计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200900013

基于代价敏感卷积神经网络的非平衡问题混合方法

黄颖琦　陈红梅
西南交通大学信息科学与技术学院　成都 611756
西南交通大学云计算与智能技术高校重点实验室　成都 611756
(Huangyingqi@my.swjtu.edu.cn)

摘　要　非平衡问题是数据挖掘领域中普遍存在的一个问题,数据的偏态分布会使得分类器的分类效果不理想。卷积神经网络作为一种高效的数据挖掘工具,被广泛应用于分类任务,但其训练过程若受到数据非平衡的不利影响,则将导致少数类的分类准确率下降。针对二分类非平衡数据分类问题,文中提出了一种基于代价敏感卷积神经网络的非平衡问题混合方法。首先将密度峰值聚类算法与SMOTE相结合,通过过采样对数据进行预处理,降低原始数据集的不平衡程度;然后利用代价敏感思想对非平衡数据中的不同类别给予不同权重,并考虑预测值与标签值之间的欧氏距离,对非平衡数据中多数类和少数类赋予不同的代价损失,构建代价敏感卷积神经网络模型,以提高卷积神经网络对少数类的识别率。选取6个不同的数据集,用于验证所提方法的有效性。实验结果表明,所提方法可以提高卷积神经网络模型对非平衡数据的分类性能。

关键词:非平衡问题;卷积神经网络;过采样;数据预处理;代价敏感损失函数
中图法分类号　TP391

Cost-sensitive Convolutional Neural Network Based Hybrid Method for Imbalanced Data Classification

HUANG Ying-qi and CHEN Hong-mei
School of Information Science and Technology, Southwest Jiaotong University, Chengdu 611756, China
Key Laboratory of Cloud Computing and Intelligent Technology, Southwest Jiaotong University, Chengdu 611756, China

Abstract　The imbalance classification is a common problem in the field of data mining. In general, the skewed distribution of data makes the classification effect of the classifier unsatisfactory. As an efficient data mining tool, convolutional neural network is widely used in classification tasks. However, if the training process is adversely affected by data imbalance, it will cause the classification accuracy of minority classes to decrease. Aiming at the classification problem of two-class unbalanced data, this paper proposes a hybrid method for unbalanced classification problems based on cost-sensitive convolutional neural networks. The proposed method first combines the density peak clustering algorithm with SMOTE, and preprocesses the data through oversampling to reduce the imbalance of the original data set. Then the cost sensitive is used to give different weights to different categories in the unbalanced data. Additionally, the Euclidean distance between the predicted value and the label value is considered. The proposed method assigns different cost losses to the majority class and the minority class in the unbalanced data to construct cost sensitivity convolutional neural network model to improve the recognition rate of convolutional neural network for minority classes. Six different datasets are used to verify the effectiveness of the proposed method. The experimental results show that the proposed method is able to improve the classification performance of the convolutional neural network model on unbalanced data.

Keywords　Imbalance classification, Convolutional neural network, Oversampling, Data preprocessing, Cost-sensitive loss function

1　引言

在现实生活的各个应用领域中,非平衡数据普遍存在,如医学疾病诊断、银行欺诈检测、网络入侵检测、机械设备故障诊断等[1-4]。非平衡问题指在一个分类任务中,某些类别的样本数远多于其他类别的样本数[5],使得数据呈现出偏态分布。其中,拥有大量样本数量的类被称为多数类,拥有少量样本数量的类被称为少数类。

在分类任务中,若将传统分类器应用于非平衡数据集,为了提高整体的分类精度,分类器会减少对少数类的关注,而偏

到稿日期:2020-09-02　返修日期:2021-01-21
基金项目:国家自然科学基金(61976182,62076171);四川省国际科技创新合作重点项目(2019YFH0097)
This work was supported by the National Natural Science Foundation of China(61976182,62076171) and Key Program for International S&T Cooperation of Sichuan Province(2019YFH0097).
通信作者:陈红梅(hmchen@swjtu.edu.cn)

向于多数类，导致少数类样本难以被识别出来，从而得不到好的分类效果。然而，在数据挖掘中，少数类往往包含更有价值的信息。因此，非平衡问题具有重要的研究意义，近年来越来越受到研究学者的广泛关注，逐渐成为了数据挖掘的研究热点。

目前，研究学者们针对非平衡问题相继提出了不同的解决方法，主要包括数据层面、特征选择和算法层面的方法。

数据层面的方法通过对原始非平衡数据集进行重采样，改变数据的分布，来降低类与类之间样本的不平衡程度，主要包括欠采样、过采样以及两者结合的混合采样方法[6-7]。欠采样方法通过移除多数类样本来调整原始数据集的类别分布，而过采样方法则通过增加少数类样本来达到平衡数据的目的。这些方法从数据层面调整数据的分布，避免了对分类算法的修改，因此独立于分类器的训练，往往更具通用性[8]。实验结果表明，将重采样方法作为一个预处理步骤用于平衡数据分布是一种积极有效的解决方案[9-10]。

特征选择[11-12]方法基于某种规则从原始数据特征集合中选择出区分能力最终的的相关特征子集，降低数据维度，从而提高学习算法的分类性能。常见的特征选择方法大致分为3类：过滤式(filter)、包裹式(wrapper)和嵌入式(embedding)[13]。

算法层面的方法主要有代价敏感学习、集成学习和单类学习[14-16]等方法，通过设计适用于非平衡数据特征的模型训练算法来解决非平衡问题。代价敏感学习方法针对传统分类算法假设所有类别的误分类代价相等而被提出，通过为不同的类赋予不同的误分类代价来构造分类器[17]。在非平衡问题中，需要对少数类予以更多关注，因此一般会为少数类赋予更高的误分类代价[18]。集成学习方法通过将多个分类器集成来构建一个新的分类器，以提高分类性能。基于Boosting和基于Bagging的算法是集成学习的两类代表方法[19-20]。单类学习方法指仅使用一个特定类别的样本进行模型学习的分类方法[21]。

算法层面的方法多种多样，除了以上介绍的方法外，神经网络也被很多学者用来处理非平衡问题。其中，卷积神经网络以其在图像分类、故障诊断、自然语言处理等各个应用领域中的出色表现而被广泛关注。与其他深度的前馈神经网络相比，卷积神经网络能用更少的参数获得更好的性能，是一种高效的数据挖掘方法。因此，越来越多的学者将其应用于数据挖掘领域中，用于解决非平衡问题。

卷积神经网络将自动特征提取和判别分类器集成在一个模型中，这是它与传统机器学习方法的主要区别[22]。标准的卷积神经网络由输入层、卷积层、池化层、全连接层和输出层等组成。其中，卷积层是卷积神经网络的核心层，每个卷积层都包含多个特征映射，并通过卷积滤波器对输入数据进行特征提取。通过对输入数据进行逐层加工，卷积神经网络把初始的“低层”特征表示转化为抽象的“高层”特征表示，进而完成分类学习任务。

神经网络的学习过程是从训练数据中自动获取最优权重参数的过程。在这个过程中，神经网络以损失函数为指标，通过不断更新权重参数来使损失函数的值尽可能小。卷积神经网络一般用交叉熵损失作为损失函数，这也是分类问题中比较广泛使用的一种损失函数。但是，在使用非平衡数据训练卷积神经网络时，训练出的模型的分类效果却并不理想。其原因在于交叉熵损失函数没有考虑数据的非平衡因素，将不同类别的代价损失设为等同。

为了解决以上问题，我们首先基于数据层面，从数据预处理角度提出了一种新的过采样方法——DPCSMOTE。该方法将密度峰值聚类算法[23]与SMOTE相结合，通过过采样对数据进行预处理，目的是降低原始数据集的不平衡程度。密度峰值聚类算法是一种基于密度的聚类算法，在精度、自动检测聚类簇数和识别中心点等方面具有较好的性能表现[24]。密度峰值聚类算法的运用使得SMOTE算法能够识别输入空间中生成人工数据最有效的区域。因此，提出的采样方法DPCSMOTE能够缓解样本的类间和类内不平衡，同时避免产生噪声样本。采用SMOTE算法进行过采样也能避免产生过拟合问题。

另一方面，在算法层面对卷积神经网络的交叉熵损失函数进行改进，针对非平衡问题提出了一种新的损失函数。该损失函数结合代价敏感思想，对非平衡数据中的不同类别赋予不同的权重，同时考虑预测值与标签值之间的欧氏距离，使得损失函数更加关注少数类，从而提高了卷积神经网络模型对少数类的识别率。

本文第2节总结了相关工作，介绍了现有的过采样算法，并对卷积神经网络损失函数的各种改进方法做了详尽的介绍；第3节详细解释了提出的方法；第4节是实验部分，介绍了实验数据集、采用的评价指标和实验设置，并给出了实验结果及其分析；最后总结全文。

2 相关工作

2.1 数据预处理

预处理对于神经网络的有效性已在提高识别性能和学习效率等实验中得到证明。目前学者们研究的数据预处理方法主要分为两种类型：第一种是通过改变数据分布得到平衡数据集；第二种是根据误分类代价信息更改训练集的分布。后者是一种代价敏感学习方法，其主要缺点是强加了定义误分类代价的需求，这通常是无法在数据集中得到的。通过改变数据分布得到平衡数据集是常用的数据预处理方法，其中重采样技术是学者们研究较多的预处理方法。通过对原始数据集进行重采样，改变数据的分布，来降低样本中类与类之间的不平衡程度。

重采样技术主要包括欠采样和过采样两种类型。过采样方法中最简单的处理方法是随机过采样，通过简单复制少数类样本使得数据平衡分布。但是，随机过采样生成的新数据集和原始数据集太相似，因此容易产生过拟合问题。为解决这一问题，学者们提出了改进方法，其中最有代表性的是Chawla等[25]提出的SMOTE方法。SMOTE方法利用线性插值的思想生成新的少数类样本，而不是简单的复制现有样本，因此能在一定程度上解决过拟合问题。但由于没有区分类重叠区域与安全区域，因此SMOTE方法可能会引入噪声数据。此外，SMOTE算法可能会加剧类内不平衡。因此，学者们不断对SMOTE算法进行改进和扩展。

针对噪声问题，Douzas等[26]在SMOTE方法的基础上提

出了一种几何 SMOTE 方法(G-SMOTE)。该方法在每个选定的少数类样本周围定义一个安全的几何区域来合成人工样本，避免产生噪声数据。在非平衡数据中，少数类样本的密度分布情况不一，难以保证安全区域内不会产生噪声数据。为此，Pan 等[27]提出了一种自适应 SMOTE 方法。该方法基于原始数据分布特征自适应地将少数类样本分为 Inner 和 Danger 两个子集，然后利用 SMOTE 方法合成新的样本来平衡原始数据集，从而防止分类边界扩展，加强了原始数据的分布特征。

针对类内不平衡问题，学者们提出将聚类方法与 SMOTE 方法相结合的解决思路。因为聚类的使用能够使 SMOTE 算法识别出合成人工样本最有效的区域，所以两者相结合能够有效解决类间和类内不平衡的问题。Douzas 等[28]提出了 SOMO 算法(Self-Organizing Map Oversampling)。这种方法应用 SOM(Self-Organizing Map)为新的样本合成指定了安全有效的区域，根据簇密度调整簇内和簇间合成数据的分布。Douzas 等[29]将 K-Means 聚类和 SMOTE 算法结合起来，提出了一种新的简易高效的过采样方法。K-Means 聚类方法的使用能够使 SMOTE 算法识别出生成人工数据最有效的区域，从而消除类间和类内的不平衡。Gong 等[30]提出了一种基于聚类的过采样方法 KMFOS。KMFOS 方法采用 K-Means 算法将少数类划分为 K 簇，然后在每两个簇的样本之间插入新样本，以缓解类内不平衡。

综上所述，近年来很多研究致力于对 SMOTE 方法进行改进。其中将聚类和 SMOTE 过采样相结合的方法在消除类间和类内不平衡的同时，避免了噪声数据的产生，是解决非平衡问题的有效方法。将其作为数据的预处理方法可以有效缓解数据集的不平衡程度，降低不平衡因素对分类器性能的影响。

2.2 卷积神经网络中的非平衡问题

卷积神经网络是深度学习的代表算法之一，近年来因其在图像分类、故障诊断、自然语言处理等应用领域中的出色表现而备受关注。卷积神经网络具有表征学习能力，是一种高效的数据挖掘工具。因此，越来越多的研究学者将卷积神经网络用于解决非平衡问题。

神经网络的学习以损失函数为指标，通过更新权重参数，使损失函数的值最小化。因此，应用卷积神经网络解决非平衡问题可从其损失函数着手。目前，针对非平衡问题，学者们提出了很多卷积神经网络损失函数的改进方案。

代价敏感方法对多数类和少数类赋予不同的误分类代价，将其与损失函数结合可有效解决非平衡问题。Khan 等[31]提出了一种代价敏感的卷积神经网络。针对传统代价矩阵存在的训练过程不稳定、容易导致误差函数不收敛等问题，文献[31]提出了一种新的代价矩阵，并将其与均方误差、交叉熵误差等常见的损失函数相结合，构造成新的损失函数。在代价敏感学习中确定误分类代价需要足够的先验知识，很难准确地设置其值。针对此问题，Geng 等[32]提出了一种自适应代价敏感学习策略，用于改进深度学习模型。通过这种策略，分类器可以自动为每个类分配误分类代价。Jia 等[33]针对机械智能故障诊断问题，提出了一种深度归一化的卷积神经网络机械故障分类框架(Deep Normalized Convolutional Neural Network，DNCNN)。在损失函数中根据数据的不平衡程度对少数类进行加权，突出少数类对卷积神经网络训练过程的影响。在非平衡问题中，少数类样本往往包含更有价值的信息。因此，提高少数类的识别率越来越受到研究学者们的关注。

此外，还有学者提出了新的损失函数。Chen 等[34]引入了三元组损失函数(triplet loss)，用于构建由锚样本、正类样本和负类样本组成的三元组样本对。通过最大化锚样本与正类样本的相似度，最小化锚样本与负类样本的相似度，使得相同类别样本之间的距离远小于不同类样本的距离。Taghanaki 等[35]将 Dice 损失和改进的交叉熵损失相结合，提出了一种 Combo Loss 损失函数，用于解决非平衡问题。Baloch 等[36]提出了一种聚焦锚点损失函数(FAL)来解决类不平衡问题。该损失函数实际上综合了 Large Margin Gaussian Mixture Loss[37]损失函数和 Focal Loss[38]损失函数。Pasupa 等[39]将卷积神经网络应用于医学领域，针对犬红细胞的不平衡分类问题，使用 Focal Loss 损失函数对卷积神经网络模型进行训练。

上述文献是近年来学者们为了解决非平衡问题对卷积神经网络的损失函数进行的研究。由此可见，改进损失函数能够提高训练模型对非平衡数据的分类性能。但是，很多改进方法的理论复杂，降低了可操作性，且神经网络训练时间增加。

3 DPCSMOTE-FCECNN

针对非平衡问题，本文提出的方法包括数据预处理和卷积神经网络中损失函数的改进两部分。数据预处理采用过采样方法，将密度峰值聚类算法(Clustering by Fast Search and Find of Density Peaks，DPC)和 SMOTE 过采样算法相结合，缓解了原始数据的不平衡程度。在训练神经网络模型时，对卷积神经网络的交叉熵损失函数进行改进，提出了基于代价敏感思想，并考虑了预测值与标签值之间距离的新的损失函数，构建了针对非平衡数据的卷积神经网络模型。图 1 给出了本文方法的框架。本文方法的具体实现过程如算法 1 所示。

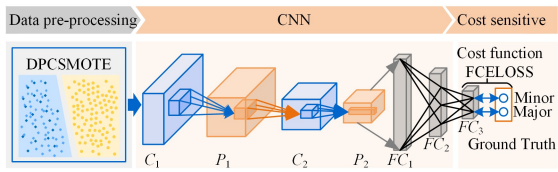


图 1 本文方法的结构

Fig. 1 Architecture of the proposed method

算法 1 DPCSMOTE-FCECNN

//数据预处理：DPCSMOTE

Input： 原始数据集 X；

目标向量 **y**；

需要合成的样本数 n；

非平衡比阈值 irt；

最近邻数 knn；

计算密度的指数 de，默认为 X 中的特征数

Output：过采样后得到的数据集 X_resampled

X_resampled 的类标 **y**_resampled

Step 1 对输入数据集进行聚类，筛选出少数类样本多于多数类样本的簇。

clusters←DPC(X)
filtered clusters←∅
for c∈clusters do
  imbalance ratio←$\frac{\text{majority count(c)}+1}{\text{minority count(c)}+1}$
if imbalance ratio<irt then
  filtered clusters←filtered clusters∪{c}
end
end
Step 2 对于每个目标簇,根据少数类密度计算采样权重。
for f∈filtered clusters do
  average minority distance (f)←mean (enclidean distances (f))
  density factor (f)←$\frac{\text{minority count(f)}}{\text{average minority distance (f)}^{de}}$
  sparsity factor (f)←$\frac{1}{\text{density factor(f)}}$
end
sparsity sum←$\sum_{\text{f∈filtered clusters}}$ sparsity factor(f)
sampling weight(f)←$\frac{\text{sparsity factor(f)}}{\text{sparsity sum}}$
Step 3 采用 SMOTE 方法在目标簇中合成新样本,根据采样权重计算需要合成的样本数。
generated samples←∅
for f∈filtered clusters do
  number of samples ← ‖ n×sampling weight(f) ‖
  generated samples←generated samples ∪ {SMOTE(f, number of samples, knn)}
end
return generated samples
//采用 FCELoss 损失函数训练 CNN
Input:X_resampled (预处理后的数据集)
y_resampled (X_resampled 的类标)
Output:trained_construct_CNN
Net←construct_CNN
Weights,biases←initialize_Net
for i∈ [1,num_steps] do
grad←FCE loss function
weights,biases←updated_NetParameters
learning_rate=0.001/(1+10 *(float(i)/num_steps) * 0.75
end
return trained_construct_CNN

3.1 DPCSMOTE 算法

作为数据的预处理方法,DPCSMOTE 算法用于减小原始输入数据的不平衡程度,使输入到神经网络中的数据分布不至于过于偏态。

DPCSMOTE 算法将密度峰值聚类算法和 SMOTE 过采样算法相结合,主要分为 3 个阶段。第一个阶段是聚类,利用密度峰值算法将输入数据聚类成簇;第二阶段筛选出需要进行过采样的簇,并确定每个簇中需要合成新样本的数量;第三阶段利用 SMOTE 算法在相应的簇中进行过采样,使其达到预定的非平衡比率。

密度峰值聚类算法是一种基于密度峰值的聚类方法,算法的核心思想在于对聚类中心点的刻画,它基于这样一种假设:在一个数据集中,聚类中心(密度峰值点)被密度较低的邻居点包围,而且聚类中心与其他密度更大的数据点之间的距离相对更大。若要找到满足以上特点的聚类中心,则需要引入两个定义:局部密度 ρ_i 和距离 δ_i。局部密度 ρ_i 的定义如下:

$$\rho_i=\sum_j \chi(d_{ij}-d_c) \tag{1}$$

其中,

$$\chi(x)=\begin{cases}1, & x<0\\ 0, & x\geqslant 0\end{cases} \tag{2}$$

参数 $d_c>0$ 为截断距离,需事先指定。由以上定义可知,数据点 i 的局部密度 ρ_i 表示数据集中与 i 点之间距离小于 d_c 的数据点的个数。距离 δ_i 的定义如下:

$$\delta_i=\min_{j:\rho_j>\rho_i}(d_{ij}) \tag{3}$$

当数据点 i 具有最大局部密度时,δ_i 表示数据集中与 i 点距离最远的数据点与 i 点之间的距离;否则,δ_i 表示在所有局部密度大于 i 点的数据点中,与 i 点距离最近的数据点与 i 点之间的距离。将每个数据点的局部密度和距离计算出来后画出决策图,将具有较大局部密度和较大距离的点确定为聚类中心,将局部密度较小而距离较大的点归为异常点。对于剩下的数据点,算法将其分配到最近邻且密度比它大的数据点所在的簇。

经过第一阶段的密度峰值聚类算法后,输入数据被划分到各个簇中,现在需要经过第二阶段,对所有的簇进行筛选判断,选出需要进行过采样的簇,并确定每个簇中需要合成的新样本的数量,为下一阶段进行 SMOTE 过采样做准备。

为了避免产生噪声样本,在安全区域进行过采样,第二阶段将在所有簇中筛选出少数类样本数量较多的簇进行下一阶段的过采样。定义每个簇的非平衡比率为(多数类样本数+1)/(少数类样本数+1)。为了筛选出目标簇,设定一个阈值 c,所有非平衡比率不超过阈值 c 的簇都将作为需要进行过采样的目标簇。

筛选出目标簇后,下一步要确定这些目标簇中需要合成的新样本数量。为了缓解类内不平衡,需要为每个筛选出的目标簇赋予权重,使少数类分布稀疏的簇可以分配到更多的合成样本。每个目标簇中少数类的稀疏度由密度的倒数计算得出,密度则是每个簇中少数类样本的个数与少数类样本平均距离的特征数次幂的比值。计算每个簇中少数类样本的平均距离时要先求出少数类样本的欧几里得距离矩阵,平均距离为距离矩阵中所有非对角元素的和与非对角元素个数的比。计算出每个目标簇中少数类的稀疏度之后,将其与所有选中的目标簇的稀疏度总和相比就可以为每个簇赋予权重了。将每个簇的权重与需要生成的合成样本总数相乘便可以确定每个目标簇中需要合成的新样本数量。

第三阶段利用 SMOTE 算法对目标簇按照计算出的合成样本数进行过采样,使得样本中多数类和少数类达到预定的非平衡比。其中需要注意的是,SMOTE 算法中超参数 *knn* 的设定,对于少数类相对稀疏的簇来说,可能出现簇中少数类样本数不足 *knn* 的情况,此时就要减小 *knn* 的设定值。

本文方法在安全区域进行过采样,并按照少数类的稀疏度分配合成样本数量,在解决噪声样本问题的同时有效缓解类内不平衡。此外,SMOTE 过采样方法的使用避免了过拟合问题的出现。通过 DPCSMOTE 算法对输入数据进行预处

理，大大降低了原始数据的非平衡程度，为后面神经网络的训练提供了更有利的数据。

3.2 FCECNN 损失函数

在神经网络的训练过程中，损失函数作为神经网络模型优化的目标函数，用于评估模型预测值和真实值之间的差异程度。神经网络训练或者优化的过程就是最小化损失函数的过程。因此，针对非平衡问题，本文将卷积神经网络中常用的交叉熵损失函数进行改进，提出了一种结合代价敏感思想，并考虑网络模型输出的预测值与标签值之间距离的新损失函数。

交叉熵损失函数是分类问题中应用比较广泛的一种损失函数。交叉熵损失函数的表达式如下：

$$CELoss=-\sum_k t_k \log y_k \tag{4}$$

其中，y_k 是卷积神经网络的输出；t_k 是正确解标签，t_k 中只有正确解标签的索引为 1，其他均为 0(one-hot 表示)。因此，交叉熵损失函数的值是由正确解标签所对应的输出结果决定的。

为应对非平衡数据，本文结合代价敏感思想在损失函数中为不同的类别赋予不同的代价损失，具体如下：给定训练集 $\{x^{(k)},y^{(k)}\}_{k=1}^{N}$，$x^{(k)}$ 表示第 k 个样本，$y^{(k)}\in\{0,1\}$ 是该样本对应的标签值；用 n_c 表示类别 c 中的样本总数，其中 $c\in\{0,1\}$，则 n_c 可表示为：

$$n_c=\sum_{k=1}^{N}1\{y^{(k)}=c\} \tag{5}$$

基于代价敏感的思想，为少数类赋予更大的代价损失，因此为其分配更大的权重，权重的计算式如下：

$$v_c=\frac{\max\{n_c\}_{c=0}^{1}}{n_c} \tag{6}$$

这样各个类别的权重就可以根据数据集中样本的分布进行自适应计算，为少数类赋予更大的代价损失，而多数类则不受影响。

除此以外，本文提出的损失函数还考虑了卷积神经网络输出层的预测值与标签值之间的距离，计算式如下：

$$d^2=(1-y_p)^2 \tag{7}$$

其中，y_p 表示输出层的预测值，式(7)计算了预测值与正确标签 $t=1$ 之间的欧氏距离的平方。因此，卷积神经网络的 softmax 损失函数的定义如下：

$$FCELoss=-\frac{1}{N}\sum_{k=1}^{N}\sum_{c=0}^{1}v_c(1-y_p)^2 1\{y^{(k)}=c\}\log y_p^{(k)} \tag{8}$$

在损失函数中考虑预测值与标签距离是为了将更多的注意力放在容易造成误分类的少数类样本的训练上。当卷积神经网络对容易分类的多数类样本进行训练时，输出层的预测值 $y_p\gg0.5$，欧氏距离的平方 $d^2\to0$，此时 $FCELoss\ll CELoss$，相当于减小了多数类的代价损失。而在对容易误分类的少数类样本进行训练时，输出层的预测值 $y_p\to0$，欧氏距离的平方 $d^2\to1$，与交叉熵损失函数相比，$FCELoss$ 值基本不受影响。因此，反过来说就相当于增加了少数类的代价损失，对卷积神经网络的训练过程产生了积极影响。

对于提出的损失函数，对其求导可以得到如下所示的导数。

$$\frac{\mathrm{d}FCE(y_p)}{\mathrm{d}y_p}=-v_c(1-y_p)\left(2\log y_p-\frac{1}{y_p}+1\right) \tag{9}$$

利用式(9)可以在卷积神经网络的训练过程中不断更新模型的参数。

4 实验及分析

4.1 数据集

在实验中使用了 6 个数据集进行性能评估，分别是 MNIST 手写数字识别数据集、Fashion-MNIST 图像数据集(F-MNIST)、CIFAR-10 数据集、17flowers 数据集、Cats vs. Dogs 数据集(Cats_Dogs)和 Weather 数据集。

MNIST 手写数字识别数据集由 60 000 个训练样本和 10 000 个测试样本组成，每个样本都是一张 28 * 28 像素的灰度手写数字图片，代表 0 到 9 中的一个数字。Fashion-MNIST 数据集与 MNIST 类似，包含 10 种类别的 70 000 张不同服饰商品的灰度图像。其图片大小、格式以及训练集/测试集划分与 MNIST 数据集完全一致。CIFAR-10 数据集是一个比 MNIST 数据集更复杂的图像分类数据集，包括 60 000 张 32 * 32 像素的三通道 RGB 彩色图片，其中训练集有 50 000 张，测试集有 10 000 张。图片分为 10 个类，每一类有 6 000 张图片样本。17flowers 数据集是牛津大学 Visual Geometry Group 选取的英国常见的 17 种花的图片数据集。Cats vs. Dogs 数据集包含训练集和测试集，训练集中包含猫和狗的图片各 12 500 张，测试集中包含猫和狗的混合乱序图片共 12 500 张。Weather 数据集是由 rain，cloudy 等天气图片构成的数据集。

由于这些数据集都不是非平衡的，因此我们选取其中的一部分数据，人为构造非平衡数据集。为了验证本文方法的有效性，我们在上述每个数据集中构造非平衡比率为 1∶5 和 1∶15 的两个非平衡数据集，并将新的数据集按照 3∶1 的比例划分训练集和测试集。本文方法只针对二分类问题，因此构造的数据集都是二分类数据集。

4.2 评价指标

在对分类器进行性能评估时，传统的评价指标在非平衡分类任务中并非都适用。因此，需要选用专门针对非平衡数据的评价指标才能对模型的性能进行更合理的评估。

精度(accuracy)是分类任务中常用的评价指标，表示分类正确的样本数占样本总数的比例。但是，当数据不平衡时，精度并不能合理评估分类器的性能。如在一个非平衡比为1∶99的数据集中，若分类器将全部多数类都准确分类，而少数类分类错误，则可以得到 99% 的精度。但事实是，看似精度很高的分类器并没有将少数类样本正确分类。

在二分类问题中，常用混淆矩阵表示分类结果。具体来说，可将样本根据其真实类别与分类器预测类别的组合划分为真正例(True Positive，TP)、假正例(False Positive，FP)、真反例(True Negative，TN)、假反例(False Negative，FN)4 种情况，令 TP，FP，TN，FN 分别表示其对应的样本数，显然有 $TP+FP+TN+FN=$样本总数。

相对于精度未能表示出少数类的分类准确度，查准率(precision)和召回率(recall)则对少数类的分类情况进行了更具体的描述。查准率和召回率的定义如下：

$$precision=TP/(TP+FP) \tag{10}$$

$$recall=TP/(TP+FN) \tag{11}$$

查准率表示在预测为正例的样本中预测正确的比例，召

回率则表示所有实际为正例的样本中被预测正确的样本所占的比例。

查准率和召回率通常此消彼长，相互制约，因此，学者们综合考虑了这两个指标，提出 F1-score 度量。F1-score 是查准率和召回率的调和平均，具体定义如下：

$$F1\text{-}score=2\times\frac{precision\times recall}{precision+recall} \tag{12}$$

G-mean 值也是基于混淆矩阵的非平衡分类评价指标，表示几何平均预测精度，其定义如下：

$$G\text{-}mean=\sqrt{\frac{TP}{TP+FN}\times\frac{TN}{TN+FP}} \tag{13}$$

此外，ROC 曲线也是非平衡分类中常用的评价指标。ROC 曲线的绘制是根据分类器的预测结果对样例进行排序，按此顺序逐个把样本作为正例进行预测，每次计算出真正例率和假正例率的值，分别以它们为横、纵坐标绘图。真正例率(Ture Positive Rate，TPR)和假正例率(False Positive Rate，FPR)的定义如下：

$$TPR=\frac{TP}{TP+FN} \tag{14}$$

$$FPR=\frac{FP}{TN+FP} \tag{15}$$

4.3 实验设置

实验以卷积神经网络为训练模型，训练过程中以本文提出的损失函数为指标更新权重参数。在使用卷积神经网络训练之前，先使用本文提出的 DPCSMOTE 方法对原始数据集进行预处理，以降低数据的非平衡度。其中，参数设置为：密度阈值 density_threshold∈{6,8,14}，距离阈值 distance_threshold∈{5,10,12}，knn∈{10,20}，非平衡比阈值 irt∈{1,∞}。为了证明预处理方法的有效性，实验中将 DPCSMOTE_FCECNN 与未使用数据预处理的 FCECNN 进行对比。

实验中使用的卷积神经网络模型的结构如表 1 所列，卷积层、池化层、全连接层、输出层在表中分别以 C，P，FC，FO 表示。

表 1 卷积神经网络的参数

Table 1 Parameters of convolutional neural network

网络层	运算	参数(以 CIFAR-10 数据为例)
输入层	输入数据	32 * 32 * 3
C1	卷积	5 * 5 * 3
P1	最大池化	4
C2	卷积	5 * 5 * 64
P2	最大池化	4
FC1	全连接	—
FC2	全连接	—
FO	输出	—

为了验证 DPCSMOTE_FCECNN (DS_FCECNN)能够提高卷积神经网络对非平衡数据的分类性能，在实验中，我们将标准交叉熵损失函数训练的卷积神经网络模型 CNN 和文献[33]提出的损失函数训练的卷积神经网络模型 DNCNN 进行了对比实验，采用 F1-score，G-mean 值和 ROC 3 个评价指标对各个模型的分类性能进行评估。

4.4 参数选择

在卷积神经网络模型中，学习率等超参数的取值会对模型的分类性能有很大影响。因此，在 MNIST 数据集上进行参数选择实验，以分别确定卷积神经网络的学习率和卷积核长度的取值。

图 2 给出了学习率的参数选择实验中模型的 F1-score 值。从箱型图可以看到，当学习率取值为 0.001 时模型的性能最好，因此将学习率设置为 0.001。

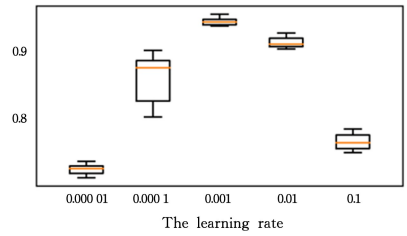


图 2 学习率参数选择实验结果

Fig. 2 Results of learning rate selection

图 3 给出了对卷积神经网络的卷积核长度进行参数选择实验的结果。从实验结果可以明显看出，当卷积核长度取值为 5 时，模型的 F1-score 值最高。此后，增加卷积核的长度，模型的性能会下降，因此将卷积核长度设置为 5。

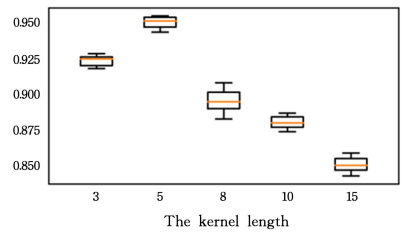


图 3 卷积核参数选择实验结果

Fig. 3 Results of convolutional kernel length selection

4.5 实验结果

为了验证本文方法的有效性，我们分别在非平衡比不同的数据集上进行实验。下文分别展示了在非平衡比为 1:5 和 1:15 的数据集上的分类结果。

4.5.1 非平衡比为 1:5

表 2 列出了非平衡比为 1:5 时 4 种模型在 6 个数据集上的 F1-score。与未使用数据预处理的 FCECNN 相比，DS_FCECNN 的 F1-score 在 6 个数据集上分别高出了 4.26%，2.15%，8.34%，10.65%，17.67% 和 8.86%，证明了本文提出的 DPCSMOTE 方法作为数据预处理方法的有效性。与 CNN 和 DNCNN 相比，DS_FCECNN 也具有最好的分类性能。尤其在 17flowers 数据集上，三者的对比更加明显，DS_FCECNN 的 F1-score 分别高出 CNN 和 DNCNN 28.36% 和 19.79%。

表 2 非平衡比为 1:5 的 6 个数据集上的 F1-score

Table 2 F1-score on the six datasets with imbalance ratio 1:5

	DS_FCECNN	FCECNN	CNN	DNCNN
MNIST	0.9403	0.8977	0.8936	0.9202
F-MNIST	0.9786	0.9571	0.9176	0.9272
CIFAR-10	0.6667	0.5833	0.5120	0.5922
17flowers	0.5913	0.4848	0.3077	0.3934
Cats_Dogs	0.5221	0.3454	0.2642	0.3256
Weather	0.8645	0.7759	0.8148	0.7800

G-mean 值的实验结果如表 3 所列。从表中可以看到，DS_FCECNN 的 G-mean 值均高于 FCECNN，进一步证明了本文提出的数据预处理方法能够提高卷积神经网络的分类性能。而相对于 CNN 和 DNCNN，DS_FCECNN 也获得了更高

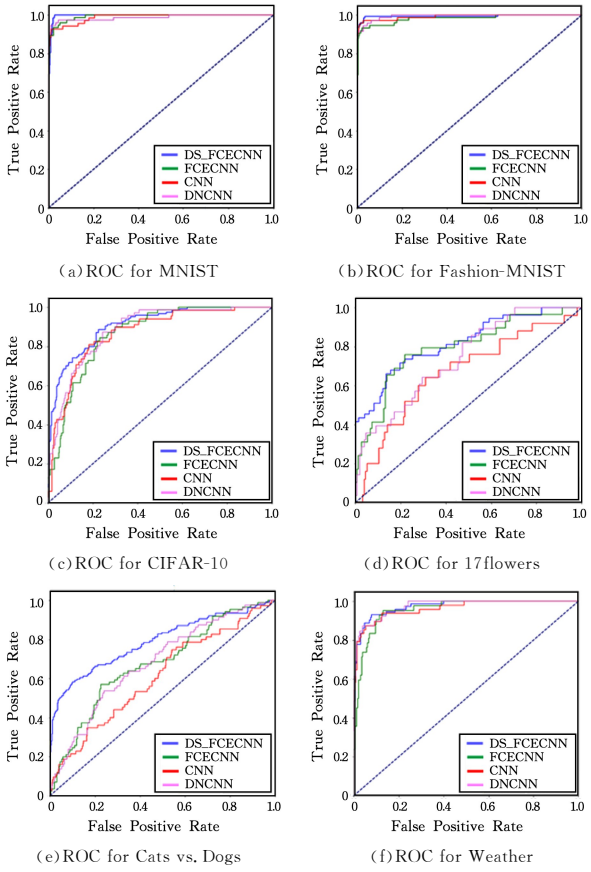
的 G-mean 值，表明 DS_FCECNN 的分类性能优于其他模型。

表 3 非平衡比为 1∶5 的 6 个数据集上的 G-mean 值

Table 3 G-mean on the six datasets with imbalance ratio 1∶5

	DS_FCECNN	FCECNN	CNN	DNCNN
MNIST	0.9733	0.9613	0.9386	0.9723
F-MNIST	0.9917	0.9905	0.9323	0.9635
CIFAR-10	0.8028	0.7746	0.6800	0.7580
17flowers	0.7501	0.6711	0.4771	0.5834
Cats_Dogs	0.6863	0.5789	0.4221	0.5596
Weather	0.9258	0.8888	0.8747	0.9054

为了直观地展示 DS_FCECNN 的分类性能，我们引入了 ROC 曲线作为分类评价度量。当数据集的不平衡比率为 1∶5 时，各个模型的 ROC 曲线如图 4 所示。从图中可以看到，DS_FCECNN 的 ROC 曲线最接近理想状态(0,1)且曲线下的面积最大，表明本文提出的 DS_FCECNN 能对不平衡数据进行更有效的分类。



(a)ROC for MNIST (b)ROC for Fashion-MNIST

(c)ROC for CIFAR-10 (d)ROC for 17flowers

(e)ROC for Cats vs. Dogs (f)ROC for Weather

图 4 非平衡比为 1∶5 的 6 个数据集的 ROC 曲线

Fig. 4 ROC curves for the six datasets with imbalance ratio 1∶5

4.5.2 非平衡比为 1∶15

当数据集的非平衡比率为 1∶15 时，4 种模型的 F1-score 和 G-mean 值如表 4 和表 5 所列。通过对比非平衡比率为 1∶5 时的实验结果可以看到，当数据集的非平衡比率增大时，F1-score 和 G-mean 值总体上呈现下降趋势，表明模型的分类性能变差。尤其在表 5 中可以看到，在 Cats vs. Dogs 数据集上 CNN 模型的 G-mean 值出现了无效结果(用 nan 表示)。这是由于 CNN 模型在学习过程中的梯度异常，使得学习的过程偏离了正常的轨迹。表明当数据的非平衡比率较高时，CNN 模型无法对数据进行有效分类。从表 4 和表 5 中可以明显看到，当数据集的非平衡比率高达 1∶15 时，DS_FCECNN 仍获得了最高的 F1-score 和 G-mean 值。一方面，通过 DS_FCECNN 与 FCECNN 的实验结果对比证明了提出的数据预处理方法 DPCSMOTE 的有效性。另一方面，通过 DS_FCECNN 与 CNN 和 DNCNN 进行对比，表明了本文提出的 DS_FCECNN 的分类性能优于其他模型。

表 4 非平衡比为 1∶15 的 6 个数据集上的 F1-score

Table 4 F1-score on the six datasets with imbalance ratio 1∶15

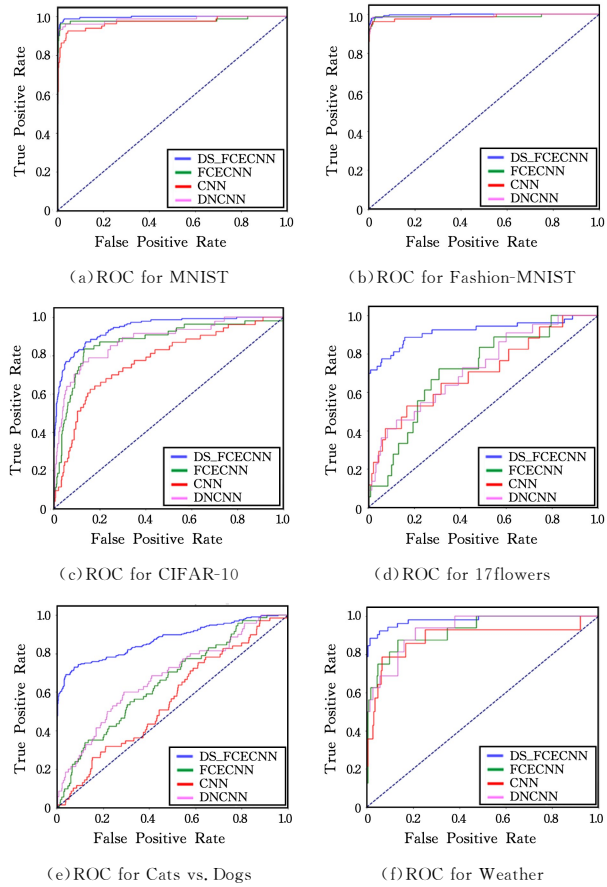
	DS_FCECNN	FCECNN	CNN	DNCNN
MNIST	0.9350	0.8372	0.7565	0.8182
F-MNIST	0.9545	0.8256	0.8668	0.8886
CIFAR-10	0.7017	0.3797	0.2188	0.4781
17flowers	0.7840	0.2381	0.1000	0.2069
Cats_Dogs	0.6586	0.1868	nan	0.1798
Weather	0.8608	0.6857	0.3636	0.6667

表 5 非平衡比为 1∶15 的 6 个数据集上的 G-mean 值

Table 5 G-mean on the six datasets with imbalance ratio 1∶15

	DS_FCECNN	FCECNN	CNN	DNCNN
MNIST	0.9836	0.9629	0.8096	0.9732
F-MNIST	0.9821	0.9739	0.8996	0.9687
CIFAR-10	0.8637	0.7785	0.3737	0.7427
17flowers	0.8624	0.5241	0.2417	0.3593
Cats_Dogs	0.8072	0.6383	nan	0.5250
Weather	0.9530	0.8288	0.4965	0.7835

图 5 给出了非平衡比率为 1∶15 时 4 种模型的 ROC 曲线。对比非平衡比率为 1∶5 时的 ROC 曲线，数据的非平衡程度越大，DS_FCECNN 的优势就越明显。



(a)ROC for MNIST (b)ROC for Fashion-MNIST

(c)ROC for CIFAR-10 (d)ROC for 17flowers

(e)ROC for Cats vs. Dogs (f)ROC for Weather

图 5 非平衡比为 1∶15 的 6 个数据集的 ROC 曲线

Fig. 5 ROC curves for the six datasets with imbalance ratio 1∶15

从图5中看到，DS_FCECNN的ROC曲线在6个数据集上几乎能将其他模型的ROC曲线包住，表明DS_FCECNN的分类性能更好。

结束语 为解决非平衡分类问题，本文提出了一种基于代价敏感卷积神经网络的混合方法DPCSMOTE-FCECNN。先通过DPCSMOTE方法对数据进行预处理，缓解原始数据集的不平衡程度，再将经过预处理后的数据输入到卷积神经网络中进行训练。为了对非平衡数据进行更有效的分类，提出了一种新的损失函数FCELoss。该损失函数基于代价敏感思想，考虑预测值与标签值之间的欧氏距离，分别赋予多数类和少数类不同的代价损失，以此来减小数据的非平衡性对卷积神经网络训练产生的不利影响。实验结果表明，本文方法在评价指标F1-score和G-mean值上有明显的优势，能够提高卷积神经网络模型在非平衡数据上的分类性能。本文方法只针对二分类问题，在未来的工作中，我们将把本文方法扩展到多分类任务中。

参 考 文 献

[1] WAHAB N, KHAN A, LEE Y S. Two-phase deep convolutional neural network for reducing class skewness in histopathological images based breast cancer detection[J]. Computers in Biology and Medicine, 2017, 85: 86-97.

[2] WEI W, LI J J, CAO L B, et al. Effective detection of sophisticated online banking fraud on extremely imbalanced data[J]. World Wide Web-internet and Webinformation Systems, 2013, 16(4): 449-475.

[3] ENGEN V, VINCENT J, PHALP K. Enhancing network based intrusion detection for imbalanced data[J]. International Journal of Knowledge-based and Intelligent Engineering Systems, 2008, 12(5/6): 357-367.

[4] MAO W T, HE L, YAN Y J, et al. Online sequential prediction of bearings imbalanced fault diagnosis by extreme learning machine[J]. Mechanical Systems and Signal Processing, 2017, 83: 450-473.

[5] CHAWLA N V, JAPKOWICZ N, KOTCZ A. Special issue on learning from imbalanced data sets[J]. ACM Sigkdd Explorations Newsletter, 2004, 6(1): 1-6.

[6] GUO H X, LI Y J, SHANG J, et al. Learning from class-imbalanced data: Review of methods and applications[J]. Expert Systems with Applications, 2017, 73: 220-239.

[7] WANG Q. A Hybrid Sampling SVM Approach to Imbalanced Data Classification[J]. Abstract and Applied Analysis, 2014, 11(6): 1-7.

[8] GALAR M, FERNANDEZ A, BARRENECHEA E, et al. A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches[J]. IEEE Transactions on Systems Man & Cybernetics Part C Applications & Reviews, 2012, 42(4): 463-484.

[9] BATISTA G E, PRATI R C, MONARD M C, et al. A study of the behavior of several methods for balancing machine learning training data[J]. Sigkdd Explorations, 2004, 6(1): 20-29.

[10] FERNANDEZ A, GARCIA S, JESUS M J, et al. A study of the behaviour of linguistic fuzzy rule based classification systems in the framework of imbalanced datasets[J]. Fuzzy Sets and Systems, 2008, 159(18): 2378-2398.

[11] PANT H, SRIVASTAVA R. A survey on feature selection methods for imbalanced datasets[J]. International Journal of Computer Engineering & Application, 2015, 9: 197-204.

[12] MOAYEDIKIA A, ONG K L, BOO Y L, et al. Feature selection for high dimensional imbalanced class data using harmony search[J]. Engineering Applications of Artificial Intelligence, 2017, 57: 38-49.

[13] MALDONADO S, LOPEZ J. Dealing with high-dimensional class-imbalanced datasets: Embedded feature selection for SVM classification[J]. Applied Soft Computing, 2018, 67: 94-105.

[14] THAINGHE N, GANTNER Z, SCHMIDTTHIEME L, et al. Cost-sensitive learning methods for imbalanced data[C]// The 2010 International Joint Conference on Neural Networks (IJCNN). IEEE, 2010: 1-8.

[15] KRAWCZYK B, WOZNIAK M, HERRERA F, et al. Weighted one-class classification for different types of minority class examples in imbalanced data[C]// 2014 IEEE Symposium on Computational Intelligence and Data Mining (CIDM). IEEE, 2014: 337-344.

[16] SUN Z B, SONG Q B, ZHU X Y, et al. A novel ensemble method for classifying imbalanced data[J]. Pattern Recognition, 2015, 48(5): 1623-1637.

[17] LI F L, ZHANG X Y, ZHANG X Q, et al. Cost-sensitive and hybrid-attribute measure multi-decision tree over imbalanced data sets[J]. Information Sciences, 2018, 422: 242-256.

[18] KRAWCZYK B, WOŹNIAK M, SCHAEFER G. Cost-sensitive decision tree ensembles for effective imbalanced classification[J]. Applied Soft Computing, 2014, 14: 554-562.

[19] WANG C, YU Q, LUO R S, et al. Adaptive Ensemble of Classifiers with Regularization for Imbalanced Data Classification.[J]. arXiv: Learning, 2019.

[20] ZHU Z H, WANG Z, LI D D, et al. Geometric Structural Ensemble Learning for Imbalanced Problems[J]. IEEE Transactions on Systems, Man, and Cybernetics, 2020, 50(4): 1617-1629.

[21] ZHU W X, ZHONG P. A new one-class SVM based on hidden information[J]. Knowledge Based Systems, 2014, 60: 35-43.

[22] 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.

[23] RODRIGUEZ A, LAIO A. Clustering by fast search and find of density peaks[J]. Science, 2014, 344(6191): 1492-1496.

[24] YU D, LIU G, GUO M, et al. Density peaks clustering based on weighted local density sequence and nearest neighbor assignment[J]. IEEE Access, 2019, 7: 34301-34317.

[25] CHAWLA N V, BOWYER K W, HALL L O, et al. SMOTE: synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research, 2002, 16: 321-357.

[26] DOUZAS G, BACAO F. Geometric SMOTE a geometrically enhanced drop-in replacement for SMOTE[J]. Information Sciences, 2019, 501: 118-135.

[27] PAN T,ZHAO J,WU W,et al. Learning imbalanced datasets based on SMOTE and Gaussian distribution[J]. Information Sciences,2020,512:1214-1233.

[28] DOUZAS G,BACAO F. Self-Organizing Map Oversampling (SOMO) for imbalanced data set learning[J]. Expert systems with Applications,2017,82:40-52.

[29] DOUZAS G,BACAO F,LAST F. Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE[J]. Information Sciences,2018,465:1-20.

[30] GONG L,JIANG S,JIANG L. Tackling Class Imbalance Problem in Software Defect Prediction Through Cluster-based Over-sampling with Filtering[J]. IEEE Access,2019(99):1.

[31] KHAN S H,HAYAT M,BENNAMOUN M,et al. Cost Sensitive Learning of Deep Feature Representations from Imbalanced Data[J]. IEEE Transactions on Neural Networks,2018,29(8):3573-3587.

[32] GENG Y,LUO X Y. Cost-sensitive convolution based neural networks for imbalanced time-series classification[J]. arXiv:1801.04396,2018.

[33] JIA F,LEI Y G,LU N,et al. Deep normalized convolutional neural network for imbalanced fault classification of machinery and its understanding via visualization[J]. Mechanical Systems and Signal Processing,2018,110:349-367.

[34] CHEN L T,XU G H,ZHANG Q,et al. Learning deep representation of imbalanced SCADA data for fault detection of wind turbines[J]. Measurement,2019,139:370-379.

[35] TAGHANAKI S A,ZHENG Y F,ZHOU S K,et al. Combo loss:Handling input and output imbalance in multi-organ segmentation[J]. Computerized Medical Imaging and Graphics,2019,75(4):24-33.

[36] BALOCH B K,KUMAR S,HARESH S,et al. Focused Anchors Loss:cost-sensitive learning of discriminative features for imbalanced classification[C]// Asian Conference on Machine Learning. 2019:822-835.

[37] WAN W T,ZHONG Y Y,LI T P,et al. Rethinking feature distribution for loss functions in image classification[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:9117-9126.

[38] GOYAL P,KAIMING H. Focal loss for dense object detection [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2018,39:2999-3007.

[39] PASUPA K,VATATHANAVARO S,TUNGJITNOB S,et al. Convolutional neural networks based focal loss for class imbalance problem:A case study of canine red blood cells morphology classification[J]. Journal of Ambient Intelligence and Humanized Computing,2020,56(4):1-17.

HUANG Ying-qi, born in 1988, postgraduate. Her main research interests include machine learning and data mining.

CHEN Hong-mei, born in 1971, Ph.D, professor, Ph.D supervisor, is a member of China Computer Federation. Her main research interests include granular calculation, rough sets and intelligent information processing.