用于高血压菜谱识别的基于遗传算法的改进 XGBoost 模型

雷雪梅 谢依彤
(北京科技大学计算机与通信工程学院 北京 100083)

摘 要 文中提出用于高血压菜谱识别的基于遗传算法的改进 XGBoost(eXtreme Gradient Boosting)模型。该模型主要包括 3 个步骤:首先,对数据集进行预处理,包括缺失值补全、数据去重和特征分析;然后,使用遗传算法自适应地优化 XGBoost 模型参数;最后,根据最优参数训练高血压菜谱识别模型,并将其应用于高血压菜谱识别。结果表明,在高血压菜谱识别效果方面,采用遗传算法优化的参数优于网格搜索所得到的参数。此外,所提出的基于遗传算法的改进 XGBoost 模型在精度、回召率、F1 值和 AUC 评估指标方面具有不错的表现,优于其他 4 种(随机森林、GBDT、Bagging 和 AdaBooster)组合分类模型,且提高了菜谱识别模型的可解释性。

关键词 XGBoost,遗传算法,高血压菜谱,数据分析
中图法分类号 TP181 **文献标识码** A

Improved XGBoost Model Based on Genetic Algorithm for Hypertension Recipe Recognition

LEI Xue-mei XIE Yi-tong
(School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China)

Abstract A novel improved XGBoost (eXtreme Gradient Boosting) model based on genetic algorithm for hypertension recipe recognition was proposed. The model consists of three steps. Firstly, data pre-processing is employed to handle missing values, remove duplicate data and analyze data feature. Then, the genetic algorithm is used to optimize the parameters of XGBoost model adaptively. At last, hypertension recipe identification model is trained according to the optimal parameters. The results show that the parameters optimized by genetic algorithm performs better than grid search. Moreover, the proposed model outperforms other four models (Random forest, GBDT, Bagging and AdaBooster) over four evaluation measures: accuracy, recall rate, F1 and the area under the curve (AUC) on average, and enhances the interpretability of credit scoring model.

Keywords XGBoost, Genetic algorithm, Hypertension recipes, Data analysis

1 引言

随着经济水平的快速发展和人们对物质生活的追求越来越高,高血压逐渐成为一种高发的疾病,同时发病人群有年轻化的趋势[1]。高血压一般通过药物控制,而错误的饮食方式会削减药物对高血压的治疗效果[2],因此在饮食方面加以控制是预防和治疗高血压的有效方法。现有的高血压食疗的研究方法多为控制患者对盐和油的摄取量[3],避免食用动物肝脏等高胆固醇的食物和甜食[4]。已有的研究案例包括基于案例推理的菜谱推荐系统研究[5]、基于协同过滤算法的用户个性化营养饮食谱推荐方法研究[6]、BP(Back Propagation)神经网络对肠胃菜谱的判定研究[7]等。但是,现有方法对食物的热量、脂肪、蛋白质等的摄入量的控制不够准确,精确判定菜谱能否供高血压人群食用的研究较少[8]。

本文提出用于高血压菜谱识别的基于遗传算法的改进 XGBoost 模型,利用遗传算法良好的全局搜索和搜索灵活性的优势来弥补 XGBoost 模型多个参数调优收敛速度慢、易陷入局部最优解和正确率波动大的缺陷,并应用所得最优参数构建高血压菜谱识别模型来判定菜谱是否为均衡的饮食方案,这样有助于高血压患者降低血压、治疗疾病和提高身体素质。

2 XGBoost

XGBoost(eXtreme Gradient Boosting)[9]是灵活、可移植的最优分布式决策梯度提升库,它能克服受限的计算速度和精度,是由华盛顿大学的陈天奇在梯度提升决策树(Gradient Boosting Decision Tree, GBDT)算法的基础上提出的。XGBoost 提供的平行提升树可以快速、准确地解决许多科学问题,如 Xia 等使用贝叶斯超参数优化 XGBoost 以提高信用评分[10],张昊等在电子商务商品推荐中应用 XGBoost 算法来准确预测用户购买行为[11],叶倩怡进行了基于 Xgboost 的销售额预测研究[12],樊鹏建立了优化的 xgboost-LMT 模型来进行供应商信用评价[13],Mustapha 等用 XGBoosting 模型对生物活性分子进行预测[14],怀浩等提出基于 XGBoost 的肽碎片离子强度建模[15]等。

相对于传统的 GBDT,XGBoost 的优势主要体现在以下几个方面:

雷雪梅(1972—),女,博士,高级工程师,主要研究方向为数据分析、数据挖掘;谢依彤(1992—),女,硕士,主要研究方向为数据分析、数据挖掘,E-mail:xin9910@163.com(通信作者)。

1）传统 GBDT 以 CART（Classification and Regression Trees）为基分类器，而 XBGboost 不仅支持 CART，还支持线性分类器。

2）XGBoost 对代价函数进行了二阶泰勒展开，使用了一阶和二阶导数，同时在代价函数中加入了正则项，降低了模型的方差，使学习到的模型更加简单，避免了过拟合[10]。引入的正则项为：

$$Obj^{(t)}=\sum_{i=1}^{n}l(y_i,\hat{y}_i^{(t-1)}+f_t(x_i))+\Omega(f_t)+constant \quad (1)$$

则得到的最终目标函数只依赖于每个数据点在误差函数上的一阶导数和二阶导数：

$$Obj^{(t)}\simeq\sum_{i=1}^{n}\left[l(y_i,\hat{y}^{(t-1)})+g_if_t(x_i)+\frac{1}{2}h_if_t^2(x_i)\right]+\Omega(f_t)+constan\ t \quad (2)$$

去掉常数项并求导，令导数等于 0，可得：

$$w_j^*=-\frac{G_j}{H_j+\lambda} \quad (3)$$

$$Obj=-\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j+\lambda}+\gamma T \quad (4)$$

Obj 代表损失函数的得分，得分越小，越逼近真实值，表明树的结构越好[9-15]。

3）XGBoost 支持列抽样，这样不仅能降低过拟合，还能减少模型的计算量；同时，XGBoost 支持并行化计算，加快了模型的训练速度。

构造简单的 XGBoost 模型，它对本文所用的 23 维高维菜谱数据的处理效果较好，尤其是在多元素的识别方面，通过赋予权值比重来构建决策树模型，能够为高血压患者提供更可靠的判断；另外，相比于传统的组合树模型（随机森林、GBDT、Bagging 和 AdaBooster），XGBoost 的训练速度快，产生的误差小。

3　遗传算法

遗传算法是一种自然适应优化方法，它将待解决的问题模拟成一个生物进化的过程，通过复制、交叉、突变等操作产生下一代的解，并逐步淘汰适应度函数值低的解[16]。遗传算法同时使用多个搜索点的搜索信息，利用概率搜索技术，能避免信息点迭代陷入局部最优解；其搜索灵活性好，可以迅速得到全局最优解[17]。另外，该算法直接以目标函数值为搜索信息，确定搜索方向和搜索范围，解决目标函数无法求导或导数不存在的优化问题。遗传算法的流程图如图 1 所示。

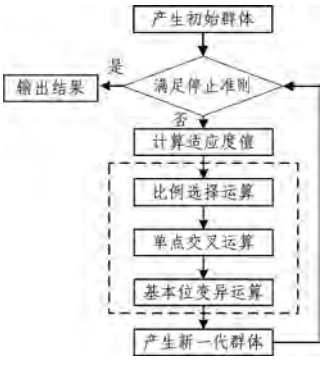

图 1　遗传算法的流程图

1）个体编码

从生物的角度看，基因型是性状染色体的内部表现，表现型是染色体决定性状的外部表现，或者说根据基因型形成个体，即基因型决定表现型。而在遗传算法中操作的对象都是基因（即 0 和 1），因此需要把输入对象编码成相应的基因序列。输入对象通常分为两种情况：离散整数值和连续浮点数值。离散取值通过二进制来编码，连续取值通过划分连续小区间来编码。根据输入对象的不同取值类型对输入对象进行编码，最终得出输入对象相应的编码基因序列。

2）适应度值计算

遗传算法以个体适应度的大小来评定各个体的优劣程度，从而决定其遗传机会的大小。适应度值高的个体有更大的机率能够生存到下一代，适应度值的计算是遗传算法的关键部分，本文计算验证集上的 AUC 值，并将其作为适应度值。

3）比例选择运算

选择的目的是从交换后的群体中选出优良的个体，使它们有机会作为父代为下一代繁殖子孙，遗传算法通过选择过程体现这一思想。选择的原则是适应性强的个体为下一代贡献的概率大，选择实现了达尔文的适者生存原则。本文通过轮盘赌算法[18]对个体进行随机选择。

4）单点交叉运算

交叉运算是遗传算法中产生新个体的主要操作过程，它以某一概率相互交换某两个个体之间的部分染色体。本文采用单点交叉的方法，首先对群体进行随机配对，其次随机设置交叉点位置，最后再相互交换配对染色体之间的部分基因。

5）基本位变异运算

首先在群体中随机选择一定数量的个体，对于选中的个体以一定的概率随机地改变串结构数据中某个基因的值。同生物界一样，在 GA 中发生变异的概率很低，一般为 0.001～0.01，变异为新个体的产生提供了机会。

6）停止准则

通常，遗传算法停止准则[19]有 3 条：①给定一个最大的遗传代数 *MAXGEN*（人为事先确定），算法迭代到 *MAXGEN* 次时停止；②给定一种下界的计算方法，当进化达到要求的偏差时，算法终止；③当监控得到的算法再进化已无法改进解的性能，即解的适应度无法再提高时，停止计算。由于本文设置的种群参数较大，种群变异数量较多，因此采用准则①，即找出适应性最强的种群，这便是最优参数的表现。

4　模型评估

4.1　交叉验证

交叉验证也称循环估计，是统计学中将数据样本切割成较小子集的实用方法。首先在一个子集上做分析，在其他子集上做后续对比分析的确认及验证。交叉验证对人工智能、机器学习、模式识别、分类器等领域的研究具有很强的指导与验证意义[20]。

只进行一次交叉验证就进行确定的最优模型存在偶然性[21]。为了使模型具有更强的稳定性，本实验采用十折交叉验证[22]来验证遗传算法中每一代群体的分类效果。

4.2　模型评估

混淆矩阵又称为可能性表格或错误矩阵，用来呈现算法性能的可视化效果，常用于监督学习[10]。混淆矩阵很明显地表明了一个类别是否被预测为另一个类别，该矩阵中的每一列代表预测类别，每一行代表实际类别，混淆矩阵中有 4 个指标：真阳性、伪阳性、伪阴性、真阴性[23]。

1)真阳性(True Positives,TP):实际是正例,被识别为正例;

2)假阴性(False Negatives,FN):实际是正例,却被识别为负例;

3)假阳性(False Positives,FP):实际是负例,却被识别为正例;

4)真阴性(True Negatives,TN):实际是负例,被识别为负例。

二分类问题混淆矩阵如图 2 所示。

		Predict	
		P	*N*
Actual	*P*	*TP*	*FN*
	N	*FP*	*TN*

图 2 二分类问题的混淆矩阵

P(Positive)表示感兴趣的值,也就是为 1 的值;*N*(Negative)表示为 0 的值。菜谱数据中 *P* 为 1 的值表示测试集中有利于高血压病人的菜谱,0 则表示不利于高血压病人的菜谱。

基于混淆矩阵,相关模型评估指标[13]如下。

准确率(Accuracy):

$$accuracy=\frac{TP+TN}{TP+TN+FN+FP}$$

误分率(Error Rate):

$$ER=\frac{FN+FP}{TP+TN+FN+FP}$$

误报率(False Acceptance Rate):

$$FAR=\frac{FP}{TP+TN}$$

即应识别为负样本而被识别为正样本的概率。

精确率(Precision):

$$P=\frac{TP}{TP+FP}$$

即在所有被识别的正样本中被正确识别的概率,也就是说,测试菜谱被判定为有利于高血压病人的菜谱中真正有利于高血压病人的菜谱的概率。

回召率(recall):

$$R=\frac{TP}{TP+FN}$$

即在所有识别为正样本中被识别为负样本的概率。

F1 值是精确值和回召率的调和均值:

$$F1=2\cdot\frac{P\cdot R}{P+R}$$ [23]

由于样本在不同类别上的不均衡分布,传统的准确率是指在给定的测试数据集中分类器正确分类的样本数与总样本数之比,其不能恰当地反映分类器的性能,比如:测试样本中有 *A* 类样本 90 个,*B* 类样本 10 个,若某个分类器简单地将所有样本都划分成 *A* 类,那么在这个测试样本中它的准确率仍为 90%,这显然是不合理的。AUC 值是一种全面度量分类模型好坏的标准[24],AUC 值是 ROC 曲线(Receiver Operating Characteristic Curve)所覆盖的区域面积,取值为 0.5~1.0,较大的 AUC 值代表模型具有较好的分类能力。

如图 3 所示,ROC 曲线的横轴是 *FAR*,纵轴是 *Recall*,每个阈值的识别结果对应一个点(*FPR*,*TPR*),当阈值最大时,所有样本都被识别成负样本,对应于右上角的点(0,0);当阈值最小时,所有样本都被识别成正样本,对应于右上角的点(1,1);随着阈值从最大值变化到最小值,*TP* 和 *FP* 都逐渐增大。好的分类模型应尽可能位于 *ROC* 曲线图的左上角[24]。

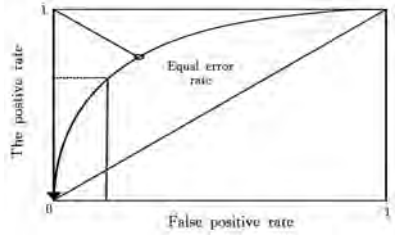

图 3 ROC 曲线图

5 基于遗传算法的改进 XGBoost 模型

由于 XGBoost 存在大量的参数,寻找最优的模型参数对高血压识别效果具有重要作用,遗传算法具有良好的全局搜索能力,能够找出全局最优的 XGBoost 参数。其次,遗传算法搜索灵活,可弥补 XGBoost 多个参数调优收敛速度慢、陷入局部最优解和正确率波动大的缺陷。另外,由于部分数据本身存在一定的误差,遗传算法鲁棒性高的优点能够使模型具有一定的稳定性,因此本文提出了一种用于高血压菜谱识别的基于遗传算法的优化 XGBoost 模型,如图 4 所示。该模型主要分为 3 个步骤:数据预处理,通过遗传算法进行参数调优,模型训练。首先,对数据进行预处理(包括缺失值补全、数据去重和特征分析),其中 0 代表不利于高血压病人的菜谱,1 代表有利于降低高血压的菜谱;然后,利用遗传算法寻找 XGBoost 模型的最优参数;最后,采用最终优化的参数构建 XGBoost 模型。

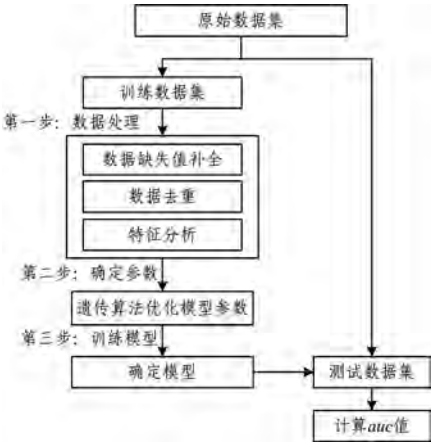

图 4 用于菜谱判定的基于遗传算法的改进 XGBoost 模型

5.1 数据预处理

本文使用 345 条样本数据,其中 146 条数据为专家已经确定的对高血压人群有益的菜谱数据,源于《降血压降血脂怎么吃》[25];114 条数据为网络爬虫数据,是一些网上公认的对高血压人群有益的菜谱;剩下的 85 条数据为营养学家判定为明显不利于高血压患者食用的菜谱。

将 345 条样本数据分为训练集和测试集,训练集包含 224 条数据,其中有益于高血压患者的菜谱数据有 177 条,不利于高血压患者的菜谱数据有 45 条;测试集数据包括 121 条,其中 81 条为有益菜谱,40 条为不利高血压患者的菜谱。

数据预处理主要是对数据集进行缺失值补全、数据去重

和特征分析。具体步骤如下：

1）对原始数据集进行特征缺失值补全，然后消除常量列和重复列。

2）对原始数据特征进行分析，得到特征数据的均值、方差、最小值、25％分位、50％分位、75％分位和最大值，如表 1 所列。

表 1　特征分析片断表

	calories	carbohydrate	fat	protein	vitamine	…
count	224	224	224	224	224	…
mean	153.76	16.44	7.04	7.24	1.30	…
std	120.98	18.13	9.93	5.16	1.68	…
min	11	0.22	0.1	0.07	0	…
25％	54.09	4.58	1.27	2.97	0.36	…
50％	124.88	7.88	3.34	5.69	0.86	…
75％	214.44	22.03	9.44	10.81	1.54	…
max	857.57	79.41	95.67	25.5	11.23	…

5.2　基于遗传算法的参数值优化

模型参数值会极大地影响模型的准确性，因此参数的最优化是模型训练中重要的步骤。XGBoost 继承 GBDT 的思想[26]，扩展和改进了 GDBT，在模型训练速度和计算准确率上都优于 GBDT。但该模型需要调试的参数较多（见表 2）[27]，且不同参数具有不同的功能，而调参通常取决于主观判断、经验和试错法，工作量大且精确度不高。因此，本文利用遗传算法解决全局优化问题，进而提升模型的准确度。

表 2　XGBoost 的部分参数及默认值解释

参数名	默认值	取值范围	解释
基分类器个数（tree_num）	—	—	基分类器个数是需要事先指定的
学习率（learning_rate）	0.3	[0,1]	通过调整每一步的权重，提高模型的鲁棒性
最大树深（max_depth）	6	[0,∞]	树深越大，模型会陷入局部最优解，控制树深，可避免过拟合
最小叶子权重（min_child_weight）	1	[0,∞]	该参数值越小，越容易过拟合，当它的值较大时，可以避免模型陷入局部最优解
gamma	0	[0,∞]	gamma 指定了节点分裂所需的最小损失函数下降值，一般数值越大，算法越保守
subsample	1	(0,1]	控制选择数据集的某部分进行训练，可以防止出现过拟合现象
colsample_bytree	1	(0,1]	太小的值会产生欠拟合现象
lambda	1	—	L2 正则的惩罚系数
alpha	0	—	L1 正则的惩罚系数
…	…	…	…

由于 XGBoost 所含参数较多，本文主要选取对模型影响较大的 4 个参数进行讨论：基分类器个数（tree_num）、学习率（learning_rate）、最大树深（max_depth）和最小叶子权重（min_child_weight）；其他参数设置为默认值。遗传算法自身有 3 个参数，即群体大小、交叉概率和变异概率。

本文根据遗传算法的特点，结合常用参数设置经验来初始化遗传算法的参数。设初始群体的大小为 100，交叉概率为 0.6，这样既有利于向前搜索，又能保护高适应值的结构；变异概率即为个体群中产生变异的概率，设其值为 0.01，这样既可产生新基因结构的动力，又能避免陷入单纯的随机搜索；表 3 列出了遗传算法结果中最优 4 组参数的参数值及其 AUC 值。从表中可以看出，第四组参数值的表现效果最好，其中最优树数目为 90，学习率为 0.17，最大树深为 3，最小叶子权重为 1。而第三组中的最大树深为 8，易产生过拟合。

表 3　遗传算法结果中最优 4 组参数的参数值及其 AUC 值

分组	第一组	第二组	第三组	第四组
代数范围	82～87	88～91	92～97	98～100
出现次数	6	4	6	3
基分类器个数	140	120	140	90
学习率	0.2	0.17	0.18	0.17
最大树深	6	3	8	3
最小叶子权重	4	2	3	1
AUC 值	0.9316	0.9323	0.9335	0.9349

5.3　XGBoost 模型训练

通过 5.1 节和 5.2 节中的数据预处理，遗传算法以多点搜索和并行性来寻找全局最优解，并快速确定最优参数群，这极大地发挥了 XGBoost 算法的有效性；在建立高血压菜谱识别的 XGBoost 模型后，将测试数据输入模型进行识别，判定其是否为有益于高血压患者的菜谱，并对输出的测试数据集进行标准评估，如准确率、回召率和 F1 值。

6　实验结果

经遗传算法调优后，得到的最好群体参数如下：基分类器个数为 90、学习率为 0.17、最大树深为 3、最小叶子权重为 1。在这些最优参数下，XGBoost 模型在高血压菜谱识别时的表现效果如表 4 所列。对高血压患者没有益处的菜谱识别 F1 值为 0.65，其中准确率为 0.79，回召率为 0.55；对高血压患者有益处的菜谱识别 F1 值为 0.86，其中准确率训 0.81，回召率为 0.93。最终得出经遗传算法调优后的 XGBoost 模型在高血压菜谱识别时的 AUC 值为 0.875，具有较高的菜谱识别性能。在参数调优过程中，遗传算法搜索出的树深为 3，说明只有少部分特征会影响模型对高血压菜谱的识别效果。

表 4　XGBoost 模型在进行高血压菜谱判定时的识别效果

类别	Precision	Recall	F1-score	个数
0	0.79	0.55	0.65	40
1	0.81	0.93	0.86	81
总体	0.80	0.80	0.79	121

本文采用信息熵来计算特征重要度，信息熵[28]被用来衡量一个随机变量出现的期望值。信息熵的计算公式为 $H(X)=-\sum_{i=1}^{n}p(x_i)\log p(x_i)$，其中 $p(x_i)$ 代表随机事件 X 产生 x_i 的发生概率。当事件发生的概率越小，发生时所包含的信息量越大，使用信息熵来描述数据的混乱程度可以有效消除系统内部分变量的不确定性，大大提高系统对数据的利用率。本文中菜谱的 23 维特征的重要度如图 5 所示。

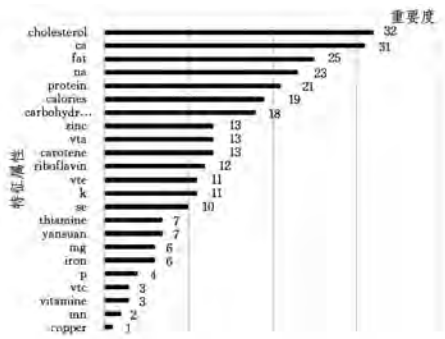

图 5　特征属性的重要度

由图可知，cholesterol，ca，fat，na，protein，calories 和 carbohydrate 这 7 个元素的重要度较大，且经专家论证这些元素

对高血压患者的影响较大[29-30]，因此选用这 7 个元素作为模型输入对高血压菜谱进行判定。

表 5 列出了这 7 个特征的表现效果，其中 AUC 值达到 0.850，与 23 维输入的模型 AUC 值 0.875 相差无几。实验结果表明，在不大量丢失准确度且选择这 7 个因子时，整体预测的准确率达到了较高水平。

表 5 7 个特征的模型性能评估

类别	Precision	Recall	F1-score	个数
0	0.73	0.60	0.66	40
1	0.82	0.89	0.85	81
总体	0.79	0.79	0.79	121

7 实验对比

本文设置两组对比实验：1)网格搜索算法参数调优和本文采用的遗传算法参数调优的对比，其中网格搜索算法调优又分为单参数网格搜索形式和全参数网格搜索形式；2)本文模型与其他 4 种常见模型(随机森林、GBDT、Bagging 和 AdaBooster)的对比。通过对比实验论证遗传算法优化 XGBoost 算法的可行性，以及优化后的模型应用于高血压菜谱判别的有效性。

7.1 遗传算法与网格搜索算法进行参数调优的对比

(1)网格搜索算法

网格搜索算法将待搜索参数在一定的空间范围内划分成网格，通过遍历网格中所有的点来寻找最优参数，这种方法在寻优区间足够大且步距足够小的情况下可以找出全局最优解[31]。在智能优化中使用网格搜索算法进行调参比较普遍，其中文献[32-33]等均利用网格搜索算法来寻找最优的参数使目标值最大化，但由于网格内多数参数组对应的分类准确率都非常低，只有较小区间内的参数组所对应的分类准确率很高，因此遍历网格内所有参数组会相当耗时[31]。

(2)参数调优实验对比

在参数调优对比实验中，网格搜索算法将 XGBoost 中的 4 个参数值(基分类器个数、学习率、最大树深和最小叶子权重)的可行区间按从小到大的顺序排列并划分出一些小区间，由计算机按顺序计算各参数变量值组合所对应的 AUC 值，从而求得本区间内最小目标值和其对应的最佳参数值。为了实验的精确性，本文设计了两种网格搜索算法形式，包括单参数网格搜索形式和全参数网格搜索形式，其中单参数网格搜索形式是指只改变目标参数值，其他参数采用模型的默认值；而全参数网格搜索形式是指对所有参数进行排列组合，缺点是模型搜索的时间复杂度高。

表 6 列出了单参数网格搜索算法、全参数网格搜索算法和本文模型采用的遗传算法优化算法对这 4 个参数调优后得到的参数结果。经对比可知，全参数网格搜索算法调优结果中的 4 个参数值除了基分类器的数目不同，其他 3 个参数值已经非常接近遗传算法优化出的参数值。

表 6 本文采用的遗传算法与网格搜索算法的参数调优对比

	本文模型	单参数网格搜索	全参数网格搜索
基分类器个数	90	40	20
学习率	0.17	0.04	0.19
最大树深	3	4	3
最小叶子权重	1	6	1

表 7 是单参数网格搜索算法、全参数网格搜索算法和遗传算法在高血压菜谱识别上的表现效果对比。由表可知，单参数网格搜索算法的调优结果在 XGBoost 模型上的 AUC 值是 0.822，全参数网格搜索算法的 AUC 值为 0.846，而本文采用的遗传算法调优结果达 0.875。

表 7 本文模型调优与网格搜索算法参数调优的对比

	本文模型	单参数网格搜索	全参数网格搜索
0 的准确率	0.79	0.83	0.74
1 的准确率	0.81	0.79	0.79
0 的回召率	0.55	0.47	0.50
1 的回召率	0.93	0.95	0.91
0 的 F1	0.65	0.60	0.60
1 的 F1	0.86	0.86	0.85
AUC 值	0.875	0.822	0.846

7.2 模型对比

7.2.1 采用遗传算法改进 XGBoost 前后模型的对比

为了说明遗传算法的有效性，本文设置一组对比实验，将遗传算法优化的 XGBoost 模型(本文模型)的表现效果与未使用遗传算法优化的 XGBoost 模型的表现效果进行对比。如表 8 所列，本文模型在测试集上的 AUC 值高于未使用遗传算法优化的 XGBoost 模型，由此说明遗传算法调优可有效寻找最优参数组合，提升 XGBoost 算法的效率。

表 8 遗传算法改进 XGBoost 算法前后 AUC 值的对比

	改进前 XGBoost	本文模型
AUC 值	0.868	0.875

7.2.2 改进的 XGBoost 与其他分类模型的对比

分类问题占机器学习和数据挖掘领域中问题总数的 70%，除了本文用到的 XGBoost 算法，还有许多被证实有效的组合分类算法，如随机森林、GBDT、Bagging 和 AdaBooster 等，这些组合分类器针对不同的数据集产生不同的效果。在高血压菜谱识别的背景下，将本文提出的基于 GA 改进的 XGBoost 模型与之相比较。

模型的好坏不能仅仅通过数据正确率的高低来评判，针对不同场景需要不同的评价指标，本文采用准确率、回召率、F1 值作为模型的对比指标，以 AUC 值作为模型最终的评价指标。

表 9 列出了是本文模型与遗传算法优化的 GBDT 模型的对比结果，在类别为 0 和 1 的各项单个指标中 GBDT 具有较高的评估指数，但是本文模型在概率评价的 AUC 值上具有较高的数值表现，说明 XGBoost 算法优于 GBDT 算法，在数据预测上保持了较好的分类能力和稳定性。

表 9 分类模型的性能对比

	本文模型	GBDT	随机森林	Bagging	AdaBooster
0 的准确率	0.79	0.85	0.92	0.82	0.77
1 的准确率	0.81	0.82	0.74	0.82	0.82
0 的回召率	0.55	0.57	0.30	0.57	0.60
1 的回召率	0.93	0.95	0.99	0.94	0.91
0 的 F1	0.65	0.69	0.45	0.68	0.68
1 的 F1	0.86	0.88	0.85	0.87	0.87
AUC 值	0.875	0.868	0.858	0.831	0.838

将本文模型与另外 3 种常见组合分类模型在准确率、回召率、F1 值上进行对比分析，随机森林在 0 的准确率上高达 0.92，但是在 1 上的准确率却明显低于其他模型。基于遗传算法的改进 GBDT 在 0 和 1 的 F1 值处取得最优值，Bagging

和 AdaBooster 模型在回召率、F1 值和 AUC 值上相差无几。

图 6 给出了其他常见组合分类模型在测试集上的 AUC 值对比结果，本文模型的 AUC 值达到最大，XGBoost 是 GBDT 算法的优化版本，也具有在 Boosting 模型基础上建立的随机森林和 AdaBooster 算法的优点，对弱学习得到的弱分类器的错误能进行适应性地调整。针对本文 23 维的菜谱数据，XGBoost 模型相比其他常见的组合树模型具有较好的数据分类能力和较高的预测准确率，可为高血压患者提供较为可靠的预测信息，因此在本文 23 维的高血压菜谱应用背景下，采用 XGBoost 算法构建菜谱识别模型具有普遍的实用价值。

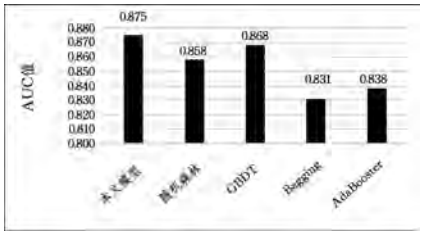

图 6　各分类模型的 AUC 值

结束语　XGBoost 模型的实际应用已经涉及到很多方面，应用力度也大为提升。本文探讨了 XGBoost 算法在菜谱识别方面的应用，通过数据预处理、参数寻优和模型训练 3 个步骤，提出了一种用于高血压菜谱识别的基于遗传算法的优化 XGBoost 模型。通过遗传算法与网格搜索算法参数调优的对比和本文模型与其他 4 种较为常见模型（随机森林、GBDT、Bagging 和 AdaBooster）的对比来对该模型进行综合评估，结果表明该模型具备良好的高血压菜谱识别效果，具有很高的实际应用价值。

本文提出的算法模型可以应用到其他疾病的菜谱或者药物方剂的识别方面，能广泛识别和处理其他二分类问题。

参 考 文 献

[1] 李小莉. 浅谈社区慢性病高血压的健康管理[J]. 环球中医药，2013(z1)：291-291，292.

[2] 王春利. 终止高血压膳食疗法对社区高血压前期人群干预效果研究[J]. 中国全科医学，2015，8(23)：2833-2836.

[3] 毕振强，梁晓峰，马吉祥，等. 遏制高血压危害，减盐行动势在必行[J]. 中华预防医学杂志，2014，48(1)：4-6.

[4] 刘雪梅，徐琳琳，王楠，等. 日常饮用洋葱汁对高血压和高血脂患者血压、血脂影响研究[J]. 中国食物与营养，2015，21(8)：84-87.

[5] 吴珊燕，许鑫. 基于案例推理的菜谱推荐系统研究[J]. 现代图书情报技术，2013，29(12)：34-41.

[6] 夏平平. 个性化营养菜谱推荐方法的研究[D]. 合肥：中国科学技术大学，2015.

[7] 张璐，雷雪梅. 基于粒子群优化 BP 神经网络的养肠胃菜谱判定[J]. 计算机科学，2016，43(11A)：63-66，72.

[8] 章艳珍，吴岚艳，李李. 膳食营养素与高血压关系的研究进展[J]. 中国食物与营养，2017，23(2)：87-89.

[9] CHEN T Q，GUESTRIN C. XGBoost：A Scalable Tree Boosting System [C]// Acm Sigkdd International Conference on Knowledge Discovery & Data Mining. 2016：785-794.

[10] XIA Y F，LIU C Z，LI Y Y，et al. A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring[J]. Expert Systems With Applications，2017，78：225-241.

[11] 张昊，纪宏超，张红宇. XGBoost 算法在电子商务商品推荐中的应用[J]. 物联网技术，2017，7(2)：102-104.

[12] 叶倩怡. 基于 Xgboost 方法的实体零售业销售额预测研究[D]. 南昌：南昌大学，2016.

[13] 樊鹏. 基于优化的 xgboost-LMT 模型的供应商信用评价研究[D]. 广州：广东工业大学，2016.

[14] MUSTAPHA I B，SAEED F. Bioactive Molecule Prediction Using Extreme Gradient Boosting [J]. Molecules，2016，21(8)：983.

[15] 怀浩，刘学，张龙波，等. 基于梯度提升决策树的肽碎片离子强度建模[J]. 山东理工大学学报(自然科学版)，2017，31(2)：64-68.

[16] 胥红敏，郭湛，李晓宇，等. 基于遗传算法优化 BP 神经网络的道口事故预测[J]. 铁路计算机应用，2016，25(3)：8-11.

[17] 史峰，王辉，胡斐，等. MATLAB 智能算法 30 个案例分析[M]. 北京：北京航空航天大学出版社，2011.

[18] 向万里，马寿峰. 基于轮盘赌反向选择机制的蜂群优化算法[J]. 计算机应用研究，2013，30(1)：86-89.

[19] 周治平，朱书伟，张道文. 分类数据的多目标模糊中心点聚类算法[J]. 计算机研究与发展，2016，53(11)：2594-2606.

[20] 李刚，高武奇，杨瑞臣. 有指导机器学习超参数的交叉验证智能优化[J]. 西安工业大学学报，2016，36(11)：906-910.

[21] 曲思杨，张秋菊，王文佶. 多次交叉验证对 PLSDA 模型的影响研究[J]. 中国卫生统计，2017，34(1)：15-17.

[22] 杨柳，王钰. 泛化误差的各种交叉验证估计方法综述[J]. 计算机应用研究，2015(5)：1287-1290.

[23] 于化龙，倪军，徐森. 基于留一交叉验证的类不平衡危害预评估策略[J]. 小型微型计算机系统，2012，33(10)：2287-2292.

[24] 吴学龙，徐维超. 基于 AUC 的非参数快速变点检测算法[J]. 计算机与现代化，2015(7)：5-8.

[25] 胡大一. 降血压降血脂怎么吃[M]. 青岛：青岛出版社，2009.

[26] 王天华. 基于改进的 GBDT 算法的乘客出行预测研究[D]. 辽宁：大连理工大学，2016.

[27] DMLC[OL]. http://xgboost. readthedocs. io/en/latest/parameter. html.

[28] 彭长根，丁红发，朱义杰，等. 隐私保护的信息熵模型及其度量方法[J]. 软件学报，2016，27(8)：1891-1903.

[29] 张亮，曹华军，李汇华. 膳食营养与高血压研究进展[J]. 中国食物与营养，2017，23(2)：78-83.

[30] 高冰. 膳食营养与高血压关系的研究进展[J]. 包头医学院学报，2013，29(6)：114-115.

[31] 王健峰，张磊，陈国兴，等. 基于改进的网格搜索法的 SVM 参数优化[J]. 应用科技，2012(3)：28-31.

[32] 刘佳，施龙青，韩进，等. Grid-Search_PSO 优化 SVM 回归预测矿井涌水量[J]. 煤炭技术，2015，34(8)：184-186.

[33] XU W，ZUO M，ZHANG M，et al. Constraint bagging for stock price prediction using neural networks [C]// International Conference on Modelling，Identification and Control. IEEE，2010：606-610.