Scientific Journal of Impact Factor(SJIF): 3.134

e-ISSN(O): 2348-4470
p-ISSN(P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 2,Issue 5, May -2015

Privacy Preserving Data Mining Using Random Rotation Based Data Perturbation Technique

AshaKatharotiya[1]Pankajkawadkar[2]

[1]*Department of Computer science &Engineering*[2]*H.O.D. computer science dept., PIES, Bhopal,*
[1,2]*PatelInstitute OfEngineering&Science, Bhopal,India*

Abstract---*To preserve privacy of data, privacy preserving data mining is the study of valid mining patterns and models which mask private information. There are many privacy preserving data mining techniques which have been studied.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. When has been modified, it would be necessary to rescan whole database, so it leads to more computation time and inability to respond the user fastly.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.*

Keywords:*Privacy;DataStreams;K-meansclustering*

I. INTRODUCTION

Inrecentyears,datamining is shown as apowerfuldataanalysistoolhasmaderemarkablecontributionsinmanyareasandhasthewideapplicationsviewpoint.Withthedevelopmentofdatabasetechnologyand networktechnology,alargenumberofusefuldata,whichcontains so muchindividualprivacyinformation,has been amassed invariousfields,suchaspatient'sconditioninformation,preferences to customer, personal backgroundinformation, account information etc.Oncetheinformationleaked,itwillbeunsafetoindividual.Iftheygivetheactualdatadirectlytotheprospectors,itwillpredictablyproduceprivateinformationdisclosure.Asthefieldofdataminingtechnologyextending,privacydisclosureproblembecomesworse,causingtheattentionofphasesofindustryandsocial.so,howtododatamining under thecircumstancesofprivacypreservinghasbecomeahotspotindatamining, so privacypreservingdatamining(PPDM)isintroduced.

Securingagainstunauthorized accesseshasbeenalong-termobjectiveofthedatabasesecurity, thegovernmentresearchstatisticalagencies and research community.Solutionstosuchadifficultyrequirecombiningseveraltechniquesandmechanisms.Inansituationwhereda tahavedifferentsensitivitystages,thisdatamaybeclassifiedat differentlevels, and it hasmade available onlytothosesubjectswithanappropriateconsent.

Technique of Clusteringisawell-knownproblematic situation instatisticsandengineering,namely,howtoarrangeasetofmeasurementsintoanumberofclusters.Clusteringisansignificantareaofapplicationforavarietyofarenasincludingdatamining, vectorquantization andstatisticaldataanalysis.Theproblemhasbeenframedinvariouswaysinthe,patternrecognition, machinelearning, optimizationandstatisticsliterature.Thefundamentalclusteringproblemisthatofassembletogetherdataitemsthatarerelatedtoeachother.Givenasetofdataitems,clusteringalgorithms can groupanalogousitemstogether.Clusteringhasmanyapplications,suchasanalysis of customerbehavior,targetedmarketing,forensics,andbioinformatics.

By miningsensitivecharacteristicsfromtheoriginaldatabase, Reconstructionbasedapproachesgenerateprivacyawaredatabase.Theseapproaches have been generatedsmallersideeffectsindatabasethanheuristicapproach.Reconstructionbasedmethodsperturbtheoriginaldatato achieveprivacypreserving.Theperturbeddatawouldmeetthetwocircumstances.First,anattackercannotdeterminetherealoriginal datafromtheissuanceofthealterationdata.Secondly, thealtereddatais stillpreservingsomestatisticalpropertiesoftheoriginal data, namelysomeof theinformationderivedfromthepartialdataareequivalenttodataacquiredfromtheoriginalinformation.Perturbingthedataforpreventingprivacyisveryfertiletechnique usedbymanyresearchers.Itisalsocapabletoreconstructthedistributionsatancumulativelevelinordertoperformthemining.

@IJAERD-2015, All rights Reserved

Therearethreetypesofdataperturbationapproaches:RotationPerturbation,ProjectionPerturbationandGeometricData Perturbation.

II.ROTATIONPERTURBATION

Inordertopreservedataprivacy inrotationperturbationmethod the datamatrixismultipliedbyarandomrotationmatrixbeforepublishing. There is Oneadvantageofthis approachisthatitcan bepreservethegeometricpropertiesofthedatamatrix,sofewcategoriesofclassifierswhicharebasedonthegeometricpropertiesofth edatacan be achievecomparableaccuracyonthealtereddataasthatontheoriginaldata.Supposeoriginaldatasethavedcolumnand N records representedasXd×N, the rotationperturbation ofthedatasetXwillbedefinedasG(X)=RX[1],here, randomrotationorthonormalmatrix is Rd×d.Akeyfeatureof rotationtransformationis preservingthe Euclidean distance, geometricshapehyperinamulti-dimensionalspace andinnerproduct.

PCA(Principalcomponentanalysis)is technique that is usedtodiscomposethemultidimensionaldataintolowerdimensions.PCAassumesthatalltheinconsistencyinaprocessshouldbeuse dintheanalysis so itbecomesdifficulttodistinguishthecrucialvariablefromthelessvital.Subsequently, Principle componentanalysis

replacestheoriginalvariablesofadatasetwiththesmallernumberofuncorrelatedvariablescalledtheprinciplecomponent[3].

PCA isastandardtoolinmoderndataanalysis task-invariousfieldsfromneurosciencetocomputergraphics.Sinceitis asimple, on-parametricmethodforextractingrelevantinformationfrommystifyingdatasets. Withminimaleffort PCAprovidesaroadmaps forhowto reduce acomposite datasettoalowerdimensiontorevealthesometimessimplified and hidden structuresthatoftenunderlieit.PrincipalComponentAnalysis(PCA)issuitablefortransformingthemultidimensionaldataintolowe rdimensions.itisastandardtoolinmoderndataanalysis.PCAassumesthatalltheinconsistencyinaprocessshouldbeusedintheanalys isso.Itbecomeschallenging tasktodifferentiatetheimportantvariablefromthelessimportant for that PCAis mostappropriateforusualdistributions(wherelinearPrinciple Component Analysisapproachprovidesthebestpossiblesolution).Accordingly,PrincipleComponentAnalysisreplacestheoriginalvariableso fadata setwithalessernumber of uncorrelatedvariablescalled as the "principlecomponents".Iftheoriginaldatasetofdimension Dcontainshighlyassociatedvariables,thenthereisanoperationaldimensionality exist as,d<D,thatexplainsalloftheta data. Thepresence ofonlya fewcomponentsofdmakes which easiertolabeleachdimensionwithanintuitivemeaning.Furthermore,its moreeffectivetooperateonfewervariablesinsubs equentanalysis.

III.RESULTSANDDISCUSSION

ino r d e r to evaluate the clusteringaccuracy ,Seriesoftrialswereperformedoverdefinesslidingwindow size (w) .Ourevaluationapproachfocusedontheinclusivequalityofgeneratedclustersafterdatasetperturbation.Experimentwasbasedonfoll owingsteps:

1. InMOAframework, Setupeachdatasetasstream.
2. To evaluate measures and cluster membership matrix, definesliding window(w)overthedatastream.
3. Alteredalltheoccurrencesinslidingwindowbyapplyingproposeddataperturbationmethodtoprotectthesensitivecharacteristic value.
4. Tofindtheclustersforourperformanceevaluation, K-Meansclusteringalgorithmisused.Ourselectionwasinfluencedby (a) K-Means isoneof the best known clustering Algorithm and it is also scalable.
   (b)The Numberofclusterfound from originalandperturbeddatasetwastakensameasnumberofcluster.

Make Comparison thathowcloselyeachclusterin theperturbeddatasetmatchesitscorresponding cluster intheOriginaldataset.Bycomputingthe F-measure, weexpressedthequalityofthe generated cluster.

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

@IJAERD-2015, All rights Reserved

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.

Table 1.2: Resultant accuracy of 5 Cluster

Dataset	Attributes	No. of Cluster	Stream Data	K-Means
Bank Mana-gement	Age	5	2000	85.21%
	Balance			89.39%
	Duration			86.81%
	Age		3000	82.96%
	Balance			83.64%
	Duration			81.02%

Table 1.3: Resultant accuracy of 3 Cluster

Dataset	Attributes	No. of Cluster Cluster	Stream Data	K-Means
Bank Mana-gement	Age	3	2000	88.13 %
	Balance			92.60 %
	Duration			89.42%
	Age		3000	85.59%
	Balance			91.22%
	Duration			89.64%

@IJAERD-2015, All rights Reserved

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

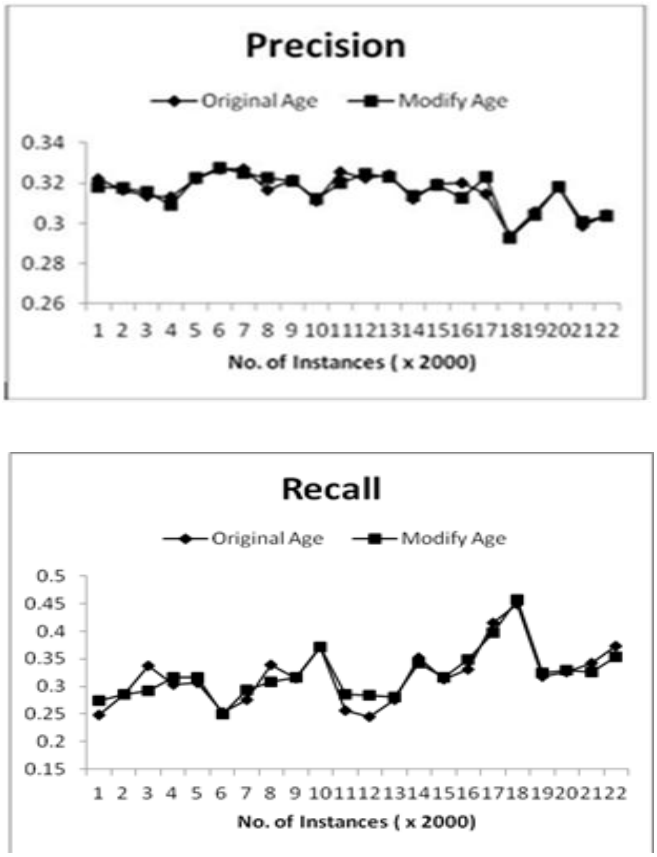
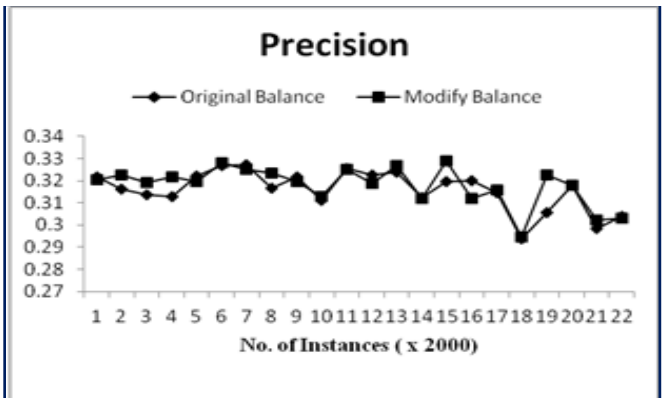


Fig 1.1: Accuracy on attribute Age in Bank Management with 5-Cluster



@IJAERD-2015, All rights Reserved

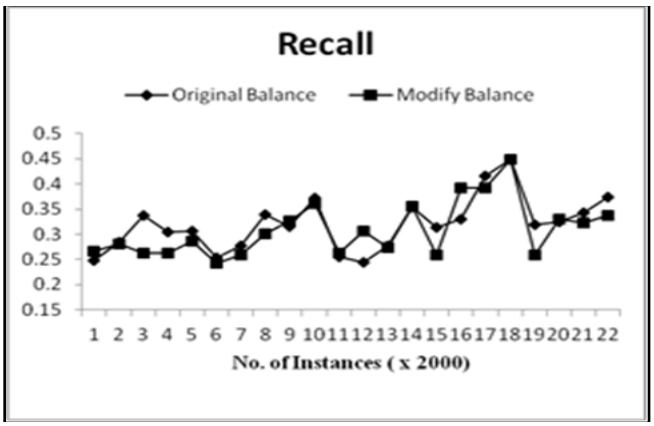
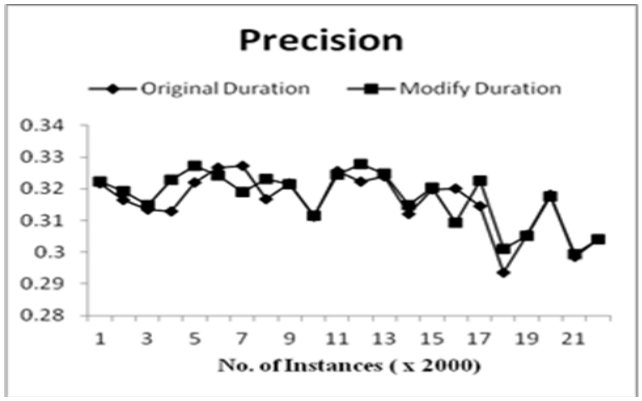


Fig.1.2: Accuracy on attribute Balance in Bank Management with 5-Cluster



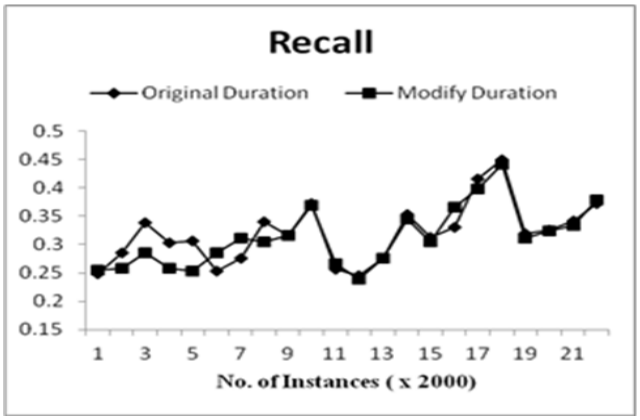


Fig 1.3: Accuracy on attribute Duration in Bank Management with 5-Cluster

@IJAERD-2015, All rights Reserved

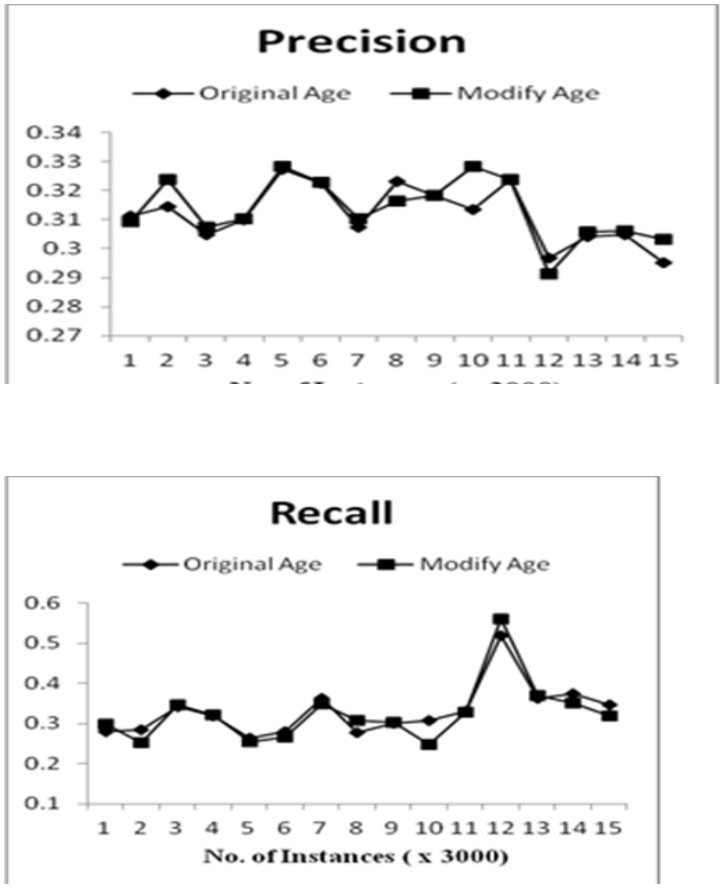
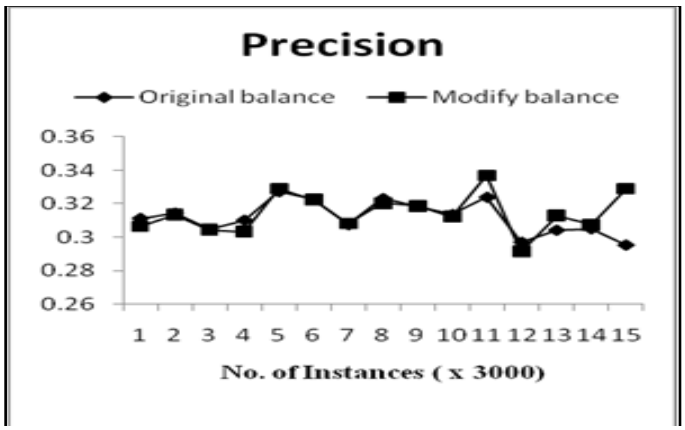


Fig.1.4: Accuracy on attribute Age in Bank Management with 5-Cluster



@IJAERD-2015, All rights Reserved

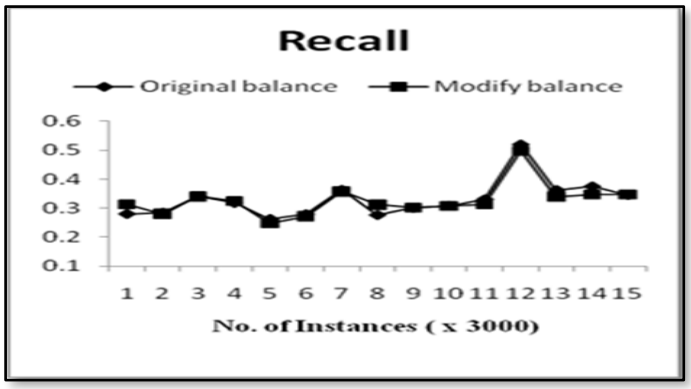
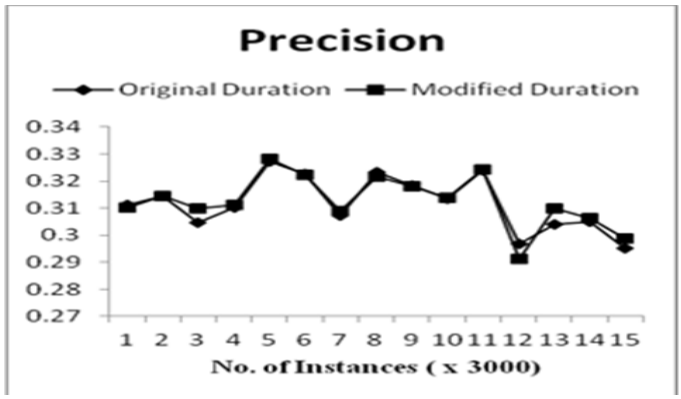


Fig.1.5: Accuracy on attribute Balance in Bank Management with 5-Cluster



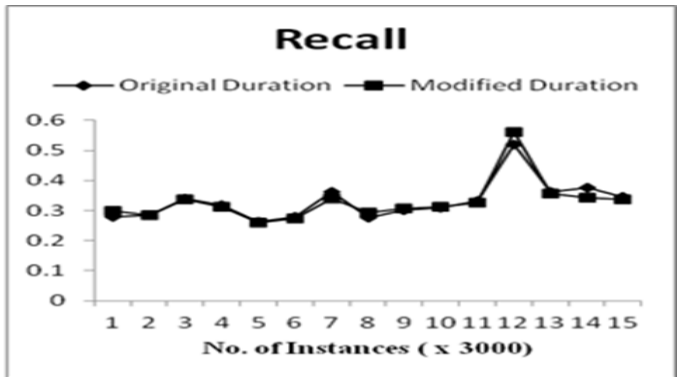


Fig. 1.6: Accuracy on attribute Duration in Bank Management with 5-Cluster

@IJAERD-2015, All rights Reserved

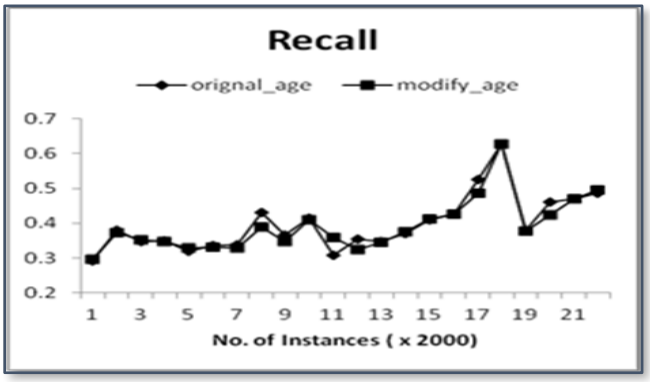
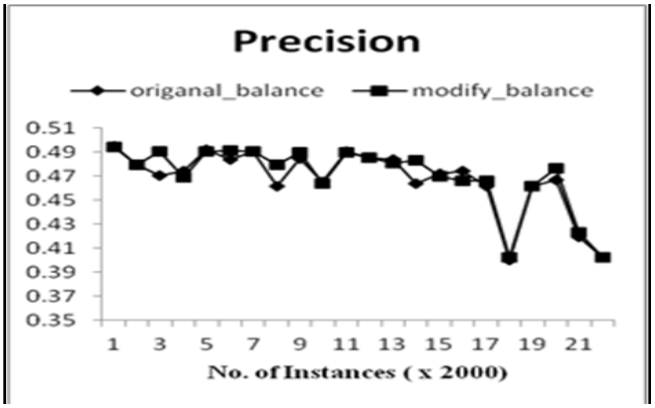


Fig. 1.7: Accuracy on attribute Age in Bank Management with 3-Cluster



@IJAERD-2015, All rights Reserved

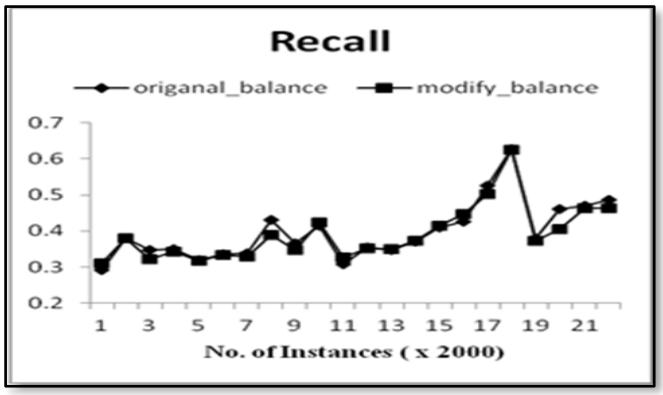
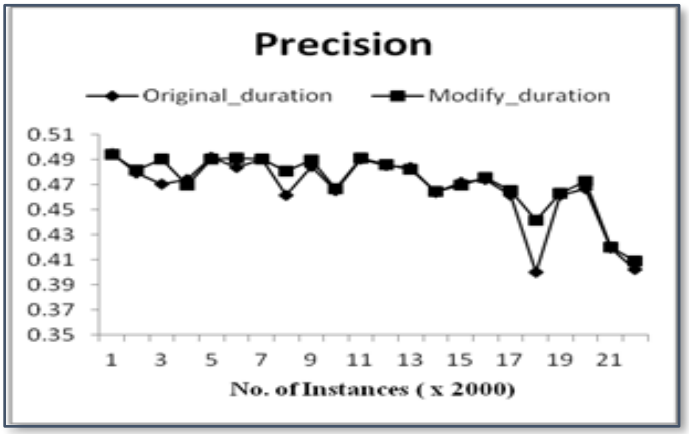


Fig. 1.8: Accuracy on attribute Balance Bank Management with 3-Cluster



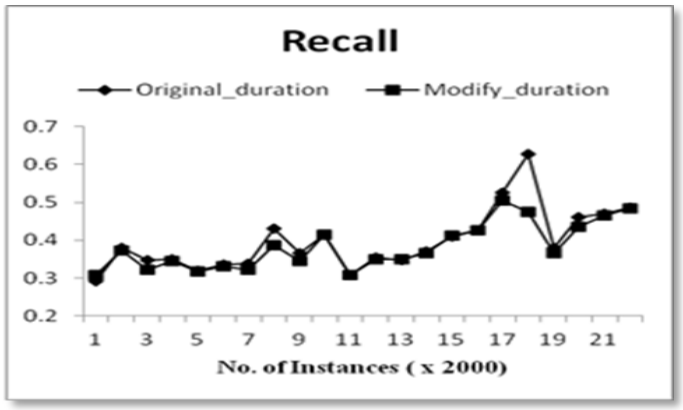
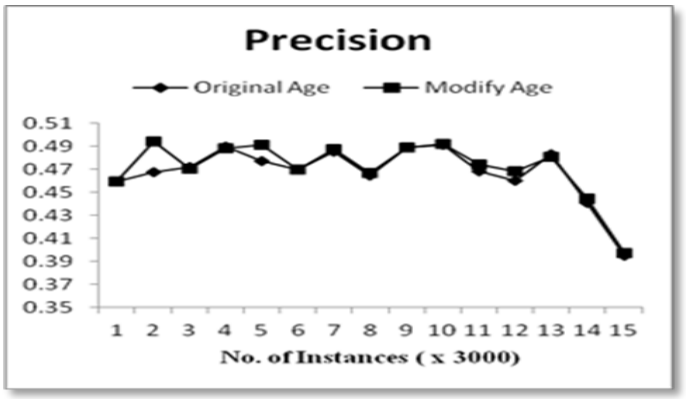


Fig. 1.9: Accuracy on attribute Duration in Bank Management with 3-Cluster

@IJAERD-2015, All rights Reserved



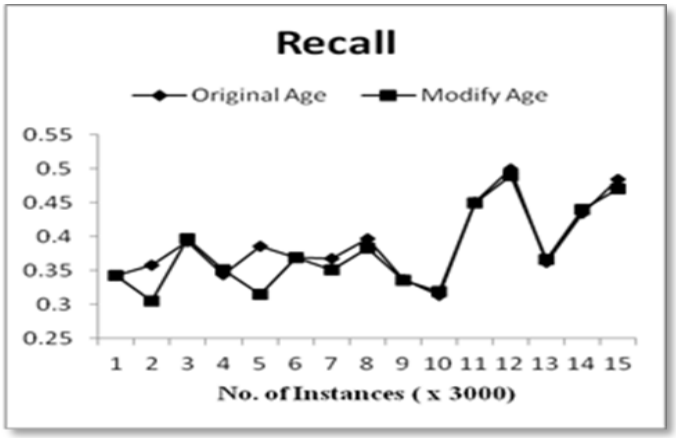
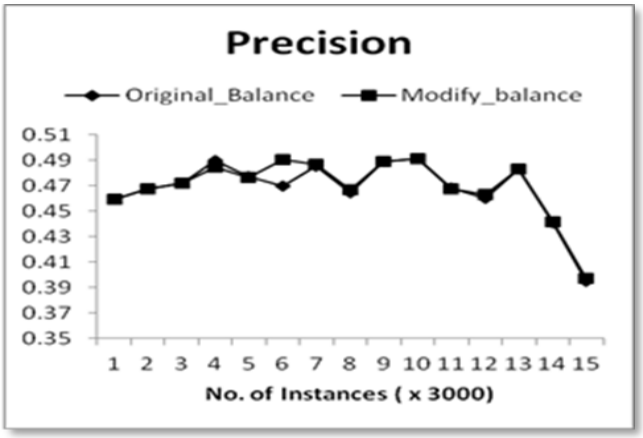


Fig. 1.10: Accuracy on attribute Age in Bank Management with 3-Cluster



@IJAERD-2015, All rights Reserved

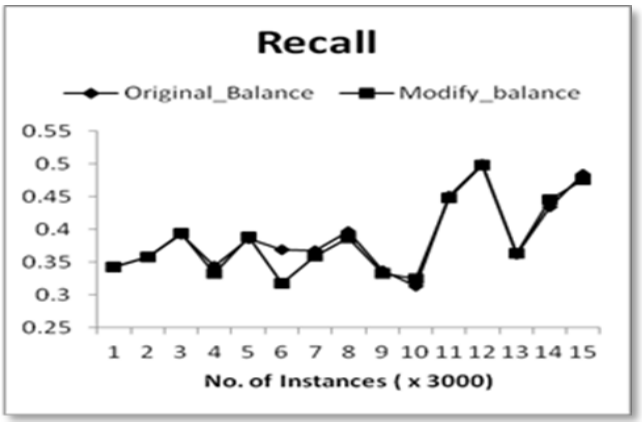
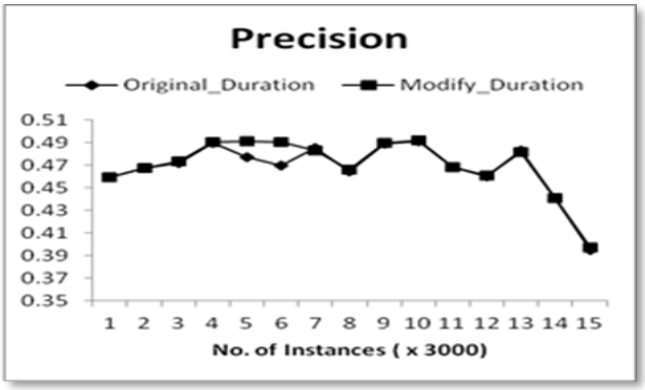


Fig. 1.11: Accuracy on attribute Balance in Bank Management with 3-Cluster



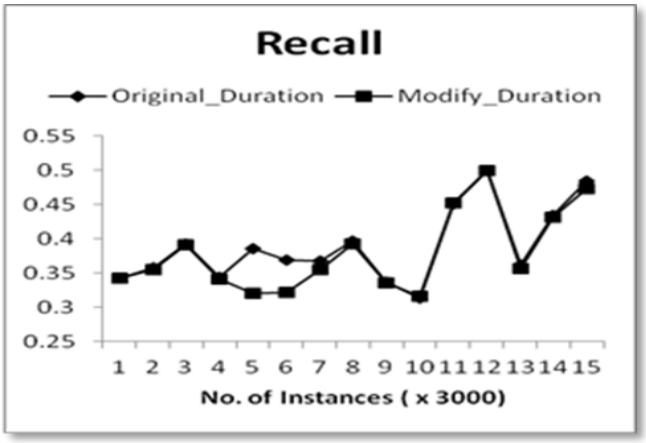


Fig 1.12: Accuracy on attribute Balance in BankManagement with 3-Cluster

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

@IJAERD-2015, All rights Reserved

[1] Majid Bashir Malik And M. Asger Ghazi And Rashid Ali ;"Privacy Preserving Data Mining Techniques: Current Scenario And Future Prospects"; Third International Conference On Computer And Communication Technology; 978-0-7695-4872-2/12 $26.00 © 2012 Ieee

[2] Hitesh Chhinkaniwala And Dr. Sanjay Garg "Privacy Preserving Data Mining Techniques: Challenges & Issues" In Proceedings Of International Conference On Computer Science & Information Technology, Cslt – 2011,P.609

[3] Chirag N. Modi,UdaiPratapRao And DhirenR.Patel "Maintaining Privacy And Data Quality In Privacy Preserving Association Rule Mining", In 2010 Second International Conference On Computing, Communication And Networking Technologies

[4] W.T. Chembian1, Dr. J.Janet, "A Survey On Privacy Preserving Data Mining Approaches And Techniques",In Proceedings Of The Int. Conf. On Information Science And Applications Icisa 2010,6 February 2010, Chennai, India

[5] Xiaolin Zhang And Hongjing Bi; "Research On Privacy Preserving Classification Data Mining Based On Random Perturbation"; International Conference On Information, Networking And Automation (Icina); 978-1-4244-8106-4/$26.00 © 2010 Ieee

[6] Ching-Ming Chao, Po-Zung Chen And Chu-Hao Sun ;"Privacy-Preserving Classification Of Data Streams"; Tamkang Journal Of Science And Engineering; Vol. 12, No. 3, Pp. 321_330 (2009)

[7] M. Naga Lakshmi, K Sandhya Rani;" Privacy
Issn: 2319-8753 Vol. 2, Issue 9, September 2013

[8] Albert Bifet, Geoff Holmes, Bernhard Pfahringer, Philipp Kranen, Hardy Kremer, Timm Jansen, And Thomas Seidl "Moa: Massive Online Analysis, A Framework For Stream Classification And Clustering."

[9] Vassilios S. Verykios, Elisa Bertino, Igor NaiFovino,LoredanaParasilitiProvenza, YucelSaygin, YannisTheodoridis "State-Of-The-Art In Privacy Preserving Data Mining"

[10] Haisheng Li;" Study Of Privacy Preserving Data Mining" ;Third International Symposium On Intelligent Information Technology And Security Informatics; 978-0-7695-4020-7/10 $26.00 © 2010 Ieee

[11] Jian Wang ,YongchengLuo ,Yan Zhao Jiajin Le;" A Survey On Privacy Preserving Data Mining"; 2009 First International Workshop On Database Technology And Applications; 978-0-7695-3604-0/09 $25.00 © 2009 Ieee

[12] MohammadrezaKeyvanpour, SomayyehSeifiMoradi;" 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

[13] Keke Chen Ling Liu;" A Random Rotation Perturbation Approach To Privacy Preserving Data Classification "

[14] S. Kasthuri, T. Meyyappan;" Detection Of Sensitive Items In Market Basket Database Using Association Rule Mining For Privacy Preserving"; Proceedings Of The 2013 International Conference On Prim, February 21-22; 978-1-4673-5845-3/13/$31.00©2013 Ieee

[15] Nikunj H. Domadiya;" Hiding Sensitive Association Rules To Maintain Privacy And Data Quality In Database"; 978-1-4673-4529-3/12/$31.00_C 2012 Ieee

RahenaAkhter, RownakJahanChowdhury, Keita Emura, Tamzida Islam, Mohammad ShahriarRahman, NusratRubaiyat;" Privacy-Preserving Two-Party K- Means Clustering In Malicious Model"; 2013 Ieee 37th Annual Computer Software And Applications Conference Workshops; 978-0-7695-4987-3/13 $26.00© 2013 Ieee

[17] JaideepVaidya , BasitShafiq ;" A Random Decision Tree Framework For Privacy-Preserving Data Mining"; 1545-5971/13/$31.00 © 2013 Ieee
2013

@IJAERD-2015, All rights Reserved