International Journal For Technological Research In Engineering
Volume 4, Issue 8, April-2017 *ISSN (Online): 2347 - 4718*

PRIVACY PRESERVATION OF STREAM DATA USING RANDOM ROTATION BASED PERTURBATION TECHNIQUE

Miss Bhagyashree Patil
M.E in Computer Engineering, Pursuing, Hashmukh Goswami Engineering College, Vehlal

Abstract: Privacy preserving data mining (PPDM) deals with protecting the privacy of individual data or sensitive knowledge without sacrificing the utility of the data..A number of algorithmic techniques have been designed for Privacy Preserving Data Mining (PPDM). one crucial concept about existing data mining privacy preserving techniques are suitable and designed for static databases and not suitable for data streams. Recently,data streams are introduced as new type of data which are differ from traditional static data. Various features of data streams are: with time, data distribution changes constantly; data is having time preferences; amount of data is extensive; flow of data with fast speed; requirement of immediate response..Further, it is observed that accuracy of data is decreases when transformation is carried out on data.so,there has been need to develop the system which preserve privacy along with accuracy. So privacy preserving on data stream mining is very crucial issue. In this paper we described rotation perturbation techniques for preserving privacy.

Keywords: Privacy; Data Streams; K-mean –clustering

I. INTRODUCTION

Data mining, with its promise to efficiently discover valuable, non obvious information from large datasets, is particularly vulnerable to misuse.[1] Data Mining refers to extracting or mining knowledge from large amounts of dataPrivacy preserving data mining (PPDM) deals with protecting the privacy of individual data or sensitive knowledge without sacrificing the utility of the data. PPDM tends to transform the original data so that the result of data mining task should not defy privacy constraints. Following is the list of five dimensions on the basis of which different PPDM Techniques can be classified [4]:

(i) Data distribution
(ii) Data modification
(iii) Data mining algorithms
(iv) Data or rule hiding
(v) Privacy preservation

The important techniques of privacy preserving data mining are[3],

- The reconstruction method
- The anonymization method
- The cryptographic method

The reconstruction method:

Reconstruction method is an important and popular method in current privacy preserving data mining techniques. It masks the values of the records by adding additional data to the original data.

The Anonymization Method:

Anonymization method is aimed at making the individual record will be indistinguishable among a group record by using generalization and suppression techniques. K-anonymity is the representative anonymization method. The motivating factor behind the k-anonymity approach is that many attributes in the data can often be considered quasi-identifiers which can be used in conjunction with public records in order to uniquely identify the records.

The cryptographic method:

The cryptographic method mainly resolves the problems that people jointly conduct mining tasks based on the private inputs they provide. These privacy mining tasks could occur between mutual un-trusted parties, or even between competitors. Therefore, to protect the privacy becomes an important concern in distributed data mining setting.

1.1 Data perturbation

Data perturbation techniques are one of the most popular models for privacy preserving data mining. It is especially convenient for applications where the data owners need to export/publish the privacy-sensitive data. A data perturbation procedure can be simply described as follows. Before the data owner publishes the data, they randomly change the data in certain way to disguise the sensitive information while preserving the particular data property that is critical for building the data models.[12] In perturbation approach, records released is synthetic i.e. it does not correspond to real world entities represented by the original data. Therefore the individual records in the perturbed data are meaningless to the human recipient as only statistical properties of the records are preserved. Perturbation can be done by using additive noise or data swapping or synthetic data generation. Since the perturbation method does not reconstruct the original values but only the distributions, new algorithms are to be developed for mining of the data . This means that a new distribution based mining algorithms need to be developed for each individual data problems like classification, clustering or association rule mining.

II. PROPOSED WORK

Data perturbation is a process of transformations on data performed by the data owners before data being published. The goal of performing such data transformation is two-sided. The owners of data want to change the data in a way that mask the sensitive information contained in the published datasets. .

The data streams pre-processing uses perturbation algorithm to perturb confidential data. Users can flexibly adjust data

attributes to be perturbed according to the security need. Thus, threats and risks from releasing data can be effectively reduced.

Standard tool in modern data analysis is principal component analysis (PCA). The aim of PCA is to distinguish the important variable from the less important uses all the variability in process for data analysis.

In PCA, data characteristics remain unchanged during whole process and behavior of data should be same as original data. For that here is derived formula based on PCA

$$f(x, v) = \mathrm{u} + (xV)V^T$$

$$R(x, V) = \frac{1^n}{n_{i=1}} \|x_i - f(x_i, V)\|$$

Where, xi= Original dataset
u= Mean of Original dataset
V = Orthogonal matrix

Algorithm

Algorithm Data Perturbation Using Principle Component Analysis (PCA)

- Procedure: PCA Based Multiplicative Data Perturbation.
- 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:

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

Data Stream Cluster Mining

The objective of this stage is to mine perturbed data streams to construct a clustering model and evaluate the clustering measures. We configure sliding window mechanism to perform clustering on incoming stream tuples. For each different cluster size we check by taking variety of stream sizes.

III. RESULTS AND DISCUSSION

To measure accuracy while protecting sensitive data, experiments were performed. Here we have presents two different results, one is analogous to clustering accuracy in terms of membership matrix which was manually plagiaristic from clustering result and another represent the equivalent graph for F1_P(precision) and F1_R(Recall) measures.

Dataset Name	Total instances	Instances processed	Attributes Protected
Account Management	42210	43k	Balance, Age Duration

Table 1.1: Dataset configuration to determine accuracy based on Membership Matrix

To determine the accuracy of our proposed method, Table 1.1 shows datasets configuration. To determine set of 3 and 5 clusters using K-Means clustering algorithm, We configured each dataset. Table 1.2, 1.3 shows the membership matrix acquired while clustering the perturbed attributes of Account Management dataset respectively. Each Matrix representing 3 and 5 clusters scenario for true dataset and discompose dataset. True dataset clustering provides information about no. of instances are actual classified in each cluster whereas perturb dataset clustering showing result of accurate assignments after attributes data perturbation and percentage of accuracy achieved.

Dataset	Attributes	No. of Cluster	Stream Data	*k-means*
Bank Management	Age	5	2000	88.59%
	Balance			88.99%
	Duration			87.41%
	Age		3000	84.28%
	Balance			88.06%
	Duration			84.39%

Table 1.2: Accuracy Results for 5-Cluster for Bank Data set.

Dataset	Attributes	No. of Cluster	Stream Data	*k-means*
Bank Management	Age	3	2000	92.30%
	Balance			93.26%
	Duration			89.46%
	Age		3000	90.31%
	Balance			89.02%
	Duration			86.69%

Table 1.3: Accuracy Result for Bank Dataset for 3-Cluster

For each modified attribute, Results are presented in terms of graphs. Here each graph comprises the measure we obtained when original data is processed without applying privacy preserving method and K-Means is applied in order to evaluate both cases by keeping number of clusters fix (K=5, K=3) ,when data is undergone through our proposed privacy preserving method. In defined sliding window size, Instances are processed. Here we representing the accuracy of our method by calculating the precision of individual cluster.

F1_R measure determine the recall of system, which take into account theclustering measure provided with MOA framework. We focused on two important measures F1_R and F1_P. F1_P.F1_P measure determine the precision of system by considering the precision of individual cluster. F1_R measure determine the recall of system, which take into account the recall of each cluster. Figure 1.1, 1.2 and 1.3 shows the precision and recall of the Bank dataset on the age, duration and balance attribute resp. The total no of instances for stream input is 2000 with the 3 cluster size.

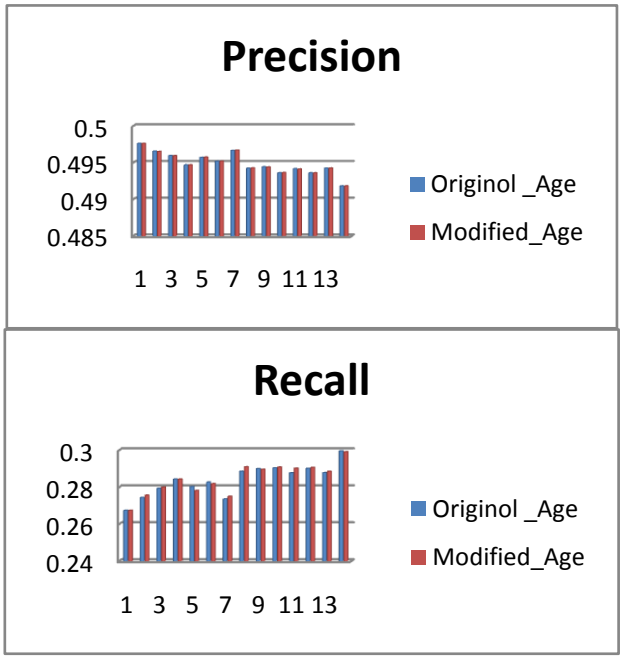

Fig 1.1: Accuracy on attribute Age in Bank Management with 3-Cluster (instances 2000)

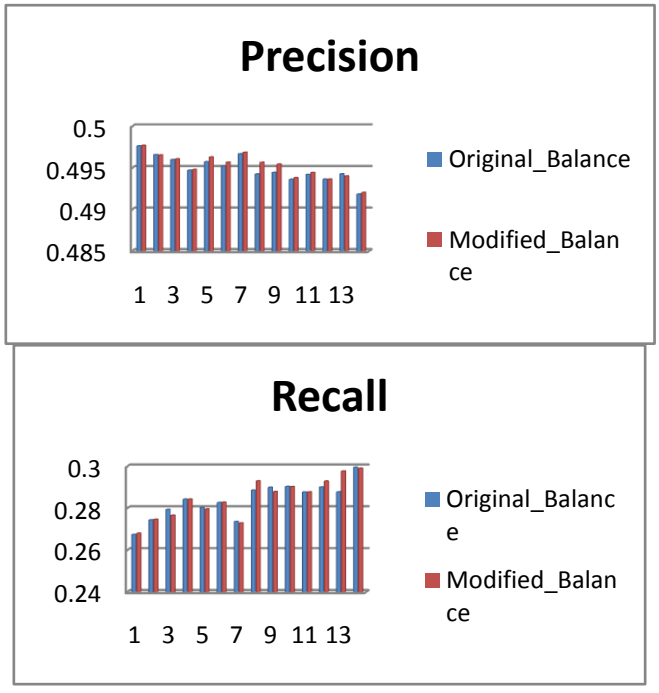

Fig 1.2: Accuracy on attribute Balance in Bank Management with 3-Cluster (instances 2000)

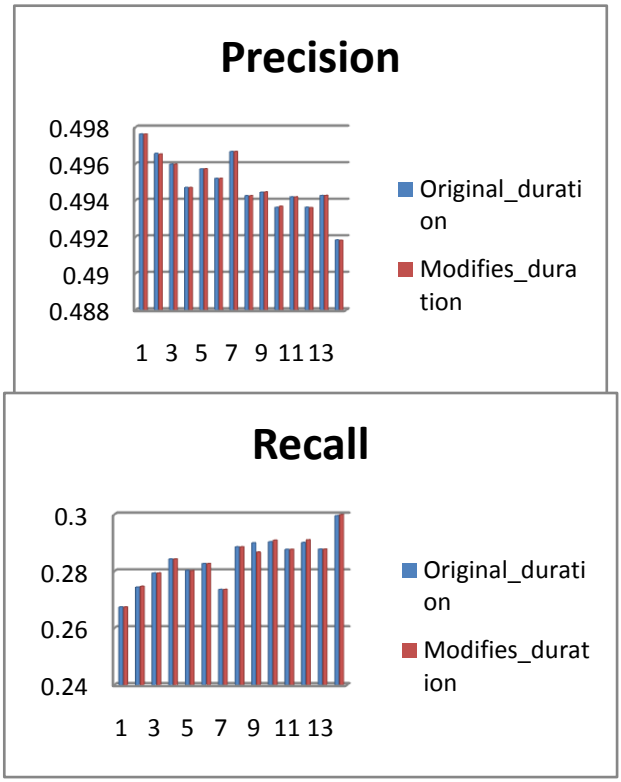

Fig 1.3: Accuracy on attribute duration in Bank Management with 3-Cluster (instances 2000)

Figure 1.4, 1.5 and 1.6 shows the precision and recall of the Bank dataset on the age, duration and balance attribute resp. The total no of instances for stream input is 3000 with the 3 cluster size

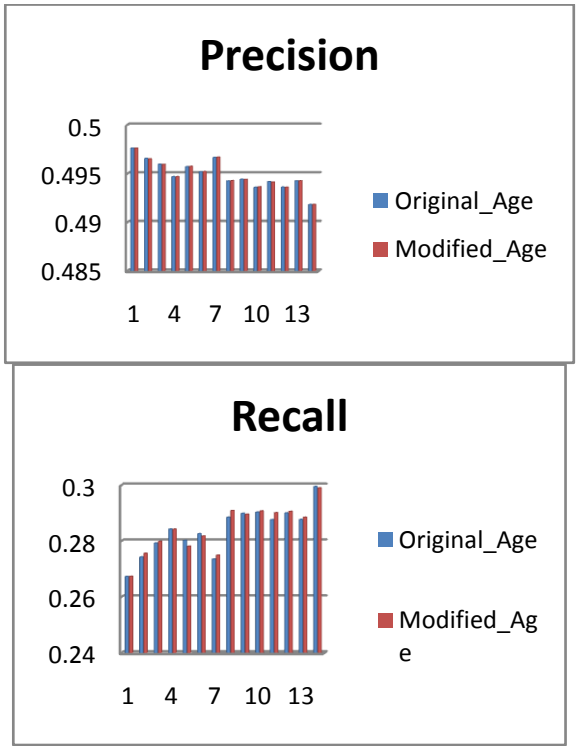

Fig 1.4: Accuracy on attribute Age in Bank Management with 3-Cluster (instances 3000)

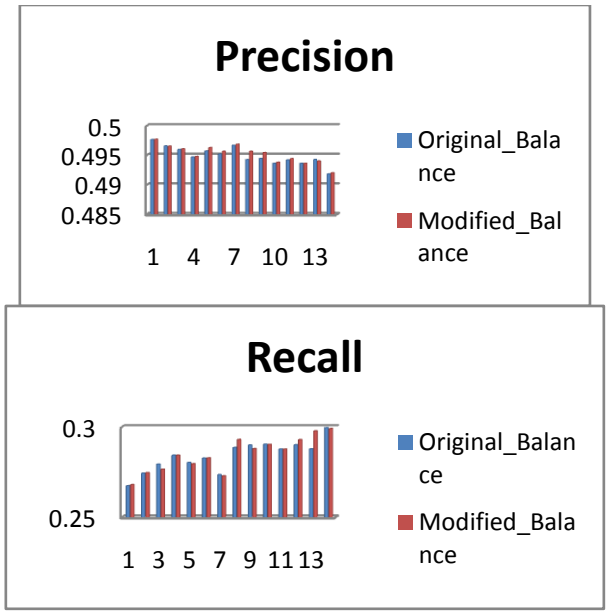

Fig 1.5: Accuracy on attribute Balance in Bank Management with 3-Cluster (instances 3000)

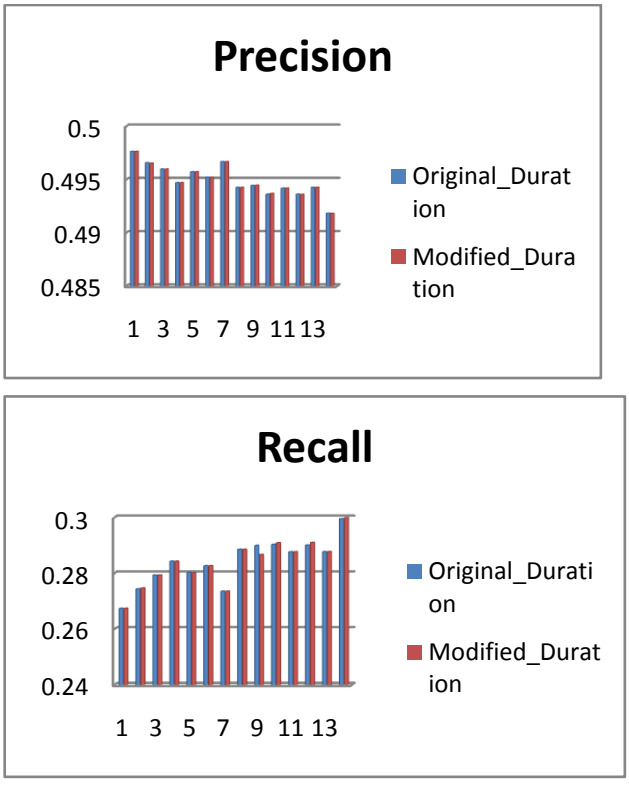

Fig 1.6: Accuracy on attribute Age in Duration Management with 3-Cluster (instances 3000)

Figure 1.7, 1.8 and 1.9 shows the precision and recall of the Bank dataset on the age, duration and balance attribute resp. The total no of instances for stream input is 5000 with the 2 cluster size

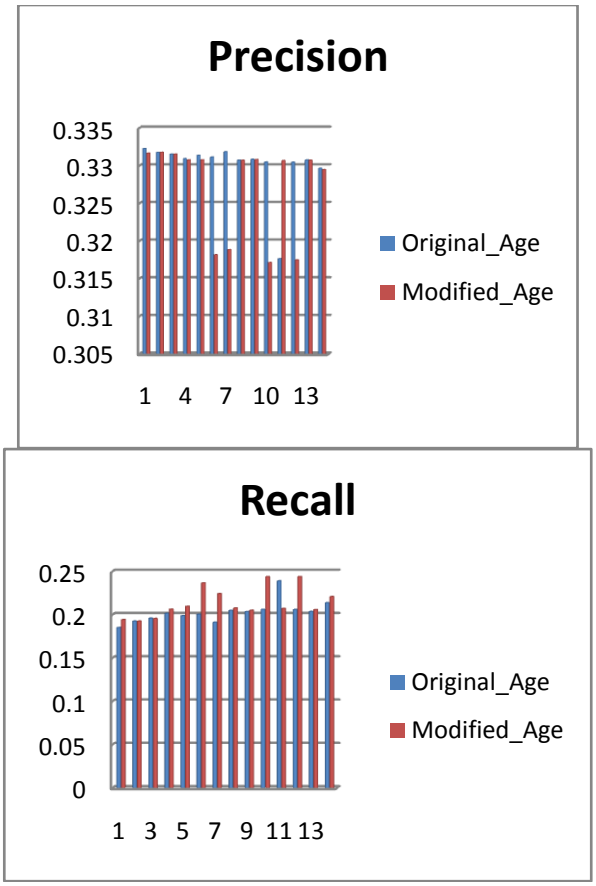

Fig 1.7: Accuracy on attribute Age in Bank Management with 5-Cluster (instances 2000)

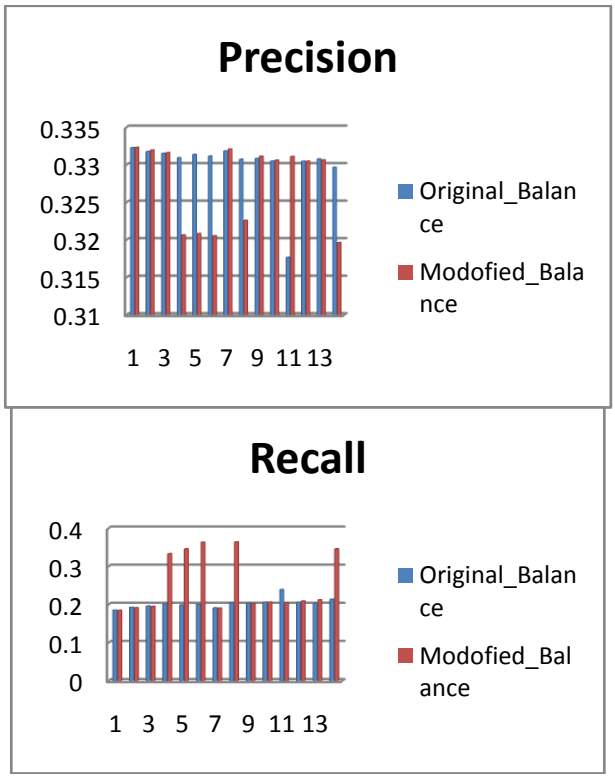

Fig 1.8: Accuracy on attribute Balance in Bank Management with 5-Cluster (instances 2000)

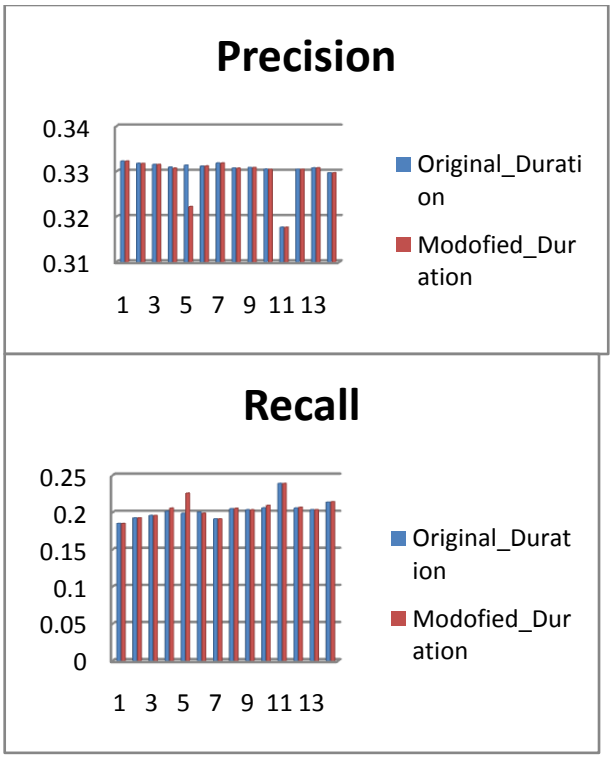

Fig 1.9: Accuracy on attribute Duration in Bank Management with 5-Cluster (instances 2000)

Figure 1.10, 1.11 and 1.12 shows the precision and recall of the Bank dataset on the age, duration and balance attribute resp. The total no of instances for stream input is 3000 with the 5 cluster size.

Fig 1.10: Accuracy on attribute Age in Bank Management with 5-Cluster (instances 3000)

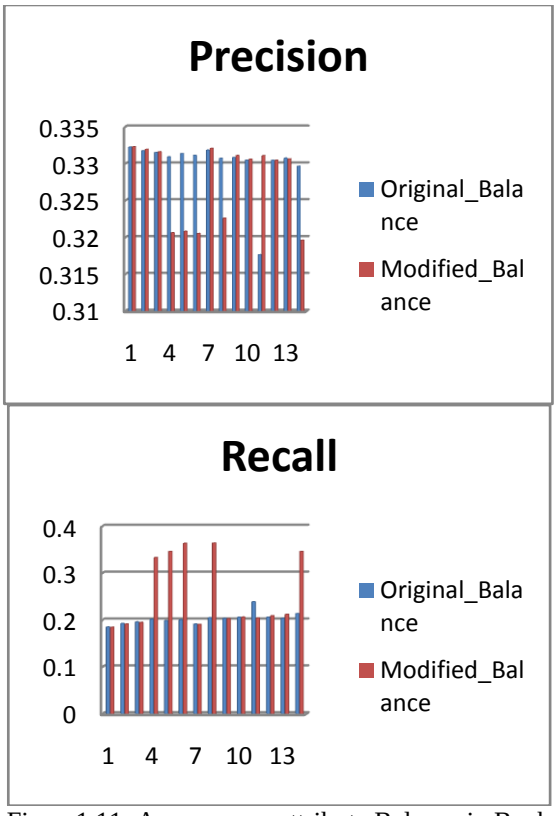

Fig 1.11: Accuracy on attribute Balance in Bank Management with 5-Cluster (instances 3000)

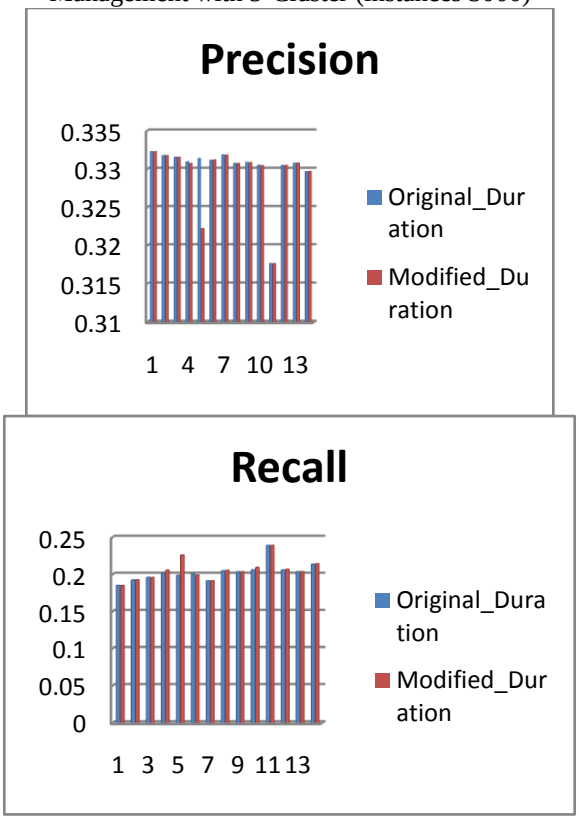

Fig 1.12: Accuracy on attribute Duration in Bank Management with 5-Cluster (instances 3000)

IV. CONCLUSION

While presenting on a publicly accessible place like internet, the proposed method can be used to hide sensitive information. The proposed privacy preserving prototype has been successfully implemented in java under Windows 7 operating system and evaluated using Massive Online Analysis (MOA). The arrived results were more substantial and promising.. Additionally, the proposed model can be used to multi party cooperative clustering development. Some of the results of earlier works have been shown, accuracy sometimes suffers as a result of security. However in the proposed method, the accuracy has been conserved and in some cases, the accuracy was almost equal to that of original data set.

REFERENCES

[1] R.Vidya Banu ,N.Nagaveni" : "Preservation of Data Privacy using PCA based Transformation" 2009 International Conference on Advances in Recent Technologies in Communication and Computing

[2] K.Saranya ,K.Premalatha, S.S.Rajasekar ," "A Survey on Privacy Preserving Data Mining", IEEE SPONSORED 2ND INTERNATIONAL CONFERENCE ON ELECTRONICS AND COMMUNICATION SYSTEM (ICECS 2015)

[3] Ms. Dhanalakshmi.M Mrs.Siva Sankari.E:" Privacy Preserving Data Mining Techniques-Survey" ICICES2014 IEEE- S.A.Engineering College, Chennai, Tamil Nadu, India

[4] Majid Bashir Malik, M. Asger Ghazi, Rashid Ali:" Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects" 2012 IEEE Third International Conference on Computer and Communication Technology.

[5] Alpa Shah, Ravi Gulati :" Contemporary Trends in Privacy Preserving Collaborative Data Mining– A Survey.

[6] Radhika Kotecha, Sanjay Garg,V.V.P. :" Data Streams and Privacy:Two Emerging Issues in Data Classification", IEEE 2015 5th Nirma University International Conference on Engineering (NUiCONE)

[7] Wee Siong Ng #, Huayu Wu #, Wei Wu #, Shili Xiang #, Kian-Lee Tan :" Privacy Preservation in Streaming Data Collection", 2012 IEEE 18th International Conference on Parallel and Distributed Systems

[8] Gaoming Yang, Jing Yang, Jianpei Zhang, Yan Chu : "Research on Data Streams Publishing of Privacy Preserving", College of computer science and technology Harbin Engineering University Harbin, China, IEEE2010

[9] Ching-Ming Chao1, Po-Zung Chen2 and Chu-Hao Sun2:" Privacy-Preserving Clustering of Data Streams", Tamkang Journal of Science and Engineering, Vol. 13, No. 3, pp. 349_358 (2010)

[10] Chen‐Yi Lin1, Yuan‐Hung Kao2, Wei‐Bin Lee2 and Rong‐Chang Chen3:" An efficient reversible privacy‐preserving data mining technology over data streams

[11] Zhenmin Lin, Jie Wang, Lian Liu, Jun Zhang,IEEE 2009:" Generalized Random Rotation Perturbation for Vertically Partitioned Data Sets",

[12] Keke Chen Ling Liu :"Privacy Preserving Data Classification with Rotation Perturbation", Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM'05), IEEE2005

[13] Junwei Zhang, Jing Yang, Jianpei Zhang,Yongbin Yuan," KIDS:K-anonymization data stream base on sliding window", IEEE 2010 2nd International Conference on Future Computer and Communication [Volume 2]

[14] Feifei Li, Jimeng Sun, Spiros Papadimitriou,† George A. Mihaila, Ioana Stanoi," Hiding in the Crowd: Privacy Preservation on Evolving Streams through Correlation Tracking", IEEE 2007 Boston University, §Carnegie Mellon University, †IBM T.J.Watson Research Center

[15] 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); Issn : 0975-3397 Vol. 3 No. 2 Feb 2011

