

文章编号:1673-5005(2012)03-0192-05

基于支持向量机的油气储量价值等级评价

杨磊¹, 王化增², 陈子凌³

(1. 中国石油大学 经济管理学院, 山东 青岛 266580; 2. 中石化胜利油田 东胜集团公司, 山东 东营 257091;
3. 中石油物资公司, 北京 100029)

摘要:针对油气储量的特点,对油气储量价值的优劣等级进行划分。选取影响油气储量价值等级的7个因素,即储量规模、储量丰度、储层埋深、原油黏度、渗透率、凝固点和采收率,采用最小二乘支持向量机模型对油气储量价值等级划分进行仿真,并运用网格搜索法确定最小二乘支持向量机模型的参数惩罚因子 C 和核函数参数 σ 。结果表明,最小二乘支持向量机是评价油气储量价值等级的有效方法,训练正判率达到95%,检验正判率达到81%。

关键词:最小二乘支持向量机; 油气储量; 价值分级

中图分类号:X 196 **文献标志码:**A doi:10.3969/j.issn.1673-5005.2012.03.033

Assessing value classification of oil and gas reserve based on support vector machine

YANG Lei¹, WANG Hua-zeng², CHEN Zi-ling³

(1. School of Economics & Management in China University of Petroleum, Qingdao 266580, China;
2. Shengli Oilfield Dongsheng Group of SINOPEC, Dongying 257091, China;
3. Materials Company of PetroChina, Beijing 100029, China)

Abstract: Based on the characteristics of oil and gas reserve, value classification of oil and gas reserve was assessed. Seven factors influencing value classification of oil and gas reserve were chosen, which were reserve scale, reserve abundance, reserve depth, oil viscosity, permeability, freezing point and recovery ratio. Least square support vector machine model was applied to simulate value degradation of oil and gas reserve. The parameters of penalty factor C and kernel function parameter σ can be decided by grid searching method. The results show that least square support vector machine is a valid method in the value classification of oil and gas reserve. The right rate of training is up to 95% and the right rate of testing is up to 81%.

Key words: least square support vector machine; oil and gas reserve; value degradation

油气储量具有经济性差异,可划分为经济储量、次经济储量和非经济储量,其中非经济储量可能占探明储量的比例较大。油气储量要实现资产化管理首先应确定其价值的优劣等级。针对油气储量价值等级评价问题前人已经做了一些研究。孟伟^[1]和周晓俊^[2]采用灰色多目标局势决策和模糊聚类理论对油气储量价值进行分级评价。袁自学等^[3]和胡健等^[4]通过一元回归分析、德尔菲法对油气储量价值进行分级评估。赵庆飞等^[5]运用层次分析法、德尔菲法对天然气储量价值进行分级。以上这些研

究多采用主观方法确定指标权重,忽视了指标数据本身对油气储量价值等级的影响。王化增和迟国泰^[6]运用BP神经网络的方法对油气储量价值等级进行了划分,该方法对油气储量价值优劣分级在参数合适和样本足够多的情况下能够得到较好的判别效果,但参数隐层个数和学习率的确定是一个难点,具体应用中这些参数需根据实际情况来确定。笔者采用最小二乘支持向量机法对油气储量价值分级进行仿真,将油气储量价值划分为优、良、中、差4个等级。

收稿日期:2011-12-29

基金项目:山东省自然科学基金项目(ZR2009HM010);中央高校基本科研业务费专项资金资助项目(09CX04085B;09CX05015B)

作者简介:杨磊(1974-),女(汉族),山东东营人,讲师,博士,研究方向为环境经济学。

1 最小二乘支持向量机基本原理

支持向量机(SVM)能较好地解决小样本、非线性、高维数和局部极小点等实际问题,已成为机器学习界的研究热点之一^[7]。最小二乘支持向量机(LS-SVM)是标准支持向量机的改进,该方法用等式约束代替标准支持向量机算法中的不等式约束,并将求解二次规划问题转化为直接求解线性方程组,降低了计算复杂度,在许多分类或回归估计问题中取得了很好的应用效果^[8]。

1.1 最小二乘支持向量机的基本模型

支持向量机的基本思想是,首先经非线性映射 $\phi(x):R^N \rightarrow H$ 把输入变量从原空间(R^N)映射到高维特征空间(H)中,然后在高维特征空间中建立优化超平面,并在该空间中进行线性分类。在线性情况下只用了原空间的点积运算,在非线性空间中也只须考虑在高维特征空间的内积运算,甚至不需要知道映射 ϕ 的具体形式^[9]。

设样本集 $S = \{(x_i, y_i) | (i = 1, 2, \dots, N), x_i \text{ 为输入变量, } y_i \text{ 为输出变量, } N \text{ 为样本个数}\}$ 。首先将输入变量从原空间映射到高维空间: $\phi(x) = (\phi_1(x), \phi_2(x), \dots, \phi_n(x), \dots)$ 。在高维特征空间中,如果训练数据集是线性可分的,则存在分类超平面。为了描述分类超平面,并考虑存在不能被分类超平面正确分类的样本,引入了松弛向量即误差量 $\varepsilon_i \geq 0$,超平面的约束条件为

$$y_i [w^T \cdot \phi(x_i) + b] - 1 + \varepsilon_i \geq 0.$$

式中, w 为超平面法线方向; b 为常数。

为了得到一个大边缘的最优超分类面,取训练集到超平面的最小距离最大。对于最小二乘支持向量机方法,高维特征空间中线性不可分相对应的优化问题转化为如下最优问题:

$$\begin{cases} \min \frac{1}{2} w^T w + C \left(\sum_{i=1}^N \varepsilon_i \right), \\ \text{s.t. } y_i [w^T \cdot \phi(x_i) + b] - 1 + \varepsilon_i \geq 0. \end{cases} \quad (1)$$

其中, C 为常数,控制错分样本惩罚的程度,实现在错分样本的比例与算法复杂度之间的折中,称为惩罚因子。要寻找最优化分类面,就是要求解式(1)所示的二次规划问题,找到唯一的极小点。最小二乘支持向量机优化目标采用 ε_i 的平方项,且不等式约束也变为如下等式约束:

$$\begin{cases} \min \frac{1}{2} w^T w + \frac{1}{2} C \left(\sum_{i=1}^N \varepsilon_i^2 \right), \\ \text{s.t. } y_i [w^T \cdot \phi(x_i) + b] = 1 - \varepsilon_i. \end{cases} \quad (2)$$

为了解决该二次规划问题,建立拉格朗日函数

$$L(w, b, \varepsilon, \alpha) = \frac{1}{2} w^T w + \frac{1}{2} C \left(\sum_{i=1}^N \varepsilon_i^2 \right) + \sum_{i=1}^N \alpha_i \{ y_i [w^T \cdot \phi(x_i) + b] - 1 + \varepsilon_i \}. \quad (3)$$

式中, α 为拉格朗日乘子。根据库恩-塔克最优条件有

$$\frac{\partial L}{\partial w} = 0, \frac{\partial L}{\partial b} = 0, \frac{\partial L}{\partial \varepsilon} = 0, \frac{\partial L}{\partial \alpha} = 0.$$

可得

$$\begin{cases} w - \sum_{i=1}^N \alpha_i y_i \phi(x_i) = 0, \\ \sum_{i=1}^N \alpha_i y_i = 0, \\ \alpha_i = C \varepsilon_i, \\ y_i [w^T \cdot \phi(x_i) + b] - 1 + \varepsilon_i = 0. \end{cases} \quad (4)$$

将式(4)消去 w 和 ε_i ,变为线性方程组,可用最小二乘法求得 b 和 α 。因此得到的分类决策函数即最小二乘支持向量机为

$$y(x) = \text{sgn} \left[\sum_{i=1}^N \alpha_i y_i K(x, x_i) + b \right].$$

其中

$$K(x, x_i) = \phi(x) \cdot \phi(x_i).$$

式中, $K(x, x_i)$ 为核函数。核函数将高维特征空间中内积运算转化为低维输入空间上一个简单的函数计算。选择不同的核函数就会得到不同的支持向量机。目前,在支持向量机中经常使用的核函数有以下3种:

$$\text{多项式内积函数 } K(x, x_i) = [(x, x_i) + 1]^q, \quad (5)$$

$$\text{径向基函数 } K(x_i, x_j) = \exp(-|x_i - x_j|^2 / \sigma^2), \quad (6)$$

$$\text{S型内积函数 } K(x, x_i) = \tanh(v(x - x_i) + c). \quad (7)$$

目前,还没有很好的方法来指导具体问题的核函数的选取。一般地,由于径向基核函数是个可适用于任意分布样本的普适函数,且仅有一个参数宽度系数 σ ,并能很好地反映模型选择的复杂度,因此被广为应用。如彭涛、杨斌^[9-10]等在分类识别时对采用不同核函数的结果做了对比分析,发现径向基核函数分类效果最好。本研究中也做了对比分析,最终选定了径向基函数作为核函数。

1.2 参数确定

核函数确定后,最小二乘支持向量机模型有两个参数需确定:惩罚因子 C 和核函数参数 σ 。惩罚因子 C 用于控制模型复杂度和逼近误差的折

中, C 越大则对数据的拟合程度越高, 学习机的复杂度就越高, 容易出现“过学习”的现象; 而 C 取值过小, 则对经验误差的惩罚小, 学习机的复杂度低, 就会出现“欠学习”的现象; 当 C 的取值大到一定程度时, 最小二乘支持向量机模型的复杂度将超过空间复杂度的最大范围, 那么 C 继续增大将几乎不会对最小二乘支持向量机的性能产生影响。径向基核函数参数 σ 的取值直接影响模型的分类精度。

目前, 还没有公认的方法来确定这两个参数。本文中采用网格搜索法^[10], 将 C 和 σ 分别作为网格的横向和纵向, 思路是首先将参数 C 和 σ 的选取范围设定得较大些, 设定较大步长进行支持向量机仿真, 比较选取不同参数数值时的训练正判率和检验正判率, 选取一组正判率最高的参数数值, 然后再以该数值为中心缩小参数范围, 同时参数选取的步长也适当地减小, 即通过这样的选取方法一步一步缩小参数选取的范围, 直到得到满意的训练正判率和检验正判率为止, 此时 C 和 σ 的数值就非常接近其最优数值。

2 油气储量价值分级的最小二乘支持向量机仿真

2.1 学习集的选取

利用最小二乘支持向量机方法进行油气储量价值等级划分的核心是建立一个训练样本集, 通过确定合适的核函数及参数, 对已知样本集进行学习训练, 最后确定最优决策函数, 并利用该函数进行价值等级评价和预测。

油气储量价值等级分级是在地质勘探、地质评价基础上, 根据已发现油气田油气资源条件对油气勘探开发经济效益的影响, 确定油气储量价值的优劣等级。油气储量价值 = 可采储量 × 油气价格 - 总开发投资 - 总经营成本。一定的市场条件下, 影响油气储量价值的因素有可采储量、开发投资和经营成本。依据油气储量资产特点和数据可得性, 本文中选取了7个影响油气储量价值优劣的因素, 即储量规模、储量丰度、储层埋深、渗透率、原油黏度、凝固点和采收率, 这7个影响因素作为油气储量价值分级支持向量机的输入; 输出为油气储量的4个价值等级, 即优、良、中、差, 分别用1、2、3、4来表示。本文中从中石化胜利油田搜集了99个已知价值等级的油气储量区块数据, 随机选取73个作为训练样本, 其余26个作为检验样本, 具体数据见表1和表

2, 建立影响油气储量价值等级因素的样本集, 即最小二乘支持向量机的学习集。

表1 学习原始数据及判别结果

Table 1 Training database and classification result

区块序号	储量规模 N/t	储量丰度 $T/(万t \cdot km^{-2})$	储层埋深 H/m	渗透率 $k/10^{-3} \mu m^2$	原油黏度 $\mu_o/(mPa \cdot s)$	凝固点 $T/^\circ C$	采收率 $E_R/\%$	价值判别结果
1	554	395.7	2050	1326	108	30	25.0	1 1
2	454	283.8	1905	323	259	38	46.0	1 1
3	542	135.5	2094	1018	51	27	46.0	1 2
4	235	635.1	999	3488	25000	12	30.0	2 2
5	282	134.3	875	3500	35000	16	41.0	2 2
6	366	135.6	1986	986	55	26	43.0	2 2
7	195	114.7	2050	1080	65	28	40.0	2 2
8	487	135.3	2275	533	1855	31	46.0	2 2
...	...	...	...	...	...	...	...	...
97	482	123.6	3000	61	26	32	12.0	4 4
98	136	138.8	875	3500	35000	15	40.0	2 2
99	58	25.2	2530	133	20	33	7.9	4 4

表2 支持向量机评价结果

Table 2 Assessment result of support vector machine

区块序号	储量规模 N/t	储量丰度 $T/(万t \cdot km^{-2})$	储层埋深 H/m	渗透率 $k/10^{-3} \mu m^2$	原油黏度 $\mu_o/(mPa \cdot s)$	凝固点 $T/^\circ C$	采收率 $E_R/\%$	价值判别结果
1	274	90.2	2050	152	135	36	18.0	2
2	50	38.7	2468	13	150	33	10.0	3
3	112	53.7	2460	13	150	33	10.0	3
4	203	92.3	1850	1500	531	-7	19.2	3
5	733	156.0	1400	500	2780	10	10.0	2
6	97	97.0	2500	24	10	32	15.1	3
7	674	132.2	2800	5	6	32	10.0	3
8	103	34.3	1900	122	306	32	10.0	3
...	...	...	...	...	...	...	...	...
57	7	23.3	2270	93	9	34	15.7	3
58	8	40.0	2450	93	12	34	15.0	3
59	58	52.7	2700	93	9	34	15.0	4

2.2 数据预处理

不同性质的指标对油气储量价值影响方向不同, 影响油气储量价值的正向指标有储量规模、储量丰度、渗透率和采收率, 逆向指标有储层埋深、原油黏度和凝固点。另外, 不同指标数据在数量级上有很大差异。因此, 有必要对原始数据做标准化处理, 将其化为 $[-1, 1]$ 之间的数。设 x_i 为原始数据, y_i 为标准化处理后的数据, 标准化方式如下:

$$\text{正向指标 } y_i = 2 \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} - 1, \quad (8)$$

$$\text{逆向指标 } y_i = 2 \frac{x_{\max} - x_i}{x_{\max} - x_{\min}} - 1. \quad (9)$$

2.3 参数 C 和 σ 的确定

在学习样本集确定后,支持向量机的参数惩罚因子 C 和核函数参数 σ 的寻求过程实质上就是预测模型的建立过程。网格搜索法确定参数的过程如下:

(1) 凭经验选定一组惩罚因子 C 和核参数 σ 的范围,如 C 取 $2^{-10} \sim 2^{15}$, σ 取 $2^{10} \sim 2^{-15}$;

(2) 设定搜索步长分别为 -1 和 1 ,训练支持向量机并开展检验,得到训练正判率和检验正判率最高的一组 C 和 σ 分别为 128 和 0.5 ;

(3) 重新选定 C 和 σ 的范围,将其范围缩小,设定 C 取 $2 \sim 200$, σ 取 $2^{-3} \sim 2^2$;

(4) 设定 C 的搜索步长为 1 , σ 的搜索步长为 0.1 ,训练支持向量机并开展检验,得出训练正判率和检验正判率最高的一组 C 和 σ ,分别为 133 和 0.57 ,对应的训练正判率为 95% ,检验正判率为 81% ,符合评价要求。因此,该 C 和 σ 的数值就是最小二乘支持向量机的最优参数。

2.4 油气储量价值分级支持向量机的训练与评价

惩罚因子 C 和核参数 σ 确定后,就可以构建出支持向量机模型,学习判别结果列在表1(其中,区块3为训练出错的区块)的最后一列。在训练样本的73个数据中,有4个出现误判,正判率达到 95% ,在检验样本的26个数据中,有5个出现误判,正判率达到 81% 。

用训练好的油气储量价值分级支持向量机模型对未知价值等级的59个油气储量区块进行评价,评价结果列在表2最后一列。结果显示,储量价值划分为良的区块有14个,占 24% ;中等的区块有42个,占 71% ;差的区块有3个,占 5% ;优级区块没有。

3 结论与讨论

(1) 最小二乘支持向量机可以应用于油气储量价值等级的划分,训练正判率达到 95% ,检验正判率达到 81% 。

(2) 网格搜索法是确定最小二乘支持向量机参数惩罚因子 C 和核参数 σ 的有效方法。最小二乘支持向量机模型中,参数的确定过程也就是向量机的训练过程。采用网格搜索法逐步缩小参数的搜索范围,使确定出的参数越来越逼近最优数值,简单实用。

(3) 在支持向量机学习集中油气储量价值优等的数据较少,只有3个,今后需要丰富价值优等的油

气储量区块数据,使支持向量机的学习样本更具有代表性。

参考文献:

- [1] 孟伟. 油气储量价值分级评价研究[J]. 新疆地质, 2000, 18(3): 43.
MENG Wei. The research of grade reserve value of oil and gas[J]. Xinjiang Geology, 2000, 18(3): 43.
- [2] 周晓俊. 油气储量价值分级评价与风险分析[D]. 天津: 天津大学管理学院, 2001.
ZHOU Xiao-jun. Graded appraisal of hydrocarbon reserve value and risk analysis[D]. Tianjin: School of Management, Tianjin University, 2001.
- [3] 袁自学, 郗君一. 油气储量资产评估方法和资产化管理探讨[M]. 北京: 石油工业出版社, 2000.
- [4] 胡健, 刘永爱, 李志学. 天然气储量的价值分级模型设计与检验[J]. 西北大学学报: 自然科学版, 2005, 35(4): 479-483.
HU Jian, LIU Yong-ai, LI Zhi-xue. A model design and checking of value classification of natural gas reserves[J]. Journal of Northwest University (Natural Science Edition), 2005, 35(4): 479-483.
- [5] 赵庆飞, 李丽娜, 徐向华, 等. 天然气储量价值评估模型[J]. 资源与产业, 2006, 8(2): 39-42.
ZHAO Qing-fei, LI Li-na, XU Xiang-hua, et al. Model of value evaluation of natural gas reserves[J]. Resources & Industry, 2006, 8(2): 39-42.
- [6] 王化增, 迟国泰, 程砚秋. 基于BP神经网络的油气储量价值等级划分[J]. 中国人口·资源与环境, 2010, 20(6): 41-46.
WANG Hua-zeng, CHI Guo-tai, CHENG Yan-qiu. Applying BP neural network to grade reserve value of oil and gas[J]. China Population, Resources and Environment, 2010, 20(6): 41-46.
- [7] 程爱辉, 高茂庭. 基于聚类的LS-SVM的入侵检测方法研究[J]. 网络安全技术与应用, 2010(5): 14-16.
CHENG Ai-hui, GAO Mao-ting. Research of the intrusion detection method based on clustering LS-SVM[J]. Network Security Technology and Application, 2010(5): 14-16.
- [8] 符杨, 张雷, 江玉蓉, 等. 基于可靠性数据分析和最小二乘支持向量机的电力变压器故障诊断[J]. 变压器, 2010, 47(9): 47-50.
FU Yang, ZHANG Lei, JIANG Yu-rong, et al. Power transformer fault diagnosis based on data reliability analysis and least squares support vector machine[J]. Transformer, 2010, 47(9): 47-50.
- [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.
- PENG Tao, ZHANG Xiang. Review of support vector machine and its applications in petroleum exploration and development [J]. Progress in Exploration Geophysics, 2007, 30(2): 91-95.
- [10] 杨斌, 匡立春, 孙中春. 一种用于测井油气层综合识别的支持向量机方法[J]. 测井技术, 2005, 29(6): 511-514.
- YANG Bin, KUANG Li-chun, SUN Zhong-chun. On support vector machines method to identify oil & gas zone with logging and mudlog information [J]. Well Logging Technology, 2005, 29(6): 511-514.
- (编辑 修荣荣)
-
- (上接第 191 页)
- [3] 田学民, 曹玉苹. 统计过程控制的研究现状及展望[J]. 中国石油大学学报: 自然科学版, 2008, 32(5): 175-180.
- TIAN Xue-min, CAO Yu-ping. Situation and perspectives of statistical process control[J]. Journal of China University of Petroleum (Edition of Natural Science), 2008, 32(5): 175-180.
- [4] CUI J, WANG Y. A novel approach of analog circuit fault diagnosis using support machines classifier [J]. Measurement, 2011, 44(1): 281-289.
- [5] 周延军, 贾江鸿, 李荣华. 基于粗糙集理论和支持向量机的套管损坏动态预报方法[J]. 中国石油大学学报: 自然科学版, 2010, 34(6): 71-75.
- ZHOU Yan-jun, JIA Jiang-hong, LI Rong-hua. Dynamic prediction method of casing damage based on rough set theory and support vector machine [J]. Journal of China University of Petroleum (Edition of Natural Science), 2010, 34(6): 71-75.
- [6] CHOI Y S. Least squares one-class support vector machine [J]. Pattern Recognition Letters, 2009, 30(13): 1236-1240.
- [7] SCHOLKOPF B, PLATT J, SHAEW-TAYLOR J, et al. Estimating the support of a high-dimensional distribution [J]. Neural Computation, 2001, 13(7): 1443-1471.
- [8] MANEVITZ L M, YOUSEF M. One-class SVMs for document classification [J]. Journal of Machine Learning Research, 2001, 2: 139-154.
- [9] 鹿卫国, 戴亚平, 高峰. 一种基于概率分布估计的水电机组故障预警方法[J]. 中国电机工程学报, 2005, 25(4): 94-98.
- LU Wei-guo, DAI Ya-ping, GAO Feng. A hydroelectric-generator unit fault early warning method based on distribution estimation [J]. Proceedings of the CSEE, 2005, 25(4): 94-98.
- [10] MAHADEVAN S, SHAH S L. Fault detection and diagnosis in process data using one-class support vector machines [J]. Journal of Process Control, 2009, 19(10): 1627-1639.
- [11] STEFATOS G, HAMZA A B. Dynamic independent component analysis approach for fault detection and diagnosis [J]. Expert Systems with Applications, 2010, 37(12): 8606-8617.
- [12] LEE J M, YOO C K, LEE I B. Statistical monitoring of dynamic processes based on dynamic independent component analysis [J]. Chemical Engineering Science, 2004, 59(14): 2995-3006.
- [13] HYVARINEN A, OJA E. Independent component analysis: algorithms and applications [J]. Neural Networks, 2000, 13(4/5): 411-430.
- [14] LEE J M, YOO C K, LEE I B. Statistical process monitoring with independent component analysis [J]. Journal of Process Control, 2004, 14(5): 467-485.
- [15] CHIANG L H, RUSSELL F L, BRAATZ R D. Fault detection and diagnosis in industrial systems [M]. New York: Springer-Verlag, 2001.
- (编辑 修荣荣)