DIFFERENTIALLY PRIVATE GREEDY DECISION FOREST

Bangzhou Xin, Wei Yang⋆*, Shaowei Wang, Liusheng Huang*

University of Science and Technology of China

ABSTRACT

As information security is increasingly valued, privacy-preserving data mining has become a research hotspot in the field of big data and signal processing. We propose a new differentially private greedy decision forest algorithm called DPGDF to help improve the accuracy of privacy-preserving data mining. Unlike previous algorithms that only employed greedy decision trees or random forests, our algorithm uses a combination of greedy trees and parallel combination theory to construct a greedy decision forest and coordinate privacy protection and prediction accuracy to achieve the best balance. Combined with smooth sensitivity, the introduction of noise is minimized, making the prediction accuracy of the algorithm notably better than the current state-of-the-art algorithms. Experiments on the UCI datasets show that the prediction accuracy of our algorithm is about 10% higher than that of those algorithms.

***Index Terms*—** information security, differential privacy, data mining, decision forest

1. INTRODUCTION

In the age of big data, knowledge discovery is rapidly driven by machine learning and data mining. The analysis and mining of collected personal data, with the help of strong computing power, can find a lot of valuable information. However, information security is also threatened, and privacy-preserving data mining shows greatly important significance.

In general, k-anonymity [1] and l-diversity [2] are common privacy protection technologies for data distribution. But their protection capabilities can not meet the general needs. The emergence of differential privacy [3] provides a new direction for privacy protection data mining. It does not require any assumptions about the attacker's background knowledge. And it only relies on the parameter called privacy budget to control the probability of privacy leakage. Decision tree [4] is one of the most popular data mining algorithms of classification, whose model is interpretable and fast. Our work focuses on decision tree data mining issues that satisfy differential privacy. Currently, there are a variety of decision tree data mining methods based on differential privacy. For a single decision tree, Friedman and Schuster proposed a differentially private ID3 decision tree induction algorithm based on the SuLQ framework [5] in [6]. In 2015, Fletcher and Islam [7] proposed a differentially private decision forest algorithm which takes advantage of the local sensitivity [8] and the Exponential Mechanism [9]. It's worth mentioning that all these works [6, 7] introduced a large amount of noise, which impaired their performance of prediction accuracy.

Considering a bunch of decision trees, Jagannathan $et\ al.$ [10] raised a different approach which utilizes random decision trees [11]. Their algorithm satisfies differential privacy by using the Laplace Mechanism [12] at the leaf. Fletcher and Islam adopt smooth sensitivity [8] to reduce the noise at the leaf in [13], increasing accuracy to some extent. Focusing on binary classification problem, Rana $et\ al.$ proposed a novel differentially private decision tree induction algorithm in [14]. They used a weaker form of differential privacy, and proved that a large ensemble of trees can get higher utility. However, because their random tree fails to make full use of the information contained in the data, the accuracy of their algorithm prediction is also undesirable.

In this paper, we propose the differentially private greedy decision forest (DPGDF), an algorithm based on parallel combination theory [15] to solve privacy-preserving data mining. By using disjoint subsets of the data to build greedy decision trees, we can apply the whole privacy budget on each tree. Maximizing the use of dataset and privacy budget allows our algorithms to perform well in any privacy budget. We conduct experiments on real-world datasets to evaluate the effectiveness of our algorithms. The experimental results show that our DPGDF outperforms the current state-of-the-art privacy-preserving data mining algorithms. In all the five datasets tested, DPGDF is about 10% ahead of those algorithms in terms of prediction accuracy.

The following definitions are the theoretical basis of our algorithm.

Definition 1. *(ϵ-Differential Privacy [3]) A randomized function $\mathcal{M}$ gives ϵ-differential privacy if for all datasets D_1 and D_2 differing on a single record, and all $S \subseteq$ Range($\mathcal{M}$),*

$$Pr[\mathcal{M}(D_1) \in S] \leq exp(\epsilon) \times Pr[\mathcal{M}(D_2) \in S] \quad (1)$$

Definition 2. *(Parallel Composition [15]) For disjoint subsets $x_i \subseteq x$, let query $f(x_i)$ satisfy ϵ-differential privacy; then applying all queries $f(x_i)$ still satisfies ϵ-differential privacy.*

⋆Corresponding author. Email:qubit@ustc.edu.cn

This work is supported by the National Natural Science Foundation of China (No. 61572456) and the Anhui Initiative in Quantum Information Technologies (No. AHY150300).

2. DIFFERENTIALLY PRIVATE GREEDY DECISION FOREST

We describe our algorithm in this section, including the main steps of the algorithm, the selection of important parameters, and the distribution of privacy budget.

2.1. Algorithm summary

The basic idea of our algorithm can be summarized as the following steps: First, we calculate the size of the forest τ and the optimal depth d of each tree, and divide the dataset into τ parts. Then we calculate the smooth sensitivity S of the dataset used by each tree and create a greedy decision tree accordingly. Repeat the previous step until we create the whole forest, as Alg. 1 shows. Finally, query the leaf nodes and output the majority class labels.

The complete algorithms are summarized in Alg. 1 and Alg. 2, respectively.

2.2. Create a single decision tree

As shown in Alg. 2, when constructing the greedy decision tree in our algorithm, we follow the conventional approach [16]: our algorithm uses Gini index [17] to choose splitting attributes. Firstly, we compute the Gini index of all attributes, then pick out one attribute a from attribute set A as splitting attribute by Exponential Mechanism. Then, we remove the attribute a from the attribute set A. The remaining layers are processed in the same way until the leaf nodes. When the algorithm gets the leaf node, it returns majority label by Exponential Mechanism with smooth sensitivity.

Algorithm 1 Differentially Private Greedy Decision Forest

Input: $tree\ max\ depth : d, tree\ number : \tau, data : D,$ $private\ budget : \epsilon, set\ of\ attributes : A,$ $class\ label : C$
Output: $decision\ forest\ classification : F$

function DPGDF($D, A, C, d, \tau, \epsilon$)
  $Data \leftarrow Divide(D, \tau)$
  for $t = 1, ..., \tau$ **do**
    $depth \leftarrow 0$
    $S \leftarrow sensitvity(Data[t])$
    $T \leftarrow Build\_Tree(Data[t], A, depth, d, \epsilon, T, S)$
    $F \leftarrow F \cup T$
  return F

2.3. Parameter selection

Building a greedy decision tree usually requires using all m attributes, but it is not necessary while building multiple trees. We empirically find there is an optimal depth range from $m/2$ to m where prediction accuracy is highest for most datasets. Intuitively, we could improve accuracy by increasing the number of trees while ensuring the same privacy. However, the experimental results negated this idea. Although privacy budget ϵ won't be divided by each individual in the forest, the accuracy of the algorithm will decline by the fact that each tree contains too few records. Because lager numbers of trees cause more leaf nodes to output a label that differs from the actual most frequent label. In other words, the most frequent label will become no longer the most frequent. They will be diluted by the large number of trees. In general, the number of records per tree is determined by the actual dataset. The number of attributes and the number of candidate values of the attribute will affect the size of the forest. Through experiments, we found that when the tree depth is not more than 10 and the privacy budget is 0.1, each tree can be allocated 400 records to train the classifier to obtain a fairly high accuracy. As the depth increases, more records will be required. And, we use a table (Table 1) to illustrate the relationship between depth and required records when privacy budget is 0.1. Another fact is that when the privacy budget is small, the smaller the forest size, the higher the accuracy of the classifier.

Algorithm 2 Build_Tree

Input: $data : D, set\ of\ attributes : A, max\ depth : d$ $current\ depth : depth, privacy\ budget : \epsilon, Tree$ $: T, sensitivity : S$
Output: $decision\ tree : T$

if $depth < d\ and\ len(D) > 10$ **then**
  if $T = None$ **then**
    $T \leftarrow Tree(D), create\ root\ node\ with\ D$
  $subset \leftarrow divide(D, C), divide\ data\ according$ $to\ class\ label$
  if $len(subset) \leq 1$ **then**
    $T\ arrives\ at\ leaf\ node\ and$ $label \leftarrow majority(Exp\ Mechanism(C, S, \frac{\epsilon}{2}))$
  for $a\ in\ A$ **do**
    $gini \leftarrow calculate\ Gini\ index\ of\ a$
    $T.attr \leftarrow Exp\ Mechanism(gini, \frac{\epsilon}{2\times(d-1)})$ $select\ splitting\ attribute$
  $subD \leftarrow divide(D, T.attr),\ divide\ data\ based$ $on\ T.attr$
  $A \leftarrow A - T.attr\ remove\ attribute\ that\ has$ $been\ used$
  for $key\ in\ subD.keys()$ **do**
    $Build\_Tree(subD[key], A, depth, d, \epsilon, T, S)$ $create\ child\ tree\ based\ on\ T.attr$
  return T
else
  $T\ arrives\ at\ leaf\ node\ and$ $label \leftarrow majority(Exp\ Mechanism(C, S, \frac{\epsilon}{2}))$
  return T

2.4. Distribution of privacy budget

Here we introduce how to allocate privacy budget. Although adopting parallel composition theory allows each tree has an

Table 1

Depth	1-10	10-20	20-30
Record Number	400	400-1000	1000-3000

entire privacy budget, all layers of one tree are composite. Every layer will cost privacy budget alone. The leaf node is the most important layer in a tree, because it decides which is the majority label. We allocate a half of privacy budget to the leaf node. As for the remaining layers, they evenly allocate the remaining half of the budget. When we calculate the majority label, we adopt smooth sensitivity [13] in the Exponential Mechanism.

2.5. Theoretical analysis

After building a greedy forest, our algorithm outputs the entire forest, including the splitting attributes of the internal nodes of each tree and the labels of the leaf nodes. Here we show that the algorithm satisfies ϵ-differential privacy in Theorem 1.

Theorem 1. *Differentially Private Greedy Decision Forest satisfies ϵ-differential privacy.*

Proof. First, we consider a single tree. A privacy budget $\frac{\epsilon}{2\times(d-1)}$ is used for each attribute selection. Each tree has a maximum of $(d-1)$ attribute selections. For each layer, the way it satisfies $\frac{\epsilon}{2\times(d-1)}$-differential privacy is to choose the splitting attribute with probability

$$Pr(a) \propto exp(\frac{-\epsilon \times Gini(a,A)}{2(d-1)\times\Delta(Gini)}) \tag{2}$$

All attribute selections follow the composition theory [9], so all (d-1) internal layers satisfy α-differential privacy. Where

$$\alpha = \sum_{i=1}^{d-1}\epsilon_i = (d-1)\times\frac{\epsilon}{2\times(d-1)} = \frac{\epsilon}{2} \tag{3}$$

The other half of the privacy budget is spent on leaf nodes. Following the composition theory with the label selection of leaf nodes, the whole tree is ϵ-differential privacy. Because each tree is built with a dataset that is not adjacent, the dataset of the whole forest used D satisfies

$$D = \bigcup_{i=1}^{\tau} D_i \tag{4}$$

all the trees follow the theory of parallel composition, thus, the algorithm satisfies ϵ-differential privacy. □

3. EXPERIMENTS

We present the experimental results and analysis of datasets from the UCI Machine Learning Repository [18]. We compare the classification accuracy of our algorithm with Smooth Random Trees (SRT) [13] and Random Decision Trees (RDT) [10] which are currently the state-of-the-art algorithms. According to a comparison of the main properties of the differentially-private decision tree algorithms in [19], the accuracy of the existing multi-class greedy decision tree algorithm is relatively low, and thus it is not listed here. Of course, we have a base-line that is the accuracy of our algorithm without privacy. All reported prediction accuracy of all algorithms are average results of performing 100 times repetition. Without loss of generality, every time we shuffled the data. And we also compared the accuracy of all algorithms at different privacy budget, $\epsilon = 0.1, 0.2, 0.3, 0.4, 0.5, 1$. We repeated experiments of SRT and RDT on different datasets using recommend parameter.

The datasets we used are all public and have the following names in the UCI Machine Learning Repository [18] : "Car Evaluation" (Fig. 1), "Mushroom" (Fig. 2), "Nursery" (Fig. 3), "Chess" (Fig. 4), "balance" (Fig. 5). After removing records with missing values, the datasets are divided into training and testing sets in a ratio of eight to two. The number of records in each dataset ranges from 625 to 12960; the number of attribute ranges from 4 to 36; and the number of class ranges from 2 to 5. Moreover, we have also tested the algorithm on other datasets. The results show that our algorithm still has obvious advantages.

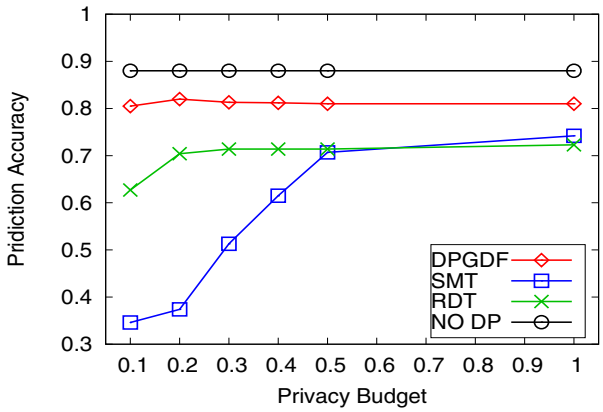

Figure 1 : Car Evaluation

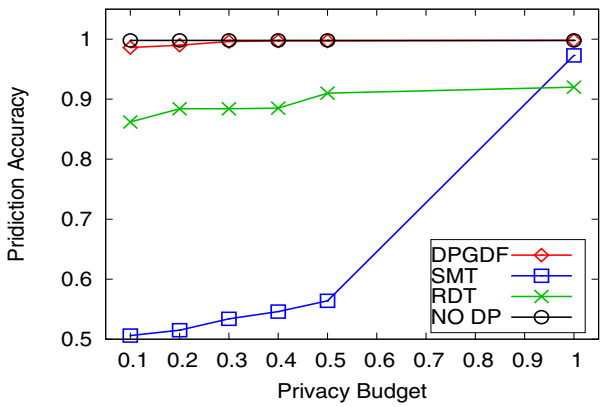

Figure 2 : Mushroom

As can be seen from the comparison chart, our algorithm is always better than the other two regardless of the privacy budget. It can get high prediction accuracy at low privacy budgets early. Fig. 1 shows the results of three algorithms with "Car Evaluation" dataset under different privacy budget. The dataset has about 1700 records, and each record includes six attributes and one label. Obviously, when our algorithm has a privacy budget of 0.1, the prediction accuracy is far ahead of the other two algorithms. Although the accuracy of their prediction has increased with the increase of the budget, our algorithm accuracy has always been close to nearly 10 percentage points. A strange phenomenon is that as the privacy budget increases, the optimal accuracy of "Car Evaluation" decreases. In common sense, the larger the privacy budget, the more accurate the results should be. However, this gives us an unexpected result. In the case of ensuring that the algorithm and parameters are correct, we try to find the cause of this phenomenon. One plausible explanation is that differential privacy can help the algorithm prevent or reduce overfitting [20]. Excessive privacy budgets result in reduced ability to suppress overfitting.

Fig. 2 shows the results of three algorithms with "Mushroom" dataset. Our algorithms maintain near-perfect accuracy rates under any size privacy budget. The prediction accuracy of the RDT algorithm is always less than 90%. The SMT algorithm only in the privacy budget of up to 1, the prediction accuracy rate barely close to our algorithm. Fig. 3 is "nursery" dataset. When the privacy budget is greater than 0.1, our algorithm accuracy begins to be far ahead of the other two algorithms. When the budget reaches 0.4, its accuracy is almost the same as the no-privacy algorithm. Fig. 4 shows the results of three algorithms with "Chess" dataset. Unlike the car dataset, which has only six attributions, the dataset has 36 attributions. In the face of such a complex data set, our algorithms still have an absolute lead in accuracy. And we find that random approach is difficult to achieve the same accuracy of greedy approach when the privacy budget is high enough. Because random methods do not use all attributes to build a tree, they use only hal f the number of attributes to build a decision tree. In this case, the algorithm will not be able to take full advantage of the information contained in the dataset to make the prediction a good result.

In the experiment, we found an interesting phenomenon: the number of trees is probably proportional to the privacy budget. After repeated experiments, we found that the number of trees is almost linear with the privacy budget. We tested multiple datasets and found that the law is basically established, especially when the privacy budget is small.

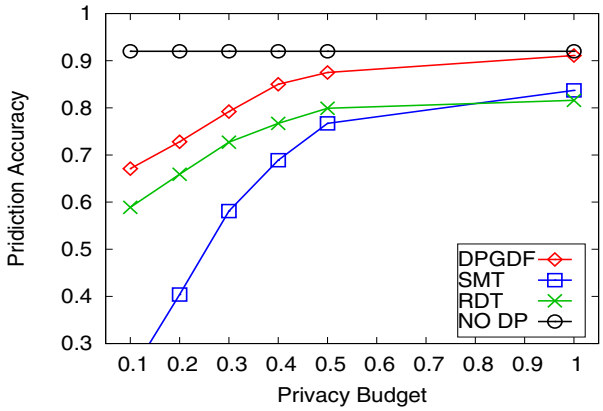

Figure 3 : Nursery

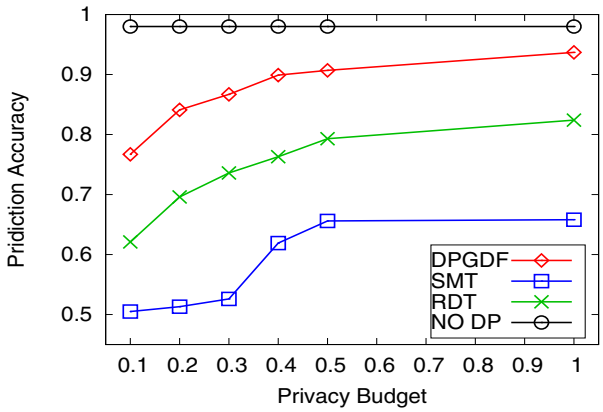

Figure 4 : Chess

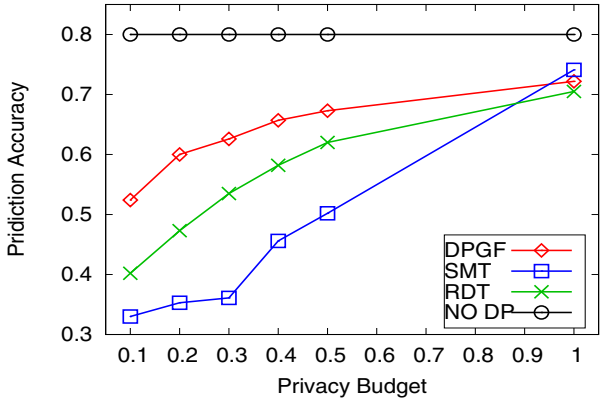

Figure 5 : Banlace

4. CONCLUSION

The DPGDF algorithm proposed in this paper is used for privacy-preserving data mining, which protects private sensitive information in data while maintaining data availability. According to the characteristics of the decision tree algorithm, while protecting privacy, the parallel composition theory and decision forest are used to improve the accuracy. The experimental results show that our algorithm has good performance even when the privacy budget is low. Its overall performance is superior to the current state-of-the-art algorithms. Future work directions include improving the accuracy of the algorithm and extending the evaluation of the proposed algorithm.

5. REFERENCES

[1] Latanya Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, 2002.

[2] Ashwin Machanavajjhala, Johannes Gehrke, Daniel Kifer, and Muthuramakrishnan Venkitasubramaniam, "\ell-diversity: Privacy beyond\kappa-anonymity," in *null*. IEEE, 2006, p. 24.

[3] Cynthia Dwork, "Differential privacy," in *International Colloquium on Automata, Languages, and Programming*, 2006, pp. 1–12.

[4] Jiawei Han, Jian Pei, and Micheline Kamber, *Data mining: concepts and techniques*, Elsevier, 2011.

[5] Avrim Blum, Cynthia Dwork, Frank McSherry, and Kobbi Nissim, "Practical privacy: the sulq framework," in *Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. ACM, 2005, pp. 128–138.

[6] Arik Friedman and Assaf Schuster, "Data mining with differential privacy," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2010, pp. 493–502.

[7] Sam Fletcher and Md Zahidul Islam, "A differentially private decision forest," in *Proceedings of the 13th Australasian Data Mining Conference*, 2015, pp. 1–10.

[8] Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith, "Smooth sensitivity and sampling in private data analysis," in *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*. ACM, 2007, pp. 75–84.

[9] Frank McSherry and Kunal Talwar, "Mechanism design via differential privacy," in *Foundations of Computer Science, 2007. FOCS'07. 48th Annual IEEE Symposium on*. IEEE, 2007, pp. 94–103.

[10] Geetha Jagannathan, Krishnan Pillaipakkamnatt, and Rebecca N Wright, "A practical differentially private random decision tree classifier," in *Data Mining Workshops, 2009. ICDMW'09. IEEE International Conference on*. IEEE, 2009, pp. 114–121.

[11] Leo Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[12] Cynthia Dwork, Aaron Roth, et al., "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[13] Sam Fletcher and Md Zahidul Islam, "Differentially private random decision forests using smooth sensitivity," *Expert Systems with Applications*, vol. 78, pp. 16–31, 2017.

[14] Santu Rana, Sunil Kumar Gupta, and Svetha Venkatesh, "Differentially private random forest with high utility," in *Data Mining (ICDM), 2015 IEEE International Conference on*. IEEE, 2015, pp. 955–960.

[15] Frank D McSherry, "Privacy integrated queries: an extensible platform for privacy-preserving data analysis," in *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*. ACM, 2009, pp. 19–30.

[16] J. Ross Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986.

[17] Leo Breiman, *Classification and regression trees*, Routledge, 2017.

[18] Catherine L Blake, "Uci repository of machine learning databases, irvine, university of california," *http://www. ics. uci. edu/˜ mlearn/MLRepository. html*, 1998.

[19] Sam Fletcher and Md Zahidul Islam, "Decision tree classification with differential privacy: A survey," *arXiv preprint arXiv:1611.01919*, 2016.

[20] Tianqing Zhu, Gang Li, Wanlei Zhou, and S Yu Philip, "Differentially private data publishing and analysis: a survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 8, pp. 1619–1638, 2017.