Privacy preservation in clustering with distributed database

Metar Javed H
Department of computer science and engineering,
B.H.Gardi College of Engineering and Technology, Rajkot, India
javedmetar@gmail.com

***Abstract*— Data mining has been a popular research area for more than a decade due to its vast spectrum of application. In this we analyzed privacy preserving techniques for clustering in distributed environment. In this work, we analyzed and compare following methods: 1) Principal Component Analysis (PCA) 2) Dimensionality Reduction Based Transformation (DRBT) 3) Shamir's Secret Sharing (SSS).**

Key words: Data mining; Privacy; distributed database; security; PCA; DRBT; secret sharing.

I. INTRODUCTION

Data mining and knowledge discovery in databases are two new research areas that investigate the automatic extraction of previously known patterns from large amounts of data. Because of this privacy comes in demand.

The main goal of the privacy is to protect data and knowledge after the mining process

II. CLASSIFICATION OF PRIVACY PRESERVATION [12]

We can classify them based on following dimensions:

1. Data distribution
2. Data modification
3. Data mining algorithm
4. Data or rule binding
5. Privacy preservation

In data distribution, data is distributed horizontally or vertically. In horizontal distribution same attributes of different entities are used (row from table). In vertical partitioned different attributes of same entities are used.

In data modification, modify the original values of database that needs to secure. Different techniques for data modification :

1. Perturbation, which a category of data modification approach that protect sensitive data contained in data set by modifying a selected portion of attributes of its transaction (I.e. changing a 1-value to 0-value,or adding a noise),
2. Blocking, which is the replacement of an existing attribute value with a"?",
3. Aggregation or merging which is the combination of several values in to coarser category.
4. Swapping, refers to interchanging values of individuals records, and
5. Sample, which refers to releasing data for only a sample of a population.

The **third dimension** refers to the data mining algorithm, for which the data modification is taking place.
The **fourth dimensions** refer to whether raw data or aggregated data should be hidden.
The **last dimension** which is the most important refers to privacy preservation technique used for the selective modification of the data .The techniques that have been applied for this reason is:
Heuristic based technique like adaptive modification that modifies only selected value that the utility loss rather than all available values.
Cryptography based techniques likes secure multi party computation where a computation is secure if at the end of the computation, no party knows anything except its own input and the result.
Reconstruction base techniques where the original distribution of the data is reconstructed from the randomized data

III. CLUSTERING

Cluster analysis is a technique for assigning data objects into related groups such that the objects within each group exhibit similar characteristics. Cluster analysis addresses the problem of arranging a set of vectors into a number of clusters. Clustering has a wide range of application in various fields, such as marketing, insurance, finance, medicine, etc.

IV. PRIVACY PRESERVING CLUSTERING

PPC is used to protect the sensitive attributes value object that is subjected to clustering analysis.oliveira and zaiane explain different types of scenarios in PPC, these scenarios are the following:

1. PPC over horizontally partitioned data

2. PPC over vertically partitioned data
3. PPC over centralized data

V. PRINCIPAL COMPONENT ANALYSIS [14]

5.1. Introduction

Principal Component Analysis (PCA) is a technique that preserves the privacy of sensitive information in a multi-party clustering scenario is proposed. The performance of this technique is evaluated further by applying a classical k-means algorithm, as well as machine learning- based clustering method on synthetic and real world data set. The accuracy and efficiency of clustering is computed before and after privacy-preserving transformation.

5.2 Pseudo code for evaluating the framework

Steps involved in implementing and evaluating the analyzed methodology is given below.

1. Create a multi-attribute synthetic data set S with dimension of d using the Gaussian distribution function. Let N=lSl and C= {C1, C2... Cc} be the known classes of S.
2. Select sample s≤S such that S∩Ci≠0, i=1, 2... c. (put true symbol)
3. Create a transformation matrix T's corresponding to S using a PCA-based transformation such that dim (T's) =d1<d.
4. Generate a shifted transformation matrix Tr from T's by using a shift factor 'r'(if necessary)
5. Project S on T's (or Tr) to obtain the new reduced dimensional data set Ss.
6. Cluster the original data set S by using K-means and SOM algorithms and let the new cluster be C1 and C2, respectively.
7. Cluster the transformed data set As by using K-means and SOM algorithms and let the new cluster be C3 and C4, respectively.
8. Obtain the Rand Index (RI) for the pairs (C, C1), (C, C2), (C, C3) and (C, C4), and estimate the accuracy of clustering
9. Repeat from step 1 by changing the parameters d and N.

5.3 Conclusion

This approach was successfully implemented for privacy-preserving clustering of centralized and horizontally partitioned data. This method provides good security. The analyzed method can be used to mask sensitive information while presenting it on publicly accessible platform such as the internet. Future research may address other scenarios such as vertically partitioned data.

VI. DIMENSIONALITY REDUCTION BASED TRANSFORMATION [17]

6.1 Introduction

Directionality reduction based transformation used random projection to protect the attribute values subjected to cluster analysis. The major features of this method are:1)this method is independent of distance based clustering algorithm;2)this method has a sound mathematical foundation; and 3)it does not require CPU-intensive operations.

6.2 Privacy preserving clustering over centralized data

DRBT performs three major steps:

1. Remove identifier**:** Attributes that are not subjected to clustering are removing.
2. Reducing the dimensional of original data set: Original data set is then transformed in to transformed data set using random projection.
3. Calculate stress function: This function is used to determine that the accuracy of the transformed data is marginally modified, which guarantee.

6.3 Privacy preserving clustering over vertically partitioned data

Privacy preserving clustering on vertically partitioned data is performed as follows:

1. Individual transformation: if k>2, share their data in a collaborative project for clustering, each party kind must transform its data according to the step in section 6.2.
2. Data exchanging or sharing: once the data are disguised by using random projection, the k parties are able to change the data among themselves. However, one party could be the central one to aggregate and cluster the data
3. Sharing clustering results: after the data have been aggregated and mined in a central party kind, the result could be shared with the other parties

6.4 Conclusion

This method provides good security, communication cost and accuracy. The main feature of DRBT are as follows:1)it is independent of distance based clustering algorithms;2)it has a sound mathematical foundation;3)it does not requires CPU-intensive operations; and 4)it can be applied to address PPC over centralized data and PPC over vertically partitioned data.

VII. SHAMIR'S SECRET SHARING [21]

7.1 Introduction

Shamir's secret sharing is a scheme where secret is dividing into parts, after that shares the secret to the selected party and then reconstruct the secret using some or all parts of secret. The scheme is formally described as follows:
The secret is D.
Divide D into n pieces such a way that:
1) Knowledge of any K or more did pieces make D easily computable;
2) Knowledge of any K-1 or fewer did pieces leaves D completely undetermined.

This scheme called a (k,n) threshold scheme. To divide it into pieces, we pick random k-1 degree polynomial(x) =ao+a1x+...+ak-1xx^k-1.and evaluate
D1=q (1)...Did=q (I)...Dan=q (n).

7.2 Example

- **Generate share**
  Let D=5678, n=5(number of secret), k=3(number of secret required for reconstruction)
  Randomly we choose two number; 2, 3. (a1=2, a2=3)
  Our Polynomial to produce secret share (points) is therefore
  F(x) =5678+2x+3x^2

  We construct 5 points from the polynomial(1,5683);(2,5694);(3,5711);(4,5734);(5,5813) ;
  We give each participant a different single point (both x and f(x))

- **Reconstruction**
  To reconstruct the secret we need three point (x0,y0)=(1;5683);(x1,y1)=(2,5694);(x2,y2)=(3,5711).

  Now calculate Lagrange basic polynomial
  l0= (x^2/2)-(5x/2) + 3
  l1= -x^2+4x-4
  l2=(x^2/2)-3x/2+1 then for

  F(x) =∑yiLi (a)
  After calculating F(x), we get our secret D=5678(free coefficient in equation).

7.3 pseudo codes for Shamir's secret sharing

D: secret value
P: set of parties p1,..,pn on to distributes the share
K: number of shares required to reconstruct the secret

Phase I

1. Select a random polynomial q(x)=a0+a1x+...+ak-1x^ (k-1) where ak-1≠0 and a0=D,
2. Choose n publicly known distinct random value x1, an such that xi≠0,
3. Compute the share of each node, where share (I) =q (xi),
4. For I=1 to n do send share I to node pi,
5. End for

Phase-II

Require every party is given a point (a pair of input to the polynomial and output)

6. Given subset of these pairs, find the coefficient of the polynomial using interpolaration,
7. The secret is the constant term (I.e.D)

7.4 conclusions

This approach provides high security. It does not required a trusted third party (TTP).It has a negligible computational overhead as compared to existing approach. Disadvantages of this approach incure more communication cost because for collaboratively computing cluster means, communication between every part is necessary.

VIII. CONCLUSIONS AND FUTURE WORK

Table given below give the comparison of three privacy preserving techniques and also its future works. From this table we conclude that Shamir's secret sharing scheme is better than other two methods.

Table I

Name of method/techniques	Computation cost	Communication cost	Accuracy	security	Future work
Shamir's Secret Sharing(SSS)	low	High(Because each party needs to communicate with each other)	high	High(very good)	1)Extend for vertical partitioned
Dimensionality Reduction Based Transformation(DRBT)	low	Low	average	High(good)	1)Extend for horizontal partitioned
Principal Component Analysis(PCA)	high	Low	High(Some cases it is almost equal to that of the original data set)	Low(Accuracy some time suffer as a result of security	1)Extend for vertical partitioned

REFERENCES

[1] W u jian and Li Xing Ming, "An Efficient Association Rule Mining Algorithm In Distributed Databases", IEEE Workshop on Knowledge Discovery and Data Mining, 2008,pp. 108-113.

[2] Xuan Canh Nguyen,Hoai Bac Le and Tung Anh Cao, "An enhanced scheme for privacy-preserving association rules mining on horizontally distributed databases",IEEE,pp.1- -4(2013)

[3] GUI Qiong and CHENG Xiao- hui , " A Privacy-Preserving Distributed Method For Mining Association Rules", IEEE International Conference On Artificial Intelligence and Computational Intelligence,pp.294– 297(2009)

[4] Marcin GORAWSKI and Zacheusz SIEDLECKI, "Optimization of Privacy Preserving Mechanisms in Homogeneous Collaborative Association Rules Mining",IEEE Sixth International Conference on Availability, Reliability and Security,2011,pp.347– 352.

[5] Chirag N. Modi and Udai Pratap Rao, "Elliptic Curve Cryptography Based Mining of Privacy Preserving Association Rules in Unsecured Distributed Environment ",IEEE International Conference on Advances in Communication, Network and Computing",2010,pp.94– 98.

[6] Vladimir Estivill-Castro and Ahmed Haj Yasien, "Fast Private Association Rules Mining by A Protocol for Securely Sharing Distributed Data" ,IEEE,pp.325– 330.

[7] Mahmoud Hussein,Ashraf EI-Sisi, and Nabil Ismail, "ast Cryptographic Privacy Preserving Association Rules Mining on Distributed Homogeneous Data Base,Loverk,R.J. Howlett and L.C.Jain(Eds):KES 2008,Part II,LNAI 5178:Springer-Verlag Berlin Heidelberg,2008,pp.607-616.

[8] Adriano A. Veloso and Wagner Neira JR.', "Efficient ,Accurate and Privacy-Preserving Data Mining for Frequent Item sets in Distributed Databases", Columbus –OH –USA srini@cis.ohio-state.edu.

[9] Marcin Gorawski and Zacheusz Siedlecki, "Implementation ,Optimization and Performance Tests of Privacy Preserving Mechanisms in Homogeneous Collaborative Association Rules", R. Meersman, T. Dillon, and P. Herrero(Eds.):OTM 2011,part I,LNCS 7044,pp.347-366,2011:5178:Springer-Verlag Berlin Heidelberg 2011.

[10] Murat Kantarcloglu and Chris Clifton, "Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned data ", IEEE, 2003.

[11] Aris Gkoulalas-Divanis and Vassilios S. Verykios, "An OVERVIEW OF PRIVACY PRESERVING DATA MINING", Summer 2009/vol. 15,No.4

[12] Vassilios S. Verykios ,Elisa Bertino , Igor Nai Fovino , "State-of-the-art in Privacy Preserving Data Mining",SIGMOD Record,Vol.33,No.1,March 2004.

[13] Daksha Agrawal and Charu C, Aggarwal, "On the Design and Quantification of Privacy Preserving Data Mining Algorithm".

[14] R. Vidya Banu and N. Nagaveni, "Evaluation of a perturbation-based technique for privacy preservation in multi-party clustering scenario", Information Sciences 232(2013), pp.437-448: Elsevier 2012.

[15] Stanely R.M. Oliveira and Osmar R. Zaiane, "A privacy-preserving clustering approach toward secure and effective data analysis for business collaboration", In: ELSEVIER COMPUTERS & SECURITY 26, 2007, PP.81-93.

[16] Vladimir Estivilla-Castro and Ahmed Haj Yasien, "Fast Private Association Rile Mining by A Protocol for Securely Sharing Distributed Data", IEEE, pp.325-330(2007).

[17] Ali Inan,Selim V. Kaya, Yucel Saygn,Erkay Savas, Ayca A. Hintoglu and Albert Levi, "Privacy preserving clustering on horizontally partitioned data", In: ELSEVIER Data & Knowledge Engineering 63,2007,pp.646-666.

[18] Alper Bilge and Huseyin Polat, "A comparison of clustering-based privacy-preserving collaborative filtering schemes", In: ELSEVIER Applied soft computing 13, 2013, pp.2478-2489.

[19] Adeela Waqar and Asad Raza, Haider Abbas and Muhannad Khurram Khan, "A framework for preservation of cloud users data privacy using dynamic reconstruction of metadata", In: ELSEVIER Journal of Network and Computer Applications 36, 2013, pp.235-248.

[20] Sheng Zhong, "Privacy-Preserving algorithms for distributed mining of frequent item sets", In: ELSEVIER Information Science 177, 2007,pp.490-503.

[21] Sankita Patel, Sweta Garasia and Devesh Jinwala," An efficient approach for Privacy Preserving Distributed K-Means Clustering based on Shamir's Secret Sharing Scheme", In: adfa, p. 1, 2011: Springer- Verlag Berlin Heidelberg 2011.