Proceedings of the 2023 International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/8/20230080

K-anonymous mathematical model based on greedy algorithm

Shuyang Chen

Faculty of science, Hangzhou Dianzi University Hangzhou Zhejiang China 310000

chenshuyangcsy@163.com

Abstract. This research propose a user-centered combinatorial data anonymization method. whereas a data matrix is said to be k-anonymous if each row occurs at least k times. Therefore, the authors propose PATTERN-GUIDED k-ANONYMITY, an improved k-anonymization problem. It allows users to designate the combinations in which suppressions may occur, building on prior work and addressing relevant shortcomings. Users of anonymous data can indicate that the aspects of the data are valued differently. The so-called K-anonymity is usually realized by Generalization and Suppression techniques. Generalization refers to Generalization and abstraction of data so that specific values cannot be distinguished, for example, the age data group can be generalized into an age group.

Keywords: K-modes clustering, greedy algorithm, K-anonymity, hiding distance.

1. Introduction

The rapid development of big data analysis technology makes people urgently require to mine more valuable information from big data, and the first step of mining is to have enough public data. However, at present, large data owners such as hospitals, governments, and big data companies that have a large amount of data will inevitably involve citizens' privacy issues when disclosing relevant data.

A traditional paradigm for (combinatorial) data privacy entails making a matrix k-anonymous, meaning that each row must appear at least k times [1][2]. The equally extremely popular "differential privacy" model, which has a statistical rather than combinatorial flavor, is not discussed in this study[3]. It is commonly recognized that the k-anonymity notion has some flaws, such as when the anonymized data is used repeatedly [1]. Due to its simplicity and good interpretability, this paper will focus on k-anonymity. The notion behind k-anonymity is that each row of the matrix symbolizes a different person, and the associated row's k-fold appearance prevents situations where the person or thing behind can be recognized. It is obvious that some information loss, or the suppression of some matrix entries, must be accepted in order to achieve this goal (blanked out). In this method, details about specific traits (represented by the matrix's columns) are lost. Thus, when converting an arbitrary data matrix into a k-anonymous one, it makes sense to try to limit this information loss. Even in special circumstances, the associated optimization problem, k-ANONYMITY, is difficult to approximate and NP-hard [4–8]. To make a matrix k-anonymous, it largely relied on heuristic methods; yet, it played a vital part in many applications [2][9][10]. The "usefulness" (also in terms of expressiveness) of the anonymized data was discovered to require caution [11][12].

DOI: 10.54254/2755-2721/8/20230080

In this paper, the whole dataset is divided into some large clusters by clustering algorithm, and then the greedy algorithm is applied to these divided clusters to K-anonymize them.

In the future, such as medical systems, transportation hub, shopping website traffic platform informatization construction and its application in business surveys and the need of scientific research, a large number of user data collection or real-time grabbing and surveyed sharing released within a certain range, although the data included in the user to interact with platform of legal information necessary, but it also contains many personal privacy information, Information sharing is inevitable. In order to ensure the effectiveness of information and avoid the disclosure of personal privacy, it is necessary to obfuscate the information in advance and hide some unnecessary information through the methods described in this article.

2. Materials and methods

2.1. Modeling ideas and formulas

The global optimum is approximated by finding the local optimum. Each time, the smallest hiding distance is selected, and its data is merged into a new data ○1, and then the hiding distance between data ○1 and other data is calculated. This process is repeated, and finally every data exists in the data group, that is, there is no independent data.

- there is no hiding spot data tuple calculation formula of the hidden distance between:

$$distance(t_i, t_j) = \sum_{k_i \in t_i, k_j \epsilon t_j} 2 \times \delta(k_i, k_j) \tag{1}$$

If different, output 1, otherwise output 0, and the distance is the total number of times two data tuples need to change into the same form.

$$\delta(k_i, k_j) = \begin{cases} 0, & k_i = k_j \\ 1, & k_i \neq k_j \end{cases} \tag{2}$$

- Calculation formula of hidden distance between data tuples and equivalent data groups

$$distance(t_i, ec_j) = distance(t_i, t^*) + w_j \times distance(ec_j, t^*) \tag{3}$$

The ti for the data tuple, ecj for the equivalent data set, wj as the hidden distance weighting, $t*$ for ti with ecj after optimizing the equivalent data set.

$$w_j = |ec_j| \times \alpha \tag{4}$$

The $|ec_j|$ for the equivalent data in data set the number of tuples, α for setting adjustment coefficient, α selection will affect the calculate method of constructing the number of tuples in the equivalent data set.

- the hidden distance formula between the equivalent data set

$$distance(ec_i, ec_j) = distance(ec_i, t^*) + w_j \times distance(ec_j, t^*) \tag{5}$$

The ec_i with ec_j for the two equivalent data set.

2.2. Symbol description

Table 1. Symbol description.

Symbol	Meaning
t_i	Data tuple i
$k_{i,j}$	The jth data in the ith row
ec_j	Equivalent data group

DOI: 10.54254/2755-2721/8/20230080

Table 1. (continued).

α	Adjustment factor, that is, the tuple into the equivalent data group difficulty
p	P heavy anonymous
w_j	Weight of stealth distance
C_{k1}	Results of hybrid clustering algorithm

2.3. Process

Data input: contain n record table data, parameter p values

Output: k after the anonymous form

Step 1: Calculate the distances between data tuples or equivalent data groups in the data table to obtain the lower triangular moment matrix for storing the distances.

Step 2: Select the element with the smallest value in the distance matrix and merge the data tuples or equivalent data groups represented by its rows and columns into a new cluster.

Step 3: Perform the optimal hiding processing on the current cluster to form an undifferentiated equivalent group, and bring the modified hidden data into the data table to update the distance matrix.

Step 4: Repeat steps 1, 2, and 3 until each individual is in an undifferentiated equivalent data group.

							0	1	2	3	4	5
0	0	0	1	0	1	0	–					
1	0	1	1	0	1	1	4	–				
0	1	1	1	0	0	2	6	6	–			
1	0	0	0	0	1	3	4	10	10	–		
0	1	1	1	0	1	4	4	4	2	8	–	
0	1	0	0	1	0	5	8	12	6	8	8	–

Figure 1. Distance matrix update process.

3. Result

It demonstrated that the greedy heuristic outperforms the optimal solution generated by the ILP implementation in three scenarios using real-world datasets in terms of solution quality. Although the heuristic in this study is significantly more efficient than the exact technique, the results are still rather close to the optimum and in many situations, they were optimal (the ILP was, on average, more than 1000 times slower). With a higher degree k of anonymity, the heuristic findings tend to be closer to the ideal number of suppressions

DOI: 10.54254/2755-2721/8/20230080

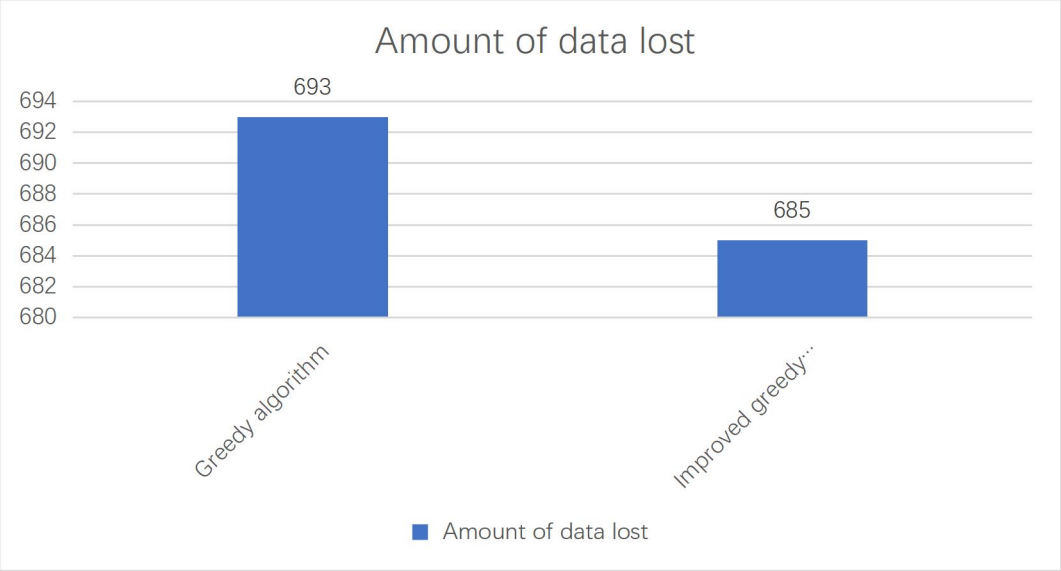

Figure 2. The amount of data lost in a data table.

4. Discussion

4.1. Advantages of the model

The solution method of this model is universal and simple, and can be applied to solve high-dimensional data sets and large data sets. In dealing with the large data set with high dimension p - heavy protection problems, still can draw better data anonymous.

Hybrid clustering greedy model has good expansibility. For high dimension, large amount of data sets, under appropriate k clustering parameters value, can at the same time of stable data loss, sharply reduce operation time; Also, for low dimensional degree, amount of data on small data sets, increasing the clustering parameters k values, the better the results of the data hiding can be.

4.2. Model shortcomings

Due to the limitation of the greedy algorithm, the model results only in the distance parameter α situation of the local optimal solution, and is not the global optimal solution, and to a large amount of data calculation, the calculation takes too long.

5. Conclusion

This research introduces the clustering algorithm K-$mode$, combined with the density-based initial point selection method, combining the hybrid clustering algorithm with the greedy algorithm, and propose a data hiding method based on the clustering and the greedy algorithm.

This result fits both hypotheses perfectly: 1. Different data refers to differences, and the case that multiple data are in the same state is not considered. 2. Assume that the local optimum is taken several times, and the final result is the global optimum.

Future work will focus on creating practical algorithms for the general k-anonymity problem using the theoretical findings from this study. More particularly, it entails connecting and transforming the Restricted k-anonymity problem and the general k-anonymity problem in order to produce a new, efficient exact method and provide a superior approximate algorithm scheme.

We have fulfilled a number of unresolved issues on the subject by proving the hardness and viability of numerous strategies employed in database privacy. The conclusion is that, even in highly special instances, the majority of these problems are challenging to answer optimally; nevertheless, in some fascinating cases, these problems can be handled more quickly. Several intriguing unanswered problems explore potential workarounds for this intractability:

DOI: 10.54254/2755-2721/8/20230080

- -How close can the difficult problems be to being solved? The best known approximation approach for k-anonymity, presented by Park and Shim [13], for instance, only suppresses O(log k) times the ideal number of entries. When there are less attributes, may approximation ratios be improved?
- -The best known running time for Simplex Matching is $O(n^3 + n^2m^2)$ steps [14]. Here, the number of nodes in the hypergraph is n, and the number of hyperedges is m. Because the author included a hyperedge for each triple, the procedure for 2-anonymity in this study uses n, which is also the number of database rows, whereas $m = \binom{n}{3} = O(n^3)$. As a result, the algorithm used in this paper for 2-Anonymity has a running time of $O(n^8)$. Is it possible to lower this exponent to a more usable running time?

This research is only theoretical at this point. It is now clearer how structural characteristics of the data matrices may affect the computational complexity of the generally NP-hard data anonymization problems explored here, but it is still not apparent whether some of the proposed solutions would be applicable in real-world settings. However, it is obvious that further reducing exponential factors.

Acknowledgment

I would like to thank professor Venkatesan Guruswami and teacher assistant, for their great support in doing implements and experiments. I'm grateful to anonymous reviewers of algorithms for constructive and extremely fast feedback.

Reference

[1] Fung, B.C.M.; Wang, K.; Chen, R.; Yu, P.S. Privacy-preserving data publishing: A survey of recent developments. ACM Comput. Surv. 2010, 42, 14:1–14:53.

[2] Navarro-Arribas, G.; Torra, V.; Erola, A.; Castellà-Roca, J. User k-anonymity for privacy preserving data mining of query logs. Inf. Process. Manag. 2012, 48, 476–487.

[3] Dwork, C. A firm foundation for private data analysis. Commun. ACM 2011, 54, 86–95.

[4] Bonizzoni, P.; Della Vedova, G.; Dondi, R. Anonymizing binary and small tables is hard to approximate. J. Comb. Optim. 2011, 22, 97–119.

[5] Han Jianmin, Cen Tingting, Yu Huiqun. [1] han j m, cen t t, yu h q. Research on micro-aggregation algorithm for k-anonymization of data tables [A]. Acta electronica sinica, 2008,36 (10) : 1-7

[6] Chakaravarthy, V.T.; Pandit, V.; Sabharwal, Y. On the complexity of the k-anonymization problem.2010, arXiv:1004.4729.

[7] Blocki, J.; Williams, R. Resolving the Complexity of Some Data Privacy Problems. In Proceedingsof the 37th International Colloquium on Automata, Languages and Programming (ICALP '10), Bordeaux, France, 6–10 July 2010; Springer: Berlin/Heidelberg, Germany, 2010; Volume 6199,LNCS, pp. 393–404.

[8] Meyerson, A.; Williams, R. On the Complexity of Optimal k-Anonymity. In Proceedings of the 23rd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS '04), Paris, France, 14–16 June 2004; ACM: New York, NY, USA, 2004; pp. 223–228.

[9] Campan, A.; Truta, T.M. Data and Structural k-Anonymity in Social Networks. In Proceedings of the 2nd ACM SIGKDD International Workshop on Privacy, Security, and Trust in KDD (PinKDD '08), Las Vegas, NV, USA, 24 August 2008; Springer: Berlin/Heidelberg, Germany, 2009; Volume 5456, LNCS, pp. 33–54.

[10] Gkoulalas-Divanis, A.; Kalnis, P.; Verykios, V.S. Providing k-Anonymity in location based services. ACM SIGKDD Explor. Newslett. 2010, 12, 3–10.

[11] Loukides, G.; Shao, J. Capturing Data Usefulness and Privacy Protection in k-Anonymisation. In Proceedings of the 2007 ACM Symposium on Applied Computing, Seoul, Korea, 11–15 March 2007; ACM: New York, NY, USA 2007, pp. 370–374.

DOI: 10.54254/2755-2721/8/20230080

[12] Rastogi, V.; Suciu, D.; Hong, S. The Boundary between Privacy and Utility in Data Publishing. In Proceedings of the 33rd International Conference on Very Large Data Bases, VLDB Endowment, Vienna, Austria, 23–27 September 2007; pp. 531–542.
[13] Park, H., Shim, K.: Approximate algorithms for K-anonymity. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 67–78 (2007)
[14] Anshelevich, E., Karagiozova, A.: Terminal backup, 3D matching, and covering cubic graphs. In: Proceedings of the 39th ACM Symposium on Theory of Computing, pp. 391–400 (2007)