International Organization of Research & Development (IORD)
ISSN: 2348-0831
Vol 10 Issue 02 | 2023

A Security and Privacy Preserving in Big Data

Vishakha V. Kharche[1], Prof. Alokkumar Shukla[2]

[1] *Post Graduate Student, Department of CE, Padm. Dr. V.B.K.C.O.E., Malkapur, S.G.B.A. University, Maharashtra, India*
[2] *Asst. professor, Department of CSE, Padm. Dr. V.B.K.C.O.E., Malkapur, S.G.B.A. University, Maharashtra, India*

ABSTRACT

In recent era's, available personal data has made security and privacy preserving on big data. Major issue and privacy preserving on data and some exist problems in this field. On big data is a broad term for data sets so large or complex that traditional data processing applications are inadequate. Challenges are the ubiquity of computing and electronic communication technologies has led to the exponential growth of data from both digital and analog sources so new capabilities to collect, analyse, spread widely, and preserve vast quantities of data raise new concerns about the nature of privacy and the means by which individual privacy might be compromised or protected include analysis, capture, curation, find, sharing, storage, transfer, visualization, and information privacy. The term frequently refers simply to the use of predictive analytics or other certain advanced methods to extract value from data, and not often to a particular size of data set.

***Keywords*— Big data, Data mining, Sensitive information, Privacy preserving, k-anonymity.**

I. INTRODUCTION

Data mining has become attractive concept in recent years. The huge amount of data stored on servers which collectively called big data. Big data are characterized by 3 V's: Volume, Velocity, and Variety. Resulting data set contain petabytes of data. For this purpose we use Data mining process that extracts some useful information from data. To extract useful information from huge amount data we need some data mining algorithms. Big data have applications in many fields like in Business Technology, Health, Smart cities etc. In data mining some privacy issue are originated. Accessing information from a large data set is difficult and time consuming process, we have to follow certain protocols, a proper algorithm and method is needed to classify the data, find appropriate pattern from them.

The main purpose of privacy protection is to hide the sensitive information from the unauthorized access. Privacy preserving technology focused on data mining and data anonymity of two domains. Data mining Privacy protection methods are focused on data distortion, data encryption, and data released and so on, such as privacy protection classification mining algorithm, privacy protection association rules mining, distributed privacy preserving collaborative recommendation, data release and so on. Many algorithms were developed based on encryption methods, such as association rules mined in horizontally partitioned and vertically partitioned data, clustering mining, classification mining, and decision tree mining etc. In data mining privacy of data means keep information safe about individual from being available to others [1]. Privacy is a most important because it may have harm full effect on someone's life.

II. LITERATURE REVIEW

This paper describe a view of different algorithms from 1994-2013 which are important for handling big data set. It gives an overview of algorithms which used in large data sets and architecture. These algorithms also define a different type of structures and methods that are implemented to handle Big Data properly and list various tools that are developed for analyzing them. It also describes about the different security issues, application and trends in a large data set [2.]

The paper shows a large overview of the topic big data mining, its current status, dispute, and predict to the future [3].

This paper describe the big data security at the environment level along with investigate the built in protections to provide security in big data. It also describes some security issues that we are facing today and provide proper security solutions and commercially accessible techniques to address the same [4].

This paper shows an overview on big data, its importance in our day to day life and some technologies to manage big data. This paper also states Big Data application in a self-organizing websites which can be extended to the field of advertising in companies [5].

III. DATA SECURITY AND PRIVACY ISSUES IN BIG DATA

One of the key issues in data mining technology is not a technological or business one, but it is a social problem. It is the issue of individual privacy. Data mining makes it possible to describe routine business transactions and glean a significant amount of information about individuals buying habits and preferences. Another issue is that integrity of data. Clearly, data analysis is good as the data that is being examined in details. A key implementation challenge is combining conflicting or redundant data from different sources. For example, a bank may handle credit cards accounts in different databases. The names of a one card of a particular person may be different in each database. Software must transfer data from one system to another and select the names which are most recently used. Finally, there is the cost issue. In recent day system hardware costs increases. The more powerful the data mining queries, the greater the usefulness of the information being acquire small quantity from the data, and increase the pressure to increase the amount of data being collected and maintained, data mining queries becomes faster, more powerful by increasing the pressure. This increases pressure for systems which is bigger and faster and that are more costly [6]. Data mining, the extraction of hidden information from huge databases, is a strong new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools foretell future behaviors and trends and, allowing businesses to make excite, knowledge-driven decisions. The automated, prospective inspect offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can solve business questions which take more time to resolve means this process is time consuming. They search databases for invisible patterns, finding predictive information that experts may miss because it lies outside their expectations [7].
Privacy preserving in data mining is mainly applied to achieve privacy protection by different data characteristics in high-level data. Data release based privacy protection is to provide a common privacy protection method in many applications, thus making designed privacy algorithm is also versatile. The research of privacy protection methods are focused on data distortion, data encryption, and data released and so on, such as privacy protection classification mining algorithm, privacy protection association rules mining, distributed privacy preserving collaborative recommendation, data release and so on. Many algorithms were developed based on encryption methods, such as association rules mined in horizontally partitioned and vertically partitioned data, clustering mining, classification mining, and decision tree mining etc. In this paper because of these issues in big data we describe some security and privacy protection technologies used in big data.

IV. RELATED WORK

A. Big data

Big data is a collection of different types both structured and unstructured data. The main reason for most appearance and increasing growth of big data is increase capacities of storage and also increase in processing power and availability of data. Big data make the use of large data sets to handle the collection of data that use for businesses or other recipients in decision making. The data may be either public or private. Big data are characterized by 3 V's: Volume, Velocity, and Variety [4].
Volume -- the size of data extend the capacity over terabytes and petabytes of data. This would make difficult to store and analyse the data by using traditional tools.
Velocity – big data should be used to mine large amount of data within a pre defined time. The traditional methods of data mining may take large amount of time to mine such a volume of data.
Variety – Big data comes from different types of sources which include both structured and unstructured data. Database systems which was traditional designed to handle smaller volumes of structured and consistent data whereas Big Data is a collection of various types of data, images, audio's and video's, and unregulated text, including log files and social media. These different types of unstructured data create problems for storage, mining and analysing of the data. Because of enormous amount of data it is difficult to manage the whole data and it is very time consuming process. So we used data mining technique to separate useful information from the big data.

B. Data mining

Data mining is the process of extracting useful patterns and knowledge from large amounts of data [8]. As a highly application-driven discipline, data mining has been successfully applied to many areas, such as business

intelligence, Web search, scientific discovery, digital libraries, etc. Data mining" is also known as a synonym for ``knowledge discovery from data" (KDD) which highlights the goal of the data mining process. To obtain useful knowledge from big data, the following steps are performed in an iterative way

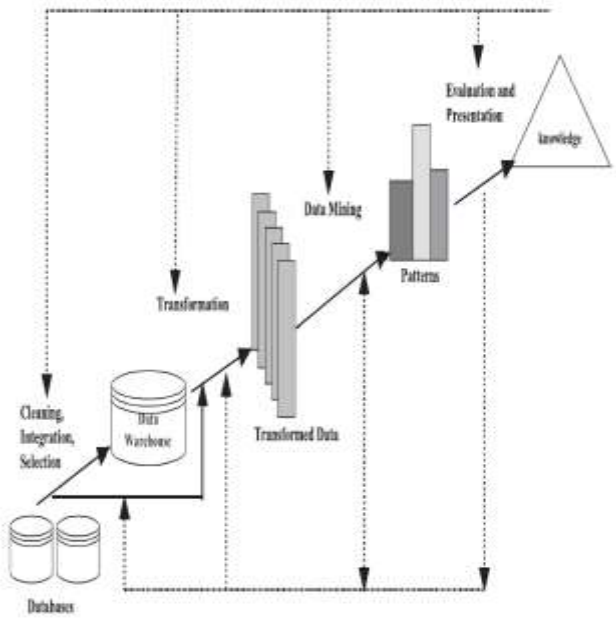

Fig.1 overview of data mining process

1) *Data preprocessing:* In the Data preprocessing basic operations are like data selection, data cleaning and data integration.
2) *Data transformation*: The aim of data transformation is to transform data into different forms which are suitable for the mining process, that is, to find useful information to represent the data.
3) *Data mining:* Data mining is a most important process where intelligent methods are use to extract some useful data from big data (e.g. association rules, clusters, classification rules, etc.)
4) *Pattern evaluation and presentation*: In this the basic operations is to identify the useful patterns which represent knowledge, and presenting the mined knowledge in an easy-to-understand fashion [9].

TABLE I
Comparison Table between Big Data and Data Mining

Big data	Data mining
Big data is a term for huge amount of data set.	Data mining is the process of separating useful information from the big data.
Big data is the asset	Data mining is the handler which provide beneficial result.
Big data" varies depending on the capabilities of the organization managing or handling the set, and on the capabilities of the applications that are traditionally used to process and analyze the data.	Data mining points to the operation that involve relatively sophisticated search operation

C. Privacy Protection Technologies

1) Data Distortion Techniques:

In order to protect privacy in released database, people proposed a lot of effective data mining technology to hide sensitive information. The purpose of privacy protection is as follow.

Hide sensitive information contained in the original data;

Data between hidden and original have the same characteristics.

Get the same data accuracy as original data set. Privacy protection data mining algorithms, such as discovery of association rule, classification, clustering, need choose data to modify or purify, and the choice of purified data is a NP hard problem. To deal with this complex problem, the methods of distortion, such as random perturbation, blocking, and condensation, are used.

- *Association Rules Mining based on Perturbation:*

Statistical significance is used to judge rules emergence in data set, and support and confidence as a metric. All association rules are greater than or equal to user defined support and confidence, but from point of view of user that some rules are sensitive, some are not. Association rules hiding technique is to use the following method to pure the original data set. All sensitive rules can only appear on original data mining, at the same time the confidence and support is not allowed to appear when the data set is purified. That non-sensitive rules can be scoop out in the original data set can also be dug on the clean data set in the same support and confidence. That sensitive rules cannot be dug out in the original data set cannot be dug out in the purification data set at the same support and confidence. The optimal purification is NP hard [10] for association rules mining to hide large item sets. The approaches adopted in this work was either to prevent the sensitive rules from being generated by hiding the frequent item sets from which they are derived, or to minimize the confidence of the sensitive rules by bringing it below a user-specified threshold. These two approaches prompt to the generation of three strategies for hiding sensitive rules. The important things to note related to these three strategies were the possibility for both a 1-value in the binary database to turn into a 0-value and a 0-value to turn into a 1-value. This extensibility in data modification had the side-effect that apart from non-sensitive association rules that were becoming hidden, and a non-frequent rule could become a frequent one.

- *Mining Association Rules using Block:*

Another perturbation for association rules of data modification method is the data block [11]. Blocking method replace a property value of data items with mark of question, That using unknown value instead of actual values rather than using false value instead of actual values is very popular in medicine. A method of association rules mining using blocking, which appropriate changes the definition on the minimum support, replace with minimum support interval and minimum confidence, and replace with confidence interval as in [10]. These are the sensitive rules below the middle of support interval, or confidence of sensitive rules below the middle of confidence interval. Whether 1-value or 0-value should be mapped to a question mark, otherwise original value of question mark will be exposed. Detailed description of effectiveness of blocking method this method is the reconstruction of the text using the rules of disturbance [12].

- *Classification Rule Mining Based on Block:*

It provides a new framework for combining classification rule analysis and parsimonious down grading, data administrator has as a goal to block values for class label. By doing this, the information receiver, will be unable to build informative models for the data that is not downgraded. Parsimonious downgrading is a framework for formalizing the phenomenon of trimming out information from a data set for downgrading information. In parsimonious downgrading a cost measure is assigned to the potential downgraded information that it is not sent to low. The main aim consider in this work is to find out whether the loss of functionality associated with not downgrading the data, is worth the extra confidentiality.

2) Distributed Privacy Preserving Mining

In the privacy preserving data mining environment, the people made a lot of encryption based approach to solve the problem with the following features. Two or more parties mine their data on the basis cooperation, but none of them consent to display their data. This is a secure multiparty computation, SMC, problems under distributed environment, which focuses on how to convert various data mining methods to secure multiparty computation issues, such as data classification, data clustering, association rules mining, data generalization, and data aggregation. Describe the Secure multiparty computation methods including, the secure set union, the secure sum, the scalar product, and secure size of set intersection the. Let us discuss the distributed association rule mining.

- *Vertically Partitioned Data Association Rules Mining:*

Vertically partitioned data set different attributes for each item in different sites. Mining private association rules from vertically partitioned data by finding the support count of an item set. If the support determines total number of such an item set can be securely calculated, then we can check if the support is greater than the

threshold, and decide whether the item set is frequent. Each party involved in the calculation by the sub-item set composed of a vector, and calculate the number of an item set support is the key to computing vector dot product. Therefore, if the dot product can be secure computing, supports can also be calculated in security.

- *Horizontally Partitioned Data Association Rules Mining:*

In horizontally partitioned database the transactions are distributed among *n* sites. The overall support count of an item set is the addition of all the local support counts. An item set X is globally supported if the global support count of X is bigger than s% of the total transaction database size.

3) Reconstructed Technology

Much privacy preserving data mining technology proposed recently use data perturbation or reconstruction in data convergence layer. We studied to construct a decision tree classifier using the individual records value of perturbation as training data [10]. Since original values of individual records can not estimate accurately, the author considers estimating original distribution accurately. In order to reconstruct the original distribution, Bayesian method is considered. It improves the Bayesian reconstruction process by using EM algorithm in the distributed data. More precisely, the author prove that the EM algorithm dictates the maximum estimated fairly as the original data on the distribution of disruption, but also proved that when large amounts of data can be obtained , EM algorithm can estimate the original distribution robust. When background was known by data miner through the reconstruction distribution then the estimation of privacy will decrease [10].

4) Anonymous Privacy Protection

Anonymous release chose to publish the raw data. In order to achieve privacy protection, sensitive data does not publish or release sensitive data with lower accuracy. The current study focused on data anonymity technical, namely, Make trade-offs between the privacy disclosure risks and data utility, which selective release of sensitive data and information that may be disclosed sensitive data, but to ensure that sensitive data and privacy disclosure risk within the tolerable range. Data anonymity focuses on two aspects: one of the principles is to design better anonymity methods, so that the data released following this principle can not only better protect privacy, but also has great practical utility. The other hand is to design more efficient anonymity algorithms for specific anonymous principle. With the research depth of anonymity, how to achieve practical application of anonymity data becomes the focus of research. Samarati and Sweeney proposed *k*-anonymity principle which requires that each record in the table released cannot distinguish from other *k*-1 records. We call *k* records, cannot be distinguished, an equivalent class. Here cannot be distinguished in terms of non-sensitive attributes. In general, greater *k* values bring about better degree of privacy protection, but the information loss increase.

It does not make any constraint for sensitive data that is flaw of *k*-anonymity. An attacker can use protocol against attack and background knowledge attack to identify sensitive data or personal relationships, which leading to privacy leaks. (α, *k*)-anonymity make a improvement on this basis, which not only ensure that *k*-anonymity publishing is satisfied but also ensure that each records related any attribute value in each released equivalence class is not higher than the percentage of alpha . Generally *k*-anonymity, *l*-diversity, *t*-closeness are the methods of publishing data [13] and other anonymous release, use generalization techniques, which reduce accuracy and data utility largely. In terms of data collection, if disclosure risks of all sensitive data, in data set D released by data owners, are less than the threshold alpha, alpha belongs to [0,1], called the disclosure risk of data set as alpha. Such as static data release *l*-diversity ensures that disclosure risk of published data sets is less than 1/*l*, and dynamic data publishing principles m-invariance ensure that the disclosure risk of published data sets is less than 1/*m*.

5) Evaluation of Privacy Protection Algorithms

An important aspect on privacy preserving data mining algorithms and tools for developing and evaluating is to select the appropriate evaluation criteria, but the reality is not a privacy protection data mining algorithms under a variety of indicators to be better than other algorithms, in general, an algorithm may be practical in terms of performance or a little better than others. It is very important to provide users with a set of metrics to enable them to choose the best appropriate algorithms for data privacy preserving. Next, we make simple introduce for performance of algorithm, utility of data, privacy degree of protection and the difficulty of data mining.

- *Algorithm Performance:*

We can see that the algorithm with $O(n^2)$ complexity polynomial time is more efficiency than those with $O(e^n)$ index of complexity. An alternative approach would be to evaluate the time requirements in terms of the average number of operations, needed to reduce the frequency of specific sensitive information appearance below a specified threshold. This values, perhaps, does not provide an absolute measure, but it can be considered in order to perform a fast comparison among *different* algorithms.

- *Data Utility:*

It is a very important issue for utility of data privacy protection. In order to hide sensitive information, false information should insert the database, or block data values. Although sample Techniques do not modify the information stored in the database, but that, since their information is incomplete, still reduces data utility. More changes to the database, less data utility of the database. So estimated parameters of data utility is data information loss applied privacy protection. Of course, the estimate of information loss related with *the* specific data mining algorithms.

- *Degree of Privacy Protection:*

Privacy protection policy is to protect the information downgrade to a certain threshold, but hidden information can be derived out by some uncertainty. The uncertainty reconstructed by hidden information can evaluate sanitation algorithm. A solution can set a maximum on perturbation information from execution point of view, and then consider achieve the degree of uncertainty by constraints of different purification method. We hope that an algorithm can achieve the greatest uncertainty, and better than all the other algorithms.

- *Difficulty of Different Data Mining:*

In order to provide the full estimation on purification method, we need to measure difficulty of data mining algorithms which is different with purification method, and this called parameter horizontal difficulty. This estimation of parameter need consider the classification of data mining which is very important on the test. Alternatively, we may need to develop a formal framework that upon testing of a sanitization algorithm against pre-selected data sets, we can transitively prove privacy assurance for the whole class of sanitization algorithms [14].

V. CONCLUSION

Big data is the collection of huge amount of data and data mining is the process of extracting some useful information from big data. In this paper first we overlook which security problems and privacy problems are occurred in big data. Security and Privacy in big data mining is most important because of sensitive information may have harm full effect on someone's life, so it is most important to provide security and privacy to our sensitive information. For this purpose we discuss some privacy preserving data mining techniques.

REFERENCES

[1] M. Prakash, G. Singaravel, "A New Model for Privacy Preserving Sensitive Data Mining", in proceedings of ICCCNT Coimbatore, India, IEEE 2012.
[2] Chanchal Yadav, Shullang Wang, Manoj Kumar, (2013) "Algorithm and Approaches to handle large Data- A Survey", IJCSN, 2(3), ISSN: 2277-5420(online), pp2277-5420
[3] Wei Fan, Albert Bifet, "Mining Big Data: Current Status, and Forecast to the Future", SIGKDD Explorations, Volume 14, Issue 2
[4] Priya P. Sharma, Chandrakant P. Navdeti, "Securing Big Data Hadoop: A Review of Security Issues, Threats and Solution", IJCSIT, Vol 5(2), 2014, 2126-2131
[5] Richa Gupta, Sunny Gupta, Anuradha Singhal, "Big Data: Overview", IJCTT, Vol 9, Number 5, March 2014
[6] Z. Ferdousi, A. Maeda, "Unsupervised outlier detection in time series data", 22nd International Conference on Data Engineering Workshops, pp. 51-56, 2006
[7] Dileep Kumar Singh and Vishnu Swaroop," Data Security and Privacy in Data Mining: Research Issues & Preparation", International Journal of Computer Trends and Technology- volume4Issue2- 2013
[8] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques. San Mateo, CA, USA: Morgan Kaufmann, 2006.
[9] LEI XU, CHUNXIAO JIANG, JIAN WANG, JIAN YUAN, AND YONG REN,"Information security in big data: Privacy and Data mining", 2014 IEEE
[10] L. Chang, and I. Moskowitz, "An Integrated Framework for Database Privacy Protection," Data and Application Security, Springer Boston, 2002, pp. 161-172.
[11] E. Dasseni, V.S. Verykios, A.K. Elmagarmid, and E. Bertino, "Hiding association rules by using confidence and support," Lecture Notes In Computer Science, vol. 2137, 2001, pp. 369-383.
[12] B.J. Ramaiah, A.R.M. Reddy, and M.K. Kumari, "Parallel privacy preserving association rule mining on pc clusters," 2009 IEEE International Advance Computing Conference, Inst. of Elec. 2009, pp. 1538-1542, doi: 10.1109/IADCC.2009.4809247.
[13] R. Wong, J. Li, A. Fu, and K. Wang, "(α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing," Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM,
2006, pp. 754-759.
[14] Xinjun Qi , Mingkui Zong," An Overview of Privacy Preserving Data Mining", 2011 International Conference on Environmental Science and Engineering (ICESE 2011)