

# 基于k近邻的多尺度超球卷积神经网络学习

刘子巍, 骆 曦, 李 克, 陈富强

(北京联合大学 智慧城市学院, 北京 100101)

**摘 要:**以卷积神经网络(CNN)为代表的深度学习模型主要面向图像、语音等均匀采样的同质欧氏空间数据,通常不适用于大量存在于工业等领域的异质、非均匀稀疏采样的结构化数据。针对异质、非均匀稀疏采样结构化数据集的预测任务,提出一种基于k近邻(kNN)算法和CNN的超球卷积神经网络学习模型。通过kNN预处理建立各样本在高维属性空间中的结构关系,将样本邻域内各样本的标记作为其属性重构样本集合,实现数据属性集从异质到同质的转化,进而通过合理设计CNN的卷积窗,有效提取和利用各样本的邻域空间中样本的标记分布特征,完成对未知样本的预测。在不同邻域尺度、软硬标记以及混淆非混淆等条件下进行实验,结果表明,该模型预测准确率达到98.04%,其准确率和召回率较FC-CNN、CNN、kNN和Radar-CNN算法分别提升0.28%~1.66%和4.78%~31.92%。

**关键词:**卷积神经网络;k近邻算法;超球卷积;结构化数据;深度学习

开放科学(资源服务)标志码(OSID):



中文引用格式:刘子巍,骆曦,李克,等.基于k近邻的多尺度超球卷积神经网络学习[J].计算机工程,2022,48(11):111-119.

英文引用格式:LIU Z W, LUO X, LI K, et al. Multi-scale hypersphere convolutional neural network learning based on k-nearest neighbor[J]. Computer Engineering, 2022, 48(11): 111-119.

## Multi-Scale Hypersphere Convolutional Neural Network Learning Based on k-Nearest Neighbor

LIU Ziwei, LUO Xi, LI Ke, CHEN Fuqiang

(College of Smart City, Beijing Union University, Beijing 100101, China)

**[Abstract]** The deep learning model represented by the Convolutional Neural Network(CNN) model is primarily used for homogeneous Euclidean domain data, such as images and speech, and is typically difficult to directly apply to a large number of heterogeneous, unevenly, and sparsely sampled structured data from industrial fields. Aiming at the prediction task of heterogeneous, nonuniform, and sparsely sampled structured datasets, a hypersphere CNN learning model based on the k-Nearest Neighbor(kNN) algorithm and CNN is proposed. Through kNN preprocessing, the structural relationship of each sample in the high-dimensional attribute space is established, and the markers of each sample in the neighborhood of the sample are used as attributes to reconstruct the sample set to realize the transformation of the data attribute set from heterogeneous to homogeneous. Subsequently, by reasonably designing the convolution window of the CNN, the marker distribution characteristics of each sample in the neighborhood space are effectively extracted and utilized, and the prediction of unknown samples is completed. Experiments are conducted under different neighborhood scales, soft markers and hard labels, and confusion and non-confusion, and the results show that the prediction accuracy of this model reached 98.04%. Compared with the FC-CNN, CNN, kNN, and Radar-CNN algorithms, the accuracy rate increased by 0.28%~1.66%, and the recall rate increased by 4.78%~31.92%.

**[Key words]** Convolutional Neural Network(CNN); k-Nearest Neighbor(kNN) algorithm; hypersphere convolution; structured data; deep learning

DOI:10.19678/j.issn.1000-3428.0062962

### 0 概述

随着深度学习技术的快速发展,其在人脸识别、机器翻译、自然语言理解、业务管理等领域得到了广

泛的应用<sup>[1-3]</sup>。以卷积神经网络(Convolutional Neural Network, CNN)为代表的深度学习模型主要面向图像<sup>[4-5]</sup>、视频<sup>[6]</sup>、语音<sup>[7]</sup>等数据类型,数据样本

基金项目:国家自然科学基金(61972040);北京联合大学校内科研专项(ZK50201911, ZB10202004)。

作者简介:刘子巍(1996—),男,硕士研究生,主研方向为机器学习、知识图谱;骆 曦,讲师、博士;李 克(通信作者),教授、博士;陈富强,硕士研究生。

收稿日期:2021-10-14 修回日期:2021-11-18 E-mail:like@bnu.edu.cn

的基本特征是均匀致密采样、各属性项性质相同的欧氏域数据,此处称之为均匀同质欧氏数据。

卷积神经网络在图像识别等领域得到有效应用,其中一个重要原因在于其能从输入数据中学习到数据的空间结构特征,即如果一个输入变量与其相邻的输入变量之间的关系比与距其较远的输入变量之间的关系更密切,则可认为这样的数据具有空间结构特征。例如在图像识别中,构成图像某个局部要素(如眼睛)的临近像素点之间的相关性比间隔较远的像素之间相关性更大。因此,可以借助CNN卷积窗的迭代训练来提取这种特征。

除了上述数据之外,在实际应用中还有一大类数据并不具备上述特征,如气象监测数据、用户购物记录以及各类工业企业经营过程中大量产生并留存的业务数据。这类数据通常以二维表的形式存储,表头为各样本的属性项,通常称之为结构化数据。此类数据各属性项的含义、数据类型、取值范围通常有很大区别,具有异质特性,且各样本是在整个高维属性空间中通过非均匀、稀疏采样获得的,但在多数应用中仍可以用欧氏距离来描述样本间的相似性特征,此处称之为非均匀异质欧氏空间数据。在深度学习兴起之前,对这类数据的分析预测通常采用以k近邻(k-Nearest Neighbor, kNN)、决策树、支持向量机(Support Vector Machine, SVM)等为代表的传统机器学习算法。

由于深度学习具有上述优良性质和性能,探索如何将深度学习技术应用于非均匀异质欧氏空间数据和相关场景中的任务是一个具有挑战性且有意义的问题,有助于扩展深度学习的应用领域。目前,已有研究者开展了这方面的尝试,提出了雷达图<sup>[8-9]</sup>、应用不规则卷积核或可变性卷积核<sup>[10-12]</sup>、将分类不平衡的数据应用在卷积神经网络<sup>[13-15]</sup>等方法。

面向异质、非均匀稀疏采样结构化数据集的预测任务,本文提出一种基于k近邻的超球卷积神经网络学习方法。给出一个朴素的假设,即2个样本在属性空间中的距离与其标记的相似性高度相关。在此基础上,利用kNN算法预处理实现异质数据的同质化改造以及邻域样本标记的空间分布特征构造,同时设计多尺度的超球卷积核,从而有效提取目标样本的邻域标记空间分布特征。

## 1 相关工作

### 1.1 非均匀异质欧氏空间数据

在各个领域中,最常见的数据是图片、视频等均匀采样的欧氏空间数据。在均匀欧氏空间数据集中,样本间可以通过一组相同维度属性,由欧氏距离表示样本间和样本内各采样点间的相似性,样本各个维度的属性具有同质特性。在非均匀欧氏空间数据集中,同样是根据样本间一组相同维度属性构成

的欧氏距离表示样本间的相关性,但样本内部属性与属性之间却通常是异质的。

由图1可以看出,在属性空间中,可以用欧氏距离来衡量样本间的相似性,但样本的空间分布为稀疏、非均匀采样的,这样的数据集通常不能满足kNN算法所要求的密采样假设。图2所示的图像是典型的均匀欧氏空间数据,样本中各像素采样点之间具有等距、致密的的空间结构特征,更便于利用卷积操作和神经网络迭代提取局部特征。

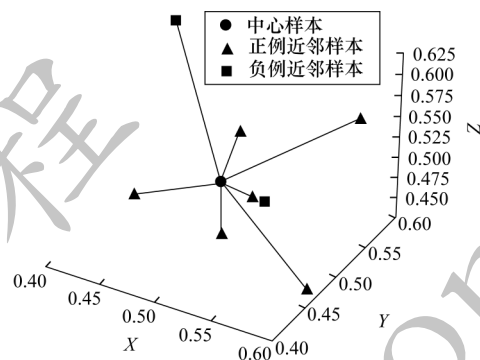


图1 非均匀欧氏空间数据特性

Fig.1 Characteristics of non-uniform Euclidean domain data

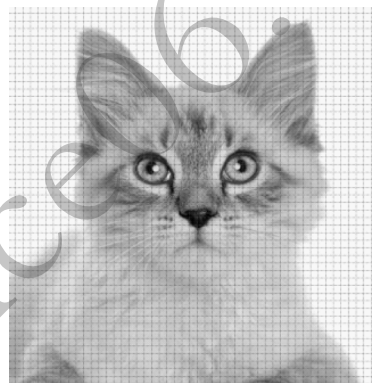


图2 均匀欧氏空间数据示例

Fig.2 Example of uniform Euclidean domain data

### 1.2 面向异质非均匀欧氏空间数据的深度学习方法

对于非均匀异质欧氏空间数据,多采用传统的机器学习方法进行处理。基于局部不变性先验和密采样假设的kNN算法对多数数据集简单有效。为了将kNN扩展到可处理多标记问题,文献[16]提出了ML-kNN算法。与kNN核心思想一致,该算法也是通过寻找一定数量的相似样本来判断测试样本标记,但与kNN不同的是,ML-kNN通过最大化后验概率的方式推理待测样本的标记项,在处理二分类问题时取得了不错的效果。然而,kNN算法在复杂分类边界情况下对于边界附近标记高度混淆样本的分类能力较差,尤其对高度不平衡数据,因为其简单的邻域投票机制无法捕捉稀疏正例样本的邻域空间分布特征,而由文献[17]提出的SVM算法在解决线性可分的二分类问题时获得了较好的效果。

深度学习在均匀欧氏空间数据集的应用中相对于传统机器学习具有明显优势。卷积神经网络由于其特有的平移不变性,在图像识别、语义分割等领域得到广泛应用。文献[18]提出的 GoogLeNet 摒弃了传统单纯依靠层数堆积来增加模型深度的方法,而是利用 inception 模块在增加模型深度与宽度的同时尽可能减少计算量。在此之后,文献[19]提出的 inception-v4 架构将 GoogLeNet 中的 inception-v1 架构与残差连接结合,进一步提升了 inception 架构的稳定性。

鉴于以 CNN 为代表的深度学习在同质均匀采样数据上的优异表现,研究者也围绕如何将这类方法应用于异质数据集展开了研究,这其中最直接的思路是先将异质属性数据进行同质化改造,以适应深度神经网络各层的处理。对此,一种方法是先利用全连接操作将异质属性值转化为同质化的神经元输入,再对神经元进行卷积池化等操作,最后利用 Softmax 分类器进行分类预测<sup>[20]</sup>,称之为全连接 CNN (Fully Connected CNN, FC-CNN)。FC-CNN 适用领域较广,并且能够得到较好的分类预测效果。文献[8]提出一种基于雷达图表示数值型数据的卷积神经网络分类方法,利用 CNN 在图像数据处理领域的优势,先对数值型数据样本进行标准化操作,再将标准化后的  $N$  维属性映射到  $N$  边形雷达图中,实现异质结构数据向同质结构数据的转化,最后将雷达图数据作为图像数据输入到常规的 CNN 模型中进行分类预测。该方法在化工过程数据集中取得了较好的故障分类结果。文献[4]利用 CNN 算法进行心脏病预测,先通过词嵌入方法将非结构化数据转换成向量,再输入 CNN 模型中进行训练和预测,取得了较好的效果。

### 1.3 卷积核的设计

在 CNN 模型中,卷积核的设计和网络模型的构造是影响分类预测效果的关键部分,其中卷积核的设计尤为重要。为了完成特定的学习任务,可以采用不同的卷积核以及卷积方式来提取感受野的数据特征。以图像识别为例,卷积层的卷积核具有良好的平移不变性,即卷积核仅对特定的特征才会有较大激活值。无论上层特征图中的某一特征平移到何处,卷积核都会在该特征处呈现较大的激活值,目前最常用的方法是矩形卷积窗。为了在扩大感受野的同时减少参数降低计算量,文献[21]提出了空洞卷积,使深层的卷积神经网络在扩大感受野的同时减少了参数数量,以这一方法代替传统增加感受野进行降采样而导致丢失分辨率的方法。文献[22]提出转置卷积这一概念,通过转置卷积使网络模型最优化的进行上采样,并且还可用于可视化卷积的过程。该方法在 GAN 等领域中得到大量应用。

卷积神经网络对于输入信息有严格要求,多是规则的矩形排列。文献[10]提出不规则卷积神经网络,将形状与参数与卷积核参数融合,一同在反向传播的过程中更新参数,以此更好地提取贴合待测目标形状的特征信息。该方法在语义分割数据集上获得了较好的效果。

在面向异质非均匀数据的分类预测任务中采用卷积神经网络方法时,一个重要的问题是如何合理设计卷积核以有效提取样本在高维属性空间中邻域样本标记的空间分布特征,本文将对此进行研究。

## 2 邻域标记特征的提取

异质非均匀数据中各个属性往往具有不同的衡量标准,为了消除各个指标之间的量纲影响,需要先进行数据标准化解决数据指标之间的可比性问题。本文利用 min-max 标准化方法对所有属性维度进行归一化处理,如式(1)所示:

$$x^* = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (1)$$

其中: $x^*$  为标准化后的属性值; $x$  为原属性值; $x_{\min}$  和  $x_{\max}$  分别为该属性的最小值和最大值。通过归一化过程,可确保后续邻域标记特征提取与邻域标记样本集构建的准确性。

邻域标记项特征提取的核心思想是利用 kNN 的近邻搜索寻找邻域样本标记,即对待测样本在训练集特征空间中根据欧氏距离寻找一定数量的邻域样本,构成邻域样本集,其中,待测样本的标记取为邻域样本集中占比最大的标记值。kNN 算法简单有效,在处理很多分类问题上性能优异,且对噪声不敏感,但是 kNN 算法在一些特定条件下效果并不好,如在邻域数据集标记项分布较均衡时预测效果就很不理想。为了定量描述近邻样本标记的混淆程度,定义混淆系数为:

$$C_f = \left| 1 - \frac{2n}{k} \right| \quad (2)$$

其中: $n$  为待测样本的同标记近邻数; $k$  为总近邻数; $C_f \in [0, 1]$ 。当近邻中正负例样本数相同时, $C_f = 0$ ,表示完全混淆;当近邻全部为正例或负例时, $C_f = 1$ ,表示无混淆。以北京市空气质量监测采样数据集<sup>[16]</sup>为例,采用 kNN 算法进行预测(取  $k=15$ ),kNN 对于同标记数相同近邻数不同的待测样本的预测准确率如图 3 所示。可以看出,当混淆系数小于 0.3 时,kNN 预测性能明显变差,将此类样本定义为重度混淆样本,对于这类样本,很难利用 kNN 实现准确预测。类似地,将混淆系数在 0.3~0.6 之间的样本定义为中度混淆样本,而将混淆系数大于 0.6 的定义为轻度混淆样本,对于轻度混淆样本,可以利用 kNN 实现准确预测。



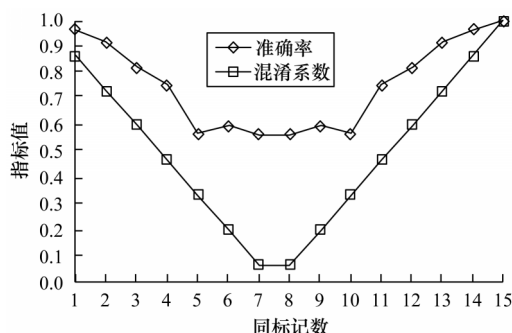


图3 不同近邻混淆条件下的kNN性能差异

Fig.3 Performance difference of kNN under different neighbor confusion conditions

然而,kNN仅利用了邻域样本标记的统计特征而忽略了其空间结构特征。因此,本文利用邻域样本标记项构成的邻域标记样本集所包含的空间结构信息,并且将数据集的属性描述由异质转变为同质,从而更好地利用CNN在处理数据的空间平移不变性特征上的优势。首先定义邻域标记样本如下:

**定义1** 邻域标记样本

已知  $d$  维样本集  $D = \{(x_i, y_i) | 1 \leq i \leq m\}$ ,  $y_i \in Y = \{1, -1\}$ , 设  $D_{ik} = \{(x_{ik}, y_{ik}, d_{ik}) | 1 \leq k \leq K\}$  为中心样本  $x_i$  在属性空间中前  $k$  个邻域样本的属性、标记及其到该中心样本的距离信息集合,各样本按距离排序即  $d_{ik} \leq d_{il} (1 \leq k < l \leq K)$ 。定义  $K$  维邻域标记样本集  $D^* = \{(x_i, y_i | 1 \leq i \leq m\}$ , 其中邻域标记样本为  $Y_i = (y_{i1}, y_{i2}, \dots, y_{iK})$ ,  $d_{ik} \leq d_{il} (1 \leq k < l \leq K)$ 。

本文将邻域标记样本集所包含的丰富的标记空间结构特征信息输入卷积神经网络模型进行分类预测,利用邻域标记样本集增加后续卷积操作的特征输入信息。图4为邻域标记样本集  $D^*$  示例,其中第1列为中心样本的标记  $y_i$ ,第2列为距中心样本最近的邻域样本的标记  $y_{i1} (1 \leq i \leq m)$ ,第2列至最后一列根据欧氏距离升序排列,依次类推,颜色越深表示距中心样本越近。

样本1	1	1	1	1	0	1	0	1	1	1
样本2	1	1	0	1	1	1	1	1	1	1
样本3	0	0	1	1	1	1	1	1	1	1
样本4	1	1	1	0	1	1	1	1	1	1
样本5	1	1	1	1	0	0	0	1	1	1
样本6	0	0	1	1	0	0	0	1	1	1
样本7	1	1	1	1	0	1	1	0	0	1
样本8	1	1	1	1	0	1	0	1	1	1
样本9	1	1	1	1	0	1	0	1	1	1
样本10	1	1	0	1	0	1	0	0	1	1
	$y_i$	$y_{i1}$	$y_{i2}$	$y_{i3}$	$y_{i4}$	$y_{i5}$	$y_{i6}$	$y_{i7}$	$y_{i8}$	$y_{i9}$

图4 邻域标记样本集示例

Fig.4 Example of neighborhood label sample set

由于邻域样本对待测样本的影响通常随距离增大而减小,为了尽可能捕捉不同感受野范围内的邻域标记空间分布特征,本文对邻域标记样本集进行划分,使用不同尺度的邻域标记样本数作为CNN模型的训练输入样本,在弥补kNN对单一  $k$  值选择局限性的同时,增加CNN模型输入样本规模和鲁棒性。图5为邻域标

记样本的邻域尺度示例(彩色效果见《计算机工程》官网HTML版)。可以看出,在原始高维属性空间中,所有邻域样本根据与中心样本间的欧氏距离进行排列并且分布在中心样本附近,图中给出了2个不同邻域尺度(4和9)下邻域样本在原始属性空间的分布和标记值,分别对应图中的黑色和蓝色球形区域。依此类推,可以提取更多尺度的邻域样本信息,构造出对应的邻域标记样本集作为CNN模型的输入数据。

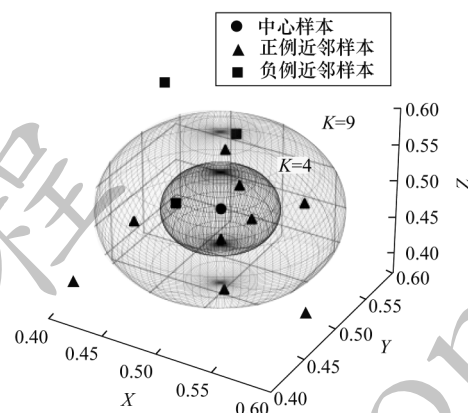


图5 邻域标记样本的邻域尺度

Fig.5 Neighborhood scale of neighborhood label samples

### 3 基于k近邻的超球卷积神经网络学习

#### 3.1 超球卷积

给出超球卷积定义如下:

**定义2** 超球卷积

对于样本集  $D$  及其邻域标记样本集  $D^* \in \mathbb{R}^K$ , 定义一个尺度为  $k (1 < k \leq K)$  的1D卷积核  $W^{(k)} = \{w_1^{(k)}, w_2^{(k)}, \dots, w_k^{(k)}\} \in \mathbb{R}^{k \times 1}$  作为对样本特征空间中各中心样本  $x_i$  的一个超球卷积,即:

$$z_i = \sum_{j=1}^k w_j^{(k)} y_{ij}, 1 \leq i \leq m \quad (3)$$

超球卷积是在  $d$  维特征空间中对中心样本  $x_i$  的  $k$  个邻域样本的标记值进行的卷积操作,具有模型简单、计算量小的优势,具体实现上可通过简单的1D卷积实现,但与常规的1D卷积具有不同的物理意义:超球卷积是在原始高维属性空间中提取待测样本与其邻域样本在标记上的统计特征与空间结构特征,其前提是必须先利用kNN近邻搜索构造出邻域标记样本集作为CNN的输入数据,而不是直接对原始样本进行卷积,其通过选取不同的尺度值  $k$  进行卷积,实现对不同范围内邻域样本的标记与空间特征的提取。

超球卷积在原始数据样本的特征空间中采用CNN而不是利用贝叶斯原理从概率的角度去估计中心(未知)样本的类标记,其通过卷积核的迭代训练来提取局部特征,从而利用CNN卷积的平移不变性来提取中心样本与邻域样本标记值的相关性特征。超球卷积中的平移不变性提取与CNN图像识别的处理不同,超球卷积是对整个属性空间进行局部采样构成邻域标记样本集,通过对每个邻域标记样本利用多个尺度卷积核进行一次性的多副本卷积,实现在整个属性空间中的特征提取。

### 3.2 HCNN 网络模型设计

基于传统CNN模型,本文引入超球卷积设计超球卷积神经网络(Hypersphere CNN, HCNN)模型。HCNN将多种不同尺度的邻域标记样本集作为输入,对每条样本的邻域按升序排列并分别提取距中心样本最近的前 $2^2, 3^2, \dots, n^2$ 个邻域标记样本构成 $n-1$ 种尺度的输入数据。图6为提取4种尺度(即 $k=4, 9, 16, 25$ )邻域标记样本的示意图(彩色效果见《计算机工程》官网HTML版)。

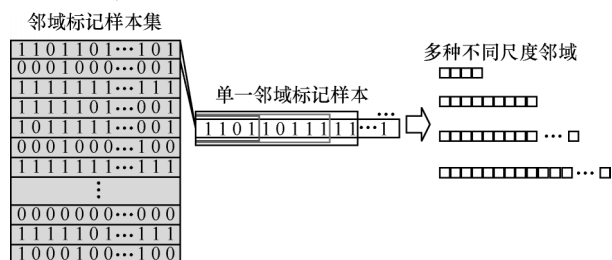


图6 多尺度邻域标记样本提取

Fig.6 Extraction of multi-scale neighborhood label samples

1)超球卷积层。超球卷积层用于提取各种不同尺度的邻域标记的空间特征,可以用1D全连接来实现,操作流程如图7所示,图中采用了4种尺度,每种尺度均按照式(3)进行400个副本的超球卷积,形成 $(n-1) \times 400$ 的输出矩阵,并将其转变为每层尺寸为 $(20, 20)$ 深度为 $n-1$ 层的张量。

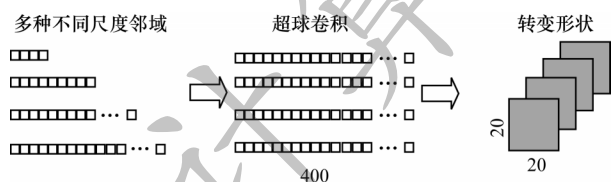


图7 超球卷积层操作流程

Fig.7 Operation procedure of hypersphere convolution layer

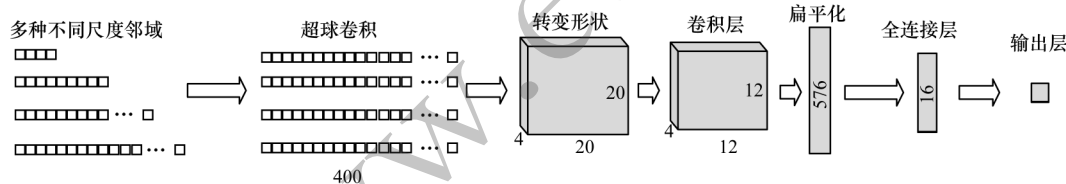


图9 HCNN模型整体架构

Fig.9 Overall architecture of HCNN model

### 3.3 HCNN 算法描述

HCNN模型中主要分为三步:构建邻域标记样本集,HCNN模型训练和利用HCNN模型进行分类预测。具体算法描述如下:

#### 算法1 HCNN

输入 原始数据集 $D$ ,迭代次数 $m$ ,最大邻域范围 $K$ , Batchsize $b$ ,样本总数 $S$

输出 预测的待测样本标记项集合 $\{\hat{y}\}$

1. 构建邻域标记样本集 $D^* = \{(x_i, y_i) | 1 \leq i \leq s\}$

2. 将 $D^*$ 划分为训练集 $D_{\text{train}}^* = \{(x_i, y_i) | 1 \leq i \leq s_1\}$ 和测试集 $D_{\text{test}}^* = \{(x_i, y_i) | s_1 < i \leq s_2\}$

2)卷积层。由于各层张量代表不同尺度的邻域标记样本,为了使各种尺度提取的特征具有不同的意义以及权重,卷积方式采用深度可分离卷积。深度可分离卷积层与传统卷积层不同之处在于一个卷积核只提取一个通道(层)的特征提取,将各个尺度的邻域分别进行超球卷积以此来对各种尺度的邻域标记进行特征提取。卷积层由依次利用 $3 \times 3, 3 \times 3$ 和 $5 \times 5$ 的卷积核且步长均为1的3个深度可分离卷积层所构成,3次深度可分离卷积具体操作如图8所示,在每层深度可分离卷积后加入一层Batch Normalization来加快学习收敛速度,同时避免梯度消失。

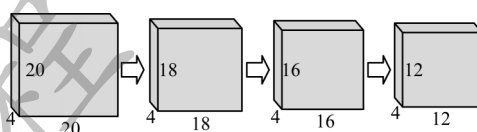


图8 3次深度可分离卷积

Fig.8 Cubic depth separable convolution

3)全连接层。全连接层的目的是利用提取出的特征进行有效分类。该层由普通的多层普通神经网络组成,激活函数采用SELU函数,并且利用Dropout防止过拟合现象。

4)输出层。输出层利用全连接层实现,用于输出模型判断出的该样本的正例概率,激活函数采用Sigmoid函数。

5)损失函数。损失函数用于衡量预测值和真实值之间的差距,并用于更新全连接层和卷积层参数,由于本文是针对二分类问题,因此采用二元交叉熵作为损失函数。

以4、9、16、25这4种尺度为例,HCNN模型整体架构如图9所示。

3.for  $i$  in range( $m$ ):

4.for  $j$  in range( $S_1$ ):

5.根据二元交叉熵损失函数计算前向网络各层残差与梯度

6.if  $s_1 \% b == 0$ :

7.根据各层计算梯度以及更新残差与窗参

8.end if

9.end for

10.end for

11.利用训练好的HCNN模型对测试集 $D_{\text{test}}^*$ 进行预测,得到各样本的正例概率集合 $\{\hat{y}\}$

12.对 $\{\hat{y}\}$ 进行硬判决,得到待测样本标记集 $\{\hat{y}\}$

## 4 实验结果与分析

### 4.1 实验数据集

本文实验所使用的数据为北京市昌平气象监测站2013年3月1日—2017年2月28日每小时的空气

污染物监测采样数据<sup>[23]</sup>。数据集总共包含32 681条样本,数据样例如表1所示。数据样本属性包括年、月、日、时、SO<sub>2</sub>浓度、NO<sub>2</sub>浓度、CO浓度、O<sub>3</sub>浓度、温度、压强、露点温度、降水量和风速。

表1 北京市空气污染物监测数据样例

Table 1 Samples of air pollutant monitoring data in Beijing

年	月	日	时	SO <sub>2</sub> 浓度 /(μg·m <sup>-3</sup> )	NO <sub>2</sub> 浓度 /(μg·m <sup>-3</sup> )	CO浓度 /(μg·m <sup>-3</sup> )	O <sub>3</sub> 浓度 /(μg·m <sup>-3</sup> )	温度/°C	压强/hPa	露点温度 /°C	降水量 /mm	风速 /(m·s <sup>-1</sup> )	PM2.5浓度 /(μg·m <sup>-3</sup> )
2013	3	1	0	13	7	300	85	-2.3	1 021	-19.7	0	0.5	3
2013	3	1	1	16	6	300	85	-2.5	1 021	-19.0	0	0.7	3
2013	3	1	2	22	13	400	74	-3.0	1 021	-19.9	0	0.2	3
2013	3	1	3	12	8	300	81	-3.6	1 022	-19.1	0	1.0	3
2013	3	1	4	14	8	300	81	-3.5	1 022	-19.4	0	2.1	3
2013	3	1	5	10	17	400	71	-4.5	1 023	-19.5	0	1.7	3
2013	3	1	6	12	22	500	65	-4.5	1 023	-19.5	0	1.8	4
2013	3	1	7	25	39	600	48	-2.1	1 025	-20.0	0	2.5	3
2013	3	1	8	13	42	700	46	-0.2	1 025	-20.5	0	2.8	9
2013	3	1	9	5	18	500	73	0.6	1 025	-20.4	0	3.8	11

取PM2.5浓度为标记项。图10为PM2.5的概率密度分布,可见出现PM2.5浓度小于200的出现天数占绝大部分,因此,将PM2.5浓度大于200(对应中重度污染)设为正例,共2 116条样本,在数据集中占6.47%,属于典型的不平衡数据集。作为对照实验,同时取门限50 μg/m<sup>3</sup>(设PM2.5浓度小于50为正例)得到一套平衡数据集,其中正例共17 221条样本(占比52.69%)。

图11为各属性与标记项之间相关系数的统计分布,可见SO<sub>2</sub>、NO<sub>2</sub>、CO浓度这三个属性与标记项之间以及这三个属性间均有较高的相关性。

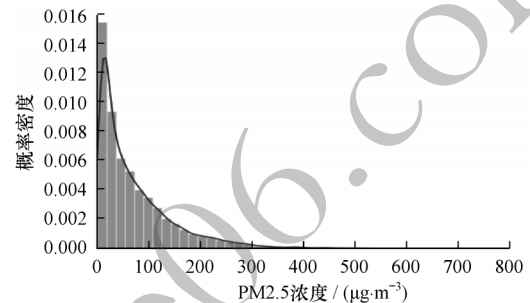


图10 PM2.5浓度的概率密度分布  
Fig.10 Probability density distribution of PM2.5 concentration

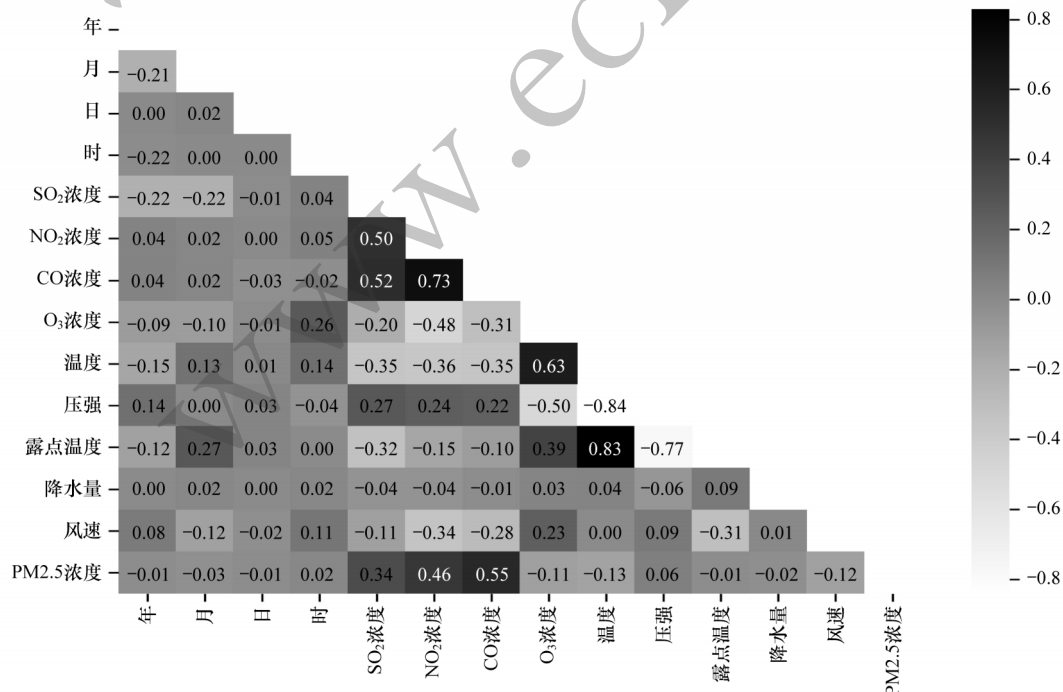


图11 属性及标记项间的相关系数统计分布

Fig.11 Statistical distribution of correlation coefficients among attributes and label items



## 4.2 性能评价准则

由于本文研究的问题属于分类问题,因此评价指标采用准确率、精度、召回率和F1测度。以下实验均使用Python语言在PyCharm平台上完成。

## 4.3 算法实验分析

本文将HCNN算法与kNN、FC-CNN、Radar-CNN等算法进行对比,并对算法各关键参数的选取进行实验对比。此外,为了对比HCNN在不采用kNN预处理时的性能变化,采用普通CNN模型进行消融实验,其神经网络模型与HCNN基本相同,区别在于对原始的异质数据不经过异质化,而是直接进行卷积处理,下文中称为CNN算法。将训练集与测试集比例设定为8:2,采用5折交叉验证的方法得到最终的性能评价结果。

### 4.3.1 HCNN实验结果

HCNN模型的训练过程以及收敛曲线如图12所示,其中,图12(a)和图12(b)为训练集的残差与准确率收敛曲线,可见训练集loss值和准确率在迭代50次左右时即快速收敛,图12(c)和图12(d)为验证集的残差与准确率收敛曲线,同样在迭代50次左右逐渐收敛趋于平滑。由此可见HCNN模型并没有产生明显的过拟合现象,可有效地对未知样本进行预测。

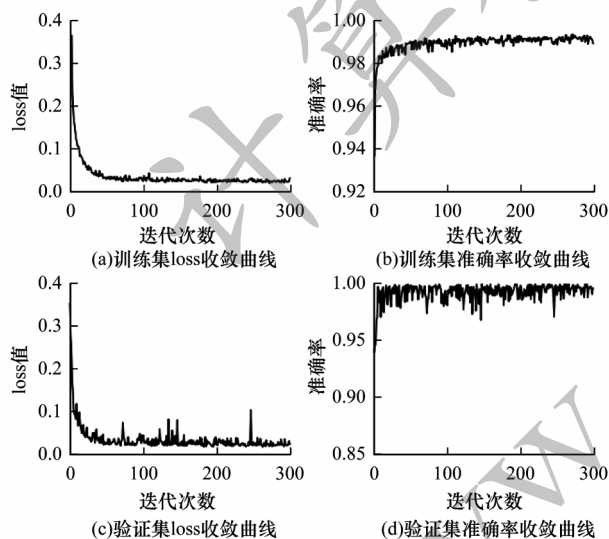


图12 HCNN模型训练的残差与准确率收敛曲线

Fig.12 Convergence curve of loss and accuracy of HCNN model training

### 4.3.2 超球卷积核最大尺度的选取

邻域样本与中心样本的距离关系如图13所示(彩色效果见《计算机工程》官网HTML版),其中,图13(a)是从全部训练数据集中随机抽取200个样本(不同颜色的曲线代表不同样本与其邻域样本间的欧式距离曲线),取最大邻域尺度为200,分别计算各样本与其邻域样本的欧式距离并按升序排列绘制的曲线。可以明显看出,当邻域尺度 $k$ 取到25以后,中心样本与邻域样本的距离差逐渐趋于平缓。为了

评估训练样本规模的影响,从训练集中随机抽取5 000个样本作为新的训练集进行近邻搜索,结果如图13(b)所示。可以看出,在减小训练集规模后,中心样本与邻域样本间的距离依旧在尺度为25左右开始趋于平缓。因此,本文实验中取最大尺度数为25。选取更大范围的邻域样本参与训练不仅会导致计算量过大,而且可能因为领域样本的区分度降低而引入不必要的干扰。对于不同的应用数据集,仍建议将最大尺度数作为一个重要的超参数,通过上述分析找到最优值。

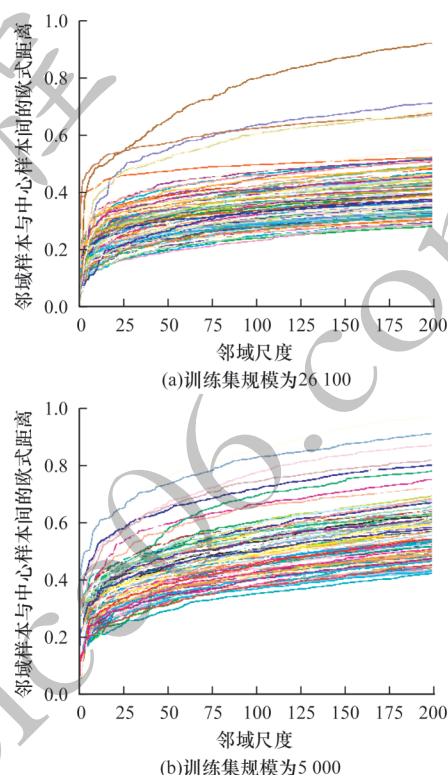


图13 邻域样本与中心样本的距离关系

Fig.13 Distance relationship between neighbor sample and center sample

图14是邻域样本不同尺度数下的预测准确率对比。实验采取的邻域尺度分别为 $k=2^2, 3^2, \dots, n^2$ 。尺度数为1就选邻域尺度为 $k=2^2$ 尺度的邻域标记样本作为输入,尺度数为2就选 $k=2^2, 3^2$ 这2个尺度的邻域标记样本作为输入,依此类推。可以看出,当模型输入取邻域尺度数为4时,HCNN模型分类预测准确率最高。

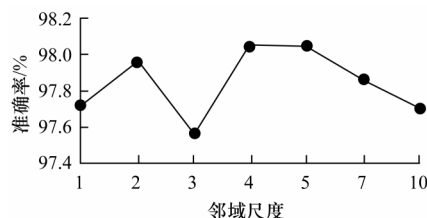


图14 不同邻域尺度下的准确率对比

Fig.14 Comparison of accuracy under different neighbor scales

#### 4.3.3 软硬邻域标记的实验对比

在邻域标记样本集中,各属性项是每个邻域样本的标记值(即PM2.5浓度是否超标,需要将实际的数值通过门限判决转为硬标记,即 $\pm 1$ ),这样输入到HCNN模型的数据中就损失了PM2.5浓度的细节信息,因此,也可以考虑保留各邻域样本的PM2.5浓度原始值,即软标记值。

图15为HCNN算法在邻域标记样本集分别采用软硬标记值的各项指标对比。可以看出,软标记输入所得的各项指标略好于硬标记输入,因此,本文使用邻域样本的软标记值作为模型训练的输入。

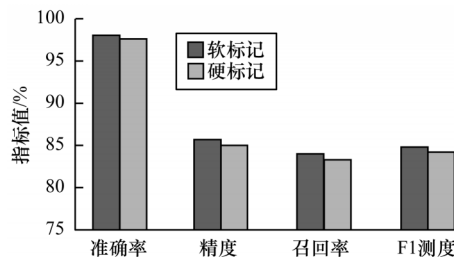


图15 邻域样本软硬标记值的HCNN性能对比

Fig.15 Comparison of HCNN performance with soft and hard labels of neighbor sample

#### 4.3.4 HCNN算法在混淆样本中的性能分析

由于kNN算法依靠邻域投票机制进行分类预测,因此往往无法解决混淆样本的分类问题(如图3所示)。在高度不平衡样本集中,这种情况更为突出。为此,分别针对平衡样本集和不平衡样本集对比HCNN和kNN算法在不同的混淆条件下的性能,实验结果如图16所示。

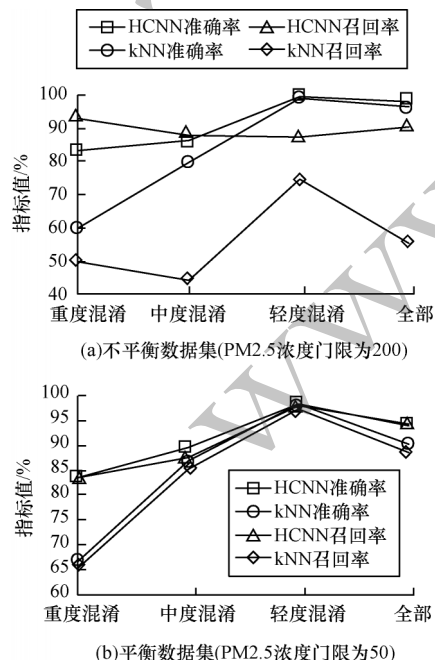


图16 HCNN与kNN在不同混淆条件下的预测性能对比

Fig.16 Comparison of prediction performance between HCNN and kNN under different confusion conditions

由图16可以看出,HCNN对混淆样本的预测效果要远好于kNN算法,体现了超球卷积在提取复杂邻域标记空间结构特征上的优异能力。

#### 4.3.5 HCNN与其他算法的性能对比

将HCNN与kNN、Radar-CNN、FC-CNN、CNN这4种算法进行性能对比,实验结果如图17所示。可以看出,HCNN的预测准确率略高于其他3种算法,预测精度kNN最高、HCNN算法排列第二。由于在不平衡数据集中准确率失真,因此主要对比召回率等指标,可以看出,HCNN算法显著优于其他算法,体现出HCNN在不平衡样本集上的性能优势。

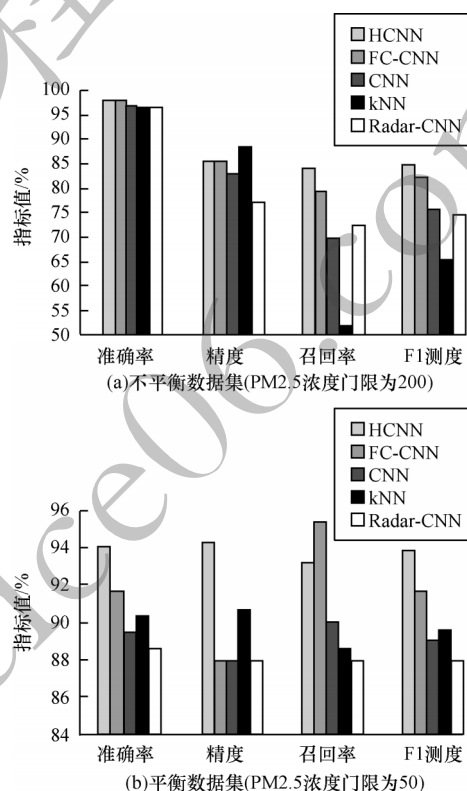


图17 不同算法的性能指标对比

Fig.17 Performance comparison of different algorithms

通过与CNN算法的消融实验对比可以发现,利用kNN预处理进行近邻结构信息的提取后能够显著提升模型性能,由此可见,对于异质数据集,利用kNN进行样本同质化是一种有效的手段。

## 5 结束语

本文提出一种结合kNN和CNN的深度学习模型。利用kNN的近邻搜索,将仅携带多维属性信息的异质数据样本集转化为携带邻域样本标记信息和属性空间结构特征的同质化数据集。在此基础上,通过对邻域样本标记进行超球卷积的方式提取未知样本的邻域标记特征,进而对该样本的标记进行预测。实验结果表明,该模型能够有效提取未知样本



的邻域样本标记特征并对该样本的标记进行准确预测, 与kNN、CNN、FC-CNN、Radar-CNN等算法相比预测性能更优。下一步将分析HCNN在多种数据噪声环境、尺度变换(平移/旋转等)条件下的性能, 并验证其在处理多标签分类和回归问题时的预测性能。

### 参考文献

- [1] 刘建伟, 谢浩杰, 罗雄麟. 生成对抗网络在各领域应用研究进展[J]. 自动化学报, 2020, 46(12): 2500-2536.  
LIU J W, XIE H J, LUO X L. Research progress on application of generative adversarial networks in various fields[J]. Acta Automatica Sinica, 2020, 46(12): 2500-2536. (in Chinese)
- [2] 付建平, 赵海燕, 曹健, 等. 面向业务过程异常检测的深度学习模型BPAD-LS[J]. 小型微型计算机系统, 2022, 43(5): 902-912.  
FU J P, ZHAO H Y, CAO J, et al. Deep learning model BPAD-LS for business process anomaly detection[J]. Journal of Chinese Computer Systems, 2022, 43(5): 902-912. (in Chinese)
- [3] 黄义妨, 魏丹丹, 武森, 等. 面向不同传感器与复杂场景的人脸识别系统防伪方法综述[J]. 计算机工程, 2021, 47(12): 1-18.  
HUANG Y F, WEI D D, WU M, et al. Overview of anti-spoofing methods of face recognition systems for different sensors and complex scenes[J]. Computer Engineering, 2021, 47(12): 1-18. (in Chinese)
- [4] 杨涵方, 周向东. 基于深度稀疏辨别的跨领域图像分类[J]. 计算机工程, 2018, 44(4): 310-316.  
YANG H F, ZHOU X D. Cross domain image classification based on deep sparse discrimination[J]. Computer Engineering, 2018, 44(4): 310-316. (in Chinese)
- [5] 申铨京, 张雪峰, 王玉, 等. 像素级卷积神经网络多聚焦图像融合算法[J]. 吉林大学学报(工学版), 2022, 52(8): 1857-1864.  
SHEN X J, ZHANG X F, WANG Y, et al. Multi-focus image fusion algorithm based on pixel-level convolutional neural network[J]. Journal of Jilin University (Engineering and Technology Edition), 2022, 52(8): 1857-1864. (in Chinese)
- [6] 杨春玲, 凌茜. 基于深度学习的两阶段多假设视频压缩感知重构算法[J]. 华南理工大学学报(自然科学版), 2021, 49(6): 88-99.  
YANG C L, LING X. Two-stage multi-hypothesis network for compressed video sensing reconstruction algorithms based on deep learning[J]. Journal of South China University of Technology (Natural Science Edition), 2021, 49(6): 88-99. (in Chinese)
- [7] 蓝天, 彭川, 李森, 等. 基于RefineNet的端到端语音增强方法[J]. 自动化学报, 2022, 48(2): 554-563.  
LAN T, PENG C, LI S, et al. RefineNet-based end-to-end speech enhancement[J]. Acta Automatica Sinica, 2022, 48(2): 554-563. (in Chinese)
- [8] 程诚, 任佳. 一种基于雷达图表示的数值型数据的CNN分类方法[J]. 信息与控制, 2019, 48(4): 429-436.  
CHENG C, REN J. A classification method of CNN for numerical data based on radar chart representation[J]. Information and Control, 2019, 48(4): 429-436. (in Chinese)
- [9] MONTI F, BOSCAINI D, MASCI J, et al. Geometric deep learning on graphs and manifolds using mixture model CNNs[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2017: 5425-5434.
- [10] MA J B, WANG W, WANG L. Irregular convolutional neural networks[C]//Proceedings of the 4th IAPR Asian Conference on Pattern Recognition. Washington D. C., USA: IEEE Press, 2017: 268-273.
- [11] SHANKAR V, KUMAR V, DEVAGADE U, et al. Heart disease prediction using CNN algorithm[J]. SN Computer Science, 2020, 1(3): 1-8.
- [12] DAI J F, QI H Z, XIONG Y W, et al. Deformable convolutional networks[C]//Proceedings of 2017 IEEE International Conference on Computer Vision. Washington D. C., USA: IEEE Press, 2017: 764-773.
- [13] BUDA M, MAKI A, MAZUROWSKI M A. A systematic study of the class imbalance problem in convolutional neural networks[J]. Neural Networks, 2018, 106: 249-259.
- [14] GUO S N, LIN Y F, LI S J, et al. Deep spatial-temporal 3D convolutional neural networks for traffic data forecasting[J]. IEEE Transactions on Intelligent Transportation Systems, 2019, 20(10): 3913-3926.
- [15] KO Y, HSU P, CHENG M, et al. Customer retention prediction with CNN[C]//Proceedings of the 4th International Conference on Data Mining and Big Data. Chiang Mai, Thailand: [s. n.], 2019: 104-113.
- [16] ZHANG M L, ZHOU Z H. ML-KNN: a lazy learning approach to multi-label learning[J]. Pattern Recognition, 2007, 40(7): 2038-2048.
- [17] VAPNIK V. The natural of statistical learning theory[M]. Berlin, Germany: Springer, 1995.
- [18] SZEGEDY C, LIU W, JIA Y Q, et al. Going deeper with convolutions[C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2015: 1-9.
- [19] SZEGEDY C, IOFFE S, VANHOUCKE V, et al. Inception-v4, inception-ResNet and the impact of residual connections on learning[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2016: 1-12.
- [20] CSDN中文开发者社区. 基于CNN的鸢尾花分类器[EB/OL]. [2020-08-15]. [https://blog.csdn.net/qq\\_38581886/article/details/107607743](https://blog.csdn.net/qq_38581886/article/details/107607743).
- [21] YU F, KOLTUN V. Multi-scale context aggregation by dilated convolutions[EB/OL]. [2020-08-15]. <https://arxiv.org/abs/1511.07122>.
- [22] RADFORD A, METZ L, CHINTALA S. Unsupervised representation learning with deep convolutional generative adversarial networks[EB/OL]. [2020-08-15]. <https://arxiv.org/abs/1511.06434>.
- [23] SONG X C. UC irvine machine learning repository, Beijing PM2.5 data set[EB/OL]. [2020-08-15]. <http://archive.ics.uci.edu/ml/datasets/Beijing+PM2.5+Data>.