局部显著单元高维聚类算法

宗 瑜[①②] 李明楚[①] 徐贯东[②] 张彦春[②]
[①](大连理工大学软件学院 大连 116621)
[②](维多利亚大学信息应用中心 墨尔本 VIC3011)

摘 要：以等宽或随机宽度网格密度单元为基础的高维聚类算法不能保证复杂数据集中的聚类结果的质量。该文在核密度估计和空间统计理论的基础上，给出一种基于局部显著单元的高维聚类算法来处理复杂数据的高维聚类问题。该方法以局部核密度估计和空间统计理论为基础定义了局部显著单元结构来捕获局部数据分布；设计了能快速发现覆盖数据分布的局部显著区域的贪婪算法；对具有相同属性子集的局部显著单元执行 Single-linkage 算法发现其中的聚类结果。实验结果表明，以局部显著单元为基础的高维聚类算法能够发现复杂数据集中隐含的高质量聚类结果。

关键词：聚类分析；高维聚类算法；核密度估计；局部显著单元
中图分类号：TP18 **文献标识码**： A **文章编号**：1009-5896(2010)11-2707-06
DOI: 10.3724/SP.J.1146.2009.01589

High Dimensional Clustering Algorithm Based on Local Significant Units

Zong Yu[①②] Li Ming-chu[①] Xu Guan-dong[②] Zhang Yan-chun[②]
[①](*School of Software*, *Dalian University of Technology*, *Dalian* 116621, *China*)
[②](*Center of Applied Information*, *Victoria University*, *Melbourne* VIC3011, *Australia*)

Abstract: High dimensional clustering algorithm based on equal or random width density grid cannot guarantee high quality clustering results in complicated data sets. In this paper, a High dimensional Clustering algorithm based on Local Significant Unit (HC_LSU) is proposed to deal with this problem, based on the kernel estimation and spatial statistical theory. Firstly, a structure, namely Local Significant Unit (LSU) is introduced by local kernel density estimation and spatial statistical test; secondly, a greedy algorithm named Greedy Algorithm for LSU (GA_LSU) is proposed to quickly find out the local significant units in the data set; and eventually, the single-linkage algorithm is run on the local significant units with the same attribute subset to generate the clustering results. Experimental results on 4 synthetic and 6 real world data sets showed that the proposed high-dimensional clustering algorithm, HC_LSU, could effectively find out high quality clustering results from the highly complicated data sets.

Key words: Clustering analysis; High dimensional Clustering (HC) algorithm; Kernel density estimation; Local Significant Unit (LSU)

1 引言

聚类分析是一种不依赖于先验知识的、从数据本身出发探寻其内部分布模式的数据分析方法，在分析和解决大规模复杂问题时体现出巨大的优越性。根据聚类定义方法的不同，研究者们提出的聚类算法大致可分为划分方法、层次方法等[1]。尽管不同聚类算法之间存在或多或少的差异，但是它们都是基于“数据的聚类依赖于数据对象之间的相似性度量”这一共同出发点的。因此以数据空间密度构造方法为基础的聚类算法展示了显著的优越性[2]。典型的以空间密度构造方法为基础的聚类算法 DENCLUS2[3]，SPARCL[4]及 DECODE[5]等，具有处理高维数据聚类的能力，在应用中表现了良好的性能。然而，这些算法的相似度是以描述数据对象的所有属性为基础的，因此无法发现仅在属性子集上相似的聚类结果，即子空间聚类结果[6]。EPCH[7]，P3C[8]及 STAPTC[9]是典型的子空间聚类算法，它们用等宽或随机宽度的空间网格单元的密度来模拟核密度估计，从而达到发现空间密度分布的目的。等

2009-12-11 收到，2010-05-20 改回
国家自然科学重点基金 (90715037)，国家 973 计划项目(2007CB714205)，澳大利亚 ARC 项目(DP0770479)和安徽省教育厅重点项目(KJ2009A54，KJ2010A325)资助课题
通信作者：宗瑜 nick.zongy@gmail.com

宽或随机宽度的空间网格对捕捉规则形状的密度区域十分有效。由于实际应用中的数据分布十分复杂，因此现有的子空间聚类算法不能保证聚类结果的质量。

本文给出一种局部显著单元(Local Significant Unit, LSU)结构来捕获高维空间中的复杂密度分布，产生聚类结果和孤立点。局部显著单元的构造方法是：(1)对任意选择的数据对象，调用局部核密度估计算法 Rodeo [10]计算该数据对象的核密度值及与其对应的窗宽向量；(2)确定与该数据对象核密度估计相关的属性子集及其对应的窗宽向量；(3)在属性子集中以窗宽向量为边长创建围绕在数据对象周围的超方体；(4)对超方体进行空间统计假设测试，通过假设检验的超方体及其对应的属性子集被定义为局部显著单元。算法 GA_LSU(Greedy Algorithm for Local Significant Unit)以贪婪搜索策略发现能覆盖数据分布的局部显著单元。HC_LSU(High dimensional Clustering algorithm based on Local Significant Unit)算法对具有相同属性子集的局部密集超方体执行 Single-linkage 算法产生聚类结果。多组仿真和实际数据集上的实验结果表明，HC_LSU 高维聚类算法具有产生高质量聚类结果的能力。

2 符号约定及背景知识

本节首先给出了相关的符号约定，然后简要地介绍局部核密度估计算法 Rodeo。

本文涉及的相关符号及其意义如表 1 所示。

表 1 符号约定

符号	含义	符号	含义
D	数据集	S	属性子集，S 的维度小于等于 d
N	样本个数或数据对象个数	$\boldsymbol{W}^S$	S 相对应的窗宽向量
d	描述数据对象的维数	$\boldsymbol{W}^*$	核密度估计方法返回的窗宽向量
$A=\{\text{attr}_1,\cdots,\text{attr}_d\}$	描述数据对象的属性全集	H	超方体

核密度函数贴切地反映了分布密度函数在一点的值对该点邻域样本点密度的正比依赖关系。高维分布函数的核密度估计不仅与给定的样本点集有关，还与核函数和窗宽参数的选择有关。在给定数据样本集的情况下，窗宽的确定成为影响核密度估计的重要因素。局部核密度估计方法 Rodeo 具有同时估计核密度和确定窗宽的功能[10]。给定数据对象 x，Rodeo 算法返回关于 x 的核密度估计值 $f_{\boldsymbol{W}^*}(x)$ 及其对应的窗宽向量 $\boldsymbol{W}^*$。假设与 $f(x)$ 核密度估计相关的属性排列在前 r 个位置，即 $1\le j\le r$，而余下属性为与 $f(x)$ 核密度估计无关的属性，即 $r+1\le j\le d$。局部核密度估计算法 Rodeo 返回的窗宽向量满足如下定理。

定理 1[11] 给定局部核密度估计算法 Rodeo 的输出窗宽向量 $\boldsymbol{W}^*=\text{diag}(w_1^*,\cdots,w_d^*)$，$\boldsymbol{W}^*$ 满足：

$$P(w_j^*=w_j^{(0)})\to 1 \quad , \quad \text{for all} \quad j>r \quad \text{and}$$

$$P(w_j^0(Nb_N)^{-1/(4+r)}\le w_j^*\le w_j^0(Na_N)^{-1/(4+r)})\to 1,$$

$$\text{for all} \quad j\le r$$

其中 $w_j^{(0)}$ 为第 j 维上 x 的初始窗宽，a_N 和 b_N 为与 N 等阶的常数。

定理 1 表明，在与核密度估计相关的属性子集上，其对应的窗宽值小，而与核密度估计无关的属性子集上，其对应的窗宽值大。因此，根据定理 1 可以确定与核密度估计相关的属性子集。

3 基于局部显著单元的高维聚类算法

本节首先在 Rodeo 算法和空间统计测试的基础上，定义局部显著单元，然后给出以贪婪策略为基础的 GA_LSU 算法，最后对同属性空间中的密集超方体进行合并形成聚类结果。

3.1 局部显著单元定义

给定由 $A=\{\text{attr}_1,\cdots,\text{attr}_d\}$ 描述的数据集 D，任意 $x_{ij}\in D$，其取值范围为 $[0,1]$。设由定理 1 确定的与 $f(x)$ 核密度估计相关属性子集为 S，且与 S 相对应的窗宽向量为 $\boldsymbol{W}^S=\text{diag}(w_1^*,\cdots,w_{|S|}^*)$，$\boldsymbol{W}^S\subseteq\boldsymbol{W}^*$。

定义1 给定任意数据对象 $x\in D$，属性子集 S 及窗宽向量 $\boldsymbol{W}^S$，在 S 中以 $\boldsymbol{W}^S$ 为宽度创建围绕在 x 周围的超方体 H：$H=I_1\times\cdots\times I_r$，其中 $I_j=[(x_j-w_j/2),(x_j+w_j/2)]$，$x_j$ 为 x 的第 j 个属性，$j=1,\cdots,r$，$r=|S|$。

定义 2 给定属性子集 S 中的超方体 $H=I_1\times\cdots\times I_r$，该超方体中包含的数据对象定义为

$$\text{remm}(H)=\{x_i\mid(x_{ij}-w_j/2)\le x_{ij}\le(x_{ij}+w_j/2)\},$$

$$i=1,\cdots,N,\quad j=1,\cdots,r$$

定义 3 给定属性子集 S 中的超方体 $H=I_1\times\cdots\times I_r$，则 H 的局部密度由其包含的数据对象个数确定，即 $LD(H)=|\text{remm}(H)|$。

d-维空间中随机分布着 N 个数据对象，落入某一密集闭合区域中的数据对象个数服从以 N 和该区域体积为参数的二项分布[12]。超方体 H 是属性子集 S 中的闭合区域，$LD(H)$ 是 H 中包含的数据对象个数，因此 $LD(H)$ 服从以 N 和 $\text{vol}(H)$ 为参数的二项分

布，即

$$LD(H) \sim \text{Binomial}(N, \text{vol}(H))$$

高维空间聚类的目的是为了发现在属性子集上相似的聚类结果，该聚类结果可以以属性子集的密集区域表示[9]。因此首先要确定由定义 1 所构造的超方体 H 是否是密集的。在空间统计测试的基础上对超方体 H 进行假设检验：

h_0： H 中至少包含了 $LD(H)$ 个数据对象。

显著性水平 α 表示 h_0 为真但在检验过程中被当作假的而丢弃的概率。给定显著性水平，可以通过单边或双边测试获得假设检验的统计临界值。对于单边测试，其统计临界值可以由 $\alpha = p(LD(H) \geq \theta)$ 计算获得，其中 p 为概率密度函数。

定义 4　给定属性子集 S 中的超方体 H，显著性水平 α， H 的临界值 θ 由单边测试获得，其中概率密度函数为 $\text{Binomial}(N, \text{vol}(H))$。如果 $LD(H) \geq \theta$，则定义 (H, S) 为局部显著单元。

从文献[12]的分析可知，如果 d-维空间中某一闭合区域是密集的，那么该闭合区域的数据对象分布服从以 N 和该区域体积的二项分布。局部显著单元的目标是从 d-维空间中发现密集的闭合区域，因此在局部显著单元中密集超方体的数据分布应该服从二项分布。本文显著性水平 α 的设置是依据二项分布或高斯分布给出的。

3.2 局部显著单元发现算法

虽然遍历搜索覆盖 D 中数据分布的方法能够保证毫无遗漏地发现其中包含的局部显著单元，但是却十分耗时且存在冗余。为了快速发现能覆盖 D 中数据分布的局部显著单元，本文给出了以贪婪策略为基础的局部显著区域算法 GA_LSU，如表 2 所示。

表 2 算法 1： GA_LSU 算法

输入：D, α

输出： LSU_Set

1 LSU_Set $= \varnothing$; %局部显著单元集合

2 Loop

2.1 从 D 中随机选择未被访问的数据对象 x ;

2.2 $[f_{\boldsymbol{W}^*}(x), \boldsymbol{W}^*]$ =Rodeo(D, x);

2.3 根据定理 1 产生与 x 核密度估计相关的属性子集 S 及其对应的窗宽向量 $\boldsymbol{W}^S$;

2.4 根据定义 1，在 S 中创建以 $\boldsymbol{W}^S$ 为宽度的超方体 H;

2.5 根据定义 2，得到包含在 H 中的数据对象集合， $\text{remm}(H)$;

2.6 根据定义 4 对 H 进行统计测试;

2.6.1 If $LD(H) \geq \theta$ % θ 是满足单边检验的统计临界值

(a) LSU_Set = LSU_Set $\bigcup (H, S)$;

(b) 标记 x 及 $\text{remm}(H)$ 中的数据对象为已访问;

2.6.2 end

3 直到 D 中所有数据对象被访问为止。

4 返回 LSU_Set 。

GA_LSU 的基本步骤：(1)随机选择未被处理的数据对象 $x \in D$； (2)调用局部核密度估计算法 Rodeo 计算 x 的核密度值及窗宽向量 $\boldsymbol{W}^*$=diag(w_1^*, $\cdots, w_d^*$)； (3)依据定理 1 产生与 x 核密度估计相关的属性子集 S 及其对应的窗宽向量 $\boldsymbol{W}^S = \text{diag}(w_1^*, \cdots, w_{|S|}^*)$; (4)在 S 中，以 W^S 为宽度创建一个围绕在 x 周围的超方体 H 并产生包含在 H 中的所有数据对象集合 $\text{remm}(H)$; (5)如果 H 通过空间统计测试，则将 (H, S) 定义为局部显著单元，并标记 x 及 H 中的数据对象为已访问；否则删除 H ; (6)迭代执行(1)-(5)直到所有数据对象被访问为止。

3.3 局部显著单元高维聚类算法(HC_LSU)如表 3 所示。

表 3　算法 2：HC_LSU 算法

输入： LSU_Set

输出：聚类结果

1 将 LSU_Set 中局部显著单元按属性子集分类, $\{\text{LSU\_}S_1, \cdots, \text{LSU\_}S_K\}$,其中 $\text{LSU\_}S_k$， $k = 1, \cdots, K$ 是具有相同属性子集的局部显著单元集合;

2 For $k = 1, \cdots, K$

2.1 对 $\text{LSU\_}S_k$ 中的密集超方体执行 Single-linkage 算法合并;

2.2 产生第 k 个属性子集中的聚类结果 $\{C_k, S_k\}$ ，其中 C_k 由 $\text{LSU\_}S_k$ 中所有超方体的 $\text{remm}(H)$ 中的数据对象产生，而 S_k 是 $\text{LSU\_}S_k$ 中任意局部显著单元的属性子集;

3 剩余数据对象重分配。

GA_LSU 算法返回一个覆盖 D 中数据分布的局部显著单元的集合 LSU_Set 。在每个局部显著单元 (H, S) 中， H 是 S 中的密集超方体。在同一个属性空间中的所有密集超方体反映了该属性空间中数据对象的密度分布情况。高维聚类算法的目的是发现高维空间中在某些属性子集上密集的数据对象，并将该属性子集和其中包含的密集数据对象看作聚类结果。因此，合并具有相同属性子集的局部显著单元中的密集超方体将会产生初始聚类结果，即属性子集和密集超方体的并集。在 LSU_Set 的基础上，给出了 HC_LSU 聚类算法(如表 3 所示)来产生高维空间聚类结果。该算法：(1)将 LSU_Set 中的所有局部显著单元按属性子集进行划分，使得具有相同属性子集的局部显著单元放在一起；(2)对具有相同属性子集的局部显著单元，执行 Single- linkage 算法合并它们，形成初始聚类结果；(3)采用文献[13]的重分配原则将未分配到初始聚类结果中的数据对象进行再分配。

4 实验及分析

本节在4组仿真和6组实际数据集中对比了算法HC_LSU与EPCH，P3C及STATPC算法的聚类质量。

4.1 评价标准

精确度[14]是广泛应用的高维聚类算法评价标准，其描述为：已知数据集的分类结果，精确度是算法产生的聚类结果与分类结果精确匹配的对象个数与总的对象个数的比值，其取值越高，说明由聚类算法发现的聚类结果就越好。本文实验是在配置为 Pentium VI 2.66 GHz CPU，2 GB RAM 的计算机上执行的。算法 HC_LSU 和 EPCH 由 matlab7.5 语言实现，而 P3C 和 STATPC 则包含在 Opensubspace 评价包中，该评价包可以从 http://dme.rwth-aachen.de/OpenSubspace/ evaluation 网站下载。

4.2 仿真数据集及实验结果

本文采用文献[15]给出的仿真数据生成器生成仿真数据。该生成器通过设置参数，如数据集大小‘N’、描述数据的维数‘d’、聚类个数‘K’、聚类平均维数‘L’等，来控制仿真数据集的大小及结构。本文实验利用该生成器创建了 4 个仿真数据集：DS1: $N10000d70L15K4$，DS2: $N15000d100L17K5$，DS3: $N20000d130L20K6$，DS4:$N25000d160L20K7$。每个数据集中的聚类结果都是由某些属性上的密集区域组成，DS1 中聚类是由每个维上宽度基本相似的密集区域组成。而其余数据集中的聚类结果则是由密度区域的宽度不同的数据分布组成。

图 1 给出了算法 EPCH，P3C，STATPC 和 HC_LSU 在 4 个仿真数据集上的精确度值对比。由于 DS1 的聚类是由规则密度区域组成的，所以 4 个算法在该数据集上的精确度值比较接近。这说明 EPCH，P3C，STATPC 和 HC_LSU 都具有能够正确发现规则聚类结果的能力。数据集 DS2，DS3 和 DS4 中包含的聚类形状比 DS1 包含的聚类形状要复杂，它们是由在相关维上宽度不同的密集区域组成。由于 ECPH 算法强制性地将每个维划分成等宽度间隔，再用这些等宽间隔捕获相关密度单元，从而无法发现宽度不同的密集区域，因此 EPCH 算法不能正确发现聚类结果。P3C 用每个维上的数据的聚类结果来确定该维上的相关密度间隔，然后用这些间隔构造超方体，该方法具有处理复杂形状聚类的能力。STATPC 算法用随机宽度从局部角度创建超方体，再用统计测试方法判断该超方体是否是密集的。P3C 和 STATPC 在一定程度上克服了等宽超方体的缺陷，但是随机宽度超方体结构不是以数据的真实分布为基础的，从而无法正确地反映它们。从图中可以看出，算法 HC_LSU 的精确度值明显高于 P3C，STATPC 及 EPCH 的精确度值，这说明 HC_LSU 算法获得的高维聚类结果的质量高于 P3C，STATPC 及 EPCH 算法的聚类质量。

为了考察 4 种算法对噪声的处理能力，我们在数据集 DS1 中加入一定数量的噪声数据。按照 DS1 规模的 0%:5%:25%向其中加入噪声数据，产生了 6 个新的带有不同规模噪声的数据集。在噪声数据集上分别执行 4 种算法，其结果如图 2 所示。由于等宽超方体对发现数据真实分布能力的限制，所以算法 P3C，STATPC 及 HC_LSU 对噪声数据的处理能力强于算法 EPCH。算法 HC_LSU 以局部显著单元来捕获数据分布的局部密集区域，具有能真实地反映数据分布的能力，因此 HC_LSU 算法对噪声数据的控制比较有效。如图 2 所示，即使在数据集中增加了 25%的噪声，算法 HC_LSU 获得的聚类结果的精确度值仍然在 0.9 以上。

图 3 给出了算法在多组不同仿真数据集上的运行时间对比，其时间用对数的形式给出。图 3(a)考察了数据集规模与算法运行时间之间的关系。在固定参数 d=10，K=2，L=2 的情况下，使参数 N 从 10000 按照步长 10000 增加到 60000。从图 3(a)中可以看出，算法 HC_LSU 和 STATPC 的时间消耗比其他两种算法的时间消耗要多，但是这个运行时间是可以接受。图 3(b)给出了算法运行时间和数据维度的关系(设定 N=300，K=2，L=2)。从维度的增加导致 4 种算法的运行时间增大的现象，可以看出

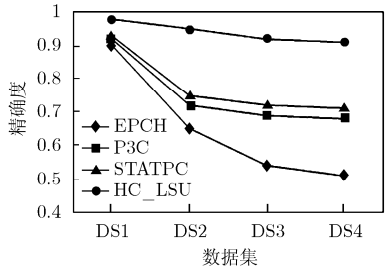


图 1 仿真数据集上 4 种算法的精确度值比较

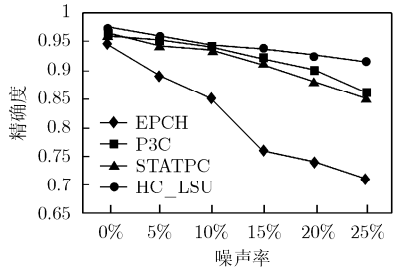
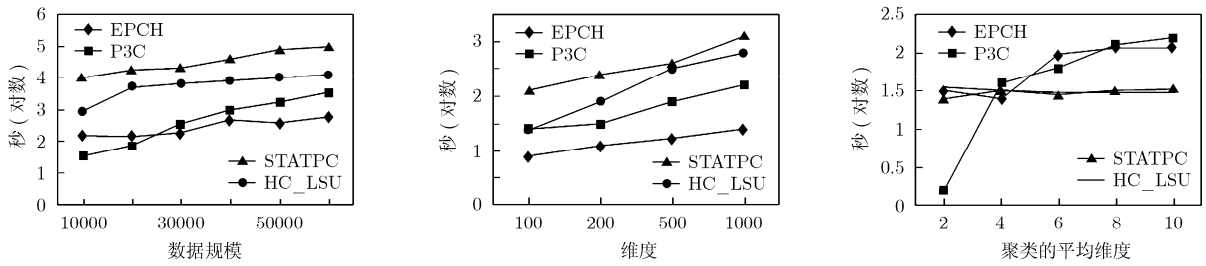


图 2 4 种算法在噪声数据集中的精确度值的比较



(a) 数据规模　(b) 数据维度　(c) 聚类平均维度

图 3 算法在多组不同仿真数据集上的运行时间对比

4 种算法的运算时间对维度变化都较敏感。图 3(c) 描述了算法运行时间和聚类平均维度之间的关系 (设定 N=300，d=50，K=2)。从图中可以看出，算法 P3C 和 EPCH 的运行时间随着聚类平均维度的增加而急剧增大，而算法 STATPC 和 HC_LSU 却对维度变化不敏感。由于 HC_LSU 的密集单元是基于局部核密度估计算法 Rodeo 估计结果和空间统计测试的，算法与聚类的平均维度之间没有直接关联，因此算法对维度变化不敏感。

局部显著单元是以局部核密度估计和空间统计为基础的，因此对于二项分布或高斯分布的数据集，局部显著单元具有正确发现密集区域的能力。但是当数据集中均匀分布的数据对象过多时，局部显著单元捕获数据分布的能力会受到影响。

4.3 实际数据集及实验结果

本节给出在 6 个实际数据集上(如表 4 所示)的实验结果与分析。这 6 个实际数据集的前 4 个数据集是从 UCI 机器学习知识库下载的(http://archive.ics.uci.edu/ml)，第 5 个数据集是由 Alon 等人研究的基因数据[15]，最后是 DePaul CTI Web server 的用户访问数据[16]。

图 4 给出 4 种算法在 6 个实际数据集上的精确度曲线对比。从图中曲线可以看出，在 6 个不同的实际数据集上 P3C，STATPC 及 HC_LSU 算法的精确度值比较接近，这是因为这 3 个算法采用的基本方法相似。由于 P3C，STATPC 及 HC_LSU 在数据分布的基础上，采用动态网格结构来反映数据的基本分布情况，因此，它们的精确度值比采用等宽 d – 维直方图为基础的 EPCH 算法的精确度值要高。从图 4 中可以看出，在 6 个实际数据集中，EPCH 算法的精确度值曲线始终处于 4 条曲线的最下方。实验用到的 6 个实验数据集中，只有 Colon Cancer 和 CTI.STD 数据集中包含子空间聚类结果，而其余 4 个数据集的聚类结果存在于描述数据对象的全维空间中。由于 HC_LSU 算法是以多维核密度估计及空间统计方法为基础的算法，因此该算法具有发现高维全空间聚类结果的能力。Colon Cancer 和 STI.STD 数据集的维数很高而且其中的数据分布十分稀疏，因此 4 种聚类算法的精确度值明显小于在其它 4 个数据集上的精确度值，这个现象说明高维数据的稀疏性对聚类算法的影响较大。即使在这种情况下，HC_LSU 算法在 Colon Cancer 和 STI.STD 数据集上的精确度值也是高于其余 3 种算法的精确度值的，这说明 HC_LSU 算法捕获数据集的真实数据分布能力较强。

表 4 实际数据集描述

数据集	大小	属性	分类
Liver Disorders	345	7	2
Wisconsin Breast Cancer Prognostic (WBCP)	198	33	2
Image Segmentation	180	19	6
Pima Indians Diabetes (PID)	768	9	2
Colon Cancer	62	2000	2
CTI.STD	13745	683	12

5 结论

核密度估计是以空间密度为基础的高维聚类算法的理论基础，但是用等宽或随机宽度模拟核宽度

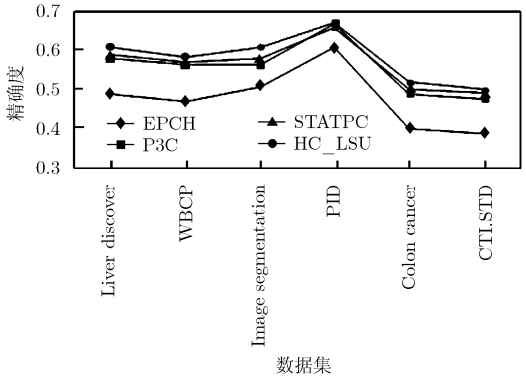


图 4 4 种算法在 6 个实际数据集上的精确度比较

的方法限制了现有高维聚类算法发现具有复杂分布的高维聚类结果的能力。在局部核密度估计方法及空间统计测试理论的基础上，本文首先定义了局部显著单元结构来捕捉局部数据分布；然后以贪婪搜索策略设计了能发现覆盖数据集分布的局部显著单元；最后对具有相同属性子集的局部显著单元执行 Single-linkage 算法产生聚类结果。在仿真及实际数据集上的实验结果表明，局部显著单元能够捕获高维空间的数据分布情况，算法 HC_LSU 具有发现高质量聚类结果的能力。

致谢 感谢大连理工大学江贺老师给予的宝贵意见和有益帮助。

参 考 文 献

[1] 孙吉贵, 刘杰, 赵连宇. 聚类算法研究[J]. 软件学报, 2008, 19(1): 48–61.
Sun Ji-gui, Liu Jie, and Zhao Lian-yu. Clustering algorithms research. *Journal of Software*, 2008, 19(1): 48–61.

[2] Hinneburg A and Keim D A. An efficient approach to clustering in large multimedia databases with noise [C]. Processing of the 4th International Conference on Knowledge Discovery and Data Mining, New York: AAAI Press, 1998: 58–68.

[3] Hinneburg A and Gabriel H H. DENCLUS2.0: Fast Clustering based on kernel density estimation[C]. IDA, 2007, LNCS 4723: 70–80.

[4] Vineet C J, Mohammad A H, and Saeed S, *et al*.. SPARCL: Efficient and effective shape-based clustering[C]. Proceedings of 8th IEEE International Conference on Data Mining, Pisa, Italy, 2008: 93–102.

[5] Tao P, Ajay J, and David J H, *et al*.. DECODE: A new method for discovering clusters of different densities in spatial data [J]. *Data Mining and Knowledge Discovery*, 2009, 18(3): 337–369.

[6] Hans H P, Peer K, and Arthir Z. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering [J]. *ACM Transactions on Knowledge Discovery from Data* (*TKDD*), 2009, 3(1): 1–58.

[7] Ng K, Fu A, and Wong C W. Projective clustering by histograms [J]. *IEEE Transactions on Knowledge and Data Engineering*, 2005, 17(3): 369–383.

[8] Moise G, Sander J, and Ester M. Robust projected clustering [J]. *Knowledge Information System*, 2008, (14): 273–298.

[9] Moise G and Sander J. Finding non-redundant, statistically significant regions in high dimensional data: a novel approach to projective and subspace clustering[C]. Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD,08) Lasvegas, 2008: 533–541.

[10] Liu H, Lafferty J, and Wasserman L. Sparse nonparametric density estimation in high dimensions using the rodeo[C]. 11th International Conference on Artificial Intelligence and Statistics, AISTATS, Florida, 2007: 1049–1062.

[11] Lafferty J D and Wasserman L A. Rodeo: Sparse nonparametric regression in high dimensions [J]. *Advances in Neural Information Processing System*, 2007(18): 1–45.

[12] Baddeley A. Spatial point processes and their applications[J]. *Lecture Notes in Mathematics*, 2007, 1892: 1–75.

[13] Aggarwal C C, Procopiuc C, and Wolf J, *et al*.. Fast algorithms for projected clustering[C]. Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD,99) (ACM SIGKDD, Philadelphia,1999), 1999: 61–72.

[14] Muller E, Assent I, and Krieger R, *et al*.. Density estimation for data mining in high dimensional spaces[C]. SDM, Nevada, USA, 2009: 173–184.

[15] Alon U, Barkai N, and Notterman K, *et al*.. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays[J]. *PNAS*, 1999(96): 6745–6750.

[16] Zhang Y C and Xu G D. On web communities mining and recommendation [J]. *Concurrency and Computation: Practice and Experience*, 2009, 21(5): 561–582.

宗 瑜： 男，1976 年生，讲师，博士生，研究方向为数据挖掘、智能算法.

李明楚： 男，1963 年生，教授，博士生导师，研究方向为图论、算法分析与设计、网格计算、数据挖掘.

徐贯东： 男，1967 年生，Research Fellow，博士，研究方向为 Web 挖掘.

张彦春： 男，1956 年生，教授，博士生导师，研究方向为数据库、数据挖掘.