

基于模式识别理论的卡钻类型判别分析

吴 军¹, 臧艳彬², 陈星星¹

(1. 中国石化重庆涪陵页岩气勘探开发有限公司, 重庆 408014; 2. 中国石化石油工程技术研究院, 北京 100101)

摘要:将模式识别理论中的支持向量机、Bayes 判别分析以及多元回归分析应用于卡钻类型判别分析, 建立了基于模式识别理论的卡钻判别模型。以近几年川东北卡钻数据为例进行了算例分析, 结果表明: 采用支持向量机、Bayes 判别法及多元回归法对卡钻类型判别的结果与实际结果的误判率分别为 1.92%, 11.52%, 61.54%。支持向量机(SVM)判别结果精度最高, 但其判别方程式较为复杂, 不能直观看出各分量对结果的影响程度; 多元回归分析判别方程形式简单, 可以直观表达各参量与卡钻之间的密切程度, 但其判别精度较低; Bayes 判别法计算精度介于二者之间, 但其判别精度与判别式的个数密切相关。

关键词:模式识别; 支持向量机; Bayes 判别法; 多元回归; 卡钻

中图分类号: P634.8 **文献标识码:** A **文章编号:** 1672-7428(2015)10-0031-04

Discriminant Analysis on Sticking Based on Pattern Recognition Theory/WU Jun¹, ZANG Yan-bin², CHEN Xing-xing¹ (1. The Exploration and Production of Shale Gas Chongqing Fuling, SINOPEC, Chongqing 408014, China; 2. SINOPEC Institute of Petroleum Engineering, Beijing 100101, China)

Abstract: The Support Vector Machine (SVM), Bayes discriminant analysis and multiple regression analysis were used for diagnosis and prediction of sticking. The discriminant model of stick type was built based on pattern recognition theory. The calculation analysis on the sticking data of northeastern Sichuan was made, it was indicated that the misjudgment rate for discriminant result of SVM, Bayes discriminant analysis and multiple regression analysis was 1.92%, 11.52% and 61.54%. The accuracy of SVM recognition result was the highest, but its discriminant equation is complex and the contribution of each component to the result could not be intuitively seen; while the equation of multiple regression analysis is simple, which could intuitively show the close degree between each component and sticking, but the accuracy of recognition result was lower. The accuracy of Bayes discriminant analysis was between the above two, but the discriminant accuracy is closely related to the number of discriminant.

Key words: pattern recognition; Support Vector Machine, Bayes discriminate method; multiple regression; sticking

作为人工智能技术基础学科的模式识别技术, 在文字识别、遥感、气象预报、军事侦察等领域都取得了巨大成效。模式识别自 20 世纪 60 年代发展成为一门独立学科以来, 共经历了 2 个阶段: 统计模式阶段和人工智能阶段。统计模式识别方法主要有判别函数法、k 近邻分类法、Bayes 判别法、主因子分析法等, 其中 Bayes 判别规则从理论上解决了最优分类器的设计问题。90 年代统计推断理论研究所取得的突破性成果促成现代统计学习(VC)理论的建立, 该理论不仅在严格的数学基础上圆满地回答了人工神经网络中出现的理论问题, 而且导出了一种新的学习方法——支撑向量机, 从而将模式识别技术推向另一个高峰^[1]。

卡钻是钻井施工过程中最常见和易发的事故之一, 卡钻事故的影响因素多种多样且关系复杂, 传统

的统计分析和摩阻系数方法预测效果不理想。目前卡钻预测与判别技术尚没有实质性的突破。卡钻的钻前预测与预防主要依靠工程技术人员的操作经验和相关技术规范进行, 预测精度较低。本文以近几年中石化四川盆地川东北地区(元坝、普光等气田)卡钻资料为样本, 采用支持向量机方法进行了卡钻判别分析, 建立了卡钻类型判别方程, 并分别与 Bayes 判别法和多元回归分析的预测结果进行了对比。计算结果表明, 利用模式识别理论建立的卡钻类型判别模型进行卡钻的预测和评估是切实可行的^[2-8]。

1 模式识别理论

1.1 多元回归

假定被解释变量 Y 与多个解释变量 X_1, X_2, \dots, X_k 之间具有线性关系, 是解释变量的多元线性函

收稿日期: 2015-02-11; 修回日期: 2015-08-01

基金项目: 中石化科技攻关项目“页岩气‘井工厂’技术研究”(编号: P13138)

作者简介: 吴军, 男, 汉族, 1981 年生, 石油工程专业, 主要从事涪陵页岩气钻井技术管理及成本分析方面的研究与管理工
作, 重庆市涪陵区焦石镇涪陵页岩气公司钻井工程项目部, zangyb.sripe@sinopec.com。

数,称为多元线性回归模型^[9]。即

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \mu \quad (1)$$

其中 Y 为被解释变量, $X_j (j=1, 2, \dots, k)$ 为 k 个解释变量, $\beta_j (j=0, 1, 2, \dots, k)$ 为 $k+1$ 个未知参数, μ 为随机误差项。

1.2 Bayes 判别法

设 G_1, G_2, \dots, G_k 为 k 个 p 元总体, 分别具有概率密度 $f_1(x), f_2(x), \dots, f_k(x)$, 在进行判别分析以前, 我们已对各总体有一定的了解。一般说来, 一个待判样品应该首先考虑判入有较大可能出现的总体之中^[10]。在 Bayes 判别中, 应该考虑误判引起的损失。

此时的 Bayes 判别法则:

$$\begin{cases} x \in G_1: \text{若 } P(G_1|x) \geq P(G_2|x) \\ x \in G_2: \text{若 } P(G_1|x) < P(G_2|x) \end{cases} \quad (2)$$

Bayes 判别(即使 L 达到最小)的最优划分为:

$$\begin{cases} x \in G_1: \text{若 } c(2|1)(G_1|x) \geq c(1|2)(G_2|x) \\ x \in G_2: \text{若 } c(2|1)(G_1|x) < c(1|2)(G_2|x) \end{cases} \quad (3)$$

1.3 支持向量机(SVM)判别法

支持向量机是 V. Vapnik 等根据统计学理论(Statistical Learning Theory 简称 SLT)提出的一种新的机器学习方法, 在解决小样本、非线性及高维模式识别问题中表现出许多特有的优势^[8]。支持向量机从本质上讲是一种前向神经网络, 通过核函数将实际问题转换到高维特征空间, 并在高维特征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。从实施的角度考虑, 训练支持向量机的核心思想等价于求解一个线性约束的二次规划问题, 从而构造一个超平面作为决策平面, 使得特征空间中两类模式之间的距离最大, 而且它能保证得到的解为全局最优解^[11]。

设线性可分样本集为 $(x_i, y_i), i=1, 2, \dots, n, x \in R^d, y \in \{+1, -1\}$ 是类别标号。 d 维空间中线性判别函数的一般形式为: $g(x) = w \cdot x + b$, 分类面方程为:

$$wx + b = 0 \quad (4)$$

将判别函数进行归一化, 使两类所有样本都满足 $|g(x)| \geq 1$, 即使离分类面最近的样本 $|g(x)| = 1$, 这样分类间隔就等于 $2/||w||$, 因此间隔最大等价于使 $||w||$ (或 $||w||^2$) 最小。而要求分类线对所有样本正确分类, 就是要求其满足:

$$y_i(wx_i + b) - 1 \geq 0, (i=1, 2, \dots, n) \quad (5)$$

因此, 满足上述条件且使 $||w||^2$ 最小的分类面就是最优分类面。这两类样本中离分类面最近的点且平行于最优分类面的超平面上的训练样本就是使

式(4)中等号成立的那些样本, 即支持向量(Support Vectors)。根据上面的讨论, 最优分类面问题可以表示成如下的约束优化问题, 即在式(4)的约束下, 求式(5)的最小值。

$$\varphi(w) = \frac{1}{2} ||w||^2 = \frac{1}{2} (ww) \quad (6)$$

这是一个二次规划问题, 可定义以下的拉格朗日(Lagrange)函数:

$$L(w, b, a) = \frac{1}{2} (ww) - \sum_{i=1}^n a_i [y_i(wx_i + b) - 1] \quad (7)$$

其中: $a_i > 0$ 为 Lagrange 系数。求式(8)的极小值就是对 w 和 b 求拉氏函数的极小值。求 L 对 w 和 b 的偏微分, 并令其等于 0, 可转化为对偶问题。

如果用内积 $K(x, x')$ 代替最优分类面中的点积, 就相当于把原特征空间变换到了某一新的特征空间, 此时优化函数变为:

$$w(a) = \sum_{i=1}^n a_i - \frac{1}{2} \sum_{i,j} a_i a_j y_i y_j (x_i, x_j) \quad (8)$$

相应的判别函数也应变为:

$$f(x) = \text{sgn} \left[\sum_{i=1}^n a_i^* y_i k(x, x_i) + b^* \right] \quad (9)$$

常用的核函数有以下几种:

线性内积函数 $K(x, y) = xy$;

多项式内积函数 $K(x, y) = (xy + 1)^d$;

径向基内积函数 $K(x, y) = \exp(-|x - y|^2 / \sigma_2)$;

二层神经网络内积函数 $K(x, y) = \tanh[k(xy) + c]$ 。

2 卡钻判别结果分析

为了对比分析 3 种判别方法, 整理了中石化川东北地区(元坝、普光等气田)卡钻事故相关数据, 这些参数包括影响卡钻的主要参数, 如地层参数、钻井液性能参数及井径扩大率的工艺参数。

2.1 数据预处理

参照现场数据, 按照支持向量机、Bayes 判别法和多元回归分析建模要求, 将所需数据作如下预处理: 卡钻类型按坍塌卡钻、压差卡钻、缩径卡钻、沉沙卡钻、键槽卡钻、其他卡钻、正常钻进分类卡钻类型值分别为(1, 2, 3, 4, 5, 6, 7); 此次研究共搜集样本 52 条, 见表 1。

2.2 计算公式

(1) 利用多元回归分析算法求出卡钻类型值与影响因素的关系式为:

表 1 卡钻类型判别样本

序号	井深/ m	地温/ ℃	地层压力/ MPa	破裂压力/ MPa	钻压/ kN	转速/ (r·min ⁻¹)	泵压/ MPa	排量/ (L·min ⁻¹)	钻时/ h	井斜角/ (°)	井径/ mm	钻头直径/ mm	密度/ (g·cm ⁻³)	粘度/ s	滤失量/ [mL·(30 min) ⁻¹]	初切力/ Pa	终切力/ Pa	PV/ (mPa·s)	YP/ Pa	类型	
1	2006	58.9	1.03	1.80	150	80	16.0	50	5.2	6.20	354.720	311.1	1.140	37.0	5.2	1.0	4.5	10.0	4.5	1	
2	2159	55.1	1.05	1.76	220	0	17.5	55	44.8	6.02	356.720	311.1	1.170	70.0	5.0	3.0	10.5	18.0	11.0	1	
3	5292	130.6	1.20	1.63	80	90	17.0	28	36.0	1.30	261.490	215.9	1.530	75.0	3.5	3.0	7.5	41.0	10.0	1	
4	5806	138.3	1.20	1.70	90	0	19.2	24	12.0	3.10	217.420	215.9	1.580	69.0	2.0	2.0	9.0	23.0	14.0	2	
5	5580	133.6	1.40	2.22	200	70	18.0	28	42.0	2.79	241.500	251.9	1.470	53.0	3.8	10.0	16.5	29.0	10.0	3	
6	2699	71.4	1.09	1.60	240	60	16.0	40	60.6	5.50	404.698	311.1	1.430	95.0	4.0	4.0	10.0	65.0	9.0	5	
7	6032	141.1	1.13	2.20	45	85	19.0	14	17.8	35.54	173.270	165.1	1.260	50.0	3.0	1.5	4.5	20.0	23.0	2	
8	5868	137.7	1.20	2.20	45	60	20.0	14	17.0	34.86	173.270	165.1	1.290	53.0	4.0	4.0	5.5	25.0	7.0	1	
9	5644	133.0	1.20	2.20	51	60	20.0	14	16.8	34.86	173.270	165.1	2.290	53.0	4.0	4.0	5.5	25.0	8.0	1	
10	6138	138.3	1.20	2.40	51	0	20.0	30	39.0	39.25	243.840	241.3	1.340	65.0	3.2	2.0	6.0	46.0	11.0	2	
11	5354	126.9	1.20	1.75	100	30	18.0	26	90.0	46.60	244.840	241.3	1.350	82.0	2.5	2.5	10.0	36.0	10.5	6	
12	213	24.5	1.02	1.68	80	70	14.0	53	12.2	0.30	626.212	444.5	1.020	37.0	8.0	1.0	2.0	0.0	0.0	2	
13	3002	72.7	1.10	1.72	200	75	18.5	46	88.2	1.42	334.650	314.0	1.370	55.0	4.0	2.5	10.5	16.0	11.0	1	
...
49	2320	65.5	1.12	1.82	124	75	2.2	46	4.5	1.02	333.760	314.3	1.100	58.5	3.7	2.3	8.5	34.5	10.5	7	
50	3120	90.2	1.16	1.82	327	60	14.0	36	70.5	4.07	320.040	314.3	1.360	60.0	3.0	3.3	9.5	36.0	10.5	7	
51	5320	135.2	1.15	1.75	90	30	17.5	26	24.0	46.22	241.300	241.3	1.350	82.0	2.5	2.5	10.0	36.0	10.5	7	
52	4540	111.9	1.75	2.16	320	45	21.0	33	104.0	0.17	323.393	314.3	1.565	65.0	3.0	9.5	19.5	38.0	10.5	7	

$$y = 0.458 - 0.002x_1 + 0.096x_2 + 9.501x_3 - 2.466x_4 + 0.011x_5 + 0.012x_6 - 0.109x_7 - 0.001x_8 - 0.012x_9 + 0.006x_{10} - 0.024x_{11} + 0.021x_{12} - 4.116x_{13} + 0.033x_{14} + 0.414x_{15} - 0.008x_{16} - 0.163x_{17} + 0.029x_{18} + 0.020x_{19} \quad (10)$$

上式中的残余方差为 340.058, 多重相关系数为 0.662, 拟合精度较低。

(2) 采用 Bayes 判别分析法求得卡钻类型值与影响因素的关系为:

$$D = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \\ x_8 \\ x_9 \\ x_{10} \\ x_{11} \\ x_{12} \\ x_{13} \\ x_{14} \\ x_{15} \\ x_{16} \\ x_{17} \\ x_{18} \\ x_{19} \end{pmatrix}^T \times \begin{pmatrix} 0.0 & 0.001 & 0.002 & 0.001 & 0.0 & 0.001 \\ -0.007 & -0.039 & -0.072 & -0.067 & -0.001 & -0.064 \\ 5.879 & -5.539 & -3.035 & -2.178 & 4.306 & 4.779 \\ -0.490 & -1.783 & 2.338 & -0.677 & -1.208 & -0.782 \\ 0.015 & -0.002 & -0.006 & 0.005 & 0.001 & 0.0 \\ 0.028 & -0.001 & -0.012 & 0.007 & -0.023 & 0.004 \\ -0.068 & -0.050 & 0.049 & 0.109 & -0.073 & 0.037 \\ 0.002 & 0.003 & 0.0 & 0.0 & 0.0 & 0.001 \\ -0.041 & 0.022 & -0.007 & -0.014 & -0.002 & -0.004 \\ 0.018 & 0.011 & 0.000 & -0.007 & -0.005 & -0.002 \\ -0.015 & -0.004 & 0.018 & 0.006 & 0.0 & 0.004 \\ 0.028 & -0.009 & -0.005 & -0.007 & 0.006 & -0.003 \\ 0.803 & 1.191 & 3.300 & 0.813 & -0.933 & 1.418 \\ 0.052 & -0.040 & -0.007 & 0.046 & 0.005 & 0.039 \\ -0.293 & 0.596 & -0.521 & 0.248 & 0.492 & -0.099 \\ -0.300 & 0.111 & -0.147 & -0.055 & -0.078 & -0.042 \\ -0.123 & 0.235 & -0.014 & -0.070 & -0.156 & -0.072 \\ -0.009 & 0.015 & -0.033 & 0.067 & 0.040 & -0.072 \\ 0.030 & -0.102 & 0.091 & -0.020 & 0.160 & -0.037 \end{pmatrix} + \begin{pmatrix} -12.134 \\ 9.204 \\ -4.920 \\ -1.950 \\ -4.770 \\ -5.570 \end{pmatrix}^T \quad (11)$$

(3) 采用支持向量机(SVM) 计算出卡钻类别值公式为:

$$y = SVM(x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, x_{16}, x_{17}, x_{18}, x_{19}) \quad (12)$$

其中 SVM 是非线性函数, 这种函数可以用显式的数学形式表示, 故称为“显式”表达式。SVM 卡钻类型判别结果如图 1 所示。

2.3 算法结果对比分析

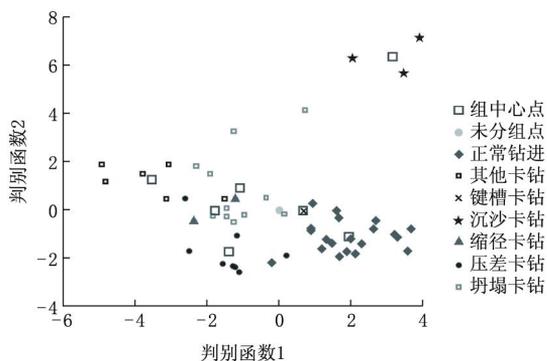


图1 SVM卡钻类型判别分析结果

以中石化川东北地区(元坝、普光等气田)近几年发生卡钻事故案例为样本,分别采用多元回归分析、Bayes 判别分析、支持向量机进行卡钻类型判别分析,其预测结果见表2。

表2 3种算法预测结果对比

样本序号	实钻	多元回归	Bayes 判别	SVM	样本序号	实钻	多元回归	Bayes 判别	SVM
1	1	4	7	1	27	2	2	2	2
2	1	3	1	1	28	2	2	2	2
3	1	5	1	1	29	1	0	1	1
4	2	2	2	2	30	4	5	4	4
5	3	5	3	3	31	2	3	1	2
6	5	5	5	5	32	1	2	1	1
7	2	3	2	2	33	7	6	7	7
8	1	3	3	1	34	7	7	7	7
9	1	-1	1	1	35	7	6	7	7
10	2	2	2	2	36	7	7	7	7
11	6	4	1	6	37	7	5	7	7
12	2	2	2	2	38	7	6	7	7
13	1	3	1	1	39	7	7	7	7
14	1	3	4	1	40	7	5	7	7
15	4	3	4	4	41	7	6	7	7
16	4	4	4	4	42	7	7	7	7
17	1	1	1	1	43	7	6	7	7
18	1	3	1	1	44	7	7	7	7
19	6	3	6	6	45	7	6	7	7
20	6	7	6	6	46	7	6	7	7
21	3	3	3	3	47	7	7	7	7
22	2	2	3	2	48	7	4	7	3
23	6	7	6	6	49	7	7	7	7
24	6	4	6	6	50	7	7	7	7
25	1	3	1	1	51	7	5	7	7
26	6	5	6	6	52	7	7	7	7

将3种算法对卡钻故障预测结果对比(见表3),可以看出其优劣排序为支持向量机 > Bayes 判别法 > 多元回归分析。卡钻判别实例表明,支持向量机作为一种非线性算法远比后两种线性算法优越。其根本原因在于卡钻类型的影响参数与卡钻类型存在复杂的非线性关系。另外,Bayes 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因素之间亲疏关系的优点,可作为辅助手段。

表3 3种算法综合评价

算法	误判率/%	拟合公式	公式特点	算法实现	运算速度	总体评价
支持向量机 SVM	1.92	显性	非线性	编程	长	优
Bayes 判别	11.52	显性	线性方程组	编程	较长	良
多元回归	61.54	显性	线性	Excel 或者编程	短	差

3 结论与建议

(1)卡钻应用实例表明,基于模式识别理论的支持向量机和 Bayes 判别法与多元回归分析相比,对卡钻类型识别准确率更高,特别是基于计算机机械学理论的支持向量机(SVM)模式识别方法,对卡钻类型的识别率达到 98.08%。

(2)将计算机模式识别理论应用于钻井工程中卡钻判别分析,取得了较好的识别效果,所建立的判别方程式具有卡钻类型判别功能,从而为卡钻类型分析与技术措施的制定提供指导意见。

(3)实际应用中可选择其中的 1~2 种进行联合判别。本研究为卡钻类型的定量判别提供了一种新的研究思路,对现场卡钻事故的预防与处理具有一定借鉴和参考价值。

参考文献:

- [1] 边肇祺,张学工. 模式识别(第二版)[M]. 北京:清华大学出版社,2007:32-51.
- [2] 阎铁,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.
- [3] 石广仁. 支持向量机在多地质因素分析中的应用[J]. 石油学报,2008,29(2):195-199.
- [4] 李建军,丁正生,张海燕. 常用判别分类方法分析[J]. 西安科技大学学报,2007,27(1):138-142.
- [5] 罗刚,艾志久,王其华,等. 基于模糊数学卡钻事故安全评价体系研究[J]. 西南石油大学学报,2007,29(6):118-122.
- [6] 陈晖,沈小翠. 卡钻事故诊断仿真系统研究[J]. 石油机械,2009,37(7):55-57.
- [7] 张林强. 井下卡钻分析及处理[J]. 海洋石油,2007,(3):112-116.
- [8] Vapnik V N. The nature of statistical learning theory[M]. Translated by zhangxuegong. Beijing: Tsinghua University Press,2000:85-205.
- [9] 严丽,王燕,范树平. 多元回归分析方法预测川东北礁滩相储层产能[J]. 新疆石油天然气,2011,7(4):37-40.
- [10] 顾和元,侯国庆,吴占伟. 基于动态贝叶斯网络的深水防喷器可靠性研究[J]. 石油机械,2013,41(3):36-39.
- [11] 韦明辉,黄海龙,韦忠良,等. 基于支持向量机的钻井风险实时预测方法[J]. 钻采工艺,2012,35(5):15-17.