基于矩阵模型的高维聚类边界模式发现

李向丽 [1] 曹晓锋 [1] 邱保志 [1]

摘 要 流形学习关注于寻找合适的嵌入方式将高维空间映射至低维空间, 但映射子空间依然可能具有较高的维度, 难以解决高维空间的数据挖掘任务. 本文建立一种简单的矩阵模型判断数据点 k 近邻空间关于该点的对称性, 并使用对称率进行边界提取, 提出一种基于矩阵模型的高维聚类边界检测技术 (Clustering boundary detection based on matrix model, MMC). 该模型构造简单、直接、易于理解和使用. 理论分析以及在人工合成和真实数据集的实验结果表明 MMC 算法能够有效地检测出低维和高维空间的聚类边界.

关键词 高维空间, 聚类边界, 矩阵模型, k 近邻, 对称率

引用格式 李向丽, 曹晓锋, 邱保志. 基于矩阵模型的高维聚类边界模式发现. 自动化学报, 2017, **43**(11): 1962−1972

DOI 10.16383/j.aas.2017.c160443

Clustering Boundary Pattern Discovery for High Dimensional Space Base on Matrix Model

LI Xiang-Li[1] CAO Xiao-Feng[1] QIU Bao-Zhi[1]

Abstract Manifold learning aims to find a reasonable embed mode to map a high-dimensional space to a low dimensional space. However, the dimension of the latter may still be so high that any data mining task cannot be effectively finished. This paper proposes a simple matrix model to judge the symmetry of data object and its k nearest neighbors space, and use the symmetry rate to extract the clustering boundary. Finally, the MMC algorithm is developed. Theoretical analysis and experimental results show that the MMC can effectively detect the clustering boundary of low and high dimensional spaces.

Key words High dimensional space, clustering boundary, martin model, k nearest neighbors, symmetry rate

Citation Li Xiang-Li, Cao Xiao-Feng, Qiu Bao-Zhi. Clustering boundary pattern discovery for high dimensional space base on matrix model. *Acta Automatica Sinica*, 2017, **43**(11): 1962−1972

数据挖掘是从大量未知数据中提取有价值的潜在知识的过程, 这些知识有助于人们更好地认识和理解数据. 聚类分析[1−5] 旨在挖掘数据内部的非平凡规律, 提取空间中自然存在的模式, 在图像分割[6−7]、生物学[8] 和医学[9] 等领域获得了广泛应用. 与聚类不同, 聚类边界是另一种新颖的具有重要价值的特殊模式. 这些数据分布于聚类的边缘, 具有明确的类归属, 但又与类内其他数据存在一定差异. 从认知事物的角度而言, 聚类边界代表着那些类归属明确但又具备脱离本类属性特征的一类特殊数据. 例如, 在医学检测中, 正常人群的聚类边界代表那些携带病毒但未患病的人群; 手写体图像的聚类边界代表着那些异于常态和不易识别的手写体图像. 随着聚类边界应用领域的不断扩大, 必然要求对聚类边界提取技术进行深入的研究.

收稿日期 2016-05-31 录用日期 2016-11-16
Manuscript received May 31, 2016; accepted November 16, 2016
河南省基础与前沿技术研究项目 (152300410191) 资助
Supported by Basic and Advanced Technology Research Project of Henan Province (152300410191)
本文责任编委 张军平
Recommended by Associate Editor ZHANG Jun-Ping
1. 郑州大学信息工程学院 郑州 450001
1. School of Information Engineering, Zhengzhou University, Zhengzhou 450001

现有的聚类和聚类边界检测技术普遍采用构造几何模型的方式对数据空间进行合理分割, 以获取空间数据的有效模式结构, 而良好的几何理论为这些研究提供了坚实基础. 但随着应用领域的不断扩展, 数据的维度迅速增长, 达到了千维、万维甚至更高, 而低维几何性质在高维空间[10−11] 内的时间性能却大幅下降. 因此, 如何打开高维数据的 “海岸围墙” 成为数据挖掘研究者亟待解决的问题, 更是聚类边界研究者所面临的严峻挑战.

本文第 1 节介绍了聚类边界的相关工作; 第 2 节介绍了矩阵模型, 提出了边界检测算法; 第 3 节是本文的实验部分, 包括在人工合成数据集和真实数据集的实验对比和分析; 第 4 节对算法的时间复杂度和参数分析; 在第 5 节给出了本文的结论.

1 相关工作

自 Ester 等[12] 提出了聚类边界的概念之后, Xia[13] 等基于聚类的边界对象的反向 k 近邻[14] 个数小于聚类内部对象的反向 k 近邻的个数这一事实提出了聚类边界算法 BORDER. 但是该算法忽视了噪声比聚类边界对象具有更少的反向 k 近邻个数这一特征, 所以其聚类边界检测结果包含了全部噪

声. 为了解决这一问题, BRIM[15] 依据边界点所在的正负半邻域内数据分布不均匀而聚类内部点的邻域内数据分布近似均匀这一事实, 有效地区分聚类的边界对象和噪声, 但易受聚类附近的噪声影响, 特别是对于变化密度和多密度的聚类, 该算法不能准确提取聚类的边界. 为此, BAND[16] 算法基于数据对象的变异系数提取聚类边界, 虽然该算法能提取变化密度和多密度数据集的聚类边界, 但由于靠近聚类边界的噪声的变异系数值可能与边界数据相同, 导致该算法将边界附近的噪声误认为聚类的边界. BRINK[17] 使用加权欧氏距离度量数据对象间的相似性, 并根据聚类边界的局部特征提出了局部质变因子的概念进行边界检测, 但随着数据维度的增加, 数据分布逐渐稀疏, 以欧氏距离作为数据对象之间的相似性度量方式逐渐失去意义, 所以 BRINK 算法不能有效提取高维数据的聚类边界. BERGE[18] 算法使用证据积累的思想检测混合属性数据集聚类边界, 通过多次统计学习进行边界对象标记, 但错误的标记结果将影响随后的迭代学习, 导致错误率快速升高.

解决高维空间的数据分析任务的方法可归纳为两类: 维度约简和空间变换. 例如, 文献 [19−23] 是高维空间数据处理的维度约简技术, 该类技术属于流形学习范畴, 关注于如何选取合适的子空间代表完整的数据空间, 使数据分析的任务在维度相对较低的子空间内完成. 但维度约简可能造成空间信息结构的损失, 并且不同的约简策略得到的子空间可能不同, 所以数据分析的结果过分依赖于约简后的子空间, 而寻找合适的约简方法又是一个相对复杂的问题. 另一种常见的高维空间处理方法是空间变换, 即将原空间通过合适的技术转换为新的空间, 如光谱聚类[24]、神经网络[25] 和支持向量机 (Support vector machine, SVM)[26] 等. 这几种方式大都关注于几何形态的构造或融入部分代数模型思想, 并广为人知.

空间几何的研究多关注于二维和三维空间内的空间变换, 对高维空间的研究较少. 这是因为高维空间具有较强的抽象性, 空间的几何形状和性质十分抽象, 空间变换难度较大, 所以很难对高维空间内的几何变换做出详尽和具体的描述, 这也是高维空间数据分析面临的难题. 但是所有的空间变换可由矩阵的初等变换表示, 例如空间的平移、旋转、缩放、侧切和错切等均可以由一定的仿射运算或矩阵运算表示. 矩阵是一种基本的代数模型, 研究者们也对矩阵的结构和性质十分了解, 其属于一种十分平凡的代数模型. 本文将脱离复杂的高维空间几何模型背景, 以矩阵模型为基础分析高维空间, 通过建立一种合理的矩阵模型判断数据点的 k 近邻空间关于该点的对称性, 使用对称率进行边界检测. 本文的创新点如下: 1) 提出了面向维度的高维空间处理思想; 2) 提出了判别数据点在其近邻空间中的对称概率模型; 3) 提出了一种提取高维空间聚类边界模式的 MMC 算法.

2 MMC 框架

2.1 动机

DBSCAN 算法是一种基于密度的聚类算法, 以空间某点为始发点进行高密度区域搜索, 直到搜索至低密度区域时停止搜索, 并转而搜索其他区域. 显然, 该搜索始于聚类内部, 并止步于聚类边界处. 而判断这一搜索过程的关键技术为密度识别. 据此, 可以认同的是边界点位于聚类的边缘处, 其近邻对象多分布于聚类内部, 另一侧为噪声. 而在无噪声聚类中, 另一侧无任何数据点分布. 对于聚类内部的数据点, 其近邻对象均匀的环绕着该点, 该邻域空间关于中心点呈现较强的对称性. 而这种对称性是一种多维对称性, 即在每个维度上该点总是位于一维射影空间的中心位置. 相反, 如果该点在每一个射影维度上均呈现出非对称性, 则该高维近邻空间必然不会关于该点对称, 所以如果高维空间关于空间内某点是标准对称的, 则每一个射影空间依然关于该点的射影位置对称. 因此, 提出了利用空间关于点的对称性的思想检测聚类边界. 显然, 聚类内部点的空间对称性较高, 边界点次之, 噪声最低. 本质上, 该方法属于密度倾斜式判断方法, 通过每一个维度内的数据分布状况判断空间整体关于某点的对称性, 称之为面向维度. 已提出的聚类和聚类边界算法多从空间整体分布形态进行模式检测和识别, 例如密度分布、统计规律和几何形态等, 而本思想将从一维射影空间角度分析高维数据空间分布特征.

2.2 MMC 算法

给定规模为 $1 \times m$ 的矩阵 $A = [e_1, e_2, e_3, \cdots, e_m]$, 若 A 关于元素 e_i 对称, 则需满足

$$Less(A, e_i) - More(A, e_i) = 0 \tag{1}$$

其中, $Less(Arry, Elem)$ 表示统计数组 $Arry$ 中小于 $Elem$ 的元素个数, $More(Arry, Elem)$ 表示统计数组 $Arry$ 中大于 $Elem$ 的元素个数 (以下 $Less$ 和 $More$ 函数均表示此意). 式 (1) 通过元素在矩阵中的位置判断矩阵关于该元素的对称性, 若矩阵关于元素标准对称, 则式 (1) 必成立, 而 $|Less(A, e_i) - More(A, e_i)|$ 的值也反映了矩阵关于元素的对称性. 将一维矩阵推广至规模为 $n \times m$ (n 为样本量, m 为维度) 的高维空间 S, 则该空间可以由规模为 $n \times m$ 的矩阵 Mar 描述:

$$Mar=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1j} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2j} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{ij} & \cdots & a_{im} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nj} & \cdots & a_{nm} \end{bmatrix} \tag{2}$$

已知空间中某点 $x_i=[x_{i1},x_{i2},\cdots,x_{im}]$, 若空间 S 关于 x_i 对称, 则该空间在每一维度上关于 x_i 的射影位置对称, 即满足如下条件:

$$\begin{cases} Less(Mar_1,x_{i1})-More(Mar_1,x_{i1})=0 \\ Less(Mar_2,x_{i2})-More(Mar_2,x_{i2})=0 \\ Less(Mar_3,x_{i3})-More(Mar_3,x_{i3})=0 \\ \qquad\qquad\vdots \\ Less(Mar_m,x_{im})-More(Mar_m,x_{im})=0 \end{cases} \tag{3}$$

其中, Mar_m 表示矩阵 Mar 的第 m 列, 其中 x_{ij} 表示 x_i 的第 j 维. 为了更好地表示空间 S 关于 x_i 对称性, 提出了对称率的概念:

$$P_{x_i}=\sum_{j=1}^{m}\frac{n-|Less(Mar_j,x_{ij})-More(Mar_j,x_{ij})|}{n} \tag{4}$$

对称率反映了矩阵关于点的对称性, 该比率越高, 说明该矩阵关于 x_i 对称的可能性越大. 在高维空间中, 常用的采样方式有超球采样、超立方体采样和 k 近邻采样. 高维空间内数据分布相对稀疏, 且大部分数据分布于空间的表层, 而超立方体采样和超球采样使用固定尺寸的采样窗口进行局部特征抓取, 背离空间的另一侧采样空间内可能包含较少的数据点, 甚至无数据点, 这造成了采样的偏见性, 不能完整地反映局部空间的分布特征. 而 k 近邻采样总是提取 k 个数据对象构成的局部空间作为动态采样空间, 可以更好地反映局部空间的分布特征. 核心点的 k 近邻对象均匀的分布在其周围, 即 k 近邻对象关于核心点的对称性较强; 边界点的 k 近邻对象分布不均匀, 近邻空间中的对象关于边界点的对称性较弱; 而噪声的 k 近邻对象关于噪声的分布可能是均匀也可能不均匀, 不能简单地从对称性上进行判别. 基于上述分析, 将高维空间 (m 维) 中的点 x_i 的近邻空间模拟为 m 个矩阵 $S_1,S_2,\cdots,S_m$, 则该高维空间关于 x_i 的对称性转换为矩阵 S_1 关于 x_{i1}, 矩阵 S_2 关于 $x_{i2},\cdots$, 矩阵 S_m 关于 x_{im} 的对称性问题, 式 (5) 给出了空间对称率的定义:

$$SP(x_i)=\sum_{j=1}^{m}\frac{k-|Less(S_j,x_{ij})-More(S_j,x_{ij})|}{k} \tag{5}$$

由式 (5) 可知, 核心点具有较小的 SP 值, 而边界点具有较大的 SP 值. 由于噪声的 k 近邻空间复杂性, 仅依据 SP 值的大小进行边界提取是不准确的. 但噪声与其 k 近邻对象的平均距离远大于核心点和边界点. 为了将噪声准确分离, 需要对该距离进行离散化, 以使噪声快速分离. 常见的离散函数及其一阶偏导数如下:

$$\begin{cases} y=x \\ y=x^2 \\ y=\lg(x) \\ y=\mathrm{e}^x \end{cases} \tag{6}$$

$$\begin{cases} \dfrac{\partial y}{\partial x}=1 \\ \dfrac{\partial y}{\partial x}=2x \\ \dfrac{\partial y}{\partial x}=\dfrac{1}{x} \\ \dfrac{\partial y}{\partial x}=\mathrm{e}^x \end{cases} \tag{7}$$

上述函数中, 随着 x 逐渐增大, 变化速度增快, 由此使用该函数进行辅助离散化, 旨在增强噪声与其余数据点的差异. 在式 (5) 中引入 $\exp(\sum_{j=1}^{k}\mathrm{dist}(x_i,c_{ij})/k)$ 进行离散化:

$$HP(x_i)=\exp\left(\sum_{j=1}^{k}\frac{\mathrm{dist}(x_i,c_{ij})}{k}\right)\times \sum_{j=1}^{m}\frac{k-|Less(S_j,x_{ij})-More(S_j,x_{ij})|}{k} \tag{8}$$

其中, $dist(x_i,c_{ij})$ 表示 x_i 与 c_{ij} 之间的欧氏距离, c_{ij} 是 x_i 的第 j 个近邻. 经过离散化, 噪声的 HP 值迅速增大, 远远大于边界点和核心点的 HP 值, 使得边界提取对噪声的敏感性迅速下降. 经过上述分析, 本文提出了一种提取高维空间聚类边界的 MMC 算法, 该算法相对于传统的聚类边界检测较为简单, 仅通过计算每个数据点的 HP 值就能提取聚类边界. 算法描述如下.

算法 1. MMC 算法

输入. 数据集 X, 近邻对象数 k, 边界起始阈值 ε_1, 边界结束阈值 ε_2.

输出. 边界集合 $Bundy$.

步骤 1. 计算每个数据对象的 k 近邻集合.

步骤 2. 按照式 (8) 计算每个数据的 HP 值, 并存储在数组 a.

步骤 3. 对 a 进行升序排序, 获取每个数据对象的排序编号, 存入数组 ID 中.

步骤 4. 遍历每个数据对象, 如果 $\varepsilon_1 < ID(i)/num < \varepsilon_2$, 将 x_i 存入 $Bundy(i)$.

在该算法中, num 是数据集样本数目, ε_1 和 ε_2 是输入参数, 控制着聚类边界的提取范围, 并满足以下定义:

1) $ID(i)/num < \varepsilon_1$, 该数据点为核心点;

2) $\varepsilon_1 < ID(i)/num < \varepsilon_2$, 该数据点为边界点;

3) $ID(i)/num > \varepsilon_2$, 该数据点为噪声.

所以, MMC 算法根据参数 ε_1 和 ε_2 就可识别出每个数据对象的类型. 需要指出的是, 对无噪数据集 $\varepsilon_2 = 1$, 只需输入 ε_1 就可以进行聚类边界检测.

3 实验与结果分析

为了验证 MMC 算法的聚类边界检测能力. 本文首先在人工合成的二维数值数据集上进行实验, 并与 BORDER 和 BRIM 等算法进行对比; 然后在多个医学数据集上进行聚类边界检测实验; 最后在手写体和多姿态人脸图像上进行边界检测实验, 并给出相关的实验分析.

算法的实验环境: CPU 为 Pentium(R) Dual-Core CPU E6700@3.20 GHz, 内存为 2.00 GB, 操作系统为 Microsoft Windows 7, 算法编写环境为 MATLAB 7.0. 实验数据集的基本信息和使用的预处理方式见表 1. 具体的预处理方式如下:

1) 每个数据对象的值除以 10^3;

2) 每个数据对象的值除以 10^4;

3) 对每张图像的灰度矩阵 (规格为 $n \times m$) 的每一维求均值, 转换成 $1 \times m$ 的矩阵, 用以表征该图像.

表 1 预处理方式

Table 1 Pretreatment methods

数据集	样本总数	维数	预处理方式
Mnist	10 000	28	3)
Colon	62	2 000	1)
Prostate	102	10 509	2)
Pointing data	2 790	384	3)

由于仅需分析检测结果的数据对象是否为真正的边界对象, 所以关注于检测结果的精度和召回率, 本文使用 F-measure 评价聚类边界检测质量, 相关定义如下:

$$\text{准确率} = \frac{\text{检测正确边界数}}{\text{检测边界数}}$$

$$\text{召回率} = \frac{\text{检测正确边界数}}{\text{真实边界数}}$$

$$\text{F-measure} = \frac{2}{\dfrac{1}{\text{准确率}} + \dfrac{1}{\text{召回率}}}$$

准确率说明算法检测聚类边界的准确程度, 准确率越小, 说明算法检测聚类边界的能力越弱; 召回率是聚类边界检测结果完整性的反映; 准确率和召回率存在相互制约的关系, F-measure 值越大, 说明算法检测聚类边界的能力越强.

3.1 合成数据集边界检测

图 1 (a) ∼ 1 (d) 是含有噪声的具有多密度聚类簇的人工合成数据集 Syn1, Syn2, Syn3 和 Syn4, 分别包含 7 832, 5 400, 5 400, 4 600 个数据点; 聚类簇数分别为 2, 5, 2, 2. 其中, Syn3 数据集的噪声紧密的分布在聚类边界处; Syn4 数据集的圆形聚类嵌套在环形聚类内, 且大量噪声均匀的分布在两个聚类之间. 为进一步分析各算法的聚类边界检测性能, 使用 DBSCAN 算法进行多次统计实验, 分别为上述四个人工合成数据集标记了 640, 538, 1 077, 1 204 个聚类边界点, 并使用 F-measure 值评价各算法的聚类最优聚类边界检测结果 (详见表 2). 作为示例, 图 1 (e) ∼ 1 (p) 呈现了部分算法的最优边界检测结果, 并给出了具体的使用参数.

从图 1 可以看出, 由于噪声比边界点具有更少的反向 k 近邻个数, BORDER 算法的检测结果含有全部噪声; BAND 算法使用变异系数进行边界检测, 但是在多噪声数据集中, 分布在边界附近的噪声的邻域分布与边界点相似, 所以该算法的检测结果包含了边界附近的噪声; BRIM 算法使用一条穿过邻域质心的直径将参考点的圆形邻域划分成两个区域, 通过对比区域之间的数据点分布判断参考点是否为边界点. 但当邻域内分布大量噪声时, 质心可能与圆心相距较近甚至重叠, 导致切割平面不一定为最优密度切割平面. 并且, 在高维空间中, 需要使用高维切割平面进行空间分割, 而一条直线仅能贯穿该空间, 不能进行切割, 所以该方法不能适用于高维空间. 因此, 后续实验将不再与 BRIM 算法进行对比. BERGE 算法使用证据积累思想多次统计和标记边界点, 但在循环标记中需要消除部分噪声点, 而算法一旦将边界点误认为噪声后, 该错误标记将影响后续循环学习. 因此, 该算法依然对噪声存在着一定的敏感性. MMC 算法将二维平面转换成有穷矩阵, 以面向维度的角度分析每个维度关于中心点的对称性问题, 并通过中心点与近邻对象之间的距离之和对噪声进行进一步离散化, 所以该算法可以有效地区分聚类的边界和噪声. 而表 2 的算法检测结果分析进一步验证了 MMC 算法更为精准的边界检

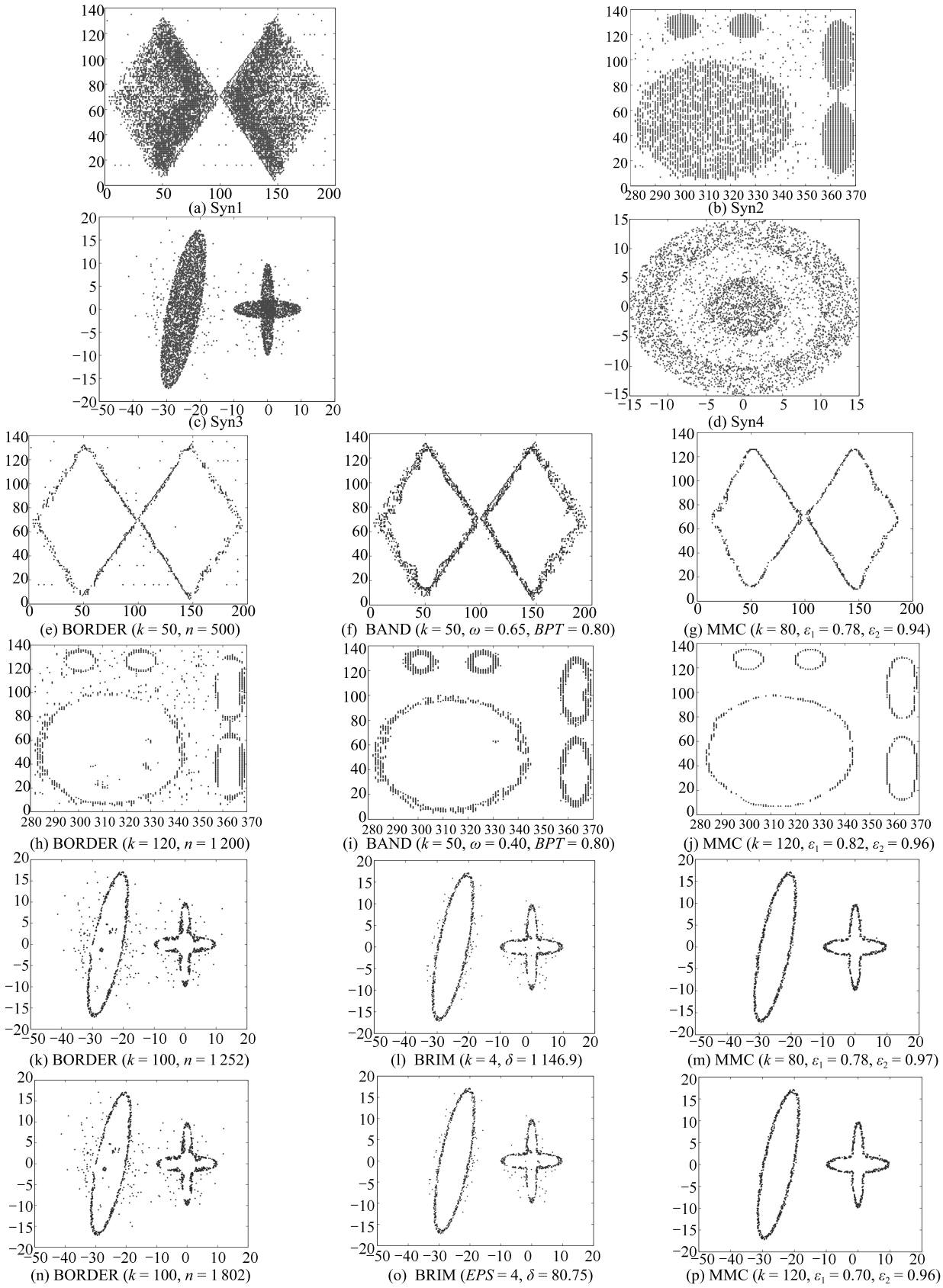


图 1 不同算法在人工合成数据集上的最佳聚类边界检测结果

Fig. 1 The best clustering boundary detection results of different algorithms on synthetic datasets

表 2 不同算法在不同数据集上聚类边界检测结果

Table 2 The clustering boundary detection results of different algorithms on different data sets

数据集	维度	算法	真实边界数	检测边界数	检测正确边界数	准确率	召回率	F-measure
DS1	2	BAND	640	823	556	0.6756	0.8688	0.7601
		BORDER		723	540	0.7469	0.8438	0.7924
		BRINK		667	520	0.7795	0.8125	0.7957
		BRIM		680	536	0.7882	0.8375	0.8121
		BERGE		662	532	0.8036	0.8313	0.8172
		MMC		630	576	0.9143	0.9000	**0.9071**
DS2	2	BAND	538	749	454	0.6061	0.8439	0.7055
		BORDER		669	445	0.6366	0.8271	0.7195
		BRINK		499	438	0.8778	0.8141	0.8447
		BRIM		562	466	0.8292	0.8661	0.8472
		BERGE		553	472	0.8535	0.8773	0.8652
		MMC		549	503	0.9162	0.9349	**0.9255**
DS3	2	BAND	1 077	1 629	961	0.5899	0.8923	0.7103
		BORDER		1 252	831	0.6637	0.7716	0.7136
		BRINK		1 540	914	0.5935	0.8478	0.6985
		BRIM		1 188	935	0.7870	0.8682	0.8256
		BERGE		1 138	942	0.8278	0.8747	0.8506
		MMC		1 016	968	0.9528	0.8988	**0.9250**
DS4	2	BAND	1 204	1 944	1 056	0.5432	0.8771	0.6709
		BORDER		1 802	1 089	0.6043	0.9045	0.7246
		BRINK		1 817	1 003	0.5520	0.8331	0.6640
		BRIM		1 355	1 062	0.7838	0.8821	0.8300
		BERGE		1 246	1 123	0.9013	0.9327	0.9167
		MMC		1 228	1 138	0.9267	0.9452	**0.9359**
Biomed	4	BAND	30	26	22	0.8462	0.7333	0.7857
		BORDER		26	23	0.8846	0.7667	0.8214
		BRINK		36	30	0.8333	1.0000	0.9089
		BERGE		26	24	0.9231	0.8000	0.8572
		MMC		30	28	0.9333	0.9333	**0.9333**
Cancer	10	BAND	37	37	25	0.6757	0.6757	0.6757
		BORDER		37	28	0.7568	0.7568	0.7568
		BRINK		37	29	0.7837	0.7837	0.7837
		BERGE		37	28	0.7568	0.7568	0.7568
		MMC		38	34	0.8947	0.9189	**0.9067**
Colon	2 000	BAND	7	6	5	0.8333	0.7143	0.7692
		BORDER		7	7	1.0000	1.0000	**1.0000**
		BRINK		6	5	0.8333	0.7143	0.7692
		BERGE		6	5	0.8333	0.7143	0.7692
		MMC		7	7	1.0000	1.0000	**1.0000**
Prostate	10 509	BAND	18	17	16	0.9412	0.8889	0.9143
		BORDER		19	18	0.9474	1.0000	0.9730
		BRINK		17	16	0.9412	0.8889	0.9143
		BERGE		17	16	0.9412	0.8889	0.9143
		MMC		18	18	1.0000	1.0000	**1.0000**

测性能.

3.2 医学数据集聚类边界检测

医学数据集中, 正常人群的聚类边界代表着那些携带某种病毒但未患病的人群. 通过对这些个体的识别, 可以有效地进行预防性治疗. 基因表达谱数据[27−28] 中的聚类边界代表着携带某些隐形基因传染病或者可能发生基因突变的个体, 对这些特殊个体的研究不仅可以有效预防和治疗携带基因遗传病的个体, 也可以获取物种可能发生突变的基因片段, 以方便进一步对物种演化进行研究.

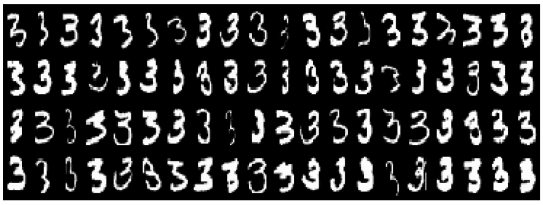
Biomed[29] 医学数据集有 209 个观测对象, 其中有 134 个正常对象 (包括 30 个病毒携带者), 75 个已患病的病毒携带者. Cancer[30] 数据集包含 699 个数据对象, 每个对象用 10 个数值属性描述, 该数据集含有恶性肿瘤 (241 人) 和良性肿瘤 (458 人) 两个类别. 良性肿瘤中存在 37 人可能发展成为恶性肿瘤患者, 所以这 37 人就定义为聚类边界. 结肠癌 (Colon)[31] 基因表达数据集包含 62 个样本, 其中包含 22 个正常样本和 40 个结肠癌样本, 每个样本有 2 000 个基因. 前列腺癌 (Prostate)[32] 基因表达数据集包含 102 个样本, 其中包含 50 个正常样本和 52 个前列腺癌样本, 每个样本含有 10 509 个基因. 这两个数据集缺少边界对象标记, 在进行实验之前, 使用 DBSCAN 算法进行多次统计, 获取这两个基因数据集的边界对象个数分别是 7 个和 18 个, 并对这两个数据集进行预处理. 表 2 是不同算法在各个数据集上的聚类边界检测结果分析.

通过观察表 2 的 F-measure 值可以看出, BAND 算法与 BRINK 算法检测能力相当; 由于真实数据集中噪声的数量相对较小, BORDER 和 BERGE 算法对于噪声敏感的问题并不突出, 所以该算法取得了相对较好的检测结果; 而在基因表达谱数据集中, 尽管数据集维度较高, 但样本量较小, 各边界检测算法均能较好地检测出聚类的边界. 总体而言, MMC 算法在处理高维空间聚类边界问题时是有效的, 且检测能力优于其他算法.

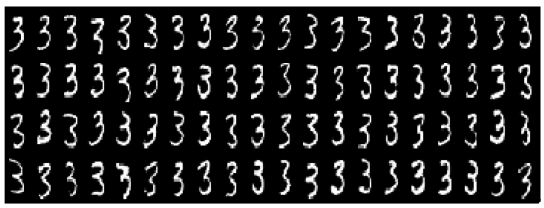
3.3 手写体数字边界检测

手写体数字识别[33−35] 是模式识别和人工智能领域的热门研究方向, 在身份识别、编号扫描、银行支票和车牌识别[36−37] 等领域有着重要应用. 由于手写体数字书写时受个人偏好和习惯的影响, 手写体数字在形状、大小和线条宽度等方面存在较大差异, 甚至出现连笔、断笔、模糊和孔洞等情况. 因此, 不同人书写的手写体数字有着迥然不同的风格. 手写体数字聚类边界具有几何性质不规则、数字之间易混淆和数字与英文字符之间难以区分等特征.

本文中采用 MNIS[38] 数据集进行聚类边界检测能力验证. 该数据集包含 10 类手写数字, 其中包括 60 000 个训练图像样本和 10 000 个检测图像样本, 并且每个图像大小均为 8 bit 位深度的 bmp 图像, 以 28 像素 × 28 像素大小的方式存储, 每个像素的灰度值范围为 0 ∼ 255. 本节在检测图像样本中选取手写体数字 3 共 1 010 张手写体图像进行边界检测实验, 并对这 1 010 张手写体图像进行预处理, 得到 1 010 × 28 的矩阵. 图 2 (a) 是 MMC 算法在手写体数字 3 中检测的边界检测结果, 使用的参数是 $k = 10$, $\varepsilon_1 = 0.9208$. 图 2 (b) 是 MMC 算法在手写体数字 3 中检测出的聚类中心图像, 使用的参数是 $k = 10$, $\varepsilon_1 = 0.0000$, $\varepsilon_2 = 0.0396$.



(a) 聚类边界
(a) Clustering boundary



(b) 聚类中心
(b) Clustering center

图 2 手写体 3 的聚类边界和聚类中心
Fig. 2 The clustering boundary and center objects of "3"

从图 2 可以看出, MMC 算法能够有效检测出手写体数字 3 中的边界图像, 并能检测出聚类中心区域相对标准且易于识别的手写体数字 3, 从而验证了 MMC 算法在手写体数字聚类边界检测中的有效性, 更进一步验证了算法在高维空间中的检测能力.

3.4 多姿态人脸边界检测

人脸识别[39−41] 是计算机图形图像学、计算机视觉、机器学习、模式识别等领域的重要研究课题. 该技术以人体脸部特征为媒介, 利用计算机图形图像的相关处理技术, 捕捉静止或者移动个体的面部局部特征进行身份匹配研究. 人脸边界是人脸图像中一些特殊的人脸图像, 具有侧脸、佩戴墨镜和胡须等特征. 这些人脸图像增加了人脸识别的难度, 所以有效地提取人脸图像的聚类边界有利于提高人脸识别的精度, 并有助于协助信息录入系统过滤掉侧脸图像.

人脸数据集 Pointing data[42] 包含 15 位志愿者不同姿态的头部图像, 其中 Vertical angle = $\{-90°$,

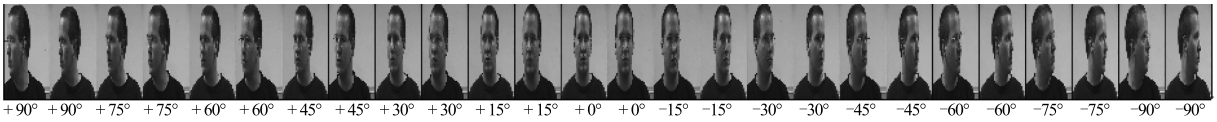


图 3　人脸簇 FC1

Fig. 3　Face cluster FC1

$-60°$, $-30°$, $-15°$, $0°$, $+15°$, $+30°$, $+60°$, $+90°$} 共 9 个变换姿势, Horizontal angle = {$-90°$, $-75°$, $-60°$, $-45°$, $-30°$, $-15°$, $0°$, $+15°$, $+30°$, $+45°$, $+60°$, $+75°$, $+90°$} 共 13 个变换姿势. 该数据集的志愿者年龄普遍处于 20～40 岁之间, 头部的位置由手工切割得到, 分为两个序列: 第 1 个序列仅包括头部姿势在水平和竖直方向上的角度变化, 第 2 个序列在第 1 个序列的基础上加入了光照、表情和佩戴眼镜等附加影响条件. 每个序列含有每个志愿者的 93 张人脸图像, 共 1 395 张人脸图像, 其中每张图像均为 8 bit 位深度的 jpg 图像, 以 384 像素 × 288 像素大小的方式存储, 每个像素的灰度值范围为 0～255. 在实验之前, 对这 2 790 张人脸图像进行转换, 得到 2 790 × 384 的矩阵, 并选择序列一进行多姿态人脸边界检测.

为了描述多姿态人脸图像的聚类边界的定义, 首先选取该数据集的一位志愿者 Vertical angle = {$0°$}, Horizontal angle = {$-90°$, $-75°$, $-60°$, $-45°$, $-30°$, $-15°$, $0°$, $+15°$, $+30°$, $+45°$, $+60°$, $+75°$, $+90°$}, 两个序列共 26 张人脸图像 (如图 3 所示) 进行聚类边界检测实验, 并称该人脸簇为 FC1. 然后选取序列 1 中的一位志愿者 (共 93 张人脸图像) 进行聚类边界检测实验, 称该人脸簇为 FC2. 图 4 给出了标记 4 张人脸图像 (Horizontal angle = {$\pm 90°$}) 作为聚类边界时, MMC 算法的检测结果. 图 5 给出了标记 8 张人脸图像 (Horizontal angle = {$\pm 90°$, $\pm 75°$}) 作为聚类边界时, MMC 算法的检测结果. 图 6 给出了 MMC 算法在 FC2 上的边界检测结果. 图 7 给出了 MMC 算法在第 1 个序列上的边界检测结果.

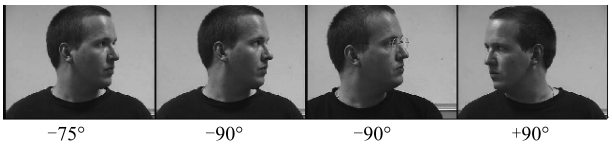


图 4　标记 FC1 的边界个数为 4 时, MMC 算法的检测结果 ($k = 5$, $\varepsilon_1 = 0.8462$)

Fig. 4　The boundary detection result of MMC when marking 4 boundaries on FC1 ($k = 5$, $\varepsilon_1 = 0.8462$)

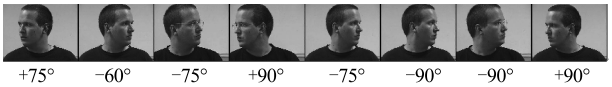


图 5　标记 FC1 的边界个数为 8 时, MMC 算法的检测结果 ($k = 5$, $\varepsilon_1 = 0.6923$)

Fig. 5　The boundary detection result of MMC when marking 8 boundaries on FC1 ($k = 5$, $\varepsilon_1 = 0.6923$)

图 6　标记 FC2 的边界个数为 8 时, MMC 算法的检测结果 ($k = 5$, $\varepsilon_1 = 0.7850$)

Fig. 6　The boundary detection result of MMC when marking 8 boundaries on FC2 ($k = 5$, $\varepsilon_1 = 0.7850$)

从图 4 和图 5 可以看出, MMC 算法可以有效检测出在一维方向上单个人脸簇的聚类边界. 其中, 图 4 的检测结果包含 4 张边界图像, 共检测出 3 张正确图像, 所以其 F-measure 值为 0.8571. 图 5 的检测结果包含 8 张人脸图像, 共检测出 7 张边界图像, 所以其 F-measure 值为 0.9333. 图 6 是 MMC 算法在单个人脸簇二维方向上的聚类边界检测结果, 从实验结果上可以观察到 MMC 算法检测出的聚类边界包含了在水平方向及竖直方向上侧脸角度较大的人脸边界图像. 图 7 是 MMC 算法在整个人脸数据集上的聚类边界检测结果, 包含了不同人脸图像在不同方向上侧脸程度较大的人脸图像. 由于二维方向上的人脸边界缺乏专业人员的标记支持, 无法标记出具体的聚类边界, 所以无法给出 F-measure 分析, 这也是下一步工作的重点. 以上实验分析可以验证 MMC 算法能够检测出人脸数据集聚类的边界, 且检测结果是有效的.

3.5　时间复杂度与参数讨论

本文中 MMC 算法步骤简单, 1) 计算每个数据对象的 k 近邻集合, 时间复杂度为 $O(kn^2)$; 2) 计算每个数据对象的 HP 值, 时间复杂度为 $O(n)$; 3) 进行降序排序, 时间复杂度为 $O(n^2)$; 4) 边界输出, 时间复杂度为 $O(n^2)$.

MMC 算法的时间复杂度为 $O(kn^2)$. BORDER 算法的时间复杂度为 $O(kn^2)$, BRIM 算法的时间复杂度为 $O(n \log n)$, BRINK 算法的时间复杂度为 $O(kn^2)$, BERGR 算法的时间复杂度为 $O(n^2)$. 所以 MMC 算法的时间复杂度与 BORDER 算法、BAND 算法、BRINK 算法、BERGE 算法的时间复杂度相当, 均高于 BRIM 算法.

图 7 MMC 在序列 1 上的聚类边界检测结果 ($k = 5$, $\varepsilon_1 = 0.9427$)

Fig. 7 The boundary detection result of MMC on the first sequence ($k = 5$, $\varepsilon_1 = 0.9427$)

MMC 算法共有 3 个参数: 近邻个数 k、阈值 ε_1 和 ε_2. 通过大量的统计实验发现, 在一般情况下, 当 $k \in [10, 100]$ 时, MMC 算法能够得到较好的边界检测结果; 当 $\varepsilon_1 \in [0.70, 0.85]$, $\varepsilon_2 \in (0.94, 0.98]$ 时, 即聚类边界点大约占据整个数据集的 15 % ∼ 30 % 的比例, 噪声大约占据整个数据集的 2 % ∼ 6 % 的比例时, 算法能取得较好的检测结果. 需要注意的是, 在无噪声数据集中, $\varepsilon_2 = 1$, 只需输入 ε_1 即可完成聚类边界检测. 为了详细分析各参数对边界检测的影响, 使用第 4.1 节和第 4.2 节中的 3 个数据集进行分析.

如图 8 所示, 当输入较小或较大的 k 值时, MMC 算法的检测性能发生较大变化, 所以图 8 中的几条折线图均先上升再下降, 但在 $k \in [10, 100]$ 的区间内, 算法性能相对稳定. ε_1 是 MMC 算法的开始阈值, 当数据点的 HP 值小于 ε_1 时, 该数据点为核心点, 所以 ε_1 的大小控制着核心点的数目. 当 ε_1 较小时, 一些核心点被误认为是边界点, 此时检出的边界点数目较多, 算法的检测性能较低. 当 ε_1 较大时, 一些边界点被误认为是核心点, 此时检出的边界点数目较少, 算法的性能也相对较低, 所以 ε_1 是核心点与边界点的临界点, 该参数影响着算法的性能. 图 9 中的折线图更加清晰地表明了 ε_1 对边界检测的影响. 当 ε_2 较小时, 一些真正边界点被误认为是噪声, 此时检出的边界点数目较少, 算法的检测性能较低. 当 ε_2 较大时, 一些噪声被误认为是边界点, 此时检出的边界点数目较多, 算法的性能也相对较低, 所以 ε_2 是边界点与噪声的临界点, 控制着噪声的数目, 并在 MMC 算法中起过滤噪声的作用. 如果 ε_2 越大, 则过滤的噪声越少, 而 ε_1 作为主要参数依然能够获得良好的检测结果, 也正如图 10 中, 随着 ε_2 的增大, 算法的性能从弱逐渐增强并稳定.

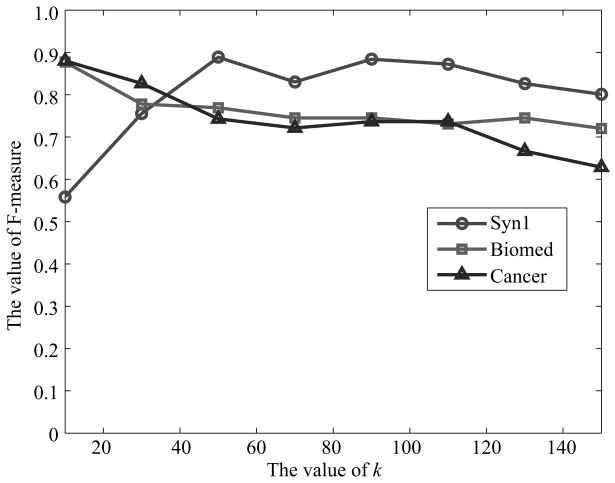


图 8 在不同数据集上输入不同 k 时, F-measure 值的变化

Fig. 8 The change of F-measure when inputting different k on some different data sets

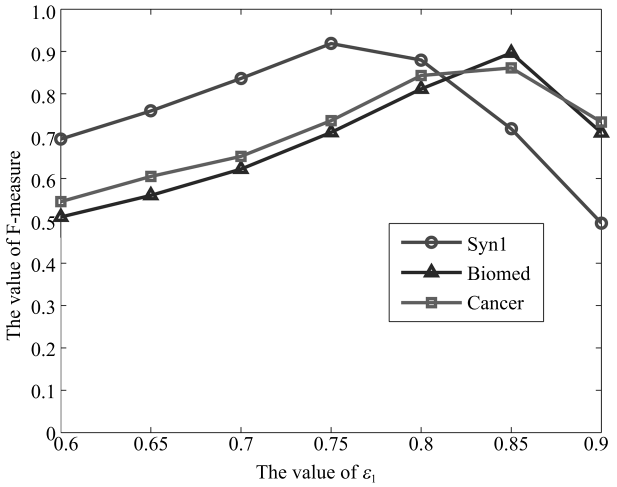


图 9 在不同数据集上输入不同 ε_1 时, F-measure 值的变化

Fig. 9 The change of F-measure when inputting different ε_1 on some data sets

4 结论

本文将高维空间转换为多个矩阵, 使用矩阵模型进行理论分析构造, 提出了检测高维空间聚类边界的 MMC 算法. 该算法构造简单, 易于理解, 能够有效地提取低维和高维空间中聚类的边界. 理论分析和实验结果验证了该算法的有效性. 未来的研究工作集中体现在以下几个方面:

1) 由于 k 近邻空间的计算开销较大, 如何降低算法的时间复杂度是一个重要问题.

2) 减少参数个数甚至无参数技术是当前亟待解决的问题之一.

3) 随着领域的扩展、聚类边界问题的多样化和

复杂化, 需要进一步研究复杂结构数据空间的边界检测技术.

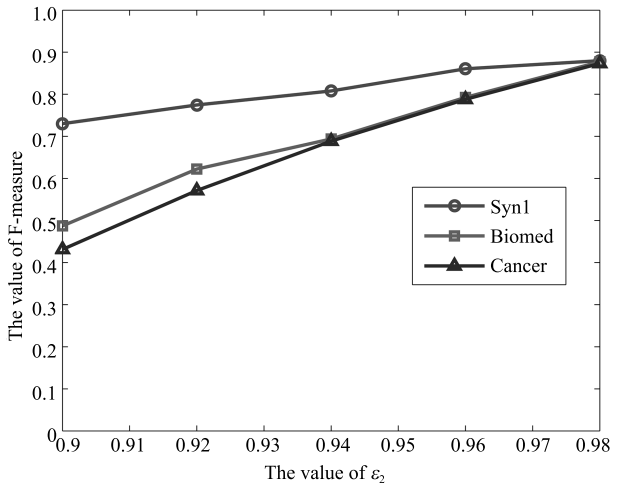


图 10 在不同数据集上输入不同 ε_2 时, F-measure 值的变化

Fig. 10 The change of F-measure when inputting different ε_2 on some data sets

References

1 Tsapanos N, Tefas A, Nikolaidis N, Pitas I. A distributed framework for trimmed kernel k-means clustering. *Pattern Recognition*, 2015, **48**(8): 2685−2698

2 Guo Y H, Sengur A. NCM: neutrosophic c-means clustering algorithm. *Pattern Recognition*, 2015, **48**(8): 2710−2724

3 Vikjord V V, Jenssen R. Information theoretic clustering using a k-nearest neighbors approach. *Pattern Recognition*, 2014,**47**(9): 3070−3081

4 Jain A K. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 2010, **31**(8): 651−666

5 Agrawal R, Gehrke J, Gunopulos D, Raghavan P. Automatic subspace clustering of high dimensional data. *Data Mining and Knowledge Discovery*, 2015, **11**(1): 5−33

6 Dai L Z, Ding J D, Yang J. Inhomogeneity-embedded active contour for natural image segmentation. *Pattern Recognition*, 2015, **48**(8): 2513−2529

7 Aja-Fernández S, Curiale A H, Vegas-Sánchez-Ferrero G. A local fuzzy thresholding methodology for multiregion image segmentation. *Knowledge-Based Systems*, 2015, **83**(1): 1−12

8 Peng P, Addam O, Elzohbi M, Özyer S T, Elhajj A, Gao S, Liu Y M, Özyer T, Kaya M, Ridley M, Rokne J, Alhajj R. Reporting and analyzing alternative clustering solutions by employing multi-objective genetic algorithm and conducting experiments on cancer data. *Knowledge-Based Systems*, 2014, **56**(3): 108−122

9 Kaur P, Soni A K, Gosain A. RETRACTED: a robust kernelized intuitionistic fuzzy c-means clustering algorithm in segmentation of noisy medical images. *Pattern Recognition Letters*, 2013, **34**(2): 163−175

10 Parsons L, Haque E, Liu H. Subspace clustering for high dimensional data: a review. *ACM SIGKDD Explorations Newsletter*, 2004, **6**(1): 90−105

11 Angiulli F, Pizzuti C. Outlier mining in large high-dimensional data sets. *IEEE Transactions on Knowledge and Data Engineering*, 2005, **17**(2): 203−215

12 Ester M, Kriegel H P, Sander J, Xu X W. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96). Portland, Oregon: Association for the Advancement of Artificial Intelligence, 1996. 226−231

13 Xia C Y, Hsu W, Lee M L, Ooi B C. BORDER: efficient computation of boundary points. *IEEE Transactions on Knowledge and Data Engineering*, 2006, **18**(3): 289−303

14 Achtert E, Böhm C, Kröger P, Kunath P, Pryakhin A, Renz M. Efficient reverse k-nearest neighbor estimation. *Informatik-Forschung und Entwicklung*, 2007, **21**(3−4): 179−195

15 Qiu B Z, Yue F, Shen J Y. BRIM: an efficient boundary points detecting algorithm. *Advances in Knowledge Discovery and Data Mining*. Berlin Heidelberg: Springer, 2007. 761−768

16 Xue Li-Xiang, Qiu Bao-Zhi. Boundary points detection algorithm based on coefficient of variation. *Pattern Recognition and Artificial Intelligence*, 2009, **22**(5): 799−802
(薛丽香, 邱保志. 基于变异系数的边界点检测算法. 模式识别与人工智能, 2009, **22**(5): 799−802)

17 Qiu Bao-Zhi, Yang Yang, Du Xiao-Wei. BRINK: an algorithm of boundary points of clusters detection based on local qualitative factors. *Journal of Zhengzhou University (Engineering Science)*, 2012, **33**(3): 117−120
(邱保志, 杨洋, 杜效伟. BRINK: 基于局部质变因子的聚类边界检测算法. 郑州大学学报 (工学版), 2012, **33**(3): 117−120)

18 Li Xiang-Li, Geng Peng, Qiu Bao-Zhi. Clustering boundary detection technology for mixed attribute data set. *Control and Decision*, 2015, **30**(1): 171−175
(李向丽, 耿鹏, 邱保志. 混合属性数据集的聚类边界检测技术. 控制与决策, 2015, **30**(1): 171−175)

19 Roweis S T, Saul L K. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 2000, **290**(5500): 2323−2326

20 Belkin M, Niyogi P. Laplacian eigenmaps and spectral techniques for embedding and clustering. In: Advances in Neural Information Processing Systems 14: Proceedings of the 2001 Conference. Cambridge MA: MIT Press, 2001. 585−591

21 He X, Niyogi P. Locality preserving projections. *Advances in Neural Information Processing Systems*, 2003, **16**(1): 186−197

22 Zhang Z Y, Zha H Y. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *SIAM Journal on Scientific Computing*, 2004, **26**(1): 313−338

23 Zheng Si-Long, Li Yuan-Xiang, Wei Xian, Peng Xi-Shuai. Nonlinear dimensionality reduction based on dictionary learning. *Acta Automatica Sinica*, 2016, **42**(7): 1065−1076
(郑思龙, 李元祥, 魏宪, 彭希帅. 基于字典学习的非线性降维方法. 自动化学报, 2016, **42**(7): 1065−1076)

24 Taşdemir K, Yalçin B, Yildirim I. Approximate spectral clustering with utilized similarity information using geodesic based hybrid distance measures. *Pattern Recognition*, 2015, **48**(4): 1465−1477

25 Hearst M A, Dumais S T, Osman E, Platt J, Scholkopf B. Support vector machines. *IEEE Intelligent Systems and their Applications*, 1998, **13**(4): 18−28

26 Chua L O, Yang L. Cellular neural networks: theory. *IEEE Transactions on Circuits and Systems*, 1998, **35**(10): 1257−1272

27 Zimdahl H, Hübner N. Gene chip technology and its application to molecular medicine. *Encyclopedic Reference of Genomics and Proteomics in Molecular Medicine*. Berlin Heidelberg: Springer, 2006. 650−655

28 Ferdous M M, Vinciotti V, Liu X H, Wilson P. Exploring the link between gene expression and protein binding by integrating mRNA microarray and ChIP-Seq data. *Statistical Learning and Data Sciences*. Switzerland: Springer International Publishing, 2015. 214−222

29 The Data and Story Library. Biomed data set [Online], available: http://lib.stat.cmu.edu/datasets/biomed.data.html, October 24, 2017

30 UCI Machine Learning Repository. Cancer data set [Online], available: http://archive.ics.uci.edu/ml/datasets.html, October 24, 2017

31 Princeton University Gene Expression Project. Colon data set [Online], available: http://genomics-pubs.princeton.edu /oncology/affydata/index.html, October 24, 2017

32 Gene Expression Model Selector. Prostate data set [Online], available: http://www.gems-system.org, October 24, 2017

33 Zhu Q, Xin H. Feature extraction and filter in handwritten numeral recognition. *Geo-Informatics in Resource Management and Sustainable Ecosystem*. Berlin Heidelberg: Springer, 2013. 58−67

34 Weber-Alonso J M, Sesmero M P, Gutierrez G, Ledezma A, Sanchis A. Input transformation and output combination for improved handwritten digit recognition. *Artificial Neural Networks*. Switzerland: Springer International Publishing, 2015, **4**: 435−443

35 Wang Y M, Peyls A, Pan Y, Claesen L, Yan X L. A fast self-organizing map algorithm for handwritten digit recognition. *Multimedia and Ubiquitous Engineering*. Berlin Heidelberg: Springer, 2013, **240**: 177−183

36 Jia W J, Zhang H F, He X J. Region-based license plate detection. *Journal of Network and Computer Applications*, 2007, **30**(4): 1324−1333

37 Zhou W G, Li H Q, Lu Y G, Tian Q. Principal visual word discovery for automatic license plate detection. *IEEE Transactions on Image Processing*, 2012, **21**(9): 4269−4279

38 THE MNIST DATABASE. Mnist data set [Online], available: http://yann.SMCun.com/exdb/mnist/, October 24, 2017

39 Huang S C, Chen J, Luo Z. RETRACTED ARTICLE: sparse tensor CCA for color face recognition. *Neural Computing and Applications*, 2014, **24**(7−8): 1647−1658

40 Bhaskar B, Mahantesh K, Geetha G P. An investigation of fSVD and ridgelet transform for illumination and expression invariant face recognition advances in intelligent informatics. *Advances in Intelligent Informatics*. Switzerland: Springer International Publishing, 2015, **320**: 31−38

41 Dang K D, Le T H. Local region partitioning for disguised face recognition using non-negative sparse coding. *Advanced Methods for Computational Collective Intelligence*. Berlin Heidelberg: Springer, 2013, **457**: 197−206

42 Head Pose Image Database. Pointing'04 dat set, [Online], available: http://www-prima. inrialpes. fr/Pointing04, October 24, 2017

李向丽 郑州大学信息工程学院副教授. 主要研究方向为计算机网络, 数据挖掘. E-mail: iexlli@zzu.edu.cn

(**LI Xiang-Li** Associate professor at the School of Information Engineering, Zhengzhou University. Her research interest covers computer network and data mining.)

曹晓锋 郑州大学信息工程学院硕士研究生. 主要研究方向为模式识别和数据挖掘. 本文通信作者. E-mail: 18739920964@163.com

(**CAO Xiao-Feng** Master student at the School of Information Engineering, Zhengzhou University. His research interest covers pattern recognition and data mining. Corresponding author of this paper.)

邱保志 郑州大学信息工程学院教授. 主要研究方向为数据库, 先进智能系统, 数据挖掘. E-mail: iebzqiu@zzu.edu.cn

(**QIU Bao-Zhi** Professor at the School of Information Engineering, Zhengzhou University. His research interest covers database, advanced intelligent system, and data mining.)