Principal Component Analysis Based Transformation for Privacy Preserving in Data Stream Mining

[1]Ms.Prachi Patel, [2]Ms.Ompriya Kale,[3]Mr.Bhavin Thakkar

[1]ME-CE Student, [2]Assistant Professor,[3] MD of Vital Solutions

[1]Computer Engineering, Gujarat Technological University, Ahmedabad, Gujarat, India.

[1]prachipatel_ce@yahoo.co.in, [2]ompriya.2007@gmail.com, [3]thakkar.bhavin@yahoo.com

***Abstract*— Data stream can be conceived as a continuous and changing sequence of data that continuously arrive at a system to store or process. Examples of data streams include computer network traffic, phone conversations, web searches and sensor data etc. The data owners or publishers may not be willing to exactly reveal the true values of their data due to various reasons, most notably privacy considerations. To preserve data privacy during data mining, the issue of privacy preserving data mining has been widely studied and many techniques have been proposed. However, existing techniques for privacy preserving data mining is designed for traditional static data sets and are not suitable for data streams. So the privacy preservation issue of data streams mining is need for the time. This paper focused on techniques for Principal Component Analysis (PCA) based transformation for stream data using Massive Online Analysis (MOA). The clustering accuracy while using the transformed data is almost equal to the original dataset.**

***Index Terms*— Data Stream, Data Perturbation, Data Perturbation, Random Function**

I. Introduction

In the field of information processing, data mining refers to the process of extracting the useful knowledge from the large volume of data. Widely used data mining techniques in such area of application includes Clustering, Classification, Regression analysis and Association rule / Pattern mining.

The data stream paradigm has recently emerged in response to the issues and challenges related with continuous data [1]. Mining data streams is concerned with extracting knowledge structures represented in models and patterns in non-stopping, continuous streams (flow) of information. Algorithms written for data streams can naturally cope with data sizes many times greater than memory, and can be extended to challenge real-time applications not previously tackled by machine learning or data mining.

But nowadays, in the field of information processing, an emergence of applications that do not fit this data model [2] Instead, information naturally occurs in the form of a sequence (stream) of data values. A data stream is a real-time, continuous, and ordered sequence of items. It is not possible to control the order in which items arrive, nor feasible to locally store a stream in its entirety. Likewise, queries over streams run continuously over a period of time and incrementally return new results as new data arrive.

II. Privacy Concern For Data Stream

Mining data streams is concerned with extracting knowledge structures represented in models and patterns in non-stopping streams of information. Motivated by the privacy concerns on data mining tools, a research area called privacy-preserving data mining has been emerged.

Verykios et al. [3] classified privacy- preserving data mining techniques based on five dimensions – data distribution , data modification, data mining algorithms, data or rule hiding, and privacy preservation. In the dimension of data distribution, some approaches have been proposed for centralized data and some for distributed data.

Du and Zhan [4] utilized the secure union, secure sum and secure scalar product to prevent the original data of each site from revealing during the mining process. The disadvantage is that the approach requires multiple scans of the database and hence is not suitable for data streams, which flows in fast and requires immediate response.

In the dimension of data modification, the confidential values of a database to be released to the public are modified to preserve data privacy. Adopted approaches include perturbation, blocking, aggregation or merging, swapping, and sampling. Agrawal and Srikant [5] used the random data perturbation technique to protect customer data and then constructed the decision tree. For data streams, because data are produced at different time, not only data distribution will change with time, but also the mining accuracy will decrease with perturb data. From the review of previous research, it can be seen that existing techniques for privacy-preserving data mining are designed for static databases with an emphasis on data security. These existing techniques are not suitable for data streams.

Perturbation techniques are often evaluated with two basic metrics: level of privacy guarantee and level of model-specific data utility preserved, which is often measured by the loss of accuracy for data classification and data clustering. An ultimate goal for all data perturbation algorithms is to optimize the data Transformation process by maximizing both data privacy and data utility achieved. Data privacy is commonly measured by the difficulty level in estimating the original data from the perturbed data. Given a data perturbation technique, the higher level of difficulty in which the original values can be estimated from the perturbed data, the higher level of data privacy this technique supports. Data utility typically refers to the amount of mining-task/model specific critical information preserved about the data set after perturbation.

III. Privacy Preserving data stream clustering

The data stream model of computation requires algorithms to make a single pass over the data, with bounded memory and limited processing time, whereas the stream may be highly dynamic and evolving over time. For effective clustering of stream data, several new methodologies have been developed, as follows: Compute and store summaries of past data: Due to limited memory space and fast response requirements, compute summaries of the previously seen data, store the relevant results, and use such summaries to compute important statistics when required.

The main idea of Perturbation- Based technique involves increasing a noise in the raw data in order to perturb the original data distribution and to preserve the content of hidden raw data. Geometric Data Transformation Methods (GDTMs) [6] is one simple and typical example of data perturbation technique, which perturbs numeric data with confidential attributes in cluster mining in order to preserve privacy.

Nonetheless Kumari et al. [7] proposed a privacy preserving clustering technique of Fuzzy Sets, transforming confidential attributes into fuzzy items in order to preserve privacy. Furthermore, the largest issue encountered when implementing a perturbation technique is the inaccurate mining result from a perturbed data.

Vaidya and Clifton [8] proposed the method of privacy preserving clustering technique over vertically partitioning data. In the vertical partitioning the attributes of the same objects are split across the partitions.

On the contrary, Meregu and Ghosh [9] proposed the method of privacy preserving cluster mining over horizontally data partitioning, whereas it is framework Of “Privacy-preserving Distributed Clustering using Generative Model.”In this approach, rather than sharing parts of the original data or perturbed data, the parameters of suitable generative models are built at each local site.

In [10] proposed a method of Privacy-Preserving Clustering of Data Stream (PPCDS), stressing the privacy-preserving process in a data stream environment while maintaining a certain degree of excellent mining accuracy. PPCDS is mainly used to combine Rotation-Based Perturbation, optimization of cluster enters and the concept of nearest neighbor, in order to solve the privacy-preserving clustering of mining issues in a data stream environment. In the phase of Rotation-Based Perturbation, rotation transformation matrix is employed to rapidly perturb with data streams in order to preserve data privacy. In the phase of cluster mining, perturbed data is primarily used to establish a micro-cluster through the optimization of a cluster center, then applying statistic calculation to update the micro-cluster.

IV. Problem Description

The initial idea of it was to extend traditional data mining techniques to work with the perturbed stream data to mask sensitive information. The key issue is to get accurate stream mining results using perturb data. The solutions are often tightly coupled with the data stream mining algorithms under consideration.

The goal is to transform a given data set D into perturbed version D’ that satisfies a given privacy requirement and loss minimum information for the intended data analysis task. In this paper data perturbation algorithms have been proposed for data set perturbation.

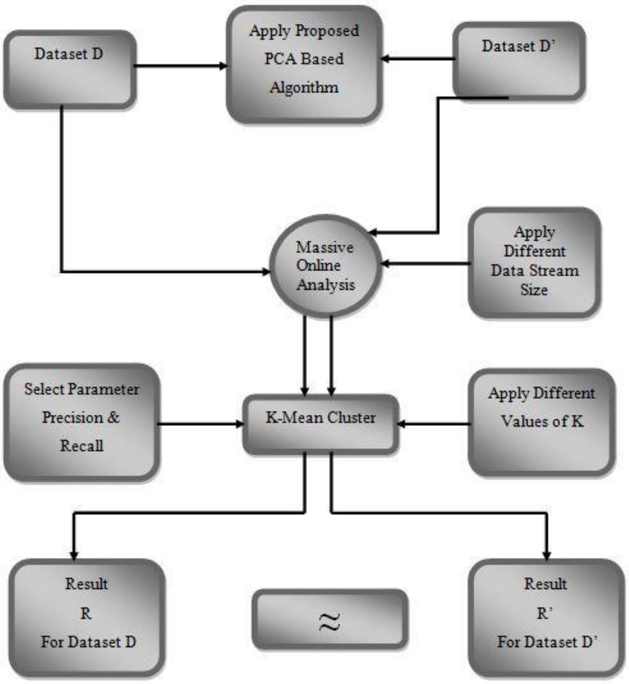
Dataset D
Apply Proposed PCA Based Algorithm
Dataset D'
Massive Online Analysis
Apply Different Data Stream Size
Select Parameter Precision & Recall
K-Mean Cluster
Apply Different Values of K
Result R For Dataset D
≈
Result R' For Dataset D'

Fig 1. Framework for privacy preserving in data stream clustering

Assuming the data stream for processing includes multiple multi-dimensional numeric data X1...XK ..., each data contains its proprietary timestamp T1…TK..., with multi-dimensional data represented by Xi = (xi1...xid). When a data stream incoming, data is represented in an m x n data matrix Dm×n, while each row represents one entry and each column represents an attribute of data.

In this work Principal Component Analysis (PCA) is used for transforming the multidimensional data into lower dimensions. PCA is a standard tool in modern data analysis .PCA assumes that all the variability in a process should be used in the analysis therefore it becomes difficult to distinguish the important variable from the less important.

Algorithm: Data Perturbation Using Principle Component Analysis (PCA)

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

Intermediate Result**: Perturbed data stream D'.**

Output**: Clustering results R and R' of Data stream D and D' respectively.**

Steps:

1. Given input data D with tuple size n, extract sensitive attribute [S]n×3.
2. Calculate Orthogonal matrix [O]n×3 .
3. Multiply [O]n×3 with [S]n×3 call as [R1]n×3 .
4. Multiply [R1]n×3 with transpose of [O]n×3 call as [R2]n×3 .
5. Calculate mean of [S]n×3 calls as µ.
6. Calculate f(x) = µ + [R2]n×3.
7. Subtract f(x) from sensitive attribute [S]n×3 call as [R3]n×3.
8. Calculate square of [R3] n×3 and divide by n call as projection matrix [P]n×3.
9. Crate perturbed dataset D' by replacing sensitive attribute [S]n×3 in original dataset D with [P]n×3.
10. Apply k-Mean clustering algorithm with different values of k on original dataset D having sensitive attribute S.
11. Apply k-Mean clustering algorithm with different values of k on perturbed dataset D' having perturbed sensitive attribute P.
12. Create cluster membership matrix of results from step 10 and step 11 and analyze.

V. Result And Discussion

Series of experiments were performed over define sliding window size (w) in order to evaluate the clustering accuracy. Our evaluation approach focused on the overall quality of generated clusters after dataset perturbation.

Experiment was based on following steps:

- Setup each dataset as stream in MOA framework.
- Define sliding window (w) over the data stream to evaluate measures and cluster membership matrix.
- Modified all the instances in sliding window by applying our proposed data perturbation method to protect the sensitive attribute value.
- K-Means clustering algorithm is used to find the clusters for our performance evaluation. Our selection was influenced by (a) K-Means is one of the best known clustering Algorithm and is scalable. (b)Number of cluster to be find from original and perturbed dataset was taken same as number of cluster.
- Compare how closely each cluster in the perturbed dataset matches its corresponding cluster in the original dataset. We expressed the quality of the generated clusters by computing the F-measure

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

Table 5.1 shows datasets configuration to determine the accuracy of our proposed method. We configured each dataset to determine 5 and 3 clusters using K-Means clustering algorithm. Table 5.2, 5.3 shows the membership matrix obtained while clustering the perturbed attributes of Bank Management dataset respectively. Each Matrix representing 5 and 3 clusters scenario for true dataset and perturb dataset. True dataset clustering gives information about no. of instances are actual classified in each cluster where as perturb dataset clustering showing result of correct assignments after attributes data perturbation and percentage of accuracy achieved.

Table 5.1 Dataset configuration to determine accuracy based on Membership Matrix

Dataset Name	Total instances	Instances processed	Attributes protected
Bank Management	45210	45k	Age, Balance, Duration

k-Mean clustering algorithm has been applied on original dataset D and perturbed dataset D' generated using proposed algorithm. Results in table 5.2 and 5.3 shows that for all tested cases almost 90% mining accuracy has been achieved. Algorithm has been tested against different values of *k* and it has been observed that accuracy has been decreasing as *k* value increases. This justifies that probability of tuple to fall into original cluster will be decreasing as number of clusters increases.

Table 5.2 accuracy of 5-Cluster

Dataset	Attributes	No. of Clusters	Stream Data	K-Means
Bank Management	Age	5	2000	89.51
	Balance			90.75
	Duration			88.10
	Age		3000	84.64
	Balance			89.05
	Duration			84.49

Table 5.3 accuracy of 3-Cluster

Dataset	Attributes	No. of Clusters	Stream Data	K-Means
Bank Management	Age	3	2000	92.77
	Balance			94.10
	Duration			90.60
	Age		3000	90.36
	Balance			91.72
	Duration			89.19

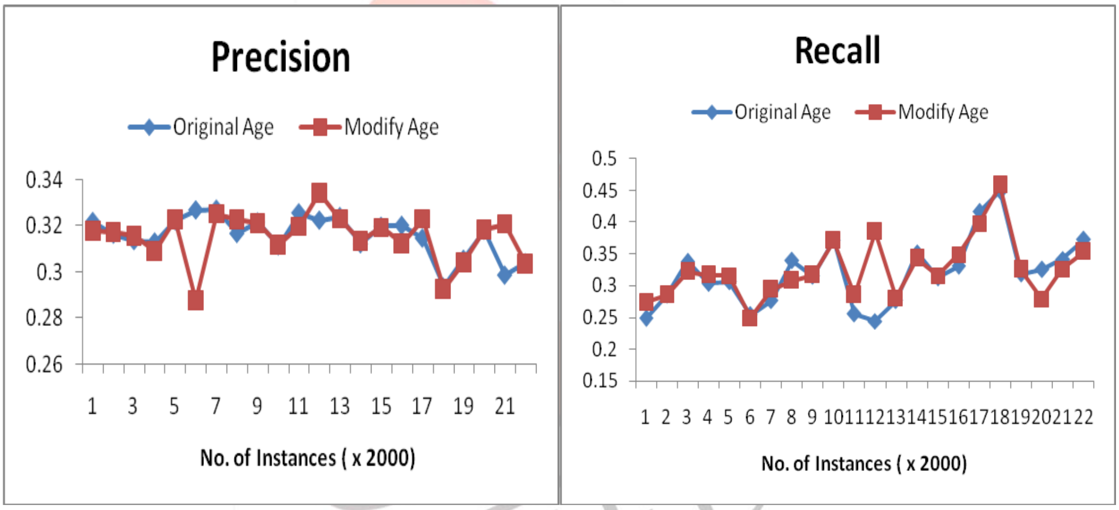

Fig.2.Accuracy on attribute Age in Bank Management with 5-Cluster

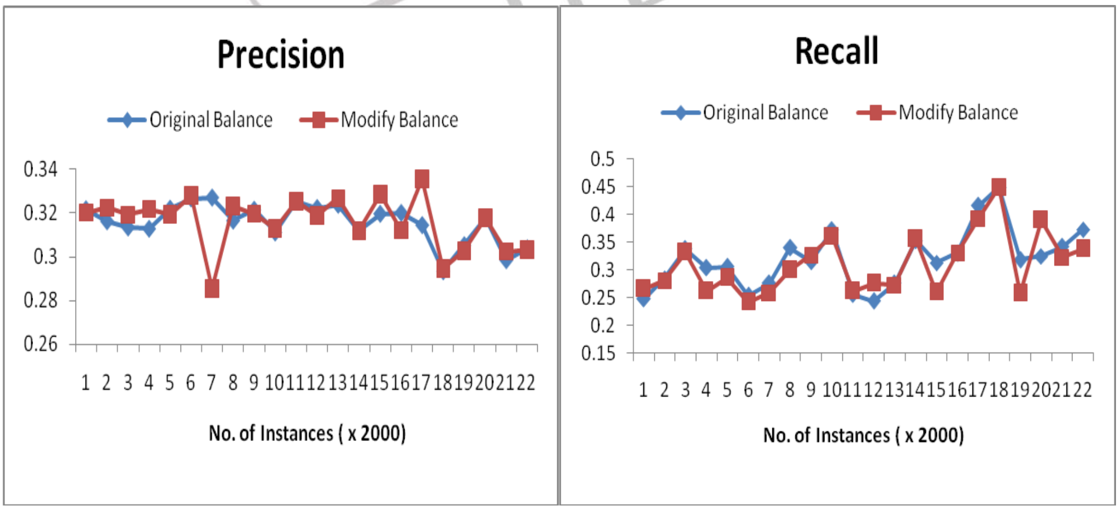

Fig.3.Accuracy on attribute Balance in Bank Management with 5-Cluster

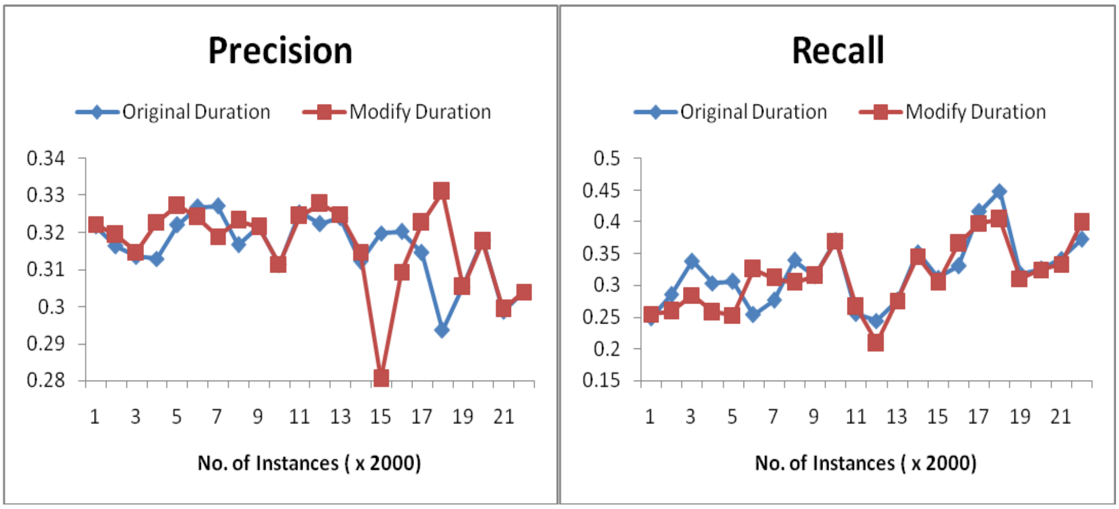

Fig.4.Accuracy on attribute Duration in Bank Management with 5-Cluster

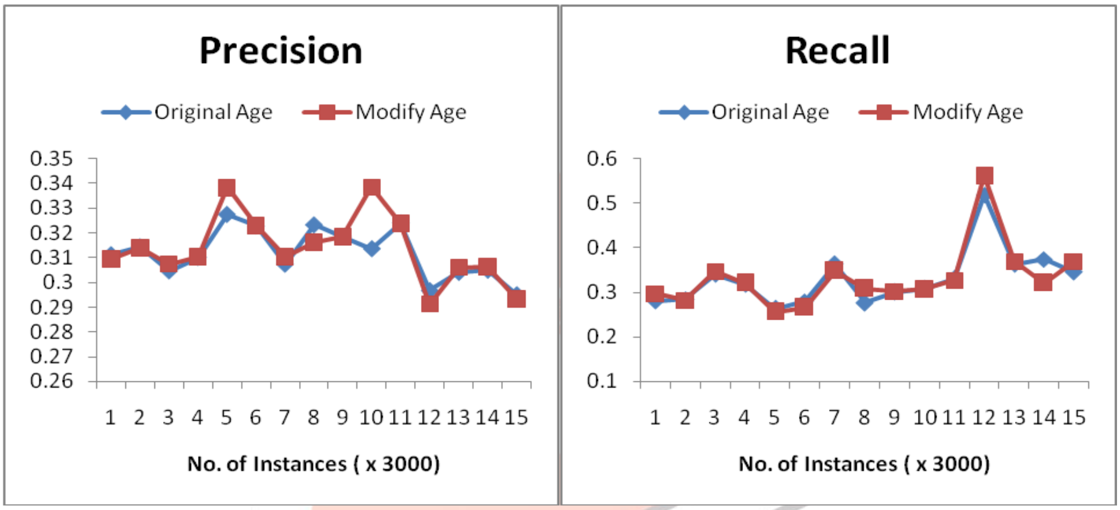

Fig.5.Accuracy on attribute Age in Bank Management with 5-Cluster

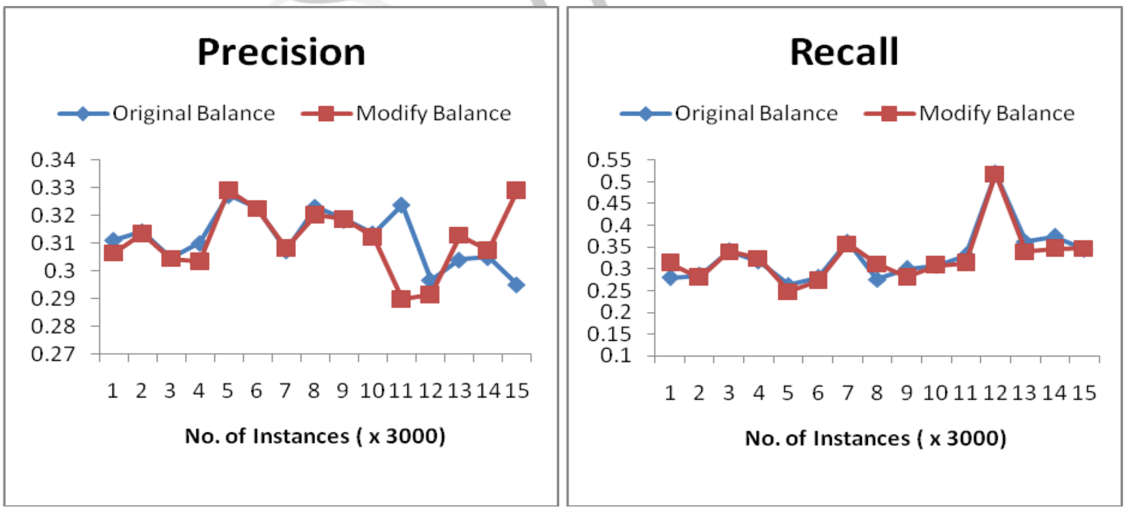

Fig.6.Accuracy on attribute Balance in Bank Management with 5-Cluster

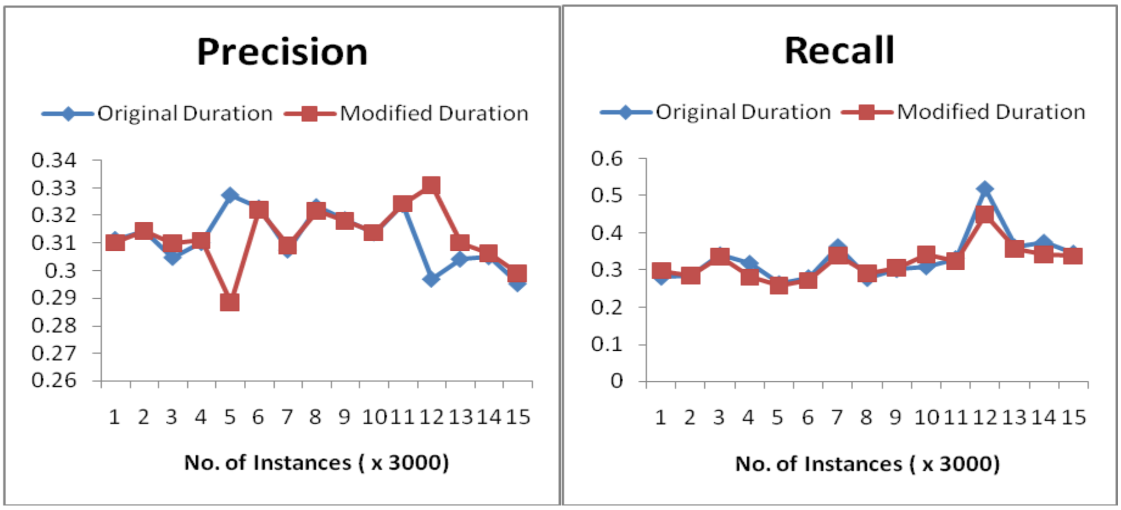

Fig.7.Accuracy on attribute Duration in Bank Management with 5-Cluster

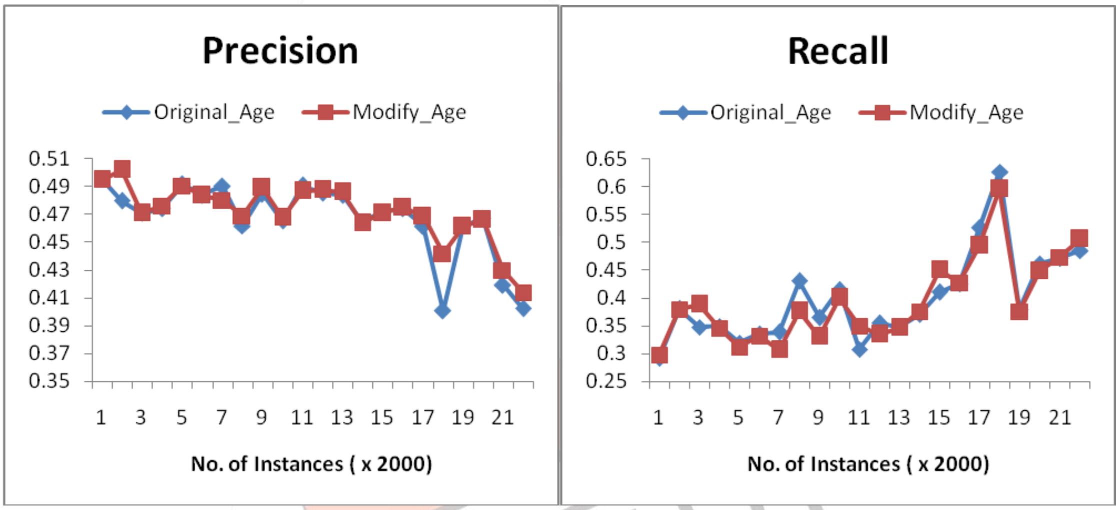

Fig.8.Accuracy on attribute Age in Bank Management with 3-Cluster

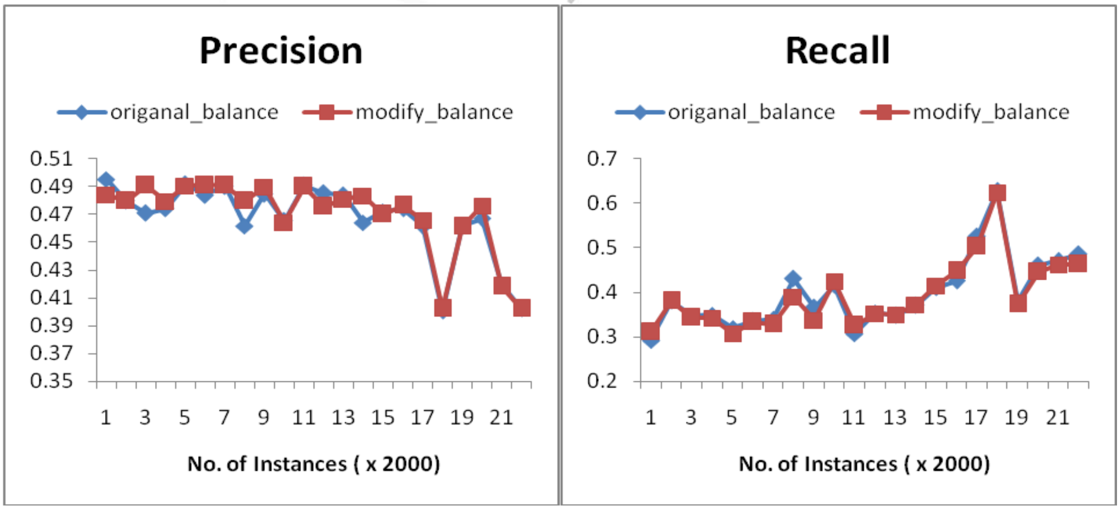

Fig.9.Accuracy on attribute Balance in Bank Management with 3-Cluster

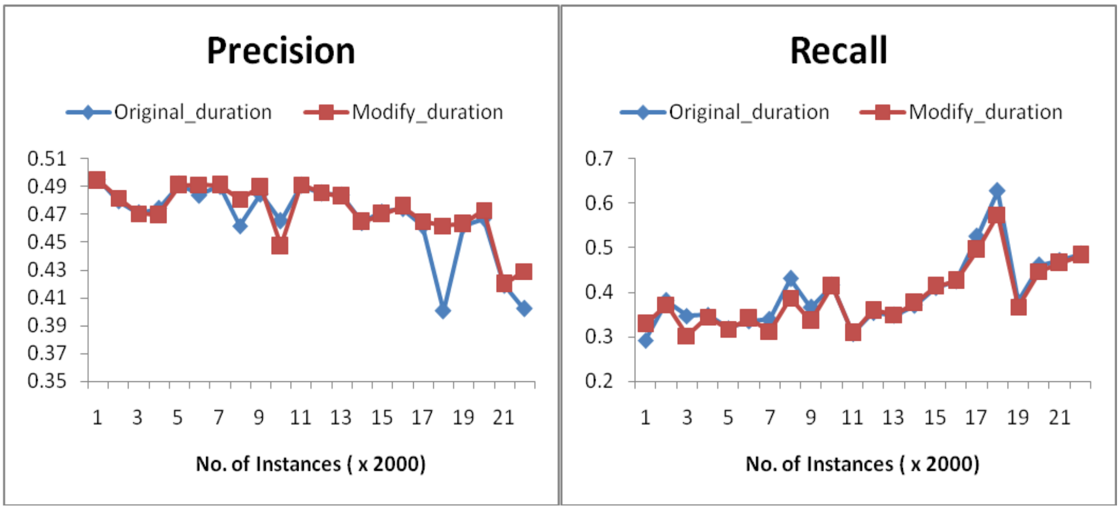

Fig.10.Accuracy on attribute Duration in Bank Management with 3-Cluster

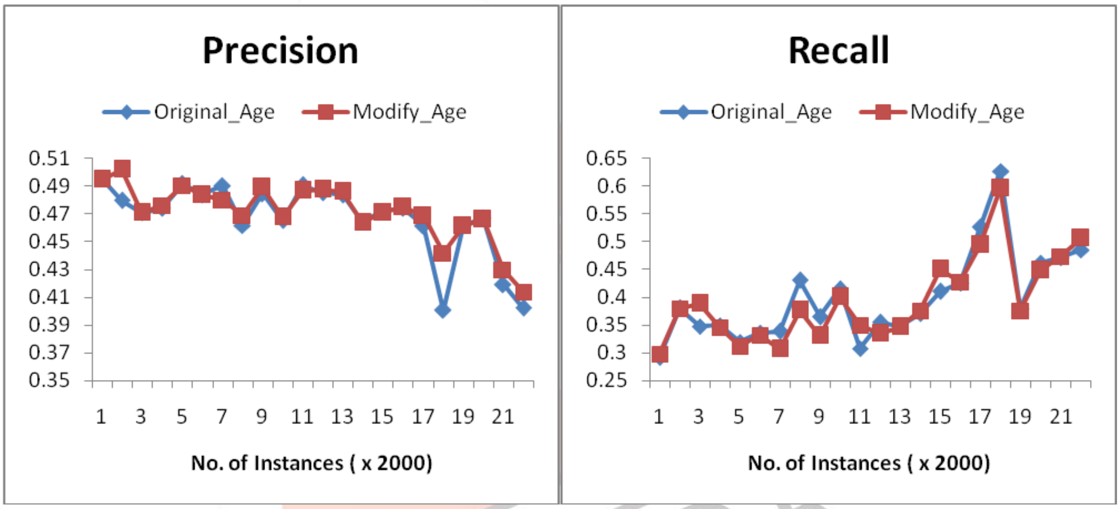

Fig.11.Accuracy on attribute Age in Bank Management with 3-Cluster

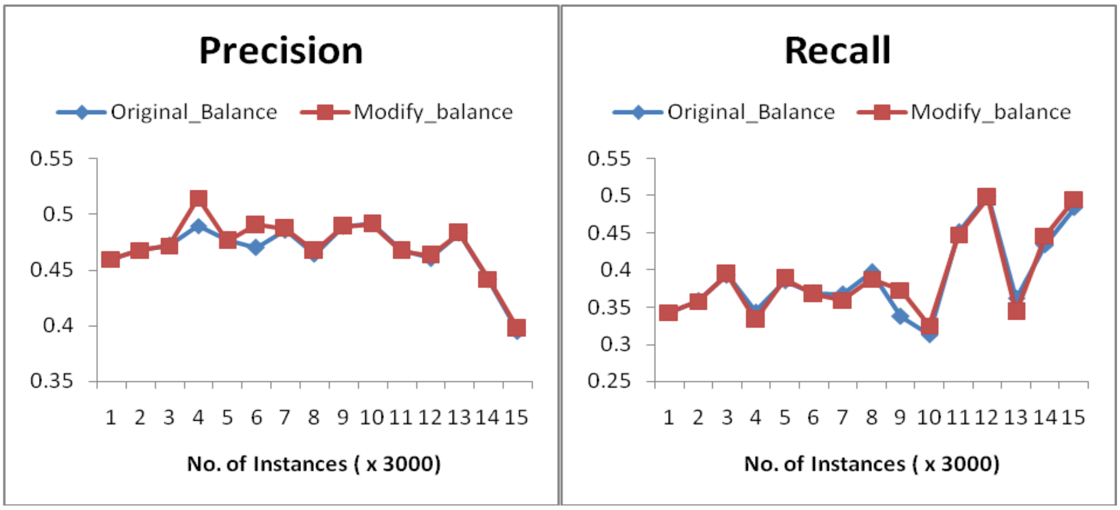

Fig.12.Accuracy on attribute Balance in Bank Management with 3-Cluster

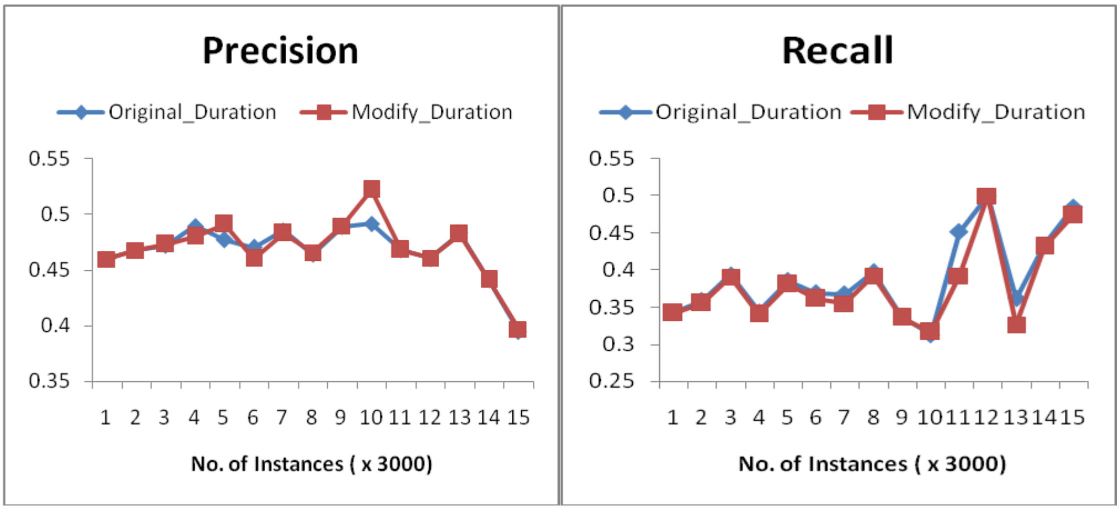

Fig.13.Accuracy on attribute Duration in Bank Management with 3-Cluster

VI. Conclusion

PCA based multiplicative data perturbation approach has been proposed for random noise addition to preserve privacy of sensitive attributes. Proposed approach has tried to keep statistical relationship among the sensitive attributes intact to mine favorable results with perturbed data. It considers sensitive attribute as dependent attribute and remaining attributes of dataset except class attribute as independent attributes. Only dependent attribute of dataset has been used to calculate tuple specific random noise. K-Mean clustering algorithm on perturbed dataset has been used in order to estimate the accuracy and effectiveness of clustering results over four standard datasets. Results show fairly good level of privacy has been achieved with reasonable accuracy in almost all tested cases. Privacy of original data after applying perturbation has been quantified using correlation analysis. Data mining accuracy due to data perturbation has been quantified by percentage of instances of dataset that are been misclassified with clustering results with original dataset. We limited experiments to protect numeric attribute only but work can be extended to nominal type attributes also.

References

[1] A. Bifet, G. Holmes, R. Kirkby and B. Pfahringer, *Data Stream Mining-A Practical approach*, 2011.

[2] L. Golab and M. T. Ozsu, Data Stream Management Issues -A Survey Technical Report, 2003.

[3] V.S. Verykios, K. Bertino, I. N. Fovino, L.P. Provenza, Y.Saygin and Theodoridis, State-of-the-Art in Privacy Preserving Data Mining, *ACM SIGMOD Record*, Vol. 33, pp. 50-57, 2004.

[4] W. Du and Z. Zhan, Building Decision Tree Classifier on Private Data, *Proceedings of IEEE International Conference on Privacy Security and Data Mining*, pp. 1-8, 2002.

[5] R. Agrawal and R. Srikant, Privacy-Preserving Data Mining, *Proceedings of ACM SIGMOD International Conference on Management of Data*, pp. 439-450, 2000.

[6] S. R. M. Oliveira and O. R. Zaiane. Privacy Preserving Clustering By Data Transformation. In *Proc. of the 18th Brazilian Symposium on Databases*, pages 304–318, Manaus, Brazil, October 2003.

[7] V. Estivill-Castro and L. Brankovic. Data Swapping: Balancing Privacy Against Precision in Mining for Logic Rules.In Proc. of Data Warehousing and Knowledge Discovery DaWaK-99, pages 389– 398, Florence, Italy, August 1999.

[8] Vaidya, J. and Clifton, C., “Privacy-Preserving KMeans Clustering over Vertically Partitioned Data,”Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and DataMining, Washington, D.C., U.S.A.,pp.206_215 (2003).

[9] Meregu, S. and Ghosh, J., “Privacy-Preserving Distributed Clustering Using Generative Models,”Proceedings of the 3th IEEE International Conference on Data Mining, Melbourne, Florida, U.S.A.,pp. 211_218 (2003).

[10] Ching-Ming Chao, Po-Zung Chen and Chu-Hao Sun, *Privacy-Preserving Clustering of Data Streams*, Tamkang Journal of Science and Engineering, Vol. 13, No. 3, pp.349 - 358(2010).