AdaBoost算法在矿井突水水源的荧光光谱识别中的研究

周孟然，李大同*，胡 锋，来文豪，王 亚，朱 松

安徽理工大学电气与信息工程学院，安徽 淮南 232001

摘 要 矿井突水是影响矿井安全生产的重要因素之一，如果矿井发生突水，能够快速、准确地判别突水水源类型是治理矿井突水灾害保证生产安全的重要环节，因此，建立一个能够快速识别矿井突水水源的模型具有重要的意义。水化学分析法作为在传统的矿井突水水源类型识别方法里应用最为广泛的识别方法，通过获得相应的pH值、离子浓度、电导率等参数，然后利用这些参数来建立突水水源的类型识别模型对矿井突水的类型进行判别。针对这种传统矿井突水水源识别方法在判别时间上耗时长和识别准确率低等不足，鉴于LIF技术具有分析速度快、灵敏度高等优点，提出了将线性判别分析(LDA)算法作为弱分类器的自适应提升(AdaBoost)算法用于激光诱导荧光(LIF)光谱识别矿井突水水源的新方法。用于实验的九种水样(每种水样各取50个样本)由淮南地区某矿的老空水、灰岩水以及按不同比例混合的老空水与灰岩水的七种混合水构成。将405 nm激光器发射的激光打入被测水体并采集荧光光谱数据，然后对采集到450组荧光光谱数据进行分析，取其中360组光谱数据(每种水样各40组)用作训练集，取剩余90组光谱数据用作测试集。分别选取三种算法针对水样的激光诱导荧光光谱的分类进行了建模并将三种结果进行对比。首先利用决策树算法对光谱进行分类识别，在节点个数为8时决策树对测试集的分类效果最好，分类准确率达到91.11%。然后针对决策树算法分类效果的不足，利用决策树算法作为弱分类器的AdaBoost算法，当选取节点个数为9的决策树作为弱分类器的时，对训练集的分类准确率为97.78%。最后针对基于决策树的AdaBoost算法的泛化性能不足和为了获得更好的分类效果，提出了基于LDA算法作为弱分类器的AdaBoost算法，在设置迭代次数为150后对水样光谱数据分类准确率可以达到100%。通过实验结果可以发现，集成学习算法的分类能力比传统的分类算法对水样的光谱的分类识别能力更强，相较于同为九个节点的决策树算法，采用节点数为9的决策树作为弱学习器的AdaBoost算法对测试集的分类准确率从88.89%提升到了97.78%，对训练集的分类准确率从99.72%提升到了100%；然后可以发现相对于使用决策树作为弱分类器的AdaBoost算法，采用LDA算法作为AdaBoost算法的弱分类器对水样的光谱的测试集的分类准确率从97.78%提升到了100%，对训练集的分类准确率达到100%，具有更好的识别效果，并且具有更好的泛化性能。实验结果证明采用Adaboost-LDA算法为激光荧光光谱的模式分类用于矿井突水水源的判别和预警是可行且有效的。

关键词 矿井突水；LIF技术；决策树；LDA；AdaBoost

中图分类号：O657.3 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2019)02-0485-06

引 言

矿井突水是危害矿井安全的五大灾害之一，造成的直接与间接经济损失位列首位[1]。矿井突水水源的快速判别是突水灾害控制的重要环节[2]，一旦矿井发生突水，快速、准确地判断突水水源类型是治理矿井突水灾害的前提和关键[3]。

目前，突水水源的识别主要有水化学分析法、地质分析法、水动力分析法、水温度分析法以及地球物理勘探法[4]。在传统的矿井突水水源类型识别方法里，水化学分析法的应用最为广泛，通过水化学分析法获得相应的pH值、离子浓度、电导率等参数，然后通过这些参数来建立突水水源的类

收稿日期：2018-01-28，**修订日期**：2018-06-12

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)资助

作者简介：周孟然，1965年生，安徽理工大学电气与信息工程学院教授 e-mail：mrzhou8521@163.com

*通讯联系人 e-mail：ldt5737@163.com

型识别模型对矿井突水的类型进行判别[5-6]。然而，实验室要测量水样中的这些参数通常需要 2 h 才能完成，由于水化分析法耗时相对过长，因此该方法不适宜用于矿井水害预警防治。

激光诱导荧光光谱分析具有分析速度快、精度高、灵敏度高等优点，在化工、医疗和生物等诸多领域有着普遍应用。近年来，矿井突水水源的识别开始采用激光荧光光谱分析技术，如闫鹏程[7]对同一矿井的五种不同的煤矿突水水源的荧光光谱通过波长范围的选取、消除噪声、主成分分析(PCA)降维、最后采用簇的独立软模式(SIMCA)算法对降维后的数据建模实现了对单一水样的识别等。AdaBoost 算法是一种可以将弱分类器提升为强分类器的常用集成学习算法，它已经有效地应用在人脸识别、医学图像识别、不平衡数据分析、大规模数据分析、图像分割、模式识别和光谱分析等领域[8-13]。考虑到矿井的复杂的实际情况，针对单一突水水样的突水水源识别已经不能够满足矿井安全生产的需求，为了满足实际的井下突水预警需求，对混合型突水水源进行分类识别才更具有实际价值。本文所采用的AdaBoost算法可以将弱分类器提升为强分类器，将其分别与决策树算法和 LDA 算法结合用于矿井混合突水水源的激光诱导荧光光谱分析的研究，寻找最合适的鉴别模型。相对于传统的需要对突水水样的荧光光谱采取截取、去噪、降维之后再使用模式识别算法对荧光光谱建模的分类方法，本文采用的AdaBoost算法不需要如此复杂的预处理过程，可直接使用AdaBoost算法对荧光光谱进行建模分析。目前，尚未有将基于AdaBoost的分类算法用于矿井混合突水水源的激光诱导荧光光谱分析的报道。

1 实验部分

1.1 材料

矿井突水是矿井安全生产中重要的隐患之一，其中老空水是矿井突水危害最大的突水水源，本实验以老空水，灰岩水以及按比例混合的老空水与灰岩水的混合水样本作为研究对象。实验材料为 2017 年 12 月在淮南地区某矿区采集的老空水和灰岩水，并通过混合得到如下的九种水样，九种实验样本依次为老空水、老空水和灰岩水按体积比为 4∶1 混合的混合水(以下简称“混合水样 A”)、老空水和灰岩水按体积比为 3∶1 混合的混合水(以下简称“混合水样 B”)、老空水和灰岩水按体积比为 2∶1 混合的混合水(以下简称“混合水样 C”)、老空水和灰岩水按体积比为 1∶1 混合的混合水(以下简称“混合水样 D”)、老空水和灰岩水按体积比为 1∶2 混合的混合水(以下简称“混合水样 E”)、老空水和灰岩水按体积比为 1∶3 混合的混合水(以下简称“混合水样 F”)、老空水和灰岩水按体积比为 1∶4 混合的混合水(以下简称“混合水样 G”)、灰岩水。每种水样分别采集 50 组光谱数据，总共 450 组。随机选取其中的 360 组(每种样本各取 40 组)作为训练集，剩余的 90 组(每种样本 10 组)作为测试集。所有采集到的水样都密封，遮光保存。

1.2 仪器

实验使用的激光诱导荧光光谱仪仪器选择是美国 Ocean Optics 生产的 USB2000＋的微型光纤光谱仪。采用 405 nm 激光器作为实验光源，入射激光的功率为 120 mW，检测荧光光谱的范围 340～1 021 nm，分辨率设定为 0.5 nm。实验探头采用 FPB-405-V3 可浸入式激光激发荧光探头(广东科思凯公司)，可以直接放入待测水体以便实时获取荧光光谱的数据。实验时将探头垂直浸入突水水样样本的接触测量。为避免环境光对实验的影响，实验在暗室下对水样进行激光诱导荧光光谱的采集。荧光光谱数据由 SpectraSuite 软件采集和记录。

1.3 算法描述

1.3.1 决策树判别方法

决策树包括一个根节点、若干个内部节点和若干个叶节点等部分，是一种基于树结构进行决策的自然的处理机制。其中叶节点对应的是决策树的决策的结果，其他的每个节点对应于一个属性测试；根据属性测试的结果每个节点所包含的样本集合会被划分到子节点中；根节点包含样本全集。从根节点到每个叶节点的路径对应了一个判定测试序列。在处理大规模的数据中，决策树不需要接受训练数据外的知识，并具有较高的分类精度。

1.3.2 LDA 算法

LDA 算法是将给定的训练集投影到一条直线上，使训练集中相同类别样本点尽可能的聚集，异类样本点尽可能远；在对测试集进行分类时，采用同样方法将他们投影到这条直线上，然后根据样本点的位置来确定测试集的类别。LDA 算法是一种经典线性学习算法。

1.3.3 AdaBoost 算法

AdaBoost 算法是集成学习的一种，通过结合多个弱分类器构成一个分类效果更好的强分类器来完成学习任务。AdaBoost 算法的实现过程如下：模型训练中，样本具有相同的初始权值，并训练学习出一个弱分类器，且计算出该分类器的误差率；然后模型的每次训练，都会在前一次的学习结果基础上调整样本权值，即，把识别错误的样本权重增大，同时降低分类正确的样本的权重。AdaBoost 的学习过程的实质，是在不停的学习中改变样本的权值，直到学习结果的误差为 0 或学习器个数达到预设值，然后把所有弱分类器学习的结果按权值综合，输出最终的结果。

AdaBoost 算法流程如下所示。

输入：

训练数据：$S=\{(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)\}$，$x_i \in X$；

标签 $y_i \in Y=\{1, 2, \cdots, k\}$；

弱分类算法：Weak Classifier；

迭代次数 T；

过程：

(1) 初始化权重值 $D_1(x)=\frac{1}{m}$

(2) for $t=1, 2, \cdots, T$ do

(3) 使用弱分类学习算法获得弱分类器 h_t：$X \rightarrow Y$

(4) 计算 h_t 的错误率：$\varepsilon_t=P_{x\sim D_t}(h_t(x)\neq f(x))$

(5) if $\varepsilon_t > 0.5$ then break

(6) 弱分类器权重 $\alpha_t = \frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$

(7) 更新训练数据权重 D_t：$D_{t+1} = \frac{D_t(x)\exp(-\alpha_t f(x) h_t(x))}{Z_t}$（$Z_t$ 为规范化因子）

(8) end for

输出：$H(x) = \mathrm{sign}\left(\sum_{t=1}^{T}\alpha_t h_t(x)\right)$

2　结果与讨论

2.1　突水水样光谱

通过一系列实验测试得到的矿井突水水样荧光光谱如图 1 所示。从上至下依次为老空水、混合水样 A、混合水样 B、混合水样 C、混合水样 D、混合水样 E、混合水样 F、混合水样 G、灰岩水。从图中可以看出当老空水和灰岩水按一定的比例混合时，突水水样的激光诱导荧光光谱无法直观的进行区分。

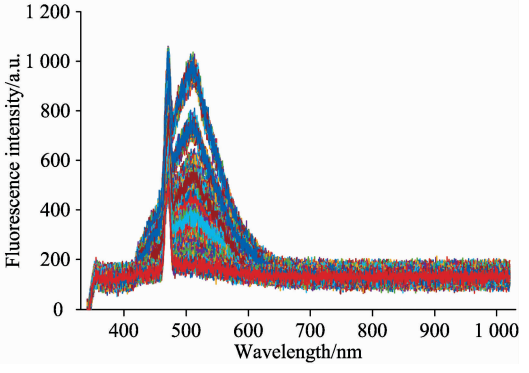

图 1　矿井突水水样光谱

Fig. 1　Spectrum of mine water inrush sample

2.2　决策树建模

使用决策树分类器将光谱数据的训练集进行建模分析，然后通过测试集测试该模型的识别效果。决策树按照属性作为树节点的分类变量，把测试变量分到各个分支中建立一颗决策分类树。因此通过手动设置不同的树节点个数，然后对比不同树节点对应的分类准确率来确定最优的树节点个数。树节点个数与分类准确率之间的关系如表 1 所示。

由表 1 可以看出当选取的节点个数越多时决策树算法对突水水样光谱数据的分类效果有所提升。当节点个数为 8 时，决策树对测试集的分类准确率达到 91.11%；此时该模型对测试集的分类准确率为 99.72%。当节点个数超过 8 个时，决策树的分类准确度开始下降。由于矿井突水的严重危害，为避免在实际应用中出现误判的行为，因此需要进一步提升对矿井突水荧光光谱的分类准确率。AdaBoost 算法可以在很大程度上提升弱分类器的分类效果，因此为了获得较高的矿井突水荧光光谱的分类准确率，将采用 AdaBoost 算法进行模型建立。

表 1　不同节点个数决策树的分类准确率

Table 1　The relational table between number of nodes and classification accuracy

节点个数	分类准确率/%	节点个数	分类准确率/%
1	22.22	7	80.00
2	33.33	8	91.11
3	44.44	9	88.89
4	55.56	10	88.89
5	64.44	11	87.78
6	71.11	12	86.67

2.3　AdaBoost 算法模型建立

2.3.1　Adaboost-Tree 建模

使用 Adaboost-Tree 对从所有光谱数据中随机选取作为训练集的 360 组光谱数据进行建模，同样的使用测试集测试模型的识别效果。初始化样本权重 $D_1(x)$，迭代次数 T 设定 150，将作为弱分类器的决策树的节点个数设置为 N。在不同节点个数下 Adaboost-Tree 分类器对应的分类准确率如表 2 所示。

表 2　不同节点个数 Adaboost-Tree 的分类准确率

Table 2　The relational table between number of nodes and classification accuracy

节点个数	分类准确率/%	节点个数	分类准确率/%
1	55.56	7	95.56
2	82.22	8	96.67
3	93.33	9	97.78
4	95.56	10	96.67
5	93.33	11	95.56
6	94.44	12	86.67

由表 2 可知当作为弱分类器决策树的节点个数为 9 时，Adaboost-Tree 分类器的分类效果最好达到 97.78%，对训练集的分类准确率达到 100%。Adaboost-Tree 分类器得到的训练周期上的替换损失如图 2 所示，Adaboost-Tree 在训练周期上的泛化误差如图 3 所示。由图可以看出分类器的替换损失在第三个周期时降为 0，但是泛化误差达到 0.161 1，因此通过不断迭代 150 次时泛化误差为 0.1，具有很强的泛化性能。

AdaBoost-Tree 算法在迭代 150 次后的对测试集分类准确率较同为 9 个节点个数的决策树算法从 88.89%提升到了 97.78%；对训练集的分类准确率从 99.72%提升到了 100%，通过对上述的分析证明了 AdaBoost 算法对弱分类器分类效果的提升。

2.3.2　Adaboost-LDA 建模

当使用 AdaBoost-Tree 算法对光谱数据进分类时，分类准确率达到 97.78%，为进一步提升 AdaBoost 算法对光谱数据的准确率，使用 LDA 算法作为新的弱分类器对光谱数据进行建模。使用 Adaboost-LDA 算法对上文相同的随机选取作为训练集的 360 组光谱数据进行建模，初始化样本权重 $D_1(x)$，设置迭代次数 $T=150$，选择 LDA 的判别方法为伪

线性。Adaboost-LDA 分类器得到的替换损失与训练周期上的关系如图 4 所示。Adaboost-LDA 在训练周期上的泛化误差如图 5 所示。由图 4 和图 5 可以看出分类器的替换损失在迭代 150 次后降为 0.005 6，泛化误差为 0.022 2，相较于 AdaBoost-Tree 算法，Adaboost-LDA 算法具有更强的泛化性能和稳定性，更具有实际应用价值。通过对实验的结果分析可知 Adaboost-LDA 算法在不断迭代 150 次后对训练集和测试集的分类准确率均达到 100%，相较于节点个数为 9 的 AdaBoost-Tree 算法对测试集的分类准确率从 97.78%提升到了 100%，通过比较三种算法的建模可知 Adaboost-LDA 算法对于突水水样的激光诱导荧光光谱的分类效果更好。

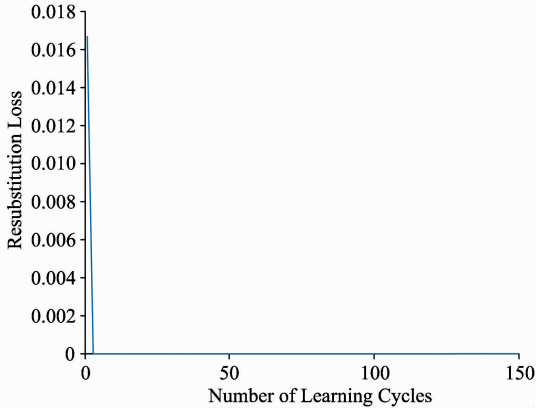

图 2 Adaboost-Tree 在训练周期上的替换损失

Fig. 2 The relationship between learning cycles and resubstitution loss

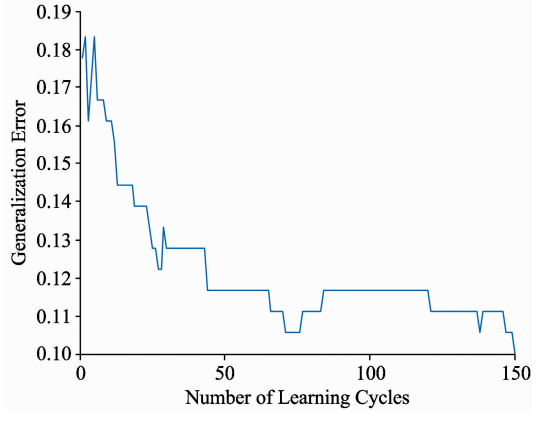

图 3 Adaboost-Tree 在训练周期上的泛化误差

Fig. 3 The relationship between learning cycles and generalization error

3 结 论

选取淮南某矿区的老空水，灰岩水和老空水与灰岩水按一定比例混合的混合水水样作为研究对象，分别选取三种算法针对水样的激光诱导荧光光谱的分类进行了建模并将三种结果进行对比，首先比较了集成了决策树作为弱分类器的集成学习相对决策树对光谱识别的提升，然后对集成了不同弱分类器的集成学习对光谱的识别分类效果进行了比较。通过实验可以发现：第一，集成学习算法的分类能力比传统的分类算法对水样的光谱的分类识别能力更强；第二，相对于使用决策树作为弱分类器的 AdaBoost 算法，采用 LDA 算法作为 AdaBoost 算法的弱分类器水样的光谱的分类具有更好的识别效果，不仅对训练集和测试集的分类准确率均可以达到 100%，并且具有更好的泛化性能。实验证明将基于 LDA 的 AdaBoost 算法用于矿井突水的激光诱导荧光光谱分析是可行的，通过分析水样的荧光光谱来判断突水水样的成分，对预测矿井突水保障矿井安全生产具有很重要的意义。本文采用 Adaboost-LDA 算法不仅可以用于老空水，灰岩水和老空水与灰岩水按一定比例混合的混合水水样分类识别，也可用于不同矿井突水的激光荧光光谱分类以及其他种类的矿井突水及混合水的识别，同时对其他种类水的分析有一定的参考价值，同时也为激光荧光光谱的模式分类在其他领域的应用提供了一种有效的解决途径。

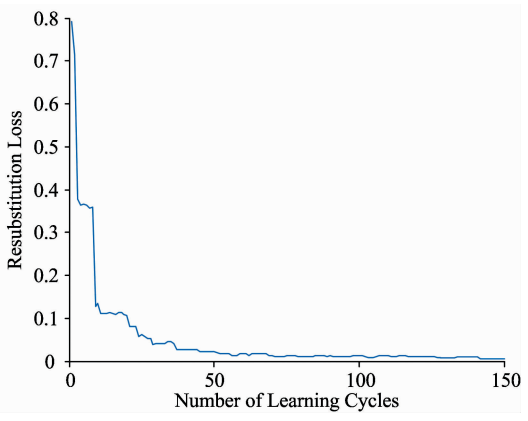

图 4 Adaboost-LDA 在训练周期上的替换损失

Fig. 4 The relationship between learning cycles and resubstitution loss

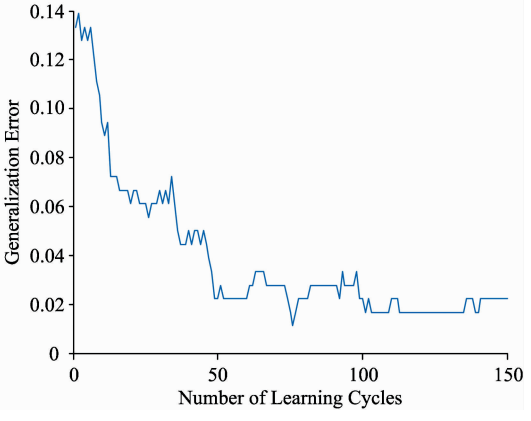

图 5 Adaboost-LDA 在训练周期上的泛化误差

Fig. 5 The relationship between learning cycles and generalization error

References

[1] MENG Lei, DING En-jie, WU Li-xin(孟 磊, 丁恩杰, 吴立新). Journal of China Coal Society(煤炭学报), 2013, 38(8): 1397.
[2] XU Xing, WANG Gong-zhong(徐 星, 王公忠). Coal Technology(煤炭技术), 2016, 35(7): 144.
[3] ZHANG Yan(张 雁). Coal Science and Technology(煤炭科学技术), 2014, 42(10): 98.
[4] SHAO Liang-shan, LI Yin-chao, XU Bo(邵良杉, 李印超, 徐 波). Journal of Safety and Environment(安全与环境学报), 2017, 17(5): 1730.
[5] LIU Jian-min, WANG Ji-ren, LIU Yin-peng, et al(刘剑民, 王继仁, 刘银朋, 等). Journal of Safety and Environment(安全与环境学报), 2015, 15(1): 31.
[6] YIN Xiao-xi, CHEN Lu-wang, XIE Wen-ping, et al(殷晓曦, 陈陆望, 谢文苹, 等). Hydrogeology & Engineering Geology(水文地质工程地质), 2017, 44(5): 33.
[7] YAN Peng-cheng, ZHOU Meng-ran, LIU Qi-meng, et al(闫鹏程, 周孟然, 刘启蒙, 等). Spectrocopy and Spectral Analysis(光谱学与光谱分析), 2016, 36(1): 243.
[8] JIANG Wei-jian, GUO Gong-de, LAI Zhi-ming, et al(江伟坚, 郭躬德, 赖智铭, 等). Journal of Shangdong University • Engineering Science(山东大学学报 • 工学版), 2014, 44(2): 43.
[9] ZHU Cheng-zhang, XIANG Yao, ZOU Bei-ji, et al(朱承璋, 向 遥, 邹北骥, 等). Journal of Computer-Aided Design & Computer Graphics(计算机辅助设计与图形学学报), 2014, 26(3): 445.
[10] LI Yi-jing, GUO Hai-xiang, LI Ya-nan, et al(李诒靖, 郭海湘, 李亚楠, 等). Systems Engineering-Theory & Practice(系统工程理论与实践), 2016, 36(1): 189.
[11] Tao CHEN, Shijian LU. IEEE Transactions on Vehicular Technology, 2016, 65(6): 4006.
[12] Ignacio Martin-Diaz, Daniel Morinigo-Sotel, Oscar Duque-Perez, et al. IEEE Transactions on Industry Applications, 2017, 53(3): 3066.
[13] Andreas Ranftl, Fernando Alonso-Fernandez, Stefan Karlsson. The Institution of Engineering and Technology, 2017, 6(6): 468.

Research of the AdaBoost Arithmetic in Recognition and Classifying of Mine Water Inrush Sources Fluorescence Spectrum

ZHOU Meng-ran, LI Da-tong*, HU Feng, LAI Wen-hao, WANG Ya, ZHU Song
College of Electrical and Information Engineering, Anhui University of Science and Technology, Huainan 232001, China

Abstract The water inrush is one of the most important elements that can influence the mining safety, and being able to recognize the category of water inrush sources accurately and rapidly will greatly enhance the mining safety condition when water inrush happens accidentally. Therefore, it is extremely important and necessary to create a model system that can recognize water inrush sources effectively. The water chemistry analytical method is the widest used method to recognize water inrush sources among traditional methods; in this method, we build a model system by using ph, ionic concentration, conductivity and so on, then use that model system to recognize water inrush sources. However, the water chemistry analytical method has disadvantages that usually be time-costing and of low accuracy. This essay will deal with this problem and introduce the AdaBoost method that uses LDA as weak classifier based on LIF technology because of the rapidnessand high sensitivity of LIF technology. In this research, there are nine kinds of waters from a certain mine in the Huainan City considered and fifty independent samples in each kind of water, limestone water, high pressure water from floor of coal seam and gob areas, and seven different proportion mixture of those two kind of water. Emit laser from the 405nm laser emitter into laboratory water samples and collect experiment statistics of fluorescence spectrum, analyze these 450 water samples by select 360 samples (40 samples of each kind of water source) as a training set first and set other 90 samples as a training set. In this essay, we use three different kinds of arithmetic to build three different model systems and compare results from each model system. First of all, we use decision-making tree to recognize and classify different fluorescence spectrum, we get the best outcome and the accuracy rate is 91.11% at that time when the node number is 8. Then, we use the AdaBoost arithmetic and set the decision-making tree as the weak classifier according to the shortage of the decision-making tree, and we get the best accuracy rate of classifying training sets of 97.78% when selecting a decision-making tree whose node number is 9 as the weak classifier. And last, we introduce a AdaBoost arithmetic base on setting LDA arithmetic as the weak classifier to get better classifying results according to the generalization shortage of AdaBoost arithmetic which bases on decision-making tree, and finally we get the spectrum accuracy rate of 100% when it-

erate 150 times. As we can get from our experiment, classifying arithmetic that integrates the learning arithmetic is much better than other traditional classifying arithmetic, for instance, compared with the decision-making tree arithmetic, AdaBoost arithmetic which sets decision-making tree as its weak classifier can enhance the accuracy rate of classifying testing set from 88.89% to 97.78% and enhance the accuracy rate of classifying training set from 99.72% to 100% when the node number is 9; then compared with the AdaBoost arithmetic which sets decision-making tree as its weak classifier, the AdaBoost arithmetic which uses LDA as its weak classifier can enhance the accuracy rate of classifying sample water fluorescence spectrum testing set from 97.78% to 100% and enhance the accuracy rate of classifying sample water fluorescence spectrum training set to 100% as well, and we can get better recognition outcomes and make our model system have better generalization by using such strategy at the same time. Therefore, it is extremely fair to say that using AdaBoost-LDA arithmetic to classify fluorescence spectrum to recognize and alarm water inrush sources is effective and feasible.

Keywords Mine water inrush; LIF technology; Decision-making tree; LDA; AdaBoost

(Received Jan. 28, 2018; accepted Jun. 12, 2018)

* Corresponding author