SHIV SHAKTI

International Journal of in Multidisciplinary and Academic Research (SSIJMAR)

Vol. 4, No. 3, June 2015 (ISSN 2278 – 5973)

PRIVACY OF DATA IN DATA MINING FOR PARTITIONED DATA

MASTER OF TECHNOLOGY
(Computer Engineering)

KAMLESH YADAV
Roll No. 2104183

Under the guidance of

Prof. PuspenderSarao
SOMANY (P.G.) INSTITUTE OF TECHNOLOGY &MANAGEMENT
June, 2015

Impact Factor = 3.133 (Scientific Journal Impact Factor Value for 2012 by Inno Space Scientific Journal Impact Factor)

Global Impact Factor (2013)= 0.326 (By GIF)

Indexing:

ABSTRACT

User data is commercially valuable. The burgeoning data science industry is predicated on the value of insights extracted from databases. At the same time, many users and politicians are concerned about Internet privacy. Intuitively, it might seem that data mining and privacy protection are mutually incompatible goals. Differential privacy, a mathematical definition of privacy invented by Cynthia Dworkat Microsoft Research Labs, offers the possibility of reconciling these competing interests. With differential privacy, general characteristics of populations can be learned while guaranteeing the privacy of any individual's records.Data mining has been a popular research area among the researchers for more than a decade because of its vast use of applications. Advances in today's technology have enabled a large collection of data in the organizations. This vast collection of data need to be mined for the purpose of knowledge discovery as Data mining is the field of extracting interesting patterns from large data collections. Data mining enables organizations to get agreed on grouping their data together for mining purpose because they know that mining results are fruitful for them. However, the popularity and wide availability of data mining tools also raised concerns about the privacy of individuals as large data collections consists sensitive information about the individual.Big data abounds. No precise definition of "big data" exists, but a good rule of thumb is data sets too large to fit in main memory on a single machine. While the buzzword may be overused, the trend is real. Cheap memory, fast Internet connections, and obsessively used, sensor-laden smart-phones have combined to generate massive datasets as well as the means to transmit and store them.While companies could amass datasets about anything measurable, generally the most coveted datasets in Silicon Valley contain individuals' personal information. The economic significance of such data is obvious. If a company can predict a user's purchasing decisions, it can advertise optimally. Google and Facebook rely upon well-chosen ads to monetize their otherwise free web services.The potential value of mining human-generated data goes beyond advertising. the collective health data generated by a large population may contain insights which could bring about better health outcomes for everyone. Medical institutions are eager to mine patient records for longitudinal observations in the hope of generating the knowledge necessary for personalized medicine. Organizations want to apply data mining on their data without leaking any sensitive information about their individuals to other organizations. Thus the aim of privacy preserving data mining researchers is to develop data mining techniques that could be applied on databases without violating the privacy of individuals. These techniques disclose nothing but the final results to all the sites. Privacy Preserving techniques are applied in many different areas like medical, bioinformatics, shopping, credit card analysis etc. And it has been a fruitful technique in all the fields. Privacy preserving techniques have been proposed for many data models like classification on centralized data then for association rules in distributed environments and clustering in vertical data partitioning. In this dissertation, we propose methods for privacy preservation in distributed environment. We construct the privacy preserving dissimilarity matrix of objects stored in different sites which can be used for privacy preserving clustering and other operations. It deals with the pair wise comparison of individual private sensitive data objects

which are distributed horizontally to multiple sites. Here all the sites taking part in mining process are supposed to be the semi-honest means these sites are in honest but curious state. In this dissertation we deal with the alphanumeric, categorical with numeric attributes as well. Dissimilarity matrix is being constructed with the help of a third party that is being involved to perform mining on over all collected data. We show communication and computation complexity of our protocol by conducting experiments over synthetically generated and real datasets. Each experiment is also performed for a basic protocol which has no privacy concern to show that the overhead comes with security and privacy by comparing the basic protocol and our protocol.

INTRODUCTION

PRIVACY IN DATA MINING:-Differential privacy addresses a very specific notion of privacy. It is suited to the situation in which one is deciding whether to allow their data to be included in database, say for a research study. If all mechanisms which access the data are proven to be differentially private, since any individual's data does not perceptibly affect the study, a guarantee of differential privacy seems a compelling argument for participation.Data are any facts, numbers, or text that can be processed by a computer. Today, organizations are accumulating vast and growing amounts of data in different formats and different databases. And from that data base fetch our needed data that process called data mining. Data mining is the analysis of data for relationships that have not previously been discovered. For example, the sales records for a particular brand of tennis racket might, if sufficiently analyzed and related to other market data, reveal a seasonal.

With the advances in technology at today's time, the amount of data stored in organizations is growing at a phenomenal rate. According to the researchers the amount of data in this world is doubling every 20 months. There are many sources of this type of data worldwide from which this data is collected and aggregated at a single place for further use. These sources include medical data, telephone data, credit card records, shopping details, university records, air line data, bank records etc. If we take an organization into consideration then we will come to know that a huge amount of data is collected by the organization about their customers, clients, projects, departments etc. This data is increasing day by day and turning into the vast quantities with the rapid growth in the technology and computer networking. Such organizational data was initially used only for record keeping and preserving historical data. But the users of this collection of data started expecting something more interesting patterns and sophisticated information that we can actually call knowledge. So the organization's people soon realized that data mining can be applied on this large amount of data for extracting interesting patterns or knowledge and these patterns are used for further improving the performance of their organization and better decision making. Here data mining attempts to solve this problem and provides effective results.

Data mining tools have been proved to be very strong and powerful because of having the power of extracting hidden information or knowledge from vast amount of data and it lead to increased data storage efforts by organizations and government agencies. Data mining has been very popular among the researchers for more than hundreds of the years due to its vast availability of

tools and applications. Data mining can extract valuable knowledge and patterns from collected data. Keeping this popularity of data mining in mind, organizations rather than only storing data for record keeping, realized that they can obtain better performance by pooling their data together from different sites and applying data mining techniques over that data. Data mining is generally defined as the process of finding meaningful trends, knowledge and new correlation patterns through non-trivial extraction of implicit, historical and previously unknown patterns from large collection of data kept in databases using mathematical, statistical techniques and pattern recognition as well.

Initially simple structured languages were used for finding trends and patterns from the collected data but with the increasing demand for sophisticated information, it is realized that these simple structured languages (like SQL) are not adequate to fulfill the needs of the users of this large collection of data. This is because a SQL1 query is generally written or constructed to retrieve data that is specific to the user. Means the researchers applying SQL query on the data collection already know what they are going to get as an output of this process. Usually the output of the SQL query is a database subset while researchers applying data mining over data collection are not exactly sure of what they require as mining result. So the output of a data mining query is said to be an analysis of the contents of the database. Data mining deals with the extraction of useful information from vast amount of data with a variety of application fields such as medical, customer relationship management, credit card analysis, market basket analysis, and bioinformatics. The output of the mining process or the extracted information could be in the form of patterns, trends, clusters or classification models. Data mining is applied on different data models like clustering, classification etc. over centralized and distributed data. Association rules on shopping records of a supermarket are a general example that can describe the relationship among the shopping items bought together. If we take a customer relationship management into consideration, then here customers could be clustered in different-different groups according to different metrics for improving customer relationship management. An another model is the classification models that can be built on the recorded data like customer shopping records and behavior to perform better decision making to do targeted marketing. If we talk about the widely acceptable data mining tasks then these include association rule mining, clustering, classification - prediction and outlier detection. Data mining being very popular and widely available, has been used in many famous areas that include financial data analysis, credit card and fraud detection analysis, identification of unusual patterns and analysis of telecommunication dataand Biomedical and DNA2 data analysis etc.

Data mining has been a very interesting and strong area for research purpose because of its wide spectrum of applications but this availability, interestingness and popularity of data mining tools and applications also raised issue about the leakage of the sensitive information that compromise the privacy of the individuals. As we discussed above, the organizations agree on pooling their data together for mining purpose but this data contains some sensitive information that an organization won't want to disclose to other organizations because it compromise individuals privacy. Here it is discussed that this term 'privacy' is overloaded with different meanings in our society and wide range of its definitions are assumed in general that show variations in meaning. For example, in the context of the HIPAA (Health Insurance Portability and Accountability Act) Privacy Rule, the meaning of privacy is the individual's ability to control the access to personal health care information. If we take organization's point of view into consideration, the definition of privacy involves the policies stating which information is collected, the way it is used, and how customers are involved in this process. In different areas privacy has different meanings and

definitions that are related to those particular areas or environment in which privacy need to be guaranteed. But in context to our work we don't need these definitions of privacy. These are described just to explain the actual meaning of privacy. But what we actually need is a generic definition of privacy that can be instantiated to different areas and situations and also easier to understand for our point of view. From a researcher's point of view, Schoeman (1984) and Walters (2001) presented three possible definitions of privacy that are similar but having difference from researcher's point of view. These definitions are stated below:

1. Privacy as the right of a person to determine which personal information about himself / herself may be communicated to others.
2. Privacy as the control over access to information about oneself.
3. Privacy as limited access to a person and to all the features related to the person.

So these three definitions explaining the term 'privacy' are very similar apart from some philosophical differences that are not under the scope of our work. One thing we take as interesting as per our point of view is the concept of "Information release control" emerging from the above definitions of privacy given by researchers. From this concept, we present a definition of privacy which has a meaning closely related to our objective. This is described as follows: "The right of an individual to be protected from unauthorized leakage of his/her sensitive information that is stored at a single location for mining purpose". A more refined definition of privacy is defined next. Privacy is considered as "The right of an entity to be protected from unauthorized leakage of sensitive information that is stored in an electronic repository or that can be derived as complex and aggregate data from data contained in an electronic repository".

Data mining technology which discloses sensitive information in large collection of data could compromise the patterns that are considered as private by individuals or organizations. When organizations want more precise and useful information from its large collection of data then there is a need of much data mining to be performed on that data. For this purpose organizations get ready to pool their data together from multiple sources. However this vast collection of data may contain some information about the organizations and their customers that is considered to be private and sensitive and when this data is shared among multiple organizations, privacy concerns get exacerbated. Privacy and secrecy concerns have emerged with the development and penetration of data mining in many popular areas and disciplines, Because of these privacy considerations the organizations seem not to be willing or able to share their data with other organizations. This problem gave birth to a new issue and forces the researchers to think about it. In response to that, data mining researchers took this problem into consideration and started to address privacy concerns. They started developing some special type of data mining techniques as opposed to normal data mining techniques having the privacy of individuals in to account. Normally data mining techniques discloses the individual's private information but in opposition to that privacy preserving data mining can be applied to the collection of data without any leakage in the privacy of individuals. Thus the aim of privacy-preserving data mining is two-fold. First is to maximize the data analysis results obtained from data mining process and second is to minimize the inferences that disclose sensitive information about organizations or its customers. In general, data mining techniques promises to extract useful information. If the data collection contains personal or organizational data, data mining shows the potential to disclose the data considered as private. This is more apparent as emerging internet technology offers the

opportunity for the users of this data to obtain records about individuals and share these records with other organizations for mining purpose to get some fruitful results. In some cases, sharing their private data with others for an analysis task may be mutually benefited for both the parties but on the other hand it compromises individual's privacy by sharing information to others however these organizations would like theirs sensitive data to remain private. In other words, a need arises to protect the sensitive or private data of the organizations at the time a data mining process is applied to the database. This is done through Privacy Preserving Data Mining (PPDM). Researchers have proposed many data mining models having privacy concerns in mind. Data is managed under this paradigm of privacy by disclosing or releasing some information to others while keeping private records as private or hidden at the same time. It is related to the issues in statistical databases and authorization and privacy access in databases as well. Privacy-preserving data mining has a wide scope and studied extensively because of having control of sensitive data or information on the internet. The aim of PPDM3 algorithms is to extract relevant and useful information from large amount of data while protecting leakage of private information at the same time. As we discussed above privacy preserving data mining is a twofold. First, sensitive information or raw data like names, addresses, identifiers etc., should be trimmed out from the original database or modified using data perturbation techniques so that the recipients of the data could not compromise the privacy of the individuals. Second, sensitive information which are considered to be mined from the database by applying data mining algorithms over that should also be excluded from the database because such type of private information can equally well compromise data privacy as discussed.

As researcher's point of view, privacy is widely known and one of the most important properties that an information system is supposed to satisfy. In response to that data mining researchers have devoted several efforts in incorporating privacy preserving data mining techniques in order to control the release of sensitive information by preventing its disclosure at the time of mining process. So the main purpose of privacy preserving data mining researchers is to develop such kind of algorithms that could alter or modify the original data records in some way, so that the sensitive information could not be revealed and remain private even after mining process. In order to modify or trim this data, data perturbation techniques are required to apply on the database in a way so as to preserve privacy of individuals. It is clear that applying "control information release" on database does not provide complete protection of data means hiding sensitive data by restricting access can only minimize risk of information release because in many cases curious parties can obtain useful information or patterns by analyzing non-sensitive data. This issue of data protection against inference has been pointed out in the literature of statistical databases since 1979. However, in the field of data mining this inference control problem has been more specific in the case of clustering technique.

Data mining researchers presented a solution to inference control problem and referred the process of protection against inference as data sanitization. In general data sanitization is referred as the process of making private data safe in non-production databases for wider visibility. Some other researchers discovered a solution that was based on collaborators who mined their data independently and then shared some of the resulting patterns. As a conclusion, the existing privacy preserving data mining techniques can be classified into following different dimensions:

(i) Data distribution structure (centralized or distributed);

(ii) The modification technique applied to the data (i.e. encryption, perturbation, generalization, and so on) in order to apply sanitization;

(iii) Data mining algorithms or techniques i.e. clustering, association rule mining, classification etc. that is to be applied to find the interesting patterns from data.

(iv)The data type (single data items or complex data correlations) that needs to be protected from leakage;

(v) The privacy preserving technique (heuristic or cryptography-base approaches).

There are two types of data distribution models addressed above: centralized and distributed. As the name indicates in centralized databases all the data is located at single place that is being aggregated from multiple sources to a single location while in distributed model rather than being locating at single place, data is located at multiple sites. In this case data holders at different sites need to share their data with each other for the global analysis task or at the time data holders wish to derive aggregate results from datasets that are distributed among all the sites. Here in distributed scenario this data may be partitioned in different ways such that it may be a horizontal partition or a vertical partition. The difference between these partitions is that in horizontal partition 'records' are distributed across multiple sites while in vertical partition 'attributes' are distributed across multiple sites. There are two types of privacy preserving techniques: heuristic based techniques and cryptographic techniques. Heuristic based techniques are mainly designed for centralized datasets while cryptography-based algorithms are developed to be used in distributed scenario to preserve privacy by using cryptographic encryption techniques. These privacy preserving techniques are used with the different existing data mining techniques that are used so far. Before applying these techniques over data, data sanitization is performed to modify or trim the original database in order to keep private records private. Sanitization approaches include randomization, K-anonymity, L-diversity, swapping etc in order to perform encryption, generalization and perturbation. Agrawal and Srikant from IBM Almaden presented data perturbation techniques for classification model of preserving privacy data mining to be applied on centralized data model.

In some cases the individual parties may not wish to share their entire data sets with each other, they may agree to share only limited data with the use of a variety of protocols. These types of methods give effect to manage the secrecy of each individual while deriving global aggregate results over the entire data. Privacy preserving data mining, introduced in, presented a solution that allows individual entities to cooperate in the extraction of information without any discloser of any of the involved site's individual data items to each other or any other parties. Privacy-preserving data mining algorithms in distributed scenario need collaboration between all the sites involved in mining process to compute the global results, while restricting the leakage of any information except the final data mining results. To achieve this goal, researchers presented a solution "secure multiparty computation (SMC)" to be applied over data in distributed scenario. SMC4 came into consideration with Yao's Millionaires' problem. The basic problem is that there are two millionaires and each of them wants to know who is richer as compared to each other but at the same time no one wants to reveal one's net worth to other while comparison takes place. Yao presented a generic circuit evaluation based technique as a solution to this problem as well as generalizing it to any efficiently computable function. Thus in these kind of problems generic circuit based solution is used while parties wish to achieve a secure multiparty solution.

Secure multiparty computation is used with the different PPDDM5 algorithms to preserve privacy of data. , Murat Kantarcioglu presented various PPDDM algorithms to be applied on horizontally partitioned data. These algorithms include K-NN6 Classification, Association Rule Mining, EM7 Clustering, Naïve Bayes Classifier, Decision Trees, and Support Vector Machine. Murat Kantarcioglu also presented common secure sub-protocols used with these algorithms in privacy preserving distributed data mining. The common secure sub protocols used in PPDDM algorithms include Secure Comparison, Secure Union, Secure Polynomial Evaluation, Secure Sum, Secure Dot Product and Secure Logarithm. To devise new privacy preserving data mining algorithms, above stated sub-protocols can be combined by the researchers. There is a large number of different privacy preserving data mining (PPDM) techniques that have been presented by the data mining researchers for several past years. But still there is an emerging need of taking this research area to another level of standard where these data mining techniques may be used to solve the emerging complex problems. One foremost step toward this approach is a "quantification technique". This approach is used for PPDM6 algorithms that make the comparison and evaluation of such algorithm easily possible. When the researchers want to choose the most appropriate privacy preserving technique, it is essential for the users to provide a number of privacy preserving related metrics that will surely help in the selection of techniques with respect to some specific parameters they choose to be interested for optimization purpose.

In general, the quantification that is used to measure the data privacy is the degree of uncertainty, according to which private and sensitive information in original database can be inferred. Protection of data privacy by privacy preserving data mining algorithm depends on the degree of uncertainty. The higher the degree of uncertainty, the better data privacy is protected by algorithm. The degree of uncertainty is estimated in many ways for different types of PPDM algorithms. Heuristic based approaches generally gives the low level of privacy in comparison to cryptography-based technique that guarantees very high level of data privacy but as opposed to the heuristic based techniques, the complexity of cryptographic based algorithms is quite high. In [2], Kantarcioglu and Clifton presented the problem of secure mining of association rules over horizontally partitioned data, using cryptographic techniques that minimize the information shared among different parties. Their solution to the above problem is based on the assumption that each site first uses cryptographic commutative encryption to encrypt its own data sets and then encrypts the pre-encrypted data sets of every other site by using the same technique. After this process, an initiating site transmits its frequency count, plus a random value, to its neighbor site. This neighbor site then adds its frequency count to the previous value and passes it on to other sites. Finally, a secure comparison is achieved between the initiating parties and the final party. Then it is to determine if the final result is greater than the threshold plus the random value. It is being noticed that level of the privacy does not only depends on the privacy preserving data mining algorithm used in the process, but also on the released information that an attacker has about the data before the data mining techniques are used and the relevance of this information in the data reconstruction process.

Researchers have presented privacy-preserving data mining solutions both with respect to both distributed data models such that horizontally and vertically partitioned databases, in which each site owns different data objects with the same attribute, or different attributes for the same data objects are owned by each site respectively.

In this dissertation, I have presented a privacy-preserving solution to a popular data mining problem, clustering technique. In general a clustering algorithm groups a set of similar data objects together into a cluster so that all objects within that cluster are having almost same property or are closely related to each other while objects in the different clusters must be dissimilar. In many different areas clustering technique has been adopted easily and is widely acceptable. In bioinformatics, clustering has been used to group genes with similar expression profile [14]. Astronomers have used clustering to create catalogs of objects in the sky [15]. In image processing, clustering algorithms are used for image segmentation, which is the problem of distinguishing objects from the background [13]. Many methods already have been developed by researchers on privacy preserving clustering over vertically partitioned data. Here we have presented privacy preserving clustering method over horizontally distributed data. Our work is based on the construction of privacy preserving dissimilarity matrix. This matrix is built of the objects that are stored with different data holders located at different sites and further it is used for clustering purpose. A pair wise comparison is made using individual sensitive and private data objects. These data objects are supposed to be horizontally distributed to different sites and all the sites which are taking part in data mining process are supposed to be the semi-honest. Means these different sites are honest in following the protocol but curious about the mining results. A third party is being involved in mining process with the help of which dissimilarity matrix of objects is being constructed and data mining is applied over aggregated data. In this dissertation we deal with different data types like alphanumeric, categorical with numeric attributes as well. We conduct experiments over both, real datasets and synthetically generated datasets and show computation and communication complexity of our method. Each experiment is also performed for a basic protocol which has no privacy concern to show that the overhead comes with security and privacy by comparing the basic protocol and our protocol. We hope that our implementation will form a benchmark for future research in the area of data mining.

Data mining is primarily used today by companies with a strong consumer focus - retail, financial, communication and marketing organizations. It enables these companies to determine relationships among "internal" factors such as price, product positioning, or staff skills, and "external" factors such as economic indicators, competition, and customer demographics. And, it enables them to determine the impact on sales, customer satisfaction, and corporate profits. Finally, it enables them to "drill down" into summary information to view detail transactional data.

Data mining technology has emerged as a means of identifying patterns and trends fromlarge quantities of data. Data mining and data warehousing go hand-in-hand: most toolsoperate by gathering all data into a central site, then running an algorithm against thatdata. However, privacy concerns can prevent building a centralized warehouse.Data mining is a greatly successful and expanding field that combinesstatistical analysis, machine learning and database technology to extract hidden patterns and relationships from large databases. Data mining as a process of nontrivial extraction of implicit, previously unknown and potentially useful information from the data stored in a database. Data mining is increasing applied to novel and non-traditional types of databases.

There are two types of data distribution models addressed above: **centralized and distributed**. As the name indicates in centralized databases all the data is located at single place that is being aggregated from multiple sources to a single location while in distributed model rather than being locating at single place, data is located at multiple sites. In this case data holders at

different sites need to share their data with each other for the global analysis task or at the time data holders wish to derive aggregate results from datasets that are distributed among all the sites. Here in distributed scenario this data may be partitioned in different ways such that it may be a horizontal partition or a vertical partition. The difference between these partitions is that in horizontal partition 'records' are distributed across multiple sites while in vertical partition 'attributes' are distributed across multiple sites.

LITERATURE SURVEY

Literature Survey provides the background material and related work that gives a required perspective for the work done in this dissertation. This part of work begins with a short summarization of privacy preserving techniques that are to be followed in our work. Section 1.2.2 presents the state of art in database partitions and defines different data models. Next section provides an overview of Secure Multiparty Computation. Section 1.2.4 gives the overview of data matrix. Section 1.2.5 defines the dissimilarity matrix used in our work.

Background

Data mining researchers have developed many methods or protocols that enable data mining techniques to be applied on data but preserving individual's privacy at the same time. Thus to preserve privacy of individuals there is an emerging need of protecting sensitive data. For this purpose researchers have devised mainly two methods: data sanitization and secure multi-party computation. As discussed in previous sections, data sanitization is the technique of trimming or modifying the original database so that sensitive data about the individuals could remain protected, while secure multiparty computation is applied in distributed scenario and mainly use cryptography based encryption to preserve privacy of individuals. From the researcher's point of view the main difference between these two approaches is that data mining on sanitized data finally gives loss of accuracy, while applying secure multi-party computation protocols in distributed scenario give accurate results but at a very high expense of computation or communication costs.

Distributed Data Mining

In practical use, there are two types of data models for the organizational or computing data to be stored. In contrast to the centralized model in which all the data is collected at a single location, the distributed data model assumes that the data sources are distributed across multiple sites. These different sites need to share their data with each other for mining purpose. In this field many efficient algorithms have been proposed which address the problem of efficiently getting the results of mining process applied over all the data across these distributed sites. But main problem arises when the sites do not want to share their data with each other because of the privacy concerns.

Vertical Partitioning

In vertical partitioning, the individual entities may have different attributes (or views) of the same set of records. Vertical partitioning is said to be a heterogeneous distribution of data in which different sites gather information about the same set of entities, they collect different feature sets. For example, financial transaction information is collected by banks, while the IRS collects tax information for everyone.

Horizontal Partitioning

Horizontal partitioning is a data distribution technique using in distributed environment. This type of partitioning is referred to as the homogeneous distribution of data in which different sites collect the same set of information, but about different entities. In horizontally partitioned data sets,the individual records are spread out across multiple entities, each of which have the same set of attributes.

Data Matrix

A data matrix is an ordered structure represented as object-by-variable form. It means each row in the data matrix represents an entity against the values of its attributes stored in columns of that matrix. Generally a data matrix is represented by m×n order which means data matrix has been constructed on the data of m objects on n attributes. It is not necessary to choose the attributes of object from the same domain.

IDENTIFICATION OF PROBLEM AND ISSUES

Advancement in technologies has enabled electronic repositories to be upgraded at a phenomenal rate. This large collection of data need to me mined for getting fruitful information. Data Mining has been a most popular research area among the researchers. Privacy preserving data mining deals with the mining process of data while preserving privacy of individuals at the same time. Data mining has been used for many data models in privacy preserving manner. These models include classification for centralized data, association rule mining in distributed environment, clustering in distributed environment. There are many points in these models that lead to identify a new problem. Thus the problem addressed in this dissertation covers following points that made us to identify this problem.

High accuracy

Privacy preserving techniques enable data mining to mine the collection of data while keeping private records as private. Means no information leak is there while applying privacy preserving techniques. Mainly two privacy preserving techniques are applied to different data models-

(1) Data Sanitization and

(2) Secure Multiparty Computation

Data mining models

There are many data models proposed for mining purpose. These involve classification, association rule mining and clustering of data. Clustering is being less studied s compared to all other techniques. It is being applied on centralized database using data sanitization technique. But in distributed environment it is not applied that much.

Clustering Shapes and Different Data Types

Many clustering techniques deal with the different type of clustering shapes. Generally there are two types of partitioning a cluster used in different algorithms-

(1) Partitioning Method

(2) Hierarchical Method

Solution approach

Previous section has addressed and discussed the problem incorporated in this dissertation. Problem states that all the sites involved in mining process first need to construct a local dissimilarity matrix on their data and a comparison function is shared among all the sites including third party. With the help of local dissimilarity matrices and comparison protocol, the third party constructs a global dissimilarity matrix that is sent to the clustering purpose later. For this purpose comparison protocols are employed on different type of attributes like numeric, alphanumeric or categorical.

Comparison Protocols

Comparison protocol is used to construct a dissimilarity matrix of objects. For each site separately, different comparison protocols are employed. As discussed in previous chapter, a dissimilarity matrix dm is constructed with the differences between the objects or we can say that a dissimilarity matrix is an object-by-object structure in which each entry is the distance between two objects. Consider an entry dm[a][b] in dm. It means dm[a][b] is the distance between objects a and b. But here two cases arise about this entry in dm. First case is: If both objects a and b are consisted by the same site. In this case there is no need of any intervene by the third party to compute the distance between these objects.

Dissimilarity Matrix Construction

In previous section we presented the comparison protocols for different data types. These protocols are the building blocks for the construction of dissimilarity matrix. So in this section, we explain how to build dissimilarity matrices for different data types i.e. numeric, alphanumeric and categorical attributes using the comparison protocols presented in previous section. We have

discussed above that all the sites agree on the list of attributes selected for the clustering technique to be applied on. So the third party T runs the appropriate matrix construction algorithm for every attribute chosen for clustering purpose in order to construct the corresponding dissimilarity matrix. We have selected three data types here. For numeric and alphanumeric attributes, construction algorithms are essentially the same. So a single protocol can be used for both data types. Each data site first constructs its local dissimilarity matrix using the algorithm described below.

Implementation

The details about the experiments conducted for evaluating the performance of proposed protocol. In our distributed clustering protocol, the value of each attribute has been blinded using a random number by adding it to attribute value and later removing it at the time of revealing results. Thus our protocol does not compromise privacy for accuracy. It gives accurate results with no loss in privacy. Therefore, only two tests have been: communication cost analysis and computation cost analysis. The experiments are conducted on an Intel Core i5 PC with 2.30GB RAM, 2MB cache and 2.53GHz clock speed. We used C# programming language to implement the algorithms.

Experimental Setup and performance evaluation

Three test cases are identified to measure the performance of the proposed distributed clustering protocol. In these test cases we use different values for (1) total number of entities (total database size), (2) average length of alphanumeric attributes, and (3) number of sites. The test cases are applied over numeric and alphanumeric attributes to show performance of our protocol over different attribute types. For numeric attributes, we use two different data types: integer and double. However test results for these two data types for numeric attributes are similar; hence due to space consideration, only test results for double data type are included.

Communication Cost Analysis

We discuss the communication cost of our protocol by providing two sets of tests on (1) communication cost of transferring dissimilarity matrices to T, and (2) communication cost of secure pair-wise entity comparisons among different sites.

For the base protocol, there is no communication cost of transferring local dissimilarity matrix since each site sends its dataset to the third party in plaintext. Overall communication cost for the base protocol is O(n+m) where *n* and *m* are dataset sizes of sites I and J, respectively. For our proposed protocol, the cost of communication for transferring the local dissimilarity matrices is $O(n^2 + m^2)$ for both numerical and alphanumeric attributes. while for alphanumeric attributes this value is $O(n*m*l_n*l_m)$ where $l_n$and $l_m$are the average lengths of attributes of sites A and B, respectively. Fig. 3.4 depicts linear behavior of the baseline protocol and quadratic behavior of our protocol with respect to varying database sizes. As seen in Fig. 3.5, communication cost increases dramatically in our protocol due to secure comparison protocol.

Conclusion

In this dissertation, we proposed a method for privacy preserving clustering over horizontally partitioned data. Our method is based on the dissimilarity matrix construction using a secure comparison protocol for numerical, alphanumeric, and categorical data. Our work can be called as an extension to the previous work as that work on privacy preserving clustering over horizontally partitioned data was on a specific clustering algorithm and only for numerical attributes. ***The main advantage of our method is its generality in applicability to different clustering methods such as hierarchical clustering etc.*** The main advantage of hierarchical clustering methods is that it can both discover clusters of arbitrary shapes and deal with different data types. We can cluster numeric data, alphanumeric data that is of outmost importance in bioinformatics researches and categorical data also. Another major contribution is that quality of the resultant clusters can easily be measured and conveyed to data owners without any leakage of private information.

We also provided the security analysis of our protocol and discussed the computation and communication complexity of the protocol. All the experimental results over real datasets and synthetically generated datasets comply with the complexity analysis, also proving that preservation of individual privacy is possible under reasonable assumptions such as non-collusion and semi-honesty of sites and the third party. However, as expected, ensuring privacy has its costs, considering the comparison against the base protocol where private data is shared with third parties. Although we used the proposed secure comparison protocols for clustering horizontally partitioned datasets, there are various other application areas of these methods such as record linkage and outlier detection problems.

Limitations & Future scope

Today in many Industries such as insurance, medical, shopping, banking etc. commonly use data mining to increase sales, reduce costs, and enhance research. While data mining in general represents a significant advance in the type of analytical tools widely available but there are limitations to its capability. These limitations may be in any form. One limitation of this is that although data mining can help reveal patterns and relationships, it never tells the user about the value or significance of these patterns. It does not tell the users which patterns are sensitive and which are not. These types of determinations must be made by the data owner with the help of experts in the domain. A second limitation is that while data mining can identify connections between behaviors, it does not necessarily identify a causal relationship between sites. A successful data mining process still requires skilled technical and analytical specialists who can structure the analysis and interpret the output and then identify the sensitive patterns.

The protocol that we proposed for clustering horizontally partitioned data can easily be adapted to vertically partitioned data in a similar fashion. Since all sites in vertically partitioned data case, possess different attributes of the same entities, each site merely constructs local dissimilarity matrices of the attributes that belongs to them. Normalization is then carried out. After normalization of dissimilarity matrices, sites send local dissimilarities matrices to the third party where each local dissimilarity matrix will be used as global dissimilarity matrix for the corresponding attribute during the clustering process.

The computation costs reduce dramatically for the vertically partitioned data case compared to horizontally partitioned data case since data holders only perform local computation, no cooperation with other sites is required, and no pseudo-random number generation is needed. Accordingly, the computation cost for each site is $O(n^2)$ for numeric and $O(n^2*l^2)$ for alphanumeric attributes, where *n* is the number of entities sites possess, and lis the average length of alphanumeric attribute. Considering communication cost, there is again a reduction with respect to horizontally partitioned data case since sites only send their local dissimilarity matrix to the third party. The communication cost is $O(p*n^2)$ for both numeric and alphanumeric attributes, where pis the number of sites.

REFERENCES

1) R. Agrawal, R. Srikant, 2000. *Privacy Preserving Data Mining*, Proc. of the 2000 ACM SIGMOD Conference on Management of Data 439-450.

2) M. Kantarcioglu, C. Clifton, 2004. *Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data*, IEEE TKDE, 16(9).

3) T. Mitchell, 1997. *Machine Learning*. McGraw Hill.

4) D. Gusfield, 1997. *Algorithms on Strings trees and Strings*. Cambridge University Press.

5) S. Benninga and B. Czaczkes, 1997. *Financial Modelling*. MIT Press.

6) R. Mattison, 1997. *Data Warehousing and Data Mining for Telecommunication*. Artech Press.

7) Januray 1998, Office of the Information and Privacy Commissioner. Data mining: Staking a claim into your privacy. Ontario, Canada.

8) M. Atallah, E. Bertino, A. Elmagarmid, M. Ibrahim, and V. Verykios, November 1999.Disclosure limitation of sensitive rules. In Proceedings of 1999 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX'99), pages 45–52, Chicago, Illinois, USA,IEEE Computer

9) Rakesh Agrawal and RamakrishnanSrikant May 14-19 2000, Privacy-preserving data mining. In Proceedings of the 2000 ACM SIGMOD Conference on Management of Data, pages 439{450, Dallas, TX, ACM.

10) A. C.-C. Yao, , 1986.How to generate and exchange secrets. In *Proc. 27th IEEE Symp. On Foundations of Computer Sciences*, pages 162 167.

11) OdedGoldreich, may 2004. The Foundations of Cryptography, volume 2, chapter 7: General Cryptographic Protocols. Cambridge University Press.

12) Michael Ben-Or, Sha_ Goldwasser, and AviWigderson, May2-4 1988. Completeness theorems for non-cryptographic fault-tolerant distributed computation. In Proceedings of the Twentieth Annual ACM Symposium on Theory of Computing, pages 1{10,Chicago, IL.