DOI:10.3969/j.issn.1005-202X.2024.06.019

医学人工智能

基于遗传算法优化LightGBM算法的医院微服务平台安全运维管理系统的流量智能化检测

卓一超,郝海宾
温州医科大学附属第一医院信息处, 浙江 温州 325000

【摘要】为提升医院微服务平台下运维管理系统的数据检测效率,提出一种新的数据检测算法。该算法以平台数据的多元特征为基础,构建运维管理系统的整体框架。通过结合遗传算法的参数寻优能力和LightGBM算法的快速检测能力,实现对运维管理系统的流量数据的有效检测。为了验证模型的有效性,增加了对照实验。实验结果表明本方法在流量智能化检测中表现最优,其准确率(0.981 0)、查全率(0.68)以及F1值(0.77)均优于传统方法。

【关键词】微服务平台;运维管理系统;遗传算法;LightGBM
【中图分类号】R197.32;R318 【文献标志码】A 【文章编号】1005-202X(2024)06-0788-05

Intelligent flow detection in hospital microservices platform security operation and maintenance management system based on genetic algorithm optimized LightGBM algorithm

ZHUO Yichao, HAO Haibin
Department of Medical Information, the First Affiliated Hospital of Wenzhou Medical University, Wenzhou 325000, China

Abstract: A novel data detection algorithm is proposed to improve the data detection efficiency of the operation and maintenance management system for the hospital microservices platform. Based on the multiple characteristics of the platform data, the algorithm constructs the overall framework of the operation and maintenance management system. By combining the parameter optimization ability of genetic algorithm and the rapid detection ability of LightGBM algorithm, the effective detection of the flow data in the operation and maintenance management system is realized. The effectiveness of the model is verified through a control test, and the results show that the proposed method performs the best in intelligent flow detection, achieving accuracy of 0.981 0, recall rate of 0.68 and F1 value of 0.77 which are all higher than those of the traditional methods.

Keywords: microservices platform; operation and maintenance management system; genetic algorithm; LightGBM

前 言

医院微服务平台安全运维管理系统的意义在于确保医院微服务平台的安全、稳定和高效运行。微服务架构在医院信息系统中的应用越来越广泛,它将整个系统拆分成多个小型、自治的服务单元,各个服务单元可以独立开发、部署和扩展。医院的业务中台落地时需要有很多的技术组件支撑,这些不同技术领域的技术组件就组成了技术中台。业务中台大多采用微服务架构,以保障各个系统的隔离性、可用性,可有效应对高频海量业务访问场景,所以技术中台会有比较多的微服务相关的技术组件。在开发中台各个微服务的时候,所涉及到各种共性技术能力的需求,整合和包装了云基础设施,以及在此基础上建立的各种技术中间件(如微服务、分布式缓存、消息队列、搜索引擎等),并在此基础上建设和封装了简单易用的能力接口。温州医科大学附属第一医院使用k8s、容器技术、Spring Cloud、消息队列、Redis、对象存储、Loki、Prometheus和CI/CD等构建应用程序基础设施。这些技术是现代应用程序开发所必需的基础设施,能够提供可用性、可扩展性和易于管理的应用程序环境。然而,微服务架构也带来了一些挑战,特别是在安全和运维管理方面。医院信息系统中包含大量敏感的医疗数据和个人隐私信

【收稿日期】2023-12-14
【基金项目】温州市基础性科研项目(Y20211158);省部级5G+医疗健康应用试点项目(2020No.78);浙江省智慧医疗工程技术研究中心项目(2016E10011)
【作者简介】卓一超,硕士,研究方向:软件工程与项目管理,E-mail:zyc@wzhospital.cn

息,医院微服务平台安全运维管理系统有着重要的意义。安全运维管理系统能够对微服务平台进行实时监控和安全审计,及时发现并阻止潜在的安全威胁和攻击,确保患者数据和系统的安全性[1-2];微服务架构中,一个系统由多个微服务组成,服务之间相互依赖。安全运维管理系统可以对微服务的运行状态进行监测,及时发现服务的异常和故障,快速进行故障定位和处理,确保系统的稳定运行[3];安全运维管理系统可以引入自动化运维工具和技术,简化运维流程,提高运维效率,降低人工操作带来的风险;通过对微服务平台的运行状态进行实时监测和分析,安全运维管理系统可以发现性能瓶颈和资源浪费问题,进行性能优化,提高系统的响应速度和吞吐量。总体而言,医院微服务平台安全运维管理系统的意义在于保障医院信息系统的安全性、稳定性和高效性,提高医疗服务的质量和效率,保障患者数据的安全和隐私保护,它是医院信息化建设中不可或缺的重要组成部分。

对医院信息系统的流量检测是医院微服务平台安全运维管理系统中的一项关键的管理和优化措施,它利用数据采集和分析技术,实时监控医院信息系统内的流量情况,为医院管理者和医护人员提供宝贵的决策支持和资源配置建议。随着信息技术的快速发展,医院信息系统流量检测技术也得到了快速的进步和应用。一种常用的医院信息系统流量检测算法是基于机器学习的方法。这些算法使用历史流量数据进行训练,构建模型预测正常流量行为,并能够检测与预期行为明显不符的异常流量。为此,王泽川等[4]利用k最近邻算法构建的自学习智能化模型,能够监测和分析信息系统的响应时延和安全事件,并生成预警和告警事件,该模型能够自动学习和适应数据的特征,以实现智能化的异常检测和警报功能。Wang等[5]讨论传统的基于支持向量机(Support Vector Machine, SVM)的异常检测算法对于高度不平衡的数据集表现不佳的问题,提出一种新的基于不平衡数据的SVM异常检测算法,该算法优于过采样技术和现有的几种不平衡算法。治晓隆等[6]提出一种基于主成分分析和tabu Tabu搜索(PCA-TS)决策树分类的异常检测方法,该方法对高维特征进行简化,选择适合于PCA-TS算法分类的最优特征子集,然后利用基于半监督学习的检出率高、错误率低的决策树进行分类和检测。Dey等[7]研究基于OpenFlow控制器流量入侵检测系统的方法,该方法克服传统网络的管理挑战,取得良好的效果。Yousef等[8]提出一种基于流量数据的两个阶段神经网络的入侵检测系统,用于对网络流量中的攻击进行检测和分类,这些算法在监测数据流时可以根据训练模型对数据进行分类和预测,从而发现异常流量和事件。另外,一些基于统计方法的算法也被应用于医院信息系统流量检测,如基于时间序列分析的方法,可以通过分析流量的周期性、趋势和季节性变化检测异常。此外,流量的频率分布和统计特征也可以用于检测异常流量。

以上方法均验证了机器学习方法在医院信息系统流量智能化检测上的有效性。然而,当前基于医院信息系统的流量异常检测方法在检测精度和计算时间方面存在一些不足,往往难以同时实现准确性和实时性的检测目标。本文使用LightGBM算法可以有效地解决机器学习算法在流量异常检测任务上所面临的问题。经过多年的研究与发展,它在机器学习领域得到了广泛应用,并在许多任务中取得了优秀的性能。Gan等[9]将LightGBM模型用于预测哥伦比亚河下游的水位,结果表明LightGBM模型可以实现较高的预测精度。Liu等[10]建立一个基于LightGBM算法的模型,以识别和分类3种主要类型的严重对流天气(即冰雹、短期暴雨、对流阵风),模型评价表明LightGBM模型在训练集(2011~2017年)和测试集(2018年)中表现良好。但LightGBM拥有大量的超参数,手动调整这些参数非常耗时且困难,而遗传算法(Genetic Algorithm, GA)可以用于搜索最佳的超参数配置,如学习率、树的深度、叶子节点数等,以及其他的模型参数,通过使用GA可以高效地找到最优的参数组合,从而提高模型的性能。本文为进一步提高模型的检测性能和泛化能力,提出一种基于GA优化LightGBM算法的医院微服务平台安全运维管理系统的流量智能化检测方法。

1 算法简介

1.1 GA

GA是一种受到自然进化理论启发的优化算法,它模拟生物进化中的遗传机制和自然选择过程,用于解决搜索和优化问题。GA是一种全局搜索算法,通常用于找到问题的最优解或近似最优解。描述如下:

Step1:将问题的解表示成一个个体,通常称为染色体。染色体是由一串基因组成的,每个基因对应解的一个特征或参数;

Step2:随机生成一组初始个体,构成初始种群;

Step3:定义适应度函数评估每个个体的优劣程度,适应度函数衡量染色体对问题的解的适应程度,是GA进行自然选择的依据;

Step4:根据适应度函数的评估结果,选择一部分

个体作为“父代”，用于产生下一代的个体；对“父代”个体进行交叉操作，产生新的个体，即“子代”；对“子代”进行变异操作，以增加种群的多样性；

Step5：根据一定规则，将新的“子代”替换掉部分“父代”，形成新的种群；

Step6：重复进行选择、交叉、变异和替换操作，直到达到设定的终止条件（如达到一定迭代次数或找到满足要求的解）。

通过不断地迭代和进化，GA逐渐找到较优的解决方案。由于GA具有并行性和全局搜索能力，它在处理复杂问题和寻找复杂优化目标时表现出色，并且在很多领域都得到了成功应用，如优化问题、函数优化、机器学习、组合优化等。

1.2 LightGBM算法

LightGBM是一种高效的梯度提升树算法，由微软亚洲研究院开发。它是GBDT（Gradient Boosting Decision Tree）算法的一种改进版本，在性能和效率上有很大的提升。LightGBM在机器学习领域得到广泛的应用，并在许多数据科学竞赛中取得了优异的成绩。LightGBM算法具有以下优势：（1）LightGBM具有较快的训练速度和较低的内存消耗，它采用直方图算法和基于直方图的决策树算法，使得在大规模数据集上也能高效运行。（2）LightGBM采用基于直方图的决策树算法，而不是传统的按层分裂，这样可以更快地找到最优的分裂点，提高训练速度。（3）LightGBM采用Leaf-wise的生长策略，它在每次选择分裂时，选择能使目标函数下降最大的叶子节点，而不是Level-wise生长的固定深度。这样可以得到更深的树，提高模型的准确性。（4）LightGBM支持特征并行化训练，在多核CPU上可以并行处理不同特征的直方图，提高训练速度。（5）LightGBM对连续特征进行离散化处理，构建直方图近似连续特征的值分布，从而减少内存开销，并提高模型的训练速度。基于以上优势，LightGBM在大规模数据集和复杂特征的场景下表现出色。

定义训练样本集为 $M=\left[\left(z_1,y_1\right),\cdots,\left(z_n,y_n\right)\right]$，则该算法的函数为：

$$\begin{aligned} O_{\text{object}}^{(k)} &= \sum_{i=1}^{n} l\left(y_i,\hat{y}_i^{(k)}\right)+\sum_{i=1}^{k}\Omega\left(f_i\right) \\ &= \sum_{i=1}^{n} l\left(y_i,\hat{y}_i^{(k-1)}+f_k\left(x_i\right)\right)+\sum_{i=1}^{k}\Omega\left(f_i\right) \end{aligned} \tag{1}$$

其中，y_i为标签真实值，$\hat{y}_i$为对应学习结果，$f(k)$表示第k棵树模型，$\Omega(f_i)$为正则项。

接下来，通过损失函数计算对应学习结果$\hat{y}_i$与标签真实值y_i之间的差异：

$$l\left(y_i,\hat{y}_i\right)=\log\left(1+e^{\hat{y}_i}\right)+y_i\log\left(1+e^{-\hat{y}_i}\right)-y_i\log\left(1+e^{\hat{y}_i}\right) \tag{2}$$

损失函数二阶泰勒公式展开式：

$$O_{\text{object}}^{(k)}=\sum_{i=1}^{n}\left[l\left(y_i,\hat{y}^{(k-1)}\right)+g_i f_k\left(x_i\right)+\frac{1}{2}h_i f_k^2\left(x_i\right)\right]+\Omega\left(f_k\right) \tag{3}$$

其中第i个样本的损失函数的一阶导数$g_i=\partial l\left(y_i,\hat{y}^{(k-1)}\right)$，第$i$个样本的损失函数的二阶导数$h_i=\partial^2 l\left(y_i,\hat{y}^{(k-1)}\right)$。目标函数进一步转变为：

$$\tilde{O}_{\text{object}}^{(k)}=\sum_{j=1}^{K}\frac{-\left(\sum_{i\in I_j}g_i\right)^2}{2\left(\sum_{i\in I_j}h_i+\lambda\right)}+\gamma\cdot K \tag{4}$$

其中，I_j表示叶子节点j的集合。

2 基于GA优化LightGBM算法的流量智能化检测方法

在当前医学领域，LightGBM模型广泛应用，然而该模型需要设置许多参数，因此优化这些参数变得至关重要。为了找到最优的模型参数组合，可以使用GA。GA通过利用已有数据，反复迭代正向输出和参数修正过程，有效地避免了陷入局部最优解的困境。基于此，能够得到最佳的模型参数组合，从而提升LightGBM模型的性能和预测能力。通过GA进行参数优化，可以更好地发掘参数空间，提高模型的泛化能力和适应性，为医学应用带来更好的效果。

本文提出的基于GA优化LightGBM算法的医院微服务平台安全运维管理系统的流量智能化检测方法，其主要流程如图1所示。

根据图1可得，本文所提出的方法可分为两个阶段，第一阶段为GA优化模型参数，在获得最优模型参数后，方法转入第二阶段，基于最优参数组合进行预测。本文中需要优化的模型参数见表1。

3 仿真及结果分析

为进一步验证本研究提出的GA-LightGBM模型的流量检测性能，建立医院微服务平台安全运维管理系统的流量检测数据集“Decetion”。同时将GA-LightGBM模型的性能与下面6种分类模型进行比较，即Quinlan[11]提出的决策树（Decision Tree）模型；Vapnik[12]提出的SVM模型；Duda等[13]提出的朴素贝叶斯（Naive Bayes）模型；Cover等[14]提出的最近邻（k-NearestNeighbor, kNN）算法；Chen等[15]提出的XGBoost（eXtreme Gradient Boosting）模型以及Ke等[16]提出的LightGBM模型。

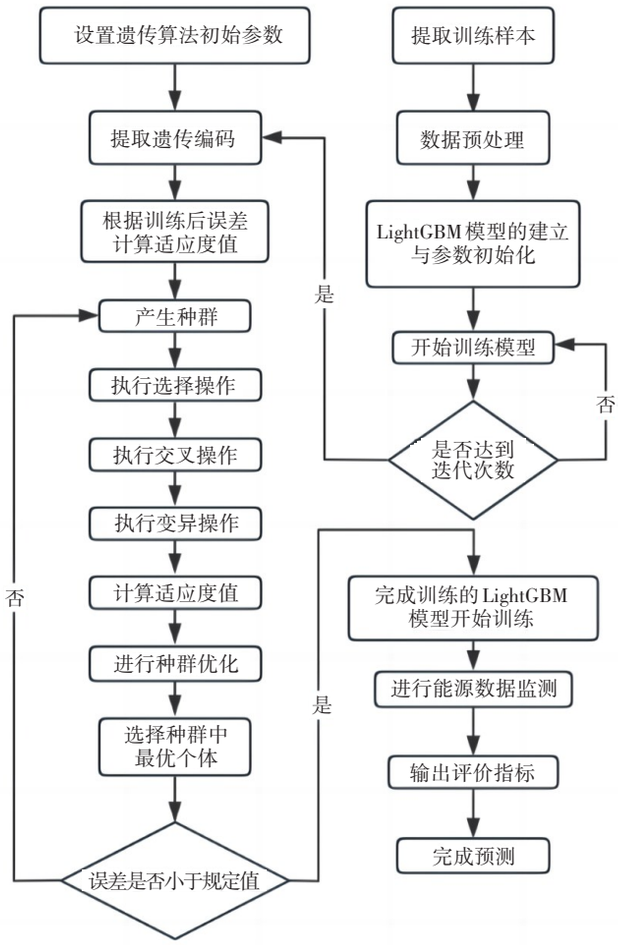

图1 算法实现流程图

Figure 1 Flowchart of the algorithm implementation

3.1 数据集简介

数据集“Decetion”是医院微服务平台安全运维管理系统的流量检测数据，该数据集待采集了一定时间范围的数据后，再根据流量检测数据的预测选择指标的基本准则，本文选择了6个检测的特征变量，包括4个连续型变量以及2个离散型变量。根据流量检测数据的标签的判定，将超出安全阈值的数据定义为流量检测异常数据。经过简单的数据预处理，共存在1 752个流量检测样本。

3.2 模型评价准则

本文中评价分类器的评价指标为准确率、查全率和F1值。准确率指被分为正例的样本中实际为正例的样本比例，查全率指被正确判定的正例占总的正例的比例。为了平衡准确率和查全率之间的关系，本文引入综合衡量指标F1值作为分类器评价指标之一。F1值是准确率和查全率的调和平均，它能够综合考虑分类器的准确性和覆盖率，从而更全面地评估分类器的性能。通过使用这3个指标，可以更全面地评估分类器在给定的测试数据集上的表现。

评价矩阵是一种可视化工具，通过评价矩阵可以直观地查看分类器在不同类别上的预测准确性和误判情况，进而对分类器的性能进行全面评估。评价矩阵通常包含4个重要的指标：真正例(TP)、假正例(FP)、真反例(TN)和假反例(FN)的数目。

本文中使用准确率、查全率和F1值作为综合评价指标，借助评价矩阵可以更好地理解这些指标。

表1 待优化的参数基本信息

Table 1 Basic information of the parameters to be optimized

参数	参数意义	取值范围
n_estimators	Number of boosted trees to fit	100~2 000
learning_rate	Boosting learning rate	0.01~0.30
max_depth	Maximum tree depth for base learners	2~30
min_child_weight	Minimum sum of instance weight (hessian) needed in a child (leaf)	0~10
num_leaves	Maximum tree leaves for base learners	10~100
random_seed	Random number seed	1~10 000

$$准确率 = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$查全率 = \frac{TP}{TP + FN} \tag{6}$$

$$F1 = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{7}$$

在上述3项指标中，准确率用于度量模型整体预测性能，即对所有样本的正确分类情况进行评估。查全率则用于衡量模型对于反例样本的识别情况，即正确识别负例样本的能力。F1值是一个综合评价指标，综合考虑了模型对于正例和反例样本的识别情况，使得能够更全面地评估分类器的性能。通过这3个指标，能够对分类器在给定的测试数据集上的表现有更深入的了解。评价矩阵则提供一种可视化方式，可以直观地查看分类器在不同类别上的预测结果，进而对其性能进行细致分析。

3.3 实验结果与分析

为了评价提出的基于GA优化LightGBM算法的医

院微服务平台安全运维管理系统的流量智能化检测方法的性能，本文将该模型与Decision Tree模型、SVM、Naive Bayes模型、kNN算法、XGBoost模型以及LightGBM模型6种分类模型进行比较。首先将数据集“Decetion”按照标签类别划分为正类和负类两个部分，然后随机抽取两个部分中的30%的样本合并得到测试集，剩余的样本即为训练集；其次7种分类模型基于相同的训练集、测试集进行实验，最后得出相应的评价指标。不同算法的结果对比见表2。

表2 不同算法结果比较

Table 2 Comparison of the results of different algorithms

算法	准确率	查全率	F1值
Decision Tree	0.941 1	0.48	0.46
SVM	0.950 6	0.22	0.32
Naive Bayes	0.952 5	0.44	0.49
kNN	0.948 7	0.41	0.45
XGBoost	0.952 5	0.41	0.47
LightGBM	0.954 4	0.41	0.48
GA-LightGBM	0.981 0	0.68	0.77

根据表2的结果，可以得出以下结论：(1)本文提出的方法相对于Decision Tree、SVM、Naive Bayes、kNN、XGBoost、LightGBM等6种算法，在准确率方面表现最优。这意味着该方法能够有效地提高流量智能化检测的准确性。(2)就查全率指标而言，GA-LightGBM模型明显优于其他算法。这表明该模型在辨别反例样本方面表现出色，有效提升了对异常流量的识别能力。在实际研究中，流量异常情况的准确识别是分类任务的重点，因此该方法更适用于流量智能化检测。(3)F1值综合考虑了各个模型对正例和反例样本的识别能力。相较于其他6种分类模型，本文提出的方法得到的F1值最高，这意味着GA-LightGBM模型在检测流量正常样本和流量异常样本方面具有更强的能力。综上所述，基于GA优化的LightGBM模型在流量智能化检测任务中表现出色，具有更高的准确率、查全率以及F1值，为流量异常识别和分类任务带来了显著改进。

4 结 论

本文提出一种用于医院微服务平台安全运维管理系统的流量智能化检测的方法，该方法利用GA优化LightGBM模型。首先，通过GA的参数寻优能力，不断迭代更新，得到LightGBM模型的最优参数组合。接着，利用这些最优参数构建LightGBM模型，并将其用于医院微服务平台安全运维管理系统的流量检测数据进行预测。研究结果表明，通过结合GA和LightGBM的优化模型(GA-LightGBM)能够有效检测流量异常数据，从而提高医疗服务的质量和效率，并保障患者数据的安全和隐私保护。该方法的优势在于简单易行，可有效地应用于医院信息系统的流量智能化检测中。通过利用GA对LightGBM模型进行优化，本研究进一步增强了医院微服务平台安全运维管理系统的流量智能化检测能力。这样的智能化检测系统有望为医院信息系统带来更高的安全性、稳定性和高效性，从而为医疗服务提供更可靠的支持。同时，该方法还能为其他领域的流量异常检测问题提供有价值的参考和借鉴。

【参考文献】

[1] 胡铁柱, 魏钰博, 雷立华. 基于人工智能的安全运维管理系统设计[J]. 现代电子技术, 2020, 43(10): 171-175.
Hu TZ, Wei YB, Lei LH. Design of security operation and maintenance management system based on artificial intelligence [J]. Modern Electronics Technology, 2020, 43(10): 171-175.

[2] 邵新民, 宫有为, 胡晓鹏, 等. 基于大数据的网络安全运维管理系统[J]. 计算机工程与应用, 2018, 54(14): 136-141.
Shao XM, Gong YW, Hu XP, et al. Network security operation and maintenance management system based on big data[J]. Computer Engineering and Applications, 2018, 54(14): 136-141.

[3] Jamil F, Qayyum F, Alhelaly S, et al. Intelligent microservice based on blockchain for healthcare applications[J]. Comput Mater Contin, 2021, 69(2): 2513-2530.

[4] 王泽川, 马存宁, 刘玉泉, 等. 基于kNN算法的流量智能化模型在医院信息系统安全运维管理中的应用[J]. 中国医疗设备, 2021, 36(6): 132-135.
Wang ZC, Ma CN, Liu YQ, et al. Application of traffic intelligent model based on kNN algorithm in the safe operation and maintenance management of hospital information system [J]. China Medical Devices, 2021, 36(6): 132-135.

[5] Wang GP, Yang JX, Li R. Imbalanced SVM-based anomaly detection algorithm for imbalanced training datasets[J]. ETRI J, 2017, 39(5): 621-631.

[6] 冶晓隆, 兰巨龙, 郭通. 基于主成分分析禁忌搜索和决策树分类的异常流量检测方法[J]. 计算机应用, 2013, 33(10): 2846-2850.
Ye XL, Lan JL, Guo T. Network anomaly detection method based on principle component analysis and tabu search and decision tree classification[J]. Journal of Computer Applications, 2013, 33(10): 2846-2850.

[7] Dey SK, Rahman MM. Effects of machine learning approach in flow-based anomaly detection on software-defined networking [J]. Symmetry, 2020, 12(1): 7.

[8] Yousef A, Goran K, Slavko G, et al. Flow-based anomaly intrusion detection system using two neural network stages[J]. Comput Sci Inf Syst, 2014, 11(2): 601-622.

[9] Gan M, Pan SQ, Chen YP, et al. Application of the machine learning LightGBM model to the prediction of the water levels of the lower Columbia river[J]. J Mar Sci Eng, 2021, 9(5): 496.

[10] Liu XW, Duan HX, Huang WB, et al. Classified early warning and forecast of severe convective weather based on LightGBM algorithm [J]. Atmos Clim Sci, 2021, 11(2): 284-301.

[11] Quinlan JR. Induction of decision trees[J]. Mach Learn, 1986, 1(1): 81-106.

[12] Vapnik VN. The nature of statistical learning theory[M]. New York: Springer, 1995.

[13] Duda RO, Hart PE, Stork DG. Pattern classification[M]. 2nd ed. New York: Wiley, 2001.

[14] Cover T, Hart P. Nearest neighbor pattern classification[J]. IEEE Trans Inf Theory, 1967, 13(1): 21-27.

[15] Chen TQ, Guestrin C. XGBoost: a scalable tree boosting system[C]// In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: Association for Computing Machinery, 2016: 785-794.

[16] Ke GL, Meng Q, Finley T, et al. LightGBM: a highly efficient gradient boosting decision tree [C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2017: 3149-3157.

(编辑：黄开颜)