Using Deep Neural Networks to predict the parameters, like K value, in K-anonymity

Group: yayaya

Members:

108753134高士傑、108753124楊芝辰、108753109陳庭軒、108753102吳映函

1. Abstract

We want to predict good K values in K-anonymity through Neural Networks. Among them, we set the "good" K value to a point where the difference is relatively high if the loss value (cumulatively) is around 30,000. After our experiments, found in the closest possible K value is 12 we assumed.

1. Introduction

When we use K-anonymity for privacy protection, it also causes data loss. This situation becomes more apparent when the K value is increased. Considering the problem of data loss, we cannot choose a K value that is too large to achieve an acceptable level of data loss. However, if the K value is lowered, the level of privacy protection will also be lowered. In this way, the purpose of using K-anonymity is not achieved. To find a trade-off between privacy protection and data loss, we use the concept of loss function to help us in determining how to choose between the two. Among them, the loss function we use is implemented by reference methods [1]. We first calculate the loss values corresponding to the K values from 1 to 14 and graph it. Use the cumulative loss value about 30,000 and the “difference” between the loss values of two continuous K values to estimate the best K value.

1. Related work

There are several algorithms that have been proposed for finding K value. Sweeney’s Datafly approach uses a heuristic method to generalize the attribute containing the most distinct sequence of values for a provided subset of quasi-identifiers [3]. Samarati’s algorithm [4] can identify all K–minimal generalizations, out of which an optimal generalization can be chosen based on certain preference information provided by the data recipient. Iyengar proposes a flexible generalization scheme and uses a genetic algorithm to perform k–anonymity on the larger search space that resulted from it [5].

1. Method description
2. Mondrian：We use Mondrian to implement k-anonymity, which is a Top-down greedy data anonymization algorithm for relational dataset, proposed by Kristen LeFevre in his papers[6]. Here is the basic workflow of Mondrian：
   1. Partition the raw dataset into k-groups using kd-tree. k-groups means that each group contains at least k records.
   2. Generalization each k-group (Fig. 1(b)), such that each group has the same QID\*.

Fig. 1 Anonymity, Privacy and Generalization

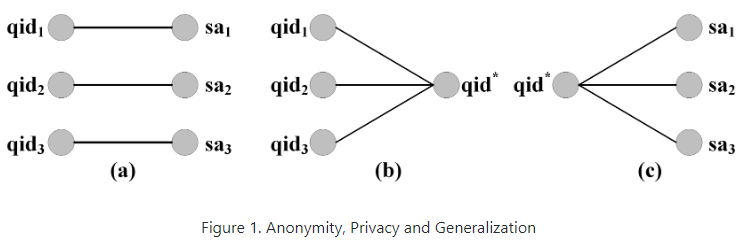
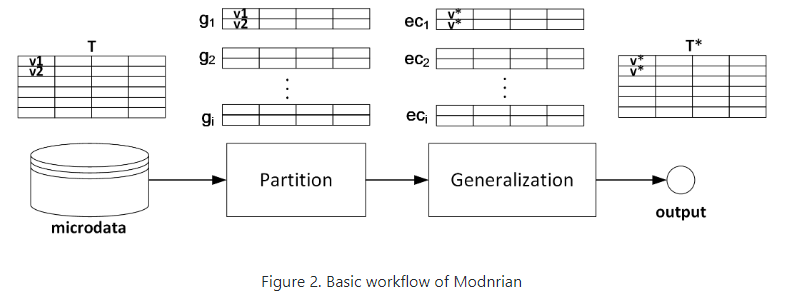
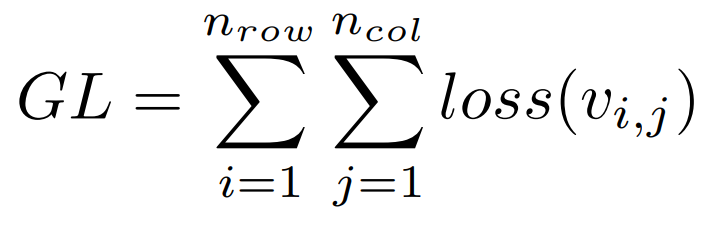


Fig. 2 Basic workflow of Modnrian



1. Loss function：We use loss function to calculate the loss value.

,

where I represent row, and j represent column. represent the ordered set. generalization Gj for an attritube aj is partitioning of the set.

We experiment the loss value when K is from 1 to 14. After calculating the loss value respectively, we took the result to draw a graph. The graph is cumulative, and we want to choose the spot near 30000 loss value. Finally, we draw the graph of the difference between the loss values of two continuous K values, and choose the spot that the difference is relatively high. After this steps, we can find the K value.

1. Experimental result

We use Fig.3 to Fig.7 to show our experiment result.

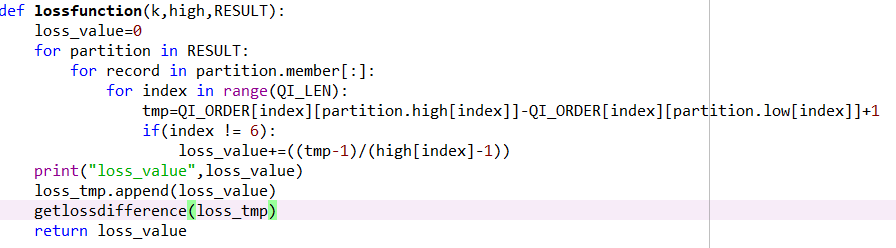


Fig.3 code screenshot 1

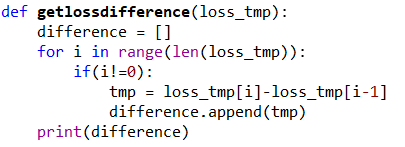


Fig. 4 code screenshot 2

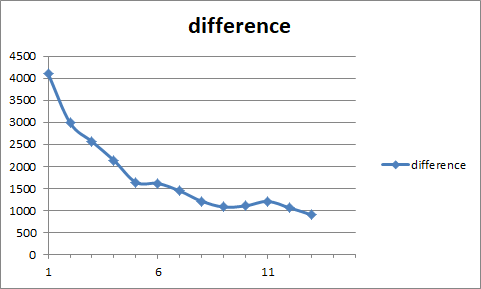


Fig. 5 the graph of the difference of the continuous loss value

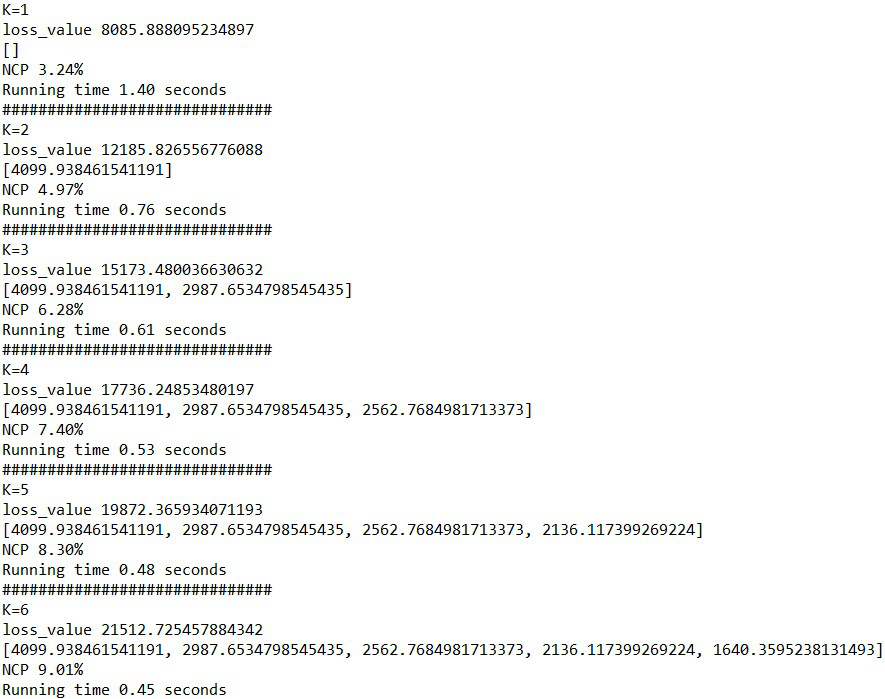


Fig. 6 result of the experiment 1

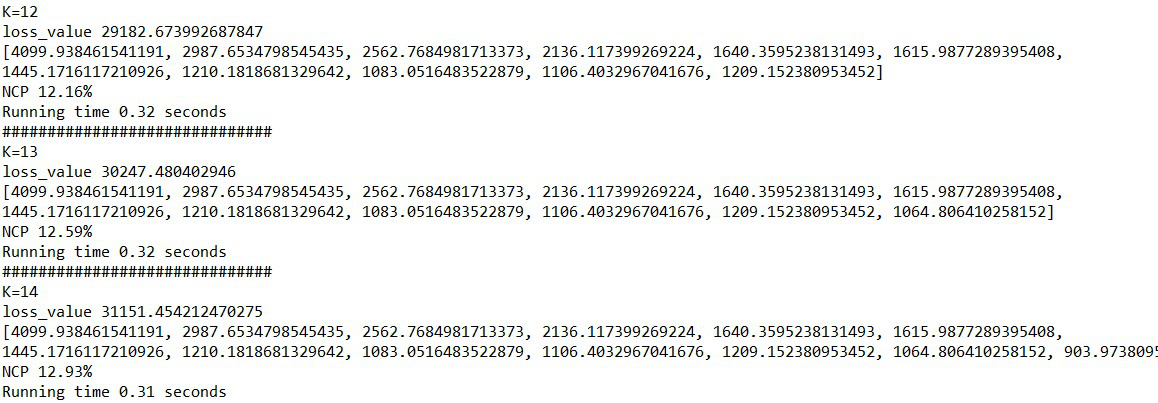


Fig. 7 result of the experiment 1

1. Conclusion

Since we cannot calculate the loss function with K values, we do not know how to predict the best K by neural networks. Although we have no way to use neural networks to predict the best K value, we use the loss value to estimate the best K value. We list the loss value one by one to show the trend of the loss value in a chart and then calculate the difference value between the loss value of the two continuous K values to analyze the best K value. Even though we do not actually use neural networks, we still try to find the better K values. The part that we can keep studying in the future is to implement the prediction of the loss value by neural networks, so that our report would be closer to this topic.

1. Reference

[1] On the Optimal Selection of k in the k–Anonymity Problem. Rinku Dewri, Indrajit Ray, Indrakshi Ray and Darrell Whitley.

<https://www.cs.colostate.edu/~genitor/2008/2008IEEEICDE.pdf>

[2] L. Sweeney, “k–Anonymity: A Model for Protecting Privacy,” International Journal on Uncertainty, Fuzziness and Knowledge-based Systems,vol. 10, no. 5, pp. 557–570, 2002.

[3] L. Sweeney, “Achieving k–Anonymity Privacy Protection Using Generalization and Suppression,” International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, vol. 10, no. 5, pp. 571–588, 2002.

[4] P. Samarati, “Protecting Respondents’ Identities in Microdata Release,” IEEE Transactions on Knowledge and Data Engineering, vol. 13, no. 6, pp. 1010–1027, 2001.

[5] V. S. Iyengar, “Transforming Data to Satisfy Privacy Constraints,” in Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Alberta, Canada, 2002, pp. 279–288.

[6] K. LeFevre, D. J. DeWitt, R. Ramakrishnan. Mondrian Multidimensional K-Anonymity ICDE '06: Proceedings of the 22nd International Conference on Data Engineering, IEEE Computer Society, 2006, 25