计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(5):247−252 [doi: 10.15888/j.cnki.csa.007882]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于逆类别注意力机制的电商文本分类①

王　维[1,2], 胡慧君[1,2], 刘茂福[1,2]

[1](武汉科技大学 计算机科学与技术学院, 武汉 430065)
[2](智能信息处理与实时工业系统湖北省重点实验室, 武汉 430081)
通讯作者: 刘茂福, E-mail: liumaofu@wust.edu.cn

摘　要: 电商数据所属类别对于分析电商数据有重要意义, 基于人力的分类无法适应如今海量的电商数据, 基于传统算法模型的分类难以提取有价值的人工特征. 本文采用 BiLSTM 模型并且引入注意力机制, 将其应用于电商数据分类中. 该模型包括 Embedding 层、BiLSTM 层、注意力机制层和输出层. Embedding 层加载 Word2Vec 开源工具训练得到的词向量, BiLSTM 层捕捉每个词语的上下文信息, 注意力机制层为每个词语分配权重, 合成新的样本特征. 实验表明, 基于逆类别率的注意力机制在电商数据的分类准确率达到 91.93%, 与不加注意力机制的 BiLSTM 模型和其他引入的注意力机制相比, 均有不同程度的提高. 此模型电商数据分类中有良好的效果, 为注意力机制的引入提供了新的思考方向.

关键词: 电商数据; 文本分类; Word2Vec; BiLSTM 模型; 注意力机制

引用格式: 王维,胡慧君,刘茂福.基于逆类别注意力机制的电商文本分类.计算机系统应用,2021,30(5):247–252. http://www.c-s-a.org.cn/1003-3254/7882.html

E-Commerce Text Classification Based on Reverse Category Attention Mechanism

WANG Wei[1,2], HU Hui-Jun[1,2], LIU Mao-Fu[1,2]

[1](School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430065, China)
[2](Key Laboratory of Intelligent Information Processing and Real-Time Industrial System of Hubei Province, Wuhan 430081, China)

Abstract: The category of e-commerce data is of great significance for its analysis. The classification based on human resources cannot adapt to the massive e-commerce data nowadays, and the classification based on traditional algorithm models can hardly extract valuable artificial features. In this study, the BiLSTM model integrated with an attention mechanism is introduced to classify e-commerce data. The model includes embedding layer, BiLTM layer, attention mechanism layer, and output layer. The embedding layer loads the word vector trained by Word2Vec; the BiLSTM layer captures the context of each word; the attention mechanism layer allocates weights for each word to synthesize new sample features. The experimental results show that the classification accuracy of the attention mechanism based on the inverse class frequency reaches 91.93%, which is improved compared with the BiLSTM model without the attention mechanism and other attention mechanisms introduced. This model has a good effect in the classification of e-commerce data and points out a new thinking direction for the introduction of attention mechanisms.

Key words: e-commerce data; text classification; Word2Vec; BiLSTM; attention

① 基金项目: 武汉科技大学大学生创新创业训练计划 (18ZRA078); 国家社会科学基金重大计划 (11&ZD189)
Foundation item: Innovation Training Program for Students of Wuhan University of Science and Technology (18ZRA078); Major Program of National Social Science Foundation of China (11&ZD189)
收稿时间: 2020-09-03; 修改时间: 2020-09-25; 采用时间: 2020-09-29; csa 在线出版时间: 2021-04-28

随着电子商务的快速发展, 电商数据呈指数式增长. 这些数据承载了各类消费群体的信息, 成为了极有价值的资产, 应用大数据正逐渐成为商业竞争的关键. 大数据的发展, 为企业带来了新的生产革命, 带来了一系列的机遇. 基于互联网实现电子商务数据收集, 大数据分析促进了企业客户服务的差异化, 强化了市场营销的针对性, 增强了电子商务企业竞争力. 而对电商数据的文本分类, 是进行一切数据分析的基础.

依靠人力对电商数据进行分类, 无法适应如今海量的电商数据且成本过高. 传统机器学习难以捕捉有用的人工特征, 深度学习可以自动提取特征. 近年来, 随着深度学习的快速发展, 深度学习在自然语言处理领域取得了良好的效果. 采用深度学习模型如卷积神经网络 (Convolutional Neural Networks, CNN)[1]、循环神经网络 (Recurrent Neural Network, RNN)[2] 以及长短期记忆神经网络 (Long Short-Term Memory, LSTM)[3]等进行文本分类, 都取得了比传统机器学习模型更好的结果.

文献 [4] 中, 作者使用深度学习模型对新闻文本进行分类. 采用的模型包括 BPNN 模型, BiLSTM 模型, TextCNN 模型和 BiLSTM+TextCNN 模型, 除了 BPNN 模型之外, 其他模型的 $F1$ 值均超过 0.9, 取得了较好的分类效果[4]. 这些模型取得的分类成果, 足以说明将深度学习模型应用于文本分类, 是一个不错的选择.

Attention 机制模拟人类注意力机制, 对信息中的关键部分进行着重关注, 最早在计算机视觉领域被提出. 2014 年, Google Mind 团队发表的论文《Recurrent Models of Visual Attention》真正让 Attention 机制大火. 作者在 RNN 模型中引入 Attention 机制进行图像分类, 取得了良好的分类效果[5]. 2015 年, Bahdanau 等在文献《Neural Machine Translation by Jointly Learning to Align and Translate》中将 Attention 机制引入机器翻译中, 这是 Attention 机制首次在 NLP 任务中应用[6]. 随后, Attention 机制被广泛应用于 NLP 的各个领域. 文献 [7] 中, 作者在 BiLSTM 模型中引入 Attention 机制对招聘信息进行分类, 分类准确率达到 93.36%, 与其他没有引入 Attention 机制的模型相比, 提高约 2%[7]. 可见, 注意力机制在文本分类中有良好的作用[8].

本文主要针对电商数据的文本分类, 分类过程中对上下文有较强依赖, 同时, 某些关键词对分类结果也有较强影响. BiLSTM 模型在对词语进行编码时, 可以充分考虑上下文信息. $Tf-idf$ 值则可以衡量一个词语对一个文档的重要性, 但忽略了文档的类别信息[9,10]. 本文由逆文档率 idf 的概念提出逆类别率 icf, 评估一个词语对一个类别的重要性, 并以此引入注意力机制. 将此模型的实验结果与未引入注意力机制的模型和以其他方式引入注意力机制的模型的实验结果进行对比, 验证基于逆类别率的注意力机制在电商文本分类中的有效性.

1 数据处理

1.1 文本预处理

本文的数据来源于第 9 届中国大学生服务外包创新创业大赛中企业方提供的真实的电商数据, 从全部数据集中取部分样本, 数量为 156 788. 一共有 24 个类别, 且每个类别的样本数量分布不均衡. 为了确保模型能够充分学习到每个类别的特征, 本文采用分层抽样的方法, 将数据集划分为训练集和测试集, 样本比例为 7:3, 划分后的数据集如图 1 所示. 对数据集使用 jieba 中文分词工具进行分词处理, 接着对分词结果进行停用词过滤, 停用词表为哈工大停用词表.

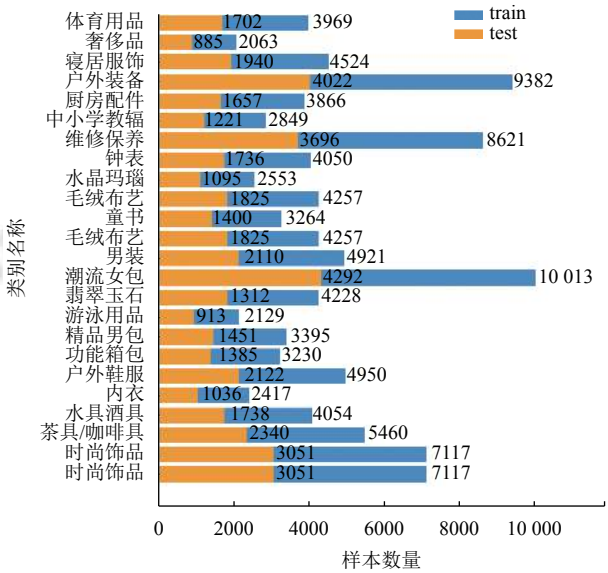

图 1 样本数量分布

1.2 生成词向量

使用 Word2Vec 开源工具的 CBOW 模型训练词向量, 可以充分考虑到每个词语的上下文信息, 训练得到的词向量维度为 64. 下文中提到的所有深度学习模型都将使用 Word2Vec 训练得到的词向量.

2 模型描述与实现

2.1 BiLSTM 模型

双向长短期记忆网络 (Bi-directional Long Short-Term Memory, BiLSTM) 是一种时序循环神经网络, 是由前向长短期记忆网络 (Long Short-Term Memory, LSTM) 和反向长短期记忆网络 (Long Short-Term Memory, LSTM) 组成. LSTM 的提出是为了解决循环神经网络 (Recurrent Neural Network, RNN) 在长距离训练过程中存在的梯度消失和梯度爆炸问题, 因此 LSTM 在结构设计中引入了门控机制, 包含 3 种门: 遗忘门、输入门、输出门. 遗忘门决定上一时刻的细胞状态有多少信息需要被遗忘, 输入门决定当前时刻的输入中有多少信息需要被添加, 输出门决定当前的细胞状态有多少信息需要被输出. 通过这 3 种门控机制可以很容易解决 RNN 在长距离训练中存在的梯度消失和梯度爆炸的问题.

LSTM 模型的第 1 步是通过遗忘门来计算f_t, 决定上一时刻的细胞状态C_{t-1}中哪些信息需要被遗忘. 具体实现方式是, 将h_{t-1}和x_t连接, 再通过遗忘门的权重矩阵W_f, 最后再经过 Sigmoid 激活函数σ, 得到一个 0–1 的值, 值越小, 表示上一时刻的细胞状态C_{t-1}中需要遗忘的信息越多.

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{1}$$

第 2 步生成新的细胞状态C_t. 首先, 与遗忘门的操作类似, 将h_{t-1}和x_t连接, 再通过输入门的权重矩阵W_i, 最后再经过 Sigmoid 激活函数σ, 求得i_t来决定更新哪些信息. 然后, 将h_{t-1}和x_t连接, 再通过权重矩阵W_C, 最后经过 tanh 激活函数, 得到新的细胞候选状态$\tilde{C}_t$. 最后, 使用上一时刻的细胞状态C_{t-1}和新的细胞候选状态$\tilde{C}_t$来生成新的细胞状态C_t.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2}$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{3}$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{4}$$

第 3 步决定细胞状态中哪些信息需要被输出. 首先, 将h_{t-1}和x_t连接, 再通过输出门的权重矩阵W_o, 最后再经过 Sigmoid 激活函数σ, 得到输出门的判断条件o_t. 最后, 将细胞状态C_t经过 tanh 层将数值规范化, 再与输出门的判断条件o_t相乘, 得到当前时刻的输出.

$$o_t = \sigma(W_t \cdot [h_{t-1}, x_t] + b_o) \tag{5}$$

$$h_t = o_t * \tanh(C_t) \tag{6}$$

以上, 就是 LSTM 模型的一次流程, LSTM 可以编码句子中从前到后的信息, 但是无法编码从后到前的信息. 本文采用的 BiLSTM 模型由前向 LSTM 模型和后向 LSTM 模型构成, 因此, 既能编码从前到后的信息同时又能编码从后到前的信息, 可以更好的捕捉双向的语义依赖.

本文使用 PyTorch 实现 BiLSTM 模型, 模型结构如图 2. 其中x_i表示文本中第 i 个词语对应的词向量, 由 Word2Vec 训练得到. 经过 BiLSTM 模型后, 取最后一个词语的两个隐藏层状态进行拼接得到向量 h, 再将向量 h 经过 Softmax 层求得样本属于每一个类别的概率.

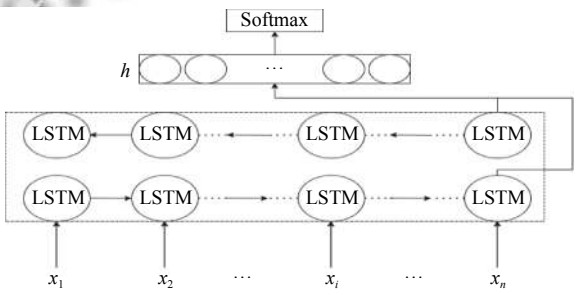

图 2 BiLSTM 模型结构图

2.2 注意力机制

近年来注意力机制被广泛应用于深度学习的各个领域, 都取得了良好的效果. 其模拟人脑中的注意力分配机制, 对信息中比较关键的部分进行着重关注. 本文以 3 种方案对 BiLSTM 模型添加注意力机制.

方案一. 按照文献 [7] 中的方式为 BiLSTM 添加注意力机制, 此模型为 BiLSTM+Attention1 模型. 计算过程如下:

$$u_t = \tanh(W_u \cdot H_t + b_u) \tag{7}$$

$$a_t = \frac{\exp\left(u_t^{\mathrm{T}} u_w\right)}{\sum_j \exp\left(u_j^{\mathrm{T}} u_w\right)} \tag{8}$$

其中, H_t是 t 时刻, 前向 LSTM 和后向 LSTM 的隐藏层状态的拼接而成. W_u, b_u以及u_w是模型需要学习的参数. a_t为计算得到的权重值, 表示 t 时刻的隐藏层状态对文本分类的贡献程度.

方案二. $tf-idf$值可以衡量一个词语对一个文档的重要性, 计算文档中每个词语$tf-idf$值, 将计算得到的$tf-idf$值经过 Softmax 函数得到一个文档中每个词语的权重, 以此来添加注意力机制, 此模型为 BiLSTM+

Attention2 模型. 计算过程如下:

$$tf_i = \text{词语}i\text{在文档中出现的次数} \tag{9}$$

$$idf_i = \log\left(\frac{1+\text{文档总数}}{1+\text{包含词语}i\text{的文档个数}}\right)+1 \tag{10}$$

$$tf-idf_i = tf_i * idf_i \tag{11}$$

$$a_i = \frac{\exp(tf-idf_i)}{\sum_j \exp\left(tf-idf_j\right)} \tag{12}$$

方案三. $Tf-idf$值可以衡量一个词语对一个文档的重要性, 但是忽略了一个词语对于一类文档的重要性. 因此, 在方案二的基础上, 提出逆类别率icf的概念, 用来衡量一个词语对一类文本的重要性, 再经过 Softmax 函数计算一个文档中每个词语的权重, 并以此来添加注意力机制, 此模型为 BiLSTM+Attention3 模型. 计算过程如下:

$$icf_i = w_1 * \frac{1+\text{类别总数}}{1+\text{包含词语}i\text{的类别个数}} + w_2 * \log\left(\frac{1+\text{类别总数}}{1+\text{包含词语}i\text{的类别个数}}\right) \tag{13}$$

$$a_i = \frac{\exp(icf_i)}{\sum_j \exp\left(icf_j\right)} \tag{14}$$

其中, w_1和w_2为模型参数, 本文中, 将w_1值设为 0.001, w_2的值设为 1.2 时, 模型的分类效果最优.

3 实验结果与分析

3.1 实验结果

本文共有 4 组对照实验, 训练次数均为 15 次, 学习率为 0.001. 模型分别为 BiLSTM 模型, BiLSTM+Attention1 模型, BiLSTM+Attention2 模型, BiLSTM+Attention3 模型, 实验结果分别见表 1 至表 4. 精确率 (Precision), 召回率 (Recall), $F1$ 值 ($F1$-score) 的加权平均值对比如图 3 所示. 加权平均值可以很好的反映模型在测试集上的分类效果. 4 个模型在每个类别上的预测准确率如图 4 所示.

4 类模型分类的准确率均超过 90%. BiLSTM+Attention2 模型和 BiLSTM+Attention3 模型均比 BiLSTM 模型的分类效果更好, BiLSTM+Attention1 模型的分类效果最差. BiLSTM+Attention3 模型在 Precision, Recall 以及 $F1$-score 值上均是最大的, 是 4 类模型中最优的, 但分类准确率的提升并不大.

表 1 BiLSTM 分类结果

类别	Precision	Recall	$F1$-score	Support
体育用品	0.9168	0.9324	0.9246	1702
奢侈品	0.8332	0.8689	0.8507	885
寝居服饰	0.8983	0.9428	0.9200	1940
户外装备	0.9142	0.9060	0.9101	4022
厨房配件	0.9058	0.9107	0.9082	1657
中小学教辅	0.9224	0.8952	0.9086	1221
维修保养	0.9603	0.9759	0.9681	3696
钟表	0.9557	0.9447	0.9502	1736
水晶玛瑙	0.8427	0.9005	0.8706	1095
毛绒布艺	0.9793	0.9841	0.9817	1825
童书	0.8322	0.9036	0.8664	1400
家装软饰	0.9728	0.9455	0.9589	2384
男装	0.8798	0.7284	0.7970	2110
潮流女包	0.8934	0.9310	0.9118	4292
翡翠玉石	0.9275	0.8830	0.9047	1812
游泳用品	0.9682	0.9682	0.9682	913
精品男包	0.8678	0.8277	0.8473	1451
功能箱包	0.9118	0.8585	0.8843	1385
户外鞋服	0.8247	0.9270	0.8729	2122
内衣	0.8712	0.8485	0.8597	1036
水具酒具	0.9155	0.9419	0.9285	1738
茶具/咖啡具	0.9628	0.9397	0.9511	2340
时尚饰品	0.9397	0.9197	0.9296	3051
妈妈专区	0.9363	0.9173	0.9267	1234
Macro avg	0.9097	0.9084	0.9083	47047
Weighted avg	0.9149	0.9141	0.9138	47047

表 2 BiLSTM+Attention1 分类结果

类别	Precision	Recall	$F1$-score	Support
体育用品	0.9454	0.9154	0.9301	1702
奢侈品	0.8235	0.8542	0.8386	885
寝居服饰	0.8916	0.9459	0.9180	1940
户外装备	0.8656	0.9224	0.8931	4022
厨房配件	0.9000	0.9016	0.9008	1657
中小学教辅	0.9086	0.9034	0.9060	1221
维修保养	0.9512	0.9765	0.9637	3696
钟表	0.9663	0.9401	0.9530	1736
水晶玛瑙	0.8649	0.8831	0.8739	1095
毛绒布艺	0.9792	0.9797	0.9795	1825
童书	0.8383	0.8921	0.8644	1400
家装软饰	0.9609	0.9484	0.9546	2384
男装	0.8552	0.7611	0.8054	2110
潮流女包	0.9201	0.8993	0.9096	4292
翡翠玉石	0.9193	0.8990	0.9090	1812
游泳用品	0.9651	0.9693	0.9672	913
精品男包	0.8661	0.8160	0.8403	1451
功能箱包	0.8460	0.8809	0.8631	1385
户外鞋服	0.8472	0.8911	0.8686	2122
内衣	0.8978	0.8137	0.8537	1036
水具酒具	0.9405	0.9102	0.9251	1738
茶具/咖啡具	0.9518	0.9457	0.9488	2340
时尚饰品	0.9234	0.9204	0.9219	3051
妈妈专区	0.9352	0.9117	0.9233	1234
Macro avg	0.9068	0.9034	0.9047	47047
weighted avg	0.9104	0.9099	0.9097	47047

表 3 BiLSTM+Attention2 分类结果

类别	Precision	Recall	*F*1-score	Support
体育用品	0.9259	0.9254	0.9257	1702
奢侈品	0.8567	0.8576	0.8571	885
寝居服饰	0.8931	0.9608	0.9258	1940
户外装备	0.8997	0.9142	0.9069	4022
厨房配件	0.9077	0.9137	0.9107	1657
中小学教辅	0.9008	0.9296	0.9150	1221
维修保养	0.9718	0.9705	0.9712	3696
钟表	0.9588	0.9384	0.9485	1736
水晶玛瑙	0.8418	0.9087	0.8740	1095
毛绒布艺	0.9787	0.9825	0.9806	1825
童书	0.8641	0.8900	0.8768	1400
家装软饰	0.9720	0.9476	0.9596	2384
男装	0.8573	0.7773	0.8153	2110
潮流女包	0.9040	0.9192	0.9115	4292
翡翠玉石	0.9519	0.8637	0.9057	1812
游泳用品	0.9642	0.9726	0.9684	913
精品男包	0.8382	0.8284	0.8333	1451
功能箱包	0.8927	0.8773	0.8849	1385
户外鞋服	0.8555	0.9015	0.8779	2122
内衣	0.8878	0.8176	0.8513	1036
水具酒具	0.9598	0.8936	0.9255	1738
茶具/咖啡具	0.9315	0.9594	0.9453	2340
时尚饰品	0.9289	0.9377	0.9333	3051
妈妈专区	0.9146	0.9287	0.9216	1234
Macro avg	0.9107	0.9090	0.9094	47047
weighted avg	0.9155	0.9151	0.9149	47047

表 4 BiLSTM+Attention3 分类结果

类别	Precision	Recall	F1-score	Support
体育用品	0.9528	0.9254	0.9389	1702
奢侈品	0.9011	0.8136	0.8551	885
寝居服饰	0.9247	0.9309	0.9278	1940
户外装备	0.903	0.9217	0.9123	4022
厨房配件	0.9195	0.9173	0.9184	1657
中小学教辅	0.9387	0.9034	0.9207	1221
维修保养	0.9524	0.9808	0.9664	3696
钟表	0.9668	0.9407	0.9536	1736
水晶玛瑙	0.8574	0.9169	0.8861	1095
毛绒布艺	0.973	0.9858	0.9793	1825
童书	0.8653	0.9129	0.8884	1400
家装软饰	0.9689	0.9534	0.9611	2384
男装	0.7946	0.8706	0.8308	2110
潮流女包	0.9157	0.9091	0.9124	4292
翡翠玉石	0.9294	0.8935	0.9111	1812
游泳用品	0.9683	0.9704	0.9694	913
精品男包	0.8091	0.8677	0.8374	1451
功能箱包	0.9086	0.8758	0.8919	1385
户外鞋服	0.9114	0.8487	0.879	2122
内衣	0.8679	0.8755	0.8717	1036
水具酒具	0.9335	0.9287	0.9311	1738
茶具/咖啡具	0.9592	0.9436	0.9513	2340
时尚饰品	0.9408	0.9266	0.9336	3051
妈妈专区	0.9514	0.9198	0.9353	1234
Macro avg	0.9172	0.9139	0.9151	47047
Weighted avg	0.9204	0.9193	0.9195	47047

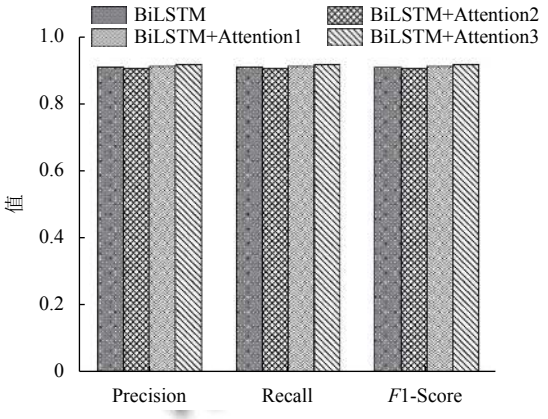

图 3 4 种方法分类效果对比

表 5 类别数据对比

潮流女包	奢侈品
Nickent 女士 化妆 手提包 BB176 黑色	华伦天奴 (Valentino) 女包 带锁 中号 皮革肩包
编织 手拿包 女士 钱包 橙色	MCM中号 双肩包 MMK5SVE38PK
淑女 贝壳包 手提包 定型包 黑色	Gucci古驰 女士 双面用 购物袋
2016春夏新品韩版单肩时尚小方包灰色	Armani jeans 阿玛尼 商务 女式 手提包 袋 Z5255V3 黑色
乔丹 休闲 时尚 儿童 双肩包 红色 其他	托里 伯奇 (Tory Burch) 女士 金属 皮革 夹趾人字拖 凉鞋 50008679_710 34.5

3.2 原因分析

取潮流女包类和奢侈品类各 5 条数据进行比较, 如表 5 所示. 两个类别样本的相关性比较高, 甚至存在包含关系, 如: 潮流女包属于奢侈品. 类似关系的还有: 童书与中小学教辅, 男装与户外鞋服, 茶具/咖啡具与水具酒具, 内衣与妈妈专区等. Attention 机制为文本中的关键词分配更多的权重, 当文本比较相近时, Attention 机制起到的效果会有折扣.

4 结论与展望

逆类别率*icf*可以评估一个词语对一个类别的重要性, 基于逆类别率*icf*引入注意力机制的 BiLSTM+Attention3 模型, 在 4 类模型中分类效果最好, *F*1 值最大, 在电商数据分类问题上表现相对最好. 但文章仍然存在不足, 逆类别率*icf*并没有考虑词语的位置信息, 词语的位置信息对于文档的语义有一定影响, 将在后续的研究中不断完善.

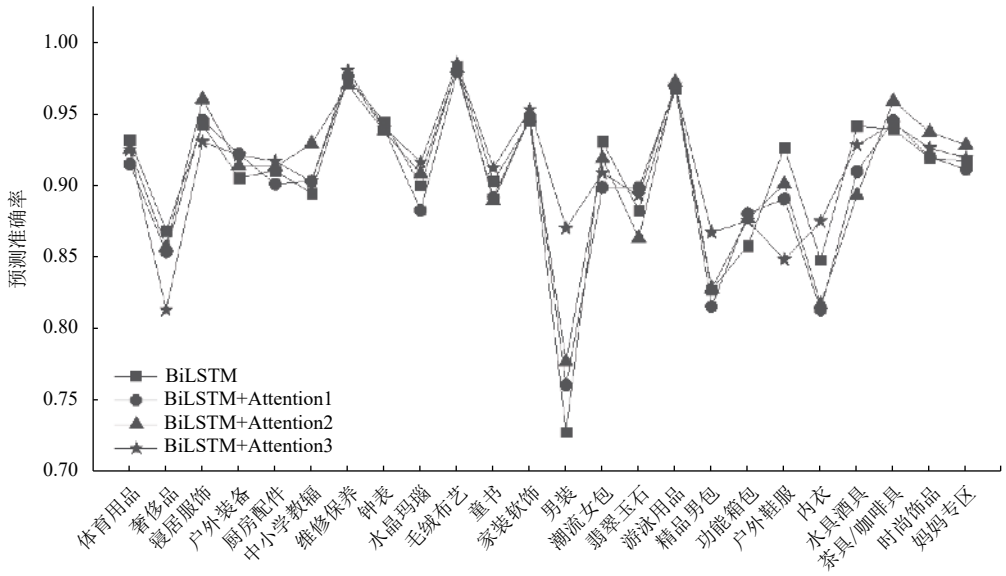

图 4 4 类模型在 24 个类别上的预测准确率

参考文献

1 Gu JX, Wang ZH, Kuen J, *et al*. Recent advances in convolutional neural networks. Pattern Recognition, 2018, 77: 354–377. [doi: 10.1016/j.patcog.2017.10.013]

2 Chien JT, Lu TW. Deep recurrent regularization neural network for speech recognition. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). South Brisbane, QLD, Australia. 2015. 4560–4564.

3 Staudemeyer RC, Morris ER. Understanding LSTM — a tutorial into long short-term memory recurrent neural networks. arXiv: 1909.09586, 2019.

4 涂文博, 袁贞明, 俞凯. 针对文本分类的神经网络模型. 计算机系统应用, 2019, 28(7): 145–150. [doi: 10.15888/j.cnki.csa.006972]

5 Mnih V, Heess N, Graves A, *et al*. Recurrent models of visual attention. Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, BC, Canada. 2014. 2204–2212.

6 Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv: 1409.0473, 2014.

7 吕飞亚, 张英俊, 潘理虎. 注意力机制的 BiLSTM 模型在招聘信息分类中的应用. 计算机系统应用, 2020, 29(4): 242–247. [doi: 10.15888/j.cnki.csa.007364]

8 Wu TH, Hsieh CC, Chen YH, *et al*. Hand-crafted attention is all you need? A study of attention on self-supervised audio transformer. arXiv: 2006.05174v1, 2020.

9 胡万亭, 贾真. 基于加权词向量和卷积神经网络的新闻文本分类. 计算机系统应用, 2020, 29(5): 275–279. [doi: 10.15888/j.cnki.csa.007391]

10 牛雪莹, 赵恩莹. 基于 Word2Vec 的微博文本分类研究. 计算机系统应用, 2019, 28(8): 256–261. [doi: 10.15888/j.cnki.csa.007030]