计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(12):243−247 [doi: 10.15888/j.cnki.csa.007206]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于机器学习的煤矿突水预测方法①

童　柔, 谢天保

(西安理工大学 经济与管理学院, 西安 710054)
通讯作者: 童　柔, E-mail: 15934845412@163.com

摘　要: 由于影响煤矿突水的因素多、相关性强, 影响模型预测精度; 数据收集工作量大, 成本较高, 如何科学地选取特征以提高模型预测准确率成为本文重点研究内容. 本文首先提出采用稳定性选择方法在已知的 22 个影响因素中选取 7 个最重要的因素, 之后构建随机森林、神经网络以及支持向量机 3 种典型机器学习分类预测模型对特征选取前后的数据进行预测分析, 实验结果表明, 特征选取后的预测模型非常稳定且预测准确率可达 100%, 同时降低了样本数据收集成本.

关键词: 煤矿突水预测; 稳定性选择; 特征选取; 机器学习算法

引用格式: 童柔,谢天保.基于机器学习的煤矿突水预测方法.计算机系统应用,2019,28(12):243–247. http://www.c-s-a.org.cn/1003-3254/7206.html

Prediction Method of Coal Mine Water Inrush Based on Machine Learning

TONG Rou, XIE Tian-Bao

(Faculty of Economics and Management, Xi'an University of Technology, Xi'an 710054, China)

Abstract: Because there are many factors affecting coal mine water inrush and they have strong correlation, the prediction accuracy of the model will be affected. Due to the heavy workload and high cost of data collection, how to select features scientifically to improve the accuracy of model prediction has become the focus of this study. At first, this study uses stability selection to select 7 factors which are more important in 22 known influence factors, and then builds three typical machine learning classification forecasting models including random forest, neural network, and support vector machine (SVM) to forecast the data before and after feature selection. The experimental results show that the prediction model is very stable after the feature selection and prediction accuracy can reach 100%, and also decrease the cost of the sample data collection.

Key words: coal mine water inrush prediction; stability selection; feature selection; machine learning algorithm

1　引言

随着我国能源行业的不断发展, 煤矿突水问题越来越成为大众值得关注的问题, 这不仅关系煤矿工人生命安全, 也关系着国家人力财力损失. 据统计, 在 2006–2016 年期间, 我国煤矿突水事故共发生 440 起, 死亡 682 人[1], 因此为了减少突水事故的发生以及人员的伤亡, 对煤矿是否会突水进行提前的预测显得越发重要.

目前煤矿突水预测方法包括很多种, 大多以神经网络为主[2–6], 这种方法针对突水量定量预测误差较大, 针对是否突水定性预测时, 开关量阈值需要靠专家经验确定. 而在构建模型前进行关键因素的选择也是很重要的一个步骤, 文献[7]使用层析分析法的一致性检验与专家评分来进行特征的选取, 主观性强, 受限于专家的经验. 文献[4]使用主成分分析进行降维, 主成分降维后的特征相互独立, 在一定程度上可以提高模型预测精度, 但对特征重要性的分析不够明确, 难以应用于

① 基金项目: 西安市科技计划 (201805037YD15CG21(5))
Foundation item: Science and Technology Program of Xi'an Municipality (201805037YD15CG21(5))
收稿时间: 2019-05-18; 修改时间: 2019-06-21; 采用时间: 2019-07-01; csa 在线出版时间: 2019-12-10

现实问题中煤矿突水特征因素的选择. 文献[7,8]首先通过理论分析确定了含水层 (包括厚度、水压及富水性)、底板隔水层厚度、地质构造等相关因素, 并通过各特征数据统计分析说明, 这些特征的重要性, 然后建模预测, 然而煤矿突水是由煤矿生产开采过程中各种复杂的因素综合作用的结果, 各特征的独立分析并不能体现各特征相互作用, 相互影响的煤矿突水机理.

基于以上分析, 本文在理论分析 (煤矿构造条件、含水层条件、开采条件、岩性组合条件) 的基础上, 收集样本数据. 然后采用稳定性选择的特征方法针对数据样本进行分析, 以预测准确率为目标对煤矿突水影响因素进行筛选, 获取与之相关的关键因素, 之后采用 3 种典型机器学习算法, 随机森林、神经网络以及支持向量机分别进行煤矿突水预测模型的构建, 结合 3 种模型的预测结果确定最终结果, 以此验证特征选取后的预测模型的准确率以及稳定性, 通过实验可帮助煤矿工作人员减少数据收集的工作量以及收集成本, 并且提高突水预测精度.

2 煤矿突水影响因素关键特征选取

常用的特征选择方法主要分为 3 种: 过滤法、包装法和嵌入法[9], 根据不同的情境及目的所使用方法也不同, 而在本次试验中, 将采取包装法中的稳定性选择方法来进行特征的选取.

2.1 稳定性选择

Meinshausen N 等人在 2009 年提出了稳定性选择这种特征选取的方法[10], 并指出其并不是一种新的算法, 而是基于 Lasso 特征选择方法并对其进行加强和改进. 具体来说, 稳定性选择是一种基于二次抽样和选择算法相结合的特征选取方法, 选择算法可以是支持向量机 SVM 或者回归等算法, 而二次抽样意味着不是使用所有的数据一次性选择出最重要的特征, 而是抽取数据子集以及特征子集来运行选择算法, 不断重复, 最终可以计算出每个特征作为重要特征出现的频率, 即使用出现的次数除以子集被测试的次数, 将其看做每个特征的得分并作为特征筛选的依据. 最重要的特征若每次都被选到, 则它的得分会高达 1, 而最不重要的特征最终得分将会为 0. 大多数实验证明, 相对比于其他的特征选择方法, 稳定性选择是性能最好的方法之一.

2.2 数据准备

煤矿突水机理具有多样性, 是指在不同的地质及水文地质条件下, 采用破坏或水压破坏表现出不同的空间组合特征, 突水机理的多样性反映了地质及水文地质条件的变化, 煤矿突水是否突水受制于诸多因素的综合影响[11]. 在本次实验中, 我们通过查阅资料以及煤矿专家的帮助, 共取得了包括构造条件、含水层条件、开采条件、岩性组合条件 4 个方面的相关影响因素, 再加上突水征兆这个因素, 共获得 22 个与煤矿突水有关的影响因素以及其所对应数据数据类型, 如表 1 所示, 并且收集与表 1 中 22 个煤矿突水相关因素以及突水结果所对应的数据 1056 例.

表 1 煤矿突水相关因素及数据类型

构造条件	含水层条件	开采条件	岩性组合条件	其他
构造 x1(离散型) 陷落柱 x2(离散型) 陷落柱充水 x3(离散型) 断层 x4(离散型) 断层充水 x5(离散型) 断层落差 x6(连续型) 裂隙带 x7(离散型) 裂隙带充水 x8(离散型)	矿井充水含水层 x9(连续型) 含水层与工作面距离 x10(连续型) 含水层厚度 x11(连续型) 含水层水压 x12(连续型)	煤层倾角 x13(连续型) 采面面积 x14(连续型) 走向长度 x15(连续型) 采高 x16(连续型)	砂性岩段 x17(连续型) 泥性岩段 x18(连续型) 灰岩段 x19(连续型) 其他厚度 x20(连续型) 煤 x21(连续型)	突水征兆 x22(离散型)

2.3 基于稳定性选择的特征选取

在 Python 的 sklearn.liner_model 库中 Randomized LogisticRegression (以下缩写 RLR) 实现了稳定性选择, 因此根据 1.1 中的相关分析, 可以使用其作为特征选择的工具. 在 RLR 中, 稳定性选择的实现主要有以下步骤:

Step 1. 对初始数据进行二次抽样, 随机选取 k 个特征以及对应 m 行数据, 统计每个特征被选次数 N;

Step 2. 使用所选数据构建逻辑回归模型;

Step 3. 对模型进行 $L1$ 正则化, 稀疏化数据使大多数不重要特征的权重变为 0, 最终筛选最重要特征, 统计每个特征被选为最重要特征的次数 n;

Step 4. 继续进行 Step 1, RLR 算法默认共构建 200 个逻辑回归模型, 直到模型构建完成进行下一步;

Step 5. 计算每个特征被选为重要特征的频率, 即

进行稳定性选择之后各特征的得分, *score* = *n*/*N*, 通过 *scores_*属性来获取每个特征的得分, 获得高分的特征就是所需选择的重要特征.

在使用 RLR 算法进行建模时, 其正则化参数 *C* 会影响最终各个特征的 *score*, 为了获取合适的正则化参数, 我们在 $(10^{-2}, 10^{2})$ 区间内取了 100 个 *C* 值进行建模, 计算出每个模型中各特征的得分情况, 以此绘制了如图 1 所示的正则化参数 *C* 与各个特征 *score* 之间的关系.

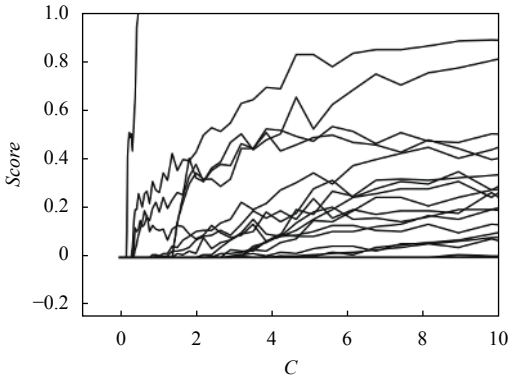


图 1 正则化参数 *C* 与 *score* 的关系

由图 1 中可以看出, 随着正则化参数 *C* 的减小, 即相当于正则化强度的增大, 各个特征的得分都将趋于 0, 而在 *C* 约为 0.18 时, 所有特征的得分都为 0, 也就是没有特征被选为重要特征, 本实验的目的是筛选出 7 个重要特征, 因此合适的 *C* 值是使得 7 个特征得分不为 0.

实验发现当 *C*=1.6681 时, 最重要的 7 个特征的得分不为 0, 其他特征得分都为 0, 因此使用此 *C* 值进行 RLR 建模进行特征选取, 将每个特征的得分从高到低进行排列后得到如下的表 2, 从表中很明显能看出来断层充水对煤矿是否突水影响最大, 得分为 1.0, 说明在每次进行特征选择时, 断层充水都会被选为最重要的特征, 在断层充水之后, 突水征兆对煤矿突水也有较大影响, 其次是裂隙带充水、陷落柱、陷落柱充水、含水层水压和裂隙带. 7 个特征选取出来后, 继续增加影响因素数量进行实验后发现, 预测准确率没有发生变化, 这说明可以继续增加因素的数量, 但因素数量过多会导致预测时计算成本过高, 收集数据时工作量变大, 因此此处仅选择 7 个最重要特征. 但此次实验结果仅针对已收集到的数据, 有新的数据增加时, 还需再进行实验来验证结果是否发生变化.

表 2 重要突水因素及其得分

突水因素	*Score*
断层充水	1.0
突水征兆	0.42
裂隙带充水	0.305
陷落柱	0.255
陷落柱充水	0.225
含水层水压	0.095
裂隙带	0.03

根据《矿区水文地质工程地质勘查规范》提出, 与煤矿突水最相关的两个因素为断层充水以及含水层水压, 文献[4]中根据 PCA 方法所筛选出的因素为断层、构造、含水层等, 文献[12–14]通过调查分析提出影响煤矿突水重要因素主要有断层、陷落柱、水压, 由此可以看出本次实验所筛选的关键因素符合煤矿突水的整体研究.

3 实验分析

经过 1.3 中的特征选取, 我们确定了与煤矿突水相关的 7 个重要的影响因素, 为了确定这 7 个因素对于预测结果的准确率是否有提高以及算法的稳定性, 本次实验使用随机森林、神经网络以及支持向量机 3 种典型机器学习算法构建煤矿突水预测模型, 使用特征选取前后的数据进行建模, 对比预测准确率.

3.1 数据处理

由表 1(见 2.1 节) 可以看出煤矿突水样本数据特征有离散型、连续型 2 种. 离散型数据的数据类型为逻辑数据 (以陷落柱充水为例, 若充水则为 1, 若无则为 0), 连续型数据则是使用浮点型数据类型来表示具体的数字. 因此在使用神经网络以及支持向量机进行建模前需要对数据进行处理, 对于离散型数据需使用独热编码, 而连续型数据由于不同特征的值大小差别太大, 因此需要进行标准化处理将数据缩放到相同的区间以提高准确率.

数据处理完成后, 使用 Python 中 *train_test_split* 函数随机将数据集划分为训练数据集和测试数据集, 函数的第 3 个参数 *test_size* 用来设定测试集数据的多少, 本实验将设定 *test_size*=0.3, 即 1056 例数据随机选取 70% 为训练集数据, 其余 30% 为测试集数据.

3.2 煤矿突水分类预测算法分析

(1) 随机森林是由许多 CART 二叉树所组成的预测模型, 其中每一棵二叉树是通过随机选取的特征以及

训练数据集建立的, 因此每一棵二叉树都是没有关联的并通过计算基尼指数来选择属性进行建立, 对于每一组测试数据, 每一个二叉树都进行预测, 最终通过投票机制得出最终的分类预测结果. 随机森林相当于一个特殊的集成学习算法, 它是由许多个弱的分类器即 CART 树所组成的一个强分类器, 因而预测准确率较高.

(2) 神经网络是一种有监督学习方法, 其结构分为输入层、隐藏层和输出层. 在进行训练的过程中, 主要分为 2 个部分: 正向传播和反向传播. 正向传播过程是在输入层进行特征属性的输入, 并设置神经元之间的权值, 通过若干隐藏层进行前向计算, 获得每个神经元的输出. 反向传播过程是将正向传播计算的结果与真实结果进行对比, 进而反向计算调整权值和误差, 反复进行调整, 提高模型准确率.

(3) 支持向量机 (SVM) 也是一种有监督的学习方法, 在二分类问题中使用较多, 对于线性可分问题, 支持向量机运用优化算法实现最大化分类间隔; 而对于非线性问题, 支持向量机通过适当的核函数将输入空间映射到高维空间, 实现高维空间线性可分, 将非线性问题转化线性问题[15], 然后在新空间中利用二次型寻优算法寻找一个最优超平面将两类样本分开, 保证分类准确率.

3.3 煤矿突水预测模型的构建

(1) 随机森林模型使用 RandomForestClassifier 来建立模型, 参数 *n_estimator* 设置为 100, 即随机森林中共建立 100 个决策树进行预测; 参数 *max_depth* 设置为 4, 即决策树深度最大为 4, 实验中对决策树进行可视化后发现, 当决策树深度为 4 时, 所有数据基本上已经分类完成, 且准确率高, 因此可将决策树深度减小来提高预测效率.

(2) 神经网络将使用 MLPClassifier 来建立模型, 参数 *hidden_layer_sizes* 设置为 (50, 50), 即设置两层隐藏层且每层神经元个数为 50, 在特征选择后其设置为 (15, 15), 这是因为输入层神经元个数不同, 隐藏层相应需要改变; 参数 *slover* 权重优化的求解器使用“lbfgs”, 它对于小型数据集可以更快的收敛并且分类表现更好.

(3) 支持向量机将使用 SVC 来建立模型, 惩罚参数 *C* 通过循环建模发现当 *C* 小于 0.4, 预测准确率将会降低, 因此设置为 *C*=0.4, 参数 kernel 通过实验得出使用‘linear’线性核函数分类更准确.

使用 3 种模型的 fit 方法分别对特征选取前后的训练数据进行训练, 之后使用训练好的模型对测试数据进行预测, 从而得到表 3 中的预测结果对比.

表 3 特征选取前后预测准确率对比

模型	特征选取前		特征选取后	
	训练集	测试集	训练集	测试集
随机森林	1.0	0.981	1.0	1.0
神经网络	1.0	0.982	1.0	0.982
支持向量机	1.0	1.0	1.0	1.0

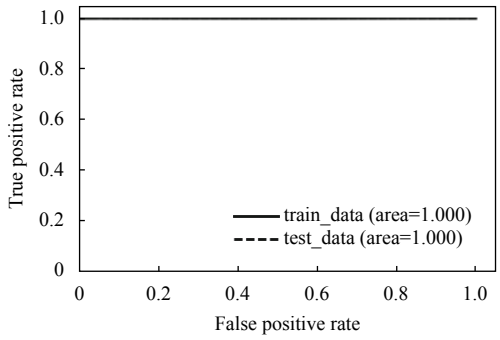
从表 3 中可以看出在特征选取前后 3 种模型预测准确率都很高, 随机森林模型在特征选取前后训练集准确率都高达 100%, 而测试集在特征选取后准确率提高至 100%, 神经网络模型和支持向量机模型在特征选取前后训练集以及测试集的准确率虽然都没有变化, 但整体准确率很高, 可以看出, 支持向量机模型是三者中最优的, 所有准确率都达到 100%. 随机森林在进行预测时, 由于在构建模型时已进行了剪枝, 从而减少了拟合且预测速度相对较快, 准确率较高; 使用神经网络进行分类时, 由于神经网络需进行反复调整权重, 因此其模型构建速率相对较慢, 使用神经网络模型预测的分类结果实际上是连续性的, 通过判断其是否大于 0.5 决定预测结果为 1 或 0, 即突水或不突水, 这种阈值判断的方法使得预测结果相对较差; 使用支持向量机进行模型构建时, 由于需要调整的参数较多, 因此在构建模型时时间较长, 但一旦选择正确的核函数之后, 其泛化能力会达到最佳, 预测准确率高.

在此基础上, 本实验继续使用交叉验证方法对特征选取的正确性进行检验, 使用 *cross_val_score* 方法对 3 种模型都进行交叉验证, 取参数 *cv*=15,即进行 15 轮的交叉验证, 取 15 次预测准确率的平均值, 实验结果如表 4 所示. 由表中可以看出, 特征选取后准确率有些许提高, 没有达到 100% 是因为 15 次交叉验证中 14 次预测准确率达到 1, 而仅有一次未达到 1, 由此可以看出利用选择后的特征建立预测模型准确率较高.

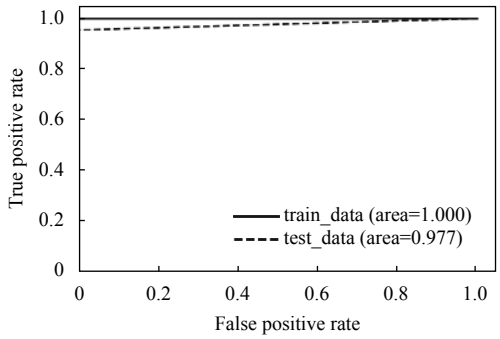
表 4 特征选取前后交叉验证结果

模型	特征选取前	特征选取后
随机森林	0.983	0.995
神经网络	0.989	0.989
支持向量机	0.995	0.995

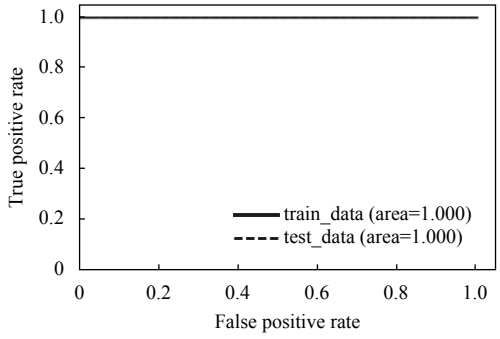
为了进一步确定特征选取后 3 种预测模型的稳定性, 我们绘制了 3 种模型的 ROC 曲线来评价模型性能, 如图 2 所示.



(a) 随机森林模型



(b) 神经网络模型



(c) 支持向量机模型

图 2 特征选取后各预测模型 ROC 曲线

从图中可以很明显看出, 图 2(a) 随机森林和图 2(c) 支持向量机模型都很稳定, 训练集以及测试集的真正类率和假正类率都达到最优, 而图 2(b) 神经网络模型相比下来略有不足, 但稳定性也算比较好.

由此可以看出在进行特征选择后, 预测模型仍然非常稳定, 因此可以证明在第 2 节中所筛选影响煤矿突水的预测模型都很稳定, 由此得出实验所筛选的关键因素在进行煤矿数据收集时是可进行参考的.

4 结论

本文通过稳定性选择的特征选取方法筛选影响煤矿突水结果的关键因素, 并通过随机森林等 3 种典型分类预测模型对选取前后预测准确率进行对比, 发现准确率都很高并且随机森林以及支持向量机模型达到 100%, 通过 ROC 曲线的绘制也可看出特征选取后的关键因素是可取的.

参考文献

1 刘艳亮. 2002–2016 年我国煤矿事故统计分析及预防措施. 陕西煤炭, 2018, (3): 64–67. [doi: 10.3969/j.issn.1671-749X.2018.03.016]

2 陶一明, 刘瑞英. 基于 BP 神经网络的煤矿突水预测系统的设计. 内蒙古煤炭经济, 2012, (12): 66–67. [doi: 10.3969/j.issn.1008-0155.2012.12.042]

3 胥良, 梁亚, 郭林. 在信息融合技术下的煤矿底板突水预测方法研究. 煤炭技术, 2015, 34(4): 190–192.

4 李培, 陈颖, 马小平, 等. 基于 PCA—ELM 的煤矿突水预测方法研究. 工矿自动化, 2013, 39(9): 46–50. [doi: 10.7526/j.issn.1671-251X.2013.09.013]

5 何风琴. 基于 PSO-WELM 模型的煤矿突水预测研究. 煤炭技术, 2017, 36(10): 124–126.

6 徐星, 孙光中, 田坤云. GA-BP 神经网络在煤矿突水水源判别中的应用. 煤炭技术, 2018, 37(10): 172–174.

7 徐星, 孙文标. 层次分析法在煤矿突水风险评价中的应用. 煤炭技术, 2016, 35(6): 143–144.

8 胡梦珂. 基于在线 SaE-ELM 的煤矿多等级突水预测方法研究[硕士学位论文]. 徐州: 中国矿业大学, 2015.

9 Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics, 2007, 23(19): 2507–2517. [doi: 10.1093/bioinformatics/btm344]

10 Meinshausen N, Bühlmann P. Stability selection. Journal of the Royal Statistical Society, 2010, 72(4): 417–473. [doi: 10.1111/j.1467-9868.2010.00740.x]

11 谢天保, 赵萌, 雷西玲. 基于非均衡样本集的煤矿突水预测模型. 计算机系统应用, 2018, 27(4): 124–130. [doi: 10.15888/j.cnki.csa.006298]

12 宋国娟. 基于极限学习机的煤矿突水预测及避险路线优化研究[硕士学位论文]. 徐州: 中国矿业大学, 2016.

13 祁春燕, 邱国庆, 张海荣. 底板突水预测模型的影响因素分析. 武汉大学学报·信息科学版, 2013, 38(2): 153–156, 247.

14 程爱平, 高永涛, 季毛伟, 等. 基于未确知测度理论的煤矿底板突水量预测. 金属矿山, 2014, (8): 157–161.

15 张松兰. 支持向量机的算法及应用综述. 江苏理工学院学报, 2016, 22(2): 14–17, 21. [doi: 10.3969/j.issn.1674-8522.2016.02.004]