A Coherent Technique for Privacy Preservation in Data Mining Using Classification

Computer Science

KEYWORDS : Privacy Preserving Data Mining, Transfiguration, Tangent Hyperbolic function, Classification, Accuracy

Sheryl Parmar Amity University, 36 km stone, Jaipur-Delhi Highway, Jaipur

Pallavi Sharma Amity University, 36 km stone, Jaipur-Delhi Highway, Jaipur

Preeti Gupta Amity University, 36 km stone, Jaipur-Delhi Highway, Jaipur

ABSTRACT

Privacy Preserving Data Mining has become a significant issue in data mining research. In case of healthcare databases, the data mining system contain a large amount of sensitive and confidential data like patient centric data, treatment data, resource data etc. Here, in this paper the representation of the protection of all these kind of data using the tangent hyperbolic transfiguration technique has been proved. The transfigured values of the original data are obtained which makes the original data unrecognizable. After obtaining the transfigured values, a data mining technique is applied to check accuracy.

INTRODUCTION

Data mining is the process to explore the task relevant data whenever and wherever required. It is the most widely carried out process in today's age. But along with data mining, it has also become necessary and obvious to protect the data from attacks i.e. the data must be secured from the intrusions. The original data needs Privacy Preserving Data Mining and therefore the data can be protected in such a way that it cannot be known to any third person. In this paper, the representation of privacy preservation technique using a tan hyperbolic function for the healthcare datasets to keep data secure has been given importance to. The attacks or intrusions take place because of linking of databases, losing the individual's basic information etc. kind of issues. Hence, it becomes necessary to apply privacy preserving data mining technique which is efficient enough to lock the original data in order not to be revealed against or to be misused or misinterpreted. We show the experimental results of the healthcare datasets by generating the protective data and then applying the data mining technique over those datasets to represent the knowledge and accuracy of the privacy parameters used in the mathematical equation.

RELATED WORK

There are a lot of privacy preserving techniques available to keep the data private. Some of the techniques directly modify the data mining algorithms to protect the data while some of the techniques modify the values of the datasets in the database to maintain privacy of data. Thus, privacy preserving data mining is well researched over the years and some of the most commonly used methods are as under:

Data perturbation was introduced where any random function or value is added to the original data so that after the addition of the function, the original values are changed to some unknown values and so the data is not recognized by any third party. Later on, a rotation based perturbation method was also introduced which assured a zero accuracy loss over the data.

The secure two-party computation problem was proposed by A.C. Yao [3], which was extended later, to the secure multi-party computation (SMC) [9] by O. Goldreich. The secure multi-party computation protocol is based on the cryptography. This secure multi-party computation protocol can compute arbitrary function in distributed networks where each participant holds his inputs, while the participants do not trust each other and also not the channels by which they communicate with each other. Each participant locally keeps his/her own data private to correctly get the results. Each party provides his input that will keep private.

Sensitive Rule Hiding method was used to hide the sensitive rues that contain sensitive data, so that the rules cannot be inferred using association rule mining on the right hand side of the rules. A method was proposed by Dr. K. Duraiswamy et al. [5] to hide the sensitive rules which was able to hide the rules automatically. Wavelet transformation was proposed by Liu et al. [6] for data distortion which was a data perturbation approach to keep the distance before and after the perturbation and to preserve the basic statistical properties of the original data while maximizing their utilities.

Data swapping was proposed by Yidong Li et al., [8] which was one of the best Privacy preservation methods used.The experimental results shown that equi-width swapping was equally efficient as equi-depth swapping according to the partitions of the data even if they are large. Here, the number of partitions is equal to the square root of P, where P is the size of the dataset.

Continuously Anonymizing Streaming data via adaptive clustering (CASTLE) was proposed, a cluster-based method k-anonymize data streams and, at the same time, ensures the freshness of the anonymized data by satisfying specified delay constraints. It describes how the CASTLE can be utilized to handle l-diversity by extending it. The experimental results show that CASTLE is effective and efficient with respect to the output data quality.

SVD (Singular Value Decomposition) and SSVD (Sparsified Singular Value Decomposition) were proposed by Jahan et al. [4] for reduced feature space Here, the various privacy parameters have been used to maintain privacy with the help of distorted data and the difference between original data, distorted data and degree of privacy protection has been proved using experimental results on real life datasets.

A secure k-means data mining method was proposed by Deepti Mittal et al., [7] assuming that the data is distributed among different hosts preserving the privacy of the data in cloud environment. This approach maintains correctness and validity of existing k-means to generate the final results and also in the distributed environment.

TRANSFIGURATION OF DATA USING TANGENT HYPERBOLIC MATHEMATICAL FUNCTION:

In order to keep the data private, the tangent hyperbolic function is used to convert the original numerical values into a tanh normalization. Further, to maintain the privacy, the converted value is multiplied by 100 so that no intruder can attack intentionally or unintentionally on the private data and the original data being not known to anyone because of the strong conversion of data. Even if a user wants to know the original values, he/she won't be able to know it because the mathematical formula applicable here has a specific format which considers the mean and the standard deviation of the original data values for particular records and thereby also multiplying those values by 100 numeric value using Java programming. So, here the encryption of data here can be called in terms of cryptography to protect the data has been carried out. The tangent hyperbolic function is given by Hampel and the equation for it is as follows:

$$X_{Tan\,h} = \frac{1}{2}\left\{tanh\left(K\frac{X-\mu}{\sigma}\right)+1\right\}$$

where μ and σ are the mean and the standard deviations of the original data respectively, k is a suitable constant, X is the original data value. After getting a value by applying this formula to the data, the same value is multiplied by 100.

The steps carried out for this process are as follows:

I. Take a random data of some particular records and find their mean and standard deviations.

II. Now, implement the tangent hyperbolic equation over each record considered for an attribute having numeric data values and find the new values for every attribute record.

III. Multiply the new value after applying the tanh equation by 100 so that a new transfigured value is a resultant value.

IV. Implement classification over the original and transfigured dataset and calculate the accuracy for the correctly classified instances in WEKA tool for mining reasons.

Here k's value is taken as 1 for testing every Healthcare dataset. The value can also be taken as negative to get some different effects over the data and it's accuracy. Hence, it can be said that the values are normalized in such a manner that the original data gives the transfiguration of data which makes impossible to determine the original data. The medical data whose sensitive and confidential data are not to be disclosed can be very efficiently made disguised and thereby the accuracy can be known for the original and transfigured data.

The privacy preservation methodology adopted here is shown using the following diagram:

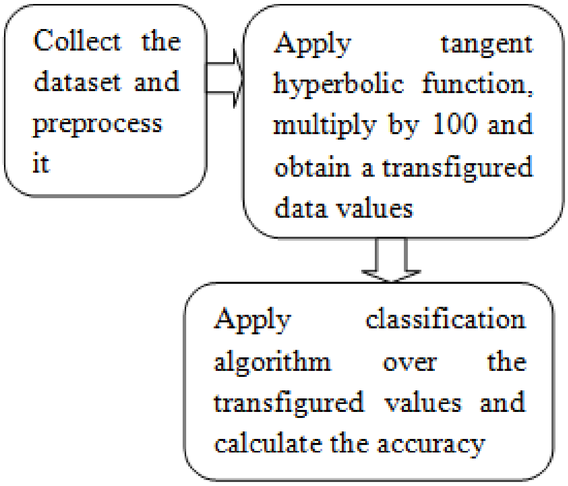
Collect the dataset and preprocess it
Apply tangent hyperbolic function, multiply by 100 and obtain a transfigured data values
Apply classification algorithm over the transfigured values and calculate the accuracy

Fig 1 The process of transfiguration for the privacy of data

EXPERIMENTAL RESULTS

Here, the results are obtained after implementation in Weka(Waikato Environment for Knowledge Analysis) where the classification is carried out for mining purpose using the J48 algorithm and the percentage values for accuracy are shown in the following table:

Datasets	Data values	Accuracy (%)
Diabetes (original)	-	90
Diabetes (transfigured)	0.78773	100
Influenza (original)	-	68
Influenza (transfigured)	60.0016	72
Heart disease (original)	-	80
Heart disease (transfigured)	4.521653333	80

Table 1 Results of datasets in WEKA

The results in the graphical form are shown below where the impact of attribute calculated for privacy is viewed in terms of accuracy. The outcomes are observed in MS Excel workbooks. For Diabetes database, the attribute BMI is modified by transfiguration and in the graph, the Y-axis indicates the values and the X-axis indicates the number of records considered. Similarly, graphs are obtained for the Influenza and Heart disease dataset where the attributes transfigured are Sample Size and max_

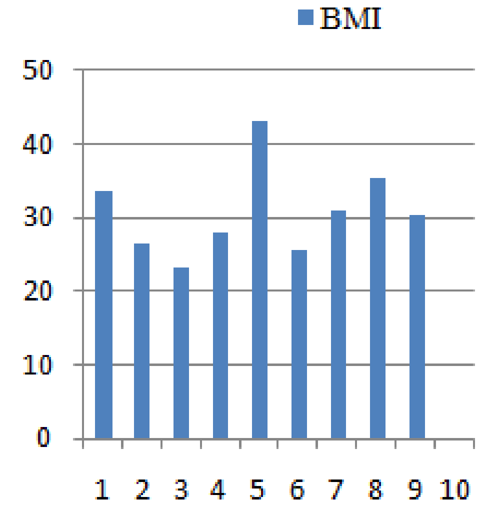

Fig 2 The accuracy of the attribute 'BMI' for the original data of Diabetes database

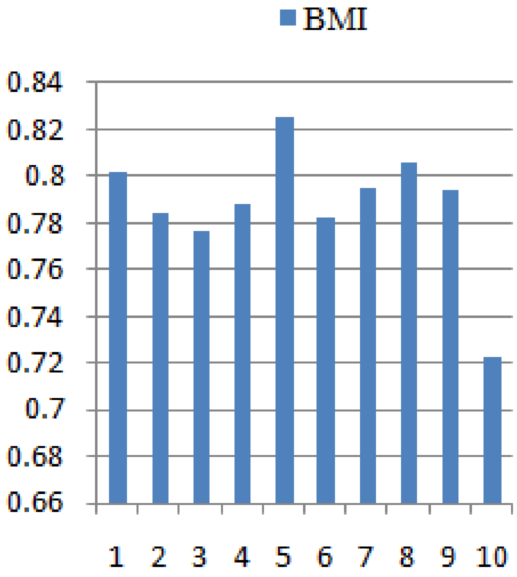

Fig 3 The accuracy of the attribute 'BMI' for the transfigured data of Diabetes database

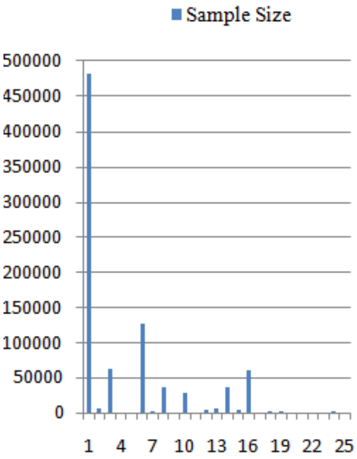

Fig 4 The accuracy of the attribute 'Sample Size' for the transfigured data of Influenza database

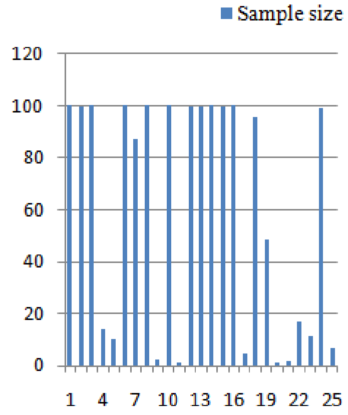

Fig 5 The accuracy of the attribute 'Sample Size' for the transfigured data of Influenza database

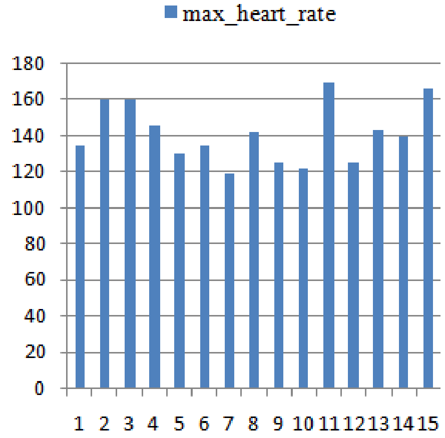

Fig 6 The accuracy of the attribute 'max_heart_rate' for the original data of Heart Disease database

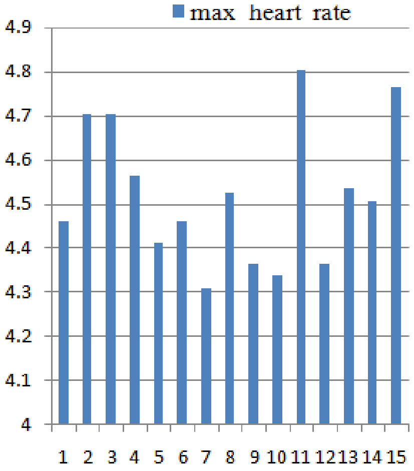

Fig 7 The accuracy of the attribute 'max_heart_rate' for the transfigured data of Heart Disease database

CONCLUSION

In this paper, we represent the transfiguration of the numerical data values by applying a tangent hyperbolic function and multiplying the same by a numeric value hundred. After getting the transfigured values, we apply the classification algorithm over those values to obtain the accuracy of the dataset and representing it for the original and transfigured values in WEKA. In future, we opt to protect our data by Privacy Preserving Data Mining using tangent hyperbolic function and using some other efficient mathematical operation instead of multiplication.

REFERENCE

[1] J. Han, and M. Kamber, "Data Mining: Concepts and Techniques", Second edition, 2006, Morgan Kaufmann, USA. | [2] The Weka Machine Learning Workbench. http://www.cs.waikato.ac.nz/ml/weka | [3] A.C.Yao, "Protocols for secure computations", In Proc Of the 23rd Annual IEEE Symposium on Foundations of computer Science, 1982. | [4] Jahan, Narsimha and Rao, "Data perturbation and feature selection in preserving privacy" Ninth International Conference on Wireless and Optical Communications Networks (WOCN), pp 1-6, 2012. | [5] Dr.K. Duraiswamy, Dr.D. Manjula , N. Maheswari (Corresponding Author), "A New Approach to Sensitive Rule Hiding", Journal on Computer and Information Science, CCSE 2008, vol. 1, No. 3, pp. 107-110, 2008. | [6] Lian Liu, Jie Wang, Jun Zhang, "Wavelet-Based Data Perturbation for Simultaneous Privacy-Preserving and Statistics-Preserving", Proceeding of IEEE International Conference on Data Mining Workshop, pp. 27-35, 2008 | [7] Mittal, Deepti ; Kaur, Damandeep ; Aggarwal, Ashish, "Secure Data Mining in Cloud Using Homomorphic Encryption", International Conference on Cloud Computing in Emerging Markets, pp 1-7, 2014. | [8] Yidong Li, Hong Shen, "Equi-Width Data Swapping for Private Data Publication", International Conference on Parallel and Distributed Computing, Applications and Technologies, pp. 231- 238, 2009. | [9] Shyue-Liang Wang, Tzung-Pei Hong, Yu-Chuan Tsai, Hung-Yu Kao3, "Hiding Sensitive Association Rules on Stars", IEEE