Indian Journal of Science and Technology, Vol 9(47), DOI: 10.17485/ijst/2016/v9i47/107360, December 2016
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Privacy Preserving Data Mining for Ordinal Data using Correlation Based Transformation Strategy (CBTS)

N. P. Nethravathi[1*], Prasanth G. Rao[1], Chaitra C. Vaidya[2], S. Geethanjali[2], P. Madhura[2], K. Neha Nandan[2], P. Deepa Shenoy[2], M. Indiramma[3] and K. R. Venugopal[2]

[1]Visvesvaraya Technological University, Belagavi - 590 018, Karnataka, India; nethravishva123@gmail.com
[2]University Visvesvaraya College of Engineering, Bangalore University, Bangalore - 560001, Karnataka, India
[3]BMS College of Engineering, Bangalore - 560019, Karnataka, India

Abstract

Objectives: Preservation of privacy is a significant aspect of data mining. The main objective of PPDM is to hide or provide privacy to certain sensitive information so that they can be protected from unauthorized parties or intruders. **Methods/ Statistical Analysis**: Though privacy is achieved by hiding the sensitive or private data, it will affect the data mining algorithms in knowledge extraction, so an effective method or strategy is required to provide privacy to the data and simultaneously protecting the quality of data mining algorithms. Instead of removing or encrypting sensitive or private data, we make use of data transformation strategies that keep the statistical, semantic and heuristic nature of data while protecting the sensitive or private data. **Findings:** In this paper we studied the technical feasibility of realizing Privacy Preserving Data Mining. In the proposed work, Correlation Based Transformation Strategy for Privacy Preserving Data Mining is used for ordinal data. We apply the method on few datasets namely soybean, Breast Cancer, Nursery dataset and Car dataset. We tabulate the end results applying the proposed strategy on both the original and the transformed dataset and observe correlation difference, Information Entropy and Classification Accuracy with different machine learning algorithms and Clustering Quality. **Application/Improvements:** As an improvement, the proposed work can be extended by use of vector marking techniques where these techniques help in increasing the efficiency by avoiding unauthorised access to the information.

Keywords: Correlation Analysis, Nominal Data, Ordinal Data, Privacy Preserving Data Mining, Transformation Strategy

1. Introduction

Data Mining is extensively used in varied areas like financial data analysis, retail industry, biological data analysis and many more. However, it has got its downsides. One of the key issues raised by data mining technology is not a business or technology one, but a social one. It is the privacy of an individual or a company. Data Mining makes it achievable to evaluate everyday business transactions and gather a considerable quantity of information about individuals buying habits and preferences. Many companies are making fortune aggregating petite pieces of information about people and putting scraps together to build a digital profile. Most of the times the

*Author for correspondence

information collected will be used to sell stuff, which is useful. However, the information extracted can be used for privacy violating purposes. Agencies, hospitals and other organizations often need to publish micro data for research and other purposes. However, the information extracted can be used for privacy violating purposes. As explained in[1] micro data is usually stored in the form of table where each row represents an individual. Here the table has three types of attributes: 1. Identity attribute (To uniquely identify an individual like name), 2. Quasi identifier (which includes demographic attributes), 3. Sensitive attributes (which include confidential information like diseases). Quasi identifiers attributes may be merged with other public databases to uniquely identify the individual and their sensitive data (Linking attack). Thus privacy is becoming a critical issue which led to a new research field called Privacy Preserving Data Mining (PPDM)[2]. PPDM comes into picture in the situations like the one described above. PPDM helps to perform data mining efficiently while preserving the private data or information about an individual or a company. Instead of hiding or encrypting, PPDM transforms the sensitive data to some other form while preserving the usefulness of the data. Many strategies have been proposed for PPDM, one of such is Correlation Based Transformation Strategy (CBTS) which is used on numerical data. The datasets also contains ordinal and nominal data; the need is to convert the ordinal and nominal data to numerical data by preserving the data utility, so that the algorithm can be applied efficiently.

In this paper, we propose a CBTS which can be applied to ordinal values. We describe a technique to convert the both ordinal and nominal data to numerical data on which the CBTS can be applied. We measure the Information Entropy values of both Original data and Transformed data and the results are comparable and also we measure Cluster Misclassification Error and prove the error is less in our approach. The paper is organized as follows: Section 2 describes Related Work. Section 3 explains Problem Definition. Architecture is presented in Section 4. Result is discussed in Section 5. We conclude this paper with future work in Section 6.

2. Related Work

In[3], the authors have used a technique called modified data transitive technique in which the sensitive numerical data item is to be protected by modifying the original data item. There is a comparison between the modified data transitive technique and the perturbative masking techniques such as additive noise, rounding and micro aggregation and performances are analyzed and results are drawn by concluding with the satisfactory results using the transitive techniques.

In[4] authors proposed a new approach which involves in preserving sensitive information using fuzzy logic. Clustering is done, in which the original dataset i.e. numerical data is transformed into fuzzy data and then noise is added to the numeric data using an S shape fuzzy membership function.

The Clusters which are generated using the fuzzified data is similar to the original cluster and privacy is also achieved.

In[5] proposed a system which makes use of a perturbative system where encryption technique is applied to sensitive data items. The information has to be changed to a considerable extent before it is made available to the public for safe guarding the confidentiality of the sensitive information. The proposed data transformation technique protects categorical sensitive data which is modified using advanced data transformation technique including cryptography technique which prevents sensitive items from public disclosure. This system gives greater results while preventing sensitive data from unauthorized disclosure and should not affect the importance of the original objective of data mining.

In[6] and[7], the authors have proposed distortion based techniques to meet the privacy requirements. In the former randomized distortion technique is applied only on confidential categorical attribute. In latter probabilistic distortion method is used on original data before using frequent item set mining on the data.

In[8] and[9], the authors have used correlation based techniques to achieve privacy in huge datasets. In paper[10] authors proposed a work which concentrates on finding an efficient solution for the classification problem over encrypted data in cloud. This work protects the privacy of sensitive data of users query and data access patterns. A k-NN classifier is developed firstly on a real-world dataset for different parameters and the efficiency is resolved.

Authors of[11] proposed a new patient centric clinical decision support system, which is of a great help for a

clinician complementary in diagnosing the risk of patient's disease without compromising its privacy. This method portrays correlation by spatial proximity. It involves the following methodologies which can handle categorical and numerical variables.

Authors of[12] proposed various methods and possible risks by the method of Random Projection. It defines a number of reconstruction techniques over the data.

In paper[13] authors concentrate on decision tree learning, without accompanying loss of accuracy. This method strives at preserving the privacy of data which are partially lost. This deals with the production of a set of unreal datasets which can be obtained as a result of conversion of original dataset. Such that, redesigning of original samples without the entire group of unreal datasets becomes impossible. From these datasets the decision tree is built precisely. And also this method is congruent with that of the other approaches which preserve the privacy of sensitive data and thereby ensuring higher protection of data.

In paper[14] the authors have proposed a method which provides an excellent spatial transformation method to protect the privacy concerns in cloud computing and this method also provides considerably good results with respect to the communication cost.

In[15] presents an erratic system based chaotic signal generator. Due to the characteristics of chaotic signal, estimators find it hard to estimate original data since they work on noise Probability Distribution Function (PDF). The issue of maintaining data privacy while publishing is resolved. Data Perturbation level depends on trust on which the data is to be generated. Due to the different levels of trusts or same levels of trust of same data, a problem on security of data arises and may cause estimation of accurate data copies by Linear Least Squares Error (LLSE), which is an advance computational algorithm.

3. Problem Definition

Given large structured data constituting of sensitive information of ordinal nature, the objective is to preserve privacy by transforming the ordinal data into an equivalent numeric representation while retaining the original statistical nature with minimal entropy.

4. Architecture

Given a huge data containing ordinal sensitive information, our solution first converts the ordinal and nominal data to numerical data and transforms the resultant numeric data in such a way that it retains the correlation structure among the data values preserving its usefulness and maintaining the level of privacy. The conversion of ordinal data is done by taking input for each data value from the concerned user and the conversion of nominal data is done by assigning random numbers to each nominal data value. The numeric attributes are retained. We consider a dataset containing mixture of ordinal, nominal and numerical data attributes, in which many attributes are private and sensitive. The dataset is subjected to clustering method like Simple K Means to group the similar rows and classification algorithm like J48. The objective of this paper is to convert and transform the ordinal sensitive data such that the correctly classified instances and the decision trees of original data and transformed data are comparable.

For the given dataset with numerical sensitive information, authors in paper[16] proposed CBTS for numerical data. Given a dataset comprising sensitive and private data, CBTS produces an outcome comprising of the subset of vectors correlated to sensitive data and produces equivalent components as substitutes. CBTS uses Pearson's correlation coefficient. The subsets generated are subjected to transformation strategies that tend to converge on the obtained similarity forming new components. Hence the components obtained are a mathematical representation of the sensitive data and used instead of sensitive data for data mining. Figure 1 gives the Architecture of CBTS for Numerical data.

Existing transformation methods PCA, SVD and NNMF have been used prior in PPDM by[17–19] demonstrate the required property of convergence. The method was able to remove the highly correlated sensitive data and transform the non correlated sensitive data. CBTS is applied to datasets which has numerical values, the information entropy values are compared for the original data and the transformed data and the results are obtained. Thorough experiment analysis proved the proposed dataset transformation method has low clustering misplacement error and minimal deviation in classifier accuracies.

In this paper we are extending CBTS to support ordinal data. The proposed architecture is shown in Figure 2. Our method first converts both ordinal and nominal data to equivalent numerical data.

Nominal and Ordinal data are defined below:

- Nominal Data or Categorical data: This is the type of data where there is no intrinsic ordering between the categories. For example, the gender has two categories male and female and there is no intrinsic ordering to the categories.
- Ordinal Data: Ordinal data is similar to a categorical data. But, there are many categories and these categories have intrinsic ordering. For example, the economic status has three categories low, medium and high which have an intrinsic ordering based on the income drawn.

The conversion step has two sub-steps. Initially, the dataset is parsed to extract the unique data values in each column which is given to next step. In the next step, based on the type of the data values of the column, conversion is done. When the column has ordinal data values, they are converted to numerical values based on the user provided ordering.

In this work we have assumed all the nominal data to have some ordinal nature. Nominal data are substituted by unsupervised statistical methods. Correlation coefficient is calculated for the respective values against the data vectors. If there exists a strong correlation, then they are converted to random numbers. If the correlation is weak, then the conversion is done by substituting categories with close ranged numbers to avoid and minimize error bias. Chi-squared test is done to determine the correlation between nominal data values. The value of the test-statistic is:

$$X^2 = \sum_{K=1}^{n} \frac{(O_k - E_k)^2}{E_k}$$

O_k - Observed frequency.
E_k - Expected frequency.

Algorithm 1: Conversion of ordinal data to numerical data.

Input: Original dataset *DS* with ordinal, nominal and numerical data.

Output: Converted dataset *DS'* with only numerical data.

Begin

Step 1: Read the dataset.

Step 2: Parse the file column wise and extract unique values from each.

Step 3: Repeat step 4 for each column.

Step 4: Based on the type of data in each column, convert the data.
If the column under consideration has ordinal data then take appropriate inputs from the user. Else if the column has nominal data, then replace the categories in that column with random numbers. If the column has numerical data, then the values in that are retained.

Step 5: The result of the above steps is the converted dataset *DS'* which is given to CBTS algorithm for transformation.

End

Algorithm 2: CBTS

Input: Converted data DS' from algorithm 1 and list of private columns P.

Output: *Transformed data DS" generated from converted data DS'.*

Begin

Step 1: Construct the Correlation matrix (*Dc*).

Step 2: Normalize the original data.

Step 3: Repeat the following steps from 3 to 6 for each private column p_i in *P*.

Step 4: Calculate threshold coefficient for selected column pi which separates the highly

correlated data of size separation factor.

Select columns whose correlation coefficient with selected private column p_i is

greater than the threshold correlation value to obtain the subset.

Step 5: Transform subset using required transformation technique or perturbation method.

Step 6: Substitute corresponding component of transformed data in place of p_i in

normalized original data.

Step 7: Denormalize the original data.

Step 8: Return *DS"* transformed data.

End

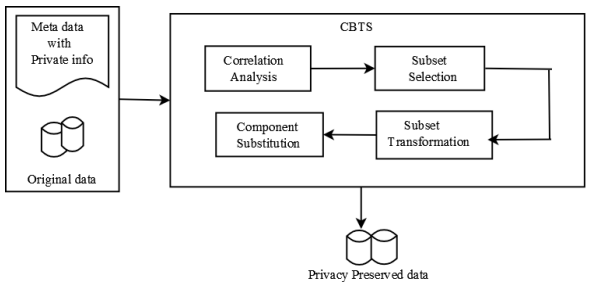

Figure 1. Architecture of CBTS for numerical data.

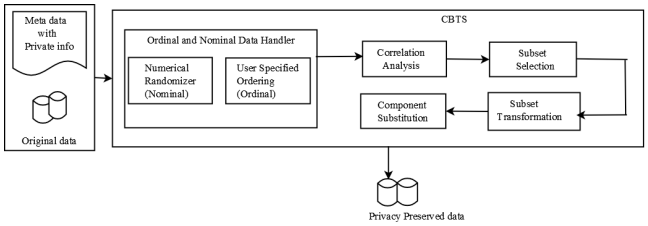

Figure 2. Architecture of CBTS for ordinal and nominal data.

The overall process in our transformation method is given in Figure 3.

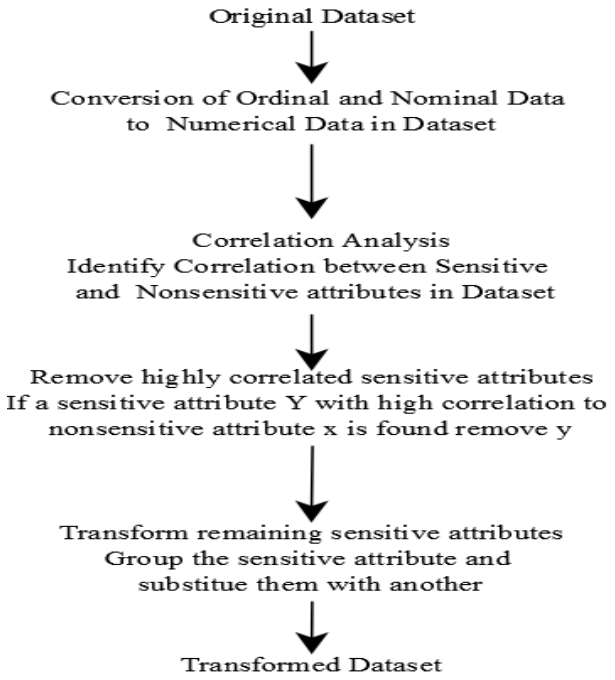

Figure 3. Transformation method.

5. Result

The datasets used in this paper are Soybean and Breast Cancer. Both the datasets are taken from UCI Machine Learning Repository. Soybean dataset is a dataset with 307 instances and 35 attributes. Among 35 attributes some are ordinal and some are nominal. Breast Cancer is another dataset with ordinal, nominal and numerical attributes. There are 286 instances and 10 attributes in this dataset. This dataset contains two classes and among 286 instances, 201 belong to one class and the other 85 belong to another class.

Information Entropy of original data against perturbed data using CBTS for Ordinal data with transformation methods is summarized in Table 1. We can infer from the table that deviation in Information Entropy is minimum using proposed CBTS method against using transformation techniques alone. Table 2 gives the comparison of classifier accuracies for various machine learning algorithms using CBTS against original data. It is clearly observable from the results the classifier performance is comparable to the original data. Table 3 shows the Misclassification Error M_E values with k-means clustering. Higher M_E values indicates lower clustering quality where as Lower M_E values indicate the higher utilization of the data.

Table 1. Comparison of information entropy

Types of Data	Original Entropy	Information Entropy(I_E) Using CBTS Method/Using existing Methods		
		PCA	SVD	NNMF
Soybean (683x36)	3.317	3.30/10.25	3.41/5.12	3.25/9.26
Car (1729x6)	2.31	2.28/5.16	2.37/2.58	2.22/9.14
Nursery Dataset (12960x7)	1.88	1.88/4.6	1.95/2.8	1.86/11.0
Breast Cancer (286x9)	3.02	3.7/6.39	3.5/3.8	3.39/8.04

6. Conclusion and Future Work

CBTS achieves accountable privacy by applying correlation transformation based methods. CBTS has applications over varied areas involving huge data. Combined with the CBTS we have presented a way of transformation by converting sensitive ordinal and nominal data to numerical data of a considered dataset simultaneously preserving the privacy and the data utility of the same. The proposed work can be extended by use of vector marking techniques where these techniques help in increasing the efficiency by avoiding unauthorised access to the information.

Table 2. Comparison of various machine learning algorithms using CBTS (M_E)

Types of Data	Machine Learning Algorithms	Observed Classifier Accuracy (%)				
		Ordinal and Nominal Data	Numerical Data	Transformation using CBTS		
				PCA	SVD	NNMF
Soybean (683x36)	Decision Tree	97.0	96.3	97.6	97.0	97.0
	Multilayer Perceptron	99.8	93.3	94.8	95.0	95.0
	Naïve Bayes	93.7	82.1	82.5	81.8	81.8
Breast Cancer(286x9)	Decision Tree	81.4	81.4	81.4	81.4	81.4
	Multilayer Perceptron	84.6	84.6	84.2	84.6	84.6
	Naïve Bayes	73.4	73.4	73.4	73.4	73.4

Table 3. Cluster misclassification error (M_E)

Types of Data	Clusters (k)	M_E(with CBTS)			M_E(without CBTS)		
		PCA	SVD	NNMF	PCA	SVD	NNMF
Soybean(683x36)	2	0.253	0.455	0.248	0.999	0.999	1.0
	3	1.22	0.88	0.74	1.09	2.6	0.9
Breast Cancer (286 x 9)	2	0.017	0.7	0.7	1.3	1.60	1.50
	3	0.7	0.74	0.74	0.5	1.91	1.54

7. References

1. Lamba S, Abbas Q. A model for preserving privacy of sensitive data. International Journal of Technical research and Applications. 2013 Jul-Aug; 1(3):7–11. ISSN: 2320-8163.
2. Naik, DP, Ghule AN. An advanced data transformation algorithm for categorical data protection**.** International Journal of Computer Science and Information Technologies. 2013; 4(6):899–902.
3. Boora RK, Shukla R, Misra AK. An improved approach to high level privacy preserving itemset mining. 2009 Dec; 6(3):216–23. ISSN: 1947 5500.
4. Sun C, Fu Y, Zhou J, Ga H. Personalized privacy-preserving frequent itemset mining using randomized response. The Scientific World Journal. 2014 Mar; 2014:10 pages.
5. Zhu T, Xiong, Li G, Zhou W. Correlated differential privacy: Hiding information in non-IID dataset. IEEE Transactions on Information Forensics and Security. 2015 Feb; 10(2):229–42.
6. Zhang Z, McDonnell K, Zadok E, Mueller K. Visual correlation analysis of numerical and categorical data on the correlation map. IEEE Transactions on Visualization and Computer Graphics. 2015 Feb; 21(2):289–303.
7. Samanthula B, Elmehdwi Y, Jiang W. k-nearest neighbor classification over semantically secure encrypted relational data. IEEE Transactions on Knowledge and Data Engineering. 2015 May; 27(5):1261–73.
8. Liu X, Lu R, Ma J, Chen L, Qin B. Privacy-preserving patient-centric clinical decision support system on naive Bayesian classification. IEEE Journal of Biomedical and Health Informatics. 2015 Jan; 20(1):1-1.
9. Sang Y, Shen H, Tian H. Effective reconstruction of data perturbed by random projections. IEEE Transactions on Computers.2012 Jan; 61(1):101–17.
10. Fong PK. Privacy preserving decision tree learning using unrealized data sets. IEEE Transactions on Knowledge and Data Engineering. 2012 Feb; 24(2):353–64.
11. Hosain AA. Shear-based spatial transformation to protect proximity attack in outsourced database. IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom); 2013.
12. Patel S, Amin KR. Privacy preserving based on PCA transformation using data perturbation technique. International Journal of Computer Science Engineering Technology. 2013; 4(35):477–84.
13. Xu S, Zhang J, Han D, Wang J. Singular value decomposition based data distortion strategy for privacy

protection. Knowledge and Information Systems. 2006; 10(3):383–97.

14. Wang J, Zhong W, Zhang J, Xu S. Selective data distortion via structural partition and SSVD for privacy preservation. IKE: Citeseer; 2006. p. 114–20.
15. Ling G. Randomization based privacy preserving categorical data analysis. Diss. The University of North Carolina at Charlotte; 2010.
16. Veryhios VS, Bertino E, Fovino IN, Provenza LP, Saygin Y, Theodoridis Y. State-of-the-art in privacy preserving data mining. SIGMOD Record. 2004 Mar; 33(1):50–7.
17. Vijayarani S, Tamilarasi A. An efficient masking technique for sensitive data protection. 2011 IEEE International Conference on Recent Trends in Information Technology (ICRTIT); 2011.
18. Singh, AP, Mathur A. A chaotic based approach for privacy preserving data mining applications with multilevel trust. 2013 IEEE International Conference on Green Computing, Communication and Conservation of Energy (ICGCE); 2013.
19. Nethravathi NP, Rao PG, Shenoy PD, Indiramma M, Venugopal KR. CBTS: Correlation Based Transformation Strategy for Privacy Preserving Data Mining. IEEE WIECON-ECE; Dhaka, Bangladesh. 2015 Dec 19–20.