ISSN:0975-9646

An Advanced Data Transformation Algorithm for Categorical Data Protection

Dnyaneshwar Pandurang Naik[#1], Anjana N. Ghule[*2]

[#]Research Scholar, Computer Science and Engineering Department, Government College of Engineering, Aurangabad [Autonomous] Station Road, Aurangabad, Maharashtra, India.
[*]Assistant Professor, Information Technology Department, Government College of Engineering, Aurangabad [Autonomous] Station Road, Maharashtra, India.

***Abstract*— The objective of data mining is extracted new, important and required information and knowledge from the large databases. Privacy preserving in the data mining process is one of the latest research areas which deals with the side effects of the data mining techniques such as viewing of personal or private or sensitive or confidential data to the public. The results of data mining techniques may produce complete data; but this resulting data contain private or sensitive information. This sensitive information should be modified before giving it to the public. For protecting the confidential or sensitive information, it is necessary to modify of the sensitive data items in a data set such that it should not affect the importance of the original objective of data mining. There are various types of masking techniques are available and can be used for protecting sensitive data items. In this research work, the proposed system makes use of perturbative system with applying encryption technique to sensitive data items.**

***Keywords*— Data Transformation, Categorical Data, Clustering, Sensitive Attribute, Privacy Preserving.**

I. INTRODUCTION

Due to the advances in information processing technology and the storage capacity, modern organizations collect a large amount of data. For extracting hidden and previously unknown information from such huge data sets the organizations rely on various data mining techniques such as classification, k-means clustering etc [1]. These data may contain sensitive information. During the whole process of data mining these data often get exposed to several organizations. If such organizations have enough supplementary knowledge about an individual having a record in the data set, then the organization can re-identify the record. Thus sensitive information stored about the individual can potentially be disclosed resulting in a breach of individual privacy. Therefore, we need techniques for protecting individual privacy while allowing data mining. Many noise addition techniques for protecting privacy have been designed for statistical databases but they do not take into account the requirements specific to data mining applications. Authors Wilson and Rosen investigated the prediction accuracy of classifiers obtained from data sets perturbed by existing noise addition techniques for statistical databases [2].

For example, a hospital may release patient's details to enable the researchers to analysis of various diseases. Table 1 shows the original datasets of patient's information. In that some sensitive numeric's as well as categorical attributes which may not be displayed to the public that should be hiding or modified by using some privacy preserving techniques. Table 2 shows that the modified data or fuzzified data [3]-[5].

Table 1 Patient's Dataset

Name	Id	Age	Gender	Marital status	Disease
Ashish Rane	01	24	M	Unmarried	Dengu
Suwarna Tale	03	22	F	Married	SwineFlu
Sandhya Patil	06	40	F	Divorce	Malaria
Nitin Gore	08	35	M	Married	Malaria
Gajendra Sane	09	21	M	Unmarried	SwineFlu
Raju Rastogi	12	30	M	Divorce	Dengu
Pooja Sakhare	14	23	F	Unmarried	Malaria
Rashi Wani	15	20	F	Unmarried	Dengu

Table 2 Modified Patient's Dataset

Id	Age	Gender	Disease
**	20-30	M	Dengu
**	20-30	F	SwineFlu
**	30-40	*	Malaria
**	30-40	M	Malaria
**	20-30	M	SwineFlu
**	30-40	*	Dengu
**	20-30	F	Malaria
**	20-30	*	Dengu

In Table 2 some attributes are directly suppressed from the dataset which may not affect on research and also maintain privacy of individual (e.g. Name attribute, marital status) and some attributes their values are modified (e.g. Id, Gender, Age).

The content in this paper is organized as follows. Section II includes an overview of existing privacy preserving techniques. Section III includes proposed system with the detailed explanation about new technique. Conclusions are given in Section IV.

II. LITERATURE SURVEY

A considerable amount of work on privacy preserving data mining methods has been reported in recent years. There are two types of masking techniques such as perturbative and non perturbative technique. In perturbative masking technique original data are modified. Resampling, random noise addition, rounding, lossy compression are some perturbative masking techniques. In non perturbative techniques the original data are not modified, but some sensitive data are removed from the dataset. Global recoding, top coding, local suppression, bottom coding are some non perturbative masking techniques [1], [5]. The most relevant work about perturbation techniques for data mining includes the random noise addition methods, the condensation-based perturbation, rotation perturbation and projection perturbation. In addition, k-anonymization, l-diversity, (n, t) - proximity can also be regarded as a perturbation technique [6]. All these perturbative techniques work on numeric sensitive attribute. In this paper we proposed new perturbative masking technique. Since our work is relevant to the only perturbation techniques. Following are the some perturbative techniques.

A. Perturbative Masking Techniques

Perturbation is nothing but altering an attribute value by a new value. The data set are distorted before publication. In the database data is distorted in some way that affects the protected data set, i.e. it may contain some mistakes. In this way the original dataset may disappear and new unique combinations of data or modified data items may appear in the perturbed dataset; in perturbation method statistics computed on the perturbed dataset do not differ from the statistics obtained on the original dataset [1], [5].

- Additive noise
- Micro aggregation
- Rounding
- Rank swapping
- Resampling etc

B. Non Perturbative Masking Techniques

Non-Perturbative methods do not modify the values of the variables rather they produce a reduction of detail in the original data set i.e. some data are suppressed or removed [1], [5].

- Top coding
- Bottom coding
- Sampling
- Local suppression
- Global recoding
- Generalization etc.

Perturbative masking techniques are as given below,

1) Noise Additive Perturbation: This technique is based on addition some noise to the attribute values of the dataset. The column-based additive randomization is the typical additive perturbation technique. This technique mainly work on attribute column so, this type of techniques relies on the facts that, first Data owners may not want to equally protect all values in a record, thus a column-based value distortion can be applied to perturb some sensitive columns of attributes. Second advantages of this technique is data classification models to be used do not necessarily require the personal records with the related to the columns, but only the column value sharing with the assumption of independent columns. The basic method is to alter the appearance of the original values by injecting certain amount of additive random noise or distinct values, while the specific information, such as the column disburse, can still be effectively reconstructed from the perturbed or bothered data. A typical random noise addition model can be precisely described as follows. We consider the original values are (p1, p2,..., pn) from a column of the dataset to be randomly drawn from a random variable P, which has some kind of distribution. The randomization technique transform the original data by adding random noises or values X to the original data values of the column of the dataset, and generates a modified data column as Y, Y = P + X. The resulting record (p1+x1, p2+x2,..., pn+xn) and the distribution of X are published [7]-[9].

2) Condensation-based Perturbation: The condensation approach aims at preserving the covariance matrix for multiple columns. In this technique masking multiple columns as a whole to generate perturb datasets. Thus, shape of decision boundary is well preserved. The data mining algorithms can be applied directly to the perturbed dataset without new development of algorithms. The condensation approach can be described as follows. This technique starts by partitioning the original data into k-record groups. Within a group, it is not possible to distinguish different records from one another. Each group has a certain minimum size k, which is referred to as the indistinguishability level of that privacy preserving approach. The greater the indistinguishability level, the greater the amount of privacy. Each group of this record is formed by two steps 1) selecting a record from the existing records randomly as the centre of group, and then finding the (k − 1) nearest neighbours of the centre to be the other (k − 1) members. The selected k records are removed from the original dataset before forming the next group in this way all the records in dataset are divided into k- records. Each record in group is distinct from others. Due to this partition of data into k groups each group has small domain, it is possible to regenerate a set of k records of the dataset to approximately preserve the distribution and covariance of the dataset values. The record regeneration algorithm tries to preserve the eigenvectors and eigenvalues or a value of a parameter of each group under the given conditions. The authors (Charu C. Aggarwal and Philip S. Yu) demonstrated that the condensation approach can well preserve the accuracy of classification models if the models are trained with the modified data. However, we have observed that the condensation approach is weak in protecting data privacy and it has some limitations. As stated by the authors, the smaller the size of the domain is in each group, the better the quality of preserving the covariance with the regenerated k records is [9], [10].

3) Rotation Perturbation: This technique first proposed for privacy preserving data clustering. Rotation perturbative technique is one of the major components in geometric based perturbative algorithm. In this we first apply rotation perturbative to privacy-preserving data classification algorithm and then addressed the general problem of privacy evaluation for multiplicative data perturbations. The rotation perturbation is simply defined by using randomly generated rotation of matrix. For example, consider one sensitive attribute ATM Pin and apply rotation perturbative is given by RP function i.e. RP (DataSet) = Ran* (DataSet)*n where Ran*(DataSet) is a randomly generated rotation matrix and (DataSet)*n is the original dataset [9], [11].

4) Random Projection Perturbation: This technique refers to projecting a set of data points from the original multidimensional space to another randomly chosen space. For example, consider Rk* d be a random projection matrix, where R's rows are orthonormal. In this RPP is the random projection perturbative function RPP (DataSet) = qd kR(DataSet) is applied to modified the dataset.

5) Modification based techniques: In these techniques original data just modify and give to the algorithms. In these techniques following are the operations on the attribute values are made [5].

- Interchange the values between records.
- Modify the original data by using some sample.
- Adding noise to the attribute values in the database.
- Sampling the result of query.

6) Refurbishing-based techniques: In these techniques the original data is constructed from the randomized data.

Algorithm:

1. Creating randomized data replica by data perturbation of entity data records.
2. Recreate distributions, not values in personal records.
3. By means of using the reconstructed distributions, fabricate the original data [5].

7) Cryptography-based techniques: In these methods data is encrypted using some techniques like Caeser cipher, hill cipher etc. This technique is used for preserving information security such as entity authentication, confidentiality, data integrity and data origin authentication. There is lager amount of sensitive and private information, which can be Bank account information, Internet banking passwords, Credit Card Information, ATM pins, Social Security Numbers, Private correspondence, Military statement above all are the sensitive information on which cryptographic techniques are applied [5].

8) K-anonymity: In this technique each attributes are suppressed or generalized until each row is identical with at least k-1 other rows thus it prevents definite database linkages. The result of this technique provides guarantees that the data released is accurate. K-anonymity works on techniques generalization and suppression. At the time of releasing micro data, some sensitive information remains contains in the micro data, such as names and social security numbers. K- Anonymity is emerging concepts in micro data protection, which has been recently proposed. Generalization technique losses amount of information. Limitations of k-anonymity are: it does not protect whether a given individual is in the database, it disclosure individuals' sensitive information, it does not protect against attacks based on background knowledge [6], [12], [13].

9) Randomization Method: It is the process of making something random. The randomization method is most preferable in current privacy preserving data mining technology. These methods also used for knowledge discovery. In this technique some noise is added to the attribute values to mask the fields of records. Randomization is used in statistics and in gambling.An easy way to comply with the conference paper formatting requirements is to use this document as a template and simply type your text into it [8].

III. Proposed System

All the existing privacy preserving data mining techniques work on the numerical attributes. Releasing microdata may contain categorical sensitive information which may violet the privacy of individuals. In proposed system we are going to work on categorical information such as marital status, gender etc. First we extract micro data from database which are going to publish for the research purposes in that all the sensitive information may be numerical or categorical information should be suppressed or removed or modified by using advanced data transformation technique. The advanced data transformation technique uses categorical attributes like gender, marital status etc [14].

Following are the steps for the advanced data transformation technique.

1. Categorical sensitive attribute selection.
2. Apply advanced data transformation algorithm.
3. Applying clustering algorithm on fuzzified data.
4. Result analysis.

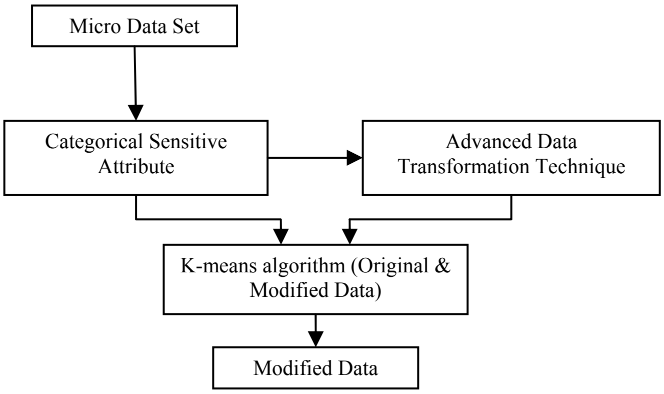
Micro Data Set
Categorical Sensitive Attribute
Advanced Data Transformation Technique
K-means algorithm (Original & Modified Data)
Modified Data

Fig. 1 proposed System Architecture

Advanced Data Transformation Algorithm:

1. Consider a database D consists of T tuples.

D= {t1,t2,…,tn}. Each tuple in T consists of set of attributes T= {A1, A2,… ,An} where Ai Ɛ T and Ti Ɛ D.

2. Consider disease name as the sensitive or confidential categorical attribute AR.

3. Divide the AR values into two clusters 'curable' and 'noncurable' disease.

4. Now apply improved polygraphic cryptography technique depending on two clusters.

5. For the curable cluster we will use ForwardCrypto() function and for noncurable cluster we will use ReverseCrypto() function.

ForwardCrypto() Function Steps:

1. Identify curable disease name and its index number from database.
2. Concatenate the index number with disease name and use it as a single string and find out its length.
3. Now we will use new string length as the key. Now apply Caesar cipher on the new string for which the key depends on the length of the key. i.e. apply Caesar cipher on the string from right end to left end i.e. from last letter to first letter with the key value equal to length of the original input string for last character & will be decreased by 1 for each letter till reached to first letter.
4. Perform left shifting operation, if there is any value repeated for sensitive item.

ReverseCrypto() Function Steps:

1. Identify non-curable disease name and its index number from database.
2. Concatenate the index number with disease name and use it as a single string and find out its length.
3. Now we will use new string length as the key. Now apply Caesar cipher on the new string for which the key depends on the length of the key. i.e. apply Caesar cipher on the string from left end to right end i.e. from first letter to last letter with the key value equal to length of the original input string for first character & will be decreased by 1 for each letter till reached to last letter.
4. Perform right shifting operation, if there is any value repeated for sensitive item.

IV. Conclusions

Protecting the sensitive information and to extracting meaningful information from the large datasets by using some data mining algorithm is a very difficult task .After the mining process secret data should be hidden from public. The proposed data transformation technique protects categorical data that produces better result than previous one. In this research first we have modified the categorical sensitive data by using advanced data transformation technique embedded with the cryptography technique which prevents sensitive items from public disclosure without affecting the objective of data mining. The proposed system maintains greater accuracy while preventing sensitive data from unauthorized disclosure.

Acknowledgment

Dnyaneshwar P. Naik is thankful to Prof. Anjana N. Ghule, Assistant Professor, Information Technology Department for her motivation and continuous support throughout this project. He is also thankful to the Principal, Government College of Engineering, Aurangabad [Autonomous] and Prof. V. P. Kshirsagar, HOD, Computer Science and Engineering Department, Government College of Engineering, Aurangabad [Autonomous] for being a constant source of inspiration.

References

[1] Han, J. and M. Kamber, 2001, *"Data Mining: Concepts and Techniques",* Morgan Kaufmann Publishers, San Francisco, CA, pp: 344- 416.

[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] M. Prakash , Dr. G. Singarave, *"A New Model for Privacy Preserving Sensitive Data Mining",* Department of CSE., K.S.R. College of Engineering, Tiruchengode, Tamilnadu, India, Department of IT., K.S.R. College of Engineering, Tiruchengode, Tamilnadu, India, IEEE-20180, July 26-28, 2012, Coimbatore, India.

[4] MIAO Yuqing, ZHANG Xiaohua, WU Kongling, SU Jie, 2011, *"An Efficient Algorithm for Privacy Preserving Maximal Frequent Itemsets Mining",* Computer Science and Engineering College, Guilin University of Electronic Technology, Guilin 541004.

[5] Ms Shalini Lamba, Dr S. Qamar Abbas, *"A Model For Preserving Privacy Of Sensitive Data",* Department Of Computer Science, Department Of Computer Science, National P.G. College, Ambalika Institute of Management & Technology, Lucknow, India, International Journal of Technical Research and Applications e-ISSN: 2320-8163, www.ijtra.com Volume 1, Issue 3 (July-August 2013), PP. 07-11.

[6] Ninghui Li Tiancheng Li, *"t-Closeness: Privacy Beyond k-Anonymity and l-Diversity",* Department of Computer Science, Purdue University.

[7] 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.

[8] Md. Zahidul Islam, *"Privacy Preservation in Data Mining Through Noise Addition",* School of Electrical Engineering and Computer Science University of Newcastle Callaghan New South Wales, 2308 Australia, November 2007.

[9] Keke Chen, Ling Liu, *"A Random Rotation Perturbation Approach to Privacy Preserving Data Classification".*

[10] Charu C. Aggarwal IBM T. J. Watson Research Center, USA and Philip S., *"Privacy Preserving Data Mining: Models and Algorithms",* Yu University of Illinois at Chicago, USA, pp: 13-120.

[11] 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.

[12] Aggarwal, G., Feder, T., Kenthapadi, K., Motwani, R., Panigrahy, R., Thomas, D., and Zhu, A., 2004, *"k-Anonymity: Algorithms and Hardness",* Technical report, Stanford University.

[13] Domingo-Ferrer, J & Torra, V (2005), 2005, *"Ordinal, Continuous and Heterogeneous k-anonymity through Micro aggregation, Data Mining and Knowledge Discovery",* vol. 11, no. 2, pp. 195-212, 2005.

[14] 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.