论 文

识别聚类间远近关系的双几何体模型

王开军*, 严宣辉, 陈黎飞

福建师范大学数学与计算机科学学院, 福州 350108

* 通信作者. E-mail: wkjwang@gmail.com

收稿日期: 2010-07-30; 接受日期: 2011-03-08

福建省教育厅 A 类资助项目 (批准号: JA09043) 和福建省省属高校科研专项 (批准号: JK2009006) 资助项目

摘要 许多实际问题的解决不仅需要聚类算法给出类标, 更依赖于类间远近关系的辨别. 对于类数较多且高维数据的困难情况, 基于降维的聚类结果可视化方法通常会出现聚类的重叠、交织或强行拉远现象, 使得一些类间的远近关系无法分辨或被错误显示; 而现有的类间距离方法则不能揭示两个聚类是远离还是靠近. 本文提出了双几何体模型方法来描述两个聚类的类间关系, 并设计了相对边界距离、绝对边界距离和区域疏密程度等测量类间远近程度的方法. 本文方法既考虑了两个聚类的最近样本集之间的绝对距离, 也考虑了聚类边界区域的疏密程度, 其优点是在上述困难情况下也能准确揭示高维空间中的类间关系. 对真实数据集的实验结果表明, 双几何体模型方法能有效地识别现有聚类可视化方法无法辨别的类间远近关系.

关键词 双几何体模型 聚类间远近关系 大类数 高维数据 划分聚类算法

1 引言

基于均方误差的划分聚类方法主要包括 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) 算法等^[1,2]. 这些聚类算法仅是给出数据集的每个样本的类别归属, 然而获取聚类间的远近信息对掌握数据集的聚类结构和解决许多实际问题是十分重要的. 例如, 采用聚类方法对 72 个白血病病例的高维基因表达数据进行分析时, 只有获得病类之间远近程度的信息, 才能辨别并发现基因表达数据空间中很靠近 ALL 病类的新的 MLL 病类^[3], 避免误诊.

已有几种类间距离的定义^[4~6]是涉及聚类间的远近关系, 例如: 最小距离定义为分别属于两类的样本之间的最近距离, 最大距离是分别属于两类的最远样本之间的距离, 类平均距离定义为分别属于两类的对象之间的平均距离, 中心距离是两类的中心之间的距离, 而先求出一个聚类的每个样本与另一个聚类的样本中最近样本的距离, 再找出这些距离中的最大距离则作为 Hausdorff 距离. 这些距离的设计目的主要是为一些聚类算法 (如凝聚的层次聚类) 服务, 以及用于计算聚类有效性指标以评价聚类的质量 (类间距离越大而类内距离越小则质量越好). 这些类间距离 (除了最小距离) 都没有描述两个聚类最靠近部分的距离; 最小距离虽是不同类的最近的两个样本之间的绝对距离, 但不考虑聚类本身疏密程度的绝对距离不能揭示两个聚类是远离还是靠近的关系.

聚类结果的可视化是获得聚类间远近信息的另一种途径. 可视化方法用图形的方式在二维或三维空间直观地显示出聚类结果 (数据点的聚集情况及聚类之间的关系). 所采用的距离测度直接影响聚

引用格式: 王开军, 严宣辉, 陈黎飞. 识别聚类间远近关系的双几何体模型. 中国科学: 信息科学, 2012, 42: 99-110

类算法产生的聚类结果, 也影响可视化效果. 利用应用问题的已知条件, 通过有监督地学习有标记的样本, 可设计出一个更好的距离测度以满足特定的聚类任务^[7]. 不失一般性, 本文涉及的距离指欧氏距离. 下文将简介常用的聚类结果的可视化方法^[8~20], 并设具有 n 个 d 维样本的数据集由 k 个聚类组成.

数据矩阵是数据在二维空间直接显示的方法, 主要包括: 按类排列的彩色数据矩阵, 它是 n 行 (样本) d 列 (维) 的数据布置方式, 数据值的大小使用颜色 (深浅) 或灰度来描述, 同一类的样本排列在一起, 不同类之间可用分隔线区分; 双向聚类数据矩阵, 其特点是两个坐标轴的数据都按聚类结果排列.

平行坐标法也是在二维空间直接显示数据, 它采用 d 个等间距且平行的 (垂直) 坐标轴表示 d 维空间, 每个样本的第 i 维数据对应到第 i 个坐标轴上, 其 d 个点用线段连接起来形成一条折线 (表示一个空间点), 并且不同的类可用不同颜色的折线显示. 由于数据量大时图形存在重叠、层次不清现象, 故适合显示少量的高维数据. 该方法的一种变化是圆形平行坐标法, 其不同之处是 d 个坐标轴的一端在圆心, 另一端在圆上等间距布置.

基于降维或映射技术的数据散点图 (简称降维可视化) 是把高维数据投影到二维或三维空间来观察详细的聚类结构. 降维可视化技术主要包括: 传统的主成分分析 (PCA) 和多维尺度变换 (MDS); 流形学习方法如等距映射 (Isomap)、局部线性嵌入 (LLE) 和 Laplacian 特征映射 (LE) 等, 都是通过高维数据点的邻域来获得局部相邻点之间的关系, 然后以试图保持这种局部关系为目标, 将高维数据点映射到一个全局的低维空间; 使用数据点的局部切空间来描述高维数据点的局部特性的局部切空间排列 (LTSA) 和综合了核方法与半正定规划的最大方差展开 (MVU) 也是流形学习方法; 以保持邻近点在观测空间和嵌套空间概率分布相似为目标的 t 随机邻域映射 (tSNE) 是概率框架下的方法; 基于参考点的核映射 (KMref) 则是利用核函数的方法, 它通过在参考点和保持局部距离约束下的核映射将高维数据映射到低维空间.

自组织映射 (SOM) 是一种神经网络模型, 能将高维数据映射到平面网络的拓扑结构上, 但提供的是网络结点之间粗略的距离关系. 为了获得更精细的可视化效果, 一些方法对 SOM 进行了改进, 例如: 生成式拓扑映射 (GTM) 是一种非线性的隐变量模型, 可视为 SOM 的概率重表达形式; 层次 GTM 方法能将 GTM 映射模型交互式分解为层次化的子模型, 可实现局部放大和调整布局的功能; 可视化诱导自组织映射 (ViSOM) 能提供数据的结构和分布情况的可视化, 而概率正则化 SOM (PRSOM) 通过引进代价函数对 ViSOM 方法进行了改进, 但它们的不足是计算时需要很大的内存.

其他方法有均值模式和树图等. 带误差棒的均值模式为计算出的每个聚类的均值模式和标准差绘制一个框图; 数据的系统树图则是以层次结构的方式显示聚类结果, 主要适用于层次聚类算法或具有层次结构的数据. 此外, 特征选择的方法也可以实现降维, 例如前向正交搜索, 但选择出的特征应能够充分代表所有的特征.

由上述方法的特点可知: 对于获取高维、大类数数据集的聚类间远近信息的任务, 数据矩阵方法在大数据量时很难观察清楚类间远近; 平行坐标法对于大量数据则难于识别类间信息; 均值模式受误差棒的影响而很难观察类间远近情况; 降维可视化方法因布局困难常常出现错误的聚类重叠、交织或拉远现象, 供观察的类间远近关系失真. 总之, 这些方法的不足是: 主要依赖视觉分析的方法是费力的、很主观的且不适合更复杂的和大尺度的情况^[21] (例如高维、大类数、大数据集).

对大尺度数据的分析和处理已成为当前和今后研究的重点和难点. 因此, 针对基于均方误差的划分聚类方法, 本文提出了识别聚类间远近的双几何体模型方法, 来获得量化的聚类间远近信息, 以克服现有方法不适用于高维、大类数情况的不足. 本文方法的 Matlab 程序可由文献 [22] 获得.

图 1 聚类 A 和 B 组成的双几何体模型

Figure 1 Double-cluster geometric model consists of cluster A and B , each cluster is divided into a border region (A_1 & B_1) and non-border region (A_0 & B_0), D_{AB} is the border distance

2 双几何体模型

首先建立两个聚类的关系模型, 再依据该模型设计两个聚类的边界区域之间远近程度的测量方法, 最后设计确认和调整边界区域的方法.

2.1 两个聚类的关系模型

将由许多样本点聚集成一个聚类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.

定义 1 (双几何体模型). 两个聚类 A 和 B 的双几何体模型由它们的聚类几何体组成. 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 A_1 中的 n_{A_1} 个样本要比聚类 A 中其他样本更靠近聚类 B ; 边界区域 B_1 中 n_{B_1} 个样本要比聚类 B 中其他样本更靠近聚类 A . (2) 非边界区域 A_0 中的 n_{A_0} 个样本由除去区域 A_1 后的聚类 A 中的样本构成; 非边界区域 B_0 中的 n_{B_0} 个样本由除去区域 B_1 后的聚类 B 中的样本构成.

边界区域的大小 (样本数目) 应适合每个聚类几何体的大小. 为了使边界区域 A_1 的样本数目 n_{A_1} 能自适应于聚类 A 的样本总数 n_A 的变化, 将 n_{A_1} 设计为 $2\sqrt{n_A}$ 的下取整. 类似地, n_{B_1} 设计为 $2\sqrt{n_B}$ 的下取整. 现设计产生 A_1 中 n_{A_1} 个样本的择近选取法: 计算聚类 A 中每个样本到聚类 B 中所有样本的距离, 将其中最小距离作为该样本到聚类 B 的间隔距离, 依据这些间隔距离选出比其他样本更靠近聚类 B 的 n_{A_1} 个样本 $\{a_i\}$ ($i = 1, \dots, n_{A_1}$) 构成 A_1 . 类似地, 可确定 B_1 的 n_{B_1} 个样本 $\{b_i\}$ ($i = 1, \dots, n_{B_1}$).

2.2 两个聚类之间远近程度的测量方法

依据上述双几何体模型, 设计对两个聚类之间的绝对边界距离 D_{AB} 进行测量的方法.

最近样本集距离法: 先找出区域 A_1 和 B_1 中相互最靠近的样本, 即设 $d(a_i, b_j)$ 为样本 a_i 与 b_j 之间的距离, 对于区域 A_1 中的每个样本 a_i , 找出区域 B_1 中最靠近该样本的样本 b_j (简称最近样本), 可获得 n_b 个最近样本的集合 $\{b_j\}$; 若 $\{b_j\}$ 只包含一个样本, 则增加一个次最近样本; 类似地, 可获得

n_a 个最近样本的集合 $\{a_j\} (a_j \in A_1)$; 最后, 计算集合 $\{a_j\}$ 与 $\{b_j\}$ 之间的距离作为两个聚类的绝对边界距离:

$$D_{AB} = \frac{1}{n_a + n_b - n_c} \left[\sum_{j=1}^{n_a} \min_{i=1, \dots, n_b} d(a_j, b_i) + \sum_{j=1}^{n_b - n_c} \min_{i=1, \dots, n_a} d(b_j, a_i) \right], \quad (1)$$

其中, 后一项的 $d(b_j, a_i)$ 中不包含与前一项 n_a 个 $d(a_j, b_i)$ 中有重复的元素 ($d(a_i, b_j)$ 和 $d(b_j, a_i)$ 视为相同的元素), 重复元素个数记为 n_c .

两个聚类之间远近程度的合理测度应当是绝对距离 D_{AB} 与边界区域内样本的疏密程度对比后的相对距离 R_{AB} . 假设 A_1 由 3 个样本 $\{a_1, a_2, a_3\}$ 组成, $d(a_1, a_2) < d(a_2, a_3) < d(a_1, a_3)$; 考虑与最近样本集距离 D_{AB} 相匹配, 取 $d(a_1, a_2)$ 和 $d(a_2, a_3)$ 的平均值作为区域内样本的疏密程度最合适. 依据此思路, 设计如下方法测量边界区域内样本的疏密程度.

最近传播法: 从区域 A_1 中的起始样本 a_1 开始, 找出 A_1 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 A_1 中的每一个样本, 完成一次传播; 选取如下 3 个起始点: 找出区域 A_1 中最靠近区域 B_1 的样本 s_1 , 以及找出区域 A_1 中最远的两个样本 s_2 和 s_3 ; 由这 3 个起始点开始分别传播一次, 可获得区域 A_1 的最近传播距离

$$D_{NA} = \frac{1}{3(n_{A_1} - 1)} \sum_{k=1}^3 \left[\sum_{i=1}^{n_{A_1}-1} d(a_i, a_{i+1}) \Big|_{a_1=s_k} \right]. \quad (2)$$

考虑到当 A_1 较大的中心区域无或极少样本点时最近传播法不准确的问题, 再设计测量 A_1 中心区域疏密的圈心传播法: 先计算区域 A_1 的中值中心 o_0 (由 A_1 中所有样本的每一维数据的中值产生); 按择近选取法找出区域 A_1 中最靠近区域 B_1 的 $n_{A_1}/2$ 个样本作为 1 层区域, 并计算 1 层区域的中值中心 o_1 和 $d(o_0, o_1)$; A_1 的其余样本作为 2 层区域, 并计算 2 层区域的中值中心 o_2 和 $d(o_0, o_2)$; 然后找出离中心 o_0 的距离小于等于 $\max\{d(o_0, o_1), d(o_0, o_2)\}$ 的那些样本组成圈心区域; 最后用最近传播法求出加上虚拟样本 o_1 和 o_2 后的圈心区域的传播距离 D_{OA} . 类似地, 可得区域 B_1 的最近传播距离 D_{NB} 和圈心区域的传播距离 D_{OB} .

两个虚拟样本点在中心区域样本点越少时起的作用就越大, 特别是当中心区域为空时传播路线仍会通过中心区域. 考虑到高维样本点近似位于以数据集的均值为中心的超球上^[23], 圈心传播法不适用于高维情况, 将只用于样本的维数小于 10 的情况. 将最近传播法和圈心传播法的结果进行综合, 则边界区域 A_1 和 B_1 内样本的疏密程度分别为: $D_{A_1} = (D_{NA} + D_{OA})/2$, $D_{B_1} = (D_{NB} + D_{OB})/2$.

最后, 基于类间绝对边界距离与聚类边界区域疏密程度比值的两个聚类之间的相对边界距离为

$$R_{AB} = \frac{D_{AB}}{\max\{D_{A_1}, D_{B_1}\}}. \quad (3)$$

2.3 双几何体模型边界区域的精炼机制

2.1 小节中确定的边界区域大小 (如 $n_{A_1} = 2\sqrt{n_A}$) 是否合适可以依据聚类中样本的分布情况进行检验, 并在必要时进行调整 (缩小或扩大边界区域). 这里设计关于边界区域与其相邻区域的疏密程度的假设检验方法来实现, 其步骤如下 (以聚类 A 为例):

(1) 按择近选取法找出相邻区域: 区域 A_0 中具有 n_{A_1} 个更靠近 A_1 的样本组成区域 A_2 , 区域 A_2 中具有 $n_{A_1}/2$ 个更靠近 A_1 的样本组成区域 A_{2h} , 区域 A_1 中具有 $n_{A_1}/2$ 个更靠近聚类 B 的样本组成区域 A_{1h} , 以及区域 A_1 除去区域 A_{1h} 后组成区域 A_{1f} .

(2) 采用 t 检验方法^[24] 比较各相邻区域的疏密程度, 如果区域 A_2 比 A_1 明显密集且区域 A_{1f} 比 A_{1h} 明显密集, 则缩小区域 A_1 为 A_{1h} , 算法停止; 否则转步骤 (3).

(3) 采用 t 检验方法继续比较, 如果区域 A_2 比 A_1 稀疏 4 倍以上 (针对 A_1 中样本几乎重叠到一个点这类情况; 下文稀疏 4 倍也简称稀疏), 则将 A_{2h} 与 A_1 合并, 组成扩大的区域 A_1 ; 否则不调整区域 A_1 .

实现步骤 (2) 和 (3) 中的比较任务, 还需要下面的具体设计. 由传播法计算区域 A_1 的疏密时可得各样本间传播距离的集合 $S_{A_1} = \{d(a_i, a_{i+1})\}$, 区域 A_2 类似地有 S_{A_2} . 首先采用 Lilliefors 检验^[25] 方法 (显著性水平取 0.05) 对 S_{A_1} 和 S_{A_2} 是否符合一维高斯分布进行检验.

若 S_{A_1} 和 S_{A_2} 均符合高斯分布, 则可看成分别来自均值 μ_{A_1} 和 μ_{A_2} 的高斯分布. 于是, 判别区域 A_2 是否比 A_1 明显密集 (或稀疏), 可具体化为判别均值 μ_{A_2} 是否显著小于均值 μ_{A_1} (或大于 $5\mu_{A_1}$). 这里设计 t 检验 (显著性水平取 0.05) 过程来实现这项判别任务 (其他区域的检验过程也类同):

(1) 对于判别区域 A_2 是否比 A_1 明显密集, 将零假设 H_0 设置给均值 μ_{A_2} 不小于均值 μ_{A_1} 的情况, 而将备择假设 H_1 设置给均值 μ_{A_2} 显著小于均值 μ_{A_1} 的情况, 即 $H_0: \mu_{A_1} \leq \mu_{A_2}, H_1: \mu_{A_1} > \mu_{A_2}$; 对于判别区域 A_2 是否比 A_1 稀疏, 则有 $H_0: 5\mu_{A_1} \geq \mu_{A_2}, H_1: 5\mu_{A_1} < \mu_{A_2}$;

(2) 按两分布总体的方差不同且未知情况下的公式^[26] 计算检验统计量 t 和自由度 v 的值;

(3) 查 t 分布表获得具有自由度 v 的临界值 $t_{\alpha, v}$;

(4) 如果 $t \leq -t_{\alpha, v}$ (对于稀疏判别应 $t \geq t_{\alpha, v}$), 拒绝假设 H_0 而接受假设 H_1 , 即 A_2 比 A_1 明显密集 (或稀疏); 否则, 接受假设 H_0 而拒绝假设 H_1 , 即 A_2 不比 A_1 明显密集 (或稀疏).

当 S_{A_1} 或 S_{A_2} 不符合高斯分布时, 则采用非参数检验方法——Mann-Whitney U 检验^[27,28] 替代 t 检验来判别区域 A_2 是否比 A_1 明显密集 (或稀疏), 其检验过程设计如下 (其他区域的检验过程也类同):

(1) 对于判别区域 A_2 是否比 A_1 明显密集, 设置: $H_0: A_2$ 不比 A_1 密集, $H_1: A_2$ 比 A_1 密集; 对于判别区域 A_2 是否比 A_1 明显稀疏, 设置: $H_0: A_2$ 不比 A_1 稀疏, $H_1: A_2$ 比 A_1 稀疏;

(2) 将 S_{A_1} 和 S_{A_2} 混合, 并按照从小到大顺序编排等级, 形成等级数据;

(3) 设 S_{A_1} 和 S_{A_2} 中样本个数分别为 n_1 和 n_2 , 用等级数据计算检验统计量 U_1 和 U_2 ;

(4) 当 n_1 和 n_2 均不大于 10 时 (小样本情况), 查临界值表得临界值 U_α 和 $U_{1-\alpha}$, 转步骤 (6); 否则转步骤 (5);

(5) 计算 $U = U_1$ 的均值 μ_U 和标准差 σ_U , 再计算检验统计量 $z = (U - \mu_U)/\sigma_U$; 查临界值表得 z_α 和 $z_{1-\alpha}$;

(6) 如果对于小样本情况 $U_1 < U_\alpha$ 或对于非小样本情况 $z < z_\alpha$ (对于稀疏判别应 $U_1 > U_{1-\alpha}$ 或 $z > z_{1-\alpha}$), 则拒绝假设 H_0 而接受假设 H_1 ; 否则, 接受假设 H_0 而拒绝假设 H_1 .

最后讨论算法的复杂度. 任两个聚类的边界区域和边界距离的确定以及最近传播法都需要使用所有样本之间的相似度, 计算相似度的复杂度为 $O(n^2)$; 计算所有聚类间边界距离的复杂度与有限的类数 k 和每个类的样本数 $n_i < n$ 有关, 即 $n_i^2 C_k^2$; 故算法的总复杂度为 $O(n^2)$.

3 实验结果

本节对双几何体模型方法^[22] 识别聚类间远近关系的性能进行实验, 同时与 Hausdorff 距离 ($dH_{i,j}$)^[5]、最小距离 ($dm_{i,j}$)^[5]、LLE^[12]、MVU^[13]、tSNE^[14] (LLE, MVU 和 tSNE 的最近邻点个数

均取 12)、KMref^[15] (最优参数依据文献 [15] 选取) 和 GTM^[16] (训练次数 20, 最优 σ 的搜索网格 [0.25, 0.5, 1, 2, 4], $\lambda=0.001$, 基函数网格 4×4 , 隐样本网格 15×15) 方法进行比较. 除 Yeast 数据集外, 实验数据集的每个样本的类标均已知.

在实验中, 先使用 AP 算法 (相似性测度采用欧式距离) 将数据集的 n 个样本聚为 k 类. 依据该聚类结果, 本文方法计算这 k 个聚类之间的边界距离等信息, 对比方法也给出计算或降维可视化的结果. 为体现公平比较, 可视化方法采用通过旋转视角找出的显示类间分离情况最好的三维图. 由于 AP 算法是发现球形 (或凸形) 的聚类, 聚类结果不会出现 2 个聚类相互交织或嵌入的情况, 据此可以判别可视化方法的效果. 当聚类结果 (与已知类标相比较) 的错误率无法准确统计时, 采用测量聚类结果与正确类标一致性的 Fowlkes-Mallows 指标 ($FM \in (0, 1)$)^[29], 当它们完全一致时 $FM=1$. 此外, 第 i 个聚类用 C_i 表示.

第 1 个数据集 Leukemia 是由 3 种病类的 72 个 5000 维样本组成的基因表达数据 (按文献 [30] 的预处理方法从 12582 个基因中选出 5000 个基因构成维数), 其中类 1 和类 2 的可分离性良好, 新发现的类 3 与类 1 和类 2 很靠近^[3]. AP 算法给出 $k=3$ 个聚类的结果的错误率为 9.7%. 本文方法求得各聚类间的远近程度信息为: $C_{1,2}=(1.03, 67.0, 65.1, 63.7)$, 其中数值依次为 C_1 与 C_2 的相对边界距离 $R_{1,2}$ 、绝对边界距离 $D_{1,2}$ 、 C_1 和 C_2 边界区域的疏密程度; 类似地, $C_{1,3}=(1.01, 83.4, 78.5, 82.6)$, $C_{2,3}=(0.98, 74.6, 64.7, 76.1)$. 可以看出, 相对边界距离 $R_{1,2}$ 和 $R_{1,3}$ 均大于 1, 说明类 1 与 2 和 3 分离良好, $C_{2,3}$ 说明类 3 与 2 很靠近且边界区域的疏密度 (64.7 和 76.1) 明显不同, 这些比较准确地反映出 3 种病类的分离情况.

最小距离和 Hausdorff 距离 $dm_{1,2}=60.4$, $dm_{1,3}=71.7$, $dm_{2,3}=71.2$, $dH_{1,2}=101.8$, $dH_{1,3}=113.8$ 和 $dH_{2,3}=110.8$ 均不能说明两个聚类是远离还是靠近的. 可视化方法的结果见图 2(MVU, tSNE 和 KMref 的结果也类似或更差, 略), 图中显示 LLE 和 GTM 结果的 C_1 与 C_2 有交织现象 (这是搜索三维图所有视角后均有交织现象的结论. 后文的类似结论都是搜索三维图所有视角后的结论, 从略), 无法分辨它们之间的远近关系.

第 2 个是关于 18 个基因的 Monocyte 数据集^[31] 的子集 Monocyte-11k. 当从 18 个基因的表达数据中选出表达特征有明显区别的 $k=11$ 个基因 (类) 的 637 个 14 维样本, 则构成该子集. 聚类结果的 $FM=0.895$ 说明各聚类间的分离性比较好. 本文方法能逐一揭示类间远近程度, 例如: $C_{2,5}=(2.43, 2.31, 0.95, 0.78)$, $C_{2,7}=(0.73, 0.70, 0.96, 0.82)$, $C_{5,7}=(0.82, 0.69, 0.84, 0.63)$. 所有 $C_{i,j}$ 信息中, 除 $R_{2,7}=0.73$ 和 $R_{5,7}=0.82$ 表明 C_7 与 C_2 和 C_5 有重叠外, 各聚类之间均有一定的间隔距离 ($R_{i,j}=1.16\sim 5.38$), 分离良好, 这些均与已知信息相一致.

最小距离和 Hausdorff 距离 $dm_{2,5}=2.28$, $dm_{5,7}=0.45$, $dH_{2,5}=4.59$, $dH_{5,7}=4.46$ 等均不能说明两个聚类是远离还是靠近. 可视化方法的结果见图 3(LLE 结果的交织情况很严重, 略); MVU 结果图中 C_2 , C_7 , C_9 , C_{10} 和 C_{11} 交织在一起, GTM 结果图中 C_7 与 C_2 等交织在一起, KMref 结果图中 C_7 , C_{10} 和 C_{11} 等交织在一起, 无法分辨它们之间的远近关系; tSNE 结果图中 C_7 有样本跑到 C_5 和 C_9 处, 它们的远近关系不准确.

第 3 个是由 2945 个基因组成的 15 维 Yeast 数据集, 其 $k=16$ 个类之间重叠情况很普遍^[32]. 本文方法能揭示各类间的远近程度, 例如: $C_{5,12}=(0.62, 1.45, 2.35, 1.96)$, $C_{7,8}=(0.80, 2.54, 2.99, 3.16)$, $C_{4,14}=(1.17, 4.18, 3.20, 3.56)$. 所有的类间距离在 0.62~1.17 之间, 其中 $R_{i,j} > 1$ (类间可分离) 有 12 个, $R_{i,j} \leq 1$ (很靠近或重叠) 有 108 个, 这些均与已知信息相一致.

最小距离和 Hausdorff 距离 $dm_{5,12}=1.06$, $dm_{7,8}=1.70$, $dm_{4,14}=3.14$, $dH_{5,12}=3.98$, $dH_{7,8}=5.25$,

图 2 Leukemia 聚类结果的可视化显示

Figure 2 Visualization of the clustering solution of Leukemia. (a) Result of LLE; (b) result of GTM, number i denotes data points of the i th cluster

图 3 Monocyte 聚类结果的可视化显示

Figure 3 Visualization of the clustering solution of Monocyte. (a) Result of MVU (small figure is the enlargement of right middle part); (b) result of GTM; (c) result of tSNE (small figure is the enlargement of center down); (d) result of KMref (small figure is the enlargement of right up part), number i denotes data points of the i th cluster

$dH_{4,14}=6.02$ 等不能说明两个聚类是远离还是重叠的. GTM 方法和 tSNE 方法的可视化结果见图 4 (LLE, MVU 和 KMref 结果中的交织情况更严重, 略), 图中显示许多聚类叠加交织在一起, 无法分辨它们之间的远近关系.

图 4 Yeast 聚类结果的可视化显示

Figure 4 Visualization of the clustering solution of Yeast. (a) Result of GTM; (b) result of tSNE, number i denotes data points of the i th cluster

图 5 COIL-10 聚类结果的可视化显示

Figure 5 Visualization of the clustering solution of COIL-10. (a) Result of GTM; (b) result of tSNE, number i denotes data points of the i th cluster

第 4 个 COIL-10 数据集是从 COIL-20 数据集^[33]的 20 个物品类中选出差别较大 (分离性较好) 的 $k=10$ 个类组成的子集, 每个类有 72 幅图像, 每一幅 32×32 的图像向量化后形成 1024 的维数. 由于是多角度旋转采样, 某些物品类之间会很靠近或有重叠 (聚类结果的错误率 27.2% 也说明了这点). 本文方法能揭示出各类间的远近程度, 例如 $C_{2,3}=(0.97, 938.9, 914.1, 971.2)$, $C_{4,5}=(0.95, 1159.9, 1164.7, 1226.3)$, $C_{4,6}=(2.01, 1691.8, 817.6, 842.2)$. 10 个聚类之间 $R_{i,j} \leq 1$ (很靠近或重叠) 的有 C_2 与 C_3 、 C_4 与 C_5 , 其余聚类之间的 $R_{i,j}=1.02-6.06$, 表明分离性比较好, 与已知信息相一致.

最小距离和 Hausdorff 距离 $dm_{2,3}=632.4$, $dm_{4,6}=752.7$, $dH_{2,3}=2543.7$, $dH_{4,6}=2880.9$ 等不能说明两个聚类是远离还是靠近的. GTM 方法和 tSNE 方法的可视化结果见图 5 (LLE, MVU 和 KMref 结果的交织情况更严重, 略), GTM 结果图中 C_2 和 C_3 交织在一起, 无法分辨它们之间的远近关系, 而 C_4 分散在多处; tSNE 结果图中 C_4 被拆分到 2 处 (C_5 也类似), 无法判别 C_4 和 C_5 与其他聚类的远近关系.

4 讨论

本文方法的类间距离是综合了两个聚类的最近样本集之间的绝对距离 D_{AB} 和聚类边界区域的疏密程度后的相对距离 R_{AB} , 能准确反映不同聚类边界之间的远近关系. 引言中介绍的类间距离与本文方法最大的不同是没有考虑聚类本身的疏密程度, 而类间绝对距离是不能说明两个聚类是远离还是靠近的. 例如, 当聚类的边界区域内样本之间的疏密程度 $D_{A_1} = D_{B_1} = 9$ 时, 类间距离 $D_{AB} = 10$ 表明两个聚类 A 和 B 很靠近; 而当 $D_{A_1} = D_{B_1} = 1$ 时, $D_{AB} = 10$ 表明两个聚类相距较远.

对于低维数据 (例如 4 或 5 维), 降维可视化可以在二维或三维空间近似地显示出聚类间的相对远近关系和数据点的相对分布情况. 对于高维数据, 这种方法会受到高维空间的维数灾难 (curse of dimensionality) 问题的影响而失效. 维数灾难的一个方面是距离集中 (现象), 即高维数据的任两点之间的距离具有几乎相等的倾向^[23], 具体地说, 一个高维数据集的数据点之间的距离集合具有集中在该集合的距离均值附近或一定范围内的倾向. 有关研究^[34,35] 显示, 在 10 维以上就会出现显著的维数灾难现象.

数据点在高维空间的几何分布与低维时具有显著差别^[34,35], 例如高维数据点都近似位于以数据集的均值为中心的超球上^[23]. 由高维数据的这种特点和距离集中现象可知: 在高维空间中不同聚类的样本点之间的距离都相差不大, 各个聚类倾向于彼此靠近, 类间距离都相差不大或倾向于相同. 那么对于高维空间中存在聚类间距趋同现象且类数较多的聚类结构, 为什么本文方法能够而降维可视化方法却通常不能准确揭示这样的聚类结构?

本文方法是在原始高维空间中测量聚类之间的远近程度, 没有涉及数据的映射或变换; 本文方法的类间距离是在欧式距离基础上设计的, 与聚类算法的测度 (欧式距离) 相一致. 因此, 准确反映出原始高维空间中聚类的远近关系 (包括聚类间距趋同情况) 是本文方法的自然优势. 第 3 节实验结果也验证了本文方法的准确性, 例如: Leukemia 数据集的 3 种病类很靠近, 本文方法给出的相对类间距离 (1.03, 1.01 和 0.98) 就倾向为 1, 表明聚类之间很靠近或几乎重叠.

降维可视化方法的关键是将高维数据映射为低维数据后, 在低维 (二维或三维) 空间显示聚类之间的远近关系. 对于高维数据 (考虑其固有维数也是高维、数据量不是很少的一般情况) 且类数较多 (如 10 个以上) 的情况, 降维可视化方法的不足主要表现在:

(1) 常常错误显示类间的远近关系: 高维空间常常存在聚类间距趋同的现象, 而二维平面上只能最多准确布局 3 个等间距的聚类 (三维空间是 4 个); 于是, 对于 5 个以上 (近似) 等间距的聚类, 其可视化布局要么出现错误的聚类重叠或交织现象, 要么出现一些聚类被不真实地拉远的情况. 一般类数越多, 这种可视化布局的错误会越严重.

(2) 不能准确显示数据的分布情况: 距离集中现象说明许多高维数据点之间的距离相差不大, 而二维和三维空间是无法准确布局众多近似等间距的数据点, 这必然导致许多数据点的错误布局.

(3) 不能准确显示原高维空间中的类间距离: 高维数据映射为低维数据后, 数据点之间的距离已经不是原高维空间中的距离, 自然显示的类间距离也不是原高维空间中的类间距离.

本文实验结果也反映了降维可视化方法的上述不足, 例如对于 Monocyte 数据集, 对比的可视化方法 (除了 tSNE) 都出现了错误的多个类的交织现象, 而 tSNE 结果中 C_7 的样本被不真实地拉远到 C_5 和 C_9 处.

总之, 本文方法能有效地识别聚类之间的远近关系, 实验结果与已知的聚类结构信息相一致也验证了这点. 而在高维数据且类数较多的情况下, 降维可视化方法通常会失效, 现有的类间距离方法则不能反映聚类之间的远近关系.

5 结论

本文提出了双几何体模型方法来获得任两个聚类之间的远近程度信息, 包括聚类之间的相对边界距离、绝对边界距离和聚类边界区域的疏密程度. 本文方法的突出优点, 一是在高维数据和大类数的复杂情况下同样能准确揭示各聚类间的远近关系, 二是给出的类间距离反映的是原始高维空间中聚类之间的远近程度. 对于这种复杂情况, 降维可视化方法由于不能准确布局 5 个以上 (近似) 等间距的聚类, 常常错误显示类间的远近关系; 现有的类间距离方法则不能反映聚类之间的远近关系. 此外, 本文方法是针对划分聚类方法和距离测度设计的, 故只适用于凸形的聚类情况, 不适用于那些不规则聚类形状的数据集 (例如有包含关系的圆环形聚类和相互嵌套的聚类).

参考文献

- 1 Xu R, Wunsch II D C. Survey of clustering algorithms. *IEEE Trans Neural Netw*, 2005, 16: 645–678
- 2 Frey B J, Dueck D. Clustering by passing messages between data points. *Science*, 2007, 315: 972–976
- 3 Armstrong S A, Staunton J E, Silverman L B, et al. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat Genet*, 2002, 30: 41–47
- 4 Han J, Kamber M. *Data Mining: Concepts and Techniques*. 2nd ed. San Francisco: Morgan Kaufmann Publishers, 2006. 300–301
- 5 Bolshakova N, Azuaje F. Cluster validation techniques for genome expression data. *Signal Process*, 2003, 83: 825–833
- 6 Wua K P, Wang S D. Choosing the kernel parameters for support vector machines by the inter-cluster distance in the feature space. *Pattern Recogn*, 2009, 42: 710–717
- 7 Yin F, Liu C L. Handwritten Chinese text line segmentation by clustering with distance metric learning. *Pattern Recogn*, 2009, 42: 3146–3157
- 8 Shamir R, Maron-Katz A, Tanay A, et al. EXPANDER—an integrative program suite for microarray data analysis. *BMC Bioinformatics*, 2005, 6: 232
- 9 Ren Y G. Study on data visualization methods and related techniques for clustering. Dissertation for PhD Degree. Shenyang: Northeastern University, 2006 [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]
- 10 Zhan D C, Zhou Z H. Ensemble-based manifold learning for visualization. *J Comput Res Develop*, 2005, 42: 1533–1537 [詹德川, 周志华. 基于集成的流形学习可视化. *计算机研究与发展*, 2005, 42: 1533–1537]
- 11 Sun M M. Study on theories and algorithms in manifold learning. Dissertation for Ph.D. Degree. Nanjing: Nanjing University of Science and Technology, 2007 [孙明明. 流形学习理论与算法研究. 博士学位论文. 南京: 南京理工大学, 2007]
- 12 Roweis S T, Saul L K. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 2000, 290: 2323–2326
- 13 Weinberger K Q, Sha F, Saul L K. Learning a kernel matrix for nonlinear dimensionality reduction. In: *Proceedings of the 21st International Conference on Machine Learning*, Banff, 2004. 839–846
- 14 Van der Maaten L J P, Hinton G E. Visualizing high-dimensional data using t-SNE. *J Mach Learn Res*, 2008, 9: 2579–2605
- 15 Suykens J A K. Data visualization and dimensionality reduction using kernel maps with a reference point. *IEEE Trans Neural Netw*, 2008, 19: 1501–1517
- 16 Bishop C, Svensen M, Williams C. GTM: the generative topographic mapping. *Neural Comput*, 1998, 10: 215–234
- 17 Tino P, Nabney I. Hierarchical GTM: constructing localized nonlinear projection manifolds in a principled way. *IEEE Trans Pattern Anal Mach Intell*, 2002, 24: 639–656
- 18 Yin H. ViSOM—a novel method for multivariate data projection and structure visualisation. *IEEE Trans Neural Netw*, 2002, 13: 237–243
- 19 Wu S, Chow T. PRSOM: A new visualization method by hybridizing multidimensional scaling and self-organizing map. *IEEE Trans Neural Netw*, 2005, 16: 1362–1380

- 20 Wei H L, Billings S A. Feature subset selection and ranking for data dimensionality reduction. *IEEE Trans Pattern Anal Mach Intell*, 2007, 29: 162–166
- 21 Datta S, Datta S. Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics*, 2003, 19: 459–466
- 22 Wang K J. Geometric double-entity model for clustering analysis. [Http://www.mathworks.com/matlabcentral/fileexchange/authors/24811](http://www.mathworks.com/matlabcentral/fileexchange/authors/24811)
- 23 Radovanovic M, Nanopoulos A, Ivanovic M. Nearest neighbors in high-dimensional data: the emergence and influence of hubs. In: *Proceedings of the 26th Annual International Conference on Machine Learning*, Montreal, Quebec, 2009. 865–872
- 24 Gong G L. *Probability Theory and Statistics*. Beijing: Tsinghua University Press, 2006 [龚光鲁. 概率论与数理统计. 北京: 清华大学出版社, 2006]
- 25 Abdi H, Molin P. Lilliefors test of normality. In: Salkind N J, ed. *Encyclopedia of Measurement and Statistics*. Thousand Oaks: Sage Publications, Inc., 2007
- 26 Walpole R E, Myers R H, Myers S L, et al. *Probability and Statistics for Engineers and Scientists*. 8th ed. Upper Saddle River: Pearson Education, Inc., 2006
- 27 Black K. *Business Statistics: Contemporary Decision Making*. 6th ed. Hoboken: John Wiley & Sons, Inc., 2010
- 28 Conover W J. *Practical Nonparametric Statistics*. 3rd ed. Beijing: Posts & Telecom Press, 2006 [Conover W J. 崔恒建, 译. 实用非参数统计. 第 3 版. 北京: 人民邮电出版社, 2006]
- 29 Wang K J, Zhang J Y, Li D, et al. Adaptive affinity propagation clustering. *Act Automat Sin*, 2007, 33: 1242–1246 [王开军, 张军英, 李丹, 等. 自适应仿射传播聚类. 自动化学报, 2007, 33: 1242–1246]
- 30 Golub T R, Slonim D K, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 1999, 286: 531–537
- 31 Hartuv E, Schmitt A, Lange J, et al. An algorithm for clustering cDNAs for gene expression analysis. *Genomics*, 2000, 66: 249–256
- 32 Dembélé D, Kastner P. Fuzzy C-means method for clustering microarray data. *Bioinformatics* 2003, 19: 973–980
- 33 Nene S A, Nayar S K, Murase H. Columbia Object Image Library (COIL-20). Technical Report CUCS-005-96. Columbia University, 1996
- 34 Clarke R, Ransom H, Wang A, et al. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nat Rev Cancer*, 2008, 8: 37–49
- 35 Verleysen M, François D. The curse of dimensionality in data mining and time series prediction. In: Cabestany J, Prieto A, Sandoval D F, eds. *Computational Intelligence and Bioinspired Systems*. Berlin: Springer, 2005. 758–770

Geometric double-entity model for recognizing far-near relations of clusters

WANG KaiJun*, YAN XuanHui & CHEN LiFei

School of Mathematics and Computer Science, Fujian Normal University, Fuzhou 350108, China

*E-mail: wkjwang@gmail.com

Abstract When solving many practical problems, we not only need sample labels given by a clustering algorithm, but also rely on the recognition of far-near relations of clusters. Under the difficult condition of many clusters in a high-dimensional data set, the clustering visualization methods based on dimension reductions usually produce the phenomena, e.g., some clusters are overlapping, interlacing, or pushed away; as a result, the far-near relations of some clusters are displayed wrongly or cannot be distinguished. The existing inter-cluster distance methods cannot determine whether two clusters are far away or near. The geometric double-entity

model method (GDEM) is proposed to describe far-near relations of clusters, and the methods such as the relative border distance, absolute border distance and region dense degree are designed to measure far-near degrees between clusters. GDEM pays attention to both the absolute distance between nearest sample sets and the dense degrees of border regions of two clusters, and it is able to uncover accurately far-near relations of clusters in a high-dimensional space, especially under the difficult condition mentioned above. The experimental results on four real data sets show that the proposed method can effectively recognize far-near relations of clusters, while the conventional methods cannot.

Keywords geometric double-entity model, far-near relations of clusters, many clusters, high-dimensional dataset, partitional clustering algorithms

WANG KaiJun was born in 1965. He received the Ph.D. degree in computer science from Xidian University, Xi'an in 2008. Currently, he is an associate professor in School of Mathematics and Computer Science, Fujian Normal University, Fuzhou. His research interests include artificial intelligence, pattern recognition and data mining.