Automation and Machine Learning (2022)
Clausius Scientific Press, Canada

DOI: 10.23977/autml.2022.030206
ISSN 2516-5003 Vol. 3 Num. 2

Research on fast clustering privacy protection of mixed data based on blockchain

Zhuohong Zhang, Guihua Huang

Guangdong University of Science and Technology, Guangzhou, 510000, China

Keywords: Cloud environment; Mixed data; Fast clustering; Block chain

Abstract: In the actual production environment, a large number of data sets to be protected are non-single type attribute data sets, that is, mixed data sets.In view of the above problems, this project proposes a fast clustering privacy protection research method based on blockchain mixed data. This method is mainly for complex data types, cloud environment for different data types with different measurement method, the difference of the hybrid data set by calculation of each sample point neighborhood density and relative distance, divided into k density and relative distance far sample points as the initial clustering center, clustering, complete and upload them to block chain. For the generated clustering results, the numerical clustering centers are calculated, and the set of attribute values of non-numerical data is generated, which ensures that each user can correctly obtain the iterative process and the final clustering centers, and reduces the error rate of data.

1. Introduction

With the continuous development of the era of big data, the amount of data is increasing day by day. Within the scope permitted by law, reasonable use of these data for data analysis to obtain the hidden information in the data has become a method followed by all walks of life. However, in the process of data analysis, all data are exposed to the network indiscriminately, which will cause huge hidden dangers and harms to the privacy of data owners. Therefore, reasonable use of data in the network and consideration of personal privacy data has become an urgent problem to be solved in cloud environment.

At the same time, with the increase of data volume, the data type also changes from single numerical data to multiple types of mixed data. The traditional privacy protection methods mostly focus on the research of numerical data, ignoring the influence of other types of data, resulting in low data privacy level and low security factor. Aiming at the above problems, this paper proposes a research method of fast clustering based on mixed data.

2. Related Work

Yan et al.[1]mainly focused on the processing of general numerical data sets by using the form of histogram in order to protect privacy. Fletcher et al.[2]proposed a differential privacy decision forest algorithm for single-type data, which can shorten the query time of data and reduce the

addition of noise. Liu et al.[3]proposed a clustering based differential privacy data publishing method, but this method is also only applicable to numerical data sets. The above methods can only protect the privacy of a single type of data source to a certain extent. In practical applications, there are a large number of other types of data. Soria et al.[4]proposed a micro-aggregation algorithm for mixed attribute datasets, which could effectively deal with differential privacy protection for mixed data types. Li et al.[5]proposed a method combining k-anonymity and differential privacy to release structural data in micro clusters, which is not applicable to the processing of large mixed data sets. Dan et al.[6]proposed a privacy-preserving k-means clustering algorithm that not only satisfies differential privacy but also has the nature of approximation error, in which the approximation error of the algorithm has a sublinear relationship with the dimension of data. Ni et al.[7]proposed a Differentially Private k-means Clustering algorithm based on Cluster Merging (DP-KCCM).

3. Basic Knowledge

3.1 Differential privacy protection

Definition 1 For random algorithm A, R_A is the set composed of all the output results of algorithm A, and R_A is any subset of R_A. For any two adjacent datasets D_0 and D_1, the algorithm satisfies the following formula[8]:

$$pr[A(D_0) \in S_A] \leq e^{\varepsilon} \times pr[A(D_1) \in S_A] \tag{1}$$

Then algorithm A satisfies ε-differential privacy, where is the privacy protection budget. The privacy protection intensity of algorithm A can be measured by ε, ε the smaller it is, the higher the privacy protection intensity is; otherwise, the lower it is.

3.2 Laplacian mechanism

Theorem 1 for the existing data set D, let it have a query function $f : D \to D^{'}$. If the algorithm K satisfies:

$$K(D) = f(D) + Lap(\frac{\Delta f}{\varepsilon}) \tag{2}$$

Then algorithm K provides ε-differential privacy protection. Where Δf represents the global sensitivity and $Lap(\frac{\Delta f}{\varepsilon})$ represents the amount of noise added to the data set.

4. Fast clustering privacy protection algorithm for mixed data based on blockchain

4.1 Measurement Method

For hybrid data sets, this paper divides the data sets into two types: one is numerical data, the other is non-numerical data divided according to attributes. Different measurement methods are designed according to the characteristics of these two types of data.

For numerical data, Minkowski distance calculation method is adopted in this paper, that is, for a given sample $x_i = (x_{i1}, x_{i2}, \cdots, x_{in}) \in R^N$ and $x_j = (x_{j1}, x_{j2}, \cdots, x_{jn}) \in R^N$, p=1 is taken in this paper, that is, the formula of the distance between samples is:

$$dist_{num}(x_i, x_j) = \sum_{k=1}^{n} \left| x_{ik} - x_{jk} \right| \tag{3}$$

For non-numerical data, this paper measures the distance between samples by measuring the dissimilarity degree.a certain type of attribute x_{ik} and x_{jk} is simply matched:

$$f(x_{ik}, x_{jk}) = \begin{cases} 1, x_{ik} = x_{jk} \\ 0, x_{ik} = x_{jk} \end{cases} \tag{4}$$

Then the distance between the two samples is defined as:

$$dist_{nonum}(x_i, x_j) = \frac{1}{n} \sum_{k=1}^{n} f(x_{ik}, x_{jk}) \tag{5}$$

According to the above method, for a mixed data set $X = \{x_1, x_2, \cdots, x_n\}$, each sample $x_i (i = 1,2, \cdots, n)$ has q attributes, that is $r_1, r_2, \cdots, r_m, r_{m+1}, \cdots, r_q$, let $r_1, r_2, \cdots, r_m$ where be numerical data and $r_{m+1}, \cdots, r_q$ be non-numerical data. The initial cluster center $C = \{c_1, c_2, \cdots, c_k\}$ is randomly selected, then the distance between the sample and the cluster center is:

$$dist(x_i, c_j) = dist_{num}(x_i, c_j) + \beta dist_{nonum}(x_i, c_j) \tag{6}$$

4.2 Initial cluster center

According to the neighborhood density calculation method, the initial cluster center of the mixed data is obtained:

1) Calculate the neighborhood density N_ρ for each sample of the initial dataset X;

2) Arrange the neighborhood density $Z = \{Z_1, Z_2, \cdots, Z_n\}$ of each sample as in descending order, and add the sorted sample with the largest neighborhood density Z_1 into a new set M;

3) If there are any $dist(Z_j, M_i) > L$ in set Z (L represents distance threshold), they will be added to set M until all sample elements in set Z are iterated. At this time, set M is the initial clustering center of this dataset, $|M|$ where the number of clusters is.

4.3 Data disturbance

The clustered data set should be perturbed to achieve the purpose of differential privacy protection. In this paper, Laplacian mechanism is adopted to perturb data for numerical data, which is defined as follows:

$$c_{iu}^{'} = c_{iu} + Lap(\frac{\Delta f}{\varepsilon}) \tag{7}$$

For non-numerical data, exponential mechanism is used for data perturbation, which is defined as follows:

$$c_{ju}^{'} = \{r \mid pr[r \in c_{ij}] \propto \exp(\frac{\varepsilon \mu(z_{ju}, r)}{2\Delta l})\} \tag{8}$$

4.4 Iterative clustering

1) According to the distance formula, calculate the distance $dist(x_i, c_j)$ between each sample x_i in the original data set X and $|M|$;

2) Re-calculate the clustering center of each cluster according to numerical type and non-numerical type;

3) According to the calculated clustering center, judge whether the data in the original cluster has changed. If there is no change, the clustering ends and the clustered data set is obtained.

5. Experimental Analysis

5.1 Data set processing

In this paper, the abalone dataset is selected for experiments. Firstly, the invalid data and attributes are processed, and 4177 data records are retained. In this paper, four attributes in the data are selected for experimental analysis, including the numerical data age and time and the non-numerical data sex and name.

This paper analyzed the algorithm by evaluating the accuracy of data clustering results, and used variance formula [9] to measure the accuracy of data set results. The specific formula is as follows:

$$NIV = \frac{1}{N}\left(\sum_{j=1}^{m}\sum_{x_i \in C_i}\left\|x_i(f_j) - C_i\right\|^2 + \frac{\sum pr(x_i(f_j))}{p}\right) \quad (9)$$

5.2 Experimental results and analysis

During the experiment, the algorithm in this paper is compared with the traditional DP-KCCM algorithm and MDAV algorithm [10]. The traditional DP-KCCM algorithm only considers the numerical data in the data set and ignores the non-numerical data. The MDAV algorithm is based on the microclustering method for differential privacy protection.

Figure 1 shows the error rate of clustering results of the three algorithms on the dataset. Under the same value, the error rate of the proposed algorithm is lower, and the error rate tends to be stable with the increase of the value.

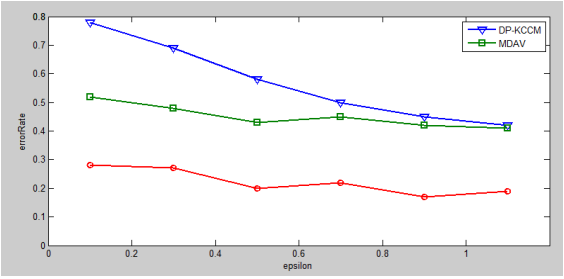

Figure 1: Error rate of clustering results

Figure 2 shows the changes of NIV value of the three algorithms. It can be seen from Figure 2 that the NIV value of the algorithm proposed in this paper is significantly lower than that of the other two algorithms, and much lower than that of DP-KCCM algorithm, indicating that the clustering effect of the algorithm in this paper is more obvious.

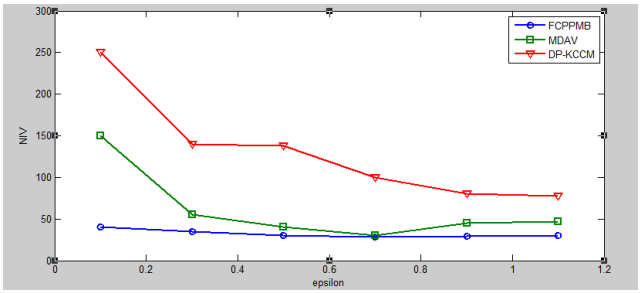

Figure 2: The changes of NIV value of the three algorithms

The experimental results show that the error rate of the data processed by the algorithm in this paper is lower and the accuracy of the clustering result is higher, which is more suitable for the privacy protection of mixed data. However, its time complexity and space complexity need to be optimized.

6. Conclusion

In view of the problems of privacy right now, in this paper, based on block chain of mixed data privacy protection fast clustering algorithm, based on the difference of privacy protection algorithm, combined with the transparent characteristics of block chain, using the clustering algorithm, the classification and clustering data processing, at the same time adding noise disturbance, implement the data so as to realize data privacy protection.

Acknowledgment

This work is supported by the PCB high-speed intelligent detection equipment project of Dongguan Science and Technology Correspondent Project in 2022(20221800500712).

References

[1] YAN F, ZHANG X, LI C, et al. Differentially private histogram publishing through fractal dimension for dynamic datasets[C]//IEEEConference on Industrial Electronics and Applications (ICIEA), 2018: 1542-1546.

[2] FLETCHER S, ISLAM M Z. Differentially private random decision forests using smooth sensitivity [J]. Expert Systems with Applications, 2017, 78(7): 16-31.

[3] LIU X Q, LI Q M. Differentially private datarelease based on clustering anonymization [J]. Journal on Communications, 2016, 37(5): 125-129.

[4] SORIA C J, DOMINGO F J, SANCHEZ D, et al. Improving the utility of differentially private data releases via k-anonymity[C]//Melbourne: 12th IEEE International Conference on Trust, Security andPrivacy in Computing and Communications(TrustCom), 2013.

[5] LI N H, QARDAJI W, SU D. Provably private data anonymization:or, k-anonymity meets differential privacy[DB/OL]. https: //arxiv.org/abs/1101.2604v1, 2011.

[6] DAN F, XIANG C Y, ZHU R H, et al. Coresets for differentially private k-means clustering and applications to privacy in mobile sensor networks [C]// Proceedings of the 2017 ACM/IEEE International Conference on Information Processing in Sensor Networks. New York: ACM, 2017: 3-15.

[7] NI T J, QIAO M H, CHEN Z L, et al. Utility-efficient differentially private K-means clustering based on cluster merging[J]. Neurocomputing, 2021, 424: 205-214.

[8] XIONG P, ZHU T Q, WANG X F, et al. A survey on differential privacy and applications[J]. Chinese Journal of Computers, 2014, 37(1): 101-122.

[9] NISSIM K, RASKHODNIKOVA S, SMITH A. Smooth sensitivity and sampling in private data analysis[C]// Proceedings of the 39th annual ACM Symposium on Theory of Computing, 2007: 75-84.

[10] NAYAHI V, KAVITHA V. Privacy and utility preserving data clustering for data anonymization and distribution on Hadoop[J]. Future Generation Computer Systems, 2016, 74(9): 393-408.