文章编号：1006-2467(2021)02-0131-10 DOI：10.16183/j.cnki.jsjtu.2020.082

基于Gaussian混合的距离度量学习数据划分方法

郑德重[1,2]， 杨媛媛[1]， 谢 哲[1,2]， 倪扬帆[1,2]， 李文涛[3]
（1. 中国科学院上海技术物理研究所 医学影像信息学实验室，上海 200080；
2. 中国科学院大学，北京 100049；3. 复旦大学附属肿瘤医院，上海 200032）

摘 要：针对有限样本情况下，多次训练模型时容易出现不稳定和偏差问题，提出一种基于Gaussian混合的距离度量学习数据划分方法，通过更合理地划分数据集来解决该问题.距离度量学习依靠深度神经网络优异的特征提取能力，将原始数据提取的特征嵌入到新的度量空间中；然后，在该新的度量空间中基于深层次特征使用Gaussian混合模型进行聚类分析和样本分布估计；最后，依据样本分布特点进行分层采样对数据进行合理划分.研究表明，该方法可以更好地理解数据分布的特点，获得更加合理的数据划分，进而提升模型的准确性和泛化性.

关键词：人工智能训练；数据集划分；深度神经网络；Gaussian混合模型
中图分类号：TP 181 **文献标志码**：A

Data Splitting Method of Distance Metric Learning Based on Gaussian Mixed Model

ZHENG Dezhong[1,2]， *YANG Yuanyuan*[1]， *XIE Zhe*[1,2]， *NI Yangfan*[1,2]， *LI Wentao*[3]
(1. Laboratory for Medical Imaging Informatics, Shanghai Institute of Technical Physics, Chinese Academy of Sciences, Shanghai 200080, China; 2. University of Chinese Academy of Sciences, Beijing 100049, China; 3. Fudan University Shanghai Cancer Center, Shanghai 200032, China)

Abstract: Aimed at the problem of instability and deviation of multiple training model in limited samples, this paper proposes a method of distance metric learning based on the Gaussian mixture model, which can solve this problem more reasonably by dividing the dataset. Distance metric learning relies on the excellent feature extraction capabilities of deep neural networks to embed the original data into the new metric space. Then, based on the deep features, the Gaussian mixture model is used to cluster the analyzer and estimate the sample distribution in this new metric space. Finally, according to the characteristics of sample distribution, stratified sampling is used to reasonably divide the data. The research shows that the method proposed can better understand the characteristics of data distribution and obtain a more reasonable data division, thereby improving the accuracy and generalization of the model.

Key words: artificial intelligence training; dataset division; deep neural networks; Gaussian mixture model

收稿日期：2020-03-24
基金项目：面向跨域协同医学影像新型服务模式解决方案(2017YFC0112900)，人工智能医学软件测评数据库和服务平台开发(2019YFC0118803)
作者简介：郑德重(1990-)，男，湖北省武汉市人，博士生，主要研究方向为机器学习、深度学习在医学影像方面的应用.
通信作者：李文涛，男，教授，博士生导师；E-mail：liwentao98@126.com.

近年来,深度学习技术在许多领域取得了巨大的成功,从计算机视觉、网络搜索、社交内容的协同过滤、电子商务的推荐系统,到消费产品中的图像识别、人脸识别等方面的应用都获得了显著提高.深度学习是一种多层次结构化的计算模型,可以从数据中学习多种不同层次的抽象表达[1].仅从数据中学习就可以获得数据不同抽象层次的特征表示,而不需要依赖于特定领域知识进行手工设计是深度学习技术进步的核心[2-3].在训练深度学习模型时,数据常常被划分成多块用于交叉验证,在模型开发中此过程是十分常见又重要的一个环节,因为交叉验证通常可以保证其良好的泛化性能.但对中等规模的数据而言,在进行交叉验证时,划分出来用于测试和验证的子集质量会在一定程度上对神经网络的训练产生一定影响,不良地数据划分会导致所获得的模型不准确,并有可能在交叉验证过程中产生较大的偏差[4-5].小样本数据更是如此,这种规模的数据较难满足统计意义上的普适性和一般随机性.例如,随着人为数据采集的方式和习惯的变化,采集来的数据可能包含着随时间变化的趋势和倾向性特点.因此在使用有限数据时,数据的统计规律是可变的.当数据规模不是足够大时,简单地以随机方式对数据进行划分是不可取的.另一方面,在收集数据的过程中,数据很少是没有噪声的,并且采集来可用的有效数据可能仅仅只占其中一小部分,并不能包含模型构建所需要的全部信息[4].尽管可以通过增加样本量来适当缓解这些问题,但是在一些特定条件下产生的稀有样本并不是那么容易获得的.因此在数据有限的情况下,简单地随机划分数据容易导致测试数据域和训练数据域的漂移,训练所得到的模型是不稳定的.

依据统计学的知识和经验容易知道,准确了解数据和目标的分布特点将有助于理解数据的内在本质,从而合理地划分训练数据和验证数据.由于深度学习技术具有强大的特征提取能力,可以提取与分类目标相关的多种层次的特征.在此基础上,分析样本在特征空间的分布特点有助于更好地理解数据本身,从而合理地划分数据.本文提出一种基于 Gaussian 混合模型的距离度量学习模型(DML-GMM)划分数据.该方法不依赖于任何特定的特征提取器,在特征提取方面,可以使用任何主流特征提取模型,如 ResNet、DenseNet 和 SENet 等模型,在性能方面强于简单随机采样、自组织映射等其他方法.

综上所述,本文做出了如下贡献:

(1) 提出一种新的数据划分方法,与其他数据划分方法相比,使用完全相同的模型结构进行训练能得到更高的准确率、更低的偏差与方差.

(2) 提供一种度量样本相似性方法.通过此方法,可以在分类任务中更准确地了解所采集样本的显著程度和样本分布特点.

(3) 所提出的度量方法还提供了一种在分类任务中获取小概率稀有样本的途径.

1 相关工作

合理的数据划分问题可以视为统计采样问题,因此,可以使用各种经典的统计采样技术来划分数据.在以往的研究中,采用的数据划分方法大致分为以下几种:简单随机采样、系统采样、DUPLEX 采样和分层采样.

1.1 简单随机采样(SRS)

简单随机采样是最常使用的方法,其具有高效且易于实现的特点.该方法随机选择分布均匀的样本,每个样本具有同等的选中概率.这种方法的优点是所获得的模型具有低偏差[4].但是,对于更复杂(非均匀分布)的数据集,若划分出来的子数据集不能完全覆盖数据分布的特点会导致模型具有较大的方差[4,6].

1.2 试错(Trial and error)法

试错法试图通过重复多次随机采样然后取平均来克服 SRS 中高方差的不足之处[7].简单的试错法表明,在有相同代表性的数据集上,其偏差具有相似的统计特性.为了最小化这种统计差异,使用较为复杂的策略,例如循环优化搜索以寻找潜在的可能拆分组合.使用各种统计学标准,例如均值、方差和 Kolmogorov-Smirnov 统计.这些方法的主要缺点是计算量大且理论基础模糊,无疑在这种情况下所获得的神经网络性能并不稳定.

1.3 系统采样方法

系统采样是为含有自然序的数据集进行采样的一种确定性方法.一种方法是沿着输出变量的维度对样本数据进行排序,以获取能够代表输出变量分布的样本[8].此方法易于实现,因为其假定可以将输出变量映射到唯一的输入状态.但是,当有多个输入状态产生相同的输出时,此假设可能不适用,并且这种做法不能保证采集得到的样本能够完全代表所有可能的输入和输出组合,因为只有输出变量被考虑到了其中.同时对于大多数类型的数据集(例如多媒体数据、基因序列等)而言,很难找到一种合适的排序.对于无序数据,系统采样的结果与 SRS 存在相

同的问题. 系统采样的另一个缺点是对数据集的周期性较为敏感.

1.4　DUPLEX 采样

DUPLEX 采样方法是由 Snee[9] 提出的基于欧几里德距离的数据划分方法. 应用该方法时，将在欧几里德距离上最远的两个点分配给第 1 个数据集，再将列表中剩余的样本之间相距最远的下一对点分配到第 2 个数据集中；重复此过程，直至所有数据分配至要划分的两个数据集中. May 等[4] 对原始 DUPLEX 算法进行了修改，将数据划分为 3 个数据集，分别为由人工神经网络模型开发生成训练、验证和测试数据三部分.

1.5　分层采样

分层采样的基本思想是探索数据集的内部结构和分布，并使用其来划分相对统一的样本组（层、簇）. 该方法可以确保训练子集完全覆盖输入空间的所有区域. 另一方面，对于分布均匀的数据集，可以将分层采样与 SRS 进行比较. 各种聚类算法[10] 可用于数据划分，包括 *C*-means 聚类，模糊 *C*-means 聚类和自组织映射（SOM）[11]. May 等[4] 提出基于自组织映射分层采样（SBSS）两步数据分割方法具有很强的稳健性，可以生成更好的人工神经网络模型，比其他技术更有效，在多元和非均匀数据集中更明显.

2　方法

本文提出的基于 Gaussian 混合模型的距离度量学习（DML-GMM）数据划分方法主要由两个主要部分组成：距离度量学习网络和估计网络，如图 1 所示. 其中：$\boldsymbol{x}$ 为输入样本；$\boldsymbol{y}$ 为输出预测值；p 为样本似然概率；$\hat{\pi}$ 为样本的对数似然；FC 为全连接层；$\boldsymbol{v}$ 为样本的向量表示；左侧虚线框图为距离度量学习（DML）网络，该网络包含特征提取（阶段 1～5）和特征向量嵌入（两层 FC），该过程将所有样本转换成可以度量的向量；右侧虚线框图为评估网络，将由距离度量学习网络获得的样本向量表征用 Gaussian 混合模型（GMM）估计其分布. DML-GMM 的工作具体来说包括如下两步：① 距离度量学习网络. 将样本送入卷积网络提取样本特征，再通过 2 层全连接嵌入. 第 1 层全连接输出用于提取样本在高维特征空间的向量表示；第 2 层全连接用于输出分类预测结果向量 $\boldsymbol{y}$，两层输出都用于该过程独立训练优化，目的是通过该网络将所有样本转换成可以度量的高维向量表示. ② 估计网络. 将第 1 步提取获得的样本高维向量表示，使用 Gaussian 混合模型估计概率密度及其分布.

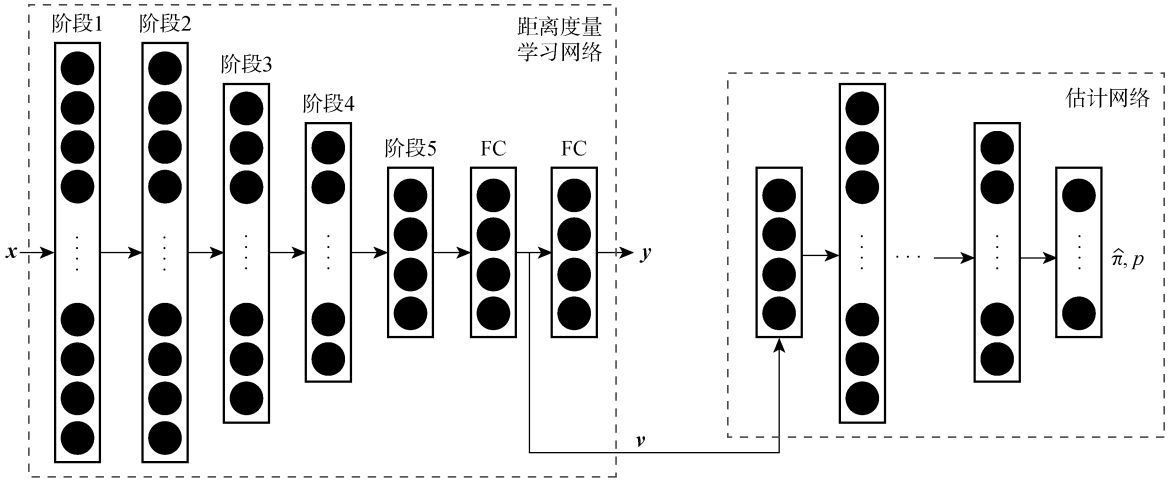

图 1　DML-GMM 框架

Fig. 1　Framework of DML-GMM

2.1　距离度量学习网络

距离度量学习问题在机器学习产生时就存在，数据和数据间的相似度差异是可以度量的. 因此，距离相近的数据将被视为相似，而相距较远的数据将被视为是不同的[12]. 样本之间的这种相似性度量需要在一个合理并且可测的空间中进行. May 等[4] 使用自组织映射方法将原始数据映射到新的嵌入空间，然后通过分层采样对数据进行划分，从而改善训练模型的偏差和方差. Fernández 等[13] 通过简单的前馈神经网络将样本转换到嵌入空间，通过计算嵌入空间特征向量的相似度来减少样本冗余并加快训练过程. Baglaeva 等[14] 通过多层感知机对嵌入空间中的原始数据进行重新划分，建立更精确的空间分布模型，用于模拟城市化地区表层土壤中 Cr 元素的

含量.将样本转换到嵌入空间进行描述有助于更准确地理解样本的特征,并且嵌入空间中样本之间的相似性可以进行度量.受此启发,DML-GMM 方法的距离度量学习网络主要分为 2 个阶段:第 1 阶段为特征提取,可以使用常见的分类模型进行特征提取;第 2 阶段为嵌入过程,将提取获得的特征图映射到一个可以度量的空间.该过程包括以下两部分,如图 2 所示.其中,左侧虚线框图为特征提取模块,可以使用常见的卷积网络提取数据的特征,如 ResNet50;右侧虚线框图为特征映射模块,通过两层全连接将提取的特征映射到嵌入高维空间中,进而进行特征转换.第 1 层全连接层用于输出样本的向量表示,第 2 层全连接层用于输出预测结果向量.

在基于内容的图像检索和人脸识别等方面获得具有稳健性高且有区分度的特征表示非常重要.但是,在特征学习中通过监督学习的交叉熵损失函数优化并不能学习到足够的区分度,因为其仅注重于找到决策边界以分离不同类别的形状,而没有考虑特征的类内紧凑性[15].为了解决这个问题,目前有许多深度度量学习算法损失函数被提出.首先,介绍两种重要且常用的损失函数:三重损失和中心损失.

三重损失在学习特征表示时,将具有相同类别样本点之间的特征距离拉得比具有不同类别样本点之间的特征距离更近.在人脸识别问题中,为了学习更多可鉴别性特征,中心损失被提出作为交叉熵的辅助损失函数一起配合使用.中心损失的主要目标是为每个类别的特征学习一个合适的中心,并将同一类别的样本更紧密地拉到相应的中心.三重损失和中心损失的作用效果如图 3 所示.由于三重损失处理时需要将数据重新组合,构造起来相对复杂,所以本文借鉴了人脸识别中采用的中心损失来优化以提取特征的距离度量学习网络.该优化过程中同时需要用到样本特征向量表示和输出预测结果,即嵌入过程中第 1 层全连接的输出 $\boldsymbol{v}$ 和第 2 层全连接的输出 $\boldsymbol{y}$(见图 2).由此,可以通过中心损失获得更好的样本向量表示,其损失函数 L_c 可表示为

$$L_c = \frac{1}{2}\sum_{i=1}^{N} D(\boldsymbol{f}(\boldsymbol{x}^i), \boldsymbol{c}_{z^i}) \tag{1}$$

式中:$\boldsymbol{c}_{z^i} \in \boldsymbol{R}_d$ 为类 z^i 样本通过网络得到的高维特征向量的向量中心;d 为特征维数;$\boldsymbol{f}(\boldsymbol{x}^i)$ 为样本 $\boldsymbol{x}^i$ 映射的高维向量,$\boldsymbol{f}$ 为映射网络;函数 $D(\cdot)$ 为欧氏距离的平方;N 为样本数量.

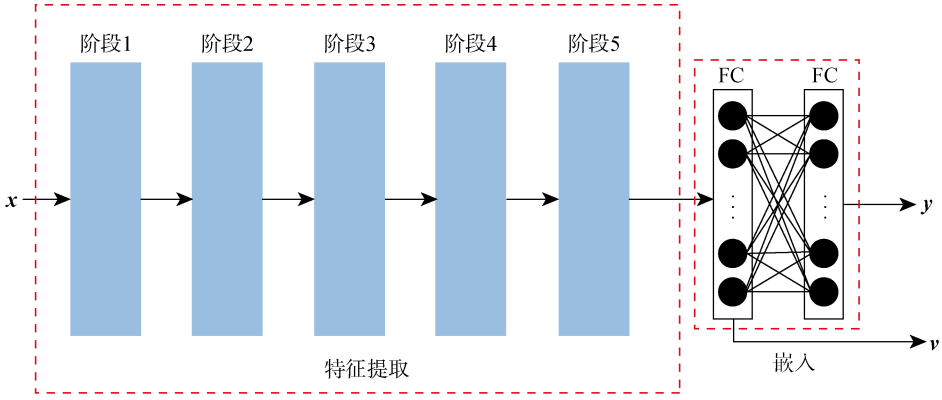

图 2 DML-GMM 中的 DML 网络示意图

Fig. 2 Schematic diagram of DML network in DML-GMM

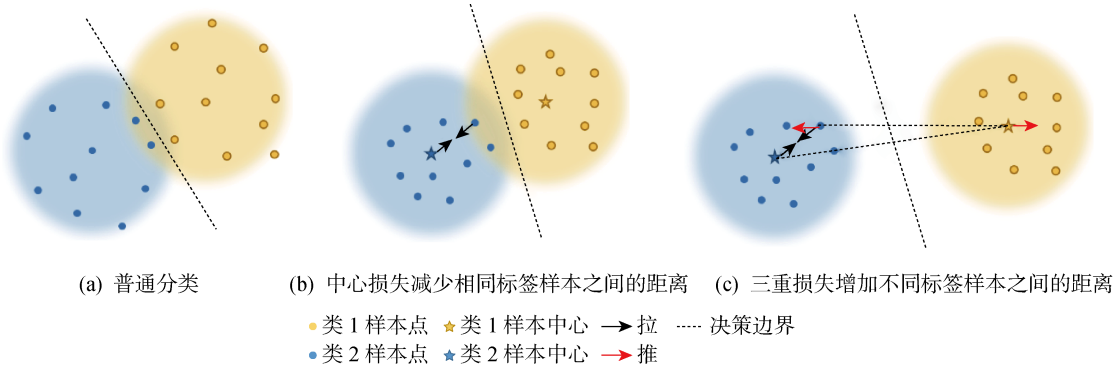

图 3 中心损失和三重损失作用示意图

Fig. 3 Schematic diagram of central loss and triple loss

2.2　基于 Gaussian 混合模型的分布估计

机器学习算法常常将数据转换到合适的度量空间，然后使用聚类算法来衡量两者之间的相似性. Alonso[16] 和 Silva 等[17] 使用 Gaussian 混合聚类模型通过对数据进行距离估计来补充缺失数据. 还有研究人员将 Gaussian 混合模型与深度编码器组合在一起，通过深度编码器将样本投射到另一个空间，保留其中异常检测所需的关键信息，将 Gaussian 混合模型用于估计和检测异常数据[18-20]. 由于 Gaussian 混合模型在高维空间中对于样本的分布和相似表示方面具有良好的性能，所以使用 Gaussian 混合模型来估计样本在高维嵌入空间中的分布. 样本的分布特点可以通过其似然概率来描述，然后通过这种分布估计来进行分层采样以获得更好的数据划分. Gaussian 混合模型由 M 个加权 Gaussian 概率密度函数和所形成的模型，可表示为

$$p(\boldsymbol{X} \mid \lambda) = \sum_{i=1}^{M} \omega_i g(\boldsymbol{X} \mid \mu_i, \boldsymbol{\Sigma}_i) \tag{2}$$

式中：$\boldsymbol{X}$ 为 d 维连续矢量(即测度或特征向量)；$\lambda = (\mu_1, \cdots, \mu_M, \boldsymbol{\Sigma}_1, \cdots, \boldsymbol{\Sigma}_M, \omega_1, \cdots, \omega_M)$ 为超参数；ω_i 为第 i 个分布的混合权重系数，且满足 $\sum_{i=1}^{M} \omega_i = 1$，$(i = 1, 2, \cdots, M)$；$g(\boldsymbol{X} \mid \mu_i, \boldsymbol{\Sigma}_i)(i = 1, 2, \cdots, M)$ 为 Gaussian 概率密度分量；μ_i 为向量均值；$\boldsymbol{\Sigma}_i$ 为协方差矩阵. 每个 Gaussian 概率密度函数都是 d 维 Gaussian 函数，可表示为

$$g(\boldsymbol{X} \mid \mu_i, \boldsymbol{\Sigma}_i) = \frac{1}{(2\pi)^{d/2} \left| \boldsymbol{\Sigma}_i \right|^{1/2}} \times \exp\left[-\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu}_i)^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_i) \right] \tag{3}$$

Gaussian 混合模型的估计过程是通过期望最大化(EM)算法实现的，该算法仅能保证到达局部最优点，不能保证该局部最优也是全局最优点. 因此，如果算法从不同的初始化点开始优化，容易生成不同的估计. 考虑到这种不确定性的情况，对其进行多次拟合，并结合评价指标的平均值和标准偏差来选择合理的参数. 这里使用 Bayesian 信息准则(BIC)来预测实际拥有的数据，此准则可以对 Gaussian 混合模型拟合的好坏程度进行评估. BIC 越低，则用于实际预测的数据(进而扩展到真实的、未知的分布)模型效果就越好[21-22].

3　实验和结果

采用几种不同数据划分方法分别在手写数字数据集 MNIST(类似于 MNIST 数据集的时尚产品图片数据集)、Fashion-MNIST、CIFAR-10 这 3 个开源数据集，以及医院实际采集的临床肺腺癌高分辨率电子计算机断层扫描 (HRCT)图像 4 个不同的数据集上对于图像分类任务的结果进行比较. 上文讨论的各种不同的数据划分方法中，简单随机采样是一种最常用的方法，试错法由于其理论模糊不便于比较，系统采样较难找到一个合理的排序，自组织映射分层采样法相较于 DUPLEX 法在受到数据分布有显著影响的某些网络上是一种更佳的采样方法[4]. 因此，本文将比较以下几种数据划分的方法，即 SRS、SBSS、DML-GMM.

3.1　实验数据

(1) MNIST：来自美国国家标准与技术研究所(NIST)手写数字开源数据库，该数据包含训练集 6×10^4 个示例，测试集 1×10^4 个示例，其是 NIST 数据集合的子集. 这些数字已进行尺寸标准化，并在固定尺寸的图像中央.

(2) Fashion-MNIST：Fashion-MNIST 数据集是德国 Zalando 公司提供的服饰图像开源数据集，包含 6×10^4 个样本的训练集和 1×10^4 个样本的测试集. 每个样本都是 28 像素×28 像素的灰度图像，与 10 个类别的标签相关联.

(3) CIFAR-10：CIFAR-10 数据集由 10 个类的 6×10^4 张 32 像素×32 像素的彩色图像组成的开源数据集，每个类有 6×10^3 张图像，有 5×10^4 张训练图像和 1×10^4 张测试图像.

(4) 肺腺癌 HRCT：为了结合实际情况进一步分析数据划分对小样本数据训练的影响，采集来自一家三甲医院的肺腺癌影像临床数据. 该 HRCT 图像一共包含 1 622 个样本的两种类型数据，即 715 例浸润性肺腺癌 CT 图像和 907 例非浸润性腺癌 CT 图像.

3.2　评价指标

交叉验证是一种用于评估模型的统计方法，也是一个重采样过程，可在有限数量的数据样本上评估学习到的模型. 通过计算 5 折交叉验证的平均准确率、平均方差和平均偏差来比较不同数据分区方法的性能. 具体流程如下：

(1) 使用不同的划分方法将数据分为 5 组；

(2) 进行 5 折交叉验证，取其中之一作为测试集，并随机选择其余 4 组中的 1 组作为验证；

(3) 确定模型训练的终止点，其余 3 组作为训练集训练模型.

计算 5 折交叉验证的平均准确率 $\bar{A}$，平均方差 $\bar{\sigma}^2$，平均偏差 $\bar{V}$，接受者操作特征曲线(ROC)下平均面积 $\overline{\mathrm{AUC}}$，表达式如下：

$$\bar{A}=\frac{1}{5}\sum_{i=1}^{5}A_i \tag{4}$$

$$\bar{\sigma}^2=\frac{1}{5}\sum_{i=1}^{5}\sigma_i^2 \tag{5}$$

$$\bar{V}=\frac{1}{5}\sum_{i=1}^{5}V_i \tag{6}$$

式中：$A_i=\frac{1}{N}\sum_{j=1}^{N}\hat{\varphi}(\boldsymbol{x}_j)$ 为每次交叉验证的准确率；$\sigma_i^2=\frac{1}{N}\sum_{j=1}^{N}\left[\hat{\varphi}(\boldsymbol{x}_j)-\frac{1}{N}\sum_{j=1}^{N}\hat{\varphi}(\boldsymbol{x}_j)\right]^2$ 为每次交叉验证的方差；$V_i=\frac{1}{N-1}\sum_{j=1}^{N}(\hat{\varphi}(\boldsymbol{x}_j)-1)^2$ 为每次交叉验证的偏差；

$$\hat{\varphi}(\boldsymbol{x}_j)=\begin{cases}1, & h(\boldsymbol{x}_j)=\boldsymbol{y}_j\\ 0, & h(\boldsymbol{x}_j)\neq\boldsymbol{y}_j\end{cases},\ j=1,\cdots,N$$

为每个样本分类问题预测输出的正确性，$h(\boldsymbol{x}_j)$ 为样本的预测标签；$\boldsymbol{y}_j$ 为样本的标签.

$$\overline{\mathrm{AUC}}=\frac{1}{5}\sum_{i=1}^{5}\mathrm{AUC}_i \tag{7}$$

式中：AUC_i 为第 i 次实验中 ROC 曲线下面积.

3.3 实验步骤

(1) 所有样本放入 DML 网络中训练 50 次，并将样本转换为嵌入空间中的高维特征向量.

(2) 数据按 SRS、SBSS、DML-GMM 3 种不同方法划分，将数据划分为训练集、验证集和测试集以进行模型训练. 验证集用于判断模型训练的终点，训练结构完全相同的网络进行比较.

(3) 采用 5 折交叉验证方法比较训练得到的模型性能指标.

3.4 结果

3.4.1 MNIST 对于 MNIST 数据，SRS 方法可以直接通过随机采样的方式划分数据集，而 SBSS 和 DML-GMM 方法需要先将样本通过度量网络(6 层卷积层和 2 层全连接层)将样本转换成高维可度量向量，再按照其对应的方法进行数据划分，将重新划分好的数据放入相同的网络中进行训练. 由于 MNIST 数据质量比较高，使用相对较浅层的多层卷积网络就能得到比较好的效果，所以为了更明显地观察出数据划分带来的影响，没有使用特征提取能力更强的网络模型(如 ResNet50)来分类效果验证，这里同样仅使用了一个 6 层卷积层和 2 层全连接层构成网络来比较分类效果. 分别使用 3 种不同的方法划分数据后，训练结构完全相同的网络，采用 5 折交叉得到的模型性能指标如图 4 所示，其在 MNIST 数据集上的性能对比如表 1 所示.

由表 1 可知，在 MNIST 数据集上使用不同的数据划分方法，采用交叉熵损失函数提取样本特征，再通过 SBSS 方法划分数据集训练模型与由 SRS 方

表 1 不同方法在 MNIST 数据集上的性能对比

Tab. 1 Performance comparison on MNIST dataset using different methods

方法	损失函数	$\bar{V}\times10^4$	$\bar{\sigma}\times10^3$	$\bar{A}/\%$	$\overline{\mathrm{AUC}}\times10^2$
SRS	交叉熵损失	4.48	20.681	97.887	98.821
SBSS	交叉熵损失	4.56	20.800	97.874	98.815
SBSS	中心损失	2.96	16.890	98.281	99.042
DML-GMM	交叉熵损失	1.84	13.293	98.652	99.244
DML-GMM	中心损失	1.20	10.761	98.912	99.392

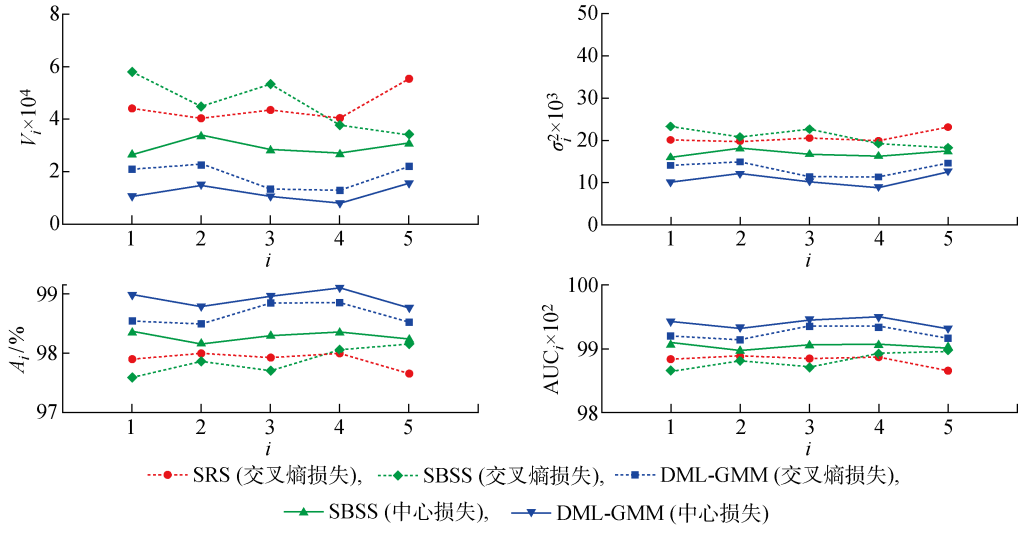

图 4 在 MNIST 数据集上由 5 折交叉验证得到的模型性能指标

Fig. 4 Model performance indicators obtained by 5 folder cross-validation on MNIST dataset

法获得的模型在偏差和方差方面没有明显差异，模型的准确率和 $\overline{\text{AUC}}$ 的差异也不大. 使用 DML-GMM 方法划分数据训练得到的模型偏差和方差更小，模型准确率和 $\overline{\text{AUC}}$ 也更高. 另外，使用中心损失函数提取样本特征，再通过 SBSS 方法和 DML-GMM 方法都可以得到偏差和方差更小、准确率和 $\overline{\text{AUC}}$ 更好的模型.

3.4.2　Fashion-MNIST　对于 Fashion-MNSIT 数据集，基于 ResNet50 模型通过交叉熵损失函数和中心损失函数提取高维特征，再通过两层全连接层将特征转到嵌入空间，分别使用 SBSS 和 DML-GMM 方法重新划分样本，对比 SRS 方法训练相同 ResNet50 模型得到的模型性能指标如图 5 所示，其在 Fashion-MNIST 数据集上的性能对比见表 2.

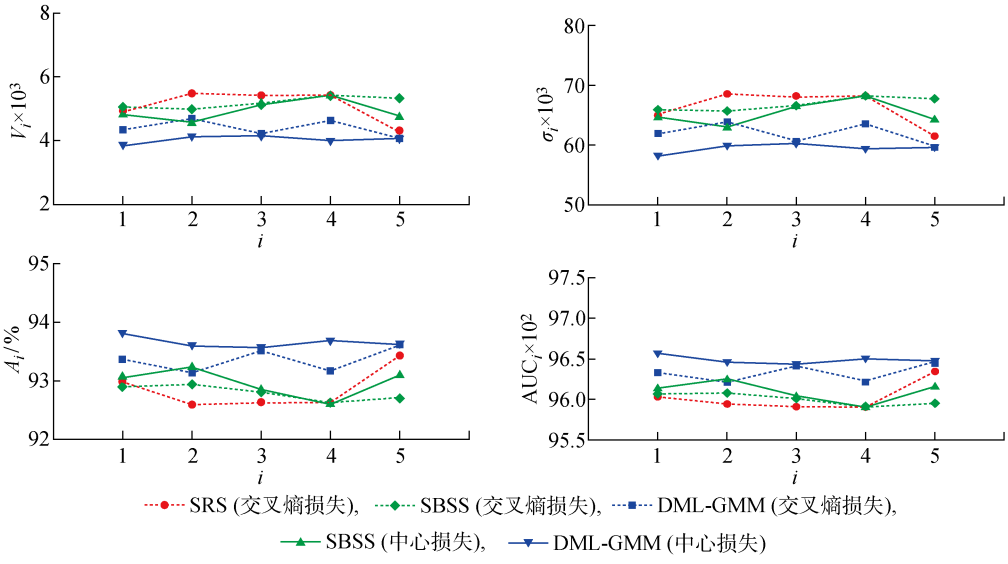

图 5　在 Fashion-MNIST 数据集上由 5 折交叉验证得到的模型性能指标

Fig. 5　Model performance indicators obtained by 5 folder cross-validationon on Fashion-MNIST dataset

表 2　不同方法在 Fashion-MNIST 数据集上的性能对比

Tab. 2　Performance comparison on Fashion-MNIST dataset using different methods

方法	损失函数	$\bar{V}\times10^3$	$\bar{\sigma}^2\times10^3$	$\bar{A}$/%	$\overline{\text{AUC}}\times10^2$
SRS	交叉熵损失	5.104	66.267	92.863	96.033
SBSS	交叉熵损失	5.181	66.790	92.803	96.002
SBSS	中心损失	4.943	65.329	92.973	96.096
DML-GMM	交叉熵损失	4.404	61.937	93.366	96.327
DML-GMM	中心损失	4.027	59.428	93.654	96.486

由表 2 可知，在 Fashion-MNIST 数据集上使用不同的数据划分方法划分得出的实验结果. 当使用交叉熵损失函数提取样本特征嵌入后，使用 SBSS 方法划分数据相较于 SRS 方法训练出来的模型偏差和方差略微高一些，而使用 DML-GMM 方法得到的偏差和方差更小. 当使用中心损失函数提取样本使用 SBSS 比 SRS 方法训练得到的模型偏差、方差更小，使用 DML-GMM 方法划分后得到的偏差和方差进一步减小，准确率和 $\overline{\text{AUC}}$ 进一步提高.

3.4.3　CIFAR-10　对于 CIFAR-10 数据集，使用交叉熵损失函数和中心损失损失函数提取特征，嵌入后分别使用 SBSS 和 DML-GMM 方法重新划分样本，对比 SRS 方法训练相同 ResNet50 网络得到的模型性能指标如图 6 所示，其在 CIFAR-10 数据集上的性能对比如表 3 所示. 由表 3 可知，在 CIFAR-10 数据集上，使用交叉熵损失函数提取样本特征嵌入后，使用 SBSS 方法划分数据相较于 SRS 方法训练出来的模型偏差和方差略微小一些，模型性能更好. 而使用 DML-GMM 方法得到的偏差和方差更小，模型性能进一步提升. 使用中心损失提取样本特征使用 SBSS 和 DML-GMM 方法比交叉熵提取样本特征得到的模型性能进一步有所提高.

表 3　不同方法在 CIFAR-10 数据集上的性能对比

Tab. 3　Performance comparison on CIFAR-10 dataset using different methods

方法	损失函数	$\bar{V}\times10^3$	$\bar{\sigma}^2\times10^3$	$\bar{A}$/%	$\overline{\text{AUC}}\times10^2$
SRS	交叉熵损失	11.013	93.754	89.523	94.181
SBSS	交叉熵损失	10.658	92.458	89.688	94.271
SBSS	中心损失	10.294	90.972	89.873	94.378
DML-GMM	交叉熵损失	9.257	86.725	90.402	94.672
DML-GMM	中心损失	8.807	84.745	90.645	94.807

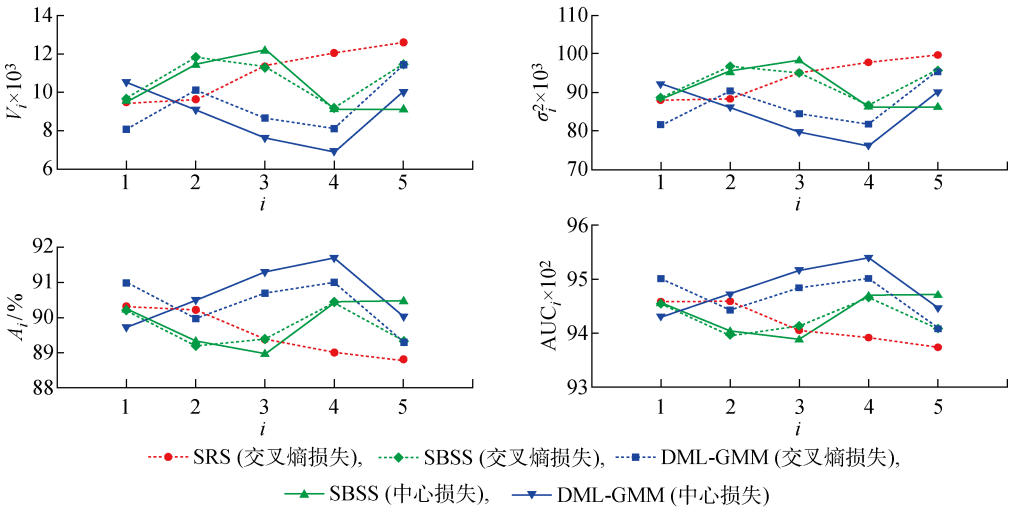

图 6 在 CIFAR-10 数据集上由 5 折交叉验证得到的模型性能指标

Fig. 6 Model performance indicators obtained by 5 folder cross-validationon on CIFAR-10 dataset

3.4.4 肺腺癌 HRCT 对于肺腺癌 HRCT 数据集，在提取特征方面使用的是 3D ResNet50 模型，损失函数分别使用的是交叉熵和中心损失函数，嵌入后分别使用 SBSS 和 DML-GMM 方法重新划分样本，对比 SRS 方法训练相同 3D ResNet50 模型得到的性能指标如图 7 所示，其在肺腺癌 HRCT 数据集上的性能对比如表 4 所示. 由表 4 可知，在肺腺癌 HRCT 数据集上，使用交叉熵损失函数提取样本特征嵌入后，使用 SBSS 方法划分数据相较于 SRS 方法训练出来的模型偏差和方差略小一些，模型性能更好. 而使用 DML-GMM 方法得到的偏差和方差更小，模型性能进一步提升. 使用中心损失提取样本特征使用 SBSS 和 DML-GMM 方法比交叉熵提取样本特征得到的模型性能进一步有所提高.

通过上述 4 个数据集的实验可以看到，在特征提取环节使用中心损失函数相较于交叉熵损失函数可以获得区分度更高的嵌入空间，更加有利于对样本之间差异性的度量. 使用 SBSS 方法划分样本相较于 SRS 方法中简单的处理样本可以提高训练得

表 4 不同方法在肺腺癌 HRCT 数据集上的性能对比

Tab. 4 Performance comparison on adenocarcinoma HRCT dataset using different methods

方法	损失函数	$\bar{V}\times10^3$	$\bar{\sigma}^2\times10^3$	$\bar{A}/\%$	$\overline{\mathrm{AUC}}\times10^2$
SRS	交叉熵损失	35.911	153.348	81.074	80.609
SBSS	交叉熵损失	35.827	152.838	81.134	80.585
SBSS	中心损失	34.633	149.745	81.562	81.159
DML-GMM	交叉熵损失	33.727	149.285	81.699	81.004
DML-GMM	中心损失	31.990	146.647	82.136	81.485

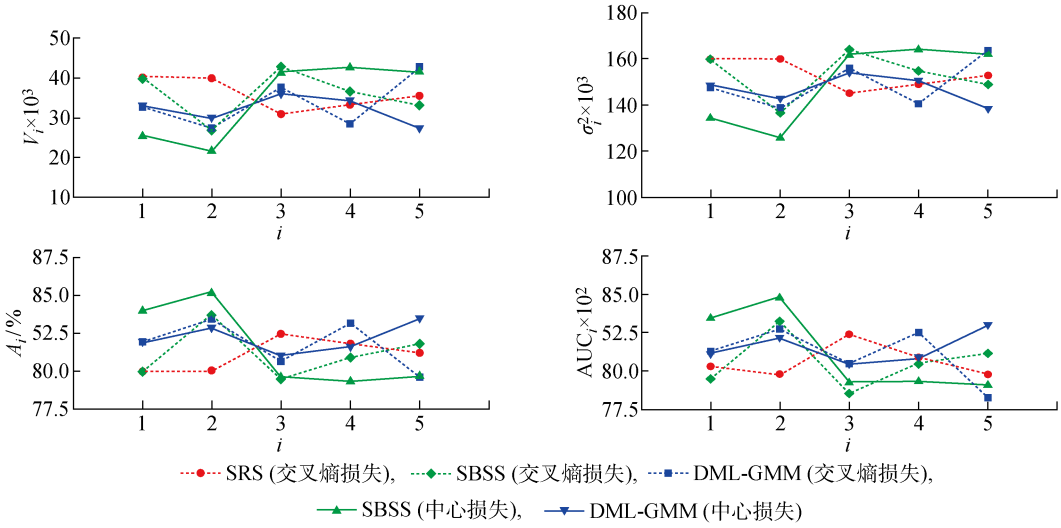

图 7 在肺腺癌 HRCT 数据集上由 5 折交叉验证得到的模型性能指标

Fig. 7 Model performance indicators obtained by 5 folder cross-validation on adenocarcinoma HRCT dataset

到的性能更好、稳定性更佳的模型，而使用 DML-GMM 方法其模型性能可以进一步获得提升.

3.5　稀有样本测试实验

综上，通过使用 Gaussian 混合模型来估计在嵌入空间中的样本，可计算出每个样本的对数似然，并用其似然概率描述分布特点和样本的典型性，进而发现特征不明显的稀有样本. 根据如下步骤进行实验：① 使用 SRS 方法随机从 MNIST 数据集中抽取 20％作为测试数据，将剩余 80％用于训练一个分类网络（6 层卷积层和 2 层全连接层）；② 提取用于训练分类网络的样本向量表示，将其特征的高维向量表示通过 Gaussian 混合模型描述其样本分布；③ 将测试数据输入训练好的网络，标记出正确分类和错误分类的样本，并放入②中建立好的 Gaussian 混合模型中计算测试样本的对数似然. 所得结果如图 8 所示，将测试数据中 10 类对数似然最高和最低的样本显示出来（见图 8(a)）. 由图 8 可知，对数似然值越高的样本其典型性越高，其特征也越显著、越容易正确分类. 样本对数似然值越低其特征显著程度也越低，越不容易分类正确. 将具有某一个相同标签的测试数据嵌入到 2D 空间并用热图对其概率密度进行描述（见图 8(b)），可以看到其对数似然值越大，其样本分布在越密集的位置，其对数似然值越低，样本分布在越稀疏的位置. 将正确分类和错误分类样本的对数似然用核密度函数描述直方图分布情况（见图 8(c)）. 其中：ρ 为样本密度；$\ln p$ 为对数似然概率. 由图 8(c)可见，正确分类的样本其 $\ln p$ 越大，错误分类的样本其 $\ln p$ 相对更低. 因此在实践中，当采集新的样本时，将该样本通过已有样本建立的 Gaussian 混合模型，计算其 $\ln p$ 就可以量化样本的显著程度和稀有程度，进而判断是否需对已有模型进行迭代更新，以提高模型的泛化能力.

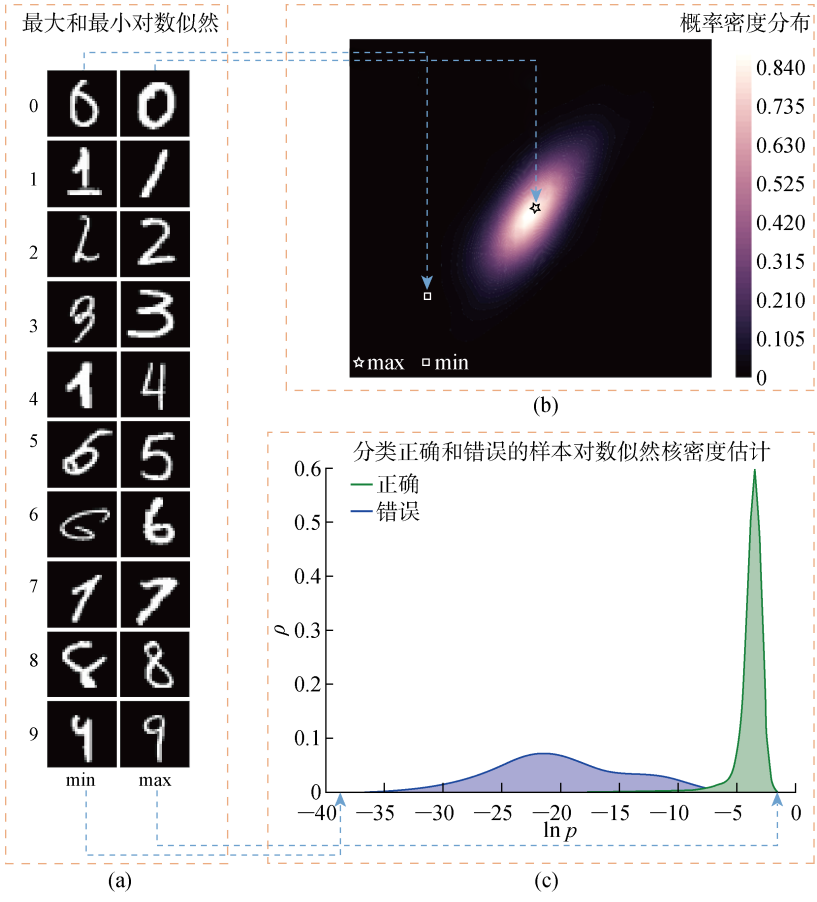

图 8　在 MNIST 数据集上，通过 GMM 获取的样本对数似然分布及其示意图

Fig. 8　Log-likelihood distribution and schematic diagram of samples by GMM on MNIST dataset

4　结语

本文提出一种基于 Gaussian 混合模型的距离度量学习数据集划分方法. 首先，将所有样本通过 DML 网络训练，将样本从图像空间转换到高维特征嵌入空间；然后通过 Gaussian 混合模型描述其分布后，分层采样划分数据集进行模型训练. 通过该方法可以更加准确地了解数据分布的特点，在这样的条

件下划分数据相比于其他方法能训练出偏差、方差更小,准确率更高,泛化性能更好的模型.另外,该方法还可以更好地理解样本的显著性,更清楚地了解哪些是最重要的样本,哪些是稀有样本.

参考文献:

[1] YU Y L, JI Z, GUO J C, *et al*. Transductive zero-shot learning with adaptive structural embedding[C]// **IEEE Transactions on Neural Networks and Learning Systems**. Piscataway, NJ, USA: IEEE, 2018: 4116-4127.

[2] SHEN D G, WU G R, SUK H I. Deep learning in medical image analysis[J]. **Annual Review of Biomedical Engineering**, 2017, 19(1): 221-248.

[3] XIONG C M. Recent progress in deep reinforcement learning for computer vision and NLP[C]// **Proceedings of the 2017 Workshop on Recognizing Families in the Wild**. New York, NY, USA: ACM Press, 2017: 1.

[4] MAY R J, MAIER H R, DANDY G C. Data splitting for artificial neural networks using SOM-based stratified sampling[J]. **Neural Networks**, 2010, 23(2): 283-294.

[5] ROBERTS D R, BAHN V, CIUTI S, *et al*. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure[J]. **Ecography**, 2017, 40(8): 913-929.

[6] PAL S K, SINGH H P, KUMAR S, *et al*. A family of efficient estimators of the finite population mean in simple random sampling[J]. **Journal of Statistical Computation and Simulation**, 2018, 88(5): 920-934.

[7] REITERMANOVA Z. Data splitting [EB/OL]. (2010-06-03) [2019-12-11]. https://www.mff.cuni.cz/veda/konference/wds/proc/pdf10/WDS10_105_i1_Reitermanova.pdf.

[8] BAXTER C W, STANLEY S J, ZHANG Q, *et al*. Developing artificial neural network models of water treatment processes: A guide for utilities[J]. **Journal of Environmental Engineering and Science**, 2002, 1(3): 201-211.

[9] SNEE R D. Validation of regression models: Methods and examples[J]. **Technometrics**, 1977, 19(4): 415-428.

[10] HADI A S, KAUFMAN L, ROUSSEEUW P J. Finding groups in data: An introduction to cluster analysis[J]. **Technometrics**, 1992, 34(1): 111.

[11] DOUZAS G, BACAO F. Self-organizing map oversampling (SOMO) for imbalanced data set learning [J]. **Expert Systems with Applications**, 2017, 82: 40-52.

[12] SUÁREZ J L, GARCÍA S, HERRERA F. A tutorial on distance metric learning: Mathematical foundations, algorithms and experiments [EB/OL]. (2018-12-14)[2019-12-11]. https://arxiv.org/abs/1812.05944.

[13] FERNÁNDEZ J J M, MAYERLE R. Sample selection via angular distance in the space of the arguments of an artificial neural network[J]. **Computers & Geosciences**, 2018, 114: 98-106.

[14] BAGLAEVA E M, SERGEEV A P, SHICHKIN A V, *et al*. The effect of splitting of raw data into training and test subsets on the accuracy of predicting spatial distribution by a multilayer perceptron[J]. **Mathematical Geosciences**, 2020, 52(1): 111-121.

[15] HE X W, ZHOU Y, ZHOU Z C, *et al*. Triplet-center loss for multi-view 3D object retrieval[C]// **2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition**. Piscataway, NJ, USA: IEEE, 2018: 1945-1954.

[16] ALONSO A G. Probability density imputation of missing data with Gaussian Mixture Models[D]. Massachusetts, USA: University of Oxford, 2017.

[17] SILVA D S F, DEUTSCH C V. Multivariate data imputation using Gaussian mixture models[J]. **Spatial Statistics**, 2018, 27: 74-90.

[18] ZONG B, SONG Q, MIN M R, *et al*. Deep autoencoding Gaussian Mixture Model for unsupervised anomaly detection[C]// **Sixth International Conference on Learning Representations**. Vancouver, Canada: ICLR, 2018: 1-19.

[19] LI L S, HANSMAN R J, PALACIOS R, *et al*. Anomaly detection via a Gaussian Mixture Model for flight operation and safety monitoring[J]. **Transportation Research Part C: Emerging Technologies**, 2016, 64: 45-57.

[20] FAN Y X, WEN G J, LI D R, *et al*. Video anomaly detection and localization via Gaussian Mixture Fully Convolutional Variational Autoencoder[J]. **Computer Vision and Image Understanding**, 2020, 195: 102920.

[21] MA J Y, JIANG J J, LIU C Y, *et al*. Feature guided Gaussian mixture model with semi-supervised EM and local geometric constraint for retinal image registration[J]. **Information Sciences**, 2017, 417: 128-142.

[22] HUANG T, PENG H, ZHANG K. Model selection for Gaussian mixture models[J]. **Statistica Sinica**, 2017: 147-169.

(本文编辑:石易文)