基于 PCA 的 LS-SVM 预测模型应用①

胡剑策

(温州医学院, 温州 325035)

摘 要：油气储层的识别和预测是当今热门的研究课题。本文以实地测井数据为基础，提出基于 PCA 的 LS-SVM 预测模型对储层油气进行分类预测，并与人工神经网络预测模型对比。结果表明，该模型的性能优于其它模型，具有一定的应用价值。

关键词：主成分；最小二乘支持向量机；预测；油气

Applications of Forecasting Model Based on PCA-LS-SVM

HU Jian-Ce

(Wenzhou Medical College ,Wenzhou 325035, China)

Abstract：Identification and prediction of oil and gas reservoirs are popular research topic today. Based on logging data, the proposed PCA-based LS-SVM forecasting model is applied identify oil and gas reservoirs, comparing with artificial neural network prediction model. The results show that the model's performance is stronger than other models, has a certain value.

Key words：principal component; least squares support vector machine; forecasts; oil and gas

储层是油气聚集的地方，是研究油气勘探的重要对象。为了深入了解油气的分布情况，首先需要建立能反映储层情况的地质模型。而测井数据是目前所能获得的连续性最好分辨率最高的地质数据，最能反映地层信息[1]。因此，测井资料在地层学研究中显得格外重要，利用测井资料进行油气储层的识别和预测也成为一项重要的研究课题。

利用测井资料识别油气层，传统的方法是人工利用长期生产实践所积累的经验来识别划分储层。虽然人工划分油气层的符合率比较高，但主要依赖测井数据解释人员丰富的实践经验，存在着一定的误差性和偶然性，且效率较低，对油气解释工作的推广极为不利。

本文以储层实地测井数据为基础，提出基于主成分分析的支持向量机预测模型对储层油气进行识别。首先采用主成分分析法对测井数据进行降维处理，再构建储层的支持向量机分类预测模型，并利用该模型对储层进行分类预测。最后通过与误差反向传播人工神经网络模型进行对比，得出该模型的精确性和实用性，具有一定的应用价值。

1 主成分分析（PCA）

主成分分析(Principal Component Analysis，PCA)是处理高维数据的方法之一，它以尽可能少的信息损失为前提，将高维数据投影到低维空间以达到简化数据结构、降维的目的。它也是将多个相关变量进行综合，化为不相关变量的方法[2,3]。

设原始数据如下:

$$X=[x_1,x_2,\cdots,x_n] \tag{1}$$

1）根据原始数据矩阵 X，求出它的协方差矩阵 S；

2）求 S 矩阵的特征值和特征向量。考虑特征方程：

$$(\lambda I-S)U=0 \tag{2}$$

式中，I 为单位矩阵，U 为特征矩阵。

3）取变换矩阵 $A=U^T$ 即得到主成分矩阵:

① 收稿时间:2011-12-06;收到修改稿时间:2012-02-17

$$Y = U^T X \tag{3}$$

得到一组（m 个）新的变量，它们依次被称为第一主成分、第二主成分、……第 m 主成分。其中第一个主成分几乎包含了总方差 80 %以上的信息量,其余各个主成分所包含的信息依次减少。实际应用中，选取前面几个主成分，因为包含了绝大部分的信息量。这样不但降低了数据维数，也不会明显损失原始数据中的信息。

2 最小二乘支持向量机（LS- SVM）

最小二乘支持向量机[4-6]（Least Square Support Vector Machine, LS-SVM）是一种改进的 SVM 算法，其不同之处在于将不等式约束条件转变成等式约束，并用训练误差的 e2 来替代松弛变量 ξ。

设训练样本集 $S=\{(x_i,y_i),i=1,\cdots,n,x\in R^d, y_i\in\{+1,-1\}\}$，其中 x_i 为输入样本，y_i 为输出类别标志，n 为样本数。与标准 SVM 算法的区别是 LS-SVM 基于 SRM 准则构造最小化目标函数及其约束条件，如式 4 所示：

$$\begin{aligned}&\min_{w,b,e} Q(w,b,e)=\frac{1}{2}\|w\|^2+\frac{\gamma}{2}\sum_{i=1}^{n}e_i^2\\&y_i(w^T\phi(x_i)+b)=1-e_i,i=1,2,...,n\end{aligned} \tag{4}$$

式 4 中，$\phi(x_i)$ 是将输入样本数据映射到高维特征空间的非线性函数，w 为权值向量，γ 为调整参数，误差变量和偏置项满足 $e_i,b\in R$。上述最优化问题对应的 Lagrange 函数为

$$L(w,b,e;\alpha)=Q(w,b,e)-\sum_{i=1}^{n}\alpha_i[y_i(w^T\phi(x_k)+b)-1+e_i] \tag{5}$$

其中，α_i 称为拉格朗日乘子。经化简，最后可求得 LS-SVM 分类函数

$$f(x)=\sum_{i=1}^{n}\alpha_i y_j K(x_i,x_j)+b \tag{6}$$

式 6 中 K 函数为 SVM 中的核函数，常用的多项式核函数、有线性核函数、Sigmoid 核函数以及高斯核函数等。本文采用的径向基核函数中的高斯核函数，其映射函数式为

$$\boldsymbol{K}(\boldsymbol{x}_i,\boldsymbol{x}_j)=\exp\{-\frac{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2}{\sigma^2}\} \tag{7}$$

3 储层油气识别预测

本文选取塔里木盆地北部某井的测井数据为研究对象，其深度段为 5000.0m - 5400.0m。该井的测井解释数据表如表 1 所示。

表 1　测井解释数据表

层号	深度 m	厚度 m	GR API	…	AC μs/ft	解释结论
1	5000~5006.5	6.5	75.7	…	76.1	水 层
2	5013.5~5020	6.5	79.2~68.6	…	72.4~73.3	水 层
3	5291~5295	4	68.6	…	73	气 层
4	5319~5322	3	81~93	…	69.5	气 层
5	5327.5~5333	5.5	78	…	74.3	气 层
6	5340~5349.5	9.5	72.7~73.3	…	71.6~73.1	气 层
7	5354.5~5358	3.5	84.9	…	71.2	水 层
8	5376~5386	10	79.5~63.8	…	72.4~70.1	油 层
9	5392~5397.0	5	77.6~71.3	…	76.4~73.9	油 层

其中 1、2、7 层为水层， 3-6 层为气层，8 和 9 层为油层，其余深度皆为非储层。该井的采样间隔为 0.125m，每个采样点都有 25 种不同属性的数据，表 2 是部分属性测井信号的统计特征。

表 2　部分属性的统计特征

属性	最小值	最大值	平均值	方差
GR	27.975	135.400	78.35826	17.183405
RT	1.443	15.866	4.43197	2.199503
SP	90.268	140.930	112.30887	11.910484
CNL	3.471	32.495	14.24334	3.682130
AC	54.552	80.638	70.36285	4.632531
…				

由表 2 可看出，具有不同量纲的各种测井曲线统计特征差异很大。

3.1 主成分分析

为了防止具有相对较大初始值域的属性与具有相对较小初始值域的属性相比权重过大，需要先对原始数据进行标准化处理。本文采用零均值规范法。

再利用 SPSS 统计分析软件进行主成分分析，求得原始数据经过标准化处理后的特征值、信息量和累计信息量值，如表 3。从表 3 中的分析结果可知，第一主成分的信息量为 62.346%，前四个主成分的累积信息量达到了 92.937%,能很好地概括原始数据。其他主成分的信息量接近于零，由此可看出原始的 25 种测井属性数据存在很大的相关性，包含大量的信息冗余。

表 3 方差分析

成分	特征值	信息量%	累计信息量%
1	15.562	62.346	62.346
2	4.614	18.486	80.832
3	1.982	7.944	88.776
4	1.038	4.161	92.937

根据前面第一节介绍的主成分分析处理步骤，利用公式 $3\ Y=U^TX$ 即可得到变换后的新数据，该式中 X 为原始的具有 25 种属性的采样点的测井数据。U^T 为主成分系数矩阵，矩阵大小为 25 行 4 列。Y 即为所求的转换后的新的数据矩阵，其属性指标数由原来的 25 个减少为现在的 4 个。表 4 为原始属性与主成分系数对比表。可见，主成分分析降低了测井属性的特征维数，压缩了原始数据。

表 4 原始属性和主成分系数

样本号	原始属性（25 种属性）				主成分			
	GR	RT	…	AC	PCA1	PCA2	PCA3	PCA4
1	76.52	1.76		74.38	0.975	0.143	0.031	-0.093
2	85.01	6.74		69.57	0.907	0.237	0.201	-0.242
3	78.24	5.77		71.63	0.887	0.206	0.327	-0.050
…								

3.2 最小二乘支持向量机预测模型

本文选取测井解释数据表 1 中第一二层、第三四五层、第八层的采样点分别为水层、气层、油层训练样本，第七层、第六层、第九层的采样点分别为水层、气层、油层测试样本。其余深度段皆为非储层样本。由于储层分四类，LS-SVM 模型需构造两个分类器，四个主成分分别作为输入，输出[-1，-1]、[-1，1]、[1，-1]和[1，1]分别对应水层、气层、油层和非储层。训练样本数各 30 个，测试样本数与训练样本数之比选取 1:1 和 2:1 分别进行实验。

LS-SVM 模型中惩罚系数 γ 与径向基核函数中 σ 对预测结果起较大作用。为获取性能优秀的支持向量机核心参数，并综合考虑各方面因素，本文采用网格搜索法优化参数。

网格搜索参数优化法的思想为：根据长期研究经验将惩罚系数 γ 和核函数参数 σ 在一定的取值范围内以一定的间隔赋值，如 γ =[2:2:10]、σ =[1:0.5:4]，那么 γ 与 σ 的组合就有 5x8=40 种。对这 40 种组合训练 LS-SVM 分类器，然后选择分类识别正确率最大的一组参数作为最优的 γ 和 σ 。如果结果均不理想，就需要重新考虑 γ 和 σ 的范围与采样间隔。本文通过该方法得出：当 $\gamma=8$，$\sigma^2=1$ 时， 模型的学习和预测能力较强。

为了更好的说明基于主成分分析的 LS-SVM 预测模型的预测效果，本文还与误差反向传播人工神经网络模型进行比较。选用 4—8—4 的 3 层结构 BP 神经网络，学习步长 0.3 ，动量项系数 0.95，最大训练次数 1000，最小均方误差 1e-8。经过 152 次迭代，人工神经网络收敛，并用于分类预测。各模型分类预测结果对比，如表 5 所示。

表 5 储层预测结果比较

预测模型	参数设置	训练耗时（s）	准确率（%）	
			1:1	1:2
LS-SVM	$\gamma=2,\sigma^2=1$	22.47	90.89	89.71
PCA-LS-SVM	$\gamma=8,\sigma^2=1$	16.13	90.83	89.65
PCA-BP 神经网络	4—8—4	37.962	91.36	83.29

从表 5 可以看出，上述方法的分类预测结果都较为理想。基于 PCA 的 LS-SVM 模型训练所需时间最短，而 BP 人工神经网络模型训练耗时最多。基于 PCA 的 LS-SVM 模型训练耗时是单一的 LS-SVM 模型的四分之三，而准确率几乎没有降低，说明主成分分析压缩了数据，但重要信息都得到保留。再从测试样本与训练样本的比例可以看出，当测试样本数目增加时，BP 人工神经网络的预测准确率大幅降低，而支持向量机模型的准确率只是略微有所下降，说明支持向量机模型比较稳定，性能明显优于 BP 人工神经网络模型。

4 结论

支持向量机是一种建立在结构风险最小化原理基础上的机器学习方法。它根据有限的样本信息，在模型学习能力和复杂度之间折衷寻求最佳的推广性能。本文在支持向量机分类算法的基础上提出基于 PCA 的最小二乘支持向量机分类预测模型，并用于储层油气分类预测。结果表明该模型具有较好的性能和较高的准确率，具有一定的实际应用价值。

参考文献

1 雍世和,张超谟.测井数据处理与综合解释.东营:石油大学出版社,1996.
2 刘国璧,孙群,孟涛,袁宏俊.基于主成分 LS-SVM 的教学质量评估模型.湖南工程学院学报,2011(3).
3 郭志钢,蒲忠,李秋.基于主成分分析的 LS-SVM 非线形预测模型应用.统计与决策,2010,(10).
4 陈帅,朱建宁,潘俊,侍洪波.最小二乘支持向量机的参数优化及其应用,华东理工大学学报(自然学版),2008(4).
5 姚凯丰,李衍达等.一种基于 SVM 特征选择的油气预测方法.地球物理勘探,2004,24(7).
6 郭辉,刘贺平,王玲.最小二乘支持向量机参数选择方法及其应用研究.系统仿真学报,2006,18(7).