文章编号: 1001-0920(2017)07-1235-06 DOI: 10.13195/j.kzyjc.2016.0629

# 基于光滑近邻表示的基因表达数据子空间聚类

陈晓云[1], 林莉媛[1], 叶先宝[2†]

(1. 福州大学 数学与计算机科学学院，福州 350116；2. 福州大学 经济与管理学院，福州 350116)

**摘 要:** 基因表达数据具有样本数少、基因维数高、非线性等特点,为能有效地处理基因表达数据,提出光滑近邻表示子空间聚类算法. 利用每个数据点的近邻线性表示刻画数据集的非线性特点,并对近邻表示添加光滑约束,使数据点与近邻的距离关系嵌入到该数据点的重构表示中. 在基因表达数据上的实验表明,所提出的方法优于其他几个现有方法,进而表明所提出方法对基因表达数据的聚类是有效的.

**关键词:** 基因表达数据；子空间聚类；光滑表示；近邻

**中图分类号:** TP311；TP371 **文献标志码:** A

# Gene expression data subspace clustering based on smooth neighbor representation

*CHEN Xiao-yun*[1], *LIN Li-yuan*[1], *YE Xian-bao*[2†]

(1. College of Mathematics and Computer Science，Fuzhou University，Fuzhou 350116，China；2. School of Economics and Management，Fuzhou University，Fuzhou 350116，China)

**Abstract:** Gene expression data has the characteristics of small sample size, high dimension, nonlinear and so on. In order to effectively deal with the gene expression data，a subspace clustering method is proposed via smooth neighbour representation(SNR). The neighborhood linear representation of data points is used to describe the nonlinear properties of data, and the smooth constraint is added on the representation which makes the relationship of distance between data point and its neighbors embed in the reconstruction representation. Experiment results on gene expression data show that the performance of SNR is superior to several existing methods, and SNR can cluster gene expression data effectively.

**Keywords:** gene expression data；subspace clustering；smooth representation；neighbor

## 0 引 言

随着DNA微阵技术的发展,人们可获得大量的基因表达数据. 面对海量且复杂的基因表达数据,如何有效识别数据中的有用信息具有重要现实意义,如肿瘤基因表达数据的分析有助于对肿瘤的发生、发展机制进行探索,有助于更好地划分肿瘤的类型及亚型,从而更有效地预防和治疗肿瘤. 聚类是分析基因表达数据的一种重要方法[1-3]. 基因表达数据的聚类可分为3类,分别是样本聚类、基因聚类和双向聚类[4],本文对样本聚类进行研究.

用于基因聚类的传统方法有层级聚类算法(HC)、$K$-means和自组织映射(SOM). 此外,基于非负矩阵分解的方法[5-7]也成功应用于基因表达数据,但只有少数子空间聚类方法应用于基因表达数据,如潜在最小二乘回归子空间分割方法[8](LatLSR).

子空间聚类也称子空间分割[9],已在图像表示和分割、计算机视觉和疾病检测等领域得到较好应用. 这些领域中的数据具有高维特征,易出现“维数灾难”问题. 实际上,高维数据在其环绕空间中往往具有低维结构[10],子空间聚类的目标是揭示环绕空间中数据的低维结构,并将数据点准确地分割到各自所属的子空间中. 现有的子空间聚类方法包括迭代方法、代数方法、统计方法和谱聚类方法[11],其中代表性方法有稀疏子空间聚类(SSC)[12]、低秩表示(LRR)[13]子空间分割、最小二乘回归(LR)子空间分割[14]等. 之后又产生一系列扩展方法,如潜在低秩表示子空间分割[15](LatLRR)、鲁棒潜在低秩表示子空间聚类[16](RLLRR)等. 这些子空间聚类方法假设样本数据具有全局线性特点且通过线性自表示方法重构,一定程度上会影响非线性数据的聚类效果.

**收稿日期:** 2016-05-20；**修回日期:** 2016-10-28.

**基金项目:** 国家自然科学基金项目(71273053, 11571074)；福建省自然科学基金项目(2014J01009).

**作者简介:** 陈晓云(1970−), 女, 教授,博士, 从事机器学习等研究；林莉媛(1991−), 女, 硕士生, 从事智能信息处理的研究.

†通讯作者. E-mail: Yexb5626@163.com

为有效处理高维非线性的基因表达数据,本文提出一种新的子空间聚类方法. 新方法中每个数据点仅用其局部近邻点线性重构,以此刻画数据的局部线性特点,并引入光滑约束项对表示系数进行约束(子空间聚类文献中称数据的重构系数为“表示[13]”,本文也沿用该称呼). 利用获得的表示构造仿射矩阵,并通过标准切割方法分割仿射矩阵得到聚类结果.

## 1 子空间聚类

**定义1** (子空间聚类)[14] 从$k$个维数未知的线性子空间采样一组数据点$X = [x_1, x_2, \cdots, x_n] \in R^{d\times n}$. 其中:$n$为数据点个数,$d$为每个数据点的维数. 子空间聚类的目标是将数据点分割到对应子空间,每个类对应一个子空间.

子空间聚类现有的算法有迭代方法、代数方法、统计方法和基于谱聚类的方法[14]. 其中基于谱聚类的方法在许多领域中都表现出不错的性能.

基于谱聚类的方法将子空间聚类问题分成两个步骤. 第1步构造仿射矩阵$W = [w_{ij}]$,其中$w_{ij} = w_{ji} \geqslant 0$,$w_{ij}$测量数据点$x_i$和$x_j$是否属于同一个子空间,理想情况下,如果属于同一个子空间,则$w_{ij} \approx 1$,否则$w_{ij} = 0$. 得到的仿射矩阵便是谱聚类方法的输入,即相似度矩阵. 第2步对仿射矩阵执行谱聚类方法. 整个过程中构造好的仿射矩阵是关键.

基于谱聚类的子空间聚类的核心是寻找数据较好的自表示,并以自表示系数构造仿射矩阵,然后对仿射矩阵用谱聚类算法如标准切割方法[17]实现分割. 在该框架下诞生了很多方法,这类方法将每个数据点表示为数据集中其余数据点的线性组合[9]

$$X = XZ,\ \mathrm{diag}(Z) = 0,$$

其中$Z \in R^{n\times n}$为表示矩阵,$Z$随约束条件不同具有不同性质. 理想条件下,当$x_i$与$x_j$不同类时,$z_{ij} = 0$,$Z$具有块对角结构. 各种方法的主要不同之处在于对表示矩阵的约束不同,如稀疏约束、低秩约束和最小二乘约束等. 有些情况下,当移除条件$\mathrm{diag}(Z) = 0$不会造成数据点仅用自身表示自身的平凡解时,可使用数据集的所有点线性表示每个数据点,称为数据的自表示[9],即

$$X = XZ.$$

有噪声时,可表示为

$$X = XZ + E,$$

其中$E \in R^{d\times n}$为噪声项. 得到表示矩阵$Z$后,通过$(Z + Z^{\mathrm{T}})/2$构造仿射矩阵. 基于该框架,文献[12]提出稀疏子空间聚类算法SSC,其目标是寻找最稀疏的表示. 文献[13]提出低秩表示子空间分割,与SSC不同的是,该算法的目标是寻找最低秩表示. 但最小化秩是NP难问题,常用核范数[13]代替秩约束. 以下主要回顾文献[14]提出的最小二乘回归子空间聚类算法(LSR).

考虑到数据的相关性,文献[14]提出利用最小二乘回归方法学习表示矩阵$Z$,有

$$\min_Z ||Z||_F;\quad \mathrm{s.t.}\ X = XZ + E, \mathrm{diag}(Z) = 0. \tag{1}$$

其中$||Z||_F = \Big(\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n} z_{ij}^2\Big)^{\frac{1}{2}}$为$Z$的Frobenius-范数.

当数据包含噪声时,有

$$\min_Z ||X - XZ||_F^2 + \lambda||Z||_F^2;\quad \mathrm{s.t.}\ \mathrm{diag}(Z) = 0, \tag{2}$$

其中参数$\lambda > 0$用于平衡目标函数中两项的影响. 因为去掉条件$\mathrm{diag}(Z) = 0$不会导致平凡解[14],所以可将式(2)改写成另一种描述LSR的方式

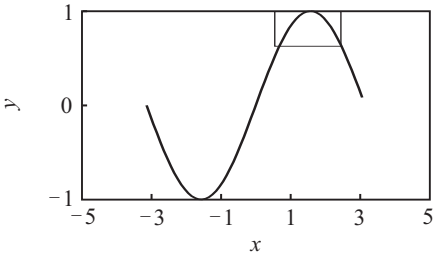
$$\min_Z ||X - XZ||_F^2 + \lambda||Z||_F^2. \tag{3}$$

文献[14]已给出了式(3)的解析解,且证明该方法具有聚集性,可以使同一类点的表示靠近.

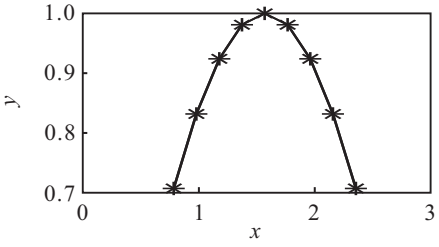
## 2 光滑近邻表示子空间聚类

### 2.1 问题提出与解决思想

基于谱聚类的子空间聚类框架下的方法采用自表示方式重构数据,默认数据具有线性关系,因此不能很好地处理非线性数据. 事实上,非线性数据在局部往往具有近似线性关系. 如图1所示,图1(a)为非线性分布的数据,将非线性曲线数据的部分(图1(a)




(a) 非线性曲线




(b) 局部近似线性

**图1 非线性图及其局部图**

的黑色矩形框部分)放大,并分成多个线段,如图1(b),每一小段的数据可以近似看成线性关系. 因此,为能有效地对非线性分布的数据聚类,本文利用每个数据点的$c$个最近邻线性表示该点.

此外,因为每个近邻与该点的距离远近不一,为使得到的表示系数能够反应其各近邻的距离关系,在模型中引入二次光滑函数约束表示系数.

### 2.2　光滑函数

二次光滑函数[18]定义为

$$\phi_{\mathrm{quad}}(y)=\sum_{i=1}^{n-1}(y_{i+1}-y_i)^2=||Sy||_2^2. \tag{4}$$

其中: $y\in R^{N\times 1}$, $S\in R^{(N-1)\times N}$是一个双对角矩阵

$$S=\begin{bmatrix}-1 & 1 & 0 & \cdots & 0 & 0 & 0\\ 0 & -1 & 1 & \cdots & 0 & 0 & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots\\ 0 & 0 & 0 & \cdots & -1 & 1 & 0\\ 0 & 0 & 0 & \cdots & 0 & -1 & 1\end{bmatrix},$$

$S$的每个元素为$s_{ij}$,有

$$s_{ij}=\begin{cases}-1,\ i=j;\\ 1,\ i=j+1;\\ 0,\ \text{otherwise.}\end{cases}$$

$||Sy||_2^2$是$y$的变化度量或光滑度[18].

### 2.3　光滑近邻表示子空间聚类(SNR)

本节给出光滑近邻表示子空间聚类(SNR)算法,其数据点由其近邻重构,并引入光滑正则项,使自表示系数可以反映原始空间中数据点与其近邻的距离关系. 光滑近邻表示子空间聚类的目标函数为

$$\min_{\widetilde{z}_i}\sum_{i=1}^{n}\Big\|x_i-\sum_{x_i\in N_c(x_i)}x_j\widetilde{z}_{ij}\Big\|_F^2+\sum_{i=1}^{n}\lambda||S_i\widetilde{z}_i||_F^2. \tag{5}$$

其中: $N_c(x_i)$为$x_i$的$c$近邻集; $\widetilde{z}_i$为$x_i$的近邻表示向量,且$\widetilde{z}_i$中的元素按照对应近邻到$x_i$的距离由近到远排列, $\widetilde{z}_{ij}$为$x_i$来自$x_j$的表示; $||\cdot||_F$为Frobenius-范数,参数$\lambda>0$. 第1项表示每个数据点可由其$c$个近邻线性重构,第2项使表示系数在子空间中依数据点与其近邻的距离由近到远变化光滑,从而使数据点距离接近的两个近邻所对应的表示也接近. 由光滑函数的定义$||S_i\widetilde{z}_i||_F^2=\sum_{m=1}^{c}(\widetilde{z}_{i,m+1}-\widetilde{z}_{i,m})^2$可看出,当数据点使用包含自身的近邻重构该点时,不会出现自身表示自身这样的平凡解,因此模型中数据点的近邻可以包含该数据点本身.

目标函数(5)对Frobenius-范数的平方和求最小,可等价表示为

$$\min_{\widetilde{z}_i}||x_i-N_i\widetilde{z}_i||_f^2+\lambda||S_i\widetilde{z}_i||_F^2. \tag{6}$$

以$x_i$的$c$个最近邻为列向量组成$x_i$的近邻矩阵$N_i$,注意近邻点在矩阵中按距离升序排列, $N_i$的列向量按其所代表近邻点与数据点距离由小到大排列. $\widetilde{z}_i$是$x_i$的近邻表示向量,矩阵$S_i\in R^{(c-1)\times c}$是双对角矩阵. 该问题是一个凸问题,有全局最优解. 令式(6)的一阶导为0,得到$x_i$的近邻表示的解析解为

$$\widetilde{z}_i^*=\Big(N_i^{\mathrm{T}}N_i+\frac{S_i^{\mathrm{T}}S_i}{\lambda}\Big)^{-1}(N_i^{\mathrm{T}}x_i). \tag{7}$$

对于每个样本点$x_i$,令$z_{ii_m}^*=\widetilde{z}_{ii_m}^*, m=1,2,\cdots,c$,其余为0. $z_{ii_m}^*$为表示矩阵$Z^*$的元素,所有的表示构成表示矩阵$Z^*$. 构造仿射矩阵$(|Z^*|+|(Z^*)^{\mathrm{T}}|)/2$,利用标准切割方法Ncut[17]切割仿射矩阵得到最终聚类结果. SNR算法描述如下.

光滑近邻表示子空间聚类(SNR).

输入: 数据$X$,类数$k$,近邻数$c$,正则化参数$\lambda$;

输出: 最终聚类结果.

Step 1: 为每个数据点选择$c$个近邻并构造近邻矩阵$N_i$;

Step 2: 通过(7)计算$n$个数据点的近邻表示,构造表示矩阵$Z^*$;

Step 3: 通过$(|Z^*|+|(Z^*)^{\mathrm{T}}|)/2$构造仿射矩阵;

Step 4: 用Ncut[17]切割仿射矩阵.

选择数据点近邻时,需先测量该点与数据集中所有点的距离,再取最近的$c$个点作为近邻. 计算距离的方法很多,本文选用欧氏距离测量距离.

## 3　实验分析

### 3.1　实验准备

所有实验均在Win7系统、内存4 GB、双核CPU的计算机上执行,实验代码用Matlab R2013a编写. 将聚类得到的类标签与数据集的类标签作比较,利用聚类准确率(Accuracy,ACC)[19]评估聚类性能.

实验选用的对比方法有: 1) 传统聚类方法: $K$-means、HC; 2) 基于非负矩阵分解的方法: 凸非负矩阵分解CNMF[6]和半非负矩阵分解SNMF[6]; 3) 子空间聚类方法: LRR[13]、LSR[14]、RLLRR[16].

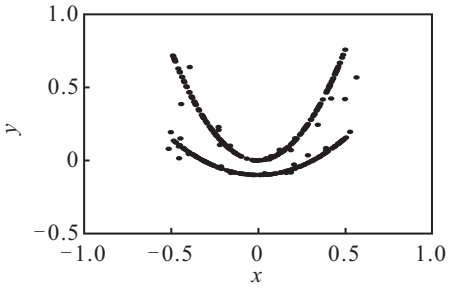
4种子空间聚类算法LSR、LRR、RLLRR和SNR的参数$\lambda$设置为$\{0.000\,1, 0.005, 0.05, 0.3, 0.5, 0.7, 1, 5, 8, 19, 50, 100\}$, SNR的近邻数$c$设置为$\{2,3,4,5,6,7,8,9,10,11,12,13,15\}$. 算法遍历所有参数取值,并选择最高的ACC作为最终结果.

### 3.2 小 实 验

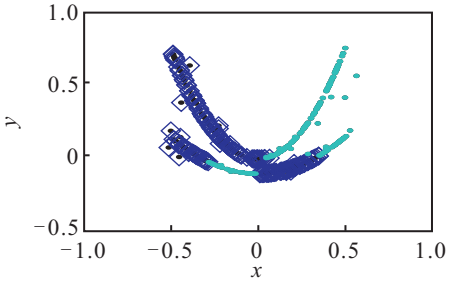
首先用人工数据验证SNR对非线性数据的聚类性能.

1) 两条不相交的二次曲线.

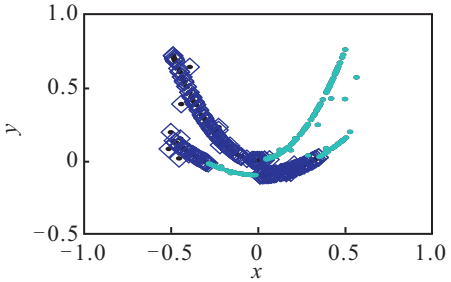
图2(a)由两条不相交非线性数据构成,图2(b)是LSR聚类后结果图,图2(c)是LRR聚类后结果图,图2(d)是SNR聚类后结果图. 由图2可见,SNR成功地将两条不相交曲线分成两类.



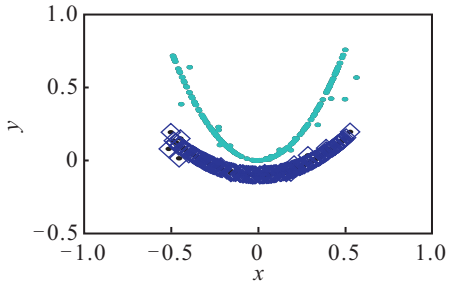

(a) original




(b) LSR




(c) LRR

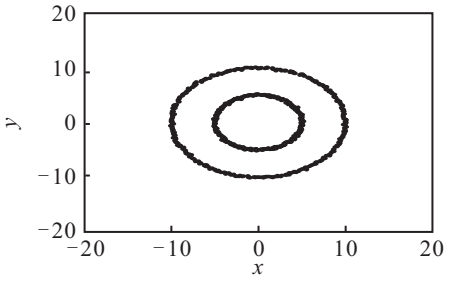



(d) SNR

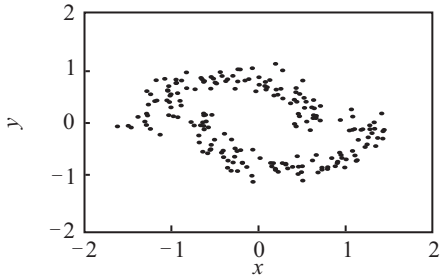
**图 2 曲线原图及聚类后的图(小实验)**

2) 双圆环和双月数据集.

在双圆环和双月型两个人工非线性数据上进行测试如图3所示,表1(单位%)给出3种子空间聚类算法的精度. 从表1可见,SNR对双圆环数据和双月数据的精度达到100%,远高于LSR和LRR算法精度,由此可看出SNR对非线性数据的聚类是有效的.




(a) 双圆环




(b) 双月

**图 3 两个人工非线性数据集**

**表 1 两个圆环和SNR聚类后的图(ACC)**

| 方法 | LSR | LRR | SNR |
|---|---|---|---|
| 双圆环 | 50.5 | 50.5 | 100 |
| 双月 | 52.5 | 52 | 100 |

### 3.3 基因表达数据实验

本实验使用公开的基因表达数据集: SRBCT[20]、Prostate1[21]、Leukemia[22]、9Tumors[23]、Prostate[20]、LUNG[24],所有数据均进行$\ell_2$-范数标准化. 为避免随机性,算法均运行10次,取平均值作为当前参数下的最终聚类结果,实验结果见表2(单位%).

**表 2 实验结果(ACC)**

| 数据集 | 传统聚类算法 | | 非负铁蛋白分解算法 | |
|---|---|---|---|---|
| | $K$-means | HC | CNMF | SNMF |
| SRBCT | 44.578 3 | 33.734 9 | 43.494 0 | 46.385 5 |
| Prostate1 | 63.333 3 | 51.960 8 | 26.034 5 | 59.080 5 |
| Leukemia2 | 63.888 9 | 40.277 8 | 65.416 7 | 61.388 9 |
| 9_Tumors | 44.000 0 | 23.333 3 | 42.166 7 | 39.500 0 |
| Prostate | 56.862 7 | 51.960 8 | 51.111 1 | 61.111 1 |
| LUNG | 70.837 4 | 67.487 7 | 59.027 8 | 65.277 8 |

| 数据集 | 子空间聚类算法 | | | |
|---|---|---|---|---|
| | LSR | LRR | RLLRR | SNR |
| SRBCT | 52.777 1 | 59.397 6 | 37.590 4 | 74.337 3 |
| Prostate1 | 63.725 5 | 61.764 7 | 56.666 7 | 65.882 4 |
| Leukemia2 | 70.555 6 | 69.722 2 | 50.694 4 | 78.472 2 |
| 9_Tumors | 48.166 7 | 39.666 7 | 41.166 7 | 46.500 0 |
| Prostate | 62.745 1 | 60.784 3 | 60.784 3 | 76.470 6 |
| LUNG | 79.458 1 | 79.359 6 | 67.9803 | 90.788 2 |

从表 2 可见, 除 9_Tumors 数据集外, SNR 在剩余数据集都取得最好的聚类结果, 其中 SRBCT、Prostate、LUNG 上的精度高于其他聚类方法 10% 以上.

与传统方法 $K$-means 和 HC 相比, 子空间聚类方法总体上取得较高的聚类准确率, 一个可能的原因是子空间聚类方法比传统方法更适用于高维数据, 因此, 作为子空间聚类的方法, SNR 比传统方法能更好地处理高维基因表达数据.

由于 SNR 考虑了数据的非线性特点, 能更好地处理非线性的基因表达数据, 反映在基因表达数据上, SNR 算法聚类性能优于 3 种对比的子空间聚类方法 LSR、LRR、RLLRR.

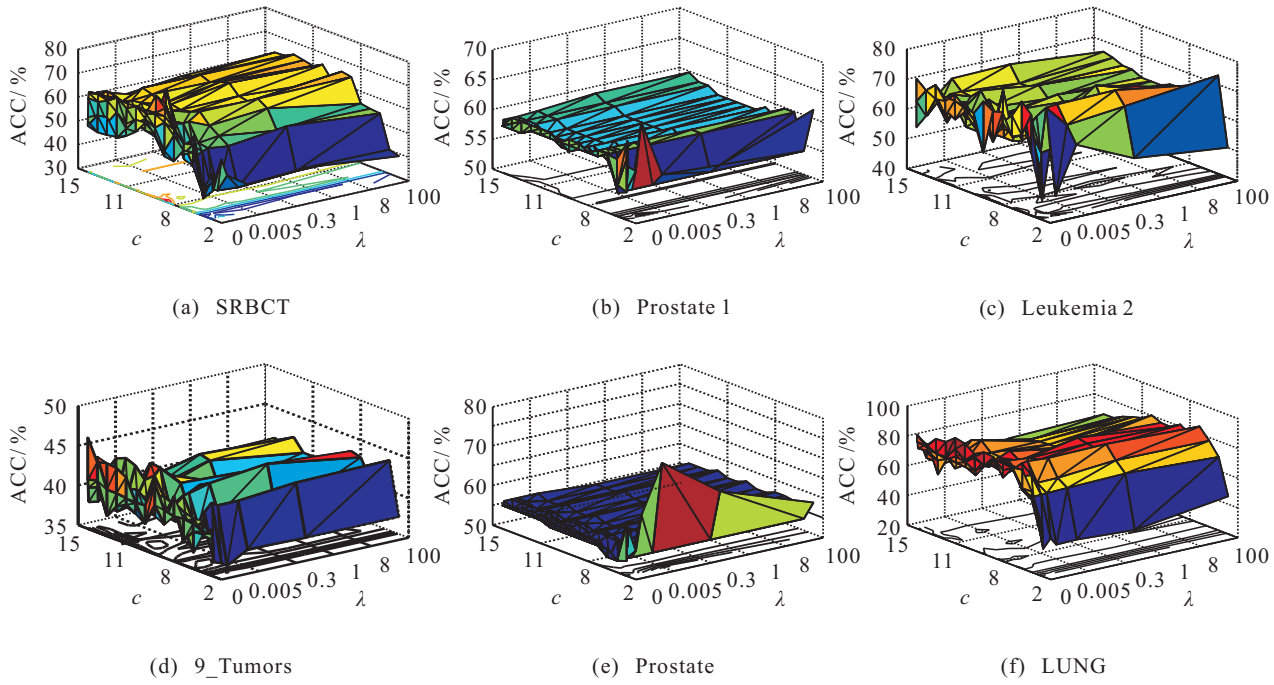
与非负矩阵分解方法相比, SNR 的聚类性能也明显更占优势. 一个可能的原因是 SNR 有解析解, 确保得到全局最优解, 而 CNMF 和 SNMF 容易得到局部最优解.

### 3.4　计算复杂度

SNR 的计算复杂度为 $O(c^2nd+nc^3)$, 影响计算复杂度的主要因素是近邻数 $c$、样本数 $n$ 和特征维数 $d$, 计算复杂度会随着样本规模、特征维度和近邻数的增大而增大. SNR 的近邻个数 $c \in [2, n]$, 可知 $n$ 对计算复杂度的影响比 $c$ 大. 另外, 基因表达数据是典型高维小样本数据, 因此特征维数 $d$ 对计算复杂度的影响远比样本数 $n$ 大.

### 3.5　参数分析

为能直观地分析 SNR 的正则参数 $\lambda$ 和近邻数 $c$ 对聚类准确率的影响, 利用图 4 显示两个参数变化时聚类精度的变化. 图 4 的 6 张子图分别对应 6 个基因表达数据集. 由图 4 可见, 近邻数 $c$ 在 2～10 时可以得到较高的 ACC. 随着正则参数 $\lambda$ 的变化, ACC 变化较为平坦, 当 $\lambda$ 取值较小时聚类准确率较高.




(a) SRBCT　(b) Prostate 1　(c) Leukemia 2

(d) 9_Tumors　(e) Prostate　(f) LUNG

**图 4　不同参数下基因表达数据聚类精度的变化**

## 4　结　论

本文针对基因表达数据的高特征维、小样本和非线性等特点提出光滑近邻表示子空间聚类 SNR, 因为全局非线性数据在局部可以近似看成具有线性关系, SNR 用数据点的近邻线性重构该数据点, 以此刻画数据非线性分布特征, 并用光滑正则项约束表示系数, 使原空间中数据点与近邻的距离关系得以保持. 实验结果表明, SNR 可以有效地对非线性的基因表达数据聚类.

参考文献(References)

[1] Pirim H, Ekşioğlu B, Perkins A D, et al. Clustering of high throughput gene expression data[J]. Computers & Operations Research, 2012, 39(12): 3046-3061.

[2] Sun J, Chen W, Fang W, et al. Gene expression data analysis with the clustering method based on an improved quantum-behaved particle swarm optimization[J]. Engineering Applications of Artificial Intelligence, 2012, 25(2): 376-391.

[3] Zhang W F, Liu C C, Yan H. Clustering of temporal gene expression data by regularized spline regression and an energy based similarity measure[J]. Pattern Recognition,

2010, 43(12): 3969-3976.

[4] Jiang D, Tang C, Zhang A. Cluster analysis for gene expression data: A survey[J]. IEEE Trans on Knowledge and Data Engineering, 2004, 16(11): 1370-1386.

[5] Zheng C H, Huang D S, Zhang L, et al. Tumor clustering using nonnegative matrix factorization with gene selection[J]. IEEE Trans on Information Technology in Biomedicine, 2009, 13(4): 599-607.

[6] Ding C, Li T, Jordan M I. Convex and semi-nonnegative matrix factorizations[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2010, 32(1): 45-55.

[7] Wang J J Y, Wang X, Gao X. Non-negative matrix factorization by maximizing correntropy for cancer clustering[J]. BMC Bioinformatics, 2013, 14(1): 1-11.

[8] 陈晓云, 陈慧娟. 潜在最小二乘回归子空间分割方法[J] 模式识别与人工智能, 2016, 29(1): 31-38.
(Chen X Y, Chen H J. Latent least square regression for subspace segmentation[J]. Pattern Recognition and Artificial Intelligence, 2016, 29(1): 31-38.)

[9] 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.

[10] Yang C, Robinson D, Vidal R. Sparse subspace clustering with missing entries[C]. Proc of the 32nd Int Conf on Machine Learning. Lille: CEUR, 2015: 2463-2472.

[11] Vidal R. A tutorial on subspace clustering[J]. IEEE Signal Processing Magazine, 2010, 28(2): 52-68.

[12] Elhamifar E, Vidal R. Sparse subspace clustering[C]. Proc of the 2009 IEEE Computer Society Conf on Computer Vision and Pattern Recognition(CVPR). Miami: IEEE, 2009: 2790-2797.

[13] Liu G, Lin Z, Yan S, et al. Robust recovery of subspace structures by low-rank representation[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2013, 35(1): 171-184.

[14] Lu C Y, Min H, Zhao Z Q, et al. Robust and efficient subspace segmentation via least squares regression[C]. European Conf on Computer Vision. Berlin: Springer, 2012: 347-360.

[15] Liu G, Yan S. Latent low-rank representation for subspace segmentation and feature extraction[C]. Proc of the IEEE Int Conf on Computer Vision(ICCV). Piscataway: IEEE, 2011: 1615-1622.

[16] Zhang H, Lin Z, Zhang C, et al. Robust latent low rank representation for subspace clustering[J]. Neurocomputing, 2014, 145(5): 369-373.

[17] Shi J, Malik J. Normalized cuts and image segmentation[J]. Trans on PatternAnalysis and Machine Intelligence, 2000, 22(8): 888-905.

[18] Boyd S, Vandenberghe L. Convex optimization[M]. New York: Cambridge University Press, 2004: 312.

[19] Cai D, He X, Wu X, et al. Non-negative matrix factorization on manifold[C]. Proc of the 8th IEEE Int Conf on Data Mining(ICDM). Hawaii: IEEE, 2008. 63-72.

[20] Yu L, Ding C, Loscalzo S. Stable feature selection via dense feature groups[C]. Proc of the 14th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: NY ACM, 2008: 803-811.

[21] Singh D, Febbo P G, Ross K, et al. Gene expression correlates of clinical prostate cancer behavior[J]. Cancer Cell, 2002, 1(2): 203-209.

[22] Armstrong S A, Staunton J E, Silverman L B, et al. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia[J]. Nature Genetics, 2002, 30(1): 41-47.

[23] Staunton J E, Slonim D K, Coller H A, et al. Chemosensitivity prediction by transcriptional profiling[J]. Proc of the National Academy of Sciences, 2001, 98(19): 10787-10792.

[24] Nie F, Huang H, Cai X, et al. Efficient and robust feature selection via joint $\ell_2$, 1-norms minimization[C]. Advances in Neural Information Processing Systems Cambridge. MA: MIT Press, 2010: 1813-1821.

（责任编辑：郑晓蕾）