doi:10.13582/j.cnki.1674-5876.2021.01.008

煤层底板突水预测的 PCA-OPF 模型

江泽华[1],袁志刚[1,2,3*],谢东海[1,2,3],邵耀华[1]

(1.湖南科技大学 资源环境与安全工程学院,湖南 湘潭 411201;2.湖南科技大学 煤矿安全开采技术湖南省重点实验室,湖南 湘潭 411201;
3.湖南科技大学 南方煤矿瓦斯与顶板灾害预防控制安全生产重点实验室,湖南 湘潭 411201)

摘 要:矿井突水是常见的突发性强烈的矿井灾害.由于矿井突水的突发性和危险性,有必要对煤层底板突水进行更快且更准确地预测.为此,本文提出了基于 PCA-OPF 模型的煤层底板突水预测方法.通过选取断层分维值、取芯率、隔水层厚度、单位涌水量、渗透系数、底板含水层总厚度、承压含水层水压作为底板突水预测的因子,利用主成分分析(PCA)将样本的 7 个因子降维为 3 个主成分,进一步简化数据结构和提高预测速度.利用最优路径森林算法(OPF)对降维后的 30 个样本数据进行训练和预测,并与实际情况进行对比.结果表明:基于 PCA-OPF 模型得到的测试集中 6 个样本的预测结果与实际情况相符,为煤层底板突水预测提供了一种新的方法.

关键词:主成分分析;最优路径森林算法;煤层底板突水;预测模型

中图分类号:TD745 **文献标志码**:A **文章编号**:1672-9102(2021)01-0049-06

Prediction of Coal Seam Floor Water Inrush Based on PCA-OPF Model

JIANG Zehua[1], YUAN Zhigang[1,2,3], XIE Donghai[1,2,3], SHAO Yaohua[1]

(1. School of Resources, Environment and Safety Engineering, Hunan University of Science and Technology, Xiangtan 411201, China;
2. Hunan Provincial Key Laboratory of Safe Mining Techniques of Coal Mines, Hunan University of Science and Technology, Xiangtan 411201, China;
3. Work Safety Key Lab on Prevention and Control of Gas and Roof Disasters for Southern Coal Mines,
Hunan University of Science and Technology, Xiangtan 411201, China)

Abstract: Mine water inrush is a common sudden and strong mine disaster. Because of the sudden and dangerous of mine water inrush, it is necessary to be predicted more quickly and accurately. A prediction method based on PCA-OPF model is proposed. The fault fractal dimension, coring rate, thickness of water barrier, unit water inflow, permeability coefficient, total thickness of bottom aquifer and water pressure of confined aquifer are selected as the factors to predict bottom water inrush. Principal component analysis (PCA) is used to reduce 7 factors in the sample to 3 principal components, which simplifies the data structure and further proposed high prediction speed.Finally, the optimal path forest algorithm (OPF) is used to train and predict the 30 sample data after dimensionality reduction and with the actual situation. The results show that based on the PCA-OPF model, the prediction results of 6 sample in the test set are consistent with the actual situation, which provides a new method for predicting water inrush from coal seam floor.

Keywords: principal component analysis;optimal-path forest algorithm;floor water inrush; prediction model

我国水文地质复杂,特别在华北型矿区,煤层底板多含奥灰含水层,矿井突水事故时有发生[1].矿井突水严重威胁矿工的生命安全,也影响了矿井的安全生产.为确保煤炭资源的安全开采,如何快速且准确地对煤层底板突水进行预测成为亟待解决的问题.针对煤矿底板突水问题我国学者做了大量研究并取得了

收稿日期:2020-10-27

基金项目:国家自然科学基金资助项目(51604111);湖南省教育厅资助科研项目(16C0654;18B213);湖南省自然科学基金资助项目(2017JJ2082)

* **通信作者**,E-mail:cquyzg@163.com

很多成果[2-7].

煤层开采底板的“下三带”与“四带”的划分对防治矿井底板突水起到了重要作用[2-3].李春元等分析了深部开采砌体梁结构失稳扰动底板破坏的动载源特征,揭示了深部开采底板突水机理[4];白峰青等开展了现场注水模拟试验,揭示了底板岩体破裂变化特征[5];王妍等用弹性力学方法求取隔水关键层的应力以及位移,为采场底板突水的预测预报提供理论支撑[6];王向前等通过在数值模拟软件模拟与突水系数法结合,实现了工作面带压开采的可行性分析[7].

但由于岩体介质的非线性、复杂性、不确定性等特点[8],传统通过理论分析、经验数值计算、相似模拟以及数值模拟满足不了对底板突水预测的需要.近年来,人工智能的发展为解决该问题提供了一种新的途径[9],即利用机器学习对煤层底板突水进行科学预测[10-15].

赵斐提出了模糊-支持向量机模型[10];施龙青等提出了 Fuzzy_PCA_PSO_SVC 以及基于灰狼算法优化的 Elman 神经网络模型[11,12];张风达提出了基于 PSO 算法优化的 SVM 模型[13];温廷新等提出了 PSO_SVM_AdaBoost 预测模型[14].以上研究为煤层底板突水预测提供了新方法,但这些算法选取的因子较多,且需要对参数进行优化,尚不能满足对煤层底板突水进行快速且准确预测的需要.为此,本文提出了基于主成分分析的最优路径森林模型(PCA-OPF),该模型通过主成分分析(PCA)将多因子减少为少数几个主成分,简化了最优路径森林算法(OPF)的数据结构,同时利用了 OPF 算法本身具有与参数无关且不需要参数优化的特点[15],能对煤层底板突水进行快速且准确地预测.

1 PCA-OPF 模型基本原理

1.1 PCA

煤层底板突水是由多个非线性因子导致的一种复杂动力现象[1].目前,煤层底板突水预测选择的因子较多,且这些因子之间存在一定的相关性,导致在数据集的分析、处理过程中往往因计算步骤过多而把问题变得更加复杂.

PCA 是一种利用降维来简化数据集的数据处理方法,其通过把高维数据投影到低维层面,使原始样本数据中的多个因子减为少数几个能包含原始样本数据大部分信息的综合性指标[16].限于篇幅原因,PCA 的具体原理、步骤不再赘述,详见文献[17].

1.2 OPF 模型

OPF 是由 Papa 等人提出的一种新的基于图的分类器[18-21].基于数据样本标签的不同,OPF 算法可分为 3 种类型:数据集有标签的监督式 OPF 算法(SupervisedOPF)[18]、数据集没有标签的无监督式 OPF 算法[19]和 2 种情况都有的半监督式 OPF 算法[20].由此可知,OPF 算法的选择主要根据数据集的标签来确定.对于煤层底板突水预测,由于数据样本较少且每个样本均能被正确标记,因此选择 SupervisedOPF 算法对数据进行处理,并对 SupervisedOPF 算法的原理进行介绍.

1.2.1 SupervisedOPF 原理

数据集 $\boldsymbol{Z}$ 是被正确用 i 标记的样本集合($i=1,2,\cdots,c$),它被分为训练数据集($\boldsymbol{Z}_1$)、测试数据集($\boldsymbol{Z}_2$).而 OPF 算法对数据的分类是通过构建完全图实现的.具体步骤如下:

1) 训练阶段

OPF 算法将 $\boldsymbol{Z}_1$ 中的每 1 个样本看做为 1 个节点,并且各节点之间两两相连,并用弧代表他们的连接关系,从而构成了一个完全图 $G_1=(V_1,\boldsymbol{A}_1)$[18].其中,$V_1$ 代表着 $\boldsymbol{Z}_1$ 各样本间的弧,$\boldsymbol{A}_1$ 代表了 $\boldsymbol{Z}_1$ 各样本的特征向量.

完全图通过生成最小生成树(MST)[21]获得原型[18]样本 s,$s\in\boldsymbol{Z}_1$,即来自不同分类的所有相邻样本.一旦原型样本被找到,它们通过路径代价函数 $f_{\max}$ 相互竞争并征服来自训练集的其他样本,进而形成一个以原型样本为根节点的最优路径树(OPT)[18],所有的最优路径树就组成了最优路径森林(OPF)[18].OPF 算法对最优路径有如下定义:

路径 π_s 是各个以样本 s 为终点的节点序列.可以通过式(1)为每条路径赋予一个代价 $f(\pi)$.

$$f_{\max}(<s>)=\begin{cases}0,\text{假如 } s\in S;\\ +\infty,\text{其他.}\end{cases}\tag{1}$$

$$f_{\max}(\pi_s\cdot<s,t>)=\max\{f_{\max}(\pi),d(s,t)\}.$$

式中:$f_{\max}(<s>)$为当路径只有一个样本 s 时的代价,若 s 为原型样本则代价为零,若为 s 他则代价为无穷大.$f_{\max}(\pi_s\cdot<s,t>)$为其他样本 t 沿着路径 $\pi_s\cdot<s,t>$到样本 s 之间的最大距离.

如果路径 π_s 的代价 $f(\pi)$比其他同样以样本 s 为终点的路径 τ_s 代价要小,则路径 π_s 为最优路径.因此最优路径的最小化代价 $C(t)$为

$$C(t)=\min_{\forall\pi_t\in(\mathbf{Z}_1,A)}\{f_{\max}(\pi_t)\}.\tag{2}$$

2)测试阶段

在测试阶段中,每一个属于测试集的样本 t 被单独分类,它连接了在训练阶段产生的各个最优路径树的所有节点,并计算连接到各最优路径树的代价,若找到路径代价最小的最优路径树,则该最优路径树根节点的标签(原型样本标签)即为测试集样本 t 的标签.

2　煤层底板突水预测的 PCA-OPF 模型实现

煤层底板突水预测的步骤:首先收集煤层顶板突水预测相关样本数据,并将样本数据导入 OPF 算法中进行解析以及分组;其次利用 PCA 进行主成分分析;最后采用 OPF 算法对降维后的样本数据进行训练和测试并得到预测结果.

2.1　样本数据的收集

通过对文献[10-14]的煤层底板突水预测样本数据调研,最终确定了本文煤层底板突水预测所采用的 30 个样本数据.根据 OPF 算法的要求,对获取的样本数据进行解析以及分组,即将 30 个样本数据分为 24 个训练集和 6 个测试集,分组后的原始样本数据如表 1 所示.表 1 中,若煤层底板未突水则标签值为 1,若突水则标签值为 2;X_1,X_2, X_3,X_4,X_5,X_6 和 X_7 分别为选取的断层分维值因子、取芯率因子、隔水层厚度因子(m)、单位涌水量因子(L/(s·m))、渗透系数因子(m/d)、底板含水层总厚度因子(m)和承压含水层水压因子(MPa).以上 7 个因子对煤层底板突水危险性的影响详见文献[12].

表 1　原始样本数据

名称	编号	标签	X_1	X_2	X_3	X_4	X_5	X_6	X_7	编号	标签	X_1	X_2	X_3	X_4	X_5	X_6	X_7
训练集	0	1	0.04	0.5	49	3.8	79	33	1.2	12	1	0.11	1.6	55	3.8	81	28	1.7
	1	1	0.13	1.9	64	4.6	94	28	1.4	13	2	0.04	1.6	54	3.8	84	23	1.2
	2	1	0.14	2.5	63	4.5	96	28	1.6	14	1	0.11	2.0	52	3.9	66	31	1.5
	3	1	0.14	2.4	65	4.5	96	28	1.5	15	1	0.12	1.7	64	4.5	92	28	1.5
	4	1	0.11	1.7	55	3.9	84	27	1.6	16	1	0.08	1.3	51	3.8	72	32	1.5
	5	2	0.04	1.3	58	4.2	85	29	1.5	17	1	0.12	1.7	58	4.2	90	28	1.6
	6	2	0.09	1.2	53	3.9	71	30	1.6	18	1	0.11	1.6	54	3.9	90	26	1.1
	7	1	0.05	0.7	50	3.8	79	33	1.3	19	1	0.12	1.9	60	4.3	92	28	1.5
	8	1	0.12	1.9	64	4.6	97	28	1.4	20	1	0.06	1.7	59	4.2	90	27	1.2
	9	1	0.12	1.8	62	4.4	92	28	1.5	21	1	0.04	1.6	55	3.9	87	24	1.3
	10	1	0.11	1.5	65	4.6	92	28	1.4	22	1	0.06	1.6	58	4.2	90	27	1.2
	11	2	0.11	1.8	51	3.9	71	32	1.5	23	1	0.03	1.2	47	3.6	77	31	1.2
测试集	24	1	0.13	1.9	54	3.8	83	28	1.6	27	1	0.01	0.05	51	3.9	85	36	1.3
	25	1	0.13	2.1	63	4.5	94	28	1.5	28	1	0.06	0.77	51	3.8	83	34	1.4
	26	1	0.04	1.6	55	3.9	89	24	1.2	29	2	0.07	0.49	53	4.0	76	28	1.6

2.2　PCA 主成分分析

为判断 PCA 主成分分析是否可行,首先须对数据进行相关性分析.由表 1 可知,原始数据中因子存在

量纲，由于量纲影响导致部分因子间的值数量级相差过大（如因子 X_1 和 X_5），对相关性分析结果产生影响，造成分析不准确[16].因此，为了消除量纲影响，首先对原始数据进行标准化处理，得到处理后的样本数据如表 2 所示（限于篇幅，表 2 只给出了部分原本数据）.

表 2　标准化处理后的部分样本数据

名称	编号	标签	X_1	X_2	X_3	X_4	X_5	X_6	X_7
测试集	0	1	−1.23	−1.87	−1.38	−0.94	−0.75	1.44	−1.35
	1	1	1.08	0.69	1.41	1.65	1.05	−0.26	−0.12
	2	1	1.34	1.76	1.22	1.33	1.29	−0.26	1.11
	3	1	1.34	1.58	1.60	1.33	1.29	−0.26	0.49
	…	…	…	…	…	…	…	…	…
	21	1	−1.23	0.14	−0.27	−0.61	0.21	−1.62	−0.74
	22	1	−0.72	0.14	0.29	0.36	0.57	−0.60	−1.35
	23	1	−1.49	−0.57	−1.76	−1.59	−0.98	0.76	−1.35
训练集	24	1	1.08	0.69	−0.45	−0.94	−0.27	−0.26	1.11
	25	1	1.08	1.05	1.22	1.33	1.05	−0.26	0.49
	…		…	…	…	…	…	…	…
	28	1	−0.72	−1.35	−1.01	−0.94	−0.27	1.78	−0.12
	29	1	−0.46	−1.85	−0.64	−0.29	−1.10	−0.26	1.11

通过对已消除量纲影响的样本数据（表 2）进行相关性分析[17]，得到相关系数矩阵如表 3 所示.

表 3　相关系数矩阵

相关性矩阵								
		X_1	X_2	X_3	X_4	X_5	X_6	X_7
相关性	X_1	1.00	0.74	0.64	0.60	0.34	−0.21	0.61
	X_2	0.74	1.00	0.65	0.53	0.44	−0.57	0.28
	X_3	0.64	0.65	1.00	0.96	0.82	−0.46	0.25
	X_4	0.60	0.53	0.96	1.00	0.76	−0.27	0.22
	X_5	0.34	0.44	0.82	0.76	1.00	−0.47	−0.12
	X_6	−0.21	−0.57	−0.46	−0.27	−0.47	1.00	0.09
	X_7	0.61	0.28	0.25	0.22	−0.12	0.09	1.00

由表 3 可知，隔水层厚度 X_3 与单位涌水量 X_4 之间的相关性系数为 0.96，而渗透系数 X_5 与隔水层厚度 X_3、单位涌水量 X_4 与渗透系数 X_5、取芯率 X_2 与断层分维值 X_1 之间的相关性系数分别为 0.82，0.76 与 0.74.结果表明，选取的 7 个突水因子之间具有较强的相关性，须对其进行主成分分析.

采用 PCA 对标准化后的数据（表 2）进行处理，得到了主成分分析的碎石图（图 1）及其分析结果（表 4）.

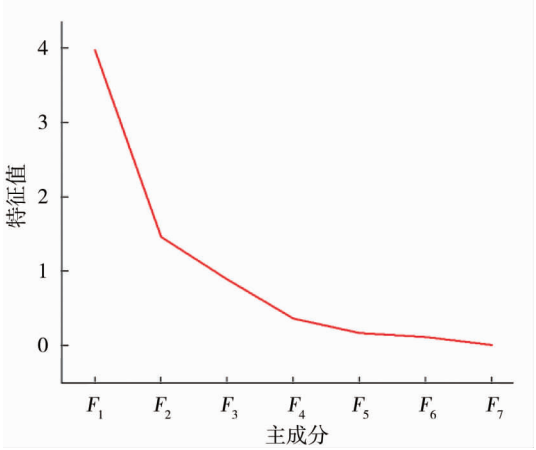

图 1　主成分分析碎石

表 4　主成分分析结果

主成分	初始特征值		
	总计	方差百分比/%	累计/%
F_1	3.971	56.727	56.727
F_2	1.464	20.922	77.649
F_3	0.893	12.759	90.408
F_4	0.367	5.247	95.655
F_5	0.173	2.474	98.129
F_6	0.119	1.700	99.829
F_7	0.012	0.171	100.000

由表 4 可知,前 3 个主成分 $F_1 \sim F_3$ 的累计贡献率为 90.4%,表明前 3 个主成分包含了预测所需要的绝大部分信息,可满足预测需求[11].因此,可将原数据的 7 个因子降维为 3 个主成分,3 个主成分因子荷载如表 5 所示.

表 5　主成分因子荷载

因子	主成分		
	F_1	F_2	F_3
X_1	0.781	0.516	−0.077
X_2	0.815	0.115	−0.431
X_3	0.954	−0.112	0.220
X_4	0.882	−0.076	0.409
X_5	0.771	−0.486	0.259
X_6	−0.550	0.430	0.647
X_7	0.334	0.863	−0.009

根据表 5,得到主成分 F_1,F_2,F_3 用数据标准化后的 7 个因子表示为

$$F_1 = 0.781X_1 + 0.815X_2 + 0.954X_3 + 0.882X_4 + 0.771X_5 - 0.55X_6 + 0.334X_7; \tag{3}$$

$$F_2 = 0.516X_1 + 0.115X_2 - 0.112X_3 - 0.076X_4 - 0.486X_5 + 0.430X_6 + 0.863X_7; \tag{4}$$

$$F_3 = -0.077X_1 - 0.431X_2 + 0.220X_3 + 0.409X_4 + 0.259X_5 + 0.647X_6 - 0.009X_7. \tag{5}$$

由式(3)、式(4)和式(5)得到采用 PCA 处理后的样本数据如表 6 所示.限于篇幅,表 6 只给出了部分样本数据.

表 6　主成分分析后的部分数据

名称	编号	标签	F_1	F_2	F_3
训练集	0	1	−6.454 85	−0.811 46	0.962 01
	1	1	5.110 12	−0.375 53	0.712 95
	…	…	…	…	…
	22	1	0.464 14	−2.116 64	−0.024 32
	23	1	−6.338 52	−0.880 23	−0.424 12
测试集	24	1	0.444 48	1.729 99	−1.108 03
	…	…	…	…	…
	29	2	−3.077 39	1.021 23	0.109 80

2.3　OPF 训练和测试

采用监督式 OPF 算法对表 6 的数据进行训练,得到了训练阶段的原型集样本(其包含的样本编号为 11,16,14,6,5,4, 13 和 21 这 8 个原型样本),并基于此原型集样本构建了最优路径森林.

其次,利用训练阶段所构建的最优路径森林对测试集中的每一个样本进行测试,得到的预测结果如表 7 所示.

表 7 OPF 预测结果与实际情况对比

样本	实际情况	预测结果	预测与实际是否符合
24	1(未突水)	1(未突水)	是
25	1(未突水)	1(未突水)	是
26	1(未突水)	1(未突水)	是
27	1(未突水)	1(未突水)	是
28	1(未突水)	1(未突水)	是
29	2(突水)	2(突水)	是

由表 7 可知,采用 PCA-OPF 模型得到的 6 个测试集样本的预测结果与实际情况相符.

3 结论

1) 采用 PCA 主成分分析法可将用于煤层底板突水预测的 7 个因子降维为 3 个主成分,3 个主成分既保留了原始数据的大部分信息以满足预测需求,同时又简化了 OPF 算法的数据结构,减少了训练和测试工作量.

2) 构建的 PCA-OPF 模型利用 PCA 对原始数据进行简化,并采用 OPF 算法进行训练和测试,训练和测试阶段与参数无关且不需进行参数寻优,可避免已有方法的局限性.

3)基于 PCA-OPF 模型的煤层底板突水预测结果表明,采用 PCA-OPF 模型得到的测试集中 6 个样本的预测结果与实际情况相符.

参考文献:

[1] 邵良杉,徐波.煤层底板突水危险性的 PNN 预测模型研究及应用[J].中国安全科学学报,2015,25(8):93-98.
[2] 李万军,杨家兵."下三带"理论和"P-h"临界曲线法预测底板突水[J].煤矿开采,2010,15(5):45-47.
[3] 施龙青,韩进.开采煤层底板"四带"划分理论与实践[J].中国矿业大学学报,2005,34(1):19-26.
[4] 李春元,张勇,左建平,等.深部开采砌体梁失稳扰动底板破坏力学行为及分区特征[J].煤炭学报,2019,44(5):1508-1520.
[5] 白峰青,王斌,刘猛.巨厚坚硬顶板工作面的底板破坏规律[J].矿业工程研究,2014,29(3):38-43.
[6] 王妍,姚多喜,鲁海峰.高水压作用下煤层底板隔水关键层弹性力学解[J].煤田地质与勘探,2019,47(1):127-132.
[7] 杨志磊,王向前,高召宁,等.承压水上采煤底板破坏流固耦合数值模拟[J].矿业工程研究,2012,27(4):61-65.
[8] 吴旋,来兴平,郭俊兵,等.综采面区段煤柱宽度的 PSO-SVM 预测模型[J].西安科技大学学报,2020,40(1):64-70.
[9] 万赟.从图灵测试到深度学习:人工智能 60 年[J].科技导报,2016,34(7):26-33.
[10] 曹庆奎,赵斐.基于模糊-支持向量机的煤层底板突水危险性评价[J].煤炭学报,2011,36(4):633-637.
[11] 施龙青,谭希鹏,王娟,等.基于 PCA_Fuzzy_PSO_SVC 的底板突水危险性评价[J].煤炭学报,2015,40(1):167-171.
[12] 施龙青,张荣遨,徐东晶,等.基于 GWO-Elman 神经网络的底板突水预测[J].煤炭学报,2020,45(7):2455-2463.
[13] 张风达,申宝宏.深部煤层底板突水危险性预测的 PSO_SVM 模型[J].煤炭科学技术,2018,46(7):61-67.
[14] 温廷新,于凤娥.基于 PSO_SVM_AdaBoost 的煤层底板突水预测研究[J].计算机应用研究,2018,35(12):3664-3667.
[15] 沈龙凤,宋万干,葛方振,等.最优路径森林分类算法综述[J].计算机应用研究,2018,35(1):7-12.
[16] 赵国彦,刘雷磊,王剑波,等.岩爆等级预测的 PCA-OPF 模型[J].矿冶工程,2019,39(4):1-5.
[17] 张文彤,董伟.SPSS 统计分析高级教程[M].北京: 高等教育出版社,2004: 213.
[18] PapaJ P, Falcão A X, SuzukiC T N. Supervised pattern classification based on optimum-path forest[J]. International Journal of Imaging Systems and Technology, 2009,19(2):120-131.
[19] Souza L A D, Afonso L C S, Ebigbo A, et al. Learning visual representations with optimum-path forest and its applications to Barrett's esophagus and adenocarcinoma diagnosis[J]. Neural Computing and Applications, 2019:1-17.
[20] Amorim W P, Rosa G H, Thomazella R, et al. Semi-supervised learning with connectivity-driven convolutional neural networks[J]. Pattern Recognition Letters, 2019,128:16-22.