(30-2))2=2(8/28)2=8/49=0.1632 and so on for other data sets. These fuzzified data would also form clusters as above mentioned for original data.

The K- means algorithm will perform the following three steps until convergence and it Iterates until stable (= no object move group):

Step 1: Determine the centric coordinate.(centroid)

Step 2: Determine the distance of each item to the centric.

Step 3: Group the item based on minimum distance.

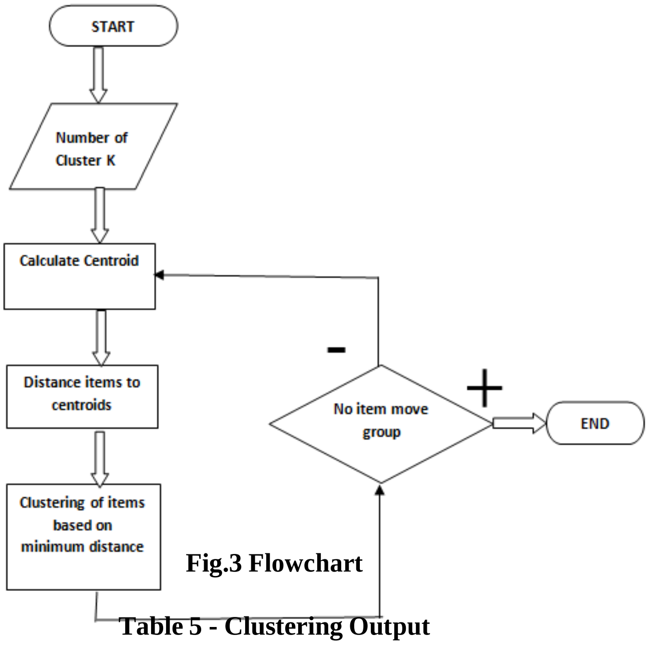
START
Number of Cluster K
Calculate Centroid
Distance items to centroids
Clustering of items based on minimum distance
-
No item move group
+
END

Fig.3 Flowchart

Table 5 - Clustering Output

Cluster 1

Data	Cluster 2 Data
{2,4,10,12,3,11}	{20,30,25}
{0, 0.0102, 0.1632, 0.2551,0.2066}	{0.7449,1,0.9362}

Finally, we are going to analyze the data utility using cluster analysis. The original and perturbed data will be separately clustered using K-Means clustering algorithm; the two outputs are compared in terms of their statistical efficiency.

From our experiment we found out that by using fuzzy approach, the processing time of the data is considerably reduced when compared to the other methods that are being used for this purpose. In the particular example that we took, we used an S-shaped fuzzy membership function in order to transform the original data into fuzzy data.

Table 6 - Comparison Table

Method Used	No of Passes	No of Clusters
K-Means	5	2
Fuzzy	4	2

V. CONCLUSION

Protecting the sensitive data and also extracting knowledge is a very complicated problem. This paper presents an approach to preserve the individual's details by transforming the original data into fuzzy data using S shaped fuzzy membership function. The main advantage of this method is that it maintains the privacy and at the same time preserves the relativity between the data values. From the results obtained by our experiments (Table 6) it is proved that performing k-means algorithm on fuzzy data increases the efficiency of the process by decreasing the number of passes required to perform the clustering. We have used numerical data for our experimentation purpose and similarly this method can be extended to categorical data. In our process, the nature of the fuzzy membership function used also affects the processing time of the algorithm and hence we can improve the working of this process by applying a different fuzzy membership function. In future this work can also be extended for data classification purpose and we would develop new masking techniques for protecting the categorical attributes.

REFERENCES

[1] Sairam et al "Performance Analysis of Clustering Algorithms in Detecting outliers",International Journal of Computer Science and Information Technologies, Vol. 2 (1) , Jan-Feb 2011, 486-488.

[2] Pinkas,."Cryptographic Techniques for Privacy-Preserving Data Mining", ACM SIGKDD Explorations, 4(2), 2002.

[3] Agrawal D, Aggarwal C.C, "On the Design and Quantification of Privacy Preserving Data mining algorithms", ACM PODS Conference,2002.

[4] Fienberg S.E. and McIntyre J. "Data Swapping:Variations on a theme by Dalenius and Reiss." In Journal of Official Statistics,21:309-323,2005.

[5] Muralidhar K. and Sarathy R. " Data Shufflinga new masking approach for numerical data",Management Science, forthcoming, 2006.

[6] Y.Li,S.Zhu,L.Wang, and S.Jajodia " A privacyenhanced micro-aggregation method", In Poc. Of 2nd International Symposium on Foundations of Information and Knowledge Systems, pp148-159, 2002.

[7] V.S. Iyengar, "Transforming data to satisfy privacy constraints", In Proc. of SIGKDD'02,Edmonton, Alberta, Canada,2002.

[8] Shuting Xu.,Shuhua Lai, "Fast Fourier Transform based data perturbation method for privacy protection", In Proc. of IEEE conference on Intelligence and Security Informatics, New Brunswick New Jersey, May 2007.

[9] Shibanth Mukharjee, Zhiyuan Chen, Arya Gangopadhyay,"A privacy preserving technique for Euclidean distance-based mining algorithms using Fourier-related transforms", The VLDB journal 2006.

[10] D. Barbara, W. DuMouchel, C. Faloutsos, P. J. Haas, J. M.Hellerstein, Y. Ioannidis, H. V.Jagadish, T. Johnson, R.Ng, V.Poosala, and K. Sevcik. The New Jersey Data Reduction Report.Data Engrg. Bull., 20:3{45, Dec. 1997.

[11] D. Dobkin, A.K. Jones, and R.J. Lipton. Secure databases:Protection against user influence. ACM TODS, 4(1):97-106, March 1979.

[12] [12] D. Meng, K. Sivakumar, and H. Kargupta. Privacy sensitive bayesian network parameter learning. In The Fourth IEEE International Conference on Data Mining(ICDM), Brighton, UK, November 2004.

[13] Zadeh L "Fuzzy sets", Inf. Control. Vol.8, PP,338 – 353, 1965.

[14] Timothy J. Ross "Fuzzy Logic with Engineering Applications", McGraw Hill International Editions, 1997.

[15] R.R.Rajalaxmi , A.M.Natarajan "An EffectiveData Transformation Approach for Privacy Preserving Clustering", Journal of Computer Science 4(4): 320-326, 2008.

[16] V.Vallikumari, S.Srinivasa Rao, KVSVN Raju, KV Ramana, BVS Avadhani "Fuzzy based approach for privacy preserving publication of data", IJCSNS, Vol.8 No.1, January 2008.

[17] B. Pinkas. Cryptographic techniques for privacy-preserving data mining. SIGKDD Explorations, 4(2), December 2002.

[18] E. Lefons, A. Silvestri, and F. Tangorra. An analytic approach to statistical databases. In 9th Int. Conf. Very Large Data Bases, pages 260-- 274. Morgan Kaufmann, Oct-Nov 1983.

[19] D.E. Denning. Cryptography and Data Security.Addison-Wesley, 1982.

[20] H.W. Engl, M. Hanke, and A. Neubaue. Regularization of Inverse Problems. Kluwer, 1996.

[21] Dakshi Agrawal and Charu C. Aggarwal, On the design andquantification of Privacy Preserving Data Mining algorithms, In Proceedings of the 20th ACM Symposium on Principles of Database Systems (2001), 247–255.

[22] J. Vaidya and C. Clifton. Privacy-preserving k-means clustering over vertically partitioned data. In Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Ddata Mining, 2003.

[23] D.E. Denning. Secure statistical databases with random sample queries. ACM TODS, 5(3):291-315, Sept. 1980.

[24] R. Agrawal and R. Srikant. Privacy Preserving Data Mining . In Proceedings of the ACM SIGMOD, pages 439–450, 2000.

[25] R. Conway and D. Strip. Selective partial access to a database.In Proc. ACM Annual Con].,pages 85-89, 1976.