ISSN No. 0976-5697

Volume 2, No. 5, Sept-Oct 2011

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

An Effective K-Anonymity Clustering Method for Less Effectiveness on Accuracy of Data Mining Results

M. El-Rashidy*
Dept. of Computer Science & Eng.
Faculty of Electronic Engineering,
Menoufiya University, Egypt.
malrashidy@yahoo.com

T. Taha
Dept. of Electronics& Electrical Communications,
Faculty of Electronic Engineering,
Menoufiya University, Egypt.
taha117@hotmail.com

N. Ayad
Nuclear Research Center
Atomic Energy Authority,
Cairo, Egypt.
n_ayad51@yahoo.com

H. Sroor
Dept. of Computer Science & Eng.
Faculty of Electronic Engineering,
Menoufiya University, Egypt.
dr.hoda_sroor@yahoo.com

Abstract: Data mining technology has interested in means of identifying patterns and trends from large collections of data. It is however evident that the collection and analysis of data that include personal information may violate the privacy of the individuals to whom data refers. The k-anonymity model is one of the most known novel privacy preserving approaches that have been extensively studied for the past few years. In this paper, effective approach that is used the idea of clustering for enforcing the k-anonymity is proposed; the goal of this approach is preserved privacy of data with less effectiveness on data mining results. A set of experiments were carried out on the database of the UC Irvine machine learning repository. The obtained results show that the proposed method keeps data privacy preservation with very low effect on accuracy of data mining results compared with greedy k-member and one pass k-means algorithms.

Keywords: Data mining, privacy preserving data, k-anonymity, greedy k-member, one pass k-means.

I. INTRODUCTION

The problem of privacy preserving data mining has become more important in recent years, because of the increasing ability to store personal data about users, and the increasing sophistication of data mining algorithms to leverage this information. A number of techniques such as randomization and k-anonymity [1-3] have been suggested in recent years in order to perform privacy preserving data mining. Privacy preserving data mining aims at providing a trade-off between sharing information for data mining analysis, on the one side, and protecting information to preserve the privacy of the involved parties on the other side.

Most methods for privacy computations use some form of transformation on the data in order to perform the privacy preservation. Typically, such methods reduce the granularity of representation in order to reduce the privacy. This reduction in granularity results in some loss of effectiveness of data management or mining algorithms [2-4]. These approaches typically are based on the concepts of: loss of privacy, measuring the capacity of estimating the original data from the modified data, and loss of information, measuring the loss of accuracy in the data. In general, the more the privacy of the respondents to which the data refer, the less accurate the result obtained by the miner and vice versa. The main goal of these approaches is therefore to provide a trade off between privacy and accuracy.

The k-anonymity model is one of the most known novel privacy preserving approaches that have been extensively studied for the past few years. K-anonymity is a property that models the protection of released data against possible re-identification of the respondents to which the data refer. Intuitively, k-anonymity states that each release of data must be such that every combination of values of released attributes that are also externally available and therefore exploitable for linking can be indistinctly matched to at least k respondents. For example, patient diagnosis records without conducting the k-anonymity model, it is clear that a diagnosis classifier can be developed using these data to predict patient's illness based on quasi identifier that is a minimal set of attributes in the table that can be joined with external information to re-identify individual records, quasi identifier is understood based on specific knowledge of the domain. In this example, quasi identifier as age, gender, zip code, height, weight and another attribute can be joined to re-identify individual records. If the hospital simply publishes the table to other organizations for classifier development, the organizations might extract patient's disease history based on quasi identifier.

Many clustering techniques have been developed to conduct the k-anonymity protected table using clustering based method. Clustering aims to grouping a set of objects into clusters so that objects in a cluster are similar to each other and are different from objects in other clusters [5]. To ensure data mining performance, usability should be taken into account when constructing the k-anonymity protected table [6]. The less of information distortion in the k-anonymity protected table makes usability of the table is larger. Therefore, a k-anonymity model must minimize the information distortion from its original table. In this paper, new method based on clustering for k-anonymity is proposed. The main goal of our proposed method is those preserve data and minimize information loss as possible that has importance effective on accuracy of data mining results. The proposed method result is compared with that greedy k-member [6] and one pass k-means [7] algorithms using dataset of UC Irvine machine learning repository [8].

We first review the basic concepts on quality metrics that precisely measures k-anonymity effect on data in Section II, and then the survey of the clustering based methods is offered in Section III. Each step of our proposed algorithms will be detailed in Section IV. The performance study of the proposed algorithms based on the extensive experimental results will be discussed in Section V. The conclusions of the work will be offered in section VI.

II. BASIC CONCEPTS

The k-anonymity model has attracted much attention for the past few years. Many approaches have been proposed for k-anonymity. This section first describes the concept of quality metrics including information loss, discernibility metric, and square error criterion which will be used throughout this paper to evaluate the effectiveness of k-anonymity approaches. Then several recently proposed clustering based k-anonymity approaches are also reviewed.

A. Information Loss:

The notion of information loss is used to quantify the amount of information that is lost due to k- anonymity. Let e = {r_1, . . . , r_k} be a cluster where the quasi identifier consists of numeric attributes N_1, . . . , N_m and categorical attributes C_1, . . . , C_n. Let T_{cj} the taxonomy tree defined for the domain of categorical attribute C_i. Let MIN_{Ni} and MAX_{Ni} be the min and max values in e with respect to attribute Ni, and let U_{Ci} be the union set of values in e with respect to attribute C_i. Then the amount of information loss occurred by generalizing e, denoted by IL (e), is defined as [6]:

$$\mathrm{IL(e)}= |e|.\left(\sum_{i=1,\ldots,m}\frac{(\mathrm{MAX}_{\mathrm{Ni}}\ \text{-}\ \mathrm{MIN}_{\mathrm{Ni}}}{|N_i|}+\sum_{j=1,\ldots,n}\frac{H(\wedge(U_{Cj}))}{H(T_{Cj})}\right) \quad (1)$$

Where |e| is the number of records in e, |N| represents the size of numeric domain N, (U_{Cj}) is the sub tree rooted at the lowest common ancestor of every value in U_{Cj}, and H (T) is the height of taxonomy tree T. Let E be the set of all equivalence classes in the anonymzed table AT. Then the amount of total information loss of AT is defined as [6]:

$$\text{Total-IL(AT)}= \sum\nolimits_{e\in E}\mathbf{IL(e)} \quad (2)$$

B. Discernibility Metric:

The discernibility metric assigns a penalty to each tuple based on how many tuples in the transformed dataset are indistinguishable from it. This penalty reflects the fact that a suppressed tuple cannot be distinguished from any other tuple in the dataset. The metric can be mathematically stated as follows [9]:

$$C_{DM}(g,k)=\sum_{\forall E s.t.|E|\geq k}|E|^2+\sum_{\forall E s.t.|E|<k}|D||E| \quad (3)$$

In this expression, the sets E refer to the equivalence classes of tuples in D induced by the anonymization. The first sum computes penalties for each non-suppressed tuple, and the second for suppressed tuples.

C. Square Error Criterion:

The k-anonymity clustering based approaches partition a set of n records into K clusters so that the resulting clusters must have high intra cluster similarity, where as the inter cluster similarity is low, to minimize the effectiveness of information loss which produced from data privacy preservation in data mining results. Clustering similarity is measured in regard to the mean value of the records in the clusters using the square error criterion metric [10].

$$E=\sum\nolimits_{i=1}^{K}\sum\nolimits_{x\in C_i}|x-m_i|^2 \quad (4)$$

Where x is the point in space representing the given record, and mi is the mean of cluster Ci both x and mi are multi dimensional. This criterion evaluates the resulting clusters as compact and as separate of k-anonymity approaches.

III. CLUSTERING-BASED APPROACHES

The greedy k-member clustering algorithm for k-anonymization is given in Byun et al [6]. This algorithm works by first randomly selecting a record r as the seed to start building a cluster, and subsequently selecting and adding more records to the cluster such that the added records incur the least information loss within the cluster. Once the number of records in this cluster reaches k, this algorithm selects a new record that is the furthest from r, and repeats the same process to build the next cluster. Eventually, when there are fewer than k records not assigned to any clusters yet, this algorithm then individually assigns these records to their closest clusters, the time complexity of this algorithm is O(n2).

Another greedy algorithm for k-anonymization Similar to the k-member algorithm is proposed [11], this algorithm builds one cluster at a time. But, unlike the k-member algorithm, this algorithm chooses the seed of each cluster randomly. Also, when building a cluster, this algorithm keeps selecting and adding records to the cluster until the diversity of the cluster exceeds a user defined threshold. Subsequently, if the number of records in this cluster is less than k, the entire cluster is deleted. With the help of the user defined threshold. The time complexity of this algorithm is O(n2log(n)/c), where c is the average number of records in each cluster.

Weighted feature C-means clustering algorithm for k-anonymization is proposed [12]. Unlike the previous two algorithms, this algorithm attempts to build all clusters simultaneously by first randomly selecting [n/k] records as seeds. This algorithm then assigns all records to their respective closest clusters, and subsequently updates feature weights to minimize information loss. This algorithm iterates these steps until the assignment of records to clusters stops changing. As some clusters might contain less than k records, a final step is needed to merge those small clusters with large clusters to meet the constraint of k-anonymity. The time complexity of this algorithm is O(cn2/k), where c is the number of iterations needed for the assignment of records to clusters to converge.

One pass K-means Algorithm for k-anonymization is introduced [7]. This algorithm derives from the K-means algorithm, but it only runs for one iteration. This method has a time complexity of O(n2/k), where n is the number of records. It first partitions all records into n groups, and then adjusts the records in each group such that each group contains at least k records [7].

IV. PROPOSED WORK

The k-anonymity model has been extensively studied for the past few years [6, 7, 11, and 12]. Three algorithms are proposed for enforcement of k-anonymity, and are derived from the k-means algorithm but the minimum number of cluster records is the threshold value for k-anonymity. We evaluate the best proposed algorithm as find the best clusters as compact and as separate as possible.

Algorithm No. 1: This algorithm proceeds in two main functions; clustering function, and adjustment function. Let T denotes the set of records to be anonymized, and K = [n/k], where n is the number of records and k is the threshold value for k-anonymity. During the clustering function, the proposed algorithm first randomly picks K records as the seeds to build K clusters. For each stage, the algorithm updates the clusters centroids. Then, for each record r $\in$T, the algorithm finds the cluster centroid that is closest to r, and assigns r to this cluster. The K clusters are constructed but some of these clusters might contain less than k records. Therefore, the algorithm calls the adjustment function as shown in Figure 4. The goal of the adjustment function is to make every cluster contains at least k records.

The adjustment function removes the records that have most distance from the centroid of the clusters with more than k records. Then, the removed records are added to their respective closest clusters with less than k records until no such cluster exists. If all clusters contain not less than k records and there are still some records not yet assigned to any cluster, then these records are simply assigned to their respective closest clusters. The clustering function repeats this stage until all centroid of the clusters do not change. Algorithm No. 1 is shown in Figure 1. The time complexity of this algorithm is estimated as:

$$Time = c.\frac{(n-1)+(n-2)+.............+k}{k} \approx c.\frac{n(n-1)}{2k} \qquad (5)$$

Therefore, time is in O(cn2/k), where c is the number of iterations needed for the assignment of records to clusters to converge.

Algorithm No. 2: This algorithm proceeds in one function, similar to the clustering function of the proposed algorithm No. 1, but differs in two points. First for each stage, the assignment of each record to the cluster that has closest centroid with it occurred under the condition that the assigned clusters must contain less than k records. So, as the constructed K clusters contain k records or greater second, this function does not call the adjustment function. The proposed algorithm No. 2 is shown in Figure 3.

Algorithm No. 3: This algorithm is similar to the proposed algorithm No. 1. It proceeds in two main functions: clustering function and adjustment function. But, the clustering function does not call the adjustment function for each stage. The adjustment function is called one time after the iterations needed for the assignment of records to clusters to converge when all centroid of the clusters do not change. The proposed algorithm No. 3 is shown in Figure 2.

```
Input  : a set T of n records; the value k for k-anonymity
      Output: a partitioning P= {$P_1$,…….., $P_K$} of T
Let K=[$\frac{n}{k}$];
Randomly select K distinct records $r_1$,….,$r_K$ $\in$ T;
Let $P_i$ ={$r_i$} for i =1 to K;
Repeat
    Let T={$r_1$,…….,$r_n$};
    Update centroid $C_i$; $C_i$ = centroid of $P_i$ for i=1 to K;
    $P_i$ ={ }; i=1 to K;
    While (T $\neq$ 0) do
        Let r be the first record in T;
        Calculate the distance between r to each $C_i$;
        Add r to its closest $P_i$;
        Let T =T/{r};
    End of while
    Call the adjustment partitioning function;
Until all centroid of $P_i$ do not change;
```

Figure 1. Proposed algorithm No. 1.

```
Input  : a set T of n records; the value k for k-anonymity
      Output: a partitioning P= {$P_1$,…….., $P_K$} of T
Let K=[$\frac{n}{k}$];
Randomly select K distinct records $r_1$,….,$r_K$ $\in$ T;
Let $P_i$ ={$r_i$} for i =1 to K;
Repeat
    Let T={$r_1$,…….,$r_n$};
    Update centroid $C_i$; $C_i$ = centroid of $P_i$ for i=1 to K;
    $P_i$ ={ }; i=1 to K;
    While (T $\neq$ 0) do
        Let r be the first record in T;
        Calculate the distance between r to each $C_i$;
        Add r to its closest $P_i$;
        Let T =T/{r};
    End of while
Until all centroid of $P_i$ do not change;
Call the adjustment partitioning function;
```

Figure 2. Proposed algorithm No. 3

```
Input   : a set T of n records; the value k for k-anonymity
Output: a partitioning P= {$P_1$,…….., $P_K$} of T
Let K=[$\frac{n}{k}$];
Randomly select K distinct records $r_1$,….,$r_K$ $\in$ T;
Let $P_i$ ={$r_i$} for i =1 to K;
Repeat
    Let T={$r_1$,…….,$r_n$};
    Update centroid $C_i$; $C_i$ = centroid of $P_i$ for i=1 to K;
    $P_i$ ={ }; i=1 to K;
    While (T $\neq$ 0) do
        Let r be the first record in T;
        If P contains cluster $P_i$ such that |$P_i$| < k then
            Calculate the distance between r to each $C_i$;
            Add r to its closest $P_i$;
        Else
            Add r to its closest cluster;
        End if;
        Let T =T/{r};
    End of while
Until all centroid of $P_i$ do not change;
```

Figure 3. Proposed algorithm No. 2

```
Input   : a portioning p= {$P_1$,…….., $P_K$} of T; the value k  for k-anonymity
Output: an adjusted partitioning P= {$P_1$,…….., $P_K$} of T
Let R = $\phi$;
For each cluster P $\in$ p with |P|>k do;
    Sort records in P by distance to centriod of $P_i$;
    While (|P|>k) do
        r $\in$ P is the record farthest from centroid of $P_i$;
        let P = P\{r};
        R=R $\cup$ {r};
    End while
End for
While (R$\neq$ $\phi$) do
    Randomly select a record r from R;
    Let R =R\{r};
    If p contains cluster $P_i$ such that |$P_i$| < k then
        Add r to its closest $P_i$;
    Else
        Add r to its closest cluster;
    End if;
End of while
```

Figure 4. The adjustment partitioning function

V. EXPERIMENTS AND DISCUSSION

We have evaluated the performance of each proposed algorithm, and compared these proposed algorithms with greedy k-member and one pass k-means algorithms. A set of extensive experiments were carried out on the database of the UC Irvine machine learning repository [8]. It has 14 attributes; nine attributes are used as the quasi-identifiers, including age, work class, education, material status, occupation, race, gender, and native country. Among these, age was treated as continuous numerical attribute, while the other attributes were treated as categorical attributes.

The square error criterion metric for the three proposed algorithms were compared to know which one is the best proposed algorithm. That is to find the one of the intra cluster similarity is high where as the inter cluster similarity is low as possible. The proposed algorithm No. 1 is less cost for all k values compared with the other proposed algorithms, as shown in Figure 5. The proposed algorithm No. 2 omitted probable founding records nearest for the clusters with more or equal than k records than intra clusters records. The proposed algorithm No. 3 omitted the highest of intra cluster similarity and the lowest of inter cluster similarity as possible for each stage and worked it after the clusters are constructed.

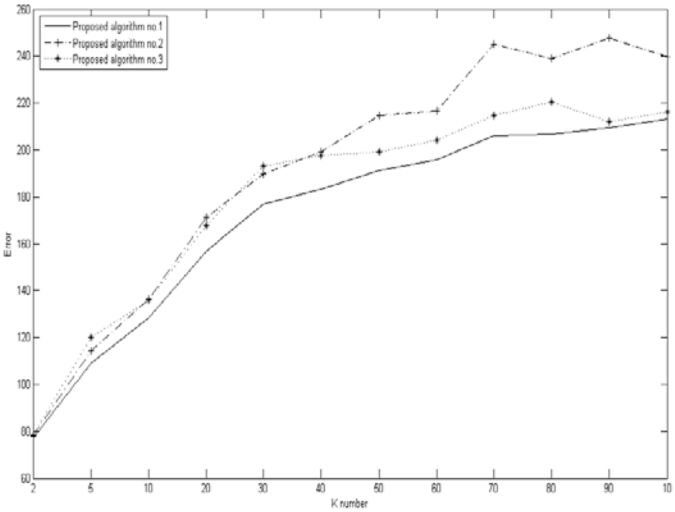

Figure 5. Square error metric of proposed algorithms

We compared the proposed algorithm No. 1 which is the best proposed algorithm of our proposed algorithms with greedy k-member and one pass k-means algorithms using the square error criterion metric (Error), the total information loss, and the discernibility metric (DM)

The proposed algorithm No. 1 is less cost of square error for all k values compared with the other proposed algorithms, as shown in Figure 6, and the results are shown in Figure 7 of the three algorithms for increasing values of k, results of k-member algorithm in the least cost of the total information loss for all k values. This enhancement of k-member algorithm was negatively effect on the accuracy of data mining results shown in Figure 6 and came on the impact of the rest of attributes in clustering process through its reliance on quasi-identifiers only. So, it is greatly importance to take square error metric as a first to know the effectiveness of these algorithms on accuracy of data results. As a result, each of one pass k-means and the proposed algorithm No.1 is interested to all attributes in clustering process, and the least of them in the information loss, and square error is our proposed algorithm.

Figure 8 shows the DM costs of the three algorithms for increasing k values. As shown, the DM cost of the three algorithms is very close to the cost of each other. In fact, the three algorithms always produce equivalence classes with sizes very close to the specified k, due to the way clusters are formed.

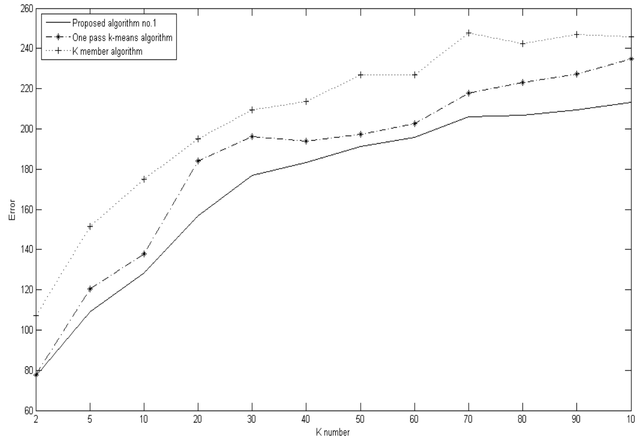

Figure 6. Square error metric of proposed algorithm no.1, k-member and one pass k-means algorithms

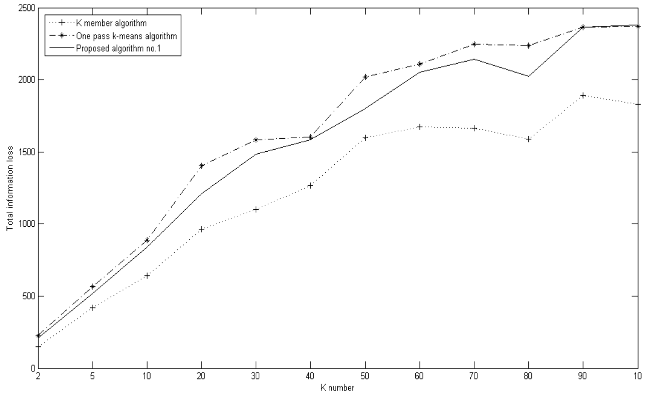

Figure 7. Information loss metric of proposed algorithm no.1, k-member and one pass k-means algorithms

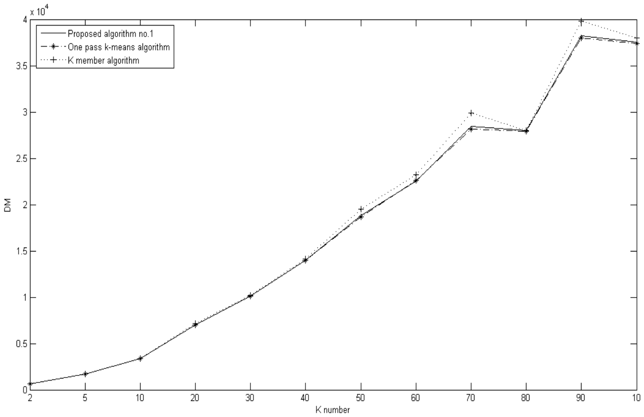

Figure 8. Discernibility metric of proposed algorithm no.1, k-member and one pass k-means algorithms.

VI. CONCLUSION

In this paper, three algorithms are proposed for minimizing data privacy preservation effectiveness as possible on accuracy of data mining results. The obtained results proved that the proposed algorithm NO. 1 keeps data privacy preservation with very low effect on accuracy of data mining results compared with other proposed algorithms, greedy k-member and one pass k-means algorithms.

VII. REFERENCES

[1] Agrawal R., and Srikant R., "Privacy Preserving Data Mining", Proc. of ACM SIGMOD Conference, 2000.

[2] Agrawal D., and Aggarwal C., "Design and Quantification of Privacy Preserving Data Mining Algorithms", Proc. of ACM PODS Conference, 2002.

[3] Samarati P., and Sweeney L., "Protecting Privacy when Disclosing Information: k-Anonymity and its Enforcement Through Generalization and Suppression", IEEE Symp. on Security and Privacy, 1998.

[4] Evfimievski A., Srikant R., Agrawal R, and Gehrke J., "Privacy preserving mining of association rules", Proc. of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Canada, 2002.

[5] Jain A., and Dube R., “Algorithms for Clustering Data”, Prentice Hall, New Jersey, 1988.

[6] Byun J., Kamra A., Bertino E., and Li N., “Efficient k-Anonymization Using Clustering Techniques”, Proc. of the International Conference on Database Systems for Advanced Applications, 2007.

[7] Lin J., and Wei M., “An Efficient Clustering Method for k-Anonymization”, Proc. of Third International Conference on Advanced Data Mining and Applications (ADMA) 2008.

[8] Hettich C., and Merz C., UCI repository of machine learning databases, 1998.

[9] Roberto J., and Agrawal B., “Data Privacy through Optimal k-Anonymization”, Proc. of the 21st International Conference on Data Engineering, 2005.

[10] Jiawei H., and Micheline K., “Data Mining: Concepts and Techniques”, Morgan Kaufmann publishers, 2005.

[11] Loukides G., and Shao J., “Capturing data usefulness and privacy protection in k-anonymisation”, Proc. of the ACM International Conference on Applied Computing, 2007.

[12] Chiu C., and Tsai C., “A k-anonymity clustering method for effective data privacy preservation”, Proc. of the Third International Conference on Advanced Data Mining and Application (ADMA), 2007.