International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 1 Issue: 12 961 - 964

A Study of Various Privacy Preserving Data Mining Algorithms for Datasets

Halkar Rachappa
Asst. Prof. & Head Dept. of Computer Science
Govt. Commerce & Management College
BALLARI (Karnataka)
e-mail :halkarrg@gmail.com

Abstract:- Privacy, security and accuracy are the major issues to be concern in field of data mining data mining when data is shared. A number of data mining algorithms are already introduced for big data when we are talking about Privacy preserving data mining. These algorithms categories data into groups. Further these groups can be used for extract useful information. Such kind of data is used in surveys, calculations etc. An election data can be considered as an example for such kind of groups. The groups are made in a, b, ab, cd category. Each group is not aware that which group has which data. In the end using data mining algorithms the desired data can be extracted.

Keywords: *PPDM, Privacy Preserving, Randomize response technique.*

*****

1. INTRODUCTION

Data Mining can be referred to as extracting the useful information from large amount of data. The goal of data mining is to improve the quality of the interaction between the organization and their customers.

According to Giudici[1], "data mining is the process of selection, exploration, and modeling of large quantities of data to discover regularities or relations that are at first unknown with the aim of obtaining clear and useful results for the owner of the database."

In earlier times, due to lack of existence of information systems companies were unable to store the data and to analyze them. Recently a new pattern of looking into data and extrapolating patterns has evolved and offered at many levels to organizations. Data mining usually denotes applications, under human control, of low-level data mining methods [2]. Large scale automated search and interpretation of discovered regularities belong to Knowledge Discovery in databases (KDD), and are typically not considered part of data mining. KDD concerns itself with knowledge discovery processes applied to databases. KDD deals with ready data, available in all domains of science and in applied domains of retailing, planning, control, etc. Typically, KDD has to deal with inconclusive data, noisy data, and sparse data [3]. Thus, KDD refers to the overall process of discovering useful knowledge from data while data mining refers to the application of algorithms for extracting patterns from data. Figure 1.1 shows data mining as a step in an iterative knowledge discovery process [4].

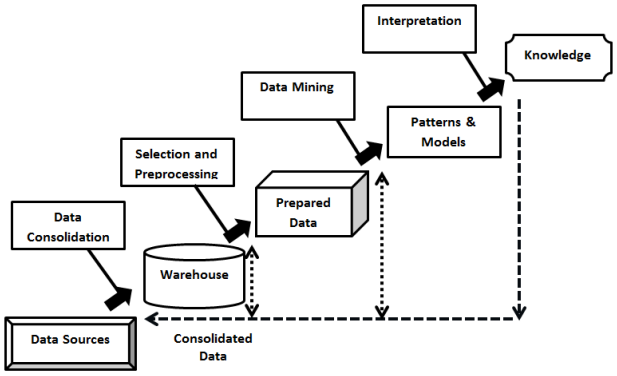

Figure 1.1 Data mining is the core of Knowledge discovery process.

Data mining is ready for application in the business community because it is supported by three mature technologies namely: massive data collection, powerful multiprocessor computers and data mining algorithms [5]. Data mining algorithms embody techniques that have existed for at least 10 years, but have only recently been implemented as mature, reliable, understandable tools that consistently outperform older statistical methods [6].

a. The Data Mining Process

The goal of identifying and utilizing information hidden in data has three requirements:

- The information contained in the integrated data must be extracted.
- The information obtained must be organized to enable decision-making.

The data mining process is composed of a series of four steps [7]. This consists of transforming the already summarized data found in a data warehouse into information producing useful results through:

- Data selection
- Data transformation

- Mining the Data
- Interpretation of Results

Data selection consists of gathering the data for analysis. Data transformation will then convert appropriate data to a particular format. As shown in Figure 1.2, the data-mining tool will extract the relevant information from the data warehouse environment. In order for the data-mining tool to work, the sub-processes of data selection and transformation must take place prior to data mining.

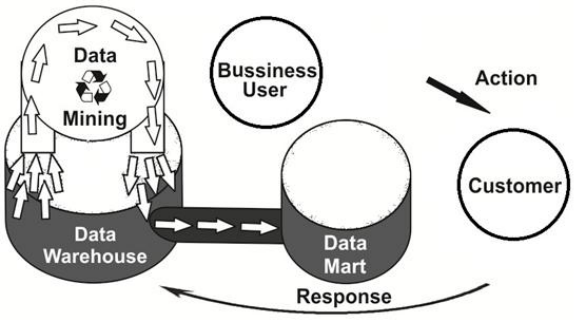

Figure 1.2: Data Mining Process

According to [8] the following terms explain data mining:

- Data, Information, and Knowledge
- Need of data mining
- Working of data mining

b. Privacy issues in data mining

It is well known that data mining is a powerful data analysis tool enabling discovery of useful patterns in several applications. The availability of large data warehouses and associated technologies reveal that the usefulness of data requires preservation of individual key attributes such as patient's condition information, customer preferences, personal background information etc. If the original data is directly released to the miner, it will inevitably produce private information of the customer. Therefore, how to do mining without sacrificing privacy is the main issue in data mining. Privacy preserving data mining (PPDM) deals with this issue. Currently, there are many approaches of privacy preserving data mining to transform the original data. The methods of privacy preserving data mining are evaluated based on applicability, privacy protection metric, the accuracy, computation, etc. [13]

2. LITERATURE SURVEY

This survey included various types of techniques used in data mining. Mainly the techniques are divided in two sections:

- Classical Techniques: This technique includes Statistics, Neighborhoods and Clustering
- Next Generation Techniques: This technique includes Trees, Networks and Rules [15]

2.1 Classical Techniques

Classical and next generation techniques have emerged with the development and maturity of the data mining. The next generation demand that the captured data must be integrated into organization-wide views instead of specific views.

2.1.1 Statistics

By the definition statistical techniques are not data mining. Because statistical techniques were used even before the data mining became important. However statistical techniques used to discover patterns and build predictive models. So user has a choice to solve the problem with statistical techniques or with other data mining techniques. Therefore it is important to have some idea of how statistical techniques work and how they can be applied.

2.2 Privacy Preserving Data Mining

Data mining and knowledge discovery in databases are two new research areas that investigate the automatic extraction of previously unknown patterns from large amounts of data. Recent advances in data collection, data dissemination and related technologies have inaugurated a new era of research where existing data mining algorithms should also be reconsidered from various point of views, such as privacy preservation. It is well documented that, this new without limits, explosion of new information through the Internet and other media, has reached to a point where threats against the privacy are very common on a daily basis and they deserve serious thinking. Privacy preserving data mining is a novel research direction in data mining and statistical databases, where data mining algorithms are analyzed for the side-effects they incur in data privacy.

2.2.1 Framework for PPDM

In data mining or knowledge discovery from databases (KDD) process the data (mostly transactional) is collected by single/multiple organization/s and stored at respective databases. Then, it is transformed to a format suitable for analytical purposes, stored in large data warehouse/s and then data mining algorithms are applied on it for the generation of information/knowledge. With the intent of protecting privacy the model has to be evolved. Privacy constraints cannot be applied at one step; it needs to be kept in mind along with the data mining process all the way from data collection to the generation of information/knowledge. There are three levels of privacy concerns. At level 1, the raw data collected from a single or multiple databases or

even data marts is transformed into a format that is well suited for analytical purposes. Even at this stage, privacy concerns are needed to be taken care of. Researchers have applied various techniques at this stage but most of them deal with making the raw data suitable for analysis.

2.2.2 Classification of PPDM

The work in PPDM can be classified according to various categories.

Data Distribution- The PPDM algorithms can be first divided into two major categories, centralized and distributed data, based on the distribution of data. In a centralized database environment, all data are stored in a single database; while, in a distributed database environment, data are stored in various databases. Distributed database scenario can be further classified into horizontal and vertical data distributions. Horizontal distributions refer to the cases where various records of the same data attributes reside in various places. While in a vertical data distribution, various attributes of the same record of data reside in various places. Earlier research has been predominately focused on dealing with privacy preservation in a centralized database. The difficulties of applying PPDM algorithms to a distributed database can be attributed to: initial, the data owners have privacy concerns so they may not willing to release their own data for others; second, even if they are willing to share data, the communication cost between the sites is too expensive [25].

Data Mining Tasks / Algorithms - Currently, the PPDM algorithms are mainly used on the tasks of classification, association rule and clustering. Association analysis involves the discovery of associated rules, showing attribute value and conditions that occur frequently in a given set of data. Classification is the process of finding a set of models (or functions) that describe and distinguish data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. Clustering Analysis concerns the problem of decomposing or partitioning a data set (usually multivariate) into groups so that the points in one group are similar to each other and are as different as possible from the points in other groups.

Privacy Preservation Techniques - PPDM algorithms can further be divided according to privacy preservation techniques used. Four techniques – sanitation, blocking, distort, and generalization -- have been used to hide data items for a centralized data distribution. The idea behind data sanitation is to remove or modify items in a database to reduce the support of some frequently used item sets such that sensitive patterns cannot be mined. The blocking approach replaces certain attributes of the data with a question mark. In this regard, the minimum support and confidence level will be altered into a minimum interval. As long as the support and/or the confidence of a sensitive rule lie below the middle in these two ranges, the confidentiality of data is expected to be protected. Also known as data perturbation or data randomization, data distortion protects privacy for individual data records through modification of its original data, in which the original distribution of the data is reconstructed from the randomized data. These techniques aim to design distortion methods after which the true value of any individual record is difficult to ascertain, but “global” properties of the data remain largely unchanged. Generalization transforms and replaces each record value with a corresponding generalized value.

2.2.3 Techniques of PPDM

Most methods for privacy computations use some form of transformation on the data in order to perform the privacy preservation. Typically, such methods reduce the granularity of representation in order to reduce the privacy. This reduction in granularity results in some loss of effectiveness of data management or mining algorithms. This is the natural trade-off between information loss and privacy. Some examples of such technique as described in [26] are:

2.2.3.1 Randomization method

The randomization technique uses data distortion methods in order to create private representations of the records. In this white noise is added to the data in order to mask the attribute values of records In most cases, the individual records cannot be recovered, but only aggregate distributions can be recovered. These aggregate distributions can be used for data mining purposes. Data mining techniques can be developed in order to work with these aggregate distributions. Two kinds of perturbation are possible with the randomization method:

- Additive Perturbation - In this case, randomized noise is added to the data records. The overall data distributions can be recovered from the randomized records. Data mining and management algorithms are designed to work with these data distributions.
- Multiplicative Perturbation- In this case, the random projection or random rotation techniques are used in order to perturb the records.

2.3 Limitations and Motivation

From the literature survey it is concluded that when only ID3 algorithm is used on dataset in privacy preserving data mining then the privacy is not preserved.

This motivates to increase the privacy of datasets by using some more data mining algorithms with ID3

algorithm. With ID3 algorithm some randomization and multi group techniques can be applied on datasets.

2.4 Problem Statement

2. The drawback of the previous work was that it was not checking the group performance at every step.
3. The privacy level was not much higher when privacy preserving data mining using ID3 algorithm is done and when it is applied on three groups datasets. So this thesis propose to work for increasing the privacy at higher level using four group dataset.

3 PROPOSED APPROACH

This work has proposed an approach for privacy preserving data mining using randomized response technique. This work uses ID3 and CART algorithm to enhance the privacy of the secret data. The problem with the previous work for three groups of data sets using ID3 algorithm was that it was not checking the group performance at every step and the privacy level was not very high [31]. The proposed work increases the level of privacy by using ID3 and CART algorithms. Previous work was giving an overall result whereas this work is implementing it in step by step manner.

CONCLUSION

A number of data mining algorithm such as randomized technique, ID3 algorithm studied. Such kind of algorithms can preserve privacy and security both for survey or huge data. In future these techniques can be applied on different behavior of data stored at multiple places.

References

[1] Giudici, P, "Applied Data-Mining: Statistical Methods for Business and Industry." John Wiley and Sons (2003) West Sussex, England.

[2] American Association for Artificial Intelligence Advances in Knowledge Discovery and Data Mining. Press/ The MIT Press. 1996.

[3] Edelstein, Herb. Data Mining News "Two Crows Releases 1999 Technology Report". Volume 2, number 18. 10 May 1999.

[4] Y. Ramamohan, K. Vasantharao, C. Kalyana Chakravarti, A.S.K.Ratnam, "A study of data mining tools in knowledge discovery process", IJSCE, Volume-2, Issue-3, July 2012.

[5] http://www.pilotsw.com/dmpaper/dmindex.htm; "An Introduction to Data Mining". Pilot Software Whitepaper. Pilot Software. 1998.

[6] "Data Mining: An Introduction", SPSS Whitepaper. SPSS. 2000.

[7] Walter Alberto, Data Mining Industry: Emerging Trends and New Opportunities: MIT, 2000 Springer book.

[8] http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm.

[9] Deependra Dwivedi, "Study Analysis of data mining Algorithms: case study" Researcher. 2012;4(2):16-19] 2012, http://www.sciencepub.net.

[10] Betts, M., "The Almanac: Hot Tech," Computer World (Nov. 17, 2003).

[11] Greg., "E-Voting Milestones," IEEE Security and Privacy, Gayatri Nayak, Swagatika Devi, "A Survey On Privacy Preserving Data Mining Approaches And Techniques" , IJEST, Vol. 3 No. 3 March 2011

[12] Reena Hooda, Nasib S. Gill "Applications And Issues of Data Mining" IJRIM Volume 2, Issue 3 March 2012

[13] G.Rama Krishna, G.V.Ajaresh, I.Jaya Kumar Naik, Parshu Ram Dhungyel, D.Karuna Prasad "A New Approach to Maintain Privacy And Accuracy In Classification Data Mining" IJCSET Volume 2, Issue 1, January 2012 Y.

[14] Ramamohan, K. Vasantharao, C. Kalyana Chakravarti, A.S.K.Ratnam, "A study of data mining tools in knowledge discovery process", IJSCE, Volume-2, Issue-3, July 2012.

[15] Gayatri Nayak, Swagatika Devi, "A Survey On Privacy Preserving Data Mining Approaches And Techniques", IJEST, Vol. 3 No. 3 March 2011.

[16] An Overview of Data Mining Techniques Excerpted from the book by Alex Berson, Stephen Smith, and Kurt Thearling. Page no 2.

[17] V.Ganti, J.Gehrke and R.Ramakrishnan, "CACTUS – Clustering Categorical Data Using Summaries", in Proceedings of ACM SIGKDD, 1999.

[18] Md. Zahidul Islam and Ljiljana Brankovicv "DETECTIVE: A Decision Tree Based Categorical Value Clustering and Perturbation Technique for Preserving Privacy in Data Mining"

[19] Berkhin Pavel, "A Survey of Clustering Data Mining Techniques", Springer Berlin Heidelberg,2006.