文章编号 1004-924X(2014)07-1921-08

基于监督保局子空间虚假近邻准则的原始特征选择

辜小花[1*],李太福[1],杨利平[2],易 军[1],周 伟[1]
(1.重庆科技学院 电气与信息工程学院,重庆 401331;
2.重庆大学 光电技术及系统教育部重点实验室,重庆 400044)

摘要:提出一种基于监督保局投影(SLPP)与虚假最近邻(FNN)准则的原始特征选择方法。该方法首先将非线性原始数据映射到监督保局子空间,消除样本数据输入变量之间的相关性;然后,利用虚假近邻点方法计算剔除每个原始特征前后输入样本在监督保局子空间里的相似性测度,获得每个原始特征对类别变量不同程度的解释力;最后,从全特征开始逐步剔除解释能力弱的特征进而获得多组特征子集,并建立最近邻分类器,识别率最高且含特征数最少的特征子集即为最优特征子集。采用合成数据对该方法进行了仿真验证,结果表明,该方法可获得与数据集本质分类特征吻合的最佳特征子集。将该方法应用于选择真实的低阻油气层特征,获得的最佳特征子集比全特征集合的特征数量减少了 50%以上,分类识别率高出 8%。结果显示该方法具有优秀的原始特征选择能力,是一种有效的非线性特征选择方法。

关 键 词:监督保局投影;虚假近邻点;特征选择;模式分类;低阻油气层识别
中图分类号:TP391.41 **文献标识码**:A **doi**:10.3788/OPE.20142207.1921

Original feature selection based on false nearest neighbor criterion in supervised locality preserving subspace

GU Xiao-hua[1*], LI Tai-fu[1], YANG Li-ping[2], YI Jun[1], ZHOU Wei[1]

(1. *College of Electrical and Information Engineering, Chongqing University of Science and Technology, Chongqing* 401331, *China*;
2. *Laboratory of Optoelectronic Technology and Systems of the Ministry of Education, Chongqing University, Chongqing* 400044, *China*)
* *Corresponding author, E-mail: xhgu@cqu.edu.cn*

Abstract: A novel method based on Supervised Locality Preserving Projection (SLPP) and False Nearest Neighbor (FNN) was proposed for selecting the most proper feature for nonlinear pattern classification. In the proposed method, nonlinear original data were mapped to the supervised locality preserving subspace to eliminate the existing multi-collinearity among the features. Then, the interpretation capability for original features was estimated through calculating the variable mapping distance in the supervised locality preserving subspace. The nearest neighbor classifier based on each subset obtained by eliminating weak features successively was constructed. Finally, the optimal

收稿日期:2013-09-09;**修订日期**:2013-11-01.
基金项目:国家自然科学基金面上项目(No. 51374268);重庆市教委科学技术研究项目(No. KJ121402;KJ1401309);重庆市基础与前沿研究计划资助项目(No. cstc2013jcyA4004);重庆科技学院校内科研基金资助项目(No. CK2011B06, No. CK2013z11);重庆市科技人才培养计划资助项目(No. cstc2013kjrc-qnrc40008)

feature subset was selected corresponding to the highest recognition accuracy and the least number of features. The experiment on synthetic dataset shows that the proposed method can obtain an optimal feature subset containing the essential features in accordance with the classification goal. The method was used to select the features of low resistivity hydrocarbon reservoir, and the result indicates that the obtained optimal feature subset contains over 50% less feature and achieves 8% higher recognition accuracy as compared to that of the all-feature set. These results validate that the proposed method can offer excellent abilities of original feature selection and nonlinear feature selection.

Key words: Supervised Locality Preserving Projection (SLPP); False Nearest Neighbor (FNN); feature selection; pattern classification; low resistivity hydrocarbon reservoir recognition

1 引　言

随着信息化与工业化深度融合的推进，越来越多的企业组织实现了自动化、数字化，并在长期的生产过程中积累了丰富详实的生产数据。如何充分利用这些数据，从中挖掘有用信息以指导生产是当前提升企业竞争力的重点方向之一。实际应用中，为了全面掌握生产状况，往往需要监控大量的参数(每个参数就是一维特征)，而这些特征通常是稀疏的、冗余的，因此容易掩盖数据的真实结构。若将这些数据全部用于构建分类器，不仅分类识别的精度不能提高，还会带来复杂度增加、可靠性降低等问题[1]。

特征提取是消除原始数据信息冗余的有效方法[2-3]。理论与实验证明，复杂模式的特征之间往往存在着高阶相关性，因此原始数据会呈现明显的非线性，而线性模型过于简单，无法反映复杂模式的内在规律[4]。目前，针对非线性系统建模问题，主要采用基于核技巧的非线性特征提取，如：核主成分分析(Kernel Principle Component Analysis，KPCA)、核独立成分分析(Kernel Independent Component Analysis，KICA)[5]、核鉴别分析(Kernel Discriminant Analysis，KDA)[6]、核典型相关分析(Kernel Canonical Correlation Analysis，KCCA)[7]和核偏最小二乘(Kernel Partial Least Squares，KPLS)[8]等。核技巧能够成功地将非线性的数据结构尽可能地线性化，但其局限性是计算复杂度高。流形学习是近年来一类非常热门而有效的解决非线性特征提取问题的新方法[9]。其中：等距映射(Isometric mapping，Isomap)[10]、局部线性嵌套(Local Linear Embedding，LLE)[11]和拉普拉斯特征映射(Laplacian Eigenmap，LE)[12]等典型的流形学习方法已经广泛应用于高维非线性模式分类的特征提取，并得到了衍生和发展[13-16]。有证据表明，基于流形学习的特征提取方法与人本身的认知机理具有某种内在的关联性，因此有着潜在的重要的研究价值。

然而，特征提取获得的新特征是原始特征在某低维空间的组合，这些组合特征往往只具有数学意义，不具有具体的物理含义。当用这些组合特征进行模式分类的时候仍需要获得所有原始特征，也就是说，仅通过非线性特征提取不能减少原始特征的数量，信息采集系统的成本和模型复杂度仍很高。因此，追溯经过特征映射的原始特征对原始数据结构的影响成为了问题的关键。

作者受混沌相空间重构的虚假近邻点法[17-18]的启示，提出了一种基于保局子空间虚假近邻(False Nearest Neighbor，FNN)准则的非线性模式分类特征选择方法。该方法首先通过监督保局投影(Supervised Locality Preserving Projection，SLPP)[19-20]把原始特征转换到监督保局子空间，然后在监督保局子空间内利用FNN思想对映射后的矩阵进行特征约简，从而间接实现了原始特征空间的特征选择。

2 监督保局投影(SLPP)

保局投影算法(Locality Preserving Projection，SLPP)[19]是拉普拉斯特征映射的线性逼近。它能够有效地描述数据非线性结构，同时是一种线性运算，可以通过求解广义特征值问题得到保局投影空间，是一种高效的非线性特征提取方法。

给定分布为一个嵌入在高维空间 $\mathbf{R}^m$ 中的流

形 M 上的数据点集 $\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n]$，保局投影旨在寻找使得下述目标函数最小的低维嵌套：

$$\min\sum_{i,j=1}^{n}(\boldsymbol{y}_i-\boldsymbol{y}_j)\boldsymbol{S}_{ij}(\boldsymbol{y}_i-\boldsymbol{y}_j)^{\mathrm{T}},\tag{1}$$

其中：$\boldsymbol{y}_i=\boldsymbol{A}^{\mathrm{T}}\boldsymbol{x}_i$；$\boldsymbol{S}=[S_{ij}]_{i,j=1}^{n}$ 为 $n\times n$ 的稀疏、对称矩阵，用于描述样本集中样本点之间的局部几何关系，可采用 k 近邻方法确定：

$$S_{ij}=\begin{cases}\exp\left\{-\dfrac{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2}{2t^2}\right\}, & \boldsymbol{x}_i\in NNk(\boldsymbol{x}_j)\text{或 }\boldsymbol{x}_j\in NNk(\boldsymbol{x}_i)\\ 0,\text{其他}\end{cases},\tag{2}$$

其中：t 为经验参数；$NNk(\cdot)$表示 $\cdot$ 的 k 近邻。

LPP 用于特征提取虽能取得较为满意的结果，但是它是一种非监督学习算法，即算法并没有考虑到样本的类别信息，因此获得的特征分类并不是最优的，为此，提出了监督保局投影 SLPP，其基本思想是用有监督保局近邻图代替原有的无监督保局近邻图，即将式(2)更改为：

$$S_{ij}=\begin{cases}\exp\left\{-\dfrac{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2}{2t^2}\right\}, & \boldsymbol{x}_i\in NNk(\boldsymbol{x}_j)\text{或 }\boldsymbol{x}_j\in NNk(\boldsymbol{x}_i),\text{且 }x_i、x_j\text{ 同类}\\ 0, & \text{其他}\end{cases}\tag{3}$$

对式(1)进行代数运算，结果如下：

$$\begin{aligned}&\sum_{i,j=1}^{n}(\boldsymbol{y}_i-\boldsymbol{y}_j)\boldsymbol{S}_{ij}(\boldsymbol{y}_i-\boldsymbol{y}_j)^{\mathrm{T}}=\\&2\sum_{i=1}^{n}\boldsymbol{y}_i\Big(\sum_{j=1}^{n}\boldsymbol{S}_{ij}\Big)\boldsymbol{y}_i{}^{\mathrm{T}}-2\sum_{i,j=1}^{n}\boldsymbol{y}_i\boldsymbol{S}_{ij}\boldsymbol{y}_j{}^{\mathrm{T}}=\\&2\sum_{i=1}^{n}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{x}_i\boldsymbol{D}_{ii}\boldsymbol{x}_i{}^{\mathrm{T}}\boldsymbol{A}-2\sum_{i,j=1}^{n}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{x}_i\boldsymbol{S}_{ij}\boldsymbol{x}_j{}^{\mathrm{T}}\boldsymbol{A}=\\&2\boldsymbol{A}^{\mathrm{T}}\boldsymbol{X}(\boldsymbol{D}-\boldsymbol{S})\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}=\\&2\boldsymbol{A}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}.\end{aligned}\tag{4}$$

其中：$\boldsymbol{D}$ 为对角阵，其对角元素为权值矩阵 $\boldsymbol{S}$ 中对应行(或列，$\boldsymbol{S}$ 为对称阵)的和，即 $\boldsymbol{D}_{ii}=\sum_{j=1}^{n}\boldsymbol{S}_{ij}$，拉普拉斯矩阵 $\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{S}$。为了保证式(1)解的唯一性，还需要加入约束条件 $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}=1$。于是，问题(1)可以进一步转化为求解如下最小特征值问题：

$$\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}=\lambda\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}.\tag{5}$$

然而在数值计算中，最小特征值由于受数值计算精度的影响，往往估计不准确，通常将其转化为等价的最大化问题来求解。由于有 $\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{S}$，则式(5)转化为求解下列最大特征值问题：

$$\boldsymbol{X}\boldsymbol{S}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}=\lambda\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}.\tag{6}$$

这样，SLPP 的最优投影矩阵 $\boldsymbol{A}$ 可由式(6)的前 d 个最大的特征值对应的特征向量构成。不难看出，保局投影的目标是保持数据的局部关系，即使原始空间内互为近邻的点在低维空间中仍然互为近邻。

至此，完成了从原始数据矩阵 $\boldsymbol{X}$ 到低维线性嵌套 $\boldsymbol{Y}$ 的映射 $\boldsymbol{A}$ 的计算，即实现了对原始数据的非线性特征提取。

3　虚假近邻点法(FNN)

虚假近邻点法(FNN)[17]是 1992 年 Kennel 等人提出的，该方法用于确定相空间重构的嵌套维数。从几何观点来看，混沌时间序列是高维相空间混沌运动轨迹在低维空间上的投影，当选择的嵌入维数太小时，经过投影的混沌运动轨迹会变得扭曲，导致原始空间离得较远的相点在重构空间成为邻点，即虚假邻点。当嵌入维数逐渐增加后，低维空间的扭曲轨迹就会被完全分离出来，重构相空间内的“虚假近邻”现象逐渐消失，这就是虚假近邻法的基本原理。

具体地，在 m 维相空间中的相点 $x(i)=\{x(i),x(i+t),\cdots,x(i+(m-1)t)\}$存在一个最近邻点 $x^{\mathrm{NN}}(i)$，它们之间的距离为：

$$R_m(i)=\|x(i)-x^{\mathrm{NN}}(i)\|.\tag{7}$$

当相空间维数从 m 变成 $m+1$ 时，这 2 个相点间的距离 $R_{m+1}(i)$与 $R_m(i)$之间存在如下关系：

$$R_{m+1}^2(i)=R_m^2(i)+\|x(i+mt)-x^{NN}(i+mt)\|^2. \quad (8)$$

不难理解，如果混沌运动轨迹没有发生扭曲，$R_{m+1}(i)$与 $R_m(i)$应该相差不大，换言之如果 $R_{m+1}(i)$比 $R_m(i)$大很多，则可认为这是由于嵌入维数 m过低导致的“虚假近邻点”。通常，若

$$(R_{m+1}(i)-R_m(i))/R_m(i)>R_\tau, \quad (9)$$

则认为 $x^{NN}(i)$是 $x(i)$的虚假近邻点，其中：R_τ为阈值，取值范围通常为[10,50]。对于有噪声的现场数据，可认为当 $R_{m+1}(i)/R_A(i)\geqslant 2$时，$x^{NN}(i)$是 $x(i)$的虚假近邻点。其中：

$$R_A=\frac{1}{N}\sum_{i=1}^{N}[x(i)-\bar{x}]^2,$$

$$\bar{x}=\frac{1}{N}\sum_{i=1}^{N}x(i). \quad (10)$$

4 融合 SLPP 和 FNN 的特征选择

相空间重构的虚假近邻点方法中，2 个相点在 m维相空间内的距离与在 $m+1$维相空间内的距离的突变可以作为衡量这 2 个相点是否是虚假近邻的标准。由此得到启示，对于包含 p个特征的样本，去掉某一个特征后的新样本与原样本之间的距离越小可以认为所去掉的那一个特征相对整个特征的冗余程度越大；反之，距离越大则可认为该特征对样本结构有着重要贡献。

根据这一启发，不难得到基于虚假近邻准则的特征提取思路：对于包含 p个特征的原始观测样本 $\boldsymbol{x}=[x_1,x_2,\cdots,x_p]^T$，将其第 i个特征置零，得到新样本 $\hat{\boldsymbol{x}}=[x_1,\cdots,x_{i-1},0,x_{i+1}\cdots,x_p]^T$，则两者之间的距离 $d(i)=\|\boldsymbol{x}-\hat{\boldsymbol{x}}\|$。若 d很小，说明特征 x_i对数据内在结构的影响不大，可以忽略，反之，说明特征 x_i对数据内在结构的影响显著，不能剔除。

传统的虚假近邻点选择方法利用原始空间内样本之间的欧氏距离来度量样本的近邻关系，其准确性严重依赖于原始空间内基于欧氏距离的近邻关系对真实数据结构描述的准确性。而企业组织通过传感器或软传感器获得的原始特征通常呈现出稀疏、非线性、冗余等特性，且很难满足欧氏空间的性质。因此，其原始空间内的欧氏距离不能很好地描述数据的内在结构。

表 1 SLPP-FNN 特征选择算法的基本流程

Tab. 1 Procedure of SLPP-FNN feature selection

输入	原始数据矩阵 $\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n]$，$\boldsymbol{x}_i=[\boldsymbol{x}_i^1,\boldsymbol{x}_i^2,\cdots,\boldsymbol{x}_i^m]^T$；原始类别向量 $\boldsymbol{L}=[l_1,l_2,\cdots,l_n]$，$i=1,2,\cdots,n$。 对数据 $\boldsymbol{X}$进行规范化后依然记为 $\boldsymbol{X}$。
SLPP 部分	1. 计算权值矩阵：如式(3)所示，若 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$同类，且互为 k近邻，则 $\boldsymbol{S}_{ij}=\exp(-\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2/2t^2)$，否则 $\boldsymbol{S}_{ij}=0$； 2. 计算对角阵 $\boldsymbol{D}$：$\boldsymbol{D}_{ii}=\sum_{j=1}^{n}\boldsymbol{S}_{ij}$； 3. 计算 SLPP 投影矩阵 $\boldsymbol{A}$：$\boldsymbol{XSX}^T\boldsymbol{A}=\lambda\boldsymbol{XDX}^T\boldsymbol{A}$； 4. 计算样本 $\boldsymbol{x}_i$ 在 SLPP 子空间的投影 $\boldsymbol{y}_i$：$\boldsymbol{y}_i=\boldsymbol{A}^T\boldsymbol{x}_i$，$i=1,2,\cdots,n$。
FNN 部分	for $j=1:m$ for $i=1:n$ a. 计算去掉第 j个特征的新样本 $\hat{\boldsymbol{x}}_i^j$，$\hat{\boldsymbol{x}}_i^j=[x_i^1,\cdots,x_i^{j-1},0,x_i^{j+1},\cdots,x_i^m]^T$，$i=1,\cdots,n$； b. 计算 $\hat{\boldsymbol{x}}_i$ 在 SLPP 子空间的投影 $\hat{\boldsymbol{y}}_i$：$\hat{\boldsymbol{y}}_i^j=\boldsymbol{A}^T\hat{\boldsymbol{x}}_i$； c. 计算距离 $d_{ij}=\|\hat{\boldsymbol{y}}_i^j-\boldsymbol{y}_i\|_2$； endfor 计算距离 d_j：$d_j=\frac{1}{n}\sum_{i=1}^{N}d_{ij}$ endfor 对 d_j 由大到小排序，记 $D=[d_{ind1},d_{ind2},\cdots,d_{indm}]$ 获得原始特征剔除顺序索引 ind：ind=[ind1,ind2,…,ind$m$]
分类部分	1. 获得剔除解释能力较弱的特征后的特征子集 $\boldsymbol{S}_i=\boldsymbol{X}(\text{ind}(1,m-i))$，$i=1,2,\cdots,m-1$； 2. 针对每个特征子集构造分类器 $\boldsymbol{S}_i$，并获得分类识别率 r_i； 3. 对 r_i 按照从大到小排序，记 $R=[r_{indc1},r_{indc2},\cdots,r_{indcm}]$；
输出	最佳特征子集 $\boldsymbol{S}_{opt}$：识别率最高，特征数目最小的特征子集

为此，本文提出了基于监督保局子空间虚假近邻准则的特征选择方法。该方法首先将原始数据 $\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n]$（$\boldsymbol{x}_i\in\mathbf{R}^m$，$1\leqslant i\leqslant n$）的第 j（$j=1,2,\cdots,m$）个特征置为零得到 $\boldsymbol{X}_j$，将 $\boldsymbol{X}$ 和 $\boldsymbol{X}_j$ 分别投影到 d 维监督保局子空间，并计算它们在该特征子空间内的距离。根据距离的大小对原始特征排序，再依次剔除影响小的原始特征，从而形成特征子集的最近邻分类器，最终将分类识别率最高且特征维度最低的特征子集作为最优特征子集。算法步骤如表 1 所示。

4　仿真实验及应用分析

为了验证所提特征选择方法的可行性和有效性，分别在合成数据集和实际生产数据集上进行仿真实验和应用研究。

4.1　仿真实验

构造一个四类数据集，其每类数据均服从标准差为 1 的二维正态分布，即“瑞士卷”高斯混合模型。该数据集的数学表达如下：

$$\mathrm{Y}=\begin{cases}1,\text{当}\begin{bmatrix}\boldsymbol{x}_1\\ \boldsymbol{x}_2\end{bmatrix}\sim \mathrm{N}_2(\boldsymbol{\mu}_1,\Sigma)\\ 2,\text{当}\begin{bmatrix}\boldsymbol{x}_1\\ \boldsymbol{x}_2\end{bmatrix}\sim \mathrm{N}_2(\boldsymbol{\mu}_2,\Sigma)\\ 3,\text{当}\begin{bmatrix}\boldsymbol{x}_1\\ \boldsymbol{x}_2\end{bmatrix}\sim \mathrm{N}_2(\boldsymbol{\mu}_3,\Sigma)\\ 4,\text{当}\begin{bmatrix}\boldsymbol{x}_1\\ \boldsymbol{x}_2\end{bmatrix}\sim \mathrm{N}_2(\boldsymbol{\mu}_4,\Sigma)\end{cases},\qquad(12)$$

其中：$\boldsymbol{\mu}_1=\begin{bmatrix}7.5\\ 7.5\end{bmatrix}$，$\boldsymbol{\mu}_2=\begin{bmatrix}7.5\\ 12.5\end{bmatrix}$，$\boldsymbol{\mu}_3=\begin{bmatrix}12.5\\ 7.5\end{bmatrix}$，$\boldsymbol{\mu}_4=\begin{bmatrix}12.5\\ 12.5\end{bmatrix}$，$\Sigma=\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$。

根据该模型构造一个每类数据集均含 400 个样本点的数据集，如图 1 所示。为了验证所提特征方法的性能，特设置多重共线性特征和噪声，分别如下：

$$z_1=x_1,z_2=x_2,z_3=x_1+10\times x_3,$$

$$z_4=x_2+15\times x_4,z_5=x_3+x_4.$$

其中：$x_i\sim\mathrm{U}(0,1)$，$i=3,4$。5 个特征中 z_1，z_2 是本质特征；z_3，z_4 是冗余特征；z_5 是噪声。

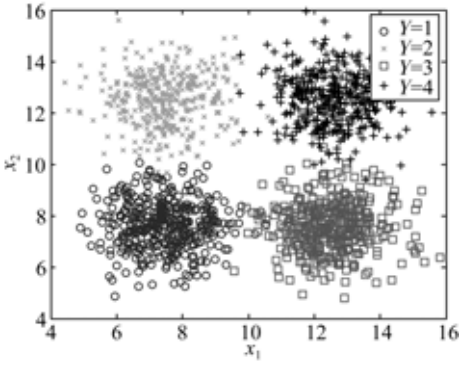

图 1　合成数据集

Fig. 1　Synthetic dataset

针对该数据集首先利用所提的基于监督保局子空间虚假近邻准则的特征选择方法，计算各原始特征的相似性测度值（欧氏距离），如表 2 所示。

表 2　合成数据各原始特征的相似性测度值

Tab. 2　Similarity measures of original features for synthetic dataset

原始特征	z_4	z_2	z_3	z_4	z_5
相似测度	0.363 8	0.709 9	0.189 9	0.072 8	0.0497

根据前面的分析可知，相似测度值越小，则去掉这个值对原始数据集的影响越小，即该原始特征对类别变量的解释能力越弱。因此，从表 2 可以看出，原始特征对类别变量的解释能力由强到弱依次是：z_2、z_1、z_3、z_4、z_5，与构造数据时的设置基本一致。

为了进一步验证所得特征子集的分类能力，依次剔除解释能力最弱的特征，共得到 5 个特征子集（含全特征），记为 $S_1=\{z_1,z_2,z_3,z_4,z_5\}$，$S_2=\{z_1,z_2,z_3,z_4\}$，$S_3=\{z_1,z_2,z_3\}$，$S_4=\{z_1,z_2\}$，$S_5=\{z_2\}$。针对每个特征子集建立最近邻分类器，分类结果如表 3 所示。

表 3　合成数据各特征子集的分类识别结果

Tab. 3　Classification results of feature subsets for synthetic dataset

特征子集	S_1	S_2	S_3	S_4	S_5
识别率/%	96.06	96.89	97.22	97.22	35.06

从表 3 可以看出，剔除掉本质特征（z_1）后，识别率急剧下降，而在冗余特征（z_3，z_4）和噪声（z_5）增加后，识别率没有提高，反倒有一定程度的下

降。综合考虑识别率(精度)和特征数目(规模)的平衡,可得到最佳特征子集。本实验中,以识别率尽量高、特征子集中包含的特征尽量少为准则,得到最佳原始特征子集 $S_4=\{z_1,z_2\}$。

4.2 应用研究

本节针对低阻油气层识别这一油田生产的难题展开实验研究。在油气开发中,低阻油气层中含油层的电阻率与水层接近,甚至低于水层的电阻率,在传统的测井解释中,常被误判为水层。为此,需要考虑融合多种测井参数的低阻油气层识别。另外,鉴于参数选择直接影响识别效果的好坏以及生产成本的高低,本文还对参数选取进行了研究。本节采用大庆油田某低阻油气地区的数据,以验证本文所提方法的效果。表 4 为部分原始数据。其中:x_1 表示井温度为 18℃时的泥浆电阻率(R_{M18});x_2 表示泥浆密度(ρ_M);x_3 表示深侧向电阻率(R_{LLD});x_4 表示浅侧向电阻率(R_{LLS});x_5 表示自然伽马(γ_{GR});x_6 表示声波时差(Δt);x_7 表示冲洗电阻率(R_{XO});x_8 表示侵入带电阻率(R_1);x_9 表示自然电位(U_{sp});Y 表示试油结果;1 表示油层;2 表示水层。从中选择前 26 组作为训练样本,后 25 组作为测试样本。

利用 SLPP-FNN 算法计算出各原始特征的相似测度值,如表 5 所示。

表 4　某低阻油气层测井、试油数据示例

Tab. 4　Illustrations of the logging and testing data of a low resistivity hydrocarbon reservoir

序号	采样点深度	储层厚度	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	Y
1	1 634	2.2	2.8	1.24	22	22	85	85	23.5	20.5	−30	1
2	1 677	3.6	2.8	1.24	12.5	14.5	95	77	18.5	12.5	−28.5	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
51	1 202	2.4	2.8	1.24	8	8	80	96	10	9	−24.5	2

表 5　低阻油气层数据各原始特征的相似性测度值

Tab. 5　Similarity measure of original feature for low resistivity hydrocarbon reservoir

原始特征	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
相似测度	0.248 0	0.220 3	0.726 3	1.290 4	0.163 4	0.559 8	0.553 6	0.359 2	0.382 8

从表 5 可以看出,原始特征对类别变量的解释能力由强到弱的顺序是:x_4,x_3,x_6,x_7,x_9,x_8,x_1,x_2,x_5。因此,依次剔除解释能力最弱的特征,可以得到如表 6 所示的特征子集及图 2 所示的分类结果。

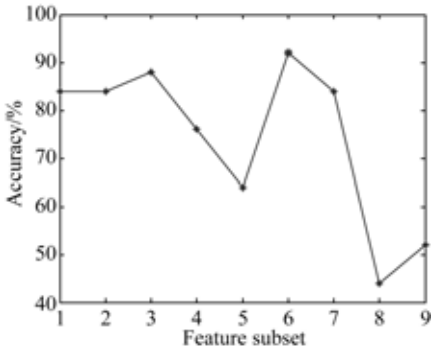

图 2　本文提出方法的各特征子集的分类识别结果

Fig. 2　Classification results of feature subsets for proposed method

从表 6、图 2 可以看出,子集 S_6 对应的识别率最高(见图 2 中圆圈标记处),特征数目也较低,为最佳特征子集。

表 6　本文提出方法的各特征子集及其分类识别结果

Tab. 6　Feature subsets and corresponding classification results of proposed method

特征子集	识别率/%
$S_1=\{x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8,x_9\}$	84
$S_2=\{x_1,x_2,x_3,x_4,x_6,x_7,x_8,x_9\}$	84
$S_3=\{x_1,x_3,x_4,x_6,x_7,x_8,x_9\}$	88
$S_4=\{x_3,x_4,x_6,x_7,x_8,x_9\}$	76
$S_5=\{x_3,x_4,x_6,x_7,x_9\}$	64
$S_6=\{x_3,x_4,x_6,x_7\}$	92
$S_7=\{x_3,x_4,x_6\}$	84
$S_8=\{x_3,x_4\}$	44
$S_9=\{x_4\}$	52

为进一步验证本方法的性能，本文将其与经典的特征选择方法——Relief 方法进行了比较。Relief 算法是一种特征权重算法，根据各个特征和类别的相关性，赋予它们不同权重，其特征和类别的相关性是基于特征对近距离样本的区分能力。表 7 和图 3 为 Relief 方法对于低阻油气层的特征子集及分类识别结果。

表 7　Relief 方法的特征子集及其分类识别结果

Tab. 7　Feature subsets and corresponding classification results of Relief algorithm

特征子集	识别率/%
$S_1=\{x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8,x_9\}$	84
$S_2=\{x_1,x_2,x_3,x_4,x_5,x_6,x_8,x_9\}$	82
$S_3=\{x_1,x_2,x_3,x_4,x_5,x_6,x_9\}$	84
$S_4=\{x_1,x_2,x_4,x_5,x_6,x_9\}$	64
$S_5=\{x_1,x_2,x_4,x_5,x_6\}$	88
$S_6=\{x_2,x_4,x_5,x_6\}$	72
$S_7=\{x_4,x_5,x_6\}$	76
$S_8=\{x_4,x_5\}$	64
$S_9=\{x_4\}$	52

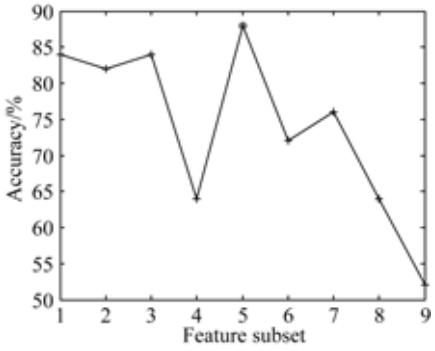

图 3　Relief 方法的各特征子集的分类识别结果

Fig. 3　Classification results of Relief algorithm

由表 7 和图 3 可知，Relief 方法获得的最佳子集是 $S_5=\{x_1,x_2,x_4,x_5,x_6\}$，对应的识别率为 88%，比本文方法的 92%低 4%，进一步验证了本文所提方法的有效性和优越性。

4.3　实验结果分析

综合 4.1 节仿真实验和 4.2 节的应用实验，有以下发现：

(1)本文所提方法获得的原始特征对类别变量的相似测度排序与数据模型中设定的原始特征对类别变量解释能力的强弱的排序吻合得较好，这说明所提方法能够捕获原始特征对类别变量的解释能力的大小；

(2)模型本质特征对类别变量的解释能力最强，多重共线性特征对类别变量的解释能力较弱，噪声对类别变量的解释能力最弱。因此从原始特征中挖掘出本质特征并剔除多重共线性特征和噪声有助于提高分类算法精度并降低其复杂度；

(3)结合原始特征相似性测度和最近邻分类识别率能够搜索出最优特征子集。以低阻油气层特征选择为例，所提方法获得的最佳特征子集的特征数量相较于全特征模型减少了 50%以上，分类识别率高出 8%，从而验证了本文所提特征提取方法的正确性；

(4)本文方法获得的最佳特征子集的识别率比 Relief 方法高出 4%，从而验证了本文方法的优越性。

5　结　论

本文针对高维数据分类中特征选择问题展开了研究，结合监督保局投影法和虚假近邻点法提出了一种基于监督保局子空间虚假近邻准则的特征选择方法。该方法通过计算剔除某特征前后监督保局子空间内的距离，获得该特征对类别变量的解释能力；接着依次剔除解释能力弱的特征，并建立相应特征子集对应的最近邻分类器；然后以分类识别率最高，包含特征数目最小为准则搜索最优特征子集。监督保局投影有效地消除了特征之间的多重共线性，大大提高了虚假近邻点法的计算准确性，在构建权值矩阵的同时，充分考虑了特征与类别变量的内在联系，从而获得了相似性测度与特征对类别变量解释能力的对应关系。在合成数据和实际应用数据上的实验验证了所提方法的有效性。

参考文献：

[1]　JUNG C S, SEO H, KANG H G. Estimating redundancy information of selected features in multi-dimensional pattern classification[J]. *Pattern Recognition Letters*, 2011, 32(4): 590-596.

[2]　王灿进，孙涛，王挺峰，等. 基于轮廓特征的神经网络目标识别研究[J]. 液晶与显示，2013，28(4)：641-648.

WANG C J, SUN T, WANG T F, *et al.*. Target recognition using Neural Network based on Contour features [J]. *Chinese Journal of Liquid Crystals and Displays*, 2013, 28(4): 641-648. (in Chinese)

[3] 杨云，岳柱. 基于融合图像轮廓和 Harris 角点方法的遮挡人体目标识别研究[J]. 液晶与显示，2013，28(2)：273-277.
YANG Y, YUE ZH. Human body target recognition under occlusion based on fusion of image Contour moment and Harris angular points [J]. *Chinese Journal of Liquid Crystals and Displays*, 2013, 28(2): 273-277.

[4] CHO H W. Nonlinear feature extraction and classification of multivariate data in Kernel feature space [J]. *Expert Systems with Applications*, 2007, 32(2): 534-542.

[5] ZHANG Y W. Enhanced statistical analysis of nonlinear processes using KPCA, KICA and SVM [J]. *Chemical Engineering Science*, 2009, 64(5): 801-811.

[6] WIDODO A, KIM E Y, SON J D, *et al.*. Fault diagnosis of low speed bearing based on relevance vector machine and support vector machine [J]. *Expert Systems with Applications*, 2009, 36(3): 7252-7261.

[7] SAKAR C O, KURSUN O, GURGEN F. A feature selection method based on kernel canonical correlation analysis and the minimum Redundancy-Maximum Relevance filter method [J]. *Expert Systems with Applications*, 2012, 39 (3): 3432-3437.

[8] ZHANG Y W, TENG Y D. Process data modeling using modified kernel partial least squares [J]. *Chemical Engineering Science*, 2010, 65 (24): 6353-6361.

[9] 杨静宇，金钟，杨健. 模式特征抽取研究进展[C]. 2009 中国自动化大会暨两化融合高峰会议，2009，杭州.
YANG J Y, JIN ZH, YANG J. The research progress of pattern feature extraction [C]. CAAC2009, 2009, *Hang Zhou*. (in Chinese)

[10] TENENBAUM J B, VIN de SILVA, LANGFORD J C. A global geometric framework for nonlinear dimensionality reduction [J]. *Science*, 2000, 290: 2319-2323.

[11] ROWEIS S T, SAUL L K. Nonlinear dimensionality reduction by locally linear embedding [J]. *Science*, 2000, 290: 2323-2326.

[12] BELKIN M, NIYOGI P. Laplacian eigenmaps and spectral techniques for embedding and clustering [C]. *Proceedings of Conference on Advances in Neural Information Processing system* 15, 2001.

[13] PARKH. ISOMAP induced manifold embedding and its application to Alzheimer's disease and mild cognitive impairment [J]. *Neuroscience Letters*, 2012, 513(2): 141-145.

[14] QIAO H, ZHANG P, WANG D. An explicit nonlinear mapping for manifold learning[J]. *IEEE Transactions on Cybernetics*, 2013, 43(1): 51-63.

[15] ZHANG J P J, HUANG H, WANG J. Manifold learning for visualizing and analyzing high-dimensional data [J]. *IEEE Intelligent Systems*, 2010, 25(4): 54-61.

[16] 黄鸿，杨媚，张满菊. 基于稀疏鉴别嵌入的高光谱遥感影像分类[J]. 光学 精密工程，2013，21(11)：2922-2930.
HUANG H, YANG M, ZHANG M J. Semi-supervised manifold learning and its application to remote sensing image classification [J]. *Opt. Precision Eng.*, 2013, 21 (11): 2922-2930. (in Chinese)

[17] KENNEL M B, BROWN R, ABARBANEL H D I. Determining embedding dimension for phase space reconstruction using a geometrical reconstruction [J]. *Phys Rev* A, 1992, 45: 3403-3411.

[18] 李太福，易军，苏莹莹，等. 基于 KPCA 子空间虚假邻点判别的非线性建模的变量选择[J]. 机械工程学报，2012，48(10)：192-198.
LI T F, YI J, SU Y Y, *et al.*. Variable selection for nonlinear modeling based on false nearest neighbours in KPCA subspace [J]. *Journal of Mechanical Engineering*, 2012, 48 (10): 192-198. (in Chinese)

[19] HE X F, YAN S C, HU Y X, *et al.*. Face recognition using Laplacianfaces [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005, 27(3): 328-340.

[20] 辜小花，龚卫国，杨利平. 有监督图优化保局投影[J]. 光学 精密工程，2011，19(3)：672-680.
GU X H, GONG W G, YANG L P. Supervised graph-optimized locality preserving projections [J]. *Opt. Precision Eng.*, 2011, 19(3): 672-680. (in Chinese)

作者简介：

辜小花(1982－)，女，四川眉山人，博士，副教授，2011 年于重庆大学获得博士学位，主要从事模式识别、图像处理、复杂系统建模与优化方面的研究。E-mail: xhgu@cqu.edu.cn

李太福(1971－)，男，四川资阳人，博士，教授，硕士生导师，1996 年、2000 年和 2004 年于重庆大学分别获得学士、硕士和博士学位，主要从事复杂系统建模与优化方面的研究。E-mail: litaifuemail@qq.com