RESEARCH ARTICLE

Editorial Process: Submission:07/05/2018 Acceptance:05/18/2019

Protecting the Privacy of Cancer Patients Using Fuzzy Association Rule Hiding

Sathiyapriya Krishnamoorthy*, Kaviya Murugesan

Abstract

Objective: Privacy protection in the medical field means the protection of individuals from being associated with undesirable conditions, diagnoses or treatments (Sensitive Attributes). The problem of knowledge discovery from health care data by applying data mining algorithms is inversely related to the privacy of individuals. Due to the tremendous growth of data in a large scale, there is a demand to protect the sensitive data accessible from medical datasets. **Methods:** This paper considers the problem of building privacy preserving association rule mining algorithm using the notion of TF * IDF derived from the information retrieval domain. The highly sensitive transaction is chosen using the product of Relative Item Frequency and Condensed Frequency. Finally, sensitive fuzzy data is perturbed to hide these refined rules. **Results:** It has been found that the number of non-sensitive rules lost as a side effect of hiding sensitive rule is 20% less and number of ghost rules is 30% less in proposed work than in previous work using Transactional Impact factor method. The execution time of hiding a rule is 26% lesser on an average in the proposed technique for various values of minimum confidence threshold. It has been observed that the number of modifications to the original dataset after hiding three rules were reduced by 66% in proposed method than in previous work. As the number of modifications to original data is less the chances of generating false association is also reduced. **Conclusion:** In this paper, a novel method was presented to hide the sensitive rule in quantitative data by decreasing the support of the RHS of the rule. Experimental results demonstrate that the proposed approach is more efficient as it facilitates better rule hiding and minimizes the number of lost rules and ghost rules. Also, this approach makes minimum modifications to the dataset.

Keywords: Privacy preservation- fuzzification- sensitive rules- rule hiding- fuzzy association rules

Asian Pac J Cancer Prev, **20 (5)**, 1437-1443

Introduction

Even though cancer research has traditionally been clinical and biological in nature, in recent years data driven analytic studies have become a common complement. A scientist alone is unlikely to discover the precise combination of variables that makes a drug work or not. But with enough data, a data mining algorithm could find a predictive correlation. Using data mining techniques, researchers can discover previously unknown associations between patient characteristics, symptoms, and medical conditions from health information datasets. And these associations enable formulation of new treatments leading to customized medical decisions. The algorithms need to be validated to ensure high-quality medicine, which require access to the same extensive medical data from which the conclusions were drawn. Medical data is one of the most private information and data mining techniques requires access to a lot of that information. These techniques also create new information like predictions, associations based on the models developed from the dataset. This information may impair individual privacy when used in marketing, sales to others, or discrimination in employment, insurance, or other decisions. Even when it is not used in these ways, its collection, disclosure, and use can infringe individual privacy. Privacy protection in the medical field means the protection of individuals from being associated with undesirable conditions, diagnoses or treatments (Sensitive Attributes).

These two problems of knowledge discovery and privacy preservation in applying data mining algorithms to healthcare data are inversely related as efforts to improve one will usually make the other worse. In order to use the conclusions/models obtained from applying data mining techniques they should be validated. This requires access to more information, which can aggravate the privacy problem. And the solution to the privacy problem is to limit the amount of information researchers, companies, and the government can use and to which they have access, but that can make it harder to validate models.

There are three approaches to an effective mining system that values patient privacy. The first approach applies restrictions on the collection, use, and distribution of medical data, so that data gathered is not also used for

*Department of Computer Science and Engineering, PSG College of Technology, Coimbatore, Tamilnadu, India. *For Correspondence: spk.cse@psgtech.ac.in*

illegitimate purposes. The second approach is to modify the patient data slightly before distributing it, so that the researchers can verify the data models and at the same time the data cannot be linked to particular patient. And the third approach is to develop robust information-security system, so that unintended outsiders cannot obtain, use, or disseminate patient data. This research employs second approach and proposes an algorithm which continues to be effective, without compromising the security.

This paper proposes an algorithm for protecting the privacy of cancer patients when association rule mining is applied on cancer data set. Association rule mining is a data mining technique that identifies associations between the attributes in a dataset. The associations generated may harm the privacy of cancer patients. Even when the data set is released after de-identifying the identity information of the patients it is possible to link the released information with other datasets and identify the patients. So the patient data slightly altered before dispensing it, so that the researchers can verify the data models and at the same time the data cannot be linked to particular patient.

Tehoretical Background

Let I= $\{i_1,...i_N\}$ be distinct literals called items. Given a database D= $\{T_1,...T_m\}$ is a set of transaction where each transaction T is a set of items as Ti ⊂ I ($1 \leq i \leq m$). An association rule is defined as X→Y, where Y ⊂ I, X ⊂ I and Y∩X= φ. X is called rule's antecedent - LHS (left hand side) and Y is called rule's consequent - RHS (right hand side). These association rules are used to discover interesting relations between items in large databases. Further strong association rules are derived using some measures of interestingness. In order to select such interesting rules from the set of all possible rules, constraints on various measures of significance and interest are used. The best-known constraints are minimum thresholds on support and confidence.

The support of rule X→Y is calculated as:

$$\text{support}(X \Longrightarrow Y) = \frac{|x \cup y|}{|D|} \quad (1)$$

where |D| define the total number of the transactions in the database D and |X ∪ Y| is the number of transactions which support item set XY. The confidence of the rule X→Y is calculated as

$$\text{confidence}(X \Longrightarrow Y) = \frac{|x \cup y|}{|x|} \quad (2)$$

where |X| is number of transactions which support item set X. A rule X→Y is mined from database if support(X→Y) ≥ MST (minimum support threshold) and confidence(X→Y) ≥MCT (minimum confidence threshold). Association rule is considered as sensitive when its confidence is above the disclosure threshold. Such rules violate the privacy of the data owners. Numerous techniques have been used to hide sensitive association rules by performing some modifications in the original dataset.

Approach for hiding association rules is almost limited to binary dataset. But, real world data mostly consists of quantitative values. The proposed system uses a method to hide sensitive association rules, in which the fuzzified database is mined using modified apriori algorithm to extract the association rules and QRIF – QCCF (Quantitative Refined Item Frequency-Quantitative Converse Catalog Frequency) algorithm is applied to hide the user preferred sensitive rules. Hence the sensitive rules are hidden by reducing confidence of these user specified sensitive rules.

Materials and Methods

To obtain meaningful rules from quantitative data, the original data is mapped into fuzzy values and fuzzy association rules are mined using algorithm proposed by Hong et al, 1999. Then the proposed QRIF- QCCF rule hiding algorithm is applied to hide the confidential rules obtained from the dataset. So, the proposed system involves two major phases. Initially the fuzzification phase and the sensitive rule hiding phase.

Fuzzification Phase

Fuzzification is the process of converting the quantitative values in the dataset into fuzzified values. Fuzzy Transaction Data-mining Algorithm (FTDA) has been used for generating the association rules (Hong et al., 1999).

The triangular membership function given in equation (3) is applied to convert the crisp values of original dataset into fuzzy values. This membership function for an attribute consists of three points, the central vertex point, b, and the two endpoints, a and c of a triangle as shown in Figure1. Hence each value x, in the dataset will be mapped across three fuzzy regions such as high, medium and low.

$$f(x;a,b,c) = \max\left(\min\left(\frac{x-a}{b-a}, \frac{c-x}{c-b}\right), 0\right) \quad (3)$$

Then the total sum of all low, medium and high fuzzy values of an item is computed. The user specifies the support threshold and confidence threshold of the association rules. Then the fuzzy regions with the total sum value larger than the support threshold is taken for further processing. Those that satisfy the minimum support threshold (MST) is considered as frequent one itemset. Then the two frequent item sets are formed by combining these frequent one itemsets. Then the two itemsets are formed by finding the minimum value across each and every transaction and then the computation of sum of these values are obtained. If this sum satisfies the threshold, it can be further considered for mining the rules. All possible combination of rules is obtained and the confidence of the rule is calculated using equation (2). If the confidence of the obtained rules satisfies the minimum confidence threshold (MCT), they are considered as the interesting rules. Among these interesting rules, the owner of the dataset specify the sensitive rules that should not be disclosed.

Sensitive Rule Hiding Phase

The QCF signifies the product of Relative Item Frequency and Condensed Frequency. The Relative Item

Frequency denotes the presence of all the sensitive items with fuzzy value greater than 0.5 in a single transaction. For each sensitive item in a sensitive rule, the Condensed Frequency represents the presence of an item with fuzzy value greater than 0.5 in all the transactions. This algorithm has been derived from the concept of TF (Term Frequency) - IDF (Inverse Document Frequency). The Quantitative Refined Item Frequency (QRIF) is applied to each and every transaction in the dataset. The Quantitative Refined Item Frequency (QCCF) is applied only to the items in the sensitive rules. This algorithm can be applied for fuzzified values obtained from the fuzzification phase.

Figure 2 shows the steps in QCF algorithm. The algorithmic steps of Sensitive Rule Hiding are as follows

Proposed Algorithm:

Quantitative Condensed Frequency (QCF)

Input: Original Database, D

Output : Modified Database, M

Step 1: Read the original Dataset D.

Step 2: Fuzzification : Convert D → F , Fuzzified Database

Step 3: Quantitative Refined Item Frequency (QRIF) is calculated as in (5):

$$QRIF = \frac{(\#\ C_{i<\mu\geq 0.5>})}{(\#\ T_{i<\mu\geq 0.5>})} \quad (5)$$

Ci - Count of number of features that has membership value $\mu \geq 0.5$ present in sensitive rules

Ti - Count of number of features having $\mu \geq 0.5$ in a single transaction

μ - Fuzzy membership value of a feature

Step 4: Quantitative Converse Catalogue Frequency (QCCF) is calculated as in (6)

i. Computation of Condensed Frequency (CF)

$$CF_i = \frac{\sigma + 1}{(\#\ C_{i<\mu\geq 0.5>})} \quad (6)$$

$$\sigma_i = \frac{(\#C_{i<\mu\geq 0.5>})}{\eta} \quad (7)$$

σ - Support of the sensitive item

η - Total number of transactions

Maximum Condensed Frequency (MCF)is calculated as in (8):

$$MCF = \max[CFi\ ] \quad (8)$$

QCCF is calculated as in (9):

$$QCCF = log\left(\frac{|\ \eta\ |}{|\ C_i - MCF\ |\ * \sigma_i}\right) \quad (9)$$

η - Number of transactions in the catalog

σi - Number of transactions that have sensitive item with$\mu \geq 0.5$

Step 5: Quantitative Condensed Frequency (QCF) ie,(QRIF * QCCF) is computed as in (10):

$$QRIF * QCCF = \frac{(\#\ C_{i<\mu\geq 0.5>})}{(\#\ T_{i<\mu\geq 0.5>})} * \log\left(\frac{|\eta|}{|C_i - MCF| * \sigma_i}\right) \quad (10)$$

Sort the transactions in descending order of QRIF* QCCF.

Step 6: Read the highly sensitive rule from the user.

Choose the feature that is in the right-hand side of the sensitive rule.

Choose the first transaction of the sensitive feature from the sorted dataset.

Step 7: Update the μ of the that sensitive feature in that particular transaction if it is greater than 0.5

$$\mu_c = 1 - \mu \quad (11)$$

μ_c- Change in membership value.

Step 8: Repeat the steps (3) to (7) till all the sensitive rules are hidden in the dataset and release the Modified database, M.

Defuzzification Phase

Defuzzification phase is the process of converting the fuzzy values into crisp values. In order to achieve this, max membership defuzzification method is used. After this process, the entire dataset is released to the outside parties. This method preserves the privacy and quality of the database.

Results

The proposed approach was tested using the Wisconsin Breast cancer Dataset obtained from the UCI Machine Learning Repository. The dataset consists of 10 attributes and 699 instances. The attributes are Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, Mitoses and Class. This QRIF * QCCF Aggregation algorithm was implemented for the nine quantitative attributes which are mapped to three fuzzy sets and the class attribute was ignored. The algorithm hides one sensitive rule at a time. The proposed fuzzy association rule hiding methods are evaluated using the following measures and compared with our previous work that uses Transactional Impact Factor (TIF)(Sathiyapriya and Sadasivam, 2016).

Number of Lost Rules

A lost rule is a non-sensitive association rule that can be discovered from the original database but cannot be mined from the released database. $\{lostRuleq \mid q \in R_n \cap q \notin R/\}$. R_n is the set of non sensitive rules and $R/$ is the set of rules mined from released dataset.

Number of Ghost Rules

A ghost rule is a non-sensitive association rule that cannot be discovered from the original database but can be mined from the released database. $\{GhostRuleq \mid q \in R/ \cap q \notin R\}$. R is the set of rules mined from original dataset.

Hiding Failure

A false rule is a sensitive association rule that cannot be hidden by hiding process. Number of false rules denotes

the hiding failure. $\{FalseRuleq \mid q \in R_h \cap q \in R'\}$. R_h is the set of sensitive rules to be hidden.

Percentage of modification

Percentage of modification refers to the distortions to the total data items in the original database expressed in percentage.

$$\text{Percentage of modification} = \frac{D'}{D} \times 100$$

Execution time

Execution time also called "running time" is the length of time required to implement hiding algorithm.

Figure 3 shows the number of lost rules as a side effect of hiding a single rule for varying values of confidence with a constant minimum support of 30 when quantitative data is fuzzified using triangular membership function values supplied by the experts. It has been found that the number of non-sensitive rules lost as a side effect of hiding sensitive rule is 20% less in proposed work than in the previous TIF work.

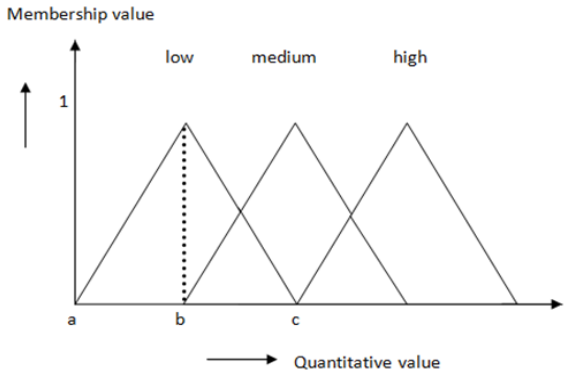

Figure 1. Fuzzification - Triangular Membership

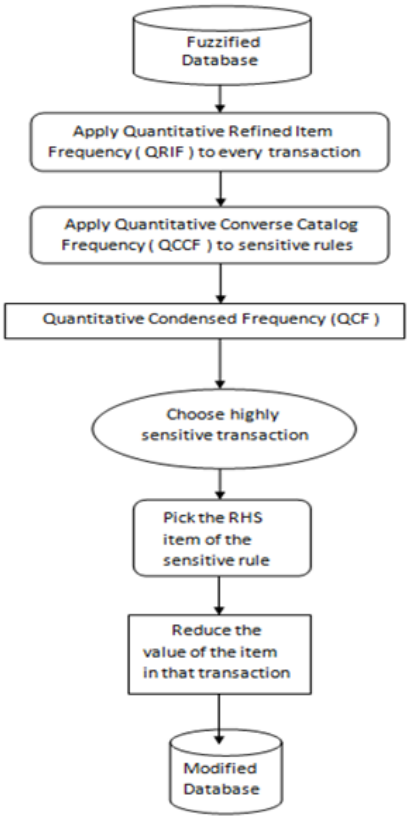

Figure 2. Steps in QCF Algorithm

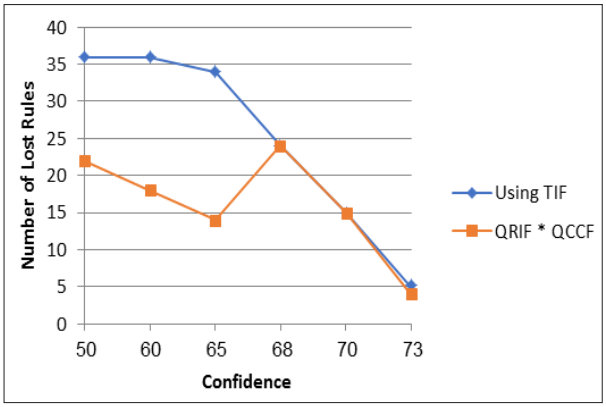

Figure 3. Rules Lost after Hiding a Rule

Figure 4 shows the number of ghost rules for varying values of confidence and a constant minimum support of 30. The number of ghost rules generated as a side effect in proposed QCF method is 30% less than in our previous work using TIF. The QCF algorithm does not show any hiding failure as the algorithm hides one rule at a time.

Figure 5 shows the execution time of hiding a rule is 26% lesser in the proposed technique than in earlier work for various values of minimum confidence threshold.

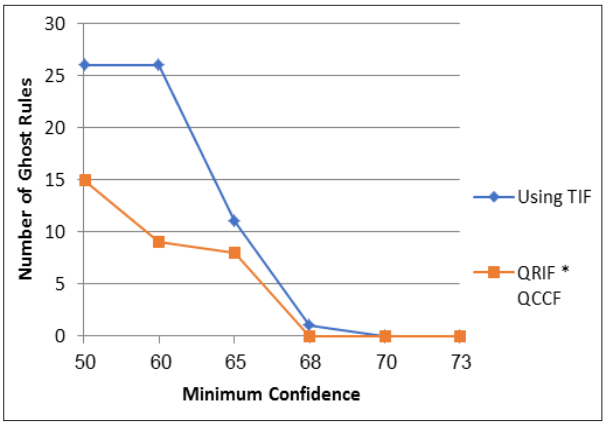

Figure 4. Ghost Rules after Hiding a Rule

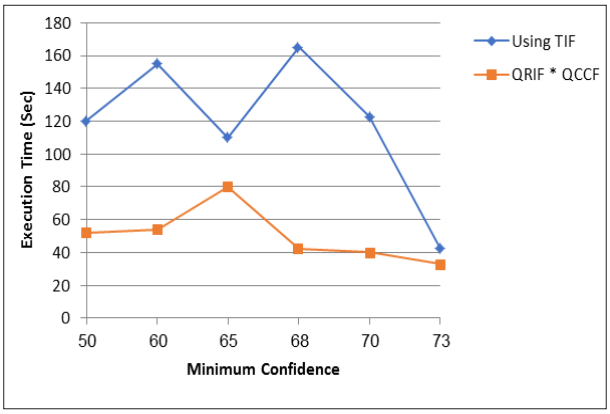

Figure 5. Comparison of Execution Time

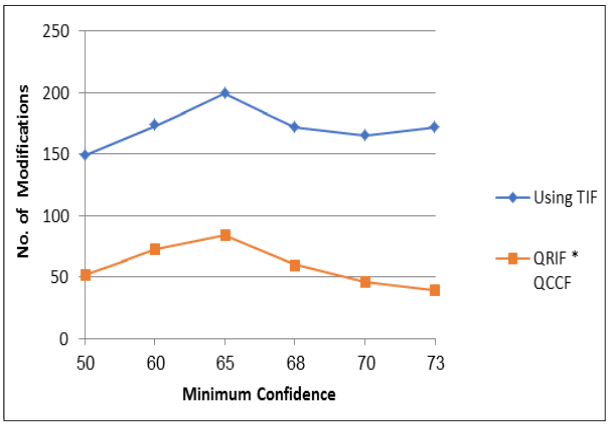

Figure 6. Number of Modifications in the Dataset after Hiding

Figure 6 shows the number of modifications to the dataset after hiding a single rule. It has been observed that modifications made in the original dataset to hide the sensitive rules were reduced by 66% in proposed method than in previous work. As the number of modifications to original data is less the chances of generating false association is also reduced. Consequently the number of lost rules and ghost rules are reduced.

Discussion

The approach for hiding sensitive association rules is broadly classified into major categories as Distortion based technique and Blocking based technique. In distortion based technique, the values of the attribute in the dataset are modified in certain transactions from 1 to 0 or vice versa which reduces the support and confidence of association rules while blocking based technique replaces the value of the sensitive attributes in some transactions with '?'. If the confidence of sensitive association rule is reduced below the threshold confidence then these association rules become uninteresting and were not analyzed. But this technique incurs side effects of 'Ghost Rules' and 'Lost Rules'.

The two common data mining algorithms for hiding sensitive items in association rules were formulated (Wang and Jafari, 2005). The first, ISLF (Increase Support of LHS (Left Hand Side) First) algorithm increases the support of left hand side of the rule. If it does not satisfy the confidence threshold then it decreases the support of the item on the rule's right hand side. In DSRF (Decrease Support of RHS First) algorithm, the confidence of the rule is reduced by replacing an item with value 1 or 0 by unknown (?) within a transaction. These algorithms are later coined as ISL and DSR. It has been proven that, these methods required minimal number of database scans and pruned a number of hidden rules. But this algorithm leads to lot of side effects.

A hiding technique based on reconstruction, which is FP-tree based method for inverse frequent set mining was proposed (Guo, 2007). This approach involved three main phases for hiding the sensitive data. In the first phase, it used frequent item set mining algorithm to generate all frequent item sets along with the support. The second phase is application of sanitization algorithm over the frequent set, which selects the appropriate hiding strategy and identifies the sensitive frequent item sets based on the sensitive rules. In the third phase, the reconstruction-based approach is used for inverse frequent set mining. This Frequent Pattern tree is drawn based on the frequency of the occurrence of the individual items in the transaction. Thus, this approach released the sanitized database with the set of non-sensitive rules to be retrieved exactly and the side effects such as lost rules and ghost rules has been reduced to some extent. The main advantage is that, reconstruction-based approach mainly deals with only frequent item set initially followed by the infrequent items. This would significantly reduce the search space, while the other approaches consider frequent and infrequent items as a whole, which makes the search space very large.

In order to mine rules from quantitative data the dataset is fuzzified. In the fuzzified dataset, the transactions that support the sensitive item in the left-hand side of the rule are perturbed. Some transactions are removed from the dataset if the confidence does not go below the threshold. In this way maximum number of sensitive rules could be made hidden by making minimum number of modifications (Berberoglu and Kaya, 2008).The SIF-IDF (Sensitive Item Frequency- Inverse Database Frequency) which is a greedy approach based on the concept of TF-IDF (Term Frequency - Inverse Document Frequency) taken from Information Retrieval domain is applied for hiding sensitive rules (Hong et al., 2011).This algorithm has been limited to binary dataset. With this algorithm sensitive association rules are efficiently hidden before disclosing the database to an external party. This approach is better and efficient when compared to Reconstruction and Cryptography based approach.

Decreasing the Support value of the Right-hand side (DSR) of the rule approach for quantitative data to hide the sensitive rules was proposed (Sathiyapriya et al., 2011). This approach first fuzzifies the quantitative values for further processing. An algorithm has been designed to derive the membership function automatically for fuzzification of the dataset. Then to generate the rules from fuzzy data, modified Apriori algorithm is used. By using DSR approach, the confidence of the restricted rules is reduced below the threshold. This approach is able to hide most of the confidential rules efficiently with the minimal side effects.

Another algorithm called BHPSP was proposed using Impact Factor (Bonam and Reddy, 2014). This algorithm calculates the impact factor of items in the sensitive association rules. Then it selects a rule which contains an item with minimum impact factor. The quality of a database can be well maintained by greedily selecting the alterations in the database with negligible side effects.

A more efficient algorithm for High Average Utility Item mining, includes three pruning strategies to provide a tighter upper bound on the average-utilities of itemsets (Lin et al., 2017). This reduces the search space more effectively to decrease the runtime. The first pruning strategy utilizes relationships between item pairs to reduce the search space for itemsets containing three or more items. The second pruning strategy provides a tighter

upper bound on the average-utilities of itemsets to prune unpromising candidates early. The third strategy reduces the time for constructing the average-utility-list structures for itemsets, which is used to calculate their upper bounds.

A novel heuristic technique ILARH (Intersection Lattice-based Association Rule Hiding) is based on the intersection lattice of frequent itemsets was proposed (Le and Arch-int, 2012). This approach helps in identifying the victim items easily and these items alone can be modified to reduce the confidence of the sensitive rules and it would have a less impact on other itemsets. This minimized the side effects to some degree.

The algorithm MDSRRC (Modified Decrease the Support of Right hand side of the rule clusters), is an extension of DSRR (Decrease the support of right hand side of the rule) (Domadiya and Rao, 2013). But this MDSRRC is suited for association rules that have multiple items in the antecedent and consequent. The sensitivity of the item is the number of sensitive rules that contain the sensitive item and the sensitivity of the transaction is the total number of sensitive items present in the transaction. In the sensitive item set, sensitive items are ordered based on their occurrence in descending order. The transactions are sorted with the help of the sensitive item set. Then the first transaction that supports the first sensitive item in the set is deleted from the transaction. Then the sensitivity of the item and transaction are updated and process continues till confidence becomes less than minimum threshold confidence. Thus, this approach modifies the database and hence maintains the data quality and efficiency.

Many privacy preserving algorithms were proposed for fields like Stream Data mining, Cloud and Big Data Mining. A framework was formed that categorizes protection approaches and encourages interdisciplinary solutions to the growing variety of privacy problems associated with knowledge discovery from data (Xu et al., 2017). The Map Reduce framework was agglomerated with adopted heuristics to overcome this challenge of scalability along with much-needed privacy preservation yielding efficient analytic results within bounded execution times (Sharma and Toshniwal, 2017). Different solutions are provided to maintain data privacy during association rule mining from the data stored in cloud (Yi et al., 2015). The solutions are built on the distributed cryptosystem and achieve item privacy, transaction privacy and database privacy, respectively, as long as at least one out of the n servers is honest.

An algorithm named Diverse and k-Anonymized HOeffding Tree (DAHOT) that is an amalgamation of popular data stream classification algorithm Hoeffding tree and a variant of k-anonymity and l-diversity principles was proposed (Kotecha and Garg, 2017).

To date, tract identifiers about geographic information have been left off the research files because they could compromise the confidentiality of patients' identities. Yu et al., (2017) present an approach to inclusion of tract identifiers based on multiply imputed, synthetic data. The idea is to build a predictive model of tract locations, given patient and tumor characteristics, and randomly simulate the tract of each patient by sampling from this model. For the predictive model, multivariate regression trees fitted to the latitude and longitude of the population centroid of each tract was used.

Acknowledgements

Our sincere thanks to the Management of PSG College of Technology and Dr R Rudramoorthy, Principal, PSG College of Technology for providing a wonderful platform to complete this research work. This project was carried out in Big Data Analysis lab, PSG College of Technology, India.

References

Berberoglu T, Kaya M (2008). Hiding fuzzy association rules in quantitative data, The 3rd International Conference on Grid and Pervasive Computing Workshops, pp 387- 92.

Bonam J, Reddy R (2014). Balanced approach for hiding sensitive association rules in data sharing environment. *Int J Inf Secur Priv (IJISP)*, **8**, 39–62.

Domadiya NH, Rao UP (2013). Hiding sensitive association rules to maintain privacy and data quality in database. 3rd IEEE International Advance Computing Conference (IACC).

George VSV, Moustakides V (2009). A maxmin approach for hiding frequent itemsets. *Data Knowl Eng*, **65**, 75–89.

Guo Y (2007). Reconstruction-based association rule hiding. In Proc. Of SIGMOD2007 Ph.D. Workshop on Innovative Database Research 2007 (IDAR 2007).

Hong TP, Kuo CS, Chi SC (1999). Mining association rules from quantitative data. *J Intell Dat Anal*, **3**, 363-76.

Hong TP, Lin TW, Chang CC, Wang SL (2011). Hiding sensitive itemsets by inserting dummy transactions, In Proceedings of the IEEE International Conference on Granular Computing, pp 246–9.

Hong TP, Lin CW, Yang KT, Wang SL (2013). Using TF-IDF to hide sensitive itemsets. *Appl Intell*, **38**, 502–10.

https://archive.ics.uci.edu/ml/datasets/BreastCancerWisconsinOriginal.

Kotecha R, Garg S (2017). Preserving output-privacy in data stream classification. *Prog Artif Intell*, **6**, 1–18.

Le HQ, Arch-int S (2012). A conceptual framework for privacy preserving of association rule mining in e-commerce, 7th IEEE Conference on Industrial Electronics and Applications (ICIEA).

Lin JC-W, Liu Q, Fournier-Viger P, et al (2016). A sanitization approach for hiding sensitive itemsets based on particle swarm optimization. *Eng Appl Artif Intell*, **53**, 1–18.

Lin JCW, Ren S, Fournier-Viger P, et al (2017). A fast algorithm for mining high average-utility itemsets. *Appl Intell*, **47**, 331–46.

Pathak K, Chaudhari NS, Tiwari A (2011). Privacy preserving association rule mining by introducing concept of impact factor, 7th IEEE Conference on Industrial Electronics and Applications (ICIEA).

Sathiyapriya K, Sudhasadasivam G, Celin N (2011). A new method for preserving privacy in quantitative association rules ysing DSR approach with automated generation of membership function", In the Proceedings of World Congress on Information and Communication Technologies, pp 148-53.

Sathiyapriya K, Sudha Sadasivam G (2016). An evolutionary approach using transactional impact factor for preserving privacy of quantitative data. *Indian J Sci Technol*, **9**, 1-9.

Sharma S, Toshniwal D (2017). Scalable two-phase co-occurring sensitive pattern hiding using mapreduce. *J Big Data*, **4**, 4.

Wang SL, Jafari A(2005). Using unknowns for hiding sensitive

predictive association rules. In IEEE International Conference on Information Reuse and Integration, pp 223 – 8.

Wang SL, Lee YH, Billis S, Jafari A (2004). Hiding sensitive items in privacy preserving association rule mining. IEEE International Conference on Systems, Man and Cybernetics.

Xu L, Jiang C, Chen Y, Wang J, Ren Y (2016). A framework for categorizing and applying privacy-preservation techniques in big data mining. *Computer*, **49**, 54–62.

Yi X, Rao F-Y, Bertino E, Bouguettaya A (2015), Privacy-preserving association rule mining in cloud computing. In: Proceedings of the 10th ACM Symposium on Information, Computer and Communications Security (ASIA CCS '15). ACM, New York.

Yu M, Reiter JP, Zhu L, et al (2017). Protecting confidentiality in cancer registry data with geographic identifiers. *Am J Epidemiol*, **186**, 83–91.