A PRIVACY PRESERVING CLUSTERING METHOD BASED ON FUZZY APPROACH AND RANDOM ROTATION PERTURBATION

M. NAGA LAKSHMI, K SANDHYA RANI

Research Scholar, Professor, Dept of Computer Science, S.P.M.V.V, Tirupati, Andhra Pradesh, INDIA

nagalaxmi.mada@gmail.com, sandhyaranikasireddy@yahoo.co.in

ABSTRACT

Individual privacy issues arise in these days when organizations using clustering as a data analysis tool. Private and sensitive data available in criminal, healthcare and financial records need to be preserved and also avoid the privacy leakage with the data mining system. In this paper, a privacy preserving clustering method is proposed for protecting the underlying sensitive attribute values when sharing the data for clustering over centralized data. The proposed method based on the concept of fuzzy logic and random rotation perturbation. This approach ensures secrecy of confidential numerical attributes without losing accuracy in results. The experiments demonstrate that the proposed method is effective and provides a feasible approach to balancing privacy and accuracy.

Keywords: Privacy preserving clustering, Fuzzy membership function, Data Perturbation, Data Mining.

I. INTRODUCTION

Data mining is a powerful technique for analyzing and collecting interesting, useful patterns from large data that reside in companies and public organizations databases. Many data mining techniques have been developed. They are association rule mining, classification and clustering. Clustering is the process of dividing a given dataset into groups (clusters), in which a group of data objects are homogeneous and heterogeneous objects are to different groups. There are several advantages with the clustering, to analyze the market effectively and also improves profitability of business collaboration. Huge volumes of detailed personal data are regularly collected and analyzed by applications using data mining. Such data include shopping habits, criminal records, medical history, credit records, among others [1]. Analyzing such data may helpful for business organizations decision making process, but this may reveal the sensitive data of individual. The data owners try to transform the original data by using some data distortion method to achieve privacy preservation of individuals and also maintain the utility of data. Therefore individual privacy concerns limit the willingness of the data custodians to share data [2]. Confidential data is not protected efficiently by standard data mining algorithms. Privacy violation occurs when confidential data revealed from the extracted patterns of data mining. So there is a need to develop data mining algorithms which efficiently protecting privacy.

Privacy preserving data mining is a new research area, which effectively extracts hidden information without including private data of individuals. Privacy preserving clustering becomes the new direction for organizations sharing the data for clustering and also achieving privacy of individuals. Privacy issues are considering in two situations. They are centralized environment and distributed environment. In centralized environment, database is available in single location. In this environment, a privacy preserving data mining techniques are used to hide sensitive data of individuals. In distributed environment, data is distributed to multiple sites. In this environment privacy preserving data mining technique is applied for integrating the data from multiple sites, without revealing the privacy of individuals... Related works about privacy preserving clustering over centralized database is discussed in the following section.

II. RELATED WORK

Privacy preserving clustering for centralized data previously introduced by the authors in [3] and [4]. They proposed geometric transformation methods based are translation data perturbation, scaling data perturbation, rotation data perturbation are used to modify sensitive data of individuals without affecting the clustering. In [5], the authors address the hybrid data transformation based on the combination of three basic transformations translation, rotation, and reflection. The authors in [6] proposed geometric transformation method in combination with random response method for privacy preserving clustering over centralized data. The authors address the concept of dissimilarity between the objects and dimensionality reduction based transformation for privacy preserving clustering [7]. In [8], authors proposed hybrid data transformation approach based on double reflecting data perturbation and rotation data perturbation for privacy preserving clustering. Privacy preserving cluster through cluster bulging have been proposed by the authors in [9]. They presented a method in which each cluster is perturbed through geometric transformation method. The authors in [10] proposed a method based on cluster rotation, in which each cluster is rotated in such a way that data is perturbed and also useful for cluster analysis. In [11], the...
authors introduced data obfuscation techniques distort sensitive data in order to achieve privacy preservation. The authors in [12] address a different approach based on the concept of fuzzy sets which are used fuzzy membership function to convert the data values into appropriate fuzzy values. In [13], authors proposed random rotation perturbation for privacy preserving classification. In [15], a new framework for privacy preserving based on the concepts of additive data perturbation and swapping. 

In this paper a new privacy preserving clustering technique is proposed based on fuzzy approach and random rotation perturbation for centralized database. This method first select the confidential numerical attributes from the given dataset and then generate random rotation matrix, multiply selected matrix with the random rotation matrix. The proposed technique gives good accuracy when dataset is shared for clustering analysis and also give high accuracy. The proposed method is discussed in the following section.

III. PROPOSED MODEL
To achieve the privacy preservation, data set is perturbed without changing the membership and shape of the clusters in the original dataset. The proposed method selects the confidential numerical attributes and then applies the data distortion technique called random rotation perturbation technique.

A. FUZZY TRANSFORMATION APPROACH
A fuzzy transformation approach distorts the sensitive numerical attributes using built in fuzzy membership function such as Z shaped fuzzy membership function, Triangular fuzzy membership function, Gaussian fuzzy membership function.

B. RANDOM ROTATION PERTURBATION:
Random rotation perturbation [13] technique is applied by first assuming dataset as an nxm matrix M. The procedure starts by selecting confidential numerical attributes and generate random rotation matrix randomly and independently and then apply rotation transformation to get the transformed data $P = M \times R$. The distorted dataset is released for clustering analysis. The distorted dataset protects the privacy of individuals and also achieves high accuracy when dataset is released for clustering analysis.

The proposed model shown in Figure 1 represents the data distortion process. The distortion process starts by taking original dataset as the input and remove the identified attributes. Next, confidential numeric attributes are extracted from the original dataset and distort by applying fuzzy based transformation approach. The resulted dataset is given as the input to the random rotation perturbation to obtain distorted dataset and released for clustering analysis.

![Figure 1: Model of the Proposed Method](image)

B. DATA PERTURBATION ALGORITHM:

| Input: | Original Dataset D of size mxn. |
| Output: | Distorted dataset D'of size mxn. |
| Method: | |
| 1) | Suppress the identifier attributes. |
| 2) | Extract the dataset that contains only confidential numeric attributes. |
| 3) | For each attribute in D do |
| 4) | Transform the attribute using z-shaped fuzzy membership function. |
| 5) | End-for |
| 6) | Generate Random rotation matrix. |
| 7) | Multiply the data matrix with rotation matrix. |
| 8) | Output the perturbed dataset D1. |

Experimental results of the proposed method are discussed in the next section.

IV. IMPLEMENTATION OF PROPOSED MODEL
The proposed model is implemented to evaluate the performance of data distortion method and experiments are conducted on two real-life datasets obtained from the University Of California Irvine (UCI), Machine Learning Repository [14]. The data sets which are considered are Wine data set with thirteen attributes and 178 record and Iris data set with four numerical attributes and 150 records. The well-known k-means clustering algorithm is used to measure the clustering quality.
After transforming the data, clusters in the original dataset should be equal to those ones in the distorted dataset. WEKA (Waikato Environment for Knowledge Analysis) software is used to test clustering accuracy of the original and modified data base. The effectiveness is measured by misclassification error. The misclassification error, denoted by \( M_{E_i} \), is measured as follows:

\[
M_{E_i} = \frac{1}{N} \sum_{j=1}^{K} \left( \left| \text{Cluster}_i(T) \right| - \left| \text{Cluster}_i(T^1) \right| \right)
\]

In the above formula \( N \) means the number of points in the original dataset, \( K \) is the number of clustering. \( \left(\text{Cluster}_i(T)\right) \) means the number data points of the \( i \)th cluster of the original data set \( \left(\text{Cluster}_i(T^1)\right) \) means the number of data points of the \( i \)th cluster of the transformed dataset.

The experimental results for two datasets are shown in Table. This Table show the \( M_E \) values obtained for the proposed method, z-shaped fuzzy membership function, novel additive approach on two datasets. Higher \( M_E \) values indicates lower clustering quality. The experiments are conducted 10 times and \( M_E \) value is taken as an average of 10. From table I, the accuracy of proposed method is better when compared to others.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Iris</th>
<th>Wine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zmf</td>
<td>0.09746</td>
<td>0.0892</td>
</tr>
<tr>
<td>Proposed Method</td>
<td>0.0867</td>
<td>0.06853</td>
</tr>
<tr>
<td>Novel Additive</td>
<td>0.196</td>
<td>0.199</td>
</tr>
</tbody>
</table>

The following figure 2 shows the comparison of misclassification error rates among the three data distortion techniques such as z-shaped fuzzy membership function, proposed method and novel additive perturbation technique for IRTS and wine datasets.

V. CONCLUSION:
The major conflicting factors for privacy preserving clustering is loss of privacy and loss of accuracy. This paper proposes a privacy preserving clustering method based on fuzzy approach and random rotation perturbation. The main objective of this method is to achieve privacy preservation and retain the information for clustering analysis. Two real life datasets are considered to implement the proposed method and the experimental results have shown that the proposed method satisfying the privacy constraints and retains the clustering quality.

VI. REFERENCES: