© 2018 IJSRSET | Volume 4 | Issue 8 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section : Engineering and Technology

Geometric Data Perturbation for Privacy Preserving in Data Stream Mining

Mayur Prajapati[1] , Aniket Patel[2]

[1]Computer Engineering Department Silver oak College of Engineering & Technology Ahmedabad, India

[2]Information Technology Department Silver oak College of Engineering & Technology Ahmedabad, India

ABSTRACT

Today as we have tendency to live within the era of information explosion. It's become important to search for helpful data from large dataset. Additionally advance in web communication and hardware technology has lead to raise within the capability of storing personal information of people. Huge quantity of data stream are generated from completely different applications like shopping record, medical, network traffic etc. Sharing such type of information is incredibly important plus to business decision but the worry is that when the non-public information is leaked it may be abused for a different purposes. Therefore some quantity of privacy preserving must be done on the information before it is free to others. Ancient ways of Privacy Preserving Data Mining (PPDM) area unit designed for static information sets that makes its unsuitable for dynamic data streams. In this paper an economical and effective information perturbation methodology is proposed that aims to protect the privacy of sensitive attributes and obtaining information bunch with minimum information loss.

Keywords: Data Mining, Data Stream Mining, Privacy, Geometric Data Perturbation

I. INTRODUCTION

Data Mining is the technique for extracting the important knowledge from the huge amount of data [1]. Various data mining techniques are classification, clustering, Association rule mining [2]. In current scenario a transaction like credit card, web browsing, sensor network lead to wide and automated data storage. All these have large and continuous data flow and it's dynamic in nature so this this type of huge volume of data leads to many mining models and challenges [3] [4]. The goal of this research work is to preserve the privacy along with the accuracy. Accuracy should be maintain with less information loss by using geometric data perturbation technique to transform the original data to transformed data.

Data Stream Mining

Data stream is recent kind of information that is completely totally different than ancient static database. Data stream is real time continuous and dynamic flow of data [5] [6]. The characteristics of data streams are: Data has temporal arrangement preference; distribution of data constantly changes with time; the size of data is extremely large; Data flows in and out with needed speed; and immediate response is required [7]. Various conventional algorithm is constructed for the static database. If the data changes, it might be necessary to rescan the database that results in a lot of computation time and inability to promptly respond to the It's sequence of real time data with high data rate and application will scan once [7].

IJSRSET184848 | Received : 05 May 2018 | Accepted : 17 May 2018 | May-June-2018 [(4) 8 : 204-210]

Various data mining algorithms are available for ancient database where data is static [8] and continuous flow. Use of traditional data mining algorithm is not applicable in data stream mining because of no management and control over dataflow. If data will change, then we've to rescan the database. This may take more computational time. In data stream mining data is not persistent but rapid and time varying. Once component of data stream is processed, it's discarded. So, it's hard to retrieve it unless if we have tendency to expressly store them in memory.

Need for privacy in data stream mining

Mostly privacy means "keep information about me from being available to others". The major goal is data not be exploited. As a result of if once the data is discharged, it'll be not possible to prevent misuse [9]. There is no issue if somebody knowing my birth date, mother's surname, or social security number; however knowing all of them allow fraud [10]. Benefits of Privacy Protection are protection of private data, protection of proprietary or sensitive data, enables participation between different data owners without reveal their data to each other [11]. Informational privacy is said to the style within which personal data is collected, used and disclosed. Some problems are there like data quality, accuracy and utility [12].

Privacy preserving in data stream mining

The main aim of privacy preserving in data stream mining is the information not be misused. Because if once information is released, it will be impossible to prevent misuse [13]. Various techniques are available for privacy preserving in data mining [14]. PPDM techniques are depicted in below figure.

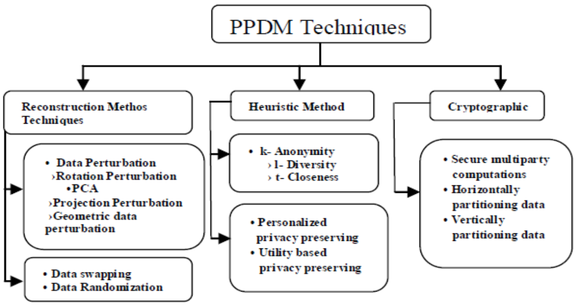

PPDM Techniques
Reconstruction Methos Techniques
Heuristic Method
Cryptographic
• Data Perturbation
›Rotation Perturbation
•PCA
›Projection Perturbation
›Geometric data perturbation
• Data swapping
• Data Randomization
• k- Anonymity
› l- Diversity
› t- Closeness
• Personalized privacy preserving
• Utility based privacy preserving
• Secure multiparty computations
• Horizontally partitioning data
• Vertically partitioning data

Figure 1. PPDM Techniques

Reconstruction Based Approach

Reconstruction based techniques construct privacy protected database by extracting the sensitive knowledge from the original database [15]. In this approach it generate the modified data from the original data by applying various methods of reconstruction based approach for preserve the privacy of sensitive knowledge of the original database [16][17]. Reconstruction based approaches generate the fewer side effects than heuristic based approach and cryptographic based approach. Reconstruction based approach perturb the sensitive knowledge of original data to achieve the privacy preserving [18]. Reconstruction based approach consists mainly three techniques in which it includes Data Perturbation, Data Swapping, and Data Randomization [19].

Data Perturbation approach consists three perturbation technique such as Rotation Perturbation, Projection Perturbation and Geometric Data Perturbation [20].

Geometric Data Peturbation Technique

Problem Description

The main objective of this proposed method is to provide privacy before release of knowledge. Perturbation method can be used for privacy preserving in data stream mining. Geometric Data Perturbations method is completely on distance base for estimating original data from the perturbed data, with addition of Gaussian noise. Geometric perturbation is an improvement to rotation perturbation by incorporating extra components like translation perturbation and noise addition to the basic form of multiplicative perturbation Y = R * X. The goal is to transform a given data set D into perturbed dataset D' that satisfies a given privacy requirement with minimum information loss for the intended data analysis task. Two step process: data stream preprocessing, data stream cluster mining. In the first step the objective is to perturb data stream to preserve data privacy. In the second step the objective is to mine perturbed data stream to cluster the data using sliding window mechanism.

Proposed Methdology Framework

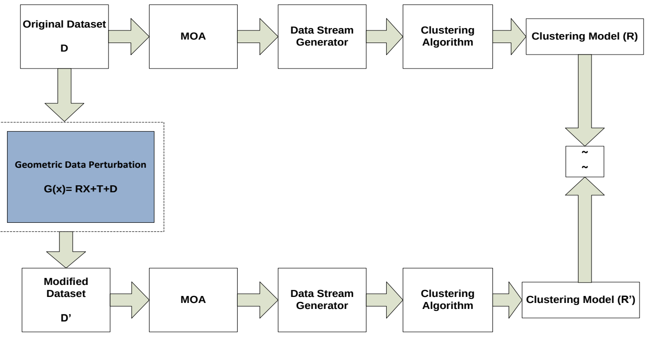


Figure 2. Proposed Methodology Framework

Algorithm

To protect the sensitive attribute, geometric data perturbation method is used. We use sliding window over the incoming data stream. The algorithm is repeated over each window. After rotation, translation Gaussian noise is added to the sensitive attribute to generate perturbed dataset. To calculate noise, all attributes except class attribute are considered. Gaussian noise is uncorrelated so, it is distributed and provide good privacy using Gaussian distribution function .Finally both datasets, original and perturbed is processed using a predefined clustering algorithm.

Algorithm: Geometric Data Perturbation for Privacy Preserving in data Stream Mining.

Input: Data Stream **D**, Sensitive Attributes **S.**

Intermediate result: Transformed Data Stream **D'.**

Output: Clustering result **R** and **R'** of the Data Stream **D** and **D'** respectively.

Steps:

1. Given input D_i^{th} tuple size n.
2. Extract the sensitive attributed $[S]_{n \times m}$.
3. Rotate the $[S]_{n \times 1}$ in to **180**· Clockwise direction and generate $[Rs]_{n \times m}$.
4. Multiply elements of $[Rs]_{n \times m}$ with $[S]_{n \times m}$ so, the transformed sensitive attributes value will be ,
   $$[X]_{n \times m} = [Rs]_{n \times m} \times [S]_{n \times m}.$$
5. Calculate Mean of the $[S]_{n \times m}$.
6. Perform Translation Transformation by adding, **Mean** + $[S]_{n \times m} = [St]_{n \times m}$.
7. Calculate the Gaussian noise,
   $$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$
8. Geometric data perturbation of sensitive attribute,
   $$[Gs]_{n \times m} = [X]_{n \times m} + [St]_{n \times m} + P(x)$$
9. Create transformed dataset **D'** by replacing sensitive attribute $[S]_{n \times m}$ in original dataset **D** with $[Gs]_{n \times m}$.
10. Apply k-mean Clustering algorithm with various values of **k** on the original data **D** having sensitive attribute **S.**
11. Apply k-mean Clustering algorithm with various values of **k** on the perturbed data **D'** having sensitive attribute **G**s.
12. Build the CMM model from step 10 and 11 and analyze it.

II. EXPERIMENT AND RESULTS

Experimental Setup

To test framework as shown in Figure 1 Massive Online Analysis (MOA) has been used. MOA framework is an open source framework for implementing algorithms and running experiments for online learning from evolving data streams [21]. It contains collection of offline and online algorithm for both classification and clustering as well as tools for evaluation [22]. In addition to this it supports bi-directional with WEKA machine learning algorithms.

To evaluate the effectiveness of proposed privacy preserving method, Experiments have been carried out on Intel Core I3 processor with 4GB memory on Windows 8. The proposed technique is implemented in Java. Simulation has been done in data stream clustering environment. The experiments were processed on two different datasets available from the UCI Machine Learning Repository [23].The brief information of chosen datasets is described below:

Table 1. Dataset Configuaration to Determine Accuracy based on Membership Matrix for Streaming Data (w=5000)

Dataset	**Total instances**	**Instances processed**	**Attributes protected**
Adult	32561	30k	Age, Fnlwgt, Education-num
Bank Marketing	45211	45k	Age, Balance, Duration
Census Income	299285	65k	Age, Industry Code, Occupation Code

k-mean Clustering algorithm using WEKA data mining tool in MOA framework has been simulated to evaluate the accuracy of proposed privacy preserving approach.

III. RESULTS

Experiments were performed to measure accuracy whereas protecting sensitive data. We here shows two different results, one is corresponding to clustering accuracy in terms of membership matrix that was manually derived from clustering result and another represent corresponding graph for F1_P (Precision) and F1_R (Recall) measures.

B.1. Cluster Membership Matrix

Table 2. Cluster Membership Matrix

	C_1'	C_2'	C_3'	…	C_n'
C_1	**Freq** $_{1,1}$	**Freq** $_{1,2}$	**Freq** $_{1,3}$	…	**Freq** $_{1,n}$
C_2	**Freq** $_{2,1}$	**Freq** $_{2,2}$	**Freq** $_{2,3}$	…	**Freq** $_{2,n}$
C_3	**Freq** $_{3,1}$	**Freq** $_{3,2}$	**Freq** $_{3,3}$	…	**Freq** $_{3,n}$
				…	
C_n	**Freq** $_{n,1}$	**Freq** $_{n,2}$	**Freq** $_{n,3}$	…	**Freq** $_{n,n}$

Using CMM accuracy can be obtained. Table III, shows the percentage of accuracy obtained when selected attributes are perturbed using our algorithm in each datasets.

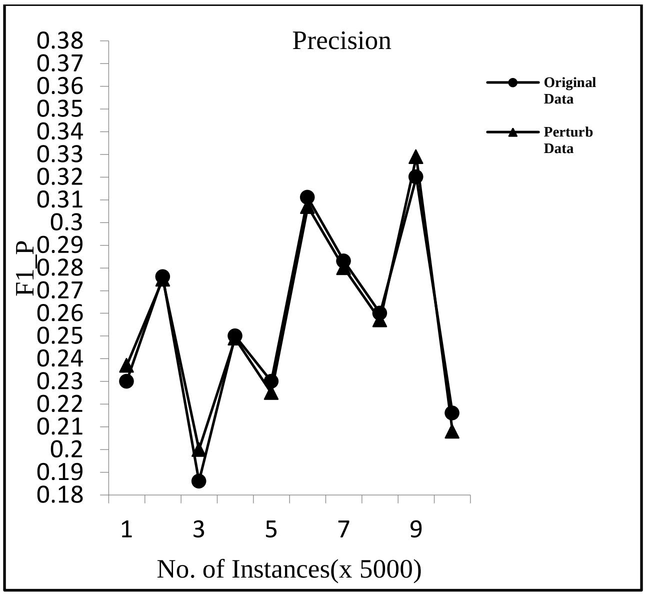
Table 3. Resultant Accuracy Of Clustering For Streaming Data (W=5000)

Dataset	**Attributes**	**k-means (k=5)**	**k-means (k=2)**
Adult	Age	87.16 %	93.82 %
	Fnlwgt	86.32 %	96.44 %
	Education-num	88.43 %	94.29 %
	Age, Fnlwgt, Eduction-num	81.56 %	89.90 %
Bank Marketing	Age	87.75 %	95.25 %
	Balance	90.84 %	93.76 %
	Duration	81.86 %	96.72 %
	Age, Balance, Duration	76.62 %	91.30 %
Census Income	Age	85.42 %	95.19 %
	Industry Code	83.65 %	94.72 %

	Occupation Code	84.87 %	95.42 %
	Age, Industry Code, Occupation	81.67 %	92.43 %

B.2. Precision and Recall Measures

We concentrated on two necessary measures F1_P and F1_R. F1_P determine the precision of system by considering the precision of individual cluster. F1_R determine the recall of system, which take into account the recall of each cluster. Results are represented in terms of graphs for each modified attribute. Each graph contains the measure we obtained when original data is processed without applying privacy preserving method and when data is undergone through our proposed privacy preserving method. k-mean is applied in order to evaluate both cases by keeping number of clusters fix (k=5 and k=2). Instances are processed in defined sliding window size. In Bank Management dataset, Data clustering after Data Perturbation in Age, Income and Duration attributes, Income provide good accuracy compare to Age and Duration. Here we present Precision and Recall graph of both attributes.



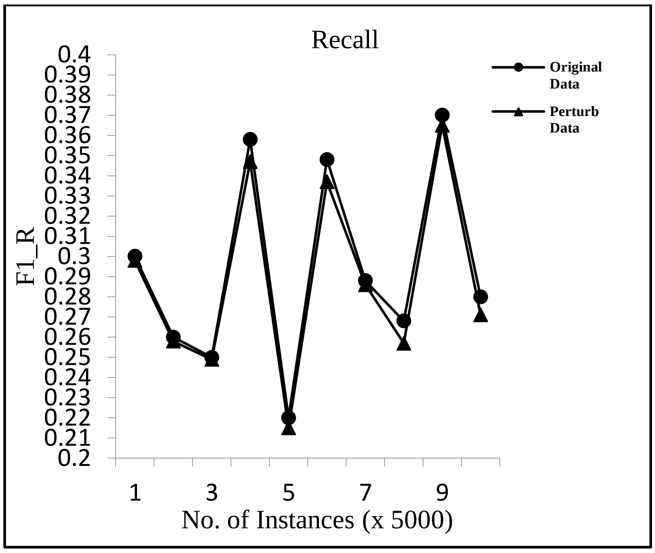


Figure 3. Accuracy on Attribute Age in Bank Marketing Dataset with 2-Cluster

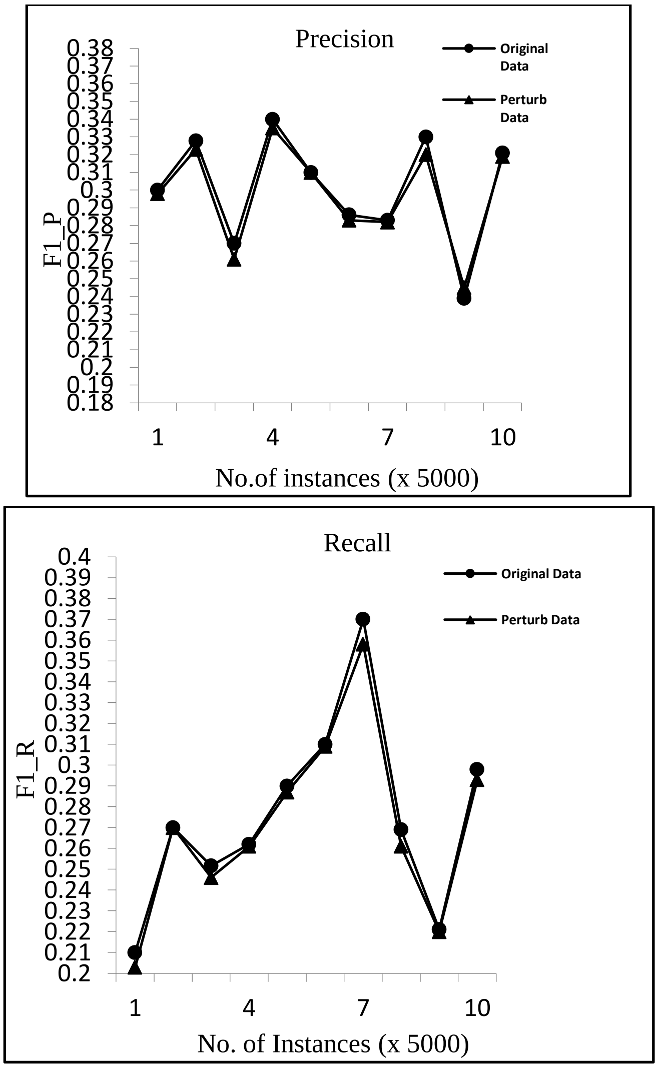


Figure 4. Accuracy on Attribute Income in Bank Marketing Dataset with 2-Cluster

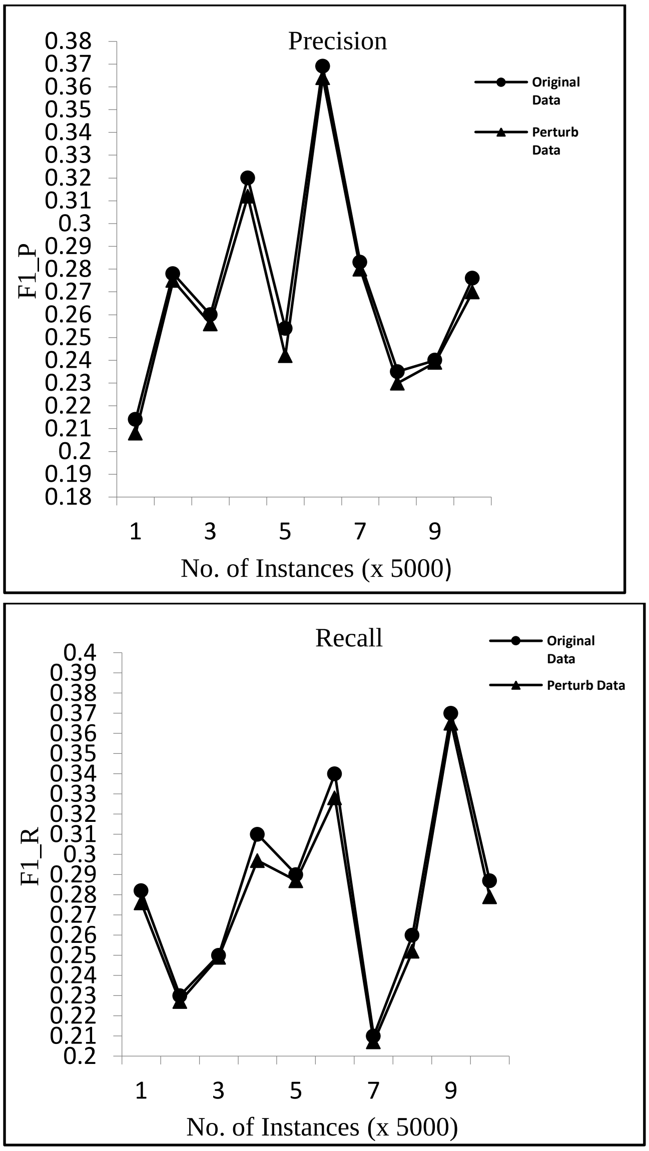


Figure 5. Accuracy on Attribute Duration in Bank Marketing Dataset with 2-Cluster

IV. CONCLUSION

An eventual goal for all data perturbation algorithm is to improve the data transformation process by maximizing both data privacy and data utility achieved. Our approach was motivated by randomization method commonly used in privacy preservation of sensitive data. Proposed approach concentrated on data perturbation by geometric data transformation and noise addition to preserve privacy of sensitive attributes. We extended existing MOA framework in which, each tuple of data stream is independently treated. We considered multiple attribute as sensitive attribute and rest are as non-private attributes. We evaluate the experiment result in terms of correctly classified instance and misclassified instances in clustering process by using k-mean clustering algorithm. Results represent adequately good level of privacy has been achieved with reasonable accuracy against evaluation measures- Precision, Recall and Cluster Membership Matrix. We limited experiments to protect only numeric attributes. Experimental results and result analysis represents that proposed method can preserve data privacy as well as it can also mine data streams accurately.

V. REFERENCES

[1]. Prof. M. Natwaria, S. Arya, "Privacy Preserving Data Mining- "A State of Art", In the Proceeding of the 2016 International Conference on Computing for Suitable Global Development (INDIACom), pp.2108-2112, 2016.

[2]. W.T.Chembian, Dr. J. Janet, "A Survey on Privacy Preserving Approaches and Techniques", In the Proceeding of the International Conference on Information Science and Applications, Chennai, India,pp.700-703, 2010.

[3]. P. Lahane, R. K. Bedi, P. Halgonkar, "Data Stream Mining", International Journal of Advances in Computing and Information Researches, Volume-1, 2012

[4]. L. Golab, M. Tamer Ozsu, "Data Stream Management Issues-A Survey", Technical Report CS-2003-08, 2003.

[5]. M. Kholghi, M. Keyvanpour, "An Analytical Framework for Data Stream Mining Techniques Based on Challenges and Requirement", International Journal of Engineering Science and Technology, Volume-3, pp.2507-2513, 2011.

[6]. M. Khalilian, N. Mustapha, "Data Steam Clustering: Challenges and Issues", In Proceeding of the international Multi

Conference of Engineering and Computer Scientists 2010, Volume-1, 2010.

[7]. N. Gupta, I. Rajput, "Preserving Privacy Using Data Perturbation in Data Stream", International Journal of Advanced Research in Computer Engineering & Technology, Volume-2, pp.1699-1704, 2013.

[8]. A. Patel, K. Dodiya, S. Patel, "A Survey on Geometric Data Perturbation in multiplicative Data Perturbation", International Journal of Research in Advent Technology, pp.603-607, 2013.

[9]. V. S. Verykois, E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin, Y. Theodoridis, "State-of-the-Art in Privacy Preserving data Mining", IEEE, pp.1-5, 2009.

[10]. A. Shah, R. Gulati, "Evaluating Applicability of Perturbation Techniques For Privacy Preserving Data Mining By Descriptive Statistics", In Proceeding of the 2016 Intl. Conference on Advance in Computing, Communications and Informatics, Jaipur, pp.607-613, 2016.

[11]. S. Chidambaram, K. G. Srinivasagam, "A Combined Random Noise Perturbation Approach for Multi Level Privacy Preserving in Data Mining", In Proceeding of the 2014 International Conference on Recent Trends in Information Technology" ,2014.

[12]. H. Li, "Study of Privacy Preserving Data Mining", Third International Symposium on Intelligent Information Technology and Security Informatics, pp.700-703, 2010.

[13]. O. Kale, P. Patel, "A Survey on Privacy Preserving Data Mining", Global Journal of Advanced Engineering Technologies, Volume-2, Issue-3, pp.143-147, 2013.

[14]. Rajesh N., Sujatha K., A. Selvakumar, "Survey on Privacy Preserving Data Mining Techniques using Recent Algorithms", International Journal of Computer Applications, Volume-113, Issue-27, pp. 30-33, 2016.

[15]. K. N. Vachhani, D. B. Vaghela, "Geometric Data Transformation for Privacy Preserving on Data Stream Using Classification", International Journal of Innovative Research in Computer and Communication Engineering, Volume-3, Issue-6, pp.6013-6019, 2015.

[16]. K. Dodiya, S. Yagnik, "Classification Techniques for Geometric Data Perturbation in Multiplicative Data Perturbation", International Journal of Engineering Development and Research, pp.2380-2383, 2014.

[17]. M. Sharma, A. Chaudhary, M. Mathuria, S. Chaudhary, S. Kumar, " An Efficient Approach for Privacy Preserving in Data Mining", In Proceeding of the 2014 International Conference on Signal Propagation and Computer Technology, pp.244-249, 2014.

[18]. H. Chhinkaniwala, A. Patel, S. Garg, "Geometric Transformation Based Multiplicative Data Perturbation for Privacy Preserving Data Mining", IEEE, 2013.

[19]. J. Liu, Y. XU, "Privacy Preserving Clustering by Random Response Method of Geometric Transformation", In Proceeding of the 2014 Fourth International Conference on Internet Computing for Science and Engineering", pp.181-188, 2010.

[20]. A. S. Shanthi, M. Karthikeyan, "A Review on Privacy Preserving Data Mining", IEEE, 2012.

[21]. A. Bifet, G. Holmes and B. Pfahringer, Massive Online Analysis, a Framework for Stream Classification and Clustering. JMLR: Workshop and Conference Proceedings 11, 2010. pp: 44-50.

[22]. MOA datasets, http://moa.cs.waikato.ac.nz/datasets.

[23]. UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/datasets/.