计算机科学
COMPUTER SCIENCE

BEML:一种面向商品隐空间表征的混合学习分析范式

郑骐健, 刘峰

引用本文

郑骐健, 刘峰. BEML:一种面向商品隐空间表征的混合学习分析范式[J]. 计算机科学, 2024, 51(11A): 240300150-6.

ZHENG Qijian, LIU Feng. BEML:A Blended Learning Analysis Paradigm for Hidden Space Representation of Commodities [J]. Computer Science, 2024, 51(11A): 240300150-6.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240300150

BEML:一种面向商品隐空间表征的混合学习分析范式

郑骐健　刘　峰
华东师范大学计算机科学与技术学院　上海 200062
(shange0403@163.com)

摘　要　随着互联网经济时代的到来,电子商务平台的高效管理日益受到学术界和工业界的广泛关注,其中,商品分类的精度与自动化水平直接影响着用户体验及运营效率的优化。鉴于此,本研究围绕商品信息的隐空间表征进行深入探讨,提出了一种面向商品隐空间表征的混合学习分析范式BEML。该框架融合了先进的双向编码器表示(BERT)技术与传统机器学习方法,旨在通过对商品信息隐空间的细致解析,显著提升商品分类的自动化处理效率及准确性。与现行主流的深度学习和机器学习算法进行对比分析的实验结果表明,BEML框架针对本次亚马逊在线分析数据集的最佳分类效果F1指标的宏平均达到了85.79%,微平均达到了84.73%,均超过了目前最佳F1指标83.3%,实现了新的SOTA。该框架不仅在理论上具有创新性,其在电子商务领域的信息管理和自动化处理实践中亦具有重要的应用价值,为科技商学领域提供了一种高效且可靠的混合学习分析范式。

关键词:隐空间表征;BERT预训练模型;自动商品分类;智能化商品分类;科技商学
中图分类号　TP311

BEML:A Blended Learning Analysis Paradigm for Hidden Space Representation of Commodities

ZHENG Qijian and LIU Feng
School of Computer Science and Technology,East China Normal University,Shanghai 200062,China

Abstract　With the advent of the Internet economy era,the efficient management of e-commerce platforms has garnered widespread attention from both academia and industry. Among various factors,the accuracy and automation level of product classification directly impact users' experience and the optimization of operational efficiency. In light of this,this study delves into the latent space representation of product information,proposing a blended learning analysis paradigm for product latent space representation(BEML). This framework integrates advanced bidirectional encoder representations from transformers(BERT) technology with traditional machine learning methods,aiming to significantly enhance the efficiency and accuracy of automated product classification through detailed analysis of the latent space of product information. By conducting comparative analysis with current mainstream deep learning and machine learning algorithms,this study validates the exceptional performance of the BEML framework in product classification tasks. Experimental results demonstrate that the BEML framework achieves a macro F1 score of 85.79% and a micro F1 score of 84.73%. Both exceed the current best F1 score of 83.3%,reaching a state of the art. Moreover,this framework not only represents a theoretical innovation but also holds significant practical application value in the realm of information management and automation processing within the e-commerce sector,providing an efficient and reliable blended learning analysis paradigm for the field of technology and business.

Keywords　Latent space representation,Pre-trained BERT model,Automated commodity classification,Intelligent commodity classification,Sci-tech driven business

1　引言

在数字化转型的推动下,电子商务作为现代经济的重要组成部分,正经历前所未有的快速发展。这一过程中,商品分类技术的进步直接影响着消费者体验和电商平台的运营效率。面对大数据背景下商品数据的爆炸式增长,传统的分类方法由于难以处理非结构化文本数据和适应商品种类的快速扩展,正面临越来越多的挑战[1]。

在这样的背景下,利用机器学习和模式识别技术对商品进行自动分类,已经成为提高分类准确性和效率的关键途径。这些方法能够有效挖掘商品数据中蕴含的复杂信息,实现更为精确的商品分类[2]。尤其是,随着预训练语言模型,如

基金项目:上海市科技计划项目(20dz2260300);华东师范大学计算机科学与技术学院"人工智能赋能心理/教育学科交叉人才培养专项基金"(2024JCRC-10);
This work was supported by the Research Project of Shanghai Science and Technology Commission(20dz2260300) and Special Fund for Talent Cultivation of Artificial Intelligence Enabled Psychology/Education Interdisciplinary Cross-disciplinary Talents(2024JCRC-10),School of Computer Science and Technology,East China Normal University.
通信作者:刘峰(lsttoy@163.com)

BERT(Bidi-reactional Encoder Representations from Transformer)的出现和发展,我们有了处理非结构化文本数据并提取深层次特征的更强大工具[3-4]。这些模型在生成商品数据的隐空间表征方面显示了其卓越的性能,显著提高了商品分类任务的准确度[5-7]。

尽管现有研究在隐空间表征的生成上取得了重要进展,但在如何有效挖掘这些表征中的信息以提高分类性能方面仍存在不足。针对这一问题,本研究提出了 BEML,一种融合预训练模型和机器学习技术的混合学习分析范式。BEML 模型以商品标题文本为输入,利用预训练的 BERT 模型进行深层次的特征提取,并结合了机器学习方法。对比现在主流的线性预测层或逻辑回归方法[8-9],BEML 模型在商品分类任务中展现出的性能优势不仅证实了预训练模型与机器学习方法进行融合的研究范式有效性,也为大数据背景下的商品智能化分类提供了新的理论和实践路径。此外,我们的研究拓宽了科技商学领域的研究视野,为处理和分析大规模多样化商品数据提供了一种高效且可靠的解决方案,具有重要的理论意义和应用价值。

本文的具体贡献如下:

(1)提出了一种融合预训练模型和机器学习技术的模型框架 BEML,大幅提升了分类的效率和准确性。该模型结合了预训练的 BERT 模型和机器学习的混合学习分析范式,并基于亚马逊在线商城的真实数据展开大规模和多样化的商品数据测试,结果达到了 SOTA。

(2)提出了一种新的基于深度学习与机器学习融合的混合学习分析范式(BEML),该范式下的模型实现了分类的高精度和良好的泛化能力,能基于开放世界数据进行业务处理。这在科技商学领域中为处理复杂数据集提供了一个高效且可靠的混合学习分析范式,展现了深度学习与传统机器学习方法融合的巨大潜力。

(3)针对电子商务领域中的大数据进行了深入分析和应用,影响了大数据驱动的商业决策,提升了用户体验,具备较高的商业潜力。这一贡献不仅在理论上推动了大数据在电子商务中的应用研究,也为实际商业操作提供了数据驱动的解决策略。

2 相关文献

商品分类任务的核心是根据商品描述信息和相关属性预测商品所属的目标目录类别。在过去,商品分类任务主要由领域专家完成,专家会根据商品的特性或参考 GPC(Global Product Classification,全球产品分类标准)对商品进行划分。但随着商品数量的增加,手动分类变得不切实际,因此,基于机器学习算法的商品自动分类任务被提出[10]。基于机器学习的商品自动分类过程可以分为两个步骤:首先是对商品描述信息的隐空间表征的映射;然后通过监督学习算法对该表征中的分类信息进行抽取[11]。其中隐空间指的是对数据进行核心特征学习得到的抽象空间,用于呈现数据的压缩形式[12]。这些压缩后的表征数据通常捕捉了原始数据的核心特征,同时剔除了噪声和不重要的细节。在隐空间映射的信息选择上,由于大部分电商平台对于商品数据严密保护,获取完整的商品描述信息往往比较困难,因此大部分研究使用了较易获得的商品标题数据作为分类的元数据[13-14]。本文所讨论的商品信息隐空间,特指基于商品标题信息所构建的语义信息隐空间。

在针对隐空间表征的映射上,早期工作主要依赖于文档/逆文档频率方法(TF/IDF)将文本标题转化为高维稀疏的隐空间向量,该方法通过统计分析语料库中的词汇使用模式来提取词义信息,并将文档信息转化为隐空间中的文档术语频率矩阵,最终根据矩阵间的相似度比较实现商品的分类[15]。之后,为了能够捕捉更细微的语义差别和复杂的语言模式。近几年的工作都采用 word2vec(词向量,word to vector)、BERT 等预训练语言模型来编码商品的标题信息[16]。这些模型不仅关注输入文本的上下文信息,解决了一词多义问题,还融合了预训练模型在商品领域的相关知识。

对于隐空间表征的分类信息提取过程,在早期研究中,由于隐空间编码方法[17]在特征表示方面存在局限,因此研究者们在分类信息提取模型上更多地使用深度神经网络一类复杂模型[18],从而实现隐空间复杂的划分效果。但在 BERT 一类预训练编码模型被提出后,标题文本信息到隐空间的提取效率大大提高,使得分类信息的提取不必再依赖复杂的深度学习模型,大部分研究便倾向于在预训练模型后加入一个全连接层用于分类信息的提取和输出[19]。并且由于预训练模型对商品分类预测准确度的巨大提升,大部分研究将重点集中于 BERT 等预训练模型的微调上,即如何更高效完整地将重要信息通过 BERT 嵌入到构造的隐空间中,从而弱化了对于分类信息提取过程的关注。然而,不同特征信息提取方法的选择可能会导致准确率等指标上的较大差异[20]。为了提高商品分类信息提取模型的准确度,本文设计的 BEML 混合学习范式提供了可替换的隐空间中分类信息提取模型架构,旨在寻找更符合商品隐空间信息分布的提取模型,并且进一步探索了预训练 BERT 生成的商品信息隐空间中的信息分布特点和信息提取方法之间的联系。

3 相关知识

在详细介绍 BEML 框架前,本章先介绍商品标题隐空间的研究问题、数据集和相关评价指标。

3.1 研究问题

本节对研究的任务进行了定义。假设在产品分类任务中,存在 M 种产品 Brick,产品 $Brick_i$($i=1,2,3,\cdots,M$)在 GPC 分类标准下归属于一个上层的品类 $Class_i$,如产品“婴儿水杯”和“摇篮”均属于“婴儿用品”类。对于产品 $Brick_i$,其在数据集中对应了一个商品实体集合 $Goods_i$,该集合中包含了 N 个具体的商品实例,每个实例由商品的标题名进行表示 $Goods_i=\{caption_{i1},caption_{i2},caption_{i3},\cdots,caption_{iN}\}$。商品的标题 $caption$ 作为元数据,对于这个商品实例所属的上层品类划分 Class 以及产品品类名 Brick 进行预测。

3.2 评测数据集

本文的数据集基于全球最大在线零售商亚马逊(Amazon.com)电商平台 2023 年 1 月至 6 月的公开商品数据。在不同的商品检索抽样下,使用爬虫工具爬取了 70 000 条真实商品标题数据,经过数据清洗去除重复、缺失值,共得到 69 919 条有效商品标题数据。在标签标注上,本文根据 2023 年 GPC 标准(Global Product Classification,全球产品分类标准)对爬取得到的商品数据在最小产品类(Brick)层级和类层

级(Class)进行了人工标注，最终得到的数据集包含 100 种不同的最小产品类(Brick)和 42 个商品大类(Class)。在最小产品类层级下，标签最少的商品数据条数为 671 条，最多的商品数据条数为 700 条，数据集标签较为平衡。而在商品大类层级下，标签最少的商品数据条数为 683 条，最多的商品数据条数为 5573 条，大部分商品类的标签数在 1 400 条左右。其中部分商品标题数据和对应的 Brick，Class 标签如表 1 所列。

表 1　商品标签数据集中的商品样例

Table 1　Product examples in products label dataset

商品标题	最小产品层级	类层级
Ossetra Sturgeon Caviar 250g(8.8 oz)	海鲜	鱼类/海鲜
Tazo Tazo Calm Chamomile Herbal Tea Bags 24 Ct,0.13 Kilogram	茶	饮品
Gold Medal Cherry Pink Glaze Pop Frosted Popcorn Mix,28 oz,12 Count	爆米花	烘焙产品
Home Dishwasher Free Installation Small Dishwashing Desktop Dishwasher Automatic Intelligent	洗碗机	厨房用品

3.3　评测指标

本文采用 F1 值的宏平均(macro average)和微平均(micro average)(即 F1_macro 和 F1_micro)两种评价指标对模型进行评估。在商品标签多分类任务下，F1 宏平均是对所有类别分别计算得出的 F1 值的平均，F1 微平均则是基于对所有类别累积计算的真正例(TP)、假正例(FP)和假负例(FN)来计算的，它计算所有类别的总 TP、总 FP 和总 FN，然后使用这些值来计算精确率和召回率，进而得到 F1 分数。F1 宏平均更加关注数据集中样本较少的类别，而 F1 微平均更关注数据集中样本较少的类别。由于最小产品类数据集标签较为均衡而商品大类数据集标签较不均衡，因此通过 F1 宏平均和微平均可以得到不同模型对分类信息的预测能力以及模型本身的泛化能力。

4　基于隐空间信息嵌入的 BEML 框架构建

本文设计的混合分析框架 BEML 由 Sentence-Bert 预训练编码模型和基于机器学习方法的隐空间信息提取层组成。如图 1 所示，对于给定的标题输入序列 $caption_i$，BEML 使用基于商品信息预训练的 sentence-BERT 得到了商品标题隐空间表征 S_i，再通过隐空间信息提取层得到商品的类别标签信息。

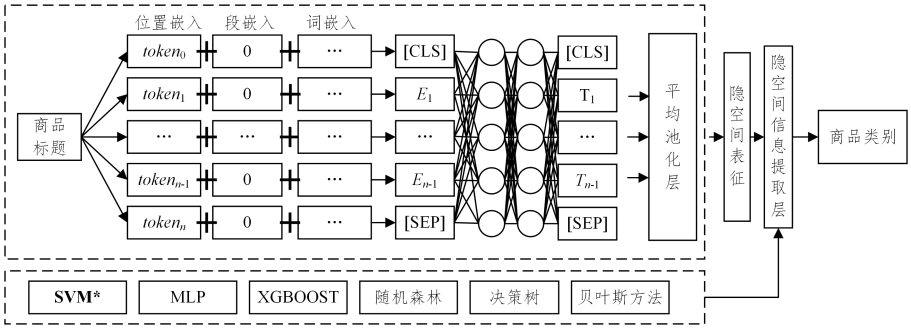


图 1　BEML 框架结构图

Fig. 1　BEML architecture

Bert 是 Google 于 2018 年提出的基于自注意力的预训练语言模型。它的主要架构为多层双向 Transformer 编码器，每一层都由多头自注意力机制和全连接前馈网络组成，通过这些层的堆叠，BERT 能够捕获上下文中丰富的语言特征。由于本文处理的数据为商品标题文本，因此选取了基于商品信息预训练的 sentence-BERT[22] 进行隐空间的表征映射。与传统 BERT 相比，sentence-BERT 通过共享权重模型进行句子嵌入并通过相似度函数进行语义比较得到训练损失，从而使得语义相似的句子编码在向量空间中彼此接近，而不相似的句子则相距较远。并且，本文使用 Siamese(孪生)网络架构进行预训练，在网络输出层后加入了一层平均池化层，从而对任意输入的句子获得相同长度的隐空间向量表示，该向量表示长度 $T=768$。

对于输入标题 $caption_i$，在经过 Sentence-BERT 的 BERT 编码和平均池化后得到了固定维度的隐空间向量表示 S_i。

$$\boldsymbol{E}_i=\{clf,w_1,w_2,w_3,\cdots,w_m \mid w_i\in R^{T\times 1}\} \tag{1}$$

$$\boldsymbol{S}_i=\left\{\frac{\sum_{k=1}^{m}\boldsymbol{w}_k}{m}\right\}=\{\boldsymbol{s}\mid \boldsymbol{s}\in R^{T\times 1}\} \tag{2}$$

在得到隐空间表征后，考虑到真实场景下不同商品分类标签层次的存在，我们将其和商品对应的类别标签 $Brick_i$、上层品类标签 $Class_i$ 分别组合得到了样本(S_i，$Brick_i$)和(S_i，$Class_i$)，用于训练和测试商品隐空间分类信息提取模型。

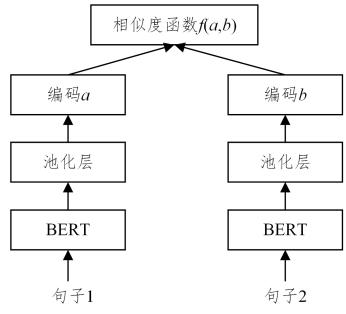


图 2　sentence-BERT 训练结构图

Fig. 2　Training architecture of sentence-BERT

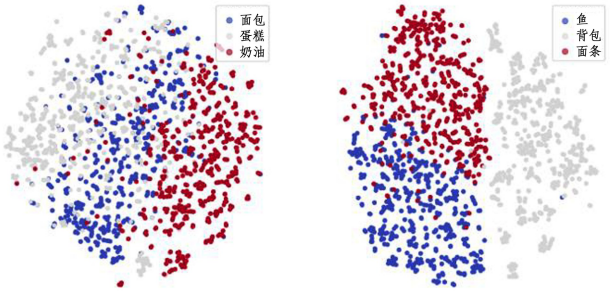
在隐空间的信息提取层上，本文共选择了 6 类在过往研究中表现卓越的机器学习算法进行对比[24]，涵盖了集成算法、贝叶斯算法、人工神经网络、支持向量机。以采用 Tanh 激活函数的多层感知机为例，在实验中该模型结构

设置为3层，则算法对于任务中的单个样本的预测过程可以描述为：

$$Brick_{\text{predict}}=L^{3}_{\text{MLP}}(Relu(L^{2}_{\text{MLP}}(Relu(L^{1}_{\text{MLP}}(\boldsymbol{S})))))\tag{3}$$

$$Class_{\text{predict}}=L^{3'}_{\text{MLP}}(Relu(L^{2'}_{\text{MLP}}(Relu(L^{1'}_{\text{MLP}}(\boldsymbol{S})))))\tag{4}$$

其中，L^{i}_{MLP}表示多层感知机的第 $i(i=1,2,3)$层，L'和 L 分别表示预测标签 Brick 和标签 Class 的模型。



(a)同一商品大类集合 A1 下产品类的信息分布　(b)同一商品大类集合 A2 下产品类的信息分布

图 3　不同商品大类下隐空间表征的 T-SNE 可视化

Fig. 3　T-SNE visualization of latent space representations at different product class levels

综上所述，BEML 混合分析框架的流程如算法 1 所示。

算法 1　BEML 混合分析框架比较

输入：商品标题〈INPUT〉

输出：商品所属的产品类别 Brick_i，商品所属产品类别的上层类别 Class_i〈OUTPUT〉

1. WHILE ML_k in MLAlgorithms：
2.　　WHILE $\text{data}_{\text{train}}$, $\text{data}_{\text{test}}$, $\text{label}_{\text{Brick}}$, $\text{label}_{\text{Class}}$ in datasets：
3.　　$\text{ML}_k'=\text{Train}(\text{data}_{\text{train}},\text{label}_{\text{Brick}},\text{MLAlgorithm})$
4.　$\text{ML}_k=\text{Train}(\text{data}_{\text{train}},\text{label}_{\text{Class}},\text{MLAlgorithm})$
5.　　$\text{Ei}=\text{BERT}_{\text{Pre}}(\text{Caption}_i)$
6. $S_i=\left\{\frac{\sum_{k=1}^{m}w_k}{m}\mid w_k\in E_i\right\}=\{s\mid s\in R^{T\times 1}\}$
7.　　$\text{Brick}_i=\text{ML}'(S_i)$
8. $\text{Class}_i=\text{ML}'(S_i)$
9.　　END WHILE
10. END WHILE

5　实验与结果分析

5.1　实验环境配置

实验运行的硬件环境为显卡：NVIDIA 4070 Ti；CPU：Intel(R) Core(TM) i5 3.50GHz；内存：32GB；硬盘：1T；系统运行环境：windows10 专业版。编程环境：Python3.8。

5.2　隐空间中基于 T-SNE 的表征可视化

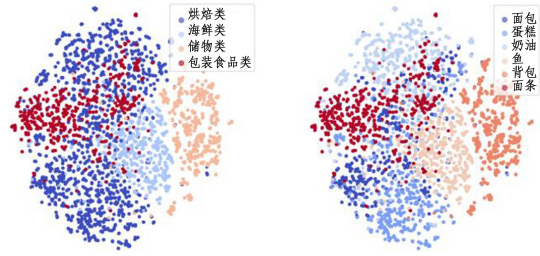
对于商品的原始文本标题数据，经过 sentence-BERT 的编码映射，转化为了隐空间中的表征信息。在进行隐空间的分类信息提取之前，为探索商品隐空间中不同商品分类层级下的信息分布特点，本文选取了部分产品类的隐空间表征，并根据这些产品在上层分类下的标签，将其分为 A1 组（上层分类标签相同产品）和 A2 组（上层分类标签不同的产品），如表 2 所列。对于这些组别中的商品在隐空间中的信息分布，本文使用 T-SNE 方法进行了可视化表示。

表 2　产品类型集合

Table 2　Product categories

CLASS 类相同（烘焙类）[A1]	CLASS 类不同[A2]
面包	鱼（海鲜类）
蛋糕	背包（储物类）
奶油	面条（预制食品类）

商品在大类和产品类两个层级隐空间中的信息分布情况如图 4 所示。图 4(a)和图 4(b)描绘了同一批商品实体在不同分类层级标签下的二维空间映射，图 4(a)中属于烘焙类的商品在图 4(b)中被分为了面包、蛋糕和奶油 3 种产品类。图 4(a)中的分布显示了不同大类下的商品间存在较严重的重叠现象，如包装食品类和烘焙类。而在产品类层级下，重叠现象主要发生在部分商品间，如面包类和与其属性相近的蛋糕、面条类。这种数据点的混合现象表明了目前使用的隐空间信息提取模型（BERT）还无法实现对不同品类下商品信息的自然划分。虽然各个商品品类在人类的认知角度上存在显著差异，但其面向机器的信息描述上的相似性仍会导致隐空间映射结果上的重叠。在 BEML 混合学习架构中，通过在隐空间信息提取层中进行谨慎的模型选择，构建特殊的分类边界，从而提升了模型整体的分类性能。



(a)商品大类层级下隐空间表征分布　(b)产品类层级下隐空间表征分布

图 4　商品大类（Class）层级和产品类（Brick）层级下隐空间表征的 T-sne 可视化

Fig. 4　T-SNE visualization of latent spaceat class levels and brick levels

5.3　实验结果分析

在隐空间的特征编码环节，本实验使用了经过预训练的 Sentence-BERT 模型。在隐空间信息挖掘的方法上，本实验选用并进行比较的机器学习模型和相应参数如表 3 所列。在模型训练过程中，实验所涉及模型的随机化参数根据最佳实践设置为 7。

表 3　机器学习算法网格化参数搜索

Table 3　Meshed parametric search of machine learning algorithms

模型类别	参数类型
MLP	隐藏层参数：3 层（参数量：[100,50,20]） 激活函数：['identity','logistic','tanh','relu'] 权重优化算法['lbfgs','sgd','adam']
XGBOOST	最大深度：[5,10,15,20,25,30,35,40,45] 预测器个数：[50,60,70,80,90]
随机森林	最大深度[5,10,15,20,25,30,35,40,45] 预测器个数：[50,60,70,80,90]
SVM	核函数类型：['linear','poly','rbf','sigmoid'] 惩罚项系数：[0.5,0.6,0.7,0.8,0.9]
决策树	最大深度：[5,10,15,20,25,30,35,40,45] 信息增益：['entropy','gini']
贝叶斯分类器	模型：高斯朴素贝叶斯（GaussianNB） 贝叶斯岭回归（MultinomialNB）

在最小产品类的分类任务下，使用不同机器学习方法的

BEML 隐空间信息提取模型效果如表 4 所列，其中各个机器学习方法均被调至最优参数。由实验结果可知，使用了 SVM 作为信息提取层的 BEML 达到了最优效果，相比多数研究中使用的多层感知机方法在宏观 F1 指标上提高了 5.84%。并且，XGBOOST 和随机森林方法也达到了接近 MLP 的效果。

表 4　商品小类分类任务下各模型最优参数及结果

Table 4　Each model's best performance and parameters in brick level classification task

提取模型	F1_micro/%	F1_macro/%	参数设置
SVM	82.61	82.48	正则化参数：0.5 核函数：poly
MLP	76.76	76.84	激活函数：identity 权重优化：adam
XGBOOST	76.56	76.71	最大深度 5 预测器个数：95
随机森林	72.25	72.67	最大深度：40 预测器个数：90
贝叶斯分类器	37.92	40.45	模型：GaussianNB
决策树	35.66	35.89	最大深度：15 信息增益：entropy

针对商品大类预测任务的实验结果如表 5 所列。在这个任务上依然是使用 SVM 作为信息提取层的 BEML 模型取得了最好的结果，在宏微观 F1 指标上达到了 85.79% 和 84.73%，相较于第二名 MLP 提升了 4.03% 和 3.21%。此外，由于分类种数从 100 种减少到了 42 种，因此模型间也出现了不同程度的 F1 指标上升的情况。其中，以 SVM，MLP，XGBOOST 作为信息提取层的 BMEL 表现最优。

表 5　商品大类分类任务下各信息提取层模型最优参数及结果

Table 5　Each model's best performance and parameters in class level classification task

信息提取层模型	F1_micro/%	F1_macro/%	参数设置
SVM	84.73	85.79	正则化参数：0.5 核函数：poly
MLP	81.52	81.76	激活函数：relu 权重优化：adam
XGBOOST	80.87	80.04	最大深度：10 预测器个数：95
随机森林	73.77	73.71	最大深度：40 预测器个数：90
贝叶斯分类器	65.83	62.64	模型：GaussianNB
决策树	44.60	43.90	最大深度：15 信息增益：gini

在两类标签情况下，以 SVM 作为信息提取层的模型都取得了最佳效果。由于 SVM 本身采用的核方法对于预测效果有着较大的影响，因此我们对实验过程中使用不同核方法的 SVM 的 BEML 框架最佳效果进行了关注，如图 5 所示。在不同核方法下，商品小类分类中宏微观 F1 指标最小为 78.81%与 78.72%，在大类分类中宏微观 F1 指标最小为 81.08%和 80.40%，均接近甚至优于其他非 SVM 信息提取层模型的最优效果。从实验的最优效果来看，多项式(poly)核方法在商品小类的宏微观 F1 指标以及商品大类预测的宏观 F1 指标上的表现效果最好，分别达到了 82.61%，82.48%和85.79%。而线性(Linear)核方法在大类分类的微观 F1 指标上效果最优，达到了 84.76%。综上，BEML 框架在亚马逊在线分析数据集上的最佳分类效果 F1 指标的宏平均达到了 85.79%，微平均达到了 84.73%，均超过了目前最佳 F1 指标 83.3%[19]，实现了新的 SOTA。

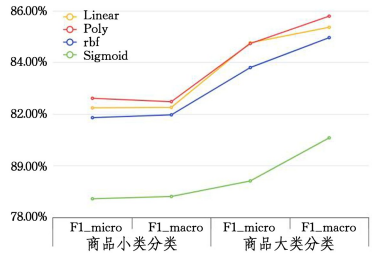


图 5　商品分类任务下不同核方法下 SVM 的预测效果

Fig. 5　SVM prediction results of different kernel methods in products classification task

5.4　讨论与展望

本节讨论不同商品标签下 BEML 框架的隐空间信息提取模型的表现。从分类标签的层次角度分析，商品小类分类的标签是在分类结构下对商品实体最准确的划分，而商品大类标签则是基于商品的用途、价值、属性等多方面的元素对产品类标签的聚集。由于不同商品大类的涵盖范围不同，因此包含的产品类标签数目也有所差异，同时，不同产品类目下的商品数目也不同。本文考虑了大类标签间最小产品类数目存在的差异，因此基于相同的元数据下形成了标签不平衡的商品大类数据集和标签平衡的产品最小类数据集。在大类预测任务下，虽然标签数减少了，但由于标签不平衡的影响，造成了部分信息提取层模型如随机森林在微观 F1 指标上的提升不明显，在其他模型最高提升 27.91%的情况下，其提升仅为 1.52%。此外，GPC 标准下的商品大类标签划分更多遵循人们对于商品的认知，如乳制奶油和面条同属于预制食品类，而预训练 BERT 模型对隐空间信息的映射由模型在预训练中获得的先验知识和输入的商品标题信息共同决定，因此面条和乳制奶油一类在标题描述信息上差异较大的商品，在隐空间中的映射距离会变得较远，这更加考验隐空间信息提取模型的边界划分能力。

在获得 BEML 信息提取层最优效果的过程中，我们对不同超参数的信息提取层模型进行了网格化搜索，其中部分参数的表现效果与隐空间信息分布的特点相关联。以表现效果较好的 SVM 方法为例，SVM 的核函数方法分为非线性核函数方法(poly，rbf，sigmoid)和线性核函数(linear)。前者采用的策略是将当前数据分布映射到更高维度的空间，让原先线性不可分的数据在高维下变得线性可分，后者则不进行高维映射操作，直接在当前的空间下进行决策边界的划分，具有计算简单、训练快速的优势。根据图 6 结果可知，在不对当前隐空间的数据进行高维映射的情况下，SVM 使用 linear 核函数进行边界划分就可以取得较好的效果，并且与 poly 核函数下的最优指标的差距均小于 0.2%。此外，在最小产品类分类任务中，效果最好的 MLP 信息提取层也使用了线性的 identity 激活函数。这说明在 BEML 框架中，大部分商品数据在预训练 BERT 所映射的商品信息隐空间下是线性可分的，不需要额外进行升维映射。基于此，在对应模型中使用线性的激活函数或核函数，可以在不损失准确度的同时带来训练和推理速度的提升。

综上，SVM，MLP，XGBOOST 模型在隐空间分类信息提取中表现出了较好的预测能力，并且在不同商品层级的数据集下均表现稳定，F1 指标均大于 76%。此外，各模型最优的超参数选择反映了商品信息隐空间下商品数据线性可分的特

点。由此,在隐空间分类信息提取模型的超参数选择中,可以更多地选择线性激活函数,从而提升 BEML 混合学习框架的效率。

结束语 本研究在大数据时代的背景下,针对电子商务领域的自动化商品分类问题,基于预训练 BERT 模型的分类架构,对比了包括 SVM 在内的 6 种机器学习方法对隐空间分类信息的提取效果,构建了覆盖 100 种最小产品类别和 42 个商品大类的商品信息隐空间,并通过可视化方法对隐空间中的信息分布特点进行了评估。此外,本文针对模型参数选择、模型分类特点以及标签层次问题进行了分析,并用实验证明了 SVM,MLP,XGBOOST 方法在隐空间分类信息提取中的能力和应用潜力。但是,目前的研究主要还在基于公开数据集的文本数据,未来我们将探索包括图像和音频在内的多模态数据,以进一步提高模型的泛化能力和准确性。我们期待本研究的成果能够激发更多关于大数据环境下商品分类的深入研究,并在实际应用中发挥重要作用。

参 考 文 献

[1] LANDAUER T K, FOLTZ P W, LAHAM D. An introduction to latent semantic analysis[J]. Discourse Processes, 1998, 25(2/3): 259-284.

[2] CHANG T Z, WILDT A R. Price, product information, and purchase intention: An empirical study[J]. Journal of the Academy of Marketing Science, 1994, 22: 16-27.

[3] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv: 1810.04805, 2018.

[4] YANG L, SHIJIA E, XU S, et al. Bert with Dynamic Masked Softmax and Pseudo Labeling for Hierarchical Product Classification[C]// MWPD@ ISWC. 2020.

[5] BELTAGY I, LO K, COHAN A. SciBERT: A pretrained language model for scientific text[J]. arXiv: 1903.10676, 2019.

[6] LEE J, YOON W, KIM S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining [J]. Bioinformatics, 2020, 36(4): 1234-1240.

[7] PEETERS R, BIZER C. Dual-objective fine-tuning of BERT for entity matching [J]. Proceedings of the VLDB Endowment, 2021, 14: 1913-1921.

[8] ZAHERA H M, SHERIF M. ProBERT: Product Data Classification with Fine-tuning BERT Model[C]// MWPD@ ISWC. 2020.

[9] MEUSEL R, PRIMPELI A, MEILICKE C, et al. Exploiting microdata annotations to consistently categorize product offers at web scale[C]// International Conference on Electronic Commerce and Web Technologies. Cham: Springer International Publishing, 2015: 83-99.

[10] YU H F, HO C H, ARUNACHALAM P, et al. Product title classification versus text classification[J]. Csie. Ntu. Edu. Tw, 2012: 1-25.

[11] ZHANG Z, SONG X. An exploratory study on utilising the web of linked data for product data mining [J]. SN Computer Science, 2022, 4(1): 15.

[12] LOUIZOS C, SWERSKY K, LI Y, et al. The variational fair autoencoder[J]. arXiv: 1511.00830, 2015.

[13] CHAVALTADA C, PASUPA K, HARDOOND R. A comparative study of machine learning techniques for automatic product categorisation [C] // Advances in Neural Networks (ISNN 2017), Part I 14. Springer International Bublishing, 2017: 10-17.

[14] RISTOSKI P, PETROVSKI P, MIKAP, et al. A machine learning approach for product matching and categorization[J]. Semantic web, 2018, 9(5): 707-728.

[15] LANDAUER T K, DUMAIS S T. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge[J]. Psychological Review, 1997, 104(2): 211-240.

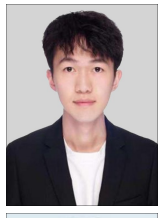
[16] LEE H, YOON Y. Engineering doc2vec for automatic classification of product descriptions on O2O applications[J]. Electronic Commerce Research, 2018, 18: 433-456.

[17] ZHANG Z, PARAMITA M. Product classification using microdata annotations[C]// The Semantic Web-ISWC 2019: 18th International Semantic Web Conference, Auckland, New Zealand, Part I 18. Springer International Publishing, 2019: 716-732.

[18] REDDY B, RAMAKANTHA R, LOKESH K. Classification of health care products using hybrid CNN-LSTM model[J]. Soft Computing, 2023, 27: 9199-9126.

[19] JAHANSHAHI H, OZYEGEN O, CEVIK M, et al. Text Classification for Predicting Multi-level Product Categories[C]// Proceedings of the 31st Annual International Conference on Computer Science and Software Engineering. 2021: 33-42.

[20] HEUNG B, HO H C, ZHANG J, et al. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping[J]. Geoderma, 2016, 265: 62-77.



ZHENG Qijian, born in 2003, master, is a student member of CCF (No. N9988G). His main research interests include deep learning technology and so on.

LIU Feng, born in 1988, Ph.D, is a senior member of CCF (No. 93542S). His main research interests include deep learning technology and blockchain technology.