Journal of Database Management
Volume 33 • Issue 2

Random Forest Algorithm Based on Linear Privacy Budget Allocation

Yanling Dong, North China University of Science and Technology, China
Shufen Zhang, North China University of Science and Technology, China*
Jingcheng Xu, North China University of Science and Technology, China
Haoshi Wang, North China University of Science and Technology, China
Jiqiang Liu, Beijing Jiaotong University, China

ABSTRACT

In the era of big data with exponential growth in data volume, how to reduce data security issues such as data leakage caused by machine learning is a hot area of recent research. The existing privacy budget allocation strategies are usually only suitable for data applications in specific spaces and cannot meet users' personalized needs for privacy budget allocation. Therefore, a linear privacy budget allocation strategy is proposed. The strategy assigns each layer a linearly increasing privacy budget from the root of the decision tree to the bottom by adjusting the coefficient or constant term. Combining this strategy with the random forest algorithm, a random forest algorithm based on linear privacy budget allocation (DiffPRF_linear) is formed. Experimental results show that the proposed algorithm can realize uniform, arithmetic, and geometric privacy budget allocation policy effects and can also achieve better classification effects than the former, which not only meets the needs of users to protect private data personalized but also maintains high classification accuracy.

KEYWORDS

Data Security, Differential Privacy, Machine Learning, Privacy Budget Allocation, Privacy Protection, Random Forest

INTRODUCTION

Currently, big data technology is more deeply and widely used (Wang et al., 2013; Tao et al., 2013), and artificial intelligence technologies such as machine learning and deep learning are also accelerating development (Wang et al., 2019). To obtain more convenient digital services, a "portrait" based on the personal data of group users is inevitable. However, the "portrait" in different fields has different requirements for data collection, combined with the different security levels of data storage by data collectors and the strong background knowledge of attackers, which makes data leakage, data selling

DOI: 10.4018/JDM.309413 *Corresponding Author

and other data security problems (Wang et al., 2019) pose significant risks to users, developers and society (Siau et al., 2020). Aiming at reducing the risk of privacy disclosure (Dumbill et al., 2013; Meng et al., 2013), researchers have proposed privacy protection technologies such as anonymization technology (Sweeney, 2002; Machanavajjhala et al., 2007; Li et al., 2007; Xiao et al., 2007), cryptography technology (Clifton et al., 2002; Rothe, 2002; Jiang et al., 2006; Ishai et al., 2006), differential privacy technology (Dwork, 2006; Dwork, 2008; Dwork et al., 2009) and blockchain technology (Turesson et al., 2021). Among them, differential privacy technology is currently the mainstream privacy protection technology. It is difficult for attackers to use background knowledge to predict the sensitive attributes of individuals who add noise compared to the primary data, so as to control the disclosure of personal privacy in a small range (Li et al., 2012; Xiong et al., 2104).

Adjusting the magnitude of the privacy budget can offer different degrees of protection for data and affect data availability (Mcsherry et al., 2007). Dwork et al. (2012) first proposed a uniform allocation strategy to distribute the privacy budget, i.e., the privacy protection budget is evenly distributed to each layer of the tree structure. However, such an allocation scheme will lead to more waste of the privacy budget, low use efficiency of the privacy budget, and the allocation of the privacy budget is relatively fixed and unadjustable. Cormode et al. (2012) proposed a geometric budget allocation strategy, namely increasing the privacy budget geometrically from the root node, but the query error may be proportional to the number of leaves in the query. Wang et al. (2016) put forward an adaptive privacy budget allocation strategy to provide privacy protection statistics which is published on unlimited timestamps, but this strategy can only be applied to specific data. Wang (2019) proposed a p-series privacy budget allocation method for infinite attacks that attackers may launch. However, when there are too many iterations, the privacy budget tends to be zero, resulting in poor data availability. In the real-time location protection of users, Li et al. (2017) proposed a privacy budget allocation strategy with adaptive adjustment according to the distribution of underlying data. However, when the total number of users is too large, the privacy budget allocated on each timestamp is too small, resulting in the addition of too much noise. Wang et al. (2018) proposed the privacy budget allocation method of arithmetic sequences and geometric sequences, but it is vulnerable to the influence of parameters and the privacy budget allocation method is not flexible enough. Ke et al. (2021) proposed a dynamic privacy allocation mechanism on the basis of the different output contributions of different input features to the model, while the accuracy of the model is low and the balance between model utility and privacy protection is not reached.

Blum et al. (2005) first combined differential privacy technology with a single learner, proposed the SuLQ-based ID3 algorithm, and applied Laplace to add noise. However, due to the excessive noise added, the accuracy of the classification results decreased significantly compared with that without noise. Since a single individual learner can no longer satisfy people's desire to have a stable model with good performance in all aspects, Breiman (2001) proposed the random forest algorithm. Patil et al. (2014) first combined the random forest algorithm with differential privacy and used the ID3 algorithm to construct the subtree of the random forest. However, this method can only handle discrete attributes, and can only preprocess continuous features when facing continuous features. The DiffPRFs algorithm proposed by Mu et al. (2016) eliminates the dependence on discrete datasets and selects split points and split attributes through the exponential mechanism. However, each iteration requires two calls to the exponential mechanism, which consumes too much privacy budget. Li et al. (2020) further proposed the RFDPP-Gini algorithm, using the CART classification tree as a subtree and invoking the exponential mechanism only once when processing continuous features to improve the use efficiency of the privacy protection budget. However, since the privacy budget allocation scheme selected is uniform allocation, the utilization rate of the privacy budget is low.

The authors present DiffPRF_linear, a stochastic forest algorithm based on linear privacy budget allocation. Through a linear allocation strategy, users can more flexibly adjust the coefficient or constant term to allocate the privacy budget of each layer, and can also realize the effect of uniform,