International Journal of Latest Trends in Engineering and Technology
Vol.(12)Issue(4), pp.060-066
DOI: http://dx.doi.org/10.21172/1.124.12
e-ISSN:2278-621X

REVIEW ON PRIVACY ISSUES IN BIG DATA USING DATA MINING TECHNIQUES

Urvashi Sangwan[1]

Abstract: The growing popularity and development of data mining technologies bring serious threat to the security of individual's sensitive information. An emerging research topic in data mining known as privacy-preserving data mining (PPDM) has been extensively studied in recent years. The basic idea of PPDM is to modify the data in such a way so as to perform data mining algorithms effectively without compromising the security of sensitive information contained in the data. Current studies of PPDM mainly focus on how to reduce the privacy risk brought by data mining operations while in fact unwanted disclosure of sensitive information may also happen in the process of data collecting data publishing and information In particular we identify four different types of users involved in data mining applications namely data provider data collector data miner and decision maker. For each type of user we discuss his privacy concerns and the methods that can be used to protect sensitive information.
Keywords: - Data mining sensitive information privacy preserving data mining.

1. INTRODUCTION

Data mining has attracted more and more attention in recent years probably because of the popularity of the ''big data'' concept. Data mining is the process of discovering interesting patterns and knowledge from large amounts of data [1]. As a highly application driven discipline data mining has been successfully applied to many domains such as business intelligence Web search scientific discovery digital libraries etc.

To deal with the privacy issues in data mining a subfield of data mining referred to as privacy preserving data mining (PPDM) has gained a great development in recent years. The objective of PPDM is to safeguard sensitive information from unsolicited or unsanctioned disclosure and meanwhile preserve the utility of the data. The consideration of PPDM is two-fold. First sensitive raw data such as individual's ID card number and cell phone number should not be directly used for mining. Second sensitive mining results whose disclosure will result in privacy violation should be excluded. After the pioneering work of [3] [4] numerous studies on PPDM have been conducted [5] [7] .We can identify four different types of users namely four user roles in a typical data mining scenario.

Data Provider: The user who owns some data that are desired by the data mining task or the owner of data.

Data Collector: The user who collects data from data providers and then publishes the data to the data miner.

Data Miner: The user who performs data mining tasks on the data.

2. APPROACHES TO PRIVACY PROTECTION

2.1. Data provider

A data provider provides his data to the collector in an active way or a passive way. By ''active'' we mean that the data provider voluntarily opts in a survey initiated by the data collector or doing in some registration forms to create an account in a website. By ''passive'' we mean that the data which are generated by the provider's routine activities are recorded by the data collector while the data provider may even have no awareness of the disclosure of his data. When the data provider provides his data actively he can simply ignore the collector's demand for the information that he deems very sensitive. If his data are passively provided to the data collector the data provider can take some measures to limit the collector's access to his sensitive data.

Suppose that the data provider is an Internet user who is afraid that his online activities may expose his privacy. To protect privacy the user can try to erase the traces of his online activities by emptying browser's cache deleting cookies clearing usage records of applications etc. Also the provider can utilize various security tools that are developed for Internet environment to protect his data. Many of the security tools are designed as browser extensions for ease of use.

1).Anti-tracking extensions. Knowing that valuable information can be extracted from the data users online activities Internet companies have a strong motivation to track the users' movements on the Internet. When browsing the Internet a user can utilize an anti-tracking extension to block the trackers from collecting the cookies. Popular anti-tracking extensions include Disconnect Do Not Track Me Ghostery etc. A major technology used for anti-tracking is called Do Not Track (DNT) [10]. Which enables users to opt out of tracking by websites they do not visit. A user's opt-out preference is signaled by an HTTP header named DNT.DNT is not only a technology but also a policy framework for how companies that receive the signal

[1] Assistant professor, Vaish college of Engg Rohtak

should respond. The W3C Tracking Protection Working Group [11] is now trying to standardize how websites should response to user's DNT request.
2) Advertisement and script blockers. This browser extensions can block advertisements on the sites and kill scripts and widgets that send the user's data to some unknown third party. Example tools include Ad Block Plus No Script Flash Block etc.
3) Encryption tools. To make sure a private online communication between two parties cannot be intercepted by third parties a user can utilize encryption tools such as Mail Cloak and Tor Chat to encrypt his emails instant messages or other types of web traffic. Also a user can encrypt all of his internet traffic by using a virtual private network service.

2.2 Data collector
A data collector collects data from data provider in order to support the subsequent data mining operations. PPDP mainly studies anonymization approaches for publishing useful data while preserving privacy. The original data is assumed to be a private table consisting of multiple records. Each record consists of the following 4 types of attributes:
Identifier (ID): Attributes that can directly and uniquely identify an individual such as name ID number and mobile number.
Quasi-identifer (QID): Attributes that can be linked with external data to re-identify individual records such as gender age and zip code.
Sensitive Attribute (SA): Attributes that an individual wants to conceal such as disease and salary.
Non-sensitive Attribute (NSA): Attributes other than ID QID and SA.
Before being published to others the table is anonymized that is identifiers are removed and quasi-identifiers are modified.

2.3 Data Miner
In order to discover useful knowledge which is desired by the decision maker the data miner applies data mining algorithms to the data obtained from data collector. The primary concern of data miner is how to prevent sensitive information from appearing in the mining results. To perform a privacy-preserving data mining the data miner usually needs to modify the data he got from the data collector. As a result the decline of data utility is inevitable. Similar to data collector the data miner also faces the privacy-utility trade-off problem. But in the context of PPDM quantifications of privacy and utility are closely related to the mining algorithm employed by the data miner.

3. PRIVACY PRESERVING ASSOCIATION RULE MINING
Association rule mining is one of the most important data mining tasks which aims at finding interesting associations and correlation relationships among large sets of data items [12]. A typical example of association rule mining is market basket analysis [1] which analyzes customer buying habits by finding associations between different items that customers place in their shopping baskets. These associations can help retailers develop better marketing strategies. The problem of mining association rules can be formalized as follows [1].Given a set of items I ={i1 i2 ... im} and a set of transactions T = {t1 t2 ... tn} where each transaction consists of several items from I. An association rule is an implication of the form: A ⇒ B where A ⊂ I B ⊂ I A 6= ∅ B 6= ∅ and A∩B 6= ∅. The rule A ⇒ B holds in the transaction set T with support s where s denotes the percentage of transactions in T that contain A∪B. The rule A⇒B has confidence c in the transaction set T where c is the percentage of transactions in T containing A that also contain B. Generally the process of association rule mining contains the following two steps:
Step 1: Find all frequent item sets. A set of items is referred to as an item set. The occurrence frequency of an item set is the number of transactions that contain the item set. A frequent item set is an item set whose occurrence frequency is larger than a predetermined minimum support count.
Step 2: Generate strong association rules from the frequent item sets. Rules that satisfy both a minimum support threshold (minsup) and a minimum confidence threshold (minconf) are called strong association rules. Given the thresholds of support and confidence the data miner can set a association rules from the transactional data set. Some of the rules are considered to be sensitive either from the data provider's perspective or from the data miner's perspective. To hiding these rules the data miner can modify the original data set to generate a sanitized data set from which sensitive rules cannot be mined while those non-sensitive ones can still be discovered at the same thresholds or higher.
Various kinds of approaches have been proposed to perform association rule hiding [13] [14]. These approaches can roughly be categorized into the following groups:
Heuristic distortion approaches which resolve how to select the appropriate data sets for data modification.
Heuristic blocking approaches which reduce the degree of support and confidence of the sensitive association rules by replacing certain attributes of some data items with a specific symbol (e.g. '?').
Probabilistic distortion approaches which distort the data through random numbers generated from a predefined probability distribution function.
Exact database distortion approaches which formulate the solution of the hiding problem as a constraint satisfaction problem (CSP) and apply linear programming approaches to its solution.
Reconstruction based approaches which generate a database from the scratch that is compatible with a given set of non-sensitive association rules.

The main idea behind association rule hiding is to modify the support or confidence of certain rules. Here we briefly review some of the modification approaches proposed in recent studies.

Transaction ID	Items	Modified Items
T1	ABC	AB
T2	ABC	ABC
T3	ABC	ABC
T4	AB	AB
T5	A	AC
T6	AC	AC

Fig 1 Altering the position of sensitive item (e.g. C to hide sensitive association rules [15].

Jain et al. [15] propose a distortion-based approach for hiding sensitive rules where the position of the sensitive item is altered so that the confidence of the sensitive rule can be reduced but the support of the sensitive item is never changed and the size of the database remains the same. For example given the transactional data set shown in Fig. 1 set the threshold of support at 33% and the threshold of confi dence at 70% then the following three rules can be mined from the data: C ⇒ A(66.67% 100%) A B ⇒ C (50% 75%) C A ⇒ B(50% 75%).. If we consider the item C to be a sensitive item then we can delete C from the transaction T1 and add C to the transaction T5. As a result the above three rules cannot be mined from the modified data set.

Zhu et al. [16] employ hybrid partial hiding (HPH) algorithm to reconstruct the support of item set and then uses Apriori [1] algorithm to generate frequent item sets based on which only non-sensitive rules can be obtained. Le et al. [17] propose a heuristic algorithm based on the intersection lattice of frequent item sets for hiding sensitive rules. The algorithm determines the victim item such that modifying this item causes the least impact on the set of frequent item sets. Then the minimum numbers of transactions that need to be modified are specified. After that the victim item is removed from the specified transactions and the data set is sanitized. Dehkoridi [18] considers hiding sensitive rules and keeping the accuracy of transactions as two objectives of some function and applies genetic algorithm to find the best solution for sanitizing original data. Bonam et al [19] treat the problem of reducing frequency of sensitive item as a non-linear and multidimensional optimization problem. They apply particle swarm optimization (PSO) technique to this problem since PSO can find high-quality solutions eficiently while requiring negligible parametrization.

Modi et al. [20] propose a heuristic algorithm named DSRRC (decrease support of right hand side item of rule clusters) for hiding sensitive association rules. The algorithm clusters the sensitive rules based on certain criteria in order to hide as many as possible rules at one time. One short-coming of this algorithm is that it cannot hide association rules with multiple items in antecedent (left hand side) and consequent (right hand side). To overcome this shortcoming Radadiya et al. [21] propose an improved algorithm named ADSRRC (advance DSRRC) where the item with highest count in right hand side of sensitive rules are iteratively deleted during the data sanitization process. Pathak et al. [22] propose a hiding approach which uses the concept of impact factor to build clusters of association rules. The impact factor of a transaction is equal to number of item sets that are present in those item sets which represents sensitive association rule. Higher impact factor means higher sensitivity. Utilizing the impact factor to build clusters can help to reduce the number of modifications so that the quality of data is less affected.

Among different types of approaches proposed for sensitive rule hiding we are particularly interested in the reconstruction-based approaches where a special kind of data mining algorithms named inverse frequent set mining (IFM) can be utilized. The problem of IFM was investigated by Mielikainen in [23]. The IFM problem can be described as follows [24]: given a collection of frequent item sets and their support a transactional data set such that the data set precisely agrees with the supports of the given frequent item set collection while the supports of other item sets would be less than the pre-determined threshold. Guo et al [25] propose a reconstruction-based approach for association rule hiding where data reconstruction is implemented by solving an IFM problem. Their approach consists of three steps.

First use frequent item set mining algorithm to generate all frequent item sets with their supports and support counts from original data set.

Second determine which item sets are related to sensitive association rules and remove the sensitive item sets.

Third use the rest item sets to generate a new transactional data set via inverse frequent set mining.

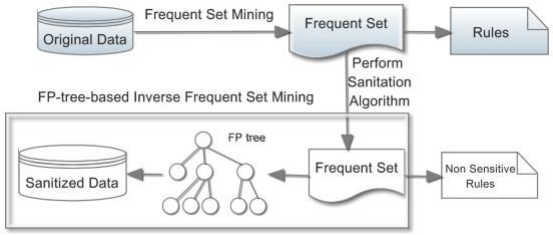

Fig 2. Reconstruction-based association rule hiding [25].

The idea of using IFM to reconstruct sanitized data set seems appealing. However the IFM problem is difficult to solve. Mielikainen [26] has proved that deciding whether there is a data set compatible with the given frequent sets is NP-complete. Researchers have made efforts towards reducing the computational cost of searching a compatible data set. Some representative algorithms include the vertical database generation algorithm [25] the linear program based algorithm [26] and the FP-tree-based method [27]. Despite the difficulty the IFM problem does provide us some interesting insights on the privacy preserving issue. Inverse frequent set mining can be seen as the inverse problem of frequent set mining. Naturally we may wonder whether we can do the inverse problems for other types of data mining problems. If the inverse problem can be clearly defined and feasible algorithms for solving the problem can be found then the data miner can use the inverse mining algorithms to customize the data to meet the requirements for data mining results such as the support of certain association rules or specific distributions of data categories. Therefore we think it is worth exploring the inverse mining problems in future research.

4. PRIVACY PRESERVING CLASSIFICATION MINING

Classification [1] is a form of data analysis that extracts models describing important data classes. Data classification can be seen as a two-step process. In the first step which is called learning step a classifi cation algorithm is employed to build a classifier (classification model) by analyzing a training set made up of tuples and their associated class labels. In the second step the classifier is used for classification i.e. predicting categorical class labels of new data. Typical classification model include decision tree Bayesian model support vector machine etc.

4.1 Decision Tree

A decision tree is a tree structure where each internal node (non-leaf node) denotes a test on an attribute each branch represents an outcome of the test and each leaf node (or terminal node) represents a class label [1]. Given a tuple X the attribute values of the tuple are tested against the decision tree. A path is traced from the root to a leaf node which holds the class prediction for the tuple. Decision trees can easily be converted to classification rules. To realize privacy-preserving decision tree mining Dowd et al. [28] propose a data perturbation technique based on random substitutions. Given a data tuple the perturbation is done by replacing the value of an attribute by another value that is chosen randomly from the attribute domain according to a probabilistic model. They show that such perturbation is immune to data-recovery attack which aims at recovering the original data from the perturbed data and repeated-perturbation attack where an adversary may repeat-edly perturb the data with the hope to recover the original data. Brickell and Shmatikov [29] present a cryptographically secure protocol for privacy-preserving construction of decision trees. The protocol takes place between a user and a server. The user's input consists of the parameters of the decision tree that he wishes to construct such as which attributes are treated as features and which attribute represents the class. The server's input is a relational database. The user's protocol output is a decision tree constructed from the server's data while the server learns nothing about the constructed tree. Fong et al. [30] introduce a perturbation and randomization based approach to protect the data sets utilized in decision tree mining. Before being released to a third party for decision tree construction the original data sets are converted into a group of unreal data sets from which the original data cannot be reconstructed without the entire group of unreal data sets. Meanwhile an accurate decision tree can be built directly from the unreal data sets. Sheela and Vijayalakshmi [31] propose a method based on secure multi-party computation (SMC) [32] to build a privacy-preserving decision tree over vertically partitioned data. The proposed method utilizes Shamir's secret sharing algorithm to securely compute the cardinality of scalar product which is needed when computing information gain of attributes during the construction of the decision tree.

4.2 Naive Bayesian Classification

Naive Bayesian classfication is based on Bayes' theorem of posterior probability. It assumes that the effect of an attribute value on a given class is independent of the values of other attributes. Given a tuple a Bayesian classifier can predict the probability that the tuple belongs to a particular class.

Vaidya et al [33] study the privacy-preserving classi cation problem in a distributed scenario where multi-parties collaborate to develop a classification model but no one wants to disclose its data to others. Based on previous studies on secure multiparty computation they propose different protocols to learn naive Bayesian classification models from vertically partitioned or horizontally partitioned data. For horizontally partitioned data all the attributes needed for classifying an instance are held by one site. Each party can directly get the classification result. therefore there is no need to hide the classification model. While for vertically partitioned data since one party does not know all the attributes of the instance he cannot learn the full model which means sharing the classification model is required. In this case protocols which can prevent the disclosure of sensitive information contained in the classification model (e.g. distributions of sensitive attributes) are desired. Skarkala et al. [34] also study the privacy-preserving classification problem for horizontally partitioned data. They propose a privacy-preserving version of the tree augmented naive (TAN) Bayesian classifier [35] to extract global information from horizontally partitioned data. Compared to classical naive Bayesian classifier TAN classifier can produce better classification results since it removes the assumption about conditional independence of attribute. Different from above work Vaidya et al. [36] consider a centralized scenario where the data miner has centralized access to a data set. The miner would like to release a classifier on the premise that sensitive information about the original data owners cannot be inferred from the

classification model. They utilize differential privacy model [36] to construct a privacy-preserving Naive Bayesian classifier. The basic idea is to derive the sensitivity for each attribute and to use the sensitivity to compute Laplacian noise. By adding noise to the parameters of the classifier the data miner can get a classifier which is guaranteed to be differentially private.

4.3 Support Vector Machine

Support Vector Machine (SVM) is widely used in classification [1]. SVM uses a nonlinear mapping to transform the original training data into a higher dimension. Within this new dimension SVM searches for a linear optimal separating hyperplane (i.e. a decision boundary separating tuples of one class from another) by using support vectors and margins (de fined by the support vectors).

Vaidya et al. [37] propose a solution for constructing a global SVM classification model from data distributed at multiple parties without disclosing the data of each party. They consider the kernel matrix which is the central structure in a SVM to be an intermediate that does not disclose any information on local data but can generate the global model. They propose a method based on gram matrix computation to securely compute the kernel matrix from the distributed data. Xia et al. [38] consider that the privacy threat of SVM-based classification comes from the support vectors in the learned classifier. The support vectors are intact instances taken from training data hence the release of the SVM classifier may disclose sensitive information about the original owner of the training data. They develop a privacy-preserving SVM classifier based on hyperbolic tangent kernel. The kernel function in the classifier is an approximation of the original one. The degree of the approximation which is determined by the number of support vectors represents the level of privacy preserving. Lin and Chen [39] also think the release of support vectors will violate individual's privacy. They design a privacy-preserving SVM classifier based on Gaussian kernel function. Privacy-preserving is realized by transforming the original decision function which is determined by support vectors to an infinite series of linear combinations of monomial feature mapped support vectors. The sensitive content of support vectors are destroyed by the linear combination while the decision function can precisely approximate the original one.

5. PRIVACY-PRESERVING CLUSTERING MINING

Cluster analysis [1] is the process of grouping a set of data objects into multiple groups or clusters so that objects within a cluster have high similarity but are very dissimilar to objects in other clusters. Dissimilarities and similarities are assessed based on the attribute values describing the objects and often involve distance measures. Clustering methods can be categorized into partitioning methods hierarchical methods density-based methods etc.

Current studies on privacy-preserving clustering can be roughly categorized into two types namely approaches based on perturbation and approaches based on secure multi-party computation (SMC).

Perturbation-based approach modifies the data before performing clustering. Oliveira and Zaiane [40] introduce a family of geometric data transformation methods for privacy-preserving clustering. The proposed transformation methods distort confidential data attributes by translation scaling or rotation (see Fig. 2) while general features for cluster analysis are preserved. Oliveira and Zaiane have demonstrated that the transformation methods can well balance privacy and effectiveness where privacy is evaluated by computing the variance between actual and perturbed values and effectiveness is evaluated by comparing the number of legitimate points grouped in the original and the distorted databases. The methods proposed in [40] deal with numerical attributes while in [40] Rajalaxmi and Natarajan propose a set of hybrid data transformations for categorical attributes. Recently Lakshmi and Rani [41] propose two hybrid methods to hide the sensitive numerical attributes. The methods utilize three different techniques namely singular value decomposition (SVD) rotation data perturbation and independent component analysis. SVD can identify information that is not important for data mining while ICA can identify that important information. Rotation data perturbation can retains the statistical properties of the data set. Compared to method solely based on perturbation the hybrid methods can better protect sensitive data and retain the important information for cluster analysis.

The SMC-based approaches make use of primitives from secure multi-party computation to design a formal model for preserving privacy during the execution of a clustering algorithm. Two pioneer studies on SMC-based clustering are presented in [42] and [43]. Vaidya and Clifton [44] present a privacy-preserving method for k-means clustering over vertically partitioned data where multiple data sites each having different attributes for the same set of data points wish to conduct k-means clustering on their joint data. At each iteration of the clustering process each site can securely find the cluster with the minimum distance for each point and can independently compute the components of the cluster means corresponding to its attributes. A check Threshold algorithm is proposed to determine whether the stopping criterion is met. Jha et al. [43] design a privacy-preserving k-means clustering algorithm for horizontally partitioned data where only the cluster means at various steps of the algorithm are revealed to the participating parties. They present two protocols for privacy-preserving computation of cluster means. The first protocol is based on oblivious polynomial evaluation and the second one uses homomorphic encryption. Based on above studies many privacy-preserving approaches have been developed for k-means clustering. Meskine and Bahloul present an overview of these approaches in [44].

Most of the SMC-based approaches deal with semi-honest model which assumes that participating parties always follow the protocol. In a recent study Akhter et al. [44] consider the malicious model where a party may substitute its local input or abort the protocol prematurely. They propose a protocol based on NIZK (non-interactive zero knowledge) proofs to conducting privacy-preserving k-means clustering between two parties in a malicious model.

In [45] Yi and Zhang identify another shortcoming of previous protocols that is each party does not equally contribute to k-means clustering. As a result a party who learns the outcome prior to other parties may tell a lie of the outcome to other parties. To prevent this attack they propose a k-means clustering protocol for vertically partitioned data in which each party equally contributes to the clustering. The basic idea is that at each iteration of k-means clustering multi-parties cooperate to encrypt k values (each corresponds to a distance between a data point and a cluster center) with a common public key and then securely compare the k values in order to assign the point to the closest cluster. Based on the assignment each party can update the means corresponding to his own attributes. Intermediate information during the clustering process such as the aforementioned k values are not revealed to any party. Under this protocol no party can learn the outcome prior to other parties.Different from previous studies which focus on k-means clustering De and Tripathy [46] recently develop a secure algorithm for hierarchical clustering over vertically partitioned data. There are two parties involved in the computation. In the proposed algorithm each party first computes k clusters on their own private data set. Then both parties compute the distance between each data point and each of the k cluster centers. The resulting distance matrices along with the randomized cluster centers are exchanged between the two parties. Based on the information provided by the other party each party can compute the final clustering result.

6. CONCLUSION

How to protect sensitive information from the security threats brought by data mining has become a hot topic in recent years. In this paper we review the privacy issues related to data mining by using a user-role based methodology. We differentiate four different user roles that are commonly involved in data mining applications i.e. data provider data collector data miner and decision maker. Each user role has its own privacy concerns; hence the privacy-preserving approaches adopted by one user role are generally different from those adopted by others:

For data provider his privacy-preserving objective is to effectively control the amount of sensitive data revealed to others. To achieve this goal he can utilize security tools to limit other's access to his data sell his data at auction to get enough compensation for privacy loss or falsify his data to hide his true identity.

For data collector his privacy-preserving objective is to release useful data to data miners without disclosing data providers' identities and sensitive information about them. To achieve this goal he needs to develop proper privacy models to quantify the possible loss of privacy under different attacks and apply anonymization techniques to the data.

For data miner his privacy-preserving objective is to get correct data mining results while keep sensitive information undisclosed either in the process of data mining or in the mining results. To achieve this goal he can choose a proper method to modify the data before certain mining algorithms are applied to or utilize secure computation protocols to ensure the safety of private data and sensitive information contained in the learned model.

7. REFERENCES

[1] J. Han M. Kamber and J. Pei Data Mining: Concepts and Techniques. San Mateo CA USA: Morgan Kaufmann 2006.

[2] L.BrankovicandV.Estivill-Castro ''Privacy issues in knowledge discovery and data mining '' in Proc. Austral. Inst. Comput. Ethics Conf. 1999 pp. 89–99.

[3] R. Agrawal and R. Srikant ''Privacy-preserving data mining '' ACM SIGMOD Rec. vol. 29 no. 2 pp. 439–450 2000.

[4] Y. Lindell and B. Pinkas ''Privacy preserving data mining '' in Advances in Cryptology. Berlin Germany: Springer-Verlag 2000 pp. 36–54.

[5] C. C. Aggarwal and S. Y. Philip A General Survey of PrivacyPreserving Data Mining Models and Algorithms. New York NY USA: Springer-Verlag 2008.

[6] S.Matwin ''Privacy-preserving data mining techniques: Survey and challenges ' 'inDiscrimination and Privacy in the Information Society.Berlin Germany: Springer-Verlag 2013 pp.209–221.

[7] O. Tene and J. Polenetsky ''To track or 'do not track': Advancing transparency and individual control in online behavioral advertising '' Minnesota J. Law Sci. Technol. no. 1 pp. 281–357 2012.

[8] R. T. Fielding and D. Singer. (2014). Tracking Preference Expression (DNT). W3C Working Draft. [Online]. Available: http://www.w3.org/TR/2014/WD-tracking-dnt-20140128/

[9] R. Agrawal T. Imieliński and A. Swami ''Mining association rules between sets of items in large databases '' in Proc. ACM SIGMOD Rec. 1993 vol. 22 no. 2 pp. 207–216.

[10] V. S. Verykios ''Association rule hiding methods '' Wiley Interdiscipl. Rev. Data Mining Knowl. Discovery vol. 3 no. 1 pp. 28–36 2013.

[11] K. Sathiyapriya and G. S. Sadasivam ''A survey on privacy preserving association rule mining '' Int. J. Data Mining Knowl. Manage. Process vol. 3 no. 2 p. 119 2013. [15] D. Jain P. Khatri R. Soni and B. K. Chaurasia ''Hiding sensitive association rules without altering the support of sensitive item(s) '' in Proc .2ndInt .Conf.Adv .Comput.Sci .Inf.Technol .Netw.Commun. 2012 pp. 500–509.

[12] J.-M.Zhu N.Zhang andZ.-Y.Li ''A new privacy preserving association rule mining algorithm based on hybrid partial hiding strategy '' Cybern. Inf. Technol. vol. 13 pp. 41–50 Dec. 2013.

[13] H. Q. Le S. Arch-Int H. X. Nguyen and N. Arch-Int ''Association rule hiding in risk management for retail supply chain collaboration ''Comput. Ind. vol. 64 no. 7 pp. 776–784 Sep.2013.

[14] M.N.Dehkordi '' A novel association rule hiding approach in OLAPdata cubes '' Indian J. Sci. Technol. vol. 6 no. 2 pp. 4063–4075 2013.

[15] J. Bonam A. R. Reddy and G. Kalyani ''Privacy preserving in association rule mining by data distortion using PSO '' in Proc. ICT Critical Infrastruct. Proc. 48th Annu. Conv. Comput. Soc. India vol. 2. 2014 pp. 551–558.

[16] C. N. Modi U. P. Rao and D. R. Patel ''Maintaining privacy and data quality in privacy preserving association rule mining ''inProc.Int.Conf. Comput. Commun. Netw. Technol. (ICCCNT) Jul. 2010 pp. 1–6.

[17] N. R. Radadiya N. B. Prajapati and K. H. Shah ''Privacy preserving in association rule mining '' Int. J. Adv. Innovative Res. vol. 2 no. 4 pp. 203–213 2013.

[18] K. Pathak N. S. Chaudhari and A. Tiwari ''Privacy preserving association rule mining by introducing concept of impact factor '' in Proc. 7th IEEE Conf. Ind. Electron. Appl. (ICIEA) Jul. 2012 pp. 1458–1461.

[19] T.Mielikäinen ' 'On inverse frequent set mining ''inProc.2nd Workshop Privacy Preserving Data Mining 2003 pp. 18–23.

[20] X. Chen and M. Orlowska ''A further study on inverse frequent set mining '' in Proc. 1st Int. Conf. Adv. Data Mining Appl. 2005 pp. 753–760.
[21] Y. Guo ''Reconstruction-based association rule hiding '' in Proc. SIGMOD Ph. D. Workshop Innovative Database Res. 2007 pp. 51–56.
[22] Y. Wang and X. Wu ''Approximate inverse frequent itemset mining: Privacy complexity and approximation '' in Proc. 5th IEEE Int. Conf. Data Mining Nov. 2005 p. 8.
[23] Y. Guo Y. Tong S. Tang and D. Yang ''A FP-tree-based method for inverse frequent set mining '' in Proc. 23rd Brit. Nat. Conf. Flexible Efficient Inf. Handling 2006 pp. 152–163.
[24] J. Dowd S. Xu and W. Zhang ''Privacy-preserving decision tree mining based on random substitutions '' in Proc. Int. Conf. Emerg. Trends Inf. Commun. Security 2006 pp. 145–159.
[25] J. Brickell and V. Shmatikov ''Privacy-preserving classifier learning '' in Proc. 13th Int. Conf. Financial Cryptogr. Data Security 2009 pp. 128–147.
[26] P. K. Fong and J. H. Weber-Jahnke ''Privacy preserving decision tree learning using unrealized data sets '' IEEE Trans. Knowl. Data Eng. vol. 24 no. 2 pp. 353–364 Feb. 2012.
[27] M. A. Sheela and K. Vijayalakshmi ''A novel privacy preserving decision tree induction '' in Proc. IEEE Conf. Inf. Commun. Technol. (ICT) Apr. 2013 pp. 1075–1079.
[28] O. Goldreich. (2002). Secure Multi-Party Computation. [Online].Available: http://www.wisdom.weizmann.ac.il/∼oded/PS/prot.ps
[29] J. Vaidya M. Kantarcıoğlu and C. Clifton ''Privacy-preserving Naïve Bayes classification '' Int. J. Very Large Data Bases vol. 17 no. 4 pp. 879–898 2008.
[30] M. E. Skarkala M. Maragoudakis S. Gritzalis and L. Mitrou ''Privacy preserving tree augmented Naïve Bayesian multi-party implementation on horizontally partitioned databases '' in Proc. 8th Int. Conf. Trust Privacy Security Digit. Bus. 2011 pp. 62–73.
[31] F. Zheng and G. I. Webb ''Tree augmented Naïve Bayes '' in Proc. Encyclopedia Mach. Learn. 2010 pp. 990–991.
[32] J. Vaidya B. Shafiq A. Basu and Y. Hong ''Differentially private Naïve Bayes classification '' in Proc. IEEE/WIC/ACM Int. Joint Conf. Web Intell. (WI) Intell. Agent Technol. (IAT) vol. 1. Nov. 2013 pp. 571–576.
[33] C.Dwork ''Differentialprivacy ''inProc. 33rdInt. Conf. Autom. Lang. Program. 2006 pp. 1–12.
[34] J.Vaidya H.Yu and X.Jiang ''Privacy-preserving SVM classification '' Knowl. Inf. Syst. vol. 14 no. 2 pp. 161–178 2008.
[35] H. Xia Y. Fu J. Zhou and Y. Fang ''Privacy-preserving SVM classifier with hyperbolic tangent kernel '' J. Comput. Inf. Syst. vol. 6 no. 5 pp. 1415–1420 2010.
[36] K.-P. Lin and M.-S. Chen ''On the design and analysis of the privacypreserving SVM classifier '' IEEE Trans. Knowl. Data Eng. vol. 23 no. 11 pp. 1704–1717 Nov. 2011.
[37] R. R. Rajalaxmi and A. M. Natarajan ''An effective data transformation approach for privacy preserving clustering '' J. Comput. Sci. vol. 4 no. 4 pp. 320–326 2008.
[38] M.N.Lakshmiand K.S.Rani ''SVD based data transformation methods for privacy preserving clustering '' Int. J. Comput. Appl. vol. 78 no. 3 pp. 39–43 2013.
[39] J. Vaidya and C. Clifton ''Privacy-preserving k-means clustering over vertically partitioned data ''inProc.9th ACMSIGKDDInt .Conf.Knowl. Discovery Data Mining 2003 pp. 206–215.
[40] S. Jha L. Kruger and P. McDaniel ''Privacy preserving clustering '' in Proc. 10th Eur. Symp. Res. Comput. Security (ESORICS) 2005 pp. 397–417.
[41] R. Akhter R. J. Chowdhury K. Emura T. Islam M. S. Rahman and N. Rubaiyat ''Privacy-preserving two-party k-means clustering in malicious model '' in Proc. IEEE 37th Annu. Comput. Softw. Appl. Conf. Workshops (COMPSACW) Jul. 2013 pp. 121–126.
[42] X. Yi and Y. Zhang ''Equally contributory privacy-preserving k-means clustering over vertically partitioned data '' Inf. Syst. vol. 38 no. 1 pp. 97–107 2013.
[43] I. De and A. Tripathy ''A secure two party hierarchical clustering approach for vertically partitioned data set with accuracy measure '' in Proc. 2nd Int. Symp. Recent Adv. Intell. Informat. 2014 pp. 153–162.