International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)
ISSN: 0976-1353 Volume 13 Issue 3 –MARCH 2015.

M-Privacy For Set Valued Data Publishing To Avoid Insider Attack

R. Vinoth
M.E. 2nd year Student
Department Of Computer Science and Engineering
RRASE College of Engineering
vinuworld@gmail.com

Prof. K. RaviKumar
Department Of Computer Science and Engineering
RRASE College of Engineering

***Abstract*— In this paper, we consider the problem of vertically published set valued data at multiple data providers. Data providers may access data records of other data providers by using their own data record. This type of attack is known as insider attack or data provider attack. To avoid this attack, we use the notation m-privacy that guarantees the privacy against group of m-colluding data providers. Then by using provider aware anonymization algorithm the system ensures data utility and efficiency. We also introduce cleaning of data and K-means clustering before applying m-privacy. Experiment results show that our approach achieves better performance than any other privacy preserving algorithms.**

***Keywords*—insider attack, m-privacy, data cleaning, K-means clustering**

I. INTRODUCTION

Everyone is sharing their data which may contain personal information. Providing security and privacy for data that is shared is a very important as well as difficult Problem. Privacy preserving data analysis and data publishing steps used to product data [2]. Data can be shared between two persons or distributed among multiple owners. In two ways data distribution takes place. One is by using TTP (Trusted Third Party) and SMC (Secure Multiparty Computation). The other way is distributing directly without any centralized party.

Centralized and Distributed Anonymization for High-Dimensional Healthcare Data [3] defines the privacy concern of sharing the patient information between the Hong Kong Red Cross BTS (Blood Transfusion Service) and public hospitals. It contains sensitive data as patient details and donor details which should not be revealed. Like that now preserving data from various attacks is a challenging task. Many types are available that are discussed in [5] and [6].

Data Provider attack

Each Data owner may infer data record of other data provider with his own data record. This phenomenon is named as insider attack. Attacks by externals may be avoided by m-privacy[1]. But attacks by data provider, needs some more care. Provider aware anonymization algorithm checks for m-privacy and avoids insider attack by using background knowledge of data provider.

II. PROPOSED WORK

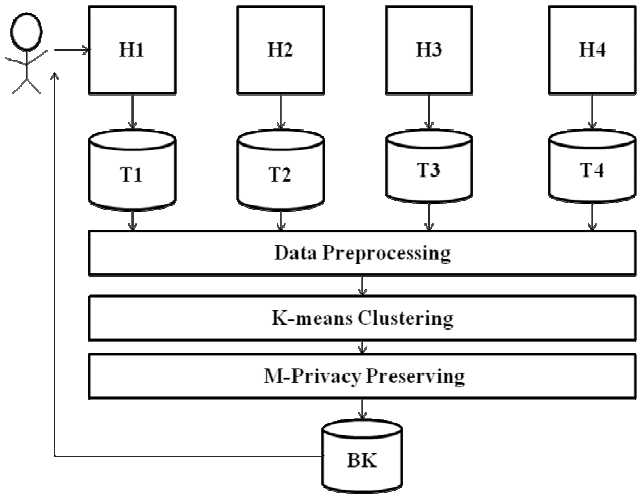

Fig 1. Proposed architecture

We propose a system that speed up the system when compared with other mechanism. Additionally it uses the few steps that are data preprocessing (data cleaning), classification and clustering. These concepts are discussed in the following sections. Fig 1 shows our proposed architecture clearly.

A. Data Preprocessing

Today's real world databases are highly susceptible to noise, missing, and inconsistent data due to their huge size and their likely origin from multiple, heterogeneous sources. Low-quality data may affect quality of mining results. There are numerous data preprocessing techniques are available.

- *Data cleaning* removes noise and corrects the inconsistency of the data.
- *Data integration* collects data from multiple sources and combines them into a single and coherent data collection, such as a data warehouse.

- *Data transformations*, such as normalization improve the accuracy and efficiency of data. It removes redundancy.
- *Data reduction* reduces the data size by clustering, eliminating redundant data, and aggregation for instance.

These techniques are not mutually exclusive meaning that they may work together. For example, data cleaning can involve transformations to correct false data. Data preprocessing techniques, when applied before mining, substantially improve the overall quality of the patterns mined and the time required for the actual mining. In other words, the data you wish to analyze by data mining techniques are incomplete, noisy and inconsistent. Data cleaning work is to "clean" the data by replacing missing values, smoothing the noisy data, identifying or removing outliers, and correcting inconsistencies of data. If the data handler believe the data are dirty, they are unlikely to trust the results of any data mining that has been applied to it. Furthermore, dirty data can cause confusion for the mining procedure, resulting in unreliable output. Data integration combines the data from multiple sources. Data transformation operations, such as normalization and aggregation, are procedures that would contribute toward the success of the mining process. Data reduction obtains a reduced representation of the data set that is much smaller in volume, yet produces the same (or almost the same) analytical results.

B. K-means Clustering

The process of grouping a set of objects into a single class of *similar* objects is called clustering. A cluster is a collection of data objects that are *similar* to one another within the same cluster and are *dissimilar* to the objects in other clusters. Similarity is commonly defined in terms of how close the objects are in space, based on distance. A cluster of data objects can be treated collectively as one group and so it may be considered as a form of data compression.

Clustering techniques apply when there is no class to be predicted but the instances are to be divided into natural groups. Cluster analysis has 4 basic concepts or steps.

1. **Feature selection or extraction**: Feature selection chooses distinguishing features from a set of candidates, while feature extraction utilizes some transformations to generate useful and novel features from the original ones. Clearly, feature extraction is potentially capable of producing features that could be of better use in uncovering the data structure. However, feature extraction may generate features that are not physically interpretable, while feature selection assures the retention of the original physical meaning of the selected features. In the literature, these two terms sometimes are used interchangeably without further identifying the difference. Both feature selection and feature extraction are very important to the effectiveness of clustering applications.

2. **Clustering algorithm design or selection:** This step usually consists of determining an appropriate proximity measure and constructing a criterion function. Intuitively, data objects are grouped into different clusters according to whether they resemble one another or not. Almost all clustering algorithms are explicitly or implicitly connected to some particular definition of proximity measure. Some algorithms even work directly on the proximity matrix. Once a proximity measure is determined, clustering could be constructed as an optimization problem with a specific criterion function. Again, the obtained clusters are dependent on the selection of the criterion function. Thus the subjectivity of cluster analysis is inescapable.

3. **Cluster validation:** Given a data set, each clustering algorithm can always produce a partition whether or not there really exists a particular structure in the data. Moreover, different clustering approaches usually lead to different clusters of data, and even for the same algorithm, the selection of a parameter or the presentation order of input patterns may affect the final results. Therefore, effective evaluation standards and criteria are critically important to provide users with a degree of confidence for the clustering results.

4. **Result interpretation.** The ultimate goal of clustering is to provide meaningful insights from the original data so that user can develop a clear understanding of the data and therefore effectively solve the problems encountered.

There are so many numbers of clusterings are available. Hierarchical clustering, Partitional clustering, Neural network based clustering, kernel based clustering and so on. The classic clustering technique is called k-means. First, user specifies in advance how many clusters are being sought: This is the parameter k. Then k points are chosen at random as cluster centers. All instances are assigned to their closest cluster center according to the ordinary Euclidean distance metric. Next the centroid, or mean, of the instances in each cluster is calculated, this is the "means" part. These centroids are taken to be new center values for their respective clusters. Finally, the whole process is repeated with the new cluster centers. Iteration continues until the same points are assigned to each cluster in consecutive rounds, at which stage the cluster centers have stabilized and will remain the same forever. This clustering method is simple and effective. It is easy to prove that choosing the cluster center to be the centroid minimizes the total squared distance from each of the cluster's points to its center. K-means clustering comes under partitional clustering.

C. M-Privacy

Goryczka, Xion and Fung [1] elaborately discussing about m-privacy in their paper. It avoids attacks by external using anonymized data. Given n data providers, a set of records T, and an anonymization mechanism A, an m-adversary I is a coalition of m-data providers, which jointly contributes a set of records. Sanitized records T* = A (T) satisfy *m*-privacy, i.e. are m-private, with respect to a privacy constraint C.

M-Privacy and Weak Privacy

Given a weak Privacy constraint C that does not consider instance level background knowledge, T* satisfying *C* will only guarantee 0-privacy with respect to C, i.e. C is not guaranteed to hold for each equivalence group after excluding records belonging to any malicious data provider[1]. Thus, each data provider may be able to breach privacy of records provided by others. Commonly more possibility of insider attacks. Privacy is defined with respect to a privacy constraint C, and so it will inherit the strengths and weaknesses of C. For example, if *C* is defined by k-anonymity, then ensuring m-privacy with respect to Constraint C will not protect against homogeneity attack [6] or deFinetti attack [5]. However, m-privacy with respect to C will protect data against a privacy attack issued by any m-adversary, if and only if, C protects against the same privacy attack by any external data recipient using anonymized data. M-Privacy constraint is substantially orthogonal to the privacy constraint C being used.

M-Privacy and Differential Privacy

Differential privacy [2], [3] does not assume specific background knowledge and guarantees privacy even if an attacker knows all records except the victim record. Thus, any statistical data that satisfying differential privacy also satisfies (n-1) privacy, i.e. maximum level of m-privacy, when any (n -1) providers can collude. While m-privacy with respect to any weak privacy notion does not guarantee unconditional privacy. In the rest of the paper, we will focus on checking and achieving m-privacy.

D. Background Knowledge

Data provider attacks are reduced by using high level security. Knowledge about the data, is an important aspect used as a security key. Set valued data increase the level of background knowledge[10]. Whenever the new patient enters his details, it is need to provide two or more unique information such as contact number, mail id and so on. Even though the attacker is also a data provider, He cannot infer other providers data. Thus system helps to identify the attacker easily at the same time it gives awareness to the data provider. Hence attackers are identified and attacks are avoided.

III. EXPERIMENT RESULTS

A. Data set

We used combination of training set and test set of the patient data set. These contain several data records with minimum of 100 entries collected from various hospitals. Here set valued data used. Set valued data provide extra security. Patient may have one or more unique contact number that is used to identify them in critical situations. It helps to avoid unauthorized access of data.

B. Data Preprocessing and classification

Data cleaning is simply known as filtering process. It performs various tasks like replacing missing values, normalization, resampling, attribute selection, transforming and combining attributes, etc. It helps to avoid null values, noisy, redundancy, dependency and inconsistency of data. Classification process used for fast retrieval of data. It minimizes searching time.

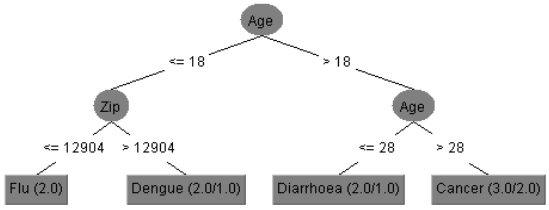

Fig.2 Example for data classification

For an example, A child specialist doctor views patient records only those who are under the age of 12. Doctor need not access all other patient records which is a time consuming process. This is shown in the above fig 2. Record is classified by patient age. Data classification process classifies the record by various attributes. Some time it may need to classify record by means of disease or by means of age. It is depending upon the application where the record to be classified.

C. K-means Clustering & M-Privacy

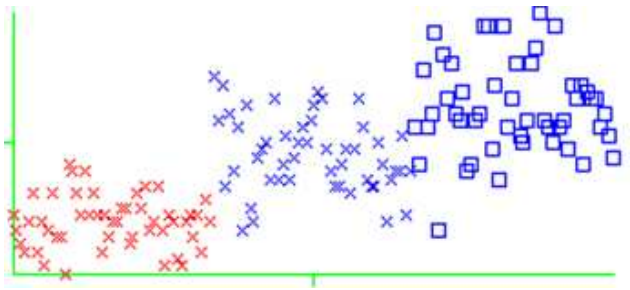

Fig 3.K-means Clustering

K-means algorithm is regarded as a staple of clustering methods due to its ease of implementation. It works well for many practical problems, particularly when the resulting clusters are compact and hyperspherical in shape. Here, this algorithm clusters the data record based on the attribute disease. For an example, the heart specialist doctor can access the patient record that is suffered by heart diseases and need not to view all

other records which is a time consuming process. Hence the clustering process reduces the response time. So, K-means is a good selection for clustering largescale data sets. Moreover, several methods have been proposed to speed up K – means clustering.

IV. RELATED WORK

M-privacy for collaborative data publishing [1] deals with collaborative data publishing which needs TTP/SMC and it is possible for data provider attacks. Differential privacy: a survey of results [2] concerning with statistical database.It does not provide an absolute guarantee of privacy and only statistical database has been taken in account which may result in possibility of realistic attacks. In Privacy-Preserving Data Publishing: A Survey of Recent Developments [4] paper differentiates the data collection process from the data publication process. Data publication process may have trusted data publisher or untrusted data publisher. If the data Publisher is a trusted party then it ensures secure transaction. unless the system leads to privacy breaching.

In the paper Centralized and Distributed Anonymization for High Dimensional Healthcare Data[3] studied that both centralized and distributed data have some demerits. They mean Centralized: Integrate then generalize and Distributed: Generalize then integrate. They proposed the new type of data collection LCM. Terrovitis [11] defined the set operations like union, intersection and element reduction. Mathew [8] defined various attacks such as masquerading, privilege abuse, abuse legitimate privileges, snooping or data-harvesting.

V. CONCLUSION

Data mining is the wide area that is connected with all other areas. Data is the basic thing needed for all types of methods. So preserving the data is becoming a very big challenge now. Here, Data preserving is achieved through some sort of process like background knowledge. There are several types of privacy methods are available. Differential privacy, m-privacy, privacy preserving techniques and so on.

Identifying attacks is the most important thing when handling data. Here, a new type of “insider attack” is identified and avoided by using background knowledge to the data provider and by using provider aware anonymization techniques. In this project m-privacy deals with set valued data. It would be also interesting to verify if this method can be adapted to other kinds of data such as tuples.

REFERENCES

[1] S.Goryczka, L.Xiong, B.Fung , "m-Privacy for Collaborative Data Publishing," Knowledge and Data Engineering, IEEE Transactions on , vol.PP, no.99, pp.1,1

[2] C. Dwork, “Differential privacy: a survey of results,” in Proc. of the 5th Intl. Conf. on Theory and Applications of Models of Computation, 2008, pp. 1–19.

[3] N. Mohammed, B. C. M. Fung, P. C. K. Hung, and C. Lee, “Centralized and distributed anonymization for high-dimensional healthcare data,” ACM Transactions on Knowledge Discovery from Data (TKDD), vol. 4, no. 4, pp. 18:1–18:33, October 2010.

[4] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, “Privacy-preserving data publishing: A survey of recent evelopments,” ACM Comput. Surv.,vol. 42, pp. 14:1–14:53, June 2010.

[5] D. Kifer, “Attacks on privacy and definetti’s theorem,” in Proc. of the 35th SIGMOD Intl. Conf. on Management of Data, 2009, pp. 127–138.

[6] Raymond Chi-Wing Wong, Adawai-Chee Fu, Ke Wang, Jian Pei ‘Minimality Attack in Privacy Preserving Data Publishing’ VLDB ‘07, September 2328, 2007, Vienna, Austria. Copyright 2007 VLDB Endowment, ACM 9781595936493/07/09.

[7] C. Dwork, “A firm foundation for private data analysis,” Commun. ACM, vol. 54, pp. 86–95, January 2011.

[8] S.Mathew, P.Michalis, Hung Q.Ngo., and S.Upadhyaya, S.Jha, R.Sommer., and C.Kreibich ‘A Data-Centric Approach to Insider Attack Detection in Database Systems,’LNCS 6307, pp. 382 -401.Springer-Verlag Berlin Heidelberg,2010.

[9] R. Burke, B. Mobasher, R. Zabicki, and R. Bhaumik, “Identifying attack models for secure recommendation,” in In Beyond Personalization: A Workshop on the Next Generation of Recommender Systems, 2005.

[10] M.Terrovitis, ‘Privacy preserving publication of set-valued data’, 2010.

[11] Lea Kissner, Dawn Song, ‘Privacy Preserving Set Operations’, CMU-CS 2005.

[12] Y. Lindell and B. Pinkas, “Secure multiparty computation for privacypreserving data mining,” The Journal of Privacy and Confidentiality, vol. 1, no. 1, pp. 59–98, 2009.