Preserving Data Clustering with Expectation Maximization Algorithm

Leila Jafar Tafreshi*
Department of Electrical and Computer Engineering, Semnan University, Semnan, Iran
leila.tafreshi66@yahoo.com
Farzin Yaghmaee
Department of Electrical and Computer Engineering, Semnan University, Semnan, Iran
f_yaghmaee@semnan.ac.ir

Received: 30/Jun/2015 Revised: 03/Aug/2016 Accepted: 13/Aug/2016

Abstract

Data mining and knowledge discovery are important technologies for business and research. Despite their benefits in various areas such as marketing, business and medical analysis, the use of data mining techniques can also result in new threats to privacy and information security. Therefore, a new class of data mining methods called privacy preserving data mining (PPDM) has been developed. The aim of researches in this field is to develop techniques those could be applied to databases without violating the privacy of individuals. In this work we introduce a new approach to preserve sensitive information in databases with both numerical and categorical attributes using fuzzy logic. We map a database into a new one that conceals private information while preserving mining benefits. In our proposed method, we use fuzzy membership functions (MFs) such as Gaussian, P-shaped, Sigmoid, S-shaped and Z-shaped for private data. Then we cluster modified datasets by Expectation Maximization (EM) algorithm. Our experimental results show that using fuzzy logic for preserving data privacy guarantees valid data clustering results while protecting sensitive information. The accuracy of the clustering algorithm using fuzzy data is approximately equivalent to original data and is better than the state of the art methods in this field.

Keywords: Privacy Preserving; Clustering; Data Mining; Expectation Maximization Algorithm.

1. Introduction

Huge volumes of individual's information are frequently collected and analyzed by various applications such as shopping habits, criminal records, medicinal history and credit records. On the other hand, data has an important role in decision making for business organizations and governments. Therefore, analyzing such data may threaten individual's privacy. Also, companies that generate a huge burden of data often need to transmit these data to third parties for their studies. As data usually contains sensitive information about people and corporations, their releasing to third parties requires mechanisms to make sure that data privacy is preserved [1].

For example, a bank may release credit records of individuals for statistical purpose or a hospital may release patient's diagnosis records. Both of these applications need the individual's data to be private while data mining process.

However, there may exist other attributes that can be used, in combination with an external database, to recover the personal identities.

Privacy preserving data mining entails two notions:

I. Extracting or mining knowledge from large amounts of data.
II. Performing data mining in such a way that data privacy is not compromised.

For these purposes, we should preserve the privacy of data or the knowledge discovered from mining results [2].

The extracted knowledge form data is generally expressed in the form of clusters, decision trees or association rules which allows one to mine the information. In this work we focus on clustering methods.

Clustering is known as identification of similar objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution patterns and correlations between data attributes.

In privacy preserving data while clustering, the main goal is to find the clusters of data without revealing the content of data elements themselves.[3] It is important to use a kind of modification technique in order to achieve higher accuracy of data mining results.

Most of works in privacy preserving clustering are developed on k-means algorithm by applying the model of secure multi-party computation on different distributions [4].

Data modification techniques for PPDM can be classified into two principle groups: perturbation-based and anonymization-based techniques according to how the protection of privacy.

Anonymization refers to an approach where identity or/and sensitive data about record owners are to be hidden. It even assumes that sensitive data should be retained for analysis.

* Corresponding Author

Perturbation of data is a very easy and effective method for protecting the sensitive information of the data from unauthorized users or hackers. There are two types of data perturbation for protecting data namely:

- Value-based Perturbation: the purpose is to preserve statistical characteristics and columns distribution. This approach perturbs data by adding noise, or other randomized processes.
- Multi-Dimensional Perturbation: aims to hold Multi-Dimensional information. This approach includes random projection or random rotation techniques. Comparing to other multi-dimensional data perturbation methods, these perturbations exhibit unique properties for privacy preserving data classification and clustering. [6]

All of these methods introduce a bit of complexity in their algorithms. Our main goal is to introduce a simple and optimum solution for privacy preserving data mining problem using fuzzy logic.

In this research we compare our clustering results with some other methods such as random projection, random rotation and noise addition.

In this work we have applied the idea of using fuzzy logic to preserve the individual's information while revealing details in public. In this regard, we used one of the fuzzy membership functions described in table (1) to transform original data. Then transformed data were clustered by EM algorithm to evaluate our method.

The rest of the paper is organized as follows: in section 2 we describe related works in privacy preserving data mining. Section 3 presents state of the art works which we used them to compare the experimental results. Section 4 explains the proposed method based on fuzzy logic. Section 5 and 6 is dedicated to experimental results and conclusion respectively.

Table 1. Fuzzy membership functions

P-shaped MF	***Z–shaped MF***	***S–shaped MF***
F (x؛a,b,c,d) = $0, x \le a$ $a \le x \le \frac{a+b}{2} \rightarrow 2\left(\frac{x-b}{b-a}\right)^2$ $\frac{a+b}{2} \le x \le b \rightarrow 1-2\left(\frac{x-b}{b-a}\right)^2$ $b \le x \le c \rightarrow 1$ $c \le x \le \frac{c+d}{2} \rightarrow 1-2\left(\frac{x-c}{d-c}\right)^2$ $\frac{c+d}{2} \le x \le d \rightarrow 2\left(\frac{x-d}{d-c}\right)^2$ $x \ge d \rightarrow 0$ The parameters a and d locate the "feet" of the curve, while b and c locate its "shoulders.	F (x؛a,b) = $x \le a \rightarrow 1$ $a \le x \le \frac{a+b}{2} \rightarrow 1-2\left(\frac{x-a}{b-a}\right)^2$ $\frac{a+b}{2} \le x \le b \rightarrow 2\left(\frac{x-b}{b-a}\right)^2$ $x \ge b \rightarrow 0$ The parameters a and b locate the extremes of the sloped portion of the curve.	F (x؛a,b) = $x \le a \rightarrow 0$ $a \le x \le \frac{a+b}{2} \rightarrow 2\left(\frac{x-b}{b-a}\right)^2$ $\frac{a+b}{2} \le x \le b \rightarrow 1-2\left(\frac{x-b}{b-a}\right)^2$ $x \ge b \rightarrow 1$ The parameters a and b locate the extremes of the sloped portion of the curve.

Gaussian MF	***Sigmoid MF***
$gaussian(x; c, \sigma) = e^{-\frac{1(\frac{x-c}{\sigma})^2}{2}}$ Parameter c represents the MFs center and σ determines the MFs width.	$sig(x; a, c) = \frac{1}{1+\exp[-a(x-c)]}$ Parameter a, controls the slope at the crossover point b.

2. Related Works

The term, privacy-preserving data mining, introduced by Agrawal and Srikant in 2000 [7]. The initial idea of PPDM is extending traditional data mining techniques to work with the modified data those mask sensitive information. The key issues is how to modify data and how to recover data mining results from the modified data. The solutions are often tightly coupled with the data mining algorithms.

One of such techniques is Rotation-Based Transformation (RBT) [8]. A novel spatial data transformation method for Privacy Preserving Clustering. This method is designed to protect the underlying attribute values subjected to clustering without jeopardizing the similarity between data objects under analysis. Releasing a database transformed by RBT, a database owner meets privacy requirements and guarantees valid clustering results. The data is shared after the transformation to preserve privacy without normalization. Researches show that having previous knowledge, the random rotation perturbation may become involved in privacy violations against different attacks including Independent Component Analysis (ICA), attack to rotation center and distance-inference attack.

Another work is Random Projection-Based, which is a dimension reduction technique, introduced by Kun Liu, Hillol Kargupta and Jessica Ryan in 2006 [9]. This work uses random projection matrices which is a tool for PPDM. It proves that, after perturbation, the distance-related statistical properties of the original data is well maintained without divulging the dimensionality and the exact data values. The experimental results demonstrate that this technique can be successfully applied to different kinds of data mining tasks such as inner product/Euclidean distance estimation, correlation matrix computation, clustering, outlier Detection and linear classification.

However, this technique can hardly preserve the distance and inner product during the modification in comparison with geometric and random rotation techniques. It has been also clarified that having previous knowledge about Random Projection-Based perturbation technique may be caught into privacy breach against the attacks. Our purposed technique does not loose data which is its benefit versus random projection.

Another work for privacy preserving clustering is double reflecting data perturbation and rotation data perturbation which is proposed in [10].

Kadampur and Somayajulu presented a method of privacy preserving clustering by cluster bulging [11]. In this method, the original values of individual objects are not revealed and the privacy of individual objects is preserved; but the perturbed dataset is still relevant for cluster analysis.

Yifeng and Harbin combined the random response technology and the geometric data transformation method in 2009 which is called random response method of geometric transformation [12]. It can protect the privacy of numerical data. Theoretical analysis and experimental results show that the algorithm improves privacy protection than the previous algorithms.

In [13] a family of geometric data transformation methods (GDTMs) which ensure that the mining process up to a certain degree of security is introduced. Their method is designed to address the privacy preservation in clustering analysis, This method distort only confidential numerical attributes to meet privacy requirements, while preserving general features for clustering analysis.

Shibnath Mukherjee, Zhiyuan Chen and Aryya Gangopadhyay proposed an integrated Dimension Reduction-based approach for data reduction and privacy for distance-based mining algorithms using Fourier-related transform [14]. Experimental results demonstrate that the proposed approach leads to much better mining quality than the existing random perturbation and random projection approaches given the same degree of privacy in both centralized and distributed cases.

Shalini Lamba present a potential approach to preserve the individual's details by transforming the original data into fuzzy data [15]. They have used only numerical data for their experimentation purpose. The main goal of their technique is reducing the run time of preserving data privacy while clustering.

3. State of the Art Methods

In this section we explain the three famous methods in this field which we used them to compare the experimental results of proposed method.

3.1 Expectation Maximization

Expectation Maximization (EM) is a well-established clustering algorithm in statistics community. EM is a distance-based algorithm which assumes that the dataset can be modeled as a linear combination of multivariate normal distributions and the algorithm finds the distribution parameters that maximizes a model quality measure, called log likelihood.

EM is linear in database size, robust to noisy data, can handle high dimensionality and has a very good quality while using huge datasets.

This algorithm assumes Apriori that tries to fit the data into 'n' Gaussian channel by expecting the classes of all data point and finding the maximum likelihood of Gaussian centers. Algorithmic steps for EM is as follows:

Let $x = \{x_1, x_2, x_3, \dots, x_n\}$ be the set of data points, $v = \{\mu_1, \mu_2, \mu_3, \dots, \mu_c\}$ be the set of means of Gaussian.

$p = \{p_1, p_2, p_3, \dots, p_c\}$ is the set of probability of occurrence of each Gaussian.

1. On the i^{th} iteration initialize:

$$\lambda_t = \{\mu_1(t), \dots \mu_c(t), \Sigma_1(t), \Sigma_2(t), \dots \Sigma_c(t), p_1(t), \dots p_c(t)\} \quad (1)$$

2. Compute the "expected" classes of all data points for each class using:

$$p(w_t | x_k, \lambda_t) = \frac{p(x_k | w_t, \lambda_t) p(w_t | \lambda_t)}{p(x_k | \lambda_t)} = \frac{p(x_k | w_t, \mu_t(t), \Sigma_t(t)) p_t(t)}{\sum_{j=1}^{c} p(x_k | w_j, \mu_j(t), \sum_j(t)) p_j(t)} \quad (2)$$

3. Compute the maximum likelihood (μ) given our data class membership distribution using:

$$\mu_t(t+1) = \frac{\sum_k p(w_t | x_k, \lambda_t) x_k}{\sum_k p(w_t | x_k, \lambda_t)} \quad (3)$$

$$p_t(t+1) = \frac{\sum_k p(w_t | x_k, \lambda_t)}{R} \quad (4)$$

Where *'R'* is the number of data points. Repeat the steps 2 and 3 while algorithm converges.

3.2 Random Projection

Random projection [9] refers to a technique of projecting a set of data points from high-dimensional space to a randomly chosen lower-dimensional subspace.

If the matrix $X\ m \times n (or\ Y\ m \times n)$ indicates original dataset, $R_{n \times k} (k < n) (or\ R'_{k \times m} (k < m))$ is a random

matrix such that each entry $r_{i\times j}$ of $R(or\ R')$ is independent and identically chosen from some unknown distribution with mean zero and variance σ_r^2, the Column-wise Projection *G(X)* and Row-wise Projection *G(Y)* will be defined as below:

$$G(X)=\frac{1}{\sqrt{k\sigma_r}}XR, G(Y)=\frac{1}{\sqrt{k}\sigma_r}R'Y \quad (5)$$

The key idea of random projection arises from the Johnson-Linden Strauss Lemma. According to this lemma, it is possible to maintain distance-related statistical properties simultaneously with dimension reduction for a dataset. Therefore, this perturbation technique can be used for different data mining tasks like including inner product/Euclidean distance estimation, correlation matrix computation, clustering, outlier detection, linear classification, etc. [16].

This method reduces the dimensionality of data by projecting it onto a lower dimensional subspace using a random matrix with columns of unit length [17].

3.3 Random Rotation

This category includes all orthonormal perturbations. $R_{d\times d}$ Represent the rotation matrix. Geometric rotation of the data X is as a function $f(X), f(X)=RX$. Transformation will not change the label of data tuples. $R_{d\times d}$ Have the following properties.

Let R^T represent the transpose of the matrix R, r_{ij} represent the (i,j) element of R, and I be the identity matrix. Both the rows and the columns of R, are orthonormal i.e., for any column $j, \sum_{i=1}^{d} r\,ij^2 = 1$ and for any two columns j and k, $\sum_{i=1}^{d} r_{ij}r_{ik} = 0$. The similar property is held for rows. The definition infers that $R^TR = RR^T = I$. It also implies that by changing the order of the rows or columns of rotation matrix, the resulting matrix is still a rotation matrix. A random rotation matrix can be efficiently generated following the "Haar" distribution. For example the Eq. (6) is a rotation matrix.

$$R=\begin{bmatrix} cos\theta & -sin\theta \\ sin\theta & cos\theta \end{bmatrix} \quad (6)$$

The goals of rotation based data perturbation are: preserving the accuracy of classifiers and preserving the privacy of data.

A key feature of rotation transformation is preserving length and the Euclidean distance between any pair of points x and y, the inner product is also invariant to rotation.

3.4 Random Noise Addition Technique

This technique is described in as follows: Consider ^{n} original data $X_1, X_2, \dots, X_N$, where X_i are variables following the same independent and identical distribution. The distribution function of X_i is denoted as F_x, n random variables $Y_1, Y_2, \dots, Y_N$ are generated to hide the real values of X_i by perturbation. Disturbed data will be generated as:

$$w_i = X_i + Y_i \text{ Where } i = 1, \dots, n \quad (7)$$

It is also assumed that the added noise variance is large enough to let an accurate estimation of main data values take place. Then, according to the perturbed dataset $w_1, \dots, w_n$ known distributional function *FY* and using a reconstruction procedure based on Bayes rule, the density function $f'x$ will be estimated by Equation:

$$f'_X(a)=\frac{1}{n}\sum_{i=1}^{n}\frac{\int_{-\infty}^{a} f_Y(w_i-a)f_X(a)}{\int_{-\infty}^{-\infty} f_Y(w_i-z)f_X(z)dz} \quad (8)$$

4. Proposed Method

In this work we propose a new approach to preserve sensitive data in general databases which have both numerical and categorical attributes with clustering by Expectation Maximization algorithm. First of all, we map categorical private data into numerical data and second the numerical private data are transformed using fuzzy membership functions described in Table (1). Finally, we cluster the transformed data using EM algorithm. For better clarity we describe our process step by step.

- Categorical private attributes are mapped to numerical values. This can be done by a simple method which works on ASCII codes of alphabet characters of each field.
- Private attribute's values are transformed using fuzzy membership functions and fuzzy data are sent back to the user.
- The received data are grouped into different clusters using EM.

As mentioned above, in second phase we transform private attribute's values to fuzzy values using a fuzzy membership function such as P-shaped, Z-shaped, S-shaped, Gaussian and Sigmoid, which are described in Table (1). Fuzzy logic is an approach to compute based on "degrees of truth" rather than usual "true or false" (0 or 1) Boolean logic. Fuzzy logic has been employed to handle the concept of partial truth, where the truth values may lie between complete truth or complete false.

By fuzzifying private data through a fuzzy membership function, each point in the input space is mapped to a value between 0 and 1. So, no one can find the real values of the private data. In the next section we show that this data transformation do not considerably change the mining results.

In the following we describe the datasets and data mining software which we used.

For our experimental purpose, we have used the Weka software which is a popular software for machine learning applications [19].

We have clustered the following described datasets by Weka software to compare the clustering accuracy of fuzzy data with state of the art techniques. We have used datasets with private attributes which should be preserved in data mining process.

The datasets are Pima Indians, German Credit, Student Evaluation, Census Income and Bank Marketing are taken from the UCI repository. These datasets have different number of instances and attributes. Also, they

have various attribute types and different number of private attributes.

Pima Indians dataset, Includes cost data (donated by Peter Turney) from National Institute of Diabetes and Digestive and Kidney Diseases. German credit describes the German people credit information. Student evaluation dataset contains 5820 evaluation scores provided by students from Gazi University in Ankara (Turkey). The Census Income (ADULT) dataset predicts whether income exceeds $50k/y or not. Bank marketing dataset is related to direct marketing campaigns of a Portuguese banking institution. The above mentioned datasets details are described in Table (2).

Table (3) shows an example of transforming a record of Census Income dataset using Gaussian fuzzy membership function. The first row shows a part of the original record, second row shows the mapped record and the third row shows the transformed record using Gaussian fuzzy membership function.

Table 2. Datasets description

Datasets	*Number of Records*	*Number of Attributes*	*Number of Private Attributes*	*Number of Clusters*
Pima Indians	768	8	2	2
German Credit	1000	20	5	2
Student Evaluation	5820	33	4	13
Census Income	11012	14	7	2
Bank Marketing	41188	20	4	2

Table 3. An example of transforming private data using Gaussian mf

	Age	*Work class*	*Education*	*Marital-status*	*Occupation*	*Relationship*
Original data	34	Local-gov	Some-college	Never-married	Protective-serve	Not-in-family
Mapped data	34	1	4	3	10	4
Transformed data with Gaussian mf	0.185	0.167	0.186	0.2	0.609	0.184

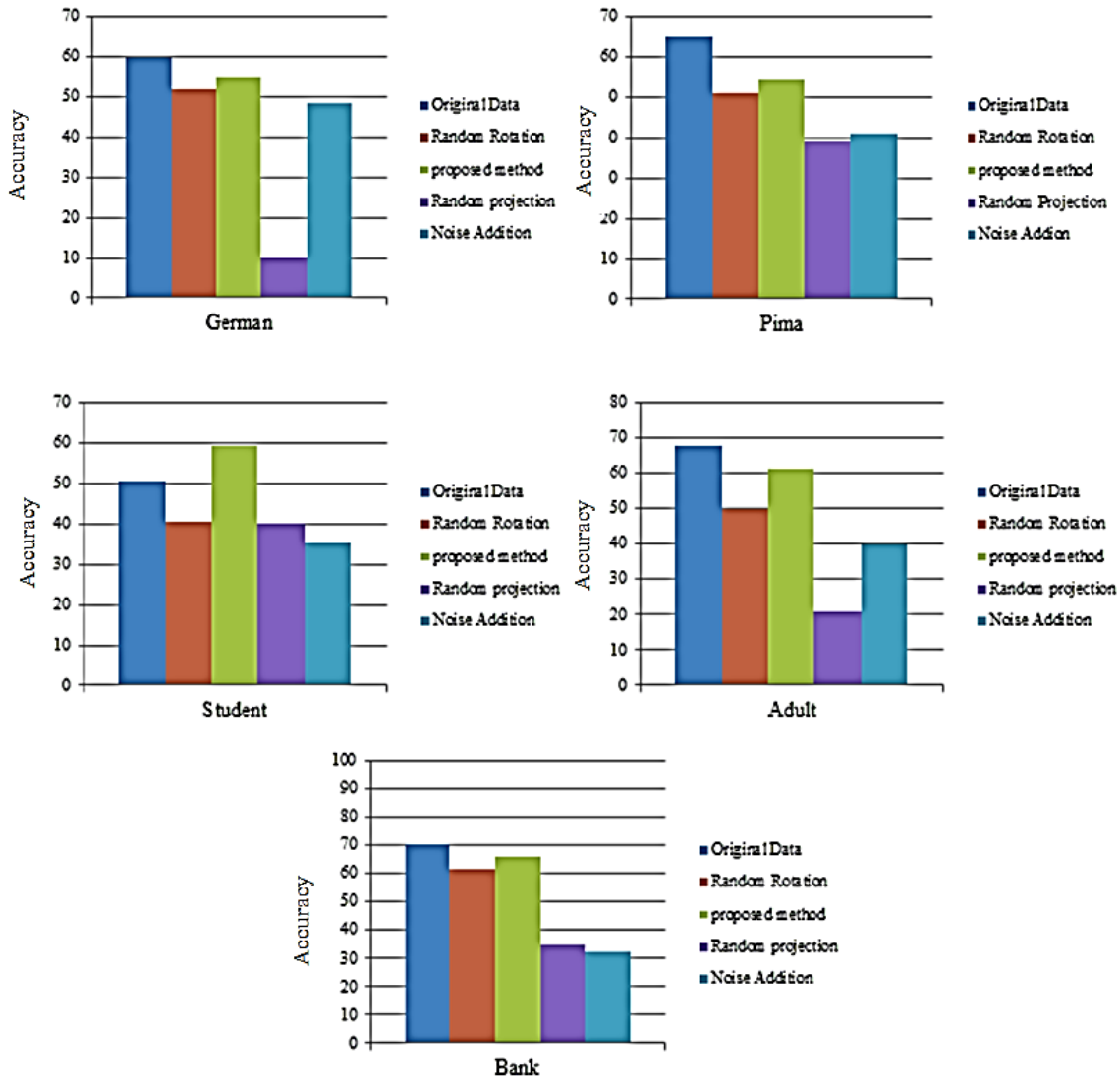

Fig. 1. Comparing the accuracy of EM clustering algorithm while transforming data using different fuzzy memberships

5. Experimental Results

Although, applying different fuzzy membership functions may lead to different clustering accuracy; in this study we used the average result of applying five membership functions described in Table (1) on private data for a fair comparison. As presented in Fig.1, we found that the accuracy of EM algorithm through fuzzy data in all datasets is better than the clustering accuracy of random rotated, random projected and noise added data.

We clustered fuzzified data using EM algorithm. As shown in fig. 1, the clustering accuracy is not changing considerably in compared with those of the original data. This can be due to the fuzzy logic feature which maps each input value to a value between 0 and 1. In addition, using fuzzy logic, data privacy is preserved and no one can guess the original data from the fuzzified data.

By using other methods such as random rotation, random projection and noise addition, clustering accuracy decreases and data privacy is not preserved appreciably.

For noise addition technique we used the average of 20%, 30%, 50% and 60% random noise addition to private data and for random projection method we used the average decreasing about 70%, 40%, 20% and 10% of the dataset's dimension. In Fig. 1 we compared the clustering accuracy of fuzzy data with these two algorithms and random rotation algorithm as well.

However we analyzed the results of proposed method with different fuzzy membership as described in Table (1). Our experimental results show that the shape of fuzzy membership function used to transform original data has not a meaningful effect on EM clustering accuracy. This means that we can use simple fuzzy membership function to reduce the computation complexity.

6. Conclusions

Releasing data and extracting knowledge without violating individual's privacy are important and complicated problems. Most of the existing methods preserve data privacy via complicated processes with disadvantages such as making essential changes or losing data; so they lose the benefits of mining. In this work we focused primarily on privacy issues in data mining, notably when data are revealed for clustering with Expectation Maximization algorithm.

We transformed private data by using fuzzy membership functions to convert a database into a new one in such a way as to preserve the main features of the original database for mining with Expectation Maximization algorithm.

Using fuzzy logic, the relationship between data is maintained while no losing data is occurred.

The results show that our method improves more than 3 percent in clustering accuracy in comparison with other conventional models, such as random rotation, random projection and noise addition.

Another important benefit of proposed algorithm is that by using fuzzy logic, user can tradeoff between privacy preserving and data mining results. This means that proposed method has better flexibility in real applications which requires different security levels.

Future works can be focuses on different clustering algorithms or using encryption algorithms to improve the security requirements.

References

[1] NIRT, RGPV, and Sajjan Singh Nagar, "A review paper on Privacy-Preserving Data Mining." Scholars Journal of Engineering and Technology (SJET), Vol. 1, No. 3, pp. 117-121, 2013.

[2] Lokesh Patel, Prof. Ravindra Gupta, "A Survey of Perturbation Technique for Privacy-Preserving of Data." International Journal of Emerging Technology and Advanced Engineering, Vol. 3, No. 6, pp. 162-166, 2013.

[3] Jharna Chopra, Sampada Satav, "Privacy preservation techniques in data mining." International Journal of Research in Engineering and Technology (IJRET), Vol. 2, No. 4, pp. 537-541, Year 2013.

[4] Tamanna Kachwala, Dr. L. K. Sharma, "A Literature analysis on Privacy Preserving Data Mining." International Journal of Innovative Research in Computer and Communication Engineering, Vol. 3, Issue 4, April 2015.

[5] A.P. Dempster, N.M. Laird, D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm." Journal of the Royal Statistical Society, Series B, Vol. 39, No. 1, pp. 1–38,1977.

[6] Ronica Raj, Veena Kulkarni, "A Study on Privacy Preserving Data Mining: Techniques, Challenges and Future Prospects." International Journal of Innovative Research in Computer and Communication Engineering, Vol. 3, Issue 11, November 2015.

[7] R. Agrawal and R. Srikant. "Privacy Preserving DataMining." In Proc. ACM SIGMOD Conference on Management of Data, Dallas, Texas, 2000, pp. 439-450.

[8] S. R. M. Oliveira, and O. R.Zaiyane, "Achieving Privacy Preservation When Sharing Data for Clustering." In Proc. Workshop on Secure Data Management in a Connected World, in conjunction with VLDB, Toronto, Ontario, Canada, 2004, pp. 67–82.

[9] Kun Liu, Hillol Kargupta, Senior Member, IEEE, and Jessica Ryan, "Random Projection-Based Multiplicative Data Perturbation for Privacy Preserving Distributed Data Mining." IEEE transactions on knowledge and data engineering, Vol. 18, No. 1, PP. 92-106, 2006.

[10] Liming Li, Sch. of Manage, Fuzhou Univ, Fuzhou, Qishan Zhang, "A Privacy preserving Clustering

Technique Using Hybrid Data Transformation Method." in In Proc. IEEE International Conference, 2009, PP. 1502 - 1506.
[11] Mohammad Ali Kadampur, D.V.L.N Somayajulu, S.S. Shivaji Dhiraj, and Shailesh G.P. Satyam, "Privacy preserving clustering by cluster bulging for information sustenance." In Proc. 4th International Conference on Information and Automation for Sustainability (ICIAfS), Colombo, Sri Lanka, 2008, pp. 158-164.
[12] Jie Liu, Yifeng XU, Harbin, "privacy preserving clustering by Random Response Method of Geometric Transformation." In Proc. Fourth international conference on internet computing for science and engineering, 2009, pp. 181-188.
[13] Keke Chen, Ling Liu, "Geometric data perturbation for privacy preserving outsourced data mining." Knowledge and Information Systems journal, Volume 29, Issue 3, pp 657-695, December 2011.
[14] Khaled Alotaibi, V. J. Rayward-Smith, Wenjia Wang, and Beatriz de la Iglesia, "Non-linear Dimensionality Reduction for Privacy-Preserving Data Classification." in Proc. ASE/IEEE International Conference on Social Computing, 2012 and ASE/IEEE International Conference on Privacy, Security, Risk and Trust, 2012, pp. 694 - 701.
[15] Ms Shalini Lamba, Dr S. Qamar Abbas, "A model for preserving privacy of sensitive data." International Journal of Technical Research and Applications Vol. 1, No. 3, PP. 07-11, 2013.
[16] MohammadReza Keyvanpour, Somayyeh Seifi Moradi, "Classification and Evaluation the Privacy Preserving Data Mining Techniques by using a Data Modification–based Framework." International Journal on Computer Science and Engineering (IJCSE), Vol. 3, No. 2, Feb 2011.
[17] Keerti Dixit, Bhupendra Pandya, "An overview of Multiplicative data perturbation for privacy preserving Data mining." International Journal for research in applied science and engineering technology (I JRAS ET), Vol. 2, Issue VII, pp 90-96, July 2014.
[18] CHEN, K., and LIU, L. "A random rotation perturbation approach to privacy preserving data classification." In Proc. Intl. Conf. on Data Mining (ICDM) 2005.
[19] http://www.cs.waikato.ac.nz/ml/weka/

Leila Jafar Tafreshi is a MSc graduate of Artificial Intelligence at Semnan University in 2015. She received her BSc degree from Kharazmi University in 2012. Her research interests include Privacy and Data mining.

Farzin Yaghmaee received his PhD in 2010 and MSc in 2002 both in Artificial Intelligence from Sharif University of Technology, Iran and received BSc from AmirKabir University of Technology. He is now a faculty member of Electrical and Computer Engineering Department of Semnan University. His research interests are: image and video processing, text mining and Persian language processing tools.