




Security Evaluation of Decision Tree Meets Data Anonymization

Ryousuke Wakabayashi¹, Lihua Wang²^a, Ryo Nojima^{3,2}^b and Atsushi Waseda^{1,*}^c

¹Department of Informatics, Tokyo University of Information Sciences,
4-1 Onaridai, Wakaba-ku, Chiba 265-8501, Japan

²Cybersecurity Research Institute, National Institute of Information and Communications Technology,
4-2-1, Nukui-Kitamachi, Koganei, Tokyo 184-8795, Japan

³College of Information Science and Engineering, Ritsumeikan University,
1-1-1 Noji-higashi, Kusatsu, Shiga 525-8577, Japan

Keywords: Privacy Preserving, Decision Tree, k -Anonymity.

Abstract: This paper focuses on the relationship between decision trees, a typical machine learning methods, and data anonymization. We first demonstrate that the information leakage from *trained* decision trees can be evaluated using well-studied data anonymization techniques. We then show that decision trees can be strengthened against specific attacks using data anonymization techniques. Specifically, we propose two decision tree pruning methods to improve security against uniqueness and homogeneity attacks, and we evaluate the accuracy of these methods experimentally.

1 INTRODUCTION

Recently, with the rapid evolution of machine learning technology and the expansion of data due to the information technology developments, it has become increasingly important for companies to determine how to best utilize big data effectively and efficiently. However, big data often includes personal and privacy information; thus, careless utilization of such sensitive information may lead to unexpected penalties.

Many privacy-preserving technologies have been proposed to utilize data while preserving user privacy. Typical privacy-preserving technologies include data anonymization and secure computation. In addition, privacy-preserving technologies inherently involve a trade-off relationship between security and usability.

Therefore, in this paper, we evaluate the privacy leakage of decision trees, which is a fundamental machine learning method, trained using data containing personal information. In particular, we examine the extent to which personal information is leaked from the model from a data anonymization perspective. Historically, data anonymization research has progressed from pseudonymization to k -

anonymity (Sweeney, 2002), l -diversity (Machanavajjhala et al., 2006), t -closeness (Li et al., 2007), and so on. Currently researchers are focusing on membership privacy and differential privacy (Stadler et al., 2022).

In this paper, we initially discuss the common structure of decision trees and data anonymization. We then demonstrate that previously proposed attacks against anonymization in the past can also be applied to decision trees. Specifically, we demonstrate that


- (1) the *uniqueness* attack against anonymization (pseudonymization),
- (2) the *homogeneity* attack and *background knowledge* attack against k -anonymity,


which are representative attacks, can also be applied to decision trees. In addition, we discuss:

- (3) how to prevent privacy information leakage from a learned decision tree using data anonymization techniques.

Specifically, we employ k -anonymity as a means to enhance the security of decision trees. It is noteworthy that similar methods have been proposed in previous studies. For example, Slijepcevic et al. provided a systematic comparison and detailed investigation into the effects of k -anonymisation data on the results of machine learning models. However, they did not investigate the impact of k -anonymization on

^a <https://orcid.org/0000-0002-7553-423X>

^b <https://orcid.org/0000-0002-2955-2920>

^c <https://orcid.org/0000-0002-3594-5704>

*Corresponding author.

Table 1: Dataset.

Zip	Age	Nationality	Disease
13053	28	Russian	Heart
13068	29	American	Heart
13068	21	Japanese	Flu
13053	23	American	Flu
14853	50	Indian	Cancer
14853	55	Russian	Heart
14850	47	American	Flu
14850	59	American	Flu
13053	31	American	Cancer
13053	37	Indian	Cancer
13068	36	Japanese	Cancer
13068	32	American	Cancer

Table 2: k -anonymity ($k = 4$).

Zip	Age	Nationality	Disease
130**	< 30	*	Heart
130**	< 30	*	Heart
130**	< 30	*	Flu
130**	< 30	*	Flu
1485*	< 30	*	Cancer
1485*	> 40	*	Heart
1485*	> 40	*	Flu
1485*	> 40	*	Flu
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer

trained decision trees (Slijepcevic et al., 2021). Nojima and Wang proposed a method that employs k -anonymity to enhance randomized decision trees, resulting in satisfactory levels of differential privacy. The advantage of this proposed method lies in its ability to achieve differential privacy without introducing Laplace noise (Nojima and Wang, 2023). Their work differs significantly from existing differentially private decision tree protocols (Friedman and Schuster, 2010; Jagannathan et al., 2012; Bai et al., 2017) which require adding noise to the tree model, although a limitation of their approach is the need for multiple trees.

These results of the current study suggest that there is a deep relationship between data anonymization and decision trees, and that investigating anonymization, including k -anonymity, is beneficial in terms of analyzing and improving the privacy protection mechanism of decision trees.

The remainder of this paper is organized as follows. Section 2 introduces relevant preliminary information, e.g., anonymization methods and decision trees. Section 3 demonstrates how to convert attack methods against data anonymization into attacks against decision trees. In addition, relevant experimental results of these attacks are also discussed in Section 3. In Section 4, we demonstrate how much security and accuracy can be practically realized when the decision tree is strengthened using a method that is similar to k -anonymity. Finally, the paper is concluded in Section 5, including a brief discussion of potential future issues.

2 PRELIMINARIES

2.1 Data Anonymization

When providing personal data to a third party, it is necessary to modify the data to preserve user privacy. Here, modifying the user’s data (i.e., a record) such that an adversary cannot re-identify a specific individual is referred to as *data anonymization*. As a basic technique, to prevent re-identification, an identifier, e.g., a name or employee number is deleted or the data holder replaces it with a pseudonym ID. This process is referred to as *pseudonymization*. However, simply modifying identifiers does not imply privacy preservation. In some cases, individuals can be re-identified by a combination of features (i.e., a quasi-identifier); thus, it is necessary to modify both the identifier and the quasi-identifier to reduce the risk of re-identification. In most of cases, the identifiers themselves are not used for data analysis; thus, removing identifiers does not sacrifice the quality of the dataset so much. However, if we modify quasi-identifiers in the same manner, although the data may become anonymous, it will also become useless. A typical anonymization technique for quasi-identifiers is to “roughen” the numerical values.

2.2 Attacks Against Data Anonymization

2.2.1 Attacks Against Pseudonymization

A simple attack is possible against pseudonymized data from which identifiers, e.g., names, have been removed. In this attack, the attacker uses the quasi-identifier of a user u . If this attacker obtains the pseudonymized data, by searching for user u ’s quasi-

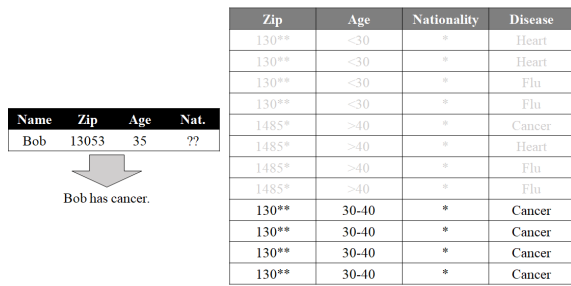


Figure 1: Homogeneity attack.

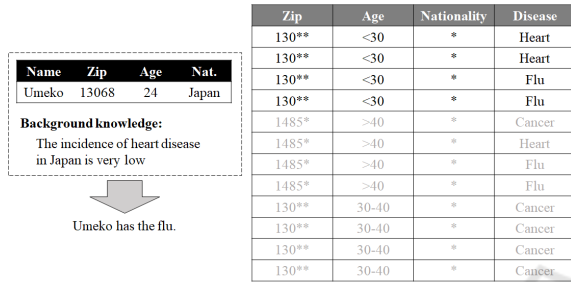


Figure 2: Background knowledge attack.

identifier in pseudonymized data, the attacker can obtain sensitive information about u . For example, if the attacker obtains the dataset shown in Table 1, and knows friend u 's Zip code is 13068, age is 29, and nationality is American, then, by searching the dataset, the attacker can identify that user u is suffering from some heart-related disease. This attack is referred to as the **Uniqueness attack**.

k -Anonymity: k -anonymity is a countermeasure against the uniqueness attack. In k -anonymity, features are divided into quasi-identifiers and sensitive information, and the same quasi-identifier is modified such that it does not become less than $k - 1$ users. Table 2 shows anonymized data that has been k -anonymized ($k = 4$) using quasi-identifiers, e.g., ZIP code, age, and nationality.

2.2.2 Homogeneity Attack

At a cursory glance, k -anonymity appears to be secure, however even if k -anonymity is employed, a *homogeneity* attack is still feasible. This attack becomes possible if the sensitive information is the same. Let's see the k -anonymized dataset shown on the right side of Figure 1, and we assume the attacker on the left side of Figure 1. Here, the attacker has the information (Zip, Age) = (13053, 35) and all sensitive information in those records is cancer, it can be revealed that Bob has cancer.

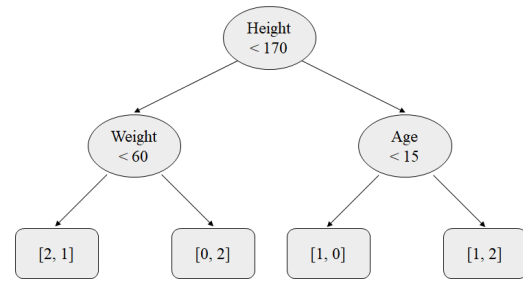


Figure 3: Decision tree.

2.2.3 Background Knowledge Attack

Homogeneous attacks suggest a problem when records with the same quasi-identifier have the same sensitive information; however, a previous study (Li et al., 2007) also argued that there is a problem even in cases where the records are not the same. The k -anonymized dataset on the right side of Figure 2 shows four records with quasi-identifiers (130, < 30, *), and two types of sensitive information, i.e., (Heart, Flu). Here, assume that the attacker has background knowledge of the data similar to that shown on the left side of Figure 2. In this case, there are certainly possibilities of Heart and Flu; however, if the probability of Japanese experiencing heart disease is extremely low, Umeko is estimated as flu. Thus, it must be acknowledged that k -anonymity does not provide a high degree of security.

2.3 Decision Trees

Decision trees are supervised learning methods that are primarily used for classification tasks, and a tree structure is created while learning from data (Figure 3). When predicting the label y of \mathbf{x} , the process begins from the root of the tree, and the corresponding leaf is searched for while referring to each feature of \mathbf{x} . Finally, by this referral process, y is predicted.

The label determined by the leaf is determined by the dataset D used to generate the tree structure. In other words, after the tree structure is created, for each element (\mathbf{x}_i, y_i) in dataset D , the corresponding leaf ℓ is found and the value of y_i is stored. If $y_i \in Y = \{0, 1\}$, then in each leaf ℓ , the number of y_i that was 0 and the number of y_i that was 1 are preserved. More precisely, $[\ell_0, \ell_1]$ are preserved for each leaf ℓ , where ℓ_0 and ℓ_1 represent the numbers of data with label y that were 0 and 1, respectively. Table 3 shows the notations used in the paper.

For the prediction given \mathbf{x} , we first search for the corresponding leaf, and it may be judged as 1 if $\frac{\ell_1}{\ell_0 + \ell_1} > \frac{1}{2}$, and 0 otherwise. Here the threshold $1/2$ can be set flexibly depending on where the decision

Table 3: Notations.

k	Anonymization parameter
D	Dataset $\{x_i, y_i\}$
\mathbf{x}	Data (x_1, \dots, x_f) with f features
y	Label of data \mathbf{x}
Y	Label space $Y = \{0, 1\}$ in the paper
ℓ	Leaf or Leaves
ℓ_i	The number of data that classed to leaf ℓ with label $i \in Y$
n_ℓ	The number of data that classed to leaf ℓ , i.e., $n_\ell = \ell_0 + \ell_1$
s	The pruning threshold, which is set to $k - 1$ in the experiments
N_u	Total number of users who can be identified by a homogeneous attack
N_ℓ	Number of leaves can perform homogeneity attack

tree is applied, and when providing the learned decision tree to a third party, it is possible to pass ℓ_0 and ℓ_1 together for each leaf ℓ . In this paper, we consider the security of decision trees in such situations.

Generally, the deeper the tree structure, the more likely it is to overfit; thus, we frequently prune the tree, and this technique is employed to realize privacy preservation in this paper.

3 SECURITY OF DECISION TREES FROM DATA ANONYMIZATION PERSPECTIVE

3.1 Security Analysis

Generally, a decision tree is constructed from a given dataset; however, we show that it is also possible to partially reconstruct the dataset from the decision tree. Table 4 shows an example of re-constructing a dataset from the decision tree shown in Figure 3. As can be seen, it is impossible to reconstruct the original data completely from a binary tree model; however, it is possible to extract some of the data. By exploiting this essential property, it is possible to mount some attacks against reconstructed data, as discussed in Section 2. In the following, using Table 4 as an example, we discuss specific cases of how each attack can be applied.

- **Uniqueness Attack:** In the dataset (Table 4) recovered from the model, there is one user whose height is greater than 170 and who is under 15 years of age (in the sixth row); thus, it is possible to perform a uniqueness attack against this user.

Table 4: Example of conversion from decision tree to anonymized data.

Height	Weight	Age	Helth
< 170	< 60	*	yes
< 170	< 60	*	yes
< 170	< 60	*	no
< 170	≥ 60	*	no
< 170	≥ 60	*	no
≥ 170	*	< 15	yes
≥ 170	*	≥ 15	yes
≥ 170	*	≥ 15	no
≥ 170	*	≥ 15	no

Note that pruning decision trees can be an effective mechanism to prevent uniqueness attacks.

- **Homogeneous Attack:** Similarly, in the fourth and fifth rows, height < 170, weight ≥ 60, and health status are the same (i.e., “unhealthy”), thus, an homogeneous attack is possible.
- **Background knowledge attack:** Similarly, in the seventh, eighth, and ninth rows there are 3 users whose data meet both height ≥ 170 and age ≥ 15. Among these users, one is healthy (yes) and two are unhealthy (no). As an attacker, we can consider the following:

- (Background knowledge of user A) Height: 173, Age: 33, Healthy
- (Background knowledge of target user B) Height: 171, age: 19.

In this case, if the adversary knows that user A is healthy, he/she can identify that user B is unhealthy.

3.2 EXPERIMENTS

In this study, we used three datasets to evaluate the vulnerability of decision trees against uniqueness and homogeneous attacks, i.e., the Nursery dataset (Rajkovic, 1997), the Loan dataset (Mahdi Navaei,), and the Adult dataset (Becker and Kohavi, 1996). In these experiments, we used Python3 and sklearn library to train the decision trees. The characteristics of each dataset are described as follows:

- **Nursery Dataset:** The Nursery dataset contains 12,960 records with 8 features, with a maximum of five values for each feature.
- **Loan Dataset:** The Loan dataset contains 5,000 records with 12 features. Each feature has many possible values, and the number of records is small.
- **Adult Dataset:** The Adult dataset contains 48,842 records with 14 features. Here, each fea-

Table 5: Number of leaves for which a uniqueness attack is possible.

Tree Depth	Nursery	Loan	Adult
3	0	0	0
4	0	0	0.5
5	0	0	1.8
6	0	3.7	5.1
7	0	6	11

Table 6: Number of leaves (N_ℓ) can perform homogeneity attacks & Total number of users (N_u) who can be identified by a homogeneous attack.

Tree Depth	Nursery		Loan		Adult	
	(N_ℓ)	(N_u)	(N_ℓ)	(N_u)	(N_ℓ)	(N_u)
3	1	3448	3.8	3653	0.3	0.7
4	1	3448	4.8	3303	2.4	210
5	3	5057	7.9	3729	7.3	830
6	11.2	6822	16.7	3746	19	1623
7	24	7863	27	3837	34.4	1918

ture has more possible values and more records than the Nursery and Loan datasets

3.2.1 Uniqueness Attack Experiment

In this experiment, the tree depths were set to 3, 4, 5, 6, and 7. We divided each dataset into a training set and an evaluation set. The training set, which was used to train the decision tree, contained 80% of the records in the dataset. Here, the decision tree was trained 10 times and the average was taken. The number of leaves for which a uniqueness attack is possible for each dataset is shown in Table 5. On the Adult dataset, there are cases where an individual can be identified by taking only four features. Thus, it is possible to perform a uniqueness attack from the trained decision tree. In other words, the risk of information leakage is possible. In addition, the Nursery dataset has a small number of value types for each feature; thus, the risk of uniqueness attacks is low.

3.2.2 Homogeneous Attack Experiments

Table 6 shows the results of the homogeneous attack experiments. As in the previous experiment, here, we set the tree depth to 3, 4, 5, 6, and 7, and we divided each dataset into a training set (80%) and an evaluation set. The decision tree was trained 10 times, and averages of the following numbers were computed.

- Number of leaves ℓ such that $(\ell_0, \ell_1) = (0, z), (z, 0)$, where $z > 0$, and
- Number of users who can be identified by a homogeneous attack.

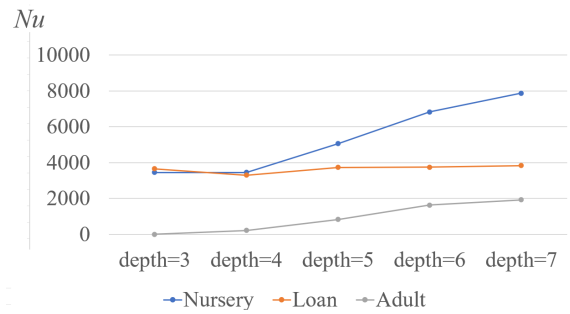


Figure 4: Tree depth & number of users who can be identified by a homogeneous attack.

On all datasets, even if the tree depth is small, information can be leaked by a homogeneous attack. In addition, similar to the uniqueness attacks, susceptibility to homogeneous attacks increases as the tree depth increases as shown in Figure 4.

4 USING ANONYMIZATION TO STRENGTHEN THE DECISION TREE

In this section, we show that the data anonymization technique can strengthen the decision tree.

4.1 Applying k -Anonymity

In a previous study (Nojima and Wang, 2023), k -anonymization was achieved by “removing leaves with a small number of users” for a randomized decision tree. Note that a similar method can be applied to the decision trees. Specifically, by setting $s = k - 1$ and pruning leaves such that $n_\ell = \ell_0 + \ell_1 \leq s$, a method that is similar to k -anonymity can be realized. Two corresponding methods are illustrated in Figures 5 and 6. Here, after training, we modify the trained decision tree as follows:

- **Method 1** (Figure 5): Leaves ℓ that result in $n_\ell \leq s$ are pruned.
- **Method 2** (Figure 6): For nodes with at least one child with $n_\ell \leq s$, both children are pruned, and the parent node is made a leaf node.

4.2 Experiments

We conducted experiments to verify the difference in accuracy and the possibility of attack for pruning Methods 1 and 2.

Here, a decision tree was trained in the manner described previously, and the impact on accuracy with

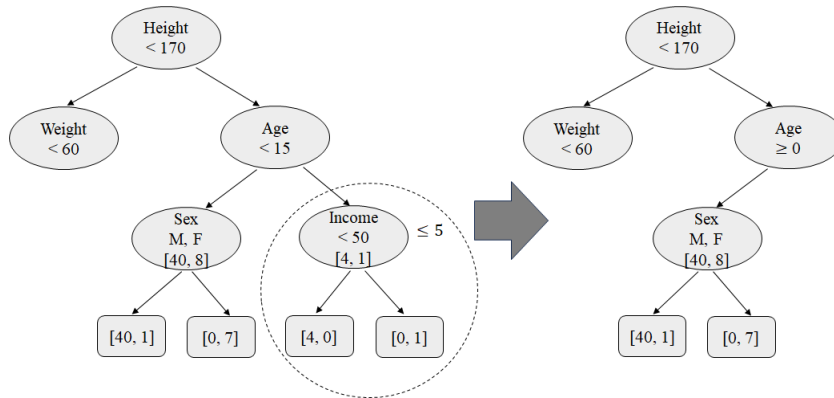


Figure 5: Decision tree pruning Method 1.

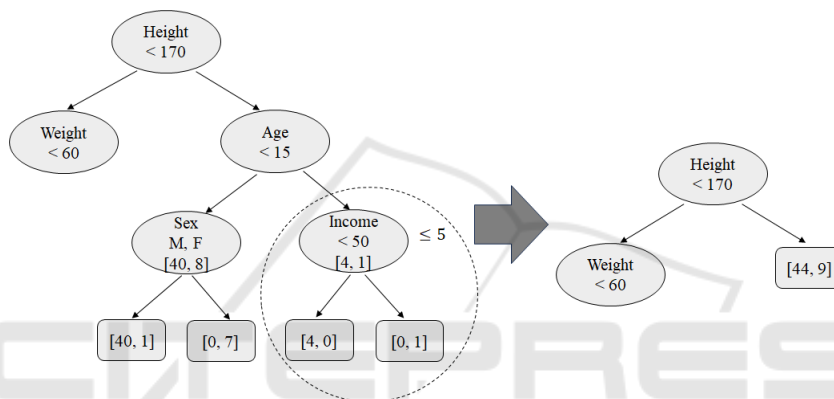


Figure 6: Decision tree pruning Method 2.

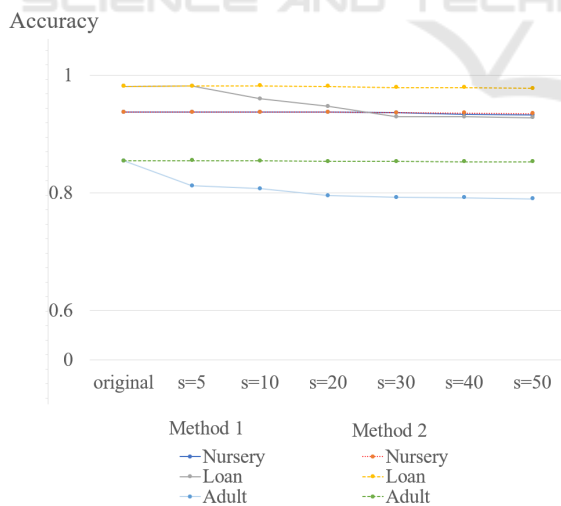


Figure 7: The pruning threshold s & accuracy.

a tree depth of 7 is shown in Table 7. For Method 1, the impact on accuracy was large when $s = 5$ on the Adult dataset. In addition, accuracy decreased as the threshold value s increased.

In terms of Method 2, the influence on accuracy

was small, and the number of users who can perform homogeneous attacks also decreased. We found that Method 2 exhibited better accuracy and effectiveness than Method 1 against homogeneous attacks as shown in Figure 7. For Method 1, the record assigned to the pruned leaf was deleted and was not used when predicting the label. In contrast, for Method 2, it was used as the predicted label at the parent node, which appeared to influence accuracy in the experiments.

5 CONCLUSION

Contribution: In this paper, we evaluated the possibility of information leakage of data from trained decision trees by discussing the relationship between the decision trees and data anonymization. We found that information leakage is possible via three distinct attacks, i.e., the uniqueness, homogeneous, and background knowledge attacks. We verified that the risk associated with uniqueness attacks was high when the total number of feature combinations was large. In contrast, we found that the risk of homogeneous at-

Table 7: Anonymization experiments for the proposed methods. * Set tree depth = 7, and the threshold value for pruning (i.e., $k - 1$) $s = 5, 10, 20, 30, 40,$ and 50 , where “original” represents the original decision tree without pruning. * Notations \mathcal{U} and \mathcal{H} denote uniqueness and homogeneity attacks, respectively; ACC denotes accuracy; N_u and N_ℓ denote numbers of users and leaves, respectively.

s ($= k - 1$)	Nursery				Loan				Adult			
	\mathcal{U}	\mathcal{H} (N_u)	\mathcal{H} (N_ℓ)	ACC	\mathcal{U}	\mathcal{H} (N_u)	\mathcal{H} (N_ℓ)	ACC	\mathcal{U}	\mathcal{H} (N_u)	\mathcal{H} (N_ℓ)	ACC
Experiment Result for Method 1												
original	0	7863.9	24	0.9370	6	3837	27	0.9809	11	1918.9	34.4	0.8545
5	0	7863	24	0.9370	0	3803.5	11	0.9815	0	1881.5	14.7	0.8120
10	0	7863	24	0.9370	0	3784.5	8.5	0.9599	0	1854.3	11.4	0.8072
20	0	7863	24	0.9370	0	3752.3	6.1	0.9469	0	1811.5	8.6	0.7954
30	0	7861	23.9	0.9363	0	3743.7	5.8	0.9294	0	1750	6.1	0.7926
40	0	7777	21.7	0.9333	0	3711.8	4.9	0.9294	0	1735.9	5.7	0.7921
50	0	7696	19.8	0.9324	0	3684.5	4.3	0.9277	0	1693.7	4.8	0.7900
Experiment Result for Method 2												
original	0	7863.9	24	0.9370	6.7	3838.3	27	0.9810	11	1918.9	34.4	0.8544
5	0	7863.9	24	0.9370	0	3763.9	7.8	0.9813	0	1269.1	10	0.8548
10	0	7860.2	23.9	0.9370	0	3648.2	5.5	0.9820	0	673.1	3.5	0.8542
20	0	7654	19.6	0.9373	0	3555.5	3.9	0.9809	0	673.1	3.5	0.8535
30	0	7601.5	18.5	0.9363	0	3169.5	3.1	0.9792	0	652	2.6	0.8535
40	0	7571.1	17.7	0.9355	0	3159.1	2.9	0.9791	0	676.1	2.2	0.8530
50	0	7528.5	17.3	0.9348	0	3108	2.7	0.9774	0	551.7	1.2	0.8527

tacks was high when the total number of feature combinations was small.

In addition, we have presented two different decision tree pruning methods. We found that, when the number of leaf samples is less than some s , to obtain stronger anonymization and high accuracy, it is better to make the parent node of a leaf into a leaf than to prune the leaf. We also confirmed that although the effect of reducing the number of records that can be attacked using homogeneity attacks can be expected to some degree, it is impossible to eliminate them *entirely*.

Extensibility: The experimental results suggest that the attacks against decision trees presented in this paper can also be applied to extended decision tree variants, e.g., randomized decision tree (Fan et al., 2003). A randomized decision tree selects features of \mathbf{x} at random, creates multiple trees, and produces a prediction result for each tree. This differs from a conventional decision tree in that the prediction is determined via a majority vote or the average value. The randomized decision tree in the literature (Fan et al., 2003) has a counter (UpdateStatistics). This information can be used to implement the attacks discussed in this paper. In addition, although there is only a single tree in the decision tree structure, a randomized decision tree utilizes multiple trees; thus, the attacks described in this paper may work more effectively.

Future Research Direction: The results of this paper demonstrate that a vulnerability similar to that of anonymization is inherent in decision trees. Historically, anonymization has progressed from pseudonymization to k -anonymization (Sweeney, 2002), l -diversity (Machanavajjhala et al., 2006), and t -closeness (Li et al., 2007). Currently, membership privacy and differential privacy are attracting increasing attention (Blum et al., 2005; Fletcher and Islam, 2017; Fletcher and Islam, 2019; Friedman and Schuster, 2010; Nojima and Wang, 2023; Patil and Singh, 2014; Stadler et al., 2022); thus, decision trees that satisfy differential privacy while maintaining sufficient accuracy will be required in the future.

ACKNOWLEDGEMENTS

This work was supported in part by JST CREST Grant Number JPMJCR21M1, and JSPS KAKENHI Grant Number JP20K11826, Japan.

REFERENCES

- Bai, X., Yao, J., Yuan, M., Deng, K., Xie, X., and Guan, H. (2017). Embedding differential privacy in decision tree algorithm with different depths. *Sci. China Inf. Sci.*, 60(8):082104:1–082104:15.

- Becker, B. and Kohavi, R. (1996). Adult. UCI Machine Learning Repository. DOI: <https://doi.org/10.24432/C5XW20>.
- Blum, A., Dwork, C., McSherry, F., and Nissim, K. (2005). Practical privacy: the SuLQ framework. In Li, C., editor, *Proceedings of the Twenty-fourth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June 13-15, 2005, Baltimore, Maryland, USA*, pages 128–138. ACM.
- Fan, W., Wang, H., Yu, P. S., and Ma, S. (2003). Is random model better? on its accuracy and efficiency. In *Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM 2003), 19-22 December 2003, Melbourne, Florida, USA*, pages 51–58. IEEE Computer Society.
- Fletcher, S. and Islam, M. Z. (2017). Differentially private random decision forests using smooth sensitivity. *Expert Syst. Appl.*, 78:16–31.
- Fletcher, S. and Islam, M. Z. (2019). Decision Tree Classification with Differential Privacy: A survey. *ACM Comput. Surv.*, 52(4):83:1–83:33.
- Friedman, A. and Schuster, A. (2010). Data mining with differential privacy. In Rao, B., Krishnapuram, B., Tomkins, A., and Yang, Q., editors, *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, July 25-28, 2010*, pages 493–502. ACM.
- Jagannathan, G., Pillaipakkamatt, K., and Wright, R. N. (2012). A practical differentially private random decision tree classifier. *Trans. Data Priv.*, 5(1):273–295.
- Li, N., Li, T., and Venkatasubramanian, S. (2007). t -Closeness: Privacy beyond k -Anonymity and l -Diversity. In Chirkova, R., Dogac, A., Özsu, M. T., and Sellis, T. K., editors, *Proceedings of the 23rd International Conference on Data Engineering, ICDE 2007, The Marmara Hotel, Istanbul, Turkey, April 15-20, 2007*, pages 106–115. IEEE Computer Society.
- Machanavajjhala, A., Gehrke, J., Kifer, D., and Venkatasubramanian, M. (2006). l -Diversity: Privacy Beyond k -Anonymity. In Liu, L., Reuter, A., Whang, K., and Zhang, J., editors, *Proceedings of the 22nd International Conference on Data Engineering, ICDE 2006, 3-8 April 2006, Atlanta, GA, USA*, page 24. IEEE Computer Society.
- Mahdi Navaei. Bank_Personal_Loan_Modelling. <https://www.kaggle.com/datasets/ngnnguythkim/bank-personal-loan-modellingcsv>.
- Nojima, R. and Wang, L. (2023). Differentially private (random) decision tree without adding noise. In Luo, B., Cheng, L., Wu, Z., Li, H., and Li, C., editors, *Neural Information Processing - 30th International Conference, ICONIP 2023, Changsha, China, November 20-23, 2023, Proceedings, Part IX*, volume 1963 of *Communications in Computer and Information Science*, pages 162–174. Springer.
- Patil, A. and Singh, S. (2014). Differentially private random forest. In *2014 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2014, Delhi, India, September 24-27, 2014*, pages 2623–2630. IEEE.
- Rajkovic, V. (1997). Nursery. UCI Machine Learning Repository. DOI: <https://doi.org/10.24432/C5P88W>.
- Slijepcevic, D., Henzl, M., Klausner, L. D., Dam, T., Kieseberg, P., and Zeppelzauer, M. (2021). k -anonymity in practice: How generalisation and suppression affect machine learning classifiers. *Comput. Secur.*, 111:102488.
- Stadler, T., Oprisanu, B., and Troncoso, C. (2022). Synthetic Data - Anonymisation Groundhog day. In Butler, K. R. B. and Thomas, K., editors, *31st USENIX Security Symposium, USENIX Security 2022, Boston, MA, USA, August 10-12, 2022*, pages 1451–1468. USENIX Association.
- Sweeney, L. (2002). k -anonymity: A model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl. Based Syst.*, 10(5):557–570.