

MIF-CNNIF:一种基于CNN的交叉特征的多分类图像数据框架

王盼红, 朱昌明

引用本文

王盼红, 朱昌明. MIF-CNNIF:一种基于CNN的交叉特征的多分类图像数据框架[J]. 计算机科学, 2022, 49(11A): 210800267-8.

WANG Pan-hong, ZHU Chang-ming. MIF-CNNIF:A Multi-classification Image Data Framework Based on CNN with Intersect Features [J]. Computer Science, 2022, 49(11A): 210800267-8.

相似文章推荐 (请使用火狐或 IE 浏览器查看文章)

Similar articles recommended (Please use Firefox or IE to view the article)

[基于机器学习的剩余使用寿命预测实证研究](#)

Empirical Research on Remaining Useful Life Prediction Based on Machine Learning
计算机科学, 2022, 49(11A): 211100285-9. <https://doi.org/10.11896/jsjcx.211100285>

[一种改进的特征选择算法在邮件过滤中的应用](#)

Application of Improved Feature Selection Algorithm in Spam Filtering
计算机科学, 2022, 49(11A): 211000028-5. <https://doi.org/10.11896/jsjcx.211000028>

[基于启发式搜索特征选择的加密流量恶意行为检测技术](#)

Detection of Malicious Behavior in Encrypted Traffic Based on Heuristic Search Feature Selection
计算机科学, 2022, 49(11A): 210800237-6. <https://doi.org/10.11896/jsjcx.210800237>

[R-YOLOv5:自动切割的旋转的文本检测模型](#)

R-YOLOv5:Auto-cutting,Rotated Text Detection Model
计算机科学, 2022, 49(11A): 210900185-6. <https://doi.org/10.11896/jsjcx.210900185>

[基于注意力机制的手写体数字识别](#)

Handwritten Digit Recognition Based on Attention Mechanism
计算机科学, 2022, 49(11A): 211100009-5. <https://doi.org/10.11896/jsjcx.211100009>

MIF-CNNIF:一种基于 CNN 的交叉特征的多分类图像数据框架

王盼红 朱昌明

上海海事大学信息工程学院 上海 201306

摘要 近年来,图像多分类任务和深度学习受到越来越多学者的重视,基于卷积神经网络(Convolutional Neural Network, CNN)的多分类图像数据框架也得到了广泛应用。传统的基于卷积神经网络的多分类图像数据学习(MIF-CNN)普遍存在图像处理复杂、特征维数大、时间复杂度高问题。针对这一问题,提出了一种基于 CNN 的交叉特征的多分类图像数据框架(MIF-CNNIF)。MIF-CNNIF 是一种基于多种特征选择算法得到相交特征并以此交叉特征代替原特征集处理图像多分类任务的框架。在 10 个多类图像数据集上进行了丰富的对比实验,结果验证了 MIF-CNNIF 的有效性。MIF-CNNIF 的贡献在于:1)使用预先训练好的 CNN 模型,避免了设置过多参数;2)与 MIF-CNN 相比,有效降低了特征维度和时间复杂度;3)具有比 MIF-CNN 更好的平均分类准确率;4)在多分类图像数据集上成功验证了组合特征算法的有效性。

关键词: 卷积神经网络,特征选择,交叉特征,图像多分类,组合特征

中图法分类号 TP389

MIF-CNNIF: A Multi-classification Image Data Framework Based on CNN with Intersect Features

WANG Pan-hong and ZHU Chang-ming

College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China

Abstract In recent years, image multi-classification task and deep learning have received increasingly attentions, and multi-classification image data framework based on convolutional neural network(MIF-CNN) has also been widely used. Traditional CNN-based multi-class image data learning generally has a problem that the image processing is complicated, the feature dimensions are large, and the time complexity is high. To solve this problem, this paper proposes a multi-classification image data framework based on CNN with intersect features(MIF-CNNIF). MIF-CNNIF is a framework for performing multi-classification tasks based on intersect features obtained by multiple feature selection algorithms. Through extensive comparative experiments on 10 multi-class image data sets, the results validate the effectiveness of MIF-CNNIF. The contributions of MIF-CNNIF are that, 1) it avoids the problem of setting too many parameters with the usage of pre-trained CNN models; 2) it keeps features dimension and time cost after comparing with MIF-CNN; 3) it has a better average recognition accuracy than MIF-CNN; 4) the effectiveness of combined feature algorithms is verified on multi-class image data sets.

Keywords Convolutional neural network, Feature selection, Intersect features, Image multi-class, Combine feature

1 引言

近年来,图像多分类问题已经渗透到各个领域,包括计算机视觉、机器学习和模式识别,受到了越来越多学者的关注,多分类图像的研究具有重要意义^[1]。

特征选择是机器学习和数据挖掘中的一个重要课题,在许多文献中都有广泛的研究^[2-5]。在分类任务中,特征选择是一个从原数据特征中选择含有大部分原数据集信息的特征子集,并用该子集替换原数据特征的过程。传统的特征选择主要分为 3 种类型,即过滤法(Filter)、包裹法(Wrapper)和嵌入法(Embedding)。过滤法^[6-8]通过应用评估指标(如距离、信

息以及单个特征和输出类别标签之间的相关性)来选择重要特征,该方法独立于任何分类器,不涉及任何学习算法。包裹法^[9]依靠预定的学习算法来确定特征子集,一般来说,包裹法比过滤法具有更好的性能,但其计算成本会更高。嵌入法^[10]旨在将特征选择集成到模型训练中,通常其比包裹法更快,并可以为学习算法提供合适的特征子集。

很多基于 CNN 的特征选择方法通过选择和保留原始特征子集并将其作为输入来降低复杂性,同时去除了不相关和冗余的特征^[11]。尽管该方式在一定程度上提高了分类性能,但特征维度仍然很大,计算代价仍旧很高。

本文提出了一种基于 CNN 的交叉特征的多分类图像数据

基金项目:国家自然科学基金(62276164,61602296);上海市“科技创新行动计划”自然科学基金项目(22ZR1427000);晨光计划(18CG54);中国博士后科学基金(2019M651576)

This work was supported by the National Natural Science Foundation of China(62276164,61602296),“Science and Technology Innovation Action Plan” Natural Science Foundation of Shanghai(22ZR1427000),Chenguang Program(18CG54) and China Postdoctoral Science Foundation(2019M651576).

通信作者:朱昌明(cmzhu@shmtu.edu.cn)

框架(MIF-CNNIF)。该方法有4个主要步骤:1)基于4个卷积神经网络,即Vgg16, Vgg19, Alexnet和Resnet50^[12-14],进行多分类图像数据的特征提取;2)基于上一步神经网络提取出的特征,使用支持向量机(Support Vector Machine, SVM)作为分类器开展分类实验;3)基于5种特征选择方法,即Recursive Feature Elimination(RFE)^[15], Gradient Boosted Decision Tree (GBDT)^[16-19], Random Forest (RF)^[20], linearSVC^[21]和Chi-squared test(Chi2)^[22],对所提取特征分别进行特征选择实验以及组合特征选择算法实验;4)基于5种特征选择方法选择出的特征,得到同一数据集的卷积神经网络的相关特征,再使用支持向量机作为分类器执行多分类任务。

MIF-CNNIF方法的优点如下:

(1)提出了一种将CNN模型和特征选择相结合的混合方法。

(2)使用一些预训练好的网络模型,避免了设置过多的参数。

(3)有效降低了特征维度和时间代价。

(4)使用多种度量并进行丰富的对比实验,验证了MIF-CNNIF相比MIF-CNN在多类分类任务中的有效性。

(5)成功验证了LinearSVC和RFE的组合特征选择方法可以有效提高多分类任务的分类性能。

本文第2节描述了相关的工作;第3节详细阐述了提出的MIF-CNNIF方法;第4节给出了实验分析和结果;最后总结全文并展望未来。

2 相关工作

目前,许多计算机技术已经被应用于图像多分类的研究。Dias等^[23]提出了一种新型基于CNN的苹果花检测算法,该算法利用计算机视觉技术有效结合颜色和形态信息来识别花卉种类,最终召回率和准确率接近80%;Ghazi等^[24]采用3种预训练好的神经网络模型,即Goog-LeNet, AlexNet和VGG-Net,结合迁移学习进行植物物种分类,最优分类准确率为80%;Sardogan等^[25]提出了一种基于CNN模型和学习矢量量化算法进行番茄叶片检测和分类;Pawara等^[26]提出基于深度学习框架进行数据增强,有效提高了植物分类的准确性;Nagasubramanian等^[27]提出了一种用于大豆炭腐病检测和感染分类的三维卷积神经网络模型,最终分类准确率达到95.73%;Gayathri等^[28]提出了一种新的CNN模型,用于提取视网膜眼底图像的特征,并使用J48分类器对其进行分类。二分类的平均准确率为99.89%,多分类的平均准确率为99.59%。

在上述大多数研究中,图像多分类问题都与神经网络相结合,但没有考虑使用交叉特征进行分类实验,导致特征维度较大,计算成本较高。本文提出了一种基于CNN的交叉特征的多分类图像数据框架。在该方法中,CNN主要用于特征提取,然后利用5种特征选择算法进行特征选择,从而找到多分类任务的交叉特征。本文在10个多分类图像数据集上进行了丰富的对比实验,验证了交叉特征与SVM分类器结合能有效降低特征维度;同时发现,将linearSVC和RFE组合的特征选择算法在一定程度上能提高分类性能。

3 MIF-CNNIF

本文提出了一种基于CNN的交叉特征的多分类图像

数据框架。实验前对数据集进行预处理操作,包括统一图像大小和格式。由于考虑到数据类别不平衡会对实验结果产生影响,因此实验中使用的10个多分类图像数据集都具有相对平衡的类别。

假设最初由CNN提取出的原始特征集合为 $X_o = \{x_{o1}, x_{o2}, x_{o3}, \dots\}$,将每个特征命名为 $N_o = \{n_{o1}, n_{o2}, n_{o3}, \dots\}$ 。原始特征集 X_o 在经过不同的特征选择算法过程后得到 $X_i = \{x_{i1}, x_{i2}, x_{i3}, \dots\}$,其中 i 代表5种不同的特征选择算法,即 $i \in \{RFE, GBDT, RF, LinearSVC, Chi2\}$,用数字分别代替5种特征选择算法,即有 $i \in \{1, 2, 3, 4, 5\}$,对应特征名称分别为 $N_i = \{n_{i1}, n_{i2}, n_{i3}, \dots\}$,则对应得到的交叉特征名称为 $N_e = \{n_{e1}, n_{e2}, n_{e3}, \dots\} = N_1 \cap N_2 \cap N_3 \cap N_4 \cap N_5$,其对应的交叉特征为 $X_e = \{x_{e1}, x_{e2}, x_{e3}, \dots\} = X_1 \cap X_2 \cap X_3 \cap X_4 \cap X_5$ 。

MIF-CNNIF的结构如图1所示,共分为4个模块。

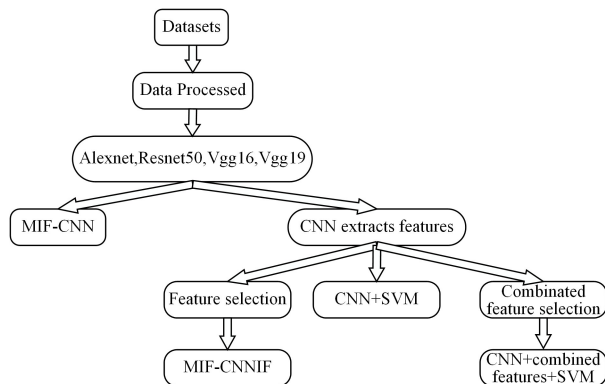


图1 MIF-CNNIF模型的结构

Fig. 1 Structure of MIF-CNNIF model

首先,4种CNN进行特征提取,然后基于Softmax分别进行图像多分类实验。

其次,基于第一步提取到的特征,使用SVM作为分类器进行多分类实验,并得到相应的实验结果。

然后,基于第一步所得特征进行特征选择。5种特征选择方法为RFE, GBDT, RF, LinearSVC和Chi2。该过程主要分为两个步骤:1)每种特征选择方法都独立进行实验,5种特征选择方法基于第一步提取的特征会分别得到一个新的特征子集,然后使用SVM作为分类器进行多分类任务;2)对5种特征选择算法进行随机组合实验,共有26种不同的组合,如RFE和GBDT的组合实验将两种特征选择算法选择的特征组合成一个新的特征集,并使用SVM作为分类器进行实验。

最后,基于上一步骤中的步骤1)获得的特征,得到5种特征选择算法的交叉特征集合,并用其代替原特征集,最后用SVM作为分类器进行实验。

4 实验

本节首先阐述关于CNN模型以及相关算法的实验设置。实验中使用的CNN是Alexnet, Resnet50, Vgg16和Vgg19。其次,为了验证MIF-CNNIF的有效性,实验在10种多分类图像数据集上执行,实验结果包括分类性能对比、训练时间对比和特征数对比。最后,本文进一步讨论了显著性分析的实验,包括配对t-test^[29]和Friedman-Nemenyi statistical test^[30]。当然,本文也给出了关于组合特征算法的实验结果。

4.1 实验设置

(1)数据集。实验中使用的数据都来源于 Kaggle 网站。为了验证所提方法的有效性,选择 10 个不同领域的多分类图像数据集,如动物、植物、天气和人体姿态等。表 1 列出了所用数据集的具体描述。

表 1 所用数据集的描述

Table 1 Description of the used data sets

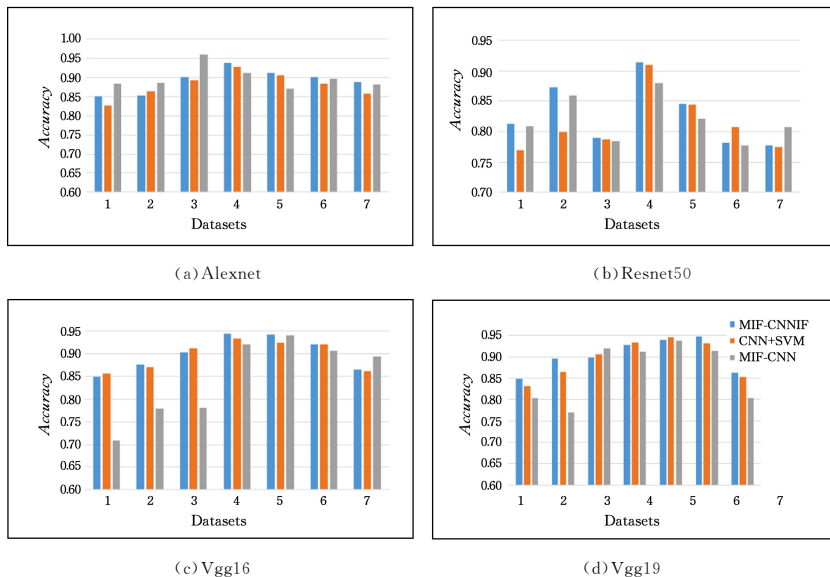
Data sets	No. class	No. samples
5_flowers	5	4323
chess	6	553
crop	5	804
dance	8	599
oxford	17	1360
pokemon	5	1000
posture	4	4800
weather	4	1125
wildlife	4	1504
yoga	5	1080

(2)参数设置。实验中采用的 4 种 CNN 模型结构都没有改变,其中 Vgg16, Vgg19 和 Resnet50 这 3 种网络使用预训练好的模型。CNN 主要用于特征提取器, SVM 作为分类器。Alexnet 的图像输入大小为 227×227 , 其余 3 种网络输入大小均为 224×224 。除此以外,为了进一步验证 MIF-CNNIF 的有效性,实验中关于特征选择过程中提取的特征数基本保持在 1000 左右。所有的实验都在 Pycharm 2019 上开展。研究中使用的硬件安装的操作系统是 64 位 Windows 10。该电脑配备了 NVIDIA GeForce RTX 2060 6 GB 显卡、英特尔 i7-Core@2.6 GHz 处理器和 16 GB 内存。

4.2 分类性能比较

为了证明所提 MIF-CNNIF 方法的有效性及其先进性,本文在 10 种多分类图像数据集上进行了丰富的对比实验,使用多种度量指标进行评价,包括准确率(accuracy)、精确度(precision)和 AUC、训练时间、特征数及显著性分析。整个实验分为 4 个过程:首先,基于 4 个神经网络提取特征并进行多分类任务;其次,基于提取的特征,使用 SVM 作为分类器进行多分类实验;然后,进行特征选择算法的随机组合实验;最后,在 SVM 分类器的基础上,用得到的交叉特征代替原始特征集进行相关多分类实验。实验指标有准确率、精确度和 AUC。为便于阐述,实验部分选择展示在 7 个数据集上的实验结果,其他数据集在相应算法上的实验也是相似的。数据集包括 dance, oxford, pokemon, posture, weather, wildlife 和 yoga。在以下的实验结果描述中,“CNN+SVM”表示相关算法基于 MIF-CNN,以 SVM 为分类器且不涉及特征选择过程。

图 2 给出了 3 种不同算法在 7 种数据集上的准确率结果,包括 MIF-CNNIF, CNN+SVM 和 MIF-CNN。根据图 2 可以得到:1)在大多数情况下, CNN+SVM 算法的准确率优于 MIF-CNN;2)从平均情况来看, Resnet50 在 4 个模型中的实验结果相对会差一些,这可能是由于其网络结构层次相对较深;3)在大多数据集上, MIF-CNNIF 相对另外两种算法在分类准确率上具有一定的优势,尤其是与 MIF-CNN 对比时,这种优势更明显。为了进一步验证所提 MIF-CNNIF 方法的有效性,我们还基于精确度和 AUC 进行了大量丰富的对比实验,结果如图 3 和图 4 所示。



注:横坐标 1-7 表示这 7 个数据集,即 1-dance, 2-oxford, 3-pokemon, 4-posture, 5-weather, 6-wildlife 和 7-yoga

图 2 Alexnet, Resnet50, Vgg16 和 Vgg19 在 7 个数据集上的 accuracy(电子版为彩图)

Fig. 2 Accuracy of Alexnet, Resnet50, Vgg16 and Vgg19 on seven data sets

如图 3 所示,可以明显得到:1)基于 4 种 CNN 模型,相比 CNN+SVM, MIF-CNNIF 在精确度上具有明显的优势;2)所提方法在某些数据集上分类效果会更好,包括 pokemon, posture, weather 和 wildlife;3)在同样的实验条件下, Resnet50 网络的实验结果相比其他网络会表现得差一点,这可能是

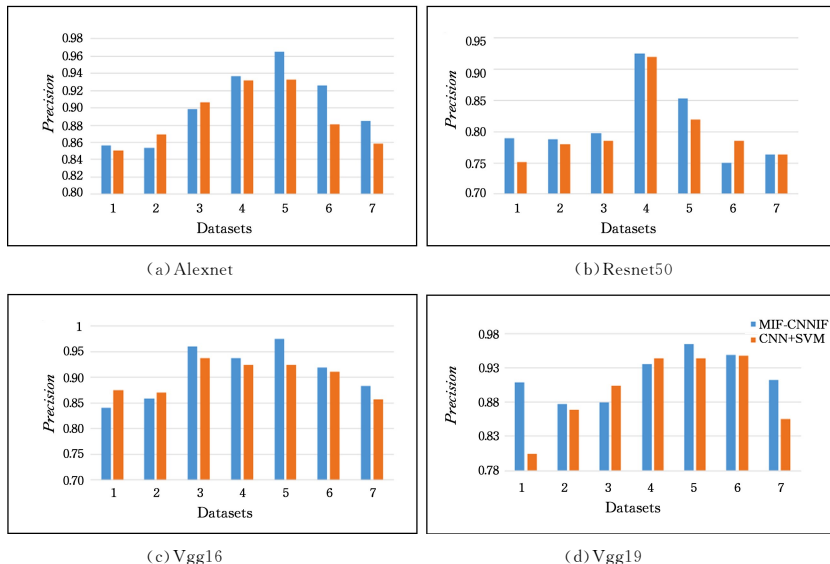
由于其本身的网络深度过大,导致在同等条件下得到的网络模型不是最优的。

如图 4 所示,比较明显的是, MIF-CNNIF 在大多情况下的 AUC 都会比 CNN+SVM 更高。AUC 的本质是比较数据集中的正样本和负样本,然后评估正样本预测的可信度。

根据图 4 结果可以认为, MIF-CNNIF 能有效提高图像多分类任务的性能。

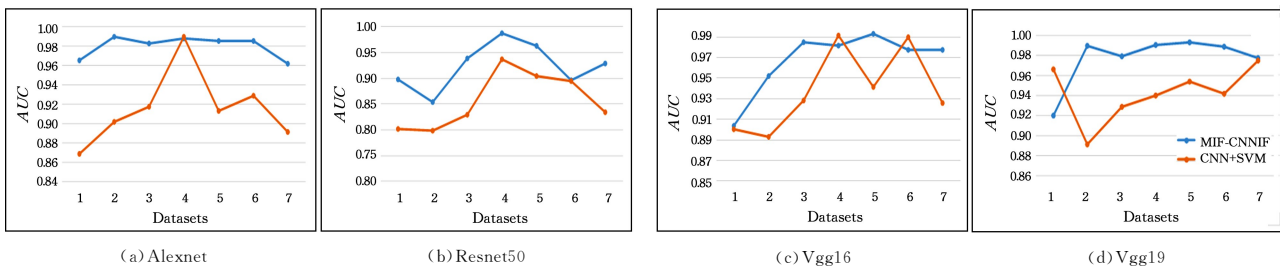
在实验过程中也发现, 5 种特征选择方法的一些随机组合也能在图像多分类任务中得到更好的识别率。由于这种随机组合共有 26 种不同的形式, 在后续的实验描述中, 为了便于描述, 会使用缩写来表示组合算法, 如表 2 所列。我们选择相同数量的平均性能最佳的组合算法进行展示。为了进一步

证明所提方法的先进性, 表 3 列出了基于 4 种 CNN 模型和 10 个多分类数据集的组合算法的平均性能结果。表 3 中, “Time”指使用对应组合特征算法得到的混合特征集进行多分类任务所需的时间。表 3 中的所有数据都是基于 10 种数据集得到的平均值。从表 3 中可以明显得到, linearSVC 与 RFE 的组合特征选择算法在任一种 CNN 上都能起到较好的分类效果。



注: 横坐标 1-7 表示 7 个数据集, 即 1-dance, 2-oxford, 3-pokemon, 4-posture, 5-weather, 6-wildlife and 7-yoga

图 3 Alexnet, Resnet50, Vgg16 和 Vgg19 在 7 个数据集上的 precision
Fig. 3 Precision of Alexnet, Resnet50, Vgg16 and Vgg19 on seven data sets



注: 横坐标 1-7 表示这 7 个数据集, 即 1-dance, 2-oxford, 3-pokemon, 4-posture, 5-weather, 6-wildlife and 7-yoga

图 4 Alexnet, Resnet50, Vgg16 和 Vgg19 在 7 个数据集上的 AUC
Fig. 4 AUC of Alexnet, Resnet50, Vgg16 and Vgg19 on seven data sets

表 2 随机组合特征选择算法的缩写

Table 2 Abbreviations of randomly combined feature selection algorithms

Combined feature selection algorithm	Abbreviation	Combined feature selection algorithm	Abbreviation
Chi2+GBDT	C_G	GBDT+linearSVC+RFE	G_L_F
Chi2+linearSVC	C_L	RF+ Chi2+linearSVC	R_C_L
Chi2+RFE	C_F	RF+ Chi2+GBDT	R_C_G
GBDT+linearSVC	G_L	RF+ Chi2+RFE	R_C_F
GBDT+RFE	G_F	RF+GBDT+linearSVC	R_G_L
linearSVC+RFE	L_F	RF+GBDT+RFE	F_G_F
RF+ Chi2	R_C	RF+linearSVC+RFE	R_L_F
RF+GBDT	F_G	Chi2+GBDT+linearSVC+RFE	C_G_L_F
RF+linearSVC	R_L	RF+ Chi2+GBDT+linearSVC	R_C_G_L
RF+RFE	R_F	RF+ Chi2+GBDT+RFE	R_C_G_F
Chi2+GBDT+linearSVC	C_G_L	RF+Chi2+linearSVC+RFE	R_C_L_F
Chi2+GBDT+RFE	C_G_F	RF+GBDT+linearSVC+RFE	R_G_L_F
Chi2+linearSVC+RFE	C_L_F	RF+Chi2+GBDT+linearSVC+RFE	R_C_G_L_F

表 3 组合特征选择算法的平均性能

Table 3 Average performance of combine feature selection algorithm

CNN	Algorithms	Performance Metrics				
		accuracy	precision	AUC	Feature numbers	Time/s
Alexnet	L_F	0.9089	0.9059	0.9830	1846	61.4032
	C_L_F	0.9032	0.9021	0.9746	2776	91.8388
	C_G_L_F	0.9025	0.9023	0.9713	3866	149.535
	R_C_G_L_F	0.9045	0.9055	0.9837	4632	170.8631
Resnet50	R_L	0.9048	0.9040	0.9781	1774	98.4518
	R_L_F	0.8151	0.8050	0.9248	2682	142.7699
	C_G_L_F	0.8138	0.8129	0.9171	3686	188.7625
	R_C_G_L_F	0.8183	0.8007	0.9103	4478	259.5693
Vgg16	L_F	0.9102	0.9084	0.9843	1852	61.4225
	C_L_F	0.9044	0.9010	0.9878	2782	91.7610
	C_G_L_F	0.9046	0.8986	0.9789	3761	144.9268
	R_C_G_L_F	0.9038	0.9010	0.9766	4661	171.007
Vgg19	L_F	0.9114	0.9062	0.9792	1846	62.5855
	C_L_F	0.9060	0.9033	0.9826	2776	93.4730
	C_G_L_F	0.9087	0.9047	0.9809	3782	151.5153
	R_C_G_L_F	0.9054	0.9019	0.9749	4625	174.3638

注:每种算法的最佳结果以粗体显示

4.3 时间性能比较

为了验证所提出的 MIF-CNNIF 的有效性,表 4 列出了 7 种不同的算法在 7 个数据集上的平均训练时间。为便于描述,我们选择 Alexnet 和 Resnet50 两个网络的实验结果进行展示,并使用一些缩写来代表相应的算法,其余网络在该条件下的实验结果也类似。例如,CNN+Chi2+SVM 在特征提取后使用 Chi2 进行特征选择得到新的特征子集,再使用 SVM 作为分类器进行多分类实验;CNN+GBDT+SVM 在特征提取后使用 GBDT 进行特征选择得到新的特征子集,再使用 SVM 作为分类器进行多分类实验。以此类推,则 CNN+linearSVC+SVM, CNN+RF+SVM 和 CNN+RFE+SVM

分别表示使用 linearSVC,RF 和 RFE 进行特征选择得到的新的特征子集。7 种不同的算法包括 MIF-CNNIF, CNN+SVM, CNN+Chi2+SVM, CNN+GBDT+SVM, CNN+linearSVC+SVM, CNN+RF+SVM 和 CNN+RFE+SVM。其他几个数据集在相关算法上的结果是相似的。从表 4 可以发现:1) 基于同一个数据集,MIF-CNNIF 相对其他算法的时间消耗最小;2) Alexnet 和 Resnet50 两种网络在 posture 数据集上的效果最差,这可能是由于数据集本身的规模相对较大,一般来说,数据规模越大,提取的特征就越多,从而消耗更多的训练时间;3) 相对 Alexnet 网络,Resnet50 网络在 7 种数据集上的时间消耗会更长,这可能是受其网络结构和深度所影响。

表 4 基于 Alexnet 和 Resnet50 的 7 个数据集的训练时间对比

Table 4 Comparison of training time of 7 data sets based on Alexnet and Resnet50

Data sets	M-C	C+S	C+C+S	C+G+S	C+L+S	C+RF+S	C+R+S
	M-C	C+S	C+C+S	C+G+S	C+L+S	C+RF+S	C+R+S
dance	3.2817	43.8335	6.1599	12.1505	7.8239	5.7537	6.4176
	14.9880	29.3446	7.2039	11.3843	7.9503	12.8951	11.5512
oxford	8.0993	133.3116	18.6520	42.1248	28.4784	17.3698	28.0224
	23.5158	61.2069	21.8871	30.0349	29.4926	16.8153	43.2425
pokemon	0.6523	31.4284	6.6784	7.6015	9.1943	5.8313	7.5906
	6.7409	8.8321	11.0266	10.2613	14.2591	10.2379	11.1454
posture	44.5556	661.1273	103.4660	131.9631	133.5356	89.7665	110.7720
	284.3855	86.3123	110.4063	163.0517	146.6502	110.8028	122.7991
weather	2.8165	43.5449	5.8463	7.6807	4.6057	4.6057	6.0465
	24.3640	23.7367	11.9916	15.8144	9.0458	9.3897	13.2847
wildlife	6.2275	89.3432	12.5243	13.3092	14.4300	10.0484	12.6536
	54.3364	21.8656	25.0841	36.9530	32.3535	22.8997	33.9108
yoga	6.1374	77.6431	12.8949	20.2969	13.0257	9.7760	11.4887
	29.4424	26.6646	14.4115	13.7587	17.4408	14.0252	14.9748

注:每一行中上半部分表示 Alexnet 的结果,下半部分表示 Resnet50 的结果。对比算法有 MIF-CNNIF, CNN+SVM, CNN+Chi2+SVM, CNN+GBDT+SVM, CNN+linearSVC+SVM, CNN+RF+SVM 和 CNN+RFE+SVM, 分别用缩写 M-C, C+S, C+C+S, C+G+S, C+L+S, C+RF+S 和 C+R+S 表示

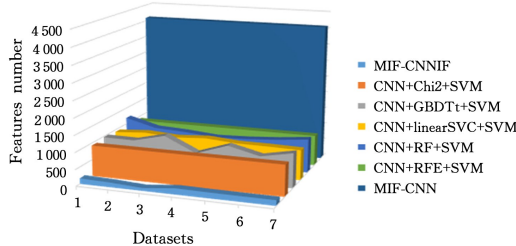
4.4 特征数比较

从 4 种 CNN 模型的结构来看,其输出特征的维度会很高。特征选择的目的是寻找一个最优的特征子集来

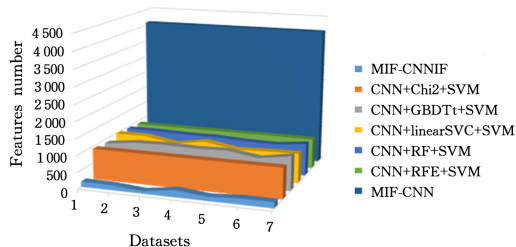
代替原始特征集,从而在不丢失原始数据信息的情况下降低维度。为描述方便,图 5 给出了 7 个数据集在 Vgg16 和 Vgg19 上的实验结果,其他数据集的实验结果在同样情况

下是相似的。

从图 5 的结果来看,比较明显的是:1)特征选择的确可以降低特征维度,相比 MIF-CNN, MIF-CNNIF 中的交叉特征维度得到了很大的降低;2)根据前面的分析, MIF-CNNIF 在有效降低维度的同时也保证了分类性能在一定程度上的提高;3)根据在表 3 中的特征数目,相比组合特征算法, MIF-CNNIF 的特征维度在一定范围内也有明显下降,从而进一步验证了该方法的有效性。



(a) Vgg16



(b) Vgg19

注:横坐标表示这 7 个数据集,即 1-dance,2-oxford,3-pokemon,4-posture,5-weather,6-wildlife and 7-yoga

图 5 基于 Vgg16 和 Vgg19 的 7 个数据集的特征数

Fig. 5 Features number of seven data sets based on Vgg16 and Vgg19

4.5 显著性分析

通过显著性分析实验,进一步有效验证了 MIF-CNNIF 方法对多分类图像任务的贡献。显著性分析主要包括配对 t-test^[29]和 Friedman-Nemenyi statistical test^[30]。就配对 t-test^[29]而言,主要用于分析在一个数据集上两种不同算法之间的差异是否显著。就 Friedman-Nemenyi statistical test^[30]而言, Friedman 检验主要用于分析所有算法在多个数据集上是否存在显著差异,而 Nemenyi 检验主要用于分析两种算法在多个数据集上的差异是否显著^[31]。

(1) 配对 t-test。配对 t-test^[29]主要用于分析两种不同算法在同一个数据集上的差异是否显著。其主要目的是判断在检验条件 1 和条件 2 情况下的平均识别率是否有显著差异。通常情况下,使用 sig 值来表示测试集的显著性差异结果。一般来说, sig 值的阈值设置为 0.05,当大于 0.05 时,表示两种算法在同一个数据集没有显著性差异,反之则代表有显著性差异^[31]。sig 值越小,表示算法之间的差异越显著。为描述方便,选择 Alexnet 和 Vgg19 两种 CNN 模型来展示相应的结果。图 6 给出了 CNN+SVM 和 MIF-CNNIF 两种算法基于 Alexnet 和 Vgg19 模型在 10 个数据集上的配对 t-test 结果。从图 6 可以得到:1) 基于 Alexnet 和 Vgg19,两种算法在大多数数据集上的差异都很

显著;2) MIF-CNNIF 相对 CNN+SVM 来说,其差异更为显著。因为对于 MIF-CNNIF 和另一种算法来说,其 sig 值在大多数数据集上都小于 0.05。因此,基于配对 t-test 成功证明了所提的 MIF-CNNIF 方法是有效的。

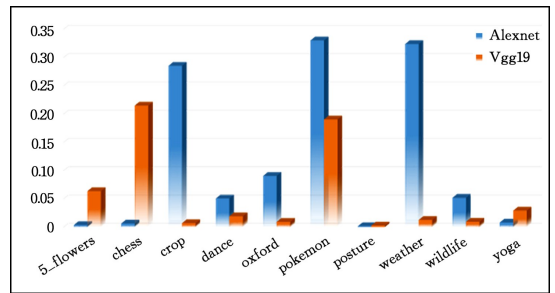


图 6 CNN+SVM 和 MIF-CNNIF 两种算法基于 Alexnet 和 Vgg19 模型在所有数据集上的配对 t-test 结果
Fig. 6 Paired t-test of CNN+SVM and MIF-CNNIF based on Alexnet and Vgg19 on all date sets

(2) Friedman-Nemenyi statistical test。此外,使用 Friedman-Nemenyi statistical test 来分析所有对比算法在多个数据集上的差异是否显著。为了更好地说明相关实验结果,选择基于 Resnet50 网络进行展示,另外 3 种 CNN 的实验结果也类似。对比算法包括 MIF-CNNIF, Resnet50+SVM, Resnet50+Chi2+SVM, Resnet50+GBDT+SVM, Resnet50+linearSVC+SVM, Resnet50+RF+SVM 和 Resnet50+RFE+SVM。Friedman 检验会基于每个数据集根据算法性能单独对算法进行排序,表现最好的算法排名为 1,表现次好的排名为 2,以此类推,如表 5 所列。此外,还会给出其平均排名和平均准确率。定义 r_i^j 为在 N 个数据集上的第 i 种数据集上, k 种算法中的第 j 种算法的排名。而对于第 j 种算法而言, Friedman 检验通过平均排名进行对比,即 $R_j = (\frac{1}{N}) \sum_i r_i^j$, Friedman 统计遵循自由度为 $k-1$ 的 χ_F^2 分布,当 N 和 k 足够大时。实际上,下面给出的另一种统计分布的效果更好^[32]:

$$\chi_F^2 = \frac{12N}{k(k+1)} \left[\sum_j R_j^2 - \frac{k(k+1)^2}{4} \right] \quad (1)$$

$$F_F = \frac{(N-1)\chi_F^2}{N(k-1) - \chi_F^2} \quad (2)$$

这种统计遵循自由度为 $k-1$ 和 $(k-1)(N-1)$ 的 F 分布。其临界值表可以在任何一本统计书中找到。在 Friedman 检验中,若 F_F 的值大于临界值 $F_{\alpha}(K-1, (K-1)(N-1))$ 时拒绝零假设,即在多个数据集上所有对比算法之间的差异是显著的,其中 α 是置信水平。而当拒绝零假设后,则需要进行另一种统计检验,如 Nemenyi 检验。当所有算法相互进行对比时,可使用其检验。如果两种算法对应的平均排名之间至少存在临界值差异,则两种算法在所有数据集上的性能都具有显著性差异。

$$CD = q_{\alpha} \sqrt{\frac{k(k+1)}{6N}} \quad (3)$$

其中,临界值 q_{α} 是通过研究范围统计除以 $\sqrt{2}$ 获得(见表 6)。

表 5 在 Resnet50 网络的基础上,不同算法在 10 个数据集上的分类准确率和对应排名

Table 5 Classification accuracy and corresponding ranking of different algorithms on ten data sets

Data sets	M+I+F rank	R+S rank	R+C+S rank	R+G+S rank	R+L+S rank	R+RF+S rank	R+R+S rank
5_flowers	0.8122	0.7705	0.7647	0.7504	0.7451	0.7821	0.7358
	1	3	4	5	6	2	7
chess	0.7889	0.7821	0.7486	0.7641	0.7694	0.7847	0.7956
	2	4	7	6	5	3	1
crop	0.8581	0.8537	0.8571	0.8385	0.8532	0.8224	0.8447
	1	3	2	6	4	7	5
dance	0.8131	0.7700	0.7967	0.7438	0.7222	0.7269	0.7500
	1	3	2	5	7	6	4
oxford	0.8735	0.7991	0.7769	0.7885	0.7582	0.7895	0.7651
	1	2	5	4	7	3	6
pokemon	0.7900	0.7875	0.7800	0.8323	0.7800	0.7810	0.7490
	2	3	6	1	5	4	7
posture	0.9144	0.9104	0.9142	0.9262	0.9070	0.9141	0.9135
	2	6	3	1	7	4	5
weather	0.8458	0.8444	0.8449	0.8444	0.8553	0.8427	0.8436
	1	5	3	4	2	7	6
wildlife	0.7815	0.8077	0.7810	0.7757	0.7811	0.7807	0.7805
	2	1	4	7	3	5	6
yoga	0.7778	0.7750	0.7730	0.7759	0.7721	0.7759	0.7731
	1	4	6	2	7	3	5
Average	0.8255	0.8100	0.8037	0.8040	0.7934	0.8000	0.7951
	1.4	3.4	4.2	4.1	5.3	4.4	5.2

注:各种算法包括 MIF-CNNIF, Resnet50+SVM, Resnet50+Chi2+SVM, Resnet50+GBDT+SVM, resnet 50+linesvc+SVM, Resnet50+RF+SVM 和 Resnet50+RFE+SVM, 分别用缩写 M+I+F, R+S, R+C+S, R+G+S, R+L+S, R+RF+S 和 R+R+S 表示, 每个模型的最佳结果用粗体标注

表 6 双尾 Nemenyi 检验标准值

Table 6 Critical values for two-tailed Nemenyi test

No. algorithms	2	3	4	5	6	7	8	9	10
$q_{0.05}$	1.960	2.343	2.569	2.728	2.850	2.949	3.031	3.102	3.164
$q_{0.10}$	1.645	2.052	2.291	2.459	2.589	2.693	2.780	2.855	2.920

根据表 5 中的实验结果, 由于实验采用 7 种对比算法和 10 个数据集, 因此使用了式(2)和式(3)进行 Friedman 检验。即 $\chi_F^2 = 22.41$ 和 $F_F = 5.37$, 此外, F_F 遵循自由度为 $7-1=6$ 和 $(7-1) \times (10-1) = 54$ 的 F 分布。临界值 $F_\alpha(6, 54)$ 的值如下:

$$F_\alpha(6, 54) = \begin{cases} 2.28, & \alpha=0.05 \\ 1.89, & \alpha=0.10 \end{cases} \quad (4)$$

当 $F_F > 2.28$ 并且 $F_F > 1.89$ 时, 则会拒绝零假设。然后再进行 Nemenyi 检验。当 $\alpha=0.05$ 时, 临界值 $q_{0.05}$ 是 2.949 (见表 6); 当 $\alpha=0.10$ 时, 临界值 $q_{0.10}$ 为 2.693 (见表 6)。对应的 CD 值如下:

$$CD = \begin{cases} 2.85, & \alpha=0.05 \\ 2.60, & \alpha=0.10 \end{cases} \quad (5)$$

当 $CD=2.85$ 时, 由于 $1.4+2.85=4.25 > 3.4, 4.25 > 4.2$ 并且 $4.25 > 4.1$, 因此 MIF-CNNIF 相比 Resnet50+SVM, Resnet50+Chi2+SVM 和 Resnet50+GBDT+SVM, 在平均排名上的差异性并不显著。但是, 相比剩下的几种算法, 这种差异性比较显著, 包括 Resnet50+linearSVC+SVM, Resnet50+RF+SVM 和 Resnet50+RFE+SVM。

当 $CD=2.60$ 时, 由于 $1.4+2.60=4.0 > 3.4$, 因此可以认为 MIF-CNNIF 相比 Resnet50+SVM 在平均排名上无显著性差异。但是相比另外几种算法, 这种差异就很显著, 包括 Resnet50+Chi2+SVM, Resnet50+GBDT+SVM, Resnet50+linearSVC+SVM, Resnet50+RF+SVM 和 Resnet50+RFE+SVM。

结束语

本文为解决传统的 MIF-CNN 方法存在的不足, 提出了一种使用交叉特征的新方法。因此, 一种基于 CNN 的交叉特征的多分类图像数据框架被提出, 即 MIF-CNNIF。通过在多个多分类图像数据集上进行大量对比实验, 证明所提方法的优点有: 1) 通过使用预训练好的 CNN 模型, MIF-CNNIF 能有效地避免设置过多的网络参数; 2) 在 MIF-CNNIF 中使用交叉特征代替原始特征, 与 MIF-CNN 相比, MIF-CNNIF 有效降低了特征维度; 3) 总体来说, MIF-CNNIF 相比其他基于 CNN 的对比算法, 在有效提高多分类的分类性能的同时减小了时间复杂度, 对比算法包括 MIF-CNN, CNN+Chi2+SVM, CNN+RF+SVM 和 CNN+GBDT+SVM; 4) 通过进行显著性分析实验, 证明 MIF-CNNIF 相对其他算法具有显著性差异; 5) linearSVC 和 RFE 的组合特征选择方法也能有效提高多分类图像的分类性能。

尽管所提方法已经在单视角数据集上被证实是有效的, 但是仍然有一个问题需要在未来的工作中解决。即多视角数据集在现实世界中被广泛使用, 因此在未来的工作中, 应考虑基于多视角数据集来验证所提出的 MIF-CNNIF 的有效性。

参考文献

- [1] GUO T, DONG J, LI H, et al. Simple convolutional neural network on image classification[C]// 2017 IEEE 2nd International Conference on Big Data Analysis(ICBDA). 2017.
- [2] ZHENG H. Online Feature Selection Based on Passive-Aggressive Algorithm with Retaining Features[C]// Asia-pacific Web Conference. 2015:707-719.
- [3] CAO M Y. Flower Image Retrieval Based on Cultural Gene Feature Selection [J]. Electronic World, 2020, 591(9):61-62.
- [4] SAEYS Y, INZA I, LARRANAGA P. A review of feature selection techniques in bioinformatics [J]. Bioinformatics, 2007, 23(19):2507-2517.

- [5] YU L, LIU H. Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution[C]// Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003). USA, 2003.
- [6] LIANG Y C. Filter Feature Selection in Confrontation Environment [D]. Guangzhou: South China University of Technology, 2020.
- [7] DASH M, GOPALKRISHNAN V. Distance Based Feature Selection for Clustering Microarray Data[C]// International Conference on Database Systems for Advanced Applications. 2008.
- [8] CHENG Y W, LI G F, JIANG G Z, et al. A bidirectional recursive EMG feature selection method based on EMG signal. CN111209857A[P]. 2020.
- [9] CHEN C J, JIANG L, LEI N, et al. Interactive feature selection method based on crowdsourcing learning[J]. Science in China: Information Science, 2020, 50(6): 20-38.
- [10] XU Z, RONG J, YE J, et al. Non-Monotonic Feature Selection [C]// Proceedings of the 26th Annual International Conference on Machine Learning. Canada, 2009.
- [11] NICOLETTA D, PES B. A Framework for Multi-class Learning in Micro-array Data Analysis[C]// Conference on Artificial Intelligence in Medicine in Europe. Berlin; Springer, 2009.
- [12] ALSALANI M G, ELREFAEI L A. Convolutional Neural Network Based Feature Extraction for IRIS Recognition[J]. International Journal of Computer Science & Information Technology, 2018, 10(2): 65-78.
- [13] ZENG ST, CAO Y C, LIN Q, et al. Classification of SPECT pulmonary perfusion images based on ResNet depth model [J]. Journal of Northwest Minzu University (Cience Dition), 2021, 42(2): 27-35.
- [14] KADHIM M A, ABED M H. Convolutional Neural Network for Satellite Image Classification [J/OL]. Intelligent Information and Database Systems: Recent Developments, 2019. https://www.researchgate.net/publication/331557018_Convolutional_Neural_Network_for_Satellite_Image_Classification.
- [15] DEWI C, CHEN R. Human Activity Recognition Based on Evolution of Features Selection and Random Forest[C]// 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC). 2019.
- [16] RAO H, SHI X, RODRIGUE A K, et al. Feature selection based on artificial bee colony and gradient boosting decision tree[J/OL]. Applied Soft Computing, 2019, 74: 634-642. <https://www.sciencedirect.com/science/article/abs/pii/S1568494618305933>.
- [17] DING P, XU A J, ZHOU S Y. Tea yield prediction based on gradient lifting decision tree and multi-feature combination [J]. Southwest Agricultural Journal, 2021, 34(7): 1556-1563.
- [18] GAO Y. Simulation of carbon dioxide flux based on adaptive neurofuzzy inference system [D]. Beijing; Beijing Forestry University, 2019.
- [19] CHEN J, WANG J, WANG X, et al. Predicting Drug Target Interactions Based on GBDT [C]// International Conference on Machine Learning and Data Mining in Pattern Recognition. Cham, 2018: 202-212.
- [20] CHEN J, ZHANG Y X, JIANG Y Y. Texture image classification method based on multi-feature and random forest [J]. Sensors and Microsystems, 2019, 38, 334(12): 64-67.
- [21] GRARLI I, ADEL M, BOURENNAME S, et al. Histogram-Based Features Selection and Volume of Interest Ranking for Brain PET Image Classification[C]// IEEE Journal of Translational Engineering in Health and Medicine. 2018.
- [22] GUYON I M, ANDR E. An introduction to variable and feature selection[J/OL]. The Journal of Machine Learning Research, 2003, 3 (Mar): 1157-1182. <https://www.jmlr.org/papers/volume3/guyon03a/guyon03a.pdf?ref=driverlayer.com/web>.
- [23] DIAS P A, TABB A, MEDEIROS H. Apple flower detection using deep convolutional networks[J/OL]. Computers in Industry, 2018, 99: 17-28. <https://www.sciencedirect.com/science/article/abs/pii/S016636151730502X>.
- [24] GHAZI M M, YANIKOGLU B, APTOULA E. Plant identification using deep neural networks via optimization of transfer learning parameters[J]. Neurocomputing, 2017, 235 (26): 228-235.
- [25] SARDOGAM M, TUNCER A, OZEN Y. Plant Leaf Disease Detection and Classification Based on CNN with LVQ Algorithm [C]// 2018 3rd International Conference on Computer Science and Engineering (UBMK). 2018.
- [26] PAWARA P, OKAFOR E, SCHOMAKER L, et al. A. Wiering. Data Augmentation for Plant Classification[C]// Advanced Concepts for Intelligent Vision Systems (ACIVS2017). 2017.
- [27] NAGASUBRAMANIAN K, JONES S, SINGH A K, et al. Explaining hyperspectral imaging based plant disease identification: 3D CNN and saliency maps[J]. arXiv:1804.08831, 2018.
- [28] GAYATHRI S, GOPI V P, PALANISAMY P A. lightweight CNN for Diabetic Retinopathy classification from fundus images [J/OL]. Biomedical Signal Processing and Control. 2020, 62. <https://www.sciencedirect.com/science/article/abs/pii/S1746809420302676>.
- [29] VAPNIK V. Statistical Learning Theory[M/OL]. DBLP, 1998. https://www.researchgate.net/publication/220694713_Statistical_Learning_Theory.
- [30] DEMIAR J, SCHUURMANS D. Statistical Comparisons of Classifiers over Multiple DataSets [J]. Journal of Machine Learning Research. 2006, 7(1): 1-30.
- [31] ZHU C, GAO D. Improved multi-kernel classification machine with Nyström approximation technique[J]. Pattern Recognition, 2015, 48(4): 1490-1590.
- [32] ZHU C, WANG P, MA L, et al. Global and local multi-view multi-label learning with incomplete views and labels[J]. Neural Computing and Applications, 2020, 371(2): 67-77.



WANG Pan-hong, born in 1998, post-graduate. Her main research interests include machine learning, online features selection.



ZHU Chang-ming, born in 1988, Ph.D. candidate at the East China University of Science and Technology. He is an associate professor in Shanghai Maritime University. His research interest covers image processing and multi-view learning.