2nd International Conference on Current Research Trends in Engineering and Technology
© 2018 IJSRSET | Volume 4 | Issue 5 | Print ISSN: 2395-1990 | Online ISSN: 2394-4099
Themed Section: Engineering and Technology

A Privacy Preserving Clustering On Alphanumeric Data

Patel Kevin[1], Patel Hiteshree[2], Patel Krishna C.[3], Patel Krishna J.[4], Mr. Kaushal Patel[5], Dr. Sheshang Degadwala[6]

[1-4]U.G.B.E. Student Computer Engineering, Sigma Institute of engineering, Bakrol, Gujarat, India
[5]Assistant Professor, Computer Engineering, Sigma Institute of engineering, Bakrol, Gujarat, India
[6]Head of Department, Computer Engineering, Sigma Institute of engineering, Bakrol, Gujarat, India

ABSTRACT

Huge volume of detailed data is collected and analyzed by applications using data mining, sharing of these data is beneficial to the unauthorized persons. On the other hand, it is an important asset to business organization and government for decision making process. At the same time analyzing such data open threats to privacy if privacy is not preserved properly. Our project aims to reveal the information by protecting sensitive data. We are using three perturbation techniques to preserve the privacy of data. Perturbed dataset does not reveal the original data and also does not change the analysis results. Hence, privacy is preserved.

Keywords: Privacy, Preserve, Data mining, Clustering, Perturbation

I. INTRODUCTION

Data mining is the process of extracting the useful information from huge amount of data. Our project aims to preserve the privacy of sensitive data and extract the useful information from that data. Clustering is the process of grouping the data in such a way that data objects in one group are more similar than to data objects in other groups.

Our system can handle both numeric and alphanumeric data. The perturbation techniques, which we are using, can be applied on numeric data only. So, we convert the alphanumeric data into numeric data first and then apply the perturbation techniques. We are using three perturbation techniques which are rotation, scaling and translation. We perturb the data in such a way that similarity between data objects in perturbed data is same as that in original data. We are using K-means algorithm for clustering.

II. LITERATURE REVIEW

Jun-Lin Lin, Meng-Cheng Wei [1] proposed the k-anonymity model. To minimize the information loss, similar data are grouped together and then anonymized each group individually.

Ali Inan, Yücel Saygın, Erkay Savaú, Ayça Azgın Hintoglu, Albert Levi [2] proposed methods for constructing the dissimilarity matrix of objects from different sites in a privacy preserving manner

which can be used for privacy preserving clustering as well as database joins, record linkage and other operations that require pair-wise comparison of individual private data objects horizontally distributed to multiple sites.

M Kalita, D K Bhattacharyya, M Dutta [3]

Proposed privacy preserving clustering technique using hybrid approach. The Hybrid Data Perturbation (HDP) method is used. The technique mainly exploits a combination of isometric transformations.

Fan-rang Meng, Bin Liu, Chu-jiaoWang [4] proposed a simple and effective privacy-preserving distributed mining method of clustering (PPD-SMD) and (PPD-JD) which is proposed to solve the issue about privacy preserving of cluster based on Binary and Nominal Attributes distance.

Fang-Yu Rao, Bharath K. Samanthula, Elisa Bertino, Xun Yi†, Dongxi Liu‡ [5] proposed a novel and efficient solution to privacy-preserving outsourced distributed clustering (PPODC) for multiple users based on the k-means clustering algorithm.

III.RESEARCH APPROACH AND METHOD

1) Rotation perturbation:

The rotation perturbation in which the object is rotated about a fixed point. The direction of rotation can be clockwise or anticlockwise.

We can rotate the object at particular angle θ (theta) from its origin. From the figure, we can see that the point P(X, Y) is located at angle φ from the horizontal X coordinate with distance r from the origin.

Suppose we want to rotate the point at the angle θ. After rotating it to a new location, we will get a new point P' (X', Y').

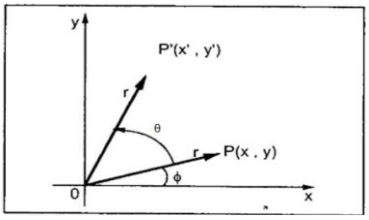


Fig 1 - rotation perturbation

Mathematical representation of the rotation perturbation is:

$$x' = x\cos\theta - y\sin\theta$$

$$y' = x\sin\theta + y\cos\theta$$

2) Translation perturbation:

The translation perturbation in which all the points are moved in a straight line and in the same direction. The size, shape and direction of the image are the same as that of the original object.

We can translate a point by adding a translation coordinate (t_x, t_y) to the original coordinate (X, Y) to get the new coordinate (X', Y').

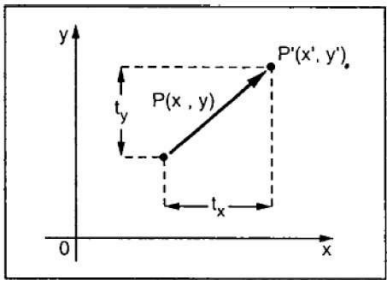


Fig 2 - translation perturbation

Mathematical representation of the translation perturbation is:

$$X'=X+t_x$$
$$Y'=Y+t_y$$

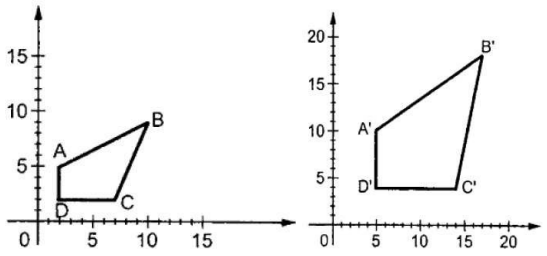
3) Scaling perturbation:

The Scaling perturbation is used to change the size of the object. It is also used to expand or compress the dimension of object. Scaling perturbation can be done by multiplying the original coordinates of the object with the scaling factor to get result.

Suppose, original coordinates are (X, Y) and the scaling factors are (S_x, S_y) and the produced new coordinates are (X', Y').

Mathematical representation of the scaling perturbation is:

$$X'=X*S_X$$
$$Y'=Y*S_Y$$



(a) (b)

Fig 3 - scaling perturbation

4) K-means clustering:

1) Specify the number of clusters (K) to be created.
2) Select randomly k objects from the dataset as the initial cluster centers or means.
3) Assigns each observation to their closest centroid based on the Euclidean distance between the object and the centroid.
4) For each of the k clusters update the cluster centroid by calculating the new mean values of all the data points in the cluster. The centroid of a K_{th} cluster is a vector of length p containing the means of all variables for the observations in the K_{th} cluster; p is the number of variables.
5) Iteratively minimize the total within sum of square. That is, iterate steps 3 and 4 until the cluster assignments stop changing or the maximum number of iterations is reached.

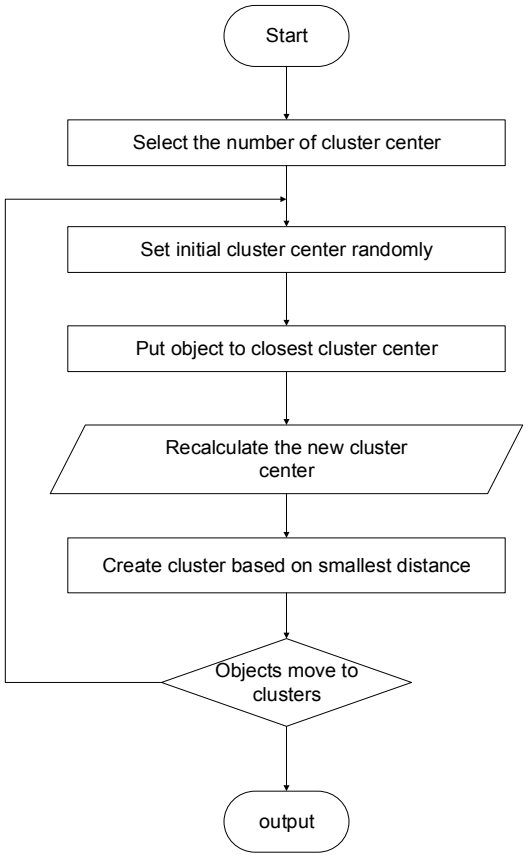


Fig. 4 K-means clustering

III. PROPOSED SYSTEM

3.1 Algorithm of proposed system

INPUT: Huge amount of raw data

OUTPUT: Clustered data

Implementation of algorithm for problem description: Preserving the privacy of data and clustering of data

Step 1: Load the data.

Step 2: Preprocess the data if data is noisy.

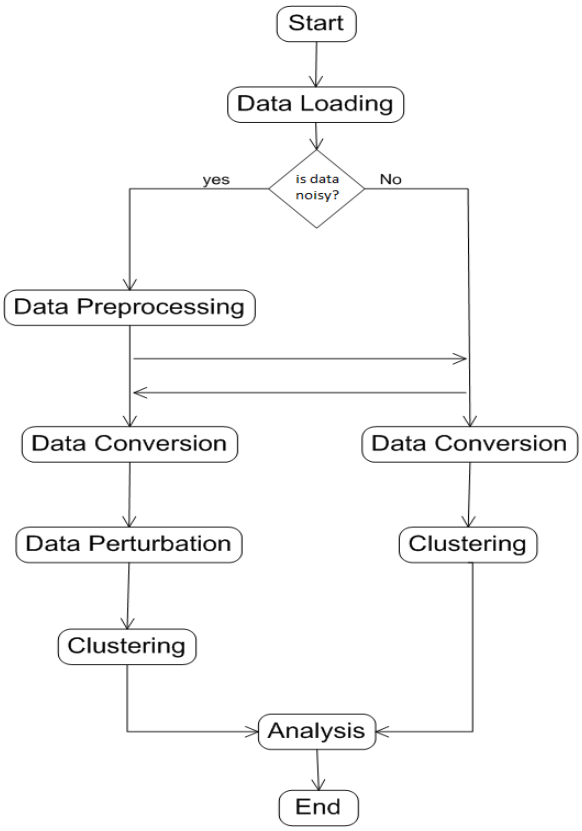
Step 3: Convert the alphabetic values into numeric values.

Step 4: Select the data transformation technique and its factors.

Step 5: Apply the selected data transformation technique.

Step 6: Apply the K-means algorithm for clustering.

Step 7: Analyze the data.



V. CONCLUSION

In the proposed system, we have implemented, different perturbation techniques which provide the privacy to sensitive data, K-means algorithm for clustering purpose.

In the proposed system, user can apply only one perturbation technique at a time. In future, user will be able to apply hybrid perturbation techniques at a time.

IV. REFERENCE

[1]. Liming Li, Qishan Zhang, "A Privacy Preserving Clustering Technique Using Hybrid Data Transformation Method", IEEE, 2009

[2]. Ali Nan, YucelSaygın, Erkay Sava, AycaAzgınHinto Lu, Albert Levi, "Privacy Preserving Clustering on Horizontally Partitioned Data", IEEE, 2006

[3]. M Kalita, D K Bhattacharyya, M Dutta, "Privacy Preserving Clustering- A Hybrid Approach", IEEE, 2008

[4]. Fang-Yu Rao, Bharath K. Samanthula, Elisa Bertino, Xun Yi, Dongxi Liu, "Privacy-Preserving and Outsourced Multi-User k-Means Clustering", IEEE, 2015

[5]. Fan-rang Meng, Bin Liu, Chu-jiao Wang, "Privacy Preserving Clustering over Distributed Data", IEEE, 2010

[6]. Keng-PeiLin,"Privacy-preserving kernek-means clustering outsourcing with random transformation", Springer, 2016

[7]. Yu Xin, Zhi-QiangXie, Jing Yang, "The privacy preserving method for dynamic trajectory releasing based on adaptive clustering", Elsevier, 2016

[8]. Alper Bilge, HuseyinPolat , "A scalable privacy-preserving recommendation scheme via bisecting k-means clustering", Elsevier, 2016

[9]. Zhi-QiangGao, Long-Jun Zhang, "DPHKMS: An Efficient Hybrid Clustering Preserving Differential Privacy in Spark", Springer, 2016