Privacy Preservation in public graph by Clustering Techniques

N.Vanitha* & Dr T. Bhuvaneswari**
*Research Scholar, ManonmanamSundaranar University, Tirunelveli, Tamilnadu.
**Supervisor & Asst.Professor, Department of Computer Applications, Queen Mary's College, Chennai -600004

Received: June 04, 2018 Accepted: July 28, 2018

ABSTRACT *In current scenario, preserving privacy is a challenging problem to tackle especially when sharing the graph data generated through social network users need to share for business analytics and social science research purposes. Analytical values are present in medical records, corporate data etc. in which definite attributes are eliminated and the data are made to be clean. Several techniques are suggested in the field of data mining such as sanitized data for subsequent mining techniques, privacy preserving and data publishing. In the last few years, researchers have been focusing on k-anonymization techniques. The essential needs for these techniques are ensuring data anonymization while at the same time minimize the information loss, that cause from data modifications. This paper suggest to minimize information loss that uses the idea of clustering to ensure good data quality. The important objective of preservation is data records that are certainly similar to each other which should be the part of identical equivalence class. The definite clustering problems are formulated by using this technique is said to be K- member clustering problem.*

Keywords: *Privacy Preservation, public graph, Clustering Techniques, k-anonymization, k-member clustering.*

1. Introduction

K-anonymity is the recent approach that addresses data privacy. These approach ensure the data privacy in the released data that any record which is indistinguishable from (K-1) other record which respect to the set of attributes known as the quasi-identifier. The concept of K anonymity is to find out an optimal solution of computational complexity for K-anonymity problem. The key objective of our method is to improve a novel K-anonymization problem that describes the limitations. The main idea of our approach is K-anonymization problem which can be viewed as the clustering problem. The requirement of K-anonymity can be instinctively transformed into clustering. It is to finda set of clusters, each of that which could be containing at least K records. Whereas, the records present in the cluster is as same as other to improve the data quality and similar quasi-identifier value.When the records in a cluster are modified it ensures that less distortion is required. Thus convey a certain clustering problem, which can be called k-member clustering problem. This problem in Non-Deterministic Polynomial-Time hardness(NP) which runs in time O(n2) and present in greedy algorithm. Though this algorithm can integrate such information which does not trust or generalization orders, if happens some natural relations among the values in a domain to find more desirable solutions. Consequently, the proposal of quality metrics have introduced for precisely measures of information loss is represented through hierarchy-free generalization which is not yet introduced. Hence, the data quality metric is used for the hierarchy-free generalization which is also known as information loss metric. This paper shows small modification in our algorithm which is able to reduce classifications of errors effectively and organized as follows. Section 2 describes the anonymity model and existing techniques. Section 3 describes the k-anonymization as a clustering problem approach. Section 4 describes the experimental results.Section 5describes the result and discussion and conclusion.

2. Concepts of Anonymity model

Anonymitymodel is most classic model that can prevent suppression or generalizing portion of the microdata so, no one can distinctly distinguish from a group of size K[1]. The concept of K-anonymity model is to convert a table in which no one can make high probability associations between the records in table and the equivalent entities. To accomplish the goal, K-anonymity model requireany one record that is identical from others (K-1) records associate with the group of attributes is said to be Quasi-identifier. The class of records which are identical with each other is mentioned as an equivalence class.

Table.1 Raw Medical Data Set

I	QI			SA
Name	**Sex**	**Age**	**Postcode**	**Illness**
Bill	M	20	13000	Flu

Ken	M	24	13500	HIV
Linda	F	26	16500	Fever
Mary	F	28	16400	HIV

Table. 2 A 2 Anonymous Data Set Of Table.1

I	QI			SA
Name	**Sex**	**Age**	**Postcode**	**Illness**
Bill	M	[20,24]	13*00	Flu
Ken	M	[20,24]	13*00	HIV
Linda	F	[26,28]	16*00	Fever
Mary	F	[26,28]	16*00	HIV

Table.1 shows storing the private information about the set of individuals. The attributes in table are classified as

1. Identifiers(I),
2. Quasi-Identifiers (QI),
3. Sensitive Attribute (SA)

Raw medical data set (Table.1) consist of age, postcode and attribute sex from QI. The patient records 1 and 2 have unique which may be re-identified easily for their combinations of age, postcode and sex are unique shown inA 2 anonymous dataset of Table.1 (Table.2). These tables make two patients and less likely to re-identified [2].

2.1 Related works

Anonymization techniques establish privacy in big data when dealing with quasi identifier attributes. This paper proposes new algorithm known as k-anonymity without prior value of the k threshold [3].Usually, usage of anonymization techniques addresses privacy issue that deals with both quasi identifier and sensitive attributes. A recent research hasdetermined some techniques which are inappropriate due to existence of quasi identifier attributes in the data set such as Sex, Age and Date of birth [4]. Attributes of quasi identifier denotes a set of datawhich can be joinedwith some background information to re-identify entities [5].According to the principle of k-anonymity, a tuple in the published dataset which is identical from k-1 other tuples in the data set. Hence, an invader who identifies the values of quasi-identifier of an individual which is not able to separate the records from (k-1) other records [6]. K-anonymity uses suppression and generalizationtechniques which are used to hide the identity of an individual [7].Zhen Tu et al. in [8] author proposes k-anonymity, t-closeness and l-diversityof tracks through a definite generalization approach, while itensures the minimum loss of Spatio-temporal granularity. The anonymization method sustain personalized privacy requirement which is present inan utility-driven adaptive clustering method has proposed with similar best data quality to partition tuples [9].This paper investigated anonymization effect due to k-anonymity on the data mining classifiers. Naïve Bayes classifier is calculated using anonymized and non-anonymized dataof results presentingthe increase anonymitythatprimes to proportional degradation of classifier performance [10].Thealgorithm used to parallelize k-anonymity by utilizing MapReduce programming paradigm with Hadoopand also observed the two attack models on k-anonymity such as background knowledge attack and homogeneity attack.The result of the proposed algorithm is capable of anonymizing big data in order to support the data mining technique of privacy preserving [11]. The main focus ofthe study is to protect against identity disclosure on k-anonymity techniques. The privacy level and mining quality of the anonymized dataset will be using decision tree classification and compared with the other data mining technique which is support vector machine and logistic regression [12].In order to prevent such attacks, K-anonymization is usedby modifying microdata with potential increase of data [13].The author extends[14] the k-anonymity model to the (α, k)-anonymity model to limit the implications from the quasi-identifier to a sensitive value within αto protect the information sensitive from being concluded by strong implications[15]. The concept of multi-dimensional k-anonymity [17][16]where the data generalization is over multi-dimension at a time, and [18] extents the multi-dimensional generalization to anonymize data for a certaintask such as grouping.

3. k-Anonymization as a Clustering Problem

Clustering problems require exact number of clusters in the solutions which does not limit the no of clusters, instead of k records in each cluster. Hence, the problem is referred to as k-member clustering or k-anonymity problem.

Definition 1.Let S represents the set of n records and k represents the specified parameter of anonymization. The optimal solution of the k-clustering problem is a set of clusters $E = \{e_1, \ldots, e_m\}$ which is

fit to the following requirements:

$\forall i \neq j = \{1,2,\dots,m\}, e_i \cap e_j = \emptyset$ ---------- (1)

$\cup_{i=1,2,\dots,m} e_i = S$ ---------- (2)

$\forall e_i \in E, |e_i| \geq k$ ---------- (3)

$\sum_{l=1,2,\dots,m} |e_l| . \max_{i,j=1,2,\dots,|e_i|} \Delta(p(l,i),p(l,j))$ ---------- (4)

Here $|e_i|$ is the size of cluster e, p represents the i-th data point in cluster e.Let Δ(X, Y) represents the distance between the two data points of X and Y. From the definition sum of all intra-cluster distance is defined as the maximum distance between any two data points in the cluster.

Definition 2. (Distance between two numeric values)

Let D be a finite numeric domain. Let $V_iV_j \in$Drepresent normalized distance between the two values and D represent a finite numeric domain.

$\delta_N(v_i, v_j) = |v_i - v_j| / |D|,$ ……(5)

where |D| is the domain size measured by the difference betweenminimum andmaximum values in D.The most straightforward solution is to assume that every value in such a domain is equally different to each other; e.g., the distance of values within the two attributes are same then the distance is said to be 0 or else the distance is said to be 1. However, some domains may have few semantic relationships among the values and it is desirable to define the distance functions based on the existing relationships.Such r elationships can be easily captured in a taxonomy tree.

Definition 3. (Distance between two categorical values)

Let D represents the categorical domain, T_D represents taxonomy tree which is defined for D. The distance between two normalized values v_i ,v_j is defined as:

$\delta_C(v_i, v_j) = H(\Lambda(v_i, v_j)) / H(T_D),$ ……(6)

where $\Lambda(v_i, v_j)$ represents the subtree rooted at lowest common ancestor of v_i and v_j, and $H(T_D)$ be the height of tree T .However, we can say taxonomy tree not as a restriction, but a user's preference.

Definition 4.(Distance between two records)

LetQ_T={$N_1, \dots, N_m, C_1, \dots, C_n$} be the quasi-identifier of table T , where N_i(i= 1, . . . , m) is anattribute with a numeric domain and C_j (j = 1, . . . , n) is an attribute with a categorical domain. The distance of two records r_1, $r_2 \in T$ is defined as:

$$(r_1, r_2) = \sum_{i=1,\dots,m} \delta_N(r1[Ni], r2[Ni]) + \sum_{j=1,2,\dots,n} \delta_C A = \pi r^2 (r1[Cj], r2[Cj]), \dots\dots\dots\dots\dots\dots (7)$$

wherer_i [A] represents the value of attribute A in r_i , and δ_N and δ_C are the distance functions defined in Definitions 2 and 3, respectively.

Definition 5. (Information loss)

For the numerical attribute, consider a table T with quasi identifier (A1.....An). Suppose a tuple t = (x1.....xn) is generalized to tuple t'= ([y1,z1],.....[yn,zn]) such that yi ≤ xi ≤ zi (1≤ i≤n). Then we de fine Normalized Certainty Penalty (NCP) of tuple t on an attribute A_i as:

$$NCP_{A_i}(t) = \frac{z_i - y_i}{|A_i|} \quad \dots\dots\dots\dots (8)$$

Where $|A_i| = \max_{t \in T} t.A_i - \min_{t \in T} t.A_i$

Hierarchical tree are used for the generalization in categorical attribute. The attribute values of different granularity are specified by the hierarchical tree. Suppose a tuple t has value v on categorical attribute A_i and it is generalized to a set of values v1,.....vm.

Definition 6.(Total information loss)

LetEbe the set of all equivalenceclasses in the anonymized table AT. The overall amount of information loss inAT is defined as:

Total-IL(AT) = $\sum_{e \in E} IL(e)$ ……(9)

The cost function of k-members problem in the cluster is defined as the maximum distance between any two data points. Consider the records in which each cluster are minimizing the total loss of information and generalized of anonymized table that can spontaneously minimize the cost function of k-members clustering problem. Therefore, total IL is defined as the cost function to minimize the clustering process.

3.1Anonymization Algorithm

The k-member clustering is potentially exponential as an optimal solution for discussing clustering problem. In order to describe the problem of the computational complexity which precisely define as a decision problem as follows.

Definition 7.(k-**member clustering decision problem**)

Theorem 1. Thek-member clustering decision problem is Nondeterministic Polynomial time(NP)-complete.

Proof. Given a set of n records, initially $r_i$record israndomly picked and make it as a cluster e_1. Later the record of$r_j$makes IL ($e1 \cup \{r_j\}$) minimal. The process get repeat until $|e_1| = k$. When $|e_1|$ reaches k, a record is most distance from $r_i$and the clustering process get repeated until less than k records left. After the iteration process is over, records are inserted into each cluster with respect to the increment of the information lossis minimal.

Theorem 2. Let us consider 'k' represents the specified anonymity parameter and 'n' represents the total number of input records. Each cluster in greedy k-member algorithm to find the k records but less than 2k-1 records.

Proof. Let '*S*' represents the set of input records. In order to find a cluster with exactly k records, the number of remaining records which is equal to or more than k, each and every cluster contains k-records.In the worst case, $k-1$ remaining records are added to a single cluster which already contains *k* records. Hence, the cluster maximum size is $2k-1$.

Theorem 3. Letnrepresents the total numbers of input records andkrepresents the specifiedanonymity parameter. The time complexity of the greedy k-member clustering algorithm is in O (n^2).

Function greedy_k_member_clustering (S, k)

Input: a set of records *S* and a threshold value *k*.

Output:a set of clusterseach of which containsat least*k* records.

```
1.  if( | S |     k )
2.   return S;
3.   end if;
4.   result = ∅; r = a randomly picked record from S;
5.  while( | S | ≥k )
6.   r = the furthest record from r;
7.   S = S – {r};
8.   c = {r};
9.   while( | c | <k )
10.    r = find_best_record(S, c);
11.    S = S – {r};
12.    c = c ∪ {r};
13.   end while;
14.   result = result ∪ {c};
15.   end while;
16.   while( | S | ≠ 0 )
17.   r = a randomly picked record from S;
18.  S = S – {r};
19.  c = find_best_cluster(result, r);
20.  c = c ∪ {r};
21.   end while;
22.   return result;
End;
```

Function *find_best_record* (S, c)

Input: a set of records *S* and a cluster *c*.

Output: a record $r \in S$ such that $IL(c \cup \{r\})$ is minimal.

```
1. n = |S|; min = ∞; best = null;
2. for(i = 1,…n)
3.   r = i-th record in S;
4.   diff = IL(c ∪ {r}) – IL(c);
5.   if( diff < min )
6.     min = diff;
7.     best = r;
```

8. end if;
9. end for;
10. return best;

End;

Function *find_best_cluster* (C, r)

Input: a set of clusters *C* and a record *r*.

Output: a cluster $c \in C$ such that $IL(c \cup \{r\})$ is minimal.

1. n = |C|; min = ∞; best = null;
2. for(i = 1,…n)
3. c = i-th cluster in C;
3. diff = IL(c ∪ {r}) – IL(c);
4. if(diff < min)
5. min = diff;
6. best = c;
7. end if;
8. end for;
9. return best;

End;

4. Experimental Results

The main goal of the experiments is to investigate the performance of proposed approach in terms of data quality, efficiency, and scalability.

In order to evaluate greedy k member clusteringwith another algorithm, namely the median partitioning algorithm.

4.1 Data Quality and Efficiency

In this section, the report experimental results on the greedy k-members algorithm for data quality and execution efficiency.Figure.1 reports the Total-IL costs of the three algorithms (median partitioning, greedy k-member, and greedy k-member modified to reduce classification error)with increasing values of *k*.

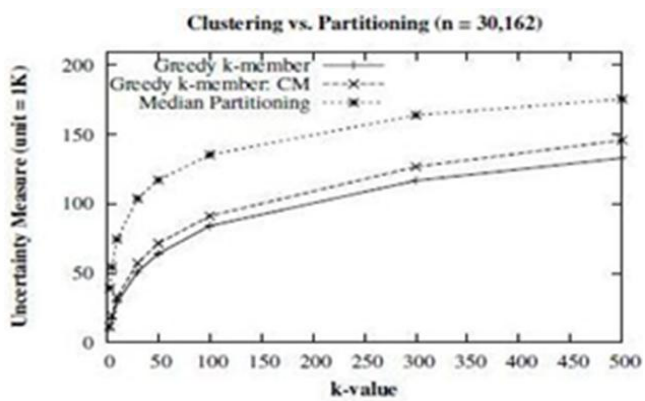

Figure.1 Information Loss Metric

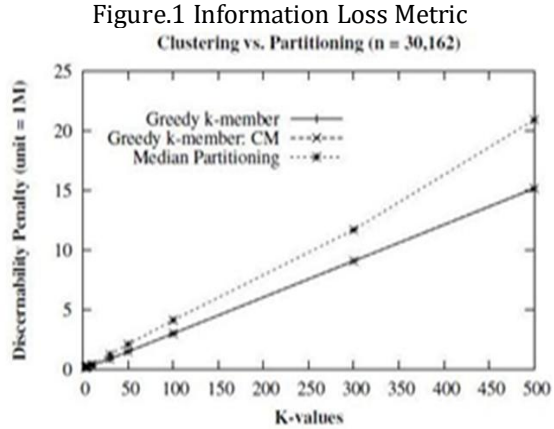

Figure.2 Discernibility Metric

The figure.1 illustrates, the greedy k-members algorithm results in the least cost of the Total-IL for all k values. The Total-IL cost of the modified greedy k-member is very close to the cost of the unmodified algorithm. The superiority of proposed algorithms over the median partitioning algorithm results from the fact that the median partitioning algorithm considers the proximity among the data points only with respect to a single dimension at each partitioning. Another metric used to measure the data quality is the Discernibility Metric (DM), which measures the data quality based on the size of each equivalence class. Intuitively data quality diminishes as more records become identical with respect to each other, and DM effectively captures this effect of the k-anonymization process which shows the DM costs of the three algorithms for increasing k values. The two greedy k-member algorithms perform better than the median partitioning algorithm which is shown in figure.2whereas, the greedy k-member algorithms always produce equivalence classes with sizes very close to the specified k due to the formation of clusters.

4.2 Scalability

The Total-IL costs and execution-time behaviors of the algorithms for various table cardinalities (for k = 5). In this proposed method,the subsets are used as the adult dataset with different sizes and Total-IL costs increase relatively linear with the size of the dataset for both algorithms. However, the greedy k-member algorithm presentsthe least Total-IL cost for any size of dataset. When compared to partitioning algorithm it is slower but still most of the cases are adequate based on the good performance with respect to Total- IL metric.

5 Conclusion

This study has proposed an effective K-anonymization algorithm through the transformation of K-anonymity problem to K-member clustering problem. The two major factors proposed for clustering are cost and distance that is accurately illustrated for the K-anonymization problem. The proposed technique highlight both IL metric and cost metric which are generally seizure the data distortion is introduced by generalization process. It is quiet enough to be used for data quality metric for K-anonymized dataset.

Reference

[1] Luo Yongcheng, Le Jiajin and Wang Jian, "Survey of Anonymity Techniques for Privacy Preserving", 2011,IACSIT Press, Singapore.

[2] Raymond Chi-Wing Wong, Jiuyong Li, Ada Wai-Chee Fu and Ke Wang, "(α,k)-Anonymity: An Enhanced k-Anonymity Model for Privacy-Preserving Data Publishing",2006,August 20–23, 2006, Philadelphia, Pennsylvania, USA.

[3] Zakariae El Ouazzani and Hanan El Bakkali, "A new technique ensuring privacy in big data: K-anonymity without prior value of the threshold k", 2018, Published by Elsevier B.V.

[4] IleasPramanik, Raymond Lau, and Wenping Zhang (2016) "K-Anonymity through the Enhanced Clustering Method"', in IEEE 13th International Conference on e-Business Engineering (ICEBE).

[5] Priyank Jain, ManasiGyanchandani, and NilayKhare. (2016) "Big data privacy: a technological perspective and review." Journal of Big Data3 (1): 25.

[6] Zakerzadeh, Hessam, Charu C. Aggarwal, and Ken Barker. "Privacy-preserving big data publishing",Proceedings of the 27th International Conference on Scientific and Statistical Database Management". ACM, 2015.

[7] Russom and Yohannes,"Privacy preserving for Big Data Analysis", MS thesis, University of Stavanger, Norway, 2013.

[8] Zhen Tu, Kai Zhao, Fengli Xu, Yong Li, Li Su, and DepengJin (2017) "Beyond K-Anonymity: Protect Your Trajectory from Semantic Attack"', in IEEE 14th International Conference on Sensing, Communication, and Networking (SECON).

[9] Jinling Song, Liming Huang, Gang Wang, Yan Kang and Haibin Liu, "The K-Anonymization Method Satisfying Personalized Privacy Preservation",2015,The Italian Association of Chemical Engineering,VOL. 46.

[10] J. ParanthamanandDr. T. Aruldoss Albert Victoire, "Performance Evaluation of K-Anonymized Data" , 2013, Global Journal of Computer Science and Technology Software & Data Engineering Volume 13 Issue 8 Version 1.0.

[11] Y. Sowmya and Dr. M. Nagaratna, "Parallelizing K-Anonymity Algorithm for Privacy Preserving Knowledge Discovery from Big Data" ,2016,International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 2 (2016) pp 1314-1321.

[12] Abdul 'Azim Mohammad and Maheyzah Md. Siraj, "Privacy Preserving Data Mining Based on K-Anonymity and Decision Tree Classification",UTM Computing Proceedings Innovations in Computing Technology and Applications Volume: 1 | Year: 2016 | ISBN: 978-967-0194-81-3.

[13] Ms. Simi M S, Mrs. SankaraNayaki K and Dr.M.SudheepElayidom, "An Extensive Study on Data Anonymization Algorithms Based on K-Anonymity"IOP Conf. Series: Materials Science and Engineering 225,2017.

[14] Wong R C, Li J, Fu A W, et a1, (α, k)-Anonymity : an enhaned k-anonymity model for privacy-preserving data publishing, Proceedings of the 12th ACM SIGKDD, New York: ACM Press, 2006, pp. 754-759.

[15] Terrovitis, M., Mamoulis, N., and Kalnis, Privacy preserving Anonymization of Set-valued Data, In VLDB, 2008, pp. 115-125.
[16] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, Incognito: Efficient full-domain k-anonymity, In Proceedings of the ACM SIGMOD International Conference on Management of Data, 2005, pp. 49-60.
[17] Xiaojun Ye, Jin, L, Bin Li, A Multi-Dimensional K-Anonymity Model for Hierarchical Data, Electronic Commerce and Security, 2008 International Symposium, Aug. 2008, pp. 327-332.
[18] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan., Workload-aware anonymization, In Proc. of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, August 2006.