Journal of Information Hiding and Privacy Protection
DOI:10.32604/jihpp.2021.012193
Article

Tech Science Press

XGBoost Algorithm under Differential Privacy Protection

Yuanmin Shi[1,2], Siran Yin[1,2], Ze Chen[1,2] and Leiming Yan[1,2,*]

[1]School of Computer & Software, Nanjing University of Information Science & Technology, Nanjing, China
[2]Engineering Research Center of Digital Forensics, Ministry of Education, Nanjing, China
[*]Corresponding Author: Leiming Yan. Email: lmyan@nuist.edu.cn
Received: 30 August 2020; Accepted: 12 December 2020

Abstract: Privacy protection is a hot research topic in information security field. An improved XGBoost algorithm is proposed to protect the privacy in classification tasks. By combining with differential privacy protection, the XGBoost can improve the classification accuracy while protecting privacy information. When using CART regression tree to build a single decision tree, noise is added according to Laplace mechanism. Compared with random forest algorithm, this algorithm can reduce computation cost and prevent overfitting to a certain extent. The experimental results show that the proposed algorithm is more effective than other traditional algorithms while protecting the privacy information in training data.

Keywords: Differential privacy; privacy protection; XGBoost algorithm; CART regression tree

1 Introduction

With the development of information technology and the multi-application of big data, the demand for data has exploded in medical systems, social networks, e-commerce and logistics systems, which promotes the release, sharing and analysis of data. The efficient use of data brings us convenience but it also has the danger of personal privacy leaks. In the era of big data, personal privacy information is flood-ing and the probability of privacy leaks is increasing, although it can achieve privacy protection by deleting unique identifiers or sensitive information. However, non-sensitive information can also be learned through certain algorithms to infer person-al identity. Moreover, during the process of using data, if there is no protection for privacy, then it will have the threat of privacy exposure. Therefore, processing data in order to protect privacy is essential before it is released.

For dealing with privacy leak, Dwork et al. first proposed a strict and provable privacy protection model in 2006—Differential Privacy Protection [1]. Differential privacy protection has several advantages over other privacy protection models. First, differential privacy protection assumes that the attacker has the most back-ground knowledge. Under this assumption, differential privacy protection can deal with various new types of attacks without considering any possible background knowledge owned by the attacker [2]. Second, it has a solid mathematical foundation, a strict definition of privacy protection and a reliable quantitative evaluation method, which makes the level of privacy protection under different parameter processing comparable [3].

While maintaining personal privacy, how to analyze from privacy-protected data, and better data mining while ensuring privacy has become a hot topic in current research. An important method in data mining analysis is classification. Xgboost is an important popular classifier, so this paper proposes a xgboost algorithm under privacy protection. Compared with the random forest algorithm, this algorithm is better in preventing over-fitting. For samples with missing feature values, it can automatically learn its split direction.

2 Related Works

Now, technologies for privacy protection mainly include random perturbation of data such as adding noise, generalizing data, not releasing certain sensitive data or thresholds, encrypting data and so on.

Samarati and Sweeney proposed the k-Anonymity model. Its main idea is to sup-press or generalize the information to ensure each record in the table has exactly the same quasi-identifier attribute value as the other k-1 records in the data table [4]. In general, the greater the anonymity parameter k, the better the privacy protection. k-Anonymity can guarantee that an attacker cannot confirm the corresponding person with a certain data. When giving a certain person, the attacker cannot confirm whether it has some kind of sensitive attribute. An attacker can use homogenous attack or background knowledge attack to obtain personal information, thus it causes privacy disclosure. This is because k-Anonymity does not restrict sensitive attributes.

Dwork et al. proposed differential privacy protection [1]. Differential privacy protection assumes that the attacker has the most background knowledge. Under this assumption, differential privacy protection can deal with a variety of new types of attacks [2], thus avoiding background knowledge attacks under k-Anonymity.

In order to better perform data mining under privacy protection, some people combine the classifier with privacy protection. The following are some different classifier algorithms based on differential privacy.

Blum et al. proposed SuLQ-based ID3 [5], its main idea is to add noise to the in-formation gain according to the Laplace mechanism when selecting the classification properties of the decision tree, but the noise greatly reduces the prediction accuracy.

Mcsherry has improved on SuLQ-based ID3, then they proposed PINQ-based ID3 [6], which uses the Partition operator to divide the data set into disjoint subsets, but because the privacy protection budget allocated to each query is not relatively large, the noise added to the information gain in ID3 cannot be reduced.

Friedman et al. proposed DiffP-ID3, which is improved on ID3 based on exponential mechanism [7], this algorithm effectively reduces noise and increases the utilization of the privacy protection budget. Then, in order to be able to process continuous attribute values, Friedman and Schuster proposed the DiffP-C4.5 algorithm [7], but this algorithm must first use exponential mechanism to select splitting points for all continuous features in each iteration, and then pass the data including results and all discrete features through the exponential mechanism again to select the final splitting scheme ,which consumes too many privacy protection budgets because it needs to call the exponential mechanism twice [8].

Mohammed et al. proposed the DiffGen algorithm, which uses generalization technique and top-down segmentation technique, then combine them with the exponential mechanism and information gain to determine the classification characteristics [9]. However, since each classification feature corresponds to a classification tree, when the feature attribute dimension of the classification in the data set is very large, this method needs to maintain a large number of classification trees, which greatly reduces the efficiency of the selection method based on the exponential mechanism, and it is possible that this method will run out of privacy budget.

Zhu et al. proposed the DT-Diff algorithm after improving the DiffGen algorithm [10], this algorithm constructs model features, assigns continuous features to a certain weight, and then groups the samples together with the discrete attributes, then adds noise. This algorithm can reduce the number of times the exponential mechanism is called, thereby reducing the consumption rate of the privacy budget.

Patil et al. applied differential privacy protection to random forests and pro-posed the DiffPRF algorithm [11], this algorithm is based on the decision tree of ID3 but it can only handle discrete features, so feature discretization is required first before processing continuous features.

From the above-mentioned several classifier algorithms base on differential privacy, it can be observed that the main problem is that the algorithm needs to be able to process continuous features and solve the consumption of privacy budget, that means it should improve the efficiency of the index mechanism call as much as possible. The algorithm proposed in this paper is based on the characteristics of xgboost to directly

process continuous features and improve the efficiency of using the exponential mechanism. Experimental results can verify that the algorithm can improve the accuracy of classification.

3 Theoretical Basis

3.1 Differential Privacy

Definition 1: ε-differential privacy [12]. A random function F, Pr [] is set to represent the risk of privacy disclosure, Range(F) represents the range of values. For any two adjacent data sets like D_1 and D_2, any subset S of $Range(F)$, if the algorithm F satisfies

$$Pr\,[F(D_1) \in S \leq exp\,(\varepsilon) \times Pr\,[F(D_2) \in S] \tag{1}$$

The algorithm F is called ε-differential privacy protection, exp is the index function at the bottom of e, and the parameter ε is called privacy protection budget, the smaller ε the higher the degree of privacy protection is.

Define 2: Global sensitivity [12]. Given any function: $f: D \to R^d$, the input is a data set D, and the output is a d dimension real number vector. For any adjacent data set D_1 and D_2

$$\Delta f = \max_{D_1, D_2} \|f(D_1) - f(D_2)\| \tag{2}$$

Called global sensitivity of f. in which R represents real space of mapping $\|f(D_1) - f(D_2)\|$ is the 1-order norm distance between $f(D_1)$ and $f(D_2)$.

Theorem 1: Laplace mechanism [13]. For a given data set D, there is a query function $f: D \to R^d$ with a sensitivity of Δf, if the following formula is met, the expression $M(D)$ satisfies ε-differential privacy protection.

$$M(D) = f(D) + \left(Laplace\left(\frac{\Delta f}{\varepsilon}\right)\right)^d \tag{3}$$

Among them, $Laplace\left(\frac{\Delta f}{\varepsilon}\right)$ is random noise, which obeys Laplace distribution with scale parameter ε. The amount of noise is directly proportional to the sensitivity of Δf and inversely proportional to the privacy protection budget ε.

Theorem 2: Index mechanism [14]. Let the input of the random algorithm F D be the data set, the output $r \in Range(F)$ be an entity object, $q(D, r)$ be the availability function, Δq be the sensitivity of the function named $q(D, r)$. If the algorithm $Range(F)$ is selected and output with a probability proportional to $exp(\varepsilon)$, then the algorithm F provides ε-differential privacy protection.

3.2 XGBoost

Xgboost (extreme gradient boosting) algorithm is an improved algorithm [15]. Xgboost is a kind of integrated learning, which is combined by classification regression tree (CART tree). The objective function of xgboost is expressed as:

$$L(\phi) = \sum_i l(\hat{y}_i - y_i) + \sum_k \Omega(f_k) \tag{4}$$

The prediction model can be expressed as follows:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \tag{5}$$

IN this, K is the total number of trees, f_k representing the k tree, $\hat{y}_i$ representing the prediction results of samples x_i.

The loss function is expressed as

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \tag{6}$$

IN this, K is the total number of trees, f_k representing the k tree, $\hat{y}_i$ representing the prediction results of samples x_i.

The loss function is expressed as

$$Obj(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{7}$$

$l(y_i, \hat{y}_i)$ is the training error of the sample x_i, $\Omega(f_k)$ representing the regular term of the kth tree.

The complexity is written as:

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{8}$$

w_j is the weight of the jth leaf.

The loss function is replaced by mean square error and expanded by Taylor formula. Finally, the final objective function can be obtained by partial derivation and simplification

$$Obj^* = -\frac{1}{2}\sum_{j=1}^{T} \frac{G_j^2}{H_j+\lambda} + \gamma T \tag{9}$$

4 XGBoost Algorithm under Differential Privacy Protection

The xgboost establishment process under differential privacy is described as follows:

First, the privacy budget $\varepsilon_a$is evenly divided among n iterations according to the number of iterations in the parameters, and then the decision tree is generated according to the same rule as below. The strategy of generating the decision tree is as follows:

Firstly, data set D is divided into training set D_i and test set D_t in a certain proportion, and then the attribute set A_D is divided into training attribute set A_{tD} and test attribute set A_{iD}. The tag set L_D performs the same operation to obtain the results L_{tD} and L_{iD}. In each iteration, the privacy budget of each tree is equally divided into each layer, then the privacy budget of each layer is divided into two halves. The classification result of the first tree is compared with the actual result to obtain difference value, which is used as the training sample of the next tree until the residual is 0 or less than a certain threshold, then the generation is completed. First, the Laplace mechanism is used to add noise to the current node. Then, determine whether the current node reach the termination condition. If so, mark this as a label and use the Laplace mechanism to add noise to the category. Otherwise, you continue to select the classification attribute. Finally, a decision tree satisfying the ε-differential privacy is obtained by repeating the above steps.

Then, for each record in the test set, the decision tree satisfying the decision tree satisfying the differential privacy generated by above steps is used for classification prediction. Starting from the root node of the current tree, determine which child node to input according to the classification result of the current node until reaching a certain leaf node. Then get the predicted result L_p of the current tree. After that, the residual is obtained by comparing the predicted result of each tree with the actual value, the residual is input to the next tree, and the steps above are repeated to select the category with the minimum residual. Finally, add residuals for each tree to get the final prediction L'_p.

The establishment process of this algorithm refers to the DiffPRFs algorithm [16], But the algorithm supposed in this paper is different from the DiffPRFs algorithm in that the exponential mechanism is not used to select the split-point, and the classification algorithm is different.

Table 1: XGBoost algorithm under differential privacy protection

Algorithm: XGBoost Algorithm under Differential Privacy Protection
Input: training data set D， attribute set A_D , label sets L_D, the times of iterations is n, privacy budget ε_a, the maximum depth of the tree d.

1) Customize a certain ratio in D to select training set D_i and verification set D_t;
2) Filling the record which has a default value with its mode of attributes;
3) Encoding non-numeric characters in data set A_D;
4) Dividing the attribute sets A_D into training attribute sets A_{iD} and testing attribute sets A_{tD} Dividing the label sets L_S into training label sets L_{tD} and testing label sets L_{iD};
5) According to feature importance, select certain features F'_{iD} for training;
6) $\varepsilon_1 = \varepsilon_a / n$;
7) for i =1 to n do

$$\varepsilon = \varepsilon_1 / 2(d+1)$$

$$N_{S_i} = Laplace\_mech(A'_{iD}, \Delta A'_{iD}, \varepsilon);$$

The classification result of the first tree is compared with the actual result to obtain difference, and the difference is used as a training sample for the next tree until the residual is 0 or less than a certain threshold, then the generation is completed.

End;

1) Achieving trees xgboost;
2) For each record of A_{tD};
3) According to classification query rules to get the prediction result of the first tree L_p;
4) The residuals are obtained by comparing the prediction results of each tree with the actual values, so that the class with the minimum residual is selected;
5) Add the residuals of each tree to get the final predicted result L'_p.

Output: the set of all classification results

5 Experiments

This study uses Python to implement the xgboost algorithm for differential privacy protection. In terms of experimental data, use the Adult data set in the UCI machine learning database to test the accuracy of algorithm in the UCI machine learning database. The Adult data set contains 48842 sample data. After deleting the samples with missing values, 45222 sample data are obtained, 30162 for training and 15060 for testing. Each sample contains 14 classification features and 1 classification result, including 6 continuous features and 8 discrete features. The classification attribute target level is divided into "≤50k" and "≥50k".

5.1Results Analysis

For the privacy budget analysis of this algorithm, this algorithm first allocates a given privacy budget B evenly to n iterations of data $\varepsilon' = B/n$. Since xgboost uses a gradient optimization model algorithm, the sample selection is not put back. Samples There is no overlap between the data and the sample data, so the privacy budget consumed by each tree does not need to be superimposed, that is, the privacy budget consumed by this iteration. The privacy budget allocation of each layer in the tree is $\varepsilon'' = \varepsilon'/(d+1)$. Because the nodes in each layer are counted and split on the non-crossing data set, the privacy protection budget allocated by each node is the privacy protection budget of this layer. Then the privacy protection budget allocated to each node is divided into two halves $\varepsilon = \varepsilon''/2$, and one half is used to estimate the number of samples in the training set, and the other half performs different operations depending on whether the node is an intermediate node or a leaf node. If the node is a leaf node, use the other half of the privacy protection budget combined with the Laplace mechanism to add noise to the count value to determine the type of the leaf node, if the node is an intermediate node, use the other half of the privacy protection budget combined with the index mechanism to select the best split point. Compared with the DiffPRFs algorithm [16], this algorithm does not need to consume an additional privacy budget when processing continuous attributes.

The feature importance ordering of the xgboost algorithm based on differential privacy is shown in Fig. 1. Due to too many features, this picture only captures the top 10 of them.

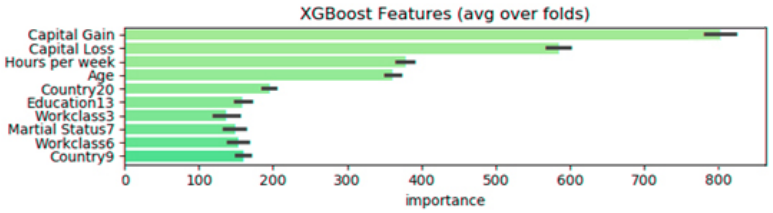

Figure 1: The importance of xgboost features

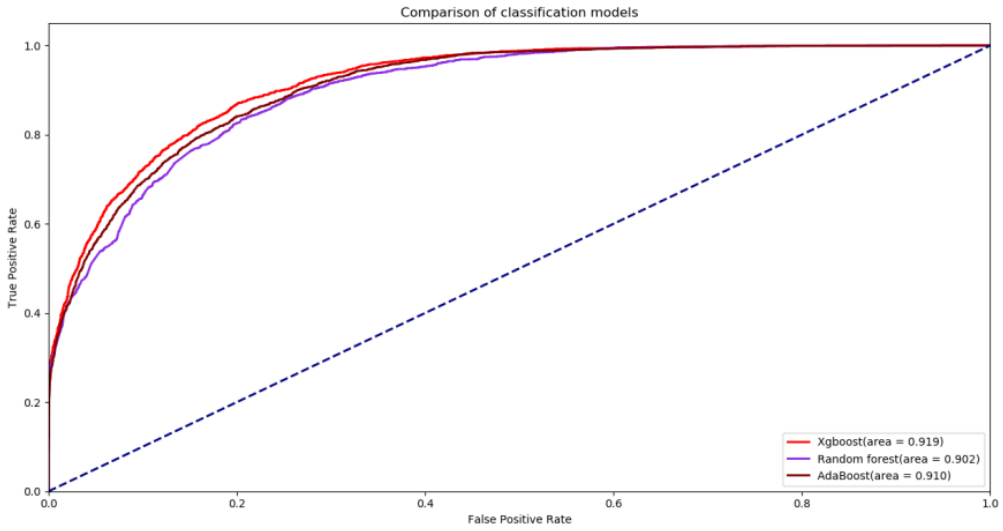

Figure 2: Auc curve of xgboost algorithm based on differential privacy

According to Fig. 3, the maximum depth of the tree affects the accuracy of classification in some extent, and with the reduction of the maximum depth of the tree, the accuracy is also decreasing, the relationship between the maximum depth and the accuracy of classification is linear. At the same time, you can see that as the privacy budget decreases, the accuracy of the classification decreases.

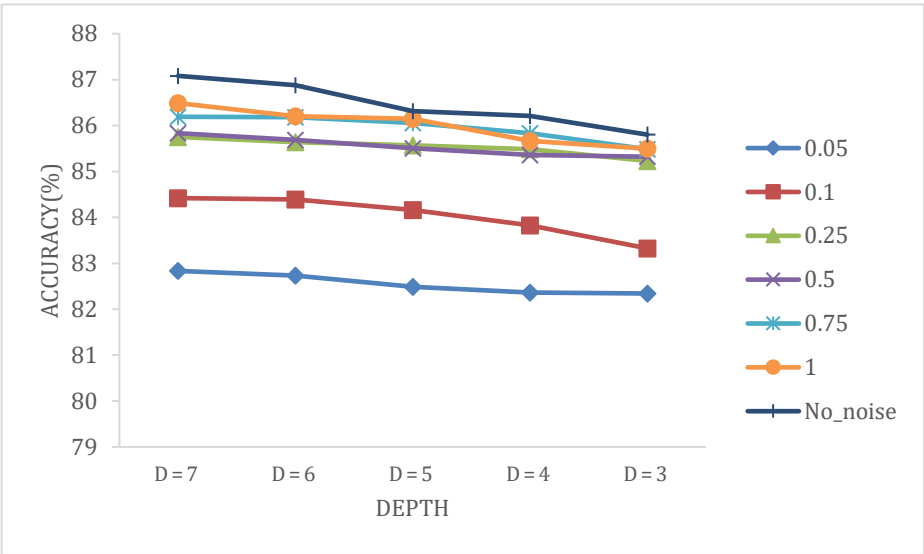

Figure 3: Classification accuracy under different conditions

Tab. 2 is a comparative experiment with Random Forest, Adaboost, Xgboost when the maximum depth of the tree is 7. It can be seen that the classification accuracy of using xgboost is higher in the same situation.

Table 2: Classification accuracy of different algorithms under the same conditions

ε	Accuracy (%)		
	Random forest	Adaboost	Xgboost
0.05	81.861	82.117	82.834
0.1	83.325	83.611	84.420
0.25	83.570	83.683	85.751
0.5	83.581	84.922	85.833
0.75	83.867	84.328	86.191
1	83.898	84.768	86.488
no_noise	86.109	85.321	87.082

6 Conclusion

In this paper, we propose the xgboost algorithm based on differential privacy. Experimental results show that the proposed method can improve the accuracy of continuous data classification. Through multiple training, the accuracy of classification can be improved. Under the same conditions, compared with the application of random forest in differential privacy classification, xgboost has better accuracy.

Funding Statement: This work is supported by the NSFC [Grant Nos. 61772281, 61703212, 61602254]; Jiangsu Province Natural Science Foundation [Grant No. BK2160968]; the Priority Academic Program Development of Jiangsu Higher Edu-cation Institutions (PAPD) and Jiangsu Collaborative Innovation Center on Atmospheric Environment and Equipment Technology (CICAEET).

Conflicts of Interest: We declare that they have no conflicts of interest to report regarding the present study.

References

[1] C. Dwork, "Differential privacy," *Lecture Notes in Computer Science*, vol. 26, no. 2, pp. 1–12, 2016.

[2] H. C. Li and X. P. Wu, "Intrusion correlation method for differential privacy protection network supporting alarm sequences," *Computer Engineering*, vol. 44, no. 5, pp. 128–132, 2018.

[3] X. J. Zhang and X. F. Meng, "Differential privacy protection for data publishing and analysis," *Chinese Journal of Computers*, vol. 37, no. 4, pp. 927–949, 2014.

[4] L. Sweeney, "K-anonymity: a model for protecting privacy," *International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 4, pp. 557–570, 2002.

[5] A. Blum, C. Dwork, F. Mcsherry and K. Nissim, "Practical privacy: the SuLQ framework," in *24th ACM SIGMOD Int. Conf. on Management of Data/Principles of Database Systems, Baltimore*, New York, USA: ACM, pp. 128–138, 2005.

[6] F. Mcsherry, "Privacy integrated queries: an extensible platform for Privacy-Preserving data analysis," *Communications of ACM*, vol. 53, no. 9, pp. 89–97, 2010.

[7] A. Friedman and A. Schuster, "Data mining with differential privacy," in *Proc.16th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, New York, USA: ACM, pp. 493–502, 2010.

[8] P. Xiong, T. Q. Zhu and X. F. Wang, "Differential privacy protection and its application," *Chinese Journal of Computers*, vol. 37, no. 1, pp. 101–122, 2014.

[9] N. Mohammed, R. Chen, B. Fung and P. S. Yu, "Differentially private data release for data mining," in *Proc.17th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, New York, USA: ACM, pp. 493–501, 2011.

[10] T. Q. Zhu, P. Xiong, Y. Xiang and W. L. Zhou, "An effective deferentially private data releasing algorithm for decision tree," in *12th IEEE Int. Conf. on Trust, Security and Privacy in Computing and Communications*, Washington DC, USA: IEEE Press, pp. 388–395, 2013.

[11] A. Patil and S. Singh, "Differential private random forest," in *ICACCI 2014*, Washington DC, USA: IEEE Press, pp. 2623–2630, 2014.

[12] C. Dwork, "A firm foundation for private data analysis," *Communications of ACM*, vol. 54, no. 1, pp. 86–95, 2011.

[13] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Foundations and Trends in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[14] C. Dwork, F. Mcsherry, K. Nissim and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography*, Springer Berlin Heidelberg, pp. 265–284, 2006.

[15] T. Chen, C. Guestrin, "Xgboost: a scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, New York, USA: ACM, pp.785–794, 2016.

[16] H. L. Mu, L. P. Ding, Y. N. Song and G. Q. Lu, "DiffPRFs: a differential privacy protection algorithm for random forest," *Journal of Communications*, vol. 37, no. 9, pp. 175–182, 2016.