计算机科学
COMPUTER SCIENCE

基于点态流形与一致正则的半监督学习算法

徐亚敏, 李晓斌, 张润

引用本文

徐亚敏, 李晓斌, 张润. 基于点态流形与一致正则的半监督学习算法[J]. 计算机科学, 2026, 53(4): 173-179.

XU Yamin, LI Xiaobin, ZHANG Run. Semi-supervised Learning Algorithm Based on Pointwise Manifold Structures and Uniform Regularity Constraints [J]. Computer Science, 2026, 53(4): 173-179.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.250300086

基于点态流形与一致正则的半监督学习算法

徐亚敏　李晓斌　张　润
西南交通大学数学学院　成都 611756
(xuyaminswjtu@foxmail.com)

摘　要　流形正则化(Manifold Regularization,MR)提供了一个有效的框架,利用有标签数据集和无标签数据集进行半监督分类。在基于流形假设的情况下,约束相似实例在样本构图上应具有相似的分类结果。值得注意的是,MR的核心在于样本构图上的成对平滑,即所有实例对中都应用了平滑约束,把每一对实例都看作一个整体。然而,平滑性本质上可以是点对点的,这意味着平滑性应当"无处不在",以关联每个点或实例与其邻近点的行为。因此,提出了一种新型的基于点态流形正则化以及一致性正则化的半监督学习算法URC-PW-MR。该方法不仅保留了平滑性的点对点特性,还通过考虑单个实例而非实例对,引入了单个实例的重要性。这种重要性可以通过局部密度等因素来描述。URC-PW-MR提供了一种新的实现流形平滑性的方法,通过约束单个局部实例并引入融合一致性正则来实现半监督学习。实证结果表明,URC-PW-MR在性能上与传统的MR相比更为精细。

关键词:流形正则化;点态流形正则化;融合一致正则;半监督学习
中图分类号　TP311

Semi-supervised Learning Algorithm Based on Pointwise Manifold Structures and Uniform Regularity Constraints

XU Yamin,LI Xiaobin and ZHANG Run
School of Mathematics,Southwest Jiaotong University,Chengdu 611756,China

Abstract　MR(Manifold Regularization) provides a powerful framework for semi-supervised classification using labeled and unlabeled datasets. Under the assumption of manifold,it enforces that similar instances should have similar classification results on the sample graph. It is notable that the core of MR lies in the pairwise smoothing on the sample graph,where smoothing constraints are applied to all instance pairs,treating each pair of instances as a whole. However,the smoothness can be point-to-point in essence,meaning that smoothness should be "everywhere" to correlate the behavior of each point or instance with that of its neighboring points. Therefore,this paper proposes a novel semi-supervised learning algorithm based on pointwise manifold and uniform regularity constraints(URC-PW-MR),which achieves semi-supervised learning by constraining individual local instances and introducing a fusion consistency regularization. This approach not only contains the pointwise nature of smoothness but also introduces the importance of individual instance by considering each instance rather than pairs of instances. The significance can be quantitatively characterized through factors such as local density. URC-PW-MR proposes a novel manifold smoothness realization approach that achieves semi-supervised learning through dual constraints:individual local instance regularization and fusion consistency regularization. Empirical evaluations demonstrate that URC-PW-MR exhibits more refined performance characteristics compared with conventional MR frameworks.

Keywords　Manifold regularization,Pointwise manifold regularization,Uniform regularity constraints,Semi-supervised learning

随着行业数据量的爆发式增长,大数据的海量性、复杂性、多样性以及多变等特性,导致传统机器学习算法在小数据集上的优势不再。因此,如何将机器学习算法高效应用于大数据环境,成为学术界和工业界共同关注的话题[1]。在机器学习领域,作为介于无监督学习和有监督学习之间的一种方法,半监督学习近年来受到了广泛关注[2]。

半监督学习在数据标注成本高和数据量大的场景中具有显著优势,但也需要在算法设计和数据质量方面进行细致的考虑和优化[3-4]。在许多实际应用中,如信用卡数据、望远镜数据等未标记数据的获取相对容易,而标记数据的获取则成

到稿日期:2025-03-17　返修日期:2025-06-16
基金资助:理科培育专项一般项目(2682021ZTPY043,2682025ZTPY001);国家自然科学基金(11501470,11426187)
This work was supported by the Fundamental Research Funds for the Central Universities(2682021ZTPY043,2682025ZTPY001) and National Natural Science Foundation of China(11501470,11426187).
通信作者:李晓斌(lixiaobin@home.swjtu.edu.cn)

本高昂[5]。深度学习的研究在大数据时代背景下蓬勃兴起，成果斐然。然而，训练深度网络模型通常需要大量优质的标记的训练样本[5-6]。因此，如何使未标记的数据得以充分利用，从而提高学习模型的泛化概括能力，成为半监督学习研究的重要课题[7]。

流形正则化(Manifold Regularization，MR)是半监督学习中的一种学习框架。它以流形假设为基础，即高维观测数据实际上分布在一个嵌入在高维空间中的低维流形上，在此流形结构上邻近或相似的数据实例应具有相似的输出。然而，传统的流形正则化方法主要关注样本对之间的平滑性，而忽略了样本点的局部特性[4,8]。点态流形正则化(Pointwise Manifold Regularization，PW-MR)则将这一概念进一步细化，将每个样本点与其近邻的行为联系起来[9]，通过逐点实现平滑性约束。这种方法不仅保留了平滑性的逐点特性，还通过引入局部密度来增强模型对数据结构的利用[10]。

与此同时，一致性正则化(Consistency Regularization)作为一种新兴的研究热点，强调模型对样本扰动的稳健性。它通过确保模型在样本邻域内的预测结果具有一致性，从而提高模型的泛化能力[11]。而半监督学习中，有一种通过标签传播机制并利用未标记数据来提高分类器的性能，然而，一致性正则化方法往往忽略样本数据集的流形结构，可能导致相近样本的输出差异性较大，进而影响分类器的性能[12-13]。

为改善传统方法的劣势，本文提出了一种基于一致性正则化与点态流形正则化的半监督学习算法。该算法引入平滑性损失等构建样本图，基于一致性约束实现了样本局部邻域及其近邻样本间的平滑性。这种综合方法充分利用了数据的流形结构信息，使模型能够向合理的低密度区域推进分类边界，同时保持了对局部预测的平滑。

实验结果表明，融合一致性正则与点态流形正则的半监督学习算法在多个图文数据上均优于其他算法。但在实际实验中，存在标记样本的稀缺性、数据分布的复杂性、样本重要性的考虑等一系列问题。然而，点态流形正则化方法可充分利用未标记数据，提升模型泛化能力。通过逐点约束，该方法能更好地保持流形的局部几何特性，并引入样本的局部密度或置信度等指标来衡量样本的重要性，进而提高分类性能。针对相关问题，本文提出了一种基于点态流形正则化与一致正则化的半监督学习算法(Semi-supervised Learning Algorithm Based on Pointwise Manifold Structures and Uniform Regularity Constraints，URC-PW-MR)。由于流形在局部与欧氏空间同胚，具有局部欧氏性质，因此使用可度量欧氏距离进行计算[14-16]。

1 研究基础

1.1 点态流形的定义与构建

1.1.1 点态流形的数学定义

点态流形(Pointwise Manifold)的数学定义可以从以下几个方面来理解。1)局部欧几里德性：流形是拓扑空间，其中每个点均有一个邻域，且欧几里德空间的一个开集同胚。这意味着在局部，流形看起来像欧几里德空间，即在小的区域内，可以使用类似于笛卡尔坐标系的坐标来描述流形上的点[17-18]。2)维度：流形具有固定的维度，这意味着在任何点的局部，它都可以用固定数量的坐标来描述。这个维度在整个流形上是一致的[18]。3)Hausdorff 空间：流形是一个 Hausdorff 空间，这意味着对于流形上的任意两个不同的点，都存在两个不相交的开邻域分别包含这两个点。这个性质强调了流形的可分性[19]。4)局部坐标覆盖：流形可以被一个开覆盖和一个连续映射族所描述，这些映射将流形的每个部分映射到欧氏空间的一个开集上，并且这些映射之间通过转换映射相互联系。这些转换映射是连续的，并且在重叠区域上是C^r映射[19]。5)点态定向：在流形的每个点上，可以指定一个切空间的定向。点态定向意味着为每个切空间T_pM指定一个定向。如果每个点p都属于某个定向的局部框架的域，则称这种定向为连续点态定向。6)流形的数学定义：一个n维流形M是一个具有A_2(第二可数)、T_2(Hausdorff)性质的拓扑空间，如果存在M的一个开覆盖$\{U_\alpha\}$和相应的连续映射族$\varphi_\alpha: U_\alpha \to R^n$，使得$\varphi_\alpha: U_\alpha \to \varphi_\alpha(U_\alpha)$是从$U_\alpha$到欧氏空间开集$\varphi_\alpha(U_\alpha)$上的同胚，并且当$U_\alpha \cap U_\beta \neq \emptyset$时，转换映射$\varphi_\beta \circ \varphi_\alpha^{-1}: \varphi_\alpha(U_\alpha \cup U_\beta) \to \varphi_\beta(U_\alpha \cap U_\beta)$为$C^r$映射，则称$M$为$C^r$流形[20]。

这些定义提供了点态流形的数学框架，它们是理解流形及其在数学、物理学和工程学等领域应用的基础。

1.2 基于一致正则的半监督学习算法

一致性正则化技术主要涵盖样本扰动与模型扰动两大类策略。虽然它们在具体操作层面各有不同，但其核心目标一致，均旨在降低模型预测结果的差异性损失。样本扰动策略通过将原始数据及其扰动版本同时输入同一模型并优化，以减少二者输出的差异，从而实现这一目标。该策略高度依赖于数据增强技术，以确保扰动样本的质量。

近年来，众多研究者致力于开发高效的数据增强技术。其中，Miyato 等[20]提出的虚拟对抗训练(VAT)模型尤为突出。该模型的核心在于确定最大化模型输出的偏差方向，并针对输入数据进行扰动。在半监督学习领域，Verma 等[21]提出了插值一致性训练(ICT)模型。该模型为了确保对插值点的预测与两个样本点预测结果的插值保持一致，在两个样本点之间进行插值，从而使模型更稳定。与此同时，Berthelot 等[22]结合多种数据增强技术，开发了 MixMatch 算法，该算法在降低分类错误率方面表现优异。

在模型扰动策略方面，Laine 等[23]设计了两种模型——“Π”模型和 Temporal ensembling 模型，均采用了不同的策略来实现一致性正则化。“Π”将训练样本同时输入两个结构相同但参数不同的网络，并利用随机 Dropout 技术为每个网络生成独特参数，最终通过最小化两个网络的预测差异来实现一致性正则化。Temporal ensembling 模型则在训练的初始阶段计算样本预测的平均值，并将其与当前周期的预测结果进行对比，以此优化模型的一致性。Tarvainen 等[24]提出了 Mean teacher 模型。区别于 Temporal ensembling 的预测平均化方法，该模型通过平均前期训练阶段的模型参数，并最小化这种预测差异，从而确保模型扰动的一致性。

传统一致正则化方法着重于样本邻域内的一致性，即通

过保证相邻数据点预测结果的相似性，来增强模型的稳定性。然而，这种方法存在一定的局限性：它仅仅考虑了局部邻域内的信息，而忽略了数据点之间的全局联系以及每个数据点的独特属性。具体来说，不同数据点可能具有不同的特征分布和内在结构，这些特性对于理解整个数据集的整体模式至关重要。因此，仅依赖于局部一致性可能会导致部分重要的数据结构信息被忽略，从而影响模型的泛化能力和准确性。为了解决这一问题，本文提出了一种新型的半监督学习算法，该算法结合了一致正则性和点流形正则性两种方法的优势。一致正则性继续用于保证局部邻域内的一致性，确保模型在小范围内保持稳定；而点态流形正则性则引入了对数据点之间全局关系的理解，特别强调了每个数据点在其所在流形上的位置和特性。通过这种方式，该算法不仅能够捕捉到局部邻域内的相似性，还能更好地反映数据的整体分布和内在结构，从而更全面地利用样本中的信息。

此外，点态流形正则性的引入使得算法可以更好地处理高维数据中的复杂模式。例如，在图像识别任务中，不同像素点之间的关系不仅仅是简单的空间邻近性，还涉及到颜色、纹理等多方面的特征关联。通过结合这两种正则化方法，模型能够在训练过程中更加充分地挖掘数据中的潜在规律，进而提升其在未标记数据上的表现。

综上所述，本文提出的基于一致正则性和点态流形正则性的半监督学习算法，不仅弥补了传统方法在信息利用上的不足，而且为解决复杂数据结构下的学习问题提供了一种新的思路。这种改进有助于提高模型的鲁棒性和泛化能力，使其在更多实际应用场景中展现出更好的性能。

1.3 点态正则化方法

基于点态流形的正则化技术是一种半监督学习方法，它将有标记数据和无标记数据结合，以提高分类的准确性。点态流形正则化(PW-MR)的核心思想是：在流形上，其结构彼此相似且相邻的实例输入应具有相似的分类输出。该原理基于流形假设，其假定数据点分布在高维空间中的低维流形上或附近。简言之，若两个数据点在这个低维结构中是近邻，则它们可能具有相似的特征，因此也有相似的分类(基于流形假设的)。

与传统的流形正则化(MR)对所有实例对施加约束不同，PW-MR 专注于单个局部实例。传统的 MR 方法施加全局平滑性约束，这意味着在执行正则化时会考虑每一对实例。然而，PW-MR 将其关注点缩小到每个实例的直接邻域，确保只有附近的点会影响正则化过程。这种局部化的方法使 PW-MR 能够更有效地保持平滑性的内在点态特性。

通过专注于单个局部实例，PW-MR 能够以更细致的方式使用每个单独实例的特征。例如，预测的置信水平或预测点的局部密度可以纳入正则化过程。如果某个实例对其分类具有较高的置信水平，那么它在确定平滑度约束时可能具有更大的权重。同样，如果流形的某个区域具有较高密度的正确预测点，那么这些点在保持流形结构的平滑度方面可以被赋予更大的重要性。

这种方法在提高预测准确率的同时，也能保证模型对数据中的局部变化保持敏感。通过考虑每个实例的特定特征，PW-MR 能够适应流形不同区域的独特属性，从而获得更稳健和可靠的分类结果。

2 算法设计

2.1 模型构建前缀

给定标记数据 $X_l=\{x_i\}_{i=1}^{l}$，其标签 $Y=\{y_i\}_{i=1}^{l}$，未标记数据 $X_u=\{x_j\}_{j=l+1}^{n}$，其中 $x_i\in R^d$，$u=n-l$。在数据集上构造近邻图 $G=\{w_{ij}\}_{i,j=1}^{n}$，其中 w_{ij} 表示连接实例 x_i 和 x_j 的权重。基于上述近邻图，对应的 MR 框架表达式如下：

$$\min_{f\in H_K}\frac{1}{l}\sum_{i=1}^{l}V(x_i,y_i,f)+C_1\|f\|_K^2+\frac{C_2}{(l+u)^2}\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}(f(x_i)-f(x_j))^2 \tag{1}$$

其中，$f(x)$ 是决策函数，C_1 和 C_2 为正则化参数，$V(x_i,y_i,f)$ 是损失函数。这种设置使得流形正则化(MR)框架能够自然地适用于特定的算法，例如损失函数可以是支持向量机(SVM)的铰链损失 $\max\{0,1-y_if(x_i)\}$，也可以是正则化最小二乘分类器(RLSC)的平方损失 $(y_i-f(x_i))^2$，或者是拉普拉斯支持向量机(LapSVM)和拉普拉斯正则化最小二乘分类器(LapRLSC)。而再生核希尔伯特空间(RHKS)中的平滑度的正则化项是 $\|f\|_K^2$。式(1)中第三项确保了近邻图上的成对平滑性，即流形结构上相似的实例应具有相似的分类输出。这一要求可以进一步表示为：

$$\sum_{i,j=1}^{l+u}w_{ij}(f(x_i)-f(x_j))^2=2\boldsymbol{f}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{f} \tag{2}$$

其中，$\boldsymbol{f}=[f(x_1),\cdots,f(x_{l+u})]^{\mathrm{T}}$；$\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{W}$，即图拉普拉斯算子；$\boldsymbol{W}$ 是图 G 的权重矩阵；$\boldsymbol{D}$ 是对角矩阵，对角分量 $D_{ii}=\sum_{j=1}^{n}\boldsymbol{W}_{ij}$。根据 Representer 定理，最小化问题方程式(1)的形式如下：

$$f^*(x)=\sum_{i=1}^{l+u}\alpha_iK(x_i,x) \tag{3}$$

其中，$K:X\times X\to R$ 是 Mercer 核(可以通过用 1-值元素增强每个实例来省略决策函数的偏差)。

2.2 模型构建

通过实现逐点平滑度，基于一致正则的点态流形正则，即 URC-PW-MR 框架的优化问题可以表述为：

$$\min_{f\in H_K}\frac{1}{l}\sum_{i=1}^{l}V(x_i,y_i,f)+C_1\|f\|_K^2+\frac{C_2}{(l+u)^2}\sum_{i=1}^{n}p(x_i)\Big(f(x_i)-\sum_{x_j\in N(x_i)}w_{ij}f(x_j)\Big)^2 \tag{4}$$

其中，$x_j\in N(x_i)$ 是邻域集 $N(x_i)$ 中 x_i 的近邻，$p(x_i)$ 表示每个实例 x_i 周围的局部密度。然而，当一个实例 x_i 出现在分类重叠的区域时，基于上述局部密度的描述，它会在 x_i 周围表现出较高的密集性，因此在式(4)的第三项中，x_i 将受到较大的影响，这并非理想情况。因此，在计算局部密度时，综合考虑了局部密度和无监督学习结构：

$$p(x_i)=\left(1-\frac{\sum_{x_j\in N(x_i)}d(x_i,x_j)}{\sum_{s=1}^{n}\sum_{x_t\in N(x_s)}(x_s,x_t)}\right)\times\max(u_{1i},u_{2i}) \tag{5}$$

其中，$d(x_i,x_j)$表示实例x_i和x_j之间的距离，$\sum_{s=1}^{n}\sum_{x_t\in N(x_s)}(x_s,x_t)$是所有实例间$d(x_i,x_j)$距离的累积和。此外，$u_{1i}$和$u_{2i}$是通过无监督学习方法（例如模糊C均值聚类算法（FCM））来确定各实例归属单一聚类的隶属度或其属于x_i的概率。式(5)的第一项着重考量了每个实例x_i与它相邻实例之间的标准化距离，当x_i周围的分布较为密集时，该值会较小，从而使$p(x_i)$的值较大。最终在式(4)中，这种密集分布会导致其受到较大的惩罚。式(5)第二项学习成果纳入了无监督学习的部分。由于监督学习方法通常侧重于关注识别分布结构的内在边界，在此背景下，$\max(u_{1i},u_{2i})$的值越大，说明x_i位于边界之外的概率越高，即越倾向于成为非边界实例。

区别于式(1)中的第三项，式(4)中的第三项聚焦于单个局部实例的平滑性，具体而言，它规定每个实例需与其周边实例拥有类似的分类输出结果，然后针对逐点平滑性进行考量，进而体现图拉普拉斯算子所包含的逐点平滑的特性。PW-MR借助局部密度对每个实例的影响进行度量。

此外，PW-MR在引入逐点平滑性的基础上，进一步结合局部密度因素，以此凸显每个实例的独特价值。进一步将式(4)中的第三项改写成如下形式：

$$J_L=\sum_{i=1}^{n}p(x_i)(f(x_i)-\sum_{x_j\in N(x_i)}w_{ij}f(x_j))^2$$
$$=\boldsymbol{f}^{\mathrm{T}}(\mathbf{I}-\boldsymbol{W})^{\mathrm{T}}\boldsymbol{P}(\mathbf{I}-\boldsymbol{W})\boldsymbol{f}=\boldsymbol{f}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{f} \tag{6}$$

其中，$\boldsymbol{W}$是邻域相似矩阵，每一个$W_{ij}=\begin{cases}w_{ij}, & \text{if } x_j\in N(x_i)\\ 0, & \text{otherwise}\end{cases}$；$\mathbf{I}\in R^{n\times n}$是单位矩阵；$\boldsymbol{P}\in R^{n\times n}$是对角矩阵，其分量为$P_{ij}=p(x_i)$。

PW-MR通过引入逐点平滑性，构建了一个图拉普拉斯矩阵的优化。由于它紧密贴合数据特性，故这一矩阵的定制化设计对提升MR的性能至关重要。值得注意的是，更优的图结构确实能够显著增强MR的表现。具体而言，逐点平滑性的融入使得新图拉普拉斯矩阵能够更精准地契合流形结构。

在PW-MR框架中使用不同的损失函数会造成分类器的差异。

2.3 损失函数

2.3.1 损失函数描述

假设数据集G里存在N个样本，其中有标记样本集$\mathcal{L}=\{(x_i,y_i)\}_{i=1}^{L}$，标记$y_i\in\mathcal{Y}=\{1,2,\cdots,K\}$，共有$K$个类别；无标记样本集$\mathcal{U}=\{x_i\}_{i=L+1}^{N}$。算法的总体损失函数如下：

$$\ell=\ell_e+\lambda_c\ell_c+\lambda_s\ell_s \tag{7}$$

总体损失由3部分构成：第一部分，针对有标记样本，模型的性能评估是通过计算模型预测结果与真实标记之间的交叉熵损失ℓ_e来完成的；第二部分，对于无标记样本，通过数据增强技术$Augement(x_u)$来计算一致性损失ℓ_c；第三部分，从无标签和有标签样本中提取样本，映射特征空间以构建图结构，并计算平滑性损失ℓ_s。同时，需要对这3项损失的权重参数进行平衡调整。本文算法整体框架如图1所示。

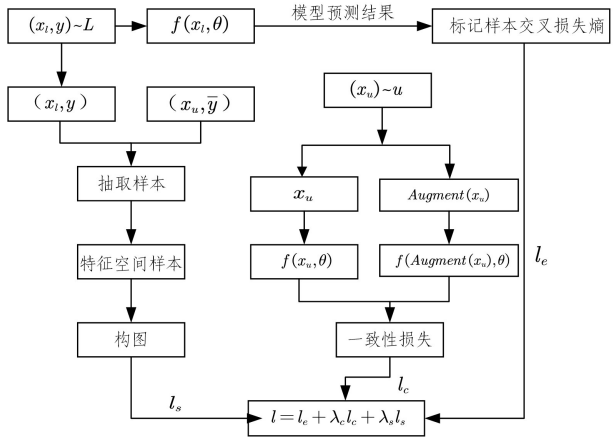

图1 损失函数示意图

Fig. 1 Schematic diagram of loss function

2.3.2 样本邻域内的一致性损失

在ICT模型框架下，运用Mixup这一数据增强技术来核算一致性损失，其基础公式如下：

$$Mix_\lambda(a,b)=\lambda a+(1-\lambda)b \tag{8}$$

其中，$Mix_\lambda(a,b)$为a和b之间的插值，λ是服从β分布的权值参数。

在小批量（Mini-batch）数据集中，对任意两个样本点x_i和x_j，以及对应模型的预测结果$f(x_i,\theta)$和$f(x_j,\theta)$，根据式(2)，可以计算出两点的插值为：

$$\hat{x}=Mix_\lambda(x_i,x_j) \tag{9}$$

则模型对该插值的预测结果为：

$$f(\hat{x},\theta)=f(Mix_\lambda(x_i,x_j),\theta) \tag{10}$$

同时，可以得到模型对样本点x_i和x_j预测结果的插值为：

$$\hat{y}=Mix_\lambda(f(x_i,\theta),f(x_j,\theta)) \tag{11}$$

一致性损失确保$f(\hat{x},\theta)$与$\hat{y}$保持一致。因此，在一个小批量（Mini-batch）数据集内的一致性损失为：

$$\ell_c(\theta,\mathcal{L},\mathcal{U})=\frac{2}{|B||B-1|}\sum_{x_i,x_j\in B}d(f(\hat{x},\theta),\hat{y}) \tag{12}$$

其中，$d(\cdot,\cdot)$是衡量分类器对预测标记（对样本插值后的标记）与真实标记（对样本标记进行插值）之间的差异性，本文采用交叉熵损失$d(f(\hat{x},\theta),y)=-\log[f(\hat{x},\theta)]\hat{y}$来评估$f(\hat{x},\theta)$和$\hat{y}$的一致性程度；$B$为一个小批量数据集。

2.3.3 样本间的平滑性损失

本文基于深度学习，使用一种基于小批量数据集的构图方法，通过计算平滑性损失来优化模型性能。在此基础上，引入动态构图策略，使图结构能够随着学习过程的推进而持续更新，进而更有效地引导学习器捕捉样本间的平滑性特征。

1）构图与邻接矩阵的计算

对于每个小批量数据集，使用数据点来构建近邻图（K-Nearest Neighbor，KNN）。区别于常规构图方法，本研究将样本的标记空间作为度量样本相似性的标准，并由此推导出权值矩阵的计算式：

$$W_{ij}=\exp\left(-\frac{\|f(x_i),f(x_j)\|_2^2}{2\delta^2}\right) \tag{13}$$

其中，$\delta>0$ 是指定的高斯函数带宽参数 $sigma$，$f(x_i)$ 是模型对样本的预测。

2）平滑性损失计算

样本特征与给定的邻接矩阵 $\boldsymbol{W}$ 进行平滑性损失计算：

$$\ell_s(\theta,\mathcal{L},\mathcal{U},\boldsymbol{W})=\frac{1}{2}\sum_{x_i,x_j\in B}W_{ij}\parallel h(x_i)-h(x_j)\parallel^2 \quad (14)$$

其中，$h:\chi\to\mathbb{R}^p$ 是输入空间到神经网络倒数第二层的映射。

2.3.4 算法实现细节

本文整体损失函数如下：

$$\ell=-\frac{1}{|B\bigcap\mathcal{L}|}\sum_{x_i\in B\cap\mathcal{L}}\log[f(x_i;\theta)]_{y_i}+$$
$$w_c(t)\lambda_c\frac{2}{|B||B-1|}\sum_{x_i,x_j\in B}d(f(\hat{x},\theta),\hat{y})+$$
$$w_s(t)\lambda_s\frac{1}{2}\sum_{x_i,x_j\in B}W_{ij}\parallel h(x_i)-h(x_j)\parallel^2 \quad (15)$$

其中，$w_c(t)$ 和 $w_s(t)$ 是权值函数，随着迭代次数线性上升。其设置依据如下。

1）动态调整策略：在训练过程中，对未标记数据的损失权重进行逐步升级，从而使其数据得到更充分的利用。

2）数据集特性：依据标记数据与未标记数据的数量比例，灵活调整权重，以平衡不同类型数据的贡献。

3）实验验证：通过在验证集上反复调整权重参数，最终确定最优的权重配置。

2.3.5 损失函数算法实现细节

首先，随机初始化模型参数 θ，然后开始一个循环，直到达到预设的迭代次数（Numepochs）。

对每个 mini-batch 执行以下操作：

1）计算有标记样本的交叉熵损失；

2）根据式（3）计算当前小批量数据内的一致性损失；

3）根据式（4）计算权值矩阵 $\boldsymbol{W}$；

4）根据式（5）计算当前 mini-batch 数据平滑性损失；

5）根据式（6）将上述 3 项损失进行加权求和，并根据总损失更新模型参数 θ。

最后结束循环周期：完成当前小批量数据的处理直到达到预设的迭代次数 numepochs，结束整个训练过程。

算法实现伪代码如算法 1 所示。

算法 1 损失函数算法

输入：有标记的样本集合，无标记样本集合 U，随 epoch 迭代线性上升的权值函数 $w_c(t)$ 和 $w_s(t)$，模型 $f(x_i;\theta)$，损失平衡项 λ_c 和 λ_s，模型迭代次数 numepochs

输出：模型参数 θ 的更新值

1. 随机初始化模型参数 θ
2. 重复
3. for each mini-batch B Do
4. 计算有标签数据的交叉熵损失
5. 由式（3）计算 mini-batch 内数据的一致性损失
6. 由式（4）计算权值矩阵 **W**
7. 由式（5）计算 mini-batch 数据的平滑性损失
8. 由式（6）得到 3 项损失的加权求和，根据损失更新模型参数
9. end for
10. 直到迭代次数>numepochs

这段伪代码的创新点主要体现在以下几个方面。

1）融合了多种损失函数。（1）交叉熵损失：针对有标记样本，确保模型能够准确分类已标注的数据。（2）一致性损失：聚焦无标记样本，增强模型在未标注数据上的稳定性，提升鲁棒性。（3）平滑性损失：借助权值矩阵 $\boldsymbol{W}$ 计算数据的平滑性，优化模型对数据分布的拟合效果。

2）权值矩阵的动态计算。权值矩阵 $\boldsymbol{W}$ 的计算（见式（4））可能是该算法的一个关键创新点。本算法通过动态调整权值矩阵，能够更好地平衡不同样本对损失函数的贡献，从而使模型的泛化能力得到提升。

3）半监督学习框架。该算法融合了有标记和无标记数据的训练机制，既利用了有限的有标记数据，又充分挖掘了大量无标记数据的潜力。在数据标注成本较高的场景下，该方法展现出显著的优势。

4）小批量训练。使用小批量（Mini-batch）训练的方式可以有效提高训练效率，同时减少内存占用，适用于大规模数据集的训练。

5）加权求和的损失函数。将 3 种损失通过式（6）进行加权求和，这种加权机制可以灵活调整不同损失在训练过程中的重要性，从而更好地平衡模型的分类性能、一致性和平滑性。

为了保证对比实验的一致性，更清晰地展示模型优化的效果，本实验的结构细节可参考文献[25]中的 Wide ResNet-28 模型，其学习衰减系数设为 0.999，权值衰减系数为 0.02。在对比实验中，选取一致性算法 UDA 以及预训练的 BERT 模型作为参考模型。针对英文文本的数据增强，采用德语作为中间语言进行回译操作。关于超参数 λ_c 和 λ_s，建议将图片数据的 λ_c 分别固定为 75，150，250，文本数据的 λ_c 固定为 100。各个数据集的 λ_s 从集合 $\left\{0,\frac{\lambda_c}{1000},\frac{\lambda_c}{100},\frac{\lambda_c}{10},\frac{\lambda_c}{3},\frac{2\lambda_c}{3},\lambda_c\right\}$ 中遍历取值，用验证集进行交叉验证，并取最优。

3 实验设计

3.1 数据描述

为验证所提算法的优越性，分别在图像数据集 CIFAR-10，CIFAR-100 和 SVHN，以及英文文本数据集 IMDB 和 Yahoo！Answers 上进行实验验证，如表 1 所列。

表 1 数据集介绍

Table 1 Datasets introduction

数据集	训练集	验证集	测试集	特征维度	类别
CIFAR-10	45000	5000	10000	32×32×3	10
CIFAR-100	45000	5000	10000	32×32×3	100
SVHN	65932	7325	26032	32×32×3	10
IMDB	63000	7000	25000	128	2
Yahoo！Answers	45000	5000	60000	128	10

选用 CIFAR-10（10 类通用物体分类，含 60000 张 32×32 像素图像），CIFAR-100（100 个细粒度子类构建层次化语义标签，样本规模与 CIFAR-10 一致）及 SVHN（超过 600000 张街景门牌数字图像，涵盖复杂背景与透视畸变场景）；文本领域采用 IMDB 电影评论数据集（50000 条标注情感极性的评论文本，用于二分类情感分析）与 Yahoo！Answers（1460000

条多主题问答数据，覆盖 10 个类别，支持长文本多分类任务)。为了评估算法在半监督学习场景中的表现，通常的做法是将大部分训练样本标记为未标记数据，然后在训练中随机选取少量样本作为标记数据。

3.2 实验环境

3.2.1 硬件环境

处理器：13th Gen Intel(R) Core(TM) i7-13700H 2.40 GHz。机带 RAM：32.0 GB(31.7 GB 可用)。设备 ID：E18EAE56-BBBF-45DF-9790-F9250C48FAA8。系统类型：64 位操作系统，基于 x64 的处理器。

3.2.2 软件环境

操作系统：Ubuntu 20.04 LTS(64 位)。Python 版本：Python 3.9.7。深度学习框架：PyTorch 1.10.0，CUDA 11.3。其他库：NumPy 1.21.2，SciPy 1.7.3，Matplotlib 3.5.0，Scikit-learn 1.0.2，OpenCV 4.5.4。实验工具：Jupyter Notebook(用于代码调试和实验记录)；TensorBoard(用于可视化训练过程)；Git(用于代码版本管理)。

3.2.3 参数设置

对于图像数据集中，相关参数 $batchsize=32$，$NUM\_LABELED\_PER\_CLASS=50$(每类有标签样本数)，$NUM\_CLASSES=10$，最近邻居个数 $k=7$，Gauss 核带宽参数 $sigma=70.0$，beta 分布参数 $alpha=0.75$。对于文本数据集，其相关参数 λ 遵循 beta 分布，$batch\ size=16$，$NUM\_LABELED=\mathrm{int}(len(x\_train\_clean)*0.1)$

3.3 实验结果与分析

在 CIFAR-10 和 SVHN 数据集上，分别选取 250，500 和 1000 个标记样本，对本文参考的原实验中的 5 种算法模型以及本文模型进行错误率评估。评估结果分别如表 2 和表 3 所列。对于 CIFAR-100 数据集，则使用 10000 个标记样本，对原实验中的 5 种算法模型与本文模型(URC-PW-MR)进行对比实验。

表 2 CIFAR-10 数据集上不同标记样本下的错误率

Table 2 Error rate on CIFAR-10 dataset with different labeled samples

(%)

Methods	样本数		
	250	500	1000
Π	53.02	41.82	31.53
VAT	36.03	26.11	18.68
Mix Match	17.06	13.77	11.16
Mean teacher	47.32	42.01	17.32
SmoothMatch	14.40	12.99	10.22
URC-PW-MR	13.03	11.99	10.02

由表 2—表 4，可以得出以下结论。

1)本文算法在 3 个图像数据集样本标记数量不同的情况下，比其他算法的准确率都要低。尤其在标记样本为[250]的情况下，其优势最为突出，这充分体现了融合一致性正则和点态流形正则策略的有效性。

2)各模型在标记样本数增加的同时，其错误率均有所降低。这是由于更多的有标记样本为分类器进行更优质的拟合提供了更丰富的监督信息。实验结果表明，在 CIFAR-10 数据集上，Mean teacher 模型的错误率显著下降，说明其对有标记数据的依赖性较强。与之相比，本文算法使得相似(相连)的样本具有相似的输出，从而减少了对有标记样本的依赖，其充分利用了数据的流形结构。因此，本文算法即使在有标记样本较少的情况下，仍能取得较好的效果。

表 3 SVHN 数据集上不同标记样本下的错误率

Table 3 Error rate on SVHN dataset with different labeled samples

(%)

Methods	样本数		
	250	500	1000
Π	17.65	11.44	8.60
VAT	8.41	7.44	5.98
Mix Match	5.58	5.46	4.45
Mean teacher	5.85	5.45	5.21
SmoothMatch	5.11	5.04	4.23
URC-PW-MR	4.98	5.01	4.14

表 4 CIFAR-100 数据集上 10000 样本数下的错误率

Table 4 Error rate on CIFAR-100 dataset with 10000 samples

算法	错误率/%
Π	39.19
VAT	35.42
Mix Match	32.88
Mean teacher	37.17
SmoothMatch	32.23
URC-PW-MR	32.12

由表 5、表 6 可知，与对比算法相比，本文算法在两个文本数据集上的表现较好。特别是在 IMDB 数据集上，当有标记样本数为 20 时，本文算法的错误率仅为 11.37%，显著低于其他模型。在有标记样本数为 100 时的结果充分体现了所提算法的优势。

表 5 4 种算法在 IMDB 上不同标记样本数下的错误率

Table 5 Error rates of the four algorithms under different numbers of labeled samples on IMDB

(%)

Methods	样本数	
	20	100
UDA	13.70	9.50
BERT	32.50	19.59
SmoothMatch	12.27	8.96
URC-PW-MR	11.37	8.56

表 6 4 种算法在 Yahoo! Answer 上不同标记样本数下的错误率

Table 6 Error rates of the four algorithms under different numbers of labeled samples on Yahoo! Answer

(%)

Methods	样本数	
	20	100
UDA	41.55	33.83
BERT	48.90	37.56
SmoothMatch	40.53	32.78
URC-PW-MR	40.46	31.78

实验结果表明，本文算法通过利用数据的流形结构，使相似样本产生相似输出，保证了样本局部预测的平滑性。

结束语 本文针对基于一致性正则的半监督深度学习算法存在的问题，即部分相近样本可能产生差异较大的输出，从而影响学习器性能，提出了一种结合一致性正则与流形正则

的改进算法。该算法通过构建样本图,引入平滑性损失,在施加一致性约束的基础上,实现了样本局部邻域的平滑性和样本间的平滑性。这一融合算法在多个图像和文本数据集上展现出显著的性能优化。

参 考 文 献

[1] HE Q,LI N,LUO W J,et al. A Survey of Machine Learning Algorithms for Big Data[J]. Pattern Recognition and Artificial Intelligence,2014,27(4):327-336.

[2] LIU J W,LIU Y,LUO X L. Semi-supervised Learning Methods [J],2015,38(8):1592-1617.

[3] YANG J C. Design and implementation of a semi-supervised continuous learning framework[D]. Nanjing:Nanjing University of Posts and Telecommunications,2020.

[4] 宋雨,许王琴,李荣鹏,等. 基于自适应流形正则化自表示的无监督特征选择算法[J]. 重庆工商大学学报(自然科学版),2023,40(6):44-52.

[5] JIA M J,LI X,KONG L,et al. Research on the Application of Deep Learning in Big Data Analysis[J]. Software Engineering and Application,2022,11(3):549-557.

[6] YAO K,CAO F,LEUNG Y,et al. Deep neural network compression through interpretability-based filter pruning[J]. Pattern Recognition,2021,119:108056.

[7] LIANG J Y,GAO J W,CHANG Y. Research Progress in Semi-supervised Learning[J]. Shanxi University(Natural Science Edition),2009,32(4):528-534.

[8] LEI Y K. Research on manifold learning algorithm and its application[D]. Hefei:University of Science and Technology of China,2011.

[9] WANG Y,HAN J,SHEN Y,et al. Pointwise manifold regularization for semi-supervised learning[J]. Frontiers of Computer Science,2021,15(1):1-8.

[10] WANG J,ZHANG S Y,LIANG J Y. Semi-supervised Deep Learning Algorithm Integrating Consensus Regulars and Manifold Regulars[J]. Big Data Research,2022,8(3):103-114.

[11] LI Y F,KWOK J T,ZHOU Z H. Semi-supervised learning using label mean[C]// Proceedings of the 26th International Conference on Machine Learning. ACM,2009.

[12] ZHU X J,GHAHRAMANI Z. Learning from labeled and unlabeled data with label propagation[C]// NIPS 2002. 2002:1-8.

[13] BELKIN M,NIYOGI P,SINDHWANI V. Manifold Regularization:A Geometric Framework for Learning from Labeled and Unlabeled Examples[J]. Journal of Machine Learning Research,2006,7(1):2399-2434.

[14] WANG J,LIANG J,CUI J,et al. Semi-supervised Learning with Mixed-order Graph Convolutional Networks[J]. Information Sciences,2021,573(8):171-181.

[15] LIANG J Y,CUI J B,WANG J,et al. Graph-based semi-supervised learning via improving the quality of the graph dynamically [J]. Machine Learning,2021,110(6):1345-1388.

[16] LU C H. Euclid and Geometry(Part II)[M]// Science World. 2019:126-127.

[17] MEI X M,HE L G. Differential manifold and Riemannian geometry[M]. Beijing:Beijing Normal University Press,1987.

[18] LIU W H,QIN B Z. Mapping properties of Meso compact space and Hausdorff space[J]. Journal of Texas College,2005,4(2):21-22.

[19] CHEN W H. Differential manifold preliminary. 2nd edition[M]. Beijing:Higher Education Press,2001.

[20] MIYATO T,MAEDASI,KOYAMA M,et al. Virtual adversarial training:aregularization method for supervised and semi-supervised learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2019,41(8):1979-1993.

[21] VERMA V,KAWAGUCHI K,LAMB A,et al. Interpolation consistency training for semi-supervised learning[J]. Neural Networks,2022,145:90-106.

[22] BERTHELOT D,CARLINI N,GOODFELLOW I,et al. Mixmatch:A holistic approach to semi-supervised learning[J]. Advances in Neural Information Processing Systems,2019,32:1-11.

[23] LAINE S,AILA T. Temporal Ensembling for Semi-Supervised Learning[C]// International Conference on Learning Representations. 2017:1-13.

[24] TARVAINEN A,VALPOLA H. Mean teachers are better role models:Weight-averaged consistency targets improve semi-supervised deep learning results[J]. arXiv:1703.01780,2017.

[25] OLIVER A,ODENA A,RAFFEL C A,et al. Realistic evaluation of deep semi-supervised learning algorithms[C]// Proceedings of the 32nd International Conference on Neural Information Processing Systems. 2018:3239-3250.

XU Yamin, postgraduate, is a member of CCF (No. Z1659G). Her main research interests include semi-supervised learning and so on.

LI Xiaobin, born in 1983, Ph.D, master's supervisor, is a member of CCF (No. Z1644M). His main research interests include geometric topology and its applications, and so on.

(责任编辑:何杨)