# Privacy Preserving Data Stream Mining Using Hybrid Geometric Data Perturbation

Paresh Solanki[1], Sanjay Garg[2], Hitesh Chhinkaniwala[3]
*[1] Research Scholar, Nirma University, Ahmedabad, India*
*[2] Department of Computer Engineering, Nirma University, Ahmedabad, India*
*[3] Adani Institute of Infrastructure Engineering, Ahmedabad, India*
*(E-mail: pmsolanki@gmail.com)*

***Abstract*—** Privacy preserving on data is crucial process in data mining and data stream mining. Data Perturbation is the well-known technique to preserve the privacy in data mining. There is always tradeoff between privacy preserving and data utility. Better privacy gaining and better utility gain is possible if perturbation is applied on selection basis. Geometric data perturbation on static dataset is one of such technique to use by various data mining model. It is not suitable for data stream or dynamic dataset because of characteristic of data stream. To preserve the sensitive values in perturbation, we have proposed the Hybrid Geometric Data Perturbation in data stream mining. Work is major focus on evaluating maximize the utility after applying Geometric data perturbation using various combination of translation, scaling and rotation. Our experimental work shows that, geometric data perturbation can not only provide the perturbation but for some combination, maximize the privacy with minimum information loss.

***Keywords*—** *Data Mining; Data Stream Mining; Privacy Preserving; Perturbation*

## I. INTRODUCTION

Databases today can range in size into the hundreds of Giga-Bytes. In recent years, data mining as a powerful data analysis tool for analyzing such huge database for getting valuable information and patterns. In the area of statistics or DBMS or Machine learning, retrieve the valuable information from large dataset using various data mining techniques or algorithms [1]. While data stream mining is the method of retrieving data from continuous, speedy information records. Mining information streams is concerned with retrieving data structures shown in patterns in nonstop streams of knowledge. Procedures written for information streams will obviously modify information sizes over and over larger than memory, and may be difficult to tackle by machine learning or data treating. The supposition of data stream processing is that training samples can be fleetingly examined a single time only, because stream reached rapidly, and then must be rejected to make space for succeeding samples. The latest advancement in hardware and software program has enabled the capture of numerous measurements of records in a huge variety of fields. Those measurements are generated continuously and in very excessive fluctuating facts charges [2]. Old-style algorithm is designed for the static dataset. If the data variations, it would be essential to rescan the dataset, which leads to lengthy computation time and inability to promptly respond to the user. In latest centuries, developments in hardware era have facilitated the capability to collect data constantly. Simple transactions of everyday life inclusive of using a credit score card, telephone or surfing the web cause computerized facts garage. Similarly, strengthen in records generation has result in big flows of data across networks. While the extent of the underlying statistics is very massive, it ends in some of computational and mining challenges [2]. Vast number of dataset has crucial information which contains individuals or corporate related sensitive information such as bank account, loan, salary, medical data, customer ID, background information, etc. among these information if it is publicly available then it will be harmful to individual or corporate. In case, if we provide the original data directly to the miners, it will unsurprisingly produce private information disclosure [3]. Because of the area of data mining, privacy disclosure issues becomes infer, triggering the attention of aspects of business and social. So we can say that PPDM is important research field to provide the privacy during the process of mining. Presently, there are several methods of PPDM is exist but they are appropriate only for static dataset. There has been less work done in privacy preserving in data stream mining (PPDSM). Provide the PPDSM is severe concerns. The goal of PPDM and PPDSM is to reduce the risk of misuse of data and at the same time maximize information gain results same as that created in the nonappearance of such privacy preserving methods. Usually when individuals talk of privacy, they say “keep information about me from being available to others”. Though, their actual alarm is that their information not be abused. The panic is that once information is out, it will be difficult to prevent abuse. For these, we need resolutions that promise sensitive data will not be out.

The rest of the paper proceeds as follows: In section 2, we have discussed different studies performed in the area of PPDM and PPDSM. In section 3, we have discussed Preliminaries of proposed algorithm. In section 4, we have discussed Problem Definition, Proposed Framework, and Proposed Algorithm. In section 5 we have shown Results, Observation and Performance analysis. Finally, we conclude in section 6.

## II. LITERATURE SURVEY

The research in PPDM extents many fields: Reconstruction, Heuristic and cryptography based techniques. In this section, we have discussed the data perturbation methods because they are more thoroughly related to our field of research. Current work in the area of Privacy preserving in data mining or privacy preserving in data stream mining has dedicated greatly determination to regulate a trade-off among privacy and privacy, which is critical in order to expand decision-making processes and other human activities. PPDM or PPDSM is worked on sensitive data for hiding sensitive knowledge and values. Sensitive knowledge or attributes are hiding via data heuristic, data modification and data cryptography methodology. These methodologies are varying from investigator to investigator since few investigators may consider merely certain attribute value should private and some may consider complete data pillar should be private. The present PPDM procedures in the literature can be categorized into three key groups: reconstruction method, heuristic method and cryptographic method [4].

Reconstruction based methods are produced privacy aware dataset by removing sensitive features from the original dataset. These methods are generating minor side-effects in dataset comparatively heuristic approach. Reconstruction based methods modify the original data to accomplish privacy preserving. The perturbed or modified data would meet the two situations. Firstly, it is not possible to retrieve original data back from the distortion data. Secondly, the distorted data is quiet to keep some statistical properties of the original data that means, information retrieved from the distorted data are same to data acquired from the original information [4]. Heuristic based methods like adaptive alteration that changes only certain values that maximize the utility gain instead of all available value [4]. There are various methods have been developed to modify choosy data for data mining methods like association, classification and clustering mining. Choosy data modification based mining is NP-hard problem and because of this, heuristic can be used to address to complexity problems. Data distortion techniques are tried to hide association rules by increasing or decreasing confidence and support. So they can produce complexity issue but they produce unnecessary side effects in database which lead to optimal solution to them. Data blocking techniques are changing the data by unknown “#” on selected transaction in place of inserting or deleting items. So it is difficult for any investigator to know the value behind “#”. In a distributed location the key issue for accomplishing privacy preserving is the security of infrastructures and encryption technology. Thus, privacy preserving based on data encryption technology usually applies on distributed application. Cryptography based methods provides clear defined model for privacy, which includes methods for demonstrating and measuring it.

Cryptography based techniques has more time complexity compare to others [4]. Data perturbation is used to provide statistical information without revealing sensitive data about individuals. In data perturbation method, the original value of the data is perturbed into a random value. The perturbation techniques employed until now consist of data replacement [5], [6], [7], data swapping [8], [9], additive value distortion [10], [11], [12], and multiplicative value distortion [13], [14], random perturbation on categorical or Boolean data [15], [16], [17], matrix multiplication [18], data shuffling [19], blocking [20] etc. k-anonymity [21], [22], [23] and sensitive rule hiding [24], [25], [26] have also been employed in PPDM. Author in [27] proposed plan for privacy preserving in data mining by joining straight data distribution and upright data distribution on dataset. This method does not permit data owners to select their preferred privacy levels. Author in [28] proposed a threshold algorithm which uses a threshold to classify a record by calculating its possibility. Without reconstructing the original data distribution, user can mine the data directly from the perturbed dataset. In [29], author proposed a noise addition plan in which create a decision tree by discovering the original data, and then each record, add a noise to get the modified data which needs to be accustomed according to decision tree. This method is harmful enough, because it can be attacked by some attack methods such as PCA [32], SF [30] and SVD [31]. In the field of matrix multiplicative perturbation, distance based privacy preserving [33], [34], [35] has gain a lot of consideration since it sureties superior accuracy. The distorted data is used as input for many significant data mining procedures, such as k-mean [36], k-nearest neighbor [37] and distance based clustering [38], and the equivalent output is accurately as similar as the outcome of examining the original data. However the security issue of how much the privacy loss has caused investigators concern. Author in [39], [40] shows that how well an attacker can recover the original data from the distorted data and past information. Random noise which is add/multiplying can be removed by the filtering attacks. Authors [41] demonstrated that the random noise preserves a very little amount of data privacy because most of noise can be removed if its variance is not large enough. For preventing the filtering attacks, [42] and [43] have exploited data correlations in generating the noise. Authors in [42] used the Principal Component Analysis (PCA) and the Bayes Estimate (BE) for data reconstruction and noise generation based on data correlations. These techniques are based on the idea that the random noise becomes difficult to be filtered from the original data if it is concentrated on principal components only. Based on the idea, they guarantee that the noise is concentrated on the principal components by making the correlations of the random noise similar to the correlations of the original data. Author in [44] proposed perturbation method which is work on categorical attributes. [45] Indicated the γ-amplification system to bind the volume of privacy breach in the categorical data sets. The main challenge of data perturbation is leveling privacy protection and data quality, which are generally measured as a pair of contradictive aspects [46].

In recent years, however, it is noticed that the perturbed or distorted datasets from certain data perturbation techniques may not be safe if an attacker has some background information about the original datasets [47],[48],[49],[50],[51]. In practice, it is unlikely that an attacker has no idea about the original dataset other than the public perturbed version. The common sense, statistical measure, reference, and even a small amount of leakage may dramatically help the attacker weaken the privacy of the dataset. [52] Indicated that it is extremely possible to differentiate the original true values from the additively perturbed data. [53],[54] calculated a useful upper bound and lower bound about the dissimilarity between the original dataset and the estimated dataset which is computed from the perturbed dataset by spectral filtering techniques. [55] Presented that, in the data additive perturbation, the privacy is susceptible from a known public dataset in a high dimensional space.

## III. PRELIMINARIES

In this section, for better understanding of the paper, we have defined the components of geometric perturbations.

### A. *Geometric Data Perturbation (G(D))*

We define a geometric perturbation as the group of the process on G(D) using random rotation perturbation say “R”, random translation perturbation say “T”, and noise addition say “N”. For each attribute of *G***(**D**),** the value is generated through *G***(**D**) =** R * X + T + N.

### B. *Scaling Data Perturbation (SDP)*

Scaling is the transformation where in x and y axis are stable straight line, the origin “O” is a stable point, dissimilar from the origin of each point measures along the axis. Scaling is done by multiplying a constant to all the values of an attribute [56]. The constant can be plus or minus. SDP process is works as following way.

```
SDP ( )
For each Aj ϵ D do (Aj = Sensitive attribute, D = database)
        N = Ej (Ej = noise for attribute Aj
        Zj ← {Mult}
For each Vi ϵ D do
For each Aj ϵ Vi = (A1, A2, A3… Ad) do (Aj = the observation of the j-th attribute)
A'j ← Transform (Aj, Zj, Ej)
End
```

### C. *Translation Data Perturbation (TDP)*

Translation is added the constant value to whole value of an attribute. The constant can be a plus or minus value. This transformation can achieve the privacy on crucial data value [56]. TDP works as following way.

```
TDP ( )
For each $A_j$ ϵ D do ($A_j$ = Sensitive attribute, D = database)
      N = $E_j$ ($E_j$ = noise for attribute $A_j$
      $Z_j$ ← {Add}
For each $V_i$ ϵ D do
For each $A_j$ ϵ $V_i$ = (A1, A2, A3… Ad) do ($A_j$ = the observation of the j-th attribute)
        $A'_j$ ← Transform ($A_j$, $Z_j$, $E_j$)
End
```

### D. *Rotation Data Perturbation (RDP)*

Arbitrarily choose the pair of attributes from dataset and rotate them according to a given angle $\theta$ with the origin as the center. If $\theta$ is positive, we rotate anti-clockwise otherwise, rotate the clockwise. Rotation can be accomplished by multiplying the matrix [56]. RDP works as following way.

```
RDP ( )
k ← |n/2| //where ‘n’ is columns
$P_k$ ← k Pairs ($A_i$,$A_j$) in D such that 1≤I, j≤n and i != j
For each selected pair $P_k$ in Pairs (D) do
V ($A'_i$,$A'_j$) ← $R\theta$ * V($A_i$,$A_j$) //V is computed as a function of $\theta$
        $\theta_k$ ← SecurityRange value of $\theta_k$ (30, 45, 60, 90)
        V ($A'_i$,$A'_j$) ← $R\theta_k$ * V($A_i$,$A_j$) //Output the distorted attributes of D'
End for
End
```

## IV. PROPOSED WORK

In this section, we present the Problem Definition, framework of our proposed approach and methodology of our work.

### A. *Problem Definition*

Here, we accept that the input consists of several nonstop streams. Without loss of generality, we may undertake that each tuple contains of a single attribute. For the purposes of our examination and without loss of generality, the input contains of N data streams indicated as D^1,D^2,D^3,D^4,….D^N . For any i^th data stream D^i, its value at the time t isD_t^i. The stream collection is printed as D=[D^i for 1≤i≤N]. Formally, the stream collection D can be considered as a T × N matrix where N is the number of streams and T is the present timestamp, which grows indefinitely. Apply the Hybrid Geometric Data Perturbation approach on data streams. Our objective is to provide privacy before release of data streams. Perturbed data streams should generate identical result as of original data stream. To preserve privacy from available data stream, online generated noise can be addition, multiplication and rotation. Next, mine perturbed data streams to construct a clustering model and evaluate the clustering measures.

### B. *Proposed Framework*

Figure 4.1 shows the framework of our proposed work. The major objective of our work is to transform the real dataset “D” into modified dataset “ D’ ” which is achieve the privacy on sensitive data and preserve the maximum information knowledge for the intended data analysis using data mining methods. We can also compare the features of both real and modified dataset in terms of information loss, privacy gain and response time therefore acquire improved accuracy of different data stream procedures against each other. Fig. 1 Framework of Privacy Preserving in Data Stream Mining using Hybrid Geometric Data Perturbation

### C. *Methodology*

This section provides detail about proposed algorithms that preserve privacy and balance trade off with data utility. Dataset statistical characteristics have been studied before applying

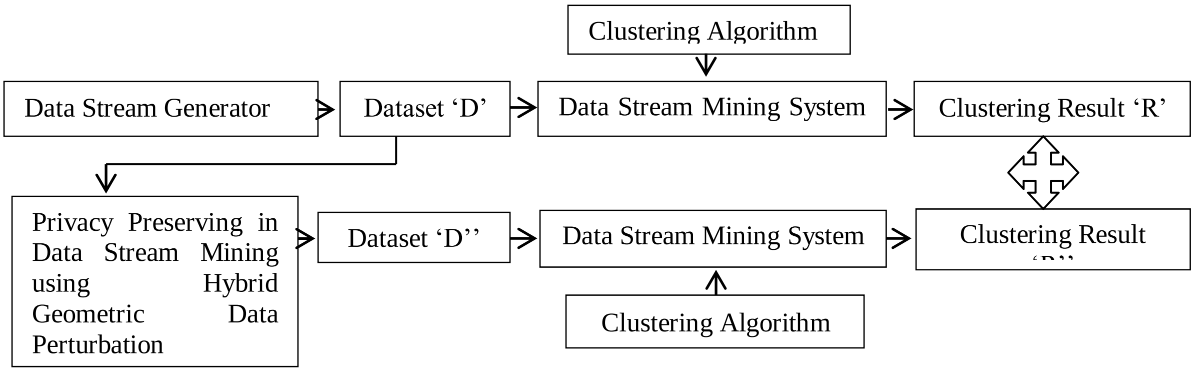

Clustering Algorithm
Data Stream Generator
Dataset 'D'
Data Stream Mining System
Clustering Result 'R'
Privacy Preserving in Data Stream Mining using Hybrid Geometric Data Perturbation
Dataset 'D''
Data Stream Mining System
Clustering Result
Clustering Algorithm

Fig. 1 Framework of Privacy Preserving in Data Stream Mining using Hybrid Geometric Data Perturbation

perturbation and values have been modified keeping such characteristics intact. The key characteristics of our method is, it is simple and easy to implement, lower complexity, required no deep mathematical calculation. Time complexity is directly proportional to no. of instances to be processed. The process is divided into two stages, which are a) Pre-process on data stream b) Data stream analysis using clustering. The primary objective of the first stage, which is controlled by the data streams preprocessing system, is to perturb data streams to protect data privacy. The primary objective of the second stage, which is controlled by the online data mining system, is to mine perturbed data streams to cluster the data. Algorithm steps are given below.

**Algorithm:** Privacy Preserving in Data Stream Mining using Hybrid Geometric Data Perturbation.
**Input:** Data Stream D and define sliding window size 'W'.
**Intermediate Result:** Perturb Data Set D'.
**Output:** Clustering Result R and R' of D and D'.

```
Steps:
For data set D
    Set SA[i]   //store sensitive attribute values in array
End for
For each instance of Dataset D
        For i=0 to W
        SA[i] = TDP() // Translation Data Perturbation
        SA[i] = SDP() // Scaling Data Perturbation
        SA[i] = RDP() // Rotation Data Perturbation
For each step
        Noise (N) = AVG (Sensitive Attribute)
End For
End For
                D' = store 'D'
End For
    Clustering (D') and Clustering (D) // Apply k-Mean clustering algorithm
```

Our main goal is to preserve privacy in Data Stream using Hybrid Geometric Data Perturbation with 1. Minimize information loss 2. More response time to construct a clustering model 3. Maximize privacy gain. 4. Maintain the accuracy of the clustering model. Our primarily focused on translation, scaling and rotation perturbation. Among them rotation data perturbation are most complicated to implement. In Hybrid Model, we are mixing all three transformations together and make a hybrid method call hybrid Geometric data perturbation. Here the operation which we perform translation (T), scaling (S) and rotation (R) to make the method as hybrid. We have applied all these transformations in random order. In our proposed approach, we are applied all six combination such as TSR, TRS, STR, SRT, RTS and RST. Among of them which of the combination give us best privacy, which is we take it granted as final output in Hybrid algorithm. Here we also add some security using rotation data perturbation method of angle of rotation. We experience that if we change the angle of rotation then privacy may be increase or decrease on based on dataset.

## V. PERFORMANCE EVALUATION

### *A. Dataset Information*

The experiments are processed on five different data sets obtainable from the UCI Machine Learning Repository [22], MOA dataset dictionary [23] and artificial data that we have created. We constrained our experiment to numeric attributes, even we can extend the implementation to categorical attributes. We performed our experiments on *Covertype, Electricity, Agrawal, Bank Marketing, and Airlines* dataset. Details about these data sets are found in table I.

TABLE I. DETAILS ABOUT DATASET

| Dataset | Covertype | Electricity | Agrawal | Bank Marketing | Airlines |
|---|---|---|---|---|---|
| Source | UCI Machine Learning Repository[57] | MOA Dataset [58] | Synthetic Dataset (WEKA) | UCI Machine Learning Repository[59] | MOA Dataset [58] |

| | | | | | |
|---|---|---|---|---|---|
| Instances | 5,81,012 | 45,312 | 50,000 | 45,211 | 5,39,383 |
| Attributes | 54 | 9 | 10 | 17 | 8 |
| Type of Attributes | Categorical, Integer | Numeric | Numeric | Real, Nominal | Numeric, Nominal |
| Characterist ic | Normalized | Normalized | - | - | - |
| Class | {1,2,3,4,5,6,7} | {UP, DOWN} | {0,1} | {No, Yes} | {0,1} |
| instances (processed) | 65000 | 45000 | 50000 | 45000 | 65000 |
| Sensitive Attributes | Elevation, Aspect, Slope | Nswprice, Nswdemand | Salary, Age | Age, Balance | Flight |

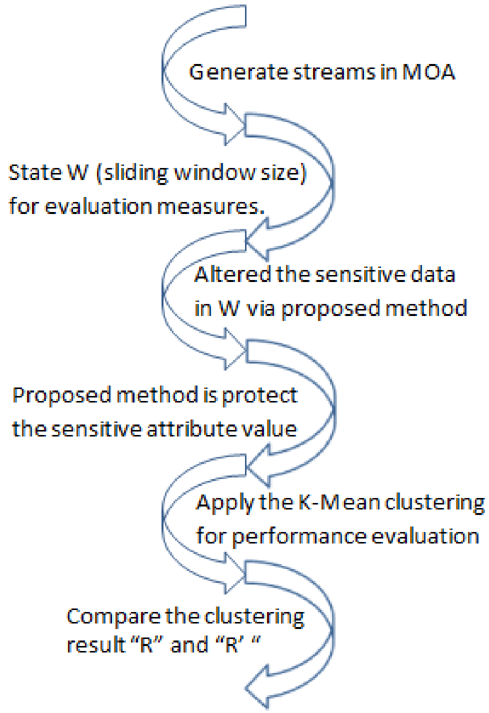

Generate streams in MOA
State W (sliding window size) for evaluation measures.
Altered the sensitive data in W via proposed method
Proposed method is protect the sensitive attribute value
Apply the K-Mean clustering for performance evaluation
Compare the clustering result "R" and "R' "

Fig. 2 Flow of Experiment in MOA Framework

### B. *Experiment Result and Analysis*

Proposed approach is divided into two stages; 1. Data stream preprocessing 2. Data stream cluster mining, respectively. In the first stage of data stream preprocessing, upon receiving data stream from data stream generator or real dataset, the Data Stream Preprocessing System (DSPS) uses perturbation algorithm to perturb confidential data. We can flexibly adjust the data attribute to be perturbed according to the security need. Therefore, threats and risks from releasing data can be effectively reduced. In the second stage of data stream mining, the online data mining system uses the sliding window mechanism to cluster perturbed data streams. Experimental results show that proposed approach not only can preserve data privacy but also can mine data stream accurately. Proposed approach has been implemented in Java and integrated with MOA framework. Experiments are performed based on sliding window size (w) concept in order to estimate the clustering accuracy. Presented work dedicated on the whole quality of produced clusters after dataset perturbation. Experimental steps are shown in Fig. 2.

Dataset D is given as an input to proposed data perturbation algorithm. Algorithm perturbs only sensitive attribute values and resultant dataset with modified values is called perturbed dataset D'. D and D' are provided to standard clustering stream learning algorithms to obtain results R and R' respectively. Proposed work focuses on obtaining close approximation between clustering results R and R' to balance in between privacy improvement and information damage. The primary objective of the second stage, which is handled by the online data mining system, is to mine perturbed data streams to cluster the data. *K-Mean* clustering algorithm over predefined sliding window size on perturbed data stream has been used in order to measure the exactness and usefulness of clustering outcomes over five different ordinary datasets. Outcomes show properly best level of privacy has been accomplished with reasonable accuracy in almost all tested cases. Accuracy between original dataset and perturbed dataset has been quantified by percentage of instances assigned to different clusters with the help of cluster membership matrix (CMM). Proposed approach shows reasonably good results against evaluation measures Precision, Recall, Misclassification and CMM (*Cluster Membership Matrix*). With the help of CMM, We matched how closely each cluster in the perturbed dataset equals its equivalent cluster in the original Dataset. We mention to such a matrix as the Clustering Membership Matrix (CMM) where the rows shows the clusters in the original dataset, the columns represent the clusters in the perturbed dataset. We constructed each dataset to define five clusters using *k-Mean* clustering procedure. Each Matrix demonstrating five clusters situation for real dataset and perturb dataset. Real dataset clustering outcomes gives information about number of occurrences are actual classified in each cluster whereas perturb dataset clustering showing outcome of correct assignments after data perturbation and percentage of accuracy accomplished. Following Table II to IV shows the some of the Membership Matrix with best Information Gain (Accuracy) after performing Geometric Perturbation on various dataset.

TABLE II. Information Gain for Covertype Dataset (w=1000, Angle = 30°, Sequence = TDP_SDP_RDP)

| Clusters | Original Dataset | Perturbed Dataset | Information Gain (%) |
|---|---|---|---|
| | No. of Instances/Cluster | No. of Instances/ Correctly Clustered | |
| C1 | 14240 | 12548 | 88.12 |
| C2 | 13440 | 11465 | 85.30 |
| C3 | 12691 | 11610 | 91.48 |
| C4 | 11788 | 09761 | 82.80 |
| C5 | 12841 | 11229 | 87.44 |
| Total | 65000 | 56613 | 87.02 |

TABLE III. Information Gain for Covertype Dataset (w=3000, Angle = 30°, Sequence = TDP_SDP_RDP)

| Clusters | Original Dataset | Perturbed Dataset | Information Gain (%) |
|---|---|---|---|
| | No. of Instances per Cluster | No. of Instances per Correctly Clustered | |
| C1 | 14984 | 12683 | 84.64 |
| C2 | 13803 | 11776 | 85.31 |
| C3 | 10710 | 09621 | 89.83 |
| C4 | 11776 | 09659 | 82.02 |
| C5 | 11727 | 10870 | 92.69 |
| Total | 63000 | 54609 | 86.68 |

TABLE IV. Information Gain for Covertype Dataset (w=5000, Angle = 60°, Sequence = RDP_SDP_TDP)

| | Original Dataset Clustering | Perturbed Dataset Clustering | Information Gain (%) |
|---|---|---|---|
| Clusters | No. of Instances per Cluster | No. of Instances per Correctly Clustered | |
| C1 | 16196 | 13422 | 82.87 |
| C2 | 12441 | 11938 | 95.96 |
| C3 | 12376 | 11084 | 89.56 |
| C4 | 11720 | 09019 | 76.95 |
| C5 | 12267 | 08515 | 69.41 |
| Total | 65000 | 53978 | 83.04 |

TABLE V. Information Gain (Covertype Dataset)

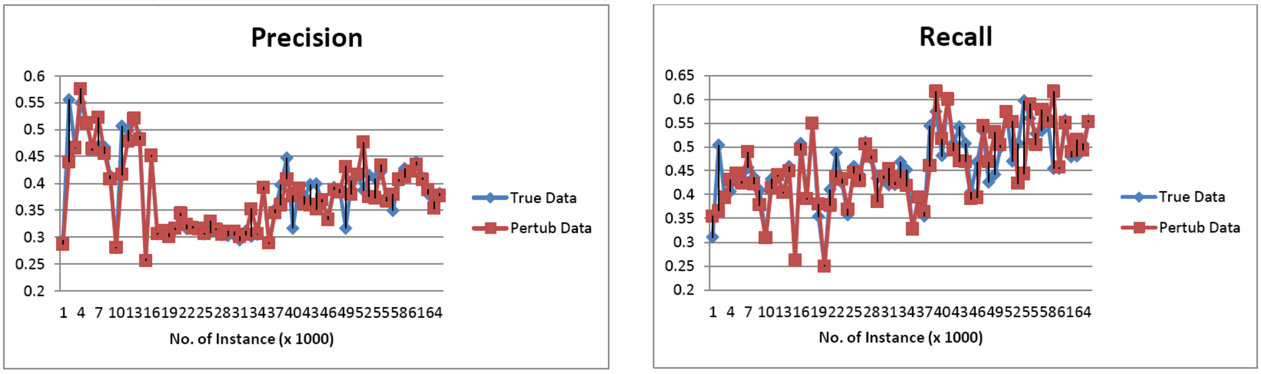
| Dataset | Sliding Window Size (w) | Angle | TSR (%) | TRS (%) | STR (%) | SRT (%) | RTS (%) | RST (%) |
|---|---|---|---|---|---|---|---|---|
| Covertype | 1000 | 30° | 87.09 | 87.09 | 87.26 | 81.84 | 82.1 | 81.84 |
| | | 45° | 83.8 | 83.8 | 83.53 | 82.82 | 82.69 | 82.82 |
| | | 60° | 82.75 | 82.75 | 82.82 | 82.9 | 82.9 | 82.91 |
| | | 90° | 81.92 | 81.92 | 81.35 | 84.86 | 83.84 | 84.86 |
| | 3000 | 30° | 86.68 | 86.68 | 89.35 | 83.93 | 82.81 | 83.93 |
| | | 45° | 82.73 | 82.73 | 82.79 | 83.65 | 82.9 | 83.65 |
| | | 60° | 82.89 | 82.89 | 82.93 | 84.07 | 83.56 | 84.07 |
| | | 90° | 83.43 | 83.43 | 81.79 | 84.42 | 82.85 | 84.42 |
| | 5000 | 30° | 82.55 | 82.55 | 82.64 | 82.65 | 81.78 | 82.65 |
| | | 45° | 81.12 | 81.12 | 81.29 | 82.04 | 82.15 | 82.04 |
| | | 60° | 82.12 | 82.12 | 81.59 | 83.04 | 82.87 | 83.04 |
| | | 90° | 79.16 | 79.16 | 80.25 | 78.331 | 78.39 | 78.31 |

Table V shows the Information gain after applied our proposed approach on Covertype perturbed Dataset. Table VI shows the best suitable transformation order which will give highest information gain among all other transformation orders. We have applied our proposed approach on five different five different datasets with different Sliding window size and rotation angle. We can also conclude that TSR (Translation, Scaling and Rotation) based Geometric Process is most suitable to achieve maximum privacy with minimum information loss.

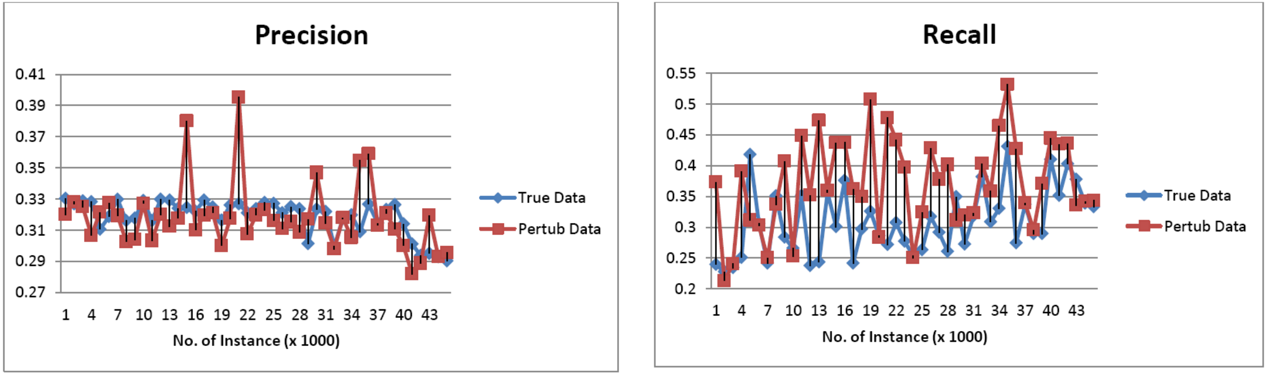
TABLE VI. Best suitable transformation order of Hybrid Geometric Perturbation based Algorithm

| W=1000 | Covertype | Electricity | Agrawal | Bank Marketing | Airlines |
|---|---|---|---|---|---|
| 30° | STR | STR | TSR | STR | TSR |
| 45° | RTS | TSR | RTS | TSR | RST |
| 60° | TSR | RST | RTS | TSR | TSR |
| 90° | RST | RST | RTS | STR | RST |
| W=3000 | | | | | |
| 30° | TSR | STR | TSR | TSR | TSR |
| 45° | STR | SRT | TSR | TSR | TSR |
| 60° | STR | SRT | RTS | TSR | TSR |
| 90° | STR | SRT | RTS | RTS | TSR |
| W=5000 | | | | | |
| 30° | SRT | SRT | TSR | TSR | TSR |
| 45° | TSR | RTS | RTS | STR | TSR |
| 60° | TSR | SRT | RTS | TSR | TSR |
| 90° | RST | STR | RTS | RST | RTS |

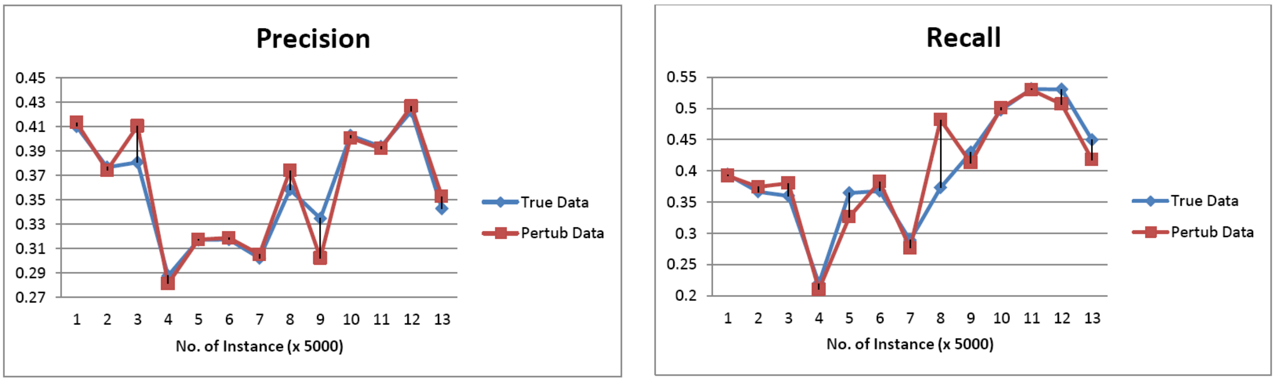
In our proposed work, analysis of Accuracy is evaluating the clustering measures with the help of MOA framework. We concentrated on two essential measures F1_P (determine the precision of system by considering the precision of individual cluster) and F1_R (determine the recall of system, which take into account the recall of each cluster). Results are presented in terms of graphs where 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. Occurrences are treated in defined sliding window size. Number of Result Graph is generated but here we are presented few of them which give best case result. Fig. 3 shows the Precision and Recall graph which is measure the accuracy.



Accuracy on Covertype Dataset (w=1000, Angle=30, Sequence=SDP_TDP_RDP)



Accuracy on Bank Dataset (w=1000, Angle=60, Sequence=TDP_SDP_RDP)



Accuracy on Covertype Dataset (w=5000, Angle=60, Sequence=SDP_RDP_TDP)
Fig. 3 Accuracy Measures using precision and Recall in MOA Framework

### C. *Quantifying Privacy*

Perturbation methods guarantee that disclosure will not occur. Revealing the data from perturbation is depends on algorithm strength. We have found that if sensitive attributes independently perturb than it is possible that complete or partially disclosure the data. Most of the reconstruction perturbation based methods are disclose if you are having the knowledge about methods. It is therefore necessary to measure the level of security provided by a specific perturbation technique when quantifying privacy by such a method. Traditionally, the privacy provided by a perturbation technique has been measured as the variance between the actual and the perturbed values. This measure is given by *Variance(X-Y)* where *X* represents a single original attribute and *Y* the distorted attribute. This measure can be made scale invariant with respect to the variance of *X* by expressing *Privacy Level = Variance(X-Y)/ Variance(X).*Clearly, the above measure to quantify privacy is based on how closely the original values of a modified attribute can be estimated. Our proposed approach maintain the variance level such that it may be not possible that adversary can disclose the original data back from perturb data.

## VI. PERFORMANCE EVALUATION

Data mining or stream mining is crucial analysis technique for business as a part of life. Mining techniques are used in various applications. Recent surveys show that privacy is major concern during the process of data mining. Most approaches towards solving the problem of data privacy are based on perturbation. In this study, we propose the Hybrid Geometric based data perturbation on stream data. We also find out the best suitable transformation order which will give highest information gain among all other transformation orders. In our proposed approach, accuracy of privacy depends on security angle, sequence of translation, scaling and rotation. In future, we can extend our work to nominal type attributes.

## REFERENCES

[1] Jiawei Han, Jian Pei and Micheline Kamber, "*Data Mining: Concepts and Techniques*", Third Edition, The Morgan Kaufmann Series in Data Management Systems Elsevier, 2012.

[2] Prashant Lahane, R K Bedi and Prasad Halgaonkar, "*Data Stream Mining*", International Journal of Advances in Computing and Information Researches, January 2012

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

[4] V.S. Verykios, E. Bertino, I. N. Fovino, L. P. Provonza, Y.Saygin and Y. Theodoridis. State-of-the-art in privacy preserving data mining. *SIGMOD Record*, 33 (1): 50-57, 2004.

[5] N. Adam and J. Worthmann, "Security-Control Methods for Statistical Databases: A Comparative Study," ACM Computing Surveys, vol. 21, no. 4, pp. 515-556, 1989.

[6] C.K. Liew, U.J. Choi, and C.J. Liew, "A Data Distortion by Probability Distribution," ACM Trans. Database Systems, vol. 10, no. 3, pp. 395-411, 1985.

[7] E. Lefons, A. Silvestri, and F. Tangorra, "An Analytic Approach to Statistical Databases," Proc. Ninth Int'l Conf. Very Large Data Bases (VLDB), 1983.

[8] T. Dalenius and S.P. Reiss, "Data-Swapping: A Technique for Disclosure Control," J. Statistical Planning and Inference, vol. 6, pp. 73-85, 1982.

[9] S.E. Fienberg and J. McIntyre, "Data Swapping: Variations on a Theme by Dalenius and Reiss," Proc. Privacy in Statistical Databases, pp. 14-29, 2004.

[10] R. Agrawal and R. Srikant, "Privacy-Preserving Data Mining," Proc. 2000 ACM SIGMOD Conf. Management of Data, pp. 439-450, 2000.

[11] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar, "On the Privacy Preserving Properties of Random Data Perturbation Techniques," Proc. Third IEEE Int'l Conf. Data Mining (ICDM '03), pp. 99-106, 2003.

[12] Z. Huang, W. Du, and B. Chen, "Deriving Private Information from Randomized Data," Proc. ACM SIGMOD Int'l Conf. Management of Data, pp. 37-48, 2005.

[13] S.R.M. Oliveira and O.R. Zaı¨ane, "A Privacy-Preserving Clustering Approach Toward Secure and Effective Data Analysis for Business Collaboration," Computers and Security, vol. 26, no. 1, pp. 81-93, 2007.

[14] K. Liu, H. Kargupta, and J. Ryan, "Random Projection-Based Multiplicative Data Perturbation for Privacy Preserving Distributed Data Mining," IEEE Trans. Knowledge and Data Eng., vol. 18, no. 1, pp. 92-106, Jan. 2006.

[15] A. Evfimievski, J. Gehrke, and R. Srikant, "Limiting Privacy Breaches in Privacy Preserving Data Mining," Proc. 22nd ACM Symp. Principles of Database Systems (PODS '03), pp. 211-222, 2003.

[16] S. Rizvi and J. Haritsa, "Maintaining Data Privacy in Association Rule Mining," Proc. 28th Int'l Conf. Very Large Databases (VLDB), Aug. 2002.

[17] S. Agrawal and J.R. Haritsa, "A Framework for High-Accuracy Privacy-Preserving Mining," Proc. 21st Int'l Conf. Data Eng. (ICDE '05), pp. 193-204, 2005.

[18] Mohammad, A. K. and Somayajulu, D.V.L.N., Privacy preserving technique for Euclidean distance based mining algorithms using a wavelet related transform. In *Proceedings of the 11th International Conference on Intelligent Data Engineering and Automated Learning. Paisley*, United Kingdom, 202-209, 2010.

[19] Muralidhar, K. and Sarathy, R., Data shuffing-a new masking approach for numerical data. Management Science, 52(5), 658-670, 2006.

[20] Agrawal, S., Krishnan, V. and Haritsa, J.R., On addressing efficiency concerns in privacy-preserving mining. In *Proceedings of the 9th International Conference on Database Systems for Advanced Applications*, 113–124, 2004.

[21] L. Sweeney, “K-Anonymity: A Model for Protecting Privacy,” Int’lJ. Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 5, pp. 557-570, 2002.

[22] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, “L-Diversity: Privacy Beyond K-Anonymity,” Proc. 22nd IEEE Int’l Conf. Data Eng. (ICDE ’06), p. 24, 2006.

[23] N. Li, T. Li, and S. Venkatasubramanian, “*T*-Closeness: Privacy Beyond K-Anonymity and L-Diversity,” Proc. IEEE 23rd Int’l Conf. Data Eng. (ICDE ’07), pp. 106-115, 2007.

[24] M. Atallah, E. Bertino, A. Elmagarmid, M. Ibrahim, and V. Verykios, “Disclosure Limitation of Sensitive Rules,” Proc. Workshop Knowledge and Data Eng. Exchange (KDEX ’99), pp. 45-52, 1999.

[25] Y. Saygin, V.S. Verykios, and C. Clifton, “Using Unknowns to Prevent Discovery of Association Rules,” ACM SIGMOD Record, vol. 30, no. 4, pp. 45-54, 2001.

[26] V. Verykios, A. Elmagarmid, B. Elisa, D. Elena, Y. Saygin, and E. Dasseni, “Association Rule Hiding,” IEEE Trans. Knowledge and Data Eng., vol. 16, no. 4, pp. 434-447, Apr. 2004.

[27] Khatri, A., Kabra, S. and Singh, S., Architecture for Preserving Privacy During Data Mining by Hybridization of Partitioning on Medical Data, 93-97, 2010.

[28] Liu, L., Kantarcioglu, M. and Thuraisingham, B., Privacy Preserving Decision Tree mining from Perturbed Data. *In proceedings of the 42th Hawaii International Conference on System Sciences*, 2009.

[29] Mohammad, A. K. and Somayajulu, D.V.L.N., A Noise Addition Scheme in Decision Tree for Privacy Preserving Data Mining. *Journal of Computing*, 2(1), 2151-9617, 2010.

[30] Kargupta, H., Datta, S., Wang, Q. and Sivakumar, K., on the privacy preserving properties of random data perturbation techniques. In *Proceeding of the IEEE International Conference on Data Mining*. Melbourne, FL, 99-106, 2003.

[31] Guo, S., Wu, X. and Li, Y., On the lower bound of reconstruction error for spectral filtering based privacy preserving data mining. In *Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases*. Berlin, Germany, 520-527, 2006.

[32] Huang, Z., Du, W. and Chen, B., Deriving private information from randomized data. In *Proceedings of the 2005 ACM SIGMOD Conference. Baltimore*, MD, 37-48, 2005.

[33] Yang, W. J., Privacy protection by matrix transformation. *IEICE Transactions on Information and Systems*, E92-D (4), 740-741, 2009.

[34] Chen, K. and Liu, L., Privacy preserving data classification with rotation perturbation. In *Proceedings of the Fifth IEEE International Conference on Data Mining*. Houston, TX, 589-592, 2005.

[35] Liu, K., Giannella, C. and Kargupta, H. An Attackers View of Distance Preserving Maps for Privacy Preserving Data Mining. In *the Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases*. Berlin, Germany, 297-308, 2006.

[36] Su, C. H., Zhan, J. and Sakurai. K., Importance of Data Standardization in Privacy-Preserving K-Means Clustering. In the *Proceedings of International Workshops on Database Systems for Advanced Applications*. Brisbane, QLD, Australia, 276-286, 2009.

[37] Chong, Z. H, Ni, W. W., Liu, T. T. and Zhang, Y., A privacy-preserving data publishing algorithm for clustering application. *Computer Research and Development*, 47(12), 2083-2089, 2010.

[38] Raaele Giancarlo, Giosue Lo Bosco, Luca Pinello., Distance functions, clustering algorithms and microarray data analysis. In *Proceedings of the 4th International Conference on Learning and Intelligent Optimization. Venice*, Italy, 125-138, 2010.

[39] Liu, K. Kargupta, H. and Ryan, J., Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. *IEEE Transactions on knowledge and Data Engineering*, 18(1), 92-106, 2006.

[40] Liu, K., Giannella, C. and Kargupta, H., A survey of Attack Techniques on Privacy-Preserving Data Perturbation Methods. In: *Privacy-Preserving Data Mining: Models and Algorithms, 2008*.

[41] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar, “Random-Data Perturbation Techniques and Privacy-Preserving Data Mining,” *Knowledge and Information Systems*, Vol. 7, No. 4, pp. 387-414, 2005.

[42] Z. Huang, W. Du, and B. Chen, “Deriving Private Information from Randomized Data,” In *Proc. of Int’l Conf. on Management of Data*, ACM SIGMOD, New York, NY, pp. 37-48, June 2005.

[43] S. Mukherjee, M. Banerjee, Z. Chen, and A. Gangopadhyay,“A Privacy Preserving Technique for Distance-based Classification with Worst Case Privacy Gaurantees,” *Data and Knowledge Engineering*, Vol. 66, No.2, pp. 264-288, Aug. 2008.

[44] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, “Privacy preserving mining of association rules,” in *Proc. KDD*, pp. 217–228, 2002.

[45] A. Evfimievski, J. Gehrke, and R. Srikant, “Limiting privacy breaches in privacy preserving data mining,” *in Proc. PODS*, pp. 211–222, 2003.

[46] K. Chen, G. Sun, and L. Liu. Towards attack-resilient geometric data perturbation. In *Proceedings of the 2007 SIAM International Conference on Data Mining (SDM’07)*, Minneapolis, MN, April 2007.

[47] S. Guo, and X. Wu. On the Use of Spectral Filtering for Privacy Preserving Data Mining. *In Proceedings of the 21st ACM Symposium on Applied Computing*, pp. 622-626, Dijon, France, 2006.

[48] S. Guo, X. Wu, and Y. Li. On the Lower Bound of Reconstruction Error for Spectral Filtering Based Privacy Preserving Data Mining. *Knowledge Discovery in Databases*: PKDD 2006, 4213: 520-527, 2006.

[49] H. Kargupta, S. Datta, Q.Wang, and K. Sivakumar. On the Privacy Preserving Properties of Random Data Perturbation Techniques. *In Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM 2003)*, pp. 99-106, Melbourne, Florida, 2003.

[50] T. Jiang. How Many Entries of a Typical Orthogonal Matrix can be Approximated by Independent Normals, Annals of Probability, 34(4): 1497-1529, 2006.

[51] D. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J. Halpern. Worst-Case Background Knowledge for Privacy-Preserving Data Publishing. *In Proceedings of the 23rd International Conference on Data Engineering (ICDE 2007)*, pp. 126-135, Istanbul, Turkey, 2007.

[52] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar. Random-Data Perturbation Techniques and Privacy-Preserving Data Mining. *Knowledge and Information Systems,* 7(4):387-414, 2005.

[53] S. Guo, and X. Wu. On the Use of Spectral Filtering for Privacy Preserving Data Mining. *In Proceedings of the 21st ACM Symposium on Applied Computing*, pp. 622-626, Dijon, France, 2006.

[54] S. Guo, X. Wu, and Y. Li. On the Lower Bound of Reconstruction Error for Spectral Filtering Based Privacy Preserving Data Mining. *Knowledge Discovery in Databases*: PKDD 2006, 4213: 520-527, 2006.

[55] C. C. Aggarwal. On Randomization, Public Information and the Curse of Dimensionality. *In Proceedings of the 23rd International Conference on Data Engineering (ICDE 2007)*, Istanbul, Turkey, April, 2007.

[56] Stanley R. M. Oliveira, Osmar R. Zaiane,Privacy Preserving Clustering by Data Transformation, *Journal of Information and Data Management*, Vol. 1, No. 1, February 2010.

[57] UCI Machine Learning Repository https://archive.ics.uci.edu/ml/datasets/Covertype.

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

[59] UCI Machine Learning Repository https://archive.ics.uci.edu/ml/datasets/BankMarketing

Paresh Solanki is received his B.E. degree in computer Engineering from Gujarat University, Ahmedabad and M.TECH Degree in Computer Science and Engineering from Nirma institute of Technology, Nirma University Ahmedabad, India. He is currently a Ph.D. candidate in the Department of Computer Science and Engineering at the Nirma University. His research interests include Data Mining, Network Security.

Sanjay Garg is working as a Professor and Head of Computer Engineering Department at Institute of Technology, Nirma University. He has successfully supervised Ph. D dissertations and currently guiding five PhD students in the field of Data Mining, Patter Recognition and Image Processing. He has completed two Research Projects funded by ISRO under RESPOND scheme and GUJCOST as Principal Investigator and one ISRO-RESPOND research project is in progress. He is recipient of 'Best Engineering College Teacher Award in Gujarat-2016' by ISTE,New Delhi. His experience includes curriculum development at various levels for the programmes under Computer Science and Engineering. He is also a Senior member of IEEE.

Hitesh Chhikaniwala is received his B.E. degree in computer Engineering from Gujarat University, Ahmedabad and M.TECH Degree in Computer Science and Engineering from Natinal institute of Technology, Surathkal, Mangaluru, India. He is receivedly a Ph.D. degree in computer engineering at Kadi Sarva Vishwavidyalaya University. His research interests include Data Mining, Network Security, Big Data Mining.