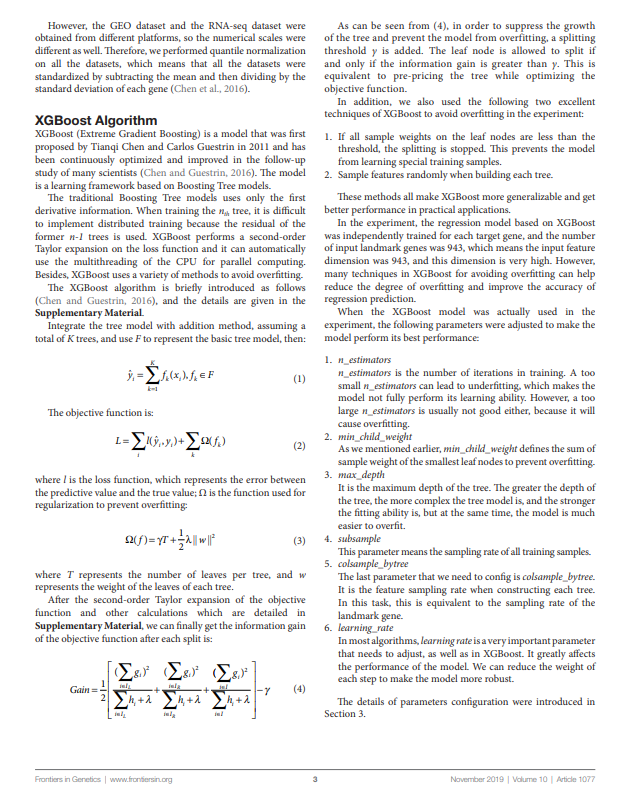
XG-BOOST

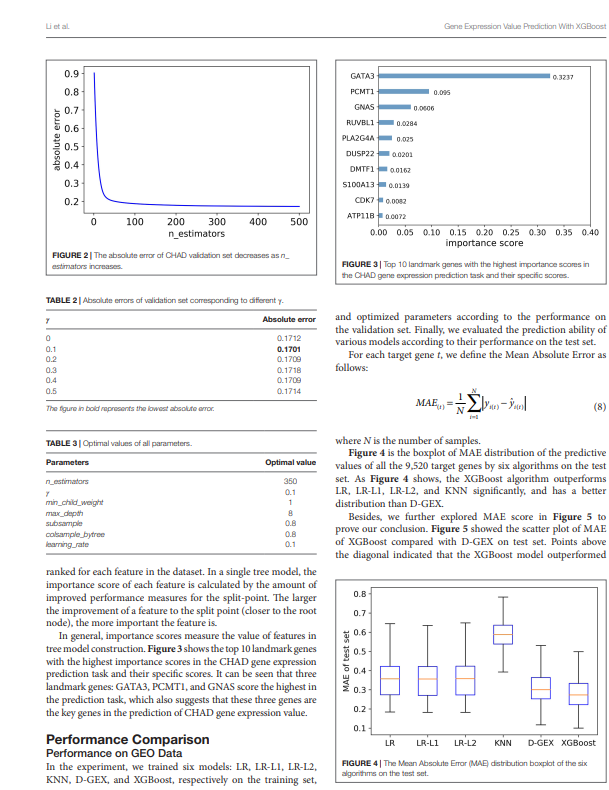
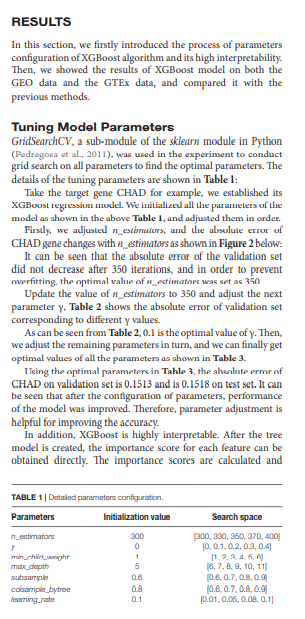
Gene expression profiling has been widely used to characterize cell status to reflect the health of the body, to diagnose genetic diseases, etc. In recent years, although the cost of genome-wide expression profiling is gradually decreasing, the cost of collecting expression profiles for thousands of genes is still very high. Considering gene expressions are usually highly correlated in humans, the expression values of the remaining target genes can be predicted by analyzing the values of 943 landmark genes. Hence, we designed an algorithm for predicting gene expression values based on XGBoost, which integrates multiple tree models and has stronger interpretability. We tested the performance of XGBoost model on the GEO dataset and RNA-seq dataset and compared the result with other existing models. Experiments showed that the XGBoost model achieved a significantly lower overall error than the existing D-GEX algorithm, linear regression, and KNN methods. In conclusion, the XGBoost algorithm outperforms existing models and will be a significant contribution to the toolbox for gene expression value prediction.

Introduction

Characterizing gene expression patterns in cells under various conditions is an important problem. Gene expression profiling is a vital biological tool commonly used to capture the response of cells to disease or drug treatments Although the cost of gene expression profiling is steadily decreasing in recent years, it is still very expensive when dozens or hundreds of samples need to processed. Inspired by this, many scholars have suggested that the expression value of the landmark gene can be used to predict the expression value of the target gene, which will greatly reduce the cost of the gene expression profiling. In 2016, Yifei Chen et al. proposed the D-GEX algorithm based on Back Propagation neural network (Chen et al., 2016), in which 943 landmark genes correspond to 943 input units, and 9,520 target genes correspond to 9,520 output units. However, the prediction accuracy of this algorithm still has a large room for improvement. Besides, deep network has poor interpretability, and for each target gene, we cannot know which landmark genes have much greater impact on its expression. Last but not the least, deep network needs to read all the data into the memory at the time of training, and therefore, the algorithm is prone to occupy excessive memory in actual use, and has high demand for GPU too. In addition to deep network, some researchers also used linear regression, KNN and other classical algorithms for target gene expression prediction (Chen, 2014), but the prediction results of these algorithms were less accurate. Among the Boosting Tree models, XGBoost (Chen and Guestrin, 2016) has a very strong expansion and flexibility. It integrates multiple tree models to build a stronger learner model. Furthermore, XGBoost is characterized by its ability to automatically use the multithreading of the CPU for parallel computing, which can speed up the calculation. Based on the above research background, we proposed a new gene expression value prediction algorithm based on XGBoost, and established a regression prediction model for each target gene independently. The results showed that the XGBoost algorithm significantly improved the prediction accuracy, which is superior to D-GEX, LR, KNN, and other algorithms. It also had better predictive ability and generalization ability. Lastly, the XGBoost algorithm had stronger interpretability than other algorithms.

XG-Boost



results