International Journal of Scientific Engineering and Research (IJSER)
ISSN (Online): 2347-3878
Index Copernicus Value (2015): 62.86 | Impact Factor (2015): 3.791

A General Survey on Methods of Privacy Preserving Data Mining

K. Sowjanya

M. Tech Scholar, Sree Vidyanikethan Engineering College (Autonomous), Sree Sainath Nagar, Tirupati, India

Abstract: ***Background/Objectives:*** *The process of removing interesting patterns or knowledge from a huge amount of information is known as Data Mining. Now a day's information like PAN No data collection is ubiquitous, and every transaction is recorded somewhere. The subsequent information sets can comprise of terabytes or even pet bytes of information, so efficiency and scalability is the primary consideration of most data mining algorithms. Increasing data collection, along with the influx of analysis tools capable of handling huge volumes of information, has led to privacy concerns. Protecting private data is an important concern for society, several laws now require explicit consent prior to analysis of an individual's data, but its importance is not limited to individuals: corporations might also need to protect their information's privacy, even though sharing it for analysis could benefit the company. Clearly, the trade-off between sharing information about analysis and keeping it secret to preserve corporate trade secrets and customer privacy is a growing challenge. In recent years, with the explosive development in Internet, data storage and data processing technologies, privacy preservation has been one of the greater concerns in data mining.* ***Findings:*** *A variety of techniques and approaches have been presented and developed for privacy preserving information mining.* ***Methods:*** *This paper provides a basic survey of different privacy preserving data mining methods and analyses the representative techniques for privacy preserving information mining.*

Keywords: PPDM, Randomization, K-Anonymity, Association Rule

1. Introduction

Now a day's, data mining has been seen as a danger to privacy on account of the across the board multiplication of electronic information kept up by associations. This has prompted expanded worries about the protection for the hidden information. In the last few decades a number of approaches and techniques such as classification, association rule mining have been proposed to modifying or transforming the data onto such a way so as to preserve the privacy. Conservations of people data are a key to the information proprietors to guarantee his protection. Security expect a basic part in data conveyed Data mining process permits an organization to utilize extensive measure of information to create connections and connections among the information to enhance the business efficiency. Therefore privacy preserving information mining has become essential field of research. The Data Mining technology can develop these analyses on its own, using to commix of statistics, artificial intelligence, machine learning algorithms, and data stores. In order to face the challenging risk, some researchers have been proposed as a remedy of this awkward situation, which focus at achieving the parity of information utility and data protection when distributed dataset. The continuous research is called Privacy Preserving Data Publishing. Balancing the privacy of the information according to the real need of the client is the significant issue. The original information is adjusted by the purification procedure to conceal sensitive knowledge before discharge 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. 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. Confidentiality issues from data mining. A key problem that arises from any mass collection of data is that of confidentiality. The need for privacy is sometimes due to law (e.g., for medical databases) or can be motivated by business interests. The irony is that data mining results rarely violate privacy. The objective of data mining is to generalize across populations, rather than reveal information about individuals.

1.1 Privacy Preserving Data Mining

Privacy preserving data mining will be accomplished in various ways specifically by using randomization methods, cryptography algorithms and anonymization methods. A modern survey on are being used on various methods using privacy preserving data mining are found in which reviews major PPDM techniques based on merits and demerits on recent trends in PPDM. The current scenario privacy preserving data mining propose some future research directions for research people. In all methods of PPDM is studied and analyzed and from the analysis of cryptography, Random data perturbation methods does better than other existing methods and specially cryptography is the best technique for encrypt the sensitive data of large data set.

1.2 Privacy Preserving Data Mining (PPDM)

Methods

Many techniques have recently been proposed to privacy preserving information mining of multidimensional information set. Many privacies preserving data mining technologies are examined in clearly and the benefits and drawbacks are analyzed such as k- anonymity, l-diversity, t-closeness, classification, association rule mining are proposed and designed to prevent identification to preserve the primary sensitive information and several application of several techniques for preserving privacy on testing dataset are expressed .

In recent situation many number of methods have been proposed to modifying or transformation of data to preserving privacy which are much needed and an effective

but without compromising security to hide the delicate information. This paper expresses a complete detailed survey on recent algorithms which are proposed to achieving privacy

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.

1.2.1 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 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. 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. In a previous research Rakesh Aggarwal and Ramakrishna Srikant presented a solution to this problem. 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.

1.2.2 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.

1.2.3 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. 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.

1.2.4 The k-anonymity model and l-diversity: The *k*-anonymity 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 *k*-anonymity model since protecting identities to the level of *k*-individuals 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 K-Anonymity 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.

2. Open Issues and Challenges

2.1 Developing a Unifying Theory of Data Mining

The current state of the art of data-mining research is too”ad-hoc“techniques are designed for individual issues no unifying theory Needs unifying research Exploration vs. explanation Longs standing theoretical issues How to avoid spurious correlations Deep research. Knowledge discovery on hidden causes. Similar to discovery of Newton’s Law?

2.2 Scaling Up for High Dimensional Data and High Speed Streams

Scaling up is required ultra-high dimensional order issues (millions or billions of features, e.g., bio data) Ultra-high speeds data streams. Continuous online processing. How to monitor network packets of intruders? Concept drifts and environment drift? RFID network and sensor system information.

2.3. Sequential and Time Series Data

How to efficiently and accurately cluster, classify and predict the patterns? Time series information used for predictions is contaminated by noise. How to do accurate short-term and long-term predictions? Signal processing techniques lags behind the filtered data, which reduces accuracy Key to

source selection, domain knowledge in rules, and optimization methods.

2.4 Mining Complex Knowledge from Complex Data

Mining graphs other, and are not of a single type. Data that are not i.e. (independent and identically distributed) many objects are not independent of each mine the rich structure of relations of objects, E.g.: interlinked Webpages, social networks, metabolic networks in the cell Integration of data mining and knowledge inference The biggest gap: unable to relate the results of mining to the real-world decisions they affect - all they can do is hand the results back to the user More research on interestingness of knowledge

2.5 Data Mining in a Network Settings

Community and Social Networks Linked data between emails, Web pages, blogs, citations, sequences and people Static and dynamic structural behavior Mining and for Computer Networks. detect anomalies (e.g., sudden traffic spikes due to a DoS (Denial of Service)attack Need to handle 10Gig Ethernet links (a) detect (b)trace back (c) drop packet

2.6 Distributed Data Mining and Mining Multi-agent Data

Need to correlate the data seen at the various probes (such as in a sensor network) Adversary data mining: deliberately manipulate the data to sabotage them (e.g., make them produce false negatives) Game theory may by need for help.

2.7 Data Mining for Biological and Environmental Problems

New problems raise new questions Large scale problems especially so Biological data mining, such as HIV vaccine design DNA, chemical properties,3D structures, and functional properties need to be fused Environmental data mining for solving the energy crisis.

2.8 Data-mining-Process Related Problems

How to automate mining process the composition of data mining operations Data cleaning, with logging capabilities Visualization and mining automation. Need a methodology: help users avoid many data mining mistakes what are canonical set of data mining operations?

2.9 Security, Privacy and Data Integrity

How to ensure the users privacy while their data are being mined? How to do data mining for protection for security and privacy Knowledge, integrity, and assessment. Data are intentionally modified from their original version, in order to misinform the recipients or for privacy and security Development of measures to evaluate the knowledge integrity of a collection of Data Knowledge and patterns.

2.10 Dealing with Non-static, Unbalanced and Cost-sensitive Data

The UCI datasets are small and not highly unbalanced Real world data are large (10^5 features) but only < 1% of the useful classes (+'ve) There is much information on costs and benefits, but no overall model of profit and loss Data may evolve with a bias introduced by sampling.

- Each test incurs a cost
- Data extremely unbalanced
- Data change in time

3. Conclusion

This paper, discussed about different approaches and techniques used in privacy preservation of data mining. Due to the large collection of information, it is important to maintain the Privacy of sensitive information. Represent survey of the different approaches to privacy preserving information mining, and analyses the major algorithms available for each method and points out the existing drawback. While all the purposed techniques are only approximate to our goal of privacy preservation.All methods are approximate to our goal of privacy preservation, we need to further perfect those approaches or develop some efficient methods.

References

[1] Shamir, ―How to share a secret,‖ Common. ACM, vol. 22, pp. 612–613, 1979.
[2] Adam, Nabil R., and John C. Worthmann. "Security-control methods of statistical databases: a comparative study." ACM Computing Surveys (CSUR) 21.4 (1989).
[3] Y. Lindell and B. Pinkas, ―Privacy preserving data mining,‖ in Proc. 20th Annu. Int. Cryptol. Conf. Adv. Cryptol., 2000, pp. 36–54.
[4] Aggarwal, Dakshi, and Charu C. Aggarwal. "On the design and quantification of privacy preserving data mining algorithms." Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems.ACM, 2001.
[5] Agrawal, Dakshi, and Charu C. Aggarwal. "On the design and quantification of privacy preserving data mining algorithms." Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. ACM, 2001.
[6] Evfimievski, Alexandre, Johannes Gehrke, and Ramakrishnan Srikant. "Limiting privacy breaches in privacy preserving data mining." Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. ACM, 2003.
[7] Vaidya, Jaideep, and Chris Clifton. "Privacy-preserving k-means clustering over vertically partitioned data." Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2003.
[8] Kantarcıoğlu, Murat, and Chris Clifton. "Privately computing a distributed k-nn classifier." European conference on principles of data mining and knowledge discovery. Springer Berlin Heidelberg, 2004.

[9] Agrawal, Shipra, and Jayant R. Haritsa. "A framework for high-accuracy privacy-preserving mining." 21st International Conference on Data Engineering (ICDE'05). IEEE, 2005.
[10] Bayardo, Roberto J., and RakeshAgrawal. "Data privacy through optimal k-anonymization." 21st International Conference on Data Engineering (ICDE'05). IEEE, 2005.
[11] Aggarwal, Charu C. "On k-anonymity and the curse of dimensionality." Proceedings of the 31st international conference on Very large data bases. VLDB Endowment, 2005
[12] Bayardo, Roberto J., and RakeshAgrawal. "Data privacy through optimal k-anonymization." 21st International Conference on Data Engineering (ICDE'05). IEEE, 2005.
[13] Aggarwal, Charu C. "On randomization, public information and the curse of dimensionality." 2007 IEEE 23rd International Conference on Data Engineering. IEEE, 2007.
[14] Aggarwal III, Charu C., and S. Yu Philip. "A survey of randomization methods for privacy-preserving data mining." Privacy-Preserving Data Mining. Springer US, 2008. 137-156.
[15] Bertino, Elisa, Dan Lin, and Wei Jiang. "A survey of quantification of privacy preserving data mining algorithms." Privacy-preserving data mining. Springer US, 2008. 183-205.
[16] Hua, Ming, and Jian Pei. "A survey of utility-based privacy-preserving data transformation methods." Privacy-Preserving Data Mining. Springer US, 2008. 207-237. storage." Proceedings of the 15th ACM conference on Computer and communications security. ACM, 2008.
[17] Haritsa, Jayant R. "Mining association rules under privacy constraints."Privacy-Preserving Data Mining. Springer US, 2008. 239-266.
[18] Verykios, Vassilios S., and Aris Gkoulalas-Divanis. "A survey of association rule hiding methods for privacy." Privacy-Preserving Data Mining. Springer US, 2008. 267-289.
[19] Fienberg, Stephen E., and Aleksandra B. Slavkovic. "A survey of statistical approaches to preserving confidentiality of contingency table entries."Privacy-Preserving Data Mining. Springer US, 2008. 291-312.
[20] Kantarcioglu, Murat. "A survey of privacy-preserving methods across horizontally partitioned data." Privacy-preserving data mining. Springer US, 2008. 313-335.
[21] Liu, Kun, Chris Giannella, and Hillol Kargupta. "A survey of attack techniques on privacy-preserving data perturbation methods." Privacy-Preserving Data Mining. Springer US, 2008. 359-381.
[22] Williams, Peter, Radu Sion, and Bogdan Carbunar. "Building castles out of mud: practical access pattern privacy and correctness on untrusted storage." Proceedings of the 15th ACM conference on Computer and communications security. ACM, 2008.
[23] Bogdanov, Dan, Sven Laur, and Jan Willemson. "Sharemind: A framework for fast privacy-preserving computations." European Symposium on Research in Computer Security. Springer Berlin Heidelberg, 2008.
[24] Ciriani, Valentina, et al. "k-anonymous data mining: A survey." Privacy-preserving data mining. Springer US, 2008. 105-136.
[25] Domingo-Ferrer, Josep. "A survey of inference control methods for privacy-preserving data mining." Privacy-preserving data mining.Springer US, 2008.53-80.
[26] Venkatasubramanian, Suresh. "Measures of anonymity." Privacy-Preserving Data Mining. Springer US, 2008. 81-103.
[27] Aggarwal, Charu C., and S. Yu Philip. "A general survey of privacy-preserving data mining models and algorithms." Privacy-preserving data mining. Springer US, 2008. 11-52.
[28] di Vimercati, Sabrina De Capitani, Sara Foresti, and Pierangela Samarati. "Managing and accessing data in the cloud: Privacy risks and approaches."2012 7th International Conference on Risks and Security of Internet and Systems (CRiSIS). IEEE, 2012.
[29] Elmehdwi, Yousef, Bharath K. Samanthula, and Wei Jiang. "Secure k-nearest neighbor query over encrypted data in outsourced environments."2014 IEEE 30th International Conference on Data Engineering. IEEE, 2014.
[30] K. Samanthula, Y. Elmehdwi, and W. Jiang, ―k-nearest neighbor classification over semantically secure encrypted relational data, e-print arXiv: 1403.5001, 2014.