基于多层正则极限学习机的煤矿突水光谱判别方法

王亚[1,2], 周孟然[1], 陈瑞云[3], 闫鹏程[1], 胡锋[1], 来文豪[1]

[1]安徽理工大学电气与信息工程学院, 安徽 淮南 232001;

[2]阜阳师范学院计算机与信息工程学院, 安徽 阜阳 236037;

[3]淮南矿业集团谢桥煤矿, 安徽 阜阳 236221

摘要 为了快速而准确地判别煤矿突水水源类型,提出了一种构建多层正则极限学习机(M-RELM)模型的方法,该模型融合了非线性特征提取和分类学习。以激光诱导荧光(LIF)技术获取水样荧光光谱,作为模型的输入;以改进的自动编码器(AE)提取荧光光谱特征,形成模型隐含层的特征空间。为了减少光谱中噪声和异常对分类结果的影响,对极限学习机(ELM)算法进行了正则化优化,根据是否利用未知样本构造训练集,进行 L2 范数正则极限学习机(L2-RELM)或基于图的流形正则极限学习机(GM-RELM)优化,实现监督或半监督的分类学习。通过不同功能的隐含层之间进行传播,构建了多层正则化模型,完成了预训练和训练两个过程的融合。以淮南区域煤矿突水水样为实验对象,与支持向量机(SVM)和单隐含层极限学习机进行性能比较。在含有混合水的样集上,该模型的平均测试准确率可达到 94%以上,训练时间为 0.2 s 左右。在含有未知样本的所有水样集上,相比于 L2-RELM 模型,采用基于图的流形正则优化的 GM-RELM 模型的测试准确率可提升 2%左右。实验结果表明,M-RELM 模型更能适应煤矿突水水源的判别要求。

关键词 光谱学; 水源判别; 荧光光谱; 非线性特征提取; 多层正则化; 极限学习机

中图分类号 O433.4 **文献标识码** A **doi**: 10.3788/AOS201838.0730002

Identification Method of Coal Mine Water Inrush Spectrum Based on Multilayer Regularization Extreme Learning Machine

Wang Ya[1,2], Zhou Mengran[1], Chen Ruiyun[3], Yan Pengcheng[1], Hu Feng[1], Lai Wenhao[1]

[1] *College of Electrical and Information Engineering, Anhui University of Science and Technology, Huainan, Anhui* 232001, *China*;

[2] *School of Computer and Information, Fuyang Normal University, Fuyang, Anhui* 236037, *China*;

[3] *Xieqiao Coal Mine, Huainan Mining Group, Fuyang, Anhui* 236221, *China*

Abstract In order to quickly and accurately identify the source types of coal mine water inrush, we propose a method of constructing a multilayer regularization extreme learning machine (M-RELM) model, which combines the functions of nonlinear feature extraction and classification learning. The fluorescence spectra of water samples are obtained by laser induced fluorescence (LIF) technique as the input of model. The features of fluorescence spectra are extracted by the improved auto encoder (AE) to form the feature space of the model hidden layer. In order to reduce the effect of noise and anomaly of spectra on classification results, the algorithm of extreme learning machine(ELM) is optimized regularly. According to whether the unknown samples are used to construct the training set, the model is optimized regularly by the L2 norm regularization (L2-RELM) or the graph manifold regularization (GM-RELM), which realizes the supervised or semi-supervised classification learning. By

收稿日期: 2018-01-10; **收到修改稿日期**: 2018-03-10

基金项目: "十二五"国家科技支撑计划(2013BAK06B01)、国家自然科学基金(51174258)、国家安全生产重大事故防治关键技术科技项目(anhui-0001-2016AQ)、安徽省自然科学基金项目(1808085MF202,1808085QE157)、安徽省自然科学研究项目(KJ2018ZD036)

作者简介: 王亚(1980—),女,博士研究生,副教授,主要从事光谱技术检测、模式识别方面的研究。E-mail: fync_wy80@163.com

导师简介: 周孟然(1965—),男,博士,教授,博士生导师,主要从事矿山机电系统监测、光电信息处理、煤矿安全监测监控方面的研究。E-mail: mrzhou8521@163.com

propagating between the hidden layers of different functions, M-RELM is constructed, and the integration of pre-training and training is completed. The water inrush samples in Huainan area coal mine as the experimental object, the performance compares with the support vector machine (SVM) and ELM with a single hidden layer. On the samples set containing mixed water, the average testing accuracy of the model can reach more than 94% and the training time is about 0.2 s. On all water samples containing the unknown samples, the testing accuracy of GM-RELM is increased by 2% than L2-RELM. The experimental results show that the M-RELM model is more suitable for the identification requirements of coal mine water inrush.

Key words spectroscopy; water source identification; fluorescence spectrum; nonlinear feature extraction; multilayer regularization; extreme learning machine

OCIS codes 300.6280; 100.4996; 120.6200

1 引　　言

在煤矿突水灾害防治过程中，需要根据不同水源类型而采取相应的防范措施，因此快速而准确地判别水源类型是防治工作的基础[1]。地下含水层中的化学物质会受环境影响而发生变化，但是在短时间内区域内含水层的化学物质则相对稳定。目前常规方法是采用水化学分析技术[2-3]，以几种代表离子的浓度为主要指标建立模型进行水源判别，虽然性能比较稳定，但效率较低、时间较长，已不能满足快速和准确性的需求。光学监测中的激光诱导荧光光谱(LIF)技术由于具有高灵敏性、快速准确的特点，近年来已在医学、环境、农业生产等领域得到了广泛应用[4-6]。文献[7]论证了 pH 值、浓度等因素对荧光强度的影响，该研究表明 LIF 可检测 pH 值或水中溶质浓度等水体参数。在煤矿水源类型研究中，文献[8]以霍州矿区深部含水层的荧光光谱综合分析了溶解性有机质(DOM)的含量和分布特征。文献[9]获取骆驼山煤矿含水层中有机质荧光光谱，以荧光强度检测和分辨不同来源的 DOM，进而利用 DOM 识别突水水源。文献[10]将 LIF 技术应用于煤矿突水水源判别中，不同含水层产生不同的荧光光谱，通过提取光谱特征进而识别不同水源。根据以上文献分析，结合 LIF 技术采集突水水样的荧光光谱，根据不同类型水源的光谱特性建立多元分类学习模型，可实现快速而准确地判别水源类型。

由于光谱携带了大量不确定、高维的信息，提取光谱特征可以提高分类性能。经典的线性方法有：主成分分析(PCA)、线性判别分析(LDA)等[11]，然而在线性方法中，存在着主成分个数难以确定、主成分对分类结果影响较大等问题，目前已逐渐向机器学习的非线性方法发展[12]。其中，自动编码器(AE)是一种通过编码方式重构输入数据的神经网络，采用网络训练方式可实现高效的特征提取。对于具有高维及非线性光谱的特征，多采用神经网络方法进行分类学习[13]。文献[14]用非线性自组织映射(SOM)神经网络分析水质的三维荧光光谱，研究荧光特征与水质之间的关系。文献[15]采用最小二乘支持向量机(LS-SVM)分析橙汁荧光光谱，获取橙汁中化学物质浓度，其中正则化参数和核参数通过优化算法得到。支持向量机(SVM)算法对处理小样本数据比较具有优越性，但建立模型需要多参数寻优调节[16]。文献[17]对不同浓度的混合物荧光特性采用优化后反向传播(BP)神经网络进行浓度质量检测。上述文献都是采用传统的非线性神经网络，训练时间较长、人工调节参数较多，并且需要结合多种优化算法选取最优参数。

Huang 等[18-19]提出了极限学习机(ELM)，其是一种新型前馈式神经网络(SLFN)，属于监督学习算法，其特点是：只需一次随机给定隐含层参数，确定激励函数和隐含层节点数，用训练数据集逼近复杂和非线性映射，具有较好的通用性，无需过多人工干预。文献[20]结合 PCA 建立了基于 ELM 算法的非线性分类模型，与 BP 和 SVM 进行比较，在降低模型训练时间和提高分类性能上取得较好效果。文献[20]采用了线性特征提取，在单纯水样集上验证了模型的性能，但对模型的抗干扰能力考虑较少。由于煤矿井下环境复杂，水样的采集难免会受到干扰出现异常，当突水发生时，多以混合水的形式存在。如何利用这些未知样本辅助训练模型，并通过模型预测未知样本结果，是急需解决的问题。为了适应实际应用需求，建立具有抗干扰能力的模型，采用正则化方式包括 L2 范数和基于图的流形框架进行优化，产生通用的正则化 ELM(RELM)模型[21-23]。文献[24]通过大量实验已经证明，RELM 在多元分类方面性能要优于 SVM 和 LS-SVM。L2 范数正则化(L2-RELM)可以增加模型的抗干扰性，但却无法利用大量的未知样本，因此，ELM 在 L2-RELM 基础上，利用图的流形正则化(GM-RELM)深入优化，得到 GM-RELM 模型。上述文献都是将

特征提取独立于分类模型之外，延长了分类判别时间，也不利于模型在应用中的推广，因此需要设计多层次融合模型。

针对以上问题，提出了多层融合学习的 RELM（M-RELM）算法，以 AE 进行非线性特征提取、以正则化优化 ELM 分类学习，建立多层监督或半监督的分类学习模型，将两种基本功能统一融入到模型中，实现分阶段、层次间传播学习能力。以淮南区域煤矿采集水样的荧光光谱为实验对象，与多种模型进行性能比较，验证该模型判别水源类型的能力。

2　正则化及特征提取理论

设样本集 $\{\boldsymbol{X},\boldsymbol{T}\}=\{(x_i,t_i)\mid x_i\in R^n,t_i\in R^m\}_{i=1}^N$ 中有 N 个样本，$\boldsymbol{X}$ 为输入矩阵，$\boldsymbol{T}$ 为目标矩阵，R 是实数集。在单隐含层的 ELM 结构中，$\boldsymbol{H}\in R^{N\times L}$ 为特征矩阵，是 L 个隐含层节点的输出矩阵，特征映射 $h_i(\boldsymbol{x})=g(a_i,b_i,\boldsymbol{x})$，其中 g 为隐含层激励函数，输入层与隐含层间的输入权值和偏置 (a_i,b_i) 为随机生成，隐含层与输出层间的输出权值 $\beta\in R^{L\times m}$ 通过计算得到。ELM 输出函数 $f_L(x)=h(\boldsymbol{x})\boldsymbol{\beta}=\sum_{i=1}^{L}\beta_i h_i(\boldsymbol{x})=\boldsymbol{H\beta}$，最小化训练误差为 $\min\limits_{\boldsymbol{\beta}\in R^{L\times m}}\|\boldsymbol{H\beta}-\boldsymbol{T}\|^2$，利用最小二乘法求解最优输出权值 $\boldsymbol{\beta}^*=\boldsymbol{H}^+\boldsymbol{T}$，其中，$\boldsymbol{H}^+$ 是 $\boldsymbol{H}$ 的 Moore-Penrose 广义逆[25]。

2.1　正则优化过程

2.1.1　L2-RELM

由于 ELM 算法中采用最小二乘法求解最优输出权值，抗干扰能力较差，容易受到噪声和异常点的影响，因此采用正则优化最小化训练误差和输出权值。以基于等式约束的 L2-RELM，引入正则参数 C 作为最小化训练误差的惩罚系数，生成 L2-RELM 算法。将约束条件转换为无约束的优化问题，可表示为：

$$\min_{\boldsymbol{\beta}\in R^{L\times m}}\frac{1}{2}\|\boldsymbol{\beta}\|^2+\frac{C}{2}\sum_{i=1}^{N}\|e_i\|^2\Rightarrow\min_{\boldsymbol{\beta}\in R^{L\times m}}\frac{1}{2}\|\boldsymbol{\beta}\|^2+\frac{C}{2}\sum\|\boldsymbol{H\beta}-\boldsymbol{T}\|^2,\ \text{s.t.}\ h(x_i)\boldsymbol{\beta}=t_i^{\mathrm{T}}-e_i^{\mathrm{T}},\tag{1}$$

当训练样本数 N 大于隐含节点数 L（即 $N>L$）时，最优输出权值 $\boldsymbol{\beta}^*$ 可以表示为：

$$\boldsymbol{\beta}^*=\left(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}+\frac{\boldsymbol{I}}{C}\right)^{-1}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{T};\tag{2}$$

否则，当 $N<L$ 时，$\boldsymbol{\beta}^*$ 可以表示为：

$$\boldsymbol{\beta}^*=\boldsymbol{H}^{\mathrm{T}}\left(\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}+\frac{\boldsymbol{I}}{C}\right)^{-1}\boldsymbol{T},\tag{3}$$

式中 $\boldsymbol{I}$ 是单位矩阵。

2.1.2　GM-RELM

在 L2-RELM 基础上，若要利用未知样本训练模型，需采用 GM-RELM 优化。首先构造一个无向近邻图 $G=<V,E>$，顶点 V 为 l 个已知样本和 u 个未知样本，边 E 为邻接节点的权值，表示样本间的相似程度[26]。在图 G 上定义函数 f，训练样本的标签 y_{label} 可以看成是函数 f 在顶点上的观测值。GM-RELM 需要同时满足：所有已知样本都应该接近于标签 y_{label}；函数 f 应该是光滑的[27]。根据图谱理论，确保函数 f 具有光滑性，并且降低函数的复杂性，可表示为：

$$f^{\mathrm{T}}\boldsymbol{L}f=\sum_{i,j=1}^{l+u}(f_i-f_j)^2w_{ij},\tag{4}$$

式中 f_i 和 f_j 分别为函数 f 在顶点 i 和 j 上的观测值，$\boldsymbol{L}$ 为 Laplacian 矩阵，$\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{W}$，其中 $\boldsymbol{D}$ 为度矩阵，$\boldsymbol{W}$ 为邻接权值矩阵，$\boldsymbol{D}$ 为 $\boldsymbol{W}$ 的对角阵 $D_{ii}=\sum_{j=1}^{l+u}W_{ij}$，顶点 i 和 $j[(i,j)\in E]$ 互为近邻节点，则邻接权值 W_{ij} 为 1，否则为 0。为了满足光滑假设，最小化(4)式得到：

$$\min\frac{\lambda}{2}\sum_{i,j=1}^{l+u}(f_i-f_j)^2w_{ij}\Rightarrow\min\frac{\lambda}{2}\mathrm{Tr}(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{F}),\tag{5}$$

式中 $\mathrm{Tr}(\cdot)$ 为矩阵的迹，流形正则参数 λ 为图光滑性的惩罚系数，$\boldsymbol{F}=\boldsymbol{H\beta}$。

对 ELM 进行 GM-RELM，结合(1)式得到：

$$\min_{\boldsymbol{\beta}\in R^{L\times m}}\frac{1}{2}\|\boldsymbol{\beta}\|^2+\frac{C}{2}\sum\|\boldsymbol{H\beta}-\boldsymbol{T}\|^2+\frac{\lambda}{2}\mathrm{Tr}(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{H}\boldsymbol{\beta}).\tag{6}$$

从(6)式可解得最优输出权值 $\boldsymbol{\beta}^*$，当 $N>L$ 时 $\boldsymbol{\beta}^*$ 可表示为：

$$\boldsymbol{\beta}^*=\left(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}+\frac{\boldsymbol{I}}{C}+\frac{\lambda}{C}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{H}\right)^{-1}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{T},\tag{7}$$

否则，当 $N<L$ 时，$\boldsymbol{\beta}^*$ 可表示为：

$$\boldsymbol{\beta}^*=\boldsymbol{H}^{\mathrm{T}}\left(\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}+\frac{\boldsymbol{I}}{C}+\frac{\lambda}{C}\boldsymbol{L}\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}\right)^{-1}\boldsymbol{T}.\tag{8}$$

2.2　AE-RELM 非线性特征提取

标准的 AE 是通过 BP 算法多次迭代以寻求训练误差最小，若采用 ELM 算法则生成 AE-ELM[28-29]。由于 RELM 具有快速学习和抗干扰能力，以 RELM 算法训练 AE 从而产生 AE-RELM。为了尽可能全面地重构输入数据信息，AE-RELM 需要经历两个阶段，其工作过程如图 1 所示。

AE-RELM 的工作过程可分为以下两个阶段：

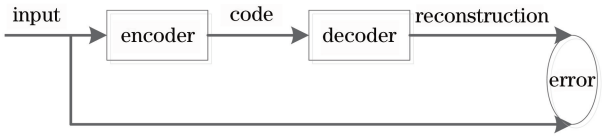

图 1　AE-RELM 工作过程

Fig. 1　Work process of AE-RELM

1)编码阶段：输入数据 $\boldsymbol{x}\in R^n$，随机正交生成隐含层节点参数(a_i,b_i)，$\boldsymbol{x}$ 经过编码器随机映射到特征空间生成编码，特征映射函数 $h_i(\boldsymbol{x})=g(a_i,b_i,\boldsymbol{x})$，隐含层输出矩阵为 $\boldsymbol{H}$，其中 $a^{\mathrm{T}}a=\boldsymbol{I}$，$b^{\mathrm{T}}b=1$。

2)解码阶段：由解码器对编码进行重构，输出 $\hat{\boldsymbol{x}}=\boldsymbol{H}\boldsymbol{\beta}\in R^n$，使得 $\hat{\boldsymbol{x}}\approx\boldsymbol{x}$，误差为重构数据 $\hat{\boldsymbol{x}}$ 与输入数据 $\boldsymbol{x}$ 之间的差异。其中，$\boldsymbol{\beta}$ 称为重构矩阵，包含了输入数据分布的重要信息。用 $\boldsymbol{X}$ 代替 $\boldsymbol{T}$ 由(2)式计算得出 $\boldsymbol{\beta}$。当 $\boldsymbol{x}$ 的维度 n 大于隐含节点个数 L(即 $n>L$)时，AE-RELM 则实现对输入数据的非线性特征提取。

3　M-RELM 模型

RELM 可通过隐含层实现多种学习能力，但单隐含层结构不能实现学习的传播。为了获得具有传播学习能力的融合模型，需要进行多层设计。M-RELM 是基于深度学习(DL)框架，并对 RELM 的扩展[30-31]。区别于传统 DL 的贪婪层次学习方式，它是以分阶段的形式进行层次学习。训练数据从输入空间到输出空间，经历了预训练和训练两个阶段。经过多次调用 AE-RELM 形成新特征空间，完成预训练阶段的特征提取；训练阶段在新特征空间下采用 RELM 算法进行分类学习。其模型框架结构如图 2 所示。

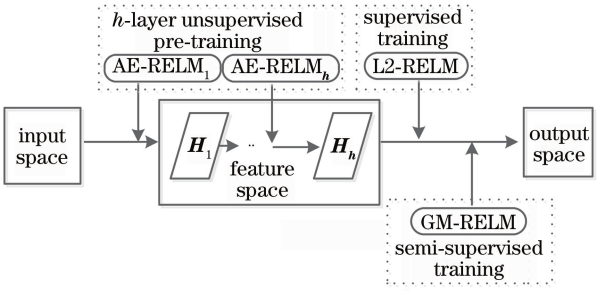

图 2　M-RELM 模型的框架结构

Fig. 2　Framework of M-RELM model

根据训练集中是否包含未知样本，分别采用 L2-RELM 和 GM-RELM 对 ELM 模型进行优化，从而形成 L2-RELM 和 GM-RELM 模型，本研究中通称为 M-RELM 模型。

L2-RELM 模型的训练过程可分为以下几个阶段：

1)输入，训练集为$\{\boldsymbol{X},\boldsymbol{T}\}=\{(x_i,t_i)\,|\,x_i\in R^n,t_i\in R^m\}_{i=1}^N$，由 N 个已知样本组成。

2)预训练阶段，由(2)式计算 AE-RELM$_i$ 的输出权值 $\boldsymbol{\beta}^i$；用$(\boldsymbol{\beta}^i)^{\mathrm{T}}$ 作为 M-RELM 模型中第 i 隐含层的输入权值，得到特征空间 $\boldsymbol{H}_i=g(\boldsymbol{H}_{i-1}\boldsymbol{\beta}^i)(i=1,\cdots,h)$，$h$ 为隐含层总层数，$\boldsymbol{H}_0$ 为输入数据 $\boldsymbol{x}$；递归调用第 h 次 AE-RELM$_h$，计算得出最后隐含层输出矩阵 $\boldsymbol{H}_h$。

3)训练阶段，采用监督学习方式的 L2-RELM 分类学习，由(2)或(3)式计算 $\boldsymbol{\beta}^h$。

4)输出，计算分类输出结果 $\boldsymbol{Y}=\boldsymbol{H}_h\boldsymbol{\beta}^h$。

GM-RELM 模型的训练过程可分为以下几个阶段：

1)输入，由 l 个已知样本的数据集$\{(x_i,t_i)\}_{i=1}^l$和 u 个未知样本的数据集$\{x_j\}_{j=l+1}^{l+u}$共同组成训练集，样本数量为 $l+u$。

2)预训练阶段，处理过程同上，得到 $\boldsymbol{H}_h$。

3)训练阶段，采用半监督学习方式的 GM-RELM 分类学习：首先，用 $l+u$ 个节点，以 k 近邻(k-NN)方法构造邻接图，采用二进制权值(0,1)为邻接边赋值，本研究中设邻接点数 v_{n} 为 10；然后，计算拉普拉斯矩阵 $\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{W}$，通常用$\boldsymbol{D}^{\frac{1}{2}}\boldsymbol{L}\boldsymbol{D}^{-\frac{1}{2}}$规范化 $\boldsymbol{L}$；最后，构造验证集，参数 C 和 λ 在$[10^{-5},10^5]$范围内寻求最优，在验证集上准确率要高于设定的阈值，由(7)式或(8)式计算 $\boldsymbol{\beta}^h$。

4)输出，计算分类输出结果 $\boldsymbol{Y}=\boldsymbol{H}_h\boldsymbol{\beta}^h$。

4　实验结果与分析

4.1　水样采集

2017 年 6 月 15 日，从淮南矿业集团谢桥煤矿采集了 4 种类型单纯水样，分别为：老空水、砂岩水、灰岩水和奥灰水，分别标识类别为 1～4。采集地点分别位于 2212(1)上顺槽处、西翼 B 组 4 煤底板皮带石门联巷、东风井井底车场和－240 m 水平奥灰供水孔。每种类型水样采集的组数分别为 20，25，15，50，共 110 组水样。煤矿突水水源类型存在多样性，其中老空水是煤矿井下采空区被地下或地表水填充而形成，一旦发生突水危害较大。因此利用 4 种类型的原始水样，以老空水为基础，与其他三种类型水样以 1∶1比例分别配制了三种类型混合水样：混合水 1(老空水∶砂岩水)25 组、混合水 2(老空水∶灰岩水)20 组、混合水 3(老空水∶奥灰水)25 组，共 70 组混合水样，分别标识类别为 5～7。由此，所有

水样集(包括 4 种类型单纯水样集和 3 种类型混合水样集)共有 180 组水样。以上 180 组水样均为已知水样,同时又采集了 160 组未知水样。

4.2　实验装置

水样的荧光光谱是在检测实验平台上采集而得,实验装置及平台搭建如图 3 所示。

激光器(LSR 405 nm,北京华源拓达激光技术有限公司)为 100 mW 蓝紫光半导体激光器,入射波长为 405 nm,激光功率在 100～130 mW 范围可调,实验中经反复测试设定为 120 mW。光谱仪(USB2000＋,美国 Ocean Optics)接收全波段为 340～1020 nm,分辨率为 0.5 nm,积分时间设置为 1 s/1000 nm。激光器和光谱仪通过光纤与荧光探头(FPB-405-V3)相连,其中光纤接口为 SMA905 接头。在采集水样的荧光光谱时,需要将探头浸入水体中,因此实验中使用了可浸入式激光激发荧光探头。荧光经过一组滤光片滤除无用波段,最后由光纤传输至光谱仪。利用光谱软件 Spectra Suite 采集并记录波段范围内荧光发射的强度值。在实验平台上,采集到的不同水样原始荧光光谱如图 4 所示。

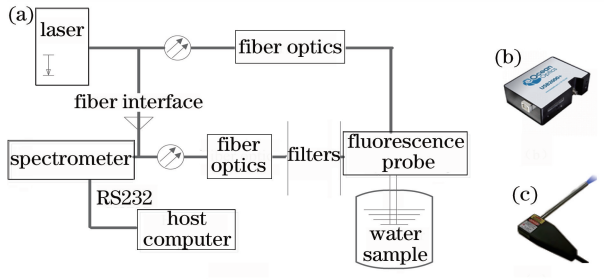

图 3　实验平台与装置。(a)平台;(b)光谱仪;(c)荧光探头

Fig. 3　Experiment platform and devices. (a) Platform; (b) spectrometer; (c) fluorescence probe

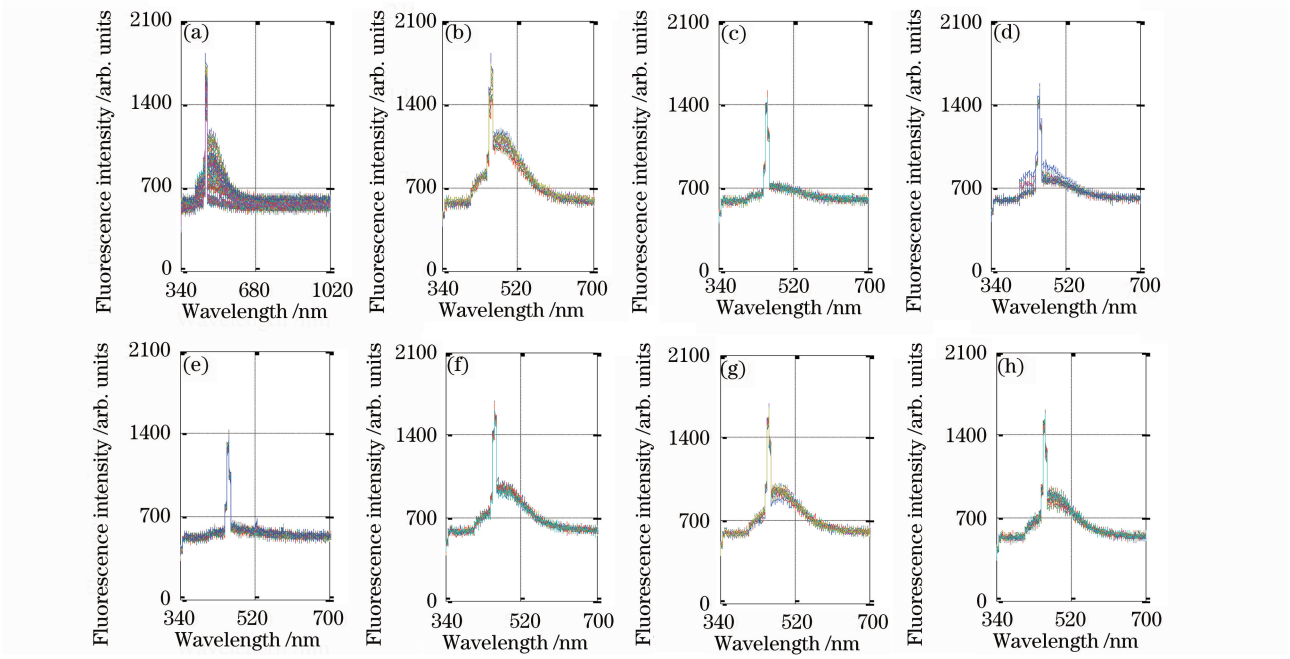

图 4　不同水样的原始荧光光谱图。(a)所有水样;(b)老空水;(c)砂岩水;(d)灰岩水;(e)奥灰水;(f)混合水 1;(g)混合水 2;(h)混合水 3

Fig. 4　Original fluorescence spectra of different water samples. (a) All samples; (b) old-kiln water; (c) sandstone water; (d) limestone water; (e) ordovician water; (f) mixed water 1; (g) mixed water 2; (h) mixed water 3

图 4(a)为所有水样在全波段上的谱图,从光谱曲线的变化趋势可以看出,波峰主要集中在 400～500 nm 波段,当波长大于 700 nm 后,光谱曲线走势平稳,基本成为平线。为了清晰显示光谱曲线的变化,并且尽可能保留谱图中的有用信息,选取 340～700 nm 波段间的荧光强度值绘制图 4(b)～(h)。从图 4(b)～(h)可以看出,由于相同含水层所含化学物质接近,水样的光谱曲线保持一致;不同含水层所含化学物质存在差异,水样的荧光光谱曲线也各不相同。虽然各种类型水样的光谱曲线存在差异,但是曲线差异性较小很难分辨。下面将结合实验结果,分析讨论如何应用模式识别的特征提取和分类学习的方法提高判别性能。

4.3　实验结果与讨论

为了验证模型在判别单纯水样和混合水样时是否具有通用性,将实验过程分为三个阶段进行,依次对 4 种单纯水样集、三种混合水样集和所有水样集进行测试。实验硬件环境为 Intel CPU(Core i7-

3687U 2.10 GHz)、8 GB 内存,算法运行在 Matlab 2011b 环境中。每种算法经过 100 次实验,统计模型训练和测试的分类准确率均值、方差和时间均值等。由于 M-RELM 模型能同时完成特征提取和分类学习,所以在统计模型的训练时间时,SVM、ELM、RELM 算法除了训练分类模型的时间,还应包括 PCA 特征提取的时间。

SVM 算法参考文献[32]以径向基函数(RBF)作为核函数,惩罚参数 C 和核参数 g 在$[2^{-5},2^{5}]$范围内采用三折交叉验证方法寻优。为了简化 M-RELM 模型结构,设置两层隐含层分别完成特征提取和分类学习功能,因此文中隐含层数设为 2,则网络结构的总层数为 4 层。ELM 算法中隐含节点激励函数选择 Sigmoid 函数,L2-RELM 算法中正则参数 C 参考文献[33]选取为10^{4};M-RELM 算法通过验证集在$[10^{-5},10^{5}]$范围内寻找最优参数 C、λ,且满足测试准确率高于 80%以上。由于本研究是建立与 SVM 算法性能相近的快速判别模型,因此实验中隐含层节点个数的选取参考 SVM 算法的支持向量总数和测试准确率而定,不需要在模型中进行优化选择。

实验 1:4 种单纯水样集

先将 4 种单纯水样集的原始荧光强度值归一化到$[-1,1]$范围,以 7∶3的比例及交叉验证的方式提取训练集和测试集。由于样本集中有 4 种类型水样,则输出类别数为 4,设输出层节点数为 4。各模型的性能分析如表 1 所示。

表 1　模型在单纯水样集上的性能比较

Table 1　Performance comparison of models on sample set of simple water

Algorithm	Training time /s		Accuracy /%		Deviation /%		Hidden nodes / nSV
	Feature extraction	Classification	Training	Testing	Training	Testing	
SVM	0.000	11.458	100	96.88	0.00	0.00	69
PCA+SVM		0.782	100	96.88	0.00	0.00	40
PCA+ELM	0.953	0.003	100	98.00	0.00	2.41	40
PCA+RELM		0.002	100	99.94	0.00	0.44	40
M-RELM	0.155		100	97.72	0.13	0.87	200-40

SVM 算法在进行 PCA 特征提取后,生成的支持向量数(nSV)从 69 个降到 40 个。ELM 算法参考 nSV 的取值,将隐含层节点设定为 40 个;RELM 算法中也取相同值;M-RELM 算法的第 2 隐含层节点也设为 40 个,第 1 隐含层需要完成特征提取的功能,将节点数设定为 200 个。

从表 1 可以看出,各种算法建立的模型都能达到良好的性能,在训练性能趋于相同的情况下,平均测试准确率都能达到 96%以上。在训练和测试准确率的方差方面,SVM 算法取得了最稳的性能,但是测试准确率达到最低、模型训练时间最长。SVM 算法的模型训练时间过长,主要是由于多个参数需要寻优选择,花费了绝大部分时间。分析测试结果表明,错误分类主要集中在第 2,3 类,即砂岩水和灰岩水。从水文地质分析,由于两种含水层位于石炭纪(较晚),水成分很相似,所以光谱曲线也非常接近。第 4 类(奥灰水)位于奥陶纪(最为久远),第 1 类老空水与地质年代无关,两者与第 2、3 类都能很好地区别。

实验 2:三种混合水样集

在三种混合水样集上,采用与实验 1 相同的方法生成训练和测试集,输出类别数为 3,设输出层节点数为 3。各模型的性能分析如表 2 所示。

表 2　模型在混合水样集上的性能比较

Table 2　Performance comparison of models on mixed water sample set

Algorithm	Training time /s		Accuracy /%		Deviation /%		Hidden nodes / nSV
	Feature extraction	Classification	Training	Testing	Training	Testing	
SVM	0.000	2.992	100.00	90.00	0.00	0.00	50
PCA+SVM		0.348	96.00	90.00	0.00	0.00	30
PCA+ELM	0.906	0.001	99.64	88.55	0.87	7.22	30
PCA+RELM		0.001	99.56	91.40	1.01	6.44	30
M-RELM	0.145		100.00	92.00	0.97	4.22	200-30

本实验中有 70 组水样,数量相比于实验 1 中 110 组水样较少,SVM 算法的 nSV 从 69 个降到 50 个,模型训练时间约缩短四分之一。经过 PCA 特征提取后,SVM 算法的 nSV 从 50 个降至 30 个。为了比较相似的网络结构,ELM 和 RELM 算法中隐含节点个数取 30,M-RELM 中第 2 隐含节点个数

也取 30，第 1 隐含节点个数仍取 200。

从表 2 可以看出，在模型的训练分类准确率及方差方面，ELM 算法都要优于 RELM 算法，但平均测试准确率却达到最低(88.55%)，出现了过拟合现象。RELM 算法的平均测试准确率(91.40%)超过了 SVM 算法的测试性能，进一步验证了 RELM 算法可以减少数据集中噪声的影响。M-RELM 算法通过 AE-RELM 特征提取，在避免噪声干扰的同时尽可能多地保留了原数据的特征信息，所以最终平均测试准确率较 RELM 算法有所提升，可达到 92%，并且模型的训练速度提升了约 6 倍。在混合水样集上，相对于其他几种算法，M-RELM 的性能表现最优。分析测试结果，错误分类集中在第 5 和 6 类，而这两类正是由砂岩水、灰岩水与老空水分别混合而成，由于砂岩水和灰岩水的相似性，使得相应的混合水也存在相似性，致使相应光谱识别也出现了误判。

实验 3：所有水样集

输出类别数为 7，设输出层节点数为 7。各模型的性能分析如表 3 所示。

表 3　模型在所有水样集上的性能比较

Table 3　Performance comparison of models on all water sample set

Algorithm	Training time /s		Accuracy /%		Deviation /%		Hidden nodes / nSV
	Feature extraction	Classification	Training	Testing	Training	Testing	
SVM	0.00	34.84	100.00	96.88	0.00	0.00	117
PCA+SVM		1.21	99.21	94.23	0.00	0.00	60
PCA+ELM	0.56	0.01	98.95	92.37	0.49	2.31	60
PCA+RELM		0.01	99.09	93.81	0.46	1.67	60
L2-RELM	0.15		100.00	94.46	0.27	1.92	200-60
GM-RELM	4.78		98.95	96.75	0.93	2.28	

随着样本数量的增多，SVM 算法的 nSV 达到了 117 个，经过特征提取后 nSV 降为 60 个。从模型的训练时间进行比较，由于 SVM 算法需要优化调节多个参数，所以训练模型耗时最长。ELM 和 RELM 的隐含层节点个数都设为 60，M-RELM 两个隐含层节点个数分别设为 200 和 60。由于增加了未知水样，因此在 M-RELM 模型训练时，需要根据训练集中是否含有未知水样而分别选择 L2-RELM 和 GM-RELM 算法。

为了验证 GM-RELM 算法的性能，以文献[34]的方法将 180 组已知样本集分成 4 块子集，分别为：带类别标签的训练集 L(50 组)、未带标签的训练集 U(35 组)、验证集 V(56 组)和测试集 T(39 组)。该处的未带标签的训练集 U 是对已知样本去除类别而形成的，当模型建立后，仍可以对 U 数据集的预测结果进行测试，主要目的是验证模型对 U 中数据预测的有效性。以 L+U 为训练集生成的模型，在验证集 V 上测试选择参数 C、λ 最优值分别为 10^{1}，10^{-2}。在数据集 L、V、T、U 上的准确率分别为 97.24%，90.85%，92.35% ，94.80%。实验结果表明，该模型在未知样本上可以获得较好的预测效果。

由于样本量较少，模型的训练准确率并不高。于是在原有的已知水样训练集上，再加入 160 组未知水样，采用 GM-RELM 算法训练模型，模型训练和测试结果如表 3 所示。与 L2-RELM 算法相比，测试准确率从 94.46%上升到 96.75%，模型的测试性能得到提升，从而进一步说明，未知样本可以辅助训练模型。

从表 3 可以看出，单隐含层 RELM 和多隐含层 L2-RELM 模型的平均训练分类准确率都比 ELM 模型高，而测试准确率的方差较低，说明 L2-RELM 可以使得模型分类结果更加稳定。但是，GM-RELM 模型的训练和测试准确率的方差有所增大，说明未知样本的加入又会影响模型的稳定性。M-RELM 通用模型与 SVM 相比，分类准确率基本接近，但模型的训练时间却大大缩短。以采用 GM-RELM 算法建立模型为例，由于需要对参数寻优、构造近邻图并规范化拉普拉斯矩阵，使得训练总时间增加到约为 4.78 s，而训练 SVM 模型时间高达 34.84 s。在所有水样集上，对 ELM、RELM、L2-RELM 和 GM-RELM 算法训练的模型性能进行直观比较，如图 5 所示。

从图 5 可以看出，图 5(b)相对于图 5(a)数据更加集中，说明 RELM 模型性能较 ELM 更加稳定。图 5(c)相对于图 5(b)数据分布高度有所提升，说明多隐含层 L2-RELM 模型训练准确率比单隐含层 RELM 模型略高，并且训练时间缩短为近四分之一。虽然 GM-RELM 模型提升了测试的准确率，但

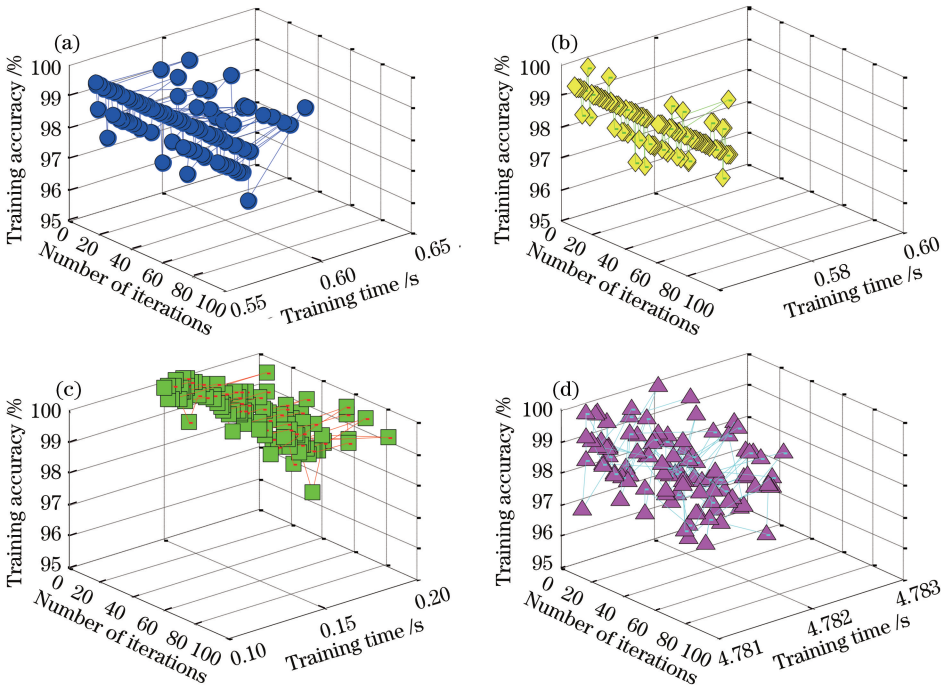

图 5　训练模型性能比较。(a) ELM；(b) RELM；(c) L2-RELM；(d) GM-RELM

Fig. 5　Performance comparison of training models. (a) ELM; (b) RELM; (c) L2-RELM; (d) GM-RELM

与图 5(c)比较，可以看出图 5(d)数据分布有些分散，说明利用未知样本训练模型影响到模型稳定性，模型训练时间也随之增加。

综合上述三个实验的结果，比较各种模型在不同水样集上的测试性能，如图 6 所示。

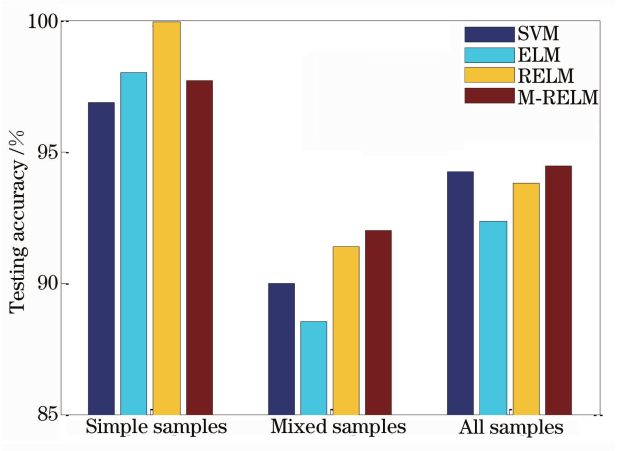

图 6　几种模型在不同水样集上测试性能比较

Fig. 6　Testing performance comparison of several models on different water sample sets

从图 6 可以看出，在单纯水样集上，M-RELM 模型并没有显示出其优势，而是 RELM 模型优势突显，说明对于单纯水样集可以用线性方式代替非线性方式进行特征提取。同时也验证了在 ELM 算法基础上进行正则化，可以有效地减少噪声对分类的干扰，避免产生过拟合问题。而在混合水样中，M-RELM 模型优势突显，说明与线性 PCA 相比，采用非线性 AE-RELM 进行特征提取能最大程度地保留混合水样光谱的特性。对所有水样进行集中分析时，M-RELM 模型的综合性能仍能保持最优。因此，可以得出 M-RELM 模型对含有混合水的突水水源判别更具有优势。

5　结　　论

结合 AE-RELM 非线性特征提取与 RELM 分类学习建立了多层次融合学习模型，对淮南区域煤矿采集水样进行荧光光谱分析，探讨了模型在不同水样集上的适用性。

通过 L2 范数正则优化 ELM 算法，得到的 RELM 算法更能减少噪声干扰，避免过拟合问题，因此对含有混合水的水样集，建立基于 RELM 算法的模型，其性能会更加稳定。

通过 RELM 算法预先训练 AE 生成 AE-RELM，进行非线性光谱特征提取。与线性 PCA 相比，提取的特征具有密集、冗余性，对提高模型的通用性能非常有效。

利用 RELM 隐含层具有独立学习能力，设计了用于特征提取和分类学习的两个隐含层，构建了融合两种功能的多层正则模型 M-RELM，实现了层次间学习能力的传播。与 L2-RELM 相比，GM-RELM 的优势在于可以更好地利用未知水样，两者

通称为 M-RELM 模型。与单隐含层 RELM 相比，M-RELM 模型结构具有适用性强、权值共享等优点，使得全局优化训练参数大大减少。所提出的 M-RELM 模型提升了对含混合水水样的判别能力，更适用于煤矿突水的实际应用。

参 考 文 献

[1] Wu Q, Cui F P, Zhao S Q, *et al*. Type classification and main characteristics of mine water disasters[J]. Journal of China Coal Society, 2013, 38(4): 561-565.
武强, 崔芳鹏, 赵苏启, 等. 矿井水害类型划分及主要特征分析[J]. 煤炭学报, 2013, 38(4): 561-565.

[2] Liu J M, Wang J R, Liu Y P, *et al*. Hydrochemistry analysis based on the source determination of coal mine water-bursts [J]. Journal of Safety and Environment, 2015, 15(1): 31-35.
刘剑民, 王继仁, 刘银朋, 等. 基于水化学分析的煤矿矿井突水水源判别[J]. 安全与环境学报, 2015, 15(1): 31-35.

[3] Chen L W, Xu D Q, Yin X X, *et al*. Analysis on hydrochemistry and its control factors in the concealed coal mining area in north China: a case study of dominant inrush aquifers in Suxian mining area[J]. Journal of China Coal Society, 2017, 42(4): 996-1004.
陈陆望, 许冬清, 殷晓曦, 等. 华北隐伏型煤矿区地下水化学及其控制因素分析——以宿县矿区主要突水含水层为例[J]. 煤炭学报, 2017, 42(4): 996-1004.

[4] Chu X L, Lu W Z. Research and application progress of near infrared spectroscopy analytical technology in China in the past five years[J]. Spectroscopy and Spectral Analysis, 2014, 34(10): 2595-2605.
褚小立, 陆婉珍. 近五年我国近红外光谱分析技术研究与应用进展[J]. 光谱学与光谱分析, 2014, 34(10): 2595-2605.

[5] Yagi I, Ono R, Oda T, *et al*. Two-dimensional LIF measurements of humidity and OH density resulting from evaporated water from a wet surface in plasma for medical use[J]. Plasma Sources Science and Technology, 2014, 24(1): 15002.

[6] Yang Y X, Kang J, Wang Y R, *et al*. Super sensitive detection of lead in water by laser-induced breakdown combined with laser-induced fluorescence technique[J]. Acta Optica Sinica, 2017, 37(11): 1130001.
杨宇翔, 康娟, 王亚蕊, 等. 水中铅元素的激光诱导击穿光谱-激光诱导荧光超灵敏检测[J]. 光学学报, 2017, 37(11): 1130001.

[7] Huang Z L, Li Y L, Yu C Z, *et al*. Analysis of effects for measurements of concentration in water by laser induced fluorescence (LIF) technique[J]. Journal of experimental mechanics, 1994(3): 232-240.
黄真理, 李玉梁, 余常昭, 等. LIF 技术测量浓度场的影响因素分析[J]. 实验力学, 1994(3): 232-240.

[8] Wang X, Yang J, Li K. Fluorescence feature of dissolved organic matters in groundwater of mining area-Ⅱ. Distribution features in the deep aquifers [J]. Journal of Safety and Environment, 2015, 15(6): 97-100.
王新, 杨建, 李凯. 煤矿区地下水中溶解性有机质荧光特征Ⅱ——深部含水层分布特征[J]. 安全与环境学报, 2015, 15(6): 97-100.

[9] Wang S D. Distribution characteristics of fluorescent dissolved organic matter in different aquifers of Luotuoshan coal mine [J]. Coal Geology & Exploration, 2015(2): 53-57.
王世东. 骆驼山煤矿不同含水层水中荧光性 DOM 分布特征[J]. 煤田地质与勘探, 2015(2): 53-57.

[10] Yan P C, Zhou M R, Liu Q M, *et al*. Research on the source identification of mine water inrush based on LIF technology and SIMCA algorithm [J]. Spectroscopy and Spectral Analysis, 2016, 36(1): 243-247.
闫鹏程, 周孟然, 刘启蒙, 等. LIF 技术与 SIMCA 算法在煤矿突水水源识别中的研究[J]. 光谱学与光谱分析, 2016, 36(1): 243-247.

[11] Bandos T V, Bruzzone L, Camps-Valls G. Classification of hyperspectral images with regularized linear discriminant analysis [J]. IEEE Transactions on Geoscience and Remote Sensing, 2009, 47(3): 862-873.

[12] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks [J]. Science, 2006, 313(5786): 504-507.

[13] Li J F, Wang Y L, Hu S, *et al*. The comparison of spectral classification based on DBN, BP neural network and SVM [J]. Spectroscopy and Spectral Analysis, 2016, 36(10): 3261-3264.
李俊峰, 汪月乐, 胡升, 等. 基于 DBN, SVM 和 BP 神经网络的光谱分类比较[J]. 光谱学与光谱分析, 2016, 36(10): 3261-3264.

[14] Wang J, Zhang F, Wang X P, *et al*. Three-dimensional fluorescence characteristics by parallel factor method coupled with self-organizing map and its relationship with water quality[J]. Acta Optica Sinica, 2017, 37(7): 0730003.
王娟, 张飞, 王小平, 等. 平行因子法结合自组织映射神经网络的三维荧光特征及其与水质关系研究

[J]. 光学学报, 2017, 37(7): 0730003.

[15] Wang S T, Zhang C X, Wang Z F, *et al*. Application of least squares support vector machine in fluorescence detection of sodium methylparaben[J]. Laser & Optoelectronics Progress, 2017, 54(7): 073001
王书涛, 张彩霞, 王志芳, 等. 最小二乘支持向量机在对羟基苯甲酸甲酯钠荧光检测中的应用[J]. 激光与光电子学进展, 2017, 54(7): 073001.

[16] Hsu C W, Chang C C, Lin C J. A practical guide to support vector classification [EB/OL]. (2016-05-19). https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf.

[17] Wang S T, Chen D Y, Wang X L, *et al*. Detection of polycyclic aromatic hydrocarbons combining fluorescence analysis with ABC-BP neural network [J]. Chinese Journal of Lasers, 2015, 42(11): 1115001.
王书涛, 陈东营, 王兴龙, 等. 荧光分析法和 ABC-BP 神经网络相结合的多环芳烃的检测[J]. 中国激光, 2015, 42(11): 1115001.

[18] Huang G B, Zhu Q Y, Siew C. Extreme learning machine: a new learning scheme of feedforward neural networks [C]. IEEE International Joint Conference on Neural Networks, 2004, 2: 985-990.

[19] Ding S F, Zhao H, Zhang Y N, *et al*. Extreme learning machine: algorithm, theory and applications [J]. Artificial Intelligence Review, 2015, 44(1): 103-115.

[20] Wang Y, Zhou M R, Yan P C, *et al*. A rapid identification model of mine water inrush based on extreme learning machine[J]. Journal of China Coal Society, 2017, 42(9): 2427-2432.
王亚, 周孟然, 闫鹏程, 等. 基于机限学习机的矿井突水水源快速识别模型[J]. 煤炭学报, 2017, 42(9): 2427-2432.

[21] Huang G, Huang G B, Song S J, *et al*. Trends in extreme learning machines: a review [J]. Neural Networks, 2015, 61: 32-48.

[22] Belkin M, Niyogi P, Sindhwani V. On manifold regularization[C]. AISTATS, 2005: 1.

[23] Liu B, Xia S X, Meng F R, *et al*. Manifold regularized extreme learning machine [J]. Neural Computing and Applications, 2016, 27(2): 255-269.

[24] Huang G B, Zhou H M, Ding X J, *et al*. Extreme learning machine for regression and multiclass classification [J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2012, 42(2): 513-529.

[25] Deng C W, Huang G B, Xu J, *et al*. Extreme learning machines: new trends and applications [J]. Science China (Information Sciences), 2015, 58(2): 1-16.

[26] Monti F, Boscaini D, Masci J, *et al*. Geometric deep learning on graphs and manifolds using mixture model CNNs[C]. Honolulu: 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[27] 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.

[28] Zhou H M, Huang G B, Lin Z P, *et al*. Stacked extreme learning machines[J]. IEEE Transactions on Cybernetics, 2015, 45(9): 2013-2025.

[29] Kasun L L C, Yang Y, Huang G, *et al*. Dimension reduction with extreme learning machine [J]. IEEE Transactions on Image Processing, 2016, 25(8): 3906-3918.

[30] Tang J X, Deng C W, Huang G B, *et al*. A fast learning algorithm for multi-layer extreme learning machine [C]. IEEE International Conference on Image Processing, 2014: 175-178.

[31] Kasun L L C, Zhou H M, Huang G B, *et al*. Representational learning with ELMs for big data[J]. IEEE Intelligent Systems, 2013, 28(6): 31-34.

[32] Chang C C, Lin C J. LIBSVM: a library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2011, 2(3): 27.

[33] Tang J X, Deng C W, Huang G B. Extreme learning machine for multilayer perceptron [J]. IEEE Transactions on Neural Networks and Learning Systems, 2016, 27(4): 809-821.

[34] Melacci S, Belkin M. Laplacian support vector machines trained in the primal [J]. Journal of Machine Learning Research, 2011, 12(3): 1149-1184.