On Analysis of Mixed Data Classification with Privacy Preservation

Sarat Kr. Chettri[1], Biplab Kr. Ray[2]

[1,2]Department of Computer Science & Engineering and Information Technology,
School of Technology, Assam Don Bosco University
Airport Road, Azara, Guwahati - 781017, Assam. INDIA.
[1]*sarat.chettri@dbuniversity.ac.in,* [2]*bip.ray11@gmail.com*

Abstract:*Privacy-preserving data classification is a pervasive task in privacy-preserving data mining (PPDM). The main goal is to secure the identification of individuals from the released data to prevent privacy breach. However, the goal of classification involves accurate data classification. Thus, the problem is, how to accurately mine large amount of data for extracting relevant knowledge while protecting at the same time sensitive information existing in the database. One of the ways is to anonymize the data set that contains the sensitive information of individuals before getting it released for data analysis. In this paper, we have mainly analyzed the proposed method Microaggregation based Classification Tree (MiCT) which use the properties of decision tree for privacy-preserving classification of mixed data. The evaluations are done based on various privacy models developed keeping in mind the various situations which may arise during data analysis.*

Keywords: Microaggregation, decision tree, mixed data, data perturbation, classification accuracy, anonymous data**.**

1. Introduction

Rapid advancement of computer technologies has enabled an organization to easily collect, store and manage large volume of data electronically. In such a scenario where there is an explosion in the amount of data availability, data mining techniques are becoming more important to assist decision making and strategic planning in various areas of medical science, marketing, official statistics and so on. However, for analyzing such tremendous amount of data, organizations are forced to share their data to third party. In such cases there exist two major problems: (1) time and cost involvement in transferring data to third party and (2) protecting the privacy of personal data which may be sensitive and confidential. It becomes a vital problem of protecting individual's privacy when data gets shared among analysts. Numerous kinds of data disclosure risks [6] increase the risk of privacy breach of the individuals. The privacy breach may be unintentional, but sharing of such personal data may raise various social and ethical issues. To protect individual's privacy, many countries have enacted strict legislation against such privacy breaches. Such issues should not curtail data mining while at the same time data privacy preservation need to be seriously looked upon. Thus, the problem is how to accurately mine large amount of data for extracting relevant knowledge while protecting at the same time sensitive information existing in the database. Some of the typical scenarios requiring both data privacy and mining accuracy are as follows:

- Smart metering system may provide data regarding usage of electricity by individual households. Such data may be collected and analyzed by an organization to plan power distribution and management. At the same time individuals may not be willing to share their data as it breaches their privacy.
- In medical domain, insurance agencies may ask hospitals to provide their patient's medical data for analysis however hospitals may not share their data due to legal obligations or because of their willingness to protect the privacy of the patients.
- Online shopping sites keep track of the items bought or browsed by customers. The customers may not be aware of such activities, and even if they know, still may not be comfortable with their personal data to be collected and mined. However, the data may be used for personalized item recommendations or making strategic planning by the shopping sites.

Taking care of such scenarios, a data owner may apply certain techniques of transforming or modifying the data before sharing it to any third party to conduct data mining. In this context, privacy-preserving data mining (PPDM) techniques has been investigated extensively in recent years. Various PPDM methods namely microaggregation, *k*-anonymity, data generalisation and suppression, data randomization, additive noise, global recoding, sampling and so on are proposed in the literature. The key idea followed is to make data anonymous before releasing it for analysis or transform individual records in such a way that the results obtained though mining the anonymized or transformed data set are (almost) the same as those obtained when using original data. Irrespective of the type of data, typical transformation process reduces the granularity of data representation and hence there is a natural trade-off between information loss and data privacy. In this paper, we have mainly analyzed the proposed microaggregation based Classification Tree (MiCT) method [4] which uses tree properties and perturbs microdata data before its release. Given a mixed data set D (continuous and categorical), classification would be performed on D ensuring the resultant classes so formed in D and its anonymized version D' are as close as possible.

2. Related Concepts

2.1 *k*-Anonymity

k-anonymity was first introduced by Samarati [9] and Sweeney [10] as an efficient approach for preserving the privacy of microdata set. To understand the concept of *k*-anonymity we start with the following definitions:

Definition 1 (Quasi-identifier): "A quasi-identifier (*QI*) is a set of attributes in a data set *D* such that the set can be used to identify individual records in *D* by linking with an external information. For e.g. job, age, gender etc."

Definition 2 (*k*-anonymity): "A data set *D* is said to satisfy *k*-anonymity for *k*>1 if, for each combination of values of quasi-identifiers (*QI*) there exist at least *k* records in the data set sharing the same combination thus making each record indistinguishable from at least (*k*-1) other records".

A common approach followed to resist re-identification of an individual by an intruder is to make data *k*-anonymous. Any adversary can at most determine *k* records from the released data set to identify the target individual. Therefore, *k*-anonymity guarantees privacy protection of microdata. Another approach followed to naturally achieve *k*-anonymity is microaggregation where the attribute values of the released data set are not suppressed or generalised leading to low information loss.

Definition 3 (Complying and Non-complying nodes of a Decision Tree): In a given decision tree *T* we develop a set of attributes S_1 by traversing *T* from its root node R_1 to leaf node L_1. Now, if a set of attributes in S_1 satisfies *k*-anonymity over quasi-identifier QI_1, we term it as a complying node else as non-complying node of *T*.

2.2 Microaggregation

Microaggregation is a data perturbation technique which belongs to the family of Statistical Disclosure Control (SDC) method [5, 7]. The approach followed here is, given a microdata set *D*, it partitions *D* into groups, each consisting of at least *k* records, where *k* is a user-defined parameter. A microaggregated data set *D* is built by replacing the original records of each group with its respective group centroid value. The microaggregated data set D' is released without jeopardizing the privacy of the microdata, as *k* records have an identical protected value. Clearly, D' is *k* anonymous over a quasi-identifier. However, the microaggregation method being perturbative in nature incurs information loss of the data set, thus the main goal is to maximize the within-group homogeneity to minimize information loss. The major challenge is how to perform data modification in such a way that both data disclosure risk and information loss are kept below certain permissible limits.

The within-group homogeneity [8] which states data utility is commonly measured with sum of square error (*SSE*). The goal of microaggregation is to partition a data set *D* with minimal SSE measure, which is defined as:

$$SSE = \sum_{i=1}^{g} \sum_{x_{ij} \in C_i} (x_{ij} - \bar{x}_i)'(x_{ij} - \bar{x}_i) \qquad (1)$$

Where *g* is the total number of groups in the data set, c_i is the *i*-th group and $\bar{x}_i$ is the centroid of c_i. Whenever, a data set *D* is microaggregated, there is always a loss of information incurred due to data modification. The information loss (*IL*) is computed as:

$$IL = \frac{SSW}{SST} . 100 \qquad (2)$$

Where, total sum of square error for entire data set *D* is computed as in equation 3 when $\bar{x}$ is the global centroid:

$$SST = \sum_{i=1}^{g} \sum_{x_{ij} \in C_i} (x_{ij} - \bar{x})'(x_{ij} - \bar{x}) \qquad (3)$$

The similarity computation [3] between any two tuples $X = \{x_1, x_2, x_{3,}x_q, x_{q+1}, x_{q+2}. . . x_m\}$ and $Y = \{y_1, y_2, y_3 y_q, y_{q+1}, y_{q+2}. . . y_m\}$ consisting of *q* continuous attributes and m categorical attributes is given as:

$$d_m(X, Y) = d_n(x_{1...q}, y_{1...q}) + \gamma_b d_c(x_{q+1...m}, y_{q+1...m}) \qquad (4)$$

3. Proposed Method

In our proposed approach as shown in table 1, given a mixed data set *D* we have first used standard C4.5 algorithm to build a classification tree. The graphical representation is shown in figure 1. On obtaining the tree, for a given *k* value, we discover the complying nodes C_n and non-complying nodes (Def. 3) NC_n for level *i*, where *i* = 1. . . *l*. We then use the MDAV2k algorithm [2] to microaggregate set *Q* of non-complying nodes. However, if *Q* is Null then we can say that the data set *D* is *k* anonymous (Def. 2)

TABLE 1: MICROAGGREGATION BASED CLASSIFICATION TREE (MICT) METHOD

```
Algorithm: MiCT (Microaggregation based Classification Tree)
Input: A classification tree T constructed over original data set D, set of virtual quasi-identifier V QID_1, VQID_2. . .VQID_l for l levels and k value.
Output: Perturbed Dataset D´

1   For each attribute in the virtual quasi-identifier $V QID_i$ of level $i = 1 . . . l$ do
2   Discover the complying nodes $C_n$ and non-complying nodes $NC_n$ for level $i$
3.  Form sets $S$ and $Q$ with complying nodes and non-complying nodes respectively
4.  End for
5.  Microaggregate set $Q$ using MDAV2k method to get $Q'$
6.  Obtain $D'$ as $S \cup Q'$
```

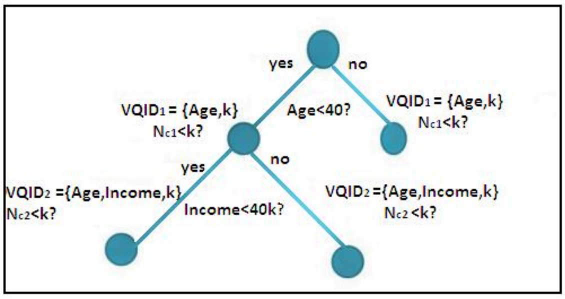

Figure1: Graphical representation of taxonomy tree construction

4. Evaluating the MiCT Method

An important aspect of development and evaluation of a method is to identify certain benchmark for its development and have suitable criteria for evaluation. Same aspect applies to the privacy-preserving data mining methods, however, no method exist in the literature which outperforms the other in every aspect. There is always a criterion of balancing a trade-off between data utility and privacy but it is up to a user to choose the most appropriate privacy-preserving data mining method based on the criteria over which they are interested. As in [11] the parameters for evaluation any privacy-preserving data mining algorithm are as follows:

- The *data disclosure risk* or level of uncertainty of discovering the sensitive information which has been hidden in the released data set.
- The *data utility* or the amount of information loss incurred on the released data set after applying the privacy preservation technique.

4.1 Privacy Models

In this paper, to evaluate how well the data mining process are carried out with the developed privacy models based on our approach of privacy preservation, we have assumed the various situations under which the privacy models will be used.

- **Perturbed Perturbed (PP) Model.** In this model, classification of perturbed data is done using data mining models built on perturbed data. This is a situation where data analysts may not get access to original data due to certain privacy concerns, instead they have to perform the data mining on perturbed data using models built on perturbed data only.
- **Perturbed Original (PO) Model.** In this model, classification of original data is done using data mining models built on perturbed data. In certain cases, a data mining model built on perturbed data may be used to analyze original data.
- **Original Perturbed (OP) Model.** In this model, classification of perturbed data is done using data mining models built on original data. We have considered certain situation where data mining models may be built by trusted party on original data held by the data owner. The developed model will be shared among various analysts for analyzing data. However, before any data release the data owner will perturb the data using certain data perturbation technique to preserve privacy. Thus, third parties use the model built on original data to analyze the released perturbed data.

5. Experimental Data and Results

For our experimental purposes we have used two standard data sets; German Credit and Adult data set form the UCI machine learning repository [1]. For classification purpose we have used the standard *k* Nearest Neighbor (*k-NN*) method to obtain the classification accuracy. All experiments are performed based on ten-fold cross validation. From the results obtained as shown in figure 2, it can be concluded that the proposed method achieves acceptable classification accuracy as per the privacy models which we have used as the evaluation parameters. Our proposed method MiCT does achieves a balance in maintaining the data utility and privacy of the microdata sets.

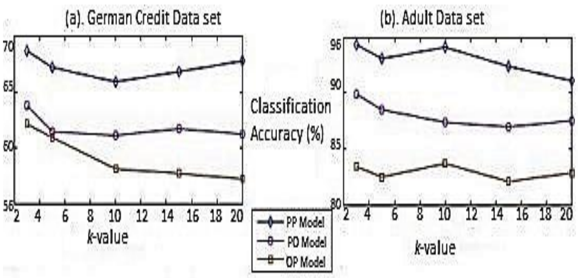

Figure2: Classification accuracy with different privacy models and k values

6. Conclusion

In this paper, we have mainly analyzed the proposed method MiCT under different privacy models as per different situations which may arise during data analysis. Classification is done on the perturbed and original mixed (continuous and categorical) data as the case may be, and measured their classification accuracy. The data sets used are the standard data sets and the results so obtained proves that being a perturbative method MiCT does not degrade the data utility of the microdata sets and achieves data privacy.

References

[1]. D. J. Newman, A. Frank, A. Asuncion. UCI machine learning repository, http://mlearn.ics.uci.edu/mlrepository.html. Irvine, CA: University of California, School of Information and Computer Science, 2010.

[2]. S. K. Chettri and B. Borah. MDAV2k : A variable-size microaggregation technique for privacy preservation. In *International Conference on Information Technology Convergence and Services*, pages 105–118, Bangalore, 2012.

[3]. S. K. Chettri and B. Borah. An efficient microaggregation method for protecting mixed data. In *Computer Networks and communications (NetCom), LNEE*, volume 131, pages 551–561, Bangalore, 2013. Springer New York.

[4]. S. K. Chettri and B. Borah. Anonymizing classification data for preserving privacy. In *Third International Symposium on Security in Computing and Communications (SSCC-2015)*, Kerala, 2015. Communications in Computer and Information Science Series (CCIS), Springer.

[5]. J. Domingo-Ferrer and J. M. Mateo-Sanz. Practical data-oriented microaggregation for statistical disclosure control. *IEEE Transactions on Knowledge and Data Engineering*, 14(1):189– 201, 2002.

[6]. J. Domingo-Ferrer and V. Torra. Ordinal, continuous and heterogeneous k-anonymity through microaggregation. *Data Mining and Knowledge Discovery*, 11(2):195– 212, 2005.

[7]. E. Fayyoumi. A survey on statistical disclosure control and microaggregation techniques for secure statistical databases. *Software: Practice and Experience*, 40:1161–1188, 2010.

[8]. J. M. Mateo-Sanz J. Domingo-Ferrer, A. Martnez-Ballest and F. Seb. Efficient multivariate data-oriented microaggregation. *The VLDB Journal*, 15:355–369, 2006.

[9]. P. Samarati. Protecting respondents' identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering*, 13(6):1010–1027, 2001.

[10]. L. Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems*, 10(5):571–588, 2002.

[11]. E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin, V. S. Verykios, S. Vassilios and Y. Theodoridis. State-of-the-art in privacy preserving data mining. *ACM Sigmod*, 33(1):50–57, 2004.

Authors Profile

Sarat Kr Chettri is an Assistant Professor in the Department of Computer Science & Engineering and IT, School of Technology, Assam Don Bosco University, Assam, India. His research interests include data mining, machine learning and data privacy preservation.

Biplab Kumar Ray: He is pursuing his M.Tech in Computer Science and Engineering from Assam Don Bosco University, Azara (India). He has completed his Bachelor degree from North Maharashtra University, Jalgaon (India). His field of research is data mining. He has published paper in the area of classification of data.