一种基于SVM特征选择的油气预测方法*

姚凯丰[1]　陆文凯[1]　丁文龙[1]　张善文[2]　肖焕钦[2]　李衍达[1]

(1.清华大学自动化系智能技术与系统国家重点实验室　2.中国石化胜利油田有限公司)

姚凯丰等.一种基于SVM特征选择的油气预测方法.天然气工业,2004;24(7):36～38

摘　要　支持向量机(SVM)是近年来发展起来的一种通用的机器学习方法,在许多分类问题和函数拟合问题上都已获得了很好的效果。对于少量样本的分类问题,SVM具有调节参数较少,运算速度快等优点。通过地震、测井等信息进行油气预测是一种典型的非线性分类器设计问题,它具有已知样本数较少、特征个数较少等特点,文章据此提出了一种基于特征扩展和特征选择的改进SVM方法。该方法将原始特征通过非线性变换到高维空间,然后应用线性SVM进行特征选择,并同时计算降维过程中各个特征子集对应的留一法错误率,最后选择错误率较小的特征子集来设计线性SVM分类器。在通用数据的实验中,这种方法仅仅用较为简单的多项式核函数就大大提高了分类器的泛化能力。与传统的模糊数学方法、神经网络方法和SVM方法相比,这种方法在四川观音场构造的碳酸岩盐储层数据的预测误差降低了50%,是一种有效的油气预测方法。

主题词　向量计算机　地震数据处理　油气藏　预测　观音场气田

Vapink从统计学习理论的角度出发,提出了支持向量机(support vector machine,SVM)方法[1]。SVM已经广泛地应用在分类和函数回归等问题,并且取得了较好的效果[1~3]。针对油气预测这种有监督的分类问题,SVM比神经网络等传统的学习方法具有以下几个特点和优点[1~3]:①它综合考虑了分类器的经验风险和置信风险,在一定概率意义下是推广能力最好的分类器。这种结构风险最小化的设计思路可以避免陷入欠学习和过学习等情况。②它有全局最优解,不会陷入局部最优。③它利用核函数的方法解决了非线性的分类问题,其算法复杂程度主要取决于训练样本的个数,而与特征维数基本无关。

油气预测需要综合少数的测井信息和和大范围地区的地震信息来预测整个地区储集层的含油气状况。这是一个典型的有监督分类问题,其特殊性在于样本(井)数量较少,并且地震特征也不能完全体现油气聚集的情况。先前的研究主要侧重于两种方法,一种是模糊数学的方法[4],一种是神经网络的方法[5~6]。模糊数学的方法需要事先通过经验建立各种地震信息与含油气之间的模糊隶属函数,再根据这些函数关系来预测油气聚集程度。神经网络的方法应用在油气预测领域同样也容易遇到在其它领域一样的困难,例如如何确定神经网络的结构,如何选择训练参数来控制过学习情况等等。与神经网络相似,SVM在进行分类器设计时也需要确定核函数K(类似于神经网络的结构)和控制经验风险和置信风险之间的折衷参数C。目前,一些SVM软件可以根据样本信息提供参考的C参数[7],并且对于大多数分类问题,应用不同的核函数基本能取得相差无几的效果[1,3]。本文中SVM特征扩展和特征选择的方法则进一步简化了核函数的选择。从理论和实验的角度均证明它的推广能力要比传统SVM有所提高。由于油气预测问题一般是小样本学习问题,加之近年来出现了SVM^{light}等许多SVM快速算法[7],使得本文提出的预测方法计算量很小。

SVM和改进的SVM算法

假设d维空间的N个两类样本的特征分别是$x_n \in \boldsymbol{R}^d$,其类别标签为$y_n \in \{+1,-1\}$($n=1,2,\cdots$,

*本研究获国家"十五"科技攻关(2001BA605A09)、中国石油天然气集团公司创新基金和国家教育部留学回国人员科研启动金资助项目资助。

作者简介:姚凯丰,1976年生,清华大学自动化系在读博士;主要从事机器学习方法和应用及地震信息处理研究。地址:(100084)北京市清华大学。电话:(010)62788078。E-mail:kfyao98@mails.tsinghua.edu.cn

N)。SVM采用类边缘最大化准则来选择分类面。对于线性不可分的情况,用带有错分惩罚的"软边缘"来代替边缘,并且以常数 C 来控制惩罚程度。这个优化问题可以变换为一个 N 维空间的受约束的二次优化问题,最终给出的决策函数为:

$$f(x)=\mathrm{sign}\Big(\sum_{n=1}^{N}\alpha_n y_n x_n{}^T x+b\Big) \tag{1}$$

式中 α_n 和 b 是 Lagrange 乘子,可以通过优化求得,sign 是取符号函数。对于非线性问题,SVM 假设将样本映射到某个高维空间之后它将变为一个线性问题。假设这个向高维空间的映射函数为 $\psi:\boldsymbol{R}^d\rightarrow\boldsymbol{R}^D$,其中 D 为映射之后空间的维数,那么只需要将所有训练和分类过程中的 x 替换为 $\psi(x)$ 即可。在 SVM 的优化和决策过程中,只涉及到样本之间的内积,所以可以用一个代表 $\boldsymbol{R}^D$ 空间的内积函数(称为核函数)$K(u,v)$,$K:\boldsymbol{R}^d\times\boldsymbol{R}^d\rightarrow\boldsymbol{R}$ 来代替 $\boldsymbol{R}^d$ 空间的内积函数 $u^{\mathrm{T}}v$,$\forall u,v\in\boldsymbol{R}^d$。这种通过核函数隐式进行的映射不会增加算法的复杂度和存贮要求,避免了"维数灾难"[1~3]。最终非线性的 SVM 分类器决策函数为:

$$f(x)=\mathrm{sign}\Big[\sum_{n=1}^{N}\alpha_n y_n K(x_n,x)+b\Big] \tag{2}$$

常用的核函数包括形如 $K(u,v)=(1+u^Tv)^p$ 的 p 阶多项式核函数和方差为 σ 的径向基(radius basis function,RBF)核函数 $K(u,v)=\exp(-\|u-v\|^2/\sigma^2)$ 等。

多项式核函数将样本映射到多项式空间,例如对于二维空间中的样本 $(x_1,x_2)^T$,2 阶多项式的映射结果为 $(x_1,x_2,x_1x_2,x_1^2,x_2^2)^T$。统计学习理论认为置信风险的上界与特征维数有关,维数越低,风险的上界越低[1,2]。实际的经验也告诉我们,特征太多容易导致过学习。本文在扩展之后的特征空间用线性 SVM 做进行特征选择。式(1)可以重写为:

$$f(x)=\mathrm{sign}(w^Tx+b) \tag{3}$$

其中 $w=\sum_{n=1}^{N}\alpha_n y_n x_n$。如果特征向量 x 已经归一化,则 w 中绝对值较大的分量对应的特征对分类器的决策有更大的贡献。为了获得较好的推广能力,需要同时控制经验风险和置信风险。实验中用留一法错误率来代表经验风险,用特征子集的维数来代表置信风险。因此可以推导出下面的特征选择过程[8,9]:①训练线性 SVM,并且计算它的留一法错误率;②计算权重 w;③剔除权重较小的一个或几个特征;④如果还有特征,则转到①,否则结束。

这个特征选择的过程最终给出留一法错误率随着特征维数变化的曲线。一般来说,错误率的变化有这样的趋势,随着特征维数的降低,错误率一开始降低,达到一定程度之后又升高[9]。因此可以选择错误率达到最小处的特征子集来做最终的训练,如果几个特征子集对应的错误率相差不大,则取维数较小的特征子集。这种方法从另外一个角度实现了经验风险和置信风险最小化,并且在一些常用的模式识别基准数据库上取得了令人满意的效果[9]。

预测效果分析

本文应用 SVM 特征选择的方法对四川观音场构造的地震数据进行了油气预测,并且与一些其它已经发表的方法结果进行了比较。这个数据集中一共有 17 口井,分属于"干井"、"低产井"和"高产井"三类。每口井给出了 3 个井旁道的各 7 个特征,分别为振幅、相位、频率、构造曲率、层速度、视极性和低速层厚,具体数据请参见文献[4]。文献[4]提出了一种模糊数学的分类器设计方法,取其中的 13 口井作为训练样本,其余 4 口井作为测试样本[4]。另外有一些基于神经网络的方法取前面 10 口井作为训练样本,采用后向传播神经网络和自组织神经网络设计分类器,并预测后 7 口井[5,6]。

本文取前 10 口井作为训练样本,采用三道数据的平均值作为原始特征,用三阶多项式把 7 维特征扩展到 119 维特征,然后将每一个特征都归一化到[-1,+1]区间。根据 SVM 特征选择过程给出的留一法错误率曲线,从中选取了 2 个和 8 个特征设计了 2 个线性分类器。选择 2 个特征时,样本在这个特征空间中的分布和根据训练样本设计的线性分类器如图 1。图中三类样本分别用三种符号表示,带有圆圈的样本是 7 个测试样本。2 个线性分类器(如虚线表示)将特征空间分为三个区域。从图中看出,三类井在这个特征空间基本可以线性分离,设计的分类器仅仅将 2 个低产井误分为高产井。本文方法结果与其它方法的比较见表 1。表中多项式 SVM 采用三阶多项式核函数,RBF SVM 核函数的参数=1。对比传统 SVM 可以看出,特征扩展和特征选择过程有效的提高了分类器的推广能力。本方法选择 2 个特征和 8 个特征均只错分 2 口井,相比其它方法

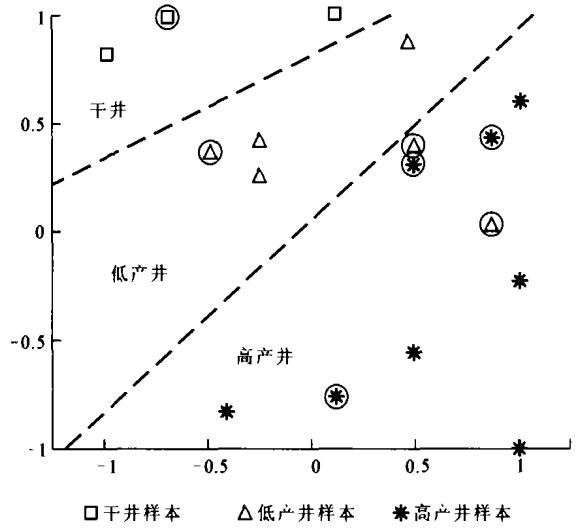

图 1　样本在二维特征空间中的分布示意图

的 3 到 5 个错分数，错误率降低了 50%左右。由于样本数较少和原始特征维数较低，整个训练和预测过程在微机上只需要运行约 10 min。

结　论

本文提出的基于 SVM 特征扩展和特征选择的油气识别方法，只需要调节多项式特征扩展的多项式阶数，而根据经验，一般可以固定这个阶数为 3。它不需要有关地质的先验知识，也无需调节很多参数，基本可以实现自动的油气预测。由于油气预测具有样本数较少、维数较低等特点，因此运算速度很快。在四川观音场构造的实际资料上的测试的结果比传统的模糊数学方法和神经网络的方法错误率降低 50%左右。

表 1　SWM 特征选择方法与其它方法进行油气预测的结果比较

井　号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	总错分个数
真实值	0	0	2	2	2	2	1	1	1	2	2	1	2	1	2	0	1	
模糊数学[4]	0	~~1~~	~~1~~	2	2	2	1	1	~~2~~	2	2	1	2	1	2	0	~~2~~	4
BP 神经网络[5]	0	~~2~~	~~1~~	2	2	2	1	~~2~~	1	2	2	1	1	1	2	0	1	3
SOM 神经网络[6]	0	~~1~~	~~1~~	2	2	2	1	1	~~2~~	2	~~1~~	1	2	1	~~1~~	0	1	5
多项式 SVM	0	0	2	2	2	2	1	1	1	2	~~0~~	1	2	~~2~~	~~0~~	0	~~0~~	4
RBF SVM	0	0	2	2	2	2	1	1	1	2	~~1~~	1	2	~~2~~	~~1~~	0	~~0~~	4
2 个特征	0	0	2	2	2	2	1	1	1	2	2	~~2~~	2	~~2~~	2	0	1	2
8 个特征	0	0	2	2	2	2	1	1	1	2	~~1~~	1	2	1	~~1~~	0	1	2

注：0、1、2 分别代表干井、低产井和高产井；数字中有横线的为错分数据。

参 考 文 献

1　Vapnik V 著，张学工译. 统计学习理论的本质. 北京：清华大学出版社，2000

2　张学工. 关于统计学习理论和支持向量机. 自动化学报，2000；26(1)：32～42

3　Burges C J C. Tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery，1998；2(2)：121～167

4　肖辞源，朱白文. 综合多种地震信息预测油气富集区的模糊数学方法. 石油地球物理勘探，1990；25(2)：191～200

5　蔡煜东，宫家文，甘骏人等. 应用人工神经网络方法预测油气. 石油地球物理勘探，1993；28(5)：634～638

6　许建华，蔡瑞. 有监督 SOM 神经网络在油气预测中的应用. 石油物探，1998；37(1)：71～76

7　Joachims T Making large-scale SVM learning practical. Advances in Kernel Methods-Support Vector Learning，Sch? lkopf B，Burges C and Smola A (ed.)，MIT Press，1999

8　Guyon I，Weston J，Barnhill S *et al*. Gene selection for cancer classification using support vector machines. Machine Learning，2002；46(1)：389～422

9　Yao Kai-Feng，Lu Wen-Kai，Zhang Shan-Wen *et al*. Feature expansion and feature selection for general pattern recognition problems. IEEE Int Conf Neural Networks & Signal Processing，Nanjing，China，2003；(12)：29～32

（收稿日期　2004-03-19　编辑　韩晓渝）