Biomedical Research 2017; 28 (3): 1043-1045 *ISSN 0970-938X*
www.biomedres.info

A new approach for ensuring medical data privacy using neural networks.

Manikandan G*, Sairam N, Sathya Priya M, Sri Radha Madhuri, Harish V, Nooka Saikumar

School of Computing, Sastra University, India

Abstract

The foremost intention of data mining is to extort unfamiliar patterns from huge set of data. There is a marvellous escalation in the quantity of data anthology with the rapid growth in technology. In recent years, with the advancement in information technology many hospitals started storing various information related to patients in the electronic format commonly referred as Electronic Health Records (EHRs). If used rightly EHRs can precisely identify diseases and add to the effectiveness of healthcare delivery. Quality of treatment given to the patients can be still improved if these records are shared among various hospitals. The key obstruction in data sharing is the sensitive medical information existing in the EHR. The revelation of sensitive data results in privacy breach of individuals, which results in the usage of many privacy preserving techniques. This work is provoked with the necessity to mutually guard private information and facilitate its use for research purpose. In this paper we present a new approach for generating noise to preserve privacy using a neural network. The pseudo identifiers in the data set (EHRs) are used as the input to the network nodes. Output generated by the neural network is added with the original data to generate the sanitized data. We have compared the output of the proposed method with the existing approaches and it seems too superior to the existing methods.

Keywords: Privacy, Security, Back propagation, Neural network, Data utility.

Accepted on July 11, 2016

Introduction

The information era has enabled many organizations to collect and store hefty volume of data. However, the implication of this data is insignificant if "meaningful information" is not extracted from it [1]. Data mining provides a solution to this problem by pulling out of new patterns or rules, from various data sources of varying dimension and with diverse nature [2].

During the entire course of data mining different sensitive data get exposed to several persons ranging from data collectors to assessment makers. Leakage of such information is regarded as privacy breach of an individual. Consciousness of privacy and lack of trust bring in extra intricacy to data accumulation [3,4].

From the recent surveys it is noticed that there is a steady rise in public awareness about privacy. The result is additional complexity in data collection process, which indirectly prevents the beneficiaries from utilizing the data mining algorithms in an efficient manner. Various privacy preserving methods were proposed as a solution to this problem.

A collection of privacy preserving system generates an artificial data set with the same characteristics as original data set and this data set is subsequently released to the data requestor as a result to their query. In order to ensure individual privacy, the other group of privacy preserving techniques adds noise to the original data and creates a perturbed data set. The advantage in this approach is that it allows a data miner to construct a decision tree with high quality and precision [5,6].

In an attempt to provide excellence in patient care many healthcare facilities started storing the patient information in the form of electronic health records (EHRs). Doctor's clinic and hospital are the prominent places where these records are stored and used as an aid in diagnosing diseases. It is possible to design and develop an efficient healthcare delivery mechanism if these stored information's are shared among various physicians and clinics [7].

One of the most important complications is the sensitivity of the information such as medications, diagnostics and demographics related to the patients are available in the EHR.

Physicians always need to maintain high standards in protecting patient privacy due to these permissible distress and medical ethics [8]. Sharing of patient information has become a main concern and has hindered the effectiveness of using EHRs.

Literature Survey

Numerous methods are available in the literature for ensuring privacy in data mining and generally these methods are divided into two major categories. Data mining algorithms are modified in the first category so that they may execute data mining operations on scattered datasets without knowing the exact values of the data. To protect privacy, methods in the second group transform the original values of the datasets.

In practical two-way data examination applications, datasets from diverse data owners have to be composed and shared to mine valuable information. In these cases, it is unfeasible that all data possessors use the identical data deformation method. Individual data holders are permitted to apply some data distortion methods to their datasets. It is vital to know if these datasets from diverse data owners using dissimilar deformation methods can be pooled.

In Decision tree classifiers are constructed by means of an additive noise based data perturbation [9]. Each data element in the input set is randomized by adding some noise selected autonomously from a recognized distribution.

K-Anonymity model is used by the data owner to protect the confidential information while sharing a collection of person-specific data. This model utilizes data generalization and suppression for generating 'k' similar records [10].

Chen et al., proposed a rotation based perturbation method which guarantees zero loss of accuracy for a lot of classifiers. Experimental results exemplify that the rotation based perturbation scheme can significantly enhance the privacy without sacrificing data truthfulness [11].

Manikandan et al., have compared the techniques like k-anonymity, Geometrical transformations and fuzzy system for preserving privacy in data mining. These methods are evaluated based on their data utility and privacy level [12].

Random noise can also be used for data modification [13]. Pseudo random generators like linear congruential generator, inverse congruential generator and RANDU can be used for generating random noise. It is observed that this approach generates modified data by adding a unique number to each data item.

The usage of various fuzzy membership functions for assuring data privacy is discussed [14]. From the experimental results it is clear that the S-Shaped membership function is more efficient than other membership functions in generating the sanitized data.

A bit wise substitution method was proposed for accomplishing data privacy [15]. The uniqueness of the proposed scheme is that it directly modifies the original data to create sanitized data. Privacy with maximum efficiency can be achieved when two bits are modified.

Proposed System

In this paper, we have generated noise using neural networks to achieve privacy. The supervised feed forward neural network is used to generate noise. Here we have constructed the Neural Network (NN) that consists of three layers namely, Input layer, hidden layer and output layer.

Every node in the network layer is connected to every other node in the successive layer. The input layer consists of various attributes of the dataset as input. The input values are processed to obtain noise in the hidden layer which has no connection with the external environment. The generated noise is obtained from the output layer, which is added to the original data to create sanitized data.

We have designed the network using back propagation algorithm as shown in Figure 1. The inputs are given to the network and then processed using the algorithm and the output is obtained from the output layer. The output is compared with the target value. If the target and the obtained output are same, then the output is considered as noise. In case of contradiction, the output value is back propagated to the hidden layer. The weights are adjusted and processed in the hidden layer. Then the value is forwarded to output layer and again the same process continues until the target and the obtained value are same.

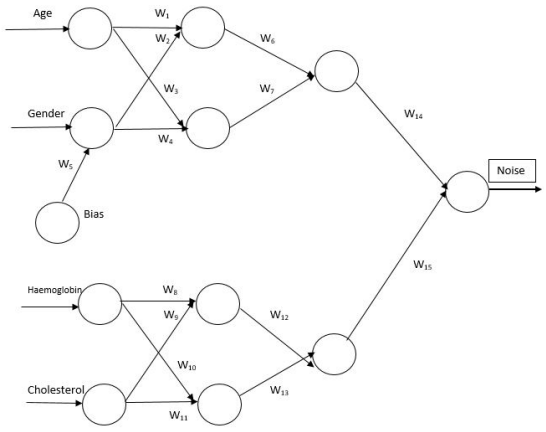

Figure 1. Proposed neural network.

Here we have taken 4 attributes as input. Initially, first two attributes namely age and gender is given as inputs to the network. Age is classified into groups with the interval of 10. Gender is classified as '1' for female and '0' for male. Weight is applied to each edge and bias is applied to the Gender.

In the next case, another two attributes namely Haemoglobin and Cholesterol is given as inputs to the network. For Haemoglobin the value in the data set is used as such. Cholesterol is classified into groups with the interval of 3. Weight is applied to each edge. Activation function is used to control the output values. In our case, we have taken the sigmoid function as the activation function to retain the output noise in the range 0 to 10. In order to get the noise in the desired range, the proposed network back propagates the output value to the hidden layers by adjusting the weights. Both the output values from the neural network are added to get the desired output. Then the resultant value is added to the original age to create sanitized age. In this way, the privacy is ensured for the sensitive attribute.

Experimental Results

For the illustration and experimental purposes, a hospital dataset is chosen and the attributes such as age, gender, cholesterol and haemoglobin are given to the input layer. The given inputs are processed using the above network. The weights in the network are chosen such that the obtained noise value is in the desired noise range. The generated noise is

passed to the output layer as the output for the network. The output is then added to the sensitive attribute age to get the sanitized data.

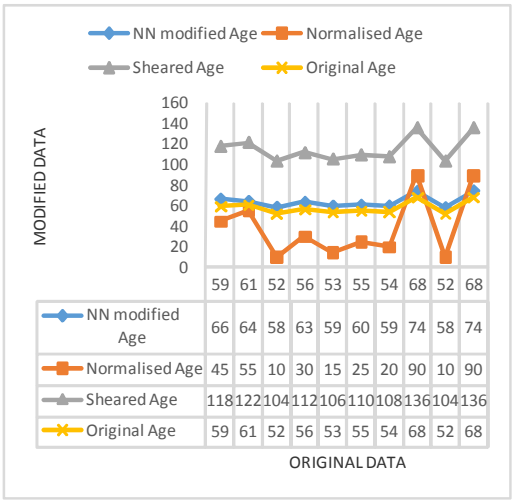

Figure 2. *Comparison graph.*

To check the effectiveness and accuracy of the resultant sanitized data, we compare the results with the data sanitized using Min-Max normalization and shearing techniques. Setting the minimum and maximum limit as 10 and 90, the modified data obtained seems to be deviated in non-uniformed manner. Whereas modified data obtained using Shearing with noise as 1 gives the doubled value of the original dataset. From the Figure 2, it can be inferred that the modified data generated using the proposed system closely resemble the original data, thus ensuring Privacy.

Conclusion

Privacy preservation plays a vital role in data mining. In this paper we have suggested a method for ensuring privacy in data mining using a neural network. We have generated the noise using a feed forward back propagation algorithm. The uniqueness in this approach is that noise is generated using more than one attribute in the data set. From the experimental results, it can be inferred that the output of the proposed method is very similar to the original data. In future this approach can be experimented with different activation functions.

References

1. Jiawei H, Micheline K. Data Mining-Concepts and Techniques. Morgan Kauffman Publ 2006.
2. Gupta GK. Introduction to data mining with case studies. Prentice Hall India 2008.
3. Margaret HD. Data Mining-Introductory and advanced topics. Pearson Edu 2003.
4. Soman KP, Shyam D, Ajay V. Insight into data mining-Theory and practice. Prentice Hall India 2006.
5. Benjamin CMF, Ke W, Ada Wai-Chee F, Philip SY. Introduction to Privacy-preserving data publishing: Concepts and techniques. Chapman Hall 2010.
6. Jaideep V, Christopher WC, Yu Michael Z. Privacy preserving data mining. Springer 2005.
7. Aggarwal CC, Philip SY. A general survey of privacy-preserving data mining models and algorithms. Springer 2008.
8. Reddy CK, Aggarwal CC. Healthcare data analytics. CRC Press 2015.
9. Agrawal R, Srikant R. Privacy-preserving data mining. Proc Acm Sigmod Conf Man Data 2000; 439-450.
10. Sweeney L. k-Anonymity-A Model for protecting privacy. Int J Uncert Fuzz Knowl Sys 2002: 557-570.
11. Chen K, Liu L. Privacy preserving data classification with rotation perturbation. Proc 5th I Int Conf D Min 2005; 589-592.
12. Manikandan G, Sairam N, Sathiya Priya M,Sri Radha M. A general critical review on privacy preserving data mining techniques. Glob J Pure Appl Math 2015; 11: 1899-1906.
13. Manikandan G, Sairam N, Rajendiran P, Balakrishnan R, Rajesh Kumar N, Raajan NR. Random noise based perturbation approach using pseudo random number generators for achieving privacy in data mining. J Comput Theor Nanosci 2015; 12: 5463-5466.
14. Manikandan G, Sairam N, Harish V, Nooka S. A substitution based approach for ensuring medical data privacy. Res J Pharma Biol Chem Sci 2016; 7: 1136-1139
15. Manikandan G, Sairam N, Harish V, Nooka S. Survey on the use of fuzzy membership functions to ensure data privacy. Res J Pharma Biol Chem Sci 2016; 7: 344-348.

*Correspondence to

Manikandan G

School of Computing

Sastra University

India