CERTAIN INVESTIGATIONS ON PRIVACY PRESERVING DATA MINING USING PERTURBATION FOR PHARMACEUTICAL DRUGS

A SYNOPSIS

Submitted by

SARANYA K

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

FACULTY OF INFORMATION AND COMMUNICATION ENGINEERING
ANNA UNIVERSITY
CHENNAI 600 025

JULY 2020

1. INTRODUCTION

Information security is one of the sensitive approaches of information sharing among the private resources. While sharing the information or transactional data few issues arise in accessing the private information, especially sensitive information. Generally the medical field is correlated with numerous sensitive information relevant to the treatment details of the patient such as drug and diagnoses information. The personal information may be accessed by the anonymous person for business motive in drug analysis field. The public accesses of transactional database consist of private information such that the unauthorized persons may exploit the transactional details and misuse the information.

Data publication is an important aspect while preserving the sensitive data. The data provider gathers the adequate information from the data owners. The collected data is to be published to the data miner, which plays a vital role in data mining operation on the collected dataset. This task is classified into two categories namely predictive and descriptive methods. The predictive method such as classification and time sequence analysis mainly focuses on present information for predictions. In descriptive method such as clustering, association rule mining (ARM) focuses on the hidden rules which may reveals the information without any predefined target.

The medical dataset contains the drug details of different patient’s medical transaction also it contains personal information. Medical organizations maintain all the information in the form of outsourcing details. But they have their own responsibility to maintain the secrecy of the patients. The medical data are exchanged between organizations which are used to perform different data mining analysis. However the data set contains private information of different patients, the entire data set cannot be shared with the third parties that are not abiding to the policy of privacy preservation. The data set may retraceable for others for verification purposes and for making them reusable beyond the original purpose for which they were collected. The data mining techniques are used to overcome the challenges and to secure the privacy.

1.1 Problem Formulation

Preserving the sensitive information in medical data is a major challenging task for the researchers, to maintain the privacy of the patient. The pharmaceutical data has number of attributes from which the most sensitive attribute has to be identified and secured before publishing the data.

Since medical transactional records are related to human resources, privacy issues are essential which is handled by data mining algorithms and techniques. But, always securing the privacy information is exists. To preserve the accuracy and information loss, task based privacy preserving techniques are developed. The organization which maintains the original data wants to provide access rights to the user to perform any data mining tasks. Even among them, only a subset of user may allow to access the data set. However the third party uses those data to infer some knowledge against the privacy policy. In this research work correlation, association rule mining and cryptographic techniques are considered to preserve the sensitive information and provides the complete privacy of the original information. The classifiers Decision Tree (DT), Random Forest(RF), Linear model, Ada Boost, Support Vector Machine and Neural Network are used to identify the performance of the proposed works.

1.2 Methodology

Figure 1 shows the overall flow of the research work using three standard privacy policies. The pharmaceutical data maintained by the medical organizations are more critical which has to be secured from the malicious users. The first approach generates the transactional graph which is used to identify correlation between the sensitive items and non-sensitive items. In the second approach associated rule mining is applied to identify the interesting patterns. The sensitive rules hiding is achieved by reducing the support below a minimum threshold or adding noise to the original confidence of the sensitive rules. In the last work, the suffix hidden crypto policy is used for securing the sensitive data by linearly verifiable method.

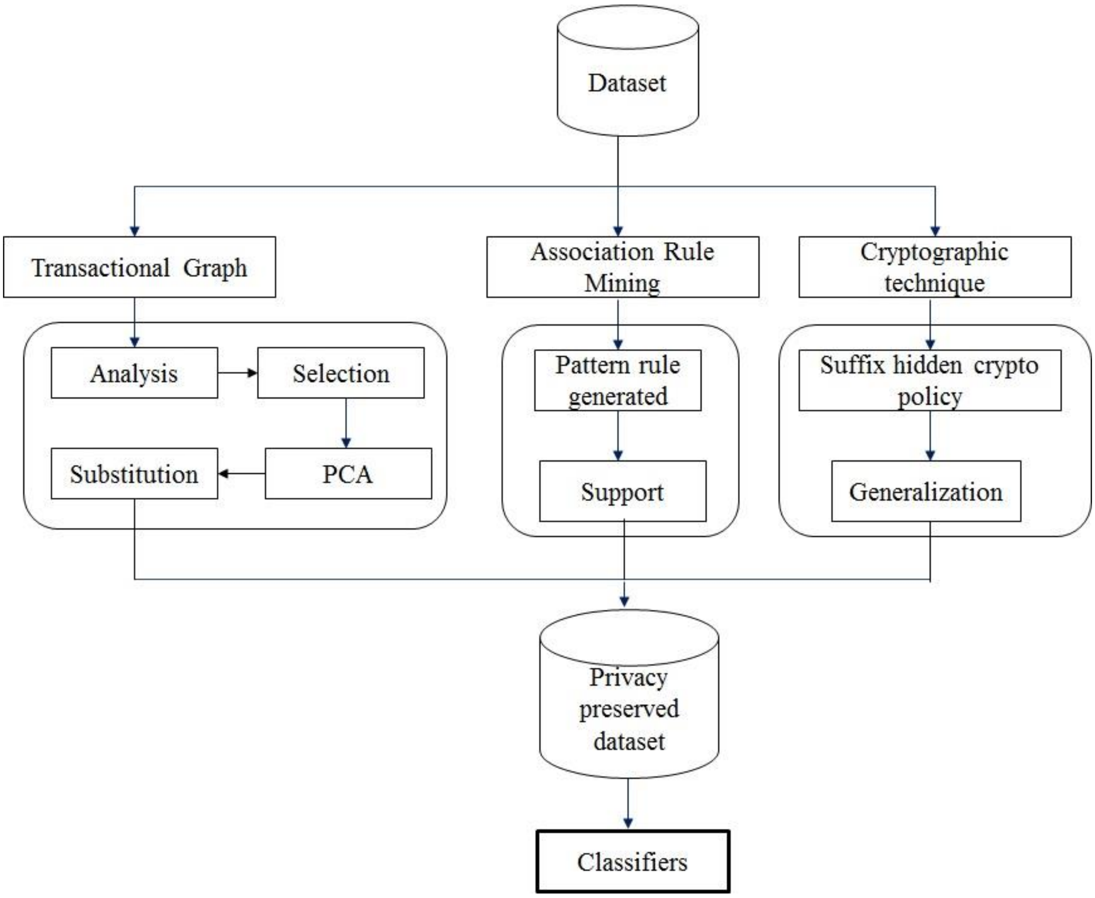

Figure 1 Block diagram for proposed system

1.3 Datasets

The experiments are analyzed on real life datasets which are obtained from pharmacopoeia. There are 10000 transactional records with 100 attributes with no missing values.

1.4 Classifiers

The performance of the proposed works are analyzed with the original dataset by the classifier models Decision Tree, Random Forest, Linear model, Ada Boost, Support Vector Machine and Neural Network.

1.4 Performance Measures

The following measures are considered in the classification process.

Accuracy

Accuracy is measured according to number of every correct prediction by total number of records present in the data set.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

Where

TP – True Positive

TN - True Negative

FP – False Positive

FN – False Negative

Sensitivity

Sensitivity is defined as the proportion of the number of correct positive prediction data by the total number of positive data.

$$Sensitivity = \frac{TP}{TP+FN} \tag{2}$$

Specificity

Specificity is defined as the proportion of the number of correct negative predictions by the total number of negatives.

$$Specificity = \frac{TN}{TN+FP} \qquad (3)$$

1.5 Research Contributions

In this research, the main contribution is privacy preservation in pharmaceutical data by the implementation of the following techniques.

1 Correlation product probability matrix using transactional graph with PCA

2 Feature influents measure applied in interesting association rules

3 Multi attribute provable portioning crypto hidden analysis for transactional data publication

2. CORRELATION PRODUCT PROBABILITY MATRIX USING TRANSACTIONAL GRAPH WITH PCA

The transactional graph is a multi-dimensional graph, in which every node represents the distinct attribute of transactional drugs purchased by various patients. The attributes and values present in the records are used to generate the transactional graph that provides link between two drugs. Similarly, the numbers of links are established between every interconnected drugs and each drug consist of two values such as convergence and divergence. The convergence represents number of links that are received from other drugs and the divergence represents number of outgoing links.

$$\text{Convergence} = \sum_{i=1}^{n} \frac{\text{NR(i)}<-NA}{\text{n}} \qquad (4)$$

$$\text{Divergence} = \sum_{i=1}^{n} \frac{\text{NR(i)}\rightarrow\text{NA}}{\text{n}} \qquad (5)$$

Whereas n- total number of transactional record

NR - number of relation exist in the transaction

NA - Distinct attributes of transaction.

Based on the occurrences in the transaction data set with the other drugs, the divergence and convergence measures are computed. The method estimates the sensitive weight for each item in the dataset. As per the sensitive weight, the product probability matrix (PPM) value is computed.

$$\mathrm{PPM} = \frac{\mathrm{Convergence(i)}*\mathrm{Divergence(j)}}{\mathrm{Divegence(i)}*\mathrm{Convergence(j)}} \tag{6}$$

From the PPM, the method identifies the similarities between sensitive and non-sensitive drugs. If the particular non-sensitive drug is highly correlated with sensitive drug, then principal component analysis is applied on those drugs. The principal component values replace the original values of a data set with a smaller number of uncorrelated values.

3. FEATURE INFLUENTS MEASURE APPLIED IN INTERESTING ASSOCIATION RULES

In this approach, similarity coefficients are the important measuring techniques used to quantify the association extent in which, the data connected with each other. Initially the transaction dataset is converted as binary transaction dataset, which represents the presence or absence of an attribute in the transactional dataset, the value 1 represents the presence of an attribute while 0 represents the absence of an attribute. Then, the list of items present in overall data set is identified. Further, for each item identified from the item set, the method generates number of patterns accordingly. With the pattern set generated, the support and confidence values are computed. Based on the support and confidence values, the method selects a least support value and the set of items that are identified as sensitive. For each sensitive items identified, the influence measure of the other items are estimated. Similarly the estimated values are added to the feature influence matrix (FIM).

The feature influence matrix is generated according to the pattern set generated. Initially the occurrence of each pattern is computed and based on that, the occurrence of item is counted in different pattern sets. For each pattern and pattern set the occurrence of item, the influence measure is estimated for each item. Similarly the estimated influence measure is added to the feature influence matrix.

The data perturbation is performed according to the feature influence matrix to sanitize the privacy information. In case the third party attempts to read the influence values of any features, the method modifies the sensitive data values by adding additive noise to the particular pattern, which is drawn from a probability distribution to the sensitive data values. However, they cannot identify the original data of any patients; still they can analyze the details to acquire the information from the sanitized data.

4. MULTI ATTRIBUTE HIDDEN CRYPTO ANALYSIS FOR TRANSACTIONAL DATA PUBLICATION

A privacy concern of information mining analysis from huge data analysis is a challenging task because it supports the secure functioning of medical organization. Using the data security, a Multi Attribute Based Provable Partition Crypto analysis technique (MAPPC) scheme is presented.

The considered data set represents the details of various patients and appropriate drug prescribed to them. It also contains the records of both general and paramedical patients. The general patient’s records consist of no sensitive information, whereas the critical patients consist of sensitive drug information. In order to maintain the privacy of the patients, the methods classify the attributes into different classes as non-sensitive, complete sensitive, partial sensitive. According to the hidden crypto policy, the method identifies the attributes and generates hidden value. For a complete sensitive attribute, the method encrypts the data entirely, whereas for the partial one, the method selects a number which represents the size of the attribute to be encrypted. The encrypted data is provided as sanitized data set.

Linear integrity represents the access rights of the user on each attribute where the attributes are sensitive. According to the user rights, the verification and validation are performed in order to acknowledge the appropriate user. A new conspire affirmed secure data sensitivity approach is proposed which increases the security much better.

5. CONCLUSION

In this work, the medical prescription transaction dataset is used. Initially, correlation product probability matrix using transactional graph with PCA approach is applied to improve the privacy preservation of sensitive data in the dataset. The method generates the transactional graph according to the number of occurrence, number of relations, convergence and divergence measures. Based on the value of convergence and divergence measures, the method estimates the weight for each item in the items set. The high correlation between the sensitive and non-sensitive items are identified. The sensitive item values are replaced with correlated non-sensitive item value with additive noise in the transaction dataset. The PCA is applied to the correlated sensitive and non-sensitive items. This method enables to attain the higher accuracy in privacy preservation.

Second, the Feature influents measure applied in interesting association rules based approach is presented. It generates possible combination of purchase patterns in the dataset. Using the patterns, for each item, towards each other item, the method estimates the influence measure according to the number of occurrence of pattern, the number of occurrence of the candidate item in other items. Using these two, the method measures the influence of items with other items. Generated influence values are estimated and it modifies the sensitive data, which is drawn from the probability distribution.

Finally, Multi attribute hidden crypto analysis based privacy preservation technique is proposed. The partitioned attributes are used to generate the case reasoning. According to the policy, the method performs suffix encoding. The proposed methods improve the performance of privacy preservation. This is measured by the classifier models with the metrics accuracy, sensitivity and specificity.

7. ORGANIZATION OF THE THESIS

The thesis on privacy preservation on pharmaceutical data set is organized as follows:

The chapter 1 presents the detailed introduction on the privacy preservation and sanitization techniques, problem statement and the proposed research methodologies.

The chapter 2 discusses the detailed review on the early methods available for the problem for sanitization of transactional and medical data sets.

The chapter 3 gives the implementation of correlation product probability matrix using transactional graph with PCA

The chapter 4 proposes the implementation of feature influents measure applied in interesting association rules

The chapter 5 discusses the implementation of multi attribute hidden crypto analysis for transactional data publication.

The chapter 6 discusses the conclusion of the research and future work to carry out.

REFERENCES

1. Asma Alnemari, Carol J. Romanowski & Rajendra K Raj 2017, ‘An Adaptive Differential Privacy Algorithm for Range Queries over Healthcare Data Sign In or Purchase’, Healthcare Informatics (ICHI).

2. Barua, M, Lu, R & Shen, X 2013, ‘SPS: Secure personal health information sharing with patient-centric access control in cloud computing’, in Proc. of GLOBECOM, pp. 647–652.

3. Chaudhuri, K & Monteleoni, C 2008, ‘Privacy-preserving logistic regression’, In Advances in Neural Information Processing Systems 21 (NIPS 2008), pp. 289-296.

4. Fung, CM, Wang, K, Chen, R & Yu, PS 2010, ‘Privacy-preserving data publishing: A survey of recent developments’. ACM Computing Survey, vol. 42.

5. Gabriel, G 2011, ‘Anonymous Publication of Sensitive Transactional Data’, IEEE Transactions on knowledge and data engineering, vol. 23, no. 2.

6. Jiang, S, Zhu, X & Wang, L 2015, ‘Epps: Efficient and privacy-preserving personal health information sharing in mobile healthcare social networks’, Sensors, vol. 15, no. 9, pp. 22 419–38.

7. Kwangsoo Seol & Young-Gab Kim 2018, ‘Privacy-Preserving Attribute-Based Access Control Model for XML-Based Electronic Health Record System’, IEEE Trans secure computing, vol. 6,pp. 9114 – 9128.

8. Lin, X, Lu, R, Shen, X, Nemoto, Y & Kato, N 2009, ‘SAGE: a strong privacy-preserving scheme against global eavesdropping for health systems’, IEEE Journal on Selected Areas in Communications, vol. 27, no. 4, pp. 365–378.

9. Machanavajjhala D Kifer, J Gehrke & Venkitasubramaniam, M 2007, ‘Diversity: Privacy beyond k-anonymity’, ACM Transactions on Knowledge Discovery from Data (TKDD), vol. 1, pp. 3.

10. Zhang, R, Liu, L & Xue, R 2014, ‘Role-based and time-bound access and management of ehr data’, Security and Communication Networks, vol. 7, no. 6, pp. 994–1015.

LIST OF PUBLICATIONS

International Journal

1. **Saranya, K** & Premalatha, K 2019,‘Privacy-preserving data publishing based on sanitized probability matrix using transactional graph for improving the security in medical environment’. Journal of Supercomputing, DOI:10.1007/s11227-019-03102-2. **(Impact Factor – 2.168)**