

Article

A Differential Privacy Budget Allocation Algorithm Based on Out-of-Bag Estimation in Random Forest

Xin Li ¹, Baodong Qin ^{1,*} , Yiyuan Luo ² and Dong Zheng ^{1,3}¹ School of Cyberspace Security, Xi'an University of Posts and Telecommunications, Xi'an 710121, China² School of Computer Science and Engineering, Huizhou University, Huizhou 516007, China³ School of Computer Science, Qinghai Normal University, Xining 810008, China

* Correspondence: qinbaodong@xupt.edu.cn

Abstract: The issue of how to improve the usability of data publishing under differential privacy has become one of the top questions in the field of machine learning privacy protection, and the key to solving this problem is to allocate a reasonable privacy protection budget. To solve this problem, we design a privacy budget allocation algorithm based on out-of-bag estimation in random forest. The algorithm firstly calculates the decision tree weights and feature weights by the out-of-bag data under differential privacy protection. Secondly, statistical methods are introduced to classify features into best feature set, pruned feature set, and removable feature set. Then, pruning is performed using the pruned feature set to avoid decision trees over-fitting when constructing an ϵ -differential privacy random forest. Finally, the privacy budget is allocated proportionally based on the decision tree weights and feature weights in the random forest. We conducted experimental comparisons with real data sets from Adult and Mushroom to demonstrate that this algorithm not only protects data security and privacy, but also improves model classification accuracy and data availability.

Keywords: differential privacy; machine learning; privacy protection; random forest; out-of-bag estimation

MSC: 68P27



Citation: Li, X.; Qin, B.; Luo, Y.; Zheng, D. A Differential Privacy Budget Allocation Algorithm Based on Out-of-Bag Estimation in Random Forest. *Mathematics* **2022**, *10*, 4338. <https://doi.org/10.3390/math10224338>

Academic Editor: Daniel-Ioan Curia

Received: 10 October 2022

Accepted: 15 November 2022

Published: 18 November 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

With the rapid progress of various emerging technologies including the Internet, cloud computing, and computer storage, the era of big data has arrived [1]. Currently, numerous personal privacy data are currently collected by medical, educational, and corporate organizations, such as patient medical records collected by hospital organizations, student learning records collected by educational institutions, and customer location information collected by taxi-hailing platforms [2]. After being collected, these personal privacy data are often used for data analysis, data mining, etc., which will certainly affect the normal life of individuals. For example, the Facebook privacy leakage in 2018 caused significant damage to users. The frequent occurrence of numerous privacy leakage events has made data privacy protection a hot topic for research in the field of information security [3].

At present, data privacy protection methods [4] are mainly based on anonymity technology, encryption technology or noise technology. Privacy protection based on anonymity technology, such as k -anonymous algorithms, relies on background knowledge, and privacy protection based on encryption technology, e.g., homomorphic encryption, has a large computational overhead, making it unsuitable for using in massive data environments. In 2006, Dwork et al. [5] presented the noise-based privacy protection technique, i.e., differential privacy, which has been widely applicable to machine learning privacy preservation because of its low computational and transmission costs.

Classification is a very important method in the field of data mining [6]. It uses a large amount of data to build algorithmic models, and uses these models to perform classification operations. There are many algorithms that can be used for classification.

Compared with classification algorithms, such as neural networks, Bayesian and genetic algorithms, decision trees have less algorithmic complexity, are resistant to noise and have strong data scalability [7]. However, traditional single classifier models such as decision trees are single and prone to problems such as overfitting. In order to improve the accuracy of classification prediction, some scholars have proposed integrated methods. Random forest is an algorithm based on decision tree integration, which not only has many advantages of decision tree classification, but also has good tolerance for unbalanced samples, noise and outliers. It has been widely used in many fields such as banking [8], medical [9], e-commerce [10], and finance [11].

In the actual classification process, decision trees, random forest models and their corresponding counting information may leak users' private information, and there is a danger of privacy leakage. Applying differential privacy techniques to random forest models to protect private data is of great importance for data security publication. Current research on this area focuses on the selection and innovation of tree building methods [12,13], the selection of availability functions [14], and methods for pre-processing attribute sets [15–17]. Data protection algorithms based on differential privacy decision trees are mainly classified into interactive framework and non-interactive framework.

In the interactive framework, users can only conduct a limited number of queries through the privacy protection interface, and each query consumes the privacy protection budget. Blum et al. [18] first fused differential privacy with decision trees to obtain the SuLQ-based ID3 algorithm, but the prediction accuracy of the decision tree model was substantially reduced because it added differential privacy noise every time the information gain was calculated. McSherry et al. [19] used the PINQ framework to improve the SuLQ algorithm to obtain the PINQ-based ID3 algorithm. The algorithm partitioned the dataset into multiple disjoint subsets. The disadvantage is that each query consumes the privacy protection budget, many queries resulting in little privacy protection budget allocated for each query, thus adding more noise when dealing with large data sets. Friedman et al. [20] in a further study designed the DiffP-ID3 algorithm, which combines ID3 algorithm with an exponential mechanism to achieve differential privacy protection and effective noise reduction. In the same paper, they proposed the DiffP-C4.5 algorithm by using exponential mechanism to split continuous attributes. However, this method required to invoke the exponential mechanism twice, consuming a disproportionate amount of the privacy protection budget.

In a non-interactive framework, privacy protection algorithms are processed and published to the database, users can process this database for any operation. In this framework, we want to improve the availability of the data publishing by allocating a reasonable privacy protection budget to reduce the overall amount of noise added to the model. Mohammed et al. [21] and Zhu et al. [22] presented the data publishing algorithms DiffGen and DT-Diff under non-interactive, respectively. Both algorithms first generalized the data sets, then performed a subdivision iteration loop, and finally used an exponential mechanism for selection. Generalization replaces calling the exponential mechanism when processing continuous features and saves the privacy protection budget, but such schemes are inefficient when the classification dimension is large. In further research, it was found that decision trees are not stable enough due to their simple structure, and the prediction results are often unsatisfactory when facing high-dimensional data sets, and random forest models were considered to replace single decision trees. Patil et al. [23] combined differential privacy with random forests to build ID3 decision trees for classification and proposed the DiffPRF algorithm, which has the disadvantage that continuous attributes need to be discretized first. Mu et al. [14] improved the DiffPRF algorithm by introducing an exponential mechanism to deal with continuous attributes and proposed the DiffPRFs algorithm. Li et al. [24] proposed the RFDPP-Gini algorithm. Exponential and Laplace mechanisms are used in the selection of split features to deal with continuous and discrete features, respectively, and the Gini index is selected to determine the best splitting feature and the best splitting point using the equivariant privacy budget allocation method.

From the above related studies, it is clear that the current innovations and improvements in this field focus on how to improve the availability of data and the accuracy of model classification. The literature [25] uses out-of-bag estimation to evaluate the classification ability of random forests. Out-of-bag data [26] is an unused asset in random forests, which reflects the classification ability, feature importance, and other data set patterns of random forests. The literature [27] adds differential privacy to the out-of-bag estimation to protect the privacy of the out-of-bag data. In this paper, considering that the importance between trees and features in random forest is not the same, in order to allocate the privacy protection budget in a more targeted way, we need to calculate the solution in advance to obtain the weight sets of tree and feature importance. Choosing out-of-bag data to solve the weight set not only can be used for pruning and preventing overfitting, but also when differential privacy protection is added to it, it does not waste the privacy protection budget and overall reduces the total amount of noise added to the dataset, thus improving the availability of data and the accuracy of model classification. This study provides a more accurate solution for privacy protection in the fields of medical diagnosis, financial decision making, personalized recommendation, and bioinformatics. The main contributions of this paper are summarized as below:

1. We propose a differential privacy budget allocation algorithm based on out-of-bag estimation in random forest.
2. We improve the algorithm for differential privacy out-of-bag estimation to obtain more accurate decision tree weights in out-of-bag forests. We introduce decision tree weights when using the VIM variable importance measure to obtain a more accurate set of feature weights and use statistical methods for classification.
3. We creatively give computational methods to allocate the overall privacy protection budget to each tree in the random forest and to each layer of each tree to achieve a more targeted privacy budget allocation.
4. We conduct a series of experiments on Adult and Mushroom datasets to demonstrate the advantages of the algorithm in this paper.

The remainder of this paper is organized as follows. In Section 2, we introduce differential privacy background knowledge. The proposed method is described in detail in Section 3. The experimental results and analysis are given in Section 4. Finally, we conclude this paper in Section 5.

2. Differential Privacy Background Knowledge

Differential privacy solves the problem of database privacy leakage. With its strict mathematical definition and flexible combination of properties, it is used in all kinds of privacy protection.

Definition 1 (Differential privacy [28]). *For any two neighboring data sets D_1 and D_2 , all differences are at most one record. Given a privacy protection algorithm F , $\text{Range}(F)$ denotes the set of all possible output ranges of F . If the algorithm F satisfies:*

$$\Pr[F(D_1) \in S] \leq e^\epsilon \Pr[F(D_2) \in S] \quad (1)$$

then it is said that algorithm F provides ϵ -differential privacy protection, where $\Pr[E_S]$ denotes the probability of event E_S occurring and ϵ is the privacy protection budget. The value of parameter ϵ should be consistent with the algorithm requirements, so as to obtain a perfect balance between data security and availability.

Definition 2 (Global sensitivity [28]). *Sensitivity is a parameter that measures the magnitude of the joining noise. For any of the functions $Q : D \rightarrow \mathbb{R}^d$, the global sensitivity of Q is:*

$$\Delta GS = \max_{D_1, D_2} \|Q(D_1) - Q(D_2)\| \quad (2)$$

where R denotes the real number space of the mapping and d denotes the query dimension of the function f .

Definition 3 (Realization mechanism). The most common implementations of adding differential privacy to the data are the Laplace mechanism and the exponential mechanism.

1. Laplace mechanism [29–31]. This mechanism is implemented by adding noise satisfying the Laplace distribution to the output result. Given any function $f : D \rightarrow R^d$, if the $F(D)$ meets Formula (3), it means that it satisfies ϵ -differential privacy.

$$F(D) = f(D) + (\text{Laplace}(\frac{\Delta GS}{\epsilon}))^d \quad (3)$$

where $\text{Laplace}(\frac{\Delta GS}{\epsilon})$, being the Laplace distribution with scale parameter $\frac{\Delta GS}{\epsilon}$.

2. Exponential mechanism [30,31]. Let the input of the randomized algorithm M be a dataset D and the output be an entity object $r \in \text{Range}$, $q(D, r)$ is the availability function, ΔGS is the sensitivity of the function $q(D, r)$. If algorithm M selects and outputs r from Range with probability proportional to $e^{q(D, r)/2\Delta GS}$, then algorithm M provides ϵ -differential privacy.

Definition 4 (Combination properties). In practical scenarios, users may make multiple queries, but the privacy protection budget needs to be kept within a given range. This problem can be cleverly solved by using the differential privacy combination properties.

1. Sequential composition [32]. Assuming that in a set of mechanisms A_1, A_2, \dots, A_n provide $(\epsilon_1, \epsilon_2, \dots, \epsilon_n)$ -differential privacy protection respectively, for a same data set D , algorithms $A(A_1(D), A_2(D), \dots, A_n(D))$ has $(\sum_{i=1}^n \epsilon_i)$ -differential privacy.
2. Parallel composition [32]. Assuming that in a set of mechanisms A_1, A_2, \dots, A_n provide $(\epsilon_1, \epsilon_2, \dots, \epsilon_n)$ -differential privacy protection respectively, for the disjoint dataset D , algorithms constitute the combination $A(A_1(D), A_2(D), \dots, A_n(D))$ with $(\max \epsilon_i)$ -differential privacy.

3. Proposed Method

3.1. Solving Decision Tree Weights

In a random forest, each decision tree is generated by randomly selecting samples. When the total number of samples selected is very large, about 1/3 were not selected. These data are out-of-bag data [26]. When these data are applied in decision making on this decision tree, the ratio of forward misclassification to reverse misclassification in the sum of the data of the respective instances is the out-of-bag estimate. Since out-of-bag estimates are closely related to feature importance, forest properties, etc., it is important to incorporate differential privacy to protect out-of-bag data.

Definition 5 (Differential privacy out-of-bag estimation). For one of the trees, the out-of-bag estimate B is:

$$B = \frac{1}{2} \left(\frac{Y}{Y_T} + \frac{N}{N_T} \right)$$

where Y and N represent the number of forward misclassifications and reverse misclassifications. Y_T and N_T indicate the total number of forward and reverse instances.

Differential privacy out-of-bag estimation is defined as:

$$\begin{aligned} B' &= \frac{1}{2} \left(\frac{Y + N(\epsilon_1)}{Y_T} + \frac{N + N(\epsilon_2)}{N_T} \right) \\ &= \frac{1}{2} \left(\frac{(Y \cdot N_T + N \cdot Y_T) + N(\epsilon)}{Y_T \cdot N_T} \right) \end{aligned} \quad (4)$$

It can be seen from Formula (4) that the noise that we add to the out-of-bag estimated data, which perturbs the true number of data, so that the differential privacy out-of-bag

estimation does not lose the dataset regularity and protects the privacy of the data at the same time.

Definition 6 (Decision tree weights). *This paper calculates the decision tree weights with reference to Paul et al. [25]. The weights of the decision tree are defined as:*

$$Q_t = \frac{1}{B'_t} \quad (5)$$

where B'_t is the differentially private out-of-bag estimate of the tree. From Formula (5), we can see that the smaller B'_t is the larger the weight of the tree is, and the better the decision tree classification ability is.

3.2. Feature Weight Calculation and Feature Selection

3.2.1. Feature Weight Calculation

The Gini index is an important method for classifying the purity of attributes. The smaller the Gini index, the better the classification method. The formula for calculating the Gini index G_m is:

$$G_m = \sum_{c=1}^C p_{mc}(1 - p_{mc}) \quad (6)$$

where C is the number of features, p_{mc} is the probability of class c at node m .

The importance of node m in feature j_i is:

$$I_{jm} = G_m - w_L G_L - w_R G_R \quad (7)$$

where G_L and G_R are the Gini indices of the left and right nodes after splitting at node m , w_L and w_R are the number of weighted samples.

If the feature j_i is selected n times in the tree T_i , the importance of the feature in this decision tree is:

$$I_{ij} = \sum_{n=1}^n I_{jm} \quad (8)$$

The importance I_j of the feature j_i in the random forest is:

$$I_j = \sum_{i=1}^t I_{ij} \cdot Q_t / \sum_{j=1}^k \sum_{i=1}^t I_{ij} \quad (9)$$

where k is the number of input features, and t is the number of decision trees.

If C features are selected to construct t decision trees, the weight W_t of the decision trees in the random forest are:

$$W_t = \sum_{n=1}^N I_j / \sum_{t=1}^t \sum_{n=1}^N I_j \quad (10)$$

3.2.2. Feature Selection

Statistical methods have good results in data pre-processing. In the article, we choose this method to divide the features into Best feature set (BFS), Pruned feature set (PFS) and Removable feature set (RFS).

Definition 7 (Feature selection). *For the set of feature weights, the following conditions are satisfied:*

$$I_j < (\mu - 3\sigma) \quad (11)$$

$$(\mu - 3\sigma) < I_j < (\mu - 1.5\sigma) \quad (12)$$

where μ is the mean of the weight set of R , and σ is the standard deviation. If the weight VIM_j satisfies the Formula (11), the features j are placed into the removable feature set (RFS). If the weight VIM_j satisfies the Formula (12), the features j are placed into the pruned feature set (PFS). The remaining features j are put into the best feature set (BFS). The features in the existing feature weight set are deleted from the removable feature set (RFS) to obtain selected weight set R' with n' features.

3.3. Method for Allocating Privacy Protection Budgets Based on Weights

The reasonable allocation of privacy protection budget has a very important impact on data availability. Previous related works such as MAXGDDP algorithm [16] and AUR-Tree algorithm [17] use class geometry, class equivariance, and equivariance to allocate privacy protection budget during the construction of decision trees. In random forests, such as DiffPRFs algorithm [14] and RFDPP-Gini algorithm [24], it allocates the privacy protection budget equally to each decision tree and then equally to each layer. However, each tree in the random forest has different strengths and weaknesses in classification ability, and each feature in the dataset has different importance in the random forest. The privacy protection budget allocation method designed in this paper is based on adaptive allocation of tree weights and feature weights.

(1) Allocation based on tree weights. The privacy protection budget allocated to each tree ε_t is:

$$\varepsilon_t = \frac{\varepsilon}{T} \cdot W_t \quad (13)$$

where ε is the sum of the privacy protection budget, T is the number of decision trees in the random forest, and W_t is the weight of the decision tree.

(2) Allocation based on tree weights. The privacy protection budget ε_t allocated to each tree is distributed proportionally according to the relative size of the feature weights. After feature selecting to obtain the feature weight set R' with n' , follow the random forest principle to randomly select a features ($a < n'$) to get feature set of the tree t and the corresponding feature weight set $R'_t \{I_{j_1}, I_{j_2}, \dots, I_{j_{n'}}\}$. Calculate the weight ratio S_{j_1} of the features j_1 in this tree:

$$S_{j_1} = \frac{I_{j_1}}{I_{j_1} + I_{j_2} + \dots + I_{j_{n'}}} \quad (14)$$

The privacy protection budget obtained from the allocation of the decision tree in selecting this feature for splitting is:

$$\varepsilon_{j_1} = \varepsilon_t \cdot S_{j_1} \quad (15)$$

Figure 1 provides an overview of the privacy protection budget allocation.

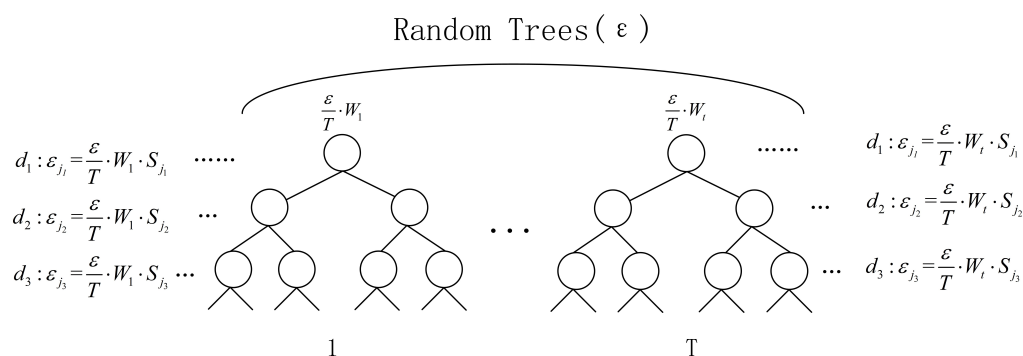


Figure 1. Overview of privacy protection budget allocation.

3.4. Pruning Based on the Divided Set of Attributes

Pruning can simplify the decision tree model in random forests, avoid overfitting, enhance generalization, and improve the accuracy of classification. In this paper, a pre-pruning strategy is used to prune using the pruning feature set (PFS) obtained in Section 3.2.2. In the process of building a tree, the features in order of importance are used as the best splitting features. For the feature set R'_t selected by the tree t , determine whether R'_t has features in the pruned feature set (PFS), and if so, remove them and allocate privacy budget. if not, allocate them according to the original feature set and construct a differential privacy random forest.

3.5. Algorithm Flow

This algorithm process in this paper is to first construct a random forest and extract the out-of-bag dataset, followed by calculating the decision tree weights and feature weights using out-of-bag estimation under differential privacy protection, and the feature weights use statistical methods to classify the features into the best feature set, pruned feature set and removable feature set. When constructing a random forest that satisfies ϵ -differential privacy protection, the randomly selected features are judged using the pruned feature set and the features belonging to the pruned feature set are removed, this method not only saves the privacy protection budget but also prevents the decision tree from overfitting. Then calculate the weights of the trees based on these features, and allocate privacy protection budget to each tree based on the tree weights. In the construction of each tree, the privacy protection budget is allocated and splits are selected sequentially according to the relative size of the feature weights in the feature set. In the construction of each tree, the privacy protection budget is allocated according to the relative size of the feature weights in the feature set and splits are selected sequentially. If continuous features are selected as splitting attributes in the tree building process, the exponential mechanism is invoked to split continuous attributes, and the Laplace mechanism is used to add noise to the count values of the leaf nodes to finally obtain a random forest that satisfies ϵ -differential privacy protection.

3.6. Description

We present a privacy budget allocation algorithm based on tree weights and feature weights in random forest, and the core strategy of the method is to allocate more privacy protection budgets to trees with better classification ability and features with higher importance in the forest.

This paper is divided into three algorithms, Algorithm 1 is the algorithm of extracting out-of-bag data.

Algorithm 1 Extract Out-of-Bag Data Algorithm

Input: Training set D , feature set $J(j_1, j_2, \dots, j_n)$, number of decision trees t , maximum depth of the decision trees d , number of randomly selected features at splitting a .

Output: Random forest *Trees*, Out-of-bag datasets D_{OOB} .

Stopping condition: All samples on a decision tree node are consistently classified, or the maximum depth of the decision tree is reached d , or the number of features set features less than a .

Step 1. Take out m samples from the training set D with put-back as D_t , and randomly select n ($n < N$) features from J .

Step 2. Calculate the Gini coefficient of each feature, and select the best splitting feature j_n as the splitting feature.

Step 3. According to the different values of feature j_n , the samples on the node are divided into different child nodes, and a decision tree is generated.

Step 4. Repeat steps 1–3 t times to get t decision trees.

Step 5. Put the data in the training set D that was not used to build the decision tree into D_{OOB} .

Step 6. Get random forest *Trees*, out-of-bag datasets D_{OOB} .

Algorithm 2 is a feature weight solving and selection algorithm using out-of-bag estimation. The out-of-bag estimation under differential privacy is used to solve feature weights and decision tree weights, select feature weights, and update the features set.

Algorithm 2 Solving Feature Weights and Selecting Algorithms Using Out-of-Bag Estimation

Input: Out-of-bag datasets D_{OOB} , random forest *Trees*, privacy protection budget ϵ_{OOB} , feature set $J(j_1, j_2, \dots, j_n)$.

Output: Selected feature weight set R' , selected feature set J' , pruned feature set (PFS), best feature set (BFS).

Stopping condition: All features are calculated and screened.

Step 1. For decision tree t , select data containing tree features from out-of-bag data set D_{OOB} , test and count.

Step 2. Add Laplacian noise to all categorical counts for each tree in a random forests, the size of the privacy protection budget for each tree is $\frac{\epsilon_{OOB}}{t}$.

Step 3. Get the differential privacy out-of-package estimation B' using the Formula (4).

Step 4. Calculate the decision tree weight Q_t using the Formula (5).

Step 5. Repeat steps 1–4 to calculate the weight of each tree in the random forests in turn, and get the weight set $Q(Q_1, Q_2, \dots, Q_t)$ of the tree.

Step 6. For each tree, use Formulas (6)–(8) to get the importance of the feature in this tree.

Step 7. Calculate the feature weight of this feature in the whole random forest using Formula (9).

Step 8. Repeat steps 6–7 to obtain the feature weight set R of all features.

Step 9. If feature weight set R satisfies Formula (11), put corresponding feature j_n into the removable feature set RFS; if it satisfies Formula (12), put corresponding feature j_n into the pruned feature set (PFS); the remaining features are best feature set (BFS).

Step 10. Delete the features and feature weights in the removable feature set RFS in turn from the initial feature set $J(j_1, j_2, \dots, j_n)$ and feature weight set R .

Step 11. Get the selected feature weight set R' , selected feature set J' , pruned feature set (PFS), best feature set (BFS).

Algorithm 3 is a differential privacy budget allocation algorithm based on tree weights and feature weights. The tree weight calculation, pruning and differential privacy budget allocation by randomly selected features in the construction makes the strategy of assigning privacy protection budgets more targeted and reasonable.

Algorithm 3 Differential Privacy Budget Algorithm Based on Tree Weight and Feature Weight

Input: Training set D , selected feature set J' , selected feature weight set R' , number of decision trees T , maximum depth of the decision trees d , number of features b randomly selected when splitting, privacy protection budget ϵ , pruned feature set (PFS), best feature set (BFS).

Output: Random forests satisfying ϵ -differential privacy protection.

Stopping condition: All samples on a decision tree node are consistently classified, or the maximum depth of the decision tree is reached d , or the feature number n' of the filtered feature set is less than b , or the privacy protection budget is exhausted.

Step 1. For decision tree t , b features are randomly selected, the feature set is J'_t , and the corresponding feature weight set calculated by Formula (10) is R'_t .

Step 2. If the feature set J'_t has features belonging to the pruned feature set (PFS), the corresponding feature weight set R'_t removes these feature weights to obtain the feature set R''_t . If the feature set J'_t belongs to the best feature set (BFS), the corresponding feature weight set R'_t does not change.

Step 3. For the feature set J'_t , the weight W_t of the decision tree is obtained according to Formula (13).

Step 4. The privacy budget of the tree is obtained from the weights of the tree as $\epsilon_t = \frac{\epsilon}{T} \cdot W_t$.

Step 5. In the feature weight set R'_t , the feature weight ratio S_{j_1} of the feature j_b is calculated according to Formula (14).

Step 6. In the tree t , according to the size of the feature weight set R'_t , the features are sequentially selected from the corresponding feature set J'_t as the best splitting feature.

Step 7. The privacy protection budget allocated by the optimal splitting feature is $\epsilon_{j_1} = \epsilon_t \cdot S_{j_1}$.

Step 8. Repeat T times to calculate the privacy protection budget required for each feature split on each tree.

Step 9. When constructing the tree to select features, if the best split feature is a continuous feature, an exponential mechanism is invoked to select the best point for splitting:

$$\frac{\exp\left(\frac{\epsilon}{2\Delta q}q(D_n, j)\right)|R_i|}{\sum_i \exp\left(\frac{\epsilon}{2\Delta q}q(D_n, j)\right)|R_i|}$$

among them, $q(D_n, j)$ is the Gini index, $|R_i|$ is the size of the interval, and Δq is the sensitivity of the Gini index.

Step 10. If stopping condition is met, the creation of node is stopped, and the node is set as a leaf node.

Step 11. Noise is added to the count value of the leaf nodes of each tree, and the classification with the most samples is selected as the label of the leaf node.

4. Experimental Results and Analysis

4.1. Privacy Analysis

(1) Privacy when computing out-of-bag data. When using the out-of-bag data for each tree calculation, the allocated privacy budget is used to add noise to the counts of the classification results, and thus differential privacy is satisfied.

(2) Privacy of random forests for training sets. There are T trees in this algorithm, and the privacy budget allocated to each tree is $\epsilon_t = \frac{\epsilon}{T} \cdot W_t$. The privacy budget divided into each layer according to feature importance is $\epsilon_{j_1} = \epsilon_t \cdot S_{j_1}$, for each level of the decision tree, the different nodes are equally divided into $\epsilon' = \epsilon_{j_1} / (d + 1)$, and the noise is eventually added using Laplace mechanism. Since the samples in each tree in a random forest are randomly selected with a put-back, the data will have crossover. From the sequential composition of differential privacy, it is clear that the consumed privacy budget is the sum of the individual tree consumption, so the training set satisfies ϵ -differential privacy in the process of constructing the random forest.

Since the two parts of the out-of-bag data and the training set are disjoint and both satisfy ϵ -differential privacy, the parallel composition of differential privacy shows that the entire process of constructing a differentially private random forest satisfies ϵ -differential privacy.

4.2. Experimental Design

The hardware environment for the experiments in this paper is Intel(R) Core (TM) i5-5200U CPU @2.20GHz processor and 8GB operating memory. The operating system is Windows 10, the experimental program development tool is Pycharm2021.3 and the programming language is Python.

Two real datasets originating from UCL were used for the experiment: the Adult and Mushroom dataset (Table 1). The Adult dataset contains U.S. Census data with discrete and continuous attributes that determine whether the category is a wage greater than 50k. The Mushroom dataset contains information about mushroom-related species, and the only discrete attributes in this dataset that determine the category is whether a mushroom is edible or not.

Table 1. Dataset Information.

Dataset	Characteristic Number (Discrete/Continuous)	Size	Class Attribute
Adult	14 (8/6)	32,561	1
Mushroom	22 (22/0)	8124	1

To test the effectiveness of Ours' algorithm, multiple sets of comparison experiments are set up in this paper: (1) comparison between different decision tree depths; (2) comparison between different number of decision trees; (3) comparison between different size of privacy budgets; (4) comparison between Ours algorithm, RFDPP-gini [24] algorithm and DiffPRFs [14] algorithms in this paper.

4.3. Experimental Results

For the Adult and Mushroom datasets, a random forest satisfying differential privacy protection is built using this paper's algorithm with different privacy protection budgets and different decision tree depths, and the test datasets are classified to obtain the accuracy of the classification results.

Figure 2 shows the classification accuracy at different tree depths for the Adult dataset with differential privacy noise added and privacy budgets of 0.10, 0.25, 0.50, 0.75, and 1.00, respectively. Figure 3 shows the classification accuracy at different tree depths for the Mushroom dataset with differential privacy noise added and privacy protection budgets of 0.10, 0.25, 0.50, 0.75, and 1.00, respectively.

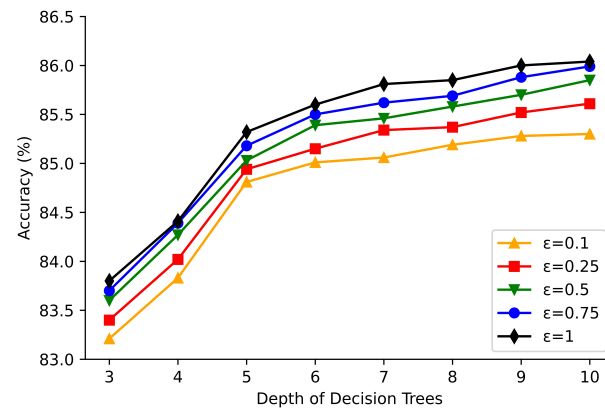


Figure 2. Variation of classification accuracy with tree depth for Adult dataset.

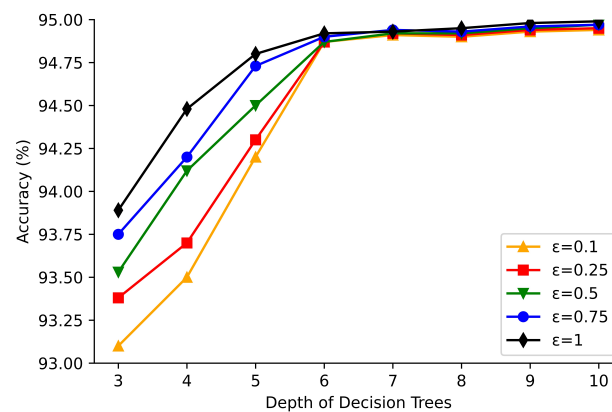


Figure 3. Variation of classification accuracy with tree depth for Mushroom dataset.

Figure 4 shows the classification accuracy for the number of decision trees in the random forest at 10, 30, 50, 70, and 90 for the Adult dataset with differential privacy noise added and privacy budgets of 0.10, 0.25, 0.50, 0.75, and 1.00, respectively. Figure 5 shows the classification accuracy of the Mushroom dataset when the number of decision trees in the random forest is 10, 30, 50, 70, and 90 when differential privacy noise is added and the privacy budget is 0.10, 0.25, 0.50, 0.75, and 1.00, respectively.

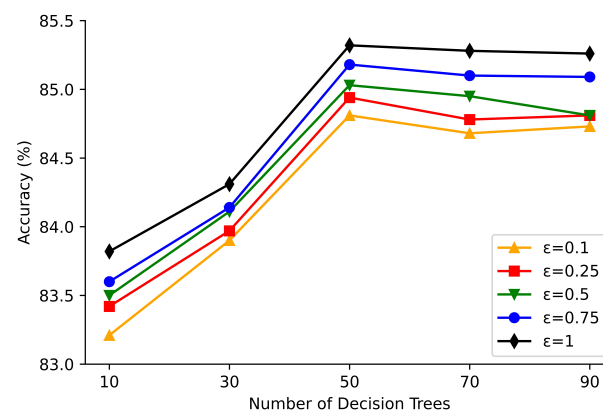


Figure 4. Variation of classification accuracy with the number of decision trees for the Adult dataset.

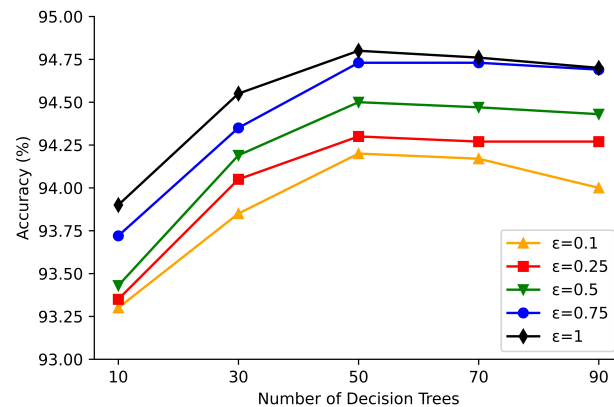


Figure 5. Variation of classification accuracy with the number of decision trees for Mushroom dataset.

To test the performance of the algorithm, the classification accuracy of Ours algorithm is compared with RFDPP-gini and DiffP-RFs algorithms on Adult and Mushroom datasets under the same conditions. Set $T = 50$, privacy budget to 0.10, 0.25, 0.5, 0.75, 1.00, the depth of the decision tree to 5, and the number of randomly selected features at node splits to 5. The experimental results are shown in Figures 6 and 7.

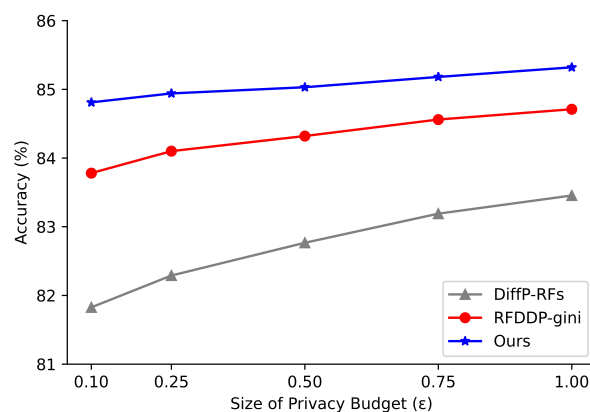


Figure 6. Comparison of classification performance of the three algorithms on Adult dataset.

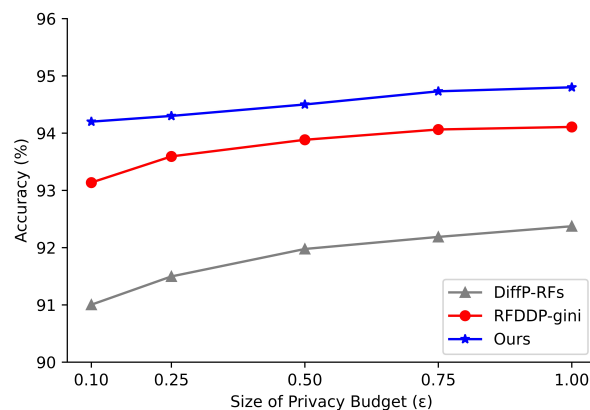


Figure 7. Comparison of classification performance of the three algorithms on Mushroom dataset.

4.4. Comparisons

The privacy budget is a measure of how much noise is added and a measure of the strength of data privacy. As the privacy budget increases, the weaker the privacy of the data, the stronger the classification accuracy of the algorithm. In the decision tree model, as the depth of tree increases, the decision tree becomes more branchy, dividing the dataset to a finer degree, and the division becomes more accurate. The performance of the model is affected by the privacy budget and the depth of the decision tree.

From Figures 2 and 3, it can be seen that when the depth of the tree is 3 and the privacy budget is 0.1, 0.25, 0.5, 0.75, and 1.00, the classification accuracy of the algorithm in this paper 83.21%, 83.40%, 83.60%, 83.70%, and 83.81% for the Adult dataset, respectively, and the accuracy for the Mushroom dataset is 93.1%, 93.38%, 93.53%, 93.75%, and 93.89%, and the accuracy all increase with increasing privacy budget. Similarly, the accuracy of algorithm for two datasets is gradually improved as the depth of the tree increases. From the above experimental results, it is concluded that the higher the depth of the tree, the higher the accuracy when the privacy budget is the same, and the larger the privacy budget, the higher the accuracy when the tree depth is the same.

From Figures 4 and 5, it can be seen that the classification accuracy of the algorithm in this paper for the Adult dataset is 83.50%, 84.11%, 85.03%, 84.95% and 84.81% when the number of decision trees is 50 and the privacy protection budget is 0.1, 0.25, 0.5, 0.75, and 1.00, respectively. The classification accuracies for Mushroom dataset were 93.43%, 94.19%, 94.50%, 94.47% and 94.43%, respectively. And the highest accuracy of classification is achieved when the number of trees is 50. This may be because when the number of trees in the random forest is too small, the generalization ability of the random forest model is poor and the model classification accuracy increases with the number of decision trees. When the number of trees is more than 70, as the privacy budget needs to be allocated to each tree, the more the number of trees the less privacy protection budget on each tree, and at this time it is the privacy budget that constrains the classification accuracy of model. Therefore, the selection of the number of trees to be built in the differential privacy random forest needs to be considered to satisfy the generalization ability of the model while building a small number of trees to improve the overall model classification accuracy.

From Figures 6 and 7 at privacy budgets of 0.1, 0.25, 0.5, 0.75, and 1.00, the classification accuracies of the algorithm in this paper are 84.81%, 84.94%, 85.03%, 85.18% and 85.32% for the Adult dataset and 94.20%, 94.30%, 94.50%, 94.73%, and 94.80% for the Mushroom dataset, respectively. It can be seen from the figure that the classification accuracy of ours, RFDPP-gini and DiffP-RFs algorithms for both Adult and Mushroom datasets increase with increasing privacy protection budget, which is satisfying our expectation. Our algorithm has better classification accuracy than RFDPP-gini and DiffP-RFs algorithms with the same privacy protection budget, which is due to the fact that ours algorithm calculates feature weights and tree weights by differential privacy out-of-bag estimation, which saves privacy budget, while performing feature selecting, etc. to make more important features prominent, and finally allocates privacy budget by feature weights and tree weights, which reduces the effect of noise on important decision trees and important features. In summary, the algorithm presented in this article makes the data more available while satisfying privacy protection.

5. Conclusions

In this paper, we propose a differential privacy budget allocation algorithm based on out-of-bag estimation of random forests, which calculates the weights and feature weights of trees under different datasets to allocate privacy protection budgets by out-of-bag estimation, selects a dynamic balance between the generalization ability of random forests and privacy protection budgets, avoids building decision trees with similar classification performance in random forests, and avoids too many decision trees making the privacy budget on each tree too small. At the same time, this paper improves the method of selecting features for each tree in the random forest, randomly selecting a certain number of

features, ranking them according to their importance and selecting splits in turn, and adding differential privacy according to the principle that the higher the importance of features, the more privacy protection budget is allocated, which makes the privacy protection budget have a higher utilization rate. However, in the course of the algorithm, the weights of each tree are found by differential privacy out-of-bag estimation, after which the weights of each tree are found in the construction of a differential privacy random forest. Follow-up consideration is given to how weights under different forests can be linked to simplify the algorithmic process, and further research is conducted to learn how to efficiently allocate differential privacy protection budgets in other models.

Author Contributions: Conceptualization, X.L., B.Q. and D.Z.; methodology, X.L., B.Q. and Y.L.; software, X.L.; validation, B.Q. and Y.L.; writing—original draft preparation, X.L. and B.Q.; writing—review and editing, B.Q. and Y.L.; project administration, B.Q. and D.Z.; funding acquisition, B.Q., Y.L. and D.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Basic Research Program of Qinghai Province (grant number 2020-ZJ-701) and by the National Natural Science Foundation of China (grant numbers 61872292 and 62072207).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Marina, S.; Stan, M. *Challenges in Computational Statistics and Data Mining*; Springer: Cham, Switzerland, 2016; pp. 365–380.
2. Hu, X.Y.; Yuan, M.X.; Yao, J.G.; Deng, Y.; Chen, L.; Yang, Q.; Guan, H.B.; Zeng, J. Differential privacy in telco big data platform. *Proc. VLDB Endow.* **2015**, *24*, 1692–1703. [\[CrossRef\]](#)
3. Abid, M.; Iynkaran, N.; Yong, X.; Guang, H.; Song, G. Protection of big data privacy. *IEEE Access* **2016**, *4*, 1821–1834.
4. Qi, X.J.; Zong, M.K. An overview of privacy preserving data mining. *Procedia Environ. Sci.* **2012**, *12*, 1341–1347. [\[CrossRef\]](#)
5. Dwork, C. Differential privacy. *Lect. Notes Comput. Sci.* **2006**, *10*, 4052.
6. Liu, H.Y.; Chen, J.; Chen, G.Q. A review of data classification algorithms in data mining. *J. Tsinghua Univ. (Nat. Sci. Ed.)* **2002**, *12*, 727–730.
7. Kotsiantis, S.B. Decision trees: A recent overview. *Artif. Intell. Rev.* **2013**, *39*, 261–283. [\[CrossRef\]](#)
8. Peter, A.; Yaw, M.M.; Ussiph, N. Predicting bank operational efficiency using machine learning algorithm: Comparative study of decision tree, random Forest, and neural networks. *Adv. Fuzzy Syst.* **2020**, *2020*, 8581202.
9. Izonin, I.; Tkachenko, R.; Shakhovska, N.; Ilchyshyn, B.; Singh, K.K. A Two-Step Data Normalization Approach for Improving Classification Accuracy in the Medical Diagnosis Domain. *Mathematics* **2022**, *10*, 1942. [\[CrossRef\]](#)
10. Huang, B.; Huang, D.R.; Mi, B. Research on E-Commerce Transaction Payment System Basedf on C4.5 Decision Tree Data Mining Algorithm. *Comput. Syst. Sci. Eng.* **2020**, *35*, 113–121.
11. Sembiring, N.S.B.; Sinaga, M.D.; Ginting, E.; Tahel, F.; Fauzi, M.Y. Predict the Timeliness of Customer Credit Payments at Finance Companies Using a Decision Tree Algorithm. In Proceedings of the 2021 9th International Conference on Cyber and IT Service Management (CITSM), Bengkulu, Indonesia, 22–23 September 2021; pp. 1–4.
12. Zhang, Y.L.; Feng, P.F.; Ning, Y. Random forest algorithm based on differential privacy protection. In Proceedings of the 20th IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom 2021), Shenyang, China, 20–22 October 2021; pp. 1259–1264.
13. Lv, C.X.; Li, Q.L.; Long, H.Q.; Ren, Y.M.; Ling, F. A differential privacy random forest method of privacy protection in cloud. In Proceedings of the 2019 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC), New York, NY, USA, 1–3 August 2019; pp. 470–475.
14. Mu, H.R.; Ding, L.P.; Song, Y.N.; Lu, G.Q. DiffPRFs: Random forest under differential privacy. *J. Commun.* **2016**, *37*, 175–182.
15. Wang, M.S.; Yao, L.; Gao, F.X.; Xu, J.C. Study on differential privacy protection for medical set-valued data. *Comput. Sci.* **2022**, *49*, 362–368.
16. Fu, J.B.; Zhang, X.J.; Ding, L.P. MAXGDDP: Decision data release with differential privacy. *J. Commun.* **2018**, *39*, 136–146.
17. Zhang, S.Q.; Li, X.H.; Jiang, X.Y.; Li, B. Aur-tree differential privacy data publishing algorithm for medical data. *Appl. Res. Comput.* **2022**, *39*, 2162–2166.

18. Blum, A.; Dwork, C.; McSherry, F.; Nissim, K. Practical privacy: The SuLQ framrk. In Proceedings of the Twenty-Fourth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS'05), Baltimore, MD, USA, 13–15 June 2005; Association for Computing Machinery: New York, NY, USA, 2005; pp. 128–138.
19. McSherry, F.D. Privacy integrated queries: An extensible platform for privacy-preserving data analysis. In Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data (SIGMOD'09), Providence, RI, USA, 29 June–2 July 2009; Association for Computing Machinery: New York, NY, USA, 2009; pp. 19–30.
20. Friedman, A.; Schuster, A. Data mining with differential privacy. In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'10), Washington, DC, USA, 25–28 July 2010; Association for Computing Machinery: New York, NY, USA, 2010; pp. 493–502.
21. Mohammed, N.; Chen, R.; Fung, B.C.M.; Yu, P.S. Data mining with differential privacy. In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'11), San Diego, CA, USA, 21–24 August 2011; Association for Computing Machinery: New York, NY, USA, 2011; pp. 493–501.
22. Zhu, T.Q.; Xiong, P.; Xiang, Y.; Zhou, W.L. An Effective Differentially Private Data Releasing Algorithm for Decision Tree. In Proceedings of the 2013 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, Melbourne, Australia, 16–18 July 2013; pp. 388–395.
23. Patil, A.; Singh, S. Differential private random forest. In Proceedings of the 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Delhi, India, 24–27 September 2014; pp. 2623–2630.
24. Li, Y.H.; Chen, X.L.; Liu, L. Random forest algorithm for differential privacy protection. *Comput. Eng.* **2020**, *46*, 93–101.
25. Paul, A.; Mukherjee, D.P.; Das, P.; Gangopadhyay, A.; Chintha, A.R.; Kundu, S. Improved random forest for classification. *IEEE Trans. Image Process.* **2018**, *27*, 4012–4024. [[CrossRef](#)] [[PubMed](#)]
26. Truex, S.; Liu, L.; Gursoy, M.E.; Yu, L. Privacy-preserving inductive learning with decision trees. In Proceedings of the 2017 IEEE International Congress on Big Data (BigData Congress), Honolulu, HI, USA, 25–30 June 2017; pp. 57–64.
27. Li, Y.Q.; Chen, Y.H.; Li, Q.; Liu, A.H. Random forest algorithm under differential privacy based on out-of-bag estimate. *J. Harbin Inst. Technol.* **2021**, *53*, 146–154.
28. Dwork, C. A firm foundation for private data analysis. *Commun. ACM* **2011**, *54*, 86–95. [[CrossRef](#)]
29. Dwork, C.; McSherry, F.; Nissim, K.; Smith, A. *Theory of Cryptography Conference*; Halevi, S., Rabin, T., Eds.; Springer: Berlin/Heidelberg, Germany, 2006; Volume 3876.
30. Geng, Q.; Viswanath, P. The optimal noise-adding mechanism in differential privacy. *IEEE Trans. Inf. Theory* **2015**, *62*, 925–951. [[CrossRef](#)]
31. Mironov, I. Rényi Differential Privacy. In Proceedings of the 2017 IEEE 30th Computer Security Foundations Symposium (CSF), Santa Barbara, CA, USA, 21–25 August 2017; pp. 263–275.
32. Xiong, P.; Zhu, T.Q.; Wang, X.F. Differential privacy protection and its application. *J. Comput. Sci.* **2014**, *37*, 101–122.