第 29 卷 第 2 期
2008 年 3 月

石 油 学 报
ACTA PETROLEI SINICA

Vol. 29 No. 2
Mar. 2008

文章编号：0253-2697(2008)02-0195-04

支持向量机在多地质因素分析中的应用

石广仁

（中国石油勘探开发研究院 北京 100083）

摘要：将支持向量机、人工神经网络、多元回归分析及参数乘积判别法 4 种算法分别应用于鄂尔多斯盆地塔巴庙地区 40 个致密砂岩储层的含气性评价，其预测结果与试气结果的平均相对误差绝对值分别为：0，4.63%，29.71%，18.75%。该实例表明：前两种非线性算法远比后两种线性算法优越；非线性算法中，支持向量机比人工神经网络优越；线性算法中，参数乘积判别法比多元回归分析优越。其根本原因在于：含气性与其相关地质因素（孔隙度、渗透率、含气饱和度）之间存在着复杂的非线性关系。因此，当描述一个研究目标与多个相关地质因素的复杂关系时，应提倡采用非线性算法，特别是在耗时巨大、多次反复进行多地质因素分析的数据处理作业中，应提倡采用支持向量机。因为它与人工神经网络相比，具有计算速度快、计算结果精度高的特点。另外，参数乘积判别法也具有简明、快速的优点，其精度远高于多元回归分析；而多元回归分析不仅计算速度快，而且还具有能表达研究目标与其相关地质因素之间亲疏关系的优点，可作为辅助手段。

关键词：支持向量机；人工神经网络；多元回归分析；参数乘积判别法；致密砂岩；含气性评价

中图分类号：TE19　　**文献标识码**：A

Application of support vector machine to multi-geological-factor analysis

SHI Guangren

(*PetroChina Exploration and Development Research Institute*, *Beijing* 100083, *China*)

Abstract: Four different methods, including support vector machine (SVM), artificial neural network (ANN), multiple regression analysis (MRA) and parameter product decision (PPD), were applied to the gassiness evaluation of forty gas-bearing layers in the tight sandstones of Tabamiao Area in Ordos Basin. Their mean absolute relative residual values between predicated results and gas test results are 0, 4.63%, 29.71% and 18.75%, respectively. This case study shows that the former two nonlinear methods (SVM, ANN) are very superior to the later two linear methods (MRA, PPD). And the SVM is superior to ANN, while PPD is in turn superior to MRA. That is because there exists a complex and nonlinear relationship between gassiness and its related geological factors such as porosity, permeability and gas saturation. Therefore, ANN and SVM should be adopted to describe any complex relationship between the study target and its related multi-geological-factors. In particular, for time-consuming tasks of data processing with repetitious multi-geological-factor analysis, it is recommended that SVM should be used, because it is much faster and more precise than ANN. On the other hand, the case study also indicates that PPD has its advantages of conciseness and high speed. PPD has more precision than MRA, while MRA is fast in processing speed and can be used as an auxiliary tool to establish the order of dependence between the target and its related multi-geological-factors which cannot be estimated using the other three methods.

Key words: support vector machine; artificial neural network; multiple regression analysis; parameter product decision; tight sandstones; gassiness evaluation

统计分析技术已应用于地震、测井、油田开发、钻井及油建等石油勘探开发领域，其中多元回归分析始于 20 世纪 70 年代，人工神经网络始于 80 年代，而支持向量机始于 90 年代。由于人工神经网络能描述非线性关系，目前其应用仍处于高潮。在石油地质方面，当描述多个地质因素的复杂关系时，人工神经网络比多元回归分析优越[1-2]。笔者以鄂尔多斯盆地塔巴庙地区实际资料[3]为样本，采用支持向量机进行含气性评价，并分别与人工神经网络、多元回归分析及陈克勇等提出的参数乘积判别法[3]的预测结果进行对比。

1 几种算法的基本原理

在下面讨论中，要求多元回归分析、人工神经网络、支持向量机和参数乘积判别法 4 种算法使用相同已知参数，并且预测的未知量也相同。

假设有 n 个样本，每个样本含有 $m+1$ 个地质量

作者简介：石广仁，男，1940 年 2 月生，1963 年毕业于西安交通大学，现为中国石油勘探开发研究院教授级高级工程师，主要从事地学定量研究工作。E-mail：grs@petrochina.com.cn

$(x_1, x_2, \cdots, x_m, y)$的成组观察值$(x_{1i}, x_{2i}, \cdots, x_{mi}, y_i)$，其中 $i=1,2,\cdots,n$。$n>m-1$，但实际应用时 n 应远大于 $m-1$。将前 m 个地质量$(x_1, x_2, \cdots, x_m)$的 n 个样本定义为 n 个向量，即

$$\boldsymbol{x}_i = (x_{1i}, x_{2i}, \cdots, x_{mi}) \tag{1}$$

式中：$\boldsymbol{x}$ 为所定义向量的一般形式，$i=1,2,\cdots,n$。

多元回归分析、人工神经网络、支持向量机这 3 种算法的基本思想都是试求一个表达式

$$y = y(\boldsymbol{x})$$

使下式成立，仅是实现途径、计算结果精度不同而已。

$$V_{\min} = \sum_{i=1}^{n}[y(\boldsymbol{x}_i) - y_i]^2 \tag{2}$$

1.1 多元回归分析

欲求的表达式是 $x_j(j=1,2,\cdots,m)$的线性组合再加上一个常数项，即

$$y(\boldsymbol{x}) = b_0 + b_1x_1 + b_2x_2 + \cdots + b_mx_m \tag{3}$$

式中：$b_0, b_1, b_2, \cdots, b_m$ 是通过逐步回归分析[4]求出的常数。

1.2 人工神经网络

利用人工神经网络可求出如下“隐式”表达式：

$$y(\boldsymbol{x}) = \mathrm{ANN}(x_1, x_2, \cdots, x_m) \tag{4}$$

式中：ANN 是非线性函数，可由 BP 模型[4]算出。这种函数不能用通常的数学公式表示，故称为“隐式”表达式。

1.3 支持向量机

支持向量机(SVM)是在统计学习理论基础上发展起来的一种新的机器学习方法，其基本实现途径是通过使用核函数将实际问题转换到高维特征空间，并在该高维空间中构造线性判别函数来实现原空间中的非线性判别函数[5-8]。从理论上说，该算法能够得到全局最优解，避免了人工神经网络在实际应用中可能出现的局部极值问题。

采用 C-支持向量二分类技术[5-6]，求出如下的非线性表达式：

$$y(\boldsymbol{x}) = \sum_{i=1}^{n}[y_i\alpha_i\exp(-\gamma\|\boldsymbol{x}-\boldsymbol{x}_i\|^2)] + b \tag{5}$$

式中：$\boldsymbol{\alpha}$ 是拉格朗日乘子向量，$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_n)$，$0 \leqslant \alpha_i \leqslant C$；$C$ 为惩罚因子；约束条件为 $\sum_{i=1}^{n} y_i\alpha_i = 0$；$\exp(-\gamma\|\boldsymbol{x}-\boldsymbol{x}_i\|^2)$为 RBF 核函数，$\gamma>0$；$\alpha_i$、$C$ 及 γ 可由下式解出，这是一个对偶最优化问题。

$$V_{\max} = \max_{\boldsymbol{\alpha}}\left\{\sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}[\alpha_i\alpha_j y_i y_j \exp(-\gamma\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2)]\right\} \tag{6}$$

b 为分类空间超平面的偏移，可通过自由 $\boldsymbol{x}_i$ 算出。自由 $\boldsymbol{x}_i$ 对应于 $\alpha_i>0$ 的样本向量，最终获得的 SVM 模型仅依赖于这些向量。

1.4 参数乘积判别法

该算法仅用于下面的实例中[3]，其他地区未见应用。样本个数 $n=40$，每个样本含有 4 个地质量的成组观察值：x_1(孔隙度)，x_2(渗透率)，x_3(含气饱和度)，y(含气性评价值)。

前 3 个地质量的乘积为

$$p_r = \phi \cdot k \cdot S_g \tag{7}$$

式中：ϕ 为孔隙度，%；k 为渗透率，$10^{-3}\mu\mathrm{m}^2$；S_g 为含气饱和度，%。

含气性评价值的表达式为[3]

$$y = \begin{cases} 1, p_r > 150 \\ 2, 150 \geqslant p_r > 50 \\ 3, 50 \geqslant p_r > 30 \\ 4, p_r \leqslant 30 \end{cases} \tag{8}$$

式中：1 表示气层，2 表示差气层，3 表示含气层，4 表示干层。

综上所述，上述 4 种算法都有一个相同的计算流程：①利用 n 个已知样本求出计算公式，即式(3)、式(4)、式(5)和式(8)；②将 n 个已知样本代入这些对应的公式，就可以得到各自的预测值 $y_1, y_2, \cdots, y_n$。

2 实际应用效果

以鄂尔多斯盆地塔巴庙地区 40 个致密砂岩储层的孔隙度、渗透率、含气饱和度及试气结果[3]为样本，分别采用多元回归分析、人工神经网络、支持向量机及参数乘积判别法进行了含气性评价，其评价参数和预测结果见表 1。

2.1 求公式

利用多元回归分析算法[4]求出的含气性评价值(y)与 3 个地质要素(x_1, x_2, x_3)的关系式为

$$y = 4.1441 - 0.1411x_1 + 0.094256x_2 - 0.027691x_3 \tag{9}$$

式(9)的残余方差为 0.39605，多重相关系数为 0.77714，说明拟合度不高。

采用人工神经网络 BP 模型[4]，输入层节点 3 个，输出层节点 1 个，隐层节点 7 个；输出层和隐层的网络学习率均为 0.6；迭代次数为 15000。计算出 y 与 x_1，x_2，x_3 的“隐式”表达式为

$$y = \mathrm{ANN}(x_1, x_2, x_3) \tag{10}$$

式(10)的网络全局误差为 0.048126。

采用支持向量机[5-6]，$C=32768$，$\gamma=0.5$，交叉检

表 1　鄂尔多斯盆地塔巴庙地区致密砂岩储层含气性评价参数及预测结果

Table 1　The parameters and predicted results for gassiness evaluation of gas-bearing layers in tight sandstones of Tabamiao Area, Ordos Basin

序号	井名	井　段/m	孔隙度 x_1/%	渗透率 x_2/($10^{-3}\mu m^2$)	含气饱和度 x_3/%	试气结果 y^*	参数乘积判别法 p_r	参数乘积判别法 y^*	参数乘积判别法 相对误差绝对值**/%	多元回归分析 y^*	多元回归分析 相对误差绝对值**/%	人工神经网络 y^*	人工神经网络 相对误差绝对值**/%	支持向量机 y^*	支持向量机 相对误差绝对值**/%
1	d3	2701.0～2708.4	10.10	0.8652	74.10	1	647.5	1	0	0.7488	25.12	1	0	1	0
2	d3	2708.4～2725.4	6.17	0.4782	50.30	3	148.4	2	33	1.9258	35.81	2.7533	8.22	3	0
3	d3	2819.5～2832.0	7.03	0.3526	66.00	1	163.6	1	0	1.3579	35.79	1	0	1	0
4	d4	2856.0～2872.0	5.57	0.3312	51.40	1	94.8	2	100	1.9662	96.62	1.1868	18.68	1	0
5	d10	2509.0～2514.4	11.06	2.0749	74.90	1	1718.8	1	0	0.7052	29.48	1	0	1	0
6	d10	2522.6～2525.0	11.65	3.9939	59.30	1	2759.2	1	0	1.2348	23.48	1	0	1	0
7	d10	2600.5～2603.8	4.43	0.1740	45.00	2	34.7	3	50	2.2894	14.47	2.0681	3.40	2	0
8	d10	2603.8～2606.3	7.05	0.4284	60.00	1	181.2	1	0	1.5284	52.84	1	0	1	0
9	d10	2607.0～2614.5	8.30	3.0923	65.00	1	1668.3	1	0	1.4646	46.46	1	0	1	0
10	d10	2672.7～2676.2	7.68	1.6651	50.00	1	639.4	1	0	1.8329	83.29	1.2297	22.97	1	0
11	d10	2676.2～2685.2	7.68	1.5102	60.00	1	695.9	1	0	1.5414	54.14	1	0	1	0
12	d10	2727.3～2730.0	11.17	1.0088	49.00	2	552.1	1	50	1.3064	34.68	1.7363	13.18	2	0
13	d10	2730.0～2747.0	11.08	2.2951	73.70	1	1874.2	1	0	0.7563	24.37	1	0	1	0
14	d14	2683.4～2687.4	5.91	0.3582	60.20	1	127.4	2	100	1.6771	67.71	1	0	1	0
15	d15	2840.6～2849.6	9.60	0.9093	54.20	1	473.1	1	0	1.3745	37.45	1.1416	14.16	1	0
16	d15	2849.6～2857.0	2.73	0.1429	0.00	4	0	4	0	3.7724	5.69	3.9983	0.04	4	0
17	d16	2647.2～2653.8	7.98	0.4096	70.00	1	228.8	1	0	1.1185	11.85	1	0	1	0
18	d16	2696.8～2703.8	6.48	0.3184	77.00	1	158.9	1	0	1.1277	12.77	1	0	1	0
19	d16	2704.5～2717.7	3.44	0.1184	58.50	4	23.8	4	0	2.0500	48.75	3.9000	2.50	4	0
20	d16	2852.9～2853.5	6.40	0.3315	47.10	2	99.9	2	0	1.9682	1.59	2.4189	20.95	2	0
21	d16	2853.5～2858.0	10.46	1.1226	64.90	1	762.1	1	0	0.9770	2.30	1	0	1	0
22	d16	2861.8～2868.5	4.42	0.1976	39.10	3	34.1	3	0	2.4564	18.12	3.1631	5.44	3	0
23	d16	2868.5～2871.0	7.17	0.4033	41.80	2	120.9	2	0	2.0130	0.65	2.1436	7.18	2	0
24	d18	2771.0～2777.8	8.94	1.6147	47.80	2	690.0	1	50	1.7113	14.43	1.9059	4.70	2	0
25	d18	2778.4～2788.1	8.65	1.5373	65.40	1	869.7	1	0	1.2576	25.76	1	0	1	0
26	d22	2763.6～2766.8	6.89	0.5337	39.20	3	144.1	2	33	2.1368	28.77	2.9023	3.26	3	0
27	d22	2766.8～2768.3	9.11	1.4718	45.70	2	612.8	1	50	1.7320	13.40	2.2566	12.83	2	0
28	d22	2768.3～2773.0	7.71	0.8055	65.10	1	404.3	1	0	1.3296	32.96	1	0	1	0
29	d22	2773.0～2774.5	8.78	2.7089	47.60	2	1132.1	1	50	1.8426	7.87	2.4785	23.93	2	0
30	dk2	2656.8～2660.0	7.22	0.5379	88.20	1	342.5	1	0	0.7338	26.62	1	0	1	0
31	dk2	2660.0～2666.5	7.60	0.6991	87.30	1	463.8	1	0	0.7203	27.97	1	0	1	0
32	dk2	2666.5～2669.4	7.86	1.1193	86.20	1	758.4	1	0	0.7537	24.63	1	0	1	0
33	dk2	2839.1～2844.0	4.74	0.1501	42.00	2	29.9	4	100	2.3265	16.32	2.0878	4.39	2	0
34	dk2	2867.6～2872.4	5.06	0.1769	63.00	1	56.4	2	100	1.7024	70.24	1	0	1	0
35	dk4	2666.1～2675.1	11.52	4.6680	87.90	1	4726.9	1	0	0.5247	47.53	1	0	1	0
36	dk4	2676.1～2680.4	10.57	3.7996	69.30	1	2783.2	1	0	1.0919	9.19	1	0	1	0
37	DT1	2737.3～2750.0	9.79	0.8721	58.36	1	498.3	1	0	1.2290	22.90	1	0.09	1	0
38	DT1	2829.1～2838.4	7.49	0.4017	59.88	1	180.2	1	0	1.4671	46.71	1	0	1	0
39	DT1	2838.4～2842.0	3.42	0.1703	14.90	3	8.7	4	33	3.2651	8.84	3.2608	8.69	3	0
40	DT1	2842.0～2846.1	8.31	0.3762	37.00	2	115.7	2	0	1.9826	0.87	1.7911	10.44	2	0
平均相对误差绝对值/%									18.75		29.71		4.63		0

注:y^* 的数值中,1 为气层,2 为差气层,3 为含气层,4 为干层;** 表示 y 与试气结果相比的相对误差绝对值。

验精度为 82.5%,17 个自由 x_i。计算出 y 与 x_1,x_2,x_3 的"显式"表达式为:

$$y = \mathrm{SVM}(x_1, x_2, x_3) \tag{11}$$

其中 SVM 是非线性函数。这种函数可以用式(5)的数学形式表示,故称为"显式"表达式。

参数乘积判别法的公式为式(8)。

2.2　4 种算法的比较

由表 2 可知,上述 4 种算法的优劣排序为:支持向量机,人工神经网络,参数乘积判别法,多元回归分析。该实例表明,前两种非线性算法远比后两种线性算法

优越。其根本原因在于:含气性与其相关地质因素(孔隙度、渗透率、含气饱和度)之间存在着复杂的非线性关系。因此,当描述多个地质因素的复杂关系时,应提倡采用人工神经网络,尤其是支持向量机。另外,参数乘积判别法也具有简明、快速的优点,其精度高于多元回归分析;而多元回归分析不仅计算速度快,而且具有能表达研究目标与其相关地质因素之间亲疏关系的优点,可作为辅助手段。

表 2 支持向量机与人工神经网络、参数乘积判别法及多元回归分析的比较

Table 2 Comparison of support vector machine with artificial neural network, parameter product decision and multiple regression analysis

算 法	拟合公式	平均相对误差绝对值/%	计算速度(在微机上运行)	预测量 y 与其相关参数 x_1、x_2、x_3 的亲密程度次序	算法评价
支持向量机(C-支持向量二分类技术)	非线性,显式	0	快(约 2 s)	计算不出	优
人工神经网络(BP 模型)	非线性,隐式	4.63	慢(约 20 s)	计算不出	良
参数乘积判别法	线性,显式	18.75	快(<1 s)	计算不出	中
多元回归分析	线性,显式	29.71	快(约 1 s)	x_3、x_1、x_2	差

3 结论与建议

实例表明,支持向量机、人工神经网络、参数乘积判别法、多元回归分析中,前两种非线性算法远比后两种线性算法优越;非线性算法中,支持向量机比人工神经网络优越;线性算法中,参数乘积判别法比多元回归分析优越,这说明在生产实践中总结出的简明快速判别算法在一定条件下是有效的。

当描述多个地质因素的复杂关系时,应提倡采用人工神经网络,尤其是支持向量机;多元回归分析可作为辅助手段。在耗时巨大、多次反复进行多地质因素分析的数据处理作业中,应提倡采用支持向量机,因为它与人工神经网络相比,具有计算速度快、计算结果精度高的特点。

参 考 文 献

[1] 石广仁,张光亚,石晓肆. 多地质因素的勘探目标优选——人工神经网络法与多元回归分析法比较研究[J]. 石油学报,2002,23(5):19-22.
Shi Guangren, Zhang Guangya, Shi Xiaofei. Application of artificial neural network and multiple regression analysis to optimization of exploration prospects[J]. Acta Petrolei Sinica, 2002, 23(5):19-22.

[2] Shi Guangren, Zhou Xingxi, Zhang Guangya, et al. The use of artificial neural network analysis and multiple regression for trap quality evaluation: A case study of the Northern Kuqa Depression of Tarim Basin in western China[J]. Marine and Petroleum Geology, 2004, 21(3):411-420.

[3] 陈克勇,张哨楠,丁晓琪,等. 致密砂岩储层的含气性评价[J]. 石油天然气学报,2006,28(4):65-68.
Chen Keyong, Zhang Shaonan, Ding Xiaoqi, et al. Gassiness evaluation of gas-bearing layers in tight sandstones[J]. Journal of Oil and Gas Technology, 2006, 28(4):65-68.

[4] 石广仁. 地学中的计算机应用新技术[M]. 北京:石油工业出版社,1999:8-21,47-61.
Shi Guangren. New computer application technologies in Earth sciences[M]. Beijing: Petroleum Industry Press, 1999: 8-21, 47-61.

[5] Vapnik V N. 统计学习理论的本质[M]. 张学工,译. 北京:清华大学出版社,2000:85-205.
Vapnik V N. The nature of statistical learning theory[M]. Translated by Zhang Xuegong. Beijing: Tsinghua University Press, 2000:85-205.

[6] Cristianini N, Shawe-Taylor J. 支持向量机导论[M]. 李国正,王猛,曾华军,译. 北京:电子工业出版社,2004:8-149.
Cristianini N, Shawe-Taylor J. An introduction to support vector machines[M]. Translated by Li Guozheng, Wang Meng, Zeng Huajun. Beijing: Publishing House of Electronics Industry, 2004: 8-149.

[7] 杨斌,匡立春,孙中春,等. 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 Lichun, Sun Zhongchun, et al. On support vector machines method to identify oil & gas zone with logging and mudlog information[J]. Well Logging Technology, 2005, 29(6): 511-514.

[8] 闫铁,毕雪亮,王长江. 基于支持向量机和聚类分析理论的钻具失效分析方法[J]. 石油学报,2007,28(3):135-140.
Yan Tie, Bi Xueliang, Wang Changjiang. Failure analysis of drill stem based on support vector machine and cluster analysis theory[J]. Acta Petrolei Sinica, 2007, 28(3):135-140.

(收稿日期 2007-10-08 编辑 王 秀)