ISSN:0975-9646

K.A. Sireesha et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (6) , 2013, 891 - 894

Privacy Preserving Clustering Over Distributed Data

K.A.Sireesha, R.Srinivas, K.Arunbhaskar

CSE Department, Sri Sai Aditya Institute of Science and Technology
Surampalem, India

Abstract **Privacy prevents sharing of data mainly in data mining applications. Privacy concerns in many application domains for not only prevent sharing of data, but also provide security to the data so that no unauthorised persons can access it. Privacy limits data mining technology to identify patterns and trends from large amount of data. The main aim of privacy preserving clustering is to develop data mining techniques that could be applied on databases without violating the privacy of individuals. In this paperKDPIPELINE algorithm is introduced to preserve privacy over distributed data.**

Keywords **K-Means Algorithm, Density based Algorithm, ε-core and noncore points**

1. INTRODUCTION

Data mining research deals with the extraction of potentially useful information from large collections of data with a variety of application areas such as customer relationship management, market basket analysis, and bioinformatics. The extracted information could be in the form of patterns, clusters or classification models. The power of data mining tools to extract hidden information from large collections of data lead to increased data collection efforts by companies and government agencies. Naturally this raised privacy concerns about collected data. Data mining researchers started to address privacy concerns by developing special data mining techniques under the framework of "privacy preserving clustering".

Clustering can be considered the most important unsupervised learning technique; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. Clustering is "the process of organizing objects into groups whose members are similar in some way". A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters [1].Consequently many clustering methods have been developed, each of which uses a different induction principle.

Privacy issues in statistical databases have been thoroughly investigated. Recently privacy-preserving clustering has been a very active area of research. Initial focus in this area was on construction of decision trees from distributed data sets. There is also a significant body of research on privacy-preserving clustering of association rules [2]. In general, there are two approaches for designing privacy-preserving machine learning algorithms. The first approach is to use transformations to perturb the data set before the algorithm is applied. This approach for designing privacy-preserving clustering algorithms is taken by several researchers. A second approach to designing privacy preserving algorithms is to use algorithmsfrom the secure multiparty computation literature. The advantage of this approach over the perturbation approach is that formal guarantees of privacy can be given for this algorithm

1.1. Data Matrix(or Object-By-Variable Structure)

A data matrix has an object-by-variable structure. Each row in the matrix represents entity values of its attributes stored in columns. An m×n data matrix has the data of m objects on n attributes as shown in Figure.

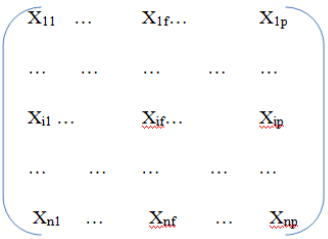

Fig1.Data matrix

Each and every horizontal partition may contain values from a different range in which another privacy preserving protocol for finding the global minimum and maximum of each attribute would be required. Normalization on the dissimilarity matrix yields the same effect, without loss of the need for another protocol and accuracy [3].

1.2 Dissimilarity Matrix

A dissimilarity matrix has object-by-object structure. An m×m dissimilarity matrix stores dissimilarity or the distance between each pair of objects as shown in fallowing figure, the distance of an object to itself is zero.

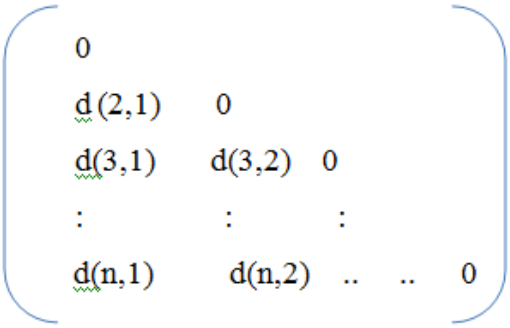

Fig 2.Dissimilarity Matrix

Only the entries below the diagonal are filled, for d[i][j] = d[j][i], Because of symmetric nature of the comparison functions. Each and every party that holds a horizontal partition of a data matrix D can construct its localdissimilarity matrix as long as comparison functions for object attributes are public. Privacy preserving comparison protocols need to be employed to calculate an entry d[i][j] of dissimilarity matrix d, if objects i and j are not held by the same party[1].

1.3 Comparision Functions

The distance between two numeric attributes is the absolute value of the difference between them. Categorical attributes are compared for equality so that any categorical value is equally distant to all other values but itself. The distance between alphanumeric attributes is measured in terms of the edit distance which is heavily used in bioinformatics [9].

Edit distance algorithm generates the number of operations required to transform a source string into a target string. Basic available operations are insertion, deletion and transformation of a character [1][3].

1.4 Comparison Protocols

Dissimilarity matrix d, is an object-by-object structure in which d[i][j] is the distance between objects i and j. Consider an entry in d, d[i][j] If both objects are held by the same data holder, and the third party need not intervene in the computation of the distance between objects i and j. The data holder that has these objects computes the distance between them and sends the result to the third party. If objects i and j are from different parties, and a privacy preserving comparison protocol must be employed between owners of these objects [8]. It follows from this distinction, in order to construct the dissimilarity matrix; each data holder must send its local dissimilarity matrix to the third party and run a privacy preserving comparison protocol with every data holder. Basic transformations are helpful to provide the privacy and security to the data. These transformations are applying to KDPIPELINE algorithm for providing privacy.

Translation A translation is applied to an object by repositioning it along a straight line path from one coordinate location to another where P'=P+T

$$P=\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \quad P'=\begin{pmatrix} x_1' \\ x_2' \end{pmatrix} \quad T=\begin{pmatrix} tx \\ ty \end{pmatrix}$$

Fig.3 Translation

Rotation A two-dimensional rotation is applied to an object by repositioning it along a circular path in the xy-plane. Rotation of a point from position(x ,y) to position (x',y') through an angle Θ relative to the coordinate origin. The original angular displacement of the point from the x axis is Ø.

$$P'=R.P$$

Where the rotation matrix $R = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$

Fig.4 Rotation

Scaling A scaling transformation alters the size of an object.

$$P'=S.P$$

$$\begin{pmatrix} X' \\ Y' \end{pmatrix} = \begin{pmatrix} Sx & 0 \\ 0 & Sy \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix}$$

Fig.5 Scaling

2. METHODOLOGY

The K-means algorithm takes the input parameter, K, Nobjects are going to be partitioned into K clusters so that the resulting intra cluster has high similarity but the inter cluster has low similarity. Cluster similarity is measured to the mean value of the objects in a cluster, which can be viewed as centroidor centre of the cluster's gravity [5].

2.1 K-Means Algorithm

The K-means algorithm for partitioning,where each and every cluster's centre is represented by the mean value of the objects in the cluster.

Input

N denotesnumber of clusters,

D denotes the data set contains n objects.

Output: A set of K number ofclusters.

Algorithm

1. Arbitrarily choose N objects from D as the initial cluster canters;
2. Repeat
3. Re-assign each and every object to the cluster, to which the object is the most similar,

Based on the mean value of the objects in the cluster;

4. Update the cluster means i.e., calculate the mean value of the objects for each and every cluster until no further change occurs.

Core point

There are a few variants of the K-means method. These can differ from the initial K means, the calculation of dissimilarity, and the strategies for calculating the cluster means. The K-means algorithm is sensitive to outliers since an object with an extremely large value may distort the distribution of data [1][4].

2.2 DBSCAN: A Density-Based Clustering Method Based onConnected Regions with Sufficiently High Density.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density based clustering algorithm. The algorithm grows with sufficiently high density into clusters

and discovers clusters of arbitrary shape in spatial databases with noise. It defines that a cluster is a maximal set of density-connected points. According to the definitions, a point p is directly density-reachable from a point q if it is in the ε-neighbourhood of q. A point p is density-reachable from q if there is a chain of points pi, where i=1 to n and pi+1 is directly density-reachable. A point p is density-connected to another point q if there is a point **o** such that both p and q are density reachable. DBSCAN can be started by bringing in a point to a temporary storage and finding its ε -neighbourhood. If the ε-neighbourhood of a data point contains less than Min Pts then it is marked as noise [1][15]. The algorithm is given as follows.

Density Based Clustering Algorithm

Input: Data of $X_n$which contains n data items.

Output: Dense based clusters

1. Clusters can be calculated fromi=1 to n
2. If $X_n$is declared as Unclassified and not noise
3. Expand the given Cluster X_n
4. Cluster centre=Clustercentre+1 until no further change is observed.

3. KD PIPELINE ALGORITHM

KD PIPELINE Algorithm is used to combine K-means and density-based clustering [4]. KD PIPELINE does the following: First perform the K-Means and then density-based clustering is performed over each and every K-means cluster. K-means is performed first instead of density-based clustering why because K-means is fast, and it is easy to find out the means of the cluster centres. Other reason is setting density threshold is difficult than setting distance threshold (K).For effective merging, each and every data point Xi has following three columns to store results of clustering[14].

For effective merging, each and every data point Xi has following three columns to store results of clustering:

1. K-means Cluster Centre,
2. Density-based Cluster Centre,
3. K-means Cluster Centre and Density-based Cluster Centre.

It is the label assigned to each and every point after running K-means based on the following definitions.

Core Distance

For each and every cluster, Core Distance is half of the distance between its centre and its closest cluster centre [7]. It is not farther from its cluster centre by Core Distance. Core region of a cluster is that in the centre which each and every data point is core [8].

Non-core point

Non-core region is that in which each and every point is non-core. There can be multiple cluster Centres for a dense cluster if it spreads more than one K-means cluster. Multiple cluster Centres are resolved by recording the different labels assigned to a single ε -core point during the two stages of clustering, i.e. searching for dense clusters in core region and in non-core region. This makes sure that dense clusters are found correctly [7][8].

The KD PIPELINE algorithm is given as follows:

Algorithm KD PIPELINE

Input: Data with K number of items.

Output: Data with K-means Cluster Centre and density-based Cluster Centre

1. Apply K-means algorithm to form cluster centres.
2. Experiments show the size of ε-core and noncore vary and is usually small compared to the size of the whole data. Compute Min Pts in the given computed cluster canters.
3. Now, apply Density based clustering for finding core and ε –core points of the cluster centres calculated above.
4. Multiple density based cluster centres are resolved by matching cluster centres of the ε –core points. Repeat until all the data points are covered in the formation of cluster centres.
5. Run K-means without noise found in density based clustering taking the earlier centres as initial points.

4. CONCLUSION

KD PIPELINE combinesthe features of two popular clustering algorithms; Distance-based clustering (K-means) and Density based clustering (DBSCAN). The resultant synergy due to merge is not only provide privacy but also improve the speed and ease in setting density threshold for density-based clustering and improvement in quality of K-means clusters by removing noise found by density-based clustering.

5. FUTURE SCOPE

The merging approach can be extended to pipeline BIRCH and density-based clustering to refine the formation of clusters efficiently.

REFERENCES

[1] Jiawei Han and MichelineKamber. Data Mining:Concepts and Techniques. Morgan Kaufmann Publishers,2000.

[2] A. K. Jain and R. C. Dubes.Algorithm for ClusteringData,chapter Clustering Methods and Algorithms.Prentice- Hall Advanced Reference Series, 1988.

[3] Ali Inan YucelSaygm, Erkay Savas, AlbertLevi. Privacy Preserving clustering on Horizontally Partitioned Data.

[4] Manoranjandash, HuanLiu, Xiaowei Xu.'1+1>2' Merging Distance and Density Based Clustering.

[5] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering:a review. ACM Computing Surveys, 31(3):264–323, 1999.

[6] N. Katayama and S. Satoh. The SR-tree: an indexstructure for high-dimensional nearest neighbour queries. In Proceedings of the ACM SIGMOD Intl.Conf.on Management of Data, pages 369–380, 1997.

[7] LKaufman and P. Rousseuw. Finding Groups in Data: An Introduction to Cluster Analysis. Wiley Series inProbability and Mathematical Statistics, 1990.

[8] J. MacQueen. Some methods for classification andanalysis of multivariate observations. In Proceedingsof the Fifth Berkely Symposium on Mathematics,Statistics, and Probability, volume 1, pages 281–296,1967.

[9] R. T. Ng and J. Han. Efficient and effective clusteringmethods for spatial data mining.In Proceedings of theVLDB Conference, Santiago, Chile, September 1994.

[10] A. K. Jain and R. C. Dubes. Algorithm for Clustering Data, chapter Clustering Methods and Algorithms. Prentice-Hall Advanced Reference Series, 1988.
[11] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering:a review. ACM Computing Surveys, 31(3):264–323, 1999.
[12] N. Katayama and S. Satoh. The SR-tree: an indexstructure for high-dimensional nearest neighborqueries. In Proceedings of the ACM SIGMOD Intl.Conf.on Management of Data, pages 369–380, 1997.
[13] L. Kaufman and P. Rousseuw. Finding Groups in Data: An Introduction to Cluster Analysis. Wiley Series in Probability and Mathematical Statistics, 1990.
[14] J. MacQueen. Some methods for classification and analysis of multivariate observations.In Proceedingsof the Fifth Berkely Symposium on Mathematics,Statistics, and Probability, volume 1, pages 281–296,1967.
[15] R. T. Ng and J. Han. Efficient and effective clustering methods for spatial data mining.In Proceedings of theVLDB Conference, Santiago, Chile, September 1994.