文章编号：2095-2163(2023)09-0085-04 中图分类号：G356;TP391.1 文献标志码：A

基于 Stacking 集成学习的中文问句分类算法

刘佳梅，丁 楷
（中国航天科工集团六院 情报信息研究中心，呼和浩特 010000）

摘 要：为提升中文问句分类的效果，改善单模型问句分类受训练数据及模型参数影响大、场景适应性差、泛化能力弱等问题，本文提出一种基于 Stacking 集成学习的中文问句分类算法。模型使用集成学习 Stacking 框架，融合 LightGBM、XGBoost 和 Random Forest 构建多基分类器，并利用 Logistic Regression 作为元分类器，实现中文问句分类，以提高模型的泛化能力，并提升分类精度。通过网络开源中文问句数据集对模型进行训练和验证，实验结果表明，本文提出的基于 Stacking 的中文问句分类模型相比于最优 LightGBM 单模型，在 *F*1 值上提升了 2.82%。因此，基于 Stacking 集成学习的中文问句分类算法能够有效提升中文问句分类的精度，支撑问答系统实现更好的性能。

关键词：问答系统；中文问句分类；集成学习；Stacking

Chinese question classification algorithm based on stacking integrated learning

LIU Jiamei, DING Kai
(Information Research Center, The Sixth Academy of China Aerospace Science and Industry Corporation, Hohhot 010000, China)

【Abstract】 In order to improve the effect of Chinese question classification, improve the problem that single model question classification is greatly affected by training data and model parameters, poor scene adaptability, and weak generalization ability. This paper proposes a Chinese question classification algorithm based on Stacking integrated learning. The model uses the integrated learning Stacking framework to integrate LightGBM, XGBoost, and Random Forest to build a multi-base classifier, and uses Logistic Regression as a meta-classifier to achieve Chinese question classification, so as to improve the generalization ability of the model and improve the classification accuracy. The model is trained and verified by using the network open-source Chinese question dataset. The experimental results show that the Chinese question classification model based on Stacking proposed in this paper improves the F1 value by 2.82% compared with the optimal LightGBM single model. Therefore, the Chinese question classification algorithm based on Stacking integrated learning can effectively improve the accuracy of Chinese question classification and support the question-answering system to achieve better performance.

【Key words】 question answering system; chinese question classification; ensemble learning; Stacking

0 引 言

问句分类作为问答系统中问题分析和处理的首要环节，是问答系统尤为重要的一部分，其分类精度会直接影响到问答系统的性能[1]。近年来，随着机器学习和深度学习的快速发展，极大的丰富了问句分类的相关研究。

问句分类所要解决的主要问题是将用户所提的问句划分到预定义的类别，从而确定问题的类型。问句分类是将问句与预定义的问句类别进行匹配的过程，如果问句分类过程用 G 表示，预定义的问句类型用 $\{C_1、C_2、C_3、C_n\}$ 表示，则该过程就是将用户所提问句 X 与某个问句类别 C_i 进行对应，如公式(1)所示：

$$G:X \rightarrow \{C_1、C_2、C_3、\cdots、C_n\} \tag{1}$$

问句分类的 3 种方法为：基于规则、基于统计和基于神经网络的问句分类方法。基于规则的问句分类方法，通过编写分类规则，从而确定问句类别[2]。这类方法往往能够取得较高的准确率，但编写规则需要花费大量的人力物力。基于统计的问句分类方法，是通过训练数据对模型进行拟合，从而实现将问句划分到具体类别。目前，关于统计的问句分类方法有贝叶斯模型[3]、支持向量机[4]、最大熵模型[5]等。基于神经网络的问句分类方法则是通过搭建神

作者简介：刘佳梅(1988-)，女，硕士，高级工程师，主要研究方向：知识管理；丁 楷(1994-)，男，硕士，助理工程师，主要研究方向：知识图谱。

收稿日期：2022-10-08

经网络模型并进行训练。如:Kalchbrenner 等人[6]采用 DCNN 模型;Le 等人[7]采用卷积神经网络;Zeng 等人[8]采用主题记忆网络;Ravi 等人[9]采用图神经网络来进行问句分类。以上关于问句分类的方法均采用单模型,模型的分类效果受训练数据及模型参数影响较大。

综合上述分析,现提出基于 Stacking 集成学习的中文问句分类算法,通过集成多个分类器,实现中文问句的分类。该模型基于集成学习的思想,采用模型融合的方式训练分类器,以提高模型的泛化能力,并提升中文问句分类的准确性。

1 集成学习模型

集成学习是指通过采用一定的策略,组合多个分类模型,形成一个性能更优的集成分类模型。通常称被组合的多个分类模型为“基分类模型”,而最终生成的分类模型为“融合分类模型”。目前,集成学习方法大致分为 3 类[10]:Bagging、Boosting 和 Stacking。其中 Stacking 相比于其它两种方法更为灵活,使其具有更为广泛的应用。在自然语言处理的分类任务中,使用单模型容易受到一些不可控因素影响,使得分类准确率不高[11]。因此,为了有效减少单个模型中随机因素的影响、提高模型的预测精度和可信度,一些学者使用不同的集成学习模型来提高分类精度。

Stacking 集成学习的主要思想是:训练一个分类器来拟合所有预测器的预测结果,而不是使用一些简单的函数(比如硬投票)来聚合所有预测器的预测。Stacking 使用交叉验证数据集来训练基分类模型,然后使用基分类模型在验证集上进行预测,将得到的预测结果作为新的训练数据,最后在新的训练数据上训练融合模型。该方法的作用能够提高模型的泛化能力,所使用的基分类模型可以是相同或不同的任意模型。1992 年 Wolpert[12]最早提出 Stacking 算法,多项研究表明,Stacking 能够得到接近最优的分类模型,但对于确定基分类模型的种类、数量、融合分类模型的选择均依赖实践。谢文涌[13]在马兜铃酸及其类似物鉴别中,采用 Stacking 集成学习分类模型比 K 近邻等单分类模型的平均鉴别正确率高出 8.23%,并且在精确率、召回率和 $F1$ 值均表现最优;胡晓丽[14]以集成学习 Stacking 融合模型,对电子商务平台新用户重复购买行为进行预测,实验结果表明 Stacking 融合模型比单一模型在准确率和 AUC 值上都平均提升了 0.4%~2%。众多研究均表明,使用 Stacking 集成学习框架来融合多模型,往往会得到比单模型更好的效果。

2 基于 Stacking 集成学习的中文问句分类

本文选择 Stacking 集成学习框架来融合多模型进行中文问句分类,来弥补单分类器对问句分类效果的影响。对于 Stacking 集成学习模型的选择包括两部分:基分类器和元分类器。

根据文献[15]可知,决策树在中文短文本分类上,能比同类算法取得更好的效果。中文问句属于中文短文本范畴,因此基分类器采用以决策树算法为基础进行优化提升后的经典机器学习算法。其中包括 LightGBM 算法、XGBoost 算法和 Random Forest 算法。在基分类器的选择上,LightGBM 是利用决策树进行迭代训练得到的强分类器,具有不易过拟合,分类效果好的特点,被广泛的应用于多分类任务中;XGBoost 是利用 boosting 方法将决策树进行集成,训练得到的强分类器;Random Forest 也是决策树的集成算法,具有泛化能力强、不易过拟合的优点。元分类器采用回归算法,选择对于线性问题有较好效果的 Logistic Regression 算法。

综上所述,将 LightGBM、XGBoost 和 Random Forest 作为 Stacking 集成学习的基分类器,元分类器选择使用最为广泛的 Logistic Regression。

(1)LightGBM 算法

LightGBM(Light Gradient Boosting Machine)是轻量级梯度提升机,其原理是将损失函数的负梯度作为当前决策树的残差近似值,拟合形成新的决策树。其优点在于减少对于内存的消耗,加快训练速度。相比于决策树,预测精度得到提升,使得该算法同时兼顾了训练速度和预测精度。

(2)XGBoost 算法

XGBoost(Xtreme Gradient Boosting)极限梯度提升算法,其实质也是基于决策树的算法。算法的优点是能够对输入数据进行并行处理,提高了处理速度,并且有多种机制去防止训练过程中的过拟合,其预测精度也得到大大提升。

(3)Random Forest 算法

Random Forest 算法是决策树的集成,通过多棵决策树对输入数据的训练,从而进行预测的一种算法。Random Forest 算法通常采用 Bagging 的方法进行训练,该算法的优点是在绝大多数的分类任务中,能够取得较好的效果,在处理高维度数据时,不容易产生过拟合现象,并且训练速度相对较快,能够适应

大数据分析。

(4) Logistic Regression 算法

Logistic Regression 算法同线性回归算法非常类似。但线性回归与 Logistic Regression 解决的问题类型存在差异。线性回归处理的是数值问题，而 Logistic Regression 处理的是分类问题。换个角度讲，线性回归输出结果是连续的，而 Logistic Regression 输出结果是离散的。例如，判断某天的天气是否会下雨，其结果只有是或否，所以 Logistic Regression 是一种较为经典的二分类算法。为了满足 Logistic Regression 输出的需要，对线性回归的计算结果加上一个 Sigmoid 函数，就可以将数值结果转化为输出 0~1 的概率，然后根据这个概率再做判断。由于 Logistic Regression 拥有很强的泛化能力，在解决分类任务的基础上，可以大大降低集成学习的过拟合风险，因此将 Logistic Regression 模型作为集成学习 Stacking 预测模型的元分类器。

基于 Stacking 的中文问句分类模型如图 1 所示，将中文问句的训练数据均等的划分为 3 份，形成训练集 1，训练集 2 及训练集 3，将其对应输入到基分类器 LightGBM 算法、XGBoost 算法和 Random Forest 算法中，将每个基分类器预测的结果作为训练集，统一输入到元分类器 Logistic Regression 算法中进行训练，并输出中文问句最终的分类结果。

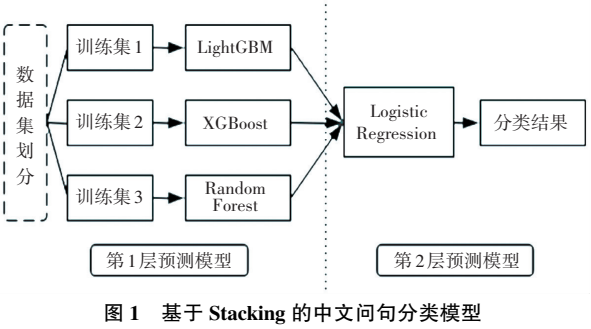

图 1　基于 Stacking 的中文问句分类模型

Fig. 1　Chinese question classification model based on Stacking

3　实验结果与分析

为验证采用集成学习 Stacking 框架融合多模型对于中文问句分类提升的效果，设置了中文问句分类实验。在实验数据相同的前提下，对比基于 Stacking 融合模型与单模型在中文问句分类上的效果。

3.1　实验过程

3.1.1　实验数据

实验采用网络公开的中文科技文献问句数据集，数据形式为“中文问句+分类标签”，如图 2 所示。其中，标签 1 代表简单问题，标签 2 代表单跳问题，标签 3 代表单重限定问题，标签 4 代表多重限定问题。该数据集包含 71 264 条数据，各类别均包含 17 816 条数据。由于实验采用的都是机器学习模型，因此将数据集按 8∶2 划分，得到训练集 56 992 条数据，测试集 14 272 条数据。

有关键词竞争情报的论文有哪些？→1
论环境科学与工程学科馆员工作管理的发表日期是什么时候？→1
中国信息服务平台研究主题在其产出的年份还发表过哪些论文？→2
2004 年发表的论文有哪些研究主题？→2
研究人员陈果和肖璐同属于哪一个机构？→3
财务情报利用模式研究和科技文献数字资源建设调查属于哪个研究领域？→3
2007 年含有企业战略决策关键词的论文有哪些？→4
黄红星写的论文在 2006 年发表有哪些？→4

图 2　基于句型的中文句数据集

Fig. 2　Chinese question data set based on sentence pattern

3.1.2　实验设置

实验采用精确率 (P)，召回率 (R) 和 $F1$ 作为评价指标，对比单模型和基于 Stacking 的集成学习模型在问句分类实验上的效果。在文本表示上，选择 TF-IDF 来提取特征，并使用交叉网络验证调参，得到各分类模型的最佳参数，见表 1。

表 1　模型最佳参数

Tab. 1　Best parameters of the model

序号	分类模型	最佳参数
1	LightGBM	{num_leaves=63, n_estimators = 100}
2	XGBoost	{n_estimators=120, learning_rate=0.08, gamma=0, subsample=0.8, colsample_bytree=0.9, max_depth=5}
3	Random Forest	{class_weight='balanced', random_state=10}
4	Logistic Regression	{penalty="l1", solver="liblinear", C=0.5, max_iter=1 000}
5	Stacking 模型	{cv=3}

3.1.3　评价指标

精确率 (*Precision*, P) 用来评测模型的查准率，正确预测为正例的样本数占预测为正例的样本数的比重。

$$P = \frac{\text{正例预测正确的样本数}}{\text{预测为正例的样本数}} \quad (1)$$

召回率(*Recall*, *R*)用来评测模型的查全率,正确预测为正例的样本数占正例样本总数的比重。

$$R = \frac{\text{正例预测正确的样本数}}{\text{正例样本总数}} \quad (2)$$

*F*1 值(*F*1 - measure, *F*1)为综合精确率和召回率的整体评价指标。

$$F1 = \frac{2 \times P \times R}{P + R} \quad (3)$$

3.2 实验结果与分析

将 Stacking 融合模型与 4 个基分类模型的实验效果进行对比,其结果见表 2。

表 2 基于 Stacking 的中文问句分类实验结果

Tab. 2 Experimental results of chinese question classification based on stacking %

模型	*P*	*R*	*F*1
LightGBM	95.56	96.52	96.03
XGBoost	97.88	96.92	95.92
Random Forest	94.84	95.12	94.98
Logistic Regression	63.56	58.68	61.02
Stacking 融合模型	98.25	97.35	97.80

由表 2 数据可见,表现最差的为 Logistic Regression 算法,其主要原因是中文问句分类属于非线性分类问题,而 Logistic Regression 属于线性分类模型,因此两者不匹配导致其效果不佳。将其余 3 个单分类模型相互进行对比发现,Random Forest 的表现最差;XGBoost 的分类效果为单分类模型中最佳,其余 3 个评价指标值仅次于 Stacking 融合模型,也是由于其本身为强分类模型。基于 Stacking 集成学习框架融合多模型取得了最优的效果,在精确率、召回率、*F*1 值均优于其它分类器。其中,*F*1 值达到 97.80%,也充分验证了本文所提基于 Stacking 的中文问句分类模型的有效性。

4 结束语

基于 Stacking 集成学习的中文问句分类方法,能够有效提升中文问句分类精度。通过使用集成学习 Stacking 框架,融合 LightGBM、XGBoost 和 Random Forest 构建多基分类器,并利用 Logistic Regression 作为元分类器,以实现中文问句分类。该模型在精确率、召回率和整体 *F*1 等指标上均优于其他模型,其中 *F*1 值达到 97.80%,能够实现较高的分类精度。该模型的基分类器由 3 个不同算法组成,在保持算法强学习能力的基础上,同时具备一定的异质性,使模型具备更好的泛化能力,支撑更多领域问答系统实现更好的性能。

参考文献

[1] 韩东方,吐尔地·托合提,艾斯卡尔·艾木都拉. 问答系统中问句分类方法研究综述[J]. 计算机工程与应用,2021,57(6):10-21.

[2] HOVY E,GERBER L,HERMJAKOB U,et al. Toward Semantics-based answer pinpointing [C]//Proceedings of the First International Conference on Human Language Technology Research. New York:Association for Computing Machinery,2003:245-256.

[3] 张宇,刘挺,文勖. 基于改进贝叶斯模型的问题分类[J]. 中文信息学报, 2005,19(2): 100-105.

[4] ALI H, MICHEL B, KRISTINE L, et al. Enhanced semantic expansion for question classification[J]. International Journal of Internet Technology and Secured Transactions, 2011, 3(2): 134-148.

[5] BO Q, GAO C, LI CUI P, et al. An evaluation of classification models for question topic categorization [J]. Journal of the American Society for Information Science and Technology, 2012, 63(5):889-903.

[6] KALCHBREBBER N, GREFENSTETTE E, BLUNSOM P. A convolutional neural network for modelling sentences [C]// Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. New York: Association for Computing Machinery,2014:655-665.

[7] PHONG L, WILLEM Z. The forest convolutional network: compositional distributional semantics with a neural chart and without binarization[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon: Association for Computational Linguistics, 2015:1155-1164.

[8] ZHANG J C, LI J. Topic memory networks for short text classification [C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Lisbon: Association for Computational Linguistics, 2018: 3120-3131.

[9] UJITH R,ZORNITSA K. Self-governing neural networks for on-device short text classification [C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Lisbon:Association for Computational Linguistics,2018: 804-810.

[10] 徐继伟,杨云. 集成学习方法:研究综述[J]. 云南大学学报(自然科学版), 2018, 40(6):1082-1092.

[11] 张元哲. 知识库问答关键技术研究[D]. 北京:中国科学院大学,2016.

[12] VAN M J, POLLEY E C, HUBBARD A E. Super learner[J]. Statistical Applications in Genetics & Molecular Biology, 2007, 6(3):1-23.

[13] 谢文涌,柴琴琴,林旎,等. 基于 Stacking 集成学习的马兜铃酸及其类似物鉴别[J]. 江苏农业学报,2021,37(2):503-508.

[14] 胡晓丽,张会兵,董俊超,等. 基于集成学习的电子商务平台新用户重复购买行为预测[J]. 现代电子技术,2020,43(11):115-119.

[15] 苑擎飚. 基于决策树中文文本分类技术的研究与实现[D]. 辽宁:东北大学,2008.