A Case Study on Issues In Privacy Preserving Data Mining

Siripuri Kiran[1], Ajmera Rajesh[2]
Assistant Professor[1], Academic Consultant[2]
[1, 2] Kakatiya Institute Of Technology and Sciences, Warangal, India.
[2]Ku College Of Engineering And Technology, Kakatiya University, Warangal. India.

Abstract- *The development in data mining technology brings serious threat to the individual information. The objective of privacy preserving data mining (PPDM) is to safeguard the sensitive information contained in the data. The unwanted disclosure of the sensitive information may happen during the process of data mining results. In this study we identify four different types of users involved in mining application i.e. data source provider, data receiver, data explorer and determiner decision maker. We would like to provide useful insights into the study of privacy preserving data mining. This paper presents a comprehensive noise addition technique for protecting individual privacy in a data set used for classification, while maintaining the data quality. We add noise to all attributes, both numerical and categorical, and both to class and non-class, in such a way so that the original patterns are preserved in a perturbed data set. Our technique is also capable of incorporating previously proposed noise addition techniques that maintain the statistical parameters of the data set, including correlations among attributes. Thus the perturbed data set may be used not only for classification but also for statistical analysis.*

Keywords- Data Mining, Security, Issues & Remedies, Privacy, Preservation, development, technology, information, process.

I. INTRODUCTION

Data mining is frequently characterized as the way toward finding important, new correlation patterns and trends through non-trifling extraction of certain, already obscure data from extensive measure of data put away in repositories utilizing design acknowledgment and additionally statistical and mathematical techniques.

A Structured Query Language (SQL) is usually stated or written to access a specific data while data miners might not even be exactly sure of what they need. So, the result of a SQL query is usually a part of the database; whereas the result of a data mining query is an analysis of full contents of the database. Data mining tasks can be classified as follows:

1) Association rule mining or market basket analysis
2) Classification and prediction
3) Cluster analysis and outlier analysis
4) Web Data mining and search engines'.
5) Evolution analysis

The main focus of this thesis is to obtain secure Clustering results. Achieving accurate clustering results by providing privacy to sensitive data is trivial task. This thesis proposes two approaches for achieving the privacy for sensitive attributes during data mining [1].

Data Mining

Data mining also called as knowledge discovery in databases (KDD). Data Mining is defined as the "process of evaluating interesting, useful and hidden patterns from large volumes of data stores and identifies the relationships among the patterns" [2-4]. Data mining task requires utilities fir statistical data and Artificial Intelligence systems (AI). AI systems includes neural networks and machine learning sometimes one can combine them with database management system for evaluating or analyzing the huge volumes of digital data, which is the derived form of data sets..

Data mining has many applications; those have been listed in the above section. They can broadly categorized in to three area's one is business (insurance company, banking corporation, retail sector), second is science research (astronomy, medicine), and government security (detection of criminals and terrorists).

The large number of organizations, government and private data bases aims to ensure that the individual records are accurate and secure from unauthorized access. The tasks of data mining are targeted towards extracting hidden predictive knowledge about a group rather than the individual.

Figure 1 shows the Data mining process. First, data is collected from various sources in Data selection step. Next, Data will be pre- processed by dealing with null values and unformatted values. Then, Data will be transformed to proper format which is suitable for data mining operation [5]. Now,

Knowledge will be extracted from data store which is nothing but data mining. Evaluation of patterns for decision making takes place finally.

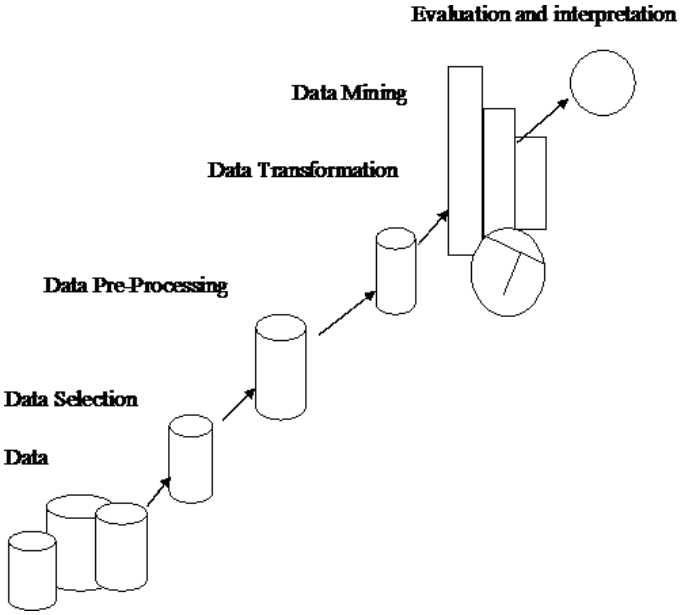

Figure 1.1 Data Mining Process

The specific goal of data mining process is to pull out the hidden information from a data set and change it into a good understandable structure for future use.

II. REVIEW OF LITERATURE

Distributed data mining privacy preserving is a field that requires close cooperation between researchers and practitioners from the fields of cryptography, data mining, public policy and law. Now, the question is how to compute the results without pooling the data in a way that reveals nothing but the final results of the data mining computation [6]. The data mining privacy-preserving is really a unique case in cryptography called multiparty secure computation. Plainly, a convention is expected to comprehend privacy-preserving distributed data mining issues.

Database keeps growing rapidly because of the availability of powerful and affordable database systems. This explosive growth in data and databases has generated an urgent need for new techniques and tools that can intelligently and automatically transform the processed data into useful information and knowledge. Consequently, data mining has become a research area with increasing importance [8]. To design an effective data mining technique several issues to be taken into account such as types of data, efficiency and scalability of data mining algorithms, usefulness, different sources of data, protection of privacy & data security and so on.

A. dvantages of Data Mining

Data mining deals with extracting inherent, historical, and hypothetically critical information from huge databases. Data mining is a very challenging task since it involves building and providing software that will manage, explore, summarize, model, analyze and interpret large datasets in order to evaluate patterns and abnormalities. The methods or techniques of data mining are widely used at a higher rate in various forms of applications. Some of the major applications are detecting fraud prevention, tax avoidance, catching drug smugglers, reducing customer churn and learning more about customers' behavior.

B. Misuses of Data Mining

There are also some (miss) uses of data mining that have little to do with any of these applications. For example, a number of news studys in early 2005 have reported results of analyzing associations between the political party that a person votes for and the car the person drives. The statistics of various branded cars used by the two key political parties of USA was analyzed. In the wake of 9/11 terrorist attacks, considerable use of personal information, provided by individuals for other purposes as well as information collected by governments including intercepted emails and telephone conversations, is being made in the belief that such information processing (including data mining) can assist in identifying persons who are likely to be involved in terrorist networks or individuals who might be in contact with such persons or other individuals involved in illegal activities (e.g. drug smuggling). Under legislation enacted since 9/11, many governments are able to demand access to most private sector data. This data can include records on travel, shopping, housing, utilities, credit, telecommunications and so on. Such data can then be mined in the belief that patterns can be found that will help in identifying terrorists or drug smugglers.

C. Data Mining for Healthcare

Data mining applications have an enormous potential and advantage in the healthcare industry. However, the quality and potential of data mining usage depends on the quality of data available in healthcare. So, keeping this in respect, the healthcare industry has the necessity to ensure quality data is captured, stored, managed, and placed. The benefit area is majorly standardization of clinical tasks and sharing the medical data among the medical organizations to enhance mining.

D. Data Mining for Market Analysis

Data Mining can also be used in market analysis. For instance, when a customer visits a store to buy certain products, then data mining helps us to identify the associated various items that the customer picks from the store. Identifying such data helps market analysis and to promote business. Such different customers and their buying patterns help identify the needs of the customers [9]. This technique helps improve the profits and to the customers to find their associate products better. So, Data mining unveils the data, which is hidden in the database, but owners will not be happy if that hidden data is confidential, and they feel very uncomfortable if this data was submitted to the public. This Problem enhances the interest in doing the research to invent different types of algorithms and protocols for privacy preserving.

E. Privacy Preserving Data Mining

Data mining privacy preserving is targeted to produce correct data mining results without revealing the sensitive information. Figure 2 shows the privacy preserving data mining architecture[10]. Data mining techniques extracts valuable information from data stores. When the techniques are applied, it not only extracts useful data, may also reveal sensitive information. So to provide the protection for sensitive information some of the privacy preserving techniques can be applied on original data then mining can be performed.

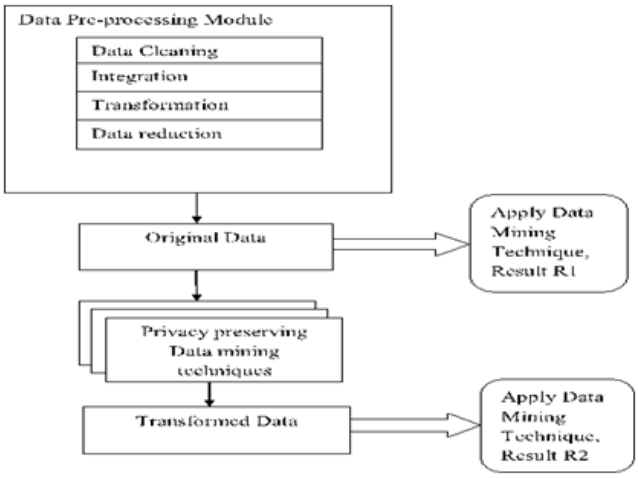

Figure: 2: Privacy Preserving Data mining in proposed approach

Data mining privacy preserving has much centrality due to the accompanying focuses:

1) Data mining causes and moral issue, since it uncovers data, which ought to requires privacy?
2) Privacy preserving data mining gives security to private data against unapproved get to is a long haul accomplishment for data mining security research group and for the administration organizations..

Hence, the security issue is one of the emerging areas that became valuable research area in data mining.

Figure: 3 explore the sequence of steps to be followed for achieving secure data mining results. This work is proposed to perform clustering task on both original data called as R1 and on transformed data called as R2. Finally R1 and R2 will be observed and analyzed for evaluating the performance of proposed approach. This work provides one of the solutions for data mining privacy preserving and the performance is measured in terms of accuracy in sensitive data and data mining result.

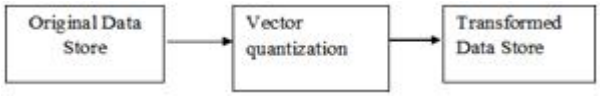

Figure: 3 Proposed Approach.

This thesis presents Vector Quantization technique with two approaches, which transforms underlined sensitive data with the help of codebook, in such a way that the patterns from the original data set are maintained more securely in transformed data set.

The experimental results are explored and analyzed; it is been observed and evaluated that the cluster objects can be extracted securely using the proposed approach. i.e., clusters obtained from the original data set and transformed data set are very similar in terms of accuracy with sufficient protection to sensitive data. It is been proposed to maintain a high quality of transformed data with privacy constraints.

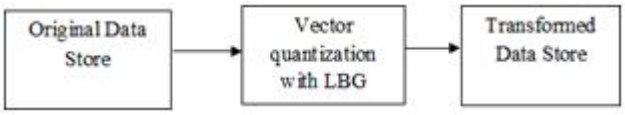

Figure: 4 Proposed Approaches 1.

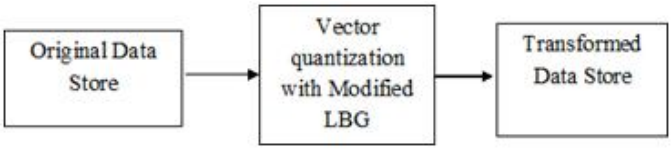

Figure: 5 Proposed Approaches 2.

Figure: 4 and Figure: 5 show the two proposed approaches. When a little amount of distortion is added to original data, one should take care about the accuracy of data

mining tasks, such as classification, association rule mining and clustering. Adding the noise is not only the technique for preserving privacy, other methods like swapping, suppression, anonymization can be used. The increasing demand of data mining privacy preservation gives the direction to research about privacy preserving data mining.

Considering the huge development in technology such as data processing methods, internet and data storage, we need to pay equal attention towards privacy preserving data mining.

Shu-Hsien Liao (2012) explained that increases in digital data have raised concerns about information privacy on a global basis. This particular research study is considered the seminal work in PPDM research. They explain that the Internet has made data collection and data storage much easier, but the potential for misuse has also risen significantly. Data mining results can show models of aggregate data, but the model's accuracy depends on the quality of data. The authors raise the concern that any changes to data affect the accuracy and output of data mining models. Their approach to the problem allows the consumer to provide a value for sensitive attributes. This allows consumers to participate in the process and hopefully gives the consumer a sense of control over his or her own information. A major drawback of this approach is that output accuracy is lost during data mining activities. However, the authors maintain that small drops in accuracy are an acceptable trade-off for privacy.

III. CONCLUSION

This paper presented a privacy preserving technique that adds noise to each and every at- tribute, both numerical and categorical, of a data set. We added noise in such a way so that a high data quality is preserved in the perturbed data set. We measured data quality through the following quality indicators: degree of similarity between two decision trees obtained from an original and a perturbed data set, prediction accuracy of the decision trees, and correlation matrices of the original and the perturbed data set. Therefore, the perturbed data set can be used for classification, prediction and correlation analyzes. More- over, since we add a little amount of noise the perturbed data set can also be used for many other data analyzes. Since noise is added to all attributes, it makes record re- identification determining the confidential class values difficult. The presented techniques for adding noise to a sensitive class attribute. We added the same amount of noise in three class attribute perturbation techniques, namely the RPT, PPT and ALPT. We compared results of our experiments on all these techniques. Our experimental results specify that the PPT and RPT preserve the patterns better than the ALPT - although the same amount of noise has been added in the techniques.

REFERENCES

[1] S. R. M. Oliveria, (2005) Data Transformation for Privacy-Preserving Data Mining, Ph. D. thesis, University of Alberta.

[2] Aggarwal,, Charu C, A General Survey of Privacy-Preserving Data Mining Models and Algorithms, Springer.

[3] Shu-Hsien Liao, Pei-Hui Chu and Pei-Yuan Hsiao, (2012) "Data mining techniques and applications–A decade review from 2000 to 2011", Expert Systems with Applications.

[4] Wei-Yin Loh (2011) "Classification and regression trees", WIREs Data Mining and Knowledge Discovery, Volume 1.

[5] Yi Peng, Yong Zhang, Yu Tang and Shiming Li, (2011) An incident information management framework based on data integration, data mining, and multi- criteria decision making, Decision Support Systems, pp.316–327.

[6] MikalaiTsytsarau and Themis Palpanas, (2012) "Survey on mining subjective data on the web", Data Mining Knowledge Discovery, pp.478–514.

[7] injun Qi "An Overview of Privacy Preserving Data Mining", International Conference of Environment Scienes and engineering.

[8] Shoban Babu Sriramoju, "Analysis and Comparison of Anonymous Techniques for Privacy Preserving in Big Data" in "International Journal of Advanced Research in Computer and Communication Engineering", Vol 6, Issue 12, December 2017, DOI 10.17148/IJARCCE.2017.61212 [ISSN(online) : 2278-1021, ISSN(print) : 2319-5940]

[9] Shoban Babu Sriramoju, " Review on Big Data and Mining Algorithm" in "International Journal for Research in Applied Science and Engineering Technology", Volume-5, Issue-XI, November 2017, 1238-1243 [ISSN : 2321-9653], www.ijraset.com

[10] Shoban Babu Sriramoju, "OPPORTUNITIES AND SECURITY IMPLICATIONS OF BIG DATA MINING" in "International Journal of Research in Science and Engineering", Vol 3, Issue 6, Nov-Dec 2017 [ISSN : 2394-8299].

[11] Shoban Babu Sriramoju, "Heat Diffusion Based Search for Experts on World Wide Web" in "International Journal of Science and Research", https://www.ijsr.net/archive/v6i11/v6i11.php, Volume 6, Issue 11, November 2017, 632 - 635, #ijsrnet

[12] Dr. Shoban Babu Sriramoju, Prof. Mangesh Ingle, Prof. Ashish Mahalle "Trust and Iterative Filtering Approaches for Secure Data Collection in Wireless Sensor Networks"

in "International Journal of Research in Science and Engineering" Vol 3, Issue 4, July-August 2017 [ISSN : 2394-8299].

[13] Sriramoju Ajay Babu, Dr. S. Shoban Babu, "Improving Quality of Content Based Image Retrieval with Graph Based Ranking" in "International Journal of Research and Applications" Vol 1, Issue 1,Jan-Mar 2014 [ISSN : 2349-0020].

[14] Dr. Shoban Babu Sriramoju, Ramesh Gadde, "A Ranking Model Framework for Multiple Vertical Search Domains" in "International Journal of Research and Applications" Vol 1, Issue 1,Jan-Mar 2014 [ISSN : 2349-0020].

[15] Mounika Reddy, Avula Deepak, Ekkati Kalyani Dharavath, Kranthi Gande, Shoban Sriramoju, "Risk-Aware Response Answer for Mitigating Painter Routing Attacks" in "International Journal of Information Technology and Management" Vol VI, Issue I, Feb 2014 [ISSN : 2249-4510]

[16] Mounica Doosetty, Keerthi Kodakandla, Ashok R, Shoban Babu Sriramoju, "Extensive Secure Cloud Storage System Supporting Privacy-Preserving Public Auditing" in "International Journal of Information Technology and Management" Vol VI, Issue I, Feb 2012 [ISSN : 2249-4510]

[17] Shoban Babu Sriramoju, "Multi View Point Measure for Achieving Highest Intra-Cluster Similarity" in "International Journal of Innovative Research in Computer and Communication Engineering" Vol 2,Issue 3,March 2014 [ISSN(online) : 2320-9801, ISSN(print) : 2320-9798]

[18] Shoban Babu Sriramoju, Madan Kumar Chandran, "UP-Growth Algorithms for Knowledge Discovery from Transactional Databases" in "International Journal of Advanced Research in Computer Science and Software Engineering", Vol 4, Issue 2, February 2014 [ISSN : 2277 128X]

[19] Shoban Babu Sriramoju, Azmera Chandu Naik, N.Samba Siva Rao, "Predicting The Misusability Of Data From Malicious Insiders" in "International Journal of Computer Engineering and Applications" Vol V,Issue II,Febrauary 2014 [ISSN : 2321-3469]

[20] Ajay Babu Sriramoju, Dr. S. Shoban Babu, "Analysis on Image Compression Using Bit-Plane Separation Method" in "International Journal of Information Technology and Management", Vol VII, Issue X, November 2014 [ISSN : 2249-4510]

[21] Shoban Babu Sriramoju, "Mining Big Sources Using Efficient Data Mining Algorithms" in "International Journal of Innovative Research in Computer and Communication Engineering" Vol 2, Issue 1,January 2014 [ISSN(online) : 2320-9801, ISSN(print) : 2320-9798]

[22] Ajay Babu Sriramoju, Dr. S. Shoban Babu, "Study of Multiplexing Space and Focal Surfaces and Automultiscopic Displays for Image Processing" in "International Journal of Information Technology and Management" Vol V, Issue I, August 2013 [ISSN : 2249-4510]

[23] Dr. Shoban Babu Sriramoju, "A Review on Processing Big Data" in "International Journal of Innovative Research in Computer and Communication Engineering" Vol-2, Issue-1, January 2014 [ISSN(online) : 2320-9801, ISSN(print) : 2320-9798]

[24] Shoban Babu Sriramoju, Dr. Atul Kumar, "An Analysis around the study of Distributed Data Mining Method in the Grid Environment : Technique, Algorithms and Services" in "Journal of Advances in Science and Technology"

[25] Vol-IV, Issue No-VII, November 2012 [ISSN : 2230-9659]

[26] Shoban Babu Sriramoju, Dr. Atul Kumar, "An Analysis on Effective, Precise and Privacy Preserving Data Mining Association Rules with Partitioning on Distributed Databases" in "International Journal of Information Technology and management" Vol-III, Issue-I, August 2012 [ISSN : 2249-4510]

[27] Shoban Babu Sriramoju, Dr. Atul Kumar, "A Competent Strategy Regarding Relationship of Rule Mining on Distributed Database Algorithm" in "Journal of Advances in Science and Technology" Vol-II, Issue No-II, November 2011 [ISSN : 2230-9659]

[28] Shoban Babu Sriramoju, Dr. Atul Kumar, "Allocated Greater Order Organization of Rule Mining utilizing Information Produced Through Textual facts" in "International Journal of Information Technology and management" Vol-I, Issue-I, August 2011 [ISSN : 2249-4510]

[29] Ramesh Gadde, Namavaram Vijay, "A SURVEY ON EVOLUTION OF BIG DATA WITH HADOOP" in "International Journal of Research in Science and Engineering", Vol-3, Issue-6, Nov-Dec 2017, 92-99 [ISSN : 2394-8299].

[30] Namavaram Vijay, S Ajay Babu, "Heat Exposure of Big Data Analytics in a Workflow Framework" in "International Journal of Science and Research", Volume 6, Issue 11, November 2017, 1578 - 1585, #ijsrnet

[31] Ajay Babu Sriramoju, Namavaram Vijay, Ramesh Gadde, "SKETCHING-BASED HIGH-PERFORMANCE BIG DATA PROCESSING ACCELERATOR" in "International Journal of Research in Science and Engineering", Vol-3, Issue-6, Nov-Dec 2017, 92-99 [ISSN : 2394-8299].

[32] Namavaram Vijay, Ajay Babu Sriramoju, Ramesh Gadde, "Two Layered Privacy Architecture for Big Data

Framework" in "International Journal of Innovative Research in Computer and Communication Engineering" Vol 5, Issue 10, October 2017 [ISSN(online) : 2320-9801, ISSN(print) : 2320-9798]

[33] Amitha Supriya. "Implementation of Image Processing System using Big Data in the Cloud Environment." International Journal for Scientific Research and Development 5.10 (2017): 211-217.

[34] SA Supriya. "A Survey Model of Big Data by Focusing on the Atmospheric Data Analysis." International Journal for Scientific Research and Development 5.10 (2017): 463-466.

[35] Siripuri Kiran, 'Decision Tree Analysis Tool with the Design Approach of Probability Density Function towards Uncertain Data Classification', International Journal of Scientific Research in Science and Technology(IJSRST), Print ISSN : 2395-6011, Online ISSN : 2395-602X,Volume 4 Issue 2, pp.829-831, January-February 2018. URL : http://ijsrst.com/IJSRST1841198

[36] Ajmera Rajesh, Siripuri Kiran, " Anomaly Detection Using Data Mining Techniques in Social Networking" in "International Journal for Research in Applied Science and Engineering Technology", Volume-6, Issue-II, February 2018, 1268-1272 [ISSN : 2321-9653], www.ijraset.com