International Journal of Application or Innovation in Engineering & Management (IJAIEM)
Web Site: www.ijaiem.org Email: editor@ijaiem.org, editorijaiem@gmail.com
Volume 2, Issue 4, April 2013 **ISSN 2319 - 4847**

Privacy preserving k-Means clustering on horizontally distributed Data

[1]Bhadresh Prajapati, [2]Prof. H. B. Jethva

[1]PG Student, Department of Computer Engg, L.D.College of Engineering, Gujarat Technological University, Ahmedabad

[2]Associate Professor, L.D.College of Engineering, Gujarat Technological University, Ahmedabad

ABSTRACT

Privacy preserving Data mining (PPDM) is the combination of information security technology and knowledge discovery technology. The aim of Privacy preserving data mining is the extraction of relevant knowledge from large amount of digital data while protecting at the same time sensitive information. Here the proposed method for k-Means Clustering techniques on horizontally partitioned data on different nodes in a privacy preserving manner. We use k-Means algorithm as the basis for a communication efficient privacy-preserving clustering on databases that are horizontally partitioned between two parties. Here the propose methods for calculating the distance of objects of two parties in a privacy preserving manner.

Keywords: Privacy, Data Mining, Clustering, PPDM,

1. INTRODUCTION

The cooperative computation of data mining algorithms without requiring the participating organizations to reveal their individual data items to each other can be possible by Privacy-preserving distributed data mining. Most of the privacy-preserving protocols available in the literature are conversion of existing (distributed) data mining algorithms into privacy-preserving protocols. The resulting protocols can sometimes leak additional information [1, 2, 3].

Traditional data mining techniques and algorithms dejectedly operated on the original data set, which will cause the leakage of privacy data. At the same time, a large amount of data implicates the sensitive knowledge that their disclosure cannot be ignored to the competitiveness of enterprise. These problems challenge the traditional data mining, so privacy-preserving data mining (PPDM) has become one of the newest trends in privacy and security and data mining research. Agarwal and Srikant [4] and Lindell and Pinkas [5] introduced the first Privacy-preserving data mining algorithms which allow parties to collaborate in the extraction of knowledge, without any party having to reveal individual data items.

Clustering is unsupervised learning deals with designing classifiers from a set of unlabeled samples. A common approach for unsupervised learning is to first cluster or group unlabeled samples into sets of samples that are “similar” to each other. Clustering is a task to group similar items in a given data set into clusters with the goal of minimizing an objective function. The error-sum-of-squares (ESS) objective function is defined as the sum of the squares of the distances between points in the database to their nearest cluster centres. The k-clustering problem requires the partitioning of the data into k clusters with the objective of minimizing the ESS. Lloyd’s (k-means) algorithm [6]

This paper presents an algorithm for I/O-efficient clustering technique with k-means. The design of algorithm is with the conversion to a privacy-preserving version in mind, examines each data item only once and uses only sequential access to the data. Privacy-preserving version of the algorithm is for two-party horizontally partitioned databases. This protocol is communication efficient and it reveals the cluster centers.

2. Preparatory

There are former two parties A and B, both own dataset.

DS1 = {r1, r2, r3, . . . , rm} and

DS2 = {rm+1, rm+2, rm+3, . . . , rn }, respectively. They wish to jointly compute clustering of both dataset DS1 and DS2. DS1 and DS2 are data which are horizontally partitioned and stored on both parties. Both parties learn the final k cluster centers, and nothing else. Alternatively, with additional computation and communication, each party could learn the cluster to which each of their data objects belongs. If there were a trusted third party to whom A and B were both willing to send their data, this party could then compute the clustering and send the cluster centers to A and B.

However, practically there is no such party. Secure multiparty computation seeks protocols that can carry out the required computation without requiring a trusted third party. Here the pair-wise distances between the centroids of the clusters are used for clustering the dataset which are horizontally partitioned on two parties.

3. PRIVACY-PRESERVING CLUSTERING ALGORITHM

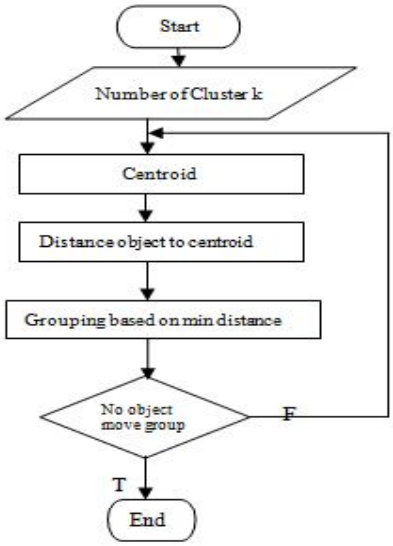
3.1 Distributed clustering algorithm DK-Means

Distributed clustering algorithm DK-Means is proposed by Zheng Miaomiao [2], ect., which improves the K-Means algorithm, that is, the site in the clustering process does not require transferring large amounts of data objects, only needs to send the clustering centers as well as the total number of clusters of data objects, which reduces data traffic of the distributed clustering process so as to improve operating efficiency. This strategy would be applied on the database into two different parties, recursively produce k cluster centers from each of the party, and then merge these 2k centers into the k final centers. We do this by repeatedly choosing a best pair of clusters Ci and Cj for merging, and replacing them in the clustering with Ci U Cj.

The k-means cluster algorithm will do the three steps below until convergence on both parties.

Iterates until stable (= no object move group):

1. Determine the centroids coordinates
2. Determine the distance of each object to the centroid
3. Group the object based on the minimum distance.



A best pair of clusters is one with least error. Suppose C_1 and C_2 be two clusters being considered for a merge and C.w denote the number of objects in cluster C then. In [7] the error of C_1 U C_2 is

$$err_w = \frac{C_1.w * C_2.w * dist^2(C_1, C_2)}{C_1.w + C_2.w}$$

where dist(C_1,C_2) is the distance between the centers of C_1 and C_2.

3.2 Subroutine MergeCenters

1. S_1= Cluster (DS_1,k)
2. S_2= Cluster(DS_2, k)
   (S_1 and S_2 are Cluster Centers)
3. S is Merge Centers (S_1U S_2, k)
4. Input Cluster centers S, Integer k
5. While (|S| > k)
   i. Compute the merge error for all pairs of centers in S.
   ii. Remove from S the pair with the lowest merge error,
   and insert the center of the merged cluster, with its weight as the sum of the weights of the pair.
6. Output Cluster centers S such that |S| = k

3.3 Algorithm to calculate distance for k-means Clustering with Privacy-preservation on horizontally distributed data

The protocol will be for three participants:

- Parties D_1 and D_2 of Dataset owner DS_1 and DS_2 respectively and the third party (TP).
- The parties D_1 and D_2 want to compute the distance between two centroid C_i and C_j, held by each site respectively, then

The protocol for computing the distance between C_i and Cj without using their original values by TP works as follows:

At site D_1
- Random number r_{12} and r_{1T} are generated then sends r_{12} to D_2, r_{1T} to TP.
  If the generated random number r_{12} is odd, then
  D_1 negates its input value as x'= - x otherwise x'=x
- D_1 send x'' = (r_{1T} + x') to D_2.

At site D_2
- If the generated random number r12 is even, then
  D_2 negates its input value as y'= -y otherwise y'= y.
- D_2 sends m=(x''+ y') to TP,

At site TP
- Distance (x , y) = (m- r_{1T})
  (x and y are values of attributes values of site D_1 and D_2 respectively.)

3.4 Algorithm to merge Clusters which found minimum error with preserving Privacy.

The protocol for merging cluster C_i and Cj without using their original values by TP works as follows:

At site D_1
- Random number r_{1T} is generated then sends r_{1T} to TP.
- D_1 send x'= (x*Ci.w - r_{1T}) to D_2.

At site D_2
- D_2 sends m=(x'+ y*Cj.w) to TP,

At site TP
- Merged value of M = (m + r_{1T})
- Average value of M is calculated using no of rows in each clusters.
  (x and y are values of attributes values of site D_1 and D_2 respectively.)

Here the random numbers used in protocol is generated by pseudo-random number generators. Our comparison protocol preserves privacy as long as a high quality pseudo-random number generator, that has a long period and that is not predictable, is used and it is secured [8].

4. CONCLUSION

This paper probes the privacy-preserving of distributed clustering data mining and proposes a horizontal partitioned clustering privacy-preserving algorithm, which, through the effective combination of secure multi-party computation and clustering algorithm, not only well hides the sensitive data in clustering mining process to realize privacy-preserving but also protects the results of clustering mining from influencing to realize the accuracy of results.

5. ACKNOWLEDGEMENT

Bhadresh G. Prajapati would like to thank to my guide Prof. H. B. Jethva for his great effort and instructive comments in this paper work. Lastly, I wish to thank to all those who helped me during my research work

References

[1] J. Vaidya and C. Clifton, Privacy-preserving k-means clustering over vertically partitioned data, In 9th KDD 2003.

[2] G. Jagannathan and R.Wright, Privacy-preserving distributed k-means clustering over arbitrarily partitioned data, In 11th KDD, pages 593–599, 2005.
[3] S. Jha, L. Kruger, and P. McDaniel, Privacy preserving clustering, In 10th ESORICS, 2005.
[4] R. Agrawal and R. Srikant, Privacy-preserving data mining, In ACM SIGMOD, pages 439–450, May 2000
[5] Y. Lindell and B. Pinkas, Privacy preserving data mining, J. Cryptology, 15(3):177–206, 2002.
[6] S. P. Lloyd, Least squares quantization in PCM, IEEE Trans. on Info. Theory, 28:129–137, 1982.
[7] J. H. Ward, Hierarchical grouping to optimize an objective function, J. Amer. Stat. Assoc., 58(2):236–244, 1963.
[8] Ali Inan, Yucel Saygin, Privacy Preserving Clustering on Horizontally Partitioned Data, IEEE, International Conference on Data Engineering Workshops, 1-7, 2006

AUTHOR

Bhadresh G. Prajapati, pursuing his Master Degree in Computer Science & Technology from Gujarat Technological University (L D College of Engg., Ahmedabad) as a sponsored candidate, received his Bachelor Degree in Computer Engg. from U.V. Patel College of Engg, Kherva From NGU in 2002. His area of interest is Database System, S/W Engineering and Information Security. He works for the State Govt. of Gujarat, since 2005, as a Lecturer in Information Technology (GES CL-II) and served various Govt. Polytechnics

Prof. Harikrishna B. Jethva received his post graduate degree in Computer engineering from Dharmsinh Desai University in 2009 and Bachelor Degree in Computer Engineering from Saurashtra University, Rajkot in 2001. He has worked as a Assistant Professor for 10 Years and presently working as a Associate Professor in L. D. College of Engineering. His area of interest is in Neural Network, Theory of Computation, Computer Network, Compiler Design, Soft Computing and Distributed Computing.