IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 7, July 2017

Methodical Clustering Slant for Privacy Preserving Data Mining

Bharath H C[1], Poornima B[2], Ashoka K[3]

M.Tech Student, Computer Science and Engineering, BIET, Davangere, India[1]

Professor & Head, Information Science and Engineering, BIET, Davangere, India[2]

Assistant Professor, Computer Science and Engineering, BIET, Davangere, India[3]

Abstract: This paper presents a clustering based k-anonymization technique to minimize the information loss while at the same time ensuring data utility. In privacy preserving data mining, anonymization based approaches have been used to preserve the privacy of an individual. However, the anonymization based approaches suffer from the issue of information loss. To minimize the information loss and ensure data quality we produce new approach called systematic clustering along with equal combination of quasi-identifier and sensitive attributes. The proposed approach first generates sub-databases by equal combination of quasi-identifier and sensitive attributes and adopts group-similar data together and then anonymizes each group individually. We also evaluate our approach empirically focusing on the information loss and execution time as vital metrics.

Keywords: quasi-identifier, sensitive attribute, sub-databases, systematic clustering, anonymization, PPDM.

I. INTRODUCTION

Presently, the volumes of generated data grow exponentially every year. Among this data, there is a growing amount of personal information contained within. This datum has attracted the attention of those fascinated in creating more tailored and personalized services. Data mining is a common methodology to retrieve and determine useful concealed knowledge and information from personal data. This violates the privacy of the individuals and leads to the apprehensions that personal data may be breached and distorted. As a result, this phenomenon has brought new challenges to protect the privacy of the people as a key issue in privacy preserving data mining [1].

Among the number of anonymization approaches, the k-anonymity model [3-8] has been considerably used in privacy preserving data mining because of its easiness and effectiveness. However, information loss and data utility are the major issues in the anonymization approaches as discussed in [2-8]. The k-anonymity model provides privacy and produces an anonymous database via generalization and/or suppression. In the case of generalization, the values in a database are replaced with some related values. For example, if the values for the Age attribute in the database are 31, 32, 33, 34, 35 and 36, then they can be represented as (31-36). Conversely, in the case of suppression, the values in a database are masked or deleted. For example, the suppressed value may be represented as 3* for the actual values 31, 32, 33, 34, 35 and 36 in a database. However, generalization is healthier as compared to suppression, since the generalization discloses at smallest amount of information as compared to suppression [7]. Though, the anonymous database generated via generalization and/or suppression outcomes in information loss.

The k-anonymity model works by ensuring that each record of a table is identical to at least (k − 1) other records with respect to a set of privacy-related features, called quasi-identifiers, that could be potentially used to identify individuals by linking these attributes to outside data sets [7].

For example, consider the patient diagnosis records in a hospital in Table 1, consists of three types of attributes. Identifier (ID), quasi-identifier (QI) and sensitive attributes (SA). The attributes ZipCode, Gender, Age and are regarded as quasi-identifiers, name as identifiers and disease as sensitive attributes. If the hospital simply publishes the table to other organizations for classifier development, those organizations might extract patient's disease histories by joining this table with other tables. By contrast, Table 2 is a 3-anonymization version where data values of Table 1 in attributes Gender and Age have been generalized and Zipcode is suppressed as common values and the numbers of records in its two equivalence classes are both equal to three features which present critical challenges to knowledge discovery and data modelling.

Unquestionably, anonymization is complemented by information loss. In order to be useful in practice, the dataset should stay as informative as possible. To minimize the information loss due to k-anonymization, all records are partitioned into several groups such that each group contains at least k similar records with respect to the quasi-identifiers and then the records in each group are generalized or suppressed such that the values of each quasi-identifier are the same.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 7, July 2017

Table I: Patient records

Identifier	Quasi-identifier			SA
Name	Age	Gender	Zip codec	Disease
ABC	31	male	55441	Typhoid
DEF	35	male	55440	Cholera
IJK	37	female	55442	Arthatis
UVW	26	female	55440	Cancer
RST	29	female	55440	Cancer
VIN	23	male	55443	Flue

Table II: Anonymization table

Equivalent Class	Age	Gender	Zip Codec	Disease
1	[31-40]	person	5544*	typhoid
	[31-40]	Person	5544*	cholera
	[31-40]	person	5544*	arthatis
2	[21-30]	Person	5544*	cancer
	[21-30]	Person	5544*	Flue
	[21-30]	person	5544*	cancer

Such similar groups are known as clusters [7]. As a result, the k-anonymity model can be addressed from the perspective of clustering.

It is highlighted that the information loss and the data utility are two incompatible goals of privacy preserving data mining. The information loss increases by hiding more data, but decreases the data utility. Conversely, information loss decreases by hiding less data, but increases the data utility. As said earlier, the problem of information has also been addressed in a clustering based anonymization approaches. The major objective of clustering based anonymization approach is to minimize the loss of information and maximize the utilization of data. The clustering based anonymization approaches make use of all the attributes of databases to produce anonymized databases [2-8] which are privacy protected. Therefore, publication of such protected databases will support the data miner in identifying the sensitive attribute of an individual.

In this paper, we propose one approach equal combination of quasi-identifiers and sensitive attribute which generate sub-databases which is a combination of quasi-identifiers and sensitive attributes using systematic clustering [7] algorithm for protecting privacy of the people. Our major aim in this research is to minimize the information loss while protecting privacy and ensure the maximum data utility.

II. RELATED WORK

The problem with preserving the privacy of an individual when data mining has gained much importance in recent years and due to this many algorithms have been proposed [2-8,10]. In privacy preserving data mining, preserving the privacy of an individual has been a prime research issue.

In order to preserve the privacy, various anonymization based approaches were proposed in the literature [2-8, 10-11]. The k-anonymity model [3] is one of the simple models used for the privacy preservation. Extending the idea of k-anonymity, a number of anonymization based clustering approaches have been proposed in [4-8, 10]. It includes Byun et al. Greedy k-member clustering algorithm [5], Loukides et al. Clustering algorithm [6], Lin et al. One passes k-means clustering algorithm [8] and Kabir et al. Systematic clustering algorithm [9].

Byun et al. [4] proposed a greedy k-member clustering algorithm. The greedy kmember clustering algorithm is sensitive to outlier records. With the presence of outlier records in the cluster, the information loss of the cluster is also increases.

Loukides et al. [5] proposed a clustering algorithm, which produce one cluster at a time. This algorithm builds a cluster with a user defined threshold value. Based on the user defined threshold value, the records are inserted and deleted in a cluster. The information loss of the generated cluster should not exceed the user defined threshold value. If the number of records in a particular cluster is less than user defined threshold, the cluster is deleted. Thus, with the use of user defined threshold, this algorithm is less sensitive to outlier records. In addition, this algorithm deletes records, and therefore, generates higher information loss.

Lin et al. [6] proposed one pass k-means clustering algorithm. This algorithm builds a cluster with lesser information loss and execution time as compared with the greedy k-member clustering algorithm [4].

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 7, July 2017

Kabir et al. presents a systematic clustering algorithm in [7]. This algorithm generates lesser information loss as compared to Byun et al. Greedy k-member clustering algorithm [4]. The systematic clustering algorithm makes a cluster of similar records. With the presence of similar kinds of records, it leads to the lesser generalization and/or suppression and hence incurs lesser information loss.

III. THE PROPOSED APPROACH

In this section, we present our approach equal combination of QI and SA using systematic clustering algorithm [7] that partition the database using a combination of quasi-identifier and sensitive attribute then generates the anonymized sub-databases by hiding the private information.
Our proposed algorithms basically first decompose and then anonymized the database into separate individual sub-databases with a set of QI and SA attributes therein. Thus, publishing such database would preserve the sensitive information of a person when joining with the external available databases.

Architecture

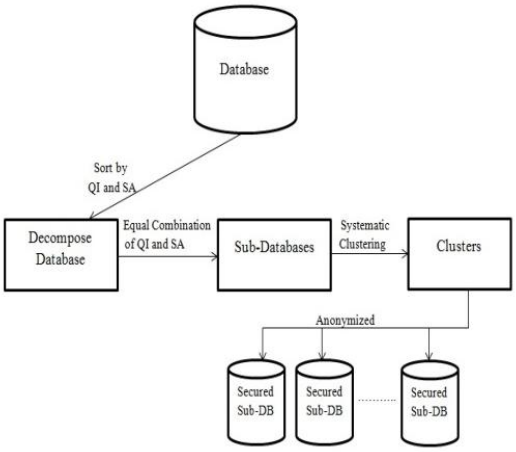

Fig.1: Architecture of proposed system

The architecture of the proposed system consists of the original database of any organizations of the distinct sectors such as Stores, Hospitals, and Banking, where privacy is required for their customer information. These databases are decomposed by using equal combination of QI and SA in the database, then sub-databases is constructed. Later clustering and anonymization is performed on these sub-databases and finally it is published.

Algorithm

Let B be a database with attributes viz. quasi-identifier (QI) and sensitive attribute (SA) as shown in Eq. (1).

$$D = \{QI, SA\} \quad (1)$$

Let QI and SA be the set of possible attribute. The possible values of quasi-identifier and sensitive attributes are represented in Eqs. (2) and (3).

$$QI = \{QI1, QI2, \ldots, QIn\} \quad (2)$$

$$SA = \{SA1, SA2, \ldots, SAn\} \quad (3)$$

Then, the resulting original database is shown in Eq. (4).

$$D = \{QI1, QI2 \ldots, QIn, \ldots, SA1, SA2, \ldots, SAn\} \quad (4)$$

Let B1 be a sub-database generated from the database B. The sub-database B1 is a combination of quasi-identifier (QI) and sensitive attribute (SA). It is shown in Eq. (5).

$$B1 = \{QI1, SA1\} \quad (5)$$

Similarly, other possible combinations {B1, B2, …, Bn} of QI and SA are constructed for the original database B.

Equal combination of QI and SA using systematic clustering algorithm.

Input : Database B with r records
Output: $\gamma=\{\sigma1, \sigma2, \ldots. \sigma p\}$ be a partitioning of r
// B is original database
// r is the number of records in the database
// γ is a partitioning of r records
// σ is a cluster

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 7, July 2017

Algorithm:
Begin
1. Identify the attributes such as identifier , quasi-identifier (QI) and sensitive attribute (SA)
2. Remove the identifier attribute and replace it with ID
3. Sort all records by their quasi-identifiers
4. Identify the number of clusters
5. Make an equal combination of QI and SA to construct the sub-database
6. Make a partition of all records into k groups
7. Select a record ri randomly from the first partition of k records
8. Similarly select another records rj from the other partition of k records
9. Calculate information loss

$$IL(\gamma)= \sum_{i-1}^{p} IL(\sigma i)$$

10. Move the records in a cluster with lowest information loss
11. Find extra element in a cluster those who exceed the k size
12. Add extra element in a cluster whose information loss is lowest
End

According to the algorithm as shown in Tables 3, we first identify and classify the attributes such as identifier, quasi-identifier and sensitive attributes in a database (step 1). Subsequently, we remove the identifier attribute from the database and sort all records using the quasi-identifiers (steps 2 and 3). Then, we find out the number of groups and clusters such that σ=r/k, where r is the number of records in a database and k is the anonymization factor (step 4).

After identifying the groups and clusters in an original database, we generate a sub-database using a combination of quasi-identifier and sensitive attribute (step 5). In our Approach, we create an equal combination of quasi-identifier (QI) and sensitive attribute (SA). From each generated sub-databases, we make a partition of all records into k groups (step 6).

Then, we used Systematic clustering algorithm in order to generate the clusters. According to the Systematic clustering algorithm, we randomly select a record from the first group for the creation of the first cluster (step 7). Similarly, we create the remaining cluster by randomly selecting the records from the remaining groups (step 8). Subsequently, we calculate the information loss of each cluster (step 9). Now, we select other records from the first group and add records in a cluster whose information loss is the lowest (step 10).

In the same way, we select and add other records in a cluster whose information loss is the lowest. During the clustering process if some cluster has exceeded to the k size, the extra element should be added in a cluster whose information loss is the lowest (step 11 and step 12).

In our proposed algorithms, we assume r, k and σ as the total number of records, the k-anonymity parameters and the number of clusters, respectively. Our algorithm takes O(nlogn) time to sort the records once in the database. The number of clusters are calculated as σ=r/k. As a result, the time complexity of our proposed clustering algorithms is a product of number of records (r) and the number of clusters (σ). Therefore, the total time complexity of our proposed algorithm is O=(r2/k).

Information Calculation

Information loss for Numerical attributes (Ln):

Let the minimum and maximum records in a cluster σ be Ri max and Ri min respectively. Let Di max and Di min be the maximum and minimum values of the records in a database D.

$$Ln = (\sum_{i=1}^{n} Ri\ max - Ri\ min\ /\ Di\ max - Di\ min)$$

Information loss for Categorical attributes (Lc):

Let Cj (j=1,2,..c) be a set of categorical attributes.

$$Lc = \sum_{j=1}^{c} H(\ A(UCj))/\ H(TCj)$$

Where, H(A(UCj)) be the sub tree rooted at the lowest common ancestor for each value of categorical attribute and H(TCj) is the height of the taxonomy tree.

Finally, the total information loss IL can be represented as

$$IL= (|\ r\ |\ (\sum_{i=1}^{n} Ri\ max - Ri\ min\ /\ Di\ max - Di\ min) + \sum_{j=1}^{c} H(\ A(UCj))/\ H(TCj)\ )$$

Where, r is the total number of records in cluster σ.

IV. EXPERIMENTAL EVALUATION

In this section, we present the effectiveness of our Approach Equal combination of QI-SA with respect to the parameters such as information loss and execution time. We compare our Approach with two state-of-the-art clustering approaches viz. Greedy k-member algorithm [4] and Systematic clustering algorithm [7]. The experiment is

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 7, July 2017

implemented in Java with JDK 1.6 in a system configured with Intel core i5 processor, 4 GB RAM and 500GB hard disk.

Experimental Setup

We use the ADULT database from the UCI Machine Learning Repository [14] for experimentation. The ADULT database contains 32561 records and 15 attributes. Out of them, we retain only attributes viz. Age, Race, Marital-status, Sex, fnlwt and Occupation. The attributes Age and fnlwt are numeric attributes, whereas Race, Marital-status, Sex and Occupation are the categorical attributes. The attribute Occupation is taken as a sensitive attribute in the database.

Methodology of Evaluation

We ran our proposed approach on the various k-values such as 20, 40, 60, 80 and 100. The total information loss and the execution time were calculated during each run of the experiment. In Fig. 2, we show that our Approach equal combination of QI and SA achieves lesser information loss as compared to state-of-the-art clustering approaches viz. Greedy k-member algorithm [4] and Systematic clustering algorithm [7].

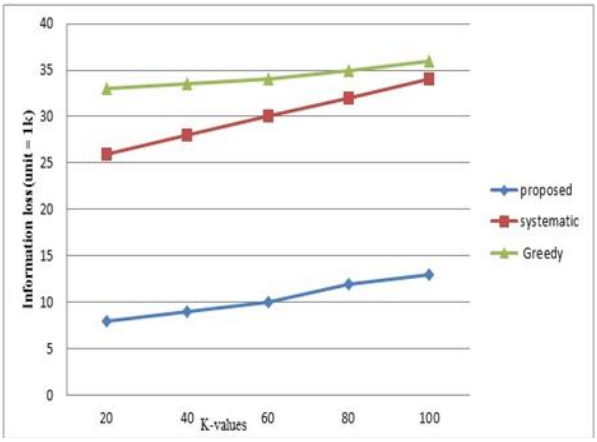

Fig.2: Information loss for Adult database

The Systematic clustering and Greedy k-member algorithm makes use of all the attributes for the construction of an anonymized database. We observed that the Systematic clustering algorithm [7] generates lesser information loss compared to Greedy k-member algorithm [4]. The Greedy k-member algorithm is slow and sensitive to outlier records. Due to the presence of outlier records, the Greedy k-member algorithm attains higher information loss [4]. Conversely, our Approach build sub-databases with a different combination of QI and SA attributes. Our approach builds the clusters using the concept of Systematic clustering algorithm [7]. By selecting a combination of QI and SA attributes, we could display minimum number of attribute in an anonymized database. Also, we use a systematic clustering algorithm to add the record in clusters whose information loss is the lowest. Therefore, our Approach achieves lesser information loss and faster in creating the cluster compared to Systematic clustering [7] and Greedy k-member algorithm [4].

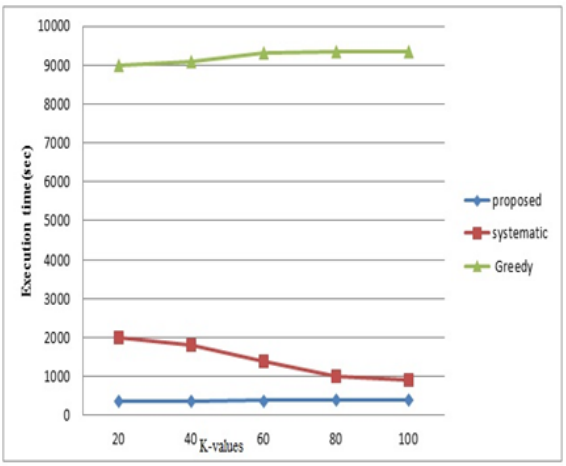

Fig.3: execution time for Adult database

DOI10.17148/IJARCCE.2017.6763

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 7, July 2017

In Fig. 3, we show that our Approach achieves lesser execution time compared with the greedy k-member algorithm [4] and systematic clustering algorithm [7]. This is because; we use the minimum number of attributes in the generated sub-databases. Subsequently, we add the records in a cluster in a systematic way using systematic clustering algorithm [7] for the production of an anonymized database. The Greedy k-member algorithm takes a large amount of time for selecting and adding the records in a cluster from the original database [4]. Therefore, our Approach takes lesser time for the execution compared to existing approaches.

V. CONCLUSION AND FUTURE WORK

In this paper, we present Approach: Equal combination of quasi-identifier and sensitive attribute. Our approach generates anonymized sub-databases with a minimum number of attribute to reduce the risk of disclosure of sensitive attribute. The proposed approaches use a concept of systematic clustering algorithm for the generation of clusters to achieve lesser information loss and execution time. The experimental result shows that our proposed approach generate lesser information loss and execution time compared to Greedy k-member algorithm and Systematic clustering algorithm.

In privacy preserving data mining, information loss is one of the prime research issue discussed in the existing literature. A number of privacy preservation models have been proposed in the literature viz. l-diversity [10], t-closeness [11] and (α, k) anonymity [12]. Therefore, our future work is to use our proposed approach on the privacy model such as l-diversity [10], t-closeness [11] and (α, k) anonymity [12].

In addition, our Approach considers one SA for generating the sub-databases with QI attributes. Thus, our further work would be to investigate the performance on a combination of multiple SA and QI attributes.

REFERENCES

1. Y. Lindell and B. Pinkas, "Privacy preserving data mining," Journal of Cryptology,Vol. 15, 2002, pp. 177-206.
2. M. Upmanyu, A. M. Namboodiri, K. Srinathan, and C. V. Jawahar, "Efficient privacy preserving k-means clustering," Intelligent and Security Informatics, LNCS,Vol. 6122, 2010, pp. 154-166.
3. L. Sweeney, "k-Anonymity: a model for protecting privacy," International Journalon Uncertainty, Fuzziness and Knowledge-based Systems, Vol. 10, 2002, pp. 557-570.
4. J. W. Byun, A. Kamra, E. Bertino, and N. Li, "Efficient k-anonymization usingclustering techniques," in Proceedings of International Conference on DatabaseSystems for Advanced Applications, 2007, pp. 188-200.
5. G. Loukides and J. Shao, "Capturing data usefulness and privacy protection in anonymization," in Proceedings of ACM Symposium on Applied Computing, 2007,pp. 370-374.
6. J.-L. Lin and M.-C. Wei, "An Efficient clustering method for k-anonymization," in Proceeding of International Workshop on Privacy and Anonymity in Information Society, 2008, pp. 46-50.
7. M. E. Kabir, H. Wang and E. Bertino, "Efficient systematic clustering method for k-anonymization," Acta Informatica, Vol. 48, 2011, pp. 51-66.
8. Pawan R. Bhaladhare and Devesh C. Jinwala, "Novel Approaches for Privacy Preserving Data Mining in k-Anonymity Model," Journal of Information Science and Engineering 32, 63-78 (2016).
9. X. Xiao and Y. Tao, "Anatomy: simple and effective privacy preservation," in Proceedings of the 32nd International Conference on Very Large Data Bases, 2006, pp. 139-150.
10. A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "l-diversity: privacy beyond k-anonymity," in Proceedings of the 22nd International Conference on Data Engineering, 2006, pp. 1-12.
11. N. Li, T. Li and S. Venkatasubramanian, "t-closeness: privacy beyond k-anonymity and l-diversity," International Conference on Data Engineering, 2007, pp. 106-115.
12. R. C.-W. Wong, J. Li, A.W.-C. Fu, and K. Wang, "(α, k) anonymity: an enhanced k-anonymity model for privacy preserving data publishing," in Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2006, pp. 754-759.
13. Dr. Poornima B and Ashok K, "A Survey of Latest Developments in Privacy Preserving Data Publishing," International Journal of Advanced Information Science and Technology (IJAIST) Vol.32, No.32, December 2014
14. UCI machine learning repository, http://archive.ics.uci.edu/ml/datasets.html.