文章编号：2095-2163(2023)12-0102-05 中图分类号：TP391.1 文献标志码：A

基于多模态的图文情感分析

孙文飞，张云华
（浙江理工大学 信息学院，杭州 310018）

摘 要：因特网飞速发展的今天，人们更倾向于将文字与图片相结合来发表自己的评论，而单一模态的情感分析精度较低，本文提出 BiGRU-ResNet 图文多模态情感分析模型，用于情感分类任务。首先，利用 BERT 将文本嵌入到词向量中；其次，通过 BiGRU 并引入注意力层对上游任务的词向量进行特征提取，图像的特征提取由 ResNet 来完成并保留更为有效的信息；最后，文本模态和图像模态使用注意力机制和张量运算来达到增益的目的，再将融合特征输入至分类器中，得到评论的情感分类。通过实验分析与对比表明，发现多模态模型相较于单模态模型可以提高情感分类任务的精确度。

关键词：多模态；BiGRU；ResNet；注意力机制；情感分析

Image-text sentiment analysis based on multi-modal

SUN Wenfei, ZHANG Yunhua
(School of Information Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: Today, with the rapid development of the Internet, people are more likely to combine words and pictures to publish their comments, the accuracy of sentiment analysis of single mode is low, therefore, this paper proposes a BiGRU-ResNet image-text multimodal sentiment analysis model for sentiment classification tasks. Firstly, BERT is used to embed the text into the word vector, secondly, BiGRU is used and the attention layer is introduced to extract the features of the word vector of the upstream task, then, the feature extraction of the image is completed by ResNet and more effective information is retained, finally, the attention mechanism and tensor operation were used to achieve the purpose of gain in the text modality and image modality, and the fused features were input into the classifier to obtain the sentiment classification of the review. Through experimental analysis and comparison, it is found that multi-modal model can improve the accuracy of emotion classification tasks compared with single modal model.

Key words: multi-modal; BiGRU; ResNet; attention mechanism; sentiment analysis

0 引 言

随着互联网的快速发展，越来越多的人们在微博、推特、新闻等各种社交平台发表评论。在发表评论的过程中，一段文本、一张图片就可以判断目标对象对客观事物的情感状况，准确地把握目标对象的情感可以在推荐服务、质量把控、事件剖析等方面提升应用价值。情感分析最早是通过情感词典来实现的，这种方法的准确度取决于词典中情感词汇的数量，随着时间的推移，词汇的数量也在不断增加，使得词典的维护越来越困难[1]。将机器学习应用于情感分析领域，通过各种机器学习算法来设计结合情感词权重的情感分析计算方法，模型训练依靠数据集标注的质量，较高质量的数据集需要较高的人工成本[2]。目前深度学习在许多领域中扮演重要角色，并且在情感分析领域中也广泛应用，提高了情感分类任务的准确度。

在单模态情感分析研究中，在文本单模态方面，Kim[3]研究一维卷积神经网络在句子层面上进行情感分类；Ma 等[4]提出在 Sentic-LSTM 的基础上引入注意力层来进行方面级的情感分析。在图像单模态方面，Chen 等[5]提出 DeepSentiBank 模型来提高视觉情感分类的准确度；Yang 等[6]提出在图像中分别提取全局和局部特征用于图像情感分类；Campos 等[7]提出在图像各个层次分别抽取深度特征，以提高图像情感分类的精确度。

在多模态情感分析研究中，各模态特征的融合会影响模型最终的情感分类性能。特征层融合方法

作者简介：孙文飞(1999-)，男，硕士研究生，主要研究方向：计算机视觉与模式识别。
通讯作者：张云华(1965-)，男，博士，研究员，主要研究方向：软件工程、智慧医疗、智能信息处理。Email：605498519@qq.com

收稿日期：2022-12-12

是通过拼接的方式将不同模态的特征进行融合，Poria 等[8]提出通过向量拼接技术将图像、文本、语音 3 种模态特征进行融合，然后将融合特征输入至多层感知机中，从而实现对情感的分类。对于中间层融合，通过编码不同模态的特征进行融合，Huang 等[9]提出在提取不同模态的特征后，通过注意力机制以一个模态为基准和另一模态进行特征融合，键(Key)和值(Value)同源，查询(Query)与 Key、Value 不同源。对于决策层融合，采用一些规则或神经网络，如通过权重加权融合、K 最邻近算法等，Song 等[10]提出将不同模态对应模型输出的情感分类输入到 K 最邻近算法或神经网络中，得到最终的情感分类结果。

为了在多模态场景下提升情感分析模型的性能，本文提出 BiGRU-ResNet 图文多模态情感分析模型，用于情感分类任务，利用引入注意力层的 BiGRU 和 ResNet 分别提取本文和图像特征；抽取特征后，通过注意力机制和张量运算交互不同模态特征，获得融合特征后进行情感分类，进一步挖掘情感信息。经过实验进行对比验证，本文构建的 BiGRU-ResNet 图文多模态情感分析模型在多模态场景下相较于单模态情感分析模型在单一模态场景下有更高的准确度。

1　相关知识

1.1　BiGRU

循环神经网络(Recurrent Neural Network，RNN)会出现短时记忆和梯度消失等问题，长短时记忆神经网络(Long Short Term Memory，LSTM)和门控循环单元网络(Gated Recurrent Unit，GRU)都是基于 RNN 进行改进的模型，两者都是通过门控机制来调节信息流[11]。GRU 相比于 LSTM 从三个门优化为两个门，门控机制由更新门、重置门两个模块组成，减少了训练参数，模型性能有所提升，GRU 神经网络结构如图 1 所示。

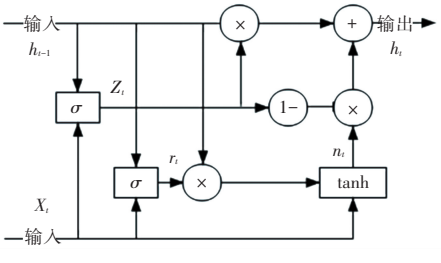

图 1　GRU 神经网络结构

Fig. 1　GRU neural network architecture

如图 1 所示，h_{t-1} 表示前一时刻的输出；x_t 表示当前时刻的输入；z_t 表示当前时刻更新门状态；r_t 表示当前时刻重置门状态；$\boldsymbol{w}_{xr}$ 和 $\boldsymbol{w}_{xz}$ 表示当前时刻的输入 x_t 分别和重置门、更新门的权重矩阵；$\boldsymbol{w}_{hr}$ 和 $\boldsymbol{w}_{hz}$ 表示前一时刻的输出 h_{t-1} 分别和重置门、更新门的权重矩阵；b_r 表示重置门的偏量；b_z 表示更新门的偏量；σ 表示 Sigmoid 激活函数；重置门和更新门如式(1)和式(2)所示：

$$r_t = \sigma(\boldsymbol{w}_{xr} x_t + \boldsymbol{w}_{hr} h_{t-1} + b_r) \tag{1}$$

$$z_t = \sigma(\boldsymbol{w}_{xz} x_t + \boldsymbol{w}_{hz} h_{t-1} + b_z) \tag{2}$$

在此基础上可以计算当前时刻信息 h_t，从而通过门控机制保留有效信息，如计算公式(3)和公式(4)所示：

$$n_t = \tanh(\boldsymbol{w}_{xn} x_t + b_{xn} + r_t \circ (\boldsymbol{w}_{hn} h_{t-1} + b_{hn})) \tag{3}$$

$$h_t = (1 - z_t) \circ n_t + z_t \circ h_{t-1} \tag{4}$$

其中，$\boldsymbol{w}_{xn}$ 和 $\boldsymbol{w}_{hn}$ 表示权重矩阵；b_{xn} 表示偏量，tanh 为激活函数；$\circ$ 表示哈达玛积。

在单向神经网络中，状态通常是单向输出的，而在文本情感分类中，采用双向结构可以获得前后两个时刻的信息，这样可以更好地抽取文本特征。双向门控循环神经网络(Bidirectional Gated Recurrent Neural Network)，模型结构如图 2 所示。BiGRU 是一个由两个方向的输出共同确定最终输出的模型[12]。

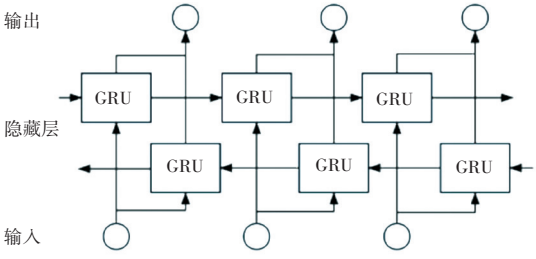

图 2　BiGRU 模型结构

Fig. 2　BiGRU model structure

1.2　ResNet

神经网络的层数越深可以提取的特征越多，而神经网络深度较深时会出现退化问题。为了解决深度神经网络在层数越深误差越大的缺陷，残差网络可以抽取到更多有效的视觉特征，如图 3 所示，残差网络中起到核心作用的是跳跃结构，跳跃结构连接着输入端和输出端，如公式(5)所示，x 表示跳跃结构的输入，$F(x)$ 表示未经过跳跃结构的输出，$H(x)$ 表示经过跳跃结构后的输出，计算量并不会因为跳跃结构的出现而增加，在神经网络的层数很深的情况下也不会出现梯度爆炸和梯度消失的问题[13]。

$$H(x) = F(x) + x \tag{5}$$

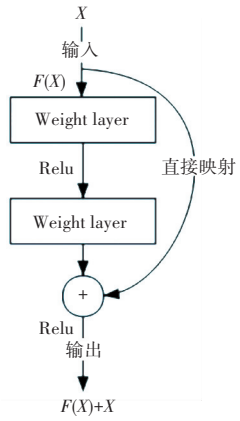

图 3 跳跃结构

Fig. 3 Skip structure

1.3 BERT

BERT 预训练模型有更加出色的表征能力，如图 4 所示。BERT 主要由输入嵌入、双向 Transformer 编码器、无监督任务 3 个模块组成[14]。

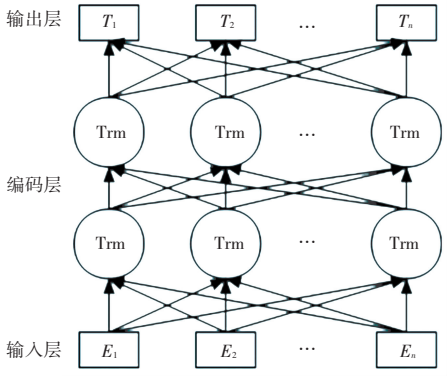

图 4 BERT 结构

Fig. 4 BERT architecture

如图 4 所示，$\boldsymbol{E}_i$ 表示第 i 个输入向量，Trm 表示 Transformer 的编码器，$\boldsymbol{T}_i$ 表示第 i 个输出向量，通过 MLM(Masked Language Model)、NSP(Next Sentence Prediction)两个无监督任务进行预训练，为下游任务提供服务，提高后续情感分类的准确度。

1.4 注意力机制

注意力机制的思想借鉴人类对客观事物的思维方式，得到客观事物中的重点关注目标以及更多细节信息[15]。注意力机制如图 5 所示，$\boldsymbol{X}_i$ 表示第 i 个输入向量，***Query*** 表示查询向量，Key_i 表示选取信息的索引下标，$F(Q,K_i)$ 表示注意力相关性计算规则，S_i 表示通过相似注意力相关性计算规则得到的数值，α_i 表示注意力分数通过 SoftMax 归一化的数值，α_i 与 X_i 进行加权求和可以得到最终的注意力数值，通过注意力机制可以提升模型分类的准确度。

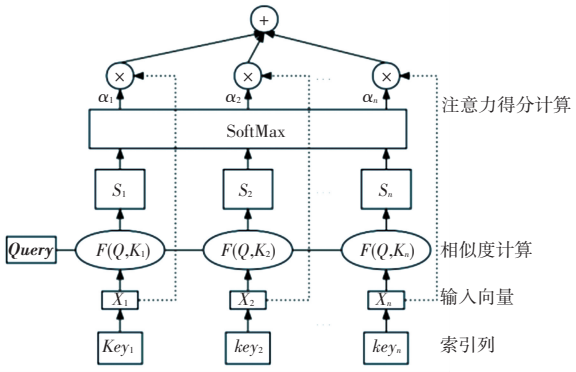

图 5 注意力机制

Fig. 5 Attention mechanism

2 模型构建

本文提出 BiGRU-ResNet 多模态情感分析模型，模型结构如图 6 所示，由图文数据输入层、文本特征提取层、图片特征提取层、模态间注意力特征融合层、分类器层 5 个模块组成。

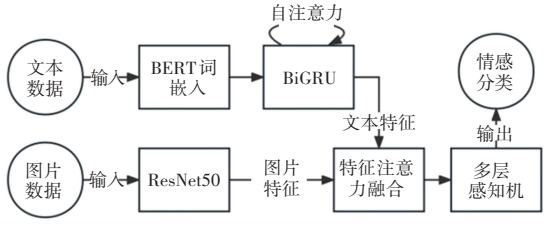

图 6 BiGRU-ResNet 模型结构

Fig. 6 BiGRU-ResNet model architecture

2.1 文本特征提取

文本特征提取结构如图 7 所示，本文方法对于文本的特征提取由 BERT 预训练层、BiGRU 层、注意力层组成。首先，在 BERT 中输入文本数据并进行预训练，然后将每个词嵌入到相应的词向量中；BERT 输出的词向量为 BiGRU 的输入，经过前向 GRU 与反向 GRU 得到抽取过后的文本特征；最后，利用自注意力得到注意力得分，得到更需要关注的文本特征。

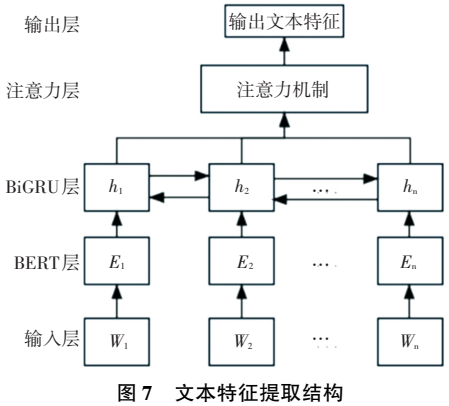

图 7 文本特征提取结构

Fig. 7 Text feature extraction architecture

2.2　图像特征提取

图像特征提取结构如图 8 所示。本文的图片特征提取以 ResNet50 为基准，提取图像有效的特征信息，发挥 ResNet50 的优势，解决网络层数较深时的退化问题。ResNet50 由输入层、4 个残差层、输出层 3 个模块组成，输入层包含卷积、批归一化、Relu 函数、最大池化，批归一化和 Relu 函数用来提高网络的拟合能力。残差层包含 Conv 块和 Identity 块两种残差，由于输入和输出的维度是不同的，用 Conv 块进行维度转换；由于输入和输出的维度是一样的，用 Identity 块加深网络层次。输出层由平均池化层、展平层、多层感知机 3 个模块组成，通过输出层得到抽取后更为有效的图像特征。

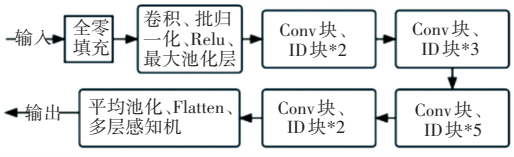

图 8　图像特征提取结构

Fig. 8　Image feature extraction architecture

2.3　特征融合

在得到文本和图像两个模态的特征后，需要解决融合不同模态的特征的问题。特征融合的优劣会影响最终的情感分类的准确度，常见的不同模态融合方式有拼接融合、注意力融合、张量融合等。拼接融合是将同一个维度的不同模态特征通过向量拼接技术进行融合，简单方便，并且节省计算时间，但会造成特征信息的丢失，影响情感分类的准确性；注意力融合是在特征拼接之前引入注意力机制，通过不同模态间信息充分交互，弥补特征拼接融合方式的缺陷；张量融合是通过张量积对不同模态特征进行交互，张量作为向量或矩阵的高阶扩展，可以充分挖掘模态间的特征。

本文基于注意力机制和张量运算相结合的方式来进行特征融合。经过注意力加权后文本特征 T_i 和图像特征 P_i 如公式(6)～公式(7)所示：

$$T_n = W_T^T \cdot \alpha_i \tag{6}$$

$$P_n = W_P^T \cdot \alpha_i \tag{7}$$

其中，α_i 表示注意力权重值。

首先，使用点积运算计算出相关性，即注意力权重，将注意力权重分别与文本特征和图像特征进行计算，可以得到两个模态特征向量的注意力值，最后，再通过张量积求出多模态融合特征，如公式(8)所示：

$$M = P_n \otimes T_n \tag{8}$$

3　实验与分析

3.1　数据集与实验环境参数

本文采用某社交平台上的评论作为数据集进行实验，每条评论都包含一段文字和一张图片，共8 268 条。为了提升本文模型的泛化能力，将样本以 4∶1 的比例拆分成训练集和测试集。参数的初始化见表 1。

表 1　参数初始化

Table 1　Parameter initialization

序号	参数	参数值
1	Batch 大小	8
2	学习率	1e-5
3	Dropout	0.2
4	Epochs	10
5	优化器	Adam

3.2　评价指标

为了检验本文所提出的 BiGRU-ResNet 多模态图文情感分析模型的性能，采用精确率(*Precision*)、召回率(*Recall*)、*F*1 值(*F*1) 作为评价指标。计算如公式(9)～公式(10)所示：

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{11}$$

其中，*TP* 表示样本中标为积极情感并且被预测为积极情感；*FP* 表示样本中标为消极情感并且被预测为积极情感；*FN* 表示样本中标为积极情感并且被预测为消极情感。

3.3　实验结果与对比

将本文提出的 BiGRU-ResNet 模型与 GRU、ResNet、BiGRU、ATT-BiGRU 这个模型进行对比实验，实验对比结果见表 2。

表 2　实验对比结果

Table 2　Experimental comparison results

序号	模型	精确率	召回率	*F*1 值
1	GRU	0.603 4	0.591 8	0.595 2
2	ResNet	0.615 0	0.601 7	0.611 6
3	BiGRU	0.628 8	0.634 3	0.633 1
4	ATT-BiGRU	0.649 3	0.641 7	0.645 9
5	BiGRU-ResNet	0.681 8	0.675 4	0.682 2

通过实验对比结果发现，GRU 和 ResNet 两个基线模型的精确率、召回率和 *F*1 值相近，因为 BiGRU 具有双向性，GRU 模型和 ResNet 模型的性能均低于 BiGRU 模型，ATT-BiGRU 模型因为加入了注意力机制，所以

性能高于 BiGRU,BiGRU-ResNet 因为分别对文本和图片进行特征提取,再进行注意力融合进行情感分类,所以 BiGRU-ResNet 相较于前面 4 个模型具有更高的性能。

4 结束语

本文针对社交平台评论多样化的特点,对评论进行多模态情感分析研究,并提出 BiGRU-ResNet 图文多模态情感分析模型用于情感分类。该模型在文本特征提取过程中利用 BERT 的预训练特性和 BiGRU 的双向性充分挖掘文本特征信息,在图像特征提取过程中利用 ResNet 在网络层数较深时也可以抽取有效特征并解决网络退化问题,考虑到不同模态特征的联系,通过注意力机制和张量运算进行特征融合并得到情感分类。经过实验对比,验证了本文模型在进行情感分类任务的精确率、召回率、*F*1 值相较于传统模型的性能有明显提升,未来还需要进一步的对比实验,使用其他数据集来检验该模型的泛化能力。

参考文献

[1] 吴杰胜,陆奎. 基于多部情感词典和规则集的中文微博情感分析研究[J]. 计算机应用与软件,2019,36(9):93-99.

[2] 戚天梅,过弋,王吉祥,等. 基于机器学习的外汇新闻情感分析[J]. 计算机工程与设计,2020,41(6):1742-1748.

[3] KIM Y. Convolutional neural networks for sentence classification [J]. arXiv preprint arXiv:1408.5882, 2014.

[4] MA Y, PENG H, CAMBRIA E. Targeted aspect-based sentiment analysisvia embedding commonsense knowledge into an attentive LSTM[J]. arXiv preprint arXiv 1609.12048,2018.

[5] CHEN T,BORTH D, DARRELL T, et al. Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks[J]. arXiv preprint arXiv:1410.8586, 2014.

[6] YANG J, SHE D, MING S, et al. Visual sentiment prediction based on automatic discovery of affective regions [J]. IEEE Transactions on Multimedia, 2018, 20(9):2513-2525.

[7] CAMPOS V, JOU B, GIRO-I-NIETO X. From pixels to sentiment: Fine-tuning CNNs for visual sentiment prediction[J]. Image and Vision Computing, 2017, 65: 15-22.

[8] PORIA S, CHATURVEDI I, CAMBRIA E, et al. Convolutional MKL based multimodal emotion recognition and sentiment analysis [C]//Proceedings of 2016 IEEE 16th International Conference on Data Mining (ICDM). IEEE, 2016: 439-448.

[9] HUANG F, ZHANG X, ZHAO Z, et al. Image-text sentiment analysis via deep multimodal attentive fusion [J]. Knowledge-Based Systems, 2019, 167(3):26-37.

[10] SONG K S, NHO Y H, SEO J H, et al. Decision-level fusion method for emotion recognition using multimodal emotion recognition information[C]//Prceedings of 2018 15th International Conference on Ubiquitous Robots (UR). IEEE, 2018: 472-476.

[11] 黄磊, 杜昌顺. 基于递归神经网络的文本分类研究[J]. 北京化工大学学报:自然科学版, 2017, 44(1):7.

[12] 程艳,尧磊波,张光河,等. 基于注意力机制的多通道 CNN 和 BiGRU 的文本情感倾向性分析[J]. 计算机研究与发展, 2020, 57(12):13.

[13] 吕梦棋,张芮祥,贾浩,等. 基于改进 ResNet 玉米种子分类方法研究[J]. 中国农机化学报,2021,42(4):92-98.

[14] KENTON J D M W C, TOUTANOVA L K. Bert: Pre-training of deep bidirectional transformers for language understanding[C]// Proceedings of naacL-HLT. 2019: 2.

[15] 李苏阳,陈富安. 基于注意力机制的双向 LSTM 锂电池 SOH 估算模型[J]. 电源技术,2022,46(7):739-742.

(上接第 101 页)

参考文献

[1] YUAN Q, YANG Z. A weight-coded evolutionary algorithm for the multidimensional knapsack problem[J]. arXiv preprint arXiv: 1302.5374, 2013.

[2] WU P, GAO L, ZOU D, et al. An improved particle swarm optimization algorithm for reliability problems [J]. ISA Transactions, 2011, 50(1): 71-81.

[3] YANG Q, GUO X, GAO X D, et al. Differential elite learning particle swarm optimization for global numerical optimization[J]. Mathematics, 2022, 10(8): 1261.

[4] MARTINS J P, RIBAS B C. A randomized heuristic repair for the multidimensional knapsack problem [J]. Optimization Letters, 2021, 15(2): 337-355.

[5] 贺毅朝,王熙照,李文斌,等. 基于遗传算法求解折扣{0-1}背包问题的研究[J]. 计算机学报,2016,39(12):2614-2630.

[6] 杨艳,刘生建,周永权. 贪心二进制狮群优化算法求解多维背包问题[J]. 计算机应用,2020,40(5):1291-1294.

[7] 刘雅文,蒋妍,潘大志. 改进二进制和声搜索算法求解多维背包问题[J]. 计算机与现代化,2022(8):13-19.

[8] 吴聪聪,赵建立,刘雪静,等. 改进的差分演化算法求解多维背包问题[J]. 计算机工程与应用,2018,54(11):153-160.

[9] YUAN Q, YANG Z. A weight-coded evolutionary algorithm for the multidimensional knapsack problem[J]. Eprint Arxiv, 2016, 6(10):659-675.

[10] 张晶,吴虎胜. 改进二进制布谷鸟搜索算法求解多维背包问题[J]. 计算机应用,2015,35(1):183-188.

[11] 王凌,王圣尧,方晨. 一种求解多维背包问题的混合分布估计算法[J]. 控制与决策,2011,26(8):1121-1125.

[12] 杨广益,欧阳智敏,全惠云. 松驰互补的分布估计算法求解多维背包问题[J]. 计算机工程与应用,2007,43(12):77-80.

[13] MARTINS J P R B C. A randomized heuristic repair for the multidimensional knapsack problem [J]. Optimization Letters, 2021, 15(2):337-355.

[14] 薛俊杰,王瑛,孟祥飞,等. 二进制反向学习烟花算法求解多维背包问题[J]. 系统工程与电子技术,2017,39(2):451-458.

[15] DENG W, SHANG S, CAI X, et al. An improved differential evolution algorithm and its application in optimization problem[J]. Soft Computing, 2021, 25(7):5277-5298.

[16] 杨艳,刘生建,周永权. 贪心二进制狮群优化算法求解多维背包问题[J]. 计算机应用,2020,40(5):1291-1294.

[17] 王志刚,郝志峰,黄翰. 混合粒子群算法求解多维背包问题[J]. 哈尔滨商业大学学报(自然科学版),2008(2):250-253.