Journal of Theoretical and Applied Information Technology
31st October 2011. Vol. 32 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A FUZZY BASED APPROACH FOR PRIVACY PRESERVING CLUSTERING

[1]B.KARTHIKEYAN, [2]G.MANIKANDAN, [3]V.VAITHIYANATHAN

[1]Assistant Professor, School of Computing, SASTRA University, TamilNadu, India.

[2]Assistant Professor, School of Computing, SASTRA University, TamilNadu, India.

[3]Associate Dean/Research, School of Computing, SASTRA University, TamilNadu, India.

ABSTRACT

Extracting previously unknown patterns from huge volume of data is the primary objective of any data mining algorithm. In recent days there is a tremendous growth in data collection due to the advancement in the field of information technology. The patterns revealed by data mining algorithm can be used in various domains like Image Analysis, Marketing and weather forecasting. As a side effect of the mining algorithm some sensitive information is also revealed. There is a need to preserve the privacy of individuals which can be achieved by using privacy preserving data mining. In this paper we propose a new approach to preserve sensitive information using fuzzy logic. First we perform clustering on the original data set then we add noise to the numeric data using a fuzzy membership function that results in distorted data. Set of Clusters generated using the distorted data is also relative to the original cluster as well as privacy is also achieved. It is also proved that the number of iterations for performing the clustering process is less in our approach when compared with the traditional approach.

Keywords: *K-Means, S-Shaped Function, Privacy, Fuzzy, Membership Function, Cluster*

1. 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. Patients name).

Table 1 – RAWDATA

ID	Attributes			
	Age	*Sex*	*Zip code*	*Disease*
1	22	F	613001	Fever
2	33	M	613002	Fever
3	44	M	613003	Headache
4	55	F	613004	Cough

Table 2 - VOTER REGISTRATION LIST

ID	Attributes			
	Name	*Age*	*Sex*	*Zip code*
1	ASHA	22	F	613001
2	DHONI	33	M	613002
3	SACHIN	44	M	613003
4	USHA	55	F	613004

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 2 shows a sample voter'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 (zip-code, age, sex).These types of attributes are called as Quasi identifier attributes.

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.

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 provides an insight on the conventional K-means algorithm. Section 4 explains about the fuzzy based membership function approach and how it can be used for privacy preserving. Section 5 shows the proposed method result and comparison with K-means algorithm.

2. LITERATURE SURVEY

In recent year's lot of papers are published to preserve data privacy while releasing the data for various research purposes which adopts various techniques like Data Auditing, Data Modification, Cryptographic methods and k-anonymity.

In Cryptographic methods [2] data is encrypted using protocols like secured multiparty computation (SMC). These protocols do not reveal any private information other than the final result to the data miners.

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. 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 .For salary attribute, instead of individual values it can be grouped as {Low, Medium, High}.

In Signal Transform methods [8] [9] Wavelet Transformation and Fourier Transformation are used to modify the data. These methods are fast when compared to its predecessors with improved time complexity

Query auditing methods preserve privacy by modifying or restricting the results of a query. [10]. Sweeney [11] introduced the k1-anonymity privacy requirement, which requires each record in an anonymized table to be indistinguishable with at least k-other records within the dataset, with respect to a set of quasi-identifier attributes. In this approach for achieving data anonymization methods like generalization and suppression are used. Unlike other privacy protection techniques such as data swapping and adding noise, information in a anonymous table through generalization and suppression remains ingenuous.. Table3 shows an example of 2-anonymous generalization for Table1. While k-anonymity prevents identity disclosure, it does not ensure any protection against attribute disclosure.

However no method is complete and satisfactory. Each method suffers from one or the other kind of bias [12].

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

Journal of Theoretical and Applied Information Technology
31st October 2011. Vol. 32 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

a fuzzy set having its own membership function.

Table 3
A 2-ANONYMOUS TABLE

ID	Attributes			
	Age	*Sex*	*Zip code*	*Disease*
1	2*	*	6130**	Fever
2	3*	M	6130**	Cough
3	4*	M	6130**	Headache
4	5*	*	6130**	Fever

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

S – shaped fuzzy membership function is given by

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

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

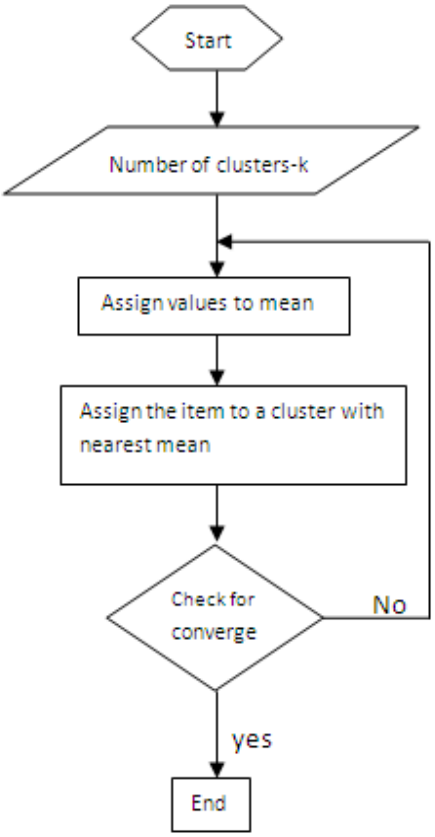
3. K-MEANS CLUSTERING ALGORITHM

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]

The algorithm is as follows

- Assign initial value for means m1,m2,……,mn
- Assign each item to the cluster which has nearest mean
- Calculate the new mean for each cluster until the convergence criteria is met



4. OUR APPROACH

The Algorithm used in our approach can be summarized as follows:

Step 1: User requests the data from the data provider.

Step 2: Data provider receives the request from the user and identifies the sensitive attribute.

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

Step 4: The received data is grouped into different clusters Using K – means algorithm.

Journal of Theoretical and Applied Information Technology
31st October 2011. Vol. 32 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

5. EXPERIMENTAL RESULTS

For our experimental purpose we have coded the k-means algorithm in Turbo C++ environment and its performance was compared using MATLAB package. We have used data sets like census details, air distance for our experiment.

Fig : 1 A Snapshot of the k-means algorithm on fuzzified data

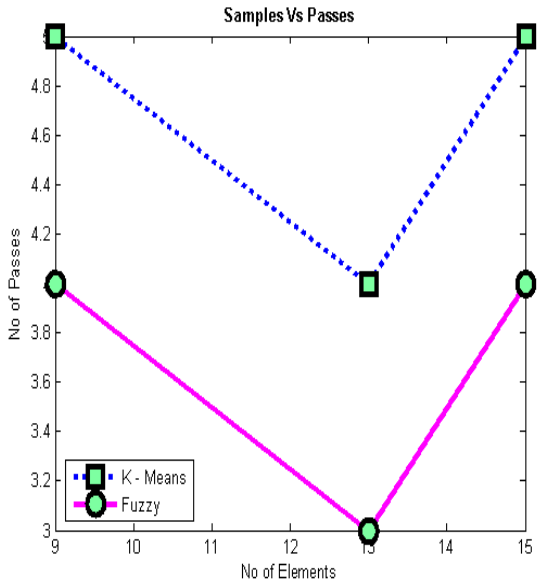


Fig: 2 A comparison between original and fuzzified data.

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	0.0025	0.7449	1	0.2066	0.9362

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 5 - Clustering Output

Cluster 1 Data	Cluster 2 Data
{2,4,10,12,3,11}	{20,30,25}
{0, 0.0102,0.1632,0.2551,0.0025,0.2066}	{0.7449,1,0.9362}

Table 6 - Comparison Table

Input Samples	No. of Elements	Method Used	No. of Passes Required	No. of Clusters
{2,4,10,12,3, 20,30,11,25}	9	K - Means	5	2
		Fuzzy	4	2
{444,62,513,267,3387,430,340,78, 57,1598,818,121, 29}	13	K - Means	4	2
		Fuzzy	3	2
{0,39,22,59,33,57,32,89,73, 29,46,16, 83,120,45}	15	K - Means	5	2
		Fuzzy	4	2

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.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

6. CONCLUSION

This paper presents a potential 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.

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 Shuffling-a new masking approach for numerical data", Management Science, forthcom-ing, 2006.

[6] Y.Li,S.Zhu,L.Wang, and S.Jajodia " A privacy-enhanced 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] Nabar S. Marthi B, Kenthapadi K, Mishra N, Motwani R., "Towards Robustness in Querry Auditing" VLDB Confer-ence, 2006.

[11] L. Sweeney, "^-anonymity: a model for protecting privacy", International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 2002, pp. 557-570.

[12] IEEE Transactions on Knowledge and Data Engineering., Vol.18,No.1,2006.

[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 Effective Data 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.