ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 4, April 2013

Privacy Preserving Data Mining

Seema Kedar[1], Sneha Dhawale[2], Wankhade Vaibhav[3],
Pavan Kadam[4] , Siddharth Wani[5] , Pavan Ingale[6]
HOD, Information Technology Department,Pune. India[1]
Student, Information Technology Department,Pune. India[2,3,4,5,6]

Abstract: There is a tremendous increase in the research of data mining. Data mining is the process of extraction of data from large database. One of the most important topics in research community is Privacy preserving data mining (PPDM). It is essential to maintain a ratio between privacy protection and knowledge discovery. The goal is to hide sensitive item sets so that the adviser cannot extract the modified database. To solve such problems there are some algorithms presented by various authors worldwide. The primary goal of this survey paper is to understand the existing privacy preserving data mining techniques and to achieve efficiency.

Keywords: Privacy Preserving, Utility Mining, Sanitization, Data Mining

I. INTRODUCTION

There are two main approaches of previous work in privacy preserving data mining. Perturbing the data values for preservation of customer privacy is the first approach. Cryptographic tools to build data mining models are the other approach. The following section contains some of the recent researches in this field.

As the data 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 efficiency of proposed algorithm is the second characteristic we analysed. The reduction of the confidence of the sensitive rules without change in the support of the sensitive item is the approach for 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 among 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.

II. RELATED WORK

Based on the concept of roles and permissions in the market, there are a number of existing systems on which we will take a brief look on. Previously two approaches of privacy preserving data mining are defined. In the first one, the aim is to preserve customer privacy by perturbing the data values and the other approach uses cryptographic tools to build various models for data mining process. In this section, some of the recent researches are described.

A. *Hiding Association Rules by Using Confidence and Support*

Author of the paper suggested some rules for hiding sensitivity by changing the support and the confidence of the association rule or frequent item set as data mining mainly deals with generation of association rules. In order to hide an association rule a new concept of '*not altering the support'* of the sensitive item(s) has been proposed in this work.

Advantages

- First advantage of proposed algorithm is that support for the sensitive item is unchanged. Instead, only the position of the sensitive itemset is changed.
- The second advantage is the proposed approach uses a different approach for modifying the database transactions so that the confidence of the sensitive rules can be reduced but without changing the support of the sensitive item.

Disadvantage

- One of the main disadvantages of the existing approaches is that the approach in tries to hide every

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 4, April 2013

single rule from a given set of rules without checking if some of the rules could be pruned after modification of some transactions from the set of all transactions. This approach hides rules having sensitive items either in the right side or in the left side.

B. *Privacy Preserving Clustering By Data Transformation*

Preserving the privacy of individuals when data are shared for clustering was a complex problem. The challenge was how to protect the underlying data values subjected to clustering without jeopardizing the similarity between objects under analysis. Stanley R. M. Oliveira, and Osmar R. Zaiane [5] revisited a family of geometric data transformation methods (GDTMs) that distort numerical attributes by scaling, rotations, translations or by the combination of all above transformations. This method was designed to specify privacy-preserving clustering, in context where data owners must meet privacy requirements as well as guarantee valid clustering results. Authors also provided a particularized, broad and advanced picture of methods for privacy-preserving clustering by data transformation.

Advantages

- The geometric data transformation methods (GDTMs) that distort confidential numerical attributes in order to meet privacy protection in clustering analysis.
- End users are able to use their own tools so that the constraint for privacy has to be applied before the mining process on the data by data transformation.
- Data owners must not only meet privacy requirements but also guarantee valid clustering results.

Disadvantages

- One major disadvantage is that the privacy preservation of individuals when data is shared for clustering is very complex.
- The protection of the underlying data values subjected to clustering without jeopardizing the similarity between objects under analysis is hard to achieve.

C. *Perturbation Based Privacy Preserving Data Mining For Real World Data*

The perturbation method has been extensively studied for privacy preserving data mining. In this method, random noise from a known distribution is added to the privacy sensitive data before the data is sent to the miner for data mining. Consequently, the data miner rebuilds an approximation to the original data distribution from the perturbed data and uses the reconstructed distribution for data mining purposes. Different individuals may have different approaches towards privacy based on cultures and customs. Unfortunately, recent privacy preserving data mining techniques based on perturbation do not allow the individuals to choose their desired privacy levels. This was a drawback as privacy was a personal choice. Li Liu *et al*. [4] proposed an individually adaptable perturbation model, which enabled the individuals to choose their own privacy levels. Based on their experiments, they have suggested a simple but effective and yet efficient technique to build data mining models from perturbed data.

Advantages

- Simple and efficient technique for building data mining models from perturbed data.
- As the distribution of the added noise is known, the data miner could rebuild the original distribution using various statistical methods and mine the rebuilt data.

Disadvantages

- Recent privacy preserving data mining techniques based on perturbation do not allow the individuals to choose their desired privacy levels.
- As the noise is added, information loss versus preservation of privacy is always a trade off in the perturbation based approaches.

III. RECENT TECHNOLOGY

A. GENETIC ALGORITHM

In Genetic Algorithms, a population consists of a group of individuals called chromosomes that represent a complete solution to a distinct dilemma. Every chromosome represents a sequence of 0s or 1s. The first set of the population is set of individuals that are randomly generated. There are two methods to generate new population: steady state Genetic Algorithm and generational Genetic Algorithm. The steady-state Genetic Algorithm replaces one or two members of the population; whereas the generational Genetic Algorithm replaces all of them at each generation of evolution. In this work a generational Genetic Algorithm is adopted as population replacement method. In this method tried to keep a certain number of the best individuals from each generation and copies them to the new generation (this approach known as elitism).

Advantages

- It provides a very high security of database as well as it keeps the utility and assurance of mined rules at highest level.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 4, April 2013

- Here a new multi-objective method for hiding sensitive association rules based on the concept of genetic algorithms is used.

Disadvantages

- One big issue is the risk of information leakage and its confidence.
- It emphasizes on alteration of original data in such a way that it would be impractical for the opponent to mine the sensitive rules from the modified data set.

B. PPDM

The privacy-preserving data mining (PPDM) has become an important issue in recent years. Tzung-Pei Hong *et al.* [8] proposed a paper, a greedy-based approach for hiding sensitive item sets by inserting dummy transactions. That computes the maximal number of transactions to be inserted into the original database for totally hiding sensitive item sets. Experimental results were also performed to evaluate the performance of that proposed approach. In recent years, the wide availability of personal data has made the problem of Privacy Preserving Data Mining an important one.

The increasing ability to track and collect large amounts of data with the use of current hardware technology has lead to an interest in the development of data mining algorithms which preserve user privacy. A number of methods have recently been proposed for privacy preserving data mining of multidimensional data records.

Advantages

- PPDM is very advantageous in development of various data mining techniques.
- It allows sharing of large amount of privacy-sensitive data for analysis purposes**.**
- It has a ability to track and collect large amounts of data with the use of current hardware technology.

Disadvantage

- One of the major problems of privacy preserving data mining is the abundant availability of personal data.

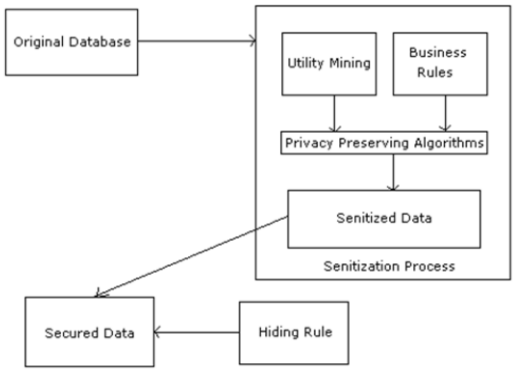

Fig. 1 **:** Block diagram of Privacy Preserving Data Mining Technique

Fig. 1 shows a block diagram of Privacy preserving data mining technique.

IV. CONCLUSION

The above discussed privacy preserving data mining techniques are remarkably good, but there is always extent for more enhancements. This survey paper on PPDM can be helpful for finding the loopholes and drawbacks of existing data mining techniques. This survey ensures efficient privacy preserving of data. The use of existing algorithms works towards the direction to reduce the impact of PPDM on the source database. A comparative study all these systems would definitely help in developing a new system that combines all the advantages and overcomes the drawbacks of these systems.

ACKNOWLEDGMENT

The authors wish to thank Mrs. Seema Kedar (HOD, Information Technology, RSCOE, Pune) for their valuable guidance and co-operation. At the outset the authors thank to Dr. D.T.Bormane (Principal, RSCOE, Pune, India) for encouragement and providing the opportunity and facilities to carry out this work.

REFERENCES

[1] Yehuda Lindell and Benny Pinkas, "Secure Multiparty Computation for Privacy-Preserving Data Mining," The Journal of Privacy and Con¯dentiality, vol. 1, no. 1, pp. 59-98, 2009.
[2] Marina Blanton, "Achieving Full Security in Privacy-Preserving Data Mining," In Proc. of the 2011 ieee third international conference on social computing (socialcom) Privacy, security, risk and trust (passat), Dame, IN, pp. 925-934, Oct 2011.
[3] Bin Yang, Hiroshi Nakagawa, Issei Sato, and Jun Sakuma, "Collusion-Resistant Privacy-Preserving Data Mining," In Proc. of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp. 483–492, 2010.
[4] Li Liu, Murat Kantarcioglu, and Bhavani Thuraisingham, "The applicability of the perturbation based privacy preserving data mining for real-world data," Journal of Data & Knowledge Engineering, vol. 65, pp. 5–21, 2008.
[5] Stanley R. M. Oliveira, and Osmar R. Zaiane, "Revisiting Privacy Preserving Clustering by Data Transformation," Journal of Information and Data Management, vol. 1, no. 1, 2010.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

6] Dhyanendra Jain, Amit sinhal, Neetesh Gupta, Priusha Narwariya, Deepika Saraswat, and Amit Pandey, "Hiding Sensitive Association Rules without Altering the Support of Sensitive Item(s),"International Journal of Artificial Intelligence & Applications (IJAIA), Vol. 3, No. 2, pp. 75-84, Mar 2012.
[7] Elena Dasseni, Vassilios S. Verkios, Ahmed K. Elmagarmid, and Elisa Bertino, "Hiding Association Rules by Using Confidence and Support," Computer Science Technical Reports, 2000.
[8] Tzung-Pei Hong, Chun-Wei Lin, Chia-Ching Chang, and Shyue-Liang Wang, "Hiding Sensitive Itemsets by Inserting Dummy Transactions," In Proc. of the IEEE International Conference on Granular Computing (GrC), Kaohsiung, Taiwan, pp. 246-249, 2011.
[9] Dr. K. Duraiswamy, Dr. D. Manjula, and N. Maheswari, "Advanced Approach in Sensitive Rule Hiding," Modern Applied Science, vol. 3, no. 2, pp. 98-107, Feb 2009.
[10] Jieh-Shan Yeh and Po-Chiang Hsu, "HHUIF and MSICF: Novel algorithms for privacy preserving utility mining," Jounal of Expert Systems with Applications, vol. 37, pp. 4779–4786, 2010.