ISSN: 2278 – 7798
International Journal of Science, Engineering and Technology Research (IJSETR)
Volume 2, Issue 2, February 2013

A SURVEY OF CLUSTERING OF PARTITONED DATA IN A DISTRIBUTED NETWORK

S.Haripriya[1], Prof.T.Kalaikumaran[2], Dr.S.Karthik[3]

***Abstract*— Data Mining plays a major role in storage of vast quantities of data. It extracts valuable knowledge, which helps organizations to obtain better results by pooling their data together. Distributed data mining is concerned about data that are shared among multiple organizations. Privacy-preserving distributed data mining deals with cooperative parties without revealing any of their individual data items. Data mining tasks include clustering, prediction, association rule mining and outlier detection. Data mining is used in biomedical and DNA data analysis, financial data analysis, identification of unusual patterns, and analysis of telecommunication data. A complementary approach to privacy-preserving data mining uses randomization techniques. Privacy-preserving data mining solutions have been presented both with respect to horizontally and vertically portioned databases, in which earlier data objects with the same attributes for the same data objects are owned by each party. Division of data into groups of similar objects is called Clustering. K-means clustering is a simple technique to group items into k clusters. The quality of a set of clusters can be measured using the value of an objective function which is taken to be the sum of the squares of the distances of each point from the centered of the cluster to which it is assigned. It's required that value of this function to be as small as possible.It is to obtain privacy preserving data mining protocols for other algorithms over arbitrarily partitioned data.**

***Index Terms*—**Data mining,privacy preserving,clustering

I. INTRODUCTION

Data mining is a technique that deals with the extraction of hidden predictive information from large database. It uses sophisticated algorithms for the process of sorting through large amounts of data sets and picking out relevant information. It requires data preparation which can uncover information or patterns which may compromise confidentiality and privacy obligations. Advancement of efficient data mining technique has increased the disclosure risks of sensitive data. Data mining is social and ethical problem by revealing the data which should require privacy. Providing security to sensitive data against unauthorized access has been a long term goal for the database security research community and for the government statistical agencies. Hence, the security issue has become, a much more important area of research in data mining.Knowledge or patterns can be extracted from large data stores while maintaining commercial or legislative privacy constraints. It is to improve the trade of between privacy and utility when mining data. The difficulties of applying Privacy Preserving Data Mining algorithms to a distributed database can be attributed to: first, 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. The Privacy Preserving Data Mining algorithms can be further classified into two types, data hiding and rule hiding, according to the purposes of hiding. Data hiding refers to the cases where the sensitive data from original database like identity, name, and address that can be linked, directly or indirectly, to an individual person are hided.

In rule hiding, the sensitive knowledge (rule) derived from original database after applying data mining algorithms is removed. Majority of the PPDM algorithms used data hiding techniques. Most PPDM algorithms hide sensitive patterns by modifying data. Four techniques of privacy preservation– 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 problem is not simply that the data is distributed, but that it must be distributed.There are several situations where this arises

Connectivity- Transmitting large quantities of data to a central site may be infeasible.

Heterogeneity of sources- Is it easier to combine results than combine sources?

Privacy of sources- Organizations may be willing to share data mining results, but not data.

Both kinds of partitioning data has different challenges to the problem to be distributed in privacy-preserving data mining. The problem of distributed privacy-preserving data mining overlaps closely with a field in cryptography for determining secure multi-party computations.The intersection between the fields of cryptography and privacy-preserving data mining may be found. The approach of cryptographic methods tends to compute functions over inputs provided by multiple recipients without actually sharing the inputs with one another. Division of data into groups of similar objects is called Clustering.A good clustering method will produce high quality clusters with high intra class and low inter class similarity.The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns.

Privacy-preserving data mining was to enable conventional data mining techniques to preserve data privacy during the mining process. Privacy-preserving data mining on horizontally and/or vertically partitioned data involving multiple parties so that no single party holds the overall data of horizontally and vertically partitioned data.For vertically partitioned data, two parties or more hold the different set of attributes for the same set of objects. In arbitrarily partitioned data, different disjoint portions are held by different parties.

Manuscript received Jan , 2013.
[1]***Haripriya.S*** *, ME Software Engineering,SNS College of Technology, Coimbatore, Tamil Nadu.*
[2]***Prof.T.Kalaikumaran,*** *HOD/CSE, SNS College of Technology,Coimbatore, Tamil Nadu.*
[3] ***Dr.S.Karthik,*** *DEAN/CSE, SNS College of Technology,Coimbatore, Tamil Nadu*

For horizontally partitioned data, two parties or more hold different objects for the same set of attributes. It means each object in the virtual database is completely owned by one party. For arbitrarily partitioned data,two parties hold data forming a (virtual) database consisting of their joint data.

II. RELATED WORK

The primary task in data mining is the development of models about aggregated data, they develop accurate models without access to precise information in individual data records. The resulting data records look very different from the original records and the distribution of data values is also very different from the original distribution. The area of statistical databases motivated by the desire to be able to provide statistical information without compromising sensitive information about individuals. The proposed techniques can be broadly classified into query restriction and data perturbation. The query restriction family includes restricting the size of query result, controlling the overlap amongst successive queries, keeping audit trail of all answered queries and constantly checking for possible compromise suppression of data cells of small size, and clustering entities into mutually exclusive atomic populations.

The perturbation family includes swapping values between records , replacing the original database by a sample from the same distribution, adding noise to the values in the database , adding noise to the results of a query, and sampling the result of a query.The proposed technique cannot satisfies the conflicting objectives of providing high quality statistics and at the same time prevent exact or partial disclosure of individual information[1]. Mining and integrating data from multiple sources, there are many privacy and security issues.The security of the full privacy-preserving data mining protocol depends on the security of the underlying private scalar product protocol. The two private scalar product protocols, one of which was proposed in a leading data mining conference, are insecure and the other one is based on homomorphic encryption and improve its efficiency so that it can also be used on massive datasets[2].The proposed solutions try to achieve information-theoretical security—that is,without relying on any computational assumption—by using additive or linear noise to mask the values. The goal is that one of the participants obtains the scalar product of the private vectors of all parties. It is often required that no information about the private vectors, except what can be deduced from the scalar product, will be revealed during the protocol. Data mining applications work with a huge amount of data, it is desirable that the scalar product protocol is also very efficient. A secure scalar product protocol has various applications in privacy preserving data mining, starting with privacy-preserving frequent pattern mining on vertically distributed database and ending with privacy-preserving cooperative statistical analysis. The secure computation of a scalar product is an important task within many data mining algorithms that require the preservation of privacy. Several protocols have been proposed to solve this task. However, that they are insecure. A private scalar product protocol based on standard cryptographic techniques and proved that it is secure.

The theoretical foundations of modern cryptography has been proposed. The focus of the course is to understand what cryptographic problems can be solved, and under what assumptions. The main focus of the cryptography is the presentation of “feasibility results” It is typically not relate to issues of efficiency. There are a number of significant differences between this course and its prerequisite. First, their will be rigorous, and so they will only present constructions that have been proven secure.Second, they will not begin with cryptographic applications like encryption and signatures, but will rather conclude with them. Rather, they start by studying one-way functions and their variants, and then show how different cryptographic primitives can be built from these. Third, the aim of the course is to provide the students with a deep understanding of how secure cryptographic solutions are achieved, rather than with a basic understanding of the important concepts and constructions. The need for a rigorous approach in cryptography is especially strong. First, intuitive and heuristic arguments of security have been known to fail dismally when it comes to cryptography. Second, in contrast to many other fields, the security of a cryptographic construction cannot be tested empirically. Finally, the potential damage of implementing an insecure solution is often too great to warrant the chance.Cryptography differs from algorithms. A rigorous approach to algorithms is also important[3]. However, heuristic solutions that almost always provide optimal solutions are often what is needed. In contrast, a cryptographic protocol that prevents most attacks is worthless, because an adversary can maliciously direct its attack at the weakest link.

The privacy preserving data mining of two parties owning confidential databases wish to run a data mining algorithm on the union of their databases, without revealing any unnecessary information. The secure multi-party computation and as such ,that can be solved by using known generic protocols. Data mining algorithms are typically complex and the input usually consists of massive data sets. The generic protocols is more efficient protocols are required. The focus on the problem of decision tree learning with the popular ID3 algorithm. Their protocol is considerably more efficient than generic solutions and demands both very few rounds of communication and reasonable bandwidth.A wide range of applications, data mining techniques also have raised a number of ethical issues. Some such issues include those of privacy, data security, intellectual property rights, and many others. The address of the privacy problem against unauthorized secondary use of information[4]. A family of geometric data transformation methods (GDTMs) which ensure that the mining process will not violate privacy up to a certain degree of security. The focus of primarily on privacy preserving data clustering, notably on partition-based and hierarchical methods.

The proposed methods distort only confidential numerical attributes to meet privacy requirements, while preserving general features for clustering analysis. The methods are effective and provide acceptable values for balancing privacy and accuracy. The methods based on the data perturbation approach fall into two main categories known as probability-distribution category and fixed-data perturbation category . In the probability-distribution category, the

security-control method replaces the original database by another sample from the same distribution or by the distribution itself. On the other hand, the fixed-data perturbation methods have been developed exclusively for either numerical data or categorical data.

III. CONCLUSION

A privacy-preserving is a solution to an important data mining problem, that of clustering data. The k-means clustering algorithm is a well known iterative algorithm that successively refines potential clusters in an attempt to minimize the k-means objective function, which measures the goodness of a given clustering. A privacy preserving protocol for k-means clustering of arbitrarily partitioned data distributed between two parties. It is efficient and provides cryptographic privacy protection. The protocol provides the first privacy-preserving solution to k-means clustering for horizontally partitioned data and also provide an analysis of the performance and privacy of our solution mapping the constraint satisfaction problem to an equivalent binary integer programming problem.

If you wish, you may write in the first person singular or plural and use the active voice ("I observed that ..." or "We observed that ..." instead of "It was observed that ..."). Remember to check spelling. If your native language is not English, please get a native English-speaking colleague to proofread your paper.

ACKNOWLEDGMENT

The authors would like to thank the Editor-in-Chief, the Associate Editor and anonymous Referees for their comments.

REFERENCES

[1]. Agrawal.R and Srikant.R (2000)'Privacy preserving data mining',In Proc.ACM SIGMID Conf. on Management of Data,pages 439-450.ACM Press

[2]. Goethals.B,Laur.S,Lipmaa.H, and Mielikainen.T,(2004)'On Secure scalar product computation for privacy-preserving data mining'.In The 7th Annual International Conf. in Information Security and Cryptology

[3]. Golreich.O(2004)'Foundations of Cryptography',Vol II Cambridge University Press

[4]. Oliveria.Sand Zaiane.O.R(2003)'Privacy preserving clustering by data transformation' In Proc.18th Brazilian Symposium on Databases, pages 304-318

[5]. Prakash. V.S,Shanmugam.A , Murugesan.P (2012) 'Efficient Cluster Based Privacy Preservation Data Perturbation Technique in Multi-Partitioned Datasets',European Journal of Scientific Research, ISSN 1450-216X Vol. 86 No 2 September, 2012, pp.254-263

[6]. Shuguo HAN, and Wee Keong NG(2007)'Multi-Party Privacy-Preserving Decision Trees for Arbitrarily Partitioned Data' International Journal Of Intelligent Control And Systems Vol. 12, No. 4, 351-358

BIOGRAPHY

S.Haripriya received B.E degree on Computer Science and Engineering from Anna University of Technology, Coimbatore, Tamilnadu,INDIA in 2007 and pursuing M.E Software Engineering in SNS College of Technology affliated to Anna University,Chennai. Her Research includes Data Mining in privacy preserving.

Prof.T.Kalaikumaran received MCA degree from the Madras University in 1996, and the M.E. degree in Computer Science and Engineering from the Anna University, Chennai, in 2006. He is a Postdoctoral fellow at the Anna University of Technology, Coimbatore. He is the Head of the Department of Computer Science & Engineering,SNS College of Technology affiliated to Anna University-Coimbatore, Tamilnadu, India. He is interested in the research areas of data mining, spatial data mining, machine learning, uncertain data classification and clustering, pattern recognition, database management system and informational retrieval system. He is a member of CSI and IEEE.

Dr.S.Karthik is presently Professor & Dean in the Department of Computer Science & Engineering, SNS College of Technology, affiliated to Anna University- Coimbatore, Tamilnadu, India. He received the M.E degree from the Anna University Chennai and Ph.D degree from Ann University of Technology, Coimbatore. His research interests include network security, web services and wireless systems. In particular, he is currently working in a research group developing new Internet security architectures and active defense systems against DDoS attacks. Dr.S.Karthik published more than 35 papers in refereed international journals and 25 papers in conferences and has been involved many international conferences as Technical Chair and tutorial presenter. He is an active member of IEEE, ISTE, IAENG, IACSIT and Indian Computer Society.