

张良良,石永奎,李俊勇.基于混合核函数支持向量机的顶板砂岩富水性研究[J].矿业安全与环保,2018,45(2):72-76.  
文章编号:1008-4495(2018)02-0072-05

# 基于混合核函数支持向量机的顶板砂岩富水性研究

张良良<sup>1</sup>,石永奎<sup>1,2</sup>,李俊勇<sup>1</sup>

(1. 山东科技大学 矿业与安全工程学院,山东 青岛 266590;  
2. 山东省矿山灾害预防控制国家重点实验室培育基地,山东 青岛 266590)

**摘要:**为了寻求一种能够较好地预测煤层顶板砂岩富水性等级的方法,以桑树坪煤矿为例,分别采用BP神经网络、K最近邻分类法、决策树和支持向量机算法对其顶板砂岩富水性进行预测。比较发现,基于支持向量机的预测模型准确率最高为87.5%,节点错误率最低,优于其他3种模型。为了进一步提高模型预测准确率,建立了煤层顶板砂岩富水性的混合核函数支持向量机预测模型,当 $\lambda_1=0.05$ 与 $\lambda_2=0.95$ 时预测准确率达到100%。研究结果表明,以条件属性作为输入、决策属性作为输出的混合核函数支持向量机预测模型能较好地预测煤层顶板砂岩富水性等级,效果较好。

**关键词:**砂岩富水性;BP神经网络;朴素贝叶斯分类器;支持向量机;混合核函数  
**中图分类号:**TD745 **文献标志码:**A

## Study on Water-Richness of Roof Sandstone Based on Hybrid Kernel Function Support Vector Machine

ZHANG Liangliang<sup>1</sup>, SHI Yongkui<sup>1,2</sup>, LI Junyong<sup>1</sup>

(1. College of Mining and Safety Engineering, Shandong University of Science and Technology, Qingdao 266590, China;  
2. State Key Laboratory Breeding Base for Mine Disaster Prevention and Control in Shandong Province, Qingdao 266590, China)

**Abstract:** In order to find a better method to predict the level of sandstone water enrichment of coal roof, taking Sangshuping Coal Mine as an example, the BP neural network, K-nearest neighbor classification, decision tree and support vector machine algorithm were used to establish the level of sandstone water enrichment of coal roof. By comparison, the accuracy of prediction model based on SVM was 87.5%, the node error rate was the lowest, better than the other three models. In order to further improve the prediction accuracy, a predictive model of mixed kernel function support vector machine was established, and the prediction accuracy was 100% when  $\lambda_1 = 0.05$  and  $\lambda_2 = 0.95$ . The results showed that: the mixed kernel function with conditional attribute as input and decision attribute as output can predict the grade of sandstone water enrichment in coal seam roof, and the effect is good.

**Keywords:** sandstone water enrichment; BP neural network; naive bayes classifier; support vector machine; mixed kernel function

我国许多煤矿水文地质条件非常复杂,导致矿井突水事故时有发生,制约了煤矿企业的健康发展。矿井水害问题一直是煤炭行业的主要攻关课题之一<sup>[1-4]</sup>。文献[5]在单因素分析的基础上,应用模糊聚类方法对含水层富水性进行评价,但未能全面体现出砂岩富水性各指标的综合影响。文献[6]建立

了基于PNN的煤层顶板砂岩含水层富水性预测模型,预测结果和实际结果相符,证明该方法是可行的,但是预测样本较少,不能全面评价整个研究区的富水性。文献[7-8]指出了砂岩裂隙型含水层富水性分布范围,但不能定量得到研究区的富水性强弱。文献[9-10]采用了综合评价方法对煤层顶板岩石富水性进行预测,但考虑的指标和样本数量较少。尽管上述文献对煤层顶板砂岩富水性进行了研究探讨,并取得了大量成果,但研究方法较为单一,选取的指标和样本数量较少,不能全面评价煤层顶板砂岩富水性。因此,笔者运用BP神经网络、K最近邻

收稿日期:2017-08-07;2017-12-25 修订

作者简介:张良良(1990—),女,山东济宁人,硕士研究生,研究方向为系统优化理论与方法。E-mail:810881323@qq.com。

分类法、决策树及支持向量机等方法,以寻求一种能够判定新样本归属的最优法则,从而使之有效地判别煤层顶板砂岩富水性。

### 1 煤层顶板砂岩富水性的指标选取

桑树坪煤矿地质条件极其复杂,开采的6#煤层厚1.08~4.76 m,可采系数为97%。煤层结构简单,顶板以泥岩、砂岩为主。全区查明断层12条,基本查明断层14条,查出断层7条,推断断层1条,这些断层的存在破坏了顶板的连续性和完整性,为顶板砂岩水的赋存提供了良好的空间。为预测6#煤层顶板砂岩的富水性,将研究区进行了剖分,确定出煤层顶板砂岩富水性的评价指标如下:

- 1) 砂岩厚度。砂岩的厚度大的地段单位面积上储水空间大,砂岩厚度越大,顶板富水性越好。
- 2) 泥岩厚度。泥岩的孔隙度小,不利于富水;砂岩的孔隙度大,有利于富水。因此砂岩越厚泥岩越薄越有利于富水。
- 3) 浆液漏失量。揭示了地层内部裂隙发育程度,漏失量越高,说明裂隙发育程度越高。
- 4) 岩芯采取率。决定钻探过程中取芯的难易与完整程度,采取率越高,说明该段岩层完整性越好。
- 5) 断层强度。表示某一预测区域受断裂破坏的程度。
- 6) 断层密度。表示单位面积内发育的断层条数。

## 2 算法简介

### 2.1 BP神经网络

神经网络中使用最多的是BP神经网络(Back propagation neural network, BPNN)模型,通过反复检验训练样本来修正权值,使得权值最好稳定于最小值<sup>[13]</sup>。其输入层到隐含层、隐含层到输出层的关系如下:

$$\text{hidden}[j] = f\left(\sum_{i=1}^n w_{ij} a_i - \theta_j\right) \quad (1)$$

$$\text{out}[i] = f\left(\sum_{j=1}^n w_{ji} b_j - r_i\right) \quad (2)$$

式中: $a_i$ 为第*i*点输入; $b_j$ 为第*j*个隐含层节点输出; $w_{ij}$ 为输入层到隐含层的权值; $w_{ji}$ 为隐含层到输出层的权值; $r_i$ 为隐含层到输出层的权值; $\theta_j$ 为隐含层的阈值。

### 2.2 K最近邻分类法

K最近邻分类法(K-nearest neighbors, KNN)是

一种操作简单、有效并且在实践中应用较成功的分类方法<sup>[14]</sup>,K最近邻分类法公式表示为:

$$\text{score}(\mathbf{P}, C) = \sum \text{sim}(\mathbf{P}, \mathbf{P}_i) u(\mathbf{P}_i, C_i) \quad (3)$$

$$I = \text{argmax}[\text{score}(\mathbf{P}, C)] \quad (4)$$

式中: $\mathbf{P}$ 为样本相应的特征向量, $\mathbf{P} = (p_{i1}, p_{i2}, \dots, p_{im})$ ;  $u(\mathbf{P}_i, C_i)$ 为样本特征向量 $\mathbf{P}_i$ 关于样本类别 $C_i$ 的类别属性。

对所有样本类的权重之和进行比较,根据分类结果将样本归类到权重和最大的类别中。

### 2.3 决策树

决策树(Decision tree, DT)是一种常用于分析数据的算法,操作简单,使用方便<sup>[15]</sup>。该算法建立在奥卡姆剃刀理论基础上,阐述了一个信息熵的概念:

$$\text{Entropy}(S) = - \sum_{i=1}^m p_i \log_2 p_i \quad (5)$$

式中: $S$ 为训练集; $p_i(i=1, 2, \dots, m)$ 为具有*m*个类别标签的类别属性*C*在所有样本中出现的频率。

### 2.4 支持向量机

支持向量机(Support vector machine, SVM)是一种通过采用风险和置信范围的最小化原则,来提高学习机泛化能力的算法<sup>[16-19]</sup>。其具有全局优化能力强和泛化性能好等优点,使得SVM广泛应用,已经成为许多学者关注的热点。

由Kuhn-Tucker条件知,最优超平面充要条件是分类超平面满足以下关系:

$$\alpha_i^0 \{ [(x_i \cdot \omega_0) - b_0] y_i - 1 \} = 0, i = 1, 2, \dots, l \quad (6)$$

式中: $\omega_0$ 为分类线性权重阵; $b_0$ 为分类的阈值; $\alpha_i$ 为拉格朗日乘数因子; $y_i$ 为可分类的类别。

## 3 建立模型

### 3.1 样本采集

为预测6#煤层顶板砂岩的富水性,将研究区进行剖分,并统计了每个预测单元的砂岩厚度( $X_1$ )、泥岩厚度( $X_2$ )、浆液漏失量( $X_3$ )、岩芯采取率( $X_4$ )、断层强度( $X_5$ )和断层密度( $X_6$ )6个预测指标,选取具有代表性样本共30组,其中22组作为训练样本,其余8组作为待预测样本(编号加上标“\*”标出)。煤层顶板砂岩富水性( $R$ )分为3个等级,即较弱(I)、中等(II)和较强(III)3类,见表1。

### 3.2 数据预处理

由于原样本数据有不同的单位制,各指标在数值上可能相差几个数量级,直接使用原数据进行预

表1 煤层顶板砂岩富水性样本

序号	$X_1/m$	$X_2/m$	$X_3/(m^3 \cdot h^{-1})$	$X_4/%$	$X_5/MPa$	$X_6/(条 \cdot m^{-2})$	R
1	0.475	0.589	0.864	0.014	0.001 3	0.111	Ⅲ
2	0.704	0.398	0.201	0.215	0.000 3	0.111	Ⅲ
3	0.296	0.818	0.543	0.072	0.000 8	0.333	Ⅲ
4	0.521	0.171	0.718	0.052	0.001 1	0.111	Ⅲ
5	0.180	0.317	0.707	0.098	0	0	Ⅲ
6	0.181	0.560	0.753	0.340	0	0	Ⅲ
7	0.210	0.606	0.234	0.063	0	0	Ⅲ
8	0.516	0.237	0.065	0.624	0	0	Ⅱ
9	0.520	0.526	0.053	0.786	0	0	Ⅱ
10	0.553	0.616	0.036	0.823	0	0	Ⅱ
11	0.527	0.581	0.041	0.789	0	0	Ⅱ
12	0.580	0.248	0.029	0.829	0	0	Ⅱ
13	0.512	0.467	0.337	0.055	0	0	Ⅱ
14	0.516	0.700	0.236	0.061	0	0	Ⅱ
15	0.513	0.123	0.611	0.048	0	0	Ⅱ
16	0.303	0.859	0.007	0.960	0	0	Ⅰ
17	0.595	0.795	0.047	0.817	0.044 0	0.500	Ⅰ
18	0.786	0.602	0.035	0.326	0.046 0	0.500	Ⅰ
19	0.575	0.573	0	0.818	0.052 0	0.500	Ⅰ
20	0.209	0.149	0	1.000	0.002 0	0.500	Ⅰ
21	0.383	0.635	0.004	0.788	0.002 0	0.500	Ⅰ
22	0.530	0.364	0.027	0.629	0.004 0	0.500	Ⅰ
23*	0.425	0.648	0.132	0.239	0.000 2	0.111	待测
24*	0.445	0.585	0.244	0.151	0.000 4	0.222	待测
25*	0.502	0.192	0.062	0.613	0	0	待测
26*	0.518	0.711	0.047	0.662	0	0	待测
27*	0.526	0.238	0.025	0.829	0	0	待测
28*	0.408	0.185	0.005	0.932	0	0	待测
29*	0.253	0.600	0	1.000	0	0	待测
30*	0.309	0.610	0.007	0.960	0.002 0	0.500	待测

测可能丢失信息或引起数值计算的不稳定,因此需要对数据进行处理,以提高模型的预测精度。数据处理步骤如下:

1) 标准化

由于选取的煤层顶板砂岩富水性指标具有不同的量纲和量纲单位,这种情况会影响到数据的分析结果,因此需要进行数据标准化处理。根据以下公式将表1的数据进行标准化处理:

$$w_{ij} = \frac{s_{ij} - \min(s_j)}{\max(s_j) - \min(s_j)} \quad (6)$$

式中:  $s_{ij}$  是标准化前样本;  $w_{ij}$  是标准化后样本;  $\min(s_j)$  是原始样本中的最小值;  $\max(s_j)$  是原始样本中的最大值。

2) 离散化

离散化可以有效地克服数据中隐藏的缺陷,减弱极端值和异常值的影响,使模型结果更加稳定。基于熵的方法是最有效的离散化方法之一<sup>[20]</sup>,具体计算公式如下:

$$e_i = - \sum_{j=1}^k p_{ij} \log_2 p_{ij} \quad (7)$$

式中:  $e_i$  为离散化后的数值;  $k$  为不同的类标号数;  $p_{ij}$  为第  $i$  个区间中类  $j$  的概率,  $p_{ij} = m_{ij}/m_i$ ;  $m_i$  为所划分的第  $i$  个区间中值的个数;  $m_{ij}$  为区间  $i$  中类  $j$  的值的个数。

3) 属性约简

属性约简可以有效消除数据之间的信息冗余,减少样本的训练时间,提高样本的预测精度。离散化后的数据进行属性约简,发现第4组、第5组和第6组数据重复,第8组和第9组数据重复,第17组和第18组数据重复,所以删除第4组、第5组、第8组及第17组数据,剩下的18组数据为训练集,其余8组作为测试集。

3.3 相关性分析

将煤层顶板砂岩富水性评价指标,即砂岩厚度、泥岩厚度、浆液漏失量、岩芯采取率、断层强度及断层密度6个指标作为分类模型的输入向量,同时将煤层顶板砂岩富水性等级(较弱、中等、较强)作为分类模型的输出类别属性,各指标之间的关系如图1所示。

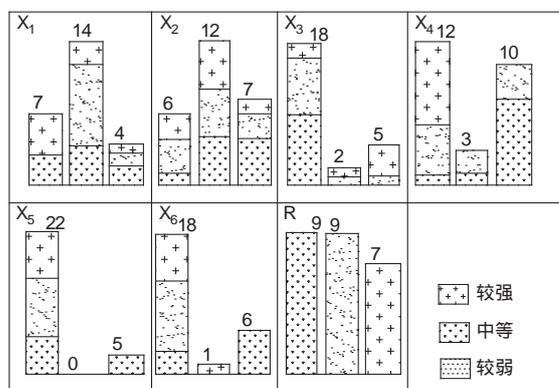


图1 煤层顶砂岩富水性评价指标之间的关系

由图1可见,煤层顶板富水性较弱与砂岩厚度(属性  $X_1$ )呈负相关性;与泥岩厚度(属性  $X_2$ )呈

正相关性;与浆液漏失量(属性  $X_3$ )呈负相关性;与岩芯采取率(属性  $X_4$ )呈正相关性;与断层强度(属性  $X_5$ )相关性不明显;与断层密度(属性  $X_6$ )呈正相关性。

煤层顶板富水性较弱—中等与砂岩厚度(属性  $X_1$ )呈负相关性;与泥岩厚度(属性  $X_2$ )呈负相关性;与浆液漏失量(属性  $X_3$ )呈正相关性;与岩芯采取率(属性  $X_4$ )呈负相关性;与断层强度(属性  $X_5$ )呈负相关性;与断层密度(属性  $X_6$ )呈负相关性。

煤层顶板富水性较弱—较强与砂岩厚度(属性  $X_1$ )呈负相关性;与泥岩厚度(属性  $X_2$ )呈负相关性;与浆液漏失量(属性  $X_3$ )呈负相关性;与岩芯采取率(属性  $X_4$ )呈正相关性;与断层强度(属性  $X_5$ )相关性不明显;与断层密度(属性  $X_6$ )相关性不明显。

### 3.4 建立机器学习算法预测模型

基于BP神经网络、K最近邻分类法、决策树和支持向量机4种模型,分别建立以砂岩厚度、泥岩厚度、浆液漏失量、岩芯采取率、断层强度、断层密度为输入向量的条件属性,煤层顶板砂岩富水性较弱、中等和较强为类别输出向量的决策属性,建立煤层顶板砂岩富水性预测模型。

通过Matlab软件编程实现BPNN预测模型:采用BP神经网络的结构为6-5-4,训练时根据预测误差调整网络权值和阈值;KNN预测模型:建立预测模型 predict(KNN.fit);DT预测模型:建立预测模型 predict(DT.fit);SVM预测模型:采用径向基(Radial basis function, RBF)核函数作为SVM核函数,建立预测模型 predict(SVM.fit)。

### 3.5 结果分析

将属性约简后的18组作为训练集,剩余8组作为测试集,利用建立的机器学习算法预测模型对测试集进行预测,预测结果见表2。

表2 煤层顶板砂岩富水性预测结果

序号	实际等级	各模型预测等级			
		BPNN	KNN	DT	SVM
1	较强	较强	较强	中等	较强
2	中等	中等	中等	中等	中等
3	中等	较弱	较强	中等	中等
4	较弱	较强	较弱	较强	中等
5	较强	较强	较强	较强	较强
6	中等	中等	中等	中等	中等
7	较弱	较弱	较弱	较弱	较弱
8	较弱	较弱	中等	中等	较弱

由表2可知, BP神经网络正确分类的实例是6个,错误的是2个,准确率为75.0%;KNN正确分类的实例是6个,错误的是2个,准确率为75.0%;DT正确分类的实例是5个,错误的是3个,准确率为62.5%;SVM正确分类的实例是7个,错误的是1个,准确率达到87.5%,高于上述其他3种模型。

各预测模型的节点错误率如表3所示。

表3 预测模型节点错误率

预测模型	平均绝对误差/%	均方根误差/%	相对误差绝对值/%	相对均方根误差/%
BPNN	7.87	17.68	34.16	36.29
KNN	15.57	29.66	33.09	50.52
DT	23.61	24.62	51.48	60.67
SVM	5.06	16.24	11.03	33.32

由表3得到各预测模型的误差大小:平均绝对误差,DT>KNN>BPNN>SVM;均方根误差, KNN>DT>BPNN>SVM;相对误差绝对值,DT>BPNN>KNN>SVM;相对均方根错误,DT>KNN>BPNN>SVM。

通过比较可知,SVM预测模型的节点错误率低于上述其他3种模型。综合考虑准确率和节点错误率,SVM的预测效果较好。

## 4 模型优化

SVM预测模型有较强的学习能力,但泛化能力较弱,预测准确率需要进一步提高。将学习能力较强的RBF核函数与泛化能力较强的Sigmoid核函数组合为混合核函数:

$$K_M(x, x_j) = \lambda_1 \tanh(\beta x^T x_j) + \lambda_2 \exp(-\gamma \|x - x_j\|^2) \quad (8)$$

式中:( $x, x_j$ )为混合核函数  $K_M$  的内积; $\lambda_1$  为Sigmoid核函数系数; $\lambda_2$  为RBF核函数系数; $\lambda_1 > 0, \lambda_2 > 0$ , 且  $\lambda_1 + \lambda_2 = 1$ ; $\beta$  为输出权重; $T$  为期望输出,  $T = H\beta$ , 其中  $H$  是隐层节点数的输出; $\gamma$  为核参数。

建立煤层顶板砂岩含水层富水性的混合核函数支持向量机预测模型,对测试集进行预测,结果见图2。

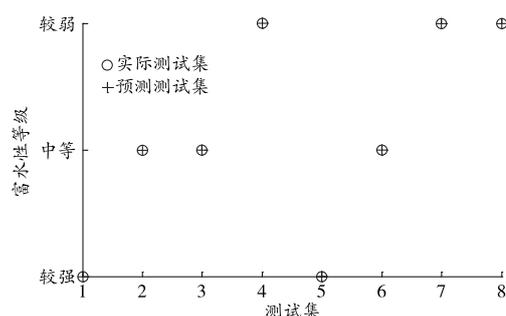


图2 模型优化后的煤层顶板砂岩富水性预测结果

运用 Matlab 软件运行得到混合核函数支持向量机预测模型的预测准确率随着  $\lambda_1$  的变化曲线, 见图 3。

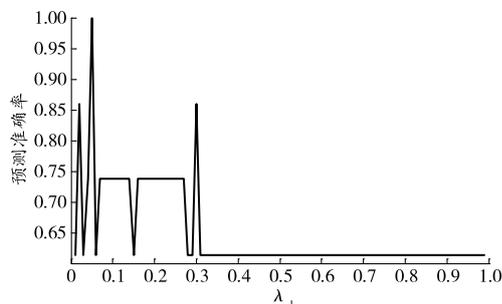


图3 预测结果准确率随着  $\lambda_1$  的变化曲线

由图 2 和图 3 可知, 当  $\lambda_1 = 0.05$ ,  $\lambda_2 = 1 - \lambda_1 = 0.95$  时, 说明混合核函数支持向量机预测模型的学习能力和泛化能力较强, 预测准确率达到 100%。

## 5 结语

以桑树坪煤矿为工程研究背景, 运用 Weka 平台, 选取 4 种算法均成功实现了对 8 组测试集的预测, 其中 SVM 预测准确率为 87.5%, 高于以上 BPNN、KNN 和 DT 预测模型; 在节点错误率方面, SVM 低于上述其他 3 种模型。综合考虑, SVM 的预测效果较好。为了提高预测准确率, 选择学习能力和泛化能力较强的混合核函数对 SVM 进行优化, 建立了煤层顶板砂岩含水层富水性的混合核函数支持向量机预测模型, 当  $\lambda_1 = 0.05$  与  $\lambda_2 = 0.95$  时预测准确率达到 100%, 效果较好。

## 参考文献:

[1] 熊晓英, 刘满才, 宋晓梅, 等. 潘谢矿区煤层顶板砂岩水害分析与防治措施[J]. 中国煤田地质, 2006, 18(1): 41-44.  
[2] 卜庆林, 陈成星, 杨成超, 等. 煤层顶板岩层富水性分区指标及其涌水量预测[J]. 山东科技大学学报(自然科学版), 2005, 24(3): 28-31.  
[3] 王尚令, 孙本魁, 洪荒. 恒源公司矿井 4 煤层顶板砂岩裂隙水突水分析及治理[J]. 矿业快报, 2005(11): 32-34.  
[4] 吴江. 分析煤层顶板砂岩含水层富水性单因素[J]. 中小企业管理与科技(上旬刊), 2014(5): 222.  
[5] 武旭仁, 魏久传, 尹会永, 等. 基于模糊聚类的顶板砂岩富水性预测研究——以龙固井田为例[J]. 山东科技大

学学报(自然科学版), 2011, 30(2): 14-18.

[6] 刘德民, 连会青, 韩永, 等. 基于概率神经网络的煤层顶板砂岩含水层富水性预测[J]. 煤炭技术, 2014, 33(9): 336-338.  
[7] 彭涛, 宣良瑞, 张海潮, 等. 卧龙湖煤矿砂岩含水层富水性预测及评价[J]. 煤矿安全, 2014, 45(8): 199-202.  
[8] 任智德, 吕玉广, 郑纲. 利用脆性岩石含量指数预测裂隙型含水层富水区[J]. 煤田地质与勘探, 2011, 39(4): 35-38.  
[9] 张文泉, 李加祥, 张红日, 等. 兴隆庄煤矿 3 煤顶板岩层富水性分区[J]. 煤田地质与勘探, 2000, 28(4): 53-56.  
[10] 崔雪丽. 综采顶板砂岩富水性多元信息预测模型与应用[D]. 徐州: 中国矿业大学, 2014.  
[11] WANG W, CHENG Y P, WANG H F, et al. Fracture failure analysis of hard - thick sandstone roof and its controlling effect on gas emission in underground ultra-thick coal extraction[J]. Eng-ineering Failure Analysis, 2015, 54(10): 150-162.  
[12] 卜庆林, 陈成星, 杨成超, 等. 煤层顶板岩层富水性分区指标及其涌水量预测[J]. 山东科技大学学报(自然科学版), 2005, 24(3): 28-31.  
[13] GHOLAM H. Mineralogy and geochemistry of the Jurassic coals from the Gheshlagh mine[J]. Eastern Alborz. Journal of Economic Geology, 2015, 7(2): 369-383.  
[14] 张庆国, 张宏伟, 张君玉. 一种基于 k 最近邻的快速文本分类方法[J]. 中国科学院研究生院学报, 2005(5): 554-559.  
[15] BUDAK C. A. Removal of impulse noise in digital images with naive bayes classifier method[J]. Turkish Journal of Electrical Engineering & Computer Sciences, 2016, 24(4): 2717-2729.  
[16] 冯少荣. 决策树算法的研究与改进[J]. 厦门大学学报(自然科学版), 2007(4): 496-500.  
[17] 张宏伟, 鹿广利, 徐路, 等. 支持向量机在煤与瓦斯突出预测中的应用[J]. 矿业安全与环保, 2013, 40(2): 55-58.  
[18] 王志辉, 舒服华. 改进的支持向量机在煤矿安全评价系统中的应用[J]. 矿业安全与环保, 2007, 34(1): 82-84.  
[19] DEMIDOVA L. Intellectual Approaches to Improvement of the Classification Decisions Quality on the Base of the SVM Classifier [J]. Procedia Computer Science, 2017, 103: 222-230.  
[20] 谭旭, 唐云岚, 陈英武. 基于粗糙集的区间型数据离散化算法[J]. 系统工程理论与实践, 2009, 29(6): 157-165.

(责任编辑: 逢锦伦)