基于 LLE 和 SVM 的地震断层自动识别方法

邹冠贵[1,2],丁建宇[1],任　珂[1],殷裁云[3],董青山[1]

(1. 中国矿业大学 (北京) 地球科学与测绘工程学院, 北京　100083; 2. 中国矿业大学 (北京) 煤炭资源与安全开采国家重点实验室, 北京　100083; 3. 华能煤炭技术研究有限公司, 北京　101100)

摘　要: 传统地震资料的断层解释主要依靠解释者的知识和经验，存在工作量大、效率低的问题。基于机器学习的断层识别方法，可以融合已有的地质资料、解释人员的知识和经验，构建高质量的数据集，增加解释的准确率。为了提高机器学习方法断层解释的准确率，构建基于局部线性嵌入 (LLE) 和支持向量机 (SVM) 算法的断层识别方法。首先，介绍了 LLE 和 SVM 算法的基本原理，说明各算法的计算过程和主要参数；然后建立断层正演模型，分析不同属性的断层响应特征，针对训练数据集中多种地震属性之间的信息冗余，分别通过 LLE 和主成分分析 (PCA)2 种算法对地震属性数据进行降维，引入的量化指标计算结果表明 LLE 算法对于非线性数据体有较好的降维效果；利用西上庄井田 6 条巷道、5 口钻井揭露的 11 854 个已知构造信息的数据点，分别训练 SVM,PCA-SVM 和 LLE-SVM 断层识别模型；以准确率 A、查全率 R、查准率 P、F 作为模型的衡量标准，对比各模型在工区数据上的预测分类性能；其中，LLE-SVM 模型综合表现最佳，查准率可达 94.4%，远高于其他模型；最后，利用构建的各模型对整个工区进行预测，并结合实际揭露情况和人机交互解释结果进行分析。综合结果表明，基于 LLE 和 SVM 的断层识别方法在去除冗余信息的同时能够有效突出断层响应特征，减少主观人为因素的影响，提高断层解释的效率。

关键词: 断层识别; 地震属性优化; 煤田三维地震; 局部线性嵌入; 支持向量机

中图分类号: P631.4　　**文献标志码:** A　　**文章编号:** 0253−9993(2023)04−1634−11

Automatic identification method of seismic fault based on LLE and SVM

ZOU Guangui[1,2], DING Jianyu[1], REN Ke[1], YIN Caiyun[3], DONG Qingshan[1]

(1. *College of Geoscience and Surveying Engineering, China University of Mining and Technology-Beijing, Beijing* 100083, *China*; 2. *State Key Laboratory of Coal Resource and Safety Mining, China University of Mining and Technology-Beijing, Beijing* 100083, *China*; 3. *Huaneng Coal Technology Research Co., Ltd., Beijing* 101100, *China*)

Abstract: The fault interpretation of traditional seismic data mainly relies on the knowledge and experience of the interpreter, which has the problems of heavy workload and low efficiency. In order to construct high-quality data sets and increase the accuracy of interpretation, machine learning can integrate the existing geological data, the knowledge and experience of the interpreter. A fault recognition method based on Local Linear Embedding (LLE) and Support Vector Machine (SVM) algorithms is constructed to improve the accuracy of fault interpretation by machine learning methods. First, the basic principles of LLE and SVM algorithms are introduced to illustrate the calculation process and main parameters of algorithms. Then a fault forward modeling model is established to analyze the fault response characteristics of different at-

收稿日期: 2022−02−23　**修回日期:** 2022−05−30　**责任编辑:** 韩晋平　**DOI:** 10.13225/j.cnki.jccs.2022.0226

基金项目: 国家重点研发计划资助项目 (2018YFC0807803)

作者简介: 邹冠贵 (1981—), 男, 福建龙岩人, 副教授。E-mail: zgg@cumtb.edu.cn

通讯作者: 任　珂 (1993—), 男, 山东寿光人, 博士研究生。E-mail: renke666@foxmail.com

引用格式: 邹冠贵, 丁建宇, 任珂, 等. 基于 LLE 和 SVM 的地震断层自动识别方法[J]. 煤炭学报, 2023, 48(4): 1634−1644.

ZOU Guangui, DING Jianyu, REN Ke, et al. Automatic identification method of seismic fault based on LLE and SVM[J]. Journal of China Coal Society, 2023, 48(4): 1634−1644.

移动阅读

tributes. Aiming at the information redundancy among various seismic attributes in the training data set, the seismic attribute data are dimensionally reduced by LLE and principal component analysis (PCA). The intersection diagram shows that the LLE algorithm has a better dimensionality reduction effect for nonlinear data volumes. The SVM, PCA-SVM and LLE-SVM recognition models of fault were trained by using 11854 known structural information data points revealed by six roadways and five drilled wells in the Xishangzhuang Coalfield. Accuracy rate *A*, recall rate *R*, precision rate *P* and *F* value were used as the measurement standards to compare the prediction and classification performance of each model in the research area. Among them, the LLE-SVM model has the best overall performance, with a precision rate of 94.4%, much higher than those of other models. Finally, the whole research area is predicted by using the models, and analyzed by combining the actual disclosure and artificial interpretation results. The comprehensive results show that the fault identification method based on LLE and SVM can effectively highlight the fault response characteristics while removing redundant information, reduce the influence of subjective factors, and improve the efficiency of fault interpretation.

Key words: fault identification; seismic attributes optimization; 3D coalfield seismic; locally linear embedding; support vector machine

断层是煤矿开采中常见的一种地质构造, 主要是由于地壳运动引发岩层断裂造成的[1]。在进行煤层开采时不可避免地遇到各种地质构造, 若在生产时忽视了地质构造或者采取的安全措施不当, 则很容易引发煤矿地质灾害, 给煤矿带来重大的经济损失和人员伤亡[2]。因此, 查明断层分布是构造解释的重要组成部分。

传统断层解释是研究人员根据地震剖面上同相轴的不连续性来判别, 这种方法不仅工作量很大, 而且很难发挥地震多属性解释的优势。为了打破传统断层解释方法的局限性, 一系列的断层增强属性从三维地震数据体中被提取出来, 如相干体属性通过道间相似性的计算, 描述地层的横向不均匀性[3]; 曲率属性通过沿层曲率值的计算, 反映地层受构造应力挤压时层面弯曲的程度[4]; 混沌体属性通过局部构造张量特征值相对大小和不同特征值的组合运算, 衡量振幅值的规律性和混乱性, 从而突出特殊地质体的边界等[5]。这些地震属性是地震数据通过数学计算得到的运动学、动力学、几何学及统计学特征, 一定程度上可以强化和反映地层的不连续性[6], 但是本质上依然是单属性解释方法。

近年来, 伴随着人工智能领域的发展, 出现了很多基于机器学习算法的断层自动识别方法, 这些方法利用多种地震属性构建训练数据集, 通过模型参数优化实现断层识别, 可以有效减少解释的多解性, 是一种真正的多地震属性断层解释方法。如 BP 神经网络算法[7]、支持向量机[8]、卷积神经网络[9] 等。支持向量机 (SVM) 作为一种新型的模式分类方法, 其本质是寻找分类平面, 在面对小样本数据时, SVM 算法构建的模型相较其他算法具有更强的鲁棒性[10]。目前支持向量机被广泛用于解决煤层气和瓦斯涌出量预测[11-12]、煤层顶底板导水断裂带高度预测[13]、底板突水量预测及突水危险性评价[14] 等问题。

已有的研究表明: 在训练数据集中, 随着地震属性数量的增加, 一是可能带来数据冗余, 造成信息的重复和浪费[15], 比如方差体属性和相干体属性的相关性很高, 这两种地震属性都可以表征断层构造; 二是大量属性中包含着许多彼此相关的因素, 带来计算效率的降低[16]。已有的机器学习训练数据集构建方法表明: 优化技术是解决此类问题的有效途径, 可以降低多解性提高预测精度[17-18]。常见的优化方法主要有主成分分析 (PCA)、局部线性嵌入 (LLE) 等。PCA 为地震属性融合过程中一种常用的属性优化方法, 其核心思想是通过坐标旋转消除原数据空间的多重共线性, 从而达到线性降维的目的[19]。JAHAN 等[20] 使用 PCA 来对地震资料多种属性进行融合的方式来进行断层识别和提取。但是, 地震属性之间不仅存在线性关系, 还存在非线性关系。相较于 PCA, 局部线性嵌入 (LLE) 可以对高维空间上的数据点进行降维, 使其低维空间的局部邻域关系与原嵌套空间相同, 更适合于解决地震数据的非线性特征降维问题[21]。

山西省西上庄煤矿小断层发育, 笔者以该矿一二分区西翼为研究靶区, 在三维地震资料的基础上, 提取多种地震属性构建特征集, 分别通过 LLE 和主成分分析 (PCA)2 种算法对地震属性数据进行降维, 对比分析 SVM 算法的断层识别效果, 从而为煤田三维地震资料解释断层分布提供了一种新的思路。

1　基本原理

1.1　LLE 算法原理

LLE 算法是 ROWEIS 和 SAUL 在 2000 年提出的非线性降维方法[22]。假设每个数据点与它近邻点位于流形的一个线性或近似线性的局部领域, 此时每

个样本点就可以通过近邻点来线性表示，在重建低维流形时，使得重构误差最小，令其每个数据点的局部近邻关系与原空间保持一致[23−25]。算法实现共需要 3 个步骤：

(1) 对样本数据集 $X=[x_1,x_2,\cdots,x_N]\in R^{D\times N}$，寻找每个样本点 x_i 的 k 个近邻点，其中，R 为实数域；D 为数据维度；N 为样本数。

(2) 计算样本点的局部重建权值矩阵。通过定义一个代价误差函数 $\varepsilon(\boldsymbol{W})$：

$$\min\ \varepsilon(\boldsymbol{W})=\sum_{i=1}^{N}\left|x_i-\sum_{j=1}^{k}w_j^i x_{ij}\right|^2 \tag{1}$$

式中，$x_{ij}(j=1,2,\cdots,k)$ 为 x_i 的 k 个近邻点；w_j^i 为 x_i 与 x_{ij} 之间的权重，且要满足条件 $\sum_{j=1}^{k}w_j^i=1$。

(3) 将所有的样本映射到低维空间中。映射条件满足：

$$\min\ \varepsilon(Y)=\sum_{i=1}^{N}\left|\boldsymbol{y}_i-\sum_{j=1}^{k}w_j^i y_{ij}\right|^2 \tag{2}$$

式中，$\varepsilon(Y)$ 为损失函数；$\boldsymbol{y}_i$ 为 x_i 的输出向量；$y_{ij}(j=1,2,\cdots,k)$ 为 $\boldsymbol{y}_i$ 的 k 个近邻点，且满足：

$$\sum_{i=1}^{k}\boldsymbol{y}_i=0,\ \ \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{y}_i\boldsymbol{y}_i^{\mathrm{T}}=\boldsymbol{I} \tag{3}$$

其中，$\boldsymbol{I}$ 为 $m\times m$ 的单位矩阵。这里 $w_j^i(i=1,2,\cdots,N)$ 存储在 $N\times N$ 的稀疏矩阵 $\boldsymbol{W}$ 中，当 x_j 为 x_i 的近邻点时，$W_{i,j}=w_j^i$，否则 $W_{i,j}=0$。损失函数可重写为

$$\min\ \varepsilon(Y)=\sum_{i=1}^{N}\sum_{j=1}^{N}\boldsymbol{M}_{i,j}\boldsymbol{y}_i^{\mathrm{T}}\boldsymbol{y}_i \tag{4}$$

其中，$\boldsymbol{M}$ 为 $N\times N$ 的对称矩阵，其表达式为

$$\boldsymbol{M}=(\boldsymbol{I}-\boldsymbol{W})^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{W}) \tag{5}$$

要使损失函数值达到最小，则取 $\boldsymbol{Y}$ 为 $\boldsymbol{M}$ 的最小 m 个非零特征值所对应的特征向量。在处理的过程中，将 $\boldsymbol{M}$ 的特征值从小到大排序，第 1 个特征值几乎接近于 0，则舍去第 1 个特征值。通常取 2～$m+1$ 间的 m 个特征值对应的特征向量作为输出结果。

1.2 SVM 算法

SVM 算法是 CORTES 和 VAPINK 于 20 世纪 90 年代提出的[26]，是使用最为广泛的核学习算法。它的基本思想为将原低维输入空间中的非线性问题映射到高维特征空间中进行求解。SVM 的研究重点是寻求最优的超平面，最大限度地减小训练数据的分类错误[10,27-28]。SVM 具体算法为

$$\min\ R_\varepsilon(\boldsymbol{w},\xi)=\frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}+C\sum_{i=1}^{N}\xi_i \tag{6}$$

$$\boldsymbol{y}_i\left(\boldsymbol{w}^{\mathrm{T}}x_i-b\right)\geqslant 1-\xi_i,i=1,2,\cdots,N \tag{7}$$

式中，C 为惩罚系数，它会对分类器错分样本数产生影响；ξ 为松弛变量。

加入核函数后得到最终的 SVM 分类函数为

$$f(x)=\mathrm{sgn}\left(\sum\nolimits_{\mathrm{SV}}a_i^*\boldsymbol{y}_iK(\boldsymbol{x}_i,x)+b_i^*\right) \tag{8}$$

其中，SV 为支持向量；a_i^*、b_i^* 为拉格朗日乘子；$K(\boldsymbol{x}_i,x)$ 为核函数，核的值分别等于特征空间 $\varphi(\boldsymbol{x}_i)$ 和 $\varphi(\boldsymbol{x}_j)$ 中 2 个向量 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 的内积，即

$$K(\boldsymbol{x}_i,\boldsymbol{x}_j)=\varphi(\boldsymbol{x}_i)\varphi(\boldsymbol{x}_j) \tag{9}$$

核函数的种类有很多，其中最常用的是高斯核函数，因为其易于实现且具有非线性的映射能力，对处理非线性数据体有较好效果。高斯核函数的表达式为

$$K(\boldsymbol{x}_i,\boldsymbol{x}_j)=\exp\left(-g\,\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2\right) \tag{10}$$

其中，g 为核函数参数。如果 g 过大，高斯分布形态又高又瘦，会造成只会作用于支持向量样本附近，模型出现过拟合；反之，g 过小，模型容易出现欠拟合。

2 断层正演模拟分析

为探讨不同地震属性对断层的响应特点，以及测试 LLE 算法对地震属性降维的效果，笔者基于交错网格有限差分法构建断层正演模型，计算并提取地震属性，观察断层对不同地震属性的响应情况，并以 PCA 线性降维方法作为参照，对比 2 种降维方法在模型数据上的降维效果。为了尽可能使模型符合实际情况，正演模型构建时参考了研究区西上庄矿的实际地质构造情况。断层正演模型 (图 1) 参数如下：模型分为 3 层，上层为砂岩层，速度均为 3 000 m/s，密度为 2.7 g/cm^3；中间为煤层，埋深 300～350 m，速度为 2 000 m/s，密度 1.5 g/cm^3，层厚为 4 m；下层为泥岩层，速度为 2 800 m/s，密度为 2.2 g/cm^3。煤层内包含 6 个断层，其中 3 个为正断层，3 个为逆断层，自左至右断层落差分别为 5、14、4、17、3、20 m。模型地震道间距为 1 m，震源为雷克子波，频率 50 Hz。采用垂直激发，自激自收，并加入了标准差 10% 的白噪声。

模型正演得到地震剖面，利用地震解释软件追踪目标层位 (图 2)。根据研究经验，提取对断层响应特征明显的属性，包括方差、混沌体、能量、倾角、瞬时频率、瞬时相位、瞬时振幅、均方根振幅、最大振幅、最小振幅和弧长，一共提取 11 种属性。将各属性值分别进行归一化处理后投影到坐标系中，如图 3(a) 所

示。可以看出: 断层信息与各属性值分布均具有一定规律性, 这表明通过这些属性可以区别断层与非断层; 同时还可以观察到部分属性与断层之间存在相似的关系, 这反映了信息的冗余问题。如果将这些属性信息全部用于断层识别, 很容易造成模型训练过程中的过拟合, 因此需要对属性数据进行降维处理。

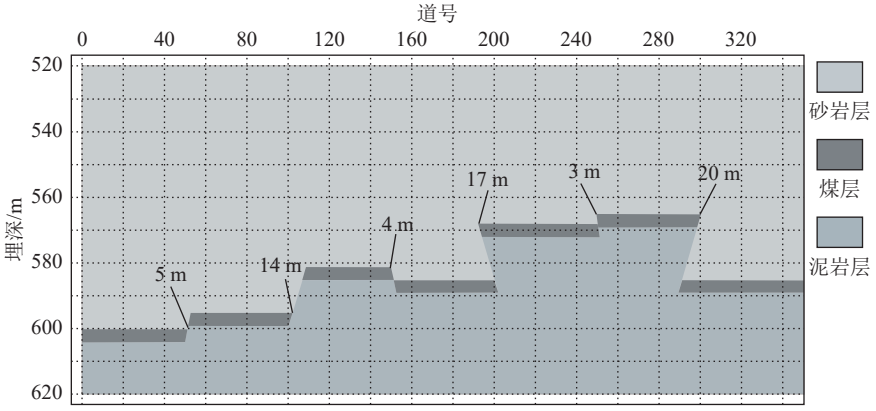


图 1　正演模型

Fig.1　Forward model

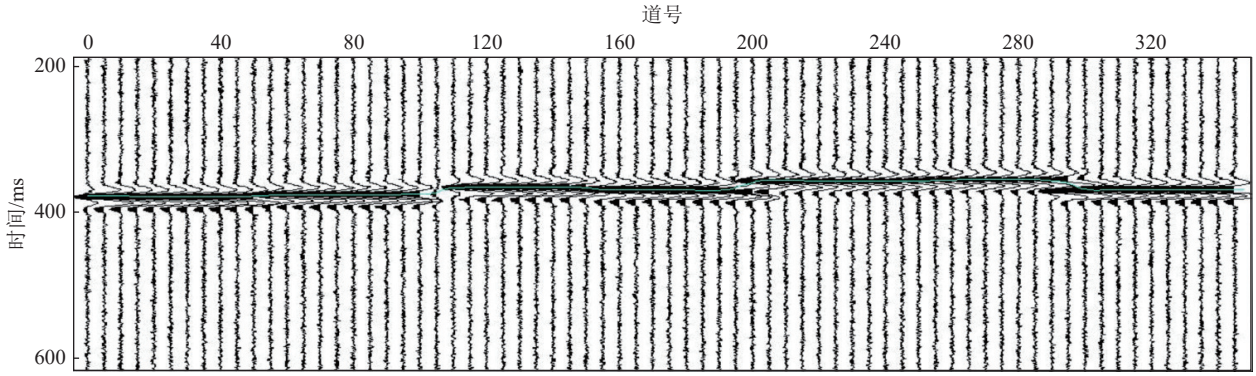


图 2　模型正演剖面

Fig.2　Cross-section of forward modeling

对提取得到的 11 种地震属性值分别进行 LLE 降维和 PCA 降维。以累计方差贡献率大于 95% 的标准, PCA 降维算法选取的主成分个数为 7; 以重构误差最小为标准, LLE 算法近邻点取值为 3, 嵌入维度取值为 2, 降维后的特征响应情况如图 3(b)、(c) 所示。为了分析 2 种算法的降维效果, 笔者引入量化指标来进行评价, 其主要思想为: 对于一个较好的降维方法而言, 任意 2 个点在高维空间中如果是近邻点, 那么降维后它们在低维空间中也应当是近邻点[29]。该指标的计算如式 (11) 所示, 它的值介于 0 和 1 之间, 指标值越小意味着降维结果中近邻信息保持得更好, 也就是降维结果更理想。

$$\text{Index}=\frac{1}{pq}\sum_{a=1}^{p}\sum_{b=1}^{q}\left[\boldsymbol{D}_{\mathrm{H}}(a,b)-\boldsymbol{D}_{\mathrm{L}}(a,b)\right] \tag{11}$$

式中, $\boldsymbol{D}_{\mathrm{H}}(a,b)$为归一化后的高维空间中的距离矩阵; $\boldsymbol{D}_{\mathrm{L}}(a,b)$为归一化后的低维空间中的距离矩阵; $a=1,2,\cdots,p$; $b=1,2,\cdots,q$。

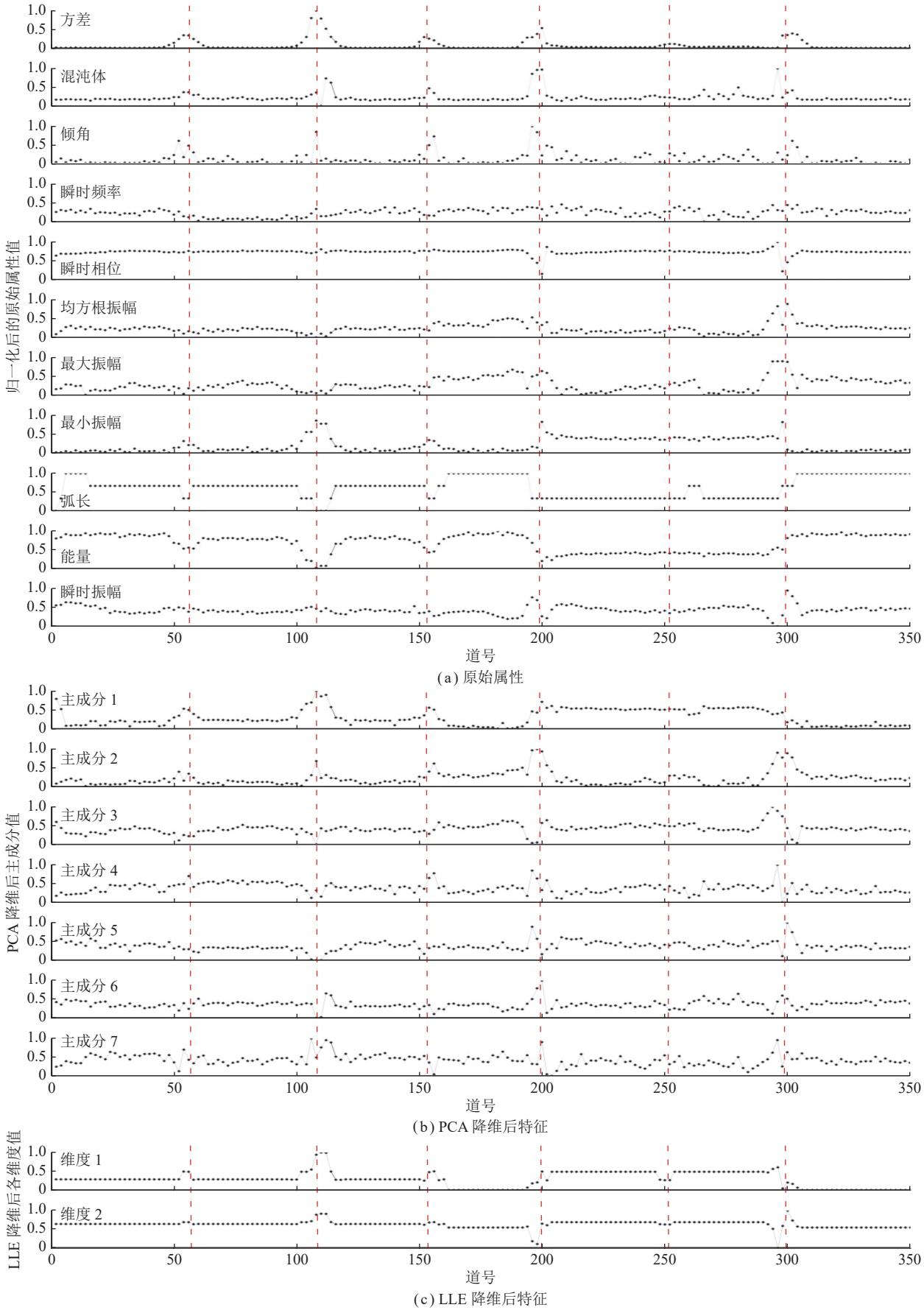
经过计算, PCA 降维的数据集 Index 值为 4.62×10^{-4}, LLE 降维的数据集 Index 值为 2.75×10^{-4}。该指标的计算结果证明, LLE 算法的降维效果更加理想。地震属性数据在经过 LLE 降维后, 既减少了数据中的冗余信息, 又很好的保留了原始数据的拓扑关系, 保证了断层和非断层点仍然可以通过新产生特征进行区分。

3　案　　例

在正演模拟结果的基础上, 为了进一步分析 LLE–SVM 断层识别方案的可行性和适用性, 尝试对实际三维地震数据进行应用。

3.1　研究区概况

本次的研究靶区是西上庄井田, 其位于山西省阳泉市及晋中市寿阳县境内。井田地处山西省黄土高原的中高山区, 井田内地势陡峻, 地形高差悬殊。一



(a) 原始属性

(b) PCA 降维后特征

(c) LLE 降维后特征

图 3 断层处的特征响应情况

Fig.3 Characteristic responses at faults

般相对高差 150～300 m, 地势西高东低, 南高北低。井田内大部为基岩裸露区, 局部为新生界地层所覆盖。井田内可采煤层有: 山西组 3、6、15 号煤层, 太原组 8、9、12、15、15 下号煤层。其中, 15 号煤层是本次解释目标层, 位于山西组中部, 煤层厚度 2.95～5.12 m, 平均 3.75 m, 煤层结构简单, 偶含 1～2 层夹矸。研究靶区的勘探面积为 4 km^2, 工作面内已有 6 条巷道、5 口钻井, 其在矿区内的分布及断层揭露情况如图 4 所示, 已揭露的 4 条断层 F1、F2、F3 和 F4 的断层信息见表 1。

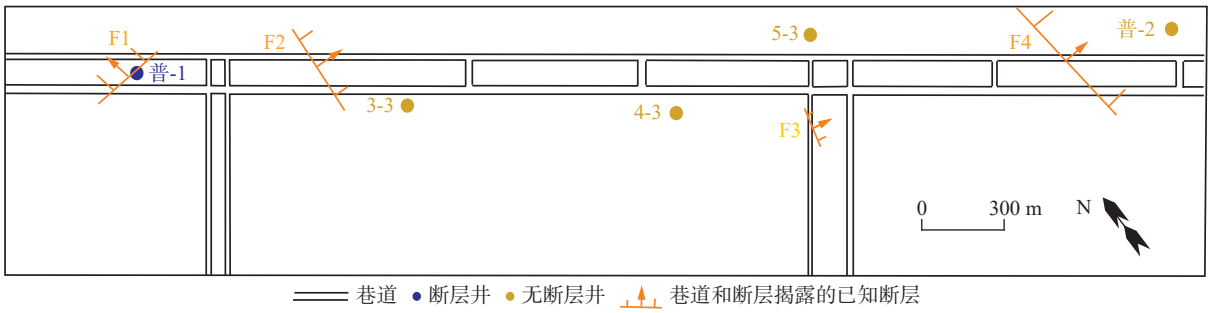


图 4　研究靶区已有巷道、钻井分布

Fig.4　Distribution of existing roadway and drilling in the research target area

表 1　已揭露断层的基本信息

Table 1　Basic information on exposed faults

断层	正逆	倾角/(°)	落差/m	延展长度/m
F1	正	70～80	0～6	398
F2	正	60～70	0～20	475
F3	正	70～80	0～5	210
F4	逆	60～70	0～17	712

3.2　地震属性提取与降维

利用工区内测井数据以及地震波的波阻抗关系在地震数据解释软件中标定并追踪目标煤层。提取和上述正演模型相同的 11 种地震属性, 全区共提取数据点 149 996 个, 图 5 为部分地震属性的可视化展示。根据巷道、测井揭露的断层和非断层点信息标记标签, 其中断层点的信息来自图 4 中的断层 F1、F2 和 F4, 断层 F3 留作验证。断层点标记为“1”, 非断层点标记为“0”, 共标记数据点 11 854 个, 包含断层数据点 4 578 个, 非断层数据点 7 276 个。

由于地震属性数据的量纲不同, 数据量差别很大, 所以在进行降维处理前, 通过式 (12) 将工区内的所有数据进行标准化, 标准化后的数据在 [0,1] 内且无量纲。将标准化后的数据复制为 3 组, 对 3 组数据分别进行不同的操作, 分别是保持原始数据不变、将数据进行 PCA 降维和进行 LLE 降维变换。

$$y_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{12}$$

式中, y_i 为归一化后样本值; x_i 为归一化前样本值; $x_{\max}$ 为样本最大值; $x_{\min}$ 为样本最小值。

在将数据集进行 PCA 降维时, 确定的主成分个数为 4, 该主成分取值下累计方差贡献率达 95%。LLE 算法主要有 2 个输入参数: 嵌入后的维数d和样本邻近点数K。降维的质量和这 2 个参数有很大关系。其中, 近邻点个数K的选取在 LLE 算法中起到关键作用, 如果K选取太大, 那么每个邻域会更趋近于整体, LLE 会丢失非线性特征, 不能体现局部特性; 如果K选取太小, LLE 则不能保持样本点在低维空间中的拓扑结构, 通常情况下K取值在 10 左右, 笔者对K在 [6,12] 上进行测试。本征维数d是指降维映射后的输出维数, 如果本征维数选取得太大, 输出数据则会受到噪声的影响; 如果本征维数选取得过小, 则不能正确地提取地震属性样本数据的固有特征。本文地震属性数据集降维的目标维度在 [4,10] 进行测试。

确定参数的范围后, 通过网格搜索法来确定最佳的K和d, 使得重构后的数据和原始数据误差最小。将邻近点数 K 和目标维度d的取值组成网格, 每一个网格就是 (K, d) 的一种取法, 计算每一对参数的重构误差, 选择重构误差最小的参数组合。数据集通过 LLE 降维后的重构误差整体上是随着邻近点数K和嵌入维度d增加而增加, 如图 6 所示。其中, 在“五角星”标记处, 存在误差最小值 $3.212\,062\,73\times10^{-16}$, 此时对应的最佳参数取值为$d = 5$, $K = 6$。

3.3　SVM 模型参数寻优

为了构建 SVM 模型, 分别将 3 组数据集中的巷道、断层已揭露的标记有标签的样本点选出, 用于训练和构建基于支持向量机算法的断层识别模型。支持向量机模型在构建过程中, 其分类性能除了和输入数据集有关外, 还取决于惩罚系数C和核函数参数 g。

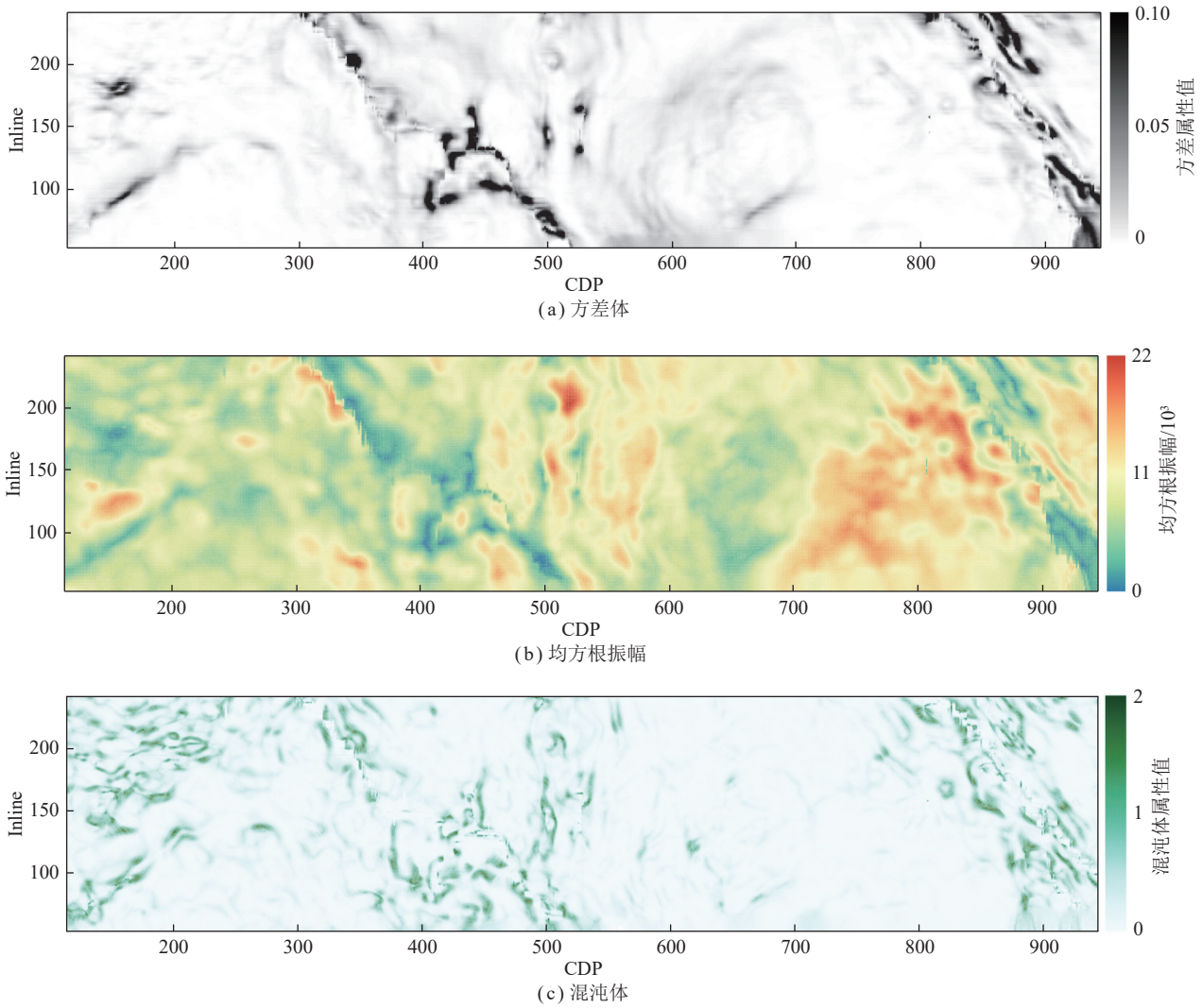


图 5 地震属性平面

Fig.5 Planar graph of seismic attributes

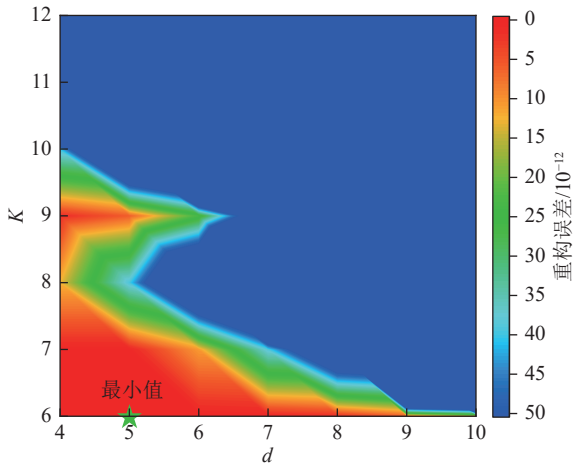


图 6 网格搜索结果

Fig.6 Results of grid search

针对支持向量机的参数选择问题，本研究采用基于粒子群优化算法的参数选择方法。粒子群优化算法 (Particle Swarm Optimization, PSO) 由 KENNEDY 和 EBERHART 在 1995 年提出，它通过个体间的协作来寻找最优解，拥有效率更高，更容易实现的优点。PSO 求解优化问题时，问题的解对应于寻找搜索空间中一只鸟的位置，这些鸟被称为“粒子”，每一个“粒子”都有自己的位置和速度 2 个属性，分别决定飞行的方向和距离，还有一个优化函数决定的适应值。每一个粒子记录并追随当前最优粒子在搜索空间中寻找最优解。PSO 初始化为一群随机粒子，通过迭代找到最优解。在每一次迭代中，粒子通过跟踪 2 个极值来更新自己，一个是粒子本身找到的最优解叫个体极值，另一个是整个种群找到的最优解叫做全局极值。在找到 2 个最优解时，粒子根据 2 个公式来更新自己的速度和位置。算法具体步骤如下：

设定 SVM 中的惩罚系数C和参数 g 作为粒子群中的粒子，将 SVM 分类正确率作为适应度函数，表达

式为

$$S = \frac{c_t}{c_t + c_f} \times 100\% \tag{13}$$

式中, c_t为支持向量机分类正确数; c_f为分类错误数。

利用建模的数据集分别计算 2 个粒子的适应度值, 并利用式 (14) 及式 (15) 对 2 个粒子的速度和位置进行更新:

$$v_{in}(t) = v_{in}(t-1) + c_1 r_{1j}[p_{in} - x_{in}(t-1)] + c_2 r_{2j}[p_{gn} - x_{in}(t-1)] \tag{14}$$

$$x_{in}(t) = x_{in}(t-1) + v_{in}(t) \tag{15}$$

式中, n为维数, $1 < n < N$; c_1和c_2为正常数; r_{1j}和r_{2j}为[0,1] 范围内的 2 个随机数; $v_{in}(t)$为t时刻、第i个粒子在第n维度上的速度; $x_{in}(t)$为t时刻、第i个粒子在第n维度上的位置; $p_{in}(t)$为t时刻、第i个粒子在第n维上的个体最优值; $p_{gn}(t)$为t时刻、所有粒子在第n维上的最优值。

将已知标签的各数据集按照 7∶3 的比例分为训练集和测试集。在利用训练集训练过程中, 通过 PSO 搜索 SVM 模型的最佳参数 C 和 g。以经过 LLE 降维的数据集为例, 参数训练过程中, 其适应度值随进化代数变化情况如图 7 所示。由图 7 可以看出, SVM 模型在进化到 30 代后, 最佳适应度值就不再变化, 此时参数 C 取值为 22.873 6, 参数 g 取值为 76.282 1。其他 2 组数据集开展同样的参数寻优过程, 最终各模型的最佳参数取值见表 2。

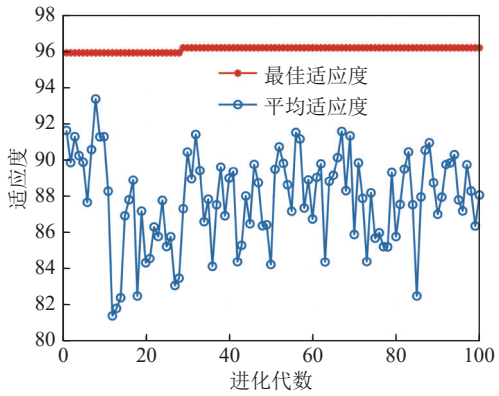


图 7　适应度值随进化代数变化 (LLE 降维数据)

Fig.7　Fitness value changes with evolutionary Algebra (LLE dimensionality reduction data)

表 2　PSO 参数寻优结果

Table 2　PSO parameter optimization results

模型	C	g
SVM	7.431 4	193.762 2
PCA-SVM	5.840 1	159.229 7
LLE-SVM	22.873 6	76.282 1

3.4　模型评价

在利用测试集进行模型评价时, 为了更好地评价各 SVM 模型的断层识别能力, 本研究选用了 4 个参数评价模型性能, 分别是准确率 A、查准率 P、查全率 R 和 F。各参数的计算方法见表 3 和式 (16)～(19)。其中, 准确率 A 是指预测正确的样本点在总样本点中占的比例, 是评价预测效果的常用指标。查准率 P 是指预测正确的断层样本在所有预测为断层样本中占的比例。查全率 R 是指预测正确的断层样本在确实为断层的样本中占的比例, 代表了模型在断层样本的预测能力。F 值通过查准率 P 和查全率 R 两项指标计算调和平均数得到。4 种指标代表了不同的意义, 在使用 4 种指标进行比较时, 好的模型并不一定在全部指标上优于其他模型, 优秀的模型是综合考虑实际应用场景和需求, 得到合适的结果。

表 3　评价指标交叉矩阵

Table 3　Cross matrix of evaluation index

样本标签(1/0)	推测断层(1)	推测非断层(0)
揭露的断层(1)	正确正例(TP)	错误的负例(FN)
揭露非断层(0)	错误的正例(FP)	正确的负例(TN)

各模型评价参数的计算结果如图 8 所示。其中准确率最高的是 LLE-SVM 模型为 0.836 895, 最低的是 PCA-SVM 模型为 0.790 051; 查准率最高的是 LLE-SVM 模型为 0.944 009, 最低的是 SVM 模型为 0.810 863; 查全率最高的是 SVM 模型为 0.626 217, 最低是 LLE-SVM 模型为 0.613 984; F 最高的是 LLE-SVM 模型为 0.744 042, 最低的是 SVM 模型为 0.706 678。综合来看 LLE-SVM 模型有最好的预测性能。

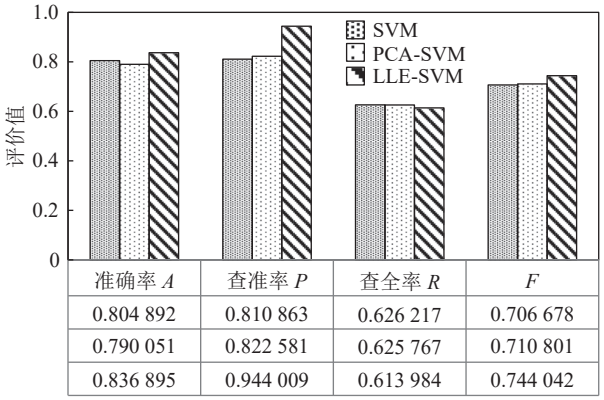


图 8　各模型评价结果

Fig.8　Evaluation results of different models

$$A = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{16}$$

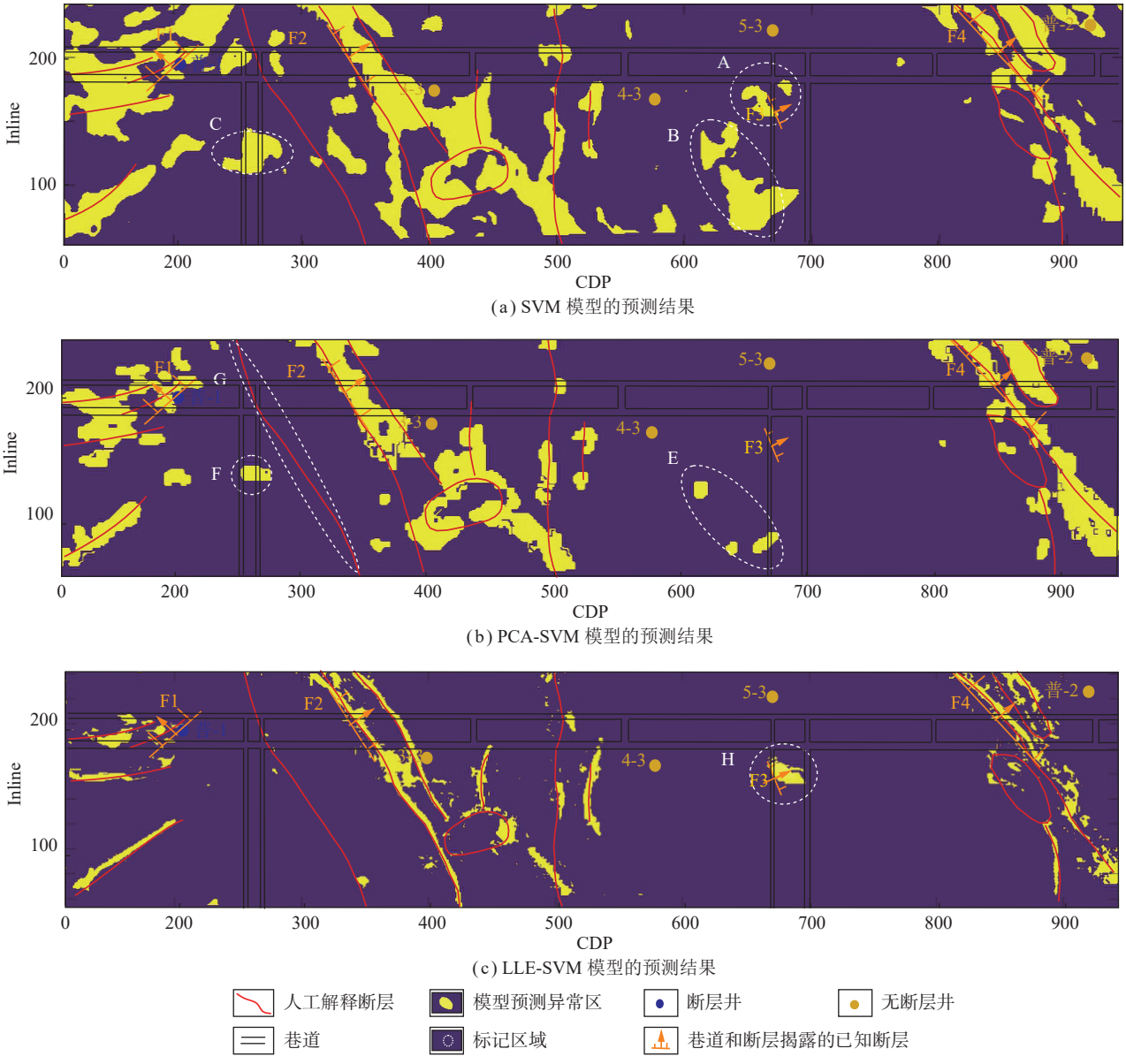
$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{17}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{18}$$

$$F = \frac{2}{1/P + 1/R} \tag{19}$$

3.5 研究区预测及分析

将训练得到的模型对整个研究区内的断层分布进行预测，对预测结果进行成图，如图 9 所示。其中图 9(a)～(c) 分别为原始数据预测结果、PCA 降维后预测结果、LLE 降维后预测结果。为了更好地对各模型预测结果进行分析，将巷道、钻井揭露的断层分布以及该研究区内人工解释的断层分布情况均标记到图 9 中。通过这些信息，对预测情况进行分析：① 整体看来，没有进行降维处理的原始数据预测结果，预测的异常区域偏大，断层连片分布严重；PCA-SVM 模型预测的异常区域比原始数据少，与人工解释的断层走向基本一致，但在断层 F3 处并没有异常响应；原始数据通过 LLE 降维后，在断层 F3 处存在异常响应。② 从图 9 中的断层展布来看，各模型预测的断层走向和断层延展长度大致相同，最大不同之处在于每个断层垂直于走向方向的分布形态，即断层的“胖瘦”情况。当断层“过胖”时，代表更多的非断层点被预测为断层点，此时查准率将会降低；反之，当模型预测的断层形态清晰且准确时，模型的查准率就会升高。LLE 模型较好的分布形态与该模型较高查准率的模型评价结



(a) SVM 模型的预测结果

(b) PCA-SVM 模型的预测结果

(c) LLE-SVM 模型的预测结果

图 9 各模型预测断层展布

Fig.9 Fault distribution predicted by each model

果相吻合。③ 图 9(a) 中的区域 A 和图 9(c) 中区域 H 处, 人工解释未解释出断层, 但从巷道实际揭露情况来看, A 区域实际存在落差 5 m 的小断层 F3。这说明相较于人工断层解释, 机器学习模型在小断层识别上具有一定优越性。④ 在区域 G 处, 尽管人工解释此区域存在断层, 但 PCA-SVM 和 LLE-SVM 两种模型在该区域响应都较弱。从巷道揭露的情况看, G 区域经过巷道的部分并没有断层, 因此判断该区域存在断层的可能性较小。

4　结　　论

(1) 利用 PCA 和 LLE 算法分别对正演模拟数据进行降维, 量化指标的评价结果表明: LLE 算法可以充分发挥非线性优势, 保留地震数据间的拓扑关系, 降维效果更加理想。

(2) 相较于 LLE-SVM 模型, SVM 模型和 PCA-SVM 具有较高的查全率 R, 并且在全区预测图上表现为大面积的异常区域。这一结果说明: 原始属性集本身存在信息冗余导致的模型过拟合问题; PCA 降维后的属性集虽然可以避免信息的重复, 但在线性降维过程中, 破坏了原有的数据结构, 导致模型分类精度相对较低。

(3) LLE-SVM 模型以牺牲小部分查全率 R 为代价得到了更高的查准率 P, 预测结果也与实际揭露更加匹配。这表明利用 LLE 算法对地震属性进行降维, 大大提高了数据的有效信息密度; 在保留原始数据集有效信息的同时, 可以有效地压制噪声。

(4) LLE-SVM 断层识别方法具有很强的可行性和适用性, 具有广泛的应用前景; 目前西上庄井田揭露的断层数量有限, 后续的回采验证可以有助于对模型进一步完善和分析。

参考文献(References):

[1] 彭苏萍, 程桦. 煤矿安全高效开采地质保障体系 [M]. 北京: 煤炭工业出版社, 2001: 103−109.

[2] 彭苏萍. 我国煤矿安全高效开采地质保障系统研究现状及展望[J]. 煤炭学报, 2020, 45(7): 2331−2345.
PENG Suping. Current status and prospects of research on geological assurance system for coal mine safe and high efficient mining[J]. Journal of China Coal Society, 2020, 45(7): 2331−2345.

[3] BAHORICH M S. Stratigraphic and structural interpretation with 3–D coherence[J]. Seg Technical Program Expanded Abstracts, 1996, 14(1): 1566.

[4] 杜文凤, 彭苏萍. 利用地震层曲率进行煤层小断层预测[J]. 岩石力学与工程学报, 2008, 27(S1): 2901−2901.
DU Wenfeng, PENG Suping. Seismic horizon curvature for predicting small fault in coalseam[J]. Chinese Journal of Rock Mechanics and Engineering, 2008, 27(S1): 2901−2901.

[5] RANDEN T, PEDERSEN S I, SNNELAND L. Automatic detection and extraction of faults from three-dimensional seismic data[J]. SEG Technical Program Expanded Abstracts, 2001, 20(1): 551.

[6] 石瑛, 王赟, 芦俊. 煤田地震多属性分析技术的应用[J]. 煤炭学报, 2008, 33(12): 1397−1402.
SHI Ying, WANG Yun, LU Jun. Application of seimic multi-attribute analysis technique in coal field[J]. Journal of China Coal Society, 2008, 33(12): 1397−1402.

[7] 董守华, 石亚丁, 汪洋. 地震多参数BP人工神经网络自动识别小断层[J]. 中国矿业大学学报, 1997, 26(3): 16−20.
DONG Shouhua, SHI Yading, WANG Yang. Automatic recognition of small fault by BP artificial nervous network from multiple seismic parameters[J]. Journal of China University of Mining & Technology, 1997, 26(3): 16−20.

[8] 孙振宇, 彭苏萍, 邹冠贵. 基于SVM算法的地震小断层自动识别[J]. 煤炭学报, 2017, 42(11): 2945−2952.
SUN Zhenyu, PENG Suping, ZOU Guangui. Automatic identification of small faults based on SVM and seismic data[J]. Journal of China Coal Society, 2017, 42(11): 2945−2952.

[9] HUANG L, DONG X, CLEE T E. A scalable deep learning platform for identifying geologic features from seismic attributes[J]. The Leading Edge, 2017, 36(3): 249−256.

[10] CHEN Y, HUANG Y, HUANG L. Suppressing migration image artifacts using a support vector machine method[J]. Geophysics, 2020, 85(5): 1−55.

[11] 邵良杉, 张宇. 基于小波理论的支持向量机瓦斯涌出量的预测[J]. 煤炭学报, 2011, 36(S1): 104−107.
SHAO Liangshan, ZHANG Yu. Mine gas gushing forecasting based on wavelet theory support vector machine[J]. Journal of China Coal Society, 2011, 36(S1): 104−107.

[12] 李艳芳, 程建远, 王成. 基于支持向量机的地震属性优选及煤层气预测[J]. 煤田地质与勘探, 2012, 40(6): 75−78.
LI Yanfang, CHENG Jianyuan, WANG Cheng. Seismic attribute optimization based on support vector machine and coalbed methane prediction[J]. Coal Geology & Exploration, 2012, 40(6): 75−78.

[13] 孙云普, 王云飞, 郑晓娟. 基于遗传–支持向量机法的煤层顶板导水断裂带高度的分析[J]. 煤炭学报, 2009, 34(12): 1610−1615.
SUN Yunpu, WANG Yunfei, ZHENG Xiaojuan. Analysis the height of water conducted zone of coal seam roof based on GA–SVR[J]. Journal of China Coal Society, 2009, 34(12): 1610−1615.

[14] 曹庆奎, 赵斐. 基于遗传–支持向量回归的煤层底板突水量预测研究[J]. 煤炭学报, 2011, 36(12): 2097−2101.
CAO Qingkui, ZHAO Fei. Forecast of water inrush quantity from coal floor based on genetic algorithm-support vector regression[J]. Journal of China Coal Society, 2011, 36(12): 2097−2101.

[15] WANG X J, HU G M, CAO J X. Application of multiple attributes fusion technology in the Su-14 Well Block[J]. Applied Geophysics, 2010, 7(3): 257−264.

[16] LIU L F, SUN Z D, YANG H J, et al. Seismic integrative prediction of fracture-cavity carbonate reservoir: Taking ZG21 well area in Tarim Basin as an example[J]. Journal of Central South

University(Science and Technology), 2011, 42(6): 1731−1737.

[17] ZOU G, REN K, SUN Z, et al. Fault interpretation using a support vector machine: A study based on 3D seismic mapping of the Zhaozhuang coal mine in the Qinshui Basin, China[J]. Journal of Applied Geophysics, 2019, 171: 103870.

[18] 印兴耀, 孔国英, 张广智. 基于核主成分分析的地震属性优化方法及应用[J]. 石油地球物理勘探, 2008, 43(2): 179−183.
YIN Xingyao, KONG Guoying, ZHANG Guangzhi. Seismic attributes optimization based on kernel principal component analysis(KPCA) and application[J]. Oil Geophysical Prospecting, 2008, 43(2): 179−183.

[19] JOLLIFFE I T. Principal component analysis [M]. SpringerVerlag, 2005.

[20] JAHAN I, CASTAGNA J, MURPHY M, et al. Fault detection using principal component analysis of seismic attributes in the Bakken Formation, Williston Basin, North Dakota, USA[J]. Interpretation-a Journal of Subsurface Characterization, 2017, 5(3): T361−T372.

[21] 李一鸣, 符世琛, 周俊莹, 等. 基于小波包熵和流形学习的垮落煤岩识别[J]. 煤炭学报, 2017, 42(S2): 585−593.
LI Yiming, FU Shichen, ZHOU Junying, et al. Collapsing coal-rock identification based on wavelet packet entropy and manifold learning[J]. Journal of China Coal Society, 2017, 42(S2): 585−593.

[22] ROWEIS S T, SAUL L K. Nonlinear dimensionality reduction by locally linear embedding[J]. Science, 2000, 290(5500): 2323−2326.

[23] LIU X F, ZHENG X D, XU G C, et al. Locally linear embedding-based seismic attribute extraction and applications[J]. Applied Geophysics, 2010, 7(4): 365−375.

[24] CHEN J, LIU Y. Locally linear embedding: A survey[J]. Artificial Intelligence Review, 2011, 36(1): 29−48.

[25] WANG J. Locally linear embedding [C]//Geometric Structure of High-Dimensional Data and Dimensionality Reduction. Springer, 2012: 203–220.

[26] CORTES C, VAPNIK V. Support-vector networks[M]. Machine Learning, 1995.

[27] NGUYEN-SY T, TO Q-D, VU M-N, et al. Predicting the electrical conductivity of brine-saturated rocks using machine learning methods[J]. Journal of Applied Geophysics, 2021, 184: 104238.

[28] YUE Y, WANG J. SVM method for predicting the thickness of sandstone[J]. Applied Geophysics, 2007, 4(4): 276−281.

[29] 田守财, 孙喜利, 路永钢. 基于最近邻的随机非线性降维[J]. 计算机应用, 2016, 36(2): 377−381.
TIAN Shoucai, SUN Xili, LU Yonggang. Stochastic nonlinear dimensionality reduction based on nearest neighbors[J]. Journal of Computer Applications, 2016, 36(2): 377−381.