International Journal of Scientific Research & Engineering Trends
Volume 7, Issue 5, Sept-Oct-2021, ISSN (Online): 2395-566X

Survey on Privacy Preserving Mining Techniques and Application

Phd Scholar Jayshree Boaddh, Dr.Shailja Sharma
Dept. of Computer Science & Engg.
Rabindranath Tagore University, Bhopal (M.P.)
jayshree.boaddh@gmail.com, shailja1901@rediffmail.com

Abstract- Digital platform increase the easiness of data organization and utility. Extraction of information from raw data was performed by data mining algorithms. This information has many applications but few of miners extract knowledge which might affect the privacy of individual, organization, community, etc. So this paper focuses on finding the techniques which provide privacy of data against data mining algorithms. Paper has performed a survey on recent methodology proposed by different researcher. Some of data mining methods were also describe in the paper which help in information extraction. Evaluation parameters were detailed for comparison of privacy preserving methods.

Keywords: Data mining, Information Extraction, Association Rule,

I. INTRODUCTION

With the development of technologies and the pervasiveness of computing devices, various types of data dramatically increase. The collected data in the big data era usually contains auxiliary information and hidden knowledge. Great advances in computing technologies bring many benefits to society, with transformative changes and financial opportunities being created in health-care, transportation, industry, education, commerce, and social interactions. During the whole process from data collection to knowledge discovery, the data typically contains sensitive and individual information such as medical and financial information, and they may be exposed to several parties including collectors, owners, users and miners. However, the collected information may contain sensitive and private information, which raises privacy concerns [1]. Up to now, there is no standard definition of privacy concept.

In [1] gave a definition of privacy, in terms of the control of the data, but explicitly incorporate the risks of privacy violation. Data mining and analytics [2], [3] have shown a powerful ability to explore data (usually large amounts of data, also known as "big data"). Therefore, data mining technologies are commonly used in many real-world applications to extract hidden valuable knowledge by analyzing large amounts of data, typically business data or other sensitive data. For the health-care data, these datasets contain confidential or secure information (i.e., personal identification number, social security number, credit card number, etc.) and lead to privacy threats if they are published in the public place or misused, especially for government agencies and commercial institutions. Data privacy and security are the key problems in data management and analytics. Transforming the data may reduce its utility, resulting in inaccurate or even infeasible extraction of knowledge through data mining. This is the paradigm known as Privacy-Preserving Data Mining (PPDM) [2], [4], [5]. The common method to protect data is encryption, but it reduces the analyzability of data and is seldom used by data mining techniques. PPDM methodologies are designed to guarantee a certain level of privacy, while maximizing the utility of data, such that data mining can still be performed on the transformed data efficiently. PPDM conducts data mining operations under the condition of preserving data privacy. Most of them focus on two aspects, such as privacy and utility, and they have conflict relationship [2], [4], [5]. With consideration of privacy and utility, PPDM aims at achieving better tradeoff between utility maximizing and privacy preserving. Both privacy and utility are critically important for PPDM. Here the term of "utility" can be referred as the availability of data.

II. LITERATURE SURVEY

In [6] privacy of Internet of Health (IOH) data was done in three module. First was the LSH (Locality-Sensitive Hashing) into multi-source IoH data fusion and integration so as to secure the sensitive information of patients hidden in the past IoH data. Second was the IoH data without patient privacy after LSH process, we bring forth a similar IoH data record search method for subsequent IoH data mining and analyses, so as to balance the IoH data availability and privacy. Finally based on a dataset collected by real-world users, we validate the advantages of the proposed work in this paper, through a set of pre-designed experiments.

In [7] author focus on *k*-nearest neighbor (*k*NN) in this study to realize classification. Although several studies have already attempted to address the privacy problems associated with *k*NN computation in a cloud environment,

© 2021 IJSRET

the results of these studies are still inefficient. In this paper, we propose a very ef_cient and privacy-preserving *k*NN classification (P*k*NC) over encrypted data. While the amount of computation (encryptions/decryptions and exponentiations) and communication of the most efficient *k*NN classification proposed in prior studies is bounded by $O(kln)$, that of the proposed P*k*NC is bounded by $O(ln)$, where l is the domain size of data and n is the number of data.

In [8] author propose an efficient protocol to evaluate whether an itemset is freq uent or not under the encrypted mining query on supermarket transactions. To improve the mining efficiency, we design a blocking algorithm. In this algorithm, we separate the encrypted transactions into blocks and only calculate bilinear pairings on ciphertexts of part blocks instead of all ciphertexts, which helps cut down the computation cost of the mining process. Finally, we evaluate the performance of our protocol by conducting theoretical analyses and simulator experiments in the aspects of computation cost, security, correctness, and running time.

In [9] author proposed a mixture-model-based label propagation algorithm against malicious adversaries with corruption abilities. Privacy constraints in this paper are mainly focused on individual privacy, which means no individual data value should be disclosed and no information can be traced back to a specific site. In addition, another constraint should be included is that no site except $P0$ shall gain the information on the task.

In [10] author work on the medical research data for the improvement by the collaborative association rule mining on vertically partitioned healthcare data. Privacy of patients must be preserved during this collaboration. Paper further proposed an efficient approach for privacy preserving association rule mining in the vertically partition healthcare data for discovering the correlation related to disease and preserving the privacy of the patients. Finally analyze the proposed scheme with the medical examination data and outpatient data. The analysis of results shows that the association between diseases and symptoms discovered using the collaborative mining as well as privacy of the patients is preserved.

III. DATA FEATURES FOR PRIVACY PRESERVING

1. Data distribution- At present, some algorithms execute privacy protection data mining on a centralized data and some on distributed data. Distributed data consist of and vertical partitioned data [11]. Different database records in different sites in horizontal partitioned data and in vertically partitioned data each database record attribute values in different sites.

2. Data distortion- This technique is to alter original data-base record before release, so as to achieve privacy protection purpose [12]. Data distortion methods include perturbation, blocking, merging or aggregation, swapping and sampling. All these techniques are accomplished by the alteration of an attribute value or granularity transformation of an attribute value.

3. Data mining algorithms- Privacy preserving data mining algorithm include classification mining, clustering, association rule mining and Bayesian networks etc.

4. Data or rules hidden- This technique refers to hide original data or rules of original data. Due to rules hidden of original data it is very complex to reform again, some person proposed heuristic method to solve this issue.

5. Privacy protection- In order to protect privacy there need to modify data carefully for achieving a high data utility. Do this for some reasons as. [13] Modify data based on adaptive heuristics methods and only modify selected values of, but not all values, which make information loss of data is minimum. [14] encryption technologies, such as secure multiparty computation. If each site know only their input and input but nothing about others, the calculations are safe. Data reconstruction method can reconstruct original data distribution from random data.

IV. TECHIQUES OF PRIVACY PRESERVING MINING

Preserving data mining using fuzzy logic, neural networks, and other asymmetric encryption methods and also comparisons are made to know the best to do further research.

Secure multi-party computation An alternative approach based on the multiparty computation is that every part of private data is validly known to one or more parties. Revealing private data to parties such as by whom the data is owned or the individual to whom the data refers to is not a condition of violating privacy. The problem arises when the private information is revealed to some other third parties. To deal with this problem, we use a specialized form of privacy preserving distributed data mining. Parties that each knows some of the private data participate in a protocol that generates the data mining results, [15] that guarantees no data items is revealed to other parties. Thus the process of data mining doesn't cause, or even increase the opportunity for breach of privacy.

Randomization: In randomization, by adding noise to hide actual data values, works because most data mining methods construct models that generalize the data. On average, adding noise preserves the data statistics, so one can reasonably expect that the data mining models will still be correct. The issue is that knowing the general

© 2021 IJSRET

qualities are not adequate for building a decision tree. Data mining must also help us figure out where to make the decision points, not just the decision on those ranges [16]. Data mining algorithms automatically find appropriate points to make such splits, but these points can be obscured by adding noise to the data. After adding noise, the data no longer has these obvious points, so a data mining algorithm is likely to pick bad decision points and produce poor results. Given the distribution of the noise added to the data, as well as the randomized data set, they could reconstruct the data set,,s distribution (but not actual data values). With this work data mining algorithm can construct a much more accurate decision tree than mining the randomized data onto, which approaches the accuracy of a decision tree constructed on the real data.

Association rules: As the information mining deals with generation of association rules, the change in support and confidence of the association rule for hiding sensitive rules is done. A new concept named not altering the support,,is proposed to hide an association rule. The support of sensitive item not being changed is the first characteristic of proposed algorithm. The position of the sensitive item is the only thing which changes.

The reduction of the confidence in the sensitive rules for change in the support of the sensitive item is the approach to modifying the database transaction. This is in contrast to this existing algorithm, which either decreases or increases the support of the sensitive item to modify the database transactions. One of the way of promotional business growth of the organization is information sharing. Intimidation of data sharing is majorly caused by recent trends in data mining. Balancing the privacy of the data as per the legitimate need of the user is the major problem. The original data is modified by the sanitization process to conceal sensitive knowledge before release so the problem can be addressed. Privacy preservation of sensitive knowledge is addressed by several researchers in the form of association rules by suppressing the frequent item sets.

Anonymization techniques: The randomization method is a simple technique which can be easily implemented at data collection time, because the noise added to a given record is independent of the behavior of other records [17]. This is also the weakness because outlier records can often be the difficult to mask. Another key weakness of the randomization frame work is that it does not consider the possibility that publicly available record can be used to identify the identity of the owners of those records. Therefore, a broad approach too many privacy transformations are to construct groups of anonymous records which are transformed to group specific way.

k-anonymity model and l-diversity: The kanonymity model was developed because of the possibility of indirect identification of records of public databases. This is because combinations of record attribute can be used to exactly identify individual records. In the k-anonymity method, we reduce the granularity of data representation of the use of techniques such as generalization and suppression. This granularity is reduced sufficiently that any given record maps onto at least k other records of the data. The l-diversity model was designed to handle some weaknesses for the kanonymity model since protecting identities to the level of kindividuals is not the same as protecting the corresponding sensitive values, especially when there is homogeneity of sensitive values within a group. To do so, the concept of intra-group diversity of sensitive values is promoted of the anonymization scheme. T-closeness: Anil Parkas, Ravindar Mogili found that KAnonymity and ℓ -Diversity was not used to prevent attribute disclosure. ℓ-Diversity would have well represented sensitive attribute value that was assigned only with certain number of limitations. T-closeness has been proposed to describe the distribution of sensitive attribute with equivalence class. Earth Mover Distance was utilized to measure the distance between the two probabilistic distributions .conjunction has been proposed to join machine learning and measurable analysis.

Sequential pattern hiding Sequential pattern hiding method [18] is necessary to conceal sensitive patterns that can otherwise be extracted from published data, without seriously affecting the data and the non sensitive interesting patterns. Sequential pattern hiding is a challenging problem, because sequences have more composite semantics than item sets, and calls for efficient solutions that offer high utility.

V. DATA MINING TECHNIQUES

1. Decision tree

Decision tree classification is the learning of decision trees from class labeled training tuples. A decision tree is a flowchart like tree structures, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node holds a class label. Advantages: Amongst other data mining methods, decision trees have various advantages. Decision trees are simple to understand and interpret. They require little data and are able to handle both numerical and categorical data. It is possible to validate a model using statistical tests. They are robust in nature, therefore, they perform well even if its assumptions are somewhat violated by the true model from which the data were generated. Decision trees perform well with large data in a short time. Large amounts of data can be analyzed using personal computers in a time short enough to enable stakeholders to take decisions based on its analysis.

2. Nearest neighbor classifier

The k-nearest neighbor's algorithm (k-NN) is a method for classifying objects based on closest training examples in the feature space. K-NN is a type of instance-based learning, or lazy learning. It can also be used for

© 2021 IJSRET

regression. The k-nearest neighbor algorithm is amongst the simplest of all machine-learning algorithms. The space is partitioned into regions by locations and labels of the training samples. A point in the space is assigned to the class c if it is the most frequent class label among the k nearest training samples. Usually Euclidean distance is used as the distance metric; however this will only work with numerical values. In cases such as text classification another metric, such as the overlap metric (or Hamming distance) can be used.

3.Artificial neural network

Neural Networks are analytic techniques modeled after the (hypothesized) processes of learning in the cognitive system and the neurological functions of the brain and capable of predicting new observations (on specific variables) from other observations (on the same or other variables) after executing a process of so called learning from existing data. Neural Networks is one of the Data Mining techniques. The first step is to design a specific network architecture (that includes a specific number of "layers" each consisting of a certain number of "neurons"). Network is then subjected to the process of "training." In that phase, neurons apply an iterative process to the number of inputs to adjust the weights of the network in order to optimally predict the sample data on which the "training" is performed. After the phase of learning from an existing data set, the new network is ready and it can then be used to generate predictions. The resulting "network" developed in the process of "learning" represents a pattern detected in the data.

4.Support vector machines

Support Vector Machines were first introduced to solve the pattern classification and regression. Support vector machines (SVMs) are a set of related supervised learning methods used for classification and regression. Viewing input data as two sets of vectors in an n-dimensional space, an SVM will construct a separating hyper-plane in that space, one which maximizes the margin between the two data sets. To calculate the margin, two parallel hyper-planes are constructed, one on each side of the separating hyper-plane, which are "pushed up against" the two data sets. A good separation is achieved by the hyper-plane that has the largest distance to the neighboring data points of both classes, since in general the larger the margin the lower the generalization error of the classifier. This hyperplane is found by using the support-vectors and margins.

5. Association rule

Association and correlation is usually to find frequent item set findings among large data sets. This type of finding helps businesses to make certain decisions, such as catalogue design, cross marketing and customer shopping behavior analysis. Association Rule algorithms need to be able to generate rules with confidence values less than one. However the number of possible Association Rules for a given dataset is generally very large and a high proportion of the rules are usually of little (if any) value.

VI. EVALUATION PARAMETERS

Direct Discrimination Prevention Degree (DDPD). This measure quantifies the percentage of discriminatory rules that are no longer discriminatory in the transformed dataset [19].

Direct Discrimination Protection Preservation (DDPP). This measure quantifies the percentage of the protective rules in the original dataset that remain protective in the transformed dataset [19].

Data Loss: As proposed work provide privacy for the sensitive item set rules with minimum data loss. As in privacy data perturbation make data loss.

Originality: As change in original data is the way to provide privacy in mining. So algorithm that will maintain maximum originality after perturbation is major expectation.

Execution time - Third parameter is to evaluate execution time time of the algorithm that is time taken by the proposed method for execution. Algorithm time is expect after the evaluation of the direct and indirect rules.

VII. CONCLUSION

Data mining has importance regarding finding the patterns, forecasting, discovery of knowledge etc., in different business domains. Data mining techniques and algorithms such as classification, clustering etc., helps in finding the patterns to decide upon the future trends in businesses to grow. Paper presented a broad survey on data mining techniques. Paper has discussed a variety of techniques and algorithms used for maintaining privacy or securing private and sensitive information. Still, there can be improvements in the defined algorithms. Better and improved algorithms must be defined which provides more security.

REFERENCES

[1]. E. Bertino, D. Lin, and W. Jiang, “A survey of quantification of privacy preserving data mining algorithms,” in Privacy-preserving data mining. Springer, 2008, pp. 183–205.

[2]. C. C. Aggarwal and P. S. Yu, “A general survey of privacy-preserving data mining models and algorithms,” in Privacy-preserving data mining. Springer, 2008, pp. 11–52.

[3]. J. Han, J. Pei, Y. Yin, and R. Mao, “Mining frequent patterns without candidate generation: A frequent-

© 2021 IJSRET

pattern tree approach," Data Mining and Knowledge Discovery, vol. 8, no. 1, pp. 53–87, 2004.
[4]. R. Agrawal and R. Srikant, Privacy-preserving data mining. ACM, 2000, vol. 29, no. 2.
[5]. Y. Lindell and B. Pinkas, "Privacy preserving data mining," in Annual International Cryptology Conference. Springer, 2000, pp. 36–54
[6]. Q. Zhang, B. Lian, P. Cao, Y. Sang, W. Huang and L. Qi, "Multi-Source Medical Data Integration and Mining for Healthcare Services," in IEEE Access, vol. 8, pp. 165010-165017, 2020.
[7]. J. Park and D. H. Lee, "Parallelly Running k-Nearest Neighbor Classification Over Semantically Secure Encrypted Data in Outsourced Environments," in IEEE Access, vol. 8, pp. 64617-64633, 2020.
[8]. C. Ma, B. Wang, K. Jooste, Z. Zhang and Y. Ping, "Practical Privacy-Preserving Frequent Itemset Mining on Supermarket Transactions," in IEEE Systems Journal, vol. 14, no. 2, pp. 1992-2002, June 2020.
[9]. Z. Li, L. Yang and Z. Li, "Mixture-Model-Based Graph for Privacy-Preserving Semi-Supervised Learning," in IEEE Access, vol. 8, pp. 789-801, 2019.
[10]. Nikunj Domadiya, Udai Pratap Rao. "Privacy Preserving Distributed Association Rule Mining Approach on Vertically Partitioned Healthcare Data". Procedia Computer Science Volume 148, 2019.
[11]. C C Aggarwal, P S Yu, "On static and dynamic methods for condensation-based privacy-preserving data mining," ACM Trans Database Syst, vol. 33, no. 1, 2008,doi: 10.1145/1331904.1331906.
[12]. M. Atallah, E. Bertino, A. Elmagarmid, M. Ibrahim and V. Verykios, "Disclosure Limitation of Sensitive Rules,"Proceedings of the IEEE Knowledge and Data Engineering Workshop, 1999, pp. 45-52.
[13]. J Lin, Y Cheng, "Privacy preserving itemset mining through noisy items," Expert Systems with Applications, vol. 36, Mar. 2009, pp. 5711-5717, doi: 10.1016/j.eswa.2008.06.052.
[14]. V.S. Verykios, E. Bertino, I.N. Fovino, L.P. Provenza, Y. Saygin and Y. Theodoridis, "State-of-the-art in privacy preserving data mining," ACM SIGMOD Record, vol. 33, no. 1, 2004, pp. 50-57, doi: 10.1145/974121.974131.
[15]. Jaideep Vaidya & Chris Clifton, "Privacy-Preserving Data Mining: Why, How, and When", the IEEE computer society, 2004.
[16]. Yu Zhu& Lei Liu, "Optimal Randomization for Privacy Preserving Data Mining", ACM, August 2004
[17]. L.Sweeney, "Achieving k-Anonymity Privacy Protection Using Generalization and Suppression", International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, vol.10, no.5.
[18]. C. C. Aggarwal, P. S. Yu, "Privacy Preserving Data Mining: Models and Algorithms". Springer, 2008.
[19]. Sara Hajian and Josep Domingo-Ferrer. "A Methodology for Direct and Indirect Discrimination Prevention in Data Mining". IEEE transaction on knowledge data engineering, VOL. 25, NO. 7, JULY 2013.

© 2021 IJSRET