ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

A Substitution Based Approach for Ensuring Medical Data Privacy.

G Manikandan*, N Sairam, V Harish, and Nooka Saikumar.

ICT Department, School of Computing, SASTRA University, India.

ABSTRACT

Data mining is defined as the process of extracting immeasurable attractive patterns from large quantity of data. The data store may contain some sensitive information related to individuals. These information needs to be preserved during data sharing and data mining. The idea of privacy-preserving data mining is to transform the original data in such a way that it doesn't reveal the confidential information to the users running the algorithm and at the same time it needs to generate the output as if the algorithm has used the original data. In this paper we put forward a substitution based process for accomplishing privacy in data mining which diverge from the traditional data perturbation methods which are used comprehensively for this purpose. The uniqueness of this approach is that it generates the sanitized data without utilizing a noise addition scheme. Bit wise substitution is used to generate the modified data from the original one. From our experiments it is established that this approach assures privacy and also enhances data utility.

Keywords: Data Utility, Data Privacy, Data Perturbation, Data accuracy, clustering.

Corresponding author

INTRODUCTION

The progression of electronic technology facilitates various public and private organizations to collect and store data as a repository from a variety of devices. The collected data may have confidential information which needs to be preserved. A detailed analysis is required in order to make the data useful for a widespread user community. Data mining is a proficient technique that hauls out momentous hidden knowledge from huge data anthology. The result is presented to the users as patterns and predictions which can be used in decision making [1]. There is a chance that the insightful information comprised in the data store may get exposed at some point in data mining. Revelation of such information is regarded as privacy breach. Public consciousness of privacy may bring in extra complication to data accumulation which forbids the organizations from utilizing the entire benefits of data mining. In the preceding years, people have attempted a lot of researches and have proposed Privacy preserving data mining as a solution for this cumbersome situation, which aims at satisfying data utility and privacy. Many realistic methods are suggested for different scenarios in privacy preserving and new methods come out ceaselessly as a result of individual's effort to achieve privacy.

Perturbation and Randomization based privacy preserving techniques generate synthetic data from an original data set that maintains some characteristics of the original data set [2–4].Data perturbation is one of the most proficient technique for preserving accuracy and privacy in data mining system. In this scheme the individual data values are modified before employing data mining technique. In this paper we discuss a substitution based data distortion approach that deviates from the traditional approaches for generating synthetic data. A bit wise substitution approach is used for generating the modified data. The uniqueness of this approach is that the modified data is generated from the original data without adding any noise as the traditional systems do. Experiments were performed using various standard benchmark data sets that are available in UCI machine repository. Experiments have shown that our proposed approach preserves privacy and results in a high degree of accuracy.

Proposed System

Even though it is easy to follow traditional approaches like transformation techniques for transforming the data, there is a difficulty in selecting a particular transformation technique based on the requirement. To make the transformation with more ease off, here we introduce a new concept of substitution technique.

The substitution technique can be done by interchanging the binary bit values of the original data. This can be done by complementing one or two or even more bits from the LSB side. The reason for choosing the LSB side is to control the range of deviation from the original value. The deviation will be larger if the MSB bits are chosen. When this technique is applied for a numerical data, whose value is to be transformed into a new value, the data value is converted into binary format, then the last one bit or even more than one bit is complemented (restricting the complementing of bits to three yields a better result as it reduces deviation range) and then the decimal equivalent of the obtained binary number gives the desired deformed data.

Here we propose three approaches. In Approach 1, the least significant bit is complemented and the corresponding decimal values are recorded. Similarly in Approach 2 and 3, two least significant bits and three least significant bits are complemented and the corresponding decimal vales are recorded. All three approaches are illustrated in the table 1.

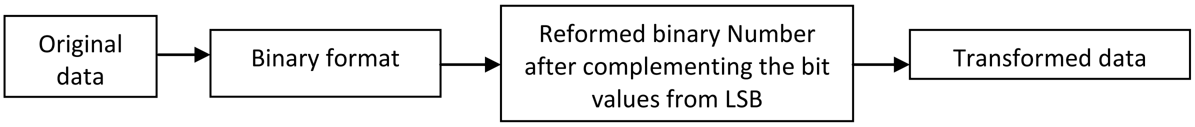
Original data
Binary format
Reformed binary Number after complementing the bit values from LSB
Transformed data

Figure 1: Steps involved in the substitution technique

EXPERIMENTAL RESULTS

For the illustration of this substitution technique, considering the age attribute of the adult dataset from the UCI repository, it is found that complementing the 'n' LSB of the binary equivalent of the decimal value yields a modified value in the range of –(2^n-1) to +(2^n-1) from the value considered to be modified. So, even though it is possible to change the entire n bits of the number, it is necessary to select the value of 'n' to yield the efficient relevant modified value. For example, if we want to alter the value of age (which probably has the range from 1 to 100) the value of n can be set to 2, so that the modified value would be in the range of -3 to +3 from the original age value. Similarly, if it is to be dealt with some sensitive data where accuracy is important, choosing the n could be limited to 1. So the redundant value can be controlled in the range of +-1 from the value considered. So, setting the value of the n is must be the prior job of the data providers to implement this technique of substitution. The k means clustering for the first 10 values of the age attribute in the adult dataset of the three approaches with respect to the actual dataset clusters with k=2 is given in the tables 2,3,4 and 5. To find the efficiency of this technique, the misclassification error (MSE) is calculated between the actual original clusters and the clusters obtained by using k means clustering for the modified values after each approach and it is briefed in table 6.

Table 1: Original dataset and the dataset after applying three approaches

DATASET	39	50	38	53	28	37	49	52	31	42
APPROACH 1	38	51	39	52	29	36	48	53	30	43
APPROACH 2	36	49	37	54	31	38	50	55	28	41
APPROACH 3	32	53	33	50	27	34	54	51	24	45

Table 2: Clustering of original dataset

DATASET	CLUSTERS	
	CLUSTER 1	CLUSTER 2
{39,50,38,53,28,37,49,52,31,42}	{39,38,28,37,31,42}	{50,53,49,52}

Table 3: Clustering of obtained approach 1 values

APPROACH 1	CLUSTERS	
	CLUSTER 1	CLUSTER 2
{38,51,39,52,29,36,48,53,30,43}	{38,39,29,36,30,43}	{51,52,48,53}

Table 4: Clustering of obtained approach 2 values

APPROACH 2	CLUSTERS	
	CLUSTER 1	CLUSTER 2
{36,49,37,54,31,38,50,55,28,41}	{36,37,31,38,28,41}	{49,54,50,55}

Table 5: Clustering of obtained approach 3 values

APPROACH 3	CLUSTERS	
	CLUSTER 1	CLUSTER 2
{32,53,33,50,27,34,54,51,24,45}	{32,33,27,34,24}	{53,50,54,51,45}

Table 6: Results of Misclassification error

MISCLASSIFICATION ERROR BETWEEN ACTUAL CLUSTERS AND	OBSERVED MSE VALUE
APPROACH 1	0.050796965695156786
APPROACH 2	0.004729584472221369
APPROACH 3	0.05380670126838856

CONCLUSION

The proposed system here is a bit wise substitution method for achieving privacy in data mining. The proposed scheme directly modifies the bit of the original data to generate the sanitized data. From the table and graph, it is observed that the privacy is achieved with maximum efficiency with negligible misclassification error in the approach 2. In future this approach can be used to modify categorical data.

REFERENCES

[1] Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, Morgan Kauffman Publishers, 2006.

[2] G.K.Gupta, Introduction to Data Mining with Case Studies, Prentice Hall of India, 2008.

[3] Benjamin C.M. Fung, Ke Wang, Ada Wai-Chee Fu; Philip S. Yu, Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques, Chapman and Hall, 2010

[4] B.Karthikeyan,G.Manikandan,Dr.V.Vaithiyanathan, “A Fuzzy Based Approach for Privacy Preserving Clustering”, Journal of Theoretical and applied information Technology , Vol 32(2), 118-122,2011.

[5] G.Manikandan, N.Sairam, R.Sudhan,Vaishnavi, “Shearing Based Data Transformation Approach for Privacy Preserving Clustering”, In Proceedings of 3rd IEEE International Conference on Computing, Communication and Networking Technologies, ICCCNT 2012

[6] G.Manikandan,N.Sairam,S.Sharmili,S.Venkatakrishnan , “Achieving Privacy in Data Mining Using Normalization” , Indian Journal of Science and Technology, Vol 6(4) , 2013,4268-4272.

[7] G.Manikandan,N.Sairam,S.Jayashree,C.Saranya , “Achieving Data Privacy in a Distributed Environment Using Geometrical Transformation”, Middle East Journal Of Scientific Research , Vol 14(1) , 2013,107-111.

[8] G.Manikandan,N.Sairam,C.Akshaya,S.Venkatakrishnan ,”An Innovative Approach for Classifying Binary Data”,International Journal of Applied Engineering Research, Vol 9(5) , 2014,589-597.

[9] G.Manikandan,N.Sairam,S.Rajarajeswari,H.Ramya,”A New Genetic Approach for Data Masking”, International Journal of Applied Engineering Research, Vol 9(7) , 2014,755-761.

[10] N.Abitha,G.Sarada,G.Manikandan,N.Sairam Presented a paper titled “ A Cryptographic Approach for Achieving Privacy in Data Mining”,In Proceedings of 3rd IEEE International Conference on Circuit, Power and Computing Technologies (ICCPCT–2015), 2015. DOI:10.1109/ICCPCT.2015.7159300

[11] G.Manikandan,N.Sairam,N.Abitha,G.Sarada,R.Pranesh,M.Vigneshwaran, “A Random Noise Based Perturbation Approach For Achieving Privacy In Data Mining”, Global Journal of Pure and Applied Mathematics, Vol 11(3) , 2015,1635-1639.

[12] G.Manikandan,N.Sairam,M.Sathiya Priya,Sri Radha Madhuri , “A General Critical Review on Privacy Preserving Data Mining Techniques”,Global Journal of Pure and Applied Mathematics, Vol 11(4) , 2015,1899-1906.

[13] UCI Data Repository http://archive.ics.uci.edu/ml/datasets.html