文章编号：1001-1749(2023)05-0566-13

基于极端随机树算法的流体识别研究

饶骁驰[1]，杨 昊[1]，喻 辉[2]，文 武[1]，周 航[1]，陈 敏[1]
(1.成都信息工程大学 计算机学院，成都 610225；
2.78111 部队，成都 610011)

摘 要： 对储层内流体进行识别存在较大的不确定性，多属性融合进行流体识别显得非常必要。机器学习方法已经日趋成熟，但在流体识别方面的应用还较少见。这里引入一种实现简单、具有较强普适性的方法——极端随机树方法对流体进行识别，比较了本方法较传统的机器方法的识别准确率优势，并通过均方误差及错误率的验证证实了本方法对于流体识别的准确性。最后将本方法应用于南海某油气田，良好的识别效果证实了本方法对于流体识别的有效性。

关键词： 流体识别；极端随机树；机器学习；多属性融合
中图分类号： P 631.4 **文献标志码：** A **DOI：**10.3969/j.issn.1001-1749.2023.05.02

0 引言

随着勘探程度的增加，一些易于发现的油气藏已勘探殆尽，深海勘探、非常规油气藏勘探已逐步成为勘探热点。十几年前，人们主要关注储层的研究，即储层物性(主要是孔隙度)究竟如何，储层厚度和范围的大小等。常规的地震属性分析、阻抗反演等方法在这些研究领域起了很大作用。近十年来，一些学者逐渐将目光转向储层内流体的识别，油、气、水的识别，这些研究目前已取得了很大地进展，叠前反演、AVO 分析、低频伴影分析等技术相继产生、发展。但随着致密油、页岩油等勘探领域逐渐进入 人们的视野，一些新的需求随之出现。相对于常规油气藏，以上提到的这些非常规油气藏孔隙度低、非均质性强、流体粘滞性偏高、地震响应特征不明显。因此，仅凭借单一属性、参数或方法进行流体识别存在较大的不确定性，多属性融合进行流体识别是大势所趋。

在储层流体识别中传统非机器学习方法限于效率低和工作量大，仅应用于对有利区或目标区的研究。同时不同的流体因子对不同区域储层含流体的敏感程度表现不同，传统非机器学习方法需要依靠人为干预，因此人为主观因素过多，干扰因素较多，不确定性强。

Fung[1](2001)在支持向量机(SVM)的基础上提出了近似支持向量机(PSVM)，该机器语言可以利用目标体的多种属性计算出反映该目标体属性特征的最优规则。在对大数据训练集进行处理时，近似支持向量机在判别准确度不低于支持向量机的前提下，在计算效率上具有明显优势，适合对具有海量数据特征的叠前地震资料进行判别处理。

近年来，机器学习已被证明在工程中具有广泛的用途(如金融领域、制造领域和零售领域)，并且正在稳步发展推进到新的领域。气象领域 Han L 等[2]利用机器学习进行实时风暴运动预测；视觉心

收稿日期：2022-05-23
基金项目：四川省重点研发项目(2020YFS0355，2020YFG0479)
第一作者：饶骁驰(1998—)，男，硕士，主要研究方向为大数据可视化与高性能计算机和机器学习，E-mail：rxc_vince@163.com。

理学领域，Robert M. French 等[3]利用机器学习进行视觉心理预测；生物学领域，Crozier T W M 等[4]利用机器学习预测蛋白质的构成；地学领域，Vedangi Godse 等[5]也利用机器学习预测地震活动。

目前在储层预测领域应用比较广泛的机器学习方法有支持向量机[6-8]、神经网络[9-12]、随机森林[13]等方法。这些方法主要是通过从测井资料中提取揭示储层特征的参数作为输入参数，利用这些智能方法建立多参数与储层物性之间的映射关系，进而开展储层预测。但值得注意的是，目前这些方法在流体识别方面的应用还较少见。为此，笔者充分比较了常见的几种机器学习算法在流体识别中的应用效果，最终选取了极端随机树方法（Extremely Randomized Trees）进行流体识别，该方法具有几个显著特征：

1）数据的准备往往是简单或者是不必要的，对于不平衡的数据集，可以平衡误差。其他的技术要求先把数据标准化（如去掉多余的或者空白的属性）。

2）易于理解和实现，在学习过程中不需要使用者了解很多的背景知识，能够直接体现数据的特点，通过解释后都有能力去理解决策树所表达的意义。

3）易于通过静态测试来对模型进行评测，可以测定模型可信度。如果给定一个观察的模型，则根据所产生的决策树很容易推出相应的逻辑表达式。

1　方法

1.1　极端随机树（ET）

1.1.1　算法介绍

Geurts P 等[14]提出 ET 或 Extra－Trees（Extremely randomized trees，极端随机树方法）。根据经典的自上而下的方法，极端随机树构建了一系列“自由生长”的回归树集合。该方法中的每一棵回归树用的都是全部训练样本，用$\{T(K,X,S)\}$表示。其中 T 表示最后的分类器模型，S 为数据样本集，K 为基分类器的数量（基分类器数量根据所要区分的结果确定，其中基分类器数量为 3，用以区分水层、气层和油层）。每个基分类器根据输入样本 $X=\{x_1,x_2\cdots,x_n\}$产生预测结果，最终通过投票确定最后的分类类别。

1.1.2　算法实现

Step1：给点原始样本数据集 S，样本数量为 N，特征数量为 M，在极端随机树的分类模型中，每个基分类器都使用全部的样本进行训练。

Step2：基于训练数据集生成决策树，生成的决策树要尽量大。

选择具有最小 Gain_GINI 的属性及其属性值，作为最优分裂属性以及最优分裂属性值。Gain_GINI 值越小，说明二分之后的子样本的“纯净度”越高，即说明选择该属性（值）作为分裂属性（值）的效果越好。

GINI 计算如下[15]：

$$GINI(S)=1-\sum P_k^2 \tag{1}$$

其中，P_k 表示分类结果中第 k 个类别出现的频率（数量在所有样本中所占的比例）。

对于含有 N 个样本的样本数据集 S，根据属性 A 的第 i 个属性值，将样本数据集 S 划分成两部分，则划分成两部分之后，Gain_GINI 计算如下[14]：

$$Gain\_GINI_{A,i}=\frac{n_1}{N}GINI(S_1)+\frac{n_2}{N}GINI(S_2) \tag{2}$$

其中：n_1、n_2 分别为样本子集 S_1、S_2 的样本个数。

对于属性 A，分别计算任意属性值将数据集划分成两部分之后的 Gain_GINI，选取其中的最小值，作为属性 A 得到的最优二分方案：

$$\min(Gain\_GINI_{A,i}(S_1)) \tag{3}$$

对于样本集 S，计算所有属性的最优二分方案，选取其中的最小值，作为样本集 S 的最优二分方案：

$$\min_{A\in Atribute}\ (\min_{i\in A}(Gain\_GINI_{A,i}(S))) \tag{4}$$

所得到的属性 A 及其第 i 属性值，即为样本集 S 的最优分裂属性以及最优分裂属性值。

Step3：用验证数据集对已生成的树进行剪枝并选择最优子树，这时损失函数最小作为剪枝的标准。

从原始决策树 T_0 开始生成第一个子树序列$\{T_0,T_1,\ldots,T_n\}$，其中 T_{i+1}从 T_i 产生，T_n 为根节点。在剪枝的过程中，计算损失函数[14]：

$$C_\alpha(T)=C(T)+\alpha\mid T\mid \tag{5}$$

$\alpha\geqslant 0$，$C(T)$ 为训练数据的预测误差，$|T|$为模型的复杂度。

将 α 在其取值空间内划分为一系列区域，在每个区域都取一个 α 然后得到相应的最优树，最终选择损失函数最小的最优树。

在选出 α 之后，计算该 α 对应的使损失函数最小的子树。即从树的根节点出发，逐层遍历每个内

部节点,计算每个内部节点处是否需要剪枝。

Step4:重复执行 Step 1、Step 2 和 Step3 迭代 K 次,生成 K 棵决策树,生成极端随机树。

Step5:将生成的极端随机树使用测试样本生成预测结果,将所有基分类器的预测结果进行统计,经过投票决策的方法产生最终的分类结果。

1.2 传统的机器学习算法

1.2.1 BP 神经网络

BP 算法是一种最有效的多层神经网络学习方法,其主要特点是信号前向传递,而误差后向传播,通过不断调节网络权重值,使得网络的最终输出与期望输出尽可能接近,以达到训练的目的。其优点是:①并行分布处理能力强,分布存储及学习能力强;②具备联想记忆的功能;③对噪声数据鲁棒性和容错性较强;④能逼近任意非线性关系。

神经网络需要大量的参数,如网络拓扑结构、权值和阈值的初始值。不能观察之间的学习过程,输出结果难以解释,会影响到结果的可信度和接受程度。学习时间过长,甚至可能达不到学习的目的。

1.2.2 支持向量机(SVM)

支持向量机是 Vapnik 等[16]在统计学习理论的基础上提出的一种采用结构风险最小化准则的新的学习方法。相对于传统采用经验风险最小化准则的学习方法,支持向量机具有更强的泛化能力。由于支持向量机是一个凸二次优化问题,所以它可以找到作为全局最优解的极值解。支持向量机可以用于对目标工区的判别分类。

此方法的优点是可用于线性/非线性分类,也可以用于回归,泛化错误率低,计算开销不大,结果容易解释。可以解决小样本情况下的机器学习问题,高维问题,避免神经网络结构选择和局部极小点问题。缺点是对缺失数据敏感,对于类域的交叉或重叠较多的待分样本集较难分类。

1.2.3 K 最近邻(KNN)

KNN 是通过测量不同特征值之间的距离进行分类。它的思路是:如果一个样本在特征空间中的 k 个最相似(即特征空间中最邻近)的样本中的大多数属于某一个类别,则该样本也属于这个类别。其中 k 通常是不大于 20 的整数。KNN 算法中,所选择的邻居都是已经正确分类的对象。该方法在定类决策上只依据最邻近的一个或者几个样本的类别来决定待分样本所属的类别。

此方法的优点是简单、有效,重新训练的代价较低。该算法比较适用于样本容量比较大的类域的自动分类。其缺点是样本不平衡时,预测偏差比较大(即某一类的样本比较少,而其它类样本比较多)。计算量大,每一次分类都会重新进行一次全局运算。

1.3 极端随机树的优点

极端随机树不同于传统机器学习方法,它对于数据的准备工作较为简单甚至不是必要的,极端随机树不需要预处理数据,并且在相对短的时间内能够对大型数据源做出可行且效果良好的结果。各种机器学习方法对比如表 1 所示[17]。

神经网络方法在进行调参时需要调整网络拓扑结构、权值和阈值的初始值、学习率、迭代次数等参数,调参过程较为复杂。支持向量机调参时需要考虑惩罚系数、核函数、核函数系数等参数。K 最近邻方法在调整参数时需要考虑选取几个邻居、邻居权值、距离等参数。而极端随机树只需要调整好合适的最大迭代次数就可以了,因此是一种简单有效的分类方法。

表 1 常用机器学习方法对比

Tab.1 Comparison of common machine learning methods

方法	优点	缺点
极端随机树	不需要预处理数据、分类方式简单有效	
神经网络	有自学习功能和联想存储功能	依赖于全部数据
决策树	直观决策规则	容易过拟合、处理缺失数据困难
SVM	不依赖所有数据	对于大量的预测样本效率低

表 2 研究区钻井的测井解释表

Tab.2 Log interpretation table of drilling in the study area

井名	平均密度/ $g \cdot cc^{-1}$	平均孔隙度/%	平均渗透率/Md	流体类型	储层厚度/m
井 A	2.5	14.3	20	天然气	31.8
井 B	2.45	19.3	62.5	气水同层	33.3
井 C	2.3	20	90	挥发油	17.5

2　测井数据测试及训练集建立

2.1　研究区概况

研究区域位于南海某油气田，储层岩性是砂岩，上方覆盖为泥岩，地质资料显示该地区为浅海三角洲沉积。储层的层系为珠江组，岩石物理试验的储层平均孔隙度为 17.41%，平均渗透率为 80 mD，平均密度为 2.17 g/cm^2，总体而言储层的物性较好。其中井 A 和井 C 在目的层顶部钻遇气层，井 B 在目的层钻遇含水层。为了检验方法的有效性，我们首先利用 A 井、B 井的测井数据进行分析，优选方法并建立训练集，C 井留作验证井。表 2 为井 A、井 B 和井 C 的测井解释表[18]。

2.2　硬件条件

CPU：intel core i7 — 7700k 4.4G；内存：4X8G；显示卡：两个 NV GTX 1070；硬盘：HDD 一个，SSD 两个。

2.3　测试模型设计

模型设计流程见图 1。

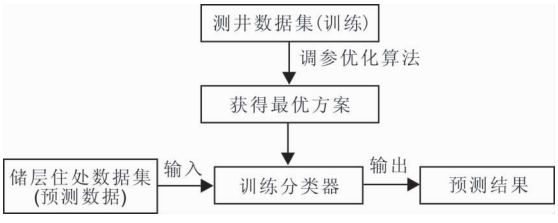

图 1　模型设计流程

Fig. 1　Model design filow

2.4　测试方法评价

2.4.1　均方误差

度量模型性能的一种方法是计算模型在测试集上的均方误差，如果 $\hat{y}^{\text{test}}$ 表示模型在测试集上的预测值，那么均方误差表示为式(6)[19]。

$$MSE_{\text{test}} = \frac{1}{m}\sum_i (\hat{y}^{\text{test}} - y^{\text{test}})_i^2 \tag{6}$$

当 $\hat{y}^{\text{test}} = y^{\text{test}}$ 时，可以看出误差降为 0，所以当预测值和目标值之间的欧几里得距离增加时，误差也会增加。

2.4.2　错误率

当数据重叠较多的时候，很难从均方误差来分辨算法的效果，因此引入了错误率(error rate)配合度量。错误率是分类错误的样本数占样本总数的比例。错误率是使用最普遍、最简单同时又是最直接的分类指标。其计算方法为[19]：

设测试样本集 $T=\{(x_1,y_1),\cdots(x_n,y_n)\}$，其中 x_n 为该样本 i 的输入特征，y_i 为样本的真实标签。

T 的预测结果：$py_n=\{py_1,py_2,\cdots,py_n\}$，其中 py_i 表示模型对 T 中第 i 个样本的预测结果。

$$error\_rate = \frac{\sum\{1 \mid y_i \neq py_i\}}{\sum\{1 \mid y_i = y_i\}} \tag{7}$$

2.5　最优方法的确定

对比极端随机树与传统的机器学习算法分类预测的效果。笔者主要利用了泊松比(σ)、高灵敏度流体识别因子(high sensitive fluid identification factor，HSFIF)、流体属性(Mobility Attribution，MA)对流体进行识别。

1)泊松比[20]：

$$\sigma = \frac{\lambda}{2(\lambda+\mu)} = \frac{(I_P/I_S)^2-2}{2[(I_P/I_S)^2-1]} \tag{8}$$

式中：σ 为泊松比；λ 为拉梅常数；μ 为剪切模量，泊松比属于波阻抗量纲的零次方类流体识别因子。大量的研究和应用表明，泊松比对储层的含气性有很好的响应，在相同的孔隙度的情况下，当砂岩储层含气时，泊松比可以从 0.3～0.4 左右降至 0.1 左右，而自然界中的岩石泊松比变化范围在 0～0.5 之间，所以泊松比经常被应用于含气储层的识别与刻画中。

2)高灵敏度流体识别因子：

$$\sigma_{HFIFI} = \frac{I_P}{I_S}I_P^2 - CI_S^2 \tag{9}$$

式(9)由贺振华等[21]提出，是从 Gassmann 公式出发所提出的流体识别因子。式(9)中的 C 为调节参数，当 $C=2\dfrac{I_P}{I_S}$ 时，式(9)变为：

$$\sigma_{HFIFI} = \frac{I_P}{I_S}(I_P^2 - 2I_S^2) \tag{10}$$

式(10)将波阻抗量纲的零次方类与流体属性的优点结合起来，突出了纵波阻抗 I_P 的作用。

3)流体属性：

$$\rho f = I_P^2 - CI_S^2 \tag{11}$$

式(11)由 Russell 等[22]提出，式(11)中的 f 代表 Gassman 方程中的流体因子项，C 为调节参数。

对评价结果进行研究，将属性两两组合进行计算并显示。

1)高灵敏度流体识别因子与泊松比交会分析。

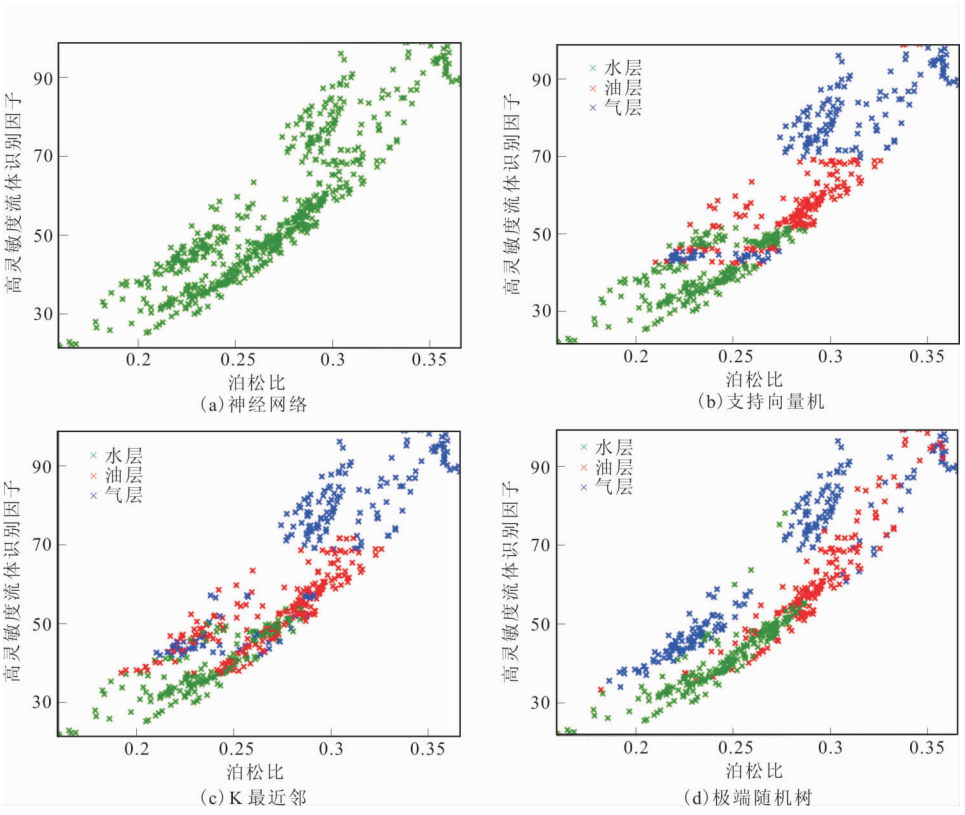

图 2 基本方法

Fig. 2 Basic method

表 3 4 种方法结果比较

Tab. 3 Comparison of 4 methods

方法名	均方误差（均值）	均方误差（方差）	错误率
神经网络	−1.456604	0.321739	0.627119
支持向量机	−0.597688	0.171491	0.384610
K 最近邻	−0.556934	0.126515	0.453925
极端随机树	−0.174359	0.056750	0.256420

表 4 正态化数据后结果比较

Tab. 3 Comparison of results after normalizing data

方法名	均方误差（均值）	均方误差（方差）	错误率
神经网络	−0.228966	0.092414	0.308136
支持向量机	−0.219591	0.071944	0.300796
K 最近邻	−0.181347	0.067014	0.251783
极端随机树	−0.177911	0.047204	0.208231

数据来自井 A、井 B 的目的层段。从表 3 及图 2 的结果显示可以看出，对于本类数据集合极端随机树的预测结果较其他几种方法优秀。支持向量机、最近邻方法对这类数据集错误率很高。而神经网络则达不到预测要求，原因主要在于原始数据中不同特征属性的度量单位不一样，相比较而言，极端随机树算法具有较强的鲁棒性。

对数据进行正态化处理，将训练数据集进行数据转换处理，所有的数据特征值转化成“0”为中位值、标准差为“1”的数据。数据优化结果如表 4、图 3 所示。

表 5 算法调参优化后结果比较

Tab. 5 Comparison of results after algorithm optimization

方法名	均方误差（均值）	均方误差（方差）	错误率
神经网络	−0.199231	0.088244	0.294686
支持向量机	−0.170269	0.079481	0.283567
K 最近邻	−0.170132	0.071498	0.272488
极端随机树	−0.159225	0.046483	0.153292

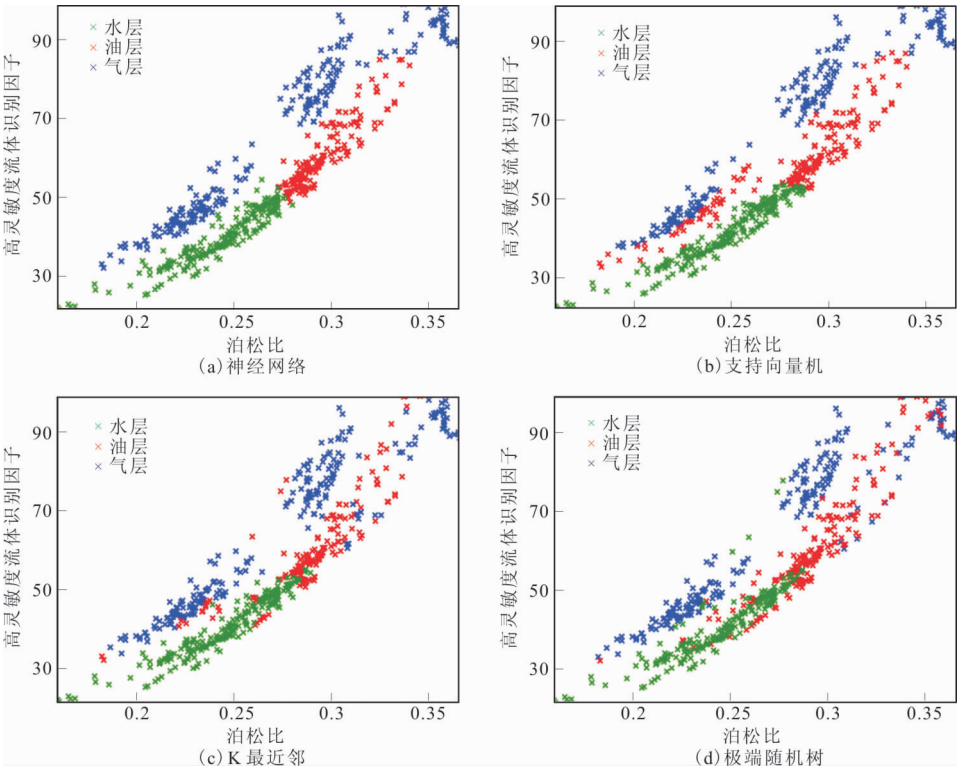

图 3　正态化数据

Fig. 3　Normalized data

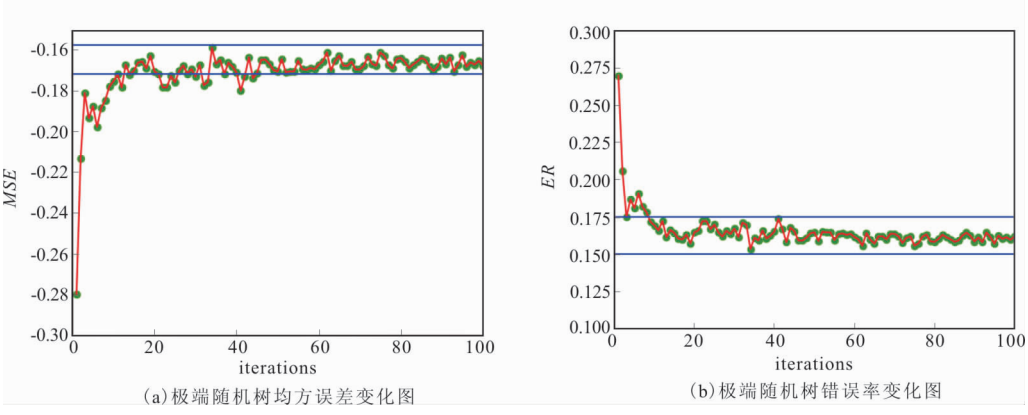

图 4　极端随机树迭代变化

Fig. 4　Extreme random tree iterative change

通过对于数据进行正态化处理以后，可以看出，支持向量机方法和神经网络方法效果得到了明显地提升，极端随机树的效果尽管提升不明显，但其分类结果仍然优于其他几种方法(表 5)。

对这些方法进行调参优化。从图 4 可以看出，当迭代次数大于 20 次后，误差基本接近稳定，迭代 34 次后，极端随机树方法达到最优解。后续均方误差在(－0.158～－0.172)附近震荡，错误率在(0.15～0.175)附近震荡，将 34 设置为本类数据应用极端随机树分类的最终迭代次数。

通过与正确分类结果比较可以看出，使用极端随机树算法不需要进行数据归一化和参数优化，就可以得到很好的预测效果，同时优化后结果依然优于其他方法。

2)流体属性与泊松比交会分析。前面分析可以看出，极端随机树算法在没有进行数据正态化处理和参数优化前，效果明显优于其他几种方法，对各个方法的最终优化结果进行分析，比较几种方法的预测效果。

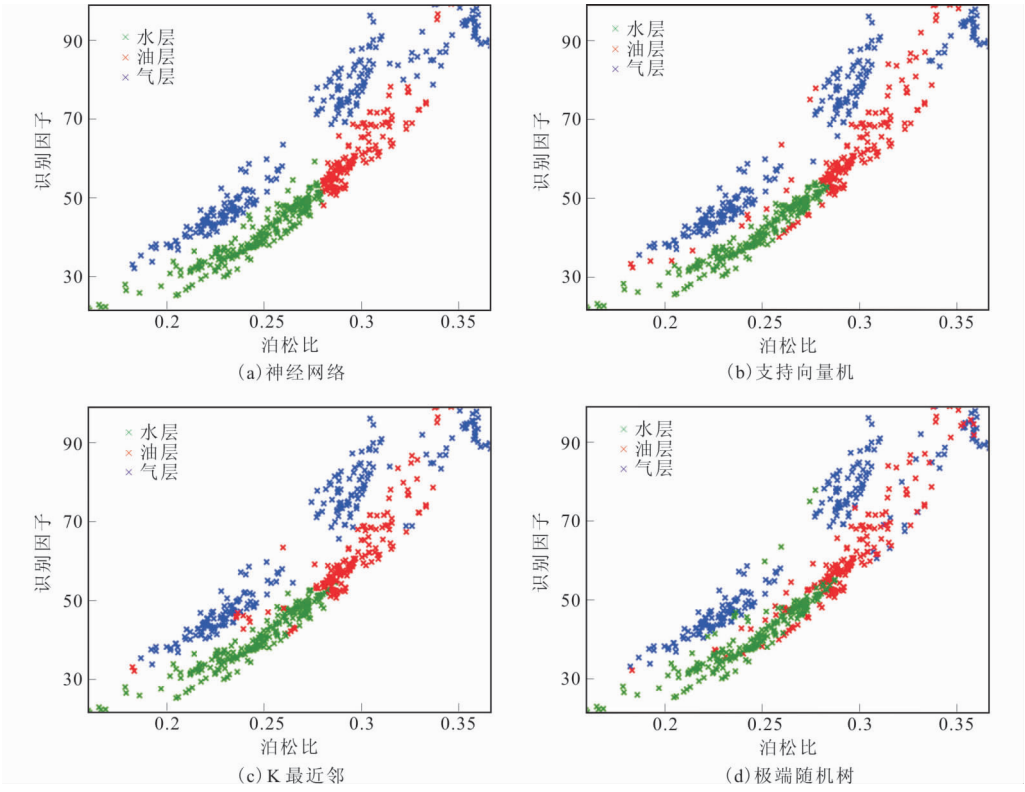

图 5 优化后高灵敏度流体识别因子与泊松比交会分析

Fig. 5 Intersection analysis of optimized high sensitivity fluid identification factor and Poisson's ratio

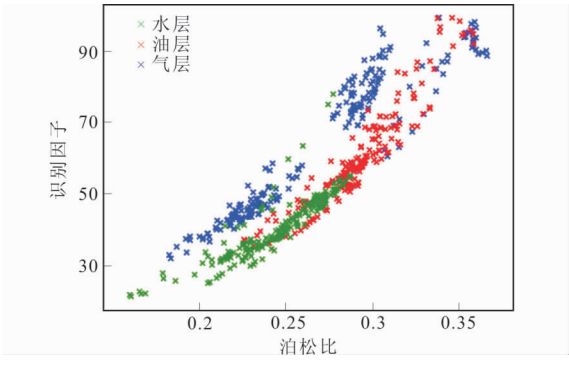

图 6 高灵敏度流体识别因子与泊松比分析真实结果

Fig. 6 Real result of high sensitivity fluid identification factor and Poisson's ratio analysis

表 6 流体属性与泊松比交会分析算法优化后结果比较

Tab. 6 Comparison of optimized results of intersection analysis algorithm for fluid properties and Poisson's ratio

方法名	均方误差（均值）	均方误差（方差）	错误率
神经网络	−0.209663	0.081992	0.169492
支持向量机	−0.177550	0.075709	0.319326
K 最近邻	−0.186270	0.067056	0.405845
极端随机树	−0.134952	0.065433	0.123211

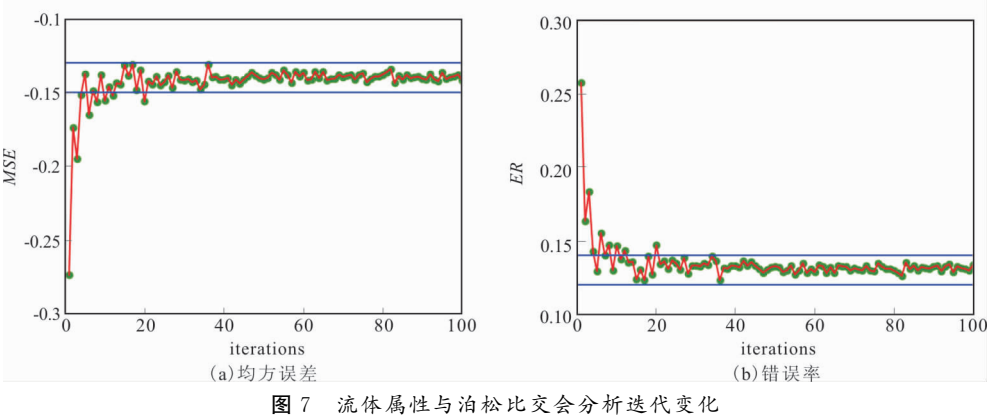

图 7 流体属性与泊松比交会分析迭代变化

Fig. 7 Iterative variation of fluid properties and Poisson's ratio intersection analysis

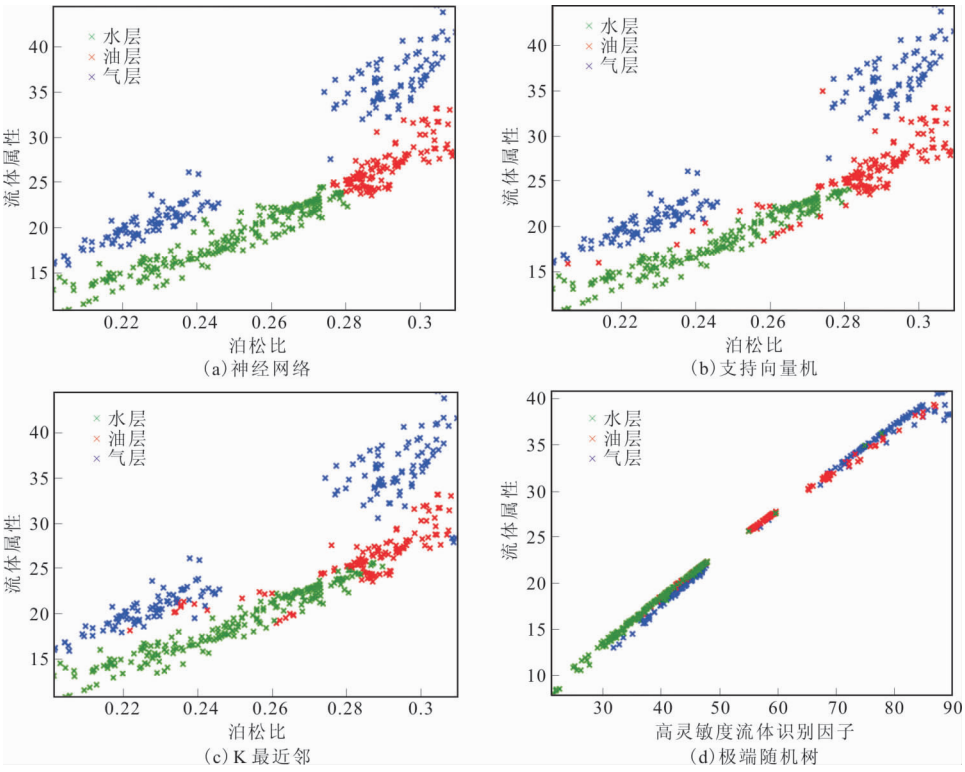

图 8　算法优化后流体属性与泊松比交会分析

Fig. 8　Intersection analysis of fluid properties and Poisson's ratio after algorithm optimization

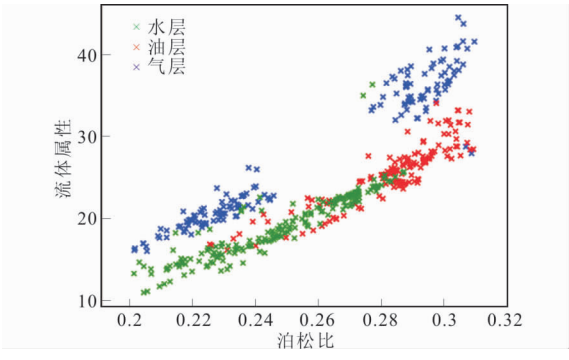

图 9　流体属性与泊松比交会分析真实结果

Fig. 9　Real results of intersection analysis of fluid properties and poisson's ratio

表 7　高灵敏度流体识别因子与流体属性交会分析优化结果比较

Tab. 7　Comparison of optimization results of high-sensitivity fluid identification factor and fluid attribute intersection analysis

方法名	均方误差（均值）	均方误差（方差）	错误率
神经网络	−0.707268	0.189877	0.352034
支持向量机	−0.268553	0.046296	0.275564
K 最近邻	−0.278191	0.063832	0.379152
极端随机树	−0.271901	0.046691	0.209111

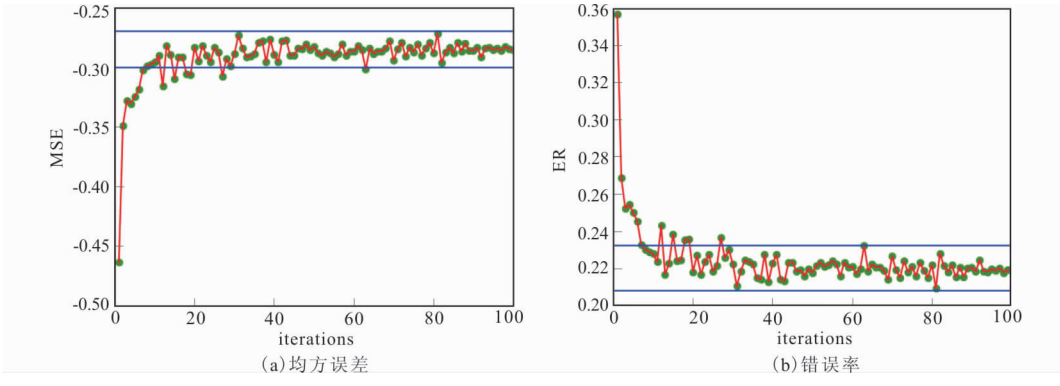

图 10　高灵敏度流体识别因子与流体属性交会分析迭代变化

Fig. 10　Iterative changes in intersection analysis of high sensitivity fluid identification factors and fluid attributes

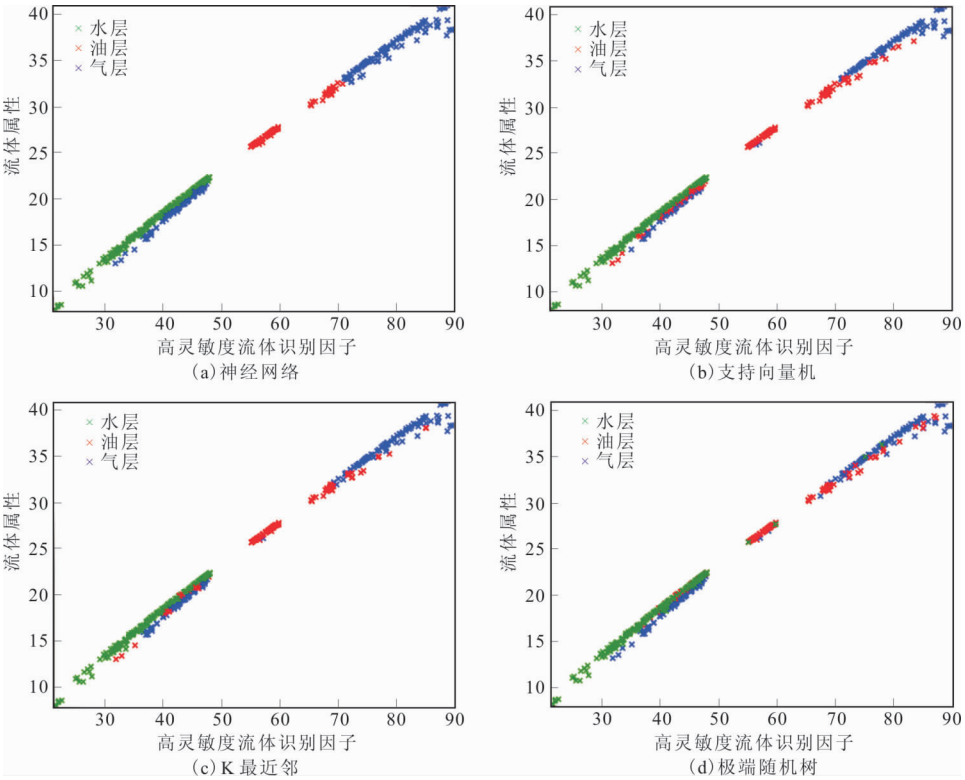

图 11 算法优化后高灵敏度流体识别因子与流体属性交会数据

Fig. 11 High sensitivity fluid identification factor and fluid attribute intersection data after algorithm optimization

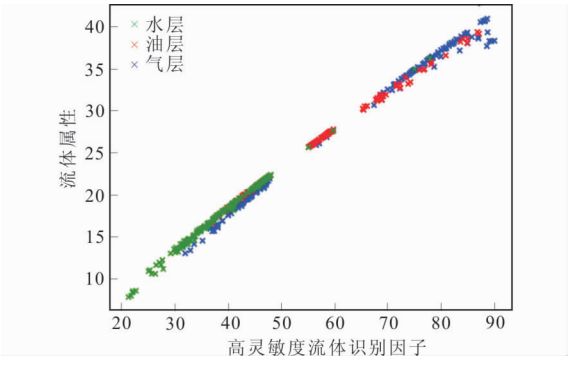

图 12 高灵敏度流体识别因子与流体属性交会数据真实结果

Fig. 12 High sensitivity fluid identification factor and real results of fluid attribute intersection data

从图 7 可以看出，迭代次数大于 20 次后，误差基本稳定，通过迭代 36 次，极端随机树方法达到最优解。后续均方误差在(－0.131～ －0.148)附近震荡，错误率在(0.12～0.14)附近震荡，将 36 设置为本类数据应用极端随机树分类的最终迭代次数。

从图 9 与图 8 可以看出，极端随机树效果优于其他几种方法。

3)高灵敏度流体识别因子与流体属性交会数据。从图 10 中可以看出，通过迭代 81 次，极端随机树方法达到最优解。为了简化计算过程，也可以将迭代 36 次的次最优解设置为最优解，与最终结果差距不大。均方误差在(－0.27～ －0.30)附近震荡，错误率在(0.208～0.232)附近震荡，将 36 设置为极端随机树的最终迭代次数。

表 8 样本训练结果比较

Tab. 8 Comparison of sample training results

过程	均方误差(均值)	均方误差(方差)	错误率
基础方法	－0.210251	0.071941	0.185506
正态化	－0.203752	0.076546	0.180579
调参优化	－0.183419	0.063197	0.163072

通过图 12 与图 11 的比较可以看出，尽管支持向量机通过参数优化，可以保证均方误差达到较优的值，但最终准确率上未能超过极端随机树算法。通过以上测试可以看出，极端随机树算法在一些类域的交叉或重叠较多的待分样本集分类有较明显的优势，且实现简单，因此选择极端随机树作为本研究的数据分类方法。

2.6 样本训练

将高灵敏度流体识别因子、泊松比、流体属性作

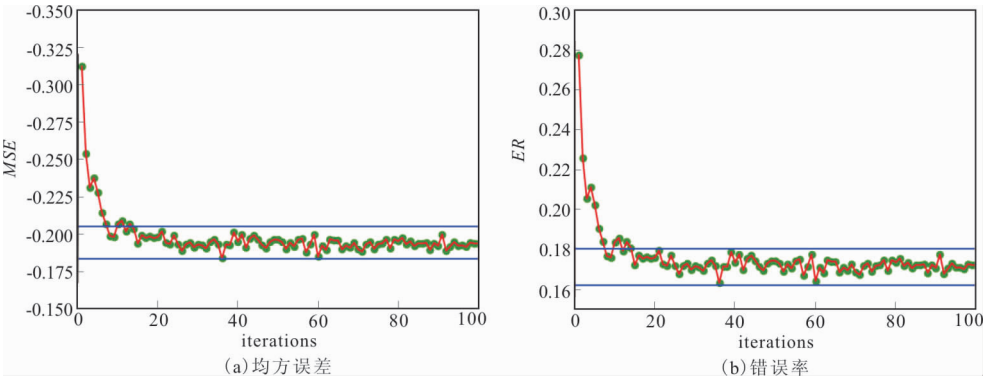

图 13　样本训练迭代变化

Fig. 13　Sample training iteration changes

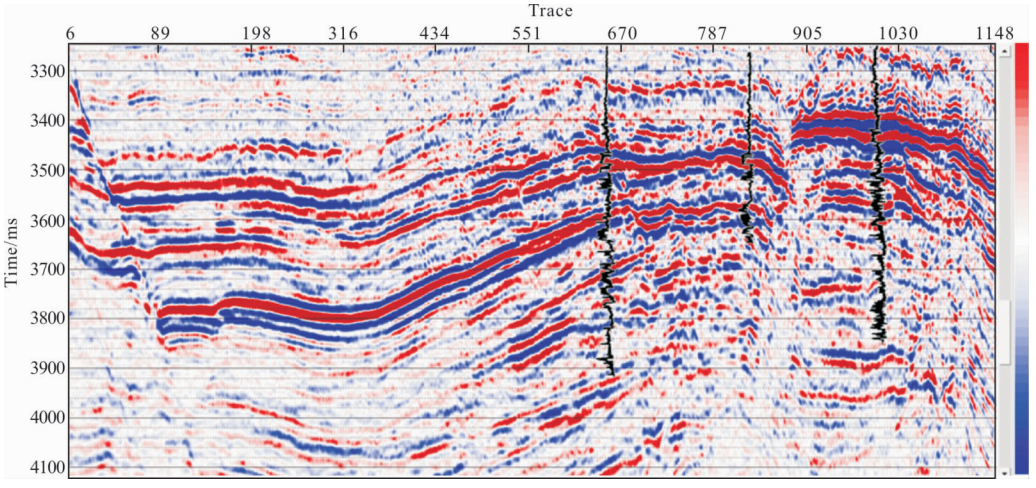

图 14　原始地震剖面

Fig. 14　Primary seismic profile

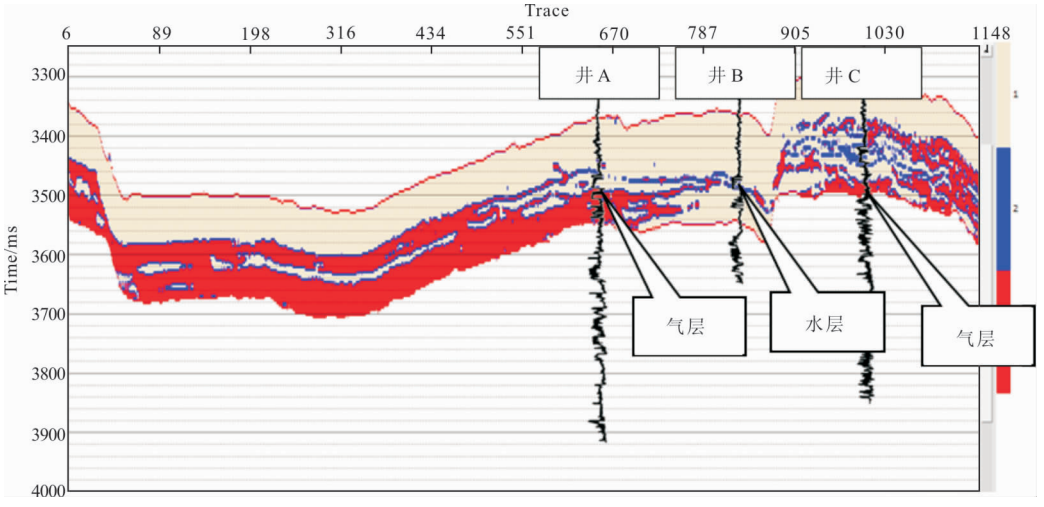

图 15　神经网络分类剖面

Fig. 15　Neural network classification profile

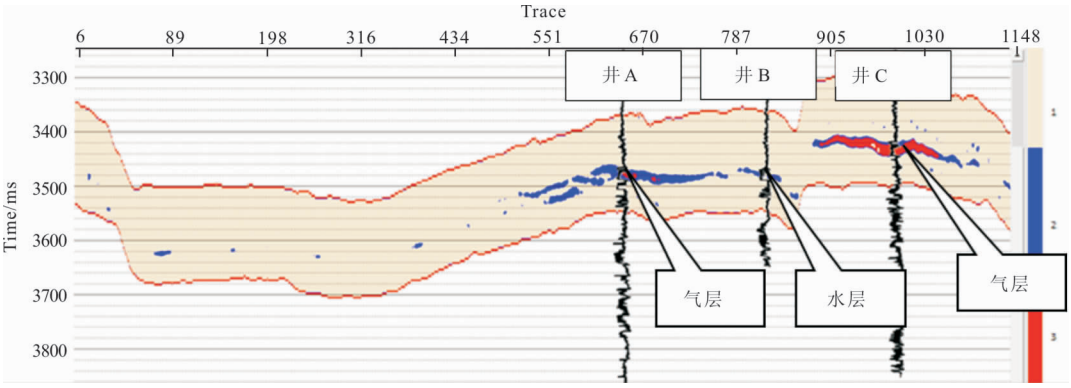

图 16 支持向量机分类剖面

Fig. 16 Classification profile of support vector machine

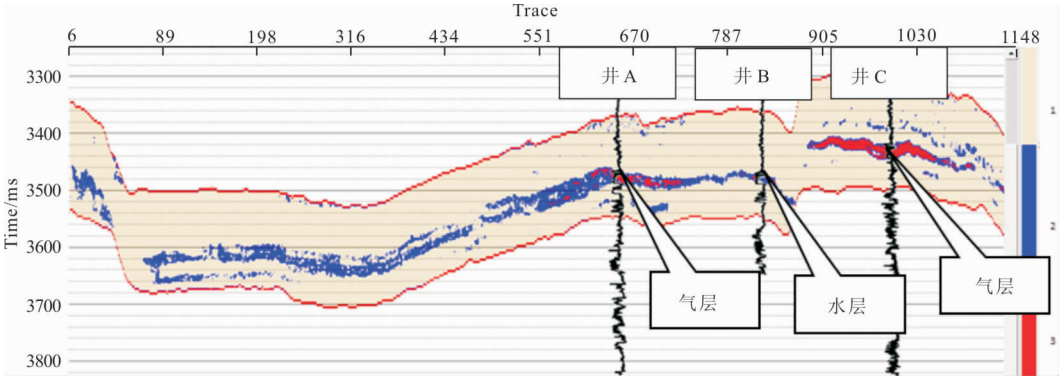

图 17 K 最近邻分类剖面

Fig. 17 K nearest neighbor classification profile

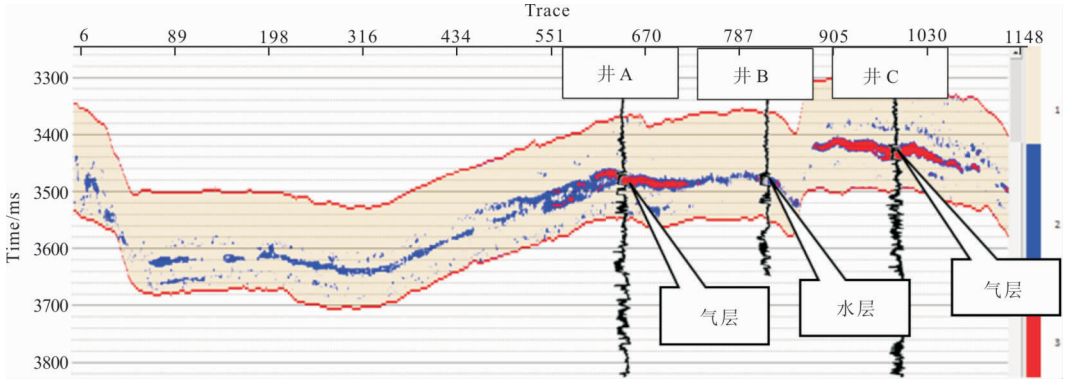

图 18 极端随机数分类剖面

Fig. 18 Extreme random number classification profile

为学习因子，流通属性作为预测因子，以井 A、井 B 数据为训练数据，井 C 数据为最终的验证数据，应用极端随机树方法进行学习(表 8)。

从图 13 可以看出，与前面 2 维参数类似，3 属性数据通过迭代 36 次可以达到最优解，将 36 设置为极端随机树的最终迭代次数，确定为最终模型，部署后应用于最终实例。

3 实例分析

图 14～图 18 分别为原始地震剖面、近似支持向量机分类剖面、神经网络分类剖面、KNN 分类剖面、极端随机数分类剖面，比较四幅图可以发现，原始剖面、支持向量机、神经网络分类剖面根本无法区分气层与水层，KNN 分类

表 9　实例预测结果比较

Tab. 9　Comparison of case prediction results

方法名	均方误差（均值）	均方误差（方差）	错误率
方法名	均方误差(均值)	均方误差(方差)	错误率
神经网络	−0.211879	0.096438	0.259132
支持向量机	−0.185936	0.078525	0.235641
K 最近邻	−0.190733	0.066260	0.201423
极端随机树	−0.183419	0.063197	0.163072

方法中气层与水层有一定差异，但差异不明显，而引入的极端随机数分类能很好地区分出气层和水层。

以上结果可以得出，在实际储层的气－水划分、气－水－油划分中，采用基于极端随机数的方法，对流体识别因子融合所划分的含气层、含水层、含油层，都与测井资料一一对应，但其他方法在做此区域的流体识别时会存在一定的不确定性，所以极端随机数所刻画的储层流体性质更加准确。从实例准确率分析数据及分类剖面图可以看出，极端随机数分类能很好地区分出气层和水层。

4　结论

在流体识别领域引入了一种有效的机器学习方法——极端随机数，本方法对于一些类域的交叉或重叠较多的待分样本集分类有较明显的优势，与传统机器学习方法相比较，极端随机树对于数据的准备往往是简单或者是不必要的，不需要预处理数据，并且在相对短的时间内能够对大型数据源做出可行且效果良好的结果。通过实例测试，可以看出本方法不仅部署简单，对于流体识别也有很好的效果。

参考文献：

[1] FUNG G, MANGASARIAN L O. Proximal support vector machine classifiers[P]. Knowledge Discovery and Data Mining, 2001.

[2] HAN L, SUN J Z, ZHANG W, et al. A machine learning nowcasting method based on real－time reanalysis data[J]. Journal of Geophysical Research Atmospheres, 2016, 122(7): 4038－4051.

[3] ROBERT M, FRENCH, YANNICK GLADY, et al. An evaluation of scanpath－comparison and machine－learning classification algorithms used to study the dynamics of analogy making[J]. Behavior Research Methods, 2017, 49(4): 1291－1302.

[4] CROZIER T W M, TINTI M, LARANCE M, et al. Prediction of protein complexes in Trypanosoma brucei by protein correlation profiling mass spectrometry and machine learning[J]. Molecular & Cellular Proteomics, MCP, 2017, 16(12): 2254－2267.

[5] VEDANGI GODSE, KARISHMA KOTHARI, VAIBHAVI NANDODE, et al. Prediction of seismic activities in coal mines using machine learning[J]. International Journal of Engineering Technology Science and Research, 2017, 4(11): 117－121.

[6] ZHANG C, LU W. Seismic attributes selection based on SVM for hydrocarbon reservoir prediction[C]. Seg Technical Program Expanded Abstracts, 2010, 1586－1590. DOI: 10.1190/1.3513144.

[7] LI J, CASTAGNA J. Support Vector Machine (SVM) pattern recognition to AVO classification[J]. Geophysical Research Letters, 2004, 31(2).

[8] YUAN Y, LIU Y, ZHANG J, et al. Reservoir prediction using multi－wave seismic attributes[J]. Earthquake Science, 2011, 24, 373－389. DOI: 10.1007/s11589－011－0800－8.

[9] 段友祥，李根田，孙歧峰. 卷积神经网络在储层预测中的应用研究[J]. 通信学报，2016，1(37)：1－9.
DUAN Y X, LI G T, SUN Q F. Research on convolutional neural network for reservoir parameter prediction[J]. Journal on Communications, 2016, 1(37): 1－9. (In Chinese)

[10] ALI, J. K. Neural networks: A new tool for the petroleum industry[J]. European Petroleum Computer Conference, 1994, 1(1): 15－17.

[11] ALI M, ADWAIT CHAWATH?. Using artificial intelligence to predict permeability from petrographic data[J]. Computers & Geoences, 2000, 26(8): 915－925.

[12] YU S, ZHU K, DIAO F. A dynamic all parameters adaptive BP neural networks model and its application on oil reservoir prediction[J]. Applied Mathematics & Computation, 2008, 195(1): 66－75.

[13] 宋建国，杨璐，高强山，等. 强容噪性随机森林算法在地震储层预测中的应用[J]. 石油地球物理勘探，2018，5(53)：954－960.
SONG J G, YANG L, GAO Q S, et al. Strong tolerance random forest algorithm in seismic reservoir prediction[J]. Oil Geophysical Prospecting, 2018, 5(53): 954－960. (In Chinese)

[14] GEURTS P, ERNST D, WEHENKEL L. Extremely randomized trees[J]. Machine Learning, 2006, 63(1): 3－42.

[15] 胡祖光.基尼系数理论最佳值及其简易计算公式研究[J].经济研究,2004(09):60－69.
HU Z G. A Study of the best theoretical value of gini coefficient and its concise calculation formula[J]. Economic Research Journalg, 2004(09):60－69. (In Chinese)

[16] VAPNIK, V. and CHERVONENKIS, A., A note on class of perceptron. Automation and Remote Control, 1964,24.

[17] 李佳楠,高兴泉,李卓,等.四种机器学习算法预测大豆蛋白质定位对比研究[J].大豆科学,2022,41(03):337－344.
LI J N,GAO X Q,LI Z,et al. Comparative study of four machine learning algorithms for soybean protein localization predicting [J]. Soybean Science ,2022,41(03):337－344. (In Chinese)

[18] 李文秀,文晓涛,李天,等.基于近似支持向量机的流体识别因子融合[J].地球物理学进展,2020,35(01):139－144.
LI W X,WEN X T,LI T,et al. Fluid identification factors fusion based on proximal support vector machine[J]. Progress in Geophysics,2020,35(01):139－144. (In Chinese)

[19] 南京大学金陵学院大学数学教研室. 概率论与数理统计简明教程[M].南京:东南大学出版社, 2014.
Department of mathematics, Jinling college, Nanjing University Concise course of probability theory and mathematical statistics [M]. Nanjing :Southeast University Press, 2014. (In Chinese)

[20] 贾云娇. AVO 分析与基于近似支持向量机的流体识别[D].成都:成都理工大学,2019.
JIA Y J. AVO analysis and fluid identification based on proximal support vector machines[D]. Chengdu : Chengdu University of Technology, 2019. (In Chinese)

[21] 宁中华,何振华,黄德基.基于地震资料的高灵敏度流体识别因子[J].石油物探,2006, 3(45):239－241.
NING Z H, HE Z H, HUANG D J. High sensitive fluid identification based on seismic data[J]. Geophysical Prospecting for Petroleum, 2006, 3(45):239－241. (In Chinese)

[22] RUSSELL B H, HEDLIN K, HILTERMAN F J, et al. Fluid－property discrimination with AVO: A Biot－Gassmann perspective[J]. Geophysics,2003(68):29－39. DOI:10.1190/1.1543192.

Research on fluid recognition with extremely randomized trees algorithm

RAO Xiaochi[a], YANG Hao[a], YU Hui[b], WEN Wu[a], ZHOU Hang[a], CHEN Min[a]
(a. Chengdu University of Information Technology, Chengdu 610225, China;
School of Computer Science, Chengdu University of Information Technology
b. 78111 troops, Chengdu 610011, China)

Abstract: Great uncertainty occurs when we make fluid identification in the reservoir, so multi－attribute fusion for fluid identification is very necessary. Machine learning methods are becoming mature but rarely used in fluid identification. This paper introduces an Extremely randomized trees(ET) algorithm for fluid identification, which is simple to implement and has strong universality. The advantages of this method and traditional machine methods are compared, and the accuracy of the method for fluid identification is confirmed by mean square error and error rate. Finally, the method is applied to actual data in the South China Sea, whose effectiveness for fluid identification is verified by the excellent result.

Keywords: fluid identification; extremely randomized trees; machine learning; multi－attribute fusion