SVM 与 BP 神经网络在石煤提钒行业清洁生产评价中的对比研究

李　佳,刘振宇

(中南民族大学　资源与环境工程学院,武汉 430074)

摘　要　为比较 BP 神经网络(ANN)和支持向量机方法(SVM)两种机器学习方法对清洁生产的评价能力,以理论原理为基础,比较了两种机器学习算法在应用原理方面的差异.并以石煤提钒生产工艺中水浸工艺为对象,对 BP 神经网络和支持向量机在清洁生产水平评价上进行了对比研究.结果表明:支持向量机方法分类精度为 100%;BP 神经网络为 90%但易陷入局部最优,因此支持向量机方法在解决小样本评价问题时具有较高的实用价值.

关键词　清洁生产;石煤提钒;支持向量机;BP 神经网络;评价方法

中图分类号　X38　**文献标识码**　A　**文章编号**　1672-4321(2018)04-0018-04

Clean Production Evaluation of Vanadium Extraction from Stone Coal by SVM and BP Neural Network: A Comparative Study

Li Jia, *Liu Zhengyu*

(College of Resource and Environmental Engineering, South-Central University for Nationalities, Wuhan 430074, China)

Abstract　In order to investigate the clean production evaluation performance of two machine learning methods, BP artificial neural net (ANN) and support vector machine (SVM), the differences between the two machine learning algorithms in application principle were compared and analyzed based on the theoretical principles. According to the water leaching process in vanadium extraction from stone coal, the performances of BP-ANN and SVM were comparatively analyzed in terms of clean production assessment. The results demonstrated that the classification accuracy of SVM reached 100%; while BP-ANN could reach 90% but was easy to fall into local optimum. So SVM method is more practical for the assessment of small samples.

Keywords　cleaner production; extraction vanadium from stone coal; SVM; BP-ANN; assessment methods

随着人工智能技术的发展,以神经网络(ANN)、模糊数学法(Fuzzy)、贝叶斯分类(Bayesian)、灰色预测模型为代表的数据挖掘技术方法成为一种非常有效的机器学习分类评价手段[1].对清洁生产评价而言,鉴于生命周期评价原则是整个清洁生产评价活动的核心依据,需据此来构建评价指标体系,并制定评估流程.本次评估对象生产流程较长,在指标体系设置及选择上较为复杂,具有模糊性、非线性、高噪声、小样本等特征[2];传统评价方法虽操作简便,应用较广泛,但由于主观性较强,评价结果存在不准确性[3,4].支持向量机方法(SVM)针对上述问题具有优势,属于分类算法的范畴,通过寻求最小结构划风险提升学习机泛化能力,能在少量样本下解决非线性及高维模式问题[5].已有学者将 ANN 运用到清洁生产工艺评价中,而 SVM 在该领域的应用尚未展开,因此,针对 SVM 在清洁生产评价领域的应用研究具有十分积极的意义.

目前我国石煤提钒行业存在问题较多,如难以有效提取钒资源、工艺设备落后、环境污染严重等,需要对该行业整体生产工艺技术制定和实施一套具有科学性、有效性的清洁生产评价方法,帮助企业

收稿日期　2017-11-24

作者简介　李　佳(1982-),女,讲师,博士,研究方向:环境管理和清洁生产,E-mail: jiajiali1982@aliyun.com

基金项目　湖北省自然科学基金资助项目(2016CFC772)

"自我评估,发现问题,制定方案"[6].因此,本文在对 BP 神经网络和 SVM 方法理论研究的基础上,以石煤提钒生产工艺中水浸工艺为对象,对比分析两种机器学习方法的异同及其应用.

1　研究方法

采用数据挖掘算法有助于从大量复杂的数据中找出数据分析模式,根据功能不同所挖掘的模式主要有描述型和预测型两种模式.BP 神经网络与 SVM 是其中典型分类预测算法模式,广泛应用于机器分类学习过程中,但 BP 神经网络与 SVM 的理论原理存在一定差异.

1.1　理论原理

BP 神经网络是基于传统统计学理论,遵循样本数目趋于无穷大时的渐近理论[7,8];SVM 是一种以统计学理论为主导的学习算法,常用于小样本机器学习规律数学框架与基本理论的研究,优势在于能很好地解决样本不足的问题.与传统神经网络等方法相比,SVM 以最小经验风险为约束条件,通过对经验风险的固定,使置信范围最小化,最小化结构风险[9].

1.2　应用原理

1.2.1　BP 神经网络

反向传播(BP)神经网络是人工神经网络中的最常见模型,原理是以误差反传误差为依托,进行学习方法的反向传播,通过训练样本对象的持续学习,不断调整不同层次间的阈值及连接权值,输入信号先后经过各隐层节点,最终实现由输入层节点向输出节点的转移,单层节点的输出仅同下一节点输出存在密切关联.假如输出层的输出达不到预期水平,将导致误差信号反向传播流程的逆向发展[10,11].在两个过程交替发展的过程中,误差函数梯度下降策略运行在有权向量空间上,动态迭代进行某组权向量的确定,获得最小化的网络误差函数,信息提取和储存任务随之完成.常见的 BP 神经网络模型共有自学习、输入输出、误差计算及作用函数共四种模型[8].

1.2.2　支持向量机

SVM 有着线性和非线性,在评价清洁生产的过程中,评价指标与清洁生产等级两者间存在着极为显著的非线性关联.本文探讨非线性 SVM.在非线性变换的支持下实现非线性问题向对应维度的线性问题的转变,通过变换过程达到分类超平面的最优化,借助核函数促进此种变换目标的达成.核函数同特定变换空间内的内积相等,也就是 $K(x_i,x_j)=\varphi(x_i)\cdot\varphi(x_j)$.“维数灾害”的问题通过 $K(x_i,x_j)$ 得到了妥善处理,其适用模式分类的理念是:在凸二次规划问题的计算过程中,在预先设定好非显现映射 φ 的支持下实现某一高维空间上的向量 x 映射,再在高维空间内计算分类超平面的最优解,使其能够尽量准确地划分两类数据点,且将划分好的数据点置于分类面的最大距离上[10].

1.3　建模数据样本生成

课题前期曾对石煤提钒工艺进行了深入研究,以 LCA 理论为基础,构建了石煤提钒工艺清洁生产评价指标体系[12],本文在前期所建立的水浸工艺清洁生产评价指标体系基础上,利用 BP 神经网络和支持向量机两种机器学习方法,对比研究两种方法在清洁生产水平评价上的应用.用随机方法[13]生成了标准清洁生产等级样本系列:

(1) 3 个清洁生产等级“清洁生产水平”、“一般水平”和“淘汰水平”分别对应清洁生产等级目标值 1,2,3;

(2) 利用均匀随机数在各评价等级每个指标变化区间内随机产生 20 个指标值;

(3) 在随机生成的 60 个样本系列中,对应每个生产等级共挑选 30 个样本构成检验集,检验集用于 SVM 和 BP 检验及对比.

1.4　数据预处理

基于数据变量间的差异,其数据级与量纲存在一定差异,必须通过归一方式进行训练样本的处理,强化指标范围的合理性,有效缓解数值差距.可通过下列公式进行规范化处理[14]:

$$x_{ij}^{*}=\frac{x_{ij}-x_{j\min}}{x_{j\max}-x_{j\min}},\qquad(1)$$

式中:x_{ij}^{*} 为训练样本 i 经归一化处理所得的 j 指标;x_{ij} 为未经归一化处理的原指标;$x_{j\max}$ 为训练样本 i 的最大值;$x_{j\min}$ 为训练样本 i 最小值.本文中 x_{ij} 指评价等级 i 上指标 j 的阈值.

2　评价模型建立

2.1　SVM 评价

分别利用 SVM 和 BP 神经网络两种方法对石煤提钒水浸工艺清洁生产水平进行评价,比较分析结果.使用 MATLAB7.8 为操作平台,SVM 选用 Libsvm 工具箱实现,BP 神经网络采用自己编制算

法程序.

2.1.1 分类器选择

SVM 算法从本质上属于两类分类器,而石煤提钒清洁生产评价结果由三个层次构成,通过 SVM 算法无法进行清洁生产的分类评价,故进一步构建了以二叉树为基础的三类分类系统(见图 1).

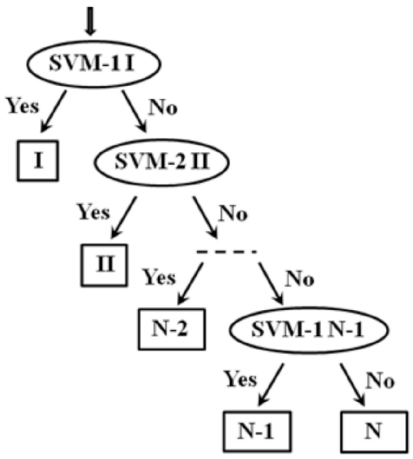

图 1 三类分类系统

Fig.1 Three kinds of classification system

2.1.2 核函数和参数

现阶段常见的 SVM 函数共 4 种,分别为 sigmoid 核函数、多项式核函数、线性核函数和径向基核函数[15].径向基核函数在某一参数的取值过程中所出现的特例是 sigmoid 核函数,它对于数值的要求相对较少;同多项式核函数对比,其参数量相对较少,还能很好地处理各分类问题,具有极为突出的适用性优势[17].本文在评价模型训练过程中选取了径向基核函数,见下式:

$$K(x_i, x_j) = \exp(-\gamma \| x_i - x_j \|^2), \gamma > 0. \quad (2)$$

确定核函数形式后,相应参数为惩罚因子 C 和核参数 g(上式中的 γ)函数参数的取值直接影响模型的分类效果,由 SVM 理论可知:若在未使用参数优化工具的情况下运用 SVM,则在参数选择过程中仅能运用试凑法的方式,此种方式规范性较弱.由于训练结果的准确性,大都需要不断重复试凑过程,少则数十、多则数百,且最终未必能获得最优化的 SVM.因此,本文选取网络搜索法(GS),对惩罚参数 C 和核参数 g 进行寻优.在设置 C 和 g 的搜索范围时,先进行粗略网格搜索,获得最佳参数位置,再在进行精细网格搜索,确定最终的参数值.具体步骤如下:

(1)以设定步长为依据,结合搜索方向,不断进行参数对的选择并进行校验验证,通过对比分析的方式重复进行各参数对的交叉验证,到网格搜索停止时结束,最终结合交叉验证率的大小选择最优化的参数.将参数 C 和 g 的搜索范围分别设为 $[\![,500]$ 和 $[\![,10]$;参数集 $C \in \{2^{-10}, 2^{-9}, \cdots 2^{10}\}$, $g \in \{2^{-10}, 2^{-9}, \cdots 2^{10}\}$,进行粗略网格搜索,确定达到最高分类效率的参数区间.

(2)以设定好的最高分类效率参数区间为依据,就 $C \in \{2^{-5}, 2^{-6}, \cdots 2^{-9}\}$, $g \in \{2^{0}, 2^{1}, \cdots 2^{4}\}$ 进行重新调整,并据此推进精细网格搜索,其结果参见图 2.分类效率最大化的平面体现为深色平面,当分类效率的结果最大时,该参数是评价模型的最优化参数,该研究过程中,分类效率达 100% 时,$C = 0.0038$, $g = 1.8$,评价模型随之形成.

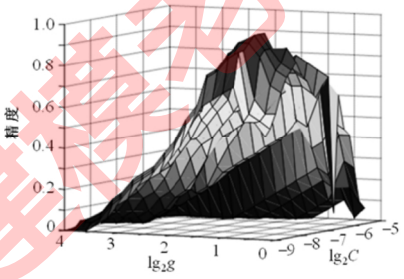

图 2 精细网格搜索结果(最佳:$C = 0.0038$, $g = 1.8$)

Fig.2 Results of fine grid search

综上设计,就 LIBSVM2.88 已有的径向基核函 SVM 网络参数的程序最优情况进行调整.确定最优化参数后,训练精度随之达到最优.图 3 为网络搜索法获得的 SVM 对训练样本和测试样本的检测结果.

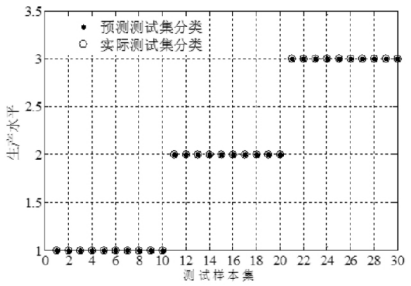

图 3 GS-SVM 的测试集实际分类和预测分类对比

Fig.3 Comparison of measured and predicted values of GS-SVM

2.2 BP 神经网络评价

该研究在函数编程计算过程中主要采用了 Matlab 神经网络工具箱,输入、输出层上的神经元分别有 29,1 个,隐层的神经元则有 20 个(该个数可自主测定,但必须控制在输入个数以内,且相对输出与输入综合的 50% 更大).当训练次数高达 30 次时,BP 网络的误差率仅为 10%,评价过程及结果如图 4 所示.

训练过程中,BP 网络会产生一定的“记忆”,该

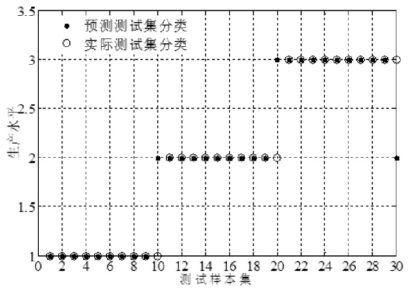

图 4　BP 神经网络评价过程及预测分类结果

Fig. 4　BP neural network evaluation process and the predicted classification results

算法不仅误差小,还在容错能力、泛化性等具有极强的优势,能够更好地满足大样本训练数据的检测要求.同 GS-SVM 对比,当样本数为 30 时,通过 BP 网络进行测试的准确率仅有 90%,代表以梯度下降法为根本的 BP 神经网络呈现出“过学习”状态,在小样本训练集较少的情况下,能很好地模拟出生产工艺现状,却无法有效地掌握评价结果的特征,使其在可推广性上存在一定劣势,证明小样本训练情况下,BP 神经网络所提出的评价模型在泛化性上的优势逐步弱化,局部极小、收敛速度不快等问题极为突出,且小样本训练结果难以满足预期条件.

3　结语

本研究分别利用 SVM 和 BP 神经网络建立了石煤提钒水浸工艺清洁生产评价模型,并进行了模型分类评价对比,结果表明: SVM 方法建模过程简单,能保证模型具有较好的泛化性能,在解决小样本即有限样本的评价问题时较 BP 神经网络有更好的适应性和推广性,不仅解决了整体工艺指标数据存在的数据不足问题,还降低了采集数据的评价成本,具有较好的准确性.而 BP 神经网络由于陷入局部最优而导致不能获得较为客观的结果.由于 SVM 不过分依赖样本数,因此,SVM 较 BP 神经网络更适合工艺清洁生产评价问题的研究,是一种具有较高实用价值的小样本评价方法.

参　考　文　献

[1] 郑明辉,吕经华. 基于机器学习的企业私有云用户行为分析模型 [J] .中南民族大学学报(自然科学版),2017,36(3):95-100.

[2] Li J, Zhang Y, Liu T. Research on pollution prevention and control technologies in the industry of vanadium extraction from stone coal [J] .IJEST,2014,17(1):83-96.

[3] Barbiroli G, Raggi A. A method for evaluating the overall technical and economic performance of environmental innovations in production cycles [J] .J Clean Prod, 2013, 11(4):365-374.

[4] 杜　栋,庞庆华.现代综合评价方法与案例精选 [M] .北京: 清华大学出版社,2008.

[5] 肖海军, 卢常景, 何　凡. 基于鸟群算法的 SVM 参数选择 [J] .中南民族大学学报(自然科学版),2017,36(3):90-94.

[6] 李　佳, 张一敏, 刘　涛. 石煤提钒行业清洁生产评价指标体系建立 [J] . 环境科学与技术, 2013,36(7):191-194.

[7] Jia L,Zhang Y,Tao L,et al.A methodology for assessing cleaner production in the vanadium extraction industry [J] . J Clean Prod, 2014, 84(1):598-605.

[8] Simonhaybn.神经网络原理 [M] .北京: 机械工业出版社,2004: 100-150.

[9] 杨道军, 王　冉, 沈　刚, 等.SVM 与 ANN 在湖泊富营养化评价中的对比研究 [J] .环境科学与技术, 2011,35(1):173-177.

[10] 张成成, 陈求稳, 徐　强, 等.基于支持向量机的太湖梅梁湾叶绿素 a 浓度预测模型 [J] .环境科学学报, 2013,33(10):2856-2861.

[11] Li J,Zhang Y,Du D,et al.Improvements in the decision making for Cleaner Production by data mining: Case study of vanadium extraction industry using weak acid leaching process [J] . J Clean Prod, 2017, 143: 582-597.

[12] 李　佳, 张一敏, 刘　涛. 石煤提钒行业清洁生产评价方法研究 [J] .环境科学与技术, 2013,36(8):200-205.

[13] 李正最, 谢悦波. 洞庭湖富营养化支持向量机评价模型研究 [J] .人民长江,2010,41(10):75-78.

[14] 毕温凯, 袁兴中, 唐清华, 等. 基于支持向量机的湖泊生态系统健康评价研究 [J] .环境科学学报, 2012, 32(8):1984-1990.

[15] 陈祖云,金　波,邬长福.支持向量机在环境空气质量评价中的应用 [J] .环境科学与技术, 2012, 35(61):395-398.

(责任编辑　刘　钊)