Privacy Preserving Distributed Extremely Randomized Trees*

Amin Aminifar[1], Fazle Rabbi[1,2], Ka I Pun[1,3], and Yngve Lamo[1]
[1]Western Norway University of Applied Sciences, [2]University of Bergen, [3]University of Oslo
{amin.aminifar,fazle.rabbi,ka.i.pun,yngve.lamo}@hvl.no

ABSTRACT

Applying machine learning and data mining algorithms over data distributed in multiple sources is challenging. One complication is to perform data analysis without compromising personal information, which is a primary concern in healthcare applications. Another issue involves communication overhead incurred from the transfer of raw data from one party to others for conducting centralized data mining. In healthcare applications, we are particularly interested in running data mining algorithms over big data without disclosing sensitive information about data subjects due to privacy and legal concerns. In this paper, we consider the classification problem and show how the Extremely Randomized Trees (ERT) algorithm could be adapted for settings where (structured) data is distributed over multiple sources. We propose the Privacy-Preserving Distributed ERT approach for privacy-preserving utilization of the ERT algorithm in a distributed setting. To the best of our knowledge, this is the first application of the ERT algorithm in the distributed setting, with privacy consideration (without sharing the raw data or intermediate training values), without any loss in classification performance.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; *Distributed artificial intelligence*; *Cooperation and coordination*; **Machine learning**; *Machine learning algorithms*;

KEYWORDS

Distributed Learning, Extremely Randomized Trees, Privacy-Preserving Data Mining, Structured Data

ACM Reference Format:
Amin Aminifar[1], Fazle Rabbi[1,2], Ka I Pun[1,3], and Yngve Lamo[1]. 2021. Privacy Preserving Distributed Extremely Randomized Trees. In *Proceedings of ACM SAC Conference (SAC'21).* ACM, New York, NY, USA, Article 4, 4 pages. https://doi.org/10.1145/3412841.3442110

1 INTRODUCTION

In many real-world applications, such as in healthcare systems, data is inherently distributed over an arbitrary number of sources instead of being stored in a central database. It is not straightforward to apply data mining algorithms in situations where distributed data cannot be transferred to a central location due to communication overheads, as well as privacy concerns. Figure 1 shows one such scenario and environment for this problem. The figure illustrates a setting where hospitals need to apply data mining methods for extracting useful patterns from patients' data. Although individual hospital information systems may be able to locally store health information and perform data mining with their limited resources, it is a necessity to share health information across hospitals to fully exploit the learning capacity of the data mining techniques. However, this is a challenging task due to privacy and legal concerns. Hospitals often need to comply with privacy regulations that restrict sharing health information about patients with other parties [13, 16, 19]. A similar problem exists when the data is stored on patients' personal devices, such as mobile phones or wearable devices with limited resources [8, 21–23]. How can we utilize large amount of healthcare data stored in an arbitrary number of sources for data mining without disclosing the private information of the subjects? In this paper, we address this problem by developing a novel approach for privacy-preserving data mining over distributed (structured) healthcare information.

Traditionally, it was assumed that all sources holding part of the data may share their information with a trusted party. However, sharing sensitive data with trusted parties is not a feasible assumption in many scenarios. In order to address the privacy concern, one solution would be to perturb data and share it. However, perturbation-based solutions do not provide absolute data privacy and utility because the privacy will not be preserved if the perturbation is not sufficient and the data utility will decrease if the perturbation is not controlled precisely [4, 26]. Similarly, anonymization techniques, e.g., [1, 14, 17, 24], share an altered version of data to prevent the re-identification of data subjects [10]. Nevertheless, there is always a trade-off between data privacy and utility in these techniques [4]. Therefore, such techniques have limited applicability. Moreover, communication and computational overheads would still be a problem for the approaches we discussed above, especially when dealing with large scale data.

There exist several data mining algorithms that utilize the indirect use of raw data. One such approach is the cryptographic technique and secure multi-party computation method for conducting privacy-preserving data mining [5, 11, 25]. However, they are inefficient when dealing with big data, due to extreme communication/computation costs [26]. Other techniques have been proposed to address communication/computational overheads of the stated privacy-preserving data mining algorithms, e.g., [7, 12, 18]. These solutions provide privacy as well as efficiency w.r.t. communication and computational overheads. Nevertheless, the data mining algorithms should be modified, depending on the possibility to support applications in distributed settings, which may negatively affect the machine learning model's performance.

*Produces the permission block, and copyright information

SAC'21, March 22-March 26, 2021, Gwangju, South Korea

ACM ISBN 978-1-4503-8104-8/21/03…$15.00
https://doi.org/10.1145/3412841.3442110

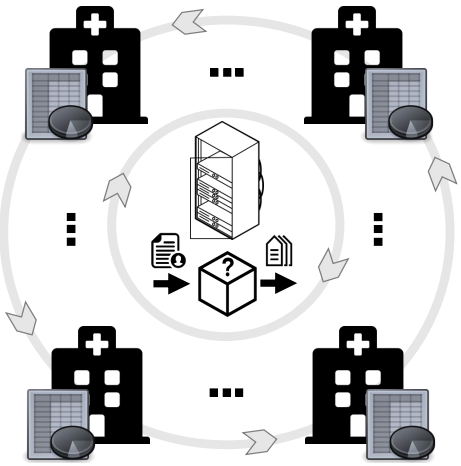

Figure 1: Overview of the environment for learning from structured data distributed over several parties

In this paper, we target the problem of learning from multiple data holders, without explicit sharing of the raw healthcare information. We assume that the learning data is horizontally partitioned, i.e., different records of data are stored on different sources. We consider the classification problem, in which each data record has one category as the target. We consider that data is structured, i.e., it can be stored in spreadsheets, and contains categorical attributes, e.g., gender or mental-disorder history, and numerical attributes, like age or frequency and duration of pathological episodes. We focus on the class of tree-based algorithms that have been shown to consistently outperform or to be on a par with the other state-of-the-art techniques when it comes to structured data [3, 15]. To learn from such horizontally-partitioned structured data, we propose privacy-preserving distributed extremely randomized trees (PPD-ERT). We first extend the ERT algorithm [9] to a distributed setting, to enable learning without explicit sharing of the raw data. We then introduce a secure aggregation technique over the distributed ERT algorithm to avoid any information leakage. We evaluate the proposed solution experimentally and compare the results against the state-of-the-art techniques.

2 PRIVACY PRESERVING DISTRIBUTED ERT

This section presents the proposed solution which is based on Extremely Randomized Trees (ERT) [9] algorithm, and discusses the procedure of learning an ensemble of decision trees based on the ERT algorithm in the discussed settings. Our main contributions w.r.t. the traditional ERT algorithm are:

- We extended ERT to the distributed setting.
- We employed a security layer by utilizing SMC techniques.

2.1 Initialization and Initiation

In the initialization phase, the mediator starts the process of leaning. The mediator initiates and mediates the overall learning process. It begins with sharing the global and personal random seeds with data holder parties. The mediator will then repeatedly learn decision trees based on our privacy-preserving distributed ERT algorithm.

In the ERT algorithm, we have two parameters of randomness for learning a weak classifier. First, we need to randomly select several attributes, among all possible data attributes, for selecting candidate decision nodes at every step of building our decision tree. Second, a random splitting point for every attribute in the candidate decision node is required. The data holder parties and the mediator are required to have the same candidate decision nodes at every step of learning a decision tree. Therefore, instead of making these randomly-made candidate decision nodes in the mediator and sharing them with all parties for further tasks, we share a common random seed that all parties, including the mediator, use to locally generate these candidate decision nodes. Since all parties use a common random seed, i.e., the global random seed, they generate the same candidate decision nodes at every step, without any communication overhead. Moreover, for the secure aggregation of partial results, described further in Section 2.3, each data holder party and the mediator share a personal random seed. These random seeds are exclusive and private for each data holder party. Only the data holder party and the mediator have access to this personal random seed.

2.2 The Process of Learning One Decision Tree

The learning of a decision tree based on the privacy-preserving distributed ERT algorithm is a recursive procedure, which is executed top-down, starting from the root and ending at the leaves.

The mediator generates the candidate decision nodes, for building the decision tree, after receiving the results from the data holder parties to select the best candidate among them. The candidate decision nodes are generated randomly based on the global random seed. Several attributes from the dataset's possible attributes are selected for candidate decision nodes. Then, each candidate decision node's splitting points are selected. We assume that all parties already have the possible categories and ranges for each attribute.

To decide the candidate decision nodes for each branch, the mediator requires the collective outcome of the classification with candidate decision nodes from all data holders on all their data. By having the combination of data record labels for each branch, the mediator can both decide if we require a leaf at that place or if we should calculate the information gain. The mediator sends a request to the first data holder party and waits for receiving the aggregated result from the last party through secure aggregation described in Section 2.3. The aggregated results are two vectors, one for each branch, representing the combination of data record labels after classification with each candidate decision node.

Having the aggregated results, the mediator determines if a decision node is required for that place in the tree. If all the labels are the same or if the number of received labels is less than the threshold parameter in the ERT algorithm, the mediator puts a label on that place, as a leaf. Otherwise, the mediator calculates the information gain of each candidate decision node based on the results from data holder parties. It then selects the candidate decision node with the highest information gain and informs all parties about this. The selected node will be used to build the tree at the mediator. After selecting the best decision node candidate, the same process is performed for each of the branches.

This process leads to learning a single decision tree; we repeat the same process for having an ensemble of decision trees.

2.3 Secure Aggregation of Results From Parties

We adopt an SMC technique in our proposed distributed ERT algorithm to avoid sharing the vectors representing the combination of the data record labels for each candidate decision node and each branch in each data holder party. In addition to the provided privacy by not sharing the raw values of data attributes, which is by construction, adoption of the SMC technique for aggregating the partial results from data holder parties contributes to privacy preservation. In an extreme example, suppose our data has one sensitive attribute in it, e.g., having previously conducted transgender surgery, and each data holder party has only one record on it. Then, sharing the partial results from one party, i.e., the vectors representing the combination of data record labels for one candidate decision node, can reveal sensitive information. If the candidate decision node is "whether the record falls into the transgender branch or not," the mediator can infer if that individual with the specified record has conducted transgender surgery. Therefore, to avoid such vulnerabilities, we adopt an SMC technique for aggregating the partial results from the data holder parties. We consider privacy among collaborating parties, but we assume no active external adversaries.

We now describe the proposed technique. The mediator shares a personal random seed with each data holder party through secure communication, to avoid sending and receiving exclusive random numbers between the mediator and each party.

Then, in the process of learning a decision tree, the mediator sends the request for secure aggregation to the first party. The party makes calculations described earlier and obtains two resulting vectors for each decision node. Afterwards, the party generates random integer masks based on its personal random seed and adds it to the results from the previous step. If the data holder party receives partial results vectors from the previous data holder party, then it also aggregates those values to the calculated vector in the previous step. Eventually, the party passes its outcome to the next party or mediator if that party is the last one.

Finally, the mediator receives the masked aggregated results from the last party. Since the mediator has the personal random seeds, it generates the same random masks as generated on the data holder parties. Then, the mediator subtracts those random masks from the received masked aggregated result. At this step, without sharing the partial information about data labels by each data holder party, the mediator has the aggregated vectors representing the combination of data record labels for each branch of each candidate decision node for all parties.

3 EVALUATION AND DISCUSSION

In this section, we evaluate our proposed approach w.r.t. classification performance, scalability and overhead, and privacy criteria [2]. We compare our approach with [7] since, similar to our approach, it is a tree-based method, employing SMC techniques for secure aggregation of partial results, to address classification problems in scenarios where data is horizontally partitioned.

Table 1: Comparison of Classification Performance

		Distributed Approaches		Centralized Approaches	
Dataset	*Metric*	*PPD-ERT*	*Distributed ID3* [7]	*ERT*[9]	*ID3*[20]
Multiple Features	*Accuracy*	98.3%	88%	98.3%	93.5%
	F1-Score	98.3%	*Not Reported*	98.3%	93.5%
Nursery	*Accuracy*	98.1%	95.7%	98.1%	99.5%
	F1-Score	95.3%	*Not Reported*	95.3%	79.2%

First, the privacy-preserving distributed ERT algorithm basically breaks the task of the centralized ERT algorithm into several parts distributed on different nodes but does not introduce any negative impact on performance by construction. Secondly, the SMC technique adopted to introduce privacy does not change the result of aggregation as opposed to the existing differential privacy techniques. The resulting vectors, representing the combination of record labels for each branch, aggregated securely by the described SMC technique, yields the same results as aggregation without adopting any SMC techniques. Therefore, the classification performance of our privacy-preserving distributed ERT remains the same as the centralized ERT. However, the proposed approach in [7] suffers from a decline in classification performance caused by its underlying learning algorithm, i.e., the ID3 algorithm.

We now evaluate the classification performance of our proposed approach. Similar to [7], we utilize Multiple Features and Nursery datasets [6] and use 2/3 of the data for learning and the rest for the test. We adopt the F1-Score and accuracy as our classification performance metrics. The accuracy of the proposed approach in [7] is also reported here for comparison. For the Multiple Features dataset, since the number of records for each class is the same, the accuracy is a proper metric for evaluating the classification performance. However, since the Nursery dataset is imbalanced, the accuracy is not a reliable measure; hence, we also consider the F1-Score. Table 1 compares the classification performance of our approach PPD-ERT with the one in [7], with their best setting where 128 parties are collaborating. Moreover, the classification performance of centralized versions of ERT [9] and ID3 [20] algorithms, i.e., the underlying standard learning algorithms for PPD-ERT and the proposed approach in [7], are also provided for comparison.

In our experiments, on the PPD-ERT, and the ERT algorithm, we learn an ensemble of 25 decision trees. For the number of candidate decision nodes' parameter in the algorithm, we used 5-fold cross-validation for the model selection (concerning classification performance measured by the F1-Score). For the Multiple Features dataset, we generate 65 candidate decision nodes (proportionate to 10% of the number of features in the dataset) at every step, and for the Nursery dataset, eight candidate decision nodes (proportionate to 90% of the number of features in the dataset) are generated. The results in Table 1 for PPD-ERT, ERT, and ID3 are the average of 10 rounds of learning and evaluation. In the case of the Multiple Features dataset, the PPD-ERT algorithm outperforms the proposed technique in [7] by 10.3%. For the Nursery dataset, the PPD-ERT outperforms the method in [7] by 2.4%. However, in the case of the Nursery dataset, since the data is imbalanced, using the accuracy metric may lead to misleading results. When considering the F1-Score metric, which is a reliable metric even for imbalanced datasets, the simple ID3 algorithm that always outperforms the

Table 2: Communication Complexity of Different SMC Approaches

Approach	Party	Communication		Total Communication
		Send	Receive	
NOSMC	Data Holders	1	0	(n-1)×1 + 1×(n-1)
	Mediator	0	n-1	
PPD-ERT	All	1	1	n×(1+1)

proposed method in [7] shows 16.1% lower performance compared to the PPD-ERT approach.

We now discuss the privacy and overhead of our proposed approach. We adopt an SMC technique to avoid direct sharing of the vectors, representing the combination of record labels for each candidate decision node, with other parties and the mediator. We compare the communication overhead and privacy of our adopted SMC technique against the NOSMC approach. Table 2 presents the communication overhead of both methods. In the table, n is the number of parties, and the communication overheads in the table are for one round of secure aggregation.

In the first approach (NOSMC), no SMC technique is adopted, and all the values are directly shared with the mediator and known to it. This approach has the lowest possible communication cost and one colluding parties, and is considered as a baseline. On the one hand, our approach's communication overhead is from order $O(n)$, which is from the same order as NOSMC. On the other hand, our technique offers interesting privacy features compared to NOSMC. Firstly, it takes three parties (or two parties in case the data holder party is the first or last) for collusion. Secondly, one of the colluding parties needs to be the mediator, which can be assumed as an honest party in many scenarios. In the case of a secret value revelation, we know that the mediator has been involved in the collusion.

We demonstrate that our proposed PPD-ERT approach provides a solution to classification of structured data distributed over multiple sources with privacy-preservation consideration. In particular, our approach does not negatively affect the classification performance compared to the centralized ERT algorithm.

4 CONCLUSION

In this paper, we have extended the ERT algorithm to ensure privacy in a distributed setting, where data is held by several parties. In our proposed algorithm, on the one hand, the data holders do not share their data values with other parties for learning. On the other hand, the required partial-information from data holders, the combination of labels after splitting their records by candidate decision nodes, which has a low risk of revealing important information, is securely aggregated to minimize the likelihood of inference of sensitive information by an adversary. We have evaluated our proposed algorithm extensively and demonstrated its efficiency in terms of prediction performance, scalability and overheads, as well as privacy. We show that our approach outperforms the state-of-the-art distributed ID3 by up to 10.3% in terms of classification performance while ensuring scalability and privacy.

ACKNOWLEDGMENTS

This research is supported by INTROducing Mental health through Adaptive Technology (INTROMAT) project. The paper is partially supported by SIRIUS: Centre for Scalable Data Access.

REFERENCES

[1] A Aminifar, Y Lamo, KI Pun, and F Rabbi. 2019. A Practical Methodology for Anonymization of Structured Health Data. In *Proceedings of the 17th Scandinavian Conference on Health Informatics*.
[2] E Bertino, D Lin, and W Jiang. 2008. A survey of quantification of privacy preserving data mining algorithms. In *Privacy-preserving data mining*. Springer.
[3] T Chen and C Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*.
[4] JS Davis and O Osoba. 2019. Improving privacy preservation policy in the modern information age. *Health and Technology* (2019).
[5] W Du and Z Zhan. 2002. Building decision tree classifier on private data. (2002).
[6] D Dua and C Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml
[7] F Emekçi, OD Sahin, D Agrawal, and A El Abbadi. 2007. Privacy preserving decision tree learning over multiple parties. *Data & Knowledge Engineering* (2007).
[8] F Forooghifar, A Aminifar, and D Atienza. 2019. Resource-aware distributed epilepsy monitoring using self-awareness from edge to cloud. *IEEE transactions on biomedical circuits and systems* (2019).
[9] P Geurts, D Ernst, and L Wehenkel. 2006. Extremely randomized trees. *Machine learning* (2006).
[10] ISO 25237:2017 2017. *Health informatics — Pseudonymization.* Standard. International Organization for Standardization, Geneva, CH.
[11] M Kantarcioglu. 2008. A survey of privacy-preserving methods across horizontally partitioned data. In *Privacy-preserving data mining*. Springer.
[12] J Konečnỳ, HB McMahan, D Ramage, and P Richtárik. 2016. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527* (2016).
[13] P Kumar and HJ Lee. 2012. Security issues in healthcare applications using wireless medical sensor networks: A survey. *sensors* (2012).
[14] N Li, T Li, and S Venkatasubramanian. 2007. t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd International Conference on Data Engineering*. IEEE, 106–115.
[15] SM Lundberg, G Erion, H Chen, A DeGrave, JM Prutkin, B Nair, R Katz, J Himmelfarb, N Bansal, and SI Lee. 2020. From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence* (2020).
[16] SD Lustgarten, YL Garrison, MT Sinnard, and AWP Flynn. 2020. Digital privacy in mental healthcare: current issues and recommendations for technology use. *Current Opinion in Psychology* (2020).
[17] A Machanavajjhala, D Kifer, J Gehrke, and M Venkitasubramaniam. 2007. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)* (2007).
[18] HB McMahan, E Moore, D Ramage, S Hampson, et al. 2016. Communication-efficient learning of deep networks from decentralized data. *arXiv preprint arXiv:1602.05629* (2016).
[19] D Pascual, A Amirshahi, A Aminifar, D Atienza, P Ryvlin, and R Wattenhofer. 2020. EpilepsyGAN: Synthetic Epileptic Brain Activities with Privacy Preservation. In *IEEE Transactions on Biomedical Engineering*.
[20] JR Quinlan. 1986. Induction of decision trees. *Machine learning* (1986).
[21] A Saeed, FD Salim, T Ozcelebi, and J Lukkien. 2020. Federated Self-Supervised Learning of Multi-Sensor Representations for Embedded Intelligence. *IEEE Internet of Things Journal* (2020).
[22] D Sopic, A Aminifar, A Aminifar, and D Atienza. 2017. Real-time classification technique for early detection and prevention of myocardial infarction on wearable devices. In *2017 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE.
[23] D Sopic, A Aminifar, A Aminifar, and D Atienza. 2018. Real-time event-driven classification technique for early detection and prevention of myocardial infarction on wearable systems. *IEEE transactions on biomedical circuits and systems* (2018).
[24] L Sweeney. 2002. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* (2002).
[25] J Vaidya. 2008. A survey of privacy-preserving methods across vertically partitioned data. In *Privacy-preserving data mining*. Springer.
[26] J Vaidya, B Shafiq, W Fan, D Mehmood, and D Lorenzi. 2013. A random decision tree framework for privacy-preserving data mining. *IEEE transactions on dependable and secure computing* (2013).