基于稀疏和近邻保持的极限学习机降维

陈晓云 [1] 廖梦真 [1]

摘 要 近邻与稀疏保持投影已被广泛应用于降维方法, 通过优化得到满足近邻结构或稀疏结构的降维投影矩阵, 然而这类方法多数只考虑单一结构特征. 此外, 多数非线性降维方法无法求出显式的映射函数, 极大地限制了降维方法的应用. 为克服这些问题, 本文借鉴极限学习机的思想, 提出面向聚类的基于稀疏和近邻保持的极限学习机降维算法 (SNP-ELM). SNP-ELM 算法是一种非线性无监督降维方法, 在降维过程中同时考虑数据的稀疏结构与近邻结构. 在人造数据、Wine 数据和 6 个基因表达数据上进行实验, 实验结果表明该算法优于其他降维方法.

关键词 极限学习机, 近邻表示, 稀疏表示, 降维

引用格式 陈晓云, 廖梦真. 基于稀疏和近邻保持的极限学习机降维. 自动化学报, 2019, **45**(2): 325−333

DOI 10.16383/j.aas.2018.c170216

Dimensionality Reduction With Extreme Learning Machine Based on Sparsity and Neighborhood Preserving

CHEN Xiao-Yun[1] LIAO Meng-Zhen[1]

Abstract Neighborhood and sparsity structure preserving projections have been widely used in dimensionality reduction, but most of them consider single structures. Moreover, existing nonlinear DR methods can not get an accurate projection function, which limits their applications. To overcome these problems, we propose a nonlinear dimensionality reduction method SNP-ELM by extending the extreme learning machine model. SNP-ELM is a nonlinear unsupervised dimensionlity reduction method, which takes both sparsity structure and neighborhood structure into account. The experimental results on toy data, wine data and six gene expression data show that our method significantly outperforms the compared dimensionality reduction methods.

Key words Extreme learning machine, neighbor representation, spare representation, dimensionality reduction

Citation Chen Xiao-Yun, Liao Meng-Zhen. Dimensionality reduction with extreme learning machine based on sparsity and neighborhood preserving. *Acta Automatica Sinica*, 2019, **45**(2): 325−333

随着大数据时代的到来, 人们对数据的处理正面临巨大挑战. 在大数据应用研究中, 高维数据分析与研究是其主要内容之一. 在现代机器学习与统计学的研究背景下, 高维数据所引发的维数灾难主要表现为: 众多低维空间中表现良好的算法在面对高维数据时性能急剧下降. 其主要原因有: 1) 维数增加导致数据空间体积急剧膨胀、同等数量样本分布非常稀疏, 难以形成有效的簇; 2) 高维空间中存在测度 "集中现象", 使样本点间距离度量的类区分性随着维数增加而减弱; 3) 样本数据包含大量冗余信息对聚类或分类无用, 甚至会降低算法的性能. 基于上述原因, 对降维方法进行研究是十分有必要的.

收稿日期 2017-04-24 录用日期 2017-10-03
Manuscript received April 24, 2017; accepted October 3, 2017
国家自然科学基金 (71273053, 11571074) 资助
Supported by National Science Foundation of China (71273053, 11571074)
本文责任编委 曾志刚
Recommended by Associate Editor ZENG Zhi-Gang
1. 福州大学数学与计算机科学学院 福州 350116
1. College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350116

总体上说, 面向聚类的降维方法均为无监督降维方法, 可分为线性降维和非线性降维. 当前, 多数无监督线性降维方法假设观测数据落在一个低维流形子空间中, 通过寻找高维空间到低维子空间的线性投影实降维, 如主成分分析 (Principal component analysis, PCA)[1]、局部保持投影 (Locality preserving projections, LPP)[2]、近邻保持嵌入 (Neighborhood preserving embedding, NPE)[3] 和稀疏保持投影 (Sparsity preserving projections, SPP)[4]. PCA 是最经典的线性降维方法, 以最大化投影散度为目标, 但未考虑样本间的近邻结构关系, 不适合分布于流形上的非线性数据; LPP 和 NPE 则考虑了样本间的近邻结构, LPP 以保持降维前后样本间的近邻关系不变为目标, 而 NPE 旨在保持降维前后样本间的局部近邻结构; SPP 的优化目标是使降维前后样本间的稀疏表示结构得以保持. 但当数据非线性分布时, 上述线性降维算法就会失效. 为弥补线性降维算法的不足, 各种非线性扩展方法被提出, 如核主成分分析 (Kernel principal

component analysis, KPCA)[5] 和局部线性嵌入 (Locally linear embedding, LLE)[6]. KPCA 是 PCA 基于核技巧的非线性推广, 用于对非线性分布数据降维; LLE 以保持投影前后局部线性关系不变为目的构造目标函数. 然而这些非线性降维方法无法求出显式的映射函数, 当有新样本加入时, 需要重新学习优化模型.

极限学习机 (Extreme learning machine, ELM)[7−8] 最早被用于训练单隐层前馈神经网络, 具有学习速度快、泛化能力强等特点, 为有监督学习如分类和回归提供了简单有效的方法[9−10]. 2014 年, Huang 等基于流形正则的思想将 ELM 推广到无监督学习任务, 提出了一种新的非线性降维方法无监督极限学习机 (Unsupervised extreme learning machine, US-ELM)[11]. 该方法很好地利用了 ELM 的逼近能力, 通过非线性映射将原数据投影到低维空间中, 并能够得到显式的非线性映射函数. 但该方法利用高斯函数描述近邻样本间的相似度, 由于高斯函数用到距离测度, 难以避免地也存在高维空间中测度 "集中现象", 即样本点间高斯相似性度量的类区分性随着维数增加而减弱, 进而影响降维算法性能. 此外, US-ELM 直接利用给定高斯函数计算样本近邻表示系数, 不具有数据自适应性.

针对上述问题, 本文对 US-ELM 进行改进, 同时考虑非线性数据的局部线性表示和全局稀疏表示. 其中, 局部线性表示用于解决非线性流形数据的刻画问题, 以获取数据的局部结构[12]; 全局稀疏表示用于描述数据的全局结构[13]; 并通过加权参数融合近邻线性表示信息和稀疏表示信息. 由此, 我们提出基于稀疏和近邻保持的极限学习机降维方法 (SNP-ELM), 使得降维前后样本间的局部近邻表示关系和全局稀疏性保持不变. SNP-ELM 通过学习得到近邻表示系数, 较之 US-ELM 具有更好的数据自适应性.

1　极限学习机

极限学习机本质上是一种单隐含层前馈神经网络, 其结构如图 1 所示[14]. ELM 网络的训练主要分为两个阶段. 第一个阶段是 ELM 网络结构构建, 隐含层将输入数据映射到 n 维的特征空间中, n_h 为隐节点个数. 定义隐含层关于 x_i 的输出向量为 $h(x) = [h_1(x), h_2(x), \cdots, h_{n_h}(x)] \in \mathbf{R}^{1\times n_h}$。其中, $x \in \mathbf{R}^m$, $h_i(x)$ 是第 i 个隐节点的输出, 其输出函数可以表示为:

$$h_i(x) = g(a_i, b_i, x), a_i \in \mathbf{R}^m, b_i \in \mathbf{R} \tag{1}$$

其中, $g(a_i, b_i, x)$ 为非线性激励函数, 常用的函数有 Sigmoid 函数和 Gaussian 函数. 本文采用 Sigmoid 函数, 其表达式为:

$$g(a_i, b_i, x) = \frac{1}{1+\exp(-a_i x + b_i)} \tag{2}$$

式中, a_i 为第 i 个隐节点的输入权值, b_i 为第 i 个隐节点的偏差, 在 ELM 网络中输入权向量 a_i 和隐节点偏差 b_i 是随机产生的.

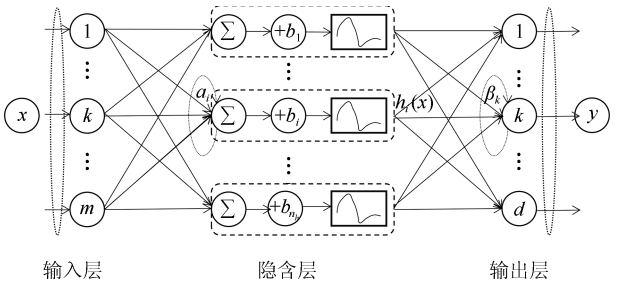

图 1　ELM 网络结构示意图

Fig. 1　ELM network structure

对于数据集 X, ELM 隐藏层输出为:

$$H(X) = \begin{bmatrix} h(x_1) \\ \vdots \\ h(x_n) \end{bmatrix} = \begin{bmatrix} g(a_1,b_1,x_1) & \cdots & g(a_{n_h},b_{n_h},x_1) \\ \vdots & \ddots & \vdots \\ g(a_1,b_1,x_n) & \cdots & g(a_{n_h},b_{n_h},x_n) \end{bmatrix} \tag{3}$$

若隐藏层到输出层的权重矩阵为 $\beta = [\beta_1, \beta_2, \cdots, \beta_m]$, 则 ELM 网络的输出为

$$Y = H(X)\beta \tag{4}$$

第二阶段是基于 ELM 网络结构求解输出权重矩阵 β, 通常根据 ELM 网络学习任务的不同构建不同的模型来求解输出权重矩阵 β. 经典的 ELM 模型用于解决有监督学习问题, 如: 分类和回归. 对于含 n 个样本的训练集 $S = \{(x_i, y_i)|x_i \in X \subseteq \mathbf{R}^m, y_i \in Y \subseteq \mathbf{R}^d, i = 1, \cdots, n\}$, x_i 为输入变量, y_i 为输出变量, 则其模型表示:

$$\begin{aligned} &\min \left(\frac{1}{2}||\beta||^2 + \frac{C}{2}\sum_{i=1}^{N} e_i^2 \right) \\ &\text{s.t.} \quad y_i = h(x_i)\beta + e_i, i = 1, \cdots, n \end{aligned} \tag{5}$$

其中, 目标函数的第一项为正则项, 用来控制模型的复杂度; 第二项为表示误差, $e_i \in \mathbf{R}^d$ 是第 i 个样本的误差向量, C 为惩罚系数.

近年来, Huang 等将 ELM 推广到无监督学习, 提出基于流形无监督极限学习机, 其模型为:

$$\begin{aligned}&\min_{\beta\in\mathbf{R}^{n_h\times d}}\left\{||\beta||^2+\lambda\mathrm{tr}(\beta^{\mathrm{T}}H(X)^{\mathrm{T}}LH(X)\beta)\right\}\\&\text{s.t.}\quad (H(X)\beta)^{\mathrm{T}}H(X)\beta=I\end{aligned}\tag{6}$$

第二项为流形正则项, 目的是使网络结构输出 Y 保持原输入数据 X 的流形结构不变, 其中, $\mathrm{tr}(\cdot)$ 表示矩阵的迹, L 为数据 X 的拉普拉斯矩阵, I 为单位阵, $H(X)\in\mathbf{R}^{n\times n_h}$ 为隐含层输出矩阵. US-ELM 将输入数据投影到 d 维空间中, 当 $d<m$ 时, US-ELM 是一种非线性降维方法.

2 基于稀疏和近邻保持的极限学习机降维

US-ELM 算法引入流形正则化的思想, 使得原始数据的流形结构经过 US-ELM 投影后得以保持, 即若在原空间近邻的两个样本在投影空间中仍然保持近邻[2]. US-ELM 算法的流形结构直接用 Gaussian 距离刻画, 随着数据维数的增加, 该距离度量的类分类性会随之减弱. 针对这一问题, 本文采用近邻表示来自适应地获取数据的低流形结构, 同时用稀疏表示来挖掘数据的全局结构. 在此基础上提出 SNP-ELM 算法, 使得数据在新的投影空间中保持其在原空间的近邻和稀疏表示结构.

2.1 近邻表示和稀疏表示

近邻表示[4]: 在样本集 $X=[x_1,x_2,\cdots,x_n]\in\mathbf{R}^{m\times n}$ 中用 x_i 的 k 近邻进行线性表示 x_i, 其表达式为:

$$\min\left\|x_i-\sum_{x_i\in N_k(x_i)}w_{ij}x_j\right\|\tag{7}$$

其中, $N_k(x_i)$ 表示 x_i 的 k 近邻, w_{ij} 为近邻表示系数, 当 $x_i\in N_k(x_i)$ 时, $w_{ij}=0$.

稀疏表示[5]: 样本 x_i 可大致由该样本集中的少量样本线性表示. 而当 x_i 由整个样本空间 X 进行线性表示时, 其表示系数是稀疏的. 其数学模型表示为:

$$\begin{aligned}&\min_{s_i}||s_i||_0\\&\text{s.t.}\quad x_i=Xs_i\end{aligned}\tag{8}$$

其中, $s_i\in\mathbf{R}^n$ 为稀疏表示系数, $||s||_0$ 是 s 非零元素个数. 由于 l_0 范数非凸且 NP 难, 因此用凸的 l_1 范数代替. 同时为了确保稀疏表示的平移不变性, 我们引入表示系数和为 1 的约束, 则式 (8) 变为:

$$\begin{aligned}&\min_{s_i}||s_i||_1\\&\text{s.t.}\quad x_i=Xs_i, 1=I_is_i\end{aligned}\tag{9}$$

其中, I_1 为所有元素均为 1 的 n 维向量. 式 (9) 是凸的, 可以利用线性规划方法求解, 如基追踪算法 (Basis pursuit, BP)[15].

2.2 基于稀疏和近邻保持的极限学习机降维算法

SNP-ELM 模型为:

$$\begin{aligned}\min_{\beta}\Big\{||\beta||_F^2+\lambda\sum_{i=1}^{n}||(h(x_i)\beta)^{\mathrm{T}}-\\(H(X)\beta)^{\mathrm{T}}(\delta w_i+\eta s_i)||_2^2\Big\}\end{aligned}\tag{10}$$

第二项的目的是使得投影后的数据保持原数据的近邻和稀疏表示结构, 其中 $W=[w_1,w_2,\cdots,w_n]$ 为近邻表示系数矩阵, w_i 表示 x_i 的近邻表示系数, 可以用模型 (7) 求解: $S=[s_1,s_2,\cdots,s_n]$ 为稀疏表示系数矩阵, s_i 表示 x_i 的稀疏表示系数, 可以用模型 (9) 求解. $\delta\in\mathbf{R}$ 和 $\eta\in\mathbf{R}$ 是权重系数, 分别反映 W 和 S 的重要性. 映射函数为 $y=f(x)=(h(x)\beta)^{\mathrm{T}}$.

令 $Z=\delta W+\eta S$, 则式 (10) 可写成:

$$\min_{\beta}\left\{||\beta||_F^2+\lambda\sum_{i=1}^{n}||(h(x_i)\beta)^{\mathrm{T}}-(H(X)\beta)^{\mathrm{T}}z_i||_2^2\right\}\tag{11}$$

通过简单的代数运算, 可以得到:

$$\begin{aligned}&\sum_{i=1}^{n}||(h(x_i)\beta)^{\mathrm{T}}-(H^{\mathrm{T}}(X)\beta)z_i||_2^2=\\&\quad\mathrm{tr}\left(\beta^{\mathrm{T}}\left(\sum_{i=1}^{n}(h^{\mathrm{T}}(x_i)-H^{\mathrm{T}}(X)z_i)\times\right.\right.\\&\quad\left.\left.(h^{\mathrm{T}}(x_i)-H^{\mathrm{T}}(X)z_i)^{\mathrm{T}}\right)\beta\right)\end{aligned}\tag{12}$$

令 e_i 为 n 维单位向量, 则式 (12) 等价于:

$$\begin{aligned}&\mathrm{tr}\left(\beta^{\mathrm{T}}\left(\sum_{i=1}^{n}(H^{\mathrm{T}}(X)e_i-H^{\mathrm{T}}(X)z_i)\right)\times\right.\\&\left.(H^{\mathrm{T}}(X)e_i-H^{\mathrm{T}}(X)z_i)^{\mathrm{T}})\beta\right)=\\&\mathrm{tr}\left(\beta^{\mathrm{T}}H^{\mathrm{T}}(X)\left(\sum_{i=1}^{n}(e_i-z_i)(e_i-z_i)^{\mathrm{T}}\right)H(X)\beta\right)=\\&\mathrm{tr}\left(\beta^{\mathrm{T}}H^{\mathrm{T}}(X)\left(\sum_{i=1}^{n}\left(e_ie_i^{\mathrm{T}}-z_ie_i^{\mathrm{T}}-\right.\right.\right.\\&\left.\left.\left.e_iz_i^{\mathrm{T}}z_iz_i^{\mathrm{T}}\right)\right)H(X)\beta\right)=\\&\mathrm{tr}(\beta^{\mathrm{T}}H^{\mathrm{T}}(X)(I-Z-Z^{\mathrm{T}}+Z^{\mathrm{T}}Z)H(X)\beta)\end{aligned}\tag{13}$$

式 (11) 可变形为:

$$\min_{\beta}\big\{||\beta||_F^2+\lambda\mathrm{tr}(\beta^{\mathrm{T}}H^{\mathrm{T}}(X)\times (I-Z-Z^{\mathrm{T}}+Z^{\mathrm{T}}Z)H(X)\beta)\big\} \tag{14}$$

为避免平凡解, 在此引入约束 $(H(X)\beta)^{\mathrm{T}}H(X)\beta=I$, 则模型变为:

$$\begin{aligned}\min_{\beta}\big\{&||\beta||_F^2+\lambda\mathrm{tr}(\beta^{\mathrm{T}}H^{\mathrm{T}}(X)\\&(I-Z-Z^{\mathrm{T}}+Z^{\mathrm{T}}Z)H(X)\beta)\big\}\\ \text{s.t.}\quad&(H(X)\beta)^{\mathrm{T}}H(X)\beta=I\end{aligned} \tag{15}$$

2.3 模型求解

为求解模型 (15), 利用拉格朗日乘子法, 得到以下拉格朗日函数:

$$\begin{aligned}L(\beta)=&\mathrm{tr}(\beta^{\mathrm{T}}\beta)+\frac{\lambda}{2}\mathrm{tr}(\beta^{\mathrm{T}}H^{\mathrm{T}}(X)AH(X)\beta)-\\&\theta\mathrm{tr}(\beta^{\mathrm{T}}H^{\mathrm{T}}(X)H(X)\beta-I)\end{aligned} \tag{16}$$

其中, $A=I-Z-Z^{\mathrm{T}}+Z^{\mathrm{T}}Z$, 令 $\frac{\partial L}{\partial\beta}=0$, 得:

$$\left(I+\frac{\lambda}{2}H^{\mathrm{T}}(X)AH(X)\right)\beta=\theta H^{\mathrm{T}}(X)H(X)\beta \tag{17}$$

求解广义特征值问题 (17) 得到最小的 d 个特征值及对应的特征向量构成最优的输出权重矩阵 β^*.

当 $n_h>n$ 时, $H^{\mathrm{T}}(X)H(X)$ 的维数比较高, 直接求解式 (17) 广义特征值问题, 需要消耗较大的内存. 为解决这个问题, 令 $\beta=H^{\mathrm{T}}(X)\alpha$, 式 (17) 两边同时左乘 $(H(X)H^{\mathrm{T}}(X))^{-1}H(X)$. 得到:

$$\left(I+\frac{\lambda}{2}AH(X)H^{\mathrm{T}}(X)\right)\alpha=\theta H(X)H^{\mathrm{T}}(X)\alpha \tag{18}$$

易知模型 (17) 与模型 (18) 具有相同特征值. 特征向量具有以下关系

$$\beta^*=H^{\mathrm{T}}(X)\alpha^* \tag{19}$$

因此解得广义特征值问题 (18) 的最小的 d 个特征值及对应的特征向量构成矩阵 α^*. 进而可获得模型 (17) 的解矩阵 $\beta^*=H^{\mathrm{T}}(X)\alpha^*$.

基于上述分析, 基于稀疏和近邻保持的极限学习机降维算法归纳如下:

算法 1. SNP-ELM 算法

输入: 数据矩阵 X, 参数 λ,δ,η.

输出: 降维后样本矩阵 Y.

1) 计算 k 近邻图.

2) 通过式 (8) 计算近邻重构矩阵 W.

3) 通过式 (10) 计算稀疏重构矩阵 S, 计算 $Z=\delta W+\eta S$, $A=I-Z-Z^{\mathrm{T}}+Z^{\mathrm{T}}Z$.

4) 初始化 ELM 网络, n_h 为隐藏层节点个数, 随机初始化输入权重 $a\in\mathbf{R}^{n\times n_h}$, 偏置 $b\in\mathbf{R}^{n_h}$ 根据式 (3) 计算隐藏层输出矩阵 $H(X)\in\mathbf{R}^{n\times n_h}$.

5) 当 $n>n_h$ 时, 利用式 (17) 计算得到输出权重矩阵 β; 否则, 利用式 (18) 计算得到 α, 再计算输出权重矩阵 $\beta=H^{\mathrm{T}}(X)\alpha$.

6) 计算降维后样本矩阵 $Y=H(X)\beta$

2.4 算法分析

SNP-ELM 算法中计算 k 近邻图的时间复杂度是 $\mathrm{O}(mn\log n)$; 计算近邻重构矩阵 W 是求解了 n 次式 (8), 其时间复杂度为 $\mathrm{O}(nk^3)$; 用 BP 算法求解式 (10) 的时间复杂度为 $\mathrm{O}(n^3)$, 因此计算稀疏重构矩阵 S 的时间复杂度为 $\mathrm{O}(n^4)$; 计算广义特征值式 (18) 的时间复杂度为 $\mathrm{O}(n_h^3)$, 求解广义特征值式 (20) 的时间复杂度为 $\mathrm{O}(n^3)$. 因此 SNP-ELM 算法的时间复杂度为 $\mathrm{O}(mn\log n+n^4+nk^3)$.

3 实验

本文提出的 SNP-ELM 降维算法有两个重要目的, 其一是便于高维数据的可视化分析, 其二是面向聚类分析的降维可有效地提高聚类准确性, 故进行数据可视化及高维基因数据降维聚类实验, 两个实验的实验环境均为 Win7 系统, 内存 4 GB, 所有方法均用 Matlab2012b 编程实现. 两个实验均采用相同的参数设置, LPP、NPE、US-ELM 和 SNP-ELM 的近邻数 k 均设为 5; US-ELM 和 SNP-ELM 的隐藏层节点个数均设为 1 000; US-ELM 的参数 λ 及 SNP-ELM 的参数 λ 统一取 $\{10^{-4},10^{-3},\cdots,10^{4}\}$, SNP-ELM 的参数 δ 和 η 的搜索范围为 $[-1,1]$, 变化步长为 0.2.

本文实验所对比的降维方法主要有以下几种: 1) 线性降维方法: PCA、LPP、NPE 和 SPP; 2) 非线性降维方法: LLE 和 US-ELM. 其中 LPP、NPE、LEE 和 US-ELM 都使得降维后的数据保持原数据的近邻结构, SPP 保持数据的稀疏表示结构, PCA 的目标是使得降维后数据方差最大.

3.1 数据可视化实验

本实验中, 我们分别用 PCA、LPP、NPE、SPP、LEE、US-ELM 和 SNP-ELM 7 种方法将一个人造数据和一个真实的 UCI 数据 Wine 分别投影到一维和二维空间, 直观地展示 SNP-ELM 算法的性能, 并选取每个降维方法的最优结果进展示.

1) 一维可视化

本实验使用的三维人造数据如图 2 所示, 该数据包含 3 类, 每类有 50 个样本, 该实验分别将数据降到一维, 实验结果如图 3 所示.

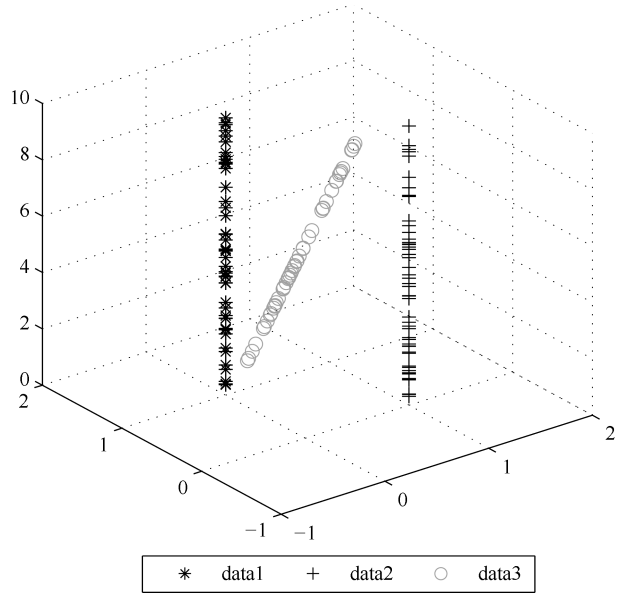

图 2 人造数据集

Fig. 2 The toy dataset

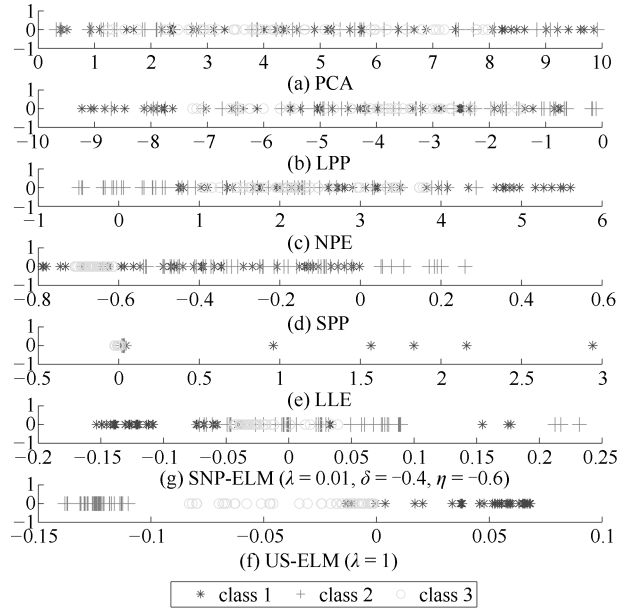

图 3 人造数据一维可视化结果

Fig. 3 The 1D visualization results of toy dataset

从图 3 可以看出 PCA 以投影后的样本方差最大为目标, 其降维结果近似于把该数据投影到 Z 轴方向, 但其将该数据投影到一维时三类数据的可分性较差. LPP、NPE、LLE 和 US-ELM 均以降维后样本保持原样本的近邻结构为目的, 因此其降维效果略有改善. 其中 LLE 和 US-ELM 是非线性降维方法, 其降维后不同类样本的分离程度较 LPP 和 NPE 高些. 稀疏保持投影方法 SPP 以降维后样本保持原样本的稀疏结构为目的, 该方法将数据投影到一维后不同类样本的分离程度与 US-ELM 相当. SNP-ELM 是一种非线性降维方法, 它使降维后样本同时保持数据的近邻结构和稀疏结构不变. SNP-ELM 虽然无法使得该数据投影到一维后三类样本完全分离, 但其降维后不同类样本可分性是 7 种降维方法中最优的, 只有少数第三类样本与第二类样本相互重叠.

2) 二维可视化

本实验使用 UCI 数据集 Wine 数据, Wine 数据包含来自 3 个类的 178 个样本, 每个样本有 14 个特征. 实验结果如图 4 所示.

由图 4 可以看出, 经 7 种降维方法将 Wine 数据投影到 2 维时仍无法完全分离 3 类样本. 但从不同类样本的重叠程度上可以看出, SPP 将数据降到二维后 3 类数据完全重叠在一起, 降维效果最差. 用 PCA、LPP、NPE、LLE 和 US-ELM 这 5 种方法降维后第一类数据能较好地分离, 而第二类和第三类数据完全重叠在一起. 本文方法将 Wine 数据降到二维后, 不同类数据的重叠程度最低, 不同类样本的可分性最好.

3.2 基因表达数据实验

本实验采用高维基因表达数据测试本文方法与对比方法面向聚类任务时的降维效果. 为了观察本文降维方法将数据投影到不同维数, 特别是投影到较低维时基因表达数据聚类效果, 将数据分别投影到 $2^1, 2^2, 2^3, 2^4, \cdots, 2^n$ 维. 该实验以降维后样本的 k-means 聚类准确率衡量降维质量, 实验中的聚类准确类采用文献 [13] 的计算方法. 计算公式如下:

$$ACC = \frac{1}{n}\sum_{i=1}^{n}\delta(s_i, \mathrm{map}(r_i)) \tag{20}$$

其中, n 为样本数, $\delta(x, y)$ 表示当 $x = y$ 时, $\delta = 1$, 否则 $\delta = 0$; s_i 和 r_i 分别为样本原始类标签和经聚类算法聚类后得到的类标签: $\mathrm{map}(r_i)$ 将聚类得到的类标签映射成与样本数据自带的类标签等价的类标签.

1) 实验数据集

实验所选用的 6 个公开的基因数据集: SBCRT、DLBCL、Leukemia2、Prostate[16]、Prostate0 和 Colon[17], 这些数据的详细描述见表 1.

2) 聚类准确率比较

为减少 k-means 初始中心随机选取以及 US-ELM 和 SNP-ELM 方法随机权重产生的随机误差. 为便于比较, 减少实验结果随机性的影响, 实验中 US-ELM 和 SNP-ELM 分别运行 10 次, 再将每次降维后数据集执行 10 次 k-means 聚类, 取 100 次聚类准确率的平均值作为各自方法的最终准确率, 而其他降维方法的聚类准确率是 10 次 k-means

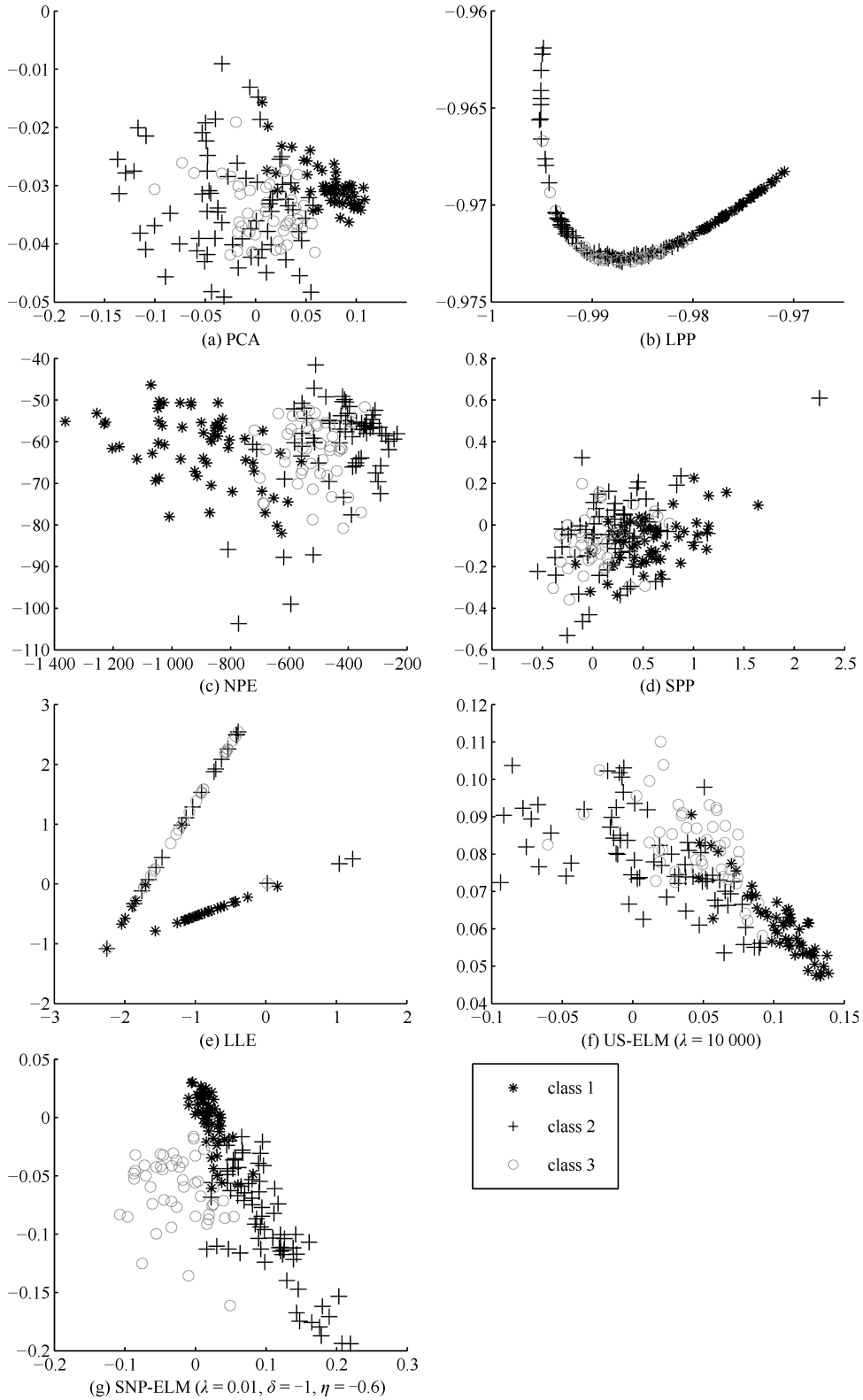

图 4　Wine 数据二维可视化结果

Fig. 4　The 2D visualization results of Wine

表 1 基因表达数据集描述

Table 1 Summary of gene expression data sets

数据集	样本数	基因数 (维数)	类别数
SRBCT	83	2 308	4
DLBCL	77	5 469	2
Prostate0	102	6 033	2
Prostate	102	10 509	2
Leukemia2	72	11 225	3
Colon	62	2 000	2

聚类准确率的平均值. 最终实验结果如表 2 所示, 表中给出聚类准确率的均值 (方差、维数), 其中维数为每个数据最优聚类结果所对应的维数. 对两种极限学习机降维方法 US-ELM 和 SNP-ELM 分别给出最优聚类结果所对应的参数. LPP、NPE、LLE、US-ELM 和 SNP-ELM 这 5 种方法都在降维时保持了原始数据的近邻结构, SPP 和 SNP-ELM 都保持了原始数据的稀疏结构, 其中 LLE、US-ELM 和 SNP-ELM 是非线性降维方法, SNP-ELM 同时保持原始数据的近邻结构和稀疏结构. 将这 5 种方法降维后的聚类准确率进行对比可以发现: 1) 将 NPE 和 LPP 分别与 LLE 和 US-ELM 的准确率进行对比, 可以发现后者的准确率比前者高, 这是因为 LEE 和 US-ELM 分别是 NPE 和 LPP 的非线性推广, 非线性降维方法更适用于非线性分布的基因表达数据. 2) SPP 与 LPP、NPE 进行比较其聚类结果各有千秋, 在 DLBCL、Prostate0 和 Colon 这 3 个数据集上 SPP 的结果较好, 而在其他数据集上 LPP 和 NPE 的结果较好, 这说明稀疏保持和近邻保持各有优势. 3) SNP-ELM 的聚类准确率是最高的, 其主要原因是 SNP-ELM 既是非线性降维方法, 又同时保持了原始数据的近邻表示结构和稀疏表示结构使得降维后低维空间的数据保持了更多的判别信息. 将表 2 中的所有方法进行对比, 可以发现基于 ELM 的 2 种降维方法的准确率普遍优于其他降维方法. 特别是 SNP-ELM 算法考虑到降维后样本局部近邻关系和全局稀疏性保持不变, 从而使其在全部 6 个基因数据降维后的聚类准确率最高, 且高于其他方法及 US-ELM 方法 10 % 以上. 这说明 SNP-ELM 是一种有效的高维非线性降维方法.

为进一步对比几种降维方法在不同维数下的聚类准确率, 分别选取目标维数 2, 4, 8, 16, 32, $\cdots$ 执行各种降维算法, 各种降维算法在不同维数下的聚类准确率如图 5 所示. 从图 5 可以看出 SNP-ELM 及其余 6 种降维算法将 6 个数据集投影到相

表 2 基因数据集上聚类准确率 (%)

Table 2 Clustering accuracy comparison (variance) on gene expression data sets (%)

Data	k-means	PCA	LPP	NPE	SPP	LLE	US-ELM (λ)	SNP-ELM (λ, η, δ)
Leukemia2	63.89 (0.00)	63.89 (0.00, 2)	70.72 (3.20, 4)	63.89 (0,32)	59.72 (0.00, 72)	65.83 (6.65,4)	64.44 (1.34, 2) (0.0001)	87.17 (3.56, 8) (0.0001,−1,−1)
SRBCT	43.61 (6.27)	48.86 (2.09, 83)	64.19 (2.21, 83)	48.43 (0.76, 8)	38.55 (0.00, 2)	49.76 (4.33, 8)	64.55 (10.29, 8) (0.1)	82.92 (6.03, 8) (0.001, −0.4, 0)
DLBCL	68.83 (0.00)	68.83 (0.00, 2)	63.55 (1.86, 8)	69.09 (0.82, 32)	74.02 (0.00, 4)	72.23 (0.00, 2)	76.62 (0.00, 32) (0.0001)	86.34 (1.78, 8) (0.001, 0.2, −0.6)
Prostate0	56.86 (0.00)	56.83 (0.00, 2)	56.86 (0.00, 2)	56.86 (0.00, 4)	59.80 (0.00, 102)	56.96 (0.93, 4)	64.09 (5.83, 2) (0.01)	82.92 (2.19, 102) (0.1, 0.2, 0.8)
Prostate	63.33 (0.83)	63.73 (0.00, 2)	59.80 (0.00, 2)	59.80 (0.00,4)	56.86 (0.00,102)	59.51 (0.93, 4)	67.57 (5.83, 2) (0.0001)	82.73 (2.19,102) (1, −1, 0.6)
Colon	54.84 (0.00)	54.84 (0.00, 2)	54.84 (0.00, 2)	56.45 (0.00, 2)	64.19 (0.68, 62)	59.52 (6.99, 32)	67.06 (4.19, 32) (0.0001)	85.95 (3.69, 8) (0.001, −0.8,1)

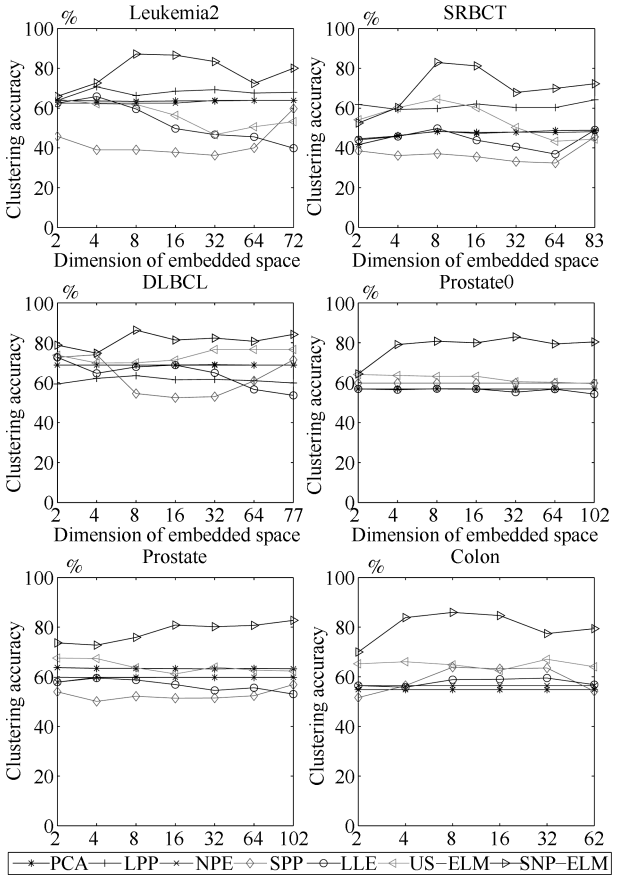

图 5　将 6 个数据集映射到不同维度特征空间时的聚类准确率

Fig. 5　Clustering accuracy on six gene expression data in different dimensions

同维数的特征空间时, SNP-ELM 的聚类准确率都是最高的. 而对于 SNP-ELM 算法, 除 Prostate 和 Prostate0 两个数据集, 在其他 4 个基因数据集上都在 8 维处得到最高的聚类准确率.

3) 参数分析

SNP-ELM 模型有 3 个参数 λ, δ 和 η, 其中 λ 为正则参数. δ 和 η 为权重系数, 分别表示近邻重构系数和稀疏重构系数的重要性. 本节讨论不同参数对实验结果的影响, 由前面的实验结果可知将基因表达数据降到 8 维时能够得到较高的聚类准确率, 因此在进行参数分析时我们固定维数为 8. 根据 3 个参数在 SNP-ELM 中的不同作用, 将其分为两组分别进行分析, 正则参数 λ 单独分析, 权重系数 δ 和 η 一起分析. 其中, λ 的取值范围为 $\{10^{-4}, 10^{-3}, \cdots, 10^{4}\}$, δ 和 η 的取值范围为 $[-1, 1]$, 取值步长为 0.2.

图 6 给出 $\delta = \eta = -0.2$ 时, SNP-ELM 降维的聚类准确率随参数 λ 不同取值的变化情况. 从图 6 可以看出, 除了 Leukema2 在 $\lambda = 10^{-4}$ 时聚类准确率达到最高, 其余 5 个基因表达数据均在 $\lambda = 10^{-3}$ 时聚类准确率达到最高. 这说明对高维基因数据而言, λ 取较小值时本文方法能达到较好效果.

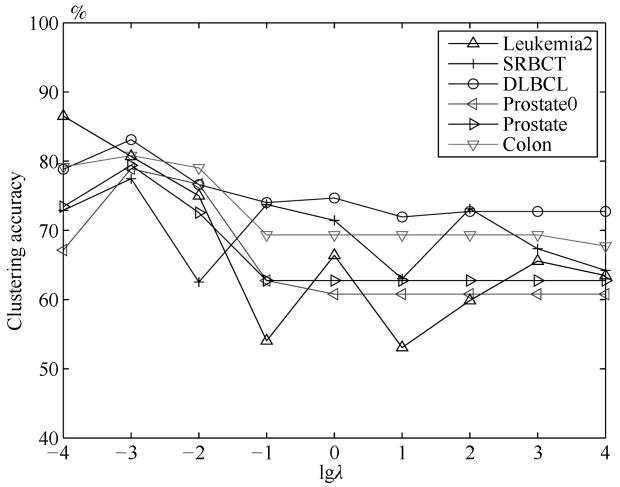

图 6　聚类准确率随参数 λ 的变化情况 ($\delta = \eta = -0.2$)

Fig. 6　Variation of accuracy with respect of parameters λ ($\delta = \eta = -0.2$))

图 7 给出 $\lambda = 0.001$ 时, 不同 δ 和 η 取值下的聚类准确率. 从图 7 可以看出当 δ 取值自 $[-0.6, -0.2]$, η 取值自 $[-0.2, 0.2]$ 时, 对高维基因表达数据而言 SNP-ELM 算法可以取得较高的聚类准确率.

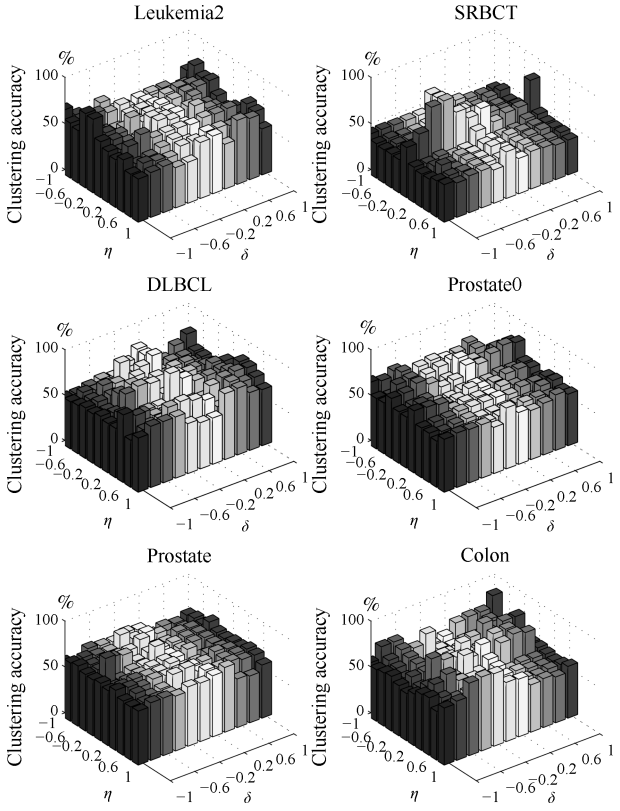

图 7　不同 δ 和 η 取值下的聚类准确率 ($\lambda = 0.001$)

Fig. 7　ariation of accuracy with respect of parameters δ and η ($\lambda = 0.001$)

4 结论

目前, ELM 模型主要用于有监督分类或回归问题, 本文则对 ELM 模型推广到无监督降维问题进行了进一步研究, 提出基于稀疏和近邻结构保持的极限学习机降维算法 SNP-ELM. SNP-ELM 通过模型优化求解计算近邻表示系数, 具有一定的数据自适应性, 实验结果表明 SNP-ELM 算法在 Wine 数据和基因表达数据集上性能优于其他对比方法. 从研究中我们可以得到以下 2 个结论: 1) 对 Wine 数据、高维基因表示数据降维时, 同时考虑稀疏结构和近邻结构比只考虑单一结构更有效; 2) 基于 ELM 的非线性降维方法在 Wine 数据和基因表达数据上优于线性降维方法.

References

1 Jolliffe I T. *Principal Component Analysis*. Berlin: Springer-Verlag, 2002.

2 He X F, Niyogi P. Locality preserving projections. In: Proceedings of 2003 Neural Information Processing Systems. Vancouver, Canada: NIPS, 2004. 153−160

3 He X F, Cai D, Yan S C, Zhang H J. Neighborhood preserving embedding. In: Proceedings of the 10th IEEE International Conference on Computer Vision. Beijing, China: IEEE, 2005. 1208−1213

4 Qiao L S, Chen S C, Tan X Y. Sparsity preserving projections with applications to face recognition. *Pattern Recognition*, 2010, **43**(1): 331−341

5 Schölkopf B, Smola A J, Müller K R. Kernel principal component analysis. In: Proceedings of the 7th International Conference on Artificial Neural Networks. Switzerland: Springer, 1997. 583−588

6 Roweis S T, Saul K L. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 2010, **290**(5500): 2323−2326

7 Huang G B, Ding X J, Zhou H M. Optimization method based extreme learning machine for classification. *Neurocomputing*, 2010, **74**(1−3): 155−163

8 Peng Y, Wang S H, Long X Z, Lu B L. Discriminative graph regularized extreme learning machine and its application to face recognition. *Neurocomputing*, 2015, **149**: 340−353

9 Peng Y, Lu B L. Discriminative manifold extreme learning machine and applications to image and EEG signal classification. *Neurocomputing*, 2016, **174**: 265−277

10 Zhang K, Luo M X. Outlier-robust extreme learning machine for regression problems. *Neurocomputing*, 2015, **151**: 1519−1527

11 Huang G, Song S J, Gupta J N D, Wu C. Semi-supervised and unsupervised extreme learning machines. *IEEE Transactions on Cybernetics*, 2014, **44**(12): 2405−2417

12 Liu Zhan-Jie, Chen Xiao-Yun. Local subspace clustering. *Acta Automatica Sinica*, 2016, **42**(8): 1238−1247
(刘展杰, 陈晓云. 局部子空间聚类. 自动化学报, 2016, **42**(8): 1238−1247)

13 Wang Wei-Wei, Li Xiao-Ping, Feng Xiang-Chu, Wang Si-Qi. A survey on sparse subspace clustering. *Acta Automatica Sinica*, 2015, **41**(8): 1373−1384
(王卫卫, 李小平, 冯象初, 王斯琪. 稀疏子空间聚类综述. 自动化学报, 2015, **41**(8): 1373−1384)

14 Kasun L L C, Yang Y, Huang G B, Zhang Z Y. Dimension reduction with extreme learning machine. *IEEE Transactions on Image Processing*, 2016, **25**(8): 3906−3918

15 Chen S S, Donoho D L, Saunders M A. Atomic decomposition by basis pursuit. *SIAM Review*, 2001, **43**(1): 129−159

16 Yu L, Ding C, Loscalzo S. Stable feature selection via dense feature groups. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Las Vegas, USA: ACM, 2008. 803−811

17 Gene expression model selector [online], available: http://www.gems-system.org, October 9, 2017.

陈晓云 福州大学数学与计算机科学学院教授. 主要研究方向为数据挖掘, 模式识别. 本文通信作者.
E-mail: c_xiaoyun@fzu.edu.cn
(**CHEN Xiao-Yun** Professor at the College of Mathematics and Computer Science, Fuzhou University. Her research interest covers data mining and pattern recognition. Corresponding author of this paper.)

廖梦真 福州大学数学与计算机科学学院硕士研究生. 主要研究方向为数据挖掘, 模式识别.
E-mail: liao_mengzhen@163.com
(**LIAO Meng-Zhen** Master student at the College of Mathematics and Computer Science, Fuzhou University. Her research interest covers data mining and pattern recognition.)