Advances in Geosciences 地球科学前沿, 2019, 9(4), 230-237
Published Online April 2019 in Hans. http://www.hanspub.org/journal/ag
https://doi.org/10.12677/ag.2019.94026

Hans汉斯

Prediction of Organic Carbon Content Based on Rendezvous Graph and Support Vector Machine Regression

—A Case Study of Maokou Formation in Southeast Sichuan Province

Meng Zhang[1], Qi Wu[1], Miao Yu[1], Yukang Xiong[1], Kun Wang[2]

[1]School of Earth Sciences and Technology, Southwest Petroleum University, Chengdu Sichuan
[2]Sinopec Southern Exploration Company Exploration and Development Research Institute, Chengdu Sichuan
Email: 1019897345@qq.com

Received: Apr. 4th, 2019; accepted: Apr. 19th, 2019; published: Apr. 26th, 2019

Abstract

The southeast area of Sichuan Basin is a shallow-deep-water shelf sedimentary environment, where source rocks are developed and rich in organic matter. At first, chemical methods are used to analyze and determine the source rocks, and the source rocks are rich in organic matter. However, the evaluation results are difficult to meet the increasing demand for production, so the (TOC) calculation model of organic carbon content has been widely used as an effective identification method. In this paper, three log curves with high contribution rate are selected by cross plot method: acoustic time difference, natural gamma and deep lateral resistivity. Then three log curves are inputted to establish organic carbon support vector machine regression prediction model. The results show that the prediction effect is better when the organic carbon content is above 0.5, and the accuracy of the model needs to be improved when the organic carbon content is lower than 0.5.

Keywords

Source Rock, Organic Carbon, Sichuan Basin, Support Vector Machine, Crossplot Method

支持向量机在回归预测有机碳含量中的应用研究

——以川东南地区为例

张　萌[1]，吴　骐[1]，于　淼[1]，熊宇康[1]，王　昆[2]

[1]西南石油大学地球科学与技术学院，四川　成都

文章引用：张萌，吴骐，于淼，熊宇康，王昆．支持向量机在回归预测有机碳含量中的应用研究[J]．地球科学前沿，2019, 9(4): 230-237. DOI: 10.12677/ag.2019.94026

[2]中石化南方勘探公司勘探开发研究院，四川 成都
Email: 1019897345@qq.com

收稿日期：2019年4月4日；录用日期：2019年4月19日；发布日期：2019年4月26日

摘 要

四川盆地川东南地区为浅水－深水陆棚沉积环境，区内烃源岩发育，烃源岩富含大量有机质，最开始采用化学方法进行分析和判定，但是评价结果难以满足日益增长的生产需求，所以有机碳含量(TOC)计算模型作为一种有效的识别方式得到了广泛的应用。本文利用交会图法选出贡献率高的三条测井曲线：声波时差、自然伽马和深侧向电阻率，然后输入三条测井曲线值建立有机碳支持向量机回归预测模型。结果表明：有机碳含量在0.5以上时预测效果较好，当有机碳含量低于0.5时模型的精确度还需提高。

关键词

烃源岩，有机碳，四川盆地，支持向量机，交会图法

1. 引言

烃源岩最开始采用化学方法进行分析和判定，但是由于岩心样本数量少，分析难度大，耗时长，评价结果难以满足日益增长的生产需求，所以有机碳含量(TOC)计算模型作为一种有效的识别方式得到了广泛的应用。总有机碳含量(TOC)是判断烃源岩的一个重要参量，我们可通过 TOC 值来确定油气资源储量。常规识别烃源岩 TOC 的测井识别标准为“高自然伽马、高电阻率、高声波时差、高中子值、低密度值”[1] [2]。在对烃源岩的研究方面，早在 80 年代 Meyer 等就提出了利用侧向电阻率、密度等测井组合法进行识别烃源岩和对烃源岩有机碳(TOC)含量的计算；Herron 等使用 C/O 测井资料来计算地层 TOC 值进而识别烃源岩，但测试结果受仪器误差和操作影响使得理论上的 C/O 值与实验数据存在误差，其结果并不准确。Passey 等提出了用 ΔLgR 的方法判断有机碳含量，进而评价烃源岩，但是成熟度水平(LOM)来估算有机碳含量(TOC)，对于不同泥岩误差太大，需要矫正。本文从上述方法中得到参考再结合四川盆地的相关资料，综合分析上述缺点并对其进行改进，决定选用支持向量机的方法来测定 TOC 值。

2. 工区概况

川东南地区(图 1)位于扬子板块中南部、黔中隆起北的北部坳陷，雪峰山古陆的西部，川中古陆的东部，构造上属于川东弧形高陡褶皱带和川南帚状低陡褶皱带的一部分[3]。晚奥陶世，由于周边挤压作用，川东南地区为海盆的边缘；早志留世，为古隆起发育的高峰阶段，造山运动强烈，造成川中隆起、雪峰隆起的范围不断扩大，使得四川盆地的川东南地区沉积环境为有古隆起带包围起来的陆棚环境。因此，在整个晚奥陶世—早志留世时期，四川盆地川东南地区为浅水—深水陆棚沉积环境，区域上沉积了一套厚度较大的海相黑色富有机质泥页岩。该区茅口组由泥晶生物碎屑灰岩、泥晶灰岩组成，含燧石结核或薄层状硅质岩。区内烃源岩发育，油气资源丰富，生物碎屑大量存在，滑石化严重。

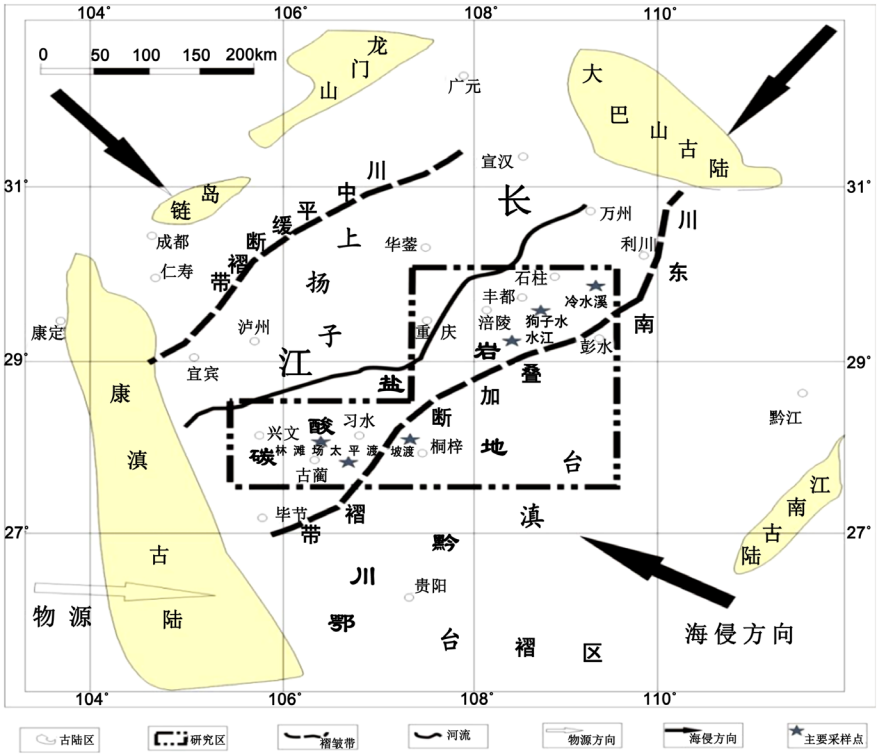

Figure 1. Overview of the study area
图 1. 研究区概况图

3. TOC 测井响应特征

总有机碳含量(TOC)是判断烃源岩的一个重要影响因素。因烃源岩富含大量有机质，所以其测井曲线在声波时差、电阻率、自然伽马、中子值和密度值等指标更为敏感，从铀、钍等元素含量测定其总自然伽马放射性强度为 136 API~200 API；有机质的存在使得岩类的许多指标发生变化，如电阻率的升高、声波时差的升高、岩石密度的减小等、其电阻率为 105 Ω·m~1015 Ω·m、声波时差大约为 571 μs/m、与普通泥岩相比烃源岩密度较小，与围岩相比烃源岩密度较大大约为 1.1~1.6 g/m^3，其中声波时差和密度受井壁和重矿物的影响较大；测得岩石干酪根含氢指数大约为 67.0% [4]。其测井响应特征可根据上述指标识别。烃源岩具有密度低和吸附性强的特点[5]。烃源岩对有机质吸附性强，比表面大，自然伽马和能谱测井曲线对这一特点有较高的辨识度，曲线表现为高值；而有机质烃的存在使得电阻率曲线高异常。并且表现为低密度曲线和高声波时差曲线则更好的分辨其他不含有机烃的泥岩，井壁和重矿物的影响。

4. 交会图法判别分析

4.1. 交会图法原理

交会图法是一种常用的岩性识别方法。将两种或多种数据通过在平面图上交会，其交会点的坐标可以比较大致地定出岩性变化的范围。在测井实际生产解释的过程中，已经有研究成果表明，对烃源岩敏感的测井曲线有自然伽马(GR)、伽马能谱(SL)、补偿中子测井(CNL)、深侧向电阻率(RLLD)、声波时差(AC)、密度(DEN)等，利用测井曲线对有机质的敏感程度不同是合理有效分辨烃源岩的资料基础。所以为了更好的对烃源岩的识别，通过对测井曲线与 TOC 含量的交会图法，得出对于烃源岩判别贡献率高的测井曲线，并根据判别结果建立 SVM 回顾预测模型。

4.2. 交会图法优选测井参数

利用交会图法分析对比烃源岩的相关测井曲线与有机碳(TOC)的联系，如图 2 所示。可以看到富含有机碳的泥岩层密度较小，声波时差相对较大；而普通泥岩层声波时差相对较小，密度较大。但密度受到黏土矿物和压实作用的影响较大，与有机碳含量相关程度并无太大关联。泥岩层易吸附有机质，有机碳含量较高吸附了高放射性的铀，所以在自然伽马上较容易分辨。对于中子测井而言，烃源岩中的大量氢聚集在页岩骨架和干酪根上，二者相互替换，所以中子曲线上不会存在较大差异，与有机碳含量相关程度低。泥岩层导电性较好，通常表现为低电阻，若含有表现为高电阻率的有机质，则会更加容易区分无关泥岩或低含量有机质泥岩，所以电法测井也是重要的参考标准之一，但是因为自然电位测井受到井径，地层等因素的影响相关程度数值波动大。所以通过对相关测井曲线的分析和交会图法识别程度，可以得到自然伽马，深侧向电阻率和声波时差对于含有机碳的烃源岩影响成分最大，对后续成分分析的回归预测具有重要意义。

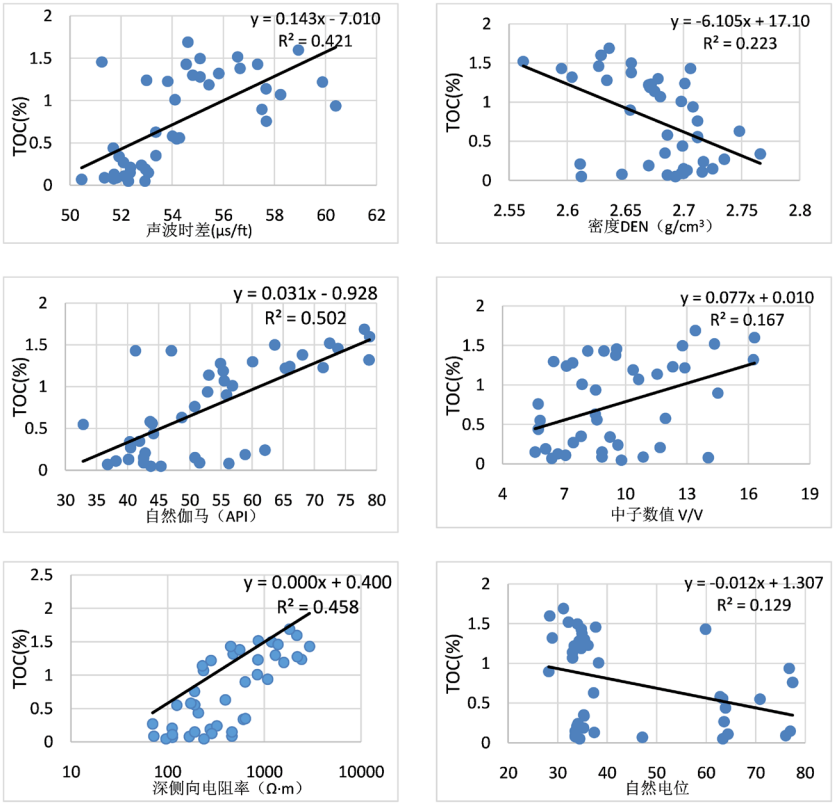

Figure 2. The preferred logging curve of the intersection chart method
图 2. 交会图法优选测井曲线

5. 交叉验证优选支持向量机参数

SVM 判别方法的准确性直接取决于核心函数参数 γ 和惩罚因子 C 的选择。核心函数 γ 能够有效地将高维度的映射转变成方便的点乘，核心函数还可以定义特征空间。选择恰当的核心函数，将研究数据映射到适当的样本空间中，才能将支持向量机性能提升到高值。惩罚因子 C 是用来权衡损失和分类间隔的权重，因子越大从而损失越大。假如不断提高因子数值，在一定程度上会实现将样本点完全正确的分类，但又会导致过度拟合使得泛化能力不足[6]。

通常情况下，通过交叉验证法的筛选可以得到最适合模型建立的相关参数[7]。在确定的数据样本中进行分组筛选，一部分作为训练集来求得模型，剩下留作测试集来对模型进行评价，并计算出测试集样本的实验误差作数据分析，从而得到准确稳定的相关模型[8]。

在具体的实验中通过程序随机地从 101 个样本数据中抽取 80 个作为训练集，其余的 21 个当作测试集。将测试数据归一化提高数据准确率后代入模型的建立，再运行程序可以得出训练集和测试集回归预测结果精度图，测试多次后寻找出最佳精度的参数 γ 和 C。算法中利用平均平方误差性能函数 mse 和拟合优度 R^2 来判断支持向量机回归预测有机碳含量的精确度。其中：

$$mse = \frac{1}{N}\sum_{T=1}^{N}(y_c - y)^2$$

得出 mse 为 0.017077，通过 mse 可以评价数据的变化程度，mse 的数值小，说明预测模型描述实验数据误差越小，更接近真实值。

R^2 是拟合优度，可以衡量回归方程整体的拟合程度，R^2 最大值为 1，R^2 的值越接近 1，说明回归直线对观测值的拟合程度越好，反之，R^2 的值越接近 0，拟合程度越差。

$$R^2 = \frac{\sum(y_c - \bar{y})^2}{\sum(y - \bar{y})^2} \text{或} R^2 = 1 - \frac{\sum(y - y_c)^2}{\sum(y - \bar{y})^2}$$

其中，$\sum(y-\bar{y})^2$ 为总偏差，$\sum(y_c-\bar{y})^2$ 为回归偏差，$\sum(y-y_c)^2$ 为剩余偏差。实验得到的 R^2 为 0.93064，说明拟合程度很好，误差很小。

通过交会图法优选测井曲线：声波时差、深侧向电阻率、自然伽马。并将以上三条曲线的数值带入支持向量机中，采用 5 折交叉验证的方法优选支持向量机的参数。最终得到最优参数 $C = 8.5635, \gamma = 0.5267$。选择出最佳的 γ 和 C 参数对后续建立回归预测模型有重要意义。

6. 建立 SVM 回归预测模型

6.1. 支持向量机与回归预测原理

支持向量机(SVM)是基于结构风险最小原理和 VC 维理论的新型机器学习方法[9]。通过寻求结构化风险最小化(SRM)来提高学习机泛化能力，实现经验风险和置信范围的最小化，从而达到在统计样本量较少的情况下，亦能获得较好统计规律的目的。支持向量机方法结构简单、适应性好、全局最优、训练速度快和泛化能力强等优点[10]。

支持向量机回归预测分析主要是分为对线性可分和线性不可分两种情况的进一步讨论[11]。根据需要嵌入到更高维度的空间中，如果该函数是一个完全分离样本的线性函数，这意味着样本是线性可分的，反之则为线性不可分。在各支持向量机算法中，不敏感函数和核函数算法是支持向量机回归预测的主要算法[12]。

6.1.1. 线性可分样本集

线性可分支持向量机处理的是严格线性可分的数据集。其分类超平面为[11] [12]：

$$w^* \cdot x + b^* = 0 \tag{1}$$

相应的决策函数为：

$$f(x) = sign(w^* \cdot x + b^*) \text{或} f(x) = sign\left(\sum_{i=1}^{N} a_i^* y_i \langle x_i, x\rangle + b^*\right) \tag{2}$$

其学习的优化问题为：

$$\min_{w,b} = \frac{1}{2}\|w\|^2 \tag{3}$$

$$s.t.\ y_i\left(w\cdot x_i+b\right)-1\geq 0, i=1,\cdots,N$$

6.1.2. 非线性可分样本集

在上述可分样本集增加一个松弛变量。其学习的优化问题为[11] [12]：

$$\min_{\omega,b,\xi}\frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{N}\xi_i \tag{4}$$

$$s.t.\ y_i\left(\omega_i+b\right)\geq 1-\xi_i; i=1,2,\cdots,N\ \ \xi_i\geq 0; i=1,2,\cdots,N$$

再引入核函数。分类决策函数变为：

$$f\left(x\right)=sign\left(\sum_{i-1}^{N}a_i^*y_ik\left(x_i,x\right)+b^*\right) \tag{5}$$

6.2. 支持向量机模型构建

$$\begin{aligned}f\left(x\right)&=\omega\cdot\phi\left(x\right)+b\\&=\sum_{i=1}^{n}\left(a_i^*-a_i\right)K\left(x_i\cdot x\right)+b^*\end{aligned} \tag{6}$$

式中：$f\left(x\right)$——预测函数；

ω——权数；

$\phi\left(x\right)$——非线性映射函数集合；

b——阈值；

a_i^*、a_i、b——可通过某点数值计算得到的模型参数 $K\left(x_i\cdot x\right)$ 为核函数且满足下式：

$$K\left(x_i\cdot x\right)=\exp\left(1-\frac{\left|x-x_i\right|^2}{\sigma^2}\right) \tag{7}$$

6.3. 训练集和测试集回归预测结果精度图

使用最优参数 $C=8.5635, \gamma=0.5267$ 进行支持向量机模型训练，再用测试集测试模型的分类识别效果，导出结果如下：

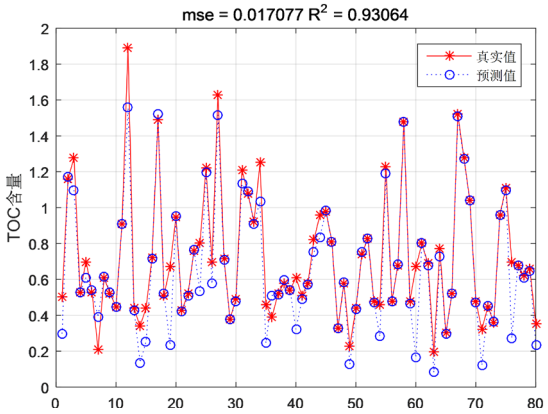

Figure 3. Comparison graph of training set results of SVM prediction model
图 3. SVM 预测模型训练集结果对比图

由图 3 可知，训练集回归预测结果拟合精度达到了 0.93064，平均平方误差只有 0.017077。TOC 含

量在 0.38%~1.56%之间拟合效果较好，预测值绝大部分在拟合曲线上近似于真实值；TOC 含量在 0.08%~0.38%之间效果不是很好，预测值和真实值之间存在误差。

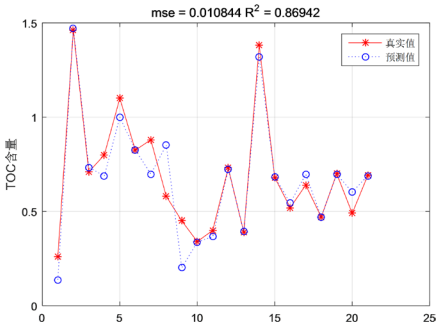

Figure 4. Comparison diagram of test set results of SVM prediction model
图 4. SVM 预测模型测试集结果对比图

由图 4 可知，测试集回归预测结果拟合精度达到了 0.86942，平均平方误差只有 0.010844。TOC 含量在 0.34%~1.47%之间拟合效果较好，预测值绝大部分在拟合曲线上近似于真实值；TOC 含量在 0.14%~0.34%之间效果不是很好，预测值和真实值之间存在误差。通过两个预测结果的对比分析可以看出，预测值和真实值在 TOC 含量小范围内存在着一定误差，程序的辨识度还有待于进一步的提高。

7. 应用实例分析

为了验证上述模型对川东南地区有机碳含量拟合的正确度，实验中选取一口该地区有机碳含量显示较好的井测试模型的精确性。XX 井在井段 1316 m~1370 m 内测试了该区域的有机碳含量，因此在实验中选取本井段作为测试井段，实验中每隔 20 cm 取一个点，总共选取了 240 个点作为测试点。将测试点对应的自然伽马，深侧向电阻率和声波时差三条测井曲线值导入模型进行有机碳拟合，然后将实地测试的有机碳含量与支持向量机拟合的有机碳含量进行对比，结果如图 5 所示。

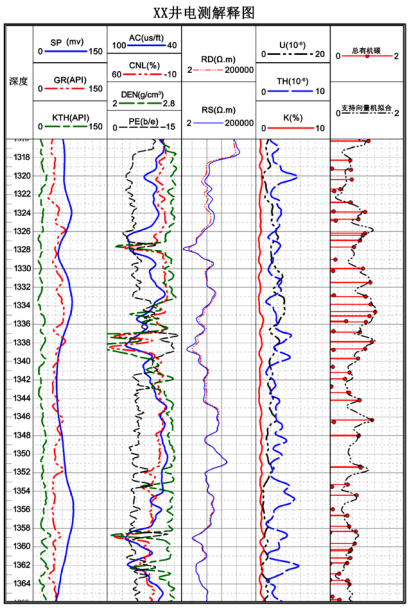

Figure 5. Comparison of results of organic carbon content in well XX
图 5. XX 井有机碳含量结果对比图

由图 5 可知，当有机碳含量在 0.5 以上时，支持向量机拟合的有机碳含量和实地测试的有机碳含量符合度较好，当有机碳含量大于 1 时，拟合程度最好。但是当有机碳含量小于 0.5 时，支持向量机拟合的有机碳含量较之实地测试的有机碳含量要高，拟合程度不是很好，与先前训练、测试支持向量机模型的结果一致。因此，想要使该模型在有机碳含量拟合方面有更好的推广价值，还需对支持向量机模型的相关程序进行进一步的改进，使其在有机碳含量较低时也有较好的拟合精确度。

8. 结论

以川东南地区茅口组一段为例，根据该地测井数据，利用交会图法选出贡献率高的三条测井曲线：声波时差、自然伽马和深侧向电阻率。将训练集和测试集通过支持向量机得到最优参数 $C=8.5635,\gamma=0.5267$，从建立的支持向量机回归预测模型测试结果中得出结论：

1) 训练集与测试集的测试结果相差不大，且拟合度普遍较高，平均平方误差小，但对于含量偏低的数据测试不佳，测试模型还有所欠缺。

2) 经过实验测试，利用支持向量机拟合一定范围内的 TOC 值有较高的准确率，支持向量机不失为探测 TOC 值的新方法。

基金项目

国家自然科学基金项目“四川盆地油钾兼探的地球物理评价方法研究”，编号“41372103”、“国家重点研发计划课题”，编号“2017YFC0602804”和“四川盆地深层钾盐勘探开发评价研究”，编号“2019YJ0312”联合资助。

参考文献

[1] 曲彦胜, 钟宁宁, 刘岩, 等. 烃源岩有机质丰度的测井计算方法及影响因素探讨[J]. 岩性油气藏, 2011, 23(2): 80-84+99.

[2] 许晓宏, 黄海平, 卢松年. 测井资料与烃源岩有机碳含量的定量关系研究[J]. 江汉石油学院学报, 1998(3): 11-15.

[3] 朱志军, 陈洪德. 川东南地区早志留世晚期沉积特征及沉积模式分析[J]. 中国地质, 2012, 39(1): 64-76.

[4] 杨涛涛, 范国章, 吕福亮, 等. 烃源岩测井响应特征及识别评价方法[J]. 天然气地球科学, 2013, 24(2): 414-422.

[5] 魏文文, 周大宇. 优质烃源岩识别标志与控制因素[J]. 内蒙古石油化工, 2010, 36(17): 10-11.

[6] 周国清. 应用 MATLAB 软件处理曲线拟合[J]. 重庆职业技术学院学报, 2003, 2(1): 38-39.

[7] 高艳芳, 陈实, 冯斌. 交叉验证在离散数据网格化时的应用[J]. 物探化探计算技术, 2012, 34(5): 619-621.

[8] 王怀亮. 交叉验证在数据建模模型选择中的应用[J]. 商业经济, 2011(10): 20-21.

[9] 彭涛, 张翔. 支持向量机及其在石油勘探开发中的应用综述[J]. 勘探地球物理进展, 2007, 30(2): 91-95.

[10] 陈科贵, 吴刘磊, 陈愿愿, 等. 基于支持向量机的川中杂卤石分类识别研究[J]. 地球科学进展, 2016, 31(10): 1041-1046.

[11] 陈金凤. 支持向量机回归算法的研究与应用[D]: [硕士学位论文]. 无锡: 江南大学, 2008.

[12] 周卫国. 基于支持向量机的大坝基础注浆量预测模型研究[J]. 水利技术监督, 2018(6): 157-160.

知网检索的两种方式：

1. 打开知网页面 http://kns.cnki.net/kns/brief/result.aspx?dbPrefix=WWJD
   下拉列表框选择：[ISSN]，输入期刊 ISSN：2163-3967，即可查询
2. 打开知网首页 http://cnki.net/
   左侧“国际文献总库”进入，输入文章标题，即可查询

投稿请点击：http://www.hanspub.org/Submission.aspx
期刊邮箱：ag@hanspub.org