NCA 降维和贝叶斯优化调参对分类模型的改进

李 斌 王卫星
(河南科技大学应用工程学院现代教育技术中心 河南 三门峡 472000)

摘 要 高校贫困生的贫困程度判定可以归属于构建分类模型对样本数据进行训练。但单个分类模型的精准度要取决于处理样本数据的大小和类型复杂度,在模型速度和准确性之间不易取舍。集成多个分类算法可以避免单个分类算法的过拟合。通过邻域分量分析(Neighborhood Component Analysis, NCA)进行特征降维降低初始分类模型的计算成本,对误判损失引入一个成本函数进行惩罚的同时采用贝叶斯优化进行超参数调优。结果表明,改进后的分类模型泛化能力得到明显提升。计算时间成本降低的同时,误判率由初始的 8% 下降到 5%,模型的准确率提升了近 4%。

关键词 分类算法 领域分量分析 贝叶斯调优 MATLAB 贫困生判别

中图分类号 TP3 **文献标识码** A DOI:10.3969/j. issn. 1000-386x. 2019. 08. 047

IMPROVEMENT OF CLASSIFICATION MODEL BY NCA DIMENSION REDUCTION AND BAYESIAN OPTIMIZATION PARAMETER ADJUSTMENT

Li Bin Wang Weixing
(*Modern Education Technology Center, College of Applied Engineering, Henan University of Science and Technology, Sanmenxia 472000, Henan, China*)

Abstract Poverty levels of poor students in the university can be attributed to build a classification model of training sample data. But the model of a single classification accuracy depends on the size of the sample data and types of complexity, and it is difficult to choose between the speed and accuracy of the model. Integrating multiple classification algorithm can avoid a single classification algorithm of fitting. Through the neighborhood component analysis (NCA) for feature dimension reduction, we reduced initial classification model of calculating cost. For misjudgment loss, we introduced a cost function to punish and used bayesian optimization to super parameter tuning simultaneously. The results show that the generalization ability of improved classification model is improved significantly. At the same time, the computation time cost decreases, misjudgment rate decreases from 8% to 5%, and the accuracy of the model increases by nearly 4%.

Keywords Classification algorithm Neighborhood component analysis(NCA) Bayesian tuning MATLAB Poor student discriminant

0 引 言

目前对高校贫困生进行判定的方法大都利用数据挖掘技术定量和定性结合。文献[1]通过能够面向多值属性的关联规则 Apriori 算法的改进提高了数据挖掘效率,为高校贫困生认定工作提供了有利依据;文献[2-4]对数据预处理并使用 C4.5 算法,将知识表示成树的形式,采用错误预测率进行修剪,分别归纳出决策树,分析并选出其中较优结果,原理简单且计算快速准确;文献[5]基于加权约束的决策树认定方法提高了贫困生认定效率;文献[6]结合 Logistic 回归、Native Bayes 和 k 近邻三种分类预测模型综合比较认为 k 近邻模型能更好地判别出学生是否是贫困生;文献[7]

收稿日期:2018-12-15。河南省 2017 年高等教育教学改革研究与实践项目(2017SJGLX636)。**李斌**,高工,主研领域:数据分析,网络规划。**王卫星**,副教授。

在相同的数据集中证明随机森林算法分类正确率较高。

上述学者针对贫困生判定的研究主要侧重于个别分类算法,对算法的计算成本、性能优化缺乏深入分析,评价方式比较单一化。本文认为高校贫困生识别可以在做好反复训练和评估模型的基础上,集成多个分类算法,运用 NCA 对特征参数降维以提升计算性能;引入成本惩罚函数并利用贝叶斯超参数调优对分类模型进行进一步优化,以提升分类模型的预测准确率。

1 分类算法的对比选择

分类算法旨在构建分类预测的模型,是人工智能、模式识别和数据挖掘领域中重要的数据处理方法[8]。

1.1 分类算法简述

1.1.1 决策树 CART

CART(Classification and Regression tree)分类回归树使用基尼指数(Gini),采用二元切分法选择特征进行训练数据切割:

$$Gini(D) = 1 - \sum_{i=0}^{n}\left(\frac{D_i}{D}\right)^2 \tag{1}$$

$$Gini(D \mid A) = 1 - \sum_{i=0}^{n}\left(\frac{D_i}{D}\right)Gini(D_i) \tag{2}$$

决策树算法的优点是计算复杂度不高,输出结果易于理解,对中间值的缺失不敏感,缺点是易会产生过拟合问题[9-10]。

1.1.2 非线性 SVM

SVM 支持向量机是将低维空间的输入数据投放到一个更高维的特征空间,用线性决策边界分割在低维空间难以区分的正例和负例。在非线性问题上,用内积 $\phi(x_i)\cdot\phi(x_j)$ 代替最优分类面中的点积。

最大化目标函数为:

$$L_D = \sum_{i=0}^{n} a_i - \frac{1}{2}\sum_{j=1}^{n} y_i y_j a_i a_j \langle \phi(x_i)\cdot\phi(x_j)\rangle \tag{3}$$

约束条件:

$$\sum_{i=1}^{n} y_i a_i = 0, 0 \leqslant a_i \leqslant C \quad i = 1,2,\cdots,n \tag{4}$$

相应的分类器函数转化为:

$$f(x) = sign\left[\left(\sum_{i=1}^{k} a_i y_i \langle \phi(x_i)\cdot\phi(x_j)\rangle + b\right)\right] \tag{5}$$

SVM 的优点是泛化错误率低,计算开销不大,结果易解释;缺点是对主要适用于处理二分类问题,参数调节和核函数的选择敏感,但经过构造可以将多分类问题转化为二分类问题[11]。

1.1.3 k-最近邻算法

k-最近邻给每个属性相等的权重进行基于距离的邻近比较。常用的邻近距离是欧几里德距离,两个点或样本 $X_1 = (x_{11}, x_{12}, \cdots, x_{1n})$ 和 $X_2 = (x_{21}, x_{22}, \cdots, x_{2n})$ 的欧几里德距离为:

$$dist(X_1, X_2) = \sqrt{\sum_{i=1}^{n}(x_{1i} - x_{2i})^2} \tag{6}$$

k-最近邻分类算法的优点是无数据输入假定、噪声数据影响不大、精度略高;缺点是计算空间复杂度高。

1.1.4 贝叶斯方法

贝叶斯是基于贝叶斯定理与特征条件独立假设的分类方法,在数据集 D 中令 $A_1, A_2, \cdots, A_{|A|}$ 为用离散值表示的属性集合,令 C 为具有 $|C|$ 个不同值的类别属性,假设所有属性都是条件独立于类别 $C = c_j$,数学表示为:

$$P = (A_1 = a_1 | A_2 = a_2, \cdots, A_{|A|} = a_{|A|}, C = c_j) = P(A_1 = a_i | C = c_j)$$

从训练数据中可以直接得到先验概率 $P(C = c_j)$ 和条件概率 $P(A_1 = a_i)$,贝叶斯的分类公式为:

$$c = \underset{c_j}{\operatorname{argmax}} P(C = c_j)\prod_{i=1}^{|A|} P(A_i = a_i \mid C = c_j) \tag{7}$$

贝叶斯法的优点即使数据较少也可高效处理多类别问题;缺点是对于数据输入假设条件较为敏感。

1.1.5 BP 神经网络

神经网络是由一个输入层、若干个隐含层和一个输出层组成的多层网络,各层之间的连接方式通过权重值调节。若模型确定训练误差的理想输出是 t_k,实际输出是 z_k,c 代表输出向量的长度,ω 代表网络的所有权值,η 是学习速率,那么总误差表示为:

$$J(\omega) = \frac{1}{2}\sum_{k=1}^{c}(t_k - z_k)^2 = \frac{1}{2}\|t - z\| \tag{8}$$

基于梯度下降的误差反向传播算法 BP 神经网络是沿着减小误差的方向来调整权值:

$$\Delta\omega = -\eta\frac{\partial J}{\partial\omega} \tag{9}$$

BP 算法对网络拓扑及初始权重敏感,泛化性能往往不能得到保证,容易陷入局部最小[12-14]。

综上所述,将几种典型的机器分类算法的对比总结如表 1 所示。

表 1 几种分类算法分析比较

算法	训练速度	内存使用	是否调优	预测速度	一般评估
逻辑回归	快	小	最小	快	擅长解决有线性决策边界小问题
决策树	快	小	有些	快	通用性好,但容易过拟合

续表 1

算法	训练速度	内存使用	是否调优	预测速度	一般评估
非线性 SVM	慢	中等	有些	慢	擅长解决二进制问题,能很好处理高维度数据
k-最近邻	最小	中等	最小	适中	精度较低,但易于使用和解释
贝叶斯	快	中等	有些	快	广泛应用于文本
神经网络	慢	中到大	很多	适中	应用于分类、压缩、识别和预测

1.2　分类算法的选择

在机器学习领域里,一方面高度灵活的模型由于拟合了噪声数据的细微变化易造成过拟合,另一方面简单的模型可能又需有更多的假设条件。在模型速度、准确性和复杂性之间的权衡本已不易,算法的选择还取决于要处理的数据的大小和类型以及如何运用从数据中获得的洞察力,因此不存在一种万能的算法可以完美解决所有问题。

在对高校贫困生预测判定建模时,需要做好反复训练和评估模型的准备。既可运行所有算法进行比较,也可从特定分类任务的经验最佳拟合算法开始。对每个训练的分类器,要保留验证数据或反复使用交叉验证对精确度进行评估,最终尝试集成多类分类算法克服训练数据的过拟合。

2　算法模型的改进优化

分类模型的改进优化意味着进一步提高其准确性和预测能力,避免模型无法区分数据和噪声时过拟合。本文在对分类模型经反复评估初步确定后,对模型的改进优化手段主要采取邻域向量分析 NCA 特征降维和贝叶斯超参数调优。

2.1　NCA 特征降维

特征降维是向模型添加变量或移除不能改进模型性能的变量,以在数据建模中提供最佳预测能力[15]。特征降维不但可以降低计算成本和存储要求,还能使预测结果更加精确。

NCA 是一种距离测度学习算法。该算法随机选择近邻,通过优化留一法(Leave-one-out, LOO)的交叉检验结果来求得马氏距离中的变换矩阵。在这个过程中完成降维,最后在低维空间对数据完成分类。

数据集 $X=\{x_1,x_2,\cdots,x_n\}$ 在 $\boldsymbol{R}^D$ 空间内分别具有类标签 $c_1,c_2,\cdots,c_n$,限定马氏距离变换矩阵 $\boldsymbol{Q}=\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$,两个样本点之间的马氏距离定义为:

$$d(x_i,x_j)=\sqrt{(x_i-x_j)^{\mathrm{T}}\boldsymbol{Q}(x_i-x_j)}=\sqrt{(\boldsymbol{A}(x_i)-\boldsymbol{A}x_j)^{\mathrm{T}}\boldsymbol{A}(x_i-\boldsymbol{A}x_j)}\quad i,j=1,2,\cdots,n \tag{10}$$

样本点 x_i 随机选择一个 x_j 近邻并继承其类标签 c_j 的概率 P_{ij},概率 P_{ij} 在变化空间中使用欧式距离定义如下:

$$\begin{cases}p_{ij}=\dfrac{\exp(-\|\boldsymbol{A}x_i-\boldsymbol{A}x_j\|)}{\sum\limits_{k\neq i}\exp(-\|\boldsymbol{A}x_i-\boldsymbol{A}x_k\|)} & j\neq 1\\ p_{ij}=0 \quad j=i\end{cases} \tag{11}$$

因为每个数据点都可以选择为近邻,因此输入数据可以继承所有的类标签,样本点 x_i 正确分类的概率为:

$$P_i=\sum_{j\in c_i}P_{ij} \tag{12}$$

NCA 搜索变换矩阵 $\boldsymbol{A}$,目标函数可以理解为要使得正确分类的点数最大化期望,也就等同于最小化类间距离:

$$f(\boldsymbol{A})=\sum_i P_i=\sum_i\sum_{j\in c_i}P_{ij} \tag{13}$$

这个无约束优化问题通过共轭梯度法或随机梯度法求出 $\boldsymbol{A}$,使用微分的变换矩阵:

$$\frac{\partial f(\boldsymbol{A})}{\partial \boldsymbol{A}}=-2A\sum_i\sum_{j\in c_i}P_{ij}(x_{ij}x_{ij}^{\mathrm{T}}-\sum_k P_{ik}x_{ik}x_{ik}^{\mathrm{T}}) \tag{14}$$

式中:$x_{ij}=x_i-x_j$,当 $\boldsymbol{A}$ 是 $d\times D$ 的非方阵时,经过 NCA 距离测度学习可以将样本降到 $\boldsymbol{R}^{\mathrm{D}}$ 空间[16-17]。

实际应用中,由于共轭梯度法通过多次迭代才能得到目标函数最优解,占用内存的同时耗时较大,因此使用等价于共轭梯度的拟牛顿法基础上的 L-BFGS(Limited-memory BFGS)算法进行计算,其中 BFGS 是四个提出这种拟牛顿法的四个人名的首字母。L-BFGS算法的核心是不再存储完整的矩阵,而是存储计算过程中的向量序列,且只利用最新的向量序列,以大幅降低运算成本。

2.2　超参数调优

识别能提供最佳模型的参数集的过程可称为超参数调优。两个常用的参数调优方法是网格搜索和贝叶斯优化。虽然网格搜索能彻底搜索参数值组合的有限集,但耗时太长并易遇到维度灾难。

贝叶斯参数优化充分利用被测试点忽略的前一个点的信息[18]。它根据先验分布假设一个搜集函数,使用每次新采样点去测试目标函数的信息来更新目标函数的先验分布。然后测试由后验分布给出的全局最值最可能出现的位置点。贝叶斯优化虽需执行更多的迭

代计算以确定下一个采样点,但可以较少的评估就找到复杂非凸函数的最小值,主要分三个步骤:

(1) 选择一个先验函数来表达关于被优化函数的假设。本文选择使用的高斯过程是一个随机变量的集合,任意有限个随机变量都满足一个联合高斯分布[9]。若 X 表示训练集 $\{x_1,x_2,\cdots,x_t\}$,f 表示未知函数值集合 $\{f(x_1),f(x_2),\cdots,f(x_t)\}$,$\boldsymbol{\Sigma}$ 表示 $k(x,x')$ 构成的协方差矩阵Ⅱ,θ 表示超参数,当存在观测噪声且假设噪声 ε 满足独立同分布的高斯分布 $p(\varepsilon)=\mathcal{N}(0,\sigma^2)$,可以得到边际似然分布为:

$$P(y|X,\theta)_2=\int p(y\mid f)p(f\mid X,\theta)\mathrm{d}f=\mathcal{N}(0,\boldsymbol{\Sigma}+\sigma^2 I) \quad (15)$$

式中:y 表示观测值集合 $\{y_1,y_2,\cdots,y_t\}$。

(2) 通过 ML 极大似然估计对边际似然分布最大化得到优化超参数 $\hat{\theta}_t$。为超参数赋予先验分布 $p(\theta)$,根据贝叶斯定理得到:

$$P(\theta|D_{1:t})=\frac{p(D_{1:t}|\theta)P(\theta)}{p(D_{1:t})} \quad (16)$$

然后通过最大后验估计 MAP 最大化式(6),得到 $\hat{\theta}_t$。

(3) 根据 $\hat{\theta}_t$ 能够得到具体的采集函数:

$$\hat{\alpha}_t(x)=\alpha(x;\theta_t) \quad (17)$$

然后选择采集函数用来从后验模型构造一个效用函数,来确定下一个采样点[20-22]。采集函数可以在具有低建模目标函数的点上对采样进行平衡,并对尚未建模区域进行搜索。

贝叶斯超参数调优的算法步骤如算法 1 所示。

算法 1 贝叶斯优化算法

Bayesian optimization:选取 n 个采样点作为先验,假设它们服从高斯分布

1: for $n=1,2,\cdots,n$, do

2: 根据最大化采集函数 α 选取下一个采集点 x_{n+1}

$x_{n+1}=\arg\max_x \alpha(x;D_n)$

3: 查询目标函数以获得 y_{n+1}

4: 整合数据集 $D_{n+1}=\{D_n,(x_{n+1},y_{n+1})\}$

5: 更新概率模型

6: end for

为提高找到最优参数值的机率,并使超参数调优更加高效,使用 MATLAB 中的贝叶斯优化工具执行超参数调优,同时引入成本函数对错误分类进行惩罚。

3 应用实证

高校贫困学生的贫困成因多集中在家庭经济情况、生活水平、家庭劳动力状况、在校消费能力水平、消费习惯、学业水平、学习主动力等方面[23]。

本文通过某高校 2016 - 2017 年度校园应用服务中积累的数据。首先选择训练数据进行分类学习,反复训练和评估分类模型后选择合适的分类算法。然后采用 NCA 特征降维和贝叶斯参数调优对模型进行优化,对某高校的贫困生的精准判定实现预测和评判。

3.1 选择训练数据和算法验证

样本数据会以各种形式和大小出现,如高校贫困生的真实数据集可能较混乱、不完整且采用格式各异。对高校各个业务子系统中得到的原始数据进行预处理需采用专业数据处理工具和不同的预处理方法。

将从高校各个应用系统中抽取出的数据进行标签标记、清理无效数据、分类汇总后得到完整的样本数据共 9 909 组。这些组样本数据初步特征值共有 21 种,其中部分特征来源于学生调查问卷等,并对部分数据进行了离散化处理,如表 2 所示。

表 2 样本特征值列表

序号	特征名称	字段	含义	描述
1	num_consump	Int	消费次数	年度内学校餐厅消费次数
2	sum_consump	Float	消费平均	平均每次消费金额
3	var_consump	Float	消费方差	消费额方差值
4	part-time job	Bit	打工	是否勤工俭学
5	num_borrow	Int	借阅数	年图书馆借阅图书数目
6	weight_average_core	Float	成绩	各科成绩加权平均分
7	repeat_core	Int	重修	课程重修门数
8	elecNum	Int	电子产品	个人拥有电子产品数量
9	income_family	Int	家庭收入	学生家庭月收入水平值
10	single_parent	Bit	单亲	是否单亲家庭
11	edu_multiple_child	Bit	子女上学	学生家庭正受教育子女数
12	cost_living	Int	生活费	学生在校每月基本生活费
13	old_poor_remote	Bit	老少边穷	学生家是否老少边穷区
14	in debt	Bit	债务	家庭是否因病或其他欠债
15	num_family	Int	家庭人口	学生家庭人口数量
16	only_child	Bit	独生	是否独生子女

续表 2

序号	特征名称	字段	含义	描述
17	disease_family	Bit	患病	家庭是否有长期患病人口
18	score_mutual	Int	评分	同学贫困水平互评得分
19	health	Int	健康情况	校医室就诊次数
20	tuition_defer	Bit	学费缓交	是否办理学费缓交手续
21	classif_academic	Int	专业类别	学生所学专业类别

在 MATLAB 中将经过初步清噪脱敏后的数据导入,对数据样本采用 k 折交叉验证,k 值取 5,每次以 $k-1$ 份作为训练集,1 份作为验证集。得到验证集性能后,将 5 次结果平均作为模型的性能指标,以最大化使用模型训练的数据量,得到泛化更好的模型。MATLAB 中多个分类器的性能比较和分类初始结果如图 1 所示。

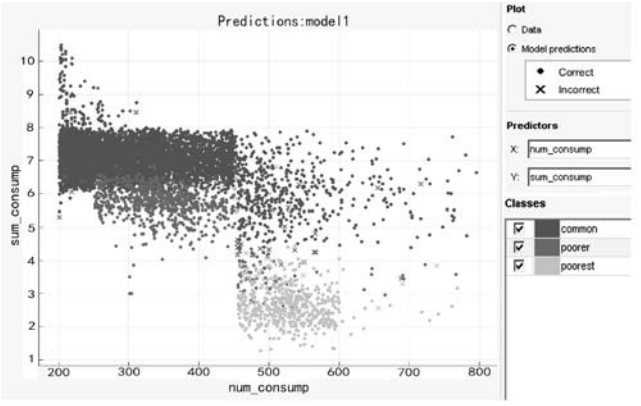


图 1　多个分类算法的初始比较图

从图 1 中可以看出,训练样本明显地被分为 common、poorer 和 poorest 三类灰度程度不同的颜色,其中的“×”为噪声数据。实证对比算法模型结果,高校贫困生预测最初显示二次支持向量机(SVM)表现良好,然后是线性支持向量机和决策树算法。不同分类器的时间消耗和准确率性能比较如表 3 所示。

表 3　不同分类算法的初始性能比较

分类器	预测速度/(obs·s^{-1})	训练时间/s	训练准确率/%
Decision Tree	~49 000	10.491 0	99.1
Linear Discriminant	~55 000	4.903 7	98.6
Linear SVM	~58 000	10.832 0	99.3
Medium KNN	~3 300	11.577 0	98.9
Quadratic SVM	~58 000	7.105 8	99.4

3.2　运用 NCA 进行特征降维

在处理高校贫困生涉及的数据集包含大量特征和有限的观察值时,运用 NCA 特征选择技术降维,具体步骤如下:

Step1　将训练数据分成 5 份,使用 CVpartition 进行交叉验证,赋值 λ 并创建一个数组阵列来存储损失函数值。

Step2　使用每部分中的训练集,为每个值训练 NCA 模型。使用 NCA 模型计算每部分中相应测试集的分类损失,记录损失值。

Step3　重复所有部分训练值和 λ 值,计算得出每个 λ 值的每个部分的平均损失。绘制平均损失值与 λ 值之间的关系,找到与最小平均损失对应的最佳 λ 值。

Step4　使用最佳 λ 值拟合 NCA 模型,使用计算效率更好的 L-BFGS 算法去求解目标函数,标准化预测值绘制特征权重。

图 2 显示了在 MATLAB 中使用邻域分量分析 NCA 识别的特征权重结果,圆圈表示对应特征的特征权重。可以看出特征指标 1(num_consump)、2(sum_consump)、3(var_consump)、9(income_family)、18(score_mutual)、12(cost_living)、6(weight_average_core)、8(elecNum)、14(indebt)、17(disease_family)、19(tuition_defer)的特征权重值高于相对阈值 0.374 6。利用 MATLAB 中自带的 NCA 降维揭示了在贫困生特征中大约一半的特征对模型没有重要作用。因此,我们可以减少特征数量,从 21 个减至 11 个。

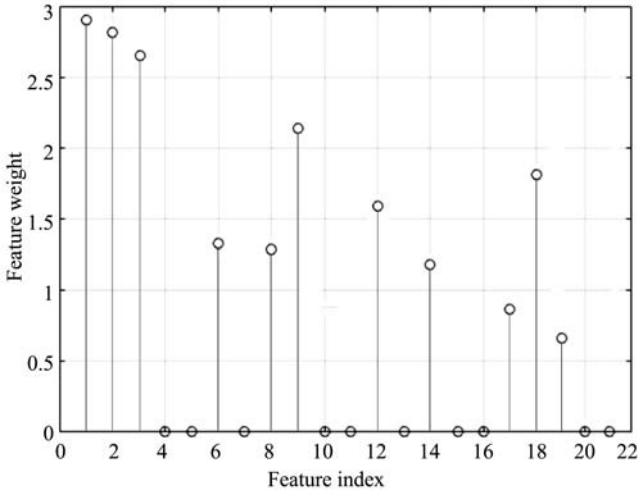


图 2　使用邻域分量分析 NCA 识别最相关的特征结果

按照 NCA 降维后的特征选择,重复前述分类算法,比较不同算法降维后的各项性能参数如表 4 所示。

表 4 不同分类算法 NCA 降维后性能比较

分类器	预测速度/(obs·s^{-1})		训练时间/s		训练准确率/%	
	降维后	变化	降维后	变化	降维后	变化
Decision Tree	51 000	+2 000	8.589 6	-1.901 4	98.9	-0.2
Linear Discriminant	83 000	+28 000	2.364 9	-2.538 8	98.2	-0.4
Linear SVM	64 000	+6 000	8.576 6	-2.255 4	98.7	-0.6
Medium KNN	54 000	+1 100	8.094 9	-3.482 1	98.6	-0.3
Quadratic SVM	57 000	-1 000	9.591 5	+2.485 7	99.0	-0.4

从表 4 的几种分类算法的性能变化值可以明显看出,NCA 降维后,整体预测速度和计算时间变化明显,特别是线性判别算法因为特征数的大幅减少而性能大幅提升,决策树分类算法表现优异。

使用单独的分类算法往往会过度拟合训练数据,为了克服这种倾向,可以尝试集成多个分类算法,典型的比如 Boosted Trees 和 Bagged Trees。测试表明这两种集成分类算法在降维后的准确率仍可以达到 99.3%。从上述算法对比中也可以看出,某些算法初始表现很好,改进后表现一般,有的反之。所以可以后退到特征提取阶段去寻找其他特征并降维,在机器学习工作流程的不同阶段之间反复实验和对比,寻找最佳模型。

3.3 引入成本函数的超参数调优

在高校贫困生预测分类模型中,单单根据总体精确度分析性能很容易产生误导,比如未能准确预测实际贫困相比错误地将正常情况学生误判为贫困要造成更大的不公平。图 3 所示的初步模型分类结果混淆矩阵,将 3% 的贫困生误报为正常学生,而将 8% 的普通学生分类为贫困和极度贫困。这将造成部分学生的评判结果失真,不需补助的学生得到补助,而急需补助的学生却失去应有的补助。

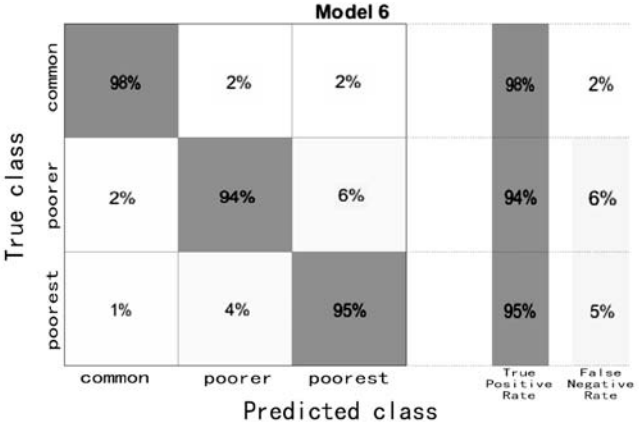


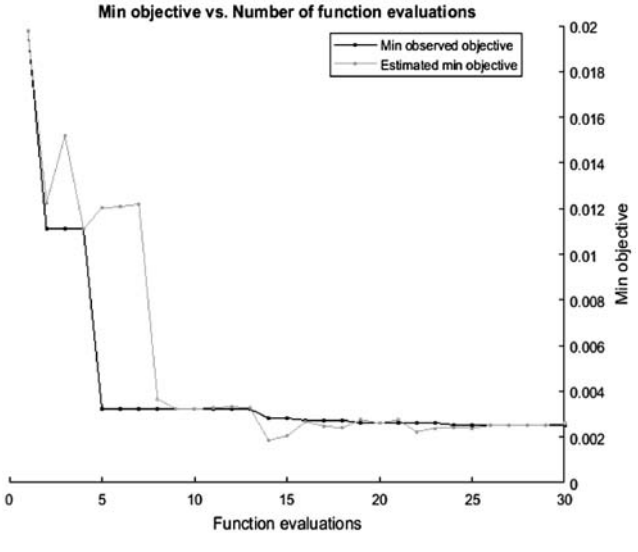
图 3 初步模型的混淆矩阵

为了改进分类器,引入成本函数对误分类进行惩罚,补偿数据中较少的“异常”观察,并使分类器偏向于较少的错误分类异常噪声,将较高的错误分类成本分配给“异常”类。同时利用贝叶斯优化方法对模型参数进行超参数调优。由于 Trees 的表现优于 SVM,本文以生成树为效果目标,步骤如下:

Step1 因为是 common、poorer 和 poorest 多分类,首先使用 AdaBoostM1 和 Trees 模型 5 倍交叉验证分类,指定每个 Trees 最多被分割 5 次。然后对“common”的误分类分配一个高成本值 20 以进行惩罚,即引入置信度的 AdaBoostM2 模型进行对比。

Step2 在 MATLAB 中选用 Bayseopt 工具箱[24],使用 fitcensemble 找到使交叉验证损失最小化 5 倍的超参数,设置随机种子值并使用“expected-improvement-plus”采集函数确定下一个要评估的点,并在置信区域内进行探索。为了重复并可视化,将它们传递到 OptimizeHyperparameters 名称-值对中,需要优化的参数默认为 KernelScale 和 BoxConstraint。

Step3 传递参数作为优化超参数的值后命令行中会出现迭代显示,超参数调优结果如图 4 所示,目标函数为回归的 log(1 + 交叉验证损失)和分类的误分类率。进行迭代以优化超参数、最小化分类器的交叉验证损失,使用经过优化超参数训练的模型预测验证集的类标签,可以看出经过迭代后泛化能力拟合。图 4 中的稍小圆点表明目标点,稍大圆点标明采集函数值最大的位置并以此作为下一个采集点。最佳估计可行点是根据最新模型估计均值最低的采集点,最佳观测可行点是目标函数评价返回值最低的采集点。



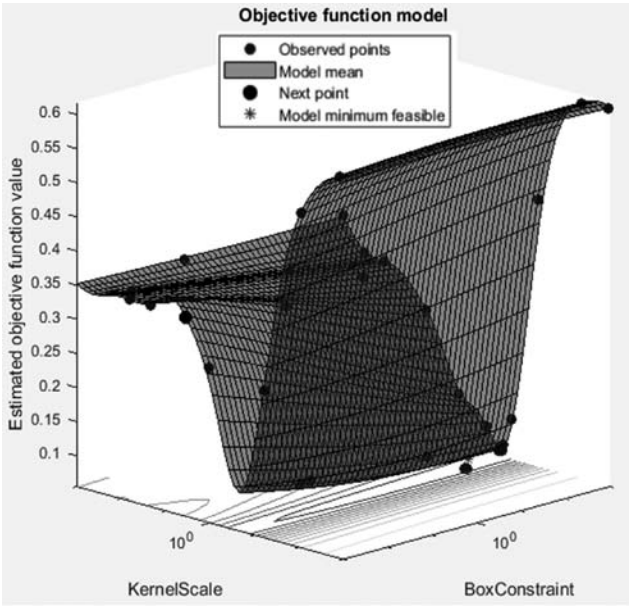


图 4　超参数调优迭代过程和结果

表 5 说明了采用集成分类 AdaBoostM2 经过贝叶斯超参数调优后最佳估计可行点和最佳观测可行点的比较结果。可以看出准确率由 93.45% 提升到了 97.49%,函数计算时间成本约降低了 14 s,优化效果明显。

表 5　超参数调优后最佳估计可行点和最佳观测可行点比较

比较值	最佳观测可行点	最佳估计可行点
集成多分类方法	AdaBoostM1	AdaBoostM2
迭代次数	39	41
准确率	0.944 51	0.974 94
目标函数值	0.002 018 4	0.002 099 6
计算时间/s	177.931	163.806 4

Step4　利用 MATLAB 中的混淆矩阵生成函数 Confusion Matrix 和热图生成函数 Heatmap 将经过训练的模型预测验证集的类标签,生成优化后的多分类混淆矩阵并可视化,如图 5 所示。

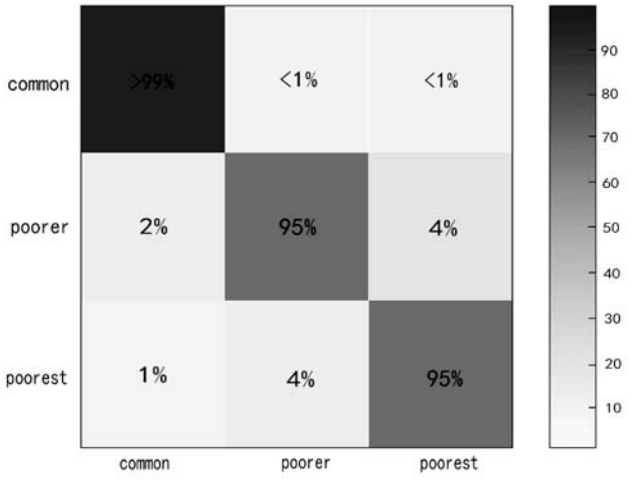


图 5　模型优化后的多分类标签混淆矩阵

从优化后的多分类标签混淆矩阵可以看出,经过 NCA 降维后引入成本函数惩罚并用贝叶斯超参数优化后的模型将初步模型 8% 的普通学生分类为贫困和极度贫困误报率减少到 5%,模型的准确率明显提升,达到了优化效果。

4　结　语

高校贫困生预测判定建模运行了多种算法训练分类器,单独的分类算法会过度拟合训练数据,而且没有一种算法是万能最优,反复训练试错才是选择最佳算法的前提。对比算法模型结果,二次支持向量机(SVM)、线性支持向量机和决策树算法表现略优。使用 NCA 方法降维后,整体预测速度和计算时间变化明显,决策树分类算法表现优异。集成分类算法 Boosted Trees 和 Bagged 是提升泛化能力的合理有效选择。

在初始模型上保留验证数据,使用 AdaBoostM1 和 Trees 模型 k 折交叉验证反复评估,与引入成本函数权重值调整的 AdaBoostM2 模型经贝叶斯超参数调优后对比。高校贫困生预测判定 AdaBoostM2 模型的准确率提升了近 4%,计算时间成本降低了 14 s,误判率由初始的 8% 改进到 5%,说明优化改进后的算法模型的泛化能力得到了一定的改进。

参　考　文　献

[1] 史甜. 数据挖掘在高校贫困生认定系统中的应用研究[D]. 西安:西安科技大学,2017.

[2] 张建明. 基于数据挖掘的高校贫困生认定系统设计和分析[D]. 南京:东南大学,2015.

[3] 杨知玲. 数据挖掘在高校贫困生评价中的应用研究[D]. 广州:华南理工大学,2015.

[4] 李明江,卢玉,刘彦. 一种基于 C4.5 决策树的贵州省高校贫困生评定方法[J]. 科技通报,2013,29(8):223 - 224,233.

[5] 陈晓,王树宝,李建晶,等. 基于加权约束的决策树方法在贫困生认定中的应用研究[J]. 计算机应用与软件,2014,31(12):136 - 139.

[6] 李明君. 基于数据挖掘的贫困助学金认定方法研究[D]. 武汉:华中师范大学,2017.

[7] 唐燕,王苹. 随机森林算法在中医药院校贫困生认定预测中的应用研究[J]. 中国医药导报,2017,14(14):164 - 168.

[8] 王正杰,杨伟丽,王喆,等. 4 种分类算法参数选择及分类特点研究[J]. 计算机与现代化,2018(2):54 - 60.

(下转第 299 页)

参 考 文 献

[1] Senthil K S, Margret A S. QoS-Based Concurrent User-Service Grouping for Web Service Recommendation[J]. Automatic Control and Computer Sciences, 2018, 52(3):220 -230.

[2] Chen W H, Paik I, Yen N Y. Discovering internal social relationship for influence-aware service recommendation[J]. Multimedia Tools and Applications, 2017, 76(18): 18193 -18220.

[3] Chen L, Wu J, Zheng Z B, et al. Modeling and exploiting tag relevance for Web service mining[J]. Knowledge and Information Systems, 2014, 39(1): 153 -173.

[4] Huang Z Z, Li T X, Xiao S. Research on Library Recommendation Reading Service System Based on Adaptive Algorithm[J]. Wireless Personal Communications, 2018, 102(2): 1963 -1977.

[5] Jayapriya K, Mary N A B, Rajesh R S. Cloud Service Recommendation Based on a Correlated QoS Ranking Prediction [J]. Journal of Network and Systems Management, 2016, 24(4):916 -943.

[6] Yu C Y, Huang L P. A Web service QoS prediction approach based on time-and location-aware collaborative filtering[J]. Service Oriented Computing and Applications, 2016, 10(2): 135 -149.

[7] Karimian K, Ashtiani M, Azgomi M A. An approach based on the transferrable belief model for trust evaluation in web services[J]. Soft Computing, 2018, 22(21): 7293 -7311.

[8] 唐明董, 张婷婷, 杨亚涛, 等. 基于因子分解机的质量感知 web 服务推荐方法[J]. 计算机学报, 2018, 41(6): 1080 -1092.

[9] Kim J, Lee D, Chung K Y. Item recommendation based on context-aware model for personalized u-healthcare service [J]. Multimedia Tools and Applications, 2011, 71(2):855 -872.

[10] 胡堰, 彭启民, 胡晓惠. 一种基于隐语义概率模型的个性化 web 服务推荐方法[J]. 计算机研究与发展, 2014, 51(8): 1781 -1793.

[11] 李鸿超, 刘建勋, 曹步清, 等. 融合多维信息的主题自适应 Web API 推荐方法[J]. 软件学报,2018,29(11): 3374 -3387.

[12] Christou I T, Amolochitis E, Tan Z H. AMORE: design and implementation of a commercial-strength parallel hybrid movie recommendation engine[J]. Knowledge and Information Systems, 2016, 47(3):671 -696.

[13] Massa P, Avesani P. Trust-aware boots trapping of recommender system[C]//Proceddings of the ECAI work shop on recommender systems. RicadelGarda, Italy, 2006: 29 -33.

[14] Wang P, Chao K M, Lo C C. Satisfaction-based Web service discovery and selection scheme utilizing vague sets theory [J]. Information Systems Frontiers,2015,17(4):827 -844.

[15] Yoo H, Chung K. Mining-based lifecare recommendation using peer-to-peer dataset and adaptive decision feedback[J]. Peer-to-Peer Networking and Applications, 2018, 11(6): 1309 -1320.

(上接第 287 页)

[9] Wiki-Pedia. ID3-algorithm[DB/OL]. 2018 -7 -6. https://en. wikipedia. org/wiki/ID3_algorithm#See_also.

[10] 卢东标. 基于决策树的数据挖掘算法研究与应用[D]. 武汉:武汉理工大学,2008.

[11] 颜会娟,秦杰. 基于非线性 SVM 模型的木马检测方法[J]. 计算机工程,2011,37(8):121 -123.

[12] Harrington P. Machine learning in Action[M]. 李锐,李鹏,曲亚东,等译. 北京:人民邮电出版社,2013.

[13] Liu Bing. Web Data Mining[M]. 俞勇,薛贵荣,韩定一,译. 北京:清华大学出版社,2009.

[14] Alpaydin E. Introduction to Machine Learning[M]. 范明,咎红英,牛常勇,译. 北京:机械工业出版社. 2009,6.

[15] 咸云浩,张恒德,谢永华,等. 多元逐步回归与卡尔曼滤波法在霾预报中应用[J]. 系统仿真学报,2018,30(4):1482 -1489.

[16] 刘丛山,李祥宝,杨煜普. 一种基于近邻元分析的文本分类算法[J]. 计算机工程,2012,38(15):139 -141.

[17] Zhou H T, Chen J, Dong G M, et al. Bearing fault recognition method based on neighbourhood component analysis and coupled hidden Markov model[J]. Mechanical Systems and Signal Processing,2016,66/67:568 -581.

[18] 邓帅. 基于改进贝叶斯优化算法的 CNN 超参数优化方法[J/OL]. 计算机应用研究,2019,36(7). [2018 -08 -02]. http://kns. cnki. net/kcms/detail/51. 1196. TP. 20180412. 0812. 030. html.

[19] Rasmussen C E, Williams C K I. Gaussian Processes for Machine Learning[M]. MIT Press, 2005.

[20] Shahriari B, Swersky K, Wang Z, et al. Taking the Human Out of the Loop: A Review of Bayesian Optimization[J]. Proceedings of the IEEE, 2016, 104(1):148 -175.

[21] 崔佳旭, 杨博. 贝叶斯优化方法和应用综述[J]. 软件学报, 2018, 29(10):176 -198.

[22] 柴慧敏, 赵昀瑶, 方敏. 利用先验正态分布的贝叶斯网络参数学习[J]. 系统工程与电子技术, 2018, 40(10): 219 -224.

[23] 黄良斌. 高校贫困生认定标准与认定模型研究[J]. 职业教育研究,2012(4):11 -12.

[24] Martinez-Cantin R. BayesOpt: A Bayesian Optimization Library for Nonlinear Optimization, Experimental Design and Bandits[J]. Journal of Machine Learning Research, 2014, 15: 3735 -3739.