结合柯西核的分类型数据密度峰值聚类算法

盛锦超,杜明晶,李宇蕊,孙嘉睿
江苏师范大学 计算机科学与技术学院,江苏 徐州 221100

摘 要:密度峰值聚类算法在处理分类型数据时难以产生较好的聚类效果。针对该现象,详细分析了其产生的原因:距离计算的重叠问题和密度计算的聚集问题。同时为了解决上述问题,提出了一种面向分类型数据的密度峰值聚类算法(Cauchy kernel-based density peaks clustering for categorical data,CDPCD)。算法首先指出分类型数据距离度量过程中有序特性(分类型数据属性值之间的顺序关系)鲜有考虑的现状,进而提出一种基于概率分布的加权有序距离度量来缓解重叠问题。通过结合柯西核函数,在共享最近邻密度峰值聚类算法基础上重新评估数据密度值,改进了密度计算和二次分配方式,增强了密度多样性,降低了聚集问题带来的影响。多个真实数据集上的实验结果表明,相较于传统的基于划分和密度的聚类算法,CDPCD都取得了更好的聚类结果。

关键词:分类型数据;有序特性;密度峰值聚类;柯西核函数;数据挖掘
文献标志码:A **中图分类号:**TP301.6 doi:10.3778/j.issn.1002-8331.2201-0440

Cauchy Kernel-Based Density Peaks Clustering Algorithm for Categorical Data

SHENG Jinchao, DU Mingjing, LI Yurui, SUN Jiarui
School of Computer Science and Technology, Jiangsu Normal University, Xuzhou, Jiangsu 221100, China

Abstract: The density peak clustering algorithm has difficulty in producing better clustering results when dealing with categorical data. To address this phenomenon, the article analyzes in detail the reasons for its generation: the overlap problem of distance calculation and the aggregation problem of density calculation. To address the above problems, this article proposes a density peak clustering algorithm for categorical data, referred to as CDPCD. The algorithm points out the ordinal feature (the order relationship between attribute values of categorical data) that rarely exists in the current categorical data distance metric process, and then proposes a weighted ordered distance measure based on probability distribution to alleviate the overlap problem. The data density values are re-evaluated by combining the method of the Cauchy kernel function on a shared nearest neighbor density peak clustering algorithm with improved density calculation and quadratic assignment, which enhances the density diversity and reduces the impact caused by the aggregation problem. Experimental results on several real datasets show that CDPCD achieves better clustering results compared to traditional division-based and density-based clustering algorithms.

Key words: categorical data; ordinal feature; density peak clustering; Cauchy kernel function; data mining

聚类[1]是一种利用数据自身属性进行归类的数据挖掘方法,可以有效处理大量无标注的数据集。聚类过程依赖数据间的距离度量,由于欧式空间的量化性质较好,因此数值型数据一般采用欧氏距离作为度量方法,但这对分类型数据并不适用。分类型数据作为一种常见的离散型数据,数值空间较为复杂。传统基于划分的聚类方法[2]在处理分类型数据过程中往往无法得到良好的划分效果,一方面考虑的距离度量相对简单,比如*K*-modes(KM)[3]算法所采用的汉明距离度量(Hamming distance measure,HDM)[4]只计算数据对应属性的差异程度,忽视了数据属性的相关性、权重等重要信息;另一方面,基于划分的聚类方法并不能有效捕捉复杂数据的数值分布,例如在属性加权*K*-modes(weighted *K*-modes,WKM)[5]、混合属性加权*K*-modes(mixed weighted *K*-modes,MWKM)[6]等算法上尽管采用了量化性质更好的距离度量方式,但其聚类效果依然有限。

基金项目:国家自然科学基金(62006104,61872168);江苏省高校自然科学基金(20KJB520012)。
作者简介:盛锦超(1997—),男,硕士研究生,CCF学生会员,研究方向为聚类分析、时间序列分析等;杜明晶(1989—),男,博士,讲师,CCF会员,研究方向为聚类分析,E-mail:dumj@jsnu.edu.cn;李宇蕊(1998—),女,硕士研究生;孙嘉睿(1997—),男,硕士研究生。
收稿日期:2022-01-27 **修回日期:**2022-03-22 **文章编号:**1002-8331(2022)18-0162-10

目前分类型数据距离度量的方法中，一般使用条件概率分布计算数据属性间的关联信息，利用信息熵计算属性的权重，比如OF[7]、Goodall[8]等，但是当前研究很少考虑到分类型数据存在的一个重要特性——有序特性[9]。事实上，分类型数据的属性可以进一步划分为有序型属性和名义型属性，其中有序是指该属性的属性值之间存在顺序关系。例如图1所示，现实情况下讲师与副教授的差距应该小于讲师与教授之间的差距，假如对这个属性进行度量，那么其对应关系的信息也应该得到保留，相反，名义型属性并没有这种顺序关系。EBDM-KM[10]是一种使用信息熵对有序特性进行度量的分类型数据聚类算法，该算法从一定程度上提高了聚类效果。HDNDW[11]同样考虑了数据属性的有序特性，并且该算法将名义型属性转换为有序型属性，统一了分类型数据的内在特性，尽管该算法的聚类表现较好，但是其内部参数计算非常复杂且算法的收敛性较差，很难应用在实际社会生产中。UDMKM[12]进一步提出了利用熵整合属性关联、权重和有序的距离度量方法，但其聚类手段仍局限于传统划分方法，聚类性能提升有限。

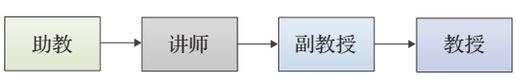
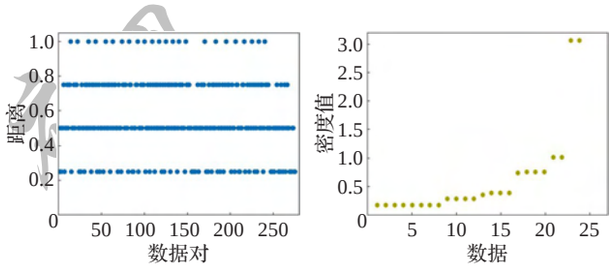


图1　职称关系

Fig.1　Title relationships

密度峰值聚类(density peaks clustering，DPC)[13]算法是一种有效应对复杂数据分布的密度聚类算法[14]，已经广泛运用在工业领域。DPC根据密度峰值点确定类中心，并将未分配数据分配给比其密度更高且距离其最近的数据所在类中，因而算法执行快速高效，但是其容易受到链式错误的影响且不能很好处理密度分布不均问题。共享最近邻密度峰值聚类算法(shared nearest neighbor density peaks clustering，SNNDPC)[15]通过共享最近邻和二次分配方式解决了数值型数据的链式错误和密度不均问题，王大刚等人[16]则利用二阶 K 近邻及多步骤的分配策略缓解了数据的不规则、密度不均匀问题，此外，丁世飞等人[17]基于度量优化，柏锷湘[18]使用自然最近邻，赵嘉等人[19]采用相互邻近度都从一定角度出发缓解甚至解决了上述问题，但是这些方法都只针对数值型数据，因而简单地将现有的分类型距离度量引入该算法中，依然无法很好地解决分类型数据的聚类问题。其中最主要的原因为在距离度量方面，现有的距离计算大多会产生“距离值的重叠问题”，例如图2(a)所示在拥有24个数据的Lenses数据集上使用HDM计算数据之间的距离，结果产生了大量相同距离结果值，这往往不利于距离远近的比较。在密度计算方面，上述提出的重叠问题会进一步导致“密度值的聚集问题”，即较多数据有相同密度值的情况，例如图2(b)所示的数据在最终分配过程中其周围数据的密度值相同，最终结果很有可能将该数据分配给一个密度更高但实际距离较远的类中，导致最终的聚类结果变差。



(a)距离值的重叠问题　(b)密度值的聚集问题

图2　分类型数据在DPC上存在的问题

Fig.2　Problems of categorical data on DPC algorithm

考虑到这些问题，本文提出一种改进的结合柯西核的分类型数据密度峰值聚类算法(Cauchy kernel-based density peaks clustering algorithm for categorical data，CDPCD)。算法首先提出了一种新的基于概率分布的加权有序距离度量方法，该方法充分考虑了分类型数据的属性相关性、属性权重和有序特性，并将有序型属性和名义型属性进行了同质计算统一，从而保证了距离值的高区分度和高信息量。其次通过结合柯西核函数[20-21]重新评估密度值，改进了SNNDPC算法中的数据密度计算和二次分配方式，使得算法不仅能够作用于分类型数据，还能有效缓解重叠和聚集问题，提高聚类性能。通过大量的对比实验和消融实验证明了CDPCD算法的有效性，CDPCD也成为了当前鲜有的面向分类型数据的密度峰值聚类算法。

1　基于概率分布的加权有序距离度量

1.1　相关定义和特性统一

本节首先给出相关定义：设 x_{ij} 为数据集合 X 中数据点 x_i 关于属性 a_j 的属性值，其中 $1\leqslant i\leqslant n, 1\leqslant j\leqslant m$ 。设 A^o 为有序型属性集合，A^n 为名义型属性集合，则 $A^o\bigcup A^n=A$ 。为了方便后续公式说明，设 O 为数据的属性值集合，$O=\{o_1,o_2,\cdots,o_m\}$ ，每个 o_j 代表 a_j 中数值的取值集合，$o_j=\{o_{j1},o_{j2},\cdots,o_{jc(j)}\}$ ，其中 $C(j)$ 表示该属性中可能取值个数，比如“性别”属性中共有“男”和“女”两个可能的属性取值，$C(j)=2$ 。

观察图3的分类型数据属性划分，从信息量层面可以认为有序型属性所提供的信息是同一内容的不同程度，而名义型属性则更侧重不同内容的信息，因此考虑属性的关联计算时难免会产生非同质的错误，为了避免这个问题，有必要将特性进行统一。有序型属性转变为名义型属性时会损失顺序信息，因此本文选择将名义型属性转换为有序型属性。转化方法的基本思想如图4所示，通过集合变换将名义型属性的每个可能取值转换为一个具有两个逻辑值的有序集合，两个逻辑值分别代表是否为当前可能取值的两种极端情况，例如“宠物”属性值中存在“猫”“狗”和“其他”三个取值，这三个取值被转换为三个“新”集合，“猫”“狗”“其他”，其中“猫”集合包含了两个极端值：是(用1表示)、否(用0表示)，同理“狗”和“其他”也包含这两个极端值。通过集合变换，名

义型属性所产生的属性集合可以被视为由多个有序型属性所组成,如上述“宠物”属性可视为由三个有序型属性组成的集合,于是可以将这个集合利用平均值等统计特征与原有的有序型属性进行计算。基于这种思想,有序型属性和名义型属性完成了同质计算统一,下节将利用这种方式辅助距离计算。

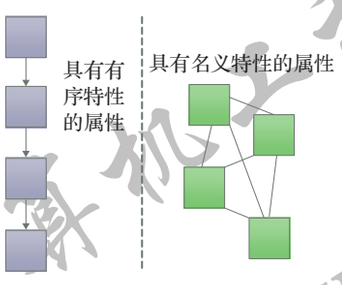


图3　分类型数据属性划分

Fig.3　Categorical data attribute division

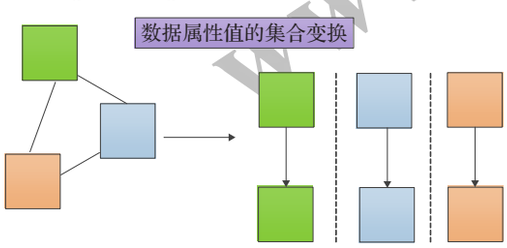


图4　名义特性变换

Fig.4　Nominal feature transformation

1.2　距离度量方法

本节将提出一个全新且完整的分类型数据距离度量,包括属性关联、权重及有序特性所反映的顺序关系。其中属性关联利用条件概率分布[22]实现,属性加权使用信息熵[10,12],而顺序关系则考虑属性值之间转换的最小移动花费代价[23]。

在属性 a_r 的属性值为 o_{rm} 条件下属性 a_s 的属性值为 o_{sp} 的条件概率如公式(1)所示:

$$p(o_{sp}|o_{rm})=\frac{\text{count}(\{x_i|x_{ir}=o_{rm}\}\bigcap\{x_i|x_{is}=o_{sp}\})}{\text{count}(\{x_i|x_{ir}=o_{rm}\})} \tag{1}$$

其中 $\{x_i|x_{ir}=o_{rm}\}$ 和 $\{x_i|x_{is}=o_{sp}\}$ 分别代表了数据点 x_i 的 a_r 和 a_s 的属性值为 o_{rm} 和 o_{sp} 的数据集合,count( )函数用于统计数据集合中数据的个数。通过公式(1),a_r 关于 a_s 所有属性值取值的条件概率分布可以用公式(2)所示的列表表示:

$$U_m^{rs}=[p(o_{s1}|o_{rm}),p(o_{s2}|o_{rm}),\cdots,p(o_{sc(s)}|o_{rm})] \tag{2}$$

传统利用 L_1 或 L_2 范式的方法计算条件概率分布会容易损失顺序关系的信息,例如在计算 $U_1=[1,0,0,0]$,$U_2=[0,1,0,0]$,$U_3=[0,0,0,1]$ 三个分布的过程中,最终得到了 $\|U_1-U_2\|_1=\|U_1-U_3\|_1=\|U_1-U_2\|_2=\|U_1-U_3\|_2$ 的结果,这与有序型属性的距离关系不符,显然第一与第二可能值之间应该比第一与第四可能值之间距离更近。本文使用了一种类似推土机算法[23]的计算方法保留了该关系,如公式(3)所示:

$$\varphi(U_m^{rs},U_h^{rs})=\sum_{t=1}^{c(s)-1}\left(\frac{\left|\sum_{g=1}^{t}(p(o_{sg}|o_{rm})-p(o_{sg}|o_{rh}))\right|\times 1}{c(s)-1}\right) \tag{3}$$

其中,U_m^{rs},U_h^{rs} 分别表示在 a_r 中取值为 o_{rm} 和 o_{rh} 关于 a_s 的条件概率分布,与公式(2)对应。方法基于条件概率分布相差较大则应该产生更多距离的基本假设,通过计算属性值之间转换的最小移动花费代价来保留离散数值的顺序关系。观察公式(3)可以发现,由于有序型属性分布各个位置取值可能不同,因而计算过程会根据前一个数值计算移动花费,数值相差越大,计算的数值包含的信息越多,最终通过移动叠加计算出的距离会更符合顺序关系的要求,例如上述提出的例子得到的结果为 $\varphi(U_1,U_2)=1/3$ 和 $\varphi(U_1,U_3)=1$。相反如果缺乏这种顺序关系,计算结果会产生相同的数值结果,必然加重距离值的重叠问题,最终可能导致聚类结果变差。进一步根据公式(3),两个数据 x_i 和 x_j 在 a_r 中关于 a_s 和整个 A 的距离度量如公式(4)与公式(5)所示:

$$dist^{rs}(x_{ir},x_{jr})=\begin{cases}\varphi(U_{x_{ir}}^{rs},U_{x_{jr}}^{rs}),s\in A^o\\ \dfrac{1}{c(s)}\displaystyle\sum_{g=1}^{c(s)}\varphi(U_{x_{ir}}^{rg},U_{x_{jr}}^{rg}),s\in A^n\end{cases} \tag{4}$$

$$dist^{r}(x_i,x_j)=\begin{cases}\dfrac{1}{m}\displaystyle\sum_{s=1}^{m}\sum_{t=\min(x_{ir},x_{jr})}^{\max(x_{ir},x_{jr})-1}dist^{rs}(t,t+1),r\in A^o\\ \dfrac{1}{m}\displaystyle\sum_{s=1}^{m}dist^{rs}(x_{ir},x_{jr}),r\in A^n\end{cases} \tag{5}$$

其中,在计算过程中使用均值完成特性的同质计算,与1.1节对应,同时出于数据属性本身的特性要求,当 $a_r\in A^o$ 时再次考虑了顺序关系。

上述讨论假设数据属性的权重一致,但实际情况中由于不同属性值的分布不同,每个属性对距离的贡献是不同的,因此有必要对属性进行加权。信息熵是一种度量权重的方法,使用信息熵对属性进行加权能够很好地衡量随机变量的不确定度,尤其是离散型随机变量,因而适合分类型数据的离散分布特征[12]。信息熵的公式如式(6)、(7)所示:

$$weigh_{a_r}=\frac{E_r}{\sum_{i=1}^{m}E_i} \tag{6}$$

$$E_r=-\frac{1}{c(r)}\sum_{i=1}^{c(r)}p(x_{*r}=i)\log(p(x_{*r}=i)) \tag{7}$$

其中,$x_{*r}=o_r$,添加 $-\frac{1}{c(r)}$ 主要为了防止 o_r 的数值过多而导致权重偏低的问题。最终两个分类型数据基于概率分布的加权有序距离计算方式如公式(8)所示:

$$dist(x_i,x_j)=\sum_{r=1}^{m}weigh_{a_r}dist^{r}(x_i,x_j) \tag{8}$$

容易验证公式(8)符合以下距离性质,是一种距离度量方法。

$$dist(x_i,x_j)\geqslant 0 \tag{9}$$

$$dist(x_i,x_j)=0\Leftrightarrow x_i=x_j \tag{10}$$

$$dist(x_i,x_j)=dist(x_j,x_i) \tag{11}$$

$$dist(x_i,x_j)\leqslant dist(x_i,x_k)+dist(x_k,x_j) \tag{12}$$

图5展示了使用HDM、OF、Goodall和本文提出的距离度量在Lenses数据集上的距离值集合,可以看到本文方法得到了更多的距离差异,保证了距离值的高区分度和高信息量,有效降低了重叠问题。

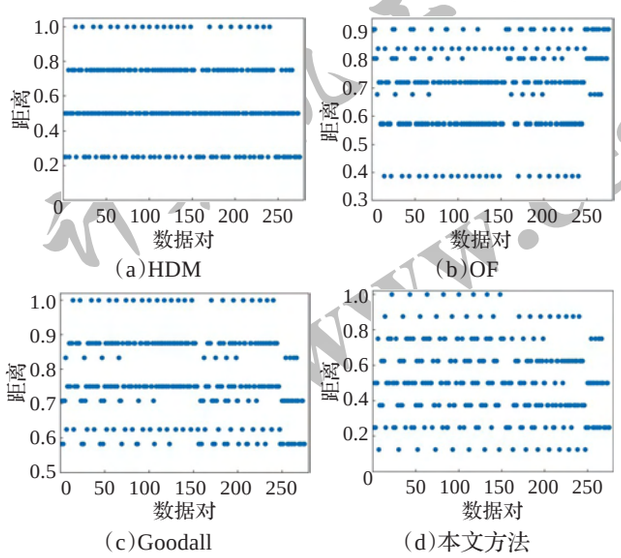


图5　距离度量结果

Fig.5　Distance measurement results

2　结合柯西核函数的聚类算法

2.1　柯西核函数

尽管提出的距离度量增加了距离差异,但是分类型数据的密度计算仍然会产生聚集问题,需要一种考虑密度差异的评估方法。

核函数方法[21]能在密度计算的过程中考虑到更多数据对待评估数据的影响,比如常见的高斯核,但是高斯核的密度计算快速衰减特性使其容易忽视某些较远邻居的贡献。根据分类型数据相对分散的分布特点,这样的结果反而不利于区分密集区域的密度峰值点。相反如图6所示,柯西核函数拥有更长的函数尾特征,因而空间中处于不同方向的密度峰值点会扩大考虑不同的较远数据的影响,从而更有利于数据密度计算结果的多样性。与此同时,柯西核函数的特性使其没有降低原先密集区域对密度计算的贡献,例如图7(a)、(b)所示经过排序后的Lenses数据集和Post数据集密度值图中,更平滑的蓝色点表示用柯西核计算出的密度值,而绿色点则使用了高斯核,可以看到绿色点区域中的密度值仍较为聚集,而蓝色点区分度较好,呈现出一个良好的上升趋势,有助于缓解聚集问题。

柯西核函数的另一个优势是运算方法简单,计算速度快。综上选择柯西核函数作为密度评估方法具有合理性和有效性。

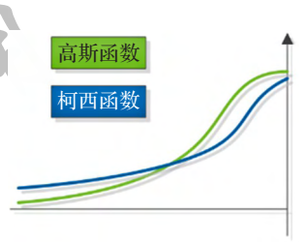


图6　高斯函数和柯西函数

Fig.6　Gaussian function and Cauchy function

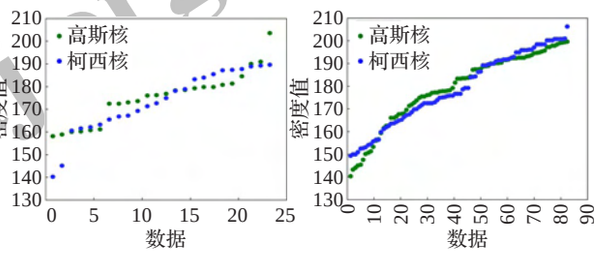


图7　数据集密度排序图

Fig.7　Density sorting diagram of dataset

两个数据结合柯西核函数的计算公式如式(13)所示,其中 υ_i 是关于 n 的超参数,本文使用 x_i 与距离 x_i 第 K 近的数据之间的距离作为 υ_i。

$$f^c(x_i,x_m)=\left(\frac{dist(x_i,x_m)^2}{\upsilon_i^2}+1\right)^{-1} \tag{13}$$

2.2　聚类算法

本节将介绍提出的CDPCD算法,出于算法描述考虑,首先简要介绍原始算法(SNNDPC)的共享最近邻、相似度矩阵和二次分配方式[15]。

共享最近邻是指两个数据最近 K 个数据点中相同数据点的集合,记为 $SNN(x_i,x_j)=\Gamma(x_i)\cap\Gamma(x_j)$,其中 $\Gamma(x_i)$ 表示距离 x_i 最近的 K 个数据点集合。

相似度矩阵是一个 n 维方阵,方阵中的值表示对应数据之间的局部相似度,计算过程如公式(14)所示:

$$Sim(x_i,x_j)=\begin{cases}\dfrac{\left|SNN(x_i,x_j)\right|^2}{\sum\limits_{p\in SNN(x_i,x_j)}(dist(x_i,x_p)+dist(x_j,x_p))},\\ \quad \text{iff } x_i,x_j\in SNN(x_i,x_j)\\ 0,\text{otherwise}\end{cases} \tag{14}$$

二次分配方式将数据按照密度比例划分为不可避免从属点和可能从属点两类,并在划分过程中对数据进行相应的二轮分配,其中第一轮分配通过划分依据将不可避免从属点进行归类,第二轮分配则通过循环统计可能从属点最近邻中已分配数据的归属进行归类。

在密度计算上,CDPCD定义的相似度矩阵在公式(14)的基础上结合了柯西核函数,基于的改进思想与2.1节对应,如公式(15)。同时关于计算数据 x_i 密度峰值参数的 ρ_{x_i} 和 δ_{x_i} 如公式(16)和公式(17)所示,其中在计算 δ_{x_i} 的过程中同样结合了柯西核函数。

$$Sim^c(x_i,x_j)=\begin{cases}\dfrac{\left|SNN(x_i,x_j)\right|^2}{\sum\limits_{p\in SNN(x_i,x_j)}(f^c(x_i,x_p)+f^c(x_j,x_p))}, \\ \quad \text{iff } x_i,x_j\in SNN(x_i,x_j) \\ 0, \text{otherwise}\end{cases} \tag{15}$$

$$\rho_{x_i}=\sum_{x_j\in\Gamma(x_i)}Sim^c(x_i,x_j) \tag{16}$$

$$\delta_{x_i}=\min_{l:\rho_{x_l}>\rho_{x_i}}\left\{dist(x_i,x_l)\left[\sum_{x_a\in\Gamma(x_i)}f^c(x_i,x_a)+\sum_{x_b\in\Gamma(x_l)}f^c(x_l,x_b)\right]\right\} \tag{17}$$

在分配方式上,CDPCD保留了从属点的命名及二次分配基本流程,但是划分依据和分配过程不同。其中划分依据方面,原始算法仅考虑数据之间共享一半数据点就划分为不可避免从属点,因此算法容易受到参数 K 的影响,过少或过多的 K 都会降低算法性能。此外分类型数据的数值分布在多数情况下会出现部分数据点集中但又与其他同类数据点分散的问题,如图8所示,同一个类中出现了两个集中的数据组,甚至有些数据是重叠的(图中放大部分),因此重叠部分的数据点很可能由于共享最近邻重合导致无法被划分为不可避免从属点,最终产生两个后果,其一是在第二轮分配过程中被分配,从而产生一定的花费,其二是最终因为遗漏进而降低算法性能。CDPCD考虑 K 与柯西核函数结合计算出的距离值并根据已分配数据点的最近邻平均距离值将数据进行划分,如定义1所示。这种划分依据会依靠数据本身最近邻点分布产生的均值作为阈值,这种阈值会保留相近点和重叠点,同时过滤远处的数据点,因而会尽可能将数据组合并,到达预期效果。与此同时CDPCD的划分依据会因为柯西核函数的加入而降低对 K 的敏感度,有利于算法的稳定。

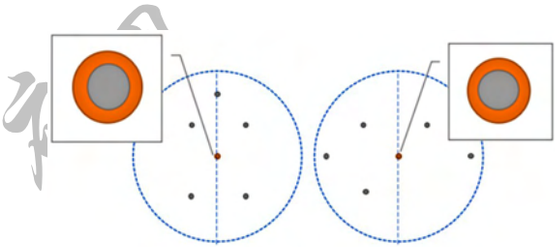


图8　分类型数据存在的分布问题

Fig.8　Distribution problems with categorical data

定义1(不可避免从属点和可能从属点)设 x_i 为已分配数据,其 $\Gamma(x_i)$ 中的任意数据点 x_l 如果满足 $\sum\limits_{x_k\in\Gamma(x_i)}f^c(x_i,x_k)\Big/K\geqslant f^c(x_i,x_l)$ 则称该数据点 x_l 是 x_i 的不可避免从属点,反之则为可能从属点。

分配过程方面,原始算法在第二轮分配过程中会出现两个问题,即重复统计某一个已分配数据点和可能从属点出现多次统计最近邻数据点所属类却未被分配的情况,为了避免这两个问题,CDPCD在第二轮分配过程中通过考虑柯西核函数计算出的密度值影响,将可能从属点依照公式(13)计算出的距离值升序排序,并使用循环依次将可能从属点 x_i 分配到 $\Gamma(x_i)$ 中距离最近的数据所在类中,在最近邻数据都未分配情况下将搜索空间变为整个数据集。这种分配过程使得每个可能从属点只计算一次就进行分配,从而能够有效解决重复统计却未被分配问题,加快第二轮的分配速度,同时基于出现就分配的原则,改进的分配方式也不会出现重复统计已分配数据点的情况。3.3.1小节将通过实验证明分配过程在有效提速的同时并没有损失聚类效果的结论。

图9给出了CDPCD算法的流程示意图,其算法步骤如下所示:

算法1　CDPCD

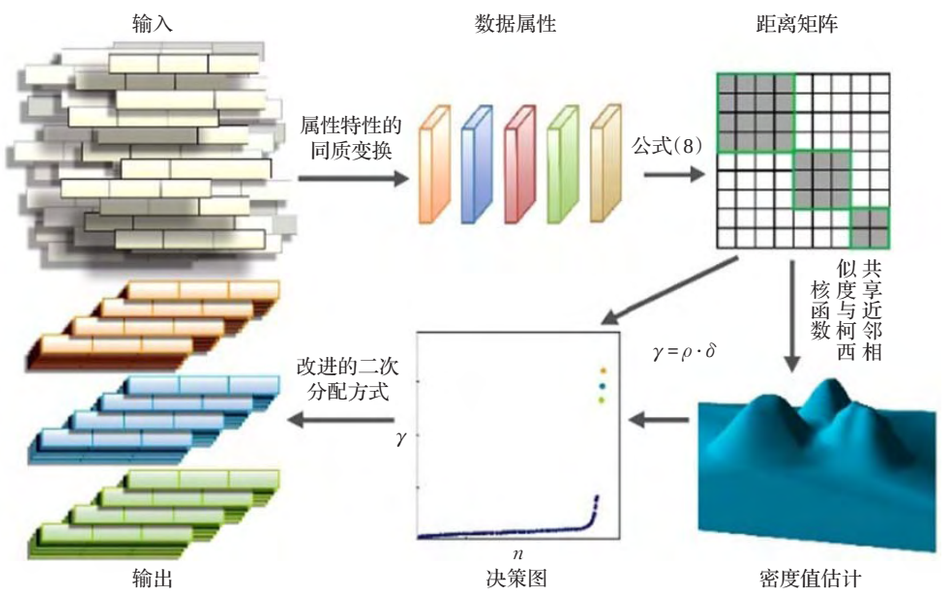


图9　算法流程示意图

Fig.9　Algorithm flow diagram

输入:数据集 X,参数 K 。

输出:聚类结果 R 。

步骤1 根据公式(8)计算 X 中任意两个数据之间的距离。

步骤2 根据公式(16)与公式(17)计算数据的 ρ_{x_i} 和 δ_{x_i} 。

步骤3 通过决策图选取密度峰值点,对数据进行第一轮分配,根据定义1将不同密度峰值点的不可避免从属点归类。

步骤4 对标记的可能从属点进行第二轮分配,计算可能从属点结合柯西核函数的距离值,并升序排列;从队列中选取每个数据,按照最近邻原则分配,如果最近邻集合中数据均未分配,则分配到距离最近的当前已经分配的数据所在类中。

步骤5 记录数据分配结果,返回 R 。

算法1的时间复杂度主要在步骤1和步骤2上。步骤1中的距离度量主要取决 $C(j)$ 的最大值,如果假设为 q ,则计算两个数据距离的时间复杂度为 $O(mq^2)$,整个数据集的距离计算复杂度为 $O(mq^2n^2)$ 。步骤2中柯西核函数产生的计算为线性运算,故不存在多余运算代价,因此时间复杂度仅与 K、m 和 n 三个参数有关,为 $O((K+m)n^2)$ 。步骤3和步骤4的计算依赖步骤1和步骤2,因此复杂度为 $O(1)$,故算法1的总时间复杂度为 $O((mq^2+K+m)n^2)$ 。与SNNDPC的 $O((K+m)n^2)$ 时间复杂度相比较,算法1中的 mq^2 绝大部分情况下会远小于 n^2 即属性值取值个数远小于数据量,因此算法1和SNNDPC的时间复杂度在数量级上是一致的,由于二次分配的优化,算法1在处理分类型数据集上会有额外增速优势。与 K-modes的 $O(E(kmn))$ 时间复杂度和HDNDW的 $O(E(kmqn+mn+mq^2k))$ 时间复杂度(注意这里的 E 为迭代次数,k 为真实标签值)等划分聚类方法相比,算法1不会受到随机结果影响,并且可以将距离计算和聚类过程分离,具有更好的灵活性。

3 实验部分

本文选用了表1所示的15组UCI公开真实数据[11]作为进行各种实验的基本数据集,其中表1的Attribute列定义为有序型属性列+名义型属性列。

实验使用KM[3]、WKM[5]、MWKM[6]、EBDM-KM[10]、HDNDW[11]、UDMKM[12]与DPC[13]作为对比算法进行对比。按照文献及对比实验要求,KM、WKM、MWKM和DPC使用HDM作为距离度量。此外,为了验证提出的距离度量以及柯西核函数方法在聚类上的有效性和重要性,在SNNDPC[15]和CDPCD算法基础上增加了相应的消融实验。最后就CDPCD算法在划分依据和分配过程的时间提速方面进行了验证,并针对参数 K 对算法的影响和柯西函数相较于其他核函数的优势进行了实验。

本文参考的聚类评价指标[24-25]分别为标准化互信息NMI、调整兰德系数ARI和福克斯马洛斯指数FMI,其中NMI和FMI的取值范围是[0,1],ARI的取值范围是[−1,1],三个指标的数值越大,聚类的效果越好。

表1 实验数据集

Table 1 Experimental data sets

Data set	Instance	Attribute	Class
Lenses	24	2+2	3
Primary	123	2+5	12
Lym	148	3+15	4
Breast	277	4+5	2
Flare	323	7+2	6
Dermatology	358	33+1	6
Soybean	47	0+16	4
Zoo	101	0+16	7
House	232	0+16	2
Tic	958	0+9	2
Post	86	8+0	2
Wisconsin	689	9+0	2
Lecturer	1 000	4+0	5
Social	1 000	10+0	4
Car	1 728	6+0	4

本文使用的编程环境为Matlab2021b和Python3.6。算法的参数设置方面,KM、WKM、MWKM、EBDM-KM、UDMKM和HDNDW使用数据集的真实类别数目,由于这些方法具有随机性,因此会随机实验50次取平均值作为最终结果;DPC算法的阈值参数设置为2%;其余算法存在最近邻参数 K ,本文选择重复实验并取NMI最优结果值,K 的最终设置范围在2~40。

3.1 对比实验

表2、表3和表4给出了各个算法在数据集上取得的聚类效果。

从表中可见在有序和名义特性都存在的数据集上(例如Lym、Flare等),进一步考虑有序特性的EBDM-KM、HDNDW和UDMKM算法比传统基于划分的MWKM、WKM和KM算法的聚类表现要好,这说明数据属性的进一步划分确实可以增加数据的信息量,提高聚类效果。CDPCD算法不仅考虑了有序特性,而且统一了特性的同质计算,因此使其在Flare、Dermatology有序特征比例较高的数据集上比HDNDW、UDMKM等先进的分类型数据聚类算法更有竞争力。

在Soybean、Zoo等纯名义型属性的数据集上,CDPCD通过将名义特性转化为有序特性的手段,增加了距离信息量,缓解了距离度量的重叠问题,因而同样取得了较好的聚类结果。在纯有序型属性的数据集上,CDPCD通过密度值计算结果的多样性有效降低了聚集问题并减少了错误分配,从而更好地处理了基于划分算法不能有效捕捉复杂分布的问题,使得算法的性能更强。

与DPC的对比结果中CDPCD展示了其作为密度聚类算法却能够克服数据的距离值重叠问题和密度值聚集问题从而对分类型数据进行聚类的可能性和有效

表2 NMI结果

Table 2 NMI result

Data sets	KM	WKM	MWKM	EBDM-KM	HDNDW	UDMKM	DPC	CDPCD
Lenses	0.237	0.217	0.248	0.251	0.297	0.257	0.263	**0.324**
Primary	0.418	0.425	0.433	0.410	0.423	0.399	0.379	**0.431**
Lym	0.091	0.150	0.163	0.201	0.242	0.120	0.172	**0.296**
Breast	0.004	0.001	0.044	0.008	0.064	0.011	0.049	**0.092**
Flare	0.223	0.255	0.211	0.312	0.310	0.233	0.354	**0.473**
Dermatology	0.665	0.594	0.728	0.671	0.814	0.802	0.531	**0.921**
Soybean	**1.000**	0.828	0.889	**1.000**	0.893	0.798	0.835	**1.000**
Zoo	0.723	0.708	0.751	0.729	0.741	0.743	0.668	**0.931**
House	0.447	0.476	0.501	0.521	**0.541**	0.511	0.530	0.539
Tic	0.024	0.011	**0.027**	0.012	0.007	0.017	0.000	0.019
Post	0.007	0.005	0.009	0.002	0.008	0.002	0.000	**0.014**
Wisconsin	0.588	0.400	0.620	0.644	0.768	0.682	0.567	**0.821**
Lecturer	0.077	0.068	0.064	0.094	0.142	0.087	0.065	**0.179**
Social	0.085	0.067	0.025	0.086	0.168	0.100	0.049	**0.184**
Car	0.127	0.055	0.133	0.106	0.126	0.090	0.081	**0.160**

表3 ARI结果

Table 3 ARI result

Data sets	KM	WKM	MWKM	EBDM-KM	HDNDW	UDMKM	DPC	CDPCD
Lenses	0.101	0.104	0.124	0.114	0.227	0.167	0.168	**0.239**
Primary	0.167	0.164	0.177	0.150	0.162	0.214	0.126	**0.285**
Lym	0.069	0.078	0.116	0.133	0.172	0.091	0.058	**0.308**
Breast	0.017	−0.002	0.098	0.009	0.139	0.027	0.083	**0.176**
Flare	0.140	0.189	0.091	0.283	0.372	0.177	0.230	**0.384**
Dermatology	0.455	0.459	0.561	0.536	0.814	0.660	0.387	**0.912**
Soybean	**1.000**	0.823	0.894	**1.000**	0.938	0.893	0.717	**1.000**
Zoo	0.588	0.576	0.598	0.578	0.701	0.526	0.411	**0.962**
House	0.548	0.558	0.575	0.599	0.602	0.587	0.600	**0.613**
Tic	0.015	0.014	0.026	0.027	0.020	0.009	−0.001	**0.048**
Post	−0.015	−0.012	0.014	−0.004	−0.003	−0.012	0.003	**0.026**
Wisconsin	0.678	0.460	0.737	0.718	0.862	0.787	0.398	**0.886**
Lecturer	0.050	0.043	0.045	0.063	0.127	0.070	0.046	**0.145**
Social	0.040	0.046	0.020	0.058	0.112	0.077	0.028	**0.132**
Car	0.023	0.026	0.027	0.029	0.118	0.054	0.075	**0.129**

表4 FMI结果

Table 4 FMI result

Data sets	KM	WKM	MWKM	EBDM-KM	HDNDW	UDMKM	DPC	CDPCD
Lenses	0.416	0.440	0.424	0.399	0.499	0.433	0.486	**0.563**
Primary	0.306	0.254	0.272	0.287	0.322	0.305	0.220	**0.443**
Lym	0.406	0.430	0.530	0.495	0.523	0.409	0.561	**0.582**
Breast	0.570	0.540	0.634	0.540	0.642	0.564	**0.767**	0.656
Flare	0.345	0.355	0.360	0.422	0.487	0.364	0.390	**0.508**
Dermatology	0.572	0.573	0.713	0.691	0.865	0.725	0.517	**0.930**
Soybean	**1.000**	0.915	0.934	**1.000**	0.963	0.927	0.803	**1.000**
Zoo	0.678	0.667	0.686	0.802	0.767	0.628	0.533	**0.972**
House	0.773	0.779	0.792	0.800	**0.810**	0.793	0.800	0.801
Tic	0.502	0.530	0.539	0.543	0.543	0.522	0.522	**0.572**
Post	0.554	0.556	0.615	0.551	0.579	0.561	0.686	**0.688**
Wisconsin	0.863	0.771	0.732	0.912	0.937	0.905	0.579	**0.947**
Lecturer	0.315	0.281	0.281	0.286	0.306	0.292	0.316	**0.414**
Social	0.333	0.324	**0.397**	0.339	0.337	0.345	0.346	0.383
Car	**0.477**	0.393	0.453	0.390	0.446	0.415	0.464	0.466

性，CDPCD通过新的距离度量和结合柯西核函数的方法让密度值的计算更符合分类型数据的需要，比如具有较高区分度的密度多样性等特征，这也为分类型数据的密度聚类方法提供了全新的解决思路。

3.2 消融实验

本节将通过消融实验进一步验证提出的基于概率分布的加权有序距离度量和结合柯西核函数方法计算密度值的重要性和有效性。实验使用SNNDPC和CDPCD作为两个基础算法，通过“控制变量”手段令SNNDPC算法的距离度量使用本文提出的方法，并将方法记作CSNNDPC；CDPCD算法的距离度量改为使用HDM、OF和Goodall，分别记作CDPCD(HDM)、CDPCD(OF)和CDPCD(Goodall)。

表5给出了最终的聚类指标结果，其中每条数据集的指标结果的下一行Δ(%)表示CDPCD与每个通过“控制变量”验证的算法在该数据集上对应指标的提高比，例如CDPCD在Primary数据集上相较于CSNNDPC的NMI提高0.567即56.7%。表5的最后两行Avg和AvgΔ(%)分别表示各个验证算法在所有数据集上的单一指标平均值和CDPCD相较于所有验证算法的单一指标在所有数据集上的平均提高比，例如CDPCD在所有的数据集上相较于CSNNDPC的NMI平均提高0.642即64.2%。

图10给出了三个聚类评价指标在各个验证算法和CDPCD算法上所有数据集上平均值的可视化结果。结果可见各个验证算法均没有达到CDPCD的平均聚类效果，CSNNDPC在缺少了柯西核函数增强密度多样性的效果后，其聚类效果明显降低。正如2.1节所述，柯西核函数提供的平滑、有差别的密度值计算结果能够很好地区分不同数据对所属类的贡献差别。此外这种密度多

表5　聚类结果

Table 5　Clustering results

Index	CSNNDPC			CDPCD(HDM)			CDPCD(OF)			CDPCD(Goodall)		
	NMI	ARI	FMI	NMI	ARI	FMI	NMI	ARI	FMI	NMI	ARI	FMI
Lenses	0.140	0.041	0.358	0.157	0.108	0.448	0.157	0.107	0.448	0.027	0.013	0.337
Δ/%	1.314	4.829	0.573	1.064	1.213	0.257	1.064	1.234	0.257	11.000	17.385	0.671
Primary	0.275	0.037	0.226	0.371	0.094	0.202	0.371	0.095	0.201	0.278	0.087	0.276
Δ/%	0.567	6.703	0.960	0.162	2.032	1.193	0.162	2.000	1.204	0.550	2.276	0.605
Lym	0.219	0.169	0.460	0.221	0.221	0.502	0.221	0.221	0.502	0.161	0.099	0.433
Δ/%	0.352	0.822	0.265	0.339	0.394	0.159	0.339	0.394	0.159	0.839	2.111	0.344
Breast	0.079	0.146	0.642	0.048	0.060	0.634	0.001	0.001	0.530	0.005	0.014	0.560
Δ/%	0.165	0.205	0.022	0.917	1.933	0.035	91.000	175.000	0.238	17.400	11.571	0.171
Flare	0.413	0.367	0.451	0.365	0.266	0.413	0.366	0.256	0.413	0.030	0.015	0.455
Δ/%	0.145	0.046	0.126	0.296	0.444	0.230	0.292	0.500	0.230	14.767	24.600	0.116
Dermatology	0.875	0.846	0.883	0.856	0.802	0.851	0.756	0.706	0.750	0.766	0.658	0.741
Δ/%	0.045	0.078	0.053	0.068	0.137	0.093	0.209	0.292	0.240	0.193	0.386	0.255
Soybean	0.948	0.953	0.965	0.893	0.891	0.918	0.750	0.622	0.719	0.645	0.390	0.569
Δ/%	0.055	0.049	0.036	0.120	0.122	0.089	0.333	0.608	0.391	0.550	1.564	0.757
Zoo	0.908	0.924	0.955	0.809	0.848	0.885	0.789	0.748	0.799	0.667	0.582	0.709
Δ/%	0.025	0.041	0.018	0.151	0.134	0.098	0.180	0.286	0.217	0.396	0.653	0.371
House	0.520	0.611	0.727	0.493	0.510	0.720	0.548	0.621	0.720	0.493	0.573	0.686
Δ/%	0.037	0.003	0.102	0.093	0.202	0.113	−0.016	−0.013	0.113	0.093	0.070	0.168
Tic	0.005	0.014	0.567	0.005	0.004	0.425	0.005	0.001	0.535	0.002	0.001	0.538
Δ/%	2.800	2.429	0.009	2.800	11.000	0.346	2.800	47.000	0.069	8.500	47.000	0.063
Post	0.004	0.002	0.540	0.005	0.030	0.565	0.007	0.030	0.666	0.011	0.005	0.757
Δ/%	2.500	12.000	0.274	1.800	−0.133	0.218	1.000	−0.133	0.033	0.273	4.200	−0.091
Wisconsin	0.678	0.737	0.876	0.667	0.763	0.837	0.767	0.860	0.938	0.788	0.874	0.939
Δ(%)	0.211	0.202	0.081	0.231	0.161	0.131	0.070	0.030	0.010	0.042	0.014	0.009
Lecturer	0.147	0.069	0.403	0.051	0.023	0.309	0.051	0.023	0.309	0.064	0.055	0.355
Δ/%	0.218	1.101	0.027	2.510	5.304	0.340	2.510	5.304	0.340	1.797	1.636	0.166
Social	0.098	0.058	0.374	0.042	0.005	0.380	0.003	0.001	0.367	0.061	0.019	0.402
Δ/%	0.878	1.276	0.024	3.381	25.400	0.008	60.333	131.000	0.044	2.016	5.947	−0.047
Car	0.121	0.104	0.464	0.101	0.062	0.450	0.003	0.002	0.435	0.006	0.004	0.464
Δ/%	0.322	0.240	0.004	0.584	1.081	0.036	52.333	63.500	0.071	25.667	31.250	0.004
Avg	0.362	0.339	0.593	0.339	0.312	0.569	0.320	0.286	0.555	0.267	0.226	0.548
AvgΔ/%	0.642	2.002	0.172	0.968	3.295	0.233	14.174	28.467	0.241	5.606	10.044	0.238

注：Δ表示在各聚类指标上CDPCD相对于相应验证算法提高的百分比。

样性的特征不仅仅体现在解决密度计算聚集问题的柯西核函数上,由于密度计算依赖距离度量,三个控制距离度量方法的算法得到的聚类结果较差,说明了提出的距离度量方法在后续计算数据密度值和进行聚类的过程中起到了关键作用,尤其在纯有序型的数据集上NMI和ARI的结果平均下降非常大,证明缺少了有序特性的距离度量方法后,聚类算法会在运行过程中损失许多有用的顺序信息,导致性能变差。

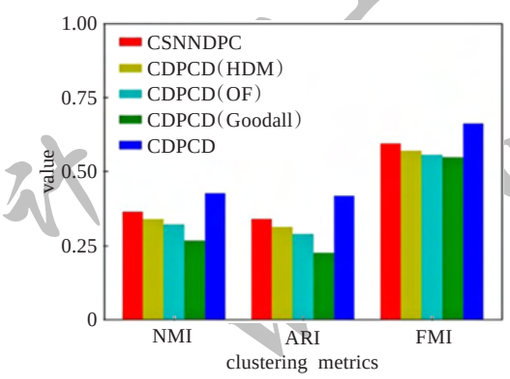


图10 验证算法和CDPCD在数据集上的平均聚类效果

Fig.10 Average clustering effect of validation algorithm and CDPCD on dataset

3.3 其他实验

3.3.1 时间对比实验

本小节将对2.2节中算法1的分配过程加速且不损失算法性能的结论进行实验验证,为了更好对比实验结果,选择Wisconsin、Tic、Lecturer、Social和Car五个较大的数据集进行实验,并将两个算法的参数 K 设置为相同值,最后统计聚类指标NMI和运行时间Time。表6给出了最终的结果,其中每个数据集分别对原始算法(Mode列标为original)和改进的算法(Mode列标为now)进行实验。从表中可见,相较于原始算法,CDPDC在基本不损失聚类性能的前提下均有一定的提速,说明该分配方式更加符合分类型数据的特点,能够提高运行速度,同时在大多数数据集上NMI指标得到了提高也侧面说明了算法改进能够提高一定的聚类性能。

表6 运行时间实验结果

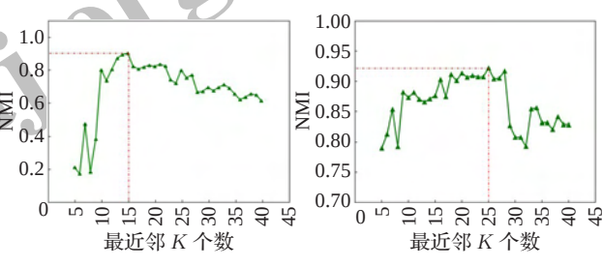
Table 6 Running time experiment results

Data sets	K	Mode	NMI	Time/s
Wisconsin	15	original	0.804	1.81
		now	**0.821**	**1.74**
Tic	15	original	0.008	2.58
		now	**0.019**	**2.56**
Lecturer	20	original	**0.180**	4.43
		now	0.179	**4.03**
Social	20	original	0.149	4.87
		now	**0.184**	**4.24**
Car	25	original	0.134	6.21
		now	**0.160**	**5.89**

3.3.2 参数实验

CDPCD算法的执行需要参数 K ,本小节通过简要举例分析该参数对聚类结果的影响。

如图11,在Zoo和Dermatology数据集上设置参数 K 的范围为[5,40]并运行算法统计其NMI,可以发现指标随着 K 的增加而上升,并逐步达到最大值,随后下降。同时在最大值参数 K 附近的指标差值较小,说明该参数还具有一定的鲁棒性。



(a)Zoo (b)Dermatology

图11 参数 K 对聚类结果NMI的影响

Fig.11 Effect of parameter K on clustering result NMI

3.3.3 核函数实验

本小节对柯西核函数的有效性进行验证,实验选取有序型+名义型、纯有序型和纯名义型的若干数据集并使用目前常见考虑数据之间距离的高斯核(GK)、二次有理核(RQK)和幂指数核(EK)进行比较。表7给出了最终结果,其中表中的最后列为柯西核(CK),可以看到由于缺乏重尾核函数提供的平滑密度值过渡特性,这些对比核函数不能够为分类型数据提供符合其数值分布的良好空间特点,因此聚类性能较差。

表7 核函数对比实验结果

Table 7 Kernel function comparison experimental results

(a)NMI

Data sets	GK	RQK	EK	CK
Lenses	0.256	0.171	0.249	**0.324**
Primary	0.352	0.227	0.411	**0.431**
Lym	0.246	0.186	0.256	**0.296**
House	0.489	0.289	0.532	**0.539**
Wisconsin	0.807	0.819	0.804	**0.821**

(b)ARI

Data sets	GK	RQK	EK	CK
Lenses	0.193	0.122	0.181	0.239
Primary	0.115	0.175	0.128	0.285
Lym	0.149	0.129	0.158	0.308
House	0.449	0.360	0.613	0.613
Wisconsin	0.849	0.713	0.885	0.886

4 结束语

本文提出了结合柯西核的分类型数据密度峰值聚类算法。算法首先针对密度聚类中距离计算产生的重叠问题,提出了一种全新的基于概率分布的加权有序距离度量来增强距离计算的高区分度和高信息量。然后

针对密度计算产生的聚集问题,使用了一种结合柯西核函数的密度估计方法重新计算数据局部密度值,同时改进了SNNDPC算法的划分依据和二次分配过程。实验证明本算法显著提高了分类型数据的聚类效果,但是本算法在处理高维分类型数据集下存在计算量较大的问题,如何有效降低计算消耗将是下一步研究的重点方向。

参考文献:

[1] ZOU H.Clustering algorithm and its application in data mining[J].Wireless Personal Communications, 2020, 110(1): 21-30.

[2] HAMERLY G,ELKAN C.Learning the k in k-means[C]// Proceedings of the 16th Advances in Neural Information Processing Systems, Vancouver and Whistler, December 8-13,2003.Cambridge:MIT Press,2003:281-288.

[3] CHATURVEDI A,GREEN P E,CAROLL J D.K-modes clustering[J].Journal of Classification,2001,18(1):35-55.

[4] BOOK A,KULYN V A,RAITA T.Generalized hamming distance[J].Information Retrieval,2002,5(4):353-375.

[5] HUANG J Z,NG M K,RONG H,et al.Automated variable weighting in k-means type clustering[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2005,27(5): 657-668.

[6] BAI L,LIANG J,DANG C,et al.A novel attribute-weighting algorithm for clustering high-dimensional categorical data[J]. Pattern Recognition,2011,44(12):2843-2861.

[7] BORIAH S,CHANDOLA V,KUMAR V.Similarity measures for categorical data:a comparative evaluation[C]//Proceedings of the SIAM International Conference on Data Mining,Georgia,April 24-26,2008.Philadelphia:SIAM,2008: 243-254.

[8] ŠULC Z,ŘEZANKOVÁ H.Comparison of similarity measures for categorical data in hierarchical clustering[J]. Journal of Classification,2019,36(1):58-72.

[9] ZHANG Y,CHEUNG Y M.An ordinal data clustering algorithm with automated distance learning[C]//Proceedings of the 34th AAAI Conference on Artificial Intelligence,NY, February 7-12,2020.Cambridge:AAAI Press,2020:6869-6876.

[10] ZHANG Y,CHEUNG Y M,TAN K C.A unified entropy-based distance metric for ordinal-and-nominal-attribute data clustering[J].IEEE Transactions on Neural Networks and Learning Systems,2019,31(1):39-52.

[11] ZHANG Y,CHEUNG Y M.Learnable weighting of intra-attribute distances for categorical data clustering with nominal and ordinal attributes[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2021,44(7): 3560-3576.

[12] ZHANG Y,CHEUNG Y M.A new distance metric exploiting heterogeneous inter-attribute relationship for ordinal-and-nominal-attribute data clustering[J].IEEE Transactions on Cybernetics,2022,52(2):758-771.

[13] RODRIGUEZ A,LAIO A.Clustering by fast search and find of density peaks[J].Science,2014,344(6191):1492-1496.

[14] DUAN L,XU L,GUO F,et al.A local-density based spatial clustering algorithm with noise[J].Information Systems, 2007,32(7):978-986.

[15] LIU R,WANG H,YU X.Shared-nearest-neighbor-based clustering by fast search and find of density peaks[J]. Information Sciences,2018,450:200-226.

[16] 王大刚,丁世飞,钟锦.基于二阶k近邻的密度峰值聚类算法研究[J].计算机科学与探索,2021,15(8):1490-1500.
WANG D G,DING S F,ZHONG J.Research of density peaks clustering algorithm based on second-order k neighbors[J].Journal of Frontiers of Computer Science and Technology,2021,15(8):1490-1500.

[17] 丁世飞,徐晓,王艳茹.基于不相似性度量优化的密度峰值聚类算法[J].软件学报,2020,31(11):3321-3333.
DING S F,XU X,WANG Y R.Optimized density peaks clustering algorithm based on dissimilarity measure[J]. Journal of Software,2020,31(11):3321-3333.

[18] 柏锣湘,罗可,罗潇.结合自然和共享最近邻的密度峰值聚类算法[J].计算机科学与探索,2021,15(5):931-940.
BAI E X,LUO K,LUO X.Peak density clustering algorithm combining natural and shared nearest neighbor[J]. Journal of Frontiers of Computer Science and Technology,2021,15(5):931-940.

[19] 赵嘉,姚占峰,吕莉,等.基于相互邻近度的密度峰值聚类算法[J].控制与决策,2021,36(3):543-552.
ZHAO J,YAO Z F,LV L,et al.Density peaks clustering based on mutual neighbor degree[J].Control and Decision,2021,36(3):543-552.

[20] VAN DER MATTEN L,HINTON G.Visualizing data using t-SNE[J].Journal of Machine Learning Research,2008,9 (11):2579-2605.

[21] DU M J,WANG R,JI R,et al.ROBP a robust border-peeling clustering using Cauchy kernel[J].Information Sciences,2021,571:375-400.

[22] AHMAD A,DEY L.A method to compute distance between two categorical values of same attribute in unsupervised learning for categorical data set[J].Pattern Recognition Letters,2007,28(1):110-118.

[23] RUBNER Y,TOMASI C,GUIBAS L J.The earth mover's distance as a metric for image retrieval[J].International Journal of Computer Vision,2000,40(2):99-121.

[24] 张忠林,赵昱,闫光辉.自然邻居密度极值聚类算法[J].计算机工程与应用,2021,57(23):200-210.
ZHANG Z L,ZHAO Y,YAN G H.Natural neighbor density extremum clustering algorithm[J].Computer Engineering and Applications,2021,57(23):200-210.

[25] MAULIK U,BANDYOPADHYAY S.Performance evaluation of some clustering algorithms and validity indices[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2002,24(12):1650-1654.