基于机器学习的贫困生分类预测研究

陆桂明[1] 张 源[1] 周志敏[2]

[1](华北水利水电大学信息工程学院 河南 郑州 450046)

[2](浙江水利水电学院信息工程与艺术设计学院 浙江 杭州 310018)

摘 要 针对高校贫困生认定工作中存在的问题,利用校园一卡通数据,综合学生消费和生活规律,结合XGBoost(Extreme Gradient Boosting)模型和主成分分析法、过采样算法,建立高校贫困生的分类预测方法。该方法在贫困生分类预测中的准确率较高。实验结果证明,采用 XGBoost 模型比其他模型预测准确率更高,为我国高校贫困学生的评定标准提供重要依据,保证了贫困学生认定工作的公正性。

关键词 主成分分析法 过采样算法 贫困生 XGBoost 模型 分类预测

中图分类号 TP3 **文献标识码** A DOI:10.3969/j.issn.1000-386x.2019.01.055

POOR STUDENTS CLASSIFICATIOIN PREDICTION BASED ON MACHINE LEARNING

Lu Guiming[1] Zhang Yuan[1] Zhou Zhimin[2]

[1](*College of Information Engineering, North China University of Water Resources and Electric Power, Zhengzhou 450046, Henan, China*)

[2](*College of Information Engineering and Art Design, Zhejiang University of Water Resources and Electric Power, Hangzhou 310018, Zhejiang, China*)

Abstract In view of the problems existing in the identification of poor students in colleges, we adopted XGBoost model, principal component analysis and the oversampling algorithm to establish classification prediction methods for poor students in colleges through campus card data, students' consumption and living patterns. This method had a higher accuracy in the classification of poor students. The experimental results prove that the accuracy of XGBoost model is better than other models. It provides important basis for the evaluation criteria of poor students in colleges in China, and ensures the fairness of the identification of poor students.

Keywords Principal component analysis Oversampling algorithm Poor students XGBoost model Classification prediction

0 引 言

贫困生是指大学生在经济上只依靠家庭资助无法完成学业者。通常情况下,学生所在的民政部门提供家庭收入等证明,然后由学生自己向学校提交申请,最终由学校评估和认证[1]。最近几年,由于高校扩招,生源面广,而各地区的经济发展又不平衡,导致高校对贫困生不公平的认定成为一个普遍性的问题。高校仅凭民政部门出具的证明来确认贫困生,其结果可想而知。所以如何确保真正的贫困生享受到国家的支持和帮助,不仅会影响到贫困学生本身之后的进步,也会影响到他的家庭,影响中国高等教育的公平性与整个社会的和谐稳定发展。

目前已有研究人员结合高校大学生一卡通消费数据展开研究。王文娟[2]结合统计学中描述性统计和非参数检验的方法,分析大学生实际消费水平和不同消费水平下的消费相异性。费小丹等[3]运用聚类分析算法建立了贫困学生的指数算法,为高校贫困生认定工作提供指导。李明君[4]通过使用 K 近邻算法,使预测

收稿日期:2018-05-23。浙江省科技厅项目(LGF18F020014)。**陆桂明**,教授,主研领域:水利水电信息化,机器学习。**张源**,硕士。**周志敏**,教授。

出真正贫困生的准确率较高,但是以上方法均忽略了分类不平衡问题。

针对上述研究中存在的问题,本文利用在校大学生的校园一卡通刷卡数据,应用 XGBoost 模型对贫困学生进行分类预测,联合 PCA 方法对数据维度进行降低。同时引入 SMO-TE(Synthetic Minority Oversampling Technique)算法对不平衡数据进行上采样。该方法取得了比较理想的结果,对贫困生认定起到积极的指导作用。

1　数据来源及预处理

1.1　数据来源及概况

校园卡记录在校大学生日常消费行为习惯,能客观、真实地展现在校生的贫困状况。本文选取了某高校 2016 级 - 2017 级学生在 2017 年 9 月到 12 月的一卡通消费数据,总计 150 万,其中包括学生的学号、消费额度、消费地点以及消费时间。从学生处获得 2016 级 - 2017 级学生的助学金发放表,可知贫困生共计 957 人。校园卡数据库表如图 1 所示。

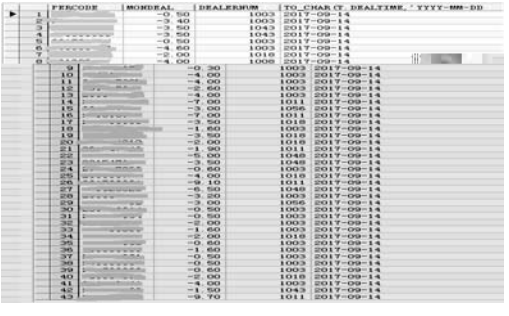

图 1　一卡通数据库

1.2　数据预处理

数据预处理是挖掘数据之前要做的第一件事,包括清理、汇总和特征创建,以及数据离散化、缺失值和异常值处理等[5]。在实际研究中不同的数据集会有不同的特征,因此对数据进行预处理时也会有不同的方式。

(1) 数据清洗　第一步是清洗原始数据。在校园一卡通数据集中,有一部分观测数据的学生 ID、消费类别、消费地点以及消费金额均相同,即出现重复观测数据,对这些出现重复观测的数据集进行删除[6]。

(2) 特征选择　第二步创建新的特征变量。本文选取一卡通数据中的消费方式和消费金额两个特征变量,消费方式共 34 种。将这 34 种消费方式进行整合分类[7],划分为 5 类消费方式:食堂、超市、水果店、咖啡店,其余的划分为其他类消费,记为 1 ~ 5。然后根据学生 ID(学号),日期和消费方法,计算每个月的总消费额度、总消费次数以及平均每月消费额度,总共生成了 61 个特征。生成的 61 个特征存在相关性,降低数据指标会丢弃必要信息,总结出不正确的论断[8]。本文采用 PCA 方法,降低数据维度。PCA 方法将 m 维特征向量映射到 l 维($l < m$),形成跟之前不同的正交特征,l 称为主成分,是重建的 l 维特征[9]。通过 PCA 方法,本文选取 $l = 6$,分别记为 V1 ~ V6。将贫困生类别记为 Class,其中贫困生记为 1,非贫困生记为 0,结合特征 V1 ~ V6,共 7 个特征。

(3) 不平衡数据处理　在 5 463 名学生中,只有 17.8% 的学生获得了补贴,如图 2 所示。样本分布是不均衡的,在算法钻研中,很多算法的根本假定就是数据散布是平均的。若数据不均衡[10],会导致算法在预测时,结果更多偏向数量多的那一类别。为解决这个问题,本文采用 SMOTE 算法对数据上采样,让两个类别的样本达到均衡,提高分类器的学习能力[11]。

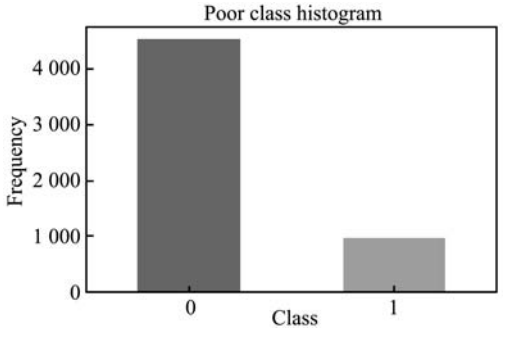

图 2　贫困与非贫困学生的比例

2　XGBoost 模型

本文使用 XGBoost 模型来预测贫困学生的分类,并为识别贫困学生提供依据。XGBoost 是一种改良的 GBDT 算法[12],与 GDBT 相比,XGBoost 将一阶导数与二阶导数结合,同时算法将树模型复杂度作为目标函数里的正则项,来防止过拟合[13]。

XGBoost 模型的目标函数:

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

用泰勒公式展开来近似原来的目标。

泰勒展开:

$$f(x + \Delta x) \simeq f(x) + f'(x)\Delta x + \frac{1}{2}f''(x)\Delta x^2 \quad (2)$$

定义:

$$\begin{cases} g_i = \partial \hat{y}^{(t-1)} l(y_i, \hat{y}^{(t-1)}) \\ h_i = \partial^2 \hat{y}^{(t-1)} l(y_i, \hat{y}^{(t-1)}) \end{cases} \quad (3)$$

决策树复杂度计算公式:

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \quad (4)$$

将式(2)－式(4)代入式(1)得:

$$Obj^{(t)} \simeq \sum_{i=1}^{n}\left[I(y_i,\hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)\right] + \Omega(f_t) + c \tag{5}$$

去掉常数项,化简为:

$$Obj^{(t)} \simeq \sum_{i=1}^{n}\left[g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)\right] + \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{T} w_j^2 \tag{6}$$

最后的目标函数为:

$$Obj^{(t)} = \sum_{j=1}^{T}\left[\left(\sum_{i\in I_j} g_i\right)w_j + \frac{1}{2}\left(\sum_{i\in I_j} h_i + \lambda\right)w_j^2\right] + \gamma T \tag{7}$$

经过对 w_j 求导等于0,能够得到最优的 w_j^*:

$$w_j^* = -\frac{\sum_{i\in I_j} g_i}{\sum_{i\in I_j} h_i + \lambda} \tag{8}$$

然后把 w_j 最优解代入,得到:

$$Obj = -\frac{1}{2}\sum_{j=1}^{T}\frac{\left(\sum_{i\in I_j} g_i\right)^2}{\sum_{i\in I_j} h_i + \lambda} + \gamma T \tag{9}$$

用 *Obj* 寻找具有最佳结构的树并将其添加到模型中,利用贪心算法来找到最优树结构[14]。

每次尝试去向现有叶子添加分割时,增益计算如下:

$$Gain(\phi) = \frac{1}{2}\left[\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{\left(\sum_{i\in I} g_i\right)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma \tag{10}$$

在树的学习中一个重要问题是依据式(10)找到最优的分割算法,又叫精确贪心算法。

3 实验结果与分析

本文应用混淆矩阵对模型结果进行可视化展示[15],如表1所示。

表1 混淆矩阵

	Negative	Positive
False	False Negative(*FN*)	False Positive(*FP*)
True	True Negative(*TN*)	True Positive(*TP*)

混淆矩阵的概念如下:

(1) True Positive(真正):将贫困学生预测为贫困学生。

(2) True Negative(真负)[5]:将贫困学生预测为非贫困学生(漏报)。

(3) False Positive(假正):将非贫困学生预测为贫困学生(误报)。

(4) False Negative(假负):将非贫困学生预测为非贫困学生[13]。

在XGBoost模型中,采用SMOTE算法对数据进行上采样,再按7:3比例划分训练、测试集。混淆矩阵如图3所示。

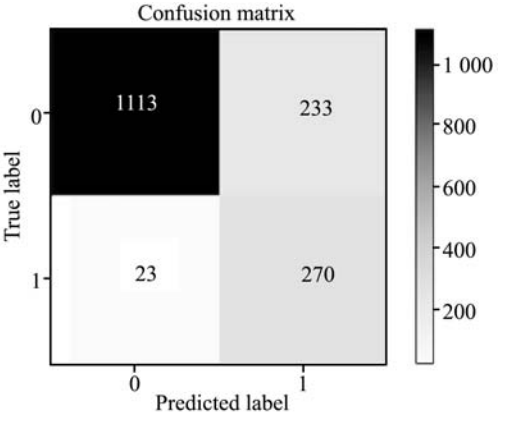

图3 XGBoost 混淆矩阵

通过计算可得 XGBoost 模型的查准率和查全率为:

$$Precison = \frac{TP}{TP+FP} = 53.68\% \tag{11}$$

$$Recall = \frac{TP}{TP+FN} = 92.15\% \tag{12}$$

基于以上模型评估方法,结合研究内容,本文采用 *FP*、*FN*、*Precision*、*Recall* 与 *F*1 来综合评价以下3种分类预测模型的性能,从中选出最优模型,如表2所示。

表2 3种分类模型的比较结果

评价值	模型1 (Logistic Regression)	模型2 (SVM)	模型3 (XGBoost)
Precision	23.23%	27.94%	53.68%
Recall	58.36%	75.43%	92.15%
*F*1	33.04%	40.77%	67.84%
FP	565	570	233
FN	781	776	1 113

查准率(Precision)表现为预测为贫困学生的样本中有多少是真的贫困生[5]。*Precision* 越高,这个模型辨认负面样本的能力就越好。从查准率来看,用来分析的数据只是校园一卡通数据,但学生消费是多元化的,比如网络消费、校外就餐等,因此会出现非贫困生类别的学生比贫困生消费更低。可以看出,模型1和模型2的准确度未达到50%,实验结果不尽人意;模

型 3 的查准率比较高,说明该模型能够以较高的精准度预测真正需要帮助的贫困生。

查全率(Recall)表现为样本中的贫困学生有多少被预测为贫困学生,*Recall* 越高,这个模型对正样本的辨认能力越高。从查全率来看,3 个模型的 *Recall* 都达到50%以上,模型2 相较模型1 来说,查全率提高了大约20%,模型3 是3 个模型中查全率最高的模型,说明该模型预测的贫困生中已发放助学金的学生比例达92%以上,结果很理想。

为了评估不同算法的优缺点,提出了基于 Precision 和 Recall 的 F1 值概念,对 *Precision* 和 *Recall* 进行综合评价,F1 值是两者的综合,F1 值越高,表示分类模型越稳定。F1 值的定义如下:

$$F1 = \frac{2TP}{2TP + FN + FP} = \frac{2Precision \cdot Recall}{Precision + Recall} \qquad (13)$$

从 F1 值来看,3 个模型的 F1 值分别为 33.04%、40.77% 和67.84%,可知,模型3 相较其他两个模型来说是最优的。

从 *FP* 和 *FN* 来看,模型 1 对非贫困生的预测结果较差,无法准确预测出非贫困生,模型 2 比模型 1 要好;模型 3 比其他两个模型更加能准确预测非贫困生,在"真正非贫困"学生的发现效果最好,而且预测的准确率较高,在样本数据中能很好地区分出非贫困生。

综合以上 5 个评价标准,可以看出模型 3 的分类预测能力最理想,说明模型 3 泛化能力最强,能够很大程度地找准并找全需要资助的贫困学生。

4　结　语

本文针对目前贫困学生认定中存有的问题,将机器学习算法与贫困生识别工作结合。在数据预处理阶段,应用 PCA 的特征降维来削减特征数,有效降低噪声、冗余和过拟合。在不平衡的数据集中用 SMOTE 算法对数据上采样,避免过度拟合分类器。采用 XGBoost 模型获得更好的分类结果,为贫困学生的认定提供了积极的指导意义。在分类模型的构造上,XGBoost 模型的 *Recall* 很高,但 *Precision* 有待继续优化,可以尝试采用组合分类模型来进行比较。同时,还应该研究各种模型的稳定性,并选择一个更好的模型。下一步将致力于相关领域的技术研发,并将不断完善和提高。

参　考　文　献

[1] 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.

[2] 王文娟. 基于一卡通数据的大学生消费分析的技术路线研究与实例分析[D]. 大连:大连医科大学,2013.

[3] 费小丹,董新科,张晖. 基于校园一卡通消费数据的高校贫困生分析[J]. 电脑知识与技术,2014,10(20):4934-4936.

[4] 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.

[5] Han J, Kamber M. Data Mining: Concepts and Techniques [M]. USA: Morgan Kaufmann Publishers, 2001.

[6] 曾军,周国富. 基于机器学习的多语言文本抽取系统实现[J]. 计算机应用与软件,2017,34(4):87-92,156.

[7] 吴玉强,田素诚. 基于代价敏感鉴别字典学习的入侵检测方法[J]. 科技通报,2017,33(12):162-166.

[8] 李建林. 一种基于 PCA 的组合特征提取文本分类方法[J]. 计算机应用研究,2013,30(8):2398-2401.

[9] 戴炳荣,王晓丽,李超,等. 一种基于 PCA-SVM 的医疗卫生数据挖掘分类方法[J]. 计算机应用与软件,2016,33(8):67-70.

[10] 秦孟梅,邱建林,陆鹏程,等. 基于 AdaBoost 的类不平衡学习算法[J]. 计算机应用研究,2017,34(11):3229-3232,3254.

[11] Prusty M R, Jayanthi T, Velusamy K. Weighted-SMOTE: A modification to SMOTE for event classification in sodium cooled fast reactors[J]. Progress in Nuclear Energy, 2017, 100:355-364.

[12] Chen T, Guestrin C. XGBoost: a scalabletree boosting system[C]//ACM SIGKDD International Conferenceon Knowledge Discovery and DataMining,2016:785-794.

[13] Torlay L, Perrone-Bertolotti M, Thomas E, et al. Machine learning—XGBoost analysis of language networks to classify patients with epilepsy[J]. Brain Informatics, 2017, 4(3): 159-169.

[14] Tharwat A, Moemen Y S, Hassanien A E. Classification of toxicity effects of biotransformed hepatic drugs using whale optimized support vector machines[J]. Journal of Biomedical Informatics, 2017, 68:132-149.

[15] 刘晓蔚. 数据挖掘预测模型在脑损伤患者认知功能康复中的应用[J]. 计算机应用与软件,2014,31(12):221-224,282.

(上接第 210 页)

[23] Rogez G, Supancic J S, Ramanan D. Understanding everyday hands in action from rgb-d images[C]//Proceedings of the IEEE International Conference on Computer Vision, 2015: 3889-3897.

[24] Henry P, Fox D, Bhowmik A, et al. Patch volumes: Segmentation-based consistent mapping with RGB-D cameras [C]//Proceedings of the International Conference on 3D Vision, 2013: 398-405.

[25] Felzenszwalb P F, Huttenlocher D P. Efficient graph-based image segmentation[J]. International Journal of Computer Vision, 2004, 59(2): 167-181.