无监督学习AE和MVO-DBSCAN结合LIF在煤矿突水识别中的应用

来文豪[1]，周孟然[1*]，李大同[1]，王 亚[2]，胡 锋[1]，赵 舜[3]，顾煜林[1]

1. 安徽理工大学电气与信息工程学院，安徽 淮南 232000
2. 阜阳师范学院计算机与信息工程学院，安徽 阜阳 236000
3. School of Electronic and Electrical Engineering，University of Leeds，Woodhouse Lane，Leeds L859JT，UK

摘 要 快速准确的识别突水类型和突水来源对煤矿安全开采具有重要意义，激光诱导荧光(LIF)技术在检测中具有快速性和灵敏性，将LIF应用于煤矿突水的检测，再结合模式识别算法，可快速识别出突水来源。目前用于识别水样光谱的算法过于依赖预先建立的水样光谱数据库，当突水水源不在该库中时，易引发误识别。无监督学习算法DBSCAN在聚类时不需样本集的标签和类别信息，能降低对未知类别的误识别，因此把DBSCAN算法用于突水的激光诱导荧光光谱识别，并将MVO用于DBSCAN的参数寻优，省去繁琐的人工参数寻优过程。实验中，从谢桥煤矿采水点获取四个水样，利用像素为2 048的USB2000+光谱仪采集水样的荧光光谱，每种水样采集30组光谱数据。首先，利用无监督学习算法自动编码器(AE)对原始光谱数据降维，以减少光谱数据中冗余信息对聚类的影响，设计的AE的结构是介于浅层和深层之间的多层网络模型，可将原始光谱数据降到2维；为使降维模型具有稀疏性，在传统的AE算法中加入一个Dropout层，由实验可知，加入Dropout层后的降维模型具有较快的收敛速度。将多元宇宙优化(MVO)算法用于DBSCAN参数寻优，在参数寻优过程中，DBSCAN对降维后的水样光谱识别率最高为97.5%，此时参数所对应的取值范围为[0.023 66 0.040 65]；为验证AE对水样光谱数据降维的有效性，把归一化后的未降维的光谱数据用于DBSCAN聚类识别，DBSCAN对原始水样光谱的识别率最高为95%，比降维后的后水样光谱识别率低了2.5%，结果表明，使用AE降维光谱数据，可提高DBSCAN对不同光谱的识别率。最后，用监督学习算法K最近邻(KNN)识别降维后的水样光谱，将识别结果和无监督学习算法DBSCAN的识别结果对比，其中训练集选用三种水样，测试集使用四种水样；在测试集中，监督学习算法只能准确地识别训练集所包含的水样类别，但把训练集没有的类别全部识别错误，而DBSCAN能准确的识别出训练集中没有的水样光谱。非线性降维算法AE能实现对高维的水样光谱数据降维，把MVO-DBSCAN用于煤矿突水水源的LIF光谱识别，可有效降低因矿井水源光谱数据库建立不完备而引起的误识别。

关键词 煤矿突水；激光诱导荧光；光谱识别；密度聚类；多元宇宙优化；自动编码器；丢失
中图分类号：O657.3 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2019)08-2437-06

引 言

矿井发生突水，不仅给煤矿开采带来巨大的经济损失，而且严重威胁井下工人的生命安全。中国有着丰富的地下水资源，伴随着开采深度和开采强度的增加，煤矿面临突水的威胁也越来越大，在煤矿发生突水后，准确快速地判断突水类型和突水来源，采取最为合适的应对措施，能最大程度减少损失。激光诱导荧光(laser induced fluorescence，LIF)技术具有很好的快速性和灵敏性，在检测中应用较为广泛，如：王文坦等[1]将LIF技术用于液体混合的定量可视化测量研究，张会书等[2]将LIF用于测量规整填料内的液体分布，李嘉铭等[3]将LIF结合激光诱导击穿光谱用于检测波玻璃中的微量元素，等。将LIF技术应用于矿井突水检测，再结合

收稿日期：2018-12-19，**修订日期**：2019-04-30
基金项目：国家“十二五”科技支撑计划重点项目(2013BAK06B01)，国家安全生产重大事故防治关键技术科技项目(anhui-0001-2016AQ)，国家自然科学基金项目(51174258)，安徽省自然科学基金面上项目(1808085MF202)，安徽省高校科学研究重大项目(KJ2018ZD036)资助
作者简介：来文豪，1992年生，安徽理工大学电气与信息工程学院博士研究生 e-mail：whlai9@163.com
*通讯联系人 e-mail：mrzhou8521@163.com

模式识别算法，可快速的判别出矿井突水类型及来源。目前已用于矿井突水光谱识别的算法可分为两类，一是监督学习算法，如 K 最邻近(KNN)，极限学习机(ELM)[4]，主成分分析(PCA)，反向传播(BP)，卷积神经网络(CNN)[5]和偏最小二乘判别分析(PLS-DA)，MVO(多元宇宙优化)，AE(自动编码器)等，二是无监督学习算法，如 FCM 算法。基于监督学习算法的突水光谱识别，虽有较高的准确率，但是需预先建立不同水源的光谱数据库，由于矿井环境较为复杂，难以预先获取井下所有水源的光谱数据；当发生的突水水源不在预先建立的光谱数据库中，利用监督学习算法识别突水光谱，就会引起误判；无监督学习算法 FCM 在聚类时，虽不需预先知道所有样本的具体标签，但是仍要准确知道被聚类样本的类别信息，因此，需要一种不依赖于样本标签和类别信息的水样光谱识别方法。

DBSCAN 算法是 Ester 等[6]提出的一种基于密度的无监督学习聚类算法，在聚类时 DBSCAN 不需样本的类别数，也不需样本的标签信息，本文将其用于煤矿突水的识别。MVO 是 Mirjalili 等提出的一种较新的启发式寻优算法[7]，具有需要调节的参数较少，适应性强，搜索效率高及优化能力好等特点。Faris 等[8]将 MVO 用于训练多层感知器神经网络，有效避免其陷入局部最优并且还具有较好的收敛速度；本研究将 MVO 用于改进传统的 DBSCAN 算法，以实现其参数自寻优，省去繁琐的人工参数寻优工作。原始光谱数据含有大量高维信息，为降低冗余信息的干扰，减少 DBSCAN 算法的计算量，把无监督学习算法自动编码器用于原始光谱数据的降维，为改善自动编码器降维模型的性能，在传统的 AE 算法中加入一个 Dropout 层。

提出一种基于无监督学习算法的矿井突水光谱识别方法，将基于密度聚类的 DBSCAN 算法用于识别突水光谱，并将 MVO 用于改进 DBSCAN 算法，省去繁琐的人工参数寻优工作；在传统的 AE 算法中加入一个 Dropout 层，用于降维原始光谱数据，减少光谱数据中冗余信息的干扰，提高了 DBSCAN 对突水光谱的识别率。

1 实验部分

1.1 LIF 技术

激光照射被测样品诱使其发出荧光即激光诱导荧光(laser induced fluorescence，LIF)，分子的荧光光谱与荧光物质的能级结构有关，水样中包含的物质不同或其浓度不同，都会有不同的荧光光谱。此外，在溶液中，pH 值和温度[9]对其荧光光谱都有影响。煤矿中不同的含水层，其水化学特征、pH 值和温度等都有差异，将激光诱导荧光用于矿井突水水源识别，能充分利用不同水源的差异性，提高对不同水源的识别率。

将激光诱导荧光技术用于煤矿突水识别，光谱仪选用的是 USB2000＋个性化配置型光谱仪(OceanOptics 公司，美国)，像素为 2048，其他可调参数设置如表 1 所示。激光器为北京华源拓达激光技术有限公司生产的 405 nm 蓝紫光半导体激光器。实验时，激光器发射的高能激光，经石英光纤由浸入式微型探头射入实验水样中，诱导水样发出荧光，再由浸入式微型探头接收荧光传送至光谱仪。

表 1 实验仪器参数

Table 1 The USB2000＋ spectrometer parameters

型号	个性化配置型
分辨率/nm	0.5
光谱范围/nm	400～800
积分时间/(ms·nm^{-1})	1
入射波长/nm	405

1.2 光谱采集

以淮南地区谢桥煤矿为实验区域，从谢桥煤矿各采集点采集四个水样，水样的采集由谢桥矿的水样采集员完成，水样采集点和坐标如表 2 所示。

表 2 水样采集地点

Table 2 Water sample collection location

水样	采集点	Z/m
水样一	东风井井底车场	−440.0
水样二	西翼 B 组 4 煤底板皮带石门联巷	−592.5
水样三	11416 采空区	−705.2
水样四	2212 上顺槽	−731.2

注：Z 采集深度

水样采集好后密封，带回实验室，利用激光诱导荧光设备采集水样的光谱数据，每个水样采集 30 组 LIF 光谱图。四个水样的全部 LIF 光谱图如图 1 所示。

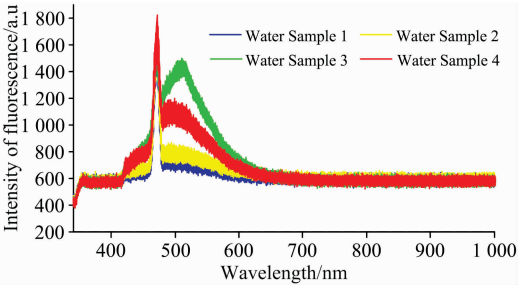

图 1 水样 LIF 光谱图

Fig. 1 LIF spectra of water samples

图 1 为共有 120 组水样 LIF 光谱图，每种颜色各代表一种水样的光谱。从图 1 中可知，四种水样的 LIF 光谱的差异主要集中在波段[400～700 nm]之间，并且水样 3 和水样 4 差异较明显，水样 1 和水样 2 差异相对较小。

2 无监督学习与多元宇宙优化理论

无监督学习(unsupervised learning)，即在学习时不赋予模型非常明确且具体的信息，让模型以无监督的方式自己学习。目前，常用的无监督学习算法可分为 4 种，分别为聚类算法[10]、自动编码器[11]、PredNet 和生成模型[12]，其中聚类

算法已广泛的应用于模式识别和机器学习领域。

2.1　自动编码器及 Dropout 层

由于选用的光谱仪像素为 2 048，获取的每条光谱曲线有 2 048 个数据，选用自动编码器(auto encoder，AE)降维水样光谱数据。AE 最初是由 Rumelhart 提出的一种无监督学习算法，主要用于高维复杂数据的处理，即数据的降维。AE 算法主要包括两部分，一是 Encoder(编码)，二是 Decoder(编码)，其中 Encoder 和 Decoder 可以是任意的学习模型，Rumelhart 曾用神经网络作为学习模型，这在当初也促进了神经网络的发展，自动编码器的模型结构简图如图 2 所示。

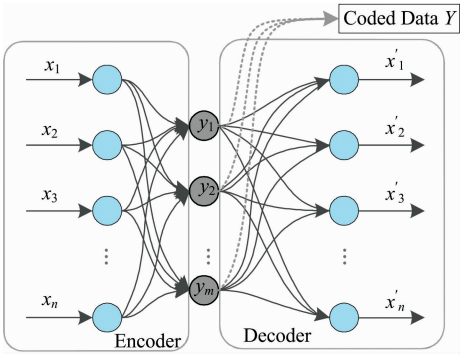

图 2　自动编码器结构图

Fig. 2　The schematic diagram of the AE

图 2 为最基本的 AE 算法结构图，模型的输入节点和输出节点数目相同，编码和解码均采用人工神经网络模型，从 X 到 Y 过程为编码，Y 到 X'过程为解码。当隐层节点数小于输入节点数时($m<n$)，通过学习使输出向量 X'等于输入向量 X，即可实现复杂数据的降维，降维后的数据为 Y。

Dropout 层是 Hinton 等[13]提出的一种用于改进卷积神经网络模型性能的方法，其作用类似于 L2 正则化，可防止卷积神经网络模型过学习，加快模型的训练速度。目前，Dropout 层已广泛应用于基于卷积神经网络的深度学习模型中，为使 AE 降维模型具有稀疏性，同时改善 AE 降维模型的训练性能，文中在传统的 AE 算法中加入一个 Dropout 层。

2.2　基于 MVO 优化的 DBSCAN

1.2.1　MVO 算法

多元宇宙优化算法是 Mirjalili 等在 2015 年提出的一种启发式优化算法[7]，灵感来源于物理学中的多元宇宙理论，主要根据多元宇宙理论的三个主要概念——白洞、黑洞和虫洞来建立数学模型。作为一种新的智能算法，MVO 算法由于需要调节的参数较少，适应性强，搜索效率高及优化能力好等特点，已经成功用于焊接梁设计、压力容器设计和悬臂梁设计等经典工程问题中。

白洞：是一个只发射不吸收的特殊天体，并且是诞生一个宇宙的主要成分；

黑洞：刚好与白洞相反，它吸引宇宙中一切事物，所有的物理定律在黑洞中都会失效；

虫洞：连结白洞和黑洞的多维时空隧道，将个体传送到宇宙的任意角落，甚至是从一个宇宙到另一个宇宙，而多元宇宙通过白洞、黑洞、虫洞相互作用达到一个稳定状态。

1.2.2　DBSCAN 算法

带有噪声的基于密度聚类(density-based spatial clustering of applications with noise，DBSCAN)是 Martin Ester 等提出的一种无监督聚类的算法，广泛应用于模式识别和数据挖掘领域。DBSCAN 的实现思想是根据密度可达关系，在样本数据集中找出最大密度相连样本的集合，被分在该子集的样本即为同一类。关于 DBSCAN 算法几个基本的概念如下：

设样本集为 $C=(c_1,c_2,c_3,\cdots,c_n)$，对于子样本 c_i，其 Eps-neighborhold 包含样本集 C 中与 c_i 的空间距离不大 Eps 的子样本集为 $N_{\mathrm{Eps}}(c_i)=\{c_j\in C\mid \mathrm{dis}(c_i,c_j)\leqslant Eps\}$，样本个数为 $|N_{\mathrm{Eps}}|$。

核心对象：任一样本 $c_i\in C$，如果其 Eps-neighborhold 对应 $N_{\mathrm{Eps}}(c_i)$的至少包含 MinPts 个样本则 c_i 是核心对象；

噪音点：不是核心对象，也不在核心对象的邻域内的样本；

密度直达：若 c_i 的 Eps-neighborhold 为 $N_{\mathrm{Eps}}(c_i)$，且 $c_j\in N_{\mathrm{Eps}}(c_i)$，则称 c_i 是 c_j 密度直达。

密度可达：对于 c_i 和 c_j，存在样本序列满足 $p_1,\cdots,p_n,\cdots,p_t$，且 $p_1=c_i$ 和 $p_t=c_j$，若 p_n 是 p_{n+1}密度直达，则 c_i 和 c_j 密度可达。

3　光谱数据的降维与识别

3.1　光谱数据降维

实验采集的水样 LIF 光谱数据的维度为 2 048，含有较多的冗余信息，利用自动编码器降维水样光谱数据。自动编码器采用多层网络结构，结构简图如图 3 所示。

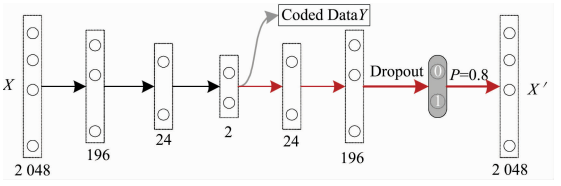

图 3　设计的 AE 降维模型

Fig. 3　The designed AE dimension reduction model

图 3 中，黑色箭头为编码过程，红色箭头为解码过程，第一层和最后一层节点数为 2 048，中间层节点数为 2，即水样 LIF 光谱数据由 2 048 维降到 2 维。为改善 AE 降维模型的训练性能，在输出层前加入 Dropout 层，使网络具有稀疏性。所设计的降维模型实现平台为 Google 公司开发的深度学习框架 Tensorflow-CPU，戴尔笔记本电脑，内存为 8G；降维模型训练时，学习率设为 0.007 5，Dropout 层的参数取值为 0.8，AE 的训练误差曲线如图 4 所示。

实验中，降维模型训练迭代了 2 400 次，为更加直观的看出加入 Dropout 层给降维模型训练速度带来的提升，图 4 中只画出其前 1 000 次迭代的误差曲线。Keep_prob 为模型参数，其越小，模型相对会越稀疏，当 keep_prob 取值为 1 时，表示网络中的 Dropout 被删除。从图 4 可知，AE 中加入 Dropout 层后，初始训练误差会有所增加，但其误差减少速

度也明显增加，大约经历 200 次迭代，其训练误差就已低于未加入 Dropout 层的 AE 的训练误差，大约经历 500 次的迭代，加入 Dropout 层的 AE 降维模型训练完毕，而此时未加入 Dropout 层的 AE 降维模型的训练误差依然较大。经对比分析，自动编码器中加入 Dropout 层可明显加快模型的收敛速度。

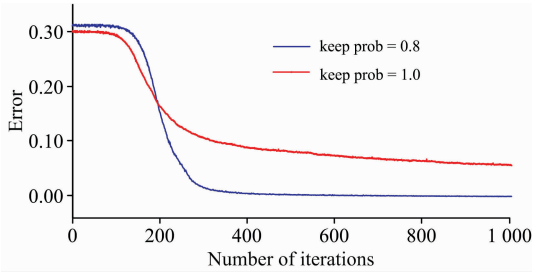

图 4 降维模型的训练误差曲线

Fig. 4 Curve of dimension reduction model training error

3.2 MVO-DBSCAN 水样聚类

DBSCAN 算法聚类时，需要合理的设置邻域距离 Eps 和核心对象的样本最小数目 MinPts。当邻域距离相同时，核心对象样本数越大，DBSCAN 聚类出的簇相对越少，当某一簇所包含的样本数小于参数 MinPts 时，该簇所有样本会被视为噪声，将 DBSCAN 的参数 MinPts 设置为 3。参数 Eps 与样本集的密度相关，最高聚类准确率对应的 Eps 往往是一个区间，将 MVO 用于 DBSCAN 算法的参数寻优，获取最高识别率下 Eps 所对应的取值区间。MVO 算法 WEP 的最大值设置为 1，最小值设置为 0.2，宇宙数设置为 15，优化时最大迭代次数设置为 300。MinPts 取值为 3 时的寻优结果如图 5 所示。

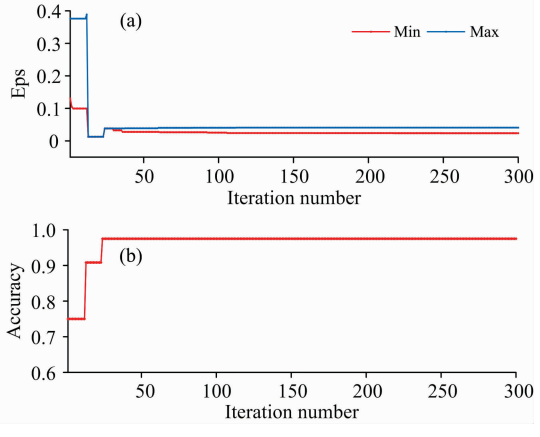

图 5 AE-MVO-DBSCAN 优化结果

(a)：参数 Eps 曲线；(b)：聚类准确率曲线

Fig. 5 Optimization results of AE-MVO-DBSCAN

(a)：The curve of parameter Eps;

(b)：The curve of clustering accuracy

图 5 中，(a)为被寻优参数 Eps 的变化曲线，(b)为寻优过程中 DBSCAN 算法对降维后的水样光谱的识别率曲线。图 5(a)中的“Max”为前 n 次迭代中，最优识别率所对应的 Eps 最大值，“Min”为其所对应的最小值。由图 5 可知，MVO 算法参数自寻优速度较快，经历 24 次迭代对降维后的水样光谱的识别率就已达到最大，为 97.5%；大约经历 120 次迭代，最高识别率所对应的的最大值和最小值已不再改变，其对应区间为[0.023 66　0.040 65]，即当取值在区间[0.023 66　0.040 65]之内时，DBSCAN 对降维后的水样光谱识别率为 97.5%。当 Eps＝(0.023 66＋0.040 65)÷2、MinPts 取值为 3 时，DBSCAN 对降维后的水样识别结果如图 6 所示。

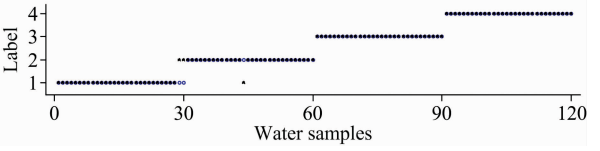

图 6 MinPts＝3，Eps＝0.0322 时 DBSCAN 识别结果

Fig. 6 DBSCAN recognition result when MinPts＝3，Eps＝0.032 2

图 6 为 Eps 取值为 0.032 2 和 MinPts 取值为 3 时，DBSCAN 对降维后的水样光谱的识别结果，蓝色“○”为样本实际标签，红色“☆”为预测标签，红色“☆”和蓝色“○”重合，表示该样本被正确识别。从图 6 中可看出，水样三和水样四全部被识别出，水样一中有 2 个样本被识别成水样二，水样二中有 1 个样本被识别成水样一，综上可知，DBSCAN 对降维后的水样光谱的识别率为 97.5%。

为验证 AE 降维算法对水样光谱数据降维的有效性，将 MVO 寻优 DBSCAN，并将结果用于识别未降维的水样光谱，在参数寻优前将原始光谱数据进行归一化处理。DBSCAN 算法的寻优结果如图 7 所示。

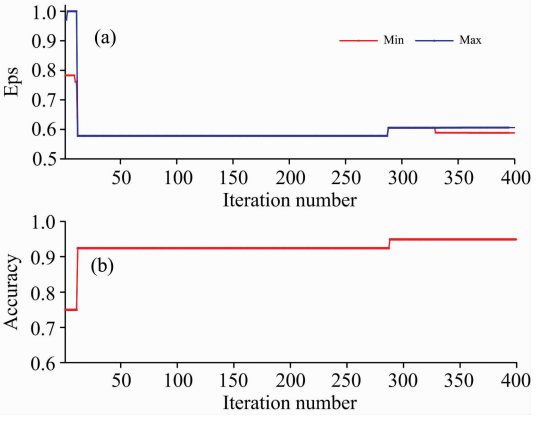

图 7 MVO-DBSCAN 优化结果

(a)：参数 Eps 曲线；(b)：聚类准确率曲线

Fig. 7 Optimization results of MVO-DBSCAN

(a)：The curve of parameter Eps;

(b)：The curve of clustering accuracy

图 7(a)为当前迭代次数最高硕别率所对应的 Eps 的最大值和最小值曲线，图 7(b)为识别率曲线。从图 7 中可知，改进 DBSCAN 对未经自动编码器降维的水样光谱的识别，经历 288 次迭代才能达到最高识别率，其最高识别率为

95%；而改进算法对降维后的水样光谱的识别，只需 24 次迭代就达到最大识别率，且最大识别率为 97.5%。综上可知，将 AE 用于降维水样光谱数据，可加快改进算法的寻优，提高其对水样光谱的识别率。

3.3　监督学习与 DBSCAN 对比

监督学习是在识别时让模型在监督条件下将训练样本映射到其所对应的标签上，在已知标签的数据集上，监督学习所表现出的性能优于无监督学习，但是新样本的标签在已有的数据中没有出现过，监督学习算法依然会将其映射到训练集已有的标签上，便发生误判。把监督学习算法用于 LIF 光谱数据的识别，将结果与 DBSCAN 的聚类结果对比，文中监督学习算法选用 K 最近邻算法。K 最近邻(K-nearest neighbors，KNN) 算法是 Cover 等提出的一种监督学习算法，主要应用场景有字符识别、文本分类、图像识别等领域。

在模型训练中，从水样一、水样二和水样三中各随机选取 20 组光谱数据作为训练集，其余的光谱数据(包含水样四的 30 组水样光谱)用于训练好的模型测试。在训练集上，DBSCAN 算法聚类准确率最高时，参数 MinPts 的取值为 3，Eps 的取值为 0.0382。将训练好的模型参数用于测试集测试，预测结果如图 7 所示。

图 8 中，黑色圆圈为测试集水样的实际标签，红色方块为 KNN 的识别结果，蓝色五角星为 DBSCAN 的识别结果。由图 8 可知，KNN 将水样一、二和三全部正确识别，把训练集中没有的水样四全部识别成水样一，DBSCAN 算法把水样一的两个样本视为噪声，其余全部正确的识别。由对比可得，监督学习算法在识别训练集已有的光谱时，具有较高的识别率，但是无法识别训练集没有的光谱，而无监督学习算法 DBSCAN 能识别出训练集中没有的水样光谱。在煤矿开采中，由于地下环境较为复杂，难以预先建立所有水样的光谱数据库，将 DBSCAN 用于矿井突水的识别，可有效降低对未知水源的误识别。

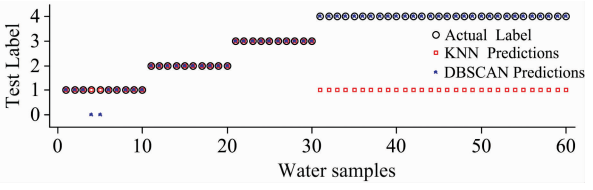

图 8　KNN 和 DBSCAN 的识别结果

Fig. 8　Recognition results of KNN and DBSCAN

4　结　论

(1) 快速准确地判别矿井突水来源，对煤矿的安全生产具有重要意义，把激光诱导荧光技术用于矿井突出检测，再结合无监督学习聚类算法 DBSCAN，快速识别突水来源的同时有效的降低了对未知水源的误识别。此外，将 MVO 用于 DBSCAN 算法的参数寻优，获取 DBSCAN 对水样光谱最高识别率所对应的参数取值范围，省去了光谱识别中繁琐的人工参数寻优过程；

(2) 把 AE 用于降维水样 LIF 光谱数据，所设计的多层网络降维模型将原始光谱数据从 2 048 维降到 2 维，大幅度减少了原始光谱数据中的冗余信息，加快了 MVO 对 DBSCAN 参数寻优的速度，同时也提高 DBSCAN 对水样光谱的识别率。在 AE 算法中引入一个 Dropout 层，使降维模型具有一定的稀疏性的同时，加快了 AE 训练收敛速度；一般原始光谱数据中都含有大量冗余信息，本工作成功地将 AE 用于 LIF 光谱数据降维，在复杂光谱数据处理中具有较重要的意义。

References

[1] WANG Wen-tan, ZHANG Meng-xue, ZHAO Shu-fang(王文坦，张梦雪，赵述芳). Journal of Chemical Industry and Engineering(化工学报), 2013, 64(3): 771.

[2] ZHANG Hui-shu, YUAN Xi-gang, Ali K M(张会书，袁希钢，Ali K M). Journal of Chemical Industry and Engineering(化工学报), 2014(9): 3331.

[3] LI Jia-ming, CHU Ying-bo, ZHAO Nan(李嘉铭，褚应波，赵　楠). Analytical Chemistry(分析化学), 2016, 44(7): 1042.

[4] WANG Ya, ZHOU Meng-ran, YAN Peng-cheng(王　亚，周孟然，闫鹏程). Journal of China Coal Society(煤炭学报), 2017, 42(9): 2427.

[5] ZHOU Meng-ran, LAI Wen-hao, WANG Ya(周孟然，来文豪，王　亚). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2018, 38(7): 2262.

[6] Ester M, Kriegel H P, Xu X. International Conference on Knowledge Discovery and Data Mining. The Association for the Advancement of Artificial Intelligence Press, 1996. 226.

[7] Mirjalili S, Mirjalili S M, Hatamlou A. Neural Computing and Applications, 2015, 27(2): 495.

[8] Faris H, Aljarah I, Mirjalili S. Applied Intelligence, 2016, 45(2): 1.

[9] Rotter F, Scholz J, Grimsel J. Applied Physics B, 2010, 101(4): 909.

[10] WU Yu-hong(伍育红). Computer Science(计算机科学), 2015, 42(z6): 491.

[11] QU Jian-ling, DU Chen-fei, DI Ya-zhou(曲建岭，杜辰飞，邸亚洲). Computer and Modernization, 2014(8): 128.

[12] Goodfellow I J, Pouget-Abadie J, Mirza M. Advances in Neural Information Processing Systems, 2014, 3: 2672.

[13] Hinton G E, Srivastava N, Krizhevsky A. Computer Science, 2012, 3(4): 212.

Application of Unsupervised Learning AE and MVO-DBSCAN Combined with LIF in Mine Water Inrush Recognition

LAI Wen-hao[1], ZHOU Meng-ran[1*], LI Da-tong[1], WANG Ya[2], HU Feng[1], ZHAO Shun[3], GU Yu-lin[1]

1. School of Electrical and Information Engineering, Anhui University of Science and Technology, Huainan 232000, China
2. School of Computer and Information Engineering, Fuyang Normal University, Fuyang 236000, China
3. School of Electronic and Electrical Engineering, University of Leeds, Woodhouse Lane, Leeds L859JT, UK

Abstract Quick and accurate identification of water inrush types and sources of water inrush is of great significance for safe mining of coal mines. Laser-induced fluorescence (LIF) technology is rapid and sensitive in detection, which applies LIF to the detection of water inrush in coal mines and uses pattern recognition algorithm to quickly identify the source of water inrush. The current algorithms for identifying water samples are too dependent on pre-established water sample spectral databases When the water source is not in the library, it is easy to cause misidentification. The unsupervised learning algorithm DBSCAN does not require the label and category information of the sample set when clustering, which can reduce the misidentification of unknown categories. Therefore, the DBSCAN algorithm is used to identify the laser-induced fluorescence spectra in water inrush, and MVO is used for the parameter optimization of DBSCAN, which can eliminate the cumbersome manual parameter optimization process. In the experiment, four water samples were taken from the water intake point of Xieqiao Coal Mine, and 30 sets of spectral data were collected for each water sample. The fluorescence spectra of the water samples were collected using a USB2000+ spectrometer with a pixel of 2 048. First, the unsupervised learning algorithm automatic encoder (AE) reduces the dimension of the original spectral data to reduce the influence of redundant information in the spectral data on the clustering. The structure of the AE designed in this paper is a multi-layer network model between shallow and deep layers, which can reduce the original spectral data to 2 dimensions. In order to make the dimensionality reduction model sparse, the author adds a Dropout layer to the traditional AE algorithm. It can be seen from the experiment that the dimensionality reduction model after adding the Dropout layer has a faster convergence speed. Then, using the multivariate optimization (MVO) algorithm to optimize the DBSCAN parameters. In the parameter optimization process, the spectral recognition rate of the water sample after DBSCAN is up to 97.5%, and the corresponding range of the parameter Eps is [0.023 66 0.040 65]. The normalized unscaled spectral data is used for DBSCAN cluster identification to verify the effectiveness of AE on the dimensionality reduction of water sample spectral data. The recognition rate of the original water sample spectrum by DBSCAN is up to 95%, which is 2.5% lower than that of the post-dimensional water sample. The results show that using AE dimensionality reduction data can improve the recognition rate of DBSCAN for different spectra. Finally, the supervised learning algorithm K nearest neighbor (KNN) is used to identify the water sample spectrum after dimension reduction, and the recognition result and the unsupervised learning algorithm DBSCAN are compared. The training set uses three water samples, and the test set uses four water samples. For the test set data, the supervised learning algorithm can only accurately identify the water sample categories contained in the training set, but all the categories that are not in the training set are identified incorrectly. On the contrary, DBSCAN can accurately identify the water sample spectrum not in the training set. The nonlinear dimensionality reduction algorithm AE can achieve dimensionality reduction on high-dimensional water spectral data. The use of MVO-DBSCAN for LIF spectral identification of coal mine water inrush can effectively reduce the misidentification caused by the incompleteness of the mine water source spectrum database.

Keywords Mine water inrush; Laser induced fluorescence; Spectral recognition; Density-based special clustering of applications with noise(DBSCAN); Multi-verse optimizer; Auto encoder; Dropout

(Received Dec. 19, 2018; accepted Apr. 30, 2019)

* Corresponding author