A Survey on A Privacy Preserving Technique using K-means Clustering Algorithm

Falguni A.Patel[1], Chetna G.Chand[2]

[1]*Student(Master of Engineering), Computer Engineering Department, Government Engineering College, Modasa, Gujarat, India*
[2] *Assistant Professor, Computer Engineering Department, Government Engineering College, Modasa, Gujarat, India*

ABSTRACT

In many organizations large amount of data are collected. These data are sometimes used by the organizations for data mining tasks. However, the data collected may contain private or sensitive information which should be protected. Privacy protection is an important issue if we release data for the mining or sharing purpose. There are various techniques that protect the sensitive data with less information loss which increase data usability and also prevent the sensitive data for various types of attack. Various comprehensive experiments on real as well as synthetic datasets show that they are effective and provide moderate privacy .Clustering based noise techniques that not only preserve the privacy but also ensure effective data mining.

Keywords: - *Data Mining, Privacy Preserving, Clustering, Privacy Preserving Techniques, K-means clustering algorithm.*

---

1. INTRODUCTION:

Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both [1]. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified [2]. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases [2].

While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries [3]. Several types of analytical software are available: statistical, machine learning, and neural networks. Generally, any of four types of relationships are sought:[8]

- **Classes**: Stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order [8]. This information could be used to increase traffic by having daily specials.

- **Clusters:** Data items are grouped according to logical relationships or consumer preferences.[8] For example, data can be mined to identify market segments or consumer affinities.

- **Associations**: Data can be mined to identify associations. The beer-diaper example is an example of associative mining.[8]

- **Sequential patterns**: Data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes [8].

Introduction of Clustering

Clustering can be considered the most important *unsupervised learning* technique; so, as every other problem of this kind, it deals with finding a *structure* in a collection of unlabeled data.[3] Clustering is "the process of organizing objects into groups whose members are similar in some way".[4]A *cluster* is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters[3] Clustering is used in many fields like pattern reorganization ,image analysis ,e-health, bio informatics[3] Major existing clustering methods are, distance based, hierarchical based, partition based and probabilistic based. In partitioning based clustering there is K-means clustering algorithm is used [9].

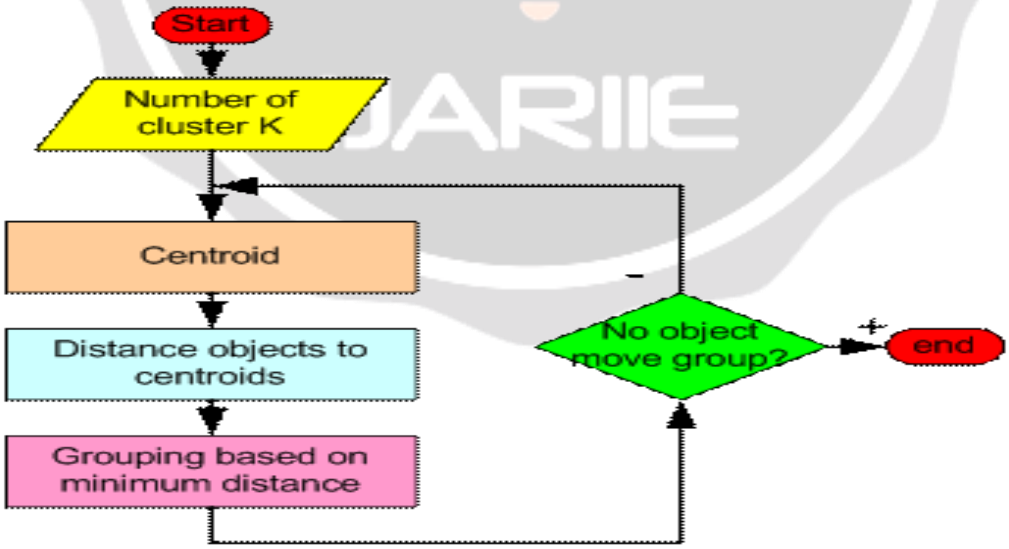
K-means clustering algorithm

- **Input:**
  Number of desired clusters, k, and a database D={d1, d2,…dn}containing n data objects.

- **Output:**
  A set of k clusters

  Steps:

  1) Randomly select k data objects from dataset D as initial cluster centers
  2) Repeat;
  3) Calculate the distance between each data object di (1 <=i<=n) and all k cluster centers cj(1<=j<=k) and assign data object di to the nearest cluster.
  4) For each cluster j (1<=j<=k), recalculate the cluster center.
  5) until no changing in the center of clusters is positive integer number.



Flowchart of k-means clustering algorithm [9]

Privacy preserving in Data mining

Data Mining is extensively used for knowledge discovery from large databases. The problem with data mining is that with the availability of non- sensitive information, one is able to infer sensitive information that is not to be disclosed [5]. Thus privacy is becoming an increasingly important issue in many data mining applications .The first approach protects the privacy of data by using an extended role based access control approach where sensitive objects identification is used to protect an individual's privacy and another approach uses cryptographic methods.[7] Privacy preserving is used in many areas like E-health, cloud ,location based services and wireless sensor networks.[10]

There are many different techniques used for privacy preserving data mining like Anonymization based ,perturbation based, randomized response based ,condensation based and cryptography based[10] .

A.Anonymization based PPDM

The basic form of the data in a table consists of following four types of attributes. (i) Explicit Identifiers is a set of attributes containing information that identifies a record owner explicitly such as name, SS number etc.[10] (ii) Quasi Identifiers is a set of attributes that could potentially identify a record owner when combined with publicly available data.[10] (iii) Sensitive Attributes is a set of attributes that contains sensitive person specific information such as disease, salary etc[10]. (iv) Non-Sensitive Attributes is a set of attributes that creates no problem if revealed even to untrustworthy [10]

Anonymization refers to an approach where identity or/and sensitive data about record owners are to be hidden.[5] It even assumes that sensitive data should be retained for analysis. It's obvious that explicit identifiers should be removed but still there is a danger of privacy intrusion [5] when quasi identifiers are linked to publicly available data. Such attacks are called as linking attacks. Sweeney proposed k-anonymity model using generalization and suppression to achieve k-anonymity i.e. any individual is distinguishable from at least k-1 other ones with respect to quasi-identifier attribute in the anonymized dataset [10]. Replacing a value less specific but semantically consistent value is called as generalization and suppression involves blocking the data. Releasing such data for mining reduces the risk of identification when combined with publically available data.[10] Limitations of the k-anonymity model stem from the two assumptions. First, it may be very hard for the owner of a database to determine which of the attributes are available or which are not available in external tables. The second limitation is that the k-anonymity model assumes a certain method of attack, while in real scenarios; there is no reason why the attacker should not try other methods.[10]

B. Perturbation based PPDM

Perturbation has an inherent property of simplicity, efficiency and ability to preserve statistical information. In perturbation the original values are replaced with some synthetic data values so that the statistical information computed from the perturbed data does not differ from the statistical information computed from the original data[5]. The perturbed data records do not correspond to real-world record owners, so the attacker cannot perform the sensitive linkages or recover sensitive information from the published data. In perturbation approach, records released is synthetic i.e. it does not correspond to real world entities represented by the original data [5],[6]. Therefore the individual records in the perturbed data are meaningless to the human recipient as only statistical properties of the records are preserved. Perturbation can be done by using additive noise or data swapping or synthetic data generation [5]. Since the perturbation method does not reconstruct the original values but only the distributions, new algorithms are to be developed for mining of the data. This means that a new distribution based mining algorithms need to be developed for each individual data 28 problems like classification, clustering or association rule mining. For example, Some approaches have been developed for distribution-based mining of data for problems such as association rules and classification, it is clear that using distributions instead of original records restricts the range of algorithmic techniques that can be used on the data. Kantar cioglu and Clifton and Rizvi and Harista have developed methods to preserve privacy of association rule mining. In the perturbation approach, any distribution based data mining algorithm works under an implicit assumption to treat each dimension independently. Relevant information for data mining algorithms such as classification remains hidden in inter-attribute correlations [5]. For example, the classification technique makes use of distribution-based analogue of single attribute split algorithm. However, other techniques such as multivariate decision tree algorithms cannot be modified accordingly to work with the perturbation approach. This is because perturbation approach treats different

attributes independently. Hence the distribution based data mining algorithms have an inherent disadvantage of loss of implicit information available in multidimensional records [10].

C. Randomized Response based PPDM

Basically, randomized response is statistical technique introduced by Warner to solve a survey problem. In Randomized response, the data is scrambled in such a way that the central place cannot tell with probabilities better than a pre-defined threshold, whether the data from a customer contains truthful information or false information.[5] The information received from each individual user is scrambled and if the number of users is significantly large, the aggregate information of these users can be estimated with good amount of accuracy. This is very useful for decision-tree classification since decision-tree classification is based on aggregate values of a dataset, rather than individual data items. [6]The data collection process in randomization method is carried out using two steps. During first step, the data providers randomize their data and transmit the randomized data to the data receiver. In second step, the data receiver reconstructs the original distribution of the data by employing a distribution reconstruction algorithm. Randomization method is relatively very simple and does not require knowledge of the distribution of other records in the data.[6] Hence, the randomization method can be implemented at data collection time. It does not require a trusted server to contain all the original records in order to perform the Anonymization process. The weakness of a randomization response based PPDM technique is that it treats all the records equal irrespective of their local density. One solution to this is to be needlessly more aggressive in adding noise to all the records in the data. But, it reduces the utility of the data for mining purposes as the reconstructed distribution may not yield results in conformity of the purpose of data mining. The randomized response approach has extended its strengths to a number of data mining problems [5]. In Agrawal and Srikant have discussed the use of randomized response approach for classification. A number of other techniques have also been proposed that work well over a variety of different classifiers. Techniques have been proposed for privacy preserving classification that enhances the effectiveness of classifiers [6]. For example, Gambs, Kegl and Aimeur have proposed methods for privacy-preserving boosting of classifiers. Methods for privacy-preserving mining of association rules have been proposed in the problem of association rules is especially challenging because of the discrete nature of the attributes corresponding to presence or absence of items[10] The randomization approach has also been extended to other applications such as OLAP, and SVD based collaborative filtering.

D. Condensation approach based PPDM

Condensation approach constructs constrained clusters in dataset and then generates pseudo data from the statistics of these clusters. It is called as condensation because of its approach of using condensed statistics of the clusters to generate pseudo data. It constructs groups of non-homogeneous size from the data, such that it is guaranteed that each record lies in a group whose size is at least equal to its anonymity level.[5] Subsequently, pseudo data is generated from each group so as to create a synthetic data set with the same aggregate distribution as the original data. This approach can be effectively used for the problem of classification [6]. The use of pseudo-data provides an additional layer of protection, as it becomes difficult to perform adversarial attacks on synthetic data. Moreover, the aggregate behavior of them data is preserved, making it useful for a variety of data mining problems.[5] This approach helps in better privacy preservation as compared to other techniques as it uses pseudo data rather than modified data. Moreover, it works even without redesigning data mining algorithms since the pseudo data has the same format as that of the original data.[10] It is very effective in case of data stream problems where the data is highly dynamic. At the same time, data mining results get affected as large amount of information is lost because of the condensation of a larger number of records into a single statistical group entity[10].

E. Cryptography based PPDM

Cryptographic techniques are ideally meant for such scenarios where multiple parties collaborate to compute results or share non sensitive mining results and thereby avoiding disclosure of sensitive information [5]. Cryptographic techniques find its utility in such scenarios because of two reasons. First, it offers a well defined model for privacy that includes methods for proving and quantifying it. Second a vast set of cryptographic algorithms and constructs to implement privacy preserving data mining algorithms are available in this domain. The data may be distributed among different collaborators vertically or horizontally.[5] In vertically partitioned data among different collaborators, the individual entities may have different attributes of same set of records and in case of horizontally partitioned data, individual records are spread out across multiple entities, each of which has the same set of

attributes. Most of the privacy preserving distributed data mining algorithms reveal nothing other than the final results[6]. Kantarcioglu and Clifton in incorporated cryptographic techniques to preserve privacy in association rule mining over horizontally partitioned data to minimize information shared and at the same time adding very little overheads to the mining task [5]. Lindell and Pinkas in have discussed how to generate ID3 decision trees on horizontally partitioned data.[10] Yang et al. in have discussed a solution for horizontally partitioned data where each customer has a private access only to their own record. Vaidya and Clifton were the first who studied how secure association rule mining can be done for vertically partitioned data [10]. Vaidya and Clifton in proposed a method for clustering over vertically partitioned data. All these methods are almost based on a special encryption protocol known as Secure Multiparty Computation (SMC) technology.[5] Yao in discussed a problem where two millionaires wanted to know who is richer with neither revealing their net worth. So, SMC was coined and developed. SMC defines two basic adversarial models namely (i) Semi-Honest model and (ii) Malicious model. Semi-honest model follows protocols honestly, but can try to infer the secret information of other parties. In Malicious model, malicious adversaries can do anything to infer secret information [5]. They can abort protocol, send spurious messages, collude with other malicious models or even spoof messages. SMC used in distributed privacy preserving data mining consists of a set of secure sub protocols that are used in horizontally and vertically partitioned data: secure sum, secure set union, secure size of intersection and scalar product [5]. Moreover, the data mining results may breach the privacy of individual records [10].

2. LITERATURE REVIEW:

2.1 The state of the art and tendency of privacy preserving data mining

- In [1] this paper two problems are considered in privacy preserving data mining .The first one is the protection of sensitive raw data, such as name, id num, address income level and other micro data. The other is the protection of sensitive knowledge showed by data mining ,which is also called knowledge hiding database(KHD)Privacy preserving data mining and traditional data mining are different in data processing, data concerning, technology application and so on. For single data record of privacy preservation Randomized technique ,Data blocking, Data perturbation techniques used. For Centralized data sets' privacy Reconstruction technique based on association rule mining and Reconstruction technique based on classification mining are used. For Distributed data's privacy Secure multiparty computation used. Standardization of privacy preserving technology is most important issue of this PPDM.

2.2 Preservation of privacy in data mining by using PCA based perturbation technique

- In [2] this paper Geometrical data transformation methods been extensively used for privacy conserving cluster, but drawback is that it is not reversible, that result in law security. The technique that preserves the privacy of delicate information in a multiparty cluster situation called guideline segment investigation based technique. K-means cluster algorithm and machine learning based methodology are used on artificial and realistic world information sets. Security preservation in data processing is a justifiably new subject of research.

2.3 Privacy preserving data classification and similarity evaluation for distributed system.

- In [3] this paper Privacy preservation data classification, Hidden data discover, Data classifier used. see there is existing data or new data. New data are not directly revealed during classification. Holomorphic cryptosystem, support vector machine, alternating direction method of multiplier are used for privacy preservation. Training data are private but trained model can be public. Correctness, feasibility, and efficiency of proposed scheme via extensive experiments.

2.4 Using K-means clustering algorithm for handling data precision.

- In [4]this paper Privacy preserving, Anonymized technique, Clustering, K-means clustering algorithm, AES encryption algorithm is used . Original patient record is transformed into Anonymized table which is identified by administrative. To keep anonymized table in cluster form using k-mean clustering algorithm. To partition anonymized table into cluster and each cluster has similar data object which is based on anonymized data attribute. After that

grouped similar data to be protected using AES encryption algorithm. More accuracy and better level of privacy of sensitive attribute.

2.5 Efficient privacy preserving classification construction model with differential privacy technology.

- In [5] this paper Decision tree, efficient classifier used to add noise via Laplace and exponential mechanism. Improving privacy budget allocation method two mode of implementing privacy protection: Interface mode, Full Access mode .Controls privacy budget and avoids various types joint visit attack that are based on background knowledge. Studies on how to apply differential privacy technology in the privacy protection of large data have a wide application value. We will continue research in the direction of meeting large data set privacy security access requirements. and practical

2.6 Modified K-means clustering algorithm

- In [11] this paper presented a modified *K*-means algorithm for iterative clustering algorithm. This procedure is based on the optimization formulation and a novel iterative method.. According to the above numerical experiment results, the proposed method is an effective clustering method..It takes less number of iterations and give more accurate result then k-means clustering algorithm.

3. COMPARATIVE TABLE:

Table -1: Comparative Table

Sr .No.	Paper Title	Method	Advantage	Disadvantage
1	The state of the art and tendency of privacy preserving data mining	Randomized technique ,Data blocking ,Data perturbation ,Secure multiparty computation	Standardization of privacy	Protection of sensitive raw data and micro data
2	Preservation of privacy in data mining by using PCA based perturbation technique	K-means cluster algorithm and machine learning based methodology	Security preservation	Law security ,not reversible
3	Privacy preserving data classification and similarity evaluation for distributed system	Support vector machine ,Alternative direction method for multiplier	Data are private and correct	New data not directly revealed.
4	Using K-means clustering algorithm for handling data precision	Anonymized technique ,k-means clustering algorithm, AES encryption	Data is accurate and private	For small data clustering is not possible
5	Efficient privacy preserving classification construction model with differential privacy technology	Decision tree ,Classifier to add noise, Privacy budget allocation method	Control privacy budget very well	Various types of attacks so maintaining privacy is difficult
6	Modified K-means clustering algorithm	K-means clustering and modified K-means clustering	Less no. of iterations and saves time	use for only numerical data

4. CONCLUSION:

Privacy preservation of data is main concern for current generation in current scenario. This paper concentrates on various privacy preserving techniques and their advantages and disadvantages. So there is a need to develop a

method to provide mechanisms for data privacy which will provide efficiency, security and accuracy. So in future, new privacy preserving techniques with improved or modified K-means clustering algorithm are to be produced which defeats the disadvantages of current systems.

5. REFERENCES:

1) Bo Wang, Jing Yang **"**The state of the art and tendency of privacy preserving data mining" Supported by grants from the National Science Foundation of China No 61073043 and No.61073041. IEEE 2011

2) C.Gokulnath, M.K.Priyam, Vishnu Balan, K.P.Ramaprabhu R . Jayanthi. . "Preservation of privacy in data mining by using PCA based perturbation technique " 2015 International conference on Smart technologies and Management for Computing, Communication, Controls, Energy and Materials(ICSTM),IEEE 2015.

3) QI Jia,Linke Guo,Zhangpeng Jin,Yuguang Fung "Privacy preserving data classification and similarity evaluation for distributed system",2016 IEEE 36TH International conference on Distributed Computing System .

4) P.Suganthi, K.Kala, DR.C.Balasubramanin. "Using K-means clustering algorithm for handling data precision ."2016 IEEE

5) Lin Zhang,Yan Liu,Ruchuan Wang,XiongFu,Qiamon Lin. "Efficient privacy preserving classification construction model with differntial privacy technology."Journal of system engineering and electronics Vol 28 no 1 February 2017,IEEE 2017

6) Zhang P, Tong YH, Tang SW, Yang DQ, Ma XL." An effective method for privacy preserving association rule mining". Journal of Software, 2006, 17(8):1764-1774.

7) Stanley R.M.Oliveira and Osmar R. Zaiane. "Achieving Privacy Preservation When Sharing Data For Clustering"[C].In Proc. of the International Workshop on Secure Data Management in a Connected World (SDM'04) in conjunction with VLDB, Canada, 2004:67-82.

8)Bhavani Thuraisingham,"A primer for understanding and applying data mining". 1520 9202/00/$10.00©2000IEEE

9) Shruti Kapil,Meenu Chawla,Mohd Dilshad Ansari, "On K-means Data Clustering Algorithm with genetic algorithm" 978-1-5090-3669-1/16/$31.00@2016 IEEE

10) Majid Mashir Malik,M.asger Ghazi, Rashid Ali, "Privacy Preserving Data Mining Techniques: Current Scenario and FutureProspects",IEEE2012

11) W. Li, "Modified K-Means Clustering Algorithm," *2008 Congr. Image Signal Process.*, pp. 618–621, 2008.