



(12) 发明专利

(10) 授权公告号 CN 114782752 B

(45) 授权公告日 2023. 09. 05

(21) 申请号 202210487571.7

G06V 10/82 (2022.01)

(22) 申请日 2022.05.06

G06N 3/0895 (2023.01)

G06N 3/096 (2023.01)

(65) 同一申请的已公布的文献号

申请公布号 CN 114782752 A

(56) 对比文件

(43) 申请公布日 2022.07.22

CN 113076437 A, 2021.07.06

CN 114387473 A, 2022.04.22

(73) 专利权人 兰州理工大学

CN 113378941 A, 2021.09.10

CN 113378942 A, 2021.09.10

地址 730050 甘肃省兰州市七里河区兰工
坪路287号

CN 108154156 A, 2018.06.12

CN 113222011 A, 2021.08.06

(72) 发明人 李晓旭 郭晓伟 刘俊 金志宇
汤卓和 曾俊瑀 李睿凡 马占宇
陶剑

CN 113408606 A, 2021.09.17

CN 113706487 A, 2021.11.26

(74) 专利代理机构 北京挺立专利事务所(普通
合伙) 11265

CN 114299362 A, 2022.04.08

CN 114329124 A, 2022.04.12

专利代理师 高福勇

叶萌. 基于集成策略的半监督小样本学习方法研究.《中国优秀硕士学位论文全文数据库》.2021, (第2期), 第6-22、37-49页.

(51) Int.Cl.

审查员 任丽娜

G06V 10/764 (2022.01)

G06V 10/774 (2022.01)

权利要求书3页 说明书8页 附图3页

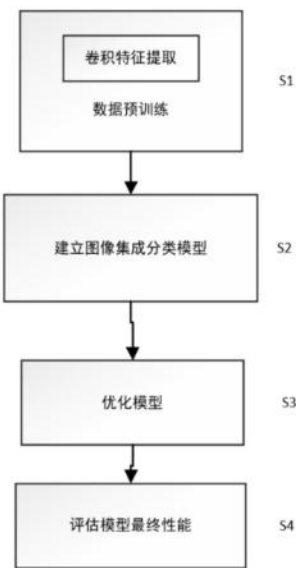
(54) 发明名称

值。

基于自训练的小样本图像集成分类方法及装置

(57) 摘要

本发明公开了一种基于自训练的小样本图像集成分类方法及装置,通过迁移基类数据的卷积特征提取,使用查询样本进行基分类器的自训练,不断地将伪标签集加到支持集扩充支持集,并运用到下一次基分类器自训练中,提高了基分类器自训练结果的可靠性。自训练过程中产生的基分类器,不断通过模型平均得到集成分类器,经过多次迭代自训练和分类器集成过程,得到最终的集成分类器,通过构建损失函数,使得基分类器在查询样本上有确定的、不同的预测,实现了基分类器的不同,解决现有小样本分类方法中,基于特征迁移和查询样本自训练的小样本图像集成分类中基学习器的多样性问题,对于提升图像的分类效果非常显著,具有很高的使用价



1.一种基于自训练的小样本图像集成分类方法,其特征在于,包括以下步骤:

S1、对数据进行预处理,其中数据包括训练集和测试集;预处理方法为:

S 1 1, 将数据分为训练集 $D_{train} = \{(x_i, y_i), y_i \in L_{train}\}_{i=1}^{N_{train}}$ 和测试集 $D_{test} = \{(x_j^*, y_j^*), y_j^* \in L_{test}\}_{j=1}^{N_{test}}$ 两部分,且这两部分的类别空间互斥,将 D_{train} 作为基类数据用于训练模型, D_{test} 作为新类数据用于测评模型性能;

S12、对于C-way K-shot分类任务,从 D_{train} 中随机选出C个类别,每个类别中随机选出M个样本,M个样本中K个样本作为支持集 S_c ,其余M-K个样本作为查询集 Q_c , S_c 和 Q_c 构成一个任务 T_i ; 同样地,对于 D_{test} 有任务 \tilde{T}_z ;

S2、构建基于自训练的小样本图像集成分类的模型:使用Baseline++网络结构构建基分类器,使用查询样本进行基分类器的自训练,不断地将伪标签集加到支持集扩充支持集,自训练过程中产生的基分类器,不断通过模型平均得到集成分类器,经过多次迭代自训练和分类器集成过程,得到最终的集成分类器;

在使用Baseline++网络结构构建基分类器阶段,建立特征提取模块、关系网络模块和基分类器模块;其中,特征提取模块基于卷积特征提取支持集的特征,特征提取模块的公式如下:

$$g_\varphi \leftarrow \arg \min_{\varphi} \left[- \sum_{i=1}^{n_i} \sum_{c=1}^C y_{ic}^s \log \left(f_{\theta_k} \left(g_\varphi \left(x_i^s \right) \right)_c \right) \right] \quad (1)$$

该公式为最大似然估计公式的变形,公式中, g_φ 是特征提取模块; f_{θ_k} 为分类器,当 $k=0$ 时, f_{θ_0} 由支持集直接训练得到;

关系网络模块基于特征提取模块输出的支持集特征和表示每个类的学习权重向量之间的余弦距离进行预测,关系网络模块的公式如下:

$$f_\psi \leftarrow \arg \min_{\psi} \left[- \sum_{i=1}^{n_i} \sum_{c=1}^C f_\psi \left(g_\varphi \left(x_i^u \right), A_c \right)_c \hat{y}_{ic}^u \log \left(f_{\theta_k} \left(g_\varphi \left(x_i^u \right) \right)_c \right) \right] \quad (2)$$

该公式为最大似然估计公式的变形,公式中, f_ψ 为关系模块,用于获取两个参数之间的距离,在 $f_\psi \left(g_\varphi \left(x_i^u \right), A_c \right)_c$ 中,第一个参数 $g_\varphi \left(x_i^u \right)$ 为提取的无标签支持集特征,第二个参数 A_c 表示当前无标签支持集到所有支持集之间的距离总和求平均;

基分类器模块由前两个模块参与构建,每次更新得到当前层次的分类器、集成分类器和伪标签预测集,其中,基分类器由有标签的支持集 S_c 和无标签的支持集 U 更新,基分类器模块的公式如下:

$$f_{\theta_k} \leftarrow \arg \min_{\theta_k} \left[- \sum_{i=1}^{n_i} \sum_{c=1}^C y_{ic}^s \log \left(f_{\theta_k} \left(g_\varphi \left(x_i^s \right) \right)_c \right) - \sum_{i=1}^{n_i} \sum_{c=1}^C f_\psi \left(g_\varphi \left(x_i^u \right), A_c \right)_c \hat{y}_{ic}^u \log \left(f_{\theta_k} \left(g_\varphi \left(x_i^u \right) \right)_c \right) \right] \quad (3)$$

该公式为最大似然估计公式的变形,公式中, f_{θ_k} 为分类器,当 $k=0$ 时, f_{θ_0} 由支持集直

接训练得到, \mathcal{G}_φ 为特征提取模块, f_ψ 为关系网络模块, C 为任务个数;

对于每次任务抽 T_i 取三个数据集: 有标签的支持集 $S_c = \{(x_i^s, y_i^s)\}_{i=1}^{n_1}$, 无标签的支持集 $U = \{(x_i^u)\}_{i=1}^{n_4}$ 和查询集 $Q_c = \{(x_i^q, y_i^q)\}_{i=1}^{n_2}$, 步骤包括:

S21, 在支持集上, 训练得到基分类器0, 同时将基分类器0赋给集成分类器0;

S22, 使用基分类器0对查询集样本进行预测, 得到预测较为准确的样本伪标签预测集1, 将其加入支持集进行训练得到基分类器1;

S23, 将集成分类器0和基分类器1通过模型平均进行集成, 得到集成分类器1;

S24, 上述步骤S21-S23迭代多次, 得到最终的集成分类器;

S3、使用测试集数据对基于自训练的小样本图像集成分类的模型进行训练, 优化模型参数;

具体包括:

S31, 对于每一个任务 \tilde{T}_z , 抽取三个数据集: 有标签的支持集 $S_c^* = \{(x_j^s, y_j^s)\}_{j=1}^{n_5}$ 、无标签的支持集 $U^* = \{(x_j^u)\}_{j=1}^{n_3}$ 和查询集 $Q_c^* = \{(x_j^q, y_j^q)\}_{j=1}^{n_6}$;

S32, 通过有标签支持集 S_c^* 和无标签支持集 U^* 进一步微调训练过的分类, 更新公式如下:

$$\begin{aligned} f_{\theta_k} \leftarrow \arg \min_{\theta_k} [& - \sum_{j=1}^{n_5} \sum_{c=1}^C y_{jc}^s \log \left(f_{\theta_k} \left(g_\varphi \left(x_j^s \right) \right)_c \right) \\ & - \sum_{j=1}^{n_3} \sum_{c=1}^C f_\psi \left(g_\varphi \left(x_j^u \right), A_c \right) \hat{y}_{jc}^{u*} \log \left(f_{\theta_k} \left(g_\varphi \left(x_j^u \right) \right)_c \right) \\ & - \sum_{j=1}^{n_5+n_3} \text{KL} \left(f_{\theta_k} \left(x_j^{su*} \right), h_{\gamma_{k-1}} \left(x_j^{su*} \right) \right) \end{aligned} \quad (4)$$

其中, $\sum_{j=1}^{n_5+n_3} (\cdot)$ 表示整个支持集, 即包括有标签支持集和无标签支持集两部分,

$\text{KL} \left(f_{\theta_k} \left(x_j^{su*} \right), h_{\gamma_{k-1}} \left(x_j^{su*} \right) \right)$ 为度量在支持集下当前迭代下生成的分类器和上一次迭代过程中生成的集成分类器两个参数的概率分布之间的差异;

S33, 集成分类器参数更新如下:

$$\gamma_k = \omega_1 \theta_k + \omega_2 \gamma_{k-1} \quad (5)$$

公式中, γ_{k-1} 为集成分类器 $k-1$ 的参数, θ_k 为基分类器 k 的参数, ω_1 和 ω_2 为集成分类器 $k-1$ 和基分类器 k 的权重, 当 $k=0$ 时, $\gamma_0 = \theta_0$;

S34、查询集的标签更新如下:

$$\hat{y}_j^{q*} = h_{\gamma_m} \left(g_\varphi \left(x_j^{q*} \right) \right), x_j^{q*} \in Q \quad (6)$$

公式中, \hat{y}_j^{q*} 为查询集的标签, h_{γ_m} 为第 m 个集成分类器的参数, g_φ 提取特征模块, x_j^{q*} 为查询集的样本, 其中 x_j^{q*} 属于有理数集;

S4、通过新类数据上所有任务中查询集标签的预测值和真实值, 评估模型的最终性能。

2. 根据权利要求1所述的基于自训练的小样本图像集成分类方法,其特征在于,在步骤S4中,通过新类数据上所有任务中查询集标签的预测值和真实值,评估集成分类方法的最终性能。

3. 一种基于自训练的小样本图像集成分类装置,其特征在于,用以实现权利要求1-2任一项所述的基于自训练的小样本图像集成分类方法,包括以下模块:

预训练模块:使用基类数据上预训练得到特征提取模块;

自训练模块:用于构建基分类器,并使用查询样本进行基分类器的自训练,不断地将伪标签集加到支持集扩充支持集;

集成分类模块:用于自训练过程中产生的基分类器,不断通过模型平均得到集成分类器。

基于自训练的小样本图像集成分类方法及装置

技术领域

[0001] 本发明涉及图像分类技术领域,尤其涉及一种基于自训练的小样本图像集成分类方法及装置。

背景技术

[0002] 近年来,随着计算机技术的发展,人们浏览的信息日益丰富,每天都有大量图片被上传到网络,由于数量巨大,人工已经无法对此进行分类。在很多大样本图像分类任务上,机器的识别性能已经超越人类。然而,当样本量比较少时,机器的识别水平仍与人类存在较大差距。因此,研究高效可靠的图片分类算法有很迫切的社会需求。

[0003] 小样本分类(Few-shot Classification)属于小样本学习(Few-shot Learning)范畴,往往包含类别空间不相交的两类数据,即基类数据和新类数据。小样本分类旨在利用基类数据学习的知识和新类数据的少量标记样本(支持样本)来学习分类规则,准确预测新类任务中未标记样本(查询样本)的类别。

[0004] 小样本图像分类是当前计算机视觉、人工智能领域中亟待解决的研究问题。现有的、较为成功的大样本图像分类方法严重依赖样本的数量,而现实世界中事物的样本量是服从长尾分布的,即大量事物的样本量都是严重不足的,例如在军事、医疗、工业,天文等领域,样本采集需要消耗大量的人力、物力、时间和经济成本,很难采集大规模的图像样本。因此,开展小样本图像分类的研究对图像分类技术的广泛应用具有重要价值。另外,小样本图像分类的研究目标就是通过少量图像样本的学习,使得机器获得与人类水平相当的识别能力。因此,开展小样本图像分类的研究对响应我国的人工智能发展规划以及突破人工智能的理论瓶颈具有重要意义。

[0005] 目前小样本图像分类的研究中,特征迁移旨在通过引入源数据特征来帮助完成目标数据特征域的机器学习任务。一个机器学习任务中,可能由于目标特征域缺少足够的标签而导致学习的效果很差。通过挖掘源数据与目标数据的交叉特征结构,或者借助中间数据进行“桥接”,可以帮助我们在目标数据特征上进行的机器学习任务实现不同特征空间之间的知识迁移。例如,我们在进行图片数据分类时缺少足够带有标记的训练数据,就可以借助已经标注好的文本数据,以及具有交叉特征的中间数据来协助提高在图片数据上的学习效果。然而,基于特征迁移的基于自训练的小样本图像集成分类中存在基学习器的多样性问题。因此,本发明主要通过迁移基类数据的卷积特征提取,并基于查询样本进行自训练,通过构建损失函数,使得基分类器在查询样本上有确定的、不同的预测,实现基分类器的不同。

发明内容

[0006] 本发明针对上述技术问题,提出一种基于自训练的小样本图像集成分类方法及装置,使用查询样本进行基分类器的自训练,不断地将伪标签集加到支持集扩充支持集。自训练过程中产生的基分类器,不断通过模型平均得到集成分类器,经过多次迭代自训练和分

类器集成过程,得到最终的集成分类器,旨在解决现有小样本分类方法中,基于特征迁移的基于自训练的小样本图像集成分类中基学习器的多样性问题。

[0007] 为了实现上述目的,本发明提供如下技术方案:

[0008] 一方面,本发明提供了一种基于自训练的小样本图像集成分类方法,包括以下步骤:

[0009] S1、对数据进行预处理,其中数据包括训练集和测试集;

[0010] S2、构建基于自训练的小样本图像集成分类的模型:使用Baseline++网络结构构建基分类器,使用查询样本进行基分类器的自训练,不断地将伪标签集加到支持集扩充支持集,自训练过程中产生的基分类器,不断通过模型平均得到集成分类器,经过多次迭代自训练和分类器集成过程,得到最终的集成分类器;

[0011] S3、使用测试集数据对基于自训练的小样本图像集成分类的模型进行训练,优化模型参数;

[0012] S4、通过新类数据上所有任务中查询集标签的预测值和真实值,评估模型的最终性能。

[0013] 进一步地,步骤S1的预处理方法为:

[0014] S11,将数据分为训练集 $D_{train} = \{(x_i, y_i), y_i \in L_{train}\}_{i=1}^{N_{train}}$ 和测试集

$D_{test} = \{(x_j^*, y_j^*), y_j^* \in L_{test}\}_{j=1}^{N_{test}}$ 两部分,且这两部分的类别空间互斥,将 D_{train} 作为基类数据用于训练模型, D_{test} 作为新类数据用于测评模型性能;

[0015] S12,对于C-way K-shot分类任务,从 D_{train} 中随机选出C个类别,每个类别中随机选出M个样本,其中K个样本作为支持样本 S_c ,其余M-K个样本作为查询样本 Q_c , S_c 和 Q_c 构成一个任务 T_i ; 同样地,对于 D_{test} 有任务 \tilde{T}_z 。

[0016] 进一步地,步骤S2中,使用Baseline++网络结构,对支持集进行训练得到m个基分类器,在此阶段,建立特征提取模块、关系网络模块和基分类器模块;其中,特征提取模块基于卷积特征提取支持集的特征,关系网络模块基于特征提取模块输出的支持集特征和表示每个类的学习权重向量之间的余弦距离进行预测,上述两个模块在此阶段共享一个,基分类器模块由前两个模块参与构建;对于每次任务 T_i 抽取三个数据集:有标签的支持集 $S_c = \{(x_i^s, y_i^s)\}_{i=1}^{n_1}$, 无标签的支持集 $U = \{(x_i^u)\}_{i=1}^{n_4}$ 和查询集 $Q_c = \{(x_i^q, y_i^q)\}_{i=1}^{n_2}$, 步骤包括:

[0017] S21,在支持集上,训练得到基分类器0,同时将基分类器0赋给集成分类器0;

[0018] S22,使用基分类器0对查询集样本进行预测,得到预测较为准确的样本伪标签预测集1,将其加入支持集进行训练得到基分类器1;

[0019] S23,将集成分类器0和基分类器1通过模型平均进行集成,得到集成分类器1;

[0020] S24,上述步骤S21-S23迭代多次,得到最终的集成分类器。

[0021] 进一步地,每次更新得到当前层次的分类器,集成分类器和伪标签预测集,其中,分类器由有标签的支持集 S_c 和无标签的支持集 U 更新,具体公式为:

$$\begin{aligned}
f_{\theta_k} \leftarrow \arg \min_{\theta_k} & \left[- \sum_{i=1}^{n_l} \sum_{c=1}^C y_{ic}^s \log \left(f_{\theta_k} \left(g_{\varphi} \left(x_i^s \right) \right)_c \right) \right. \\
& \left. - \sum_{i=1}^{n_u} \sum_{c=1}^C f_{\psi} \left(g_{\varphi} \left(x_i^u \right), A_c \right) \hat{y}_{ic}^u \log \left(f_{\theta_k} \left(g_{\varphi} \left(x_i^u \right) \right)_c \right) \right] \quad (1)
\end{aligned}$$

[0023] 该公式为最大似然估计公式的变形,其中, f_{θ_k} 为分类器,当 $k=0$ 时, f_{θ_0} 由支持集直接训练得到, g_{φ} 为特征提取模块, f_{ψ} 为关系网络模块, C 为任务个数, A_c 为当前无标签支持集到所有支持集之间的距离总和求平均。

[0024] 进一步地,特征提取模块采用如下公式:

$$g_{\varphi} \leftarrow \arg \min_{\varphi} \left[- \sum_{i=1}^{n_l} \sum_{c=1}^C y_{ic}^s \log \left(f_{\theta_k} \left(g_{\varphi} \left(x_i^s \right) \right)_c \right) \right] \quad (2)$$

[0026] 该公式为最大似然估计公式的变形,特征提取时,在基分类器下对每个查询样本进行特征提取,它们的和进行最大似然估计处理,得到的是特征提取模块需要的结果。

[0027] 进一步地,关系网络模块采用如下公式:

$$f_{\psi} \leftarrow \arg \min_{\psi} \left[- \sum_{i=1}^{n_u} \sum_{c=1}^C f_{\psi} \left(g_{\varphi} \left(x_i^u \right), A_c \right) \hat{y}_{ic}^u \log \left(f_{\theta_k} \left(g_{\varphi} \left(x_i^u \right) \right)_c \right) \right] \quad (3)$$

[0029] 该公式为最大似然估计公式的变形, f_{ψ} 为关系网络模块,用于获取两个参数之间的距离,在 $f_{\psi} \left(g_{\varphi} \left(x_i^u \right), A_c \right)$ 中,第一个参数 $g_{\varphi} \left(x_i^u \right)$ 为提取的无标签支持集特征,第二个参数 A_c 表示当前无标签支持集到所有支持集之间的距离总和求平均,将所有样本的数据进行最大似然估计处理,得到的是关系网络模块需要的结果。

[0030] 进一步地,步骤S3具体包括:

[0031] S31,对于每一个任务 \tilde{T}_z ,抽取三个数据集:有标签的支持集 $S_c^* = \{(x_j^*, y_j^*)\}_{j=1}^{n_5}$ 、无标签的支持集 $U^* = \{(x_j^{u*})\}_{j=1}^{n_3}$ 和查询集 $Q_c^* = \{(x_j^{q*}, y_j^{q*})\}_{j=1}^{n_6}$;

[0032] S32,通过有标签支持集 S_c^* 和无标签支持集 U^* 进一步微调训练过的分类器,更新公式如下:

$$\begin{aligned}
f_{\theta_k} \leftarrow \arg \min_{\theta_k} & \left[- \sum_{j=1}^{n_5} \sum_{c=1}^C y_{jc}^{s*} \log \left(f_{\theta_k} \left(g_{\varphi} \left(x_j^{s*} \right) \right)_c \right) \right. \\
& - \sum_{j=1}^{n_3} \sum_{c=1}^C f_{\psi} \left(g_{\varphi} \left(x_j^{u*} \right), A_c \right) \hat{y}_{jc}^{u*} \log \left(f_{\theta_k} \left(g_{\varphi} \left(x_j^{u*} \right) \right)_c \right) \\
& \left. - \sum_{j=1}^{n_5+n_3} \text{KL} \left(f_{\theta_k} \left(x_j^{su*} \right), h_{\gamma_{k-1}} \left(x_j^{su*} \right) \right) \right] \quad (4)
\end{aligned}$$

[0034] 其中, $\sum_{j=1}^{n_5+n_3} (\cdot)$ 表示整个支持集,即包括有标签支持集和无标签支持集两部分,

$\text{KL} \left(f_{\theta_k} \left(x_j^{su*} \right), h_{\gamma_{k-1}} \left(x_j^{su*} \right) \right)$ 为度量在支持集下当前迭代下生成的分类器和上一次迭代过程生成的集成分类器两个参数的概率分布之间的差异;

[0035] S33,集成分类器参数更新如下:

$$[0036] \quad \gamma_k = \omega_1 \theta_k + \omega_2 \gamma_{k-1} \quad (5)$$

[0037] 其中, γ_{k-1} 为集成分类器k-1的参数, θ_k 为基分类器k的参数, ω_1 和 ω_2 为集成分类器k-1和基分类器k的权重, 当k=0时, $\gamma_0 = \theta_0$ 。

[0038] S34, 查询集的标签更新如下:

$$[0039] \quad \hat{y}_j^{q*} = h_{\gamma_m}(g_\phi(x_j^{q*})), x_j^{q*} \in Q \quad (6)$$

[0040] 其中, \hat{y}_j^{q*} 为查询集的标签, h_{γ_m} 为第m个集成分类器的参数, g_ϕ 为提取特征模块, x_j^{q*} 为查询集的样本, 其中 x_j^{q*} 属于有理数集。

[0041] 进一步地, 步骤S4中, 通过新类数据上所有任务中查询集标签的预测值和真实值, 评估集成分类方法的最终性能。

[0042] 另一方面, 本发明还提供了一种基于自训练的小样本图像集成分类装置, 用以实现上述的任一项方法, 包括以下模块:

[0043] 预训练模块: 使用基类数据上预训练得到特征提取模块;

[0044] 自训练模块: 用于构建基分类器, 并使用查询样本进行基分类器的自训练, 不断地将伪标签集加到支持集扩充支持集;

[0045] 集成分类模块: 用于自训练过程中产生的基分类器, 不断通过模型平均得到集成分类器。

[0046] 与现有技术相比, 本发明的有益效果为:

[0047] 本发明建立了一种基于自训练的小样本图像集成分类方法及装置, 通过迁移基类数据的卷积特征提取, 使用查询样本进行基分类器的自训练, 不断地将伪标签集加到支持集扩充支持集, 并运用到下一次基分类器自训练中, 提高了基分类器自训练结果的可靠性。自训练过程中产生的基分类器, 不断通过模型平均得到集成分类器, 经过多次迭代自训练和分类器集成过程, 得到最终的集成分类器, 通过构建损失函数, 使得基分类器在查询样本上有确定的、不同的预测, 实现了基分类器的不同, 解决现有小样本分类方法中, 基于特征迁移和查询样本自训练的小样本图像集成分类中基学习器的多样性问题, 对于提升图像的分类效果非常显著, 具有很高的使用价值。

附图说明

[0048] 为了更清楚地说明本申请实施例或现有技术中的技术方案, 下面将对实施例中所需要使用的附图作简单地介绍, 显而易见地, 下面描述中的附图仅仅是本发明中记载的一些实施例, 对于本领域普通技术人员来讲, 还可以根据这些附图获得其他的附图。

[0049] 图1为本发明实施例提供的基于自训练的小样本图像集成分类方法的流程图。

[0050] 图2为本发明实施例提供的基于查询样本自训练和模型平均的基于自训练的小样本图像集成分类的模型结构图。

[0051] 图3为本发明实施例提供的Baseline++网络结构图。

[0052] 图4为本发明实施例提供的基于自训练的小样本图像集成分类装置功能模块示意图。

具体实施方式

[0053] 下面结合本发明实施例中的附图,对本发明实施例中的技术方案进行清楚、完整地描述,显然,所描述的实施例是本发明一部分实施例,而不是全部的实施例。本发明中的实施例,本领域技术人员在没有做出创造性劳动前提下所获得的所有其他实施例,都属于本发明保护的范围。

[0054] 根据本文公开的一个方面,本发明拟使用查询样本进行基分类器的自训练,不断地将伪标签集加到支持集扩充支持集,自训练过程中产生基分类器,不断通过模型平均得到集成分类器。如图1所示,本发明提供了一种基于自训练的小样本图像集成分类方法,包括以下步骤:

[0055] S1、对数据进行预处理,其中数据包括训练集和测试集;

[0056] 具体地,步骤S1的预处理方法为:

[0057] S11,将数据分为训练集 $D_{train} = \{(x_i, y_i), y_i \in L_{train}\}_{i=1}^{N_{train}}$ 和测试集

$D_{test} = \{(x_j^*, y_j^*), y_j^* \in L_{test}\}_{j=1}^{N_{test}}$ 两部分,且这两部分的类别空间互斥,将 D_{train} 作为基类数据用于训练模型, D_{test} 作为新类数据用于测评模型性能;

[0058] S12,对于C-way K-shot分类任务,从 D_{train} 中随机选出C个类别,每个类别中随机选出M个样本,其中K个样本作为支持样本 S_c ,其余M-K个样本作为查询样本 Q_c , S_c 和 Q_c 构成一个任务 T_i ; 同样地,对于 D_{test} 有任务 \tilde{T}_Z 。

[0059] S2、使用Baseline++网络结构构建基分类器,引入基于查询样本自训练进行联合训练的方法,构建基于查询样本自训练和模型平均的基于自训练的小样本图像集成分类的模型。如图2所示。

[0060] 具体地,基分类器主要采用Baseline++网络结构进行构建,也可以使用其他的网络结构进行构建,例如采用baseline网络结构来构建基分类器。如图3所示,其中关于特征提取网络模块,使用基类数据上预训练得到的特征提取模块;对于新类数据上的分类任务,抽取三个数据集:有标签的支持集 $S_c^* = \{(x_j^{s*}, y_j^{s*})\}_{j=1}^{n_s^*}$, 无标签的支持集 $U^* = \{(x_j^{u*})\}_{j=1}^{n_u^*}$ 和查询集 $Q_c^* = \{(x_j^{q*}, y_j^{q*})\}_{j=1}^{n_q^*}$;

[0061] 如图2所示优化基分类器和集成分类器。具体地,步骤S2中,使用Baseline++网络结构,对支持集进行训练得到m个基分类器,在此阶段,建立特征提取模块、关系网络模块和基分类器模块;其中,特征提取模块基于卷积特征提取支持集的特征,关系网络模块基于特征提取模块输出的支持集特征和表示每个类的学习权重向量之间的余弦距离进行预测,上述两个模块在此阶段共享一个,基分类器模块由前两个模块参与构建;对于每次任务抽取三个数据集:有标签的支持集 $S_c = \{(x_i^s, y_i^s)\}_{i=1}^{n_1}$, 无标签的支持集 $U = \{(x_i^u)\}_{i=1}^{n_4}$ 和查询集 $Q_c = \{(x_i^q, y_i^q)\}_{i=1}^{n_2}$, 步骤具体包括:

[0062] S21,在支持集上,训练得到基分类器0,同时将基分类器0赋给集成分类器0;

[0063] S22,使用基分类器0对查询集样本进行预测,得到预测较为准确的样本伪标签预测集1,将其加入支持集进行训练得到基分类器1;

[0064] S23,将集成分类器0和基分类器1通过模型平均进行集成,得到集成分类器1;

[0065] S24,上述步骤S21-S23迭代多次,得到最终的集成分类器。

[0066] S25,该模型结构类似于迭代模型,每次更新得到当前层次的分类器,集成分类器和伪标签预测集,其中,分类器由有标签的支持集 S_c 和无标签的支持集 U 更新,具体公式为:

$$f_{\theta_k} \leftarrow \arg \min_{\theta_k} [-\sum_{i=1}^{n_l} \sum_{c=1}^C y_{ic}^s \log(f_{\theta_k}(g_{\varphi}(x_i^s)))_c] \quad (1)$$

$$-\sum_{i=1}^{n_u} \sum_{c=1}^C f_{\psi}(g_{\varphi}(x_i^u), A_c) \hat{y}_{ic}^u \log(f_{\theta_k}(g_{\varphi}(x_i^u)))_c]$$

[0068] 该公式为最大似然估计公式的变形,其中, f_{θ_k} 为分类器,当 $k=0$ 时, f_{θ_0} 由支持集直接训练得到, g_{φ} 为特征提取模块, f_{ψ} 为关系网络模块, C 为任务个数, A_c 为当前无标签支持集到所有支持集之间的距离总和求平均。

[0069] S26,特征提取模块采用如下公式:

$$g_{\varphi} \leftarrow \arg \min_{\varphi} [-\sum_{i=1}^{n_l} \sum_{c=1}^C y_{ic}^s \log(f_{\theta_k}(g_{\varphi}(x_i^s)))_c] \quad (2)$$

[0071] 该公式为最大似然估计公式的变形,特征提取时,在基分类器下对每个查询样本进行特征提取,它们的和进行最大似然估计处理,得到的是特征提取模块需要的结果。

[0072] S27,关系网络模块采用如下公式:

$$f_{\psi} \leftarrow \arg \min_{\psi} [-\sum_{i=1}^{n_u} \sum_{c=1}^C f_{\psi}(g_{\varphi}(x_i^u), A_c) \hat{y}_{ic}^u \log(f_{\theta_k}(g_{\varphi}(x_i^u)))_c] \quad (3)$$

[0074] 该公式为最大似然估计公式的变形, f_{ψ} 为关系网络模块,用于获取两个参数之间的距离,在 $f_{\psi}(g_{\varphi}(x_i^u), A_c)$ 中,第一个参数 $g_{\varphi}(x_i^u)$ 为提取的无标签支持集特征,第二个参数 A_c 表示当前无标签支持集到所有支持集之间的距离总和求平均,将所有样本的数据进行最大似然估计处理,得到的是关系网络模块需要的结果。

[0075] S28,将S26的公式(2)和S27的公式(3),经过最大似然估计处理,得到的结果就是S25的公式(1)。

[0076] S3、使用测试集数据对基于自训练的小样本图像集成分类的模型进行训练,优化模型参数;

[0077] 具体地,使用的数据是测试集 D_{test} 。其中,步骤S3具体包括:

[0078] S31,对于每一个任务 \tilde{T}_z ,抽取三个数据集:有标签的支持集 $S_c^* = \{(x_j^*, y_j^*)\}_{j=1}^{n_5}$ 、无标签的支持集 $U^* = \{(x_j^*)\}_{j=1}^{n_3}$ 和查询集 $Q_c^* = \{(x_j^*, y_j^*)\}_{j=1}^{n_6}$;

[0079] S32,通过有标签支持集 S_c^* 和无标签支持集 U^* 进一步微调训练过的分类器,更新公式如下:

$$\begin{aligned}
f_{\theta_k} \leftarrow \arg \min_{\theta_k} & \left[- \sum_{j=1}^{n_5} \sum_{c=1}^C y_{jc}^{s*} \log \left(f_{\theta_k} \left(g_{\varphi} \left(x_j^{s*} \right) \right)_c \right) \right. \\
[0080] \quad & \left. - \sum_{j=1}^{n_3} \sum_{c=1}^C f_{\varphi} \left(g_{\varphi} \left(x_j^{u*} \right), A_c \right) \hat{y}_{jc}^{u*} \log \left(f_{\theta_k} \left(g_{\varphi} \left(x_j^{u*} \right) \right)_c \right) \right. \\
& \left. - \sum_{j=1}^{n_5+n_3} \text{KL} \left(f_{\theta_k} \left(x_j^{su*} \right), h_{\gamma_{k-1}} \left(x_j^{su*} \right) \right) \right] \quad (4)
\end{aligned}$$

[0081] 该公式基于S25中的公式(1),通过S3中的 S_c^* 和 U^* 进一步微调训练过的分类器,在此基础上进一步考虑各分布之间的交叉熵,即KL散度这一因素,得到此处公式(4)。

[0082] 其中, $\sum_{j=1}^{n_5+n_3} (\cdot)$ 表示整个支持集,即包括有标签支持集和无标签支持集两部分,

$\text{KL} \left(f_{\theta_k} \left(x_j^{su*} \right), h_{\gamma_{k-1}} \left(x_j^{su*} \right) \right)$ 为度量在支持集下当前迭代下生成的分类器和上一次迭代过程生成的集成分类器两个参数的概率分布之间的差异;

[0083] S33,集成分类器参数更新如下:

$$[0084] \quad \gamma_k = \omega_1 \theta_k + \omega_2 \gamma_{k-1} \quad (5)$$

[0085] 其中, γ_{k-1} 为集成分类器k-1的参数, θ_k 为基分类器k的参数, ω_1 和 ω_2 为集成分类器k-1和基分类器k的权重,当k=0时, $\gamma_0 = \theta_0$ 。

[0086] S34,查询集的标签更新如下:

$$[0087] \quad \hat{y}_j^{q*} = h_{\gamma_m} \left(g_{\varphi} \left(x_j^{q*} \right) \right), x_j^{q*} \in Q \quad (6)$$

[0088] 其中, \hat{y}_j^{q*} 为查询集的标签, h_{γ_m} 为第m个集成分类器的参数, g_{φ} 为提取特征模块, x_j^{q*} 为查询样数据的数据,其中 x_j^{q*} 属于有理数集。

[0089] S4、通过新类数据上所有任务中查询集标签的预测值和真实值,评估模型的最终性能。

[0090] 另一方面,本发明还提供了一种基于自训练的小样本图像集成分类装置,用以实现上述的任一项基于自训练的小样本图像集成分类方法,如图4所示,包括以下模块:

[0091] 预训练模块:使用基类数据上预训练得到特征提取模块;

[0092] 自训练模块:用于构建基分类器,并使用查询样本进行基分类器的自训练,不断地将伪标签集加到支持集扩充支持集;

[0093] 集成分类模块:用于自训练过程中产生的基分类器,不断通过模型平均得到集成分类器。

[0094] 经过自训练模块多次迭代自训练和集成分类模块分类器集成过程,得到最终的集成分类器。

[0095] 从上述技术方案可以了解到,本发明的基于查询样本自训练和模型平均的基于自训练的小样本图像集成分类装置具有以下有益效果:

[0096] 融合了查询样本自训练和模型平均的小样本集成分类装置,拟迁移基类数据的卷积特征提取,并基于查询样本自训练,通过构建损失函数,使得基分类器在查询样本上有确定的、不同的预测,实现了基分类器的不同,解决了现有的小样本分类方法中,基于特征迁移和查询样本自训练的小样本图像集成分类中,基学习器的多样性问题,对于提升图像的

分类效果非常显著,具有很高的使用价值。

[0097] 以上结合附图对所提出的基于自训练的小样本图像集成分类方法及模型的具体实施方式进行了阐述。通过以上实施方式的描述,所属领域的技术人员可以清楚的了解该方法以及装置的实施。

[0098] 在此提供的算法和显示不与任何特定计算机、虚拟系统或者其他设备固有相关。各种通用系统也可以与基于在此的启示一起使用。根据上面的描述,构造这类系统所要求的结构是显而易见的。此外,本文公开的也不针对任何特定的编程语言。但是应当了解,可以利用各种编程语言实现在此描述的本文公开的内容,并且上面对特定语言所做的描述是为了披露本文公开的最佳实施方式。

[0099] 类似的,应当理解,为了使本文尽量精简并且帮助理解各个公开方面中的一个或多个,在上面对本文公开的示例性实施例的描述中,本文公开的各个特征有时被一起分组到单个实施例、图、或者对其的描述中。然而,并不应将该公开的方法解释成反映如下示意图:即要求所保护的本文公开的要求比在每个权力要求中所明确记载的特征具有更多的特征。更确切地说,如下面的权力要求书所反映的那样,公开方面在于少于前面公开的单个实施例的所有特征。因此,遵循具体实施方式的权利要求书由此明确地并入该具体实施方式,其中每个权利要求本身都作为本公开的单独实施例子。

[0100] 以上所述实施例,仅为本申请的具体实施方式,用以说明本申请的技术方案,而非对其限制,本申请的保护范围并不局限于此,尽管参照前述实施例对本申请进行了详细的说明,本领域的普通技术人员应当理解:任何熟悉本技术领域的技术人员在本申请揭露的技术范围内,其依然可以对前述实施例所记载的技术方案进行修改或可轻易想到变化,或者对其中部分技术特殊进行等同替换;而这些修改、变化或者替换,并不使相应技术方案的本质脱离本申请实施例技术方案的精神和范围。都应涵盖在本申请的保护范围之内。因此,本申请的保护范围应所述以权利要求的保护范围为准。

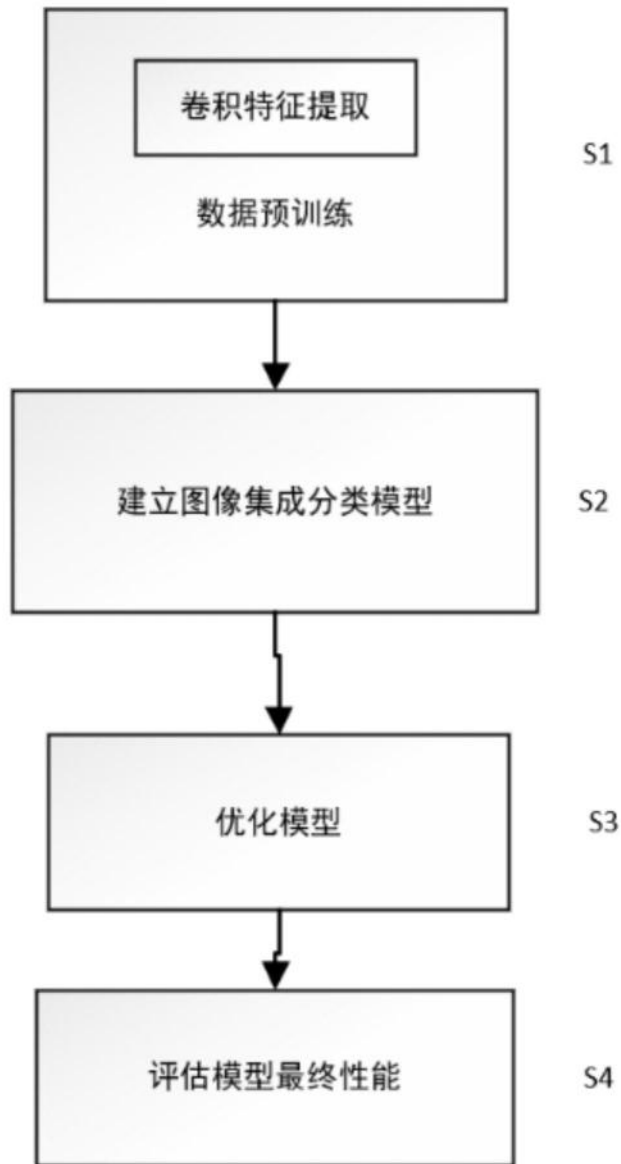


图1

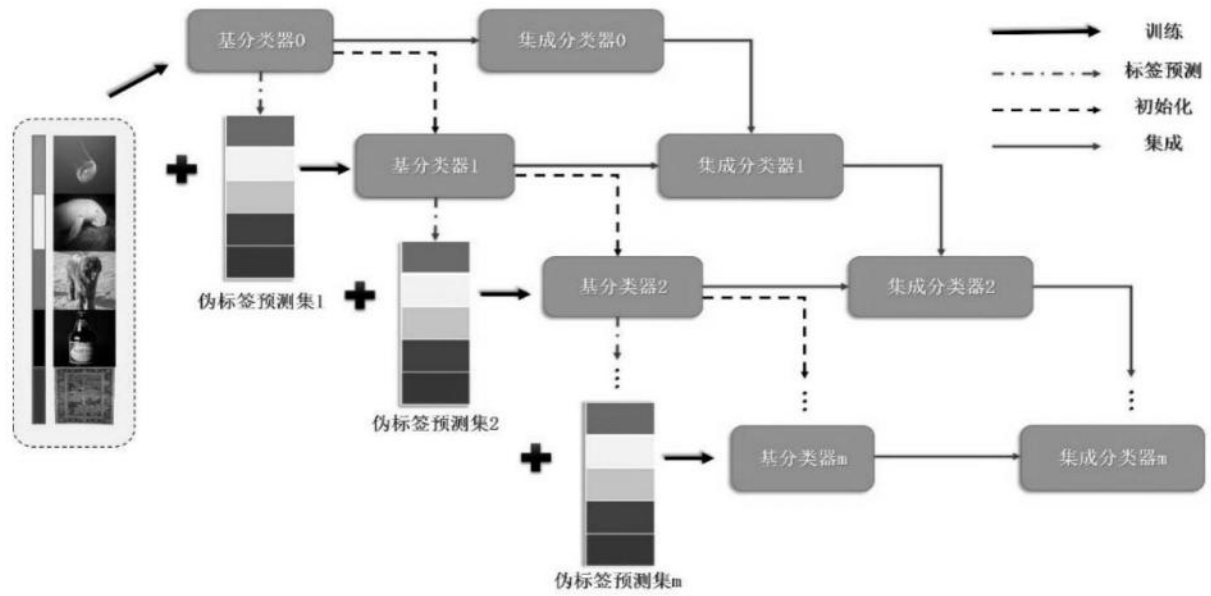


图2

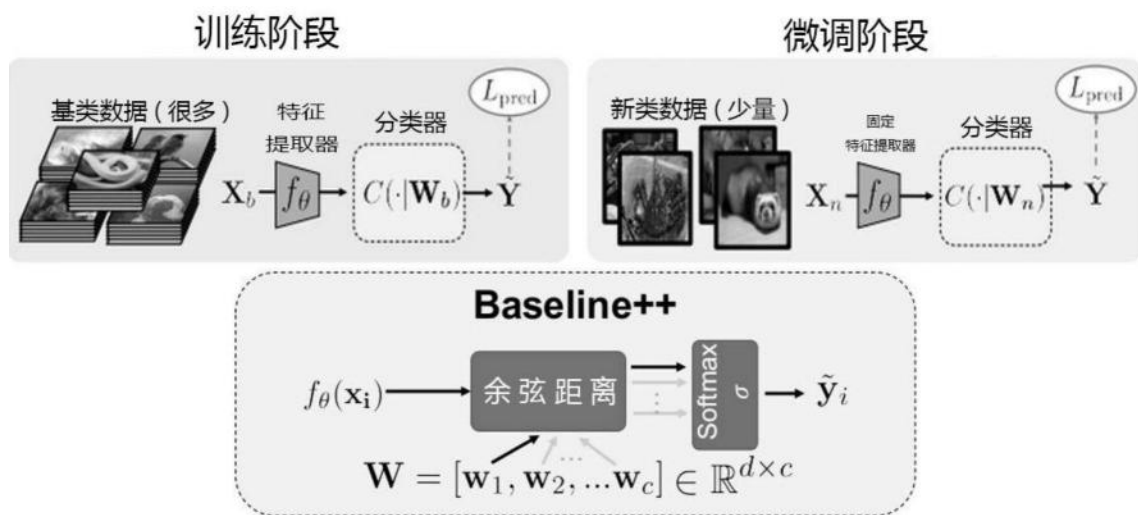


图3

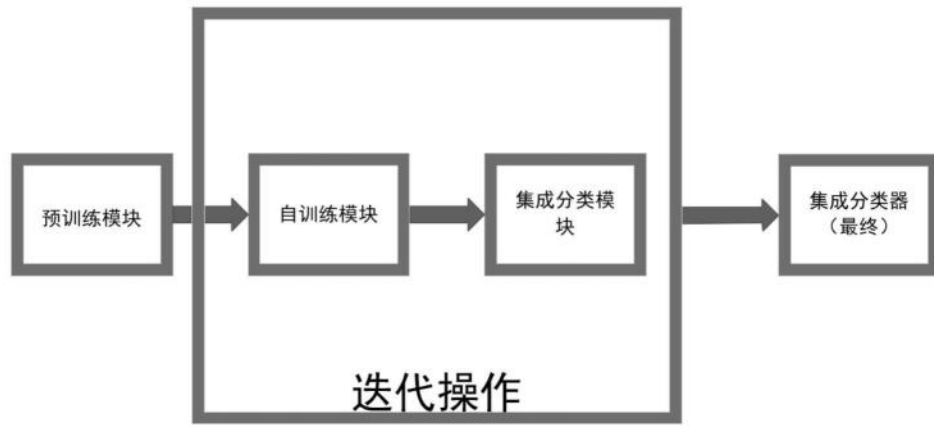


图4