文章编号：2095-2163(2022)08-0165-06 中图分类号：P413 文献标志码：A

三种机器学习算法在回归应用中的对比分析

蔡 明[1,2]，孙 杰[1,2]，李培德[1]，鲍 清[1]
（1 湖北省气象信息与技术保障中心，武汉 430074；2 暴雨监测预警湖北重点实验室，武汉 430074）

摘 要：梯度提升方法(Gradient Boosting Machine，GBM)是一种经事实证明并被广泛应用的集成学习方法。许多成功的机器学习解决方案都是使用 XGBoost 及其衍生算法实现的。本文针对 XGBoost、LightGBM、CatBoost 这 3 种梯度方法及原理进行简要介绍，通过实验对比了 3 种方法在回归应用中的效率，并将 3 种梯度方法应用于由 Kaggle 的 NYC Taxi fares 数据集构建的车费回归预测模型中。实验结果表明，LightGBM 在使用不同数量特征和样本的情况下，模型训练速度和最终预测精度均优于 XGBoost 和 CatBoost。

关键词：GBM；XGBoost；LightGBM；CatBoost；回归预测

Comparative analysis of three machine learning algorithms in regression application

CAI Ming[1,2]，SUN Jie[1,2]，LI Peide[1]，BAO Qing[1]
（1 Hubei Meteorological Information and Technical Support Center，Wuhan 430074，China；
2 Hubei Key Laboratory for Heavy Rain Monitoring and Warning Research，Wuhan 430074，China）

【Abstract】Gradient boosting machine（GBM）is a proven and widely used ensemble learning method. Many successful machine learning solutions are implemented using XGBoost and its derivative algorithms. This paper briefly introduces the gradient methods and principles of XGBoost，LightGBM and CatBoost，compares the efficiency of the three methods in regression application through experiments，and applies the three gradient methods to the fare regression prediction model constructed by Kaggle′s NYC taxi fares dataset. The implementation indicates that LightGBM is faster and more accurate than CatBoost and XGBoost using variant number of features and samples.

【Key words】GBM；XGBoost；LightGBM；CatBoost；regression prediction

0 引 言

目前，对于中等数据集来说，与人工神经网络(Artificial Neural Network，ANN)相比，boosting 方法则有着较为明显优势。相对来说，boosting 的训练时间会更短，参数调整时也不会耗费太多时间。

Boosting 是一种集成学习策略，致力于从各种弱分类器中生成准确的分类器。通过划分训练数据，并使用每个部分来训练不同的模型或用一个具有不同设置的模型来实现，最后再用多数票将结果组合在一起。AdaBoost 是 Freund 等人[1]提出的第一个用于二元分类的有效 boosting 方法。当 AdaBoost 进行第一次迭代时，所有记录的权重相同，但在下一次迭代中，却会为错误分类的记录赋予更高的权重，模型迭代将继续、直到构造出有效的分类器。AdaBoost 发布后不久，就有研究发现，即使迭代次数增加，模型误差也不会变大[2]。因此，AdaBoost 模型十分适用于解决过拟合问题。近些年来，学者们基于梯度提升决策树(Gradient Boosting Decision Tree，GBDT)提出了 3 种基于决策树的有效梯度方法，分别是：XGBoost、CatBoost 和 LightGBM。这些方法均已成功应用于工业界、学术界和竞争性机器学习的研究中[3]。

1 相关算法介绍

1.1 GBDT

梯度提升树是一种利用加法模型与前向分步算法实现学习的优化过程。当损失函数为平方误差损失函数和指数损失函数时，每一步的优化较为简单。但对一般损失函数而言，往往每一步优化并不容易。针对这一问题，Freidman 提出了梯度提升(Gradient Boosting)算法。Gradient Boosting 是 Boosting 中的一

基金项目：湖北省气象局科技发展基金课题(2020Z07)；湖北省气象局重点科研项目(2022Z04)。

作者简介：蔡 明(1987-)，男，硕士，工程师，主要研究方向：气象装备保障、气象信息技术；孙 杰(1981-)，男，硕士，高级工程师，主要研究方向：气象装备保障、气象信息技术；李培德(1991-)，男，硕士，工程师，主要研究方向：气象装备保障；鲍 清(1981-)，男，学士，助理工程师，主要研究方向：气象装备保障。

通讯作者：孙 杰 Email：3037998@qq.com

收稿日期：2022-03-06

类算法,设计思想参考自梯度下降法,基本原理是根据当前模型损失函数的负梯度信息,来训练新加入的弱分类器,并将训练好的弱分类器以累加的形式结合到现有模型中。采用决策树作为弱分类器的 Gradient Boosting 算法被称为 GBDT,有时也称为 MART(Multiple Additive Regression Tree)。

梯度提升方法以分段方式构造解,并通过优化损失函数来解决过拟合问题。例如:假设有一个定制的基学习器 $h(x,\theta)$(如决策树)和一个损失函数 $\psi(y,f(x))$。若直接估计参数会十分困难,因此在每次迭代时使用迭代模型。每次迭代模型都将被更新、并重选一个新的基学习器 $h(x,\theta_t)$,其中增量可表示为:

$$g_t(x)=E_y\left[\frac{\partial\psi(y,f(x))}{\partial f(x)}\,|\,x\right]_{f(x)=f^{t-1}(x)} \tag{1}$$

这样就可以将难解的优化问题转化为常用的最小二乘优化问题,即:

$$(\rho_t,\theta_t)=\operatorname{argmin}_{\rho,\theta}\sum_{n=1}^{N}\left[-g_t(x_n)+\rho h(x_n,\theta)\right]^2 \tag{2}$$

这里,对梯度提升算法的实现步骤可做阐释表述如下。

步骤 1 令 f_0 为常数;

步骤 2 对于 $m=1$ 到 M 有:

1:利用式(1)计算 $g_m(x)$;

2:训练函数 $h(x,\theta_m)$;

3:利用式(2)寻找最优 ρ_m;

4:更新函数 $\hat{f}_m=\hat{f}_{m-1}+\rho_m h(x,\theta_m)$。

步骤 3 结束。

该算法从一片叶子开始,接着将针对每个节点和每个样本优化学习速率[4-6]。

1.2 XGBoost

XGBoost(eXtreme Gradient Boosting)是一种高度可扩展、灵活且通用的梯度提升工具[7],其设计目的在于正确使用资源,并克服以往梯度提升算法的局限性。XGBoost 和其它梯度提升算法的主要区别是,XGBoost 使用了一种新的正则化技术,控制过拟合现象的产生。因此,在模型调整期间,XGBoost 会更快、更健壮。正则化技术是通过在损失函数中添加一个新项来实现的,此处的数学公式可写为:

$$L(f)=\sum_{n=1}^{N}L(\hat{y}_n,y_n)+\sum_{m=1}^{N}\Omega(\delta_m) \tag{3}$$

其中,$\Omega(\delta)=\alpha|\delta|+0.5\beta\|\omega\|^2$;$|\delta|$ 为树枝的数量;ω 为每片叶子的值;$\Omega(f)$ 为正则化函数。

XGBoost 使用了新的增益函数,相应的函数形式具体如下:

$$G_j=\sum_{i\in I_j}g_i$$

$$H_j=\sum_{i\in I_j}h_i$$

$$Gain=\frac{1}{2}\left[\frac{G_L^2}{H_L+\beta}+\frac{G_R^2}{H_R+\beta}-\frac{(G_L+G_R)^2}{H_L+H_R+\beta}\right]-\alpha \tag{4}$$

其中,$g_i=\partial_{\hat{y}_i}L(\hat{y}_i,y_i)$、$h_i=\partial^2_{\hat{y}_i}L(\hat{y}_i,y_i)$;$G$ 是右子节点的分数;H 是左子节点的分数;$Gain$ 是在没有新子节点情况下的分数[8]。

文中,对 XGBoost 基本核心算法流程拟做阐释如下。

(1)不断地添加树,并不断地进行特征分裂来生长一棵树。每次添加一个树,其实是学习一个新函数 $f(x)$,去拟合上次预测的残差。

(2)当训练完成得到 k 棵树,需要预测一个样本的分数,即根据这个样本特征,在每棵树中会求得对应的一个叶子节点,每个叶子节点就对应一个分数。

(3)基于此,只需将每棵树对应的分数加起来,就得到了该样本的预测值。

1.3 LightGBM

为了提高 GBDT 算法效率、避免 XGBoost 的缺陷、并且能够在不损害准确率的条件下加快 GBDT 模型的训练速度,微软研究团队于 2017 年 4 月开发了 LightGBM[9-10]。LightGBM 在传统 GBDT 算法上进行了如下优化:

(1)基于 Histogram 的决策树算法。一个叶子的直方图可以由其父亲节点直方图与其兄弟直方图做差得到,在速度上可以提升一倍。

(2)单边梯度采样(Gradient-based One-Side Sampling,GOSS)。使用 GOSS 可以减少大量只具有小梯度的数据实例,使其在计算信息增益时只利用余下的具有高梯度的数据即可。相比 XGBoost 而言,既遍历所有特征值,也节省了不少时间和空间上的开销。GOSS 算法从减少样本的角度出发,排除大部分小梯度的样本,仅用剩下的样本计算信息增益,这样做的好处是在减少数据量和保证精度上取得平衡。

(3)互斥特征捆绑(Exclusive Feature Bundling EFB)。使用 EFB 可以将许多互斥的特征绑定为一个特征,这样达到了降维的目的。

(4)带深度限制的 Leaf-wise 叶子生长策略。大多数 GBDT 工具使用低效的按层生长(level-

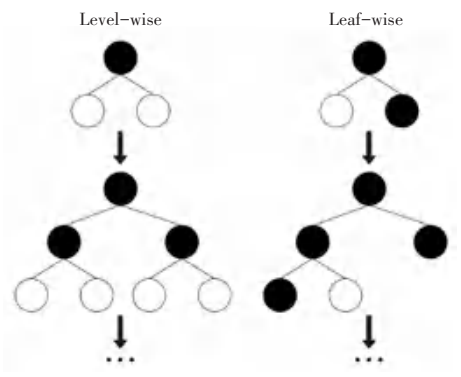
wise）的决策树生长策略，且由于不加区分地对待同一层的叶子，带来了很多额外开销。实际上很多叶子的分裂增益较低，没必要进行搜索和分裂。LightGBM 使用了带有深度限制的按叶子生长（leaf-wise）算法，在分裂次数相同的情况下，Leaf-wise 可以降低误差，得到更好的精度。并且，还能做到：

① 直接支持类别特征（Categorical Feature）；

② 支持高效并行；

③ Cache 命中率优化。

上述优化使得 LightGBM 具有更好的准确性、更快的训练速度、以及大规模处理数据能力，同时还能支持 GPU 学习的优点。按层生长与按叶子生长的设计示意如图 1 所示。



（a）按层生长　　（b）按叶子生长

图 1　按层生长与按叶子生长示意图

Fig. 1　Schematic diagram of level-wise growth and leaf-wise growth

1.4　CatBoost

CatBoost 是 Yandex 在 2017 年提出的开源的机器学习库[11]，同前面介绍的 XGBoost 和 LightGBM 类似，依然是在 GBDT 算法框架下的一种改进算法，是一种基于对称决策树（oblivious trees）算法的 GBDT 框架，不仅参数少、准确性高，还能支持类别型变量，高效合理地处理类别型特征（Categorical features）也是其主要亮点及优势。由其名称就可以看出，CatBoost 是由 categorical 和 boost 组成，并改善了梯度偏差（Gradient bias）及预测偏移（Predictionshift）问题，提高了算法准确性和泛化能力[12]。

CatBoost 可以利用各种统计上的分类特征和数值特征的组合，将分类值编码成数字，并通过在当前树的新拆分处，使用贪婪方法解决特征组合的指数增长问题。同均值编码类似，重点是通过以下步骤防止过拟合：

（1）将记录随机划分为子集。

（2）将标签转换为整数的同时，将分类特征转化为数字特征，研究求得的数学公式为：

$$avgTarget = \frac{countinClass + prior}{totalCount + 1} \tag{5}$$

其中，*countinClass* 是给定分类特征在目标中的个数；*totalCount* 是之前对象的个数；*prior* 由初始参数指定[13-14]。

与 XGBoost、LightGBM 相比，CatBoost 的创新点体现在如下方面：

（1）嵌入了将类别型特征自动处理为数值型特征的创新算法。先对 categorical features 做一些统计，计算某个类别特征（category）出现的频率，此后加上超参数，生成新的数值型特征（numerical features）。

（2）Catboost 使用了组合类别特征，可以用到特征之间的联系，极大地丰富了特征维度。

（3）采用排序提升的方法对抗训练集中的噪声点，这就避免了梯度估计的偏差，进而解决预测偏移的问题。

（4）采用完全对称树作为基模型。

2　前期准备

2.1　数据集

选择 Kaggle 比赛中的 NYC Taxi fares 数据集作为 3 种模型对比实验的数据集，以此来对比 3 种算法的性能。数据集共有1 108 477条数据，数据集的前 5 行数据样本见表 1。特征变量数目为 8，目标特征为 fare_amount。

表 1　初始数据集快照

Tab. 1　Snapshot of initial dataset

特征	*Pickup datetime*	*Pickup longitude*	*Pickup latitude*	*Dropoff longitude*	*Dropoff latitude*	*Passenger count*	*Fare amount*
1	2013-01-29 12:26:00	-73.99	40.74	-73.98	40.76	1	9.0
2	2011-06-09 00:53:00	-73.98	40.72	-73.98	40.73	3	5.7
3	2015-04-19 22:21:12	-73.99	40.75	-74.01	40.74	1	8.0
4	2013-01-05 21:36:00	-73.99	40.74	-73.98	40.74	1	5.5
5	2009-04-21 22:59:27	-73.97	40.75	-73.99	40.74	1	7.3

在对特征变量进行处理时,将 *Pickup datetime* 拆分生成新特征变量年、月、星期、年积日、时;通过 *Pickup longitude*、*Pickup latitude*、*Dropoff longitude*、*Dropoff latitude* 和 NYC 内机场经纬度坐标,计算乘车距离 Distance 和到各个机场的距离作为新的特征变量。同时对数据集进行处理,去除 *Passenger count* ≥ 5 或记录为空的数据。

最终,将经过预处理和特征工程加工的数据集按照 7∶3 的比例划分为训练集和测试集。

2.2 实验设计

为了从性能表现、效率等方面对比最具代表性的 3 种基于 GBDT 的研发算法在回归应用中的情况,文中将按照以下步骤进行实验:

(1)使用相同的初始参数训练 XGBoost、CatBoost、LightGBM 算法的基准模型。

(2)使用超参数自动搜索模块 GridSearch CV 训练 XGBoost、CatBoost 和 LightGBM 算法的调整模型。

(3)从训练和预测时间、预测得分两方面比较算法性能的表现情况。

3 实验结果对比分析

3.1 预测精度对比

为了研究不同数据样本量对模型性能的影响,分别按照全部、1/2、1/5 和 1/10 的比例,从样本数据集中随机抽取样本形成新的样本集。对新的样本集,按照 7∶3 的比例划分训练集和测试集,从模型预测精度和训练、预测用时等方面,对比 3 种算法的回归预测性能。

本文使用均方根误差 *RMSE* 对模型的预测精度进行评价。均方根误差的数学定义的公式表述可写为:

$$RMSE = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (y_o^i - y_e^i)^2} \tag{6}$$

3 种模型回归预测的 *RMSE* 见表 2。表 2 中,XGBoost、LightGBM、CatBoost 代表建立的基准模型,XGBoost_CV、LightGBM_CV、CatBoost_CV 代表在基准模型基础上,经过网格搜索和交叉验证后的优化模型。观察表 2 可以看出,3 种算法经过参数调优后的 *RMSE*,相比各自基准模型的 *RMSE* 都有所降低,说明参数优化提高了模型的预测精度。随着样本规模的降低,3 种算法的 *RMSE* 皆有不同程度的增长,说明样本规模的减小,降低了模型的预测精度。但是,CatBoost 算法在样本规模由总样本数目的 1/5 降至 1/10 时,模型预测结果的 *RMSE* 并没有出现增长。说明样本规模降低至总样本数目的 1/5 后,CatBoost 对样本规模的降低已不再敏感,样本规模与模型预测精度的具体联系有待进一步研究。

表 2 模型预测精度 *RMSE*

Tab. 2 Prediction accuracy of the models *RMSE*

样本规模	XGBoost	LightGBM	CatBoost	XGBoost_CV	LightGBM_CV	CatBoost_CV
N	3.38	3.26	3.43	3.12	2.84	3.29
N/2	3.44	3.35	3.50	3.20	2.99	3.36
N/5	3.46	3.39	3.55	3.19	3.00	3.36
N/10	3.48	3.43	3.54	3.20	3.08	3.38

由此可见,LightGBM 在基准模型和优化模型上都比其它 2 种算法的 *RMSE* 要小,说明 LightGBM 算法在实验数据集上的预测效果优于其它 2 种算法。

3.2 运行时间对比

通过记录 3 种模型训练和预测用时,进行 3 种模型的运行用时对比,对比结果见表 3。从表 3 中也可以看出,对于同一模型,使用网格搜索交叉检验模型的运行用时远高于其基准模型,这是由于网格搜索和交叉检验操作用时较多。同时,从表 3 中也可以看出,样本规模和模型运行时间成正比,模型样本规模越大,训练和预测用时越多。不同模型间进行对比时,LightGBM 无论是基准模型、还是经过网格搜索交叉检验后的优化模型,在运行用时上都是最少,CatBoost 模型的运行时间次之,XGBoost 模型运行耗时最多,这与前文论述中对 3 种模型的特性介绍相符。

表 3　模型运行时间

Tab. 3　Running time of the models　　s

样本集	样本规模	XGBoost	LightGBM	CatBoost	XGBoost_CV	LightGBM_CV	CatBoost_CV
训练	N	55.362 4	2.048 7	5.128 9	2 473.152 4	127.442 3	186.142 8
	$N/2$	30.965 0	1.458 1	3.925 6	1 185.534 9	64.578 3	76.611 0
	$N/5$	10.043 8	0.591 6	1.804 9	429.886 4	31.008 4	34.739 4
	$N/10$	5.159 8	0.344 8	1.241 2	238.761 0	18.053 0	22.091 4
测试	N	0.271 8	0.291 8	0.077 9	0.712 6	0.059 0	0.687 6
	$N/2$	0.181 9	0.187 9	0.047 9	0.315 8	0.038 9	0.346 8
	$N/5$	0.057 9	0.073 9	0.025 8	0.151 9	0.001 0	0.097 9
	$N/10$	0.046 0	0.038 0	0.012 0	0.087 9	0.008 9	0.078 0

3.3　参数重要性评价

通过比较 3 种模型的 feature_importances_属性，研究这些属性中哪些对模型的预测影响最大，对比结果如图 2～图 4 所示。由图 2～图 4 分析可知，虽然 3 种模型中的各个变量重要性排序不尽相同，但订单距离 *distance*、*jfk* 机场订单距离 *to_ jfk*、订单年份 *year* 和乘客下车时的经度 *dropoff_longitude* 的变量重要性均排名前 4，说明无论是采用哪种模型，这 4 个变量均是决定模型预测效果的关键变量。4 个变量中，订单距离 *distance*、*jfk* 机场订单距离 *to_ jfk* 和订单年份 *year* 均是通过特征工程从原始数据集中生成的变量，这也说明对原始数据集进行特征工程加工是提升模型训练效果的一种有效手段。

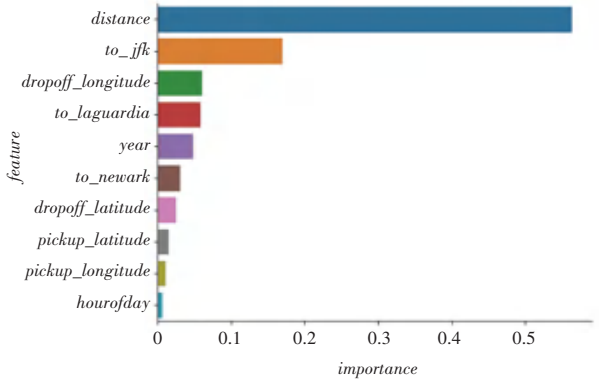


图 2　使用 XGBoost 模型的特征重要性排序图

Fig. 2　Ranking diagram of feature importance using XGBoost model

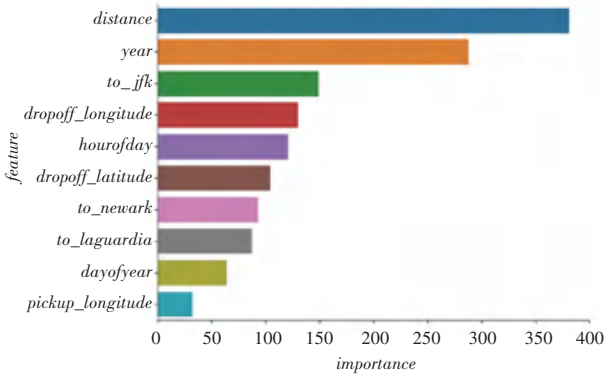


图 3　使用 LightGBM 模型的特征重要性排序图

Fig. 3　Ranking diagram of feature importance using LightGBM model

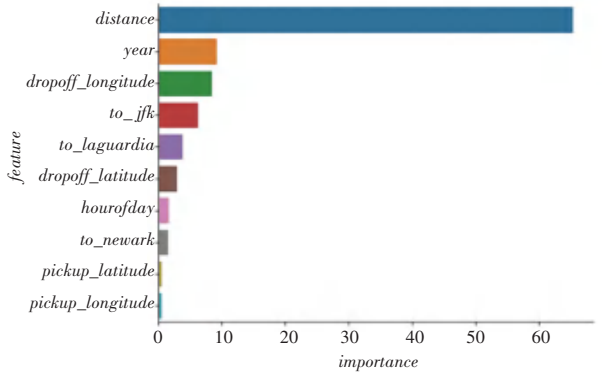


图 4　使用 CatBoost 模型的特征重要性排序图

Fig. 4　Ranking diagram of feature importance using CatBoost model

4　结束语

本文比较了 3 种最先进的梯度增强方法（XGBoost、LightGBM 和 CatBoost）的回归预测精度和运行时间。LightGBM 在实验数据集上的表现较

其他梯度增强方法要快得多,而且在同样经过超参数优化后,可以取得更好的回归预测结果;可以通过对原始数据集进行新特征生成和最佳特征选择等特征工程操作,提升模型预测性能。综合前文论述可知,由于 LightGBM 模型在预测精度和运行速度上的优势,可以作为回归应用的首选模型。

参考文献

[1] FREUND Y, SCHAPIRE R E. A decision-theoretic generalization of online learning and an application to boosting[J]. Journal of Computer and System Sciences, 1997, 55(1): 119-139.

[2] KONTSCHIEDER P, FITERAU M, CRIMINISI A, et al. Deep neural decision forests[C]//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015: 1467-1475.

[3] WANG J C, HASTIE T. Boosted varying-coefficient regression models for product demand prediction [J]. Journal of Computational and Graphical Statistics, 2014, 23(2): 361-382.

[4] DAOUD E A. Intrusion detection using a new particle swarm method and Support Vector Machines[J]. World Academy of Science, Engineering and Technology, 2013, 77: 59-62.

[5] DAOUD E A, TURABIEH H. New empirical nonparametric kernels for support vector machine classification[J]. Applied Soft Computing, 2013, 13(4): 1759-1765.

[6] DAOUD E A. An efficient algorithm for finding a fuzzy rough set reduct using an improved harmony search[J]. Modern Education and Computer Science, 2015, 7(2): 16-23.

[7] 郭颖婕, 李傲, 刘晓燕,等. 基于 XGBoost 的质量性状基因互作检测方法[J]. 智能计算机与应用, 2020, 10(03): 202-208.

[8] ZHANG Y, HAGHANI A. A gradient boosting method to improve travel time prediction [J]. Transportation Research Part C, Emerging Technologies, 2015, 58: 308-324.

[9] GUOLIN K, QI M, THOMAS F, et al. LightGBM: A highly efficient gradient boosting decision tree[J]. Advances in Neural Information Processing Systems. Long Beach, CA, USA: NIPS Foundation, 2017, 30: 3149-3157.

[10] 谭晓, 孙全明, 曲志坚. 基于多模态特征融合的个性化视频推荐方法[J]. 智能计算机与应用, 2020, 10(12): 209-213.

[11] DOROGUSH A, ERSHOV V, GULIN A. CatBoost: Gradient boosting with categorical features support [C]// Neural Information Processing Systems. Long Beach, CA, USA: NIPS Foundation, 2017: 1-7.

[12] QI M, GUOLIN K, TAIFENG W, et al. A Communication-Efficient Parallel Algorithm for Decision Tree[C]//Advances in Neural Information Processing Systems. Barcelona, Spain: NIPS Foundation, 2016, 29: 1279-1287.

[13] KLEIN A, FALKNER S, BARTELS S, et al. Fast bayesian optimization of machine learning hyperparameters on large datasets [J]. Proceedings of Machine Learning Research (PMLR), 2017, 54: 528-536.

[14] ABOOBYDA J H, TARIG M A. Developing Prediction Model of Loan Risk In Banks Using Data Mining[J]. Machine Learning and Applications: An International Journal (MLAIJ), 2016, 3(1): 1-9.

(上接第 164 页)

构设计和参数设置上仍有提升的空间,在未来的肺癌图像识别工作中,将完善肺部 CT 数据集、改善网络结构和优化实验参数。

参考文献

[1] SUN T, WANG J, LI X, et al. Comparative evaluation of support vector machines for computer aided diagnosis of lung cancer in CT based on a multi-dimensional data set[J]. Computer methods and programs in biomedicine, 2013, 111(2): 519-524.

[2] CHENG Jiezhi, CHOU Y H, HUANG C S, et al. Computer-aided US diagnosis of breast lesions by using cell-based contour grouping[J]. Radiology, 2010, 255(3): 746-754.

[3] WAY T W, SAHINER B, CHAN H P, et al. Computer - aided diagnosis of pulmonary nodules on CT scans: improvement of classification performance with nodule surface features[J]. Medical Physics, 2009, 36(7): 3086-3098.

[4] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[5] XU Jun, XIANG Lei, LIU Qingshan, et al. Stacked sparse autoencoder (SSAE) for nuclei detection on breast cancer histopathology images [J]. IEEE Transactions on Medical Imaging, 2015, 35(1): 119-130.

[6] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE, 2016: 770-778.

[7] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks [C]// Advances in Neural Information Processing Systems. Lake Tahoe, NV: NIPS, 2012: 1097-1105.