



文章编号:1001-9081(2018)02-0483-08

DOI:10.11772/j.issn.1001-9081.2017082053

## 基于密度峰值的混合型数据聚类算法设计

李 眯<sup>1,2</sup>, 陈奕延<sup>1\*</sup>, 张淑芬<sup>2</sup>

(1. 中国市场学会服务质量专业委员会, 北京 100048; 2. 河北省数据科学与应用重点实验室(华北理工大学), 河北 唐山 063210)

(\* 通信作者电子邮箱 townjam\_sovietnia@163.com)

**摘要:**针对  $k$ -prototypes 算法无法自动识别簇数以及无法发现任意形状的簇的问题,提出一种针对混合型数据的新方法:寻找密度峰值的聚类算法。首先,把 CFSFDP(Clustering by Fast Search and Find of Density Peaks)聚类算法扩展到混合型数据集,定义混合型数据对象之间的距离后利用 CFSFDP 算法确定出簇中心,这样也就自动确定了簇的个数,然后其余的点按照密度从大到小的顺序进行分配。其次,研究了该算法中阈值(截断距离)及权值的选取问题:对于密度公式中的阈值,通过计算数据场中的势熵来自动提取;对于距离公式中的权值,利用度量数值型数据集和分类型数据集聚类趋势的统计量来定义。最后通过在三个实际混合型数据集上的测试发现:与传统  $k$ -prototypes 算法相比,寻找密度峰值的聚类算法能有效提高聚类的精度。

**关键词:**聚类分析;混合型数据;数据场;聚类趋势;密度峰值

**中图分类号:** TP301.6    **文献标志码:**A

### Design of mixed data clustering algorithm based on density peak

LI Ye<sup>1,2</sup>, CHEN Yiyan<sup>1\*</sup>, ZHANG Shufen<sup>2</sup>

(1. Service Quality Specialty Committee, Chinese Association of Market Development, Beijing 100048, China;

2. Hebei Key Laboratory of Data Science and Application (North China University of Science and Technology), Tangshan Hebei 063210, China)

**Abstract:** Focusing on the issue that  $k$ -prototypes algorithm is incapable of identifying automatically the number of clusters and discovering clusters with arbitrary shape, a mixed data clustering algorithm based on searching for density peaks was proposed. Firstly, CFSFDP (Clustering by fast Search and Find of Density Peaks) clustering algorithm was extended to mixed datasets in which the distances between mixed data objects were calculated to determine the cluster centers by using CFSFDP algorithm, that is, the number of clusters was determined automatically. The rest points were then assigned to the cluster in order of their density from large to small. Secondly, the selection method of threshold and weight in the proposed algorithm was introduced. In the density formula, the threshold (cutoff distance) was extracted automatically by calculating potential entropy of data field; in the distance formula, the weight was defined through certain statistic which can measure clustering tendency of numeric datasets and categorical datasets. Finally, experimental results on three real mixed datasets show that compared with  $k$ -prototypes algorithm, the proposed algorithm can effectively improve the accuracy of clustering.

**Key words:** cluster analysis; mixed data; data field; clustering tendency; density peak

### 0 引言

随着互联网技术的迅速发展,人们采集与获取信息的能力大大提高。当海量信息出现时,人们希望从繁多复杂的数据中得到有价值的信息。聚类分析是数据挖掘中一个极其重要的分支。文献[1]指出聚类是一个把数据对象划分成子集的过程,每个子集是一个簇,使得簇中的对象彼此相似,但与其他簇中的对象不相似。

根据数据对象(数据点)每个属性值的数据类型,可把数据分为数值型数据、分类型数据(无序分类变量构成的数据)及混合型数据。很多科研工作者为保证数据分析的普适性和全面性,采用的调查数据大多为既含数值型属性又含分类型属性的混合型数据。由于面对的数据类型日益多样化,聚类分析也需要处理不同类型的数据,然而目前的聚类算法大多是针对数值型数据提出的,也有一部分可用来处理分类型数

据,但能够处理混合型数据的聚类算法却相对较少。文献[2]指出这两种属性在属性值的取值范围、分布和特点上差异较大,许多研究人员认为针对单一属性数据的传统聚类算法已不适用于混合属性数据。因此,设计更多针对混合型数据的聚类算法是聚类分析中一项极具意义的工作。

最早用于混合型数据的聚类算法是  $k$ -prototypes 算法<sup>[3]</sup>,该算法结合了  $k$ -means 和  $k$ -modes 两种方法,故而不像很多传统的聚类算法一样只能处理单一属性的数据集。另外,该算法具备了  $k$ -means 算法高效性的特点,应用十分广泛,尤其是对于大型数据集也十分有效。虽然  $k$ -prototypes 算法有很多优点且被人们广泛使用,但仍存在以下一些不足:1) 无法实现对数据分布的适应性。对于簇中的数值部分,  $k$ -prototypes 算法同  $k$ -means 算法一样只能发现球状分布的簇。2) 无法自动确定簇的个数。3) 没有考虑聚类过程中的模糊性问题。很多情况下,数据对象很可能同时分属多个簇,有时簇边界附

收稿日期:2017-08-10;修回日期:2017-09-11。    基金项目:河北省数据科学与应用重点实验室开放课题资助项目(20170320002)。

**作者简介:**李眩(1992—),女,河北保定人,博士研究生,CCF 会员,主要研究方向:机器学习、数据分析; 陈奕延(1986—),男,北京人,工程师,经济师,博士研究生,CCF 会员,主要研究方向:统计建模、技术经济; 张淑芬(1972—),女,河北唐山人,教授,硕士,CCF 会员,主要研究方向:云计算、数据安全、隐私保护。



近的数据对象的归属问题会很模糊,而  $k$ -prototypes 算法并未考虑这一点。4)对混合属性数据簇原型(中心)的表示可能造成严重的信息丢失。5)没有考虑每个属性对聚类结果影响的差异性。6)受初始值的影响很大。由于  $k$ -prototypes 算法是一种迭代算法,只能收敛于局部最优解,因此对初值的选取十分敏感。

近年来有不少研究者纷纷对上述缺点进行了不同程度的改进。针对缺点 3):文献[4]在  $k$ -prototypes 算法的基础上提出了 fuzzy  $k$ -prototypes 聚类算法,通过在  $k$ -prototypes 算法中引入模糊理论的概念,增加了数据对象分配到簇原型时的不确定性,使改进的算法具有分析模糊性和不确定性问题的能力;文献[5]中则认为常用于混合型数据模糊聚类的 fuzzy  $k$ -prototypes 算法仅仅是在原始的模糊 C 均值(Fuzzy C-Means, FCM)聚类算法中使用了不同的相异性函数,从而使其可用于同时具有数值型属性和分类型属性的混合型数据,于是该文中提出了一个全新的 FCM-type 算法,采用全概率相异性函数来处理混合属性数据,通过交叉熵使得模糊目标函数正则化,最终达到提高聚类精度的目的。针对缺点 4),文献[6]提出了 fuzzy  $k$ -prototypes 聚类算法的一种改进算法,该算法改进了簇原型的选取方式,对每个有  $p$  种不同取值的分类型属性,将其看成是一个  $p$  维的属性,迭代过程中原型的计算也要将每个分类型属性看成一个  $p$  维的属性,按每个分类型属性的可能取值在所属簇中的比例来定义簇原型,从而也间接改变了分类型属性的相异性度量方式。这样的原型选取方式包含了更多的数据信息,从而提高了聚类的精度。针对缺点 5):文献[7]提出了基于聚类相似性的算法——SBAC 算法,它是一个凝聚层次聚类算法,引入了一个相似度的度量方式来计算数据对象间的相似性;文献[8]提出了基于熵权法的针对混合属性数据的聚类算法,改进了数据对象之间距离的度量方式,利用信息熵作为各个属性的权值,从而提高了聚类的精度和稳定性;文献[9]中利用类内和类间信息熵来度量各个属性在聚类过程中的作用,由此给不同的属性分配不同的权重,从而使得数据对象可以在统一的框架下更客观地度量彼此之间以及对象与簇原型之间的相异性。针对缺点 6),文献[10]对  $k$ -prototypes 聚类算法初始点选取方法作了改进,通过对模糊  $k$ -prototypes 的分析,分别对数值型属性部分和分类型属性部分进行划分,在每个划分中选取初值,最后将两部分和在一起组成初始的簇原型。该方法降低了数据对初值的敏感度,从而减少了聚类算法的迭代次数,同时还能避免选取到相同的初始簇原型。

虽然目前国内外针对缺点 3)~6)有诸多改进,但却鲜有针对缺点 1)和 2)的改进算法,故本文主要针对这两个不足之处提出一种可用于混合型数据的新型聚类算法。2014 年,文献[11]提出一种用于数值型数据的 CFSFDP(Clustering by Fast Search and Find of Density Peaks)聚类算法,该算法具有能发现任意形状数据集且能自动确定簇数的优点,但使用范围局限于数值型数据。之后文献[12]进行了基于 CFSFDP 算法的模糊聚类研究;文献[13]提出了基于近邻距离曲线和类合并优化 CFSFDP 的聚类算法;文献[14]提出一种基于簇中心点自动选择策略的密度峰值聚类算法;文献[15]提出一种基于粒子群算法的 CFSFDP 算法。这些改进虽然提高了 CFSFDP 算法的性能,但仍然不能用于混合型数据的聚类。

因此,本文首先重新定义了混合型数据之间的距离,接着

把 CFSFDP 算法扩展到混合型数据,这样可以克服  $k$ -prototypes 算法相应的 1)、2)两个缺点,使得该算法能够自动确定簇数,并且对于任意形状的簇都有一个比较满意的聚类效果。接着对算法复杂度进行分析,并且研究了算法中的阈值  $d_c$  及权值  $\gamma$  的选取问题,分别利用数据场中的势熵和可度量数值型及分类型数据集聚类趋势的统计量来确定这两个参数。最后用文献[16]中的三个混合型数据集作为实验对象,通过和  $k$ -prototypes 算法的比较来验证针对混合型数据的寻找密度峰值算法的有效性。

## 1 寻找密度峰值的聚类算法

本文把 CFSFDP 算法扩展到混合型数据,提出一种可用于混合型数据的寻找密度峰值聚类算法,该算法的基本思想是簇中心应该同时满足以下两点:1)簇中心的密度比它周围的点的密度更大;2)簇中心离比自身密度大的点的距离较远,即不同的簇中心之间的距离相对较远。

该想法是整个聚类过程的基础,找到簇中心以后,也就自动确定了簇的个数。对于任意一个非簇中心数据点  $i$ ,认为点  $i$  跟所有比它密度更大的点中与之距离最近的那个点属于同一个簇。该算法中簇的分配在一步中完成,这与那些通过迭代来优化目标函数的算法是不同的。

### 1.1 相关定义

接下来介绍针对混合型数据的寻找密度峰值聚类算法中涉及的一些相关定义。考虑给定的混合型数据集  $S = \{x_i\}_{i=1}^N, I_S = \{1, 2, \dots, N\}$  为相应指标集,  $d_{ij} = dist(x_i, x_j)$  表示混合型数据集中点  $x_i$  和  $x_j$  之间的距离,利用以下公式来度量:

$$d_{ij} = \sum_{k=1}^{m_r} (x_{ik}^r - x_{jk}^r)^2 + \gamma \sum_{k=1}^{m_c} \delta(x_{ik}^c, x_{jk}^c) \quad (1)$$

这里当  $p = q$  时,  $\delta(p, q) = 0$ ; 当  $p \neq q$  时,  $\delta(p, q) = 1$ 。对于对象  $i$  和  $j$ ,  $x_{ik}^r$  和  $x_{jk}^r$  表示数值型属性的取值,而  $x_{ik}^c$  和  $x_{jk}^c$  表示分类型属性的取值。 $m_r$  和  $m_c$  分别表示数值型和分类型属性的个数。 $\gamma$  是分类型属性的权重。对于  $S$  中的任何一个数据点  $x_i$ ,可为其定义  $\rho_i$  和  $\delta_i$  两个量,利用这两个量可以确定出簇中心。

#### 1.1.1 计算 $\rho_i$

计算密度时常用的核函数包括截断核(Cut-off kernel)、高斯核(Gaussian kernel)和指数核(Exponential kernel),它们的定义分别如下:

$$\rho_i = \sum_j \chi(d_{ij} - d_c); \chi(x) = \begin{cases} 1, & x < 0 \\ 0, & x \geq 0 \end{cases} \quad (2)$$

$$\rho_i = \sum_j \exp\{- (d_{ij}/d_c)^2\} \quad (3)$$

$$\rho_i = \sum_j \exp\{- (d_{ij}/d_c)^2\} \quad (4)$$

式(2)~(4)中的参数  $d_c > 0$  为阈值(截断距离),需由使用者事先指定。

#### 1.1.2 计算 $\delta_i$

设  $\{q_i\}_{i=1}^N$  为  $\{\rho_i\}_{i=1}^N$  的一个降序排列下标序,满足式(5):

$$\rho_{q_1} \geq \rho_{q_2} \geq \dots \geq \rho_{q_N} \quad (5)$$

然后定义

$$\delta_{q_i} = \begin{cases} \min_{q_j < i} \{d_{q_i q_j}\}, & i \geq 2 \\ \max_{j \geq 2} \{\delta_{q_j}\}, & i = 1 \end{cases} \quad (6)$$



至此,对于 $S$ 中的每个数据点 $x_i$ ,可利用式(2)、(3)或(4)以及式(6)为其计算 $(\rho_i, \delta_i)$  $(i \in I_S)$ 。

根据 $\rho_i, \delta_i$ 的定义可知,当某个点的这两个量同时取值较大时,这个点就满足前文提到的成为簇中心的两个条件,因此可以利用以 $\rho$ 为横轴, $\delta$ 为纵轴的图(称为决策图)来判断哪些点同时具有较大的 $\rho$ 值和 $\delta$ 值,即哪些点可以被当成是簇中心。

## 1.2 算法步骤及流程

本节给出寻找密度峰值算法的具体流程,为了方便描述,先引入一些记号。仍考虑给定的混合型数据集 $S = \{x_i\}_{i=1}^N$ ,设其包含 $n_c$ ( $> 1$ )个簇。

1)  $\{m_j\}_{j=1}^{n_c}$ :各个簇中心对应的数据点编号,即 $X_{m_j}$ 为第 $j$ 个簇的中心。

2)  $\{c_i\}_{i=1}^N$ :记录数据点所属的簇,即 $c_i$ 表示 $S$ 中第 $i$ 号数据点归属于第 $c_i$ 个簇。

3)  $\{n_i\}_{i=1}^N$ : $n_i$ 表示 $S$ 中所有比 $x_i$ 的密度大的点中,与之距离最近的数据点的编号。其定义为:

$$n_{q_i} = \begin{cases} \arg \min_{q_j: j < i} |d_{q_i q_j}|, & i \geq 2 \\ 0, & i = 1 \end{cases}$$

其中 $\{q_i\}_{i=1}^N$ 的含义同式(5),表示 $\{\rho_i\}_{i=1}^N$ 的一个降序排列下标序。注意, $\{n_i\}_{i=1}^N$ 是为非簇中心的数据点而定义的,确定好簇中心后,将利用 $\{n_i\}_{i=1}^N$ 来确定非簇中心的数据点的归类。

4)  $\{h_i\}_{i=1}^N$ :数据点中“簇核心”(cluster core)和“簇光晕”(cluster halo)的标识。一个簇中的数据点可分为“簇核心”和“簇光晕”两部分,核心部分的局部密度较大,光晕(边缘)部分的局部密度较小,常说的“离群点”就分布在“簇光晕”中。这里,若 $h_i = 1$ ,则表示 $x_i$ 属于“簇光晕”中;若 $h_i = 0$ ,则表示 $x_i$ 属于“簇核心”中。

下面利用上述符号给出寻找密度峰值聚类算法的具体步骤。

### 步骤1 初始话及预处理。

1) 给定用于确定阈值(截断距离) $d_c$ 的参数 $t \in (0, 1)$ 。

2) 给定权值 $\gamma$ 。

3) 计算距离 $d_{ij}$ ,并令 $d_{ji} = d_{ij}(i < j, i, j \in I_S)$ 。

4) 确定阈值 $d_c$ 。

将上一步计算的距离 $d_{ij}(i < j)$ 共 $M = 2^{-1}N(N-1)$ 个进行升序排列,设得到的序列为 $d_1 \leq d_2 \leq \dots \leq d_M$ ,取 $d_c = d_{f(M)}$ ,其中 $f(M)$ 表示对 $M$ 四舍五入后得到的整数。

5) 按照式(2)、(3)或(4)计算 $\{\rho_i\}_{i=1}^N$ ,并生成其降序排列的下标序 $\{q_i\}_{i=1}^N$ 。

6) 计算 $\{\delta_i\}_{i=1}^N$ 及 $\{n_i\}_{i=1}^N$ 。

步骤2 确定簇中心 $\{m_j\}_{j=1}^{n_c}$ ,并利用 $\{c_i\}_{i=1}^N$ 标记簇中心所属的簇,具体如下:

$$c_i = \begin{cases} k, & \text{若 } x_i \text{ 为簇中心, 且归属于第 } k \text{ 个簇} \\ -1, & \text{其他} \end{cases}$$

### 步骤3 对非簇中心数据点进行归类。

注1:当处理某个 $x_{q_i}$ 时,若所有比 $x_{q_i}$ 密度大的点到 $x_{q_i}$ 的距离都相等,则把 $x_{q_i}$ 随机分配到一个 $x_{q_j}(j < i)$ 所属的簇即可。

注2:对非簇中心的数据点进行簇归类时,是按 $\rho$ 值从大到小进行遍历的。之所以这样做,是想借助 $\{n_i\}_{i=1}^N$ 逐层扩充每个簇。

步骤4 若 $n_c > 1$ ,则将每个簇中的数据点进一步分为“簇核心”和“簇光晕”。

1) 初始化标记 $h_i = 0(i \in I_S)$ 。

2) 确定每个簇的边界区域。这个区域中的数据点定义如下:它们属于该簇,且在与其距离不超过 $d_c$ 的范围内,存在属于其他簇的数据点。

3) 利用边界区域计算每个簇的平均局部密度上界 $\{\rho_i^b\}_{i=1}^{n_c}$ 。

4) 局部密度小于 $\{\rho_i^b\}_{i=1}^{n_c}$ 的点标识为 cluster halo。

## 1.3 算法复杂度分析

在针对混合型数据的寻找密度峰值算法中,给定阈值 $d_c$ 及权值 $\gamma$ 后,该算法的时间复杂度主要包括距离矩阵的计算,本文算法需计算 $n(n-1)/2$ 个距离值, $n$ 是数据集中的数据对象数,因此算法时间复杂度为 $O(n^2)$ 。 $k$ -prototypes 算法的时间复杂度为 $O((l+1)kn)$ ,其中 $k$ 是簇数, $n$ 是数据集中的数据对象数, $l$ 是 $k$ -prototypes 算法收敛所需的迭代次数。当 $n$ 很大时,本文算法的时间复杂度可能会远高于 $k$ -prototypes 算法。

## 2 阈值和权值的选取

从以上论述可知,新算法中可能影响最终聚类结果的因素主要有两个:1)密度公式中的阈值(截断距离) $d_c$ ;2)距离公式中的权值 $\gamma$ 。接下来需进一步讨论并验证这两个因素对聚类结果的影响,并给出一个相对合理的方式来选取这两个参数,以达到最理想的效果。

### 2.1.1 阈值 $d_c$ 的选取

对于 $d_c$ 的选取,文献[11]给出了一种一般性的方法:选取一个 $d_c$ ,使每个数据点的平均“邻居”个数大约为数据点总数的1%~4%,这里的“邻居”是指与其距离不超过 $d_c$ 的点,具体的比例通常根据研究员的经验和实际的数据集而定。

以上方法需要根据个人实际经验来确定 $d_c$ ,但通过下文的实验(这些实验的细节会在2.1.3节中介绍)可以看到,聚类结果会受到阈值 $d_c$ 的影响,而根据经验很难估计出最优的阈值 $d_c$ 。对同一数据集,若阈值的经验估计值不同,则聚类的结果可能也不同。

为了解决这一问题,使用数据场中的势熵从混合型数据集中自动提取最优的阈值 $d_c$ 。文献[17]提出用数据场去描述数据空间中对象之间共同的交互作用,同时文献[18~19]指出对于同样的参数,数据点在稠密区域有较高的势而在稀疏区域有较低的势。

可用于计算势的势函数有很多种,如高斯核、指数核、截断核等。接下来介绍如何通过数据场中的势熵计算出在不同势函数和不同数据集下的最优阈值。

#### 2.1.1 数据场

假设在数据空间 $\Omega$ 中有一个混合型数据集 $X = \{x_1, x_2, \dots, x_n\}$ 。文献[19]指出受到物理上场论的影响,将 $X$ 中的一个数据对象当成一个在给定的任务中传播它的数据分布的物理对象,这就形成了数据场。对于一个任意的点 $x_i \in \Omega(1 = 1, 2, \dots, n)$ ,场函数按如下公式定义:

$$\varphi_i = \sum_{j=1}^n m_j \times K[(x_i - x_j)/\sigma] \quad (7)$$

其中: $\sigma$ 是一个影响因子; $m_j$ 是 $x_j$ 的质量; $K(x)$ 是一个单位势函数, $x_i - x_j$ 是点 $i$ 和另一个点 $j$ 之间的方位距离。 $\sigma$ 对最终的势分布有一定的影响。 $K(x)$ 给出了数据对象把它的数据分布扩散到数据场中的规则,通常情况下 $K(x)$ 选择为高



斯核函数。

### 2.1.2 利用数据场提取最优的截断距离

在式(7)中,如果数据场是一个标量场,则  $m_j = 1$ ,当  $K(x)$  选择高斯核函数时,  $x_i - x_j$  变成  $d_{ij}$ ,  $d_{ij}$  表示点  $x_i$  和  $x_j$  之间的某种距离,那么每一个点的势  $\varphi_i$  由如下公式计算:

$$\varphi_i = \sum_{j=1}^n \exp[-(d_{ij}/\sigma)^2] \quad (8)$$

如果式(8)中的  $d_{ij}$  和新算法中  $d_{ij}$  的度量方式相同,并且如果新算法在计算密度时也选取高斯核函数,那么点  $x_i$  在数据场中的势  $\varphi_i$  和在新算法中的密度  $\rho_i$  是等价的,也就是说式(8)与式(3)是等价的。

同理,当数据场中的势函数和密度公式都取为截断核函数或指数核函数时,点  $x_i$  在数据场中的势  $\varphi_i$  和在新算法中的密度  $\rho_i$  是等价的。因此在寻找密度峰值的聚类算法中,阈值  $d_c$  的最优化问题可以转化为数据场中影响因子  $\sigma$  的最优化问题。我们希望找到一个影响因子  $\sigma$  使得随机变量的不确定性达到最小。

在信息论与概率统计中,熵<sup>[20]</sup>是随机变量不确定性的度量,熵越大,随机变量的不确定性越大。

设  $X$  是一个取有限个值的离散随机变量,其概率分布为:

$$P(X = x_i) = p_i; i = 1, 2, \dots, n \quad (9)$$

则随机变量  $X$  的熵定义为:

$$H(X) = - \sum_{i=1}^n p_i \log(p_i) \quad (10)$$

因此,可以使用熵来最优化  $\sigma$ 。对于数据集  $X$ ,如果数据场中每个点的势为  $\{\varphi_1, \varphi_2, \dots, \varphi_n\}$ ,则熵  $H$  的定义如式(11):

$$H = - \sum_{i=1}^n \frac{\varphi_i}{Z} \log(\varphi_i/Z) \quad (11)$$

其中  $Z = \sum_{i=1}^n \varphi_i$  是一个标准化因子。

由于取相同的核函数时,数据场中的影

响因子  $\sigma$  与新算法中密度公式里的阈值  $d_c$  等价,因此通过计算使熵  $H$  达到最小的影响因子  $\sigma$ ,就可以得到寻找密度峰值聚类算法中的最优阈值  $d_c$ 。

### 2.1.3 实验和分析

为了更加直观地看到选取不同阈值时的聚类结果,且为了验证基于数据场的方法提取最优阈值的合理性,本节使用具有两个数值型和一个分类型属性的混合型数据集进行模拟,数据的记录按照如下方法产生。

首先生成四组二维数据点。第一组包含四个正态分布并且包含 300 个点(如图 1(a));第二组包含七个正态分布并且包含 840 个点(如图 2(a));第三组是两个环状分布并且包含 400 个点(如图 3(a))。然后给每个点加入一个分类值,从而把这些点扩展到 3 维,分别记这三个数据集为 I、II 和 III(如图 1(b)、2(b)、3(b))(文献[21]展示了本文所有彩色原图)。

需要注意的是,一个点的分类值不能表明它是哪个类成员。事实上,这些点完全没有类,分类值仅仅代表对象在第三

维既不是连续的也不是有序的。因此对每个数据集加入分类值时,我们使分类型属性可能的取值个数等于仅考虑数值型属性时二维平面中簇的个数,这样做是为了分别利用数值型属性和分类型属性进行聚类时簇的个数相等,把两个属性放在一起考虑时也可以有一个统一的聚类簇数(在此先不考虑数值型属性和分类型属性的权重问题),同时也可以在模拟实验中看到该算法自动选取的簇数是否合理。

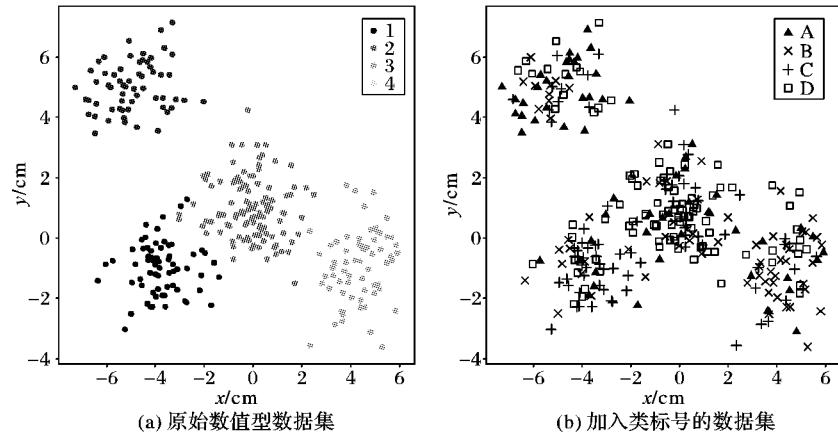


图 1 原始数值型数据集和加入类标号的数据集 I

Fig. 1 Original numeric data set and data set with the class label I

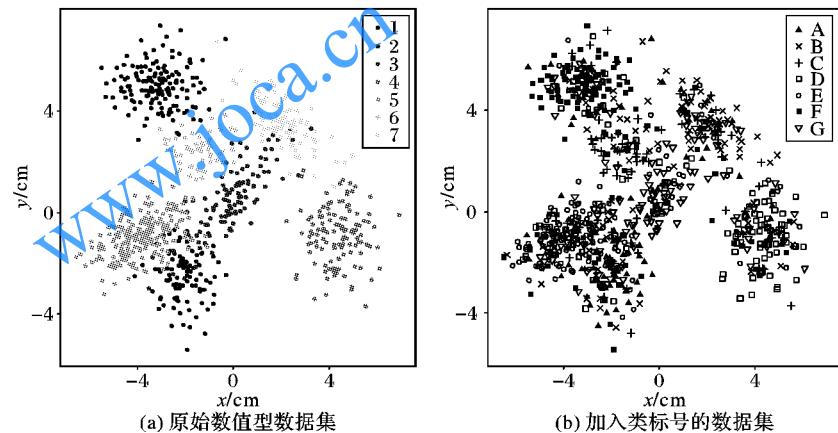


图 2 原始数值型数据集和加入类标号的数据集 II

Fig. 2 Original numeric dataset and dataset with class label II

对于第一个数据集,把每个正态分布看成一部分,并且给每部分的大多数点分配其中一个分类型属性值,使得拥有这个分类型属性值的点数与拥有其余的分类型属性值的点数大致相等。例如在图 1(b)左上角的部分中大多数点被分配属性值 A,并且这部分其余的点被分配属性值 B、C 或者 D。所有的分配都是随机的。剩下两个数据集的分类值的分配情况类似(如图 2(b)、3(b))。

对于同样的数据集选择相同类型的核函数和相同距离公式,如果聚类结果不同,说明阈值  $d_c$  对该聚类算法的聚类结果有重要的影响。在本实验中使用数据集 I,并且度量数据集中数据点的密度时选用高斯核函数。将数据集中所有对象之间的距离  $d_{ij}$  ( $i < j$ ) 按升序排列,分别取大约前 1%、2%、3% 和 4% 处对应的距离作为阈值  $d_c$ 。聚类结果如图 4 所示,其中不同的颜色表示聚类后得到的不同的簇。从图 4 中可以看到,当检测数据集、核函数类型及距离度量公式都相同时,选用不同的阈值其最终的聚类结果是不同的,因此阈值  $d_c$  的取值是影响聚类效果的一个非常关键的因素,那么对其选取



的讨论也是很必要的。

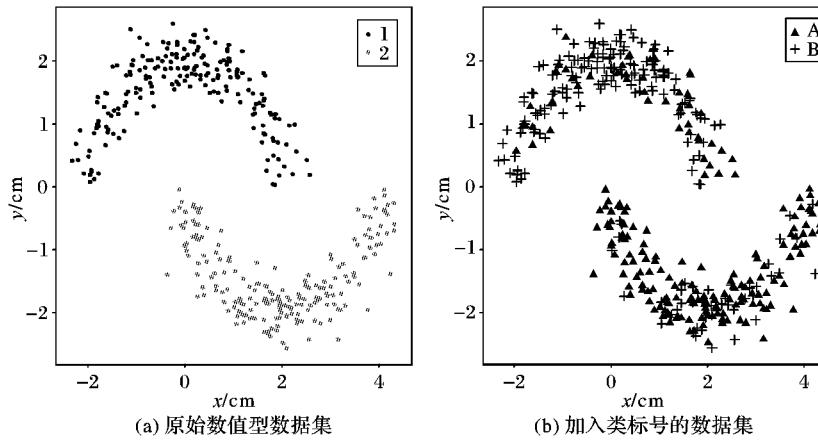


图3 原始数值型数据集和加入类标号的数据集III  
Fig. 3 Original numeric dataset and dataset with class label III

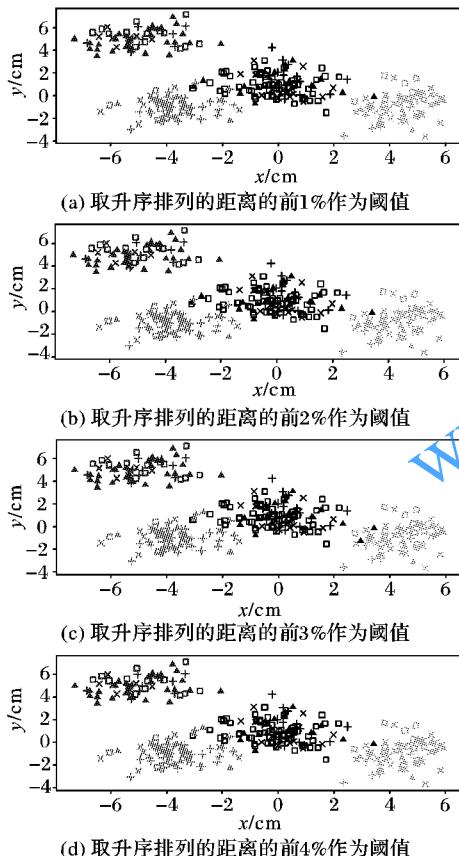


图4 数据集I对于选取不同阈值的聚类结果  
Fig. 4 Clustering results with different thresholds for dataset I

为了印证利用数据场得到阈值 $d_c$ 是一种较合理的方法,接下来仍继续使用数据集I、II和III进行比较性模拟实验,实验中计算数据场中势的势函数和计算数据集中点密度的核函数均选取高斯核函数,距离度量公式中数值型和分类型属性的权值取为 $\gamma = 2$ 。对于利用经验选取阈值的方法,将所有的 $d_{ij}$  ( $i < j$ )按升序排列,取大约前1%处对应的距离作为阈值 $d_c$ 。对于利用数据场选取阈值的方法,通过计算数据集的势熵来得到最优的影响因子,从而也就得到了最优的阈值。

由前述可知,寻找密度峰值的聚类算法可识别出数据集中的噪声点(异常值),噪声点会对聚类算法产生一定干扰,因此,须尽可能多地识别出噪声点从而提高算法的稳定性。表1

中列出了利用两种不同方法选取阈值时,算法移除噪声点的个数。从表1可以看到,对于这三个数据集,利用数据场提取阈值的方法比利用经验的方法移除的噪声点更多。同时对于这三个数据集,无论是用一般性的方法选取阈值还是利用数据场自动提取阈值都与每个数据集对应的合理簇数一致(进行模拟的数据集包含的簇数),所以至少可以说明利用数据场来确定阈值的方法不会比利用经验来确定的方法表现更差,而且它对于任何的数据集都适用,也就是说对于任意的数据集都可以自动计算出一个阈值,而不需要根据用户的经验来指定,因此该方法的适用性更广。

## 2.2 权值 $\gamma$ 的选取

对于混合型数据的聚类,距离公式中分类型属性权值 $\gamma$ 的选取也会直接影响到最终的聚类效果,通常为了简单起见,可以取 $\gamma = 1$ ,即两种属性的权重相等。文献[3]指出混合型数据距离公式中权值 $\gamma$ 的选取依赖于数值型属性的分布,即权值 $\gamma$ 的选取依赖于数值型属性的平均标准差 $\sigma$ ,但该文献并未给出具体应如何选取以及这样选取的依据。因此本节对权值 $\gamma$ 的选取提出一个新的方法。

表1 利用两种不同方法选取阈值时移除噪声点的个数

Tab. 1 Number of removed noise points with different threshold selected by two different methods

数据集	一般方法		数据场方法	
	选取的阈值	移除噪声点的个数	选取的阈值	移除噪声点的个数
I	0.55	61	0.89	68
II	0.67	119	0.81	131
III	0.05	0	0.07	76

### 2.2.1 聚类趋势

对于给定的数据集,几乎每个聚类算法都可以为该数据集返回簇,然而如果数据集中根本不存在自然的簇,那么通过聚类算法所产生的簇很可能不具有实际的意义。考虑一个完全随机结构的数据集,如数据空间中均匀分布的点,尽管聚类算法仍然可以人工地把这些点划分成簇,但这些簇是随机的,不具有任何意义。所以只有当数据集有明显的聚类趋势时,聚类算法返回的簇才有实际意义。

因此,可以提出这样的方法来确定权值 $\gamma$ :1)分别计算混合型数据集中数据对象的数值部分和分类部分的聚类趋势。数值部分的聚类趋势可以利用霍普金斯(Hopkins)统计量的变化形式来计算,分类部分的聚类趋势利用下文2.2.3节中提出的统计量来计算。2)如果数值部分的聚类趋势更明显,那么希望 $\gamma$ 小一些;如果分类部分的聚类趋势更明显,那么希望 $\gamma$ 大一些。

### 2.2.2 数值型数据集的聚类趋势

对于给定的混合型数据集,将其中的数值部分取出构成一个数值型数据集,然后将该数据集与其在数据空间中的均匀分布进行比较,以此评估该混合型数据集中数值部分的聚类趋势。霍普金斯统计量可用来度量数值型数据集的聚类趋势。

霍普金斯统计量<sup>[1]</sup>是一种空间统计量,检验空间分布的变量的空间随机性。给定混合型数据集 $D$ ,取出其中的数值型部分组成一个数值型数据集 $D^r$ ,它可以看作随机变量 $Z$ 的



一个样本,想要确定  $Z$  在多大程度上不同于数据空间中的均匀分布,可以根据文献[1]中总结的步骤来计算:

1) 均匀地从数据集  $D^r$  的空间中抽取  $n$  个点  $p_1^r, p_2^r, \dots, p_n^r$ , 也就是说,  $D^r$  的空间中的每个点都以相同的概率包含在这个样本中。对于每个点  $p_i^r (1 \leq i \leq n)$ , 找出  $p_i^r$  在  $D^r$  中的最近邻, 并令  $x_i^r$  为  $p_i^r$  与它在  $D^r$  中的最近邻之间的距离, 即:

$$x_i^r = \min_{v \in D^r} \{dist(p_i^r, v)\} \quad (12)$$

2) 均匀地从  $D^r$  中抽取(按数据编号均匀地来抽取) $n$  个点  $q_1^r, q_2^r, \dots, q_n^r$ 。对于每个点  $q_i^r (1 \leq i \leq n)$ , 找出  $q_i^r$  在  $D^r - \{q_i^r\}$  中的最近邻, 并令  $y_i^r$  为  $q_i^r$  与它在  $D^r - \{q_i^r\}$  中的最近邻之间的距离, 即:

$$y_i^r = \min_{v \in D^r, v \neq q_i^r} \{dist(q_i^r, v)\} \quad (13)$$

3) 计算可度量数值型数据集聚类趋势的统计量  $H_r$ (霍普金斯统计量的变形)。

$$H_r = \begin{cases} \frac{1}{\sum_{i=1}^n x_i^r + 1}, & \sum_{i=1}^n y_i^r = 0 \\ \frac{n}{\sum_{i=1}^n y_i^r + 1}, & \sum_{i=1}^n x_i^r = 0 \\ \frac{\sum_{i=1}^n y_i^r / \sum_{i=1}^n x_i^r}{n}, & \text{其他} \end{cases} \quad (14)$$

为了方便后边的使用, 将原始的霍普金斯统计量作一些变形。从式(14)可以看到, 对于一般的情况而言, 如果  $D^r$  是均匀的, 则  $\sum_{i=1}^n x_i^r$  会明显小于  $\sum_{i=1}^n y_i^r$ , 因而  $H_r$  会比较大; 如果  $D^r$  是高度倾斜的, 则  $\sum_{i=1}^n y_i^r$  会明显小于  $\sum_{i=1}^n x_i^r$ , 因而  $H_r$  会很小。对于  $\sum_{i=1}^n y_i^r = 0$  的情况,  $H_r$  不会超过 1,  $\sum_{i=1}^n x_i^r$  与  $\sum_{i=1}^n y_i^r$  越接近,  $H_r$  越接近 1; 对于  $\sum_{i=1}^n x_i^r = 0$  的情况,  $H_r$  一定比 1 大,  $\sum_{i=1}^n y_i^r$  与  $\sum_{i=1}^n x_i^r$  越接近,  $H_r$  越接近 1。因此, 可以看到这样的定义使得  $H_r$  整体统一起来:  $\sum_{i=1}^n y_i^r$  越小,  $H_r$  越小,  $\sum_{i=1}^n x_i^r$  越小,  $H_r$  越大, 并且  $\sum_{i=1}^n y_i^r = \sum_{i=1}^n x_i^r$  时,  $H_r = 1$ 。同时, 这样的定义可以与之后提出的可度量分类型数据集聚类趋势的统计量  $H_e$  统一起来, 且使得  $H_r$  和  $H_e$  都不等于 0, 这样  $\gamma$  的计算不会出现没有意义的情况。

### 2.2.3 分类型数据集的聚类趋势

受到可度量数值型数据集聚类趋势的霍普金斯统计量的启发, 对于分类型数据集, 同样可以提出一个用来度量其聚类趋势的统计量  $H_e$ 。

$H_e$  是一个可以评估分类型数据分布随机性的统计量。给定混合型数据集  $D$ , 取出其中的分类型部分组成一个分类型数据集  $D^e$ , 然后按以下步骤计算:

1) 找出数据集  $D^e$  中每个属性的可能取值, 从每个属性的可能取值中随机抽取一个, 构成一个从  $D^e$  空间中随机抽取的样本, 需均匀地从中抽取  $n$  个这样的样本  $p_1^e, p_2^e, \dots, p_n^e$ 。对于每个点  $p_i^e (1 \leq i \leq n)$ , 找出  $p_i^e$  在  $D^e$  中的最近邻, 并令  $x_i^e$  为  $p_i^e$  与它在  $D^e$  中的最近邻之间的距离, 即:

$$x_i^e = \min_{v \in D^e} \{dist(p_i^e, v)\} \quad (15)$$

2) 随机地从  $D^e$  中抽取  $n$  个点  $q_1^e, q_2^e, \dots, q_n^e$ 。对于每个点

$q_i^e (1 \leq i \leq n)$ , 找出  $q_i^e$  在  $D^e - \{q_i^e\}$  中的最近邻, 并令  $y_i^e$  为  $q_i^e$  与它在  $D^e - \{q_i^e\}$  中的最近邻之间的距离, 即:

$$y_i^e = \min_{v \in D^e, v \neq q_i^e} \{dist(q_i^e, v)\} \quad (16)$$

3) 计算可度量分类型数据集聚类趋势的统计量  $H_e$ :

$$H_e = \begin{cases} \frac{1}{\sum_{i=1}^n x_i^e + 1}, & \sum_{i=1}^n y_i^e = 0 \\ \frac{n}{\sum_{i=1}^n y_i^e + 1}, & \sum_{i=1}^n x_i^e = 0 \\ \frac{\sum_{i=1}^n y_i^e / \sum_{i=1}^n x_i^e}{n}, & \text{其他} \end{cases} \quad (17)$$

从式(17)可以看到, 对于一般的情况, 如果  $D^e$  是均匀的(完全随机的), 则  $\sum_{i=1}^n x_i^e$  会明显小于  $\sum_{i=1}^n y_i^e$ , 因而  $H_e$  会比较大; 如果  $D^e$  是高度倾斜的, 则  $\sum_{i=1}^n y_i^e$  会明显小于  $\sum_{i=1}^n x_i^e$ , 因而  $H_e$  会很小。对于  $\sum_{i=1}^n y_i^e = 0$  的情况,  $H_e$  不会超过 1,  $\sum_{i=1}^n x_i^e$  与  $\sum_{i=1}^n y_i^e$  越接近,  $H_e$  越接近 1; 对于  $\sum_{i=1}^n x_i^e = 0$  的情况,  $H_e$  一定比 1 大,  $\sum_{i=1}^n y_i^e$  与  $\sum_{i=1}^n x_i^e$  越接近,  $H_e$  越接近 1。因此, 可以看到这样的定义使得  $H_e$  整体统一起来:  $\sum_{i=1}^n y_i^e$  越小,  $H_e$  越小,  $\sum_{i=1}^n x_i^e$  越小,  $H_e$  越大, 并且  $\sum_{i=1}^n y_i^e = \sum_{i=1}^n x_i^e$  时,  $H_e = 1$ 。

### 2.2.4 权值 $\gamma$ 的计算

从以上分析可以看到,  $H_r$  越小说明混合型数据中数值部分的聚类趋势越明显; 同理,  $H_e$  越小说明混合型数据中分类部分的聚类趋势越明显。因此, 如果混合型数据集  $D$  的  $H_r$  越小, 则希望  $\gamma$  的取值越小, 即希望聚类时更多的考虑数值型属性; 相反, 如果混合型数据集  $D$  的  $H_e$  越小, 则希望  $\gamma$  的取值越大, 即希望聚类时更多地考虑分类型属性。因此可以按以下公式定义权值  $\gamma$ :

$$\gamma = H_r / H_e \quad (18)$$

### 2.2.5 实验与分析

为了更加直观地看到选取不同权值时的聚类结果并且验证利用聚类趋势定义权值的合理性, 使用一个新的仅有三个属性的混合型数据集 IV 进行模拟, 属性中有两个数值型和一个分类型, 这些数据记录按如下的方式产生。

首先生成一组二维数据点, 这组数据包含 400 个点且是一个正态分布(图 5(a)); 然后通过给每个点加入一个分类值把这些点扩展到三维(图 5(b))。对于这个数据集我们人为地把它分成四个部分: 用以(0,0)为坐标原点的坐标轴把区域分成左上、左下、右上、右下, 因此给每部分中的大多数点分配一个分类值, 使得拥有这个分类值的点数与拥有其余分类值的点数大致相等。举个例子, 在图 5(b)中右上角的部分大多数点被分配属性值 A 并且这部分其余的点被分配 B、C 或者 D。

模拟的主要目的是验证在该算法中数值型和分类型属性如何相互影响, 如果只考虑数值型属性, 那么可知这个二维数据只包含一个自然簇; 当加入第三维分类型属性时, 可知那些空间上离得比较近且有相同分类值的点更倾向于分到同一个簇。图 6 为数据集 IV 对于不同权值  $\gamma$  的聚类结果, 其他的参



数完全一样,其中不同的颜色表示聚类后得到的不同的簇。从图6可以看到,对于不同的权值 $\gamma$ ,聚类结果的差异很大,所以对权值选取的讨论很有必要。当权值很小( $\gamma = 0.1$ )时,分类型属性的贡献较小,虽然难以解释此时的聚类结果,但可以看到空间上离得比较近的点会被分到同一个簇;随着权值 $\gamma$ 的增加,可以看到第三维上类标号相同的点会逐渐被分到同一个簇中。对于数据集IV,当权值 $\gamma = 7.5$ 时,所有类标号相同的点被分到同一个簇。

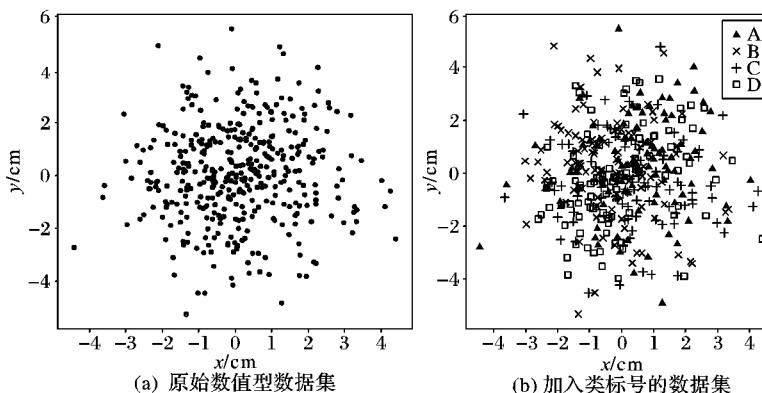


图5 原始数值型数据集和加入类标号的数据集IV  
Fig. 5 Original numeric dataset and the dataset with class label IV

接下来对数据集IV使用本文提出的方法计算一个较为合理的 $\gamma$ 。利用式(18)得到 $\gamma = 0.37$ ,聚类结果见图7,其中不同的颜色表示聚类后得到的不同的簇。从图7可以看到, $\gamma = 0.37$ 时的聚类结果与之前人为的划分较为吻合,大致分

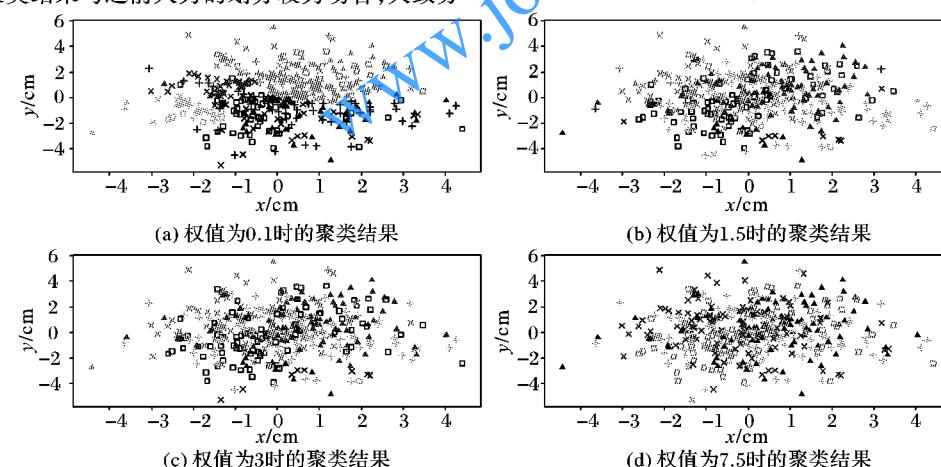


图6 数据集IV对于不同的权值的聚类结果  
Fig. 6 Clustering results of different weights in dataset IV

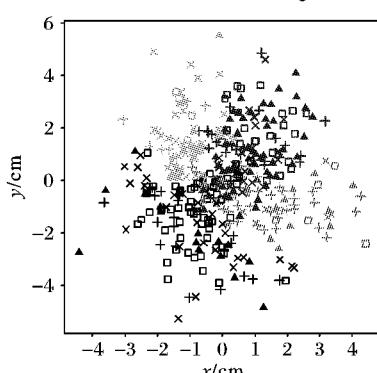


图7 数据集IV的一个合理权值的聚类结果  
Fig. 7 Clustering result of reasonable weight of the dataset IV

成左上、左下、右上、右下四个部分,因此也从另一个角度说明了 $\gamma$ 选取的合理性。

### 3 实证分析

为了进一步对提出的寻找密度峰值聚类算法进行验证,本章使用一些实际数据集进行测试,并对实验结果进行分析。

#### 3.1 数据集的选取

为了验证本文算法的有效性,选用UCI机器学习库<sup>[16]</sup>中的Cylinder Bands(以下简称Cylinder)、German Credit Data(以下简称German)和Postoperative Patient Data(以下简称Patient)这三组混合型数据集,数据描述见表2。分别在这三组数据集上对新算法和原始的k-prototypes算法进行测试,聚类结果的分析主要依靠聚类的准确率。

#### 3.2 评价指标

为了比较不同算法的正确率,引用文献[22]提出的聚类正确率定义,算法在数据集上的聚类正确率如下:

$$ac\_rate(D/f) = \frac{1}{|D|} \sum_{i=1}^k corr_{c_i} \quad (19)$$

其中: $k$ 为该数据集真实的类的个数, $corr_{c_i}$ 表示第*i*个类中被正确聚类的样本个数, $|D|$ 为该数据集的样本个数。

由此看出, $ac\_rate$ 值越大,聚类效果越好。当聚类结果与已知类标号完全一致时,取得最大值1。

表2 实际数据集描述

Tab. 2 Real dataset description

数据集	对象数	属性总数	数值型属性数	分类型属性数	类别数
Cylinder	277	39	20	19	2
German	1 000	20	7	13	2
Patient	90	8	1	7	3

### 3.3 实验结果

对于k-prototypes算法,使用式(1)作为数据对象之间的相异性度量, $\gamma$ 取为1(使得数值型属性和分类型属性有同等的地位)。对于新提出的算法,按照常规参数选取和本文新提出的参数选取分别进行验证,这样不仅可以测试新型算法



的有效性,同时可以测试新提出的参数选取方法的有效性。对于新型算法的常规参数选取,  $\gamma$  取为 1, 阈值  $d_c$  取为所有的  $d_j$  ( $i < j$ ) 按升序排列后大约前 1% 处对应的距离; 对于本文新提出的参数选取, 用 2.1 和 2.2 节的方法分别计算出每个数据集的  $d_c$  和  $\gamma$ , 所有的核函数和势函数均选用高斯核, 具体的参数取值和聚类正确率见表 3 所示。从表 3 中可以看到, 新提出的寻找密度峰值聚类算法无论是常规的参数选取还是新提出的参数选取, 在三个数据集上的测试效果都要优于原始的混合型数据聚类算法  $k$ -prototypes。对于新提出的算法, 在数据集 Cylinder 上常规的参数选取法的表现要优于新提出的参数选取法, 在其他两个数据集上新提出的参数选取法的聚类效果更好。因此可以看出, 本文新提出的算法是一个比较有效的针对混合型数据集的算法, 而对于该算法中参数的选取方法本文也给出了一个可供参考的建议。

表 3 不同算法的聚类正确率比较

Tab. 3 Comparison of clustering accuracy of different algorithms

数据集	$k$ -prototypes		本文算法 (常规的参数选取)			本文算法 (新提出的参数选取)			
	算法	$\gamma$	正确率	$d_c$	$\gamma$	正确率	$d_c$	$\gamma$	正确率
Cylinder	1	0.495	1061	1	0.693	53 058.17	5.13	0.549	
German	1	0.329	6 660.64	1	0.534	165 336.90	89.87	0.627	
Patient	1	0.437		1	1	0.598	3.13	2.26	0.667

## 4 结语

本文打破了对混合型数据进行聚类时最常用的  $k$ -prototypes 算法的框架, 把用于数值型数据的 CFSFDP 算法扩展到混合型数据, 提出了一种新的针对混合型数据的寻找密度峰值聚类算法。该算法不像  $k$ -prototypes 算法一样仅仅考虑对象到簇中心(原型)的距离, 而是考虑了对象之间的某种关系, 认为密度高的点“控制”了与它距离最近的低密度点, 这样可以使那些距离较远但原本属于同一个自然簇的对象在聚类过程中能被正确地分配, 由此实现了该算法对簇分布的适应性, 而不像  $k$ -prototypes 算法那样只能识别球状分布的簇。同时该算法在确定簇中心的同时也就自动确定了簇的个数, 而不需要像  $k$ -prototypes 算法根据经验事先指定簇数。

虽然本文提出的针对混合型数据的聚类算法在很多情况下是很有效的, 但仍存在一些不足之处, 通过算法复杂度分析可以看出寻找密度峰值聚类算法的时间复杂度较高, 当采用本文提出的方法选取阈值和权值时会进一步增加其复杂度, 因此对于超大型数据集, 其有效性还需进一步验证, 在未来的研究中需要针对该算法的复杂度问题作更深入的讨论。

## 参考文献:

- [1] HAN J W, KAMBER M, PEI J. 数据挖掘: 概念与技术 [M]. 范明, 孟小峰, 译. 3 版. 北京: 机械工业出版社, 2012: 288, 315 – 316. (HAN J W, KAMBER M, PEI J. Data Mining: Concept and Techniques [M]. FAN M, MENG X F, translated. 3rd ed. Beijing: China Machine Press, 2012: 288, 315 – 316.)
- [2] 冀进朝. 针对多维混合属性数据的聚类算法研究 [D]. 长春: 吉林大学, 2013: I. (JI J C. Research on clustering algorithms for the data with multidimensional mixed attributes [D]. Changchun: Jilin University, 2013: I.)
- [3] HUANG Z. Clustering large data sets with mixed numeric and categorical values [C]// PAKDD 1997: Proceedings of the First Pacific-Asia Knowledge Discovery and Data Mining Conference. Singapore: World Scientific, 1997: 21 – 34.
- [4] CHEN N, CHEN A, ZHOU L. Fuzzy  $k$ -prototypes algorithm for clustering mixed numeric and categorical valued data [J]. Journal of Software, 2001, 12(8): 1107 – 1119.
- [5] CHATZIS S P. A fuzzy  $c$ -mean-types algorithm for clustering of data with mixed numeric and categorical attributes employing a probabilistic dissimilarity functional [J]. Expert Systems with Application, 2011, 38(7): 8684 – 8689.
- [6] 王宇, 杨莉. 模糊  $k$ -prototypes 聚类算法的一种改进算法 [J]. 大连理工大学学报, 2003, 43(6): 849 – 852. (WANG Y, YANG L. An improved algorithm for fuzzy  $k$ -prototypes algorithm [J]. Journal of Dalian University of Technology, 2003, 43(6): 849 – 852.)
- [7] LI C, BISWAS G. Unsupervised learning with mixed numeric and nominal data [J]. IEEE Transactions on Knowledge and Data Engineering, 2002, 14(4): 673 – 690.
- [8] 孙浩军, 高玉龙, 闪光辉, 等. 基于熵权法的混合属性聚类算法 [J]. 汕头大学学报(自然科学版), 2013, 28(4): 58 – 65. (SUN H J, GAO Y L, SHAN G H, et al. A clustering algorithm based on entropy weight for mixed numeric and categorical data [J]. Journal of Shantou University (Natural Science), 2013, 28(4): 58 – 65.)
- [9] 赵兴旺, 梁吉业. 一种基于信息熵的混合型数据属性加权聚类算法 [J]. 计算机研究与发展, 2016, 53(5): 1018 – 1028. (ZHAO X W, LIANG J Y. An attribute weighted clustering algorithm for mixed data based on information entropy [J]. Journal of Computer Research and Development, 2013, 53(5): 1018 – 1028.)
- [10] 周才英, 黄龙军. 模糊  $K$ -Prototype 聚类算法初始点选取方法的改进 [J]. 计算机科学, 2010, 37(7A): 69 – 75. (ZHOU C Y, HUANG L J. Improvement of the method to choosing the initial value of fuzzy  $K$ -prototypes clustering algorithm [J]. Computer Science, 2010, 37(7A): 69 – 75.)
- [11] RODRIGUEZ A, LAIO A. Clustering by fast search and find of density peak [J]. Science, 2014, 344(6191): 1492 – 1496.
- [12] BIE R, MEHMOOD R, RUAN S, et al. Adaptive fuzzy clustering by fast search and find of density peaks [J]. Personal and Ubiquitous Computing, 2016, 20(5): 785 – 793.
- [13] 蒋礼青, 张明新, 郑金龙, 等. 快速搜索与发现密度峰值聚类算法的优化研究 [J]. 计算机应用研究, 2016, 33(11): 3251 – 3254. (JIANG L Q, ZHANG M X, ZHENG J L, et al. Optimization of clustering by fast search and find of density peaks [J]. Application Research of Computers, 2016, 33(11): 3251 – 3254.)
- [14] 马春来, 单洪, 马涛. 一种基于簇中心点自动选择策略的密度峰值聚类算法 [J]. 计算机科学, 2016, 43(7): 255 – 258, 280. (MA C L, SHAN H, MA T. Improved density peaks based clustering algorithm with strategy choosing center automatically [J]. Computer Science, 2016, 43(7): 225 – 258, 280.)
- [15] 詹春霞, 王荣波, 黄孝喜, 等. 基于改进 CFSFDP 算法的文本聚类方法及其应用 [J]. 数据分析与知识发现, 2017, 1(4): 94 – 99. (ZHAN C X, WANG R B, HUANG X X, et al. Application of text clustering method based on improved CFSFDP algorithm [J]. Data Analysis and Knowledge Discovery, 2017, 1(4): 94 – 99.)
- [16] UCI database [EB/OL]. [2017-01-20]. <http://archive.ics.uci.edu/ml/datasets.html>.

(下转第 496 页)



行套间种植,利用不同对象间的相互作用,提高模式的整体收益,具有一定应用价值。从邻近关系中计算出每个模式的表实例,并根据增益率阈值挖掘出高增益率模式,为科学指导套间种植提供理论依据。在未来的研究工作当中,可以继续研究高效的剪枝策略和基于 top- $k$  的高增益率 co-location 模式挖掘。

#### 参考文献:

- [1] 王丽珍,周丽华,陈红梅,等.数据仓库与数据挖掘原理及应用 [M].2 版.北京:科学出版社,2009:1~19. (WANG L Z, ZHOU L H, CHEN H M, et al. The Principle and Applications of Data Warehouse and Data Mining [M]. 2nd ed. Beijing: Science Press, 2009: 1~19.)
- [2] HUANG Y, SHEKHAR S, XIONG H. Discovering co-location patterns from spatial data sets: a general approach [J]. IEEE Transactions on Knowledge and Data Engineering, 2004, 16(12): 1472~1485.
- [3] YOO J S, SHEKHAR S, SMITH S, et al. A partial join approach for mining co-location patterns [C]// GIS '04: Proceedings of the 12th Annual ACM International Workshop on Geographic Information Systems. New York: ACM, 2004: 241~249.
- [4] YOO J S, SHEKHAR S, CELIK M. A join-less approach for co-location pattern mining: a summary of results [C]// ICDM '05: Proceedings of the 5th IEEE International Conference on Data Mining. Washington, DC: IEEE Computer Society, 2005: 813~816.
- [5] WANG L, BAO Y, LU J, et al. A new join-less approach for co-location pattern mining [C]// CIT 2008: Proceedings of the 8th IEEE International Conference on Computer and Information Technology. Washington, DC: IEEE Computer Society, 2008: 197~202.
- [6] 曾新,杨健.带时间约束的 co-location 模式挖掘[J].计算机科学,2016,43(2):293~296,301. (ZENG X, YANG J. Co-location patterns mining with time constraint [J]. Computer Science, 2016, 43(2): 293~296, 301.)
- [7] LIU M, QU J. Mining high utility itemsets without candidate generation [C]// CIKM '12: Proceedings of the 21st ACM International Conference on Information and Knowledge Management. New York: ACM, 2012: 55~64.
- [8] KRISHNAMOORTHY S. Pruning strategies for mining high utility itemsets [J]. Expert Systems with Applications, 2015, 42(5): 2371~2381.
- [9] TSENG V S, SHIE B-E, WU C-W, et al. Efficient algorithms for mining high utility itemsets from transactional databases [J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(8): 1772~1786.
- [10] FOURNIER-VIGER P, WU C W, ZIDA S, et al. FHM: faster high-utility itemset mining using estimated utility co-occurrence pruning [C]// ISMIS 2014: Proceedings of the 21st International Symposium Foundations of Intelligent System, LNCS 8502. Cham: Springer, 2014: 83~92.
- [11] LIN J C-W, LI T, FOURNIER-VIGER P, et al. An efficient algorithm to mine high average-utility itemsets [J]. Advanced Engineering Informatics, 2016, 30(2): 233~243.
- [12] 杨世晟,王丽珍,芦俊丽,等.空间高效用 co-location 模式挖掘技术初探[J].小型微型计算机系统,2014,35(10):2302~2307. (YANG S S, WANG L Z, LU J L, et al. Primary exploration for spatial high utility co-location patterns [J]. Journal of Chinese Computer Systems, 2014, 35(10): 2302~2307.)
- [13] 王敬华,罗相洲,吴倩.基于效用表的快速高平均效用挖掘算法[J].计算机应用,2016,36(11):3062~3066. (WANG J H, LUO X Z, WU Q. Fast high average-utility itemset mining algorithm based on utility-list structure [J]. Journal of Computer Applications, 2016, 36(11): 3062~3066.)
- [14] 江万国,王丽珍,方圆,等.领域驱动的高效用 co-location 模式挖掘方法[J].计算机应用,2017,37(2):322~328. (JIANG W G, WANG L Z, FANG Y, et al. Domain-driven high utility co-location pattern mining method [J]. Journal of Computer Applications, 2017, 37(2): 322~328.)

This work is partially supported by the National Natural Science Foundation of China (71462001), the Application Foundation Youth Project of Yunnan Provincial Science and Technology Department (2016FD071), the Project of Yunnan Provincial Education Department (2016ZZX192).

**ZENG Xin**, born in 1986, M. S., lecturer. His research interests include spatial data mining.

**LI Xiaowei**, born in 1985, Ph. D., lecturer. His research interests include information safety, computer network.

**YANG Jian**, born in 1976, Ph. D, associate professor. His research interests include cloud computing, data security, privacy protection.

(上接第 490 页)

- [17] WANG S, CAN W, LI D, et al. Data field for hierarchical clustering [J]. International Journal of Data Warehousing and Mining, 2011, 7(4): 43~63.
- [18] WANG S, CHEN Y. HASTA: a hierarchical-grid clustering algorithm with data field [J]. International Journal of Data Warehousing and Mining, 2014, 10(2): 39~54.
- [19] WANG S, WANG D, LI C, et al. Clustering by fast search and find of density peaks with data field [J]. Chinese Journal of Electronics, 2016, 25(3): 397~402.
- [20] 李航.统计学习方法 [M].北京:清华大学出版社,2012:60. (LI H. Method of Statistical Learning [M]. Beijing: Qinghua University Press, 2012: 60.)
- [21] 陈奕延.《基于密度峰值的混合型数据聚类算法设计》——聚类效果彩色图[EB/OL].[2017-09-09].[http://blog.csdn.net/dr\\_chenyiyan/article/details/77914036](http://blog.csdn.net/dr_chenyiyan/article/details/77914036). (CHEN Y Y. Design of hybrid data clustering algorithm based on density peak: Chromatic

effect diagrams in clustering [EB/OL]. [2017-09-09]. [http://blog.csdn.net/dr\\_chenyiyan/article/details/77914036](http://blog.csdn.net/dr_chenyiyan/article/details/77914036).)

- [22] AL-SHAMMARY D, KHILI I, TARI Z, et al. Fractal self-similarity measurements based clustering technique for SOAP Web messages [J]. Journal of Parallel and Distributed Computing, 2013, 73(5): 664~676.

This work is partially supported by the Open Project Program of Hebei Key Laboratory of Data Science and Application (20170320002).

**LI Ye**, born in 1992, Ph. D. candidate. Her research interests include machine learning, data analysis.

**CHEN Yiyian**, born in 1986, Ph. D. candidate, engineer, economic engineer. His research interests include statistical modeling, technological economics.

**ZHANG Shufen**, born in 1972, M. S., professor. Her research interests include cloud computing, data security, privacy protection.