

Privacy Preserving Based on Geometric Transformation Using Data Perturbation Technique

Hirva Divecha

Computer Engineering

Parul Institute Of Engg. & Technology

Waghodiya, Vadodara, India

Sheetal Mehta

Assistant Professor, Computer Science & Engg

Parul Institute Of Engg. & Technology

Waghodiya, Vadodara, India

Abstract— Maintain confidentiality, privacy and security research in data mining (PPDM) is one of the biggest trends. Recent advances in data collection, data dissemination and related technologies have inaugurated a new era of research where existing data mining algorithms should be reconsidered from a different point of view, this of privacy preservation. We propose a simple PCA based transformation approach for various datasets to preserve privacy and maintain accuracy based on clustering. A privacy soil, the proposal to convert the data into nature, conservation saves. The accuracy of clustering before and after privacy preserving transformation was estimated.

Keywords- Privacy; Geometric; K-means clustering

I. INTRODUCTION

Data Mining efficiently discover valuable, non-obvious information from large datasets, is particularly vulnerable to abuse. A fruitful future research leadership in data mining is the development of technology that incorporates the concern for privacy. A recent survey of web users 17% of respondents as privacy fundamentalists, the unclassified data on a site, even if privacy measures are in place [1]. A more recent study of web users found that 86% of respondents believe that information for participation in benefits programs is a matter of individual choice privacy [2]. Nowadays organisms around the world are dependent on mining gigantic datasets. These datasets typically contain delicate individual information Inevitably All is exposed to the various parties. Consequently privacy issues are constantly in the limelight and the public dissatisfaction May well threaten the exercise of data mining. It is of great importance used technical security to protect the confidentiality of individual values for data mining for the development of appropriate Malthus.

There is much research on privacy-preserving data mining (PPDM) [6] malfunctioning, randomization and secure multi-party system based calculations. More recently, there has been much research on anonymity

Including k-anonymity and l-diversity. As a result, we now have numerous privacy and anonymity preserving algorithms.

Many government agencies, businesses and non-profit organizations to support their short-and long-term schedule activities, to collect for a way to store, analyze and report data on persons, households or businesses looking. Information systems therefore contain confidential information such as social security numbers, income, credit ratings, type of illness, customer purchases, etc., that 'need to be adequately protected. With the Web revolution and the emergence of data mining, have privacy concerns provided technical challenges fundamentally different from those that occurred before the information age [3].

The understanding of privacy in data mining requires understanding how privacy can be violated [5], the can means clustering and clustering on the prevention of invasion of privacy. Usually carries a significant factor in breach of Private in data mining :the misuse of data. Users' privacy can be violated in different ways and with different intentions. Although data mining can be very useful in many applications, it is also on the lack of adequate safeguards may violate informational privacy. Privacy can be violated are when personal data for other purposes after the original transaction brokers are an individual and an organization if the information was collected, used. Malthus when personal data are exposed to mining, the attribute values associated with private persons and must be protected from disclosure.

Users' privacy can be violated in different ways and with different intentions. Although data mining can be extremely valuable in many applications it can also, in the absence of adequate safeguards, violate informational privacy. Privacy can be violated if personal data are used for other purposes subsequent to the original transaction between an individual and an organization when the information was collected. Thus, when personal data are subjected to mining, the attribute values associated with individuals are private and must be protected from disclosure [4].

One of the sources of privacy violation is

called data magnets. Data magnets are techniques and tools used to collect personal data. Examples of data magnets include explicitly collecting information through on-line registration, identifying users through IP addresses, software downloads that require registration, and indirectly collecting information for secondary usage. In many cases, users may or may not be aware that information is being collected or do not know how that information is collected. In particular, collected personal data can be used for secondary usage largely beyond the users' control and privacy laws. This scenario has led to an uncontrollable privacy violation not because of data mining itself, but fundamentally because of the misuse of data.

Securing against unauthorized accesses has been a long-term goal of the database security research community and the government research statistical agencies. Solutions to such a problem require combining several techniques and mechanisms. In an environment where data have different sensitivity levels, this data may be classified at different levels, and made available only to those subjects with an appropriate clearance.

Clustering is a well-known problem in statistics and engineering, namely, how to arrange a set of vectors (measurements) into a number of groups (clusters). Clustering is an important area of application for a variety of fields including data mining, statistical data analysis and vector quantization. The problem has been formulated in various ways in the machine learning, pattern recognition optimization and statistics literature. The fundamental clustering problem is that of grouping together (clustering) data items that are similar to each other. Given a set of data items, clustering algorithms group similar items together. Clustering has many applications, such as customer behavior analysis, targeted marketing, forensics, and bioinformatics [7].

Small companies have recognized the value in data, especially with the introduction of the knowledge discovery process. However, small companies do not have enough expertise for doing data analysis, although they have good domain knowledge and understand their data.

II. GEOMETRIC DATA PERTURBATION:

Def.: Geometric data perturbation consists of a sequence of random geometric transformations, including multiplicative transformation (R), translation transformation (Ψ), and distance perturbation Δ. $G(X) = RX + Ψ + Δ$ [8].

The data is assumed to be a matrix $A_{p \times q}$, where each of the p rows is an observation, O_i , and each observation contains values for each of the q attributes, A_i . The matrix may contain categorical and numerical attributes. However, our Geometric Data Transformation Methods rely on d numerical attributes, such that $d \leq q$. Thus, the $p \times d$ matrix, which is subject to transformation, can be thought of as a vector subspace V in the Euclidean space such that each vector $v_i \in V$ is the form $v_i = (a_1; \dots; a_d)$, $1 \leq i \leq p$, where $\forall i a_i$ is one instance of A_i , $a_i \in R$, and R is the set of real numbers. The vector subspace V must be transformed before releasing the data for clustering analysis in order to preserve privacy of individual data

records. To transform V into a distorted vector subspace V' , we need to add or even multiply a constant noise term e to each element v_i of V [9].

Translation Transformation: A constant is added to all value of an attribute. The constant can be a positive or negative number. Although its degree of privacy protection is 0 in accordance with the formula for calculating the degree of privacy protection, it makes we cannot see the raw data from transformed data directly, so translation transform also can play the role of privacy protection [10].

Translation is the task to move a point with coordinates (X; Y) to a new location by using displacements (X0; Y0). The translation is easily accomplished by using a matrix representation $v' = Tv$, where T is a 2 x 2 transformation matrix depicted in Figure 1(a), v is the vector column containing the original coordinates, and v' is a column vector whose coordinates are the transformed coordinates. This matrix form is also applied to Scaling and Rotation .

Rotation Transformation: For a pair of attributes arbitrarily chosen, regard them as points of two dimension space, and rotate them according to a given angle θ with the origin as the center. If θ is positive, we rotate them along anti- clockwise. Otherwise, we rotate them along the clockwise.

Rotation is a more challenging transformation. In its simplest form, this transformation is for the rotation of a point about the coordinate axes. Rotation of a point in a 2D discrete space by an angle is achieved by using the transformation matrix depicted in Figure 1(b). The rotation angle is measured clockwise and this transformation effects the values of X and Y coordinates [11].

$$\begin{bmatrix} 1 & 0 & X_0 \\ 0 & 1 & Y_0 \end{bmatrix} \quad \begin{bmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{bmatrix}$$

Figure 2 (a) Translation Matrix (b) Rotation Matrix

The above two components, translation and rotation preserve the distance relationship. By preserving distances, a bunch of important classification models will be "perturbation-invariant", which is the core of geometric perturbation. Distance preserving perturbation may be under distance-inference attacks in some situations. The goal of distance perturbation is to preserve distances approximately, while effectively increasing the resilience to distance-inference attacks. We define the third component as a random matrix $\Delta_{d \times n}$, where each entry is an independent sample drawn from the same distribution with zero mean and small variance. By adding this component, the distance between a pair of points is disturbed slightly [12]

III. RESULTS 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 3.1: Dataset configuration to determine accuracy based on Membership Matrix.

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

Table 3.1 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 3.2, 3.3 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 3.2: Resultant accuracy of 5-Cluster

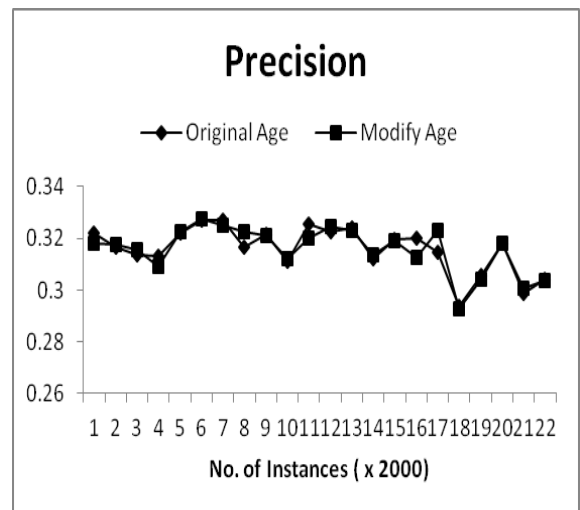
Dataset	Attributes	No. of	Stream	K-
	Age		2000	84.21%
	Balance			89.37%

Bank Management	Duration	5	3000	86.83%
	Age			81.96%
	Balance			84.64%
	Duration			81.01%

Table 3.3: Resultant accuracy of 3-Cluster

Dataset	Attributes	No. of	Stream	K-
Bank Management	Age	3	2000	87.13%
	Balance			92.18%
	Duration			89.41%
	Age	3000	85.56%	
	Balance		90.22%	
	Duration		87.64%	

Here we represent the accuracy of our method by evaluating the clustering measure provided with MOA framework. We focused on two important 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 presented 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-Means is applied in order to evaluate both cases by keeping number of clusters fix (K=5, K=3). Instances are processed in defined sliding window size.



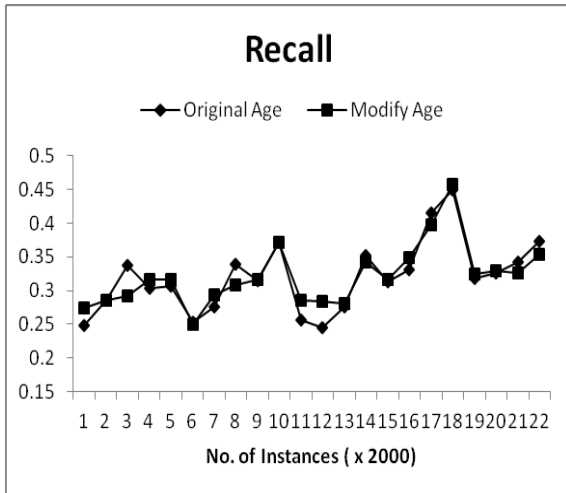


Figure 3.1: Accuracy on attribute Age in Bank Management with 5-Cluster

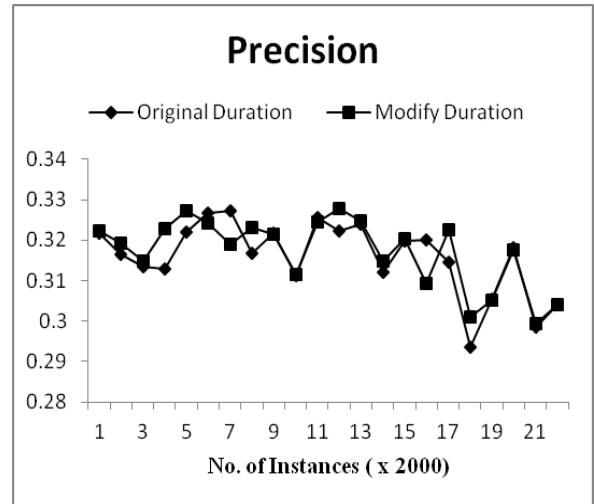


Figure 3.3: Accuracy on attribute Duration in Bank Management with 5-Cluster

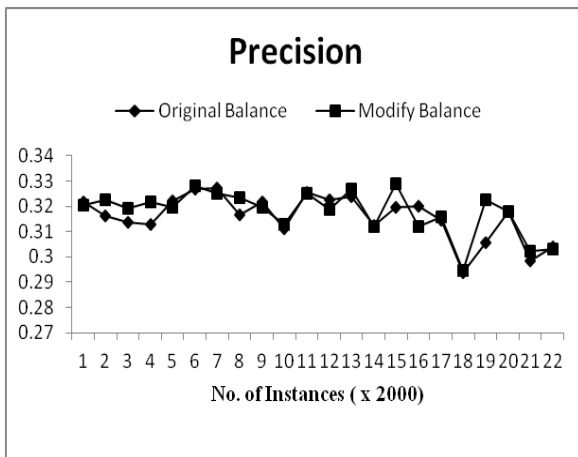
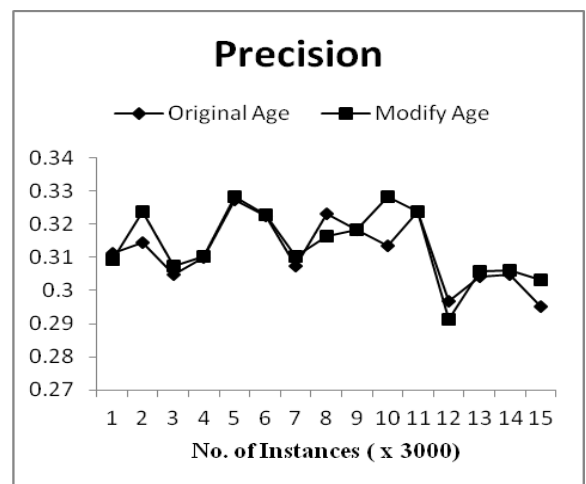
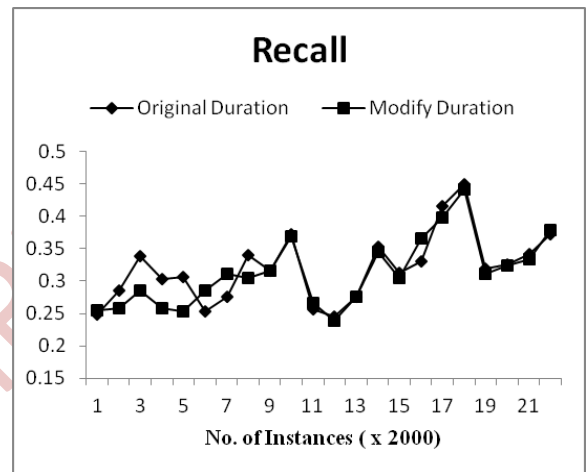


Figure 3.2: Accuracy on attribute Balance in Bank Management with 5-Cluster



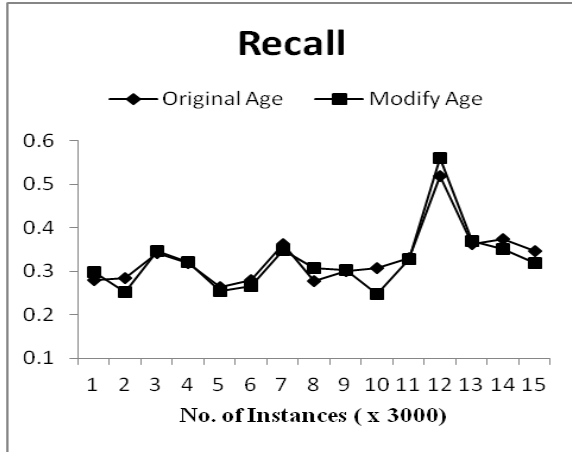


Figure 3.4: Accuracy on attribute Age in Bank Management with 5-Cluster

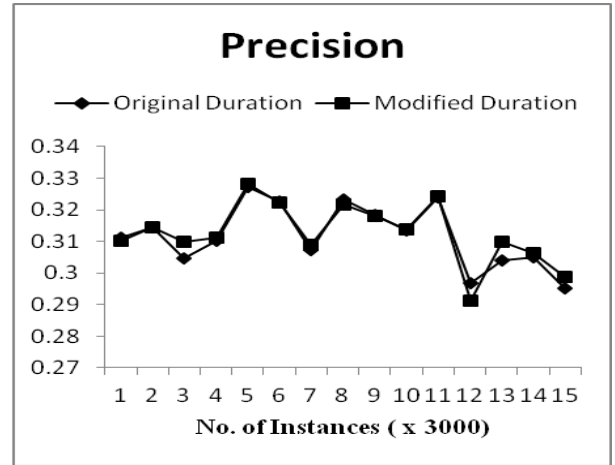


Figure 3.6: Accuracy on attribute Duration in Bank Management with 5-Cluster

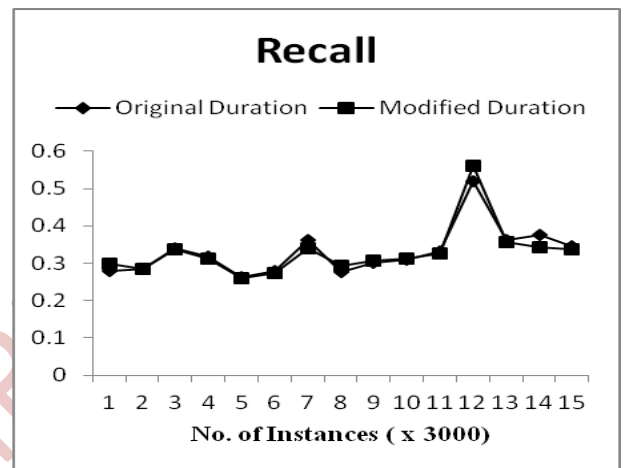
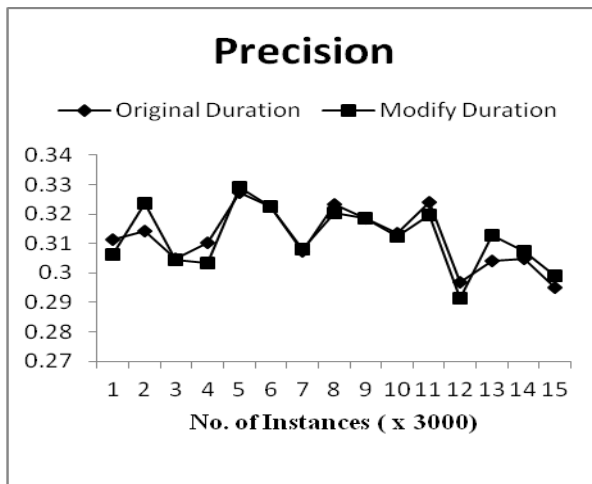
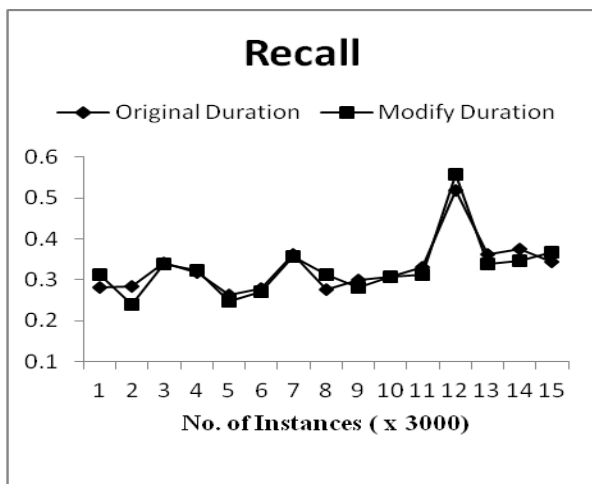


Figure 3.5: Accuracy on attribute Balance in Bank Management with 5-Cluster



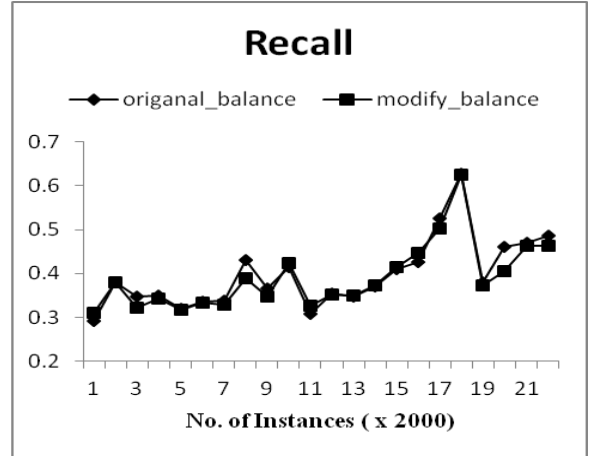
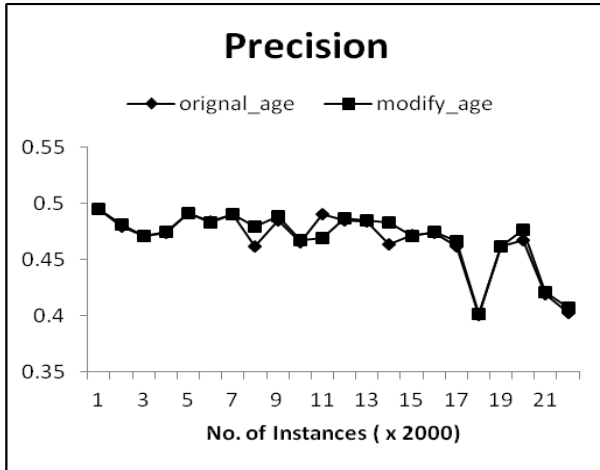


Figure 3.8: Accuracy on attribute Balance in Bank Management with 3-Cluster

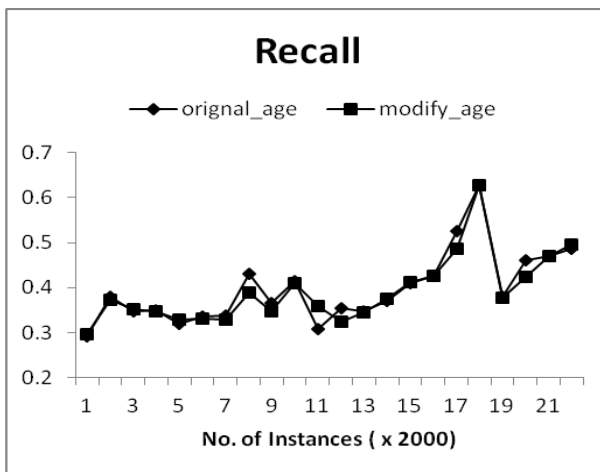


Figure 3.7: Accuracy on attribute Age in Bank Management with 3-Cluster

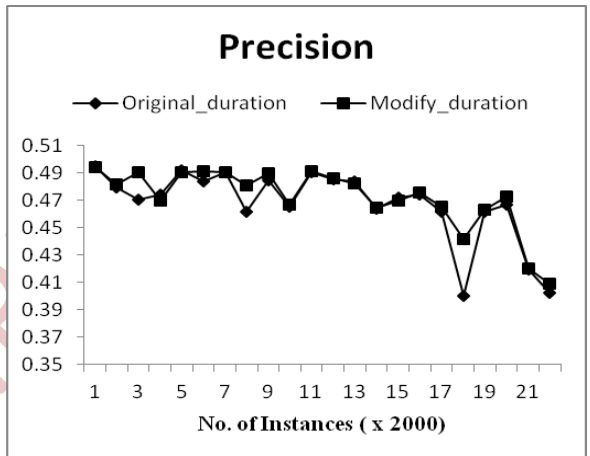


Figure 3.9: Accuracy on attribute Duration in Bank Management with 3-Cluster

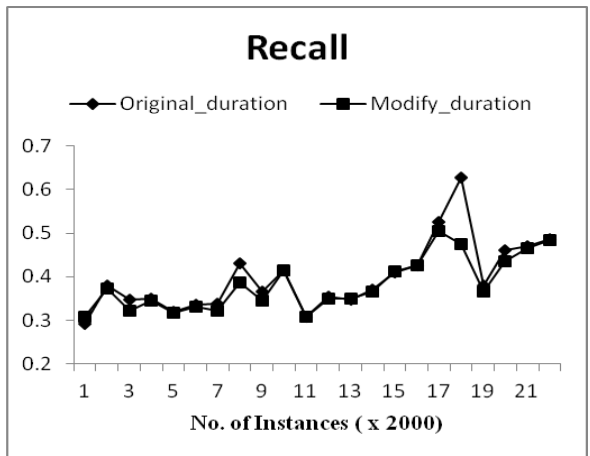
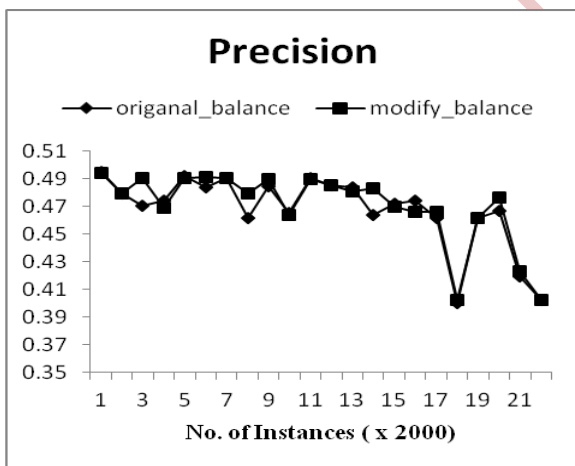


Figure 3.9: Accuracy on attribute Duration in Bank Management with 3-Cluster

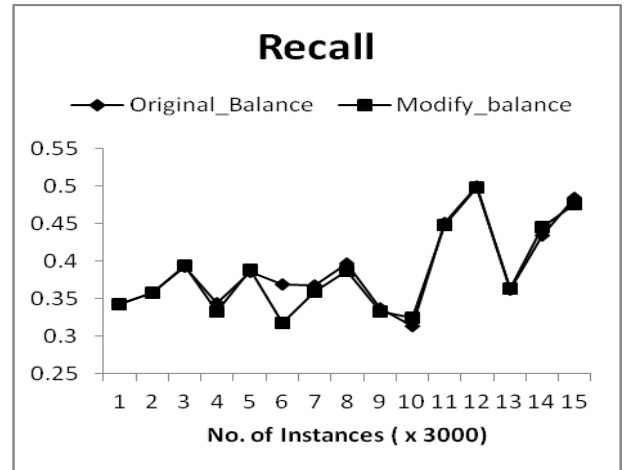
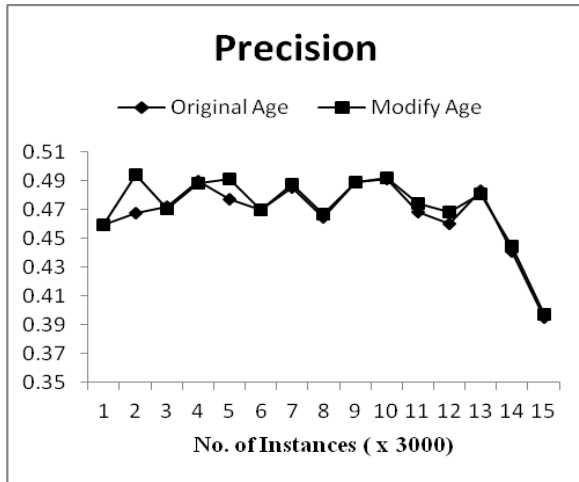


Figure 3.11: Accuracy on attribute Balance in Bank Management with 3-Cluster

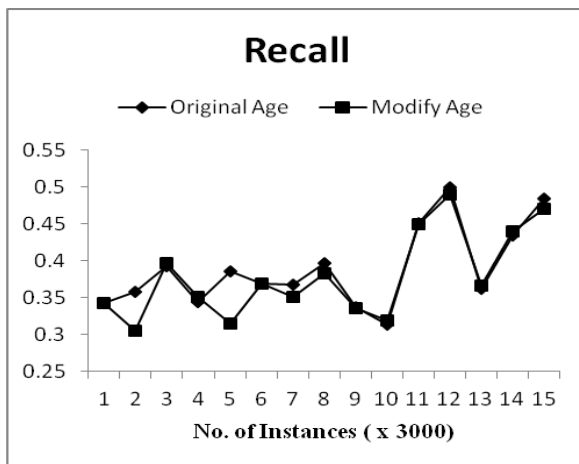
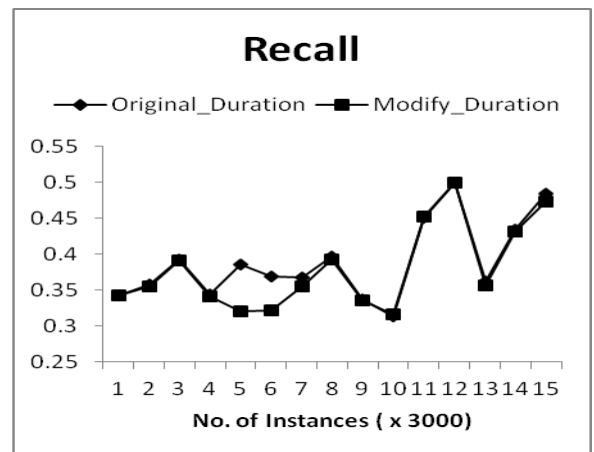
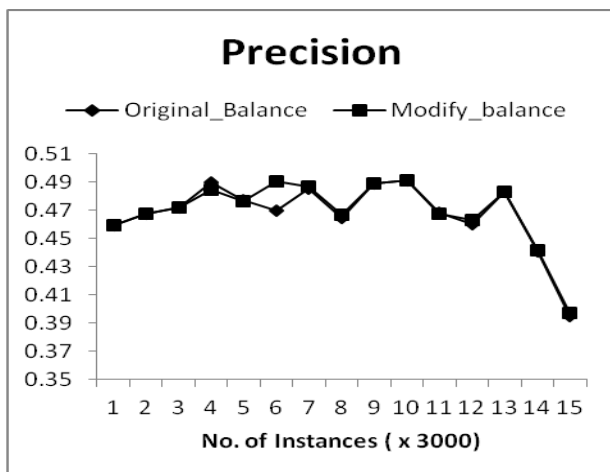
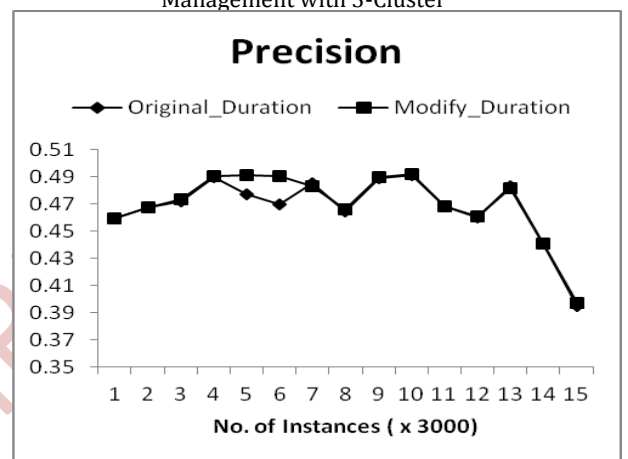


Figure 3.10: Accuracy on attribute Age in Bank Management with 3-Cluster



IV. CONCLUSION

The proposed privacy preserving paradigm has been successfully implemented in java and evaluated using Massive Online Analysis (MOA) under Windows 7 operating system. The arrived results were more significant and promising. The proposed method can be used to hide valuable information while presenting it on a publicly accessible place like internet. Further, the proposed model can be used to multi party collaborative clustering scenario. Some of the results of earlier works have shown, accuracy sometimes suffers as a result of security. But in the proposed method, the accuracy has been preserved and in some cases, the accuracy was almost equal to that of original data set.

REFERENCES

- [1] L.F.Cranor, J.Reagle and M.S.Ackerman. "Be-yond concern: Understanding netusers' attitudes about on line privacy". Technical Report TR99.4.3, AT&TLabs Research, April 1999.
- [2] A.F.Westin. "Freebies and privacy: What net users think". Technical report, Opinion ReSearch Corporation, July 1999. Available from <http://www.privacyexchange.org/iss/surveys/sr990714.html>.
- [3] Stanley R.M Oliveira, Osmar R. Zaiane, "Towards Standardization in privacy Preserving data Mining" In the proceeding of IEEE International Conference on data Mining 2006
- [4] Rakesh Agarwal, Ramakrishnan Srikant, "Privacy Preserving Data Mining "In Proceedings of the 1st International Conference on Knowledge Discovery and Databases
- [5] J P. Samarati, "Protecting respondent's privacy in micro data release", In IEEE Transaction on Knowledge and Data Engineering, 2001, pp.1010-1027.
- [6] V. S. Verykios, E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin, and Y. Theodoridis, "State-of-the-art in privacy preserving data mining", In Proc of ACM SIGMOD, 2004, pp. 50-57.
- [7] Ackerman, M. S., Cranor, L. F., and Reagle, J, "Privacy in e- commerce: examining user scenarios and privacy preferences", In Proc. EC99, 1999, pp. 1-8.
- [8] Vassilios S.Verylios, E.Bertino, Igor N,"State -of-the art in Privacy preserving Data Mining", published in SIGMOD 2004 pp.121-154.
- [9] L.Golab and M.T.Ozsu ,Data Stream Management issues-"A Survey Technical Report", 2003.
- [10] Keke Chen, Ling Liu," Geometric data perturbation for privacy preserving outsourced data mining", Springer, 2010.
- [11] Stanley R. M. Oliveira, Osmar R. Zaiane, Privacy Preserving Clustering by Data Transformation, February 2010
- [12] Jie Liu, Yifeng XU, "Privacy Preserving Clustering by Random Response Method of Geometric Transformation", IEEE 2010