计算机科学
COMPUTER SCIENCE

基于分层伪标签的图像聚类方法

蔡少填, 陈小军, 陈龙腾, 邱莉萍

引用本文

蔡少填, 陈小军, 陈龙腾, 邱莉萍.基于分层伪标签的图像聚类方法[J]. 计算机科学, 2023, 50(6): 225-235.

CAI Shaotian, CHEN Xiaojun, CHEN Longteng, QIU Liping. Stratified Pseudo-label Based Image Clustering [J]. Computer Science, 2023, 50(6): 225-235.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220900197

基于分层伪标签的图像聚类方法

蔡少填　陈小军　陈龙腾　邱莉萍
深圳大学计算机与软件学院　广东 深圳 518060
大数据系统计算技术国家工程实验室(深圳大学)　广东 深圳 518060
(cai.st@foxmail.com)

摘　要　图像聚类是图像处理中一个重要且开放的问题。最近,一些方法利用联合对比学习的良好表征能力来进行端到端聚类学习,利用伪标签技术来生成高质量的伪标签以提升聚类模型的鲁棒性。伪标签方法通常需要设置一个较大的概率阈值,并对满足要求的样本生成one-hot的标签,同时利用生成的标签来更新模型。但是,这种简单的伪标签生成方法难以获得足够数量的高质量伪标签。为了解决以上问题,提出了一种基于分层伪标签的图像聚类方法,它旨在利用结构化信息与伪标签信息对分类模型进行训练和精炼。引入3个假设来指导聚类方法的设计,包括局部平滑假设、自训练假设及低密度分离假设。新方法包含两个阶段:1)基于流形的一致性学习,利用近邻一致性学习来初始化聚类模型;2)基于分层伪标签的模型精炼,基于第一阶段的结果生成伪标签,并利用其来提升聚类模型的鲁棒性。首先,将基于第一阶段的结果生成强伪标签数据集及弱伪标签数据集;然后,提出了基于标签传播及分层混合的伪标签提升技术来提升弱伪标签数据集的质量;最后,同时利用强伪标签数据集及弱伪标签数据集来提升分类模型的泛化能力。相较于最优结果,SPC算法在STL10和Cifar100-20基准数据集上,ACC平均结果分别提升了7.6%和5.0%。

关键词:深度聚类;一致性学习;伪标签;标签传播;自训练学习
中图法分类号　TP391

Stratified Pseudo-label Based Image Clustering

CAI Shaotian,CHEN Xiaojun,CHEN Longteng and QIU Liping
College of Computer Science & Software Engineering,Shenzhen University,Shenzhen,Guangdong 518060,China
National Engineering Laboratory for Big Data System Computing Technology(Shenzhen University),Shenzhen 518060,Guangdong 518060,China

Abstract　Image clustering is an important and open problem in image processing. Recently,some methods combine the powerful representation ability of contrastive learning to carry out end-to-end clustering learning and utilize the pseudo-label technique to improve the robustness of clustering methods. In the existing pseudo-label methods,we need to set a large threshold parameter to obtain highly confident samples to generate one-hot pseudo-labels and often cannot obtain enough highly confident samples. To make up for these defects,we propose a stratified pseudo-label clustering(SPC) method,which aims to train and refine the classification model using both structure and pseudo-labels information. We first introduce three assumptions for designing of deep clustering methods,i.e.,local smoothing assumption,self-training assumption,and low-density separation assumption. The method consists of two stages:1)manifold based consistency learning,which is used to initialize the classification model in the training stage;and 2)stratified pseudo-label based model tefinement,which generates stratified pseudo-labels to improve the robustness of the clustering model. We first generate a strong pseudo-label dataset and a weak pseudo-label dataset with a threshold parameter,and then propose a label-propagation method and a mix-up method to improve the weak pseudo-label dataset. Finally,we use both strong pseudo-label dataset and weak pseudo-label dataset to refine the clustering model. Compared with the best baseline,the averaged ACC of SPC improves by 7.6% and 5.0% on STL10 and CIFAR100-20 benchmark datasets,respectively.

Keywords　Deep clustering,Consistency learning,Pseudo-labels,Label propagation,Self-training learning

1　引言

图像聚类是一项重要且具有挑战性的任务,其目的是在没有标签的条件下将图像划分成不同的图像簇。目前的深度聚类方法在这个领域已经展现出优越的性能。早期,深度聚类工作[1-6]通常是将自编码器(Auto-Encoders,AE)或卷积

到稿日期:2022-09-21　返修日期:2022-12-02
基金项目:深圳市基础研究面上项目(JCYJ20210324093000002)
This work was supported by the Shenzhen Research Foundation for Basic Research,China(JCYJ20210324093000002).
通信作者:陈小军(xjchen@szu.edu.cn)

神经网络(Convolutional Neural Network,CNN)与传统聚类算法相结合。然而,它们通常借助一些初始化或者后处理(如 k-means)来获得最终的聚类概率指示,并且适应于一些特殊的数据分布而非任意的数据分布,因此很难被应用到不同的数据中。

近年来,表征学习,尤其是对比学习[7-8],取得了很大的成功。一些方法[9-11]通过最大化图像及增强其间的互信息来捕获更具区分性的表征,并使用多层感知机(Multilayer Perceptron,MLP)来直接获得聚类结果。然而,实例级对比损失将其他所有样本都作为负样本,这可能会将相似的样本分开,从而破坏聚类结构。一些方法[11-12]首先通过前置任务[13-15]学习高质量的表征,接着最大化样本及其近邻之间的聚类指示概率的相似度,来获得最终的聚类结果。然而,在实际应用中近邻关系可能并不准确,因此此类方法严重依赖近邻的质量。同样,近期提出的聚类方法也利用伪标签技术来生成高质量的伪标签,以提升聚类模型的鲁棒性[16]。这类方法通常设置一个较大的概率阈值并对满足要求的样本生成 one-hot 的标签,最终利用生成的标签来更新模型。但是,这种简单的伪标签生成方法难以获得高质量的伪标签。如图 1 所示,根据训练阶段在 STL10 数据集上得到的聚类指示概率,采用不同的阈值对其截断可得到相应的高置信样本。结果表明获得的高置信样本即高质量伪标签数量有限。具体地,将阈值设置为 0.99 得到的高置信样本仅占训练集的 30%,这意味着只能利用数据集中的少量样本来对模型进行提升。

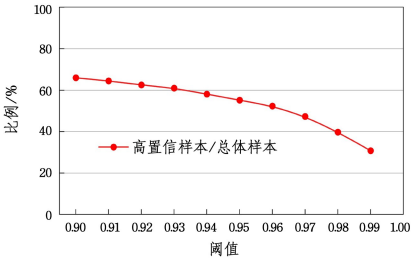


图 1 定义聚类概率指示大于阈值的样本为高置信样本

Fig. 1 Samples whose clustering probability indicates are greater than the threshold are defined as high confidence samples

为了解决以上问题,提出了一种新颖的图像聚类方法(见图 2)——基于分层伪标签的图像聚类方法(Stratified Pseudo-label Clustering,SPC)。首先,SPC 算法引入 3 个假设来指导聚类方法的设计,包括局部平滑假设、自训练假设及低密度分离假设。接着,SPC 算法利用两个阶段来学习聚类:1)基于流形的一致性学习,即构建一个静态的邻接图,通过近邻一致性来迫使相邻的样本具有相似的聚类指示概率;2)基于分层伪标签的模型精炼,即使用阈值截断的方法来将训练集划分为一个包含高置信度样本的强伪标签数据集及对应的强伪标签集,以及一个包含其他样本的弱伪标签数据集及对应的弱伪标签集。与只使用强伪标签数据集来提升模型性能的传统聚类方法[16]不一样,SPC 算法提出了一系列的方法来提升弱伪标签数据集的标签质量,最终同时使用两个数据集来对模型进行更新以提升其鲁棒性。在 Cifar-10,Cifar100-20 和 STL10 基准数据集上测试了 SPC 算法,并将其与现有的方法进行了对比。

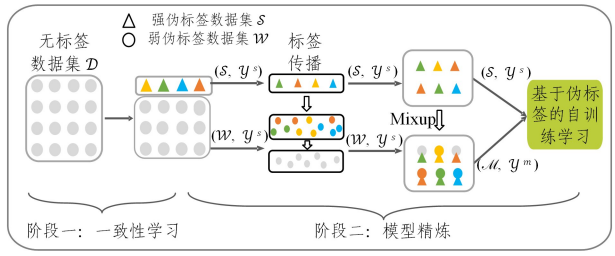


图 2 SPC 算法框架

Fig. 2 Framework of SPC

本文的主要贡献包括:

(1)提出了基于标签传播的弱伪标签修正方法。新方法计算样本之间的相关性矩阵,利用样本相关性并使用谱学习方法来获得更好的预测概率。对于弱伪标签样本,将利用学到的预测概率生成新的伪标签以对弱伪标签进行修正。由于有效利用了强伪标签及样本之间的关系,因此弱伪标签数据集的伪标签质量将得到有效提升。

(2)提出了基于分层数据混合的弱伪标签修正方法。新方法利用数据混合方法,即对每个弱伪标签样本引入一个强伪标签样本来进行线性组合,以获得混合训练样本。混合训练样本有效利用了强伪标签数据集及弱伪标签数据集,扩充了训练样本,并有助于提升模型的鲁棒性。

(3)对所提算法进行了相关实验。结果表明,SPC 算法的聚类结果超过了现有的图像聚类方法,特别是在 STL10 和 Cifar100-20 数据集上,平均准确度分别提升了 7.6% 和 5.0%。

2 相关工作

2.1 深度聚类

早期,深度聚类通常简单地将特征学习与浅层聚类相结合。例如,一些方法将堆叠式自动编码器(Stacked Autoencoder,SAE)或 CNN 与传统的聚类算法(如 k-means[1-3]、子空间聚类[17-18]和谱聚类[5,19])相结合。然而,上述方法通常需要用传统聚类算法进行初始化,或后处理,才能得到聚类指示概率,这限制了聚类的性能表现。

为了实现表征学习与聚类学习同时进行优化的目标,人们提出了“表征即标签”的思想,即图像表征直接映射为标签。例如,一些方法通过最大化原始与增强图像之间标签的互信息[10-11,20],或者最大化样本及其近邻样本之间聚类指示的似然估计[11-12],以端到端的形式将图像用分类模型直接映射为聚类指示概率并得到标签。然而,这种方法对初始化参数敏感或受劣质初始嵌入特征的影响,聚类性能往往不佳。此外,文献[10]中的实例级对比损失通常将其他所有样本作为负样本,这可能会将相似的样本推开并破坏语义簇结构。

为了弥补这些不足,一些学者提出以“训练-精炼(Train-and-Refinement)”的方式来进行聚类。在训练阶段,这些方法通常借助具备学习高质量特征能力的前置任务(Pretext Task)来进行特征学习[7-8];接着,通过最大化近邻样本之间聚类概率指示相似度的方式来训练分类模型[16,21-23]。在精炼阶段,通常选择聚类指示概率逼近 one-hot 向量的置信样本,

并用 one-hot 形式的伪标签对这些样本进行标记。最后，利用标记过的样本[16,22]来对分类模型进行精炼。文献[24]通过独立训练多个聚类算法来生成多组伪标签，将其中共同的伪标签设置为高质量的伪标签，但这存在计算量大以及匈牙利算法难以对多组伪标签有效对齐的问题。SPC 算法采取“训练-精炼”的方式，并提出分层伪标签学习，以此获得更多高质量的伪标签，进而提升对分类模型的精炼。此外，一些方法提供了插入附加模块的思路，通过标签清理和精炼标签的使用对分类模型重新进行训练[25-26]。

2.2 表征学习

图像表征学习的早期工作使用手工制作的特征，如尺度不变特征变换(Scale-invariant Feature Transform，SIFT)[27]和定向梯度直方图(Histogram of Oriented Gradient，HOG)[28]。后期，一些工作采用深度神经网络(如自动编码器[29]和对抗学习[30])来提取图像的有效特征。最近，随着自监督学习的兴起，表征学习得到了很大的发展。自监督学习旨在构建一系列前置任务，并利用图像的不同先验来探索数据样本的内在分布，这已成为一种越来越流行的深度图像表征学习方法。这些工作主要集中在前置任务上，基于图像变换前后标签不变形的特征，通过对输入图像应用变换(如图像着色[13]、计算视觉单元[31]和图像旋转[14])来学习良好的数据表达。具体来说，对比学习最近已成为图像处理学习方法的主要组成部分，通常优化对比损失(例如最大化不同视图之间的互信息[32]或动量对比学习[8])来拉近邻居和推开非邻居[33]。一些最新的对比学习方法通过结合数据增强策略(如 SimCLR[7]，BYOL[34]，PCL[35]和 HCSC[36])，在图像表征学习方面也取得了很大成功。以 SimCLR[7]为例，基于一致性正则化原则[37]，其将样本 x_i 以及它的增强样本 $T[x_i]$以如下方式进行对比学习。

$$\mathscr{L}_{\text{simclr}}=-\log\frac{\exp(sim(x_i,\mathcal{T}[x_i])/\tau)}{\sum_{j=1}^{2B}\mathbf{1}_{j\neq i}\exp(sim(x_i,\mathcal{T}[x_j])/\tau)} \tag{1}$$

其中，sim 是相似性函数，如余弦相似度；B 为批量大小；τ 为可调节的温度系数；$\mathbf{1}_{k\neq i}$是指示函数，当 $k\neq i$ 时其值为 1，否则为 0。对于 x_i，式(1)中将 $\mathcal{T}[x_i]$作为正样本，其余$(2B-1)$个样本为负样本。InfoNCE[38]中已证明负样本越多则互信息下界越小，在足够大的批量下，式(1)中损失最小化可达到表示学习的互信息最大化的目标。

2.3 伪标签技术

伪标签是已训练的分类模型对无标签数据进行概率预测并加以筛选的结果。它是一种常见的半监督分类模型训练方法，即通过联合标签样本和预测得到的伪标签来训练模型。在伪标签学习过程中，迫使模型对无标签样本进行预测，有利于最小化无标签样本预测的熵[39]，并将决策边界移动到低密度区域。目前常见的构造伪标签损失项的方法有两种：1)创建一个关于“样本-伪标签”的指定损失项[40-41]；2)将相同数量的有标签样本和伪标签样本合并为一个批量，共用标准的半监督分类损失[42-43]。Lee[44]首次将伪标签应用于深度学习，并将模型 $f(x)$的输出视为离散概率分布，把概率向量中最大元素指定的类设置为无标签样本的伪标签 $\hat{y}_i=\text{argmax}\ f(x_i)$，最后用如下损失函数对有标签样本和伪标签样本进行联合学习。

labelled loss unlabelled loss

$$\hat{\mathscr{L}}_{\text{ssl}}=\frac{1}{n_l}\sum_{i=1}^{n_l}\ell_s(f(x),y)+\lambda_u\frac{1}{n_u}\sum_{i=1}^{n_u}\ell_s(f(x),\tilde{y}) \tag{2}$$

其中，ℓ_s 表示某种损失函数，λ_u 是权重参数。

近年来，越来越多基于伪标签的半监督学习算法被提出，包括 MixMatch[45]，ReMixMatch[46]，FixMatch[40]，LaplaceNet[47]以及 FlexMatch[48]等优秀工作。这些算法很好地促进了伪标签学习技术的发展。

3 基于分层伪标签的图像聚类

本节将详细介绍 SPC 算法。如图 2 所示，SPC 算法主要分为两个阶段：1)基于流形的一致性学习(3.2 节)；2)基于标签传播的模型精炼(3.3 节)，包含分层伪标签生成、基于标签传播的弱伪标签提升、基于分层混合的弱伪标签提升以及分层伪标签的模型精炼过程。

为了使新的方法在技术上更合理，SPC 算法首先为本聚类工作引入 3 个假设，它们来自于一些半监督工作的启发[49-50]。

(1)局部平滑假设：如果两个样本位于低维流形中的一个局部领域，它们将具有相似的软聚类指示。

(2)自训练假设：如果一个数据集包含清晰分离的簇，那么簇中聚类指示概率接近于 1 的样本属于该簇。

(3)低密度分离假设：预测决策边界应位于低密度区域，而不是经过高密度区。

全文采用加粗大写字母如 $\boldsymbol{A}$ 来表示矩阵，$\boldsymbol{a}_i$ 表示矩阵 $\boldsymbol{A}$ 的第 i 行，a_{ij} 表示矩阵 $\boldsymbol{A}$ 第 i 行第 j 列所在位置的元素，$\boldsymbol{A}^{\mathrm{T}}$ 表示 $\boldsymbol{A}$ 的转置。

3.1 框架

假定 $\mathscr{D}=\{x_1,x_2,\cdots,x_n\}$为一个包含 n 张无标签图像的数据集。聚类的目标是将图像数据集 $\mathscr{D}$ 划分为 c 个清晰的簇，使得簇内图像相似，簇间图像不相似。令 $f:\mathscr{D}\rightarrow\mathbb{R}^d\times\mathbb{R}^c$ 表示一个映射图片为语义特征和聚类概率指示的函数。在 SPC 算法中，f 由一个参数为 θ 的主干网络 ResNet-18 和一个参数为 ϕ 的分类模型(MLP)组成，则语义特征矩阵可表示为 $\boldsymbol{Z}=f_Z(\mathscr{D};\theta)$，聚类概率指示矩阵可表示为 $\boldsymbol{Q}=f_Q(\mathscr{D};\theta,\phi)$。

如图 2 所示，SPC 算法第一阶段的目标是初始化分类模型，具体做法是基于流形的一致性学习训练分类模型，最后获得每张图像 x_i 对应的聚类概率指示 $\boldsymbol{q}_i$；第二阶段的目标是对分类模型进行精炼，以提升模型的鲁棒性。首先，设置阈值 δ 来对 $\boldsymbol{Q}$ 进行截断，从而得到两个数据集，即强伪标签数据集 $\mathscr{S}$ 及弱伪标签数据集 $\mathscr{W}$，其满足 $\mathscr{S}\cup\mathscr{W}=\mathscr{D}$，$\mathscr{S}\cap\mathscr{W}=\emptyset$。同时，强伪标签数据集 $\mathscr{S}$关联一个强伪标签集 $\mathscr{Y}^s$，弱伪标签数据集 $\mathscr{W}$ 关联一个弱伪标签集 $\mathscr{Y}^w$。由于这里 $\mathscr{Y}^w$的质量较低，我们提出基于标签传播及数据混合两种方法来利用 $\mathscr{Y}^s$及样本之间的相关性来提升 $\mathscr{Y}^w$。最终，同时利用$(\mathscr{S},\mathscr{Y}^s)$及$(\mathscr{W},\mathscr{Y}^w)$来对模型进行精炼。

3.2 基于流形的一致性学习

基于局部平滑假设，构造近邻样本对来指导分类模型的

训练，其中近邻样本的聚类指示概率在样本对中可被视为一种伪标签。为了与 SCAN[16] 进行公平的比较，SPC 算法在一致性学习阶段与其保持一致。给定样本 x_i，首先使用前置任务[7]进行预训练，以此保持原样本和增强样本在特征空间中的一致，使模型特征对数据增强转换具有不变性，从而获得高质量的语义特征 $\boldsymbol{z}_i$；接着，调用 Faiss 库[51]，计算得到最近的 k_1 个近邻样本集合 $\mathcal{N}^1(x_i)$；然后，随机选取 x_i 及其对应 k_1 个近邻样本中的一个 x_j，对它们使用强增强[52]，并通过最小化式(3)来强制 $\boldsymbol{q}_i$ 和 $\boldsymbol{q}_j$ 的一致性。

$$\begin{aligned}\mathcal{L}_c &= -\frac{1}{B}\sum_{i=1}^{B}\sum_{j=rn(\mathcal{N}^1(x_i))}\ell_{ce}(1,\boldsymbol{q}_i^{\mathrm{T}}\boldsymbol{q}_j)\\ &= -\frac{1}{B}\sum_{i=1}^{B}\sum_{j=rn(\mathcal{N}^1(x_i))}\log\boldsymbol{q}_i^{\mathrm{T}}\boldsymbol{q}_j\end{aligned}\tag{3}$$

其中，B 是批量的大小，$rn(.)$ 表示从给定集合中随机选择一个数据，ℓ_{ce} 是一个二分交叉熵函数。另外，与文献[11,16,53]一样，SPC 算法引入一个流形的正则项来联合优化模型，以防止陷入局部最优(如大多数样本聚到一个簇中)。具体公式如下所示：

$$\mathcal{L}_b=\sum_{l}^{c}\hat{\boldsymbol{q}}_l\log\hat{\boldsymbol{q}}_l\tag{4}$$

其中，$\hat{\boldsymbol{q}}_l=\frac{1}{B}\sum_{i}^{B}q_{il}$ 表示一个批量集合中 l 类的平均聚类指示。因此，第一阶段总的目标函数可以表示如下：

$$\mathcal{L}_{sl}=\mathcal{L}_c+\lambda_b\mathcal{L}_b\tag{5}$$

其中，λ_b 是控制正则项的权重系数。

3.3 基于分层伪标签模型精炼

为了提升模型的鲁棒性，本文将基于一致性学习得到的预测概率来生成伪标签，并基于自训练假设和低密度分离假设来对伪标签进行分层提升，最终利用所有的伪标签对模型进行精炼。

3.3.1 分层伪标签生成

给定数据集 $\mathcal{D}$，经过一致性学习阶段可以获得数据集的预测概率 $\boldsymbol{Q}$。由于此时的预测概率存在较大的不准确性，故设定一个阈值 δ 来对 $\boldsymbol{Q}$ 进行截断，从而得到两个数据集——强伪标签数据集 $\mathcal{S}$ 及弱伪标签数据集 $\mathcal{W}$，其满足 $\mathcal{S}\cup\mathcal{W}=\mathcal{D}$，$\mathcal{S}\cap\mathcal{W}=\varnothing$。给定一个样本 $x_i\in\mathcal{D}$，如果 $\max_l q_{il}>\delta$，则对 $\boldsymbol{q}_i$ 进行硬化以得到 one-hot 标签 $\boldsymbol{y}_i^s$，并将 x_i 及 $\boldsymbol{y}_i^s$ 分别加入 $\mathcal{S}$ 及 $\mathcal{Y}^s$ 中；否则将 x_i 及 $\boldsymbol{y}_i^w=\boldsymbol{q}_i$ 分别加入 $\mathcal{W}$ 及 $\mathcal{Y}^w$ 中。直觉来讲，强伪标签数据集 $\mathcal{S}$ 关联的强伪标签集 $\mathcal{Y}^s$ 比弱伪标签数据集 $\mathcal{W}$ 中样本关联的弱伪标签集 $\mathcal{Y}^w$ 更可靠。为了有效提升 $\mathcal{Y}^w$ 的质量并利用 $\mathcal{W}$ 来对模型进行提升，提出基于标签传播及数据混合的两种弱伪标签提升策略。

3.3.2 基于标签传播的弱伪标签提升

局部平滑假设假设标签具有一定的平滑性质。同时，样本之间的关系也将影响这种平滑性。为了对生成的伪标签进行提升，借鉴谱聚类算法[54]，提出基于标签传播的伪标签修正方法。具体地，根据语义特征空间 f_Z 中近邻节点之间特征的相似程度 s 计算一个相似性矩阵 $\boldsymbol{W}$，并以此构建一个稀疏的加权图；同时，$\boldsymbol{W}$ 将在每次迭代完成后进行更新，以更好地捕捉样本之间的相关性。$\boldsymbol{W}$ 的构造方法如下：

$$w_{ij}=\begin{cases}s(\boldsymbol{z}_i,\boldsymbol{z}_j), & j\in\mathcal{N}^2(x_i)\\ 0, & j\notin\mathcal{N}^2(x_i)\end{cases}\tag{6}$$

其中，$\mathcal{N}^2(x_i)$ 表示节点 x_i 的所有近邻节点集合，$s(,)$ 表示一个点积相似性函数。$\mathcal{N}^2(x_i)$ 通常可以使用 Faiss 库[51]计算得其近邻样本集合 $\mathcal{N}^2(x_i)$ 并保留 x_i 的最近的 k_2 个样本。

基于相似度矩阵 $\boldsymbol{W}$，可以得到度矩阵 $\boldsymbol{D}=\mathrm{diag}(\boldsymbol{W}\boldsymbol{1}_n)$，并进一步获得归一化矩阵 $\widetilde{\boldsymbol{W}}=\boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}$。为了提升弱伪标签集 $\mathcal{Y}^w$，根据谱学习算法[54]，可以使用 $\mathcal{Y}^s$ 并利用 $\widetilde{\boldsymbol{W}}$ 来对 $\mathcal{Y}^w$ 进行修正。具体地，提出如下目标函数来学习数据集 $\mathcal{D}$ 中样本的预测概率矩阵 $\boldsymbol{P}$。

$$\mathcal{J}(\boldsymbol{P})=\frac{1}{2}\sum_{i,j=1}^{n}\widetilde{w}_{ij}\left\|\frac{\boldsymbol{p}_i}{\sqrt{d_{ii}}}-\frac{\boldsymbol{p}_j}{\sqrt{d_{jj}}}\right\|^2+\frac{\mu}{2}\sum_{i=1}^{n}\|\boldsymbol{p}_i-\boldsymbol{Y}_i\|^2\tag{7}$$

其中，第一项根据相似度矩阵 $\boldsymbol{W}$ 强制近邻节点之间共享相似的标签，而第二项鼓励学到的预测概率 $\boldsymbol{P}$ 逼近初始的伪标签 $\boldsymbol{Y}$。μ 是平衡两项的系数。为了加速计算，将以上目标函数的优化转化为使用共轭梯度方法来求解线性方程[54]$(\boldsymbol{I}-\gamma\boldsymbol{W})\boldsymbol{P}=\boldsymbol{Y}$，其中 $\gamma=\frac{1}{1+\mu}$。最后，本算法保持强伪标签不变，对于任意的 $x_i\in\mathcal{W}$，根据 $\boldsymbol{P}$ 来更新其伪标签 y_i^w 为：

$$y_i^w=\text{one-hot}(\arg\max_j p_{ij})\tag{8}$$

其中，one-hot(j) 将得到第 j 个位置为 1 的硬标签。

3.3.3 基于分层混合的弱伪标签提升

受半监督学习启发，将 $\mathcal{S}$ 近似看成有标签数据集，将 $\mathcal{W}$ 视为无标签数据集，并使用 Mixup 算法[55]对强、弱伪标签数据进行混合，再应用半监督目标损失对分类模型进行精炼。使用 Mixup 数据增强方法，可以根据 $\mathcal{S}$ 及 $\mathcal{W}$ 获得线性组合后的新样本来提高模型的鲁棒性。对 $\mathcal{S}$ 中的每个样本进行弱增强得到 $\hat{\mathcal{S}}$，同时对 $\mathcal{W}$ 中的每个样本进行强增强得到 $\hat{\mathcal{W}}$。随后，将 $\hat{\mathcal{S}}$ 和 $\hat{\mathcal{W}}$ 组合，生成的批数据将作为分层混合的数据源。

分层混合的具体操作如下，对任意的 $\hat{u}_i\in\hat{\mathcal{W}}$，从 $\hat{\mathcal{S}}$ 中取一个样本 $\hat{x}_i\in\hat{\mathcal{S}}$：

$$\begin{aligned}x_i' &= \lambda'\hat{x}_i+(1-\lambda')\hat{u}_i\\ y_i' &= \lambda'\hat{y}_i^s+(1-\lambda')\hat{y}_i^w\end{aligned}\tag{9}$$

其中，λ' 是两个样本混合程度的平衡系数，计算如下：

$$\lambda\sim Beta(\alpha,\alpha)\tag{10}$$

其中，$Beta$ 表示一个连续概率分布函数，α 为该分布的系数。

$$\lambda'=\max(\lambda,1-\lambda)\tag{11}$$

这里对 λ 平衡参数进行调整，使带有强伪标签的数据获得较大的混合比重，以更好地监督模型的训练。

经过上述分层混合操作之后，可以得到新的混合数据集 $\mathcal{M}$ 及混合标签集 $\mathcal{Y}^m$。利用这种混合数据来对模型进行更新，将有利于提升模型的泛化能力。

3.3.4 基于分层伪标签的模型精炼

给定强伪标签数据集及其伪标签集 $(\mathcal{S},\mathcal{Y}^s)$，以及混合数据集及其混合标签集 $(\mathcal{M},\mathcal{Y}^m)$，下一步的目标是通过自训练学习来精炼聚类模型 ϕ，以进一步提升模型的性能。

在每个 mini-batch 操作中,将针对 $(\mathscr{S},\mathscr{Y}^s)$ 设计如下的损失函数来进行学习:

$$\mathscr{L}_S=\frac{1}{B}\sum_{(x',y')\in(\mathscr{S},\mathscr{Y}^s)}\ell_{ce}(y',f_\vartheta(x';\theta,\phi)) \tag{12}$$

针对 $(\mathscr{M},\mathscr{Y}^m)$ 设计如下的损失函数来进行学习:

$$\mathscr{L}_{\mathscr{M}}=\frac{1}{B}\sum_{(x',y')\in(\mathscr{M},\mathscr{Y}^m)}\ell_{ce}(y',f_\vartheta(x';\theta,\phi)) \tag{13}$$

最终在每个 mini-batch 操作中的精炼损失函数为:

$$\mathscr{L}_{plsl}=\mathscr{L}_{\mathscr{I}}+\lambda_{\mathscr{M}}\mathscr{L}_{\mathscr{M}} \tag{14}$$

其中,$\ell_{ce}(,)$表示交叉熵损失函数,$\lambda_{\mathscr{M}}$是平衡混合伪标签损失部分的权重系数。

本文提出的基于分层伪标签的图像聚类方法(SPC)的流程如算法 1 所示。

算法 1 SPC 算法

输入:训练集 $\mathscr{D}$,簇数量 c,前置任务 τ,主干网络和分类模型 $f_{\mathscr{Z}}(;\theta)$ 和 $f_\vartheta(;\theta,\phi)$,近邻集合 $\mathscr{N}_D^1=\{\}$ 和 $\mathscr{N}_D^2=\{\}$,置信度阈值 δ

输出:$f_\vartheta(\mathscr{D};\theta,\phi)$ //$\mathscr{D}$ 划分为 c 个簇

1. 基于前置任务 τ 优化 $f_{\mathscr{Z}}(;\theta)$ //预训练
2. for $x_i\in D$ do
3. $\mathscr{N}_D^1\leftarrow\mathscr{N}_D^1\cup\mathscr{N}^1(x_i)$,$\mathscr{N}^1(x_i)$包含 x_i 在预训练空间 f_θ 中的 k_1 个近邻样本。
4. end for
5. while $\mathscr{L}_{sl}$下降 do //训练阶段
6. 用式(5)更新 $f_{\mathscr{Z}}(;\theta)$和 $f_\vartheta(;\theta,\phi)$。
7. end while
8. while $\mathscr{L}_{plsl}$下降 do //精炼阶段
9. 生成强伪标签集 $\mathscr{Y}^s$;
10. 更新 $\mathscr{N}_D^2$,利用标签传播方法得到新的弱伪标签集 $\mathscr{Y}^w$;
11. 利用分层混合方法得到混合数据集 $(\mathscr{M},\mathscr{Y}^m)$;
12. 用式(14)更新 $f_{\mathscr{Z}}(;\theta)$和 $f_\vartheta(;\theta,\phi)$。
13. end while

3.4 复杂度分析

3.4.1 空间复杂度

在训练和精炼阶段中,存储近邻矩阵的空间复杂度分别为 $O(k_1N)$和 $O(k_2N)$。式(7)中伪标签矩阵 $\boldsymbol{Y}$ 的空间复杂度为 $O(cN)$。其中,c,k_1 和 k_2 均为远小于 N 的常数。因此,SPC 算法的总体空间复杂度为 $O(N)$。

3.4.2 时间复杂度

SPC 算法的时间复杂度取决于以下几个方面:1)计算相似度矩阵,其时间复杂度为 $O(N^2d)$;2)knn 搜索算法,调用 Faiss[51] 库中简单的排序搜索,搜索第 i 样本的 top k 时间复杂度为 $O(k\log N)$,因此训练和精炼阶段总的时间复杂度为 $O((k_1+k_2)N\log N)$;3)一致性学习,在训练阶段中,式(3)和式(4)的时间复杂度分别为 $O(T_1Nc)$和 $O(T_1(N/B)c)$,其中 T_1 是整个数据集的训练次数;4)分层伪标签模型精炼,式(7)利用共轭梯度法求解的时间复杂度为 $O(tNk_2c)$(t 为迭代次数),式(8)更新伪标签的时间复杂度为 $O(Nc)$,式(14)求解交叉熵函数的时间复杂度为 $O(Nc)$(其中 T_2 是整个数据集的训练次数,d,k_1,k_2,c,t,T_1 以及 T_2 等均为远小于 N 的常数)。除此以外,算法训练的时间复杂度也与网络模型的规模大小有关[56],这里假设网络模型总的参数规模为 E,则样本经过网络的时间复杂度可以近似为 $O(NE)$。因此,SPC 总的时间复杂度为 $O(N)$。

4 实验及结果分析

4.1 实验设置

4.1.1 实验数据集

将 SPC 算法在 Cifar10[57],Cifar100-20[57] 和 STL10[58] 这 3 个不同的基准数据集上进行了评测。基准数据集特征信息如表 1 所列,在后续实验中均使用原始的图像尺寸。

表 1 基准数据集的主要特征

Table 1 Characteristics of benchmark datasets

数据集名称	图像大小	# 训练集	# 测试集	# 类数量
STL10	96×96	5000	8000	10
Cifar10	32×32	50000	10000	10
Cifar100-20	32×32	50000	10000	20

4.1.2 实现细节

为公平起见,SPC 算法中图像的预训练特征学习和 SCAN[16] 方法的实验设置保持一致,在 Cifar10,Cifar100-20 和 STL10 数据集上均采用 SimCLR[7] 方法作为实例判别任务;与 SCAN[16] 一致,使用 ResNet-18[59] 作为主干网络,并在训练阶段和精炼阶段学习同一个分类模型。该分类模型由一个大小为 $d\times c$ 的全连接层组成,其中 d 和 c 分别表示预训练特征的维度和类别的数量。在训练阶段中,Cifar10,Cifar100-20 和 STL10 数据集训练所需的 Epoch 次数均设置为 100,批量大小设置为 128,近邻数量 k_1 为 20,正则项权重系数 λ_b 设置为 5,采用 Adam 优化器进行迭代优化,学习率和权重衰减分别设置为 0.0001 和 0.0001。在模型精炼阶段,对于 Cifar10,Cifar100-20 和 STL10 数据集,Epoch 次数均设置为 500。强伪标签数据集 $\mathscr{S}$ 和弱伪标签数据集 $\mathscr{W}$ 的批量大小分别设置为 1024 和 256,1024 和 256,256 和 64。置信度阈值 δ 分别设置为 0.99,0.98 和 0.95,标签传播中近邻数量 k_2 分别设置为 20,20 和 5。Beta 分布的系数 α 设置为 1。采用 SGD 优化器进行迭代优化,学习率为 0.03,权重衰减为 0.0005,nesterov 动量为 0.9。两个阶段均使用 Faiss[51] 库来搜索最近邻样本。

4.1.3 数据增强实现

如表 2 所列,SPC 算法中所用到的数据增强方式可以划分为强增强和弱增强两种策略。弱增强中使用随机水平翻转、随机裁剪以及 RandAugment[52] 的方式,尽量不改变图像的原始形式。而强增强策略在弱增强基础上结合了 Cutout[60] 生成增强样本。对于标签传播后生成的强伪标签数据集 $\mathscr{S}$ 和弱伪标签数据集 $\mathscr{W}$,本文结合其生成伪标签的可靠性分别选择了弱数据增强和强数据增强。

表 2 弱数据增强和强数据增强策略

Table 2 Strategies of weak augmentation and strong augmentation

弱数据增强	强数据增强
随机水平翻转+随机裁剪	
RandAugment×1	RandAugment×4 Cutout
正则化	

4.1.4 验证

基于聚类准确性(ACC)、归一化互信息(NMI)[61]和调整后的 rand 指数(ARI)[62]进行评估。训练和测试数据集中的图像分别进行训练和模型测试,即在训练阶段屏蔽测试集图像。最后,取 5 次不同运行结果的平均值和标准差作为最终实验结果,且在训练过程中所有数据集均使用相同的主干网络、增强方法和前置任务。

4.2 对比实验

在训练数据集与测试数据集分隔(训练数据集只用于训练,测试数据集只用于测试)的情况下,将提出的 SPC 算法在所列举的 3 种数据集上进行评测。将训练数据集与测试数据集分隔开,有利于研究算法对未见数据的泛化能力。将 SPC 算法与以下 22 种最先进的聚类方法进行了比较,包括 k-means[63], SC[64], NMF[65], JULE[6], SAE[66], DAE[67], AE[68], DCGAN[69], VAE[70], DEC[1], ADC[71], DeepCluster[72], DAC[23], DDC[73], DCCM[22], IIC[9], PICA[53], GCC[21], CC[10], SCAN[16], NNM[12]和 CRLR[11]。对于 SC, NMF, AE, DAE, DCGAN 和 VAE 这些方法,聚类结果是从图像中提取特征后进行 k-means 聚类得到。

表 3 列出了 SPC 算法和最先进的方法在 3 种数据集上的聚类结果。在后 7 行采用分隔数据集单独进行训练和评测的方法中,SPC 算法均超过了对比方法的最优结果。具体地说,在 STL10, Cifar10 和 Cifar100-20 数据集上,对于 ACC, NMI 和 ARI, SPC 算法的平均结果相较于方法 SCAN[16]的精炼阶段提升了 7.6%, 6.5%, 10.0%, 0.2%, 5.0%, 5.1%和 5.9%。前 19 行的算法均采用整体数据集同时进行训练和测试的形式,SPC 算法基本上都超过了对比方法。在 STL10, Cifar10 和 Cifar100-20 数据集的 ACC, NMI 和 ARI 指标中,即使在使用分隔数据集的情况下,相较于最优结果,SPC 的最优结果仍分别提升了 0.1%, 1.5%, 3.5%, 2.8%, 4.5%, 4.7%和 5.5%。这验证了 SPC 算法的有效性。

表 3 3 个基准数据集的最优结果比较

Table 3 Comparison of optimal results on three benchmark datasets

(单位:%)

Dataset	STL10			Cifar10			Cifar100-20		
Metrics	*ACC*	*NMI*	*ARI*	*ACC*	*NMI*	*ARI*	*ACC*	*NMI*	*ARI*
k-means[63]	19.2	12.5	6.1	22.9	8.7	4.9	13.0	8.4	2.8
SC[64]	15.9	9.8	4.8	24.7	10.3	8.5	13.6	9.0	2.2
NMF[65]	18.0	9.6	4.6	19.0	8.1	3.4	11.8	7.9	2.6
JULE[6]	27.7	18.2	16.4	27.2	19.2	13.8	13.7	10.3	3.3
SAE[66]	32.0	25.2	16.1	29.7	24.7	15.6	15.7	10.9	4.4
DAE[67]	30.2	22.4	15.2	29.7	25.1	16.3	15.1	11.1	4.6
AE[68]	30.3	25.0	16.1	31.4	23.4	16.9	16.5	10.0	4.7
DCGAN[69]	29.8	21.0	13.9	31.5	26.5	17.6	15.1	12.0	4.5
VAE[70]	28.2	20.0	14.6	29.1	24.5	16.7	15.2	10.8	4.0
DEC[1]	35.9	27.6	18.6	30.1	25.7	16.1	18.5	13.6	5.0
ADC[71]	53.0	—	—	32.5	—	—	16.0	—	—
DeepCluster[72]	33.4	—	—	37.4	—	—	18.9	—	—
DAC[23]	47.0	36.6	25.6	52.2	40.0	30.1	23.8	18.5	8.8
DDC[73]	48.9	37.1	26.7	52.4	42.4	32.9	—	—	—
DCCM[22]	48.2	37.6	26.2	62.3	49.6	40.8	32.7	28.5	17.3
IIC[9]	59.6	49.6	39.7	61.7	51.1	41.1	25.7	22.5	11.7
PICA[53]	71.3	61.1	53.1	69.6	59.1	51.2	33.7	31.0	17.1
GCC[21]	78.8	68.4	63.1	85.6	76.4	72.8	47.2	47.2	30.5
CC[10]	85.0	76.4	72.6	79.0	70.5	63.7	42.9	43.1	26.6
SCAN*(Avg±Std)	75.5±2.0	65.4±1.2	59.0±1.6	81.8±0.3	71.2±0.4	66.5±0.4	42.2±3.0	44.1±1.0	26.7±1.3
SCAN†(Avg±Std)	76.7±1.9	68.0±1.2	61.6±1.8	87.6±0.4	78.7±0.5	75.8±0.7	45.9±2.7	46.8±1.3	30.1±2.1
SCAN†[16](Best)	80.9	69.8	64.6	**88.3**	79.7	**77.2**	50.7	48.6	33.3
NNM[12]	76.8±1.2	66.3±1.3	59.6±1.5	83.7±0.3	73.7±0.5	69.4±0.6	45.9±0.2	48.0±0.4	30.2±0.4
CRLC[21]	78.7±1.1	68.4±1.7	62.7±1.8	84.2±0.1	74.7±0.3	70.6±0.5	45.0±0.7	44.8±0.8	28.7±0.9
SPC†(Avg±Std)	84.3±0.8	73.5±0.9	70.1±1.3	87.0±0.1	79.6±0.3	75.4±0.2	50.9±1.1	50.9±1.1	35.1±1.1
SPC†(Best)	**85.1**	**74.5**	**71.6**	87.1	**79.9**	75.6	**51.7**	**51.9**	**36.0**

注:表中数据包括 5 次不同运行结果的平均值和标准差以及最优的结果;前 19 行表示方法在整体数据集上同时进行训练和评价,后 7 行表示方法在分割的数据集上分别进行训练和评价;最优的结果采用黑色加粗表示

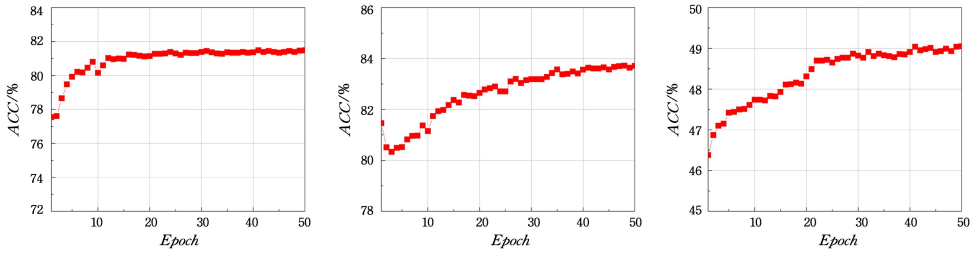
4.3 消融分析

为了验证各个损失项对聚类性能的影响,分别在 STL10, Cifar10 和 Cifar100-20 这 3 个数据集上进行了消融实验。具体来说,为了能够体现每个损失项带来的性能增益,以损失项 $\mathscr{L}_c$ 构造的退化模型作为基线,然后引入 $\mathscr{L}_b$ 进行正则约束,最后在此基础上加入 $\mathscr{L}_{plsl}$ 验证伪标签监督学习的作用。如表 2 所列,$\mathscr{L}_b$ 的引入能够控制样本均匀分布在所有簇中,在 STL10 数据集上所提模型性能比基线提高了 50.9%(*ACC*), 30.1%(*NMI*), 41.7%(*ARI*),这表明 $\mathscr{L}_b$ 有效防止了模型陷入平凡解。而在 Cifar10 数据集上,SPC 算法比基线提高了 71.7%(*ACC*), 66.6%(*NMI*)和 71.3%(*ARI*),在 Cifar100-20 数据集上比基线提高了 38.6%(*ACC*), 28.4%(*NMI*)和 45%(*ARI*),这表明 $\mathscr{L}_b$ 有效避免了退化解的出现。最后,引入 $\mathscr{L}_{plsl}$ 给模型在 STL10 数据集上带来了 7.5%(*ACC*), 6.8%(*NMI*)和 9.3%(*ARI*)的提升,在 Cifar10 数据集上提升了 5.3%(*ACC*), 8.3%(*NMI*)和 8.8%(*ARI*),在

Cifar100-20 数据集上提升了 7.3%（*ACC*），5.9%（*NMI*）和 6.7%（*ARI*），这个结果说明了分层伪标签监督学习的有效性，增加的高质量伪标签数量，提升了模型输出置信预测的能力。

4.4 伪标签质量评估

为了验证伪标签质量的有效性，SPC 算法分别在 Cifar10，Cifar100-20 和 STL10 数据集上对标签传播后生成的弱伪标签集 y^w 的准确度进行了评估，如图 3 所示。在标签传播开始之前，对 Cifar10，Cifar100-20 和 STL10 数据集中 $\mathcal{W}$ 部分数据的聚类指示概率 $f_2(\mathcal{W};\theta,\phi)$ 进行评测，得到相应的准确率，即 81.3%，44.3% 和 77.2%。随着训练 Epoch 次数的增加，通过标签传播产生的弱伪标签集 y^w 的准确度不断上升，并在大约 50 个 Epoch 之后，y^w 的准确度趋于稳定，在 Cifar-10，Cifar100-20 和 STL10 数据集上分别提升了 2.8%，4.9% 和 4.2%。



(a)STL10 (b)Cifar10 (c)Cifar100-20

图 4 STL10，Cifar10 以及 Cifar100-20 数据集中 $\mathcal{W}$ 对应数据的弱伪标签 $\tilde{y}^{\mathcal{W}}$ 正确率随标签传播的变化

Fig. 4 Change of the weak pseudo-label $\tilde{y}^{\mathcal{W}}$ accuracy of corresponding data of $\mathcal{W}$ o STL10，Cifar10 and Cifar100-20 datasets with label propagation

4.5 参数敏感性分析

4.5.1 近邻数量 k_1

在第一阶段中，基于流形的一致性学习涉及使用构造的近邻关系进行一致性学习，因此通过设置不同的近邻数量观察其对 STL10 数据集聚类结果的影响。将 k_1 的取值设置为 0，5，10，20，30，50，其中 $k_1=0$ 表示只有样本和它的增强样本之间进行一致性学习。从图 4 可以看出，在 $k_1=20$ 时聚类效果最好；随着 k_1 值的继续增大，近邻样本中噪声样本也逐渐增多，因此当 $k_1>20$ 时聚类效果趋于稳定。

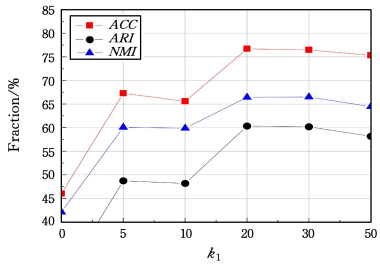


图 4 STL10 数据集训练阶段中近邻数量 k_1 的敏感性分析

Fig. 4 Sensitivity analysis of k_1 on STL10 dataset in training stage

4.5.2 置信度阈值 δ

阈值 δ 控制了初步筛选出的伪标签的置信度，因此设定合适的阈值有助于获得足够置信的伪标签。以 STL10 数据集为例，通过设置不同的参数值验证 δ 对伪标签的影响，其中 $\delta\in\{0.8,0.83,0.86,\cdots,0.98,0.985,0.99\}$。从图 5 中不同阈值对应的聚类结果可以得出，当 $\delta\leqslant 0.98$ 时整体呈现波动式增长态势，当 $\delta=0.98$ 时获得最佳性能；当 $\delta>0.98$ 时性能下降，原因是置信度阈值过高导致了伪标签数量大幅减少，从而致使后续伪标签监督学习的过程缺乏足够的样本，进而影响了聚类结果。此外，由于不同数据集的数据特征与第一阶段的聚类效果不同，因此将 Cifar10，Cifar100-20 和 STL10 这 3 个数据集上的置信度阈值 δ 分别设置为 0.99，0.95 和 0.98。

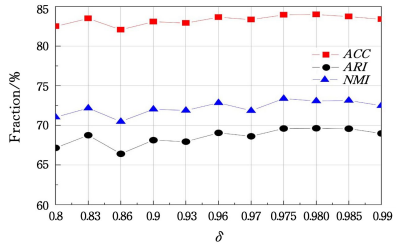


图 5 STL10 数据集精炼阶段阈值 δ 的敏感性分析

Fig. 5 Sensitivity analysis of δ on STL10 dataset in refining stage

表 4 SPC 算法在 3 个数据集上的消融分析

Table 4 Ablation analysis of SPC on three datasets

（单位：%）

Dataset	STL10			Cifar10			Cifar100-20		
Metrics	*ACC*	*NMI*	*ARI*	*ACC*	*NMI*	*ARI*	*ACC*	*NMI*	*ARI*
$\mathcal{L}_c$	25.9±5.2	36.6±6.2	19.1±3.7	10.0	0	0	5.0	0	0
$\mathcal{L}_c+\mathcal{L}_b$	76.8±1.1	66.7±0.4	60.8±0.8	81.7±0.3	71.3±0.3	66.6±0.4	43.6±2.7	45.0±0.8	28.4±1.5
$\mathcal{L}_c+\mathcal{L}_b+\mathcal{L}_{plsl}$	**84.3**±0.8	**73.5**±0.9	**70.1**±1.3	**87.0**±0.1	**79.6**±0.3	**75.4**±0.2	**50.9**±1.1	**50.9**±1.1	**35.1**±1.1

4.5.3 近邻数量 k_2

为了从已标记的强伪标签节点中预测未标记节点的标签，根据语义特征空间近邻节点之间的相似程度进一步构建相似度矩阵。参数值 k_2 的设置对相似度矩阵的构建有直接影响，因此设置了相关实验来观察模型对不同 k_2 值的敏感性，其中 k_2 设置为 1，5，…，80，100。STL10 在模型精炼阶段

取得的效果如图 6 所示，在 $k_2=5$ 附近取得了最优聚类性能，原因是 STL10 数据集的训练集样本较少。同理，Cifar10 和 Cifar100-20 的最佳近邻数量 k_2 为 20。

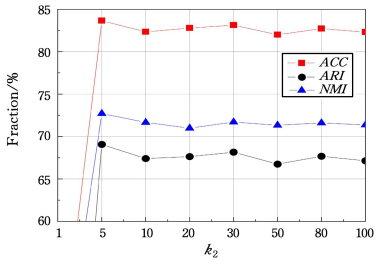


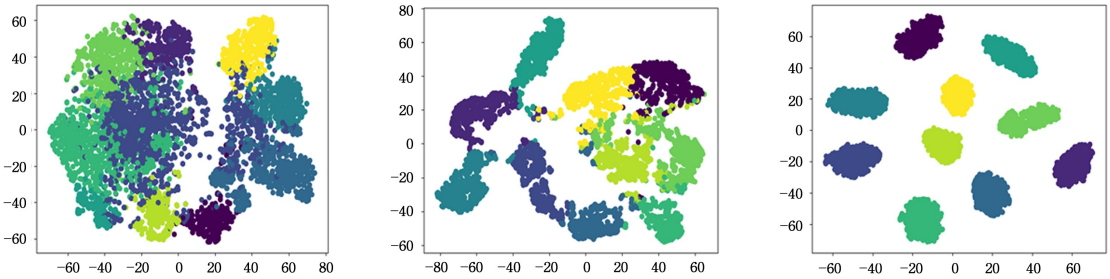
图 6 STL10 精炼阶段近邻数量 k_2 的敏感性分析

Fig. 6 Sensitivity analysis of k_2 on STL10 dataset in refining stage

4.5.4 可视化分析

(1)簇结构的演变。为了解 SPC 模型收敛到目标的过程，分别在预训练、基于流形的一致性阶段和最后的模型精炼阶段，利用 T-SNE 方法来对聚类结果进行可视化。结果如图 7 所示，其中不同的颜色表示特征预测的不同标签。结果表明，预训练特征分散，而一致性学习的加入，使得每个样本与其最近邻的类别差异尽可能小，因而此时聚集更加明显。模型精炼阶段进行标签传播之后，将伪标签数据集和弱伪标签数据集进行自训练学习，如图 7(c)所示。可以看到，随着阶段的不断进行，聚集效果更加明显，聚类分配更加合理。

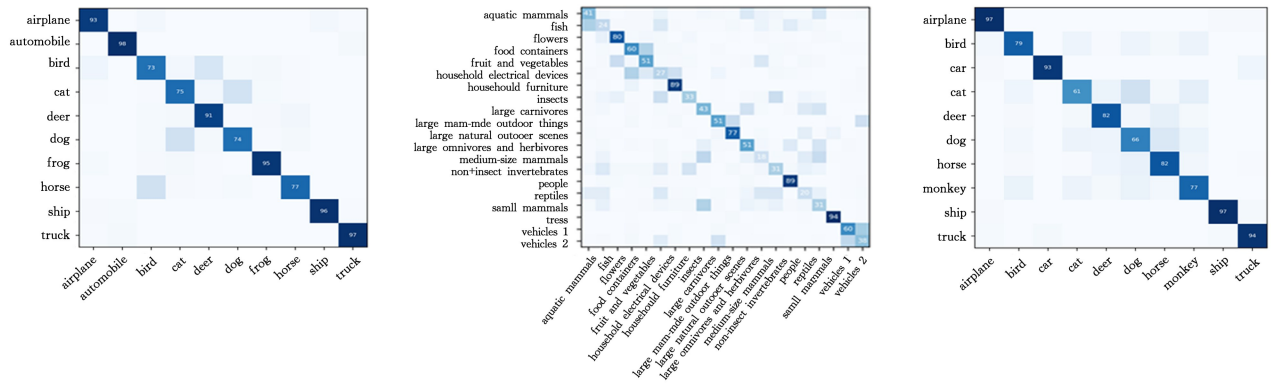
(2)混淆矩阵。图 8 给出了 SPC 算法分别在 Cifar10，STL10 和 Cifar100-20 数据集上得到的混淆矩阵。混淆矩阵具有块对角结构。能观察到，大多数错误示例都可以在难以聚类的类之间找到(如 Cifar10 和 STL10 数据集中的“猫”和“狗”)。由于超类的模糊性，Cifar100-20 数据集上的结果并不好。可以使用更深的 CNN 或更大的 Epoch 来进一步提升 SPC 的性能。



(a) Pre-trained (b) Stage1 (c) Stage2

图 7 STL10 数据集预训练特征以及第一、二阶段后特征在 T-SNE 上的可视化效果(电子版为彩图)

Fig. 7 T-SNE visualization of pre-trained features, stage1 features and stage2 features on STL10 datasets



(a) Cifar10 (b) Cifar100-20 (c) STL10

图 8 Cifar-10，Cifar100-20 以及 STL10 数据集的混淆矩阵

Fig. 8 Confusion matrices of Cifar10, Cifar100-20 and STL10

4.5.5 训练及精炼过程分析

为了验证所提的方法的收敛性，在 STL10 数据集的训练阶段运行 100 个 Epoch，在精炼阶段运行 300 个 Epoch，并对整个过程的 ACC 及 Loss 值进行记录。训练阶段结果如图 9 所示，模型在训练的前 20 个 Epoch 的准确度快速提升，对应的 Loss 值迅速下降，在 20 个 Epoch 后曲线趋于平缓。如图 10 所示，精炼阶段中 ACC 与 Loss 值的变化曲线呈现出明显的负相关，模型在 150 个 Epoch 后曲线趋于平缓，经历 260 个 Epoch 后几乎达到收敛。

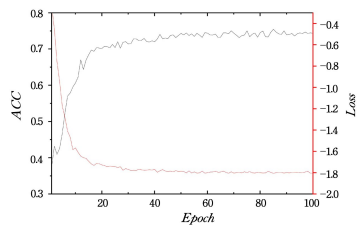


图 9 STL10 数据集训练阶段 ACC 和 Loss 值随 Epoch 的变化

Fig. 9 Changes of ACC and Loss values with Epoch in the training stage on STL10 dataset

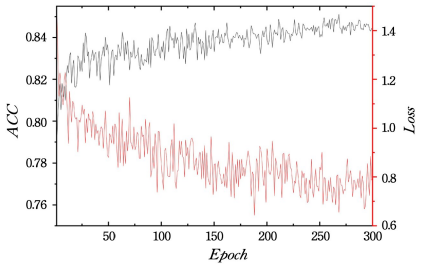


图 10 STL10 数据集精炼阶段 ACC 和 Loss 值随 Epoch 的变化

Fig. 10 Changes of ACC and Loss values with Epoch in the refining stage on STL10 dataset

结束语 本文基于分层伪标签提出了一种新的图像聚类方法。与其他简单依赖于表征学习或通过阈值截断来获取伪标签的方法不同，SPC 算法更关注结构化信息和伪标签信息来提高分类模型的泛化能力。过度依赖对比表征学习不利于学到更好的聚类语义结构，如何提高对伪标签信息的利用在聚类学习上也较少研究。因此，提出基于分层伪标签的模型精炼方法，将训练集分为强伪标签数据集及弱伪标签数据集，并有效地利用强伪标签数据集及样本之间的相关性来改进弱伪标签集质量。最终，同时利用强伪标签数据集及弱伪标签数据集来提升模型。在 3 个基准数据集上的实验结果表明，SPC 算法能够在大部分情况下获得更优的聚类性能。

在未来的研究工作中，将进一步研究结构化信息，包括邻接图中边的权重关系表示，以及标签传播过程如何调节不同质量样本的传播能力，并改进分层伪标签方法，以此来提高 SPC 算法的鲁棒性。

参 考 文 献

[1] XIE J，GIRSHICK R，FARHADI A. Unsupervised deep embedding for clustering analysis[C]// International Conference on Machine Learning. PMLR，2016：478-487.

[2] YANG B，FU X，SIDIROPOULOS N D，et al. Towards k-means-friendly spaces：Simultaneous deep learning and clustering[C]// International Conference on Machine Learning. PMLR，2017：3861-3870.

[3] TIAN K，ZHOU S，GUAN J. DEEPCLUSTER：A general clustering framework based on deep learning[C]// Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Cham：Springer，2017：809-825.

[4] AGGARWAL C C，WOLF J L，YU P S，et al. Fast algorithms for projected clustering[J]. ACM SIGMoD Record，1999，28(2)：61-72.

[5] SHAHAM U，STANTON K，LI H，et al. Spectralnet：Spectral clustering using deep neural networks[J]. arXiv：1801. 01587，2018.

[6] YANG J，PARIKH D，BATRA D. Joint unsupervised learning of deep representations and image clusters[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016：5147-5156.

[7] CHEN T，KORNBLITH S，NOROUZI M，et al. A simple framework for contrastive learning of visual representations[C]// International Conference on Machine Learning. PMLR，2020：1597-1607.

[8] HE K，FAN H，WU Y，et al. Momentum contrast for unsupervised visual representation learning[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020：9729-9738.

[9] JI X，HENRIQUES J F，VEDALDI A. Invariant information clustering for unsupervised image classification and segmentation[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019：9865-9874.

[10] LI Y，HU P，LIU Z，et al. Contrastive clustering[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2021：8547-8555.

[11] DO K，TRAN T，VENKATESH S. Clustering by maximizing mutual information across views[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021：9928-9938.

[12] DANG Z，DENG C，YANG X，et al. Nearest neighbor matching for deep clustering[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021：13693-13702.

[13] ZHANG R，ISOLA P，EFROS A A. Colorful image colorization[C]// European Conference on Computer Vision. Cham：Springer，2016：649-666.

[14] GIDARIS S，SINGH P，KOMODAKIS N. Unsupervised representation learning by predicting image rotations[J]. arXiv：1803. 07728，2018.

[15] PATHAK D，KRAHENBUHL P，DONAHUE J，et al. Context encoders：Feature learning by inpainting[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016：2536-2544.

[16] VAN GANSBEKE W，VANDENHENDE S，GEORGOULIS S，et al. Scan：Learning to classify images without labels[C]// European Conference on Computer Vision. Cham：Springer，2020：268-285.

[17] JI P，ZHANG T，LI H，et al. Deep subspace clustering networks[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017：23-32.

[18] YAMAGUCHI M，IRIE G，KAWANISHI T，et al. Subspace structure-aware spectral clustering for robust subspace clustering[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019：9875-9884.

[19] LAW M T，URTASUN R，ZEMEL R S. Deep spectral clustering learning[C]// International Conference on Machine Learning. PMLR，2017：1985-1994.

[20] YANG X，DENG C，ZHENG F，et al. Deep spectral clustering using dual autoencoder network[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019：4066-4075.

[21] ZHONG H，WU J，CHEN C，et al. Graph contrastive clustering[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021：9224-9233.

[22] WU J，LONG K，WANG F，et al. Deep comprehensive correlation mining for image clustering[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019：8150-8159.

[23] CHANG J, WANG L, MENG G, et al. Deep adaptive image clustering[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017: 5879-5887.

[24] MAHON L, LUKASIEWICZT. Selective Pseudo-Label Clustering [C] // German Conference on Artificial Intelligence (Künstliche Intelligenz). Cham: Springer, 2021: 158-178.

[25] GUPTA D, RAMJEE R, KWATRA N, et al. Unsupervised clustering using pseudo-semi-supervised learning[C]// International Conference on Learning Representations. 2020.

[26] PARK S, HAN S, KIM S, et al. Improving unsupervised image clustering with robust learning[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 12278-12287.

[27] LOWE D G. Object recognition from local scale-invariant features[C]// Proceedings of the seventh IEEE International Conference on Computer Vision. 1999: 1150-1157.

[28] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]// 2005 IEEEComputer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). 2005: 886-893.

[29] VINCENT P, LAROCHELLE H, BENGIO Y, et al. Extracting and composing robust features with denoising autoencoders [C]// Proceedings of the 25th International Conference on Machine Learning. 2008: 1096-1103.

[30] DONAHUE J, SIMONYAN K. Large scale adversarial representation learning[C]// Proceedings of the 33rd International Conference on Neural Information Processing Systems. 2019: 10542-10552.

[31] NOROOZI M, PIRSIAVASH H, FAVARO P. Representation learning by learning to count[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017: 5898-5906.

[32] BACHMAN P, HJELM R D, BUCHWALTER W. Learning representations by maximizing mutual information across views [C]// Proceedings of the 33rd International Conference on Neural Information Processing Systems. 2019: 15535-15545.

[33] HADSELL R, CHOPRA S, LECUN Y. Dimensionality reduction by learning an invariant mapping[C]// 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). IEEE, 2006: 1735-1742.

[34] GRILL J B, STRUB F, ALTCHÉ F, et al. Bootstrap your own latent—a new approach to self-supervised learning[J]. Advances in Neural Information Processing Systems, 2020, 33: 21271-21284.

[35] LI J, ZHOU P, XIONG C, et al. Prototypical Contrastive Learning of Unsupervised Representations[C]// ICLR. 2021.

[36] GUO Y, XU M, LI J, et al. HCSC: Hierarchical Contrastive Selective Coding[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 9706-9715.

[37] SAJJADI M, JAVANMARDI M, TASDIZEN T. Regularization with stochastic transformations and perturbations for deep semi-supervised learning[C]// Proceedings of the 30th International Conference on Neural Information Processing Systems. 2016: 1171-1179.

[38] OORD A, LI Y, VINYALS O. Representation learning with contrastive predictive coding[J]. arXiv: 1807. 03748, 2018.

[39] CHAPELLE O, SCHOLKOPF B, ZIEN A. Semi-supervised learning[J]. IEEE Transactions on Neural Networks, 2009, 20(3): 542-542.

[40] SOHN K, BERTHELOT D, CARLINI N, et al. Fixmatch: Simplifying semi-supervised learning with consistency and confidence[J]. Advances in Neural Information Processing Systems, 2020, 33: 596-608.

[41] HU Z, YANG Z, HU X, et al. Simple: similar pseudo label exploitation for semi-supervised classification[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 15099-15108.

[42] ARAZO E, ORTEGO D, ALBERT P, et al. Pseudo-labeling and confirmation bias in deep semi-supervised learning[C]// 2020 International Joint Conference on Neural Networks(IJCNN). IEEE, 2020: 1-8.

[43] TARVAINEN A, VALPOLA H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017: 1195-1204.

[44] LEE D H. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks[C]// Workshop on Challenges in Representation Learning. 2013: 896.

[45] BERTHELOT D, CARLINI N, GOODFELLOW I, et al. Mixmatch: A holistic approach to semi-supervised learning [C]// Proceedings of the 33rd International Conference on Neural Information Processing Systems. 2019: 5049-5059.

[46] BERTHELOT D, CARLINI N, CUBUK E D, et al. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring[J]. arXiv: 1911. 09785, 2019.

[47] SELLARS P, AVILES-RIVERO A I, SCHÖNLIEB C B. Laplacenet: A hybrid energy-neural model for deep semi-supervised classification[J]. arXiv: 2106. 04527, 2021.

[48] ZHANG B, WANG Y, HOU W, et al. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling[J]. Advances in Neural Information Processing Systems, 2021, 34: 18408-18419.

[49] CHAPELLE O, SCHOLKOPF B, ZIEN A. Semi-supervised learning[J]. IEEE Transactions on Neural Networks, 2009, 20(3): 542-542.

[50] ZHU X, GOLDBERG A B. Introduction to semi-supervised learning[J]. Synthesis Lectures on Artificial Intelligence and Machine Learning, 2009, 3(1): 1-130.

[51] JOHNSON J, DOUZE M, JÉGOU H. Billion-scale similarity search with gpus[J]. IEEE Transactions on Big Data, 2019, 7(3): 535-547.

[52] CUBUK E D, ZOPH B, SHLENS J, et al. Randaugment: Practical automated data augmentation with a reduced search space [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020: 702-703.

[53] HUANG J, GONG S, ZHU X. Deep semantic clustering by partition confidence maximisation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 8849-8858.

[54] ZHOU D, BOUSQUET O, LAL T, et al. Learning with local

and global consistency[C]// Proceedings of the 16th International Conference on Neural Information Processing Systems. 2003:321-328.

[55] ZHANG H,CISSE M,DAUPHIN Y N,et al. mixup:Beyond empirical risk minimization[J]. arXiv:1710.09412,2017.

[56] HE K,SUN J. Convolutional neural networks at constrained time cost[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015:5353-5360.

[57] KRIZHEVSKY A,HINTON G. Learning multiple layers of features from tiny images[J/OL]. Handbook of Systemic Autoimmune Diseases,2009,1(4). https://scholar.google.com/scholar?hl=zh-CN&as_sdt=0%2C5&q=Learning+multiple+layers+of+features+from+tiny+image&btnG=.

[58] KINGMA D P,BA J. Adam:A method for stochastic optimization[J]. arXiv:1412.6980,2014.

[59] WANG F,JIANG M,QIAN C,et al. Residual attention network for image classification[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017:3156-3164.

[60] DEVRIES T,TAYLOR G W. Improved regularization of convolutional neural networks with cutout[J]. arXiv:1708.04552,2017.

[61] MCDAID A F,GREENE D,HURLEY N. Normalized mutual information to evaluate overlapping community finding algorithms[J]. arXiv:1110.2515,2011.

[62] HUBERT L,ARABIE P. Comparing partitions[J]. Journal of Classification,1985,2(1):193-218.

[63] MACQUUEN J B. Some methods for classification and analysis of multivariate observation[C]// Proceedings of the 5th Berkley Symposium on Mathematical Statistics and Probability. 1967:281-297.

[64] ZELNIK-MANOR L,PERONA P. Self-tuning spectral clustering[C]// Proceedings of the 17th International Conference on Neural Information Processing Systems. 2004:1601-1608.

[65] CAI D,HE X,WANG X,et al. Locality preserving nonnegative matrix factorization[C]// 21st International Joint Conference on Artificial Intelligence(IJCAI 2009). 2009:1010-1015.

[66] NG A. Sparse autoencoder [J]. CS294A Lecture Notes,2011,72(2011):1-19.

[67] VINCENT P,LAROCHELLE H,LAJOIE I,et al. Stacked denoising autoencoders:Learning useful representations in a deep network with a local denoising criterion[J]. Journal of Machine Learning Research,2010,11(12):3371-3408.

[68] BENGIO Y,LAMBLIN P,POPOVICI D,et al. Greedy layer-wise training of deep networks[C]// Proceedings of the 19th International Conference on Neural Information Processing Systems. 2006:153-160.

[69] RADFORD A,METZ L,CHINTALA S. Unsupervised representation learning with deep convolutional generative adversarial networks[J]. arXiv:1511.06434,2015.

[70] KINGMA D P,WELLING M. Auto-encoding variational bayes [J]. arXiv:1312.6114,2013.

[71] HAEUSSER P,PLAPP J,GOLKOV V,et al. Associative deep clustering:Training a classification network with no labels[C]// German Conference on Pattern Recognition. Cham:Springer,2018:18-32.

[72] CARON M,BOJANOWSKI P,JOULIN A,et al. Deep clustering for unsupervised learning of visual features[C]// Proceedings of the European Conference on Computer Vision (ECCV). 2018:132-149.

[73] CHANG J,GUO Y,WANG L,et al. Deep discriminative clustering analysis[J]. arXiv:1905.01681,2019.

CAI Shaotian, born in 1999, postgraduate, is a member of China Computer Federation. His main research interests include data mining and machine learning.

CHEN Xiaojun, born in 1981, associate professor, is a member of China Computer Federation. His main research interests include subspace clustering, topic model, feature selection and massive data mining.

(责任编辑:柯颖)