RESEARCH ARTICLE OPEN ACCESS

Privacy Preservation in Collaborative Data Mining As Goal Oriented Attack Model

Mr. J.RangaRajesh [1], Mrs. Srilakshmi.Voddelli [2]
Mr. G.Venkata Krishna [3], Mrs. T.Hemalatha [4]
Department of Computer Science and Engineering
DVR & Dr. HS MIC College of Technology [1] & [2]
PSCMR College of Engineering & Technology [3]
Kanchikacherla
K L University, Vaddeswaram [4]
India

ABSTRACT

Data collection is an essential task in data mining process. Data Mining is the process of extracting important knowledge from large databases. Data can be shared between various users but sometimes these collections are split among various parties. While considering security, the data collection from different parties becomes too difficult. Protecting private data is an important concern for society - several laws now require explicit consent prior to analysis of an individual's data. A simple notion of privacy is to protect only the actual data values within any transaction – as long as none of the data is known exactly, privacy is preserved. Privacy concerns may prevent the parties from directly sharing the data and some types of information about the data. How multiple parties collaboratively conduct data mining without breaching data privacy presents a challenge. The main challenge is to maintain the privacy in data mining among the collaboratively distributed system. In particular, we illustrate how to conduct privacy-preserving naive Bayesian classification which is one of the data mining tasks. To measure the level of privacy we propose a define privacy and show that our solutions preserve data privacy. The methods incorporate cryptographic techniques to minimize the information shared, while adding little overhead to the mining task. We are going to addresses secure mining of association rules over vertically partitioned data.

Keywords:- Privacy Preserving, Cryptographic Techniques, Homomorphic, Data mining.

I. OBJECTIVE

The main objective of privacy preserving data mining is to hide sensitive information across populations, rather than reveal information about individuals. The problem is that data mining works by evaluating individual data that leads to privacy issues. Thus, the true problem is not data mining, but the way mining of data is done. However, bringing data together to support data mining makes misuse easier. Much of this information has already been collected, however it is held by various organizations. Separation of control and individual safeguards prevent correlation of this information, providing acceptable privacy in practice. However, this separation also makes it difficult to use the information for purposes that would benefit society, such as monitoring criminal activity. Proposals to share information across different parties, most recently to combat terrorism, would eliminate the safeguards imposed by separation of the information.

1.2. OVERVIEW

The justification of "Privacy Preservation in Collaborative Data Mining as Goal Oriented Attack Model" is done by the implementation of Homomorphic Encryption to secure the mined data between the user and the server. The goal of data mining is to extract or mine knowledge from larger information. However, data is often retrieved by several different sites. Privacy, legal and commercial factors restrict centralized access to the data. Survey results from the area of secure multiparty computation in cryptography proved that assuming the existence of trapdoor permutations; one may provide secure rules for any two party's computation as well as for any multiparty computation with honest majority. However, the normal methods are not efficient and impractical for computing complex functions on inputs consisting of large data sets. The available option is to come up with a set of techniques to achieve data efficiently within a quantifiable security framework. The distributed data model considered is the heterogeneous database scenario with different features of the same set of data being collected by

different sites. This paper projects that it is indeed possible to have efficient and practical techniques for useful privacy-preserving mining of knowledge from large amounts of data. The dissertation presents several privacy preserving data mining algorithms operating over vertically partitioned data. The set of underlying techniques solving independent sub-problems are also presented. Together, these enable the secure "mining" of knowledge.

In today's environment, data collection is ubiquitous, and every transaction is recorded in some location. The resulting data sets can consist of terabytes or even peta bytes of data, so efficiency and growth is the primary consideration for most data mining algorithms. Data mining technology has emerged as a means of identifying patterns and trends from large quantities of data. Most tools operate by gathering all data into a central site, then running an algorithm against that data. However, privacy concerns can prevent building a centralized warehouse and data may be shared among several clients, none of which are allowed to transfer their personal data to another site. The goal is to produce association rules that are globally used while limiting the information shared about each site. Previous analysis in privacy preserving data mining has addressed two issues. Preserving customer privacy by distorting the data values which does not reveal private information and thus is safe to use for mining. The key result is that the distorted data, and information on the distribution of the random data used to distort the data, can be used to generate an approximation to the original data values. Recent enhancements in data collection, data dissemination and related technologies have launched a new era of research where existing data mining algorithms should be reconsidered from the point of view of privacy preservation. The need for privacy is sometimes due to law (e.g., for medical databases) or can be motivated by business interests. However, sharing of data can lead to mutual benefit. Despite the benefit, this is often not possible due to the privacy issues which arise. The goal of this paper is to present techniques to solve privacy-preserving collaborative data mining problems over large data sets with reasonable efficiency. The contributions of this paper contain the following: (1) a projected description of privacy for privacy-preserving collaborative data mining; (2) a solution for naive Bayesian classification with vertical collaboration; and (3) an proficiency analysis to show the performance mounting up with various factors.

1.2.1. DESCRIPTION

Extracting or mining knowledge from large amounts of vertically partitioned data within quantifiable security restrictions is efficiently possible. Knowledge Discovery in Databases (KDD) is the term used to represent the process of extracting knowledge from large amounts of data. The KDD process agree to take that all the data is easily accessible at a central location or through centralized access mechanisms such as collective databases and virtual warehouses. Moreover, developments in information technology and the ubiquity of networked computers have made personal information much more available. Privacy activists have been challenging attempts to loop more and more information into unified collections. Efforts to combine data have even resulted in public disapproval; witness Japan's creation of a national registry containing information previously held by the zones. Data mining in exact has come under blockade, such as the outline of U.S.

1.2.2 PHASES

PHASE – I

COMMUTATIVE ENCRYPTION

Every single party encrypts its own frequent item sets. The encrypted item sets are then handed over to a common party to eliminate duplicates and to begin decryption. This set is then handed to each party and each party decrypts each item set. The final result is the shared item sets.

PHASE- II

RANDOMIZATION

Each of the locally supported item sets is tested to see if it is supported globally. The item set is known to be supported at one or more sites and each computes their local support. The first site chooses a random value R and adds to R the amount by which its support for item set exceeds the minimum support threshold. This value is passed to site 2, which again adds its excess support. The resulting value is tested using a secure comparison to see if it exceeds the Random value. If so, item set is supported globally.

1.2.3. PROCESS

Our technique follows the straightforward methodology except that values are passed between the local data mining sites rather than to a central combiner. The two levels are realizing candidate item sets (those that are frequent on one or more sites), and determining which of the candidate item- sets meet the global support/confidence thresholds. The first phase practices commutative encryption. Each party encrypts its own item sets, and then the (already encrypted) item sets of every other party. These are passed from one place to another, with every site decrypting, to obtain the complete set. In the second phase, an originating party passes its support count, plus a random value, to its neighbour. The neighbour adds its support count and passes it on. The concluding party then engages in a secure comparison with the originating party to determine if the final result is greater than the threshold plus the random value.

1.2.3.1. KNOWLEDGE DISCOVERY

In this module we just show the data collections and the available sites. Also we discover the important knowledge from the data collections. Secret information and some important data cannot be shared at all sites. The data base is maintained under only one organization. Because this organization maintaining the secret information for different sites. Also each site updates their information in database whenever they want. Extract knowledge from this database based on secret required for sites. This privacy information should be considered secrets and that have to be protected from others.

1.2.3.2. SPLIT DATA AMONG VARIOUS SITES

The extracted information is a privacy or secret information combined with data, and these are all distributed among various sites. Each site receives information from collections called secret and data sharing. Data collections shared for all sites to share the information. But we have to restrict secret data from the sites. Each party at different sites shares the information from the data collections.

1.2.3.3. ENCRYPTION

Sites are having extracted data from collections. Each party encrypts its own frequent item sets. This encryption is done instantly after getting data. The encrypted item sets are then passed to other parties until all parties have encrypted all item sets. So, all sites item sets are in encrypted format. These are passed to a common party to eliminate duplicates and to begin decryption. Duplicate items are eliminated at last and this process has been done by common party. So every party is encrypted its own frequent item sets in this module. Choose any random number for discovering support from various sites. Random number will be added by each sites then this number is considered to discover global support. The authorized site should have supported globally to share secret information. Other sites will not support globally then we can identify these parties should not get secret. At last using this technique we just identify he valuable site for sharing the secret information.

1.2.3.4. DECRYPTION

The parties decrypt the information which was encrypted already and their frequent item sets were globally supported. These valid parties have the key to decrypt the data. One important concern is that only the valid parties can share secret from the data collections. Secrets will be shared among varies parties only who are all valid. These secrets are highly secured and only accessible by the valid users.

1.3. PROJECT CATEGORY:

1.3.1. DATA MINING

Data mining is "the process of extraction of implicit, formerly unfamiliar, and theoretically useful information from data" and "the art of mining useful information from large data sets or databases". Even if it is usually used in relation to inquiry of data, data mining, like artificial intelligence, is an umbrella term and is used with varied meaning in a wide range of circumstances. It is usually connected with a business or other organization's need to classify trends. Data mining includes the process of analyzing data to show patterns or relationships; arranging through large amounts of data; and picking out parts of relative information or patterns that occur e.g., picking out facts from some data. A simple example of data mining is its use in a retail sales department. If a store tracks the purchases of a customer and notices that a customer buys a lot of T-shirts, the data mining system will make a correlation between that customer and T-shirts. The sales department will look at that information and may begin direct mail marketing of T-shirts to that customer, or it may alternatively attempt to get the customer to buy a wider range of products. In this case, the data mining system used by the retail store discovered new information about the customer that was previously unknown to the company. Another widely used example is that of a very large South Indian chain of supermarkets. Through rigorous analysis of the transactions and the goods bought over a period of time, analysts found that energy drinks and diapers were often bought together. Though explaining this interrelation might be difficult, taking advantage of it, on the other hand, should not be hard (e.g. placing the high-profit diapers next to the high-profit E-drinks). This technique is often referred to as Market Basket Analysis. In statistical analysis, in which there is no underlying theoretical model, data mining is often approximated via stepwise regression methods wherein the space of 2^k possible relationships between a single outcome variable and k potential explanatory variables is smartly searched. With the arrival of parallel computing, it became possible to inspect all 2^k models. This technique is called all subsets or exhaustive regression. Some of the first applications of exhaustive regression involved the study of plant data.

Usually, data mining (also called data or knowledge discovery) is the process of analyzing and grouping data from different data stores and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of many analytical tools for analysis of data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns

among dozens of fields in large relational databases. Companies have used powerful computers to inspect large volumes of supermarket scanner data and analyze market research reports for a long time. However, continuous modernizations in computer processing power, disk storage, and statistical software are dramatically increasing the accuracy of analysis while reducing the cost. For example, one mythical Midwest grocery chain used the data mining capacity of Oracle software to analyze local purchasing patterns. They discovered that when men bought diapers on Tuesdays and Saturdays, they also tended to buy Energy drinks. Further analysis showed that these shoppers typically did their weekly grocery shopping on Saturdays. On Tuesdays, however, they only bought a few items. The retailer concluded that they purchased the Energy drinks to have it available for the upcoming weekend. The grocery chain could use this newly discovered information in various ways to increase revenue. For example, they could move the beer display closer to the diaper display. And, they could make sure Energy drinks and diapers were sold at full price on Thursdays.

1.3.2. DISTRIBUTED DATA MINING

In distinction to the centralized model, the Distributed Data Mining (DDM) model assumes that the data sources are distributed across multiple sites. Algorithms developed within this area address the problem of efficiently getting the mining results from all the data across these distributed sources. Since the key focus is on efficiency, most of the algorithms which have been already developed do not take security factor into account. However, they are still useful in formulating the context of the problem. A simple approach in multiple sources will not share data to run existing data mining tools at each site independently and combine the results. However, this will often fails to give global convincing results. Issues that cause a inequality between local and global results include: Values for a single entity may be divided across sources. Data mining at individual sites will not be able to detect cross-site correlations. The same data item may be duplicated at muliple sites, and will be over-weighted in the outcomes. Data at a single site is likely to be from an identical population.

1.3.3. DATA DREDGING

Used in the technical context of data warehousing and analysis, the term "data mining" is neutral. However, it sometimes has a more pejorative usage that implies imposing patterns (and particularly causal relationships) on data where none exist. This imposition of irrelevant, misleading or trivial attribute correlation is more properly criticized as "data dredging in the statistical literature. Another term for this misuse of statistics is data fishing. Used in this latter sense, data dredging implies scanning the data for any relationships, and then when one is found coming up with an interesting explanation. The problem is that large data sets invariably happen to have some exciting relationships peculiar to that data. Therefore any conclusions reached are likely to be highly suspect. In spite of this, some exploratory data work is always required in any applied statistical analysis to get a feel for the data, so sometimes the line between good statistical practice and data dredging is less than clear. One common approach to evaluating the fitness of a model generated via data mining techniques is called cross validation. Cross validation is a technique that produces an estimate of generalization error based on resampling. In simple terms, the general idea behind cross validation is that dividing the data into two or more separate data subsets allows one subset to be used to evaluate the generalize ability of the model learned from the other data subset(s). A data subset used to build a model is called a training set; the evaluation data subset is called the test set. Common cross validation techniques include the holdout method-fold validation and the leave-one –out method. Another pitfall of using data mining is that it may lead to discovering correlations that exist due to chance rather than due to an underlying relationship. "There have always been a considerable number of people who busy themselves examining the last thousand numbers which have appeared on a roulette wheel, in search of some repeating pattern. Sadly enough, they have usually found it.". However, when properly done, determining correlations in investment analysis has proven to be very profitable for statistical operations (such as pairs trading strategies), and furthermore correlation analysis has shown to be very useful in risk management. Indeed, finding correlations in the financial markets, when done properly, is not the same as finding false patterns in roulette wheels. Most data mining efforts are focused on developing highly detailed models of some large data set. Other researchers have described an alternate method that involves finding the minimal differences between elements in a data set, with the goal of developing simpler models that represent relevant data. [8]

1.3.4 PRIVACY CONCERNS:

Privacy concerns are also associated with data mining - specifically regarding the source of the data analyzed.

For example, if an insurance employee r has access to medical records, they may screen out people who have diabetes or have had a heart attack or Kidney problem. Identifying such employees will cut costs for insurance, but it creates ethical and legal problems. Another Example is mining government or commercial data sets for national security or law enforcement purposes have also raised privacy concerns. There are many real uses of data mining. For example, a database of prescription drugs taken by a group of individuals could be used to find combinations of drugs exhibiting harmful interactions. Since any particular combination may occur in only 1 out of 100 people, a great deal of data would need to be examined to discover such an interaction. A project involving pharmacies could reduce the number of drug reactions and potentially save lives. Unfortunately, there is also a huge potential for misuse of such a database. Essentially, data mining gives information that would not be available otherwise. It must be properly interpreted to be useful. When the data collected involves individual people, there are many questions concerning confidentiality, rightfulness and ethics.

1.3.4.1. Where we need privacy?

Identifying public health problem out breaks (e.g, epidemics, and biological warfare instances). There are many data collectors (insurance companies, HMOs, public health agencies). Individual privacy concerns will limit the willingness of the data custodians to share data, even with government agencies such as the Centers for Disease Control. Can we accomplish the desired results while still preserving privacy of individual entities?[9]

II. EXISTING SYSTEM

- Data collection, data dissemination and related technologies in the existing data mining algorithms should be reconsidered for privacy preservation. Privacy is not possible because of intruders within the network. Threats against privacy are very common. Lack of security. Common Accessibility. Maintain the privacy of individuals from others.

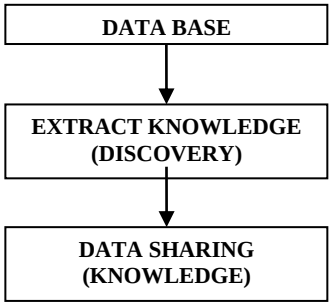
2.2. PROPOSED SYSTEM:

Privacy is main factor for privacy-preserving collaborative data mining. Provides a solution for naive Bayesian classification with vertical collaboration. Mine distributed association rules on vertically partitioned data. An efficiency analysis to show the performance scaling up with various factors. Homomorphic encryption, digital envelope technique and the attack models are employed. Privacy-preserving collaborative data mining solves problems over large data sets with reasonable efficiency.

III. MODULES

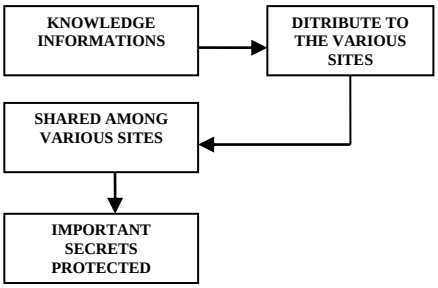
3.1. DATA COLLECTIONS AND KNOWLEDGE DISCOVERY:

In this module we just show the data collections and the available sites. Also we discover the important knowledge from the data collections. Secret information and some important data cannot be shared at all sites. The data base is maintained under only one organization. Because this organization maintaining the secret information for different sites. Also each site updates their information in database whenever they want. Extract knowledge from this database based on secret required for sites. This privacy information should be considered secrets and that have to be protected from others.



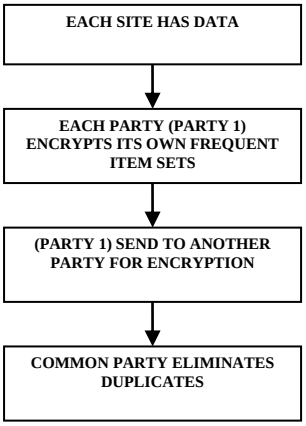
3.2. SPLIT DATA FROM COLLECTIONS AMONG VARIOUS SITES

The extracted information is a privacy or secret information combined with data, and these are all distributed among various sites. Each site receives information from collections called secret and data sharing. Data collections shared for all sites to share the information. But we have to restrict secret data from the sites. Each party at different sites shares the information from the data collections.



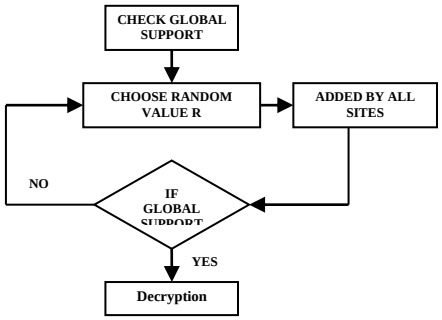
3.3. COMMUTATIVE ENCRYPTION

Sites are having extracted data from collections. Each party encrypts its own frequent item sets. This encryption is done instantly after getting data. The encrypted item sets are then passed to other parties until all parties have encrypted all item sets. So, all sites item sets are in encrypted format. These are passed to a common party to eliminate duplicates and to begin decryption. Duplicate items are eliminated at last and this process has been done by common party. So every party is encrypted its own frequent item sets in this module.



3.4 CHECK THE ITEM SETS FOR GLOBAL SUPPORT

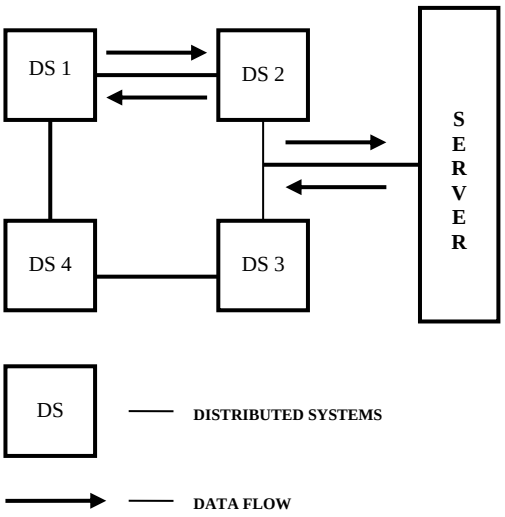
The encrypted item sets are checked for its global support. This global support is identified because we have to know which site has this support to share secret information. Choose any random number for discovering support from various sites. Random number will be added by each sites then this number is considered to discover global support. The authorized site should have supported globally to share secret information. Other sites will not support globally then we can identify these parties should not get secret. At last using this technique we just identify the valuable site for sharing the secret information.



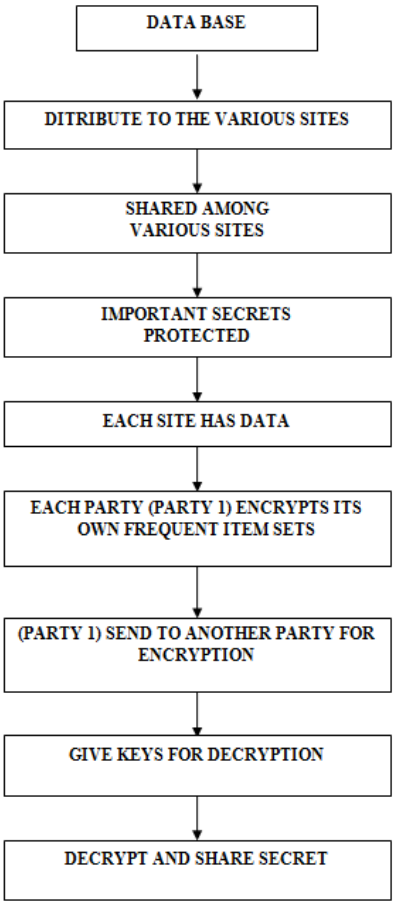
3.5 DECRYPTION AND SECRET SHARING MODULE

The parties decrypt the information which was encrypted already and their frequent item sets were globally supported. These valid parties have the key to decrypt the data. One important concern is that only the valid parties can share secret from the data collections. Secrets will be shared among varies parties only who are all valid. These secrets are highly secured and only accessible by the valid users.

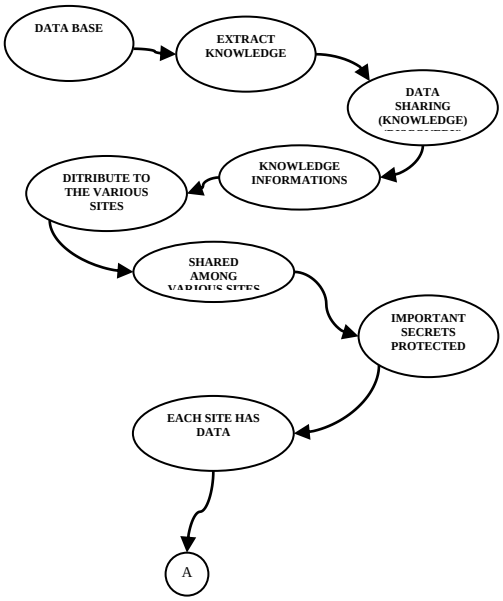
IV. SYSTEM DESIGN

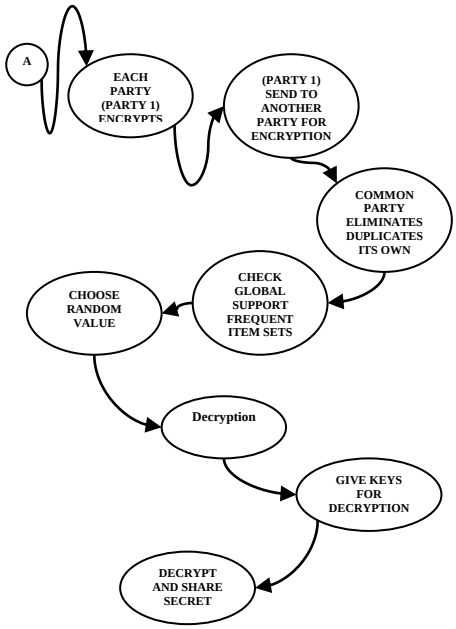


V. OVERALL BLOCK DIAGRAM



VI. DATAFLOW DIAGRAM





VII. ENCRYPTION

Sites are having extracted data from collections. Each party encrypts its own frequent item sets. This encryption is done instantly after getting data. The encrypted item sets are then passed to other parties until all parties have encrypted all item sets. So, all sites item sets are in encrypted format.

These are passed to a common party to eliminate duplicates and to begin decryption. Duplicate items are eliminated at last and this process has been done by common party. So every party is encrypted its own frequent item sets in this module.

Choose any random number for discovering support from various sites. Random number will be added by each sites then this number is considered to discover global support. The authorized site should have supported globally to share secret information. Other sites will not support globally then we can identify these parties should not get secret. At last using this technique we just identify the valuable site for sharing the secret information.

7.1. DECRYPTION

The parties decrypt the information which was encrypted already and their frequent item sets were globally supported. These valid parties have the key to decrypt the data. One important concern is that only the valid parties can share secret from the data collections. Secrets will be shared among varies parties only who are all valid. These secrets are highly secured and only accessible by the valid users.

7.2 DISADVANTAGES OF PREVIOUS TECHNIQUES

Privacy is not possible because of intruders within the network..Threats against privacy are very common and lack of security..Common Accessibility.

7.3 ADVANTAGES

Privacy is main factor for privacy-preserving collaborative data mining..Provides a solution for naive Bayesian classification with vertical collaboration..Mine distributed association rules on vertically partitioned data..An efficiency analysis to show the performance scaling up with various factors. Homomorphic encryption, digital envelope technique and the attack models are employed. Privacy-preserving collaborative data mining solves problems over large data sets with reasonable efficiency.

VIII. LITERATURE SURVEY

Two parties own confidential databases wish to run a data mining algorithm on the union of their databases, without revealing any unnecessary information. Our previous work is motivated by the need to both protect privileged information and enable its use for research or other purposes. The above problem is a specific example of secure multiparty computation and as such, can be solved using known generic protocols. However, data mining algorithms are typically complex and furthermore, the input usually consists of massive data sets. The generic protocols in such a case are of no practical use and therefore more efficient protocols are required. This focuses on the problem of decision tree learning with the popular ID3 algorithm. Our protocol is considerably more efficient than generic solutions and demands both very few rounds of communication and reasonable bandwidth. Secure two party computations were first investigated by Yao and were later generalized to multi-party computation. Expectation Maximization (EM) algorithm for distribution reconstruction is more effective than the currently available method in terms of the level of information loss. EM algorithm converges to the maximum likelihood estimate of the original distribution based on the perturbed data. When a large amount of data is available, the EM algorithm provides robust estimates of the original distribution. We propose metrics for quantification and measurement of privacy preserving data mining algorithms. Thus, this paper provides the foundations for measurement of the effectiveness of privacy preserving data mining algorithms.

Our privacy metrics illustrate some interesting results on the relative effectiveness of different perturbing distributions. Data are altered in such a way that actual individual data values cannot be recovered in order to avoid exposure, while certain computations can still be applied to the data. Due to the fact that the actual data are not provided for the mining, the privacy of data is preserved. This is the core idea of randomization-based techniques. The random perturbation technique is usually realized by adding noise or uncertainty to actual data such that the actual values are prevented from being discovered. Since the data no longer contains the actual values, it cannot be misused to violate individual privacy.

Randomization approaches proposed by Agrawal and Srikant [2] to solve the privacy-preserving data mining problem addresses access to precise information in individual data records, since the primary task in data mining is the development of models about aggregated data. The underlying assumption is that a person will be willing to selectively divulge information in exchange for useful information that such a model can provide.

Agrawal and Aggarwal [1] showed that the EM algorithm converges to the maximum likelihood estimate of the original distribution based on the perturbed data. Evfimievski et.al. presented a framework for mining association rules from transactions consisting of categorical items where the data has been randomized to preserve privacy of individual transactions. While it is feasible to recover association association rules and preserve privacy using a straightforward *uniform* randomization, the discovered rules can unfortunately be exploited to find privacy breaches. They analyzed the nature of privacy breaches and proposed a class of randomization operators that are much more effective than uniform randomization in limiting the breaches. Du and Zhan proposed a technique for building decision trees using randomized response techniques which were developed in the statistics community for the purpose of protecting surveyeesí privacy. The randomization- based methods have the benefits of efficiency.

The drawbacks are that post-randomization data mining results are only an approximation of pre-randomization results. Secure Multi-party Computations (SMC) deal with computing any function on any input in a distributed network. Each participant holds one of the inputs while ensuring that no more information is revealed to a participant in the computation than can be inferred from that participant's input and output. The SMC problem was introduced by Yao. It has been proved that for any polynomial function, there is a secure multi-party computation solution. The approach used is as follows: the function *F* to be

computed is first represented as a combinatorial circuit, and then the parties run a short protocol for every gate in the circuit. Every participant gets corresponding shares of the input wires and the output wires for every gate. This approach, though appealing in its generality and simplicity, is highly impractical for large data sets. Following the idea of secure multiparty computation, people designed privacy-oriented protocols for the problem of privacy- preserving collaborative data mining.

Lindell and Pinkas [5] first introduced a secure multi-party computation technique for classification using the ID3 algorithm, over horizontally partitioned data. Specifically, they consider a scenario in which two parties owning confidential databases wish to run a data mining algorithm on the union of their databases, without revealing any unnecessary information. Du and Zhan proposed a protocol for making the ID3 algorithm privacy-preserving over vertically partitioned data. Vaidya and Clifton presented protocols for privacy-preserving association rule mining over vertically partitioned data. Encryption is known for preserving the confidentiality of information. In comparison with the other techniques described, a strong encryption scheme can be more effective in protecting the data privacy. An encryption system normally requires that the encrypted data should be decrypted before making any operations on it. For example, if the value is hidden by a randomization-based technique, the original value will be disclosed with certain probability. If the value is encrypted using a semantic secure encryption scheme, the encrypted value provides no help for an attacker to find the original value. One such scheme is the homomorphic encryption scheme which was originally proposed in with the aim of allowing certain computations performed on encrypted data without preliminary decryption operations. To date, there are many such systems. Homomorphic encryption is a very powerful cryptographic tool and has been applied in several research areas such as electronic voting, on-line auctions, etc. is based on homomorphic encryption where Wright and Yang applied homomorphic encryption to the Bayesian networks induction for the case of *two* parties.

Zhan et. al. [7] proposed a cryptographic approach to tackle collaborative association rule mining among multiple parties. In this paper, we will apply homomorphic encryption and digital envelope techniques [3] to privacy-preserving data mining and use them to design privacy-oriented protocols for privacy-preserving naïve Bayesian classification problem. The preliminary idea of this paper has been published in [6].

REFERENCES

[1] D. Agrawal and C. Aggarwal, "On the design and quantification of privacy preserving data mining algorithms," In *Proceedings of the 20th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, Santa Barbara, CA, pp. 247–255, May 21–23, 2001.

[2] R. Agrawal and R. Srikant, "Privacy-preserving data mining," In *Proceedings of the ACM SIGMOD Conference on Management of Data*, ACM Press, pp. 439–450, May 2000.

[3] D. Chaum, "Security without identification," In *Communication of the ACM*, vol. 28, no. 10, pp. 1030–1044, Oct. 1985.

[4] D. Dolev, D. Dwork, and M. Naor, "Non-malleable cryptography," In *Proceedings of the Twenty-third Annual ACM Symposium on Theory of Computing*, New Orleans, Louisiana, USA, pp. 542–552, 1991.

[5] Y. Lindell and B. Pinkas, "Privacy preserving data mining," In *Advances in Cryptology— Crypto2000, Lecture Notes in Computer Science*, vol. 1880, 2000.

[6] Z. Zhan and S. Matwin, "A crypto-approach to privacy preserving data mining," In *IEEE International Workshop on Privacy Aspect of Data Mining*, Hong Kong, Dec. 18–22, 2006.

[7] Z. Zhan, S. Matwin, and L. Chang, "Privacy-preserving collaborative association rule mining," In *19th Annual IFIP WG 11.3 Working Conference on Data and Applications Security*, University of Connecticut, Storrs, CT, U.S.A., Aug. 7–10, 2005.

[8] http://dataminingnora.blogspot.in/

[9] https://www.cs.purdue.edu/homes/clifton/icdm02 /psdm.pdf