聚类结果可视化的线珠模式①

王开军[1] 李 健[2](1.福建师范大学 数学与计算机学院 福建 福州 350108;
2.西北政法大学 网络信息中心 陕西 西安 710061)

摘 要: 对于揭示困难情况(高维且大类数)下的类间远近关系，现有的聚类结果可视化方法的效果均不理想。提出了以线珠模式和(非)线性坐标表示类间远近关系的可视化方法，其优点是在上述困难情况下也能准确地显示各个聚类之间的远近关系或距离。对模拟和真实数据集的实验结果表明，线珠模式方法十分有效，能准确地显示基于降维可视化方法无法正确显示的类间远近关系。

关键词: 聚类结果可视化；类间远近关系；线珠模式；affinity propagation clustering

Line-Pearl Pattern for Visualization of Clustering Results

WANG Kai-Jun[1], LI Jian[2]
(1. School of Mathematics and Computer Science, Fujian Normal University, Fuzhou 350108, China;
2. Net Information Center, Northwest University of Political Science and Law, Xi'an 710061, China)

Abstract: To uncover far/near relations between clusters accurately, existing clustering visualization methods usually have poor effects under the difficult condition of many clusters in a high-dimensional dataset. A new visualization method — line-pearl pattern with (non-) linear coordinates is proposed to show far/near relations or distances between clusters. Its advantage is that it can show far/near relations between clusters accurately under the difficult condition mentioned above. Experimental results on simulated and real data sets show that the proposed method is effective and much better than the clustering visualization methods.

Keywords: visualization of clustering results; relations between clusters; line-pearl pattern; affinity propagation clustering

聚类分析在各行业和科技领域中应用十分广泛，其中的划分聚类方法仅能给出数据集的每个样本的类别归属(即聚类结果)，而实际中更重要且困难的任务—准确掌握聚类之间的远近关系则通常由聚类结果的可视化来完成。聚类结果可视化的主要方法[1-3]包括：数据矩阵、平行坐标法、带误差棒的均值模式、系统树图以及基于降维或映射技术(例如主分量分析、等距映射和基于参考点的核映射等)的数据散点图等。对于获取类间远近信息的任务，上述方法对大尺度数据(例如高维、大类数)的效果均不理想，并且均不能给出量化的类间远近信息。故此，针对划分聚类方法的聚类结果，例如K-均值算法、围绕中心点划分(PAM)算法和仿射传播聚类(affinity propagation clustering, AP)算法[4,5]的聚类结果，如何准确地显示大尺度数据的类间远近关系仍是一个具有挑战性的问题。本文在前期工作[6]的基础上，提出了以线珠模式和(非)线性坐标表示类间远近关系的可视化方法，以解决现有可视化方法不能准确地量化显示类间远近关系且在高维、大类数的困难情况下效果不佳的问题。本文方法的程序可由文献[6]获得。

1 类间远近关系的线珠模式

首先介绍测量两个聚类之间远近程度的双几何体

① 基金项目:福建省教育厅项目(JA09043)
收稿时间:2009-09-08;收到修改稿时间:2009-10-20

模型[6]，再设计线珠模式的可视化方法。

1.1 双几何体模型

由许多样本点聚集成的一个聚类可视为在空间中具有几何形状和空间范围的一个聚类几何体。双几何体模型是由二个聚类 A 和 B 的聚类几何体组成，每个聚类几何体被划分为边界区域和非边界区域(参见图 1)，其中：

① 边界区域 A1 中的 $n_{A1}(=2\sqrt{n_A}$ 的下取整)个样本要比聚类 A 中其它样本更靠近聚类 B；边界区域 B1 中 $n_{B1}(=2\sqrt{n_B}$ 的下取整)个样本要比聚类 B 中其它样本更靠近聚类 A。

② 非边界区域 A0 中的 n_{A0} 个样本由除去区域 A1 后的聚类 A 中的样本构成；非边界区域 B0 中的 n_{B0} 个样本由除去区域 B1 后的聚类 B 中的样本构成。

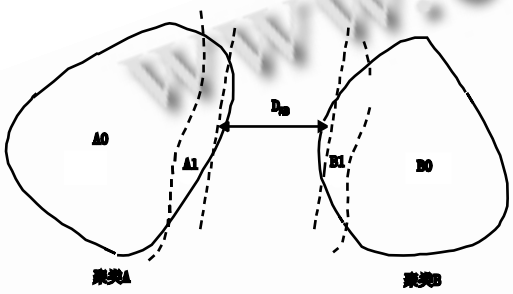

图 1　双几何体模型的聚类 A 与 B 被分别划分为边界区域 A1 与 B1 和非边界区域 A0 与 B0

两个聚类之间的绝对边界距离 D_{AB} 由最近样本集距离法进行测量：设 $d(a_i,b_j)$为样本 a_i 与 b_j 之间的距离；对于区域 A1 中的每个样本 a_i，找出区域 B1 中最靠近 a_i 的样本 b_j(简称最近样本)，可形成 n_b 个最近样本的集合$\{b_j\}$；对于区域 B1 中的每个样本 b_i，找出区域 A1 中最靠近 b_i 的样本 a_j，可形成 n_a 个最近样本的集合$\{a_j\}$；于是，

$$D_{AB}=\begin{cases}\dfrac{1}{n_a}\displaystyle\sum_{j=1}^{n_a}\min_{i=1\sim n_b} d(a_j,b_i), & 若\quad n_a > n_b\\ \dfrac{1}{n_b}\displaystyle\sum_{j=1}^{n_b}\min_{i=1\sim n_a} d(b_j,a_i), & 若\quad n_a \le n_b\end{cases} \tag{1}$$

边界区域内样本的疏密程度由最近传播法测量：从区域 A1 中的启始样本 a_1 开始，找出 A1 中与 a_1 距离最近的样本 a_2(称为最近传播到 a_2)，计算 a_1 与 a_2 之间的距离 $d(a_1,a_2)$；再由 a_2 最近传播到 a_3，计算 $d(a_2,a_3)$；类似地传播到 A1 中的每一个样本，完成一次传播。选取如下三个启始点：找出区域 A1 中最靠近区域 B1 的样本 s1，以及找出区域 A1 中最远的两个样本 s_2和 s_3。于是，区域 A1 的最近传播距离 D_{NA}(区域 B1 类似有 D_{NB})为：

$$D_{NA}=\frac{1}{3(n_{A1}-1)}\sum_{k=1}^{3}\left[\sum_{i=1}^{n_{A1}-1} d(a_i,a_{i+1})\Big|_{a_1=S_k}\right] \tag{2}$$

再将区域 A1 沿 A1 与 B1 中心的大致连线方向进行分层，靠近区域 B1 的 $n_{A1}/2$ 个样本作为 1 层区域，其余样本作为 2 层区域，3 层区域则由 1 层和 2 层区域各一半相互更靠近的样本组成。将每层内最近传播法测量的结果进行平均可得 D_{LA}，于是边界区域 A1 内样本的疏密程度为：$D_{A1}=(D_{NA}+D_{LA})/2$，类似地有 $D_{B1}=(D_{NB}+D_{LB})/2$。最后，两个聚类之间的相对边界距离为：

$$R_{AB}=D_{AB}/\max(D_{A1},D_{B1}) \tag{3}$$

1.2 显示聚类结果的线珠模式和对数坐标

考虑到类间远近程度显示的直观性，不采用相对边界距离的公式(3)，而是作如下新设计。由于边界区域内样本之间的疏密程度直接影响两个聚类之间远近关系的辨别，两个聚类之间远近程度的合理测度设计为绝对距离 D_{AB} 减除边界区域内样本疏密程度后的相对距离 H_{AB}(或记为 H(A,B))：

$$H_{AB}=D_{AB}-\max(D_{A1},D_{B1}) \tag{4}$$

设平面框图上的垂直坐标轴 y 为分度均匀的线性轴(普通坐标)，每个聚类 C_i 用带数字的圆圈表示(例如一个圆圈和其边上的数字 2 代表 C_2)，先将每个聚类等间距地放置到平面框图的底部作为观察点；对每个观察点 C_i，绘制其余聚类 C_j，并将聚类 C_j 沿垂直方向按与 C_i 的远近程度 $H(C_i,C_j)$放置，再用直线将 C_i 与 C_j 连接起来，形成显示类间远近关系的线珠模式(参见图 2)。

当一些聚类之间相对距离 H_{AB} 的数量级有差别时，上述线性线珠模式的观察效果会变差，某些间距较小的聚类粘连到一起而很难辨别。为解决这种粘连情况，不采用线性坐标系表示类间远近程度，而设计分度不均匀的非线性坐标系显示相对距离 H_{AB}：设 H_{AB} 在线性的垂直坐标轴 y 上的值为 y_1，若采用以 e 为底的对数坐标轴(ln(y))替代线性坐标轴 y，则在对数坐标轴上 H_{AB} 的对应值为对数 $\ln(y_1)$。对一些聚类之间 H_{AB}

的数量级相差几个级别的更复杂情况，再设计超对数坐标轴(ln(ln(y)))来显示类间的远近关系，则在超对数坐标轴上 H_{AB} 的对应值为 $\ln(\ln(y_1))$。在实际应用中，可通过比较在线性坐标系、对数坐标系或超对数坐标系下类间关系的显示效果确定一种最合适的显示坐标系。

在上述非线性坐标系中，两个类之间的位置关系(距离)是按实际类间距离的(超)对数值绘制的。当 $H_{AB} \leqslant 1$ 时，其对数值为负数或零，不能表示已有的距离。这里设计平移方法解决这个问题：将所有类间距离同时加 e 值再取对数，同时坐标分度的显示值减 e 值。

为便于直观理解非线性坐标系中的距离 H_{AB}，将(超)对数坐标轴上的坐标分度仍标注量本身的值，即标注值 y_1，而不标注 $\ln(y_1)$或 $\ln(\ln(y_1))$。例如图 2 的下图中采用超对数坐标轴显示 H_{AB}，最大坐标分度标注为 1343，而不标注 $\ln(\ln(1343)) = 1.97$。

2 实验结果

本节实验将检验线珠模式显示类间远近关系的性能，同时与基于主分量分析[3]和基于核映射[3](按该文献中的算法步骤选取最优参数)的可视化方法进行比较。设一个数据集有 n 个样本$\{x_i\}$，关于 x_i 和 x_j 之间相似性测度，对一般数据采用欧式距离，对基因表达数据采用 Pearson 相关系数[7]。实验步骤是先采用 AP 算法将 n 个样本聚为 k 类，再依据双几何体模型按公式(4)计算这 k 个聚类之间的距离 HAB，最后用线珠模式将聚类结果可视化。

第一个数据集 14k10far2 是将模拟数据集 y14c[8](具有 480 个 10 维样本和 14 个相距较远的聚类)的第 5 类和第 10 类的每个样本 x_i 的每一维数值 x_{ij} 分别加以 400 和 420 而产生的，各个聚类相距较远或很远。实验中聚类结果的错误率为 0。本文方法显示的各聚类间的远近信息见图 2，其中采用了超对数坐标轴来显示 H_{AB}。可以看出，C_5 和 C_{10} 与其它聚类相距很远(距离 1000 以上)，其余聚类之间的间隔距离大多在 9 至 26 的范围，相距较远(包括最近的 C_6 和 C_9 相距不到 6)。这些均与已知信息相一致。核映射方法的可视化结果见图 2(主分量方法的结果类似，略)，图中错误显示 C_9 与 C_{14} 有重叠，这是映射变形造成的假象；这也说明该方法不能真实或准确反映类间距离。

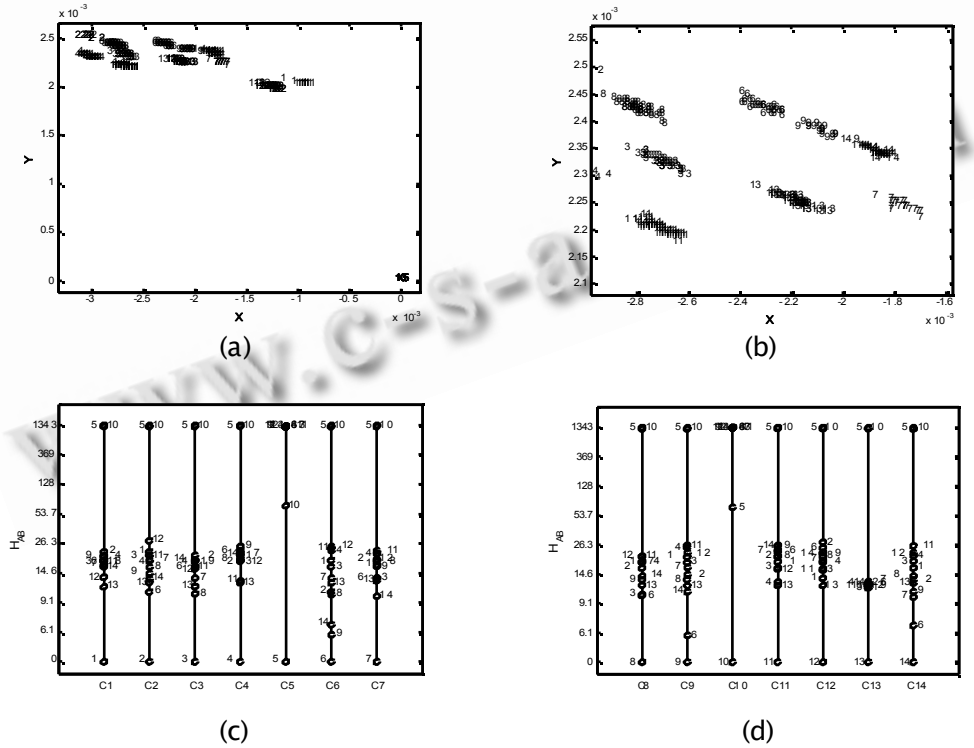

图 2 (a)：核映射方法显示的 14k10far2 数据集的聚类结果((b)是局部放大图)；第 i 类的数据点用数字 i 显示。(c)和(d)：超对数坐标线珠模式显示聚类结果的聚类 C_i 与其它聚类间的远近程度，其中圆圈表示聚类，圆圈边的数字 i 表示第 i 个聚类

第二个数据集 M637-11k14 是关于 18 个基因的数据集 Monocyte[9]的子集，是由 11 个基因(类)的 637 个 14 维样本构成的；聚类结果的错误率为 0 说明各聚类之间的分离性良好。图 3 显示本文方法能逐一揭示各聚类之间的远近程度，除 C_6 与 C_9 快要重叠或很靠近外，每个聚类均与其它聚类有一定的间隔距离，分离良好；这些均与已知信息相一致。主分量与核映射方法的可视化结果见图 3，图中错误地显示 C_1、C_2、C_4、C_6、C_{10}、C_{11} 叠加或交织在一起，无法分辨其远近关系；这也说明该方法不能正确显示类间关系和距离。

上述线珠模式的结果均与已知类间远近关系相一致，验证了本文方法能有效地显示出聚类之间的远近关系，而且能准确地显示类间的远近程度。而对比方法在类数较多且高维数据的困难情况下，由于高维数据的降维映射所带来的数据间远近关系的变形，一些聚类出现叠加或交织现象，显示出错误的类间远近关系。

3 结论

本文提出了线性和非线性线珠模式来表示类间远近关系的可视化方法。线珠模式是以每个聚类 C_i 为观察点，绘制并显示其余聚类与 C_i 的远近程度的分布式可视化方法，其优点是在高维数据和大类数的复杂情况下也能准确地揭示各聚类之间的远近关系。

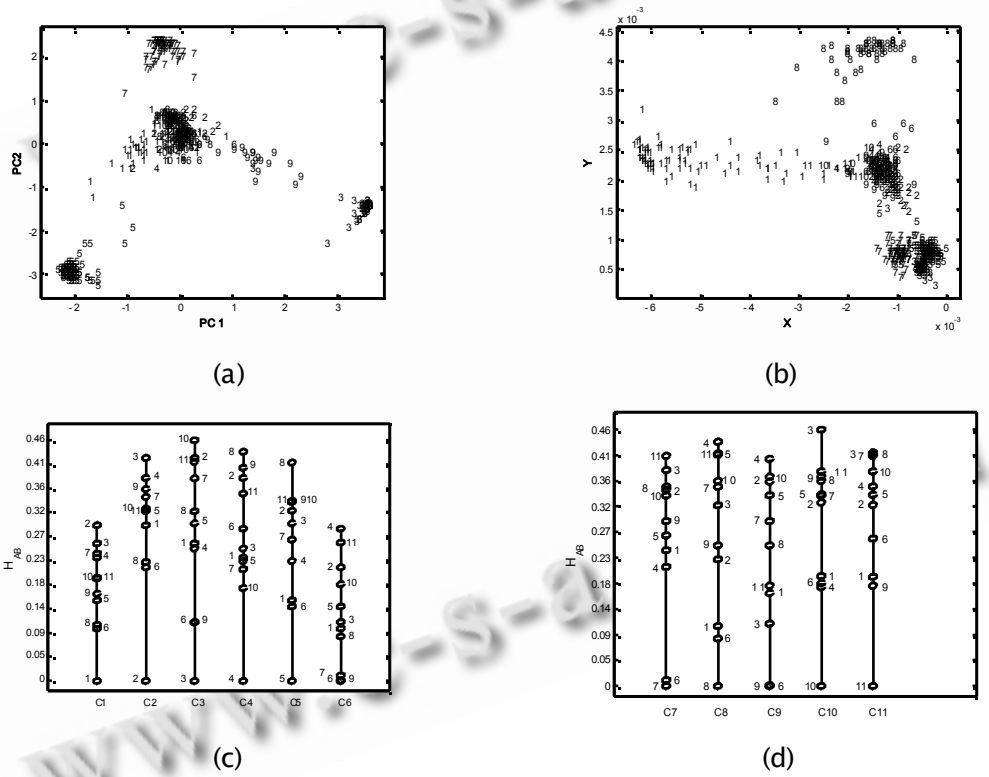

图 3 (a)基于第 1、第 2 主分量和(b)核映射方法显示的 M637-11k 数据集的聚类结果；第 i 类的数据点用数字 i 显示。(c)和(d)：线性线珠模式显示的聚类结果中聚类 C_i 与其它聚类间的远近程度，其中圆圈表示聚类，圆圈边的数字 i 表示第 i 个聚类

参考文献

1 Sharan R, Maron-Katz A, Shamir R. CLICK and EXPANDER: A System for Clustering and Visualizing Gene Expression Data. Bioinformatics, 2003,19(14): 1787－1799.

2 任永功.面向聚类的数据可视化方法及相关技术研究[博士学位论文].沈阳:东北大学, 2006.

3 Suykens JAK. Data visualization and dimensionality reduction using kernel maps with a reference point. IEEE Trans on Neural Networks, 2008,19(9):1501－1517.

4 Xu R, Wunsch II D C. Survey of clustering algorithms.

(下转第 182 页)

(上接第 170 页)

IEEE Transactions on Neural Networks, 2005,16(3): 645－678.

5 Frey BJ, Dueck D. Clustering by passing messages between data points. Science, 2007,315(5814):972－976.

6 王开军.识别聚类间远近关系的双几何体模型.技术报告. 福建师范大学, 2009. http://www.mathworks.com/matlabcentral/fileexchange/authors/24811

7 王开军,张军英,李丹,张新娜,郭涛.自适应仿射传播聚类.自动化学报, 2007,33(12):1242－1246.

8 Dembélé D, Kastner P. Fuzzy C-means method for clustering microarray data. Bioinformatics, 2003,19 (8):973－980.

9 Hartuv E, Schmitt A, Lange J, Meier-Ewert S, Lehrach H, Shamir R. An algorithm for clustering cDNAs for gene expression analysis. Genomics, 2000,66(3):249－256.