

基于近邻图改进的块对角子空间聚类算法

王丽娟¹, 陈少敏¹, 尹明^{2*}, 许跃颖³, 郝志峰^{1,4}, 蔡瑞初¹, 温雯¹

(1. 广东工业大学 计算机学院, 广州 510006; 2. 广东工业大学 自动化学院, 广州 510006;
3. 北京师范大学珠海分校 信息技术学院, 广东 珠海 519000; 4. 佛山科学技术学院 数学与大数据学院, 广东 佛山 528000)
(* 通信作者电子邮箱 yiming@gdut.edu.cn)

摘要:块对角表示(BDR)模型可以通过利用线性表示对数据有效地进行聚类,却无法很好地利用高维数据常见的非线性流形结构信息。针对这一问题,提出了基于近邻图改进的块对角子空间聚类(BDRNG)算法来通过近邻图来线性拟合高维数据的局部几何结构,并通过块对角约束来生成具有全局信息的块对角结构。BDRNG同时学习全局信息以及局部数据结构,从而获得更好的聚类表现。由于模型包含近邻图算子和非凸的块对角表示范数,BDRNG采用了交替最小化来优化求解算法。实验结果如下:在噪声数据集上,BDRNG能够生成稳定的块对角结构系数矩阵,这说明了BDRNG对于噪声数据具有鲁棒性;在标准数据集上,BDRNG的聚类表现均优于BDR,尤其在人脸数据集上,相较于BDR,BDRNG的聚类准确度提高了8%。

关键词:近邻图;块对角表示;稀疏表示;子空间聚类;高维数据
中图分类号:TP181 **文献标志码:**A

Improved block diagonal subspace clustering algorithm based on neighbor graph

WANG Lijuan¹, CHEN Shaomin¹, YIN Ming^{2*}, XU Yueying³,
HAO Zhifeng^{1,4}, CAI Ruichu¹, WEN Wen¹

(1. School of Computers, Guangdong University of Technology, Guangzhou Guangdong 510006, China;
2. School of Automation, Guangdong University of Technology, Guangzhou Guangdong 510006, China;
3. School of Information Technology, Beijing Normal University, Zhuhai, Zhuhai Guangdong 519000, China;
4. School of Mathematics and Big Data, Foshan University, Foshan Guangdong 528000, China)

Abstract: Block Diagonal Representation (BDR) model can efficiently cluster data by using linear representation, but it cannot make good use of non-linear manifold information commonly appeared in high-dimensional data. To solve this problem, the improved Block Diagonal Representation based on Neighbor Graph (BDRNG) clustering algorithm was proposed to perform the linear fitting of the local geometric structure by the neighbor graph and generate the block-diagonal structure by using the block-diagonal regularization. In BDRNG algorithm, both global information and local data structure were learned at the same time to achieve a better clustering performance. Due to the fact that the model contains the neighbor graph and non-convex block-diagonal representation norm, the alternative minimization was adopted by BDRNG to optimize the solving algorithm. Experimental results show that: on the noise dataset, BDRNG can generate the stable coefficient matrix with block-diagonal form, which proves that BDRNG is robust to the noise data; on the standard datasets, BDRNG has better clustering performance than BDR, especially on the facial dataset, BDRNG has the clustering accuracy 8% higher than BDR.

Key words: neighbor graph; Block Diagonal Representation (BDR); sparse representation; subspace clustering; high-dimensional data

0 引言

高维数据聚类是图像表示和压缩^[1]、基因表达分析^[2]和计算机视觉领域^[3]中重要的研究课题之一。高维数据聚类会导致严重的维度灾难,同时花费较大内存和较长的运算时间。

近年来,子空间聚类被视为高维数据聚类分析的有效方法之一。子空间聚类假设高维数据是多个低维子空间的并集,数据可以在低维空间中进行重构,从而降低高维数据对内存和时间复杂度的要求^[4]。

稀疏表示(Sparse representation)旨在通过特定空间数据

收稿日期:2020-05-31;修回日期:2020-09-02;录用日期:2020-09-03。

基金项目:国家自然科学基金资助项目(61502108,61876042,61876043);NSFC-广东联合基金资助项目(U1501254)。

作者简介:王丽娟(1978—),女,河北邢台人,副教授,博士,主要研究方向:数据挖掘、机器学习;陈少敏(1994—),女,广东汕头人,硕士研究生,主要研究方向:数据挖掘、子空间聚类;尹明(1975—),男,湖南永州人,副教授,博士,主要研究方向:机器学习、模式识别、图像处理;许跃颖(1984—),男,广东汕头人,讲师,硕士,主要研究方向:互联网应用;郝志峰(1968—),男,江苏苏州人,教授,博士,主要研究方向:机器学习、人工智能;蔡瑞初(1983—),男,浙江温州人,教授,博士,主要研究方向:机器学习、数据挖掘;温雯(1981—),女,江西赣州人,副教授,博士,主要研究方向:支持向量机、模式识别。

的稀疏表示重构数据或揭示数据本质特征^[5]。稀疏表示在新的基或字典基础上使用尽可能少的非零系数重构数据。Elhamifar 等^[6]基于一维稀疏性提出了稀疏子空间聚类 (Sparse Subspace Clustering, SSC), 使每个数据仅用一个子空间其他数据的线性组合来表示。SSC 仅考虑每个数据的稀疏表示, 缺乏对数据全局结构的描述。为获得数据集全局结构, 2010 年, Liu 等^[7]进一步利用了二维稀疏性提出了基于低秩表示 (Low-Rank Representation, LRR) 模型的子空间聚类方法。然而, LRR 通过奇异值分解的方式来求解秩的最小化, 复杂度较高, 不利于实际应用。在理想情况下, 即子空间相互独立的情况下, 这些方法及其拓展方法通过稀疏或者低秩约束保证了系数表示矩阵的块对角结构。块对角的系数表示矩阵能够使得: 同一子空间内数据靠近, 不同子空间数据远离, 从而获得较好的聚类结果。因此, 系数矩阵呈块对角结构被认为是解决子空间聚类问题重要性质。由于实际数据集往往含有噪声和异常数据, 无法满足子空间聚类的理想条件。针对这些问题, 人们提出了一系列的改进模型和方法。Feng 等^[8]提出了显式的块对角约束的 SSC (Block Diagonal-SSC, BD-SSC) 和 LRR (Block Diagonal-LRR, BD-LRR) 通过在稀疏表示模型的基础上加入块对角约束使得稀疏表示模型的系数表示矩阵块对角更加稳定。Lu 等^[9]提出了最小二乘回归子空间聚类 (Least Squares Regression for subspace clustering, LSR), 可以将高度相关的数据聚合在一起, 与此同时, LSR 对于噪声具有一定的鲁棒性。在 LSR 提出的强制块对角 (Enforced Block Diagonal, EBD) 的基础上, Lu 等^[10]继续拓展了 EBD, 提出了块对角表示 (Block Diagonal Representation, BDR) 模型, 在数据稀疏优化同时增加改进块对角约束, 简化了优化过程, 提升了优化效率。与 SSC 和 LRR 相比, BDR 的稀疏优化和块对角约束同时得以满足实现, 能够确保子空间聚类结果更优, 使得同一类数据彼此靠近, 不同类数据相互远离。

稀疏表示的子空间聚类方法希望用尽可能少的数据点线性组合来表达高维数据。实际上, 高维数据集内嵌不同的流形几何结构, 而且数据点通常分布在这些流形结构上。这些非线性流形几何结构中, 往往蕴含着一些关于图像的细节内容的描述, 可以有助于模型对于数据的结构学习。然而, 线性的子空间表示方法无法有效利用高维数据中非线性的流形几何结构。针对这一问题, 研究者提出基于核的子空间聚类方法, 学习高维数据中的非线性信息。Patel 等^[11]提出了核稀疏子空间聚类 (Kernel Sparse Subspace Clustering, KSSC) 以及 Xiao 等^[12]提出了核低秩表示 (Kernel Low-Rank Representation, KLRR) 模型。这些方法通过在稀疏表示模型中引入核方法, 强化模型学习非线性流形结构的能力。Ji 等^[13]在自适应低秩核子空间聚类 (adaptive Low-Rank Kernel Subspace Clustering, LKSC) 中提出了一种能够处理非线性模型的内核子空间聚类方法, 使得隐式特征空间中的映射数据不仅低秩, 而且具有自表达能力; Xie 等^[14]在隐式块对角线低秩表示 (Implicit Block Diagonal Low-rank Representation, IBDLR) 模型中提出将隐式特征表示和块对角线合并到 LRR 模型中, 既解决了解决隐式特征空间中的聚类问题, 同时使得获得的稀疏表示矩阵呈块对角状。但是, 通过核方法求解高维数据聚类的方式往往需要较大时间和内存开销, 因此无法应用于实际应用之中。

为了解决上述问题, 本文提出一种基于近邻图的块对角

子空间聚类算法, 即基于近邻图改进的块对角表示 (improved Block Diagonal Representation based on Neighbor Graph, BDRNG) 模型。通过近邻图学习, 可以通过局部线性来拟合局部流形结构, 从而获得局部流形结构。相比起复杂耗时的核方法, 近邻图学习简单, 通过更少的时耗获得数据集内的流形结构信息。BDRNG 通过块对角约束来揭示高维数据的全局结构; 通过近邻图来学习高维数据内在的局部流形结构信息。本文通过仿真实验和真实数据集的验证表明, 本文方法优于其他相同类型的子空间聚类方法。

1 相关工作

1.1 子空间聚类

设数据集 $X = [x_1, x_2, \dots, x_n] \in \mathbb{R}^{D \times n}$, 数据属于 k 个子空间 $\{S_i\}_{i=1}^k$ 的并集。稀疏子空间聚类将数据表示为同一子空间内其他数据的线性组合, 即, 同一类数据可以相互表示 $X = XZ$, 其中 $Z \in \mathbb{R}^{n \times n}$ 为稀疏系数矩阵。换言之, 当 $x_i \in X_j (i \neq j)$ 时, $Z_{i,j} \neq 0$; 而不同类数据的表示系数 $Z_{i,j}$ 则为 0, 即当 $x_i \notin X_j (i \neq j)$ 时, $Z_{i,j} = 0$ 。在实际情况中, 数据往往受到噪声和奇异样本影响, 所以, 数据常表示为 $X = XZ + E$, E 为奇异样本或者噪声。一般稀疏表示的子空间聚类模型可以统一描述为:

$$\min_Z J(Z) = F(E) + \lambda R(Z)$$

s.t. $Z \in C$

其中: $J(Z)$ 为目标函数, $F(E)$ 为数据项, $R(Z)$ 为正则项或者惩罚项; C 为系数矩阵 Z 的约束集合, $\lambda \geq 0$ 为正则参数。稀疏表示的子空间聚类算法对系数表示矩阵 $\|Z\|$ 引入不同的稀疏约束来发现数据内在的结构, 例如基于一维稀疏性的 SSC 或者基于二维稀疏性的 LRR 等。在理想子空间条件下, Z 具有理想的块对角结构。最后, 利用 Z 构造数据的相似度矩阵 $Y = \frac{\|Z\| + \|Z\|^T}{2}$, 并通过谱聚类算法^[15]获得最后的聚类结果。

1.2 块对角约束

当子空间聚类的系数表示矩阵具有块对角结构, 聚类结果较好。从直观意义上理解, 块对角结构可以很好地描述聚类目标——类内数据密集、类间数据稀疏。Feng 等^[8]提出了 BD-SSC 和 BD-LRR, 将块对角约束引入 SSC 和 LRR。但是, 块对角约束属于 NP-hard 问题, 通过投影迭代法优化, 计算效率不高。Lu 等^[10]提出了块对角表示模型 BDR。在 BDR 中, 块对角约束通过 ℓ_1 -范数松弛, 正则项可以有效优化。BDR 直接约束系数表示矩阵为块对角形式, 具有较高的优化效率, 同时系数表示矩阵的块对角形式更加稳定。

假设矩阵 $Z \in \mathbb{R}^{n \times n}$ 是系数表示矩阵, 满足 $Z \geq 0$ 且 $Z = Z^T$ 。 Z 的拉普拉斯矩阵记为 L_Z , $L_Z = \text{diag}(Z\mathbf{1}) - Z$, 其中, $\mathbf{1}$ 表示全为 1 的向量。已知数据集为 k 类, k -块对角正则项定义如下:

$$\|Z\|_k = \sum_{i=n-k+1}^n \lambda_i(L_Z) \quad (1)$$

其中, $\lambda_i(L_Z)$ 是 L_Z 的降序特征值, $i = 1, 2, \dots, n$, 且因为 $L_Z > 0$ 所以 $\lambda_i(L_Z) \geq 0$, k 代表类数。

1.3 近邻图学习

高维空间下, 数据点沿着高维数据内嵌的流形结构分布。因此, 流形学习被视作为高维数据非线性降维的常用方法。

流形学习的局部不变性假定,对于某一特定的数据点以及在原始空间中与该点接近的邻域点,在投影空间中依然接近^[16]。这些局部几何结构可以由近邻图表达,通过近邻图局部线性拟合数据点,从而学习到局部几何结构。对于给定的数据集 X , 每个数据点 x_i 的 k 个近邻信息记录在邻接矩阵 $N(i)$ 中, $W_{i,j}$ 表示第 j 数据点对于第 i 数据点局部线性拟合重建的权重贡献。权重矩阵 W 学习的目标函数旨在能够通过局部线性表达拟合几何结构来尽可能多地保留局部信息,从而使得数据与其线性拟合重构数据的误差尽可能小,定义如下:

$$\begin{aligned} & \min_{W} \left\| x_i - \sum_{x_j \in N(i)} W_{i,j} x_j \right\|_F^2 \\ \text{s.t. } & \sum_{x_j \in N(i)} W_{i,j} = 1 \end{aligned} \quad (2)$$

因此,通过利用近邻图捕获的局部几何结构信息,可以弥补以往线性子空间聚类表示模型对于高维数据中流形几何结构信息处理的不足,从而提高子空间聚类表示模型对高维数据的聚类准确度^[17]。

2 近邻图增强块对角子空间聚类

近邻图能够有效地学习数据局部结构信息,块对角表示可以获得全局的数据结构信息,因此,本文提出在块对角表示模型中引入了近邻图学习,从而改善目前子空间聚类算法无法有效利用高维数据的局部几何结构信息的问题。本文将式(2)改写为矩阵形式:

$$\begin{aligned} & \text{tr}(X(I - W^*)^T(I - W^*)X^T) \\ \text{s.t. } & \sum_{x_j \in N(i)} W_{i,j} = 1 \end{aligned} \quad (3)$$

当获得权重矩阵的最优解 W^* , 设 $L_M = (I - W^*)^T(I - W^*)$, 可以构建数据局部几何结构正则项—— $\text{tr}(XL_M X^T)$ 。

2.1 近邻图正则化的块对角表示模型

在基于块对角表示模型的子空间算法引入数据局部结构信息,从而增强模型对于高维数据内嵌的流形结构的学习。本文将 $\text{tr}(XL_M X^T)$ 作为近邻图正则项加入到本文模型中,提出 BDRNG 模型如下:

$$\begin{aligned} & \min_{Z,B} \frac{1}{2} \|E\|_F^2 + \beta \text{tr}(ZL_M Z^T) + \gamma \|Z\|_k \\ \text{s.t. } & X = XZ + E, \text{diag}(Z) = 0, Z = Z^T, Z \geq 0 \end{aligned} \quad (4)$$

其中: Z 为系数表示矩阵, β, γ 为用于平衡对应项的正系数。由于数据集 X 学到的流形结构对其对应的块对角表示系数矩阵 Z 也有相似的流形结构,可以将 X 替换为 Z ^[18]。块对角约束需要强制 Z 为非负和对称,从而限制 Z 的表示能力。根据文献[10],引入中间项 B , 减少块对角约束对 B 的影响,并将模型改写为:

$$\begin{aligned} & \min_{Z,B} \frac{1}{2} \|E\|_F^2 + \frac{\lambda}{2} \|Z - B\|_F^2 + \beta \text{tr}(ZL_M Z^T) + \gamma \|B\|_k \\ \text{s.t. } & X = XZ + E, \text{diag}(Z) = 0, Z = Z^T, Z \geq 0 \end{aligned} \quad (5)$$

其中: λ 为平衡松弛项的正系数。 $\frac{\lambda}{2} \|Z - B\|_F^2$ 可以将 Z, B 的求解子问题变为凸问题,从而可以得到唯一的稳定解。 $\|B\|_k$ 改写为 $\min_G \langle L_B, G \rangle$, 其中 $0 < G < 1, \text{tr}(G) = k$ 。根据本文 1.2 节中的内容,可以得到 BDRNG 最终的模型为:

$$\begin{aligned} & \min_{Z,B,G} \frac{1}{2} \|E\|_F^2 + \frac{\lambda}{2} \|Z - B\|_F^2 + \beta \text{tr}(ZL_M Z^T) + \\ & \lambda \langle \text{diag}(B1) - B, G \rangle \end{aligned} \quad (6)$$

$$\begin{aligned} \text{s.t. } & X = XZ + E, \text{diag}(Z) = 0, Z = Z^T, Z \geq 0, \\ & 0 < G < 1, \text{tr}(G) = k \end{aligned}$$

2.2 求解 BDRNG

本文将当前迭代次数记作 t , 那么 B^t 就代表第 t 次迭代时矩阵 B 的值。BDRNG 可以通过交替更新 G, Z, B 来求解, 获得最后的系数表示矩阵。

固定 $B = B^t$, 更新 Z^{t+1}, G^{t+1} :

$$\begin{aligned} \{G^{t+1}, Z^{t+1}\} = \arg \min_{G,Z} & \frac{1}{2} \|E\|_F^2 + \frac{\lambda}{2} \|Z - B\|_F^2 + \\ & \beta \text{tr}(ZL_M Z^T) + \gamma \langle \text{diag}(B1) - B, G \rangle \end{aligned} \quad (7)$$

$$\begin{aligned} \text{s.t. } & X = XZ + E, \text{diag}(Z) = 0, Z = Z^T, Z \geq 0, \\ & 0 < G < 1, \text{tr}(G) = k \end{aligned}$$

而式(7)可以简化为分别更新 Z^{t+1}, G^{t+1} 。对于 G^{t+1} :

$$G^{t+1} = \arg \min_G \langle \text{diag}(B1) - B, G \rangle \quad (8)$$

$$\text{s.t. } 0 < G < 1, \text{tr}(G) = k$$

其中, 式(8)通过 $G^{t+1} = UU^T$ 求解 G^{t+1} , 其中 $U \in \mathbb{R}^{n \times k}$ 是由 $\text{diag}(B1) - B$ 的 k 个最小的特征值对应的特征向量组成^[10]。

对于更新 Z^{t+1} 的子问题, 更新计算如式(9):

$$Z^{t+1} = \arg \min_Z \frac{1}{2} \|E\|_F^2 + \frac{\lambda}{2} \|Z - B\|_F^2 + \beta \text{tr}(ZL_M Z^T) \quad (9)$$

$$\text{s.t. } X = XZ + E, \text{diag}(Z) = 0, Z = Z^T$$

把式(9)记为二次函数 $f(Z)$, 令 $f(Z)$ 的导数为 0, 得到式(10):

$$(X^T X + \lambda I)Z + \beta ZL_M = X^T X + \lambda B \quad (10)$$

可以用西尔维斯特方程求得唯一解;

固定 $Z = Z^{t+1}, G = G^{t+1}$, 更新 B :

$$B^{t+1} = \arg \min_B \frac{\lambda}{2} \|Z - B\|_F^2 + \gamma \langle \text{diag}(B1) - B, G \rangle \quad (11)$$

根据参考文献[10]中的证明, 式(11)可以通过 $B^{t+1} = Z - \frac{\gamma}{\lambda} (\text{diag}(G)1^T - G)$ 求解 B 。

利用谱聚类方法^[15]将数据集划分成 k 个簇, 输出最后聚类的结果。

算法 1 BDRNG 算法。

输入 数据集矩阵 $X \in \mathbb{R}^{D \times n}$, 聚类簇数 k , 最大迭代次数 T_{\max} , 迭代终止阈值 tol 以及近邻图正则项 L_M ;

步骤 1 初始化 B^0, Z^0, G^0 。

步骤 2 更新 G 。

步骤 3 更新 Z 。

步骤 4 更新 B 。

步骤 5 计算 $\max(|Z^{t+1} - Z^t|, |B^{t+1} - B^t|)$ 判断是否小于或等于迭代终止阈值 tol ; 如果小于 tol , 则结束循环, 进入步骤 7; 否则进入步骤 6。

步骤 6 重复步骤 2~5, 达到最大迭代次数 T_{\max} 时, 结束循环, 进入下一步。

步骤 7 根据 Z, B 计算数据相似性矩阵 $Y, Y = \frac{\|Z\| + \|Z^T\|}{2}$ 或 $Y = \frac{\|B\| + \|B^T\|}{2}$ 。

输出 利用 Y , 通过谱聚类划分成 k 个簇, 输出最后聚类的标签划分。

2.3 时间复杂度分析

在BDRNG中, L_M 的计算以及和 Z 的求解的时间复杂度较高, 当数据集为 $X \in \mathbf{R}^{n \times n}$, 局部线性表示的权重矩阵 $W^* \in \mathbf{R}^{n \times n}$ 的计算时间复杂度为 $O(Dn^2 + (D+K)K^2n)$, 其中, K 为KNN算法所设定的邻域大小^[18]。根据 W^* , L_M 的计算时间复杂度为 $O(n^2)$ 。 $Z \in \mathbf{R}^{n \times n}$ 通过标准西尔维斯特方程进行求解, 时间复杂度为 $O(T_{\max}(n^2))$, T_{\max} 为程序最大迭代次数。在模型学习中, L_M 只需要一次计算, 不需要迭代更新, 所以 L_M 的计算时间复杂度不计入程序总时间复杂度。因此, 算法总时间复杂度为 $O(T(n^2))$, 其中 T 为BDRNG迭代次数。BDRNG与其他算法运行时间结果对比详见本文3.4节。

2.4 收敛性证明

在本文中, 本文通过交替最小化来更新模型。本文将式(3)标记为 $f(Z, B, G)$ 。并设, $S_1 = \{\text{diag}(B) = 0, B = B^T, B \geq 0\}$, S_1 的指示函数记为 l_{S_1} , 以及 $S_2 = \{0 < G < 1, \text{tr}(G) = k\}$, S_2 的指示函数记为 l_{S_2} 。

在更新 G^{t+1} 子问题中, 有:

$$f(Z^t, G^t, B^t) + l_{S_2}(G^{t+1}) \leq f(Z^t, G^t, B^t) + l_{S_2}(G^t)$$

其中, $G^t \in S_2$, 可得 $\|G^t\| \leq 1$, 所以 G^t 是收敛的。

在更新 Z^{t+1} 子问题中, 有:

$$f(Z^t, G^t, B^t) \leq f(Z^t, G^t, B^t) - \frac{\lambda}{2} \|Z^{t+1} - Z^t\|_F^2$$

这说明 Z^t 的更新也是收敛的。

在更新 B^{t+1} 子问题中, 有

$$f(Z^t, G^t, B^t) + l_{S_1}(B^{t+1}) \leq$$

$$f(Z^t, G^t, B^t) - \frac{\lambda}{2} \|B^{t+1} - B^t\|_F^2 + l_{S_1}(B^t)$$

综上, 本文将所有的子问题合并之后, 可以得到:

$$f(Z^{t+1}, G^{t+1}, B^{t+1}) + l_{S_1}(B^{t+1}) + l_{S_2}(G^{t+1}) \leq$$

$$f(Z^t, G^t, B^t) + l_{S_1}(B^t) + l_{S_2}(G^t)$$

$$- \frac{\lambda}{2} \|B^{t+1} - B^t\|_F^2 - \frac{\lambda}{2} \|Z^{t+1} - Z^t\|_F^2$$

因此, $f(Z^t, G^t, B^t) + l_{S_1}(B^{t+1}) + l_{S_2}(G^{t+1})$ 单调递减。同时, G^t 和 $\text{diag}(B^t) - B^t$ 均为半正定, 所以, 本文有 $\langle G^t, \text{diag}(B^t) - B^t \rangle \geq 0$ 。因此, 本文有 $f(Z^t, G^t, B^t) + l_{S_1}(B^{t+1}) + l_{S_2}(G^{t+1}) \geq 0$ 。

综上, 当 $t = 0, 1, 2, \dots, \infty$ 时, 本文有 $\sum_{t=0}^{+\infty} \frac{\lambda}{2} (\|B^{t+1} - B^t\|_F^2 + \|Z^{t+1} - Z^t\|_F^2) \leq f(Z^0, G^0, B^0)$ 。所以, 本文有:

$$G^{t+1} - G^t \rightarrow 0$$

$$Z^{t+1} - Z^t \rightarrow 0$$

$$B^{t+1} - B^t \rightarrow 0$$

综上所述, 本文的模型是收敛的。

3 实验结果与分析

为了验证BDRNG的有效性, 本文分别在仿真数据集和标准数据集上进行了实验。本文选用的测试对比算法包括: 本文提出的算法BDRNG、稀疏子空间聚类(Sparse Subspace Clustering, SSC)^[6]、低秩表示(Low-Rank Representation, LRR)模型^[7]、块对角表示(Block Diagonal Representation, BDR)模型^[10]、自适应低秩核子空间聚类(adaptive Low-Rank Kernel Subspace Clustering, LKSC)^[13]、隐式块对角线低秩表示(Implicit Block Diagonal Low-rank Representation, IBDLR)

模型^[14]6种算法。由于BDR中会生成2个可以作为最后聚类划分的系数表示矩阵BDR(Z)和BDR(B), 本文在具体对比结果评价中, 将分别提供BDR(Z)和BDR(B)最后聚类结果作为对比方法的评价结果之一。由于其中一个矩阵作为另外一个矩阵的求解的中间项, 例如BDR(B)作为BDR(Z)求解的中间项引入到模型中, 两者的迭代更新共享同一个更新过程。BDRNG同理。

本文采用聚类正确率(clustering Accuracy, ACC), 归一化互信息(Normalized Mutual Information, NMI)^[19]和调整兰德指数(Adjusted Rand Index, ARI)^[20]作为评价指标。

ACC是衡量聚类表现最通用及简单的方法, 可以作为评价整体聚类结束之后, 数据集中正确聚类的数据点个数; 但ACC无法说明更多关于聚类结果好坏的原因。NMI可以测量聚类结果和原本类的预标记标签之间的相互关系。如果聚类的结果和原来数据集标签非常相似, 则说明模型可以很好地对数据点进行划分。但是在聚类过程中, 得到的类别标签无法很好地说明聚类的准确度。聚类更多考察的是这些类别的分布和样本真实类别分布的相似度, 换言之, 同一个簇内, 样本的相似度比其他不同类别的样本相似度更好, 才能说明聚类的效果更好。兰德指数(Rand Index, RI)将聚类的过程视为一个决策的过程, 对数据集上所有的点进行两两配对, 当且仅当两个数据点相似时, 才将其归为一类。RI的取值范围在[0, 1]区间, 0代表聚类效果非常差。但当类数越来越多时, RI对于两个数据点随机地划分, 其RI值不为一个接近于0的常数, 即RI值无法很好地反映真实的聚类结果。所以本文采用的是ARI作为评价依据, ARI可以很好解决RI的随机划分问题, ARI取值范围在[-1, 1]区间, ARI越接近1, 说明模型聚类的效果更符合真实样本的分布^[20]。

ACC、NMI、ARI的值越大说明聚类的表现效果越好。所有实验结果为各个模型在数据集上面调优之后, 分别运行20次聚类的平均结果作为最后评价依据。

3.1 仿真数据集

本文在500维空间内随机选取5个不相交的子空间, 每个子空间的秩为75维, 每个子空间随机选取150个样本点。鲁棒性实验数据在干净数据集上随机生成10%、20%、30%的高斯噪声, 即从无噪声到含有30%噪声的数据集。

如果系数表示矩阵如果有很好的块对角属性, 那么子空间聚类算法将得到较好的聚类结果。图1中对比了BDRNG在仿真数据库上从干净数据集到30%噪声的表现效果。在图1中可以看到, 即使噪声等级上升, BDRNG所生成的系数表示矩阵依然保持不错的块对角形式。

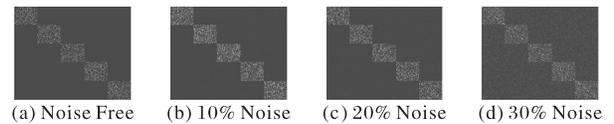


图1 BDRNG在不同噪声比例下生成的系数表示矩阵可视化图

Fig. 1 Visualization of coefficient representation matrices generated by BDRNG with different noise ratios

BDRNG和LRR、SSC、IBDLR、BDR四种方法在仿真数据集上的实验结果如表1所示。从表1中可以看出: BDRNG具有较好的聚类性能。图2为在20%噪声下各方法的系数表示矩阵可视化图。可以看出BDRNG在20%噪声的仿真数据库上相较于其他方法有更好的块对角形式的表现。在20%噪声下, LRR的系数矩阵块对角形式被严重破坏, SSC的系数矩阵块对角形状隐约可见, BDR的系数矩阵块对角边界模糊而

IBDLR 的系数矩阵比上述 3 种算法较好,然而依旧存在着块对角形式外模糊的现象。由此可见,在 20% 噪声下,对比方法都存在被噪声不同程度干扰的现象,而 BDRNG 可以使得系数表示矩阵保持较好的块对角形式。BDRNG 在 BDR 的基础上增加非线性学习正则算子。当处理噪声数据时,BDRNG 比 BDR 系数矩阵获取的块对角结构更具有鲁棒性。这说明了引入的正则算子可以有效地利用非线性结构数据,有助于模型对于数据内在的结构学习和提升了模型对于数据集的表达能力。

随着噪声比例变高,聚类问题就会变得更加困难。为了更好地观察、分析各个算法在不同噪声程度下的影响,本文对比 LRR、SSC、BDR、IBDLR 和 BDRNG 在 0~20% 噪声条件下的聚类表现。随机初始化 50 组参数进行实验。所有参数均在 $\{1E-5, 1E+5\}$ 内变化。在无噪声的情况下,除了 LRR 以外,其他所有方法最大聚类准确度均可以保持 100% 完美的

聚类表现。随着噪声等级上升的过程中,其他方法的平均聚类准确度快速下降,而 BDRNG 的下降幅度相对较为平缓。在 10% 噪声条件下,LRR、SSC、IBDLR 的平均聚类准确度快速下降到 50% 及以下。由表 1 可以看出,在噪声比例上升的过程中,BDRNG 的聚类准确度的中位数(Med)依旧可以保持在 85% 以上,而其他方法的 50 次聚类准确度的中位数已经快速下降。由此可以看出,本文提出的方法在引入非线性结构学习算子后,可以增强线性的子空间聚类算法的鲁棒性。

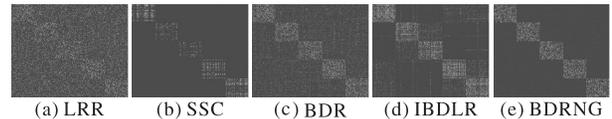


图 2 各算法在 20% 噪声条件下的仿真数据集下生成的系数表示矩阵可视化图

Fig. 2 Visualization of coefficient representation matrices generated by different algorithms with 20% noise

表 1 各算法在不同噪声条件的仿真数据集上的聚类准确度

单位:%

Tab. 1 ACCs of different algorithms on synthetic dataset with different noise ratios

unit:%

Method	Noise-Free			10% Noise			20% Noise		
	Max	Med	Mean	Max	Med	Mean	Max	Med	Mean
LRR	98.77	91.06	64.79	95.70	22.47	26.47	90.07	20.92	22.41
SSC	<u>100(1)</u>	96.13	76.12	96.53	29.53	30.14	93.13	24.12	25.23
BDR(Z)	<u>100(22)</u>	99.74	<u>98.12</u>	<u>100(12)</u>	98.40	73.28	<u>100(3)</u>	33.34	47.90
BDR(B)	<u>100(20)</u>	99.74	85.73	<u>100(7)</u>	97.67	71.49	<u>100(1)</u>	29.47	48.62
IBDLR	<u>100(22)</u>	79.79	62.04	<u>100(16)</u>	79.79	57.29	<u>100(4)</u>	30.04	50.90
BDRNG(Z)	<u>100(26)</u>	<u>100</u>	95.80	<u>100(19)</u>	<u>99.13</u>	<u>89.82</u>	<u>100(2)</u>	<u>90.27</u>	<u>71.17</u>
BDRNG(B)	<u>100(20)</u>	99.74	91.79	<u>100(14)</u>	98.54	84.54	<u>100(2)</u>	87.13	69.16

注:*括号中的数字为聚类为 100% 出现的次数;Max 表示最大值,Med 表示中值,Mean 表示均值。

3.2 标准数据集

本节选用标准数据集:ORL^[21]、COIL-20^[22]和 CIFAR-10^[23],通过这 3 个标准数据集对比分析各个算法的聚类性能。ORL 数据集包含 40 个人在不同光照条件、面部细节和面部表情下 10 幅人脸图像。COIL20 数据集是由 20 个物体组成的总共 1 440 幅灰色图像。每个物体以 5° 为拍摄固定间隔角度进行拍摄,总共拍摄 72 幅图片。CIFAR-10 数据集包含 10 个类别:飞机、汽车、鸟类、猫、鹿、狗、青蛙、马、船和卡车。每个类别包含 6 000 幅图像。为了降低内存和计算成本,在每个类别中随机选择 100 个样本,通过使用 ILSVRC (ImageNet Large Scale Visual Recognition Challenge)^[24]上的卷积神经网络作为特征提取器,将每个图像表示为 2 048 维向量,因此,本文实验中使用的 CIFAR-10 数据集大小为 2 048×1 000。部分数据集示例样本如图 3 所示。

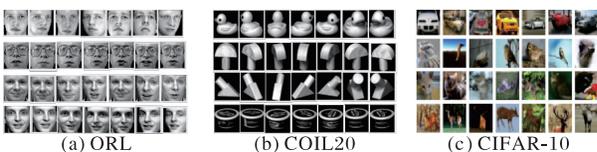


图 3 部分数据集示例样本

Fig.3 Partial samples of datasets

在本节实验中,选用了 LRR、SSC、BDR (包括 BDR(Z) 和 BDR(B))、IBDLR、LRKSC 作为实验的对比方法。各个对比方法先通过 $\{1E-5, 1E+5\}$ 的区间内进行调整参数实验,具体实验设置可参考 3.3 节;其次,再进行随机参数实验,即随机进行了 2 000 次的随机参数设置,进行实验;最后,选择了聚类

效果最佳的实验结果作为最后对比评价依据。

从表 2 可以得到如下结论:

1) BDR、IBDLR 和 BDRNG 为块对角表示模型为基础或是引入块对角约束正则项。实验结果表明,在这 3 个数据集上,BDR、IBDLR、BDRNG 均有明显优于 LRR 和 SSC。这说明了块对角约束确实有效提高了聚类效果。此外,BDRNG 又在 BDR 的基础上,强化了数据局部几何结构的学习,补充了局部几何结构信息。相比起 BDR,本文提出的 BDRNG,包括 BDRNG(Z) 和 BDRNG(B),均取得了更好的聚类效果。这说明学习局部几何结构信息之后的子空间聚类算法对于最后聚类效果确实有所助益。

表 2 在各标准数据集上各方法聚类表现比较

单位:%

Tab. 2 Comparison of clustering performance of different methods on different datasets

unit:%

方法	ORL			COIL 20			CIFAR-10		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
LRR	56.84	75.87	30.07	68.60	82.51	63.51	74.89	59.97	55.31
SSC	60.27	78.77	41.17	60.09	71.34	46.53	75.30	61.80	57.42
BDR(Z)	75.95	86.95	65.31	81.38	92.39	80.71	79.66	66.92	62.88
BDR(B)	76.20	87.07	66.36	81.38	92.63	80.77	79.04	65.96	61.07
IBDLR	75.41	86.23	64.91	78.37	88.02	76.60	79.11	66.33	61.59
LRKSC	80.25	87.66	70.78	83.63	91.05	81.68	74.60	63.27	54.63
BDRNG(Z)	82.25	89.95	72.72	83.50	92.57	83.05	<u>81.16</u>	<u>67.98</u>	<u>66.42</u>
BDRNG(B)	<u>84.00</u>	<u>90.59</u>	<u>76.26</u>	<u>84.61</u>	<u>93.57</u>	<u>83.57</u>	<u>80.21</u>	67.02	64.27

2) ORL 为人脸数据集,人脸数据有强非线性结构。在 ORL 上,BDRNG 的表现明显优于其他算法。这是因为 BDRNG 通过近邻图学习了数据集的非线性结构。通过

BDRNG 和 BDR 的比较,可以看出 BDRNG 有效地提升了非线性结构信息的学习效果。除 BDRNG 以外,ORL 数据集上表现最佳为 LRKSC 算法。该方法中引入了核方法的学习,同样有效提高了这类非线性结构信息的学习效果。然而,通过后续运行时间的分析,发现 LRKSC 的耗时非常突出,不适用于大规模的高维数据的学习。BDRNG 通过近邻图得到数据局部非线性结构信息,优于核方法聚类性能,且时间性能更优。

3.3 参数分析

在本节里,以 ORL 数据库为例,分析各个参数对于模型的影响程度,实验结果如图 4 所示。本文提出的模型 BDRNG 总共有 3 个平衡正则项: λ 、 γ 和 β 。将这 3 个参数分别分组,在 $\{1E-5, 1E+5\}$ 的取值范围内变化其中两个参数,保持另外一个参数不变(当某一个参数固定时,分别固定变量值设置为 $\lambda = 0.1, \gamma = 1, \beta = 1$)。

观察图 4,可以发现不同的参数对于 BDRNG 学习影响各有强弱,其中, λ 和 β 对聚类效果影响更大。首先,当 λ 取值过小会导致模型学习效果不佳。在模型迭代更新的初期, B 、 Z 还没有充分学习。当 λ 取值过小,该正则项对于迭代更新的影响变小,无法得到一个好的学习效果^[10]。其次,当 β 取值较大时,聚类结果下降明显,这是因为非线性结构学习影响过大,影响到了原有块对角表示模型的学习,导致最后聚类效果下降明显。当 β 非常小的时候, BDRNG 获得的数据局部几何结构信息非常少,对于模型学习贡献较少。当 λ 、 β 均比较小的时候, BDRNG 表现非常差,这是因为块对角表示和数据内在非线性结构信息均没有有效学到数据信息,导致模型表现效果差。最后,通过结合各组参数影响图发现,聚类效果变化趋势由 λ 或者是 β 主导。相比起 λ 的变化对于模型学习效果的影响相对较弱, γ 的变化没有引起模型聚类效果的明显变化。从直观上来理解, γ 对于模型的影响只有在更新矩阵 B 的过程中,同时, γ 的作用还受到了 λ 的制约。所以,相比起 λ 和 β , γ 对于模型最后效果的影响不是非常的明显。

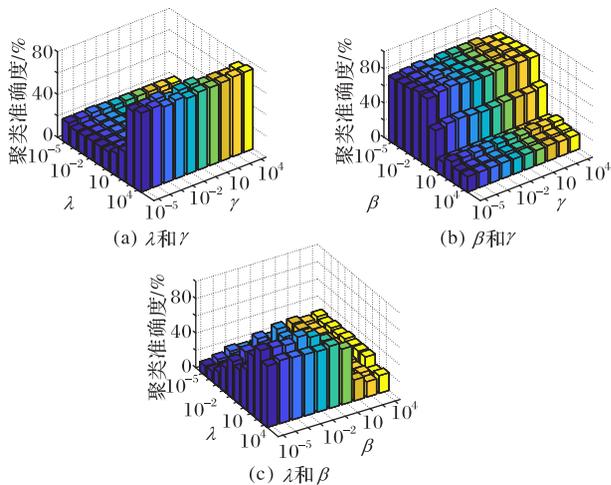


图 4 各参数在 ORL 数据集上对聚类准确度的影响
Fig. 4 Impact of parameters on ACC on ORL dataset

此外,在本文中,除了 λ 、 γ 和 β 以外,还有一个隐藏的参,即在构造近邻图时使用到的近邻数 K 。在多次实验中,发现构造近邻图时所选择的近邻数 K 有一定的影响。 K 越小,则所构造的近邻图会越集中于局部数据,近邻点对于中心的影响也会越大,其中,受到噪声的干扰影响也就越大;反之, K

越大,近邻会包括更多的数据点,在其中数据点越多,包含噪声点的可能性也就越大,导致中心点的重改易受到更多噪声点的干扰。以 ORL 数据集为例,从图 5 可以看出,随着近邻数 K 的变大,聚类效果呈先升后降的趋势,近邻数 K 在样本数的 1/4 处时,可以获得较好的聚类结果。

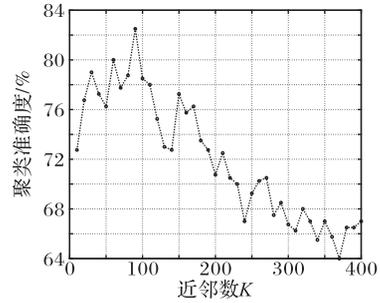


图 5 在 ORL 数据集上构造近邻图的近邻数 K 对聚类结果的影响
Fig. 5 Impact of the number of neighbors K on clustering performance on ORL dataset

3.4 时间复杂度和收敛性分析

本节以 ORL 数据库为例,对比分析 BDRNG 和其他方法的计算耗时和以及可视化 BDRNG 模型迭代更新过程的收敛曲线。由于 BDRNG(Z)、BDRNG(B)是共享同一个更新过程,本实验仅记录模型求解的总耗时,不单独计算 BDRNG(Z)或 BDRNG(B)耗时。BDR 同理。本节实验随机设置各个对比方法的参数,进行 100 次实验,最后将实验结果平均耗时作为该对比方法的评价依据(参数变化范围在 $\{1E-5, 1E+5\}$ 之中)。

根据图 6(a)可以看出, LRR(ACC-56.84%)的计算耗时最少,但是聚类准确度是所有对比方法中最低。IBDLR(ACC-75.41%)和 LRKSC(ACC-80.25%)是耗时最长的两个方法。这是因为这些方法中还涉及到了核方法的运算,耗时相对于其他方法长。然而,高的时间开销并未换来更优的聚类效果。本文提出的模型 BDRNG(ACC-84.00%)耗时相对多,但是在 ORL 数据集上, BDRNG 提升效果最明显。这说明,相比起核方法, BDRNG 通过更少的时间代价学习数据非线性结构信息,显著提升了聚类准确度。

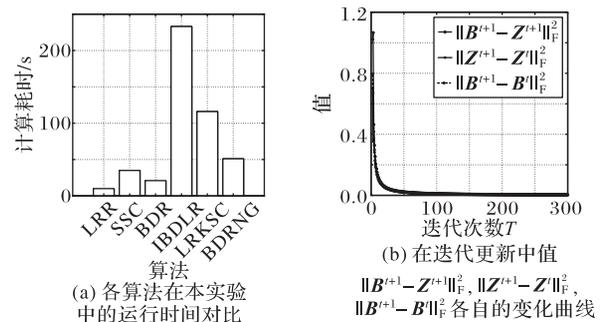


图 6 ORL 实验结果
Fig. 6 Results on ORL dataset

根据本文 2.3 节中的内容,算法总时间复杂度为 $O(T(n^2))$,其中 T 为 BDRNG 迭代次数, n 为数据样本个数。图 6(b)为 BDRNG 迭代更新中,各项值变化趋势,可以看出 BDRNG 的目标函数值呈单调递减,在有限的循环次数内可以快速达到收敛阈值。综合考虑耗时和聚类表现, BDRNG 是性

价比最高的聚类模型。

4 结语

针对高维流形数据聚类问题,本文提出了基于近邻图的块对角子空间聚类算法——BDRNG。该方法不仅考虑原始数据的局部几何结构,并且结合了数据的全局结构信息。BDRNG通过近邻图拟合高维数据原始空间中局部几何结构;同时,通过块状稀疏范数优化生成稳定结构的块对角系数表示矩阵,描述全局结构信息。实验表明,本文通过学习高维数据中的局部结构,有效获取高维数据中的流形结构信息,提升线性子空间表示模型处理高维非线性结构数据的能力。未来的工作将更进一步研究如何快速和有效获取高维数据流形信息。

参考文献 (References)

- [1] HONG W, WRIGHT J, HUANG K, et al. Multiscale hybrid linear models for lossy image representation [J]. IEEE Transactions on Image Processing, 2006, 15(12): 3655-3671.
- [2] D' HAESLEER P. How does gene expression clustering work? [J]. Nature Biotechnology, 2005, 23(12): 1499-1501.
- [3] COSTEIRA J P, KANADE T. A multibody factorization method for independently moving objects [J]. International Journal of Computer Vision, 1998, 29(3): 159-179.
- [4] 鲁全茂. 面向高维数据的聚类算法研究[D]. 北京:中国科学院大学, 2018: 1-4, 57-75. (LU Q M. Research on clustering algorithms for high-dimensional data [D]. Beijing: University of Chinese Academy of Sciences, 2018: 1-4, 57-75.)
- [5] 王卫卫, 李小平, 冯象初, 等. 稀疏子空间聚类综述[J]. 自动化学报, 2015, 41(8): 1373-1384. (WANG W W, LI X P, FENG X C, et al. A survey on sparse subspace clustering [J]. Acta Automatica Sinica, 2015, 41(8): 1373-1384.)
- [6] ELHAMIFAR E, VIDAL R. Sparse subspace clustering: algorithm, theory, and applications [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(11): 2765-2781.
- [7] LIU G, LIN Z, YU Y. Robust subspace segmentation by low-rank representation [C]// Proceedings of the 27th International Conference on Machine Learning. Madison, WI: Omnipress, 2010: 663-670.
- [8] FENG J, LIN Z, XU H, et al. Robust subspace segmentation with block-diagonal prior [C]// Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, IEEE 2014: 3818-3825.
- [9] LU C, MIN H, ZHAO Z, et al. Robust and efficient subspace segmentation via least squares regression [C]// Proceedings of the 2012 European Conference on Computer Vision, LNCS 7578. Berlin: Springer, 2012: 347-360.
- [10] LU C, FENG J, LIN Z, et al. Subspace clustering by block diagonal representation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41(2): 487-501.
- [11] PATEL V M, VIDAL R. Kernel sparse subspace clustering [C]// Proceedings of the 2014 IEEE International Conference on Image Processing. Piscataway: IEEE, 2014: 2849-2853.
- [12] XIAO S, TAN M, XU D, et al. Robust kernel low-rank representation [J]. IEEE Transactions on Neural Networks and Learning Systems, 2016, 27(11): 2268-2281.
- [13] JI P, REID I, GARG R, et al. Adaptive low-rank kernel subspace clustering [EB/OL]. [2019-11-20]. <https://arxiv.org/pdf/1707.04974.pdf>.
- [14] XIE X, GUO X, LIU G, et al. Implicit block diagonal low-rank representation [J]. IEEE Transactions on Image Processing, 2018, 27(1): 477-489.
- [15] NG A Y, JORDAN M I, WEISS Y. On spectral clustering: analysis and an algorithm [C]// Proceedings of the 14th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2001: 849-856.
- [16] ROWEIS S T, SAUL L K. Nonlinear dimensionality reduction by locally linear embedding [J]. Science, 2000, 290(5500): 2323-2326.
- [17] 叶东升. 多流形嵌入子空间聚类方法研究[D]. 哈尔滨: 哈尔滨工程大学, 2019: 6-10, 12-17 (YE D S. Research on multi-manifold embedded subspace clustering method [D]. Harbin: Harbin Engineering University, 2019: 6-10, 12-17.)
- [18] YANG Y, HU Y, WU F. Sparse and low-rank subspace data clustering with manifold regularization learned by local linear embedding [J]. Applied Sciences, 2018, 8(11): No. 2175.
- [19] MANNING C D, RAGHAVAN P, SCHÜTZE H. Introduction to Information Retrieval [M]. Cambridge: Cambridge University Press, 2008: 356-360
- [20] YEUNG K Y, RUZZO W L. Details of the adjusted rand index and clustering algorithms, supplement to the paper an empirical study on principal component analysis for clustering gene expression data [J]. Bioinformatics, 2001, 17(9): 763-774.
- [21] SAMARIA F S, HARTER A C. Parameterisation of a stochastic model for human face identification [C]// Proceedings of 1994 IEEE Workshop on Applications of Computer Vision. Piscataway: IEEE, 1994: 138-142.
- [22] NENE S A, NAYAR S K, MURASE H. Columbia Object Image Library (COIL-20) [R]. New York: Columbia University, 1996.
- [23] KRIZHEVSHY A. Learning multiple layers of features from tiny images [R]. Toronto: University of Toronto, 2009.
- [24] RUSSAKOVSKY O, DENG J, SU H, et al. ImageNet large scale visual recognition challenge [J]. International Journal of Computer Vision, 2015, 115(3): 211-252.

This work is partially supported by the National Natural Science Foundation of China (61502108, 61876042, 61876043), the NSFC-Guangdong Joint Found (U1501254).

WANG Lijuan, born in 1978, Ph. D., associate professor. Her research interests include data mining, machine learning.

CHEN Shaomin, born in 1994, M. S. candidate. Her research interests include data mining, subspace clustering.

YIN Ming, born in 1975, Ph. D., associate professor. His research interests include machine learning, pattern recognition, image processing.

XU Yueying, born in 1984, M. S., lecturer. His research interests include Internet application.

HAO Zhifeng, born in 1968, Ph. D., professor. His research interests include machine learning, artificial intelligence.

CAI Ruichu, born in 1983, Ph. D., professor. His research interests include machine learning, data mining.

WEN Wen, born in 1981, Ph. D., associate professor. Her research interests include support vector machine, pattern recognition.