SW-SDF based privacy preserving for k-means clustering

Kiran P and Kavya N P

Abstract— Privacy preserving data mining is a new direction in data mining which ensures the privacy of individual and company related information even after mining. Personalized privacy preservation uses flag for differentiating actual records which require privacy and records which does not require privacy such that overall information loss can be reduced. SW-SDF based privacy preservation uses sensitive weight (SW) and sensitive disclosure flag (SDF). Sensitive weight is used for differentiating between records which actually require privacy and which does not require privacy. Among records which require privacy, SDF value is accepted. SDF=0 indicate record owner is not ready to disclose his information and SDF=1 indicate record owner is ready to reveal his identity. The major drawback of this approach is that methods were not defined for specific mining algorithms. In this paper we have defined the representation of SW-SDF based privacy method on k-means clustering. Experiment results indicate that the mean of the original cluster is almost similar to the original mean and privacy is retained.

Index Terms— privacy preserving data mining(PPDM), privacy preserving data publishing(PPDP), SW-SDF privacy, security, clustering , K-means, unsupervised learning.

1 INTRODUCTION

Data mining is a prerequisite for most of the present day application. The major goal is for decision making and prediction. Applications include patient data base, transactional database, financial data base and many more applications. Data miner and data publisher are two important entities of this environment and usually belong to different organization. Data publisher accepts information from multiple locations. This data in its original format is not secure since the data miner may be untrusted and the information can be misutilized there by privacy is lost. The objective of data publisher is to transform data in a way which ensures confidentiality of individual or company related even after mining process.

Privacy preserving data mining (PPDM)[1] is a novel approach in data mining which defines methods that can be used for preserving privacy. Privacy preserving data publishing(PPDP) concentrates on methods from data publisher end. Different methods exist in PPDP among them the most important are anonymization, data randomization and data swapping Information. Anonymization is an approach which concentrates on generalization of data there by privacy of individual is restored. The most important anonymization techniques include k-anonymity[4], l-diversity[5] and t-closeness[6]. Information loss was more in PPDP methods so personalized privacy preservation was defined. The basic assumption of personalized privacy is that in the given data base not all records require privacy or the actual records which require privacy is very less. For example in patient data base not all patient records require privacy. Patient details are not so much important if the sensitive attribute is flu as compared with Heart disease.

SW-SDF personalized privacy preservation[2] uses two flags SW and SDF. SW indicates the sensitive weight and SDF indicates sensitive disclosure flag. Records within the data base can be divided in to non-sensitive and sensitive records. Sensitive records are records which actually require privacy and non-sensitive records are records where privacy is not a concern. SW is used for differentiating between sensitive records and non-sensitive records. Statistical based approach or clustering based approach can be used for identifying the SW[3]. SW=0 indicate a record which is not sensitive and SW=1 indicates a record which requires privacy. SDF=1 is assigned to all records where SW=0 indicating that the record owner details are non-sensitive. For SW=1, SDF is accepted from user. SDF=0 indicates the record owner is not ready to reveal his identity and SDF=1 is ready to reveal his identity. This method uses anonymization approach for preserving privacy and was defined without considering how this will be used for mining which was a major drawback.

In this paper we propose a method SW-SDF personal privacy for k-means clustering. k-means[7,8] is one of the most widely used unsupervised clustering technique for statistical data analysis. This algorithm groups objects in to k-clusters. Each item is placed in to the closest cluster based on the distance measures computed. In our method we propose an algorithm in such a way that the resultant clusters are almost equal to the original cluster and the privacy is retained. This paper is organized as follows. In section 2 we discuss the related work on privacy preserving data clustering. Existing k-means algorithm for data clustering has been discussed in section 3. Proposed method for SW-SDF based personalized privacy for k-means clustering in section 4. Result analysis and conclusion in section 5 and section 6.

- *Kiran P, Research scholar, VTU,Belgaum,E-mail: kiranmys@rediffmail.com*
- *Kavya N P, Prof & Head, Dept of MCA, RNSIT, Bangalore, India, E-mail: npkavya@yahoo.com*

2 Related work

In literature various methods have been defined for privacy preserving data clustering. These methods can be categorized in to two major techniques i)data perturbation[9] and ii)secure multiparty computation(SMC)[10]. In data perturbation data is modified in such a way that it preserves clustering and privacy. SMC methods define a way in which data is accepted from multiple locations. This data between two parties must be merged in order to identify the relevant clusters. This is achieved by sharing data and there by the overall clusters is generated without revealing the identify. The first method concentrates on the techniques for hiding sensitive information and second technique deals with finding common cluster details without revealing the identity. These methods has been used in most of the privacy preserving clustering and has been indicated below.

In [11] author has defined a method which addresses unauthorized secondary use of information. Geometric data transformation method has been used for converting data in to a format which ensures privacy. This method clusters the objects focusing on partional based hierarchical method. Data standardization was used in [12] to make the result more reasonable to clustering. SMC drawbacks were overcome in [13] by using secret sharing. The data is divided into multiple shares and was processed by different servers. In [14] author has discussed the dual goal of privacy and valid clustering by using object similarity based representation and dimensionality reduction based transformation. Protocols that provide privacy preservation for the computation of covariance and means was discussed in [15]. Intermediate clusters distance was not revealed by using the method of secret sharing among the two parties were the data is distributed horizontally was discussed in [16]. In [17] data distribution there by utilization and identification of cluster was made. The basic idea was that each location learns the cluster of its own attributes, but learns nothing about the attributes at other locations.

3. Existing k-means algorithm

k-means conventional algorithm is indicated in algorithm 1. This algorithm is one of the most popular and most used for various applications. K-means can be applied for numerical data which has a mean value among the given set of data. It is a partitional based method. The pre-requisite of this algorithm is that the k value must be indicated. This k value defines the number of clusters in the given data. This algorithm works as follows. A random set of k value will be selected from the given number of points which is initialized as the mean value. Each point in the database symbolizes a record. Remaining records are inserted in to the clusters based on distance measure. Recalculation of mean and there by the change in the position of points with clusters also gets changed. This change continues until the error function does not change considerably or points within the cluster are not getting changed.

Let T be a relation with n records and let Clust1, Clust2, …, Clustk be k disjoint clusters of T. then error function is defined in equation 1.

$$\text{Errfunct} = \sum_{i=1}^{k} \sum_{x \in \text{clust}_i} d(X, \text{mean}(\text{clust}_i)), \quad (1)$$

Where mean(clusti) is the centroid of cluster clusti. $d(X, \text{mean}(\text{clust}_i))$ indicates the distance between point x and the mean of clusti. There are different measures available in literature among them let us assume we are using Euclidean distance.

Algorithm 1.k-means algorithm.
Input: T data base, Number of clusters k, distance measure d
Output: k clusters
//Initialization phase
1: (Clust1, Clust2,…,Clustk) are initialized.
//Repetition phase
2: repeat
2.1:dij = d(Ti ,mean(clustj)) ;
2.2: clustj=min(dij) 1≤j≤k is determined;
2.3: Assign Ti to clustj ;
2.4: Cluster means of any changed clusters is calculated;
3: while (mean value of the cluster does not get changed or error function does not change considerably)

The algorithm of k-means can be divided into initialization phase and repetition phase. In the initialization phase randomly k points in the data base are initialized as clusters. In repetition phase the distance is calculated for each point with the cluster. The minimum distance determines the insertion of point into the cluster. After all points are inserted the mean value of the cluster in which point was inserted is recalculated.

4. SW-SDF based privacy preservation k-means algorithm

Usually for clustering the attributes are independent and does not have a sensitive attribute. The basic assumption of this approach is that for all records SW=1 and SDF is accepted from user. SDF=0 indicates that the record owner is not ready to reveal his identity and SDF=1 indicates the user is ready to reveal his identity. The first phase of the algorithm is similar to k-means. Restructuring phase tries to identify the sensitive information among individual clusters and searches for similar records in nonsensitive records. In the next phase anonymization of this QIDB is done. it finds the same number of records which are almost similar in the next subsequent cluster. The records are selected based on a predefined threshold which is used to find the maximum displacement of mean value and is represented as THPvar. SW-SDF based privacy preservation k-means clustering is indicated in algorithm 2.

Algorithm 2. *SW-SDF* based privacy preservation *k*-means

clustering
Input: T data base, Number of clusters k, distance measure d, threshold THP_{var}
Output: k clusters
//Initialization phase
1: $(Clust_1, Clust_2,\ldots,Clust_k)$ are initialized.
//Repetition phase
2: repeat
2.1:$dij = d(T_i, mean(clust_j))$;
2.2: $clust_j=min(d_{ij})$ $\forall 1\leq j\leq k$ is determined;
2.3: Assign T_i to $clust_j$;
2.4: Cluster means of any changed clusters is calculated;
3: while (mean value of the cluster does not get changed or error function does not change considerably)
//Restructuring phase
4: for each $clust_j$ $\forall 1\leq j\leq k$ records where $SDF=0$
4.1: Search and identity $clust_x$ $\forall 1\leq x\leq k$ and $x\neq j$ which is almost similar to the records where $SDF=1$ or $SDF=0$;
4.2: recalculate mean$(clust_x)$ by removing the records ;
4.3: if $old\_mean(clust_x)$-$new\_mean(clust_x)$<THP_{var} then $found\_flag=true$;
4.4: if NOT $found\_flag$ suppress;
4.5: else generalize
5:end for

Generalization identifies the highest value in each column. The entire column values are generalized to the same level. In each cluster of QIDB outliers are identified and those records are suppressed from the data base.

5. Experimental Results

Experiment was conducted using the standard UCI machine learning database. The data base that was considered was Breast Cancer Wisconsin. There were 9 attributes among them we have considered only 3 attribute for better visualization. The attributes are Uniformity of Cell Shape, Clump Thickness and Uniformity of Cell Size. Number of records that were considered was 699. Number of records which are suppressed were 3. Initial cluster is shown in figure 1 and mean value of this cluster is shown in table 1. SDF=0 was assigned with a probability of 0.25 randomly in the given database. clusters formed by removing SDF=0 from the given data base is shown in figure 2 and mean values in table 2. After applying SW-SDF based privacy preserving for k-means clustering additional two clusters are formed which is shown in figure 3 and mean values in table 3. Experimental results show that the mean cluster position has not been drastically changed and the impact of this in the overall clusters is minimum.

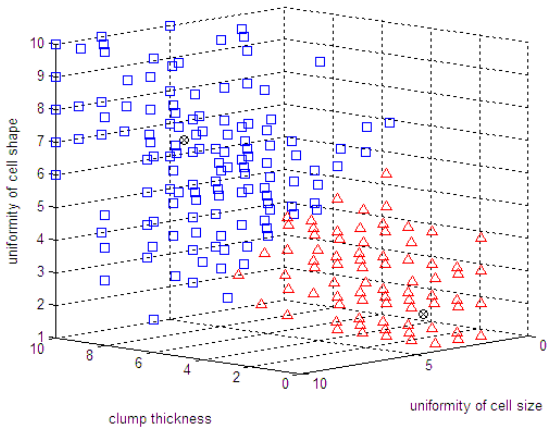

Figure 1: Details of the cluster before SW-SDF based privacy preserving for k-means clustering

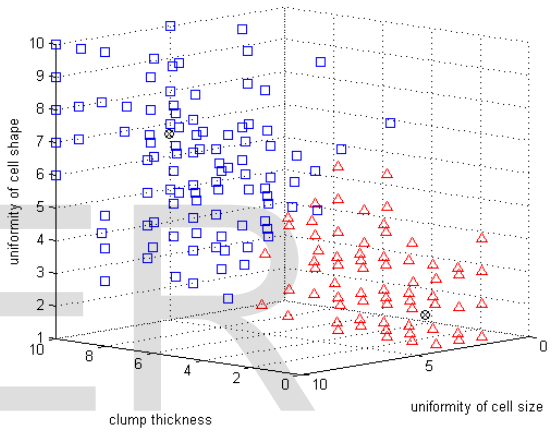

Figure 2: Details of the cluster after removing SDF=0 from data base

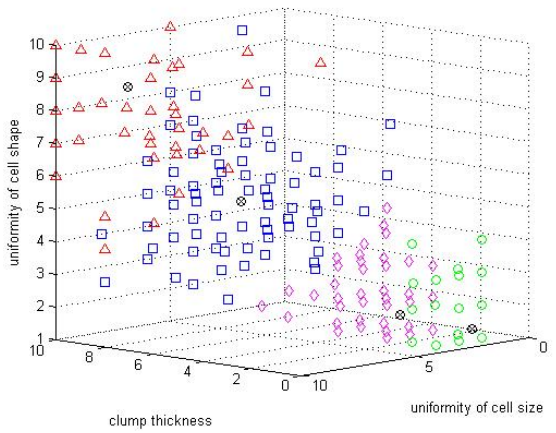

Figure 3: Details of the cluster after SW-SDF based privacy preserving for k-means clustering

TABLE 1: DETAILS OF THE MEAN VALUE OF CLUSTER BEFORE SW-

SDF BASED PRIVACY PRESERVING FOR K-MEANS CLUSTERING

Cluster/Attribute	Uniformity of cell shape	Clump thickness	Uniformity of cell size
Cluster 1	3.06	1.34	1.49
Cluster 2	7.45	7.12	7.03

TABLE 2: DETAILS OF THE MEAN VALUE OF CLUSTER AFTER REMOVING SDF=0 FROM DATA BASE

Cluster/Attribute	Uniformity of cell shape	Clump thickness	Uniformity of cell size
Cluster 1	2.96	1.37	1.48
Cluster 2	7.71	7.48	7.23

TABLE 3: DETAILS OF THE MEAN VALUE OF CLUSTER AFTER SW-SDF BASED PRIVACY PRESERVING FOR K-MEANS CLUSTERING

Cluster/Attribute	Uniformity of cell shape	Clump thickness	Uniformity of cell size
Cluster 1	1.24	1.17	1.25
Cluster 2	7.79	9.22	8.88
New cluster 1	7.17	4.92	4.96
New cluster 2	4.03	1.33	1.41

6. CONCLUSION AND FUTURE WORK

SW-SDF based privacy preserving for k-means clustering provides a means for identifying clusters which are almost equivalent to the original cluster and privacy of those records are also retained. In our data base we have used a single release but other releases must also be considered. Future work may be done on faster retrieval of similar records with SDF=0. Experiment can also be done taking into account the actual application. Identification of appropriate number of clusters for SDF=0 can be investigated.

REFERENCES

[1] Kiran P, S Sathish Kumar and Kavya N P, Modelling Extraction Transformation Load embedding Privacy Preservation using UML. In International Journal of Computer Applications (IJCA), vol. 50, No. 6, July 2012.

[2] Kiran P and Kavya N P, SW-SDF based personal privacy with QIDB-Anonymization method. International Journal of Advanced Computer Science and Applications (IJACSA), vol 3, issue 8, August 2012.

[3] Kiran P, S Sathish kumar, Hemanth S and Kavya N P, Assignment of SW using statistical based data model in SW-SDF based personal privacy with QIDB-anonymization method, In Proc of second IEEE International conference on Parallel, Distributed and Grid computing,pp. 816 – 821, Dec 06-08 2012.

[4] L. Sweeney, k-Anonymity: A Model for Protecting Privacy. Int'l J.Uncertain. Fuzz., vol. 10, no. 5, pp. 557-570, 2002.

[5] Machanavajjhala A, Gehrke J, Kifer D and Venkitasubramaniam M, l-diversity: Privacy beyond k-anonymity. In Proceedings of the 22nd IEEE International Conference on Data Engineering(ICDE), 2006.

[6] Ninghui Li , Tiancheng Li , Suresh Venkatasubramanian, t-Closeness: Privacy Beyond k-Anonymity and l–Diversity. ICDE Conference, 2007.

[7] Duda, R., Hart, P.: Pattern Classification and Scene Analysis. John Wiley and Sons, Chichester,1973.

[8] Fukunaga, K.: Introduction to Statistical Pattern Recognition. Academic Press, London,1990.

[9] Kargupta, H., Datta, S., Wang, Q., Sivakumar, K.: On the privacy preserving properties of random data perturbation techniques. In: ICDM, pp. 99–106,2003.

[10] Bunn, P., Ostrovsky, R.: Secure two-party k-means clustering. In: CCS, pp. 486–497,2007.

[11] Stanley R. M. Oliveira , Osmar R. Zaïane,Privacy Preserving Clustering By Data Transformation, In proc. of the 18th brazilian symposium on databases,pp 304-31,2003.

[12] Chunhua Su; Feng Bao; Jianying Zhou; Takagi, T.; Sakurai, K., Privacy-Preserving Two-Party K-Means Clustering via Secure Approximation, Advanced Information Networking and Applications Workshops, 2007, AINAW '07. 21st International Conference on , vol.1, no., pp.385,391, 21-23 May 2007

[13] Maneesh Upmanyu, Anoop M. Namboodiri, Kannan Srinathan, and C.V. Jawahar, Efficient Privacy Preserving K-Means Clustering, PAISI 2010, LNCS 6122, pp. 154–166, 2010

[14] Stanley R. M. Oliveira and et al., Privacy-Preserving Clustering by Object Similarity-Based Representation and Dimensionality Reduction Transformation, In Proc. of the workshop on privacy and security aspects of data mining (psadm'04) ,pp 21-30,2004

[15] Luong The Dung; Ho Tu Bao, Privacy Preserving EM-Based Clustering, Computing and Communication Technologies, RIVF '09. International Conference on , vol., no., pp.1,7, 13-17 July 2009.

[16] Geetha Jagannathan and Krishnan Pillaipakkamnatt and Rebecca N. Wright, A New Privacy-Preserving Distributed k-Clustering Algorithm, SDM 2006

[17] Jaideep Vaidya and Chris Clifton, Privacy-Preserving K-Means Clustering over Vertically Partitioned Data, IN SIGKDD, 2003.