Hans Journal of Data Mining 数据挖掘, 2020, 10(4), 254-260
Published Online October 2020 in Hans. http://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2020.104027

Hans汉斯

基于Stacking算法实现信贷不平衡数据分类

郑利沙，黄　浩

对外经济贸易大学，北京
Email: 1742332599@qq.com

收稿日期：2020年9月6日；录用日期：2020年9月20日；发布日期：2020年9月28日

摘　要

随着大数据技术在应用层面的日渐普及，机器学习、深度学习相关算法在金融风控行业的应用得到了积极的探索。本文基于开源的信用卡数据(该数据具有样本比例极度不平衡的特点)，比较不同采样方法对类别不平衡数据分类结果的影响，并应用集成学习算法Stacking融合多个基分类器训练数据，得到更为稳健的分类模型，有效避免了过拟合现象的发生。

关键词

样本不平衡数据，集成学习，Stacking

Classification of Credit Imbalance Data Based on Stacking Algorithm

Lisha Zheng, Hao Huang

Foreign Economic and Trade University, Beijing
Email: 1742332599@qq.com

Received: Sep. 6th, 2020; accepted: Sep. 20th, 2020; published: Sep. 28th, 2020

Abstract

With the increasing popularity of big data technology at the application level, the application of machine learning and deep learning related algorithms in the financial risk control industry has been actively explored. Based on open source credit card data (the data has the characteristics of extremely unbalanced sample ratios), this paper compares the impact of different sampling methods on the classification effect of different classification algorithms in the binary classification problem of unbalanced data, and applies ensemble learning algorithm to fuse multiple base clas-

文章引用: 郑利沙, 黄浩. 基于 Stacking 算法实现信贷不平衡数据分类[J]. 数据挖掘, 2020, 10(4): 254-260.
DOI: 10.12677/hjdm.2020.104027

sifier training data. A more robust classification model is obtained, effectively avoiding the occurrence of overfitting.

Keywords

Sample Unbalanced Data, Integration Learning, Stacking

1. 引言

信用卡欺诈模型是用于检测用户信用卡被盗用的风险，信用卡欺诈识别问题实质上是一个二分类问题。随着 2016 年央行发布的“信用卡新规”执行以来，信用卡市场再一次被激活，信用卡的使用率保持增长。信用卡欺诈识别是长久以来亟待解决的问题之一，传统的信用卡欺诈方式主要是伪卡盗刷，但在新兴互联网时代这一方式不再是主流。信用卡欺诈行为的检测已成为一个不容忽视的问题，这一问题不但会给银行造成巨大的经济损失，而且还会影响银行的信誉，给银行带来潜在的风险。

当前，银行业风险防控得到了广泛的关注，作为银行系统风险防控中的关键一环，信用卡欺诈行为的检测发挥着重要的作用，引起了众多学者的关注。经典的机器学习算法，如逻辑斯蒂回归、决策树、支持向量机以及集成学习方法等被应用到信用卡欺诈行为的研究中；深度学习网络，如 GAN 网络也被应用于预测银行信用卡欺诈。分析已有的论文发现，基于 Stacking 的集成学习算法在信用卡欺诈检测上的应用较少，大部分论文是基于 bagging 或基于 boosting，受此现象启发，本文基于 Stacking 集成学习算法的思想，融合多种常用分类算法实现信用卡欺诈数据二分类，有效解决了不平衡数据容易过拟合的问题。

2. 相关工作

信用卡欺诈检测是典型的样本不平衡分类问题，到目前为止有不少学者已经对信用卡欺诈检测和样本不平衡数据分类问题做出了相关的阐述。在信用卡欺诈检测中，徐永华[1]提出一种基于支持向量机的信用卡欺诈检测算法，用以提高检测的准确率。李赛虎等人[2]基于特征工程提出一个取决于样本实际财务成本的代价矩阵，并针对决策树(DT)、Logistic 回归和随机森林三种代价敏感性分类算法检测信用卡欺诈。陈冠宇[3]提出一种基于 KNN-Smote-LSTM 的信用卡欺诈检测模型，通过 KNN 判别分类器筛选出样本以提升模型的性能，克服了 Smote 算法在生成新样本时的盲目性和局限性。Brause [4]提出一种基于规则，并结合自适应神经数据挖掘方法的信用卡欺诈检测方法。Kokkinaki [5]利用决策树和布尔逻辑函数构造单用户模型，结合聚类分析挖掘正常和异常交易的区别，从而来判断用户的交易是否正常。Maes 等人[6]进行了人工神经网络和贝叶斯网络在信用卡欺诈检测中的性能对比。除此之外，还有很多学者对不平衡数据的处理进行了相关研究。相关研究人员提出了经典的少类样本合成过采样技术(Synthetic minority oversampling technique, SMOTE)，随后为了解决新样本带来的噪声、样本重叠等问题，有学者相继提出了一系列 SMOTE 的改进算法，如 N-SMOTE、Borderline-SMOTE 等；Batista 等[7]出了 SMOTE 与 Tome Link 相结合的算法，能很好地克服 SMOTE 带来的噪声问题。利用采样解决样本不平衡问题是分类算法中最常使用的方法，一般而言，采样方法能够很好地提升常见分类算法的性能，但对于集成学习算法来

说这样的说法还有商榷。

3. 模型介绍

3.1. 集成学习

集成学习是使用一系列学习器进行学习，并使用某种规则把单个学习器的结果进行整合，从而获得比单个学习器更好的效果的一种及其学习算法，集成学习算法的分类有基于 bagging 的集成学习方法、基于 boosting 的集成学习方法和基于 Stacking 的集成学习方法。

3.1.1. Bagging (Bootstrap Aggregating，装袋法)

Bagging 即是套袋法，使用有放回的抽样方法从原始样本中随机抽取训练集，进行 k 轮抽取，得到 K 个训练集，每个训练集得到一个模型，共 K 个模型。Bagging 的性能依赖于其基分类器的稳定性，通过降低基分类器的方差，能够有效改善模型的泛化误差。常见的基于 bagging 的集成学习方法有随机森林。

3.1.2. Boosting

Boosting 算法的核心思想是将多个弱分类器组合成强学习器，通过提高在前一轮训练中被弱分类器错分样例的权值，较小前一轮正确分类的样本的权值，使分类器对误分的数据有很好的分类效果。常见的基于 boosting 的集成学习方法有 adaboost、GBDT (梯度提升树)算法。

3.1.3. Stacking

Stacking 方法是指先训练多个模型，然后把训练所得的输出作为输入来训练一个模型，得到最终的输出，在实际应用中，我们通常使用 logistic 回归作为组合策略。

3.2. 采样方法

由于数据分布不均匀，少数类样本仅占很小一部分，模型很难辨别少数类，容易造成过拟合。目前解决样本不平衡数据分类的手段主要是采样，包括过采样、欠采样和综合采样。

3.2.1. 过采样

过采样是对原始数据中的少数类进行采样，以合成新样本来缓解类别不平衡问题。常见的过采样方法有随机过采样、SMOTE (Synthetic Minority Oversampling Technique)，即合成少数类过采样技术、Border-line SMOTE、自适应合成抽样 ADASYN 等。

3.2.2. 欠采样

欠采样是对原始数据中的多数类进行采样，较少多数类样本以使数据类别分布达到平衡。欠采样的分类有随机欠采样、EasyEnsemble、BalanceCascade、NearMiss、Tomek Link、Edited Nearest Neighbours (ENN)等。

3.2.3. 综合采样

综合采样方法是将欠采样和过采样结合，利用欠采样中的数据清洗技术如 ENN 处理过采样中出现的重叠样本，得到更为准确的分类数据。综合采样是处理不平衡分类数据最常使用的方法，常见的综合采样方法有 SMOTE + ENN 和 SMOTE + Tomek Link Removal。

本文选取 smote + ENN 作为不平衡数据的采样方法。ENN 针对类别比例大的样本进行操作，如果某一个样本的大部分 K 邻近样本都跟它自身的类别不一致，说明这个样本处于类别边缘交界处或者少数类别簇中，即可以删除这个样本。

3.3. 基分类器

3.3.1. K 近邻算法

K 近邻算法是一种基本的分类和回归算法，其原理是给定一个训练数据集，当新实例输入时，首先在给定的训练集中找到与该实例最邻近的 K 个实例，然后判断这 K 个实例中的多数实例属于哪一类，把新实例划分为同样的类中。

3.3.2. 逻辑斯蒂回归

逻辑斯蒂回归是针对线性可分问题的常用分类算法，通过训练数据样本，学习样本特征到样本标签之间的假设函数，是机器学习算法容易实现且表现优异的算法。对于给定的输入实例 x，条件概率分布如下

$$P\left(Y=1|x\right)=\frac{\exp\left(w*x+b\right)}{1+\exp\left(w*x+b\right)}$$

$$P\left(Y=0|x\right)=\frac{\exp\left(w*x+b\right)}{1+\exp\left(w*x+b\right)}$$

根据样本数据求出 $P\left(Y=1|x\right)$ 和 $P\left(Y=0|x\right)$，通过比较两个条件概率值的大小，将 x 划分为概率值大的那一类。

3.3.3. 贝叶斯算法

贝叶斯模型是以概率论上的贝叶斯原理为基础来进行分类的算法。朴素贝叶斯算法的假设前提是：每一个属性在目标值相同时是相互独立的。朴素贝叶斯算法的优点是时间和空间复杂度相比较于其他算法而言较低，算法的逻辑清晰，容易理解。但由于朴素贝叶斯分类的最小错误率是在条件独立的假设前提下发生的。在现实情况下进行分类时，数据之间往往是不独立的，因此朴素贝叶斯算法的分类效果会受到一定程度的影响。

3.3.4. 决策树和随机森林

决策树是一种树形结构，使用决策树算法进行分类的过程就是从根节点开始，测试待分类项中相应的特征属性，并按照其值选择输出分支，直到到达叶子节点，将叶子节点存放类别作为决策结果。

随机森林(Random Forest, RF)是对决策树算法的改进，算法的本质是由多棵决策树整合形成的多分类器，通过决策树的投票结果决定数据的类别。随机森林是采用重采样方法，从训练数据中有放回地抽取固定数量的样本生成样本子集，使用生成的样本子集生成多棵决策树，进行训练。当输入测试集时，每一棵决策树都进行判断，最后统计所有决策树判断的类别总数，最多的分类类别即是测试集的类别。随机森林解决了决策树算法泛化能力差的缺点。

4. 模型评价标准

分类任务中常用的评价指标有准确度、精确率、召回率、F1 Score、P-R 曲线等，针对不同的训练集，需要根据数据的特性采用不同的评价指标。本文采用的数据集存在样本类别不平衡的特性，故选取精确率和 F1 值的平均值作为 Stacking 集成学习模型的分类评价指标。

算法分类的精确率是直观地衡量一个算法分类结果好坏的指标。对于一个分类模型，其训练过程是首先用划分好的训练集进行训练，然后使用测试集测试模型对于新样本的分类结果，在测试集上算法分类的准确度被认为是该算法的分类准确度。准确度作为分类算法中最常使用的指标，对于正负样本数量相当的数据集的分类结果的衡量比较可靠，本文通过采样平衡样本数据，使用交叉验证训练模型，并求

解交叉验证后的平均精确率作为评价指标。

F1-score 是精确率(precision)和召回率(recall)的调和平均值，是分类结果的综合评价指标。F1 值的适用范围很广，无论是类别不平衡数据还是样本比例相当的数据都可适用。

$$F1_{\text{score}} = 2\frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

5. 实验过程

5.1. 数据描述及预处理

本文使用的是开源的信用卡交易数据进行训练，数据集中包含正常交易样本 284,315 条，异常交易的样本 492 条，异常交易比率为 0.173%。该数据集是信用卡反欺诈模型最常使用的数据之一，具有代表性。

数据预处理过程包括：数据清洗，包括重复数据的删除、缺失数据的填充；数据转换，对数据集中数值相差较大的特征列进行标准化处理，并使用主成分分析(Principal Components Analysis, PCA)对高维特征进行降维处理。

5.2. 实验结果

5.2.1. 不同单模型的分类结果

Stacking 集成学习算法的常使用 Logistic 回归作为元模型，基模型可以是任何一种分类模型。模型的选取将影响分类结果，为此我们将训练集输入到不同的分类模型中进行训练，得到各个单模型的分类结果，如表 1 所示。

Table 1. Single model classification
表 1. 单模型分类结果

模型	F1-score	Accuracy
KNN	0.6	1
SVM	0.5	1
贝叶斯	0.62	0.99
决策树	0.88	1
逻辑回归	0.85	1
随机森林	0.85	1

如上表，在此数据集下，K 邻近、支持向量机、贝叶斯算法分类的 F1 值分别为 0.6，0.5，0.62，相比较于决策树、逻辑回归和随机森林算法(F1 值分别为 0.88，0.85，0.85)而言，分类表现欠佳；而六种单模型除贝叶斯算法外分类的准确度都为 1，存在过拟合的现象，验证了准确度并不适合作为类别不平衡数据分类评价指标的观点。

5.2.2. 不同采样方法下各单模型的分类结果

采样方法在一定程度上会对数据集的分类结果产生影响，但从目前查询的资料来看，并未有学者对采样方法在具体分类算法上的表现进行阐述。因此，本文通过比较不同采样方法在 5 种常见分类算法中的性能表现，选择效果最优的采样方法(结果见表 2)。

Table 2. Sampling with single model classification
表 2. 采样 + 单模型分类结果

模型	未采样	SMOTE	SMOTE + ENN	SMOTE + Tmoek link
KNN	0.6	0.91	0.94	0.91
贝叶斯	0.62	0.54	0.52	0.53
决策树	0.88	0.93	0.96	0.93
逻辑斯蒂回归	0.85	0.54	0.55	0.54
随机森林	0.85	0.54	0.55	0.54

我们选取过采样技术 SMOTE、综合采样技术 SMOTE + ENN 和 SMOTE + Tomek Link Removal 作为三种主要的采样方法。如表 2 所示，KNN 和决策树算法采样之后的分类结果优于未采样时；除贝叶斯算法外，其余四种算法在 SMOTE + ENN 采样方法下的性能表现略优于 SMOTE 和 SMOTE + Tomek Link Removal。随机森林算法经过采样之后的分类性能比采样前更差，这在一定程度上可以说明采样方法并不会对集成学习算法在类别不平衡数据上的分类结果产生提升作用。

5.2.3. Stacking 集成学习算法分类结果

基于 Stacking 集成学习算法的思想，对比分析采样与否对样本不平衡数据分类效果的影响。Stacking 的优点在于能够“取长补短”，故选取在同一数据集下在二分类中表现不一样的单模型，尽可能地提高模型的泛化能力。我们以 logistic 回归模型作为 Stacking 的组合策略，使用不同的基分类器进行组合，同时根据 5.2.2 的实验结果选取综合采样方法 SMOTE + ENN 作为对照组，对不同组合的分类结果进行分析。我们将预处理后的数据输入 Stacking 集成学习算法中进行训练，得到的分类结果如表 3。

Table 3. Stacking learning algorithm
表 3. Stacking 集成学习算法

方法	平均 F1 值	平均准确度
KNN + 贝叶斯 + 决策树(SMOTEENN)	0.34	0.50
KNN + 贝叶斯 + 决策树(未采样)	0.67	0.67
KNN + 决策树 + 随机森林(SMOTEENN)	0.42	0.55
KNN + 决策树 + 随机森林(未采样)	0.75	0.76

如上表所示，未采样的 KNN + 决策树 + 随机森林组合模型得到的平均准确度和平均 F1 值相比其余三种组合策略来说是最高的，其值分别达到了 0.75 和 0.76，分别比采样后的 KNN + 决策树 + 随机森林组合模型高 0.27 和 0.21。采样后两种组合策略的 Stacking 模型平均 F1 值分别为 0.5 和 0.55，平均准确度都不超过 0.5，分类表现并不突出，这意味着采样方法对集成学习算法 Stacking 分类性能的提升并没有显著的促进作用。未采样的数据在 Stacking 模型上进行训练得到的结果相较而言更稳健。

6. 结论

使用常见的分类模型进行不平衡数据的分类容易造成过拟合问题，分类模型很难识别少数类样本，在信用卡欺诈这类更重视少数类样本识别的实际应用问题中，常见的分类算法并不具备现实应用意义。本文通过不同的对照实验，验证了准确度不适合作为类别不平衡数据分类评价指标的观点，并发现在 Stacking 集成学习方法下是否对不平衡数据采样对分类结果没有太大的影响。Stacking 集成学习算法虽然

牺牲了时间复杂性，但在很大程度上提高了类别不平衡数据分类的准确性，得到了更为稳健的分类模型。未来可以尝试缩短 Stacking 的运行时间，研究效率更高的分类模型。

参考文献

[1] 徐永华. 基于支持向量机的信用卡欺诈检测[J]. 计算机仿真, 2011, 28(8): 376-379.

[2] 李赛虎, 张丽娟. 基于特征工程的信用卡欺诈检测策略研究[J]. 现代电子技术, 2019, 42(15): 175-180.

[3] 陈冠宇. 基于 kNN-Smote-LSTM 的信用卡欺诈风险检测网络模型[D]: [硕士学位论文]. 杭州: 浙江工商大学, 2018.

[4] Brause, R.W., Langsdorf, T.S. and Hepp, H.M. (1999) Credit Card Fraud Detection by Adaptive Neural Data Mining. Universitätsbibliothek Frankfurt am Main, Frankfurt am Main.

[5] Kokkinaki, A.I. (1997) On Atypical Database Transactions: Identification of Probable Frauds Using Machine Learning for User Profiling. *IEEE Knowledge and Data Engineering Exchange Workshop, Proceedings*, Newport Beach, CA, 4 November 1997, 107-113. https://doi.org/10.1109/KDEX.1997.629848

[6] Maes, S., Tuyls, K., Vanschoenwinkel, B., *et al.* (2002) Credit Card Fraud Detection Using Bayesian and Neural Networks. *Proceedings of the* 1*st International NAISO Congress on Neuro Fuzzy Technologies*, Havana, 16-19 January 2002, 261-270.

[7] Batista, G.E., Bazzan, A.L. and Monard, M.C. (2003) Balancing Training Data for Automated Annotation of Keywords: A Case Study. *Conference*: *II Brazilian Workshop on Bioinformatics*, Macaé, 3-5 December 2003, 10-18.