Journal of Computer Science 4 (4): 320-326, 2008
ISSN 1549-3636
© 2008 Science Publications

An Effective Data Transformation Approach for Privacy Preserving Clustering

[1]R.R. Rajalaxmi and [2]A.M. Natarajan
[1]Kongu Engineering College, Perundurai, Tamil Nadu, India
[2]Bannari Amman Institute of Technology, Sathiamangalam, Tamil Nadu, India

Abstract: A new stream of research privacy preserving data mining emerged due to the recent advances in data mining, Internet and security technologies. Data sharing among organizations considered to be useful which offer mutual benefit for business growth. Preserving the privacy of shared data for clustering was considered as the most challenging problem. To overcome the problem, the data owner published the data by random modification of the original data in certain way to disguise the sensitive information while preserving the particular data property. Data transformation techniques played a vital role to preserve privacy in data mining. We put forward an effective approach which defeats the problem of addressing privacy of confidential categorical data in clustering. A set of hybrid data transformations are introduced (HDTTR and HDTSR) and the effectiveness of the approach has been analyzed. A complete analysis of the proposed approach and a formal study of the problem have been done. Our proposed approach illustrates the effectiveness of clustering of sensitive categorical data before and after the transformation.

Key words: Clustering, categorical data, data transformation, translation, rotation, scaling

INTRODUCTION

Due to the ever increasing use of information technology, large volumes of detailed personal data are regularly collected. Such data include shopping habits, criminal records, medical history and credit records, among others[2,3]. These data can be analyzed by applications which make use of data mining techniques. Hence such data is an important asset to business organizations and governments for decision making processes and also to offer social benefits, such as medical research, crime reduction, national security, etc.[10]. On the other hand, analyzing such data opens new threats to privacy and autonomy of the individual if not done properly.

With the conventional data analysis methods there is a limited threat to privacy. Also these techniques mainly present the results based on the mathematical characteristics associated with the data. Making use of such techniques may not reveal some interesting patterns which are hidden in the data. By using appropriate data mining techniques it is possible to explore the hidden patterns. But the threat to privacy becomes real since data mining techniques are able to derive highly sensitive knowledge from unclassified data which is not even known to database holders[6]. In order to overcome this issue the data owners may decide not to share or release such data for analysis provided they should make a compromise for exploring hidden knowledge[7]. The privacy becomes worst when they decided to have secondary usage of data when they are unaware of behind the scenes use of data mining techniques[11]. As an example in point, Culnan[5] made a particular study of secondary information use which she defined as the use of personal information for other purposes subsequent to the original transaction between an individual and an organization when the information was collected. The key finding of this study was that concern over secondary use was correlated with the level of control the individual has over the secondary use. As a result, individuals felt that they are losing control over their own personal information that may reside on thousands of file servers largely beyond the control of existing privacy laws.

The challenging problem that we address in this study is: how can we protect against the misuse of the knowledge discovered from secondary usage of data and meet the needs of organizations to support decision making

In order to address this issue, we focus on privacy preserving confidential categorical data clustering, particularly when personal or confidential data are

Corresponding Author: R.R. Rajalaxmi, Department of Computer Science and Engineering, Kongu Engineering College, Perundurai, Erode-638052, Tamil Nadu, India

shared before clustering analysis. To address privacy concerns in clustering analysis, we need to design specific data transformation methods that enforce privacy without loosing the benefit of mining.

Literature survey: The primary goal in privacy preserving clustering is to protect the sensitive data before it is released for analysis. However the data may reside within an organization or in different places a distributed data. In such a scenario appropriate algorithms or techniques should be used which does not reveal any sensitive information in the knowledge discovery process. To address this issue there are many approaches adopted for privacy preserving data mining. It can be classified based on the following dimensions: Data distribution, Data modification, Data mining algorithm, Data or rule hiding and Privacy preservation[15].

In[18], this problem is addressed by transforming a database using Object Similarity-Based Representation (OSBR) which uses the similarity between objects and Dimensionality Reduction-Based Transformation (DRBT) which uses random projection. Here the dissimilarity matrix is shared for the analysis purpose. Privacy preserving clustering is addressed[16,17] based on either vertically partitioned data or horizontally partitioned data. Protecting privacy for numerical data is addressed[14] by using geometric data transformation. Selective modification of data can be performed to achieve higher utility for the modified data given that the privacy is not jeopardized[15]. Uniform randomization approach is applied in preserving the privacy of association rules[8]. Privacy preservation using rotation is performed for classification[4]. The authors proposed[13] data transformation approaches with binary representation of the data, which does not reveal experimental analysis. Oliveria *et al.*[14] proposed an approach to perform privacy preserving clustering of numerical data using geometric data transformation. Although our work is also based on geometric data transformation methods, there are two significant differences between our work and their work: first, our work deals with hybrid data transformation. Second, in their solution, each sensitive attribute is numeric whereas we have considered categorical attributes. Our work considers selective modification of confidential categorical data such that the perturbed data is released for secondary use which maintains appropriate level of privacy.

Problem definition: Let us consider an organization A**.** It owns a dataset D and wants to cluster it. However A does not have the expertise to do the clustering process. Hence it is decided to release the dataset to the any other organization B to perform clustering. Since organization A has confidential data, the original dataset cannot be released as such to B**.** Also the dataset D may contain different type of attributes. For our problem we have taken the dataset consisting of sensitive categorical attributes. Before sharing the dataset D with B, organization A must transform D to preserve privacy of individual data records. However, the transformation applied to D must not affect the similarity between objects. The problem can be stated as follows:

Let D be a relational database and the set of clusters generated from D is *C*. The goal is to transform D into D' so that the following limitations hold:

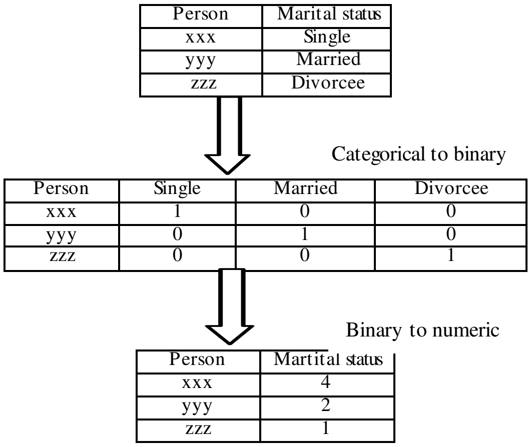
- A transformation T when applied to D must preserve the privacy of individual records, so that the released database D' conceals the values of confidential attributes, such as sex, marital status, credit rating and others
- The similarity between objects in D' must be the same as that one in D, or slightly altered by the transformation process. Although the transformed database D' looks very different from D, the clusters in D and D' should be as close as possible

Proposed approach: In order to address the above problem, the original database consisting of categorical data is transformed using the following steps.

- The categorical attribute is converted into binary attribute and mapped to numeric value
- Hybrid geometric data transformation approach is used to transform the converted categorical attribute

A. Categorical data conversion: The Geometric data transformation methods can not be applied for the categorical value. Categorical variable can be converted into asymmetric binary variable by creating a new binary variable for each of the M nominal states[9]. For an object with a given state value, the binary variable representing that state is set to 1 while the remaining binary variable are set to 0. After the conversion the binary value is mapped to the corresponding numeric value.

For example, to encode the nominal variable marital status, a binary variable can be created for each of the three values listed in Fig. 1. For a person having the marital status 'married', the married variable is set to 1, while the remaining two variables are set to 0 (Fig. 1).



Person	Marital status
xxx	Single
yyy	Married
zzz	Divorcee

Person	Single	Married	Divorcee
xxx	1	0	0
yyy	0	1	0
zzz	0	0	1

Person	Martital status
xxx	4
yyy	2
zzz	1

Fig. 1: Converting categorical data to binary and mapping to numeric value

The data transformation approach:

Geometric data transformation methods: In this research, we consider the family of geometric data transformation methods (GDTM) specified in[14]. The inputs for the GDTMs are the vectors of V, composed of confidential converted categorical attributes only and the random noise vector N, while the output is the transformed vector subspace V. The data transformation algorithms have essentially two major steps:

- Choose a noise term and the operations that must be applied to each confidential attribute. In this step random noise vector N is created
- Using the random noise vector N, transform V into V' using a geometric transformation function

Translation data transformation: In this method the noise term applied to each confidential attribute is constant and can be either positive or negative[14]. The set of operations takes only the value {Add} corresponding to an additive noise applied to each confidential attribute.

Scaling data transformation: In this method the noise term applied to each confidential attribute is constant and can be either positive or negative[14]. The set of operations takes only the value {Multi} corresponding to a multiplicative noise applied to each confidential attribute.

Rotation data transformation: This method works differently from the previous methods. In this case, the noise term is an angle θ. The rotation angle θ, measured clockwise, is the transformation applied to the observations of the confidential attributes[14]. The set of operations takes only the value {Rotate} that identifies a common rotation angle between the attributes Ai and Aj. Unlike the previous methods, RDP may be applied more than once to some confidential attributes.

Data reconstruction methods can be used to deduce original data from the randomized data. By applying the above transformations separately to the original data, the privacy breach is high. In order to overcome this issue, we have applied hybrid transformation to the original data which makes it difficult to construct the sensitive data.

Noise level: In order to measure the effectiveness of our approach with respect to varying noise range, we define noise level for the attributes. Let us consider an attribute Ai. Let n be the number of categories in the attribute represented as $a_{i1},a_{i2},\ldots,a_{in}$. Let e be a noise level. When the noise level is low, the probability of moving a record from original category to a new category in the distorted database is less. However when the percentage is high the probability of moving the record to a new category is also high. Hence it is essential to choose a suitable noise level such that the privacy level is high and the misclassification of the records in the clusters is low.

Hybrid Data Transformation using Translation and Rotation (HDTTR): In this scheme, we select randomly one operation for each confidential attribute that can take the values {Add, Rotate} in the set of operations. Thus, each confidential attribute is perturbed using an additive noise followed by rotation.

Algorithm:

- Input: V, N
- Output: V

```
Step 1: For each confidential attribute $A_j$ in V, where $1 \leq j \leq d$ do
        Get the noise level e
        Accordingly calculate the noise range $el_1$ to $el_2$
        Select the noise term $e_j$ in N for the confidential attribute Aj randomly within the range
        The j-th operation opj ← {Add}
        The k-th operation $op_j$ ← {Rotate}

Step 2: For each $v_i \in$ V do
        For each aj in $v_i = (a_1,\ldots,a_d)$, where aj is the observation of the j-th attribute do
        aj ← Transform. $(a_j, op_j, e_j)$
End
```

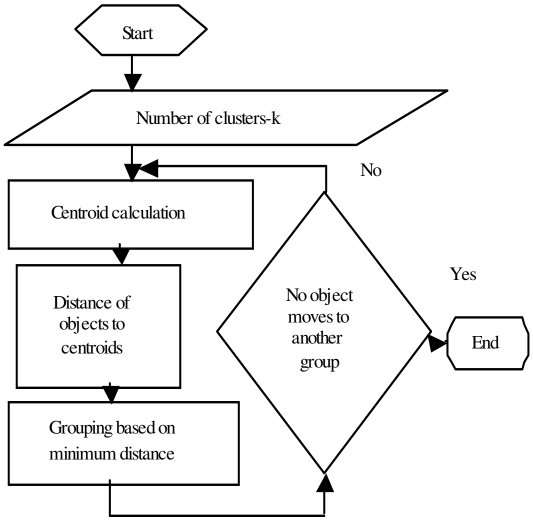

Start
Number of clusters-k
Centroid calculation
Distance of objects to centroids
Grouping based on minimum distance
No object moves to another group
No
Yes
End

Fig. 2: Clustering process

Hybrid Data Transformation using Scaling and Rotation (HDTSR): In this scheme, we select randomly one operation for each confidential attribute that can take the values {Mult, Rotate} in the set of operations. Thus, each confidential attribute is perturbed using multiplicative noise term followed by a rotation.

Clustering technique: In order to compare the results of clustering before and after the data transformation we have used K-means clustering algorithm. It is used to group the objects based on attributes/features into K number of groups where K is positive integer. The grouping is done by minimizing the sum of squares of distances between data and the corresponding cluster centroid. Thus, the purpose of K-mean clustering is to group the data.The basic steps of k-means clustering are as shown in Fig. 2:

Iterate until stable (= no object moves to another group):

- Determine the centroid coordinate
- Determine the distance of each object to the centroids
- Group the object based on minimum distance

RESULTS AND DISCUSSION

To evaluate the significance of the proposed approach we have adopted the measures specified in[14].

Effectiveness: While measuring the effectiveness in clustering, it is essential to consider the number of legitimate points grouped in the original and the distorted databases. After transforming the data, the clusters in the original databases should be equal to those ones in the distorted database. But, this is not always the case since the original data is transformed using the approach. After the transformation: a point from a cluster becomes a noise point, or a point from a cluster migrates to a different cluster. Misclassification Error is measured in terms of the percentage of legitimate data points that are not well-classified in the distorted database. Ideally, the misclassification error should be 0%. The misclassification error, denoted by M_E, is measured as

$$M_E = 1/n\sum_{1}^{k} |\,\mathrm{CLUSTER(D)}\,| - |\,\mathrm{CLUSTER(D')}\,|$$

Quantifying privacy: Traditionally, the privacy provided by a perturbation technique has been measured as the variance between the actual and the perturbed values[12]. This measure is given by Var(X-Y) where X represents a single original attribute and Y the distorted attribute. Privacy level can be specified by expressing security as Sec = Var(X-Y)/Var(X).

Identification of sensitive attributes: The sample dataset chosen for analysis is the census dataset which contains weighted census data extracted from the 1994 and 1995 current population surveys conducted by the US Census Bureau. The data contains demographic and employment related attributed. The attributes present are

- Age
- Class of worker
- Marital status
- Sex
- Tax filer status
- Country of birth self
- Citizenship

The input for the system is a set of sensitive attributes of a dataset. The attributes Marital status, Sex and Tax filer status are identified as sensitive since they denote the sociological and economical aspects of an individual which should not be revealed to third parties.

Cluster groups: The clusters are grouped according to number of clusters chosen ranging from Cluster 1 (Nonfiler, Never married) to Cluster N (Head of household, Married-civilian spouse present) into various groups where N is the number of clusters chosen. We chose our clustering parameters in order to achieve an optimal performance considering the number of records misclassified between these groups which should be less for a good preservation approach.

The clustering process is done on the original data set before transformation and the results are recorded. Then, after transforming the data set using any one of the above hybrid data transformation methods again the clustering process is done over the preserved data. Both the results are analyzed for minimal deviation.

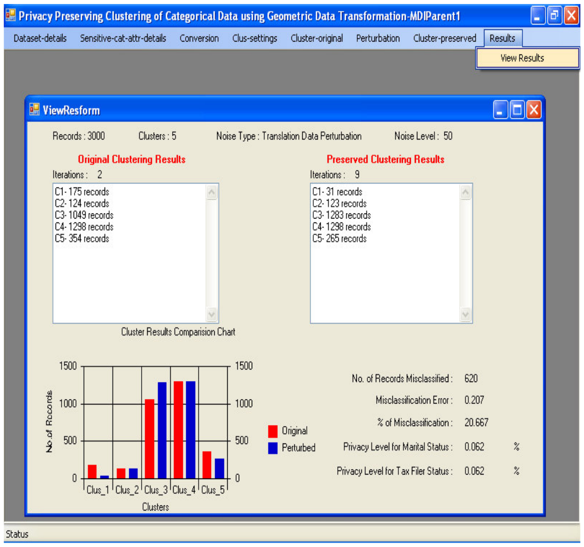
In order to customize the process of data transformation the proposed system has been developed using VB.Net and uses MS-Access as the backend to store the database. The system allows the user to do the transformation which involves the following steps:

- Select the source database to be transformed and display the details
- Suppress the numerical attributes which are not required for clustering and show the relevant attributes
- Choose the specific categorical attributes which are considered as sensitive. The system displays the details about the various categories present in the attribute
- Select the noise level for the attribute to be transformed
- Choose the type of hybrid data transformation
- Create the modified database
- Compare the privacy preserving measures for varying noise level

By following the above steps the user can decide which noise level is suitable for the attribute and can create the transformed database D. A screenshot of our developed Privacy preserving clustering system is given below in Fig. 3.

To our knowledge this is a new effort to protect privacy in categorical data clustering. Hence we compared the GDTMs against each other and with respect to the following benchmarks: (1) the results of clustering analysis without transformation, (2) the results of HDTTR, (3) the results of HDTSR.

The various experimental results are shown for census dataset in Tables 1-2. The misclassification errors obtained after applying the HDTs are shown in Table 1. The noise level is varied form lower level to higher level to record the effectiveness of the transformation. For a noise level of 75%, it is observed that the misclassification error is less and also the distortion of the original data is high which makes it difficult for an adversary to predict the original data.



Records : 3000 Clusters : 5 Noise Type : Translation Data Perturbation Noise Level : 50

Original Clustering Results

Iterations : 2

C1- 175 records
C2- 124 records
C3- 1049 records
C4- 1298 records
C5- 354 records

Preserved Clustering Results

Iterations : 9

C1- 31 records
C2- 123 records
C3- 1283 records
C4- 1298 records
C5- 265 records

No. of Records Misclassified : 620
Misclassification Error : 0.207
% of Misclassification : 20.667
Privacy Level for Marital Status : 0.062 %
Privacy Level for Tax Filer Status : 0.062 %

Fig. 3: A screenshot of Privacy preserving clustering system

Table 1: Misclassification error

Noise level (%)	12	25	50	75
HDTTR	0.14	0.691	0.951	0.569
HDTSR	0.141	0.703	0.951	0.569

Table 2a: Privacy level for sex

Noise level (%)	12	25	50	75
HDTTR	0.059	0.112	0.227	0.322
HDTSR	0.131	0.05	0.012	0.041

Table 2b: Privacy level for marital status

Noise level (%)	12	25	50	75
HDTTR	0.031	2.12	4.365	8.861
HDTSR	0.032	2.041	4.365	8.631

Table 2c: Privacy level for tax filer status

Noise level (%)	12	25	50	75
HDTTR	1.006	1.029	1.171	0.997
HDTSR	1.002	1.03	1.175	0.997

The privacy level for the attribute Sex is shown in Table 2a. It is observed that the attribute has more privacy level by using HDTTR than HDTSR. The privacy level for the attribute Marital Status is shown in Table 2b. For this attribute the effect of privacy level is having a marginal difference. The privacy level for tax filer status is shown in Table 2c. Here also there is a marginal difference between the two transformations.

Table 3 shows the effect of misclassification error based on the number of clusters by applying the proposed transformations. The results show that HDTSR yields considerably low misclassification error

Table 3: Misclassification provided by HDTMs

No. of clusters	HDTTR	HDTSR
Misclassification error		
K = 2	0.001	0.000
K = 4	0.049	0.050
K = 6	0.168	0.114
K = 8	0.440	0.400
K = 10	0.511	0.538

CONCLUSION

The family of hybrid data transformation methods introduced ensures privacy preservation in clustering analysis, notably on categorical data. The proposed methods distort only confidential categorical attributes to meet privacy requirements, while preserving general features for clustering analysis. Hence the data owner can decide to select an appropriate noise level for distortion based on the categories present in the sensitive attributes. To our best knowledge this is the first effort to provide a solution for the problem of privacy preserving clustering of categorical data. The experiments demonstrated that the methods are effective and provide practically acceptable values for balancing privacy and accuracy. The transformed database is available for secondary use such that the distorted database preserves the main features of the clusters mined from the original database and an appropriate balance between clustering accuracy and privacy is guaranteed.

REFERENCES

1. Agrawal, D. and C.C. Aggarwal, 2001. On the design and quantification of privacy preserving data mining algorithms, In: Proceedings of ACM SIGMOD/PODS, California, USA, pp: 247-255. DOI:10.1145/375551.375602
2. Agrawal, R. and R. Srikant, 2000, Privacy preserving data mining. In: Proceedings of ACM SIGMOD Conference on Management of Data, New York, USA, pp: 439-450.
3. Brankovic, V. and V. Estivill-Castro, 1999. Privacy issues in knowledge discovery and data mining, In: Proceeding of Australian Institute of Computer Ethics Conference, Melbourne, Victoria, Australia.
4. Chen, K. and L. Liu, 2005. Privacy preserving data classification with rotation perturbation, In: Proceedings of the 5th IEEE International Conference on Data Mining, IEEE Computer Society, USA, pp: 589-592. DOI:10.1109/ICDM.2005.121
5. Culnan, M.J., 1993. How did they get my name? An exploratory investigation of consumer attitudes toward secondary information. MIS Q., 17: 41-363.
6. Elisa Bertino and Ravi Sandhu, 2005. Database security-concepts, approaches and challenges. IEEE Trans. Dependable Secure Computing.,Vol 2, No.1, pp:2-19. DOI:10.1109/TDSC.2005.9
7. Estivill-Castro, V. and L. Brankovic, 1999. Data swapping: Balancing privacy against precision in mining for logic rules. In: Proceedings of Data Warehousing and Knowledge Discovery Dawak-99, Aug.30-Sept.1, Florence, Italy, pp: 389-398.
8. Evfimievski, A., R. Srikant, R. Agrawal and J. Gehrke, 2002. Privacy preserving mining of association rules. In: Proceedings of the 8th ACM SIGKDD Italian Conference on Knowledge Discovery And Data Mining, Edmonton, AB, Canada, pp: 217-228. DOI:10.1145/775047.775080
9. Han, J. and M. Kamber, 2001. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, San Francisco, CA,
10. Jefferies. V, 2000. Multimedia, cyberspace and ethics. In: Proceedings of International Conference on Information Visualization (Iv2000), IEEE, England, London, pp: 99-104.
11. John, G.H., 1999. Behind-the-scenes data mining. Newsletter, ACM SIGKDDM, 1: 9-11. DOI:10.1145/846170.846176.
12. Muralidhar, K., R. Parsa and R. Sarathy, 1999. A general additive data perturbation method for database security, J. Mgmt. science, Vol:45,pp: 1399-1415.
13. Natarajan, A.M., Rajalaxmi R.R, Uma N and Kirubakar G, 2007. A Hybrid Data Transformation Approach for Privacy Preserving Clustering of categorical Data, In: Proceedings of International Conference on Systems, Computing Sciences and Software Engineering (SCSS), Vol 1:pp 403-408
14. Stanley R. M. Oliveira, Osmar and R. Zaiyane, 2003. Privacy preserving clustering by data transformation. In: Proceedings of 18th Brazilian Conference on Databases, October 2003, Manaus, Brazil, pp 304—318.
15. Verykios, V.S., E. Bertino, I.N. Fovino, L.P. Provenza, Y. Saygin and Y. Theodoridis, 2004, State-of-the-art in privacy preserving data mining, ACM SIGMOD Rec., 3: 50-57. DOI:10.1145/974121.974131.

16. A. Inan, Y. Saygyn, E. Savas, A.A. Hintoglu, A. Levi, 2006, Privacy Preserving Clustering on Horizontally Partitioned Data, In: Proceedings of the 22nd International Conference on Data Engineering Workshops, pp: 95-98. DOI:10.1109/ICDEW.2006.115
17. Geetha Jagannathan_ Krishnan Pillaipakkamnatt Rebecca N. Wright, 2006, A New Privacy-Preserving Distributed k-Clustering Algorithm. In: Proceedings.of the SIAM International Conference on Data Mining, pp:494-498.
18. Stanley R.M. Oliveira,Osmar R Zaiane, 2004. Privacy-preserving clustering by object similarity-based representation and dimensionality reduction transformation, In: Proceedings of the ICDM Workshop on Privacy and Security Aspects of Data Mining, pp: 40-46.