软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2020,31(11):3492−3505 [doi: 10.13328/j.cnki.jos.005819]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

类属型数据核子空间聚类算法*

徐鲲鹏[1,2], 陈黎飞[1,2], 孙浩军[3], 王备战[4]

[1](福建师范大学 数学与信息学院,福建 福州 350117)

[2](数字福建环境监测物联网实验室(福建师范大学),福建 福州 350117)

[3](汕头大学 工学院,广东 汕头 515063)

[4](厦门大学 软件学院,福建 厦门 361005)

通讯作者: 陈黎飞, E-mail: clfei@fjnu.edu.cn

摘 要: 现有的类属型数据子空间聚类方法大多基于特征间相互独立假设,未考虑属性间存在的线性或非线性相关性.提出一种类属型数据核子空间聚类方法.首先引入原作用于连续型数据的核函数将类属型数据投影到核空间,定义了核空间中特征加权的类属型数据相似性度量.其次,基于该度量推导了类属型数据核子空间聚类目标函数,并提出一种高效求解该目标函数的优化方法.最后,定义了一种类属型数据核子空间聚类算法.该算法不仅在非线性空间中考虑了属性间的关系,而且在聚类过程中赋予每个属性衡量其与簇类相关程度的特征权重,实现了类属型属性的嵌入式特征选择.还定义了一个聚类有效性指标,以评价类属型数据聚类结果的质量.在合成数据和实际数据集上的实验结果表明,与现有子空间聚类算法相比,核子空间聚类算法可以发掘类属型属性间的非线性关系,并有效提高了聚类结果的质量.

关键词: 聚类;类属型数据;核方法;非线性度量;子空间

中图法分类号: TP181

中文引用格式: 徐鲲鹏,陈黎飞,孙浩军,王备战.类属型数据核子空间聚类算法.软件学报,2020,31(11):3492−3505. http://www.jos.org.cn/1000-9825/5819.htm

英文引用格式: Xu KP, Chen LF, Sun HJ, Wang BZ. Kernel subspace clustering algorithm for categorical data. Ruan Jian Xue Bao/Journal of Software, 2020,31(11):3492−3505 (in Chinese). http://www.jos.org.cn/1000-9825/5819.htm

Kernel Subspace Clustering Algorithm for Categorical Data

XU Kun-Peng[1,2], CHEN Li-Fei[1,2], SUN Hao-Jun[3], WANG Bei-Zhan[4]

[1](College of Mathematics and Informatics, Fujian Normal University, Fuzhou 350117, China)

[2](Digital Fujian Internet-of-Things Laboratory of Environmental Monitoring (Fujian Normal University), Fuzhou 350117, China)

[3](College of Engineering, Shantou University, Shantou 515063, China)

[4](College of Software, Xiamen University, Xiamen 361005, China)

Abstract: Currently, the mainstream subspace clustering methods for categorical data are dependent on linear similarity measure and the relationship between attributes is overlooked. In this study, an approach is proposed for clustering categorical data with a novel kernel soft feature-selection scheme. First, categorical data is projected into the high-dimensional kernel space by introducing the kernel function and the similarity measure of categorical data in kernel subspace is given. Based on the measure, the kernel subspace clustering objective function is derived and an optimization method is proposed to solve the objective function. At last, kernel subspace clustering algorithm

* 基金项目: 国家自然科学基金(U1805263, 61672157); 福建省科技厅项目(JK2017007); 福建师范大学创新团队项目(IRTL1704)

Foundation item: National Natural Science Foundation of China (U1805263, 61672157); Project of Science and Technology Bureau, Fujian Province (JK2017007); Program of Innovative Research Team of Fujian Normal University (IRTL1704)

收稿时间: 2018-01-10; 修改时间: 2018-05-16; 采用时间: 2019-01-15

for categorical data is proposed, the algorithm considers the relationship between the attributes and each attribute assigned with weights measuring its degree of relevance to the clusters, enabling automatic feature selection during the clustering process. A cluster validity index is also defined to evaluate the categorical clusters. Experimental results carried out on some synthetic datasets and real-world datasets demonstrate that the proposed method effectively excavates the nonlinear relationship among attributes and improves the performance and efficiency of clustering.

Key words: clustering; categorical data; kernel method; nonlinear measure; subspace

聚类分析作为数据挖掘研究中的一种重要手段,目的是从给定数据集中寻找数据之间的相似性,并以此划分多个子集(每个子集为一个簇),使得不同簇中的数据尽可能相异,而同一簇中的数据尽可能相似,即“物以类聚”[1].目前,聚类分析已经在许多领域获得广泛应用,包括模式识别、文本挖掘、机器学习、图像处理和基因表达等.

在聚类分析中,子空间聚类因其能够从数据集的不同子空间中发现相应的簇[2]而得到广泛研究.现有子空间聚类方法依据提取子空间方法的不同大致可以分为两类.

- 以谱聚类[3]为代表的第 1 类方法,只需将拉普拉斯矩阵生成的特征向量通过 k-means[4]或其他经典方法进行聚类,其本质是将聚类问题转化为图的最优划分问题,是一种点对聚类方法,代表算法包括 PF 算法[5]等.然而此类方法依赖于相似度矩阵,并且在聚类过程中需要求解拉普拉斯矩阵的特征值与特征向量,不仅耗时,而且需要很大的内存空间,算法的时间复杂度和空间复杂度分别为 $O(N^3)$和 $O(N^2)$,其中,N 是数据集样本的数量.
- 第 2 类方法给簇内的每个属性赋予相应的权值,在聚类过程中嵌入特征选择,代表算法包括 FWKM[6]等.该类型算法中,簇内数据分布的方差越大,则属性权重就越小.其中,Chen 等人[7]通过子空间尺度的方法来优化投影子空间,提出 ASC 算法.

本文着重于第 2 类方法,因为第 1 类方法具有较大的时间和空间复杂度.

在第 2 类方法中,大多数研究针对于连续型(数值型)数据.与连续型数据相比,类属型数据属性取值为离散的符号,这导致适用于连续型数据的子空间聚类算法不能直接迁移到类属型数据.本文研究类属型数据的子空间聚类问题,因为类属型数据在现实世界中应用更为广泛,例如,投票数据(vote 数据集)包含移民(immigration)和教育支出(education-spending)等 16 个属性,其中每一个属性值都由 Y 或 N 这两个离散的符号组成.

由于类属型数据属性取值离散的特点,使得类属性数据间的相似性度量比连续型数据更为困难.目前,主流方法采取简单匹配系数(simple matching coefficient,简称 SMC)[8]及其一些改进方法[9],但是这些方法都是基于特征之间相互独立这个假设,而在许多实际应用中,这种假设往往是不成立的[10].比如,临床肺癌诊断数据(breastcancer 数据集)中,肿块密度(clump thickness)随着细胞大小(cell size)的增大以一种非线性的形式降低.当前,发掘属性间非线性关系的主要方法包括深度神经网络和核(kernel)方法等.其中,使用 Mercer 核函数隐式刻画属性间非线性关系的核方法,因其数学表达的简洁性和计算的高效性,已得到广泛研究和应用,例如 Kernel k-means(核 k-均值)[11]算法.目前,研究者已提出很多核聚类的方法,但是它们同样多针对于连续型数据,这是由于核函数中的内积计算在类属型数据中没有意义.

为了解决上述问题,本文提出一种类属型数据核子空间聚类算法.该算法将原作用于连续型数据的核函数用来发掘类属型数据属性间的关系,并且在非线性空间中进行自动的类属型属性加权,实现特征选择.最后,本文还定义了一种基于 AIC 准则的聚类有效性指标,以验证聚类质量并估计类属型数据集划分的簇数目.

本文第 1 节主要是相关工作介绍.第 2 节介绍类属型数据核子空间聚类模型及聚类目标优化函数,并提出一种高效求解该目标函数的方法.第 3 节给出算法的原理及具体描述,并给出聚类有效性指标.第 4 节介绍实验环境和实验结果分析.第 5 节对本文进行总结,并指出进一步的研究方向.

1 相关工作

本节介绍类属型数据子空间聚类的相关背景知识,并分析若干代表性的相关工作.下面,首先给出本文使用

的符号定义.

令给定的数据集 $DB=\{x_1,x_2,\ldots,x_i,\ldots,x_N\}$,其中,$N$ 为数据样本点数目.$\boldsymbol{x}_i=\{x_{i1},x_{i2},\ldots,x_{ij},\ldots,x_{iD}\}$为一个 D 维数据对象,表示 DB 中第 i 个样本点($i=1,2,\ldots,N$),x_{ij} 为 $\boldsymbol{x}_i$ 的第 j 维属性($j=1,\ldots,D$).O_d 是第 d 个属性取值的集合,$o\in O_d$ 表示其中的任一符号(离散值),并用$|O_d|$表示属性 d 中离散取值的总数.典型的硬聚类算法是将 DB 划分成 K 个簇的集合$\Pi=\{\pi_1,\pi_2,\ldots,\pi_K\}$,$\pi_k$ 为 DB 的第 k 个簇($k=1,2,\ldots,K$),且任意两个簇的交集是空集,$K(K>1)$是给定的簇数目.簇π_k 包含的数据对象数目记为$|\pi_k|$.

目前,许多基于特征选择的类属型数据子空间聚类方法相继被提出,主要区别在于属性加权方式的不同.

- WKM[12]算法根据簇内对象到类属属性模的平均距离赋予属性相应的权重.
- wk-modes[13]利用熵来计算各个簇中每个属性的权重.
- 新近出版的 SCC[14]根据核密度估计方法定义了概率距离函数,并基于簇内类别平滑差异进行加权.

这些方法虽然属性加权方式不同,但都根据以下形式计算数据间相似度.

$$sim(\boldsymbol{x}_i,\boldsymbol{x}_j)=\sum_{d=1}^{D}w_{kd}\cdot sim_d(\boldsymbol{x}_i,\boldsymbol{x}_j) \tag{1}$$

$sim_d(\boldsymbol{x}_i,\boldsymbol{x}_j)$表示样本 $\boldsymbol{x}_i$ 和样本 $\boldsymbol{x}_j$ 在第 d 维属性上的相似度;w_{kd} 表示属性 d 对簇π_k 的贡献程度,越大说明越重要,其中,$w_{kd}\geqslant 0$ 并且 $\sum_{d=1}^{D}w_{kd}=1$.

从上式可以看出,这类子空间聚类方法度量两个样本间的相似性时都是独立地计算每个维度上的相似性并累加起来,优势在于较高的聚类效率.然而如前所述,这种方法基于特征之间相互独立这个假设,使得特征之间的关系被忽略不计,而这个假设丢失了很多特征之间的信息.

为了解决上述问题,研究者提出了核聚类方法,将核方法引入到了聚类中.从度量学习的角度来看,核方法就是一种度量方法,通过直接度量数据间的相似性,而不是将每个维度间的相似性累加起来,隐式地考虑了特征之间的关系.

所谓核聚类就是把核方法和聚类算法进行耦合,通过把输入空间的数据利用 Mercer 核非线性映射到高维特征空间,增加了数据点的线性可分概率,即扩大数据之间的差异,在高维特征空间达到线性可聚的目的,例如 Kernel k-means(核 k-均值)[11]算法.核 k-均值算法首先通过“核化”的方式,即利用一个非线性映射$\phi:R^n\to H,x\to\phi(x)$,将原始空间 R^n 中的样本 $\boldsymbol{x}$ 映射到一个高维的核空间 H 中,通过黑盒的方式对原空间中样本的特征进行组合,使得样本在核空间中变得线性可分(或近似线性可分);然后,在这个高维的核空间中进行传统的 k-means 聚类.其中,核函数$\kappa(\boldsymbol{x},\boldsymbol{z})=\phi(\boldsymbol{x})\cdot\phi(\boldsymbol{z})$.

然而,现有核方法主要作用于连续型数据.进行类属型数据核子空间聚类的困难主要包括两个方面.

- 首先,Mercer 核函数是为连续型数据定义的,类属型数据并不能直接进行“核化”.
- 其次,当前核方法基于原始特征是同等重要这一假设的,没有区分特征对不同簇类的重要程度.

本文提出一种类属型数据核子空间聚类模型,以克服上述方法存在的缺陷.该方法不仅考虑了属性间的关系,而且在核空间中能够进行自动的属性加权.本文基于新的核子空间相似性度量定义了一个聚类目标优化函数,并提出了一种高效求解该目标函数的方法,最后定义了一种类属型数据核子空间聚类算法,称该算法为 KSCC(kernel subspace clustering algorithm for categorical data).

2 类属型数据核子空间聚类模型

本节介绍类属型数据核子空间聚类模型以及聚类目标优化函数,并提出了一种高效求解该目标函数的方法.如同其他子空间聚类方法一样,首先需要解决类属型数据间相似性问题,下面给出核子空间相似性度量.

2.1 核子空间相似性度量

如前所述,现有主流子空间方法未考虑属性间的关系,而核方法虽然在非线性空间中挖掘了属性间的关系,但并未区分属性的重要程度.因此,引入原作用于连续型数据的核函数将类属型数据投影到核空间,并且在核空间中为每个簇类π_k 引入一个权值向量 $w_k=\{w_{kd}|d=1,2,\ldots,D\}$用于原始特征选择.$w_{kd}$ 表示属性 d 对簇π_k 的贡献程

度,越大说明越重要.这里,w_{kd}满足约束条件:

$$\begin{cases}\forall k,d: w_{kd} \geqslant 0 \\ \forall k: \sum_{d=1}^{D} w_{kd} = 1\end{cases} \tag{2}$$

此外,为 w_{kd}引入指数$\theta(\theta\neq 0)$控制多属性聚类中的激励强度,并假定θ为已知的常数[14].θ越大,权值分布越平滑.形式上,定义核子空间相似性度量为

$$sim(\boldsymbol{x}_i,\boldsymbol{x}_j)=\kappa_w(\boldsymbol{x}_i,\boldsymbol{x}_j) \tag{3}$$

$\kappa_w(\boldsymbol{x}_i,\boldsymbol{x}_j)$表示特征加权的核函数,为两个数据对象在每个属性上的组合形式引入权重.

以多项式核函数为例:

- 原多项式核函数为 $\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)=(\boldsymbol{x}_i\cdot\boldsymbol{x}_j+1)^p=\left(\sum_{d=1}^{D}x_{id}x_{jd}+1\right)^p$.
- 经过特征加权变为 $\kappa_w(\boldsymbol{x}_i,\boldsymbol{x}_j)=(\boldsymbol{x}_i\cdot\boldsymbol{x}_j+1)^p=\left(\sum_{d=1}^{D}w_{kd}^{\theta}x_{id}x_{jd}+1\right)^p$.

为了使类属型数据能够通过核函数映射到高维空间,这里用符号向量化技术定义.

$$x_{id}=\langle I(x_{id}=O_{d1}),I(x_{id}=O_{d2}),...,I(x_{id}=O_{d|O_d|})\rangle.$$

公式(3)与现有的类属型数据相似性度量相比,不仅借助核方法来进行类属型数据的“核化”,在非线性空间中考虑了特征之间的联系,而且在映射后的核空间中进行了特征选择,区分了特征对簇类有区别的重要.

2.2 类属型数据核子空间聚类目标函数

在聚类分析中,定义簇为紧凑度(或分散度最小)的样本集合,其中,紧凑度以样本到簇中心的相似度来衡量.结合核子空间相似性度量公式,定义类属型数据的核子空间聚类优化目标函数为

$$J(\Pi,W)=\sum_{k=1}^{K}\sum_{x_i\in\pi_k}Sim(\boldsymbol{x}_i,\boldsymbol{v}_k)=\sum_{k=1}^{K}\sum_{x_i\in\pi_k}\kappa_w(\boldsymbol{x}_i,\boldsymbol{v}_k) \tag{4}$$

$\boldsymbol{v}_k$是簇π_k的中心,簇π_k在第 d 维上的中心为 v_{kd}.由于一个类属属性值由一个向量表示,簇π_k在第 d 维上的中心也应该由一个向量表示.令 $v_{kd}=\langle f_k(O_{d1}),f_k(O_{d2}),...,f_k(O_{d|O_d|})\rangle$,这里, $f_k(o)=\frac{1}{|\pi|}\sum_{x_i\in\pi_k}I(x_{id}=o)$ 表示簇π_k 中属性值 $o\in O_d$的频度估计;$I(\cdot)$为指示函数,I(true)=1,I(false)=0.从优化角度分析,划分型聚类算法是求解目标函数的最优值过程.由于是以相似性度量为基础的目标函数,所以本文以最大化公式(4)为优化目标.在计算过程中,求和函数在核函数的内部运算(比如上述多项式核子空间函数),导致 w_{kd}难以求解,这大大增加了求解目标函数的难度.

本文提出一种高效求解该目标函数的优化方法,将目标函数转换为现有主流方法的形式(如公式(1))以提高计算效率.下面对公式(4)定义的优化目标进一步分析.定理 1 表明,对于所有上凸的核函数($\theta\geqslant 1$ 时),求解公式(4)最大值等价于求解公式(5)目标函数最大值.

$$J(\Pi,W)=\sum_{k=1}^{K}\sum_{x_i\in\pi_k}\sum_{d=1}^{D}w_{kd}^{\theta}\kappa_d(\boldsymbol{x}_i,\boldsymbol{v}_k) \tag{5}$$

公式(5)中,$\kappa_d(\boldsymbol{x}_i,\boldsymbol{v}_k)$是指 $\boldsymbol{x}_i$ 和 $\boldsymbol{v}_k$ 在 d 维上映射函数的内积,即第 d 维属性上的核函数.以多项式核函数为例:

$$\kappa_d(x_i,v_k)=(x_{id}v_{kd}+1)^p.$$

定理 1. 当$\theta\geqslant 1$ 时,对于所有上凸的核函数$\kappa(\cdot,\cdot)$,最大化公式(4)与最大化公式(5)同解.

证明:首先定义 z_d 为核子空间相似性度量中两个输入对象在第 d 个属性上的组合,当核函数输入对象为样本 $\boldsymbol{x}_i$ 和簇中心 $\boldsymbol{v}_k$ 时,z_d 表示 $\boldsymbol{x}_i$ 与 $\boldsymbol{v}_k$ 在第 d 个属性上的组合.令 $f\left(\sum_{d=1}^{D}w_{kd}^{\theta}z_d\right)=\kappa_w(\boldsymbol{x}_i,\boldsymbol{v}_k)$,$f(\cdot)$是新定义的函数.可以发现,$f(z_d)=\kappa_d(\boldsymbol{x}_i,\boldsymbol{v}_k)$.下面用数学归纳法证明 $\sum_{d=1}^{D}w_{kd}^{\theta}f(z_d)\leqslant f\left(\sum_{d=1}^{D}w_{kd}^{\theta}z_d\right)$,s.t. Eq(2).

a. 当 D=1,2 时,不等式显然成立.

b. 假设当 $D=n$ 时,不等式成立,即 $\sum_{d=1}^{n}w_{kd}^{\theta}f(z_d)\leqslant f\left(\sum_{d=1}^{n}w_{kd}^{\theta}z_d\right)$.$D=n+1$ 时,令 $p_n=\sum_{d=1}^{n}w_{kd}$,则

1. $\sum_{d=1}^{n+1} w_{kd}^{\theta} f(z_d) = w_{k(n+1)}^{\theta} f(z_{n+1}) + \sum_{d=1}^{n} w_{kd}^{\theta} f(z_d)$

2. $= w_{k(n+1)}^{\theta} f(z_{n+1}) + p_n^{\theta} \sum_{d=1}^{n} \left(\frac{w_{kd}}{p_n} \right)^{\theta} f(z_d)$

3. $\leqslant w_{k(n+1)}^{\theta} f(z_{n+1}) + p_n^{\theta} f\left(\sum_{d=1}^{n} \left(\frac{w_{kd}}{p_n} \right)^{\theta} z_d \right)$

4. $\leqslant f\left(w_{k(n+1)}^{\theta} z_{n+1} + p_n^{\theta} \sum_{d=1}^{n} \left(\frac{w_{kd}}{p_n} \right)^{\theta} z_d \right)$

5. $= f\left(w_{k(n+1)}^{\theta} z_{n+1} + \sum_{d=1}^{n} w_{kd}^{\theta} z_d \right)$

6. $= f\left(\sum_{d=1}^{n+1} w_{kd}^{\theta} z_d \right)$.

所以 $\sum_{d=1}^{D} w_{kd}^{\theta} f(z_d) \leqslant f\left(\sum_{d=1}^{D} w_{kd}^{\theta} z_d \right)$ 得以证明.特别地,当θ=1 时,不等式即为 Jesson 不等式.通过拉伸下界 $\sum_{d=1}^{D} w_{kd}^{\theta} f(z_d)$ 使其上升至与 $f\left(\sum_{d=1}^{D} w_{kd}^{\theta} z_d \right)$ 相等位置,然后调整 w_{kd},使 $\sum_{d=1}^{D} w_{kd}^{\theta} f(z_d)$ 达到最大值.逐步迭代,最终达到 $f\left(\sum_{d=1}^{D} w_{kd}^{\theta} z_d \right)$ 的最大值处.w_{kd} 的推导在第 3 节中给出. □

定理 1 中,当 $D=n+1$ 时,首先令 $p_n = \sum_{d=1}^{n} w_{kd}$.注意到,由于此时 $D=n+1$,所以 $p_n \neq 1$, $p_n + w_{k(n+1)} = \sum_{d=1}^{n+1} w_{kd} = 1$. 步骤 2 中,由于 $p_n = \sum_{d=1}^{n} w_{kd}$,所以 $\sum_{d=1}^{n} \frac{w_{kd}}{p_n} = 1$ 符合约束条件,因此可以把 $\sum_{d=1}^{n} \left(\frac{w_{kd}}{p_n} \right)^{\theta} f(z_d)$ 利用情形 b.中假设的不等式进行转换,即步骤 3 中的不等式;在步骤 3 中, $w_{k(n+1)} + p_n = \sum_{d=1}^{n+1} w_{kd} = 1$,再次符合约束条件.这其实等价于 $D=2$ 时的情况,因此可以把两项通过不等式合并为一项,即步骤 4 中的不等式.定理 1 定义 z_d 作为两个输入对象在第 d 个属性上的组合,$f(\cdot)$是新定义的函数,目的是为了转换核函数的表示形式,例如高斯核子空间函数:

$$\kappa_w(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\sum_{d=1}^{D} w_{kd}^{\theta} \frac{(x_{id} - x_{jd})^2}{2\sigma^2} \right) = f\left(\sum_{d=1}^{D} w_{kd}^{\theta} z_d \right),$$

其中, $z_d = -\frac{\| x_{id} - x_{jd} \|^2}{2\sigma^2}$,$f(x)=\exp(x)$.

在核函数中,高斯核函数无论对于大样本还是小样本都有较好的性能,且比其他核函数参数较少,是应用最广的核函数.选择高斯核函数代入目标函数,则公式(5)可改写成:

$$J(\Pi, W) = \sum_{k=1}^{K} \sum_{x_i \in \pi_k} \sum_{d=1}^{D} w_{kd}^{\theta} \exp\left(-\frac{(x_{id} - v_{kd})^2}{2\sigma^2} \right) = \sum_{k=1}^{K} \sum_{x_j \in \pi_k} \sum_{d=1}^{D} w_{kd}^{\theta} \exp\left(-\frac{\sum_{o \in O_d} [I(x_{id} = o) - f_k(o)]^2}{2\sigma^2} \right) \tag{6}$$

高斯核函数中的参数σ^2定义为数据集 DB 的全局方差, $\sigma^2 = \frac{1}{ND} \sum_{i=1}^{N} \sum_{d=1}^{D} \sum_{o \in O_d} [I(x_{id} = o) - f_k(o)]^2$.

3 KSCC 聚类算法

本节介绍算法的原理及具体描述,并提出一个新的聚类有效性指标,用于验证聚类质量和估计类属型数据集划分的簇的数目.

针对公式(6),这是一个带约束的非线性优化问题,应用拉格朗日乘子法,结合上述定义,优化目标函数可转换为

$$\max J(\Pi, W) = \sum_{k=1}^{K} \sum_{x_i \in \pi_k} \sum_{d=1}^{D} w_{kd}^{\theta} \exp\left(-\frac{\sum_{o \in O_d} [I(x_{id} = o) - f_k(o)]^2}{2\sigma^2} \right) + \sum_{k=1}^{K} \lambda_k \left(1 - \sum_{d=1}^{D} w_{kd} \right) \tag{7}$$

本文采用 EM 型算法对其优化,即通过迭代法求 J 的局部最优值.根据这个原理,每次迭代过程首先设定

$\Pi=\hat{\Pi}$ 以求解最大化 $J(\hat{\Pi},W)$ 的 W,记为 $\hat{W}$;其次,设定 $W=\hat{W}$,通过最大化 $J(\Pi,\hat{W})$ 求解最优的Π,即 $\hat{\Pi}$.后者可以通过将每个对象 x_i 划分到相似性最高的簇来实现.算法根据如下规则将 $\boldsymbol{x}_i$ 划分到簇π_k 中去.

$$k=\arg\max_{\forall k}(\kappa_w(\mathrm{x}_i,\mathrm{v}_k))=\arg\max_{\forall k}\left(\exp\left(-\sum_{d=1}^{D}w_{kd}{}^{\theta}\frac{\sum_{o\in O_d}[I(x_{id}=o)-f_k(o)]^2}{2\sigma^2}\right)\right) \tag{8}$$

从而生成新的聚类划分 $\hat{\Pi}$.求解 $\hat{W}$ 可以通过定理 2 进行.

定理 2. 设定当 $\Pi=\hat{\Pi}$ 时,最大化公式(7)当且仅当

$$\hat{w}_{kd}=\left[\sum_{x_i\in\pi_k}\exp\left(-\frac{\sum_{o\in O_d}[I(x_{id}=o)-f_k(o)]^2}{2\sigma^2}\right)\right]^{\frac{1}{1-\theta}}\Bigg/\sum_{d=1}^{D}\left[\sum_{x_i\in\pi_k}\exp\left(-\frac{\sum_{o\in O_d}[I(x_{id}=o)-f_k(o)]^2}{2\sigma^2}\right)\right]^{\frac{1}{1-\theta}} \tag{9}$$

证明:根据公式(7),可以定义 K 个独立的子优化目标函数:

$$\begin{aligned}J_k(w_k,\lambda_k)&=\sum_{d=1}^{D}w_{kd}^{\theta}\exp\left(-\frac{(x_{id}-x_{jd})^2}{2\sigma^2}\right)+\lambda_k\left(1-\sum_{d=1}^{D}w_{kd}\right)\\&=\sum_{d=1}^{D}w_{kd}^{\theta}\exp\left(-\frac{\sum_{o\in O_d}[I(x_{id}=o)-f_k(o)]^2}{2\sigma^2}\right)+\lambda_k\left(1-\sum_{d=1}^{D}w_{kd}\right).\end{aligned}$$

设 $(\hat{w}_k,\hat{\lambda}_k)$ 最大化 $J_k(\hat{w}_k,\hat{\lambda}_k)$,有:

$$\frac{\partial J_k(\hat{w}_k,\hat{\lambda}_k)}{\partial\hat{w}_{kd}}=\theta\exp\left(-\frac{\sum_{o\in O_d}[I(x_{id}=o)-f_k(o)]^2}{2\sigma^2}\right)\hat{w}_{kd}-\hat{\lambda}_k=0,$$

$$\frac{\partial J_k(\hat{w}_k,\hat{\lambda}_k)}{\partial\hat{\lambda}_k}=1-\sum_{d=1}^{D}\hat{w}_{kd}=0.$$

合并以上两个公式,可以得到公式(9). □

具体算法过程如下.

算法. **KSCC**.

输入:类属型数据集 DB,簇数目 K,激励强度θ.

输出:簇划分Π及属性权重集合 W.

1. 初始化:生成数据集初始划分,记为$\Pi^{(0)}$;算法迭代次数 p,p=0.

REPEAT:

2. 更新 W:设定 $\hat{\Pi}=\Pi^{(p)}$,根据公式(9)更新属性权重,得到 $W^{(p+1)}$.
3. 更新Π:设定 $\hat{W}=W^{(p+1)}$,利用公式(8)将每个数据对象划分到簇,生成新的聚类集合$\Pi^{(p+1)}$.
4. 迭代次数:p=p+1.

UNTIL 聚类集合不发生变化,即$\Pi^{(p)}$=$\Pi^{(p+1)}$.

EM 型算法对其初始状态具有一定依赖性,因此我们借助 k-modes[15]算法对 KSCC 算法第一步中的数据集进行初始划分.从算法结构可知,KSCC 算法与 k-means[16]算法相似,算法中每一次迭代过程的时间复杂度为 $O(KND)$,假设算法迭代次数为 P,KSCC 算法时间复杂度为 $O(KNDP)$.算法在每步迭代过程中提高目标函数值,且数据间相似度最大为 1,所以目标函数存在上界.因此在有限的迭代次数下,KSCC 算法是收敛的.

3.1 关于参数θ的讨论

在 KSCC 聚类过程中,通过核函数直接度量数据间的相似性,在核空间中每个属性都被自动赋予一个衡量其重要程度的权值,通过特征选择寻找到相应的子空间.根据公式(9),簇π_k 中属性 d 的权值计算为

$$w_{kd}^{\theta}\sim\left(\sum_{x_i\in\pi_k}\exp\left(-\frac{\sum_{o\in O_d}[I(x_{id}=o)-f_k(o)]^2}{2\sigma^2}\right)\right)^{\frac{\theta}{1-\theta}}=\left(\sum_{x_i\in\pi_k}\exp\left(\frac{-2(1-f_k(x_{id}))+(1-\sum_{o\in O_d}[f_k(o)]^2)}{2\sigma^2}\right)\right)^{\frac{\theta}{1-\theta}}.$$

上式中,$1-\sum_{o\in O_d}[f_k(o)]^2$ 是表示类属型数据分布离散程度的基尼系数.θ作为激励强度,是控制权重的分配

参数.图 1 给出了参数对于设定的 3 个属性权重的变化,这里设定 3 个属性的离散程度从属性 1 开始依次递增.

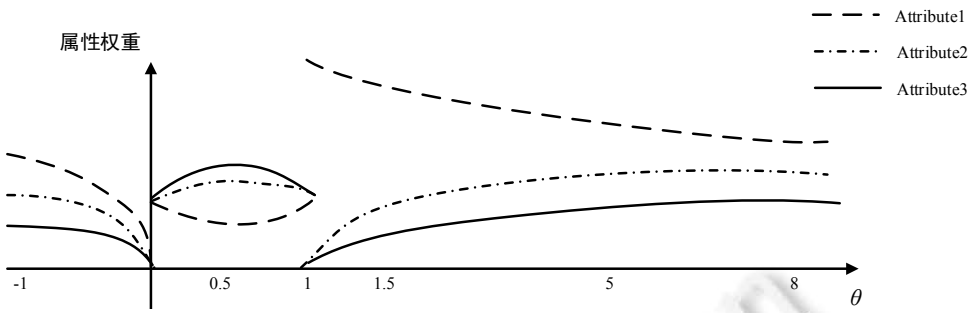

Fig.1 Analysis of weight with different θ

图 1 不同θ值下,属性权重的分析

如图 1 所示,θ取值分以下几种情况讨论.

(1) 当θ=0 时,w_{kd}^{θ} 为常数,每个属性将被分配相等的权重.

(2) 当θ=1 时,$\frac{\theta}{1-\theta}$ 趋近于无穷大,又因为受到 $\sum_{d=1}^{D} w_{kd}=1$ 的限制,所以$\theta\to 1^{+}$时,样本最小偏差的属性将得到加权,而其余属性被赋予零权重.这样其实是在每个簇中仅仅选择一个属性,其他属性都被忽略;当 $\theta\to 1^{-}$时,所有属性的重要程度趋于一致.

(3) 当 0<θ<1 时,离散程度越大的属性,其权重越大.

(4) 当θ<0 和θ>1 时,属性权重与数据分布的离散程度成反比.因此结合定理 1,设定θ>1.然而注意到,当θ过大时,属性权重之间的差异被降低.

实验将基于聚类质量结果来选择θ,具体在第 4 节给出.

3.2 聚类有效性指标

为了估计类属型数据集划分的簇数目,本节给出了一个新的聚类有效性指标.传统的试错过程[17]认为,当簇数目 K 从最小取到最大的过程中,有效性指标最小的 K 值即为最佳的簇数目.新的聚类有效性指标基于有限样本修正的 AIC(akaike information criterion)准则(简称 AICc)[18].

$$AICc=\frac{2NP}{N-P-1}-2\ln(\hat{L}).$$

这里,P 是聚类模型中自由参数的个数,N 为数据集中数据样本点数目,$\hat{L}$ 定义为模型似然函数的最大值.

对于类属型数据而言,应用 AICc 的难点在于似然函数的估计,原因是根据正态分布,模型误差是独立同分布的假设条件下,类属型数据的似然函数是没有被明确定义的[19].由于我们定义了对象到中心的相似性,所以似然函数可以通过用类属型数据间的距离来替换正态分布中的欧式距离来实现.我们定义第 i 个对象在第 k 个簇中的似然函数为

$$L_{ki}(\delta^2)=\frac{1}{\sqrt{2\pi}\delta}\exp\left(-\frac{1}{2\delta^2}\sum_{d=1}^{D}(1-\kappa_d(\boldsymbol{x}_i,\boldsymbol{v}_k))\right).$$

上式中,δ^2 是高斯函数的方差,区别于公式(6)中的方差.定义 $\hat{\delta}^2$ 是δ^2 的最大似然估计,可以得到:

$$\hat{\delta}^2=\frac{1}{N-K}\sum_{k=1}^{K}\sum_{d=1}^{D}\sum_{x_i\in\pi_k}(1-\kappa_d(\boldsymbol{x}_i,\boldsymbol{v}_k)).$$

通过 $\ln(\hat{L})=\sum_{k=1}^{K}\sum_{x_i\in\pi_k}\ln L_{ki}(\hat{\sigma}^2)$ 计算最大对数似然,对于 K 个类属型簇的平均 AICc 为

$$\frac{1}{N}\left(\frac{2NP}{N-P-1}-2\ln(\hat{L})\right)=\frac{N+P-1}{N-P-1}+\ln(2\pi\delta^2)-\frac{K}{N}.$$

由于 $\ln(2\pi)$是一个与 K 无关的常数,由此我们可以得出聚类有效性指标 V_{KC} 如下.

$$V_{KC}(K)=\frac{N+P-1}{N-P-1}+\ln(\hat{\delta}^2)-\frac{K}{N}.$$

实际上,$\frac{K}{N}-\ln(\hat{\delta}^2)$ 在此聚类有效性指标中度量的是模型的拟合度;而 $\frac{N+P-1}{N-P-1}$ 作为模型复杂性的惩罚,随着参数数量的增加变大.因此,最好的模型是使得 V_{KC} 值最小的,也就是使得 $V_{KC}(K)$最小的 K 值即为最佳簇数目.

4 实验分析

实验部分主要验证 KSCC 算法对类属型数据的聚类性能,在合成数据集与真实数据集上进行了实验,并与若干当前主流的类属型聚类算法相比较.实验平台如下:Core(TM)i5-4590 3.30GhzCPU,4.00GB 内存,操作系统为 Windows 7.

4.1 实验设置

如前所述,本文选择径向基核中的高斯核函数用于挖掘类属属性间非线性的关系,不仅因为其适应于各种样本以及参数较少,而且在于这类核函数提供的映射空间是无穷维的,以至于原空间中线性不可分的数据可以被直接映射成线性可分的点.在实际应用中,高斯核函数也是应用最广的核函数.本文高斯核函数中的参数σ^2 定义为数据集 DB 的全局方差: $\sigma^2=\frac{1}{ND}\sum_{i=1}^{N}\sum_{d=1}^{D}\sum_{o\in O_d}[I(x_{id}=o)-f_k(o)]^2$,由数据驱动学习得出.实验选择了WKM[12]、MWKM[20]和 KKM(核 k-means)[11]这 3 种算法进行对比.WKM[12]在 k-modes[15]的基础上引入了基于距离的属性加权,MWKM[20]算法根据模的频度进行属性加权.这两种算法代表了目前类属性数据子空间聚类的两种主流方法,然而这两种算法都是基于特征相互独立来计算样本间的相似性(相异性),选择这两种算法可以和 KSCC 非线性相似性度量作对比.KKM[11]算法与 KSCC 一样经过了“核化”的过程,考虑了属性间的关系,但未能区分特征的重要程度,因此选择 KKM[11]来验证 KSCC 在非线性空间中的子空间聚类效果.WKM[12]算法的参数β根据作者的建议值设置为 2.MWKM[20]算法的参数根据作者的建议值设置,即β=2 和 T_s=T_v=1.

合成数据能够从簇的数目、大小等控制数据集的簇结构,便于分析算法的性能以及对各样数据集的适应性.本文首先在多个合成数据集上进行测试,然后在若干真实数据集上实验.由于各数据集已知类别标签,本文选择两个外部评价指标 *Accuracy* 和 *F-Score*[21]来评估新的算法的聚类性能.两个指标的值越大,表明聚类的效果越好.其中,*F-Score* 定义如下.

$$F\text{-}Score=\sum_{k=1}^{K}\frac{n_k}{N}\max_{1\leqslant i\leqslant K}\left(\frac{2\times R(class_k,\pi_i)\times P(class_k,\pi_i)}{R(class_k,\pi_i)+P(class_k,\pi_i)}\right).$$

这里,$class_k$表示数据集中第 k 个真实的类,n_k表示 $class_k$包含的样本数目,$P(class_k,\pi_i)$和 $R(class_k,\pi_i)$分别表示数据集中真实的类 $class_k$与聚类结果中的簇π_i相比较的准确率和召回率.

4.2 合成数据及结果分析

实验采用的合成数据由 MATLAB 生成.首先利用 *mvnrnd*(·)函数分别生成了 3 个多维数值型数据集,通过设定属性的方差来控制属性的权重;通过调整协方差矩阵的参数来控制属性间的相关程度.然后对合成得到的数值型数据进行等宽离散化处理[22],转化为类属型数据.合成的数据集都包含着正确的类别标签.具体参数见表 1.

其中,

- DataSet1 数据集中设定属性 1 与属性 2 的协方差为–2,使其属性相关;且设定属性 3、属性 4 的方差与其余属性相比较小,使得这两个属性相对于其他属性更为重要.通过 DataSet1 对比 KSCC 算法和没有进行特征选择的 KKM[11]算法.
- DataSet2 与 DataSet3 数据集中分别抽取 10 个和 20 个属性设定彼此之间的协方差;并设定各属性上方差相等.通过这两个数据集来验证 KSCC 与目前主流子空间聚类方法相比,在处理属性关系上的优势.

Table 1 Synthetic datasets parameters

表 1 合成数据集参数

	属性数目 D	簇数目 K	样本数目 N
DataSet1	6	2	1 000
DataSet2	30	4	1 000
DataSet3	60	8	1 000

在KSCC聚类过程中,第1步要确定聚类簇数K和权重参数θ.由于数据集已经给定簇数目K,所以这一步骤可以用来验证指标V_{KC}的有效性.首先,通过固定参数θ来估计K,然后通过估计出来的K来确定最佳的θ.为了估计K,我们设定θ=5并且$K\in\left[2,\sqrt{N}\right]$[23].然后,我们选择使$V_{KC}$值最小的$K$作为最佳簇数.图2显示了每个合成数据集在不同的$K$上运行30次的平均$V_{KC}$结果.图3显示了6个数据集上固定簇数目$K$,不同$\theta$取值下的$V_{KC}$值.$\theta$的取值为1~10,每次增加0.5.每个$V_{KC}$值对应每个$\theta$在数据集上50次实验的平均$V_{KC}$.

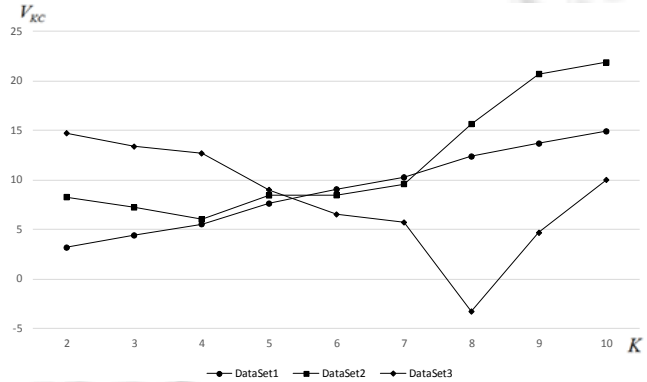

Fig.2 Change in the cluster validity index V_{KC} with various K on the Synthetic datasets

图2 合成数据集上不同K值对应的V_{KC}变化

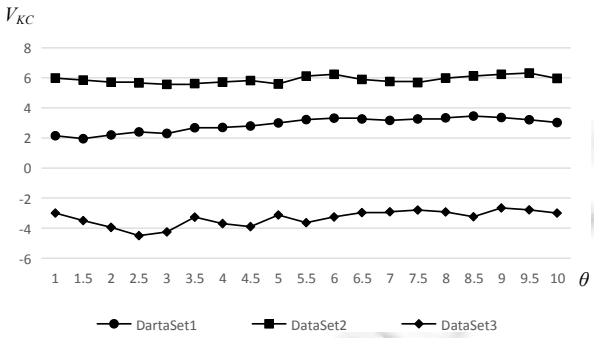

Fig.3 Change in the V_{KC} with various θ on the synthetic datasets

图3 合成数据集上,不同θ值对应的V_{KC}变化

结果表明,这3个合成数据集的最佳簇数目分别为2,4,8,这些簇数目恰恰是已知数据集中的簇数目;并且观察得到最佳的V_{KC}值对应的θ为1.5(DataSet1),3(DataSet2),2.5(DataSet3).表2列出了4种算法在合成数据集上独立运行100次的平均聚类结果,以“均值±方差”的形式提供.

表2所报告的聚类精度均值反映了各个聚类算法的总体性能,而判断各个算法聚类性能的稳定性可以依据所列的方差.聚类精度方差越小,说明算法聚类性能的稳定性越好.针对表2中所列的每行聚类结果,将最大的指标值加黑显示.

Table 2 Comparison of *F*-score and Accuracy by different algorithms on the synthetic datasets

表 2 合成类属型数据集上不同算法的 *F*-score 和 Accuracy 指标对比

指标	数据集	KSCC	WKM	MWKM	KKM
F-score	DataSet1	**0.9516±0.02**	0.9312±0.18	**0.9503±0.06**	0.8869±0.04
	DataSet2	**0.8023±0.05**	0.7231±0.12	0.7623±0.06	0.7882±0.13
	DataSet3	**0.7133±0.13**	0.6156±0.15	0.63796±0.08	0.6723±0.03
Accuracy	DataSet1	**0.9654±0.02**	0.9314±0.12	0.9631±0.03	0.8952±0.03
	DataSet2	**0.8305±0.04**	0.7319±0.11	0.7824±0.06	0.8054±0.06
	DataSet3	**0.7232±0.13**	0.6346±0.14	0.6526±0.05	0.6833±0.03

从表 2 可以看出,由于 DataSet1 中只有两个属性相关,所以除 KKM[11]之外,其余算法聚类质量相差不大,与 KKM[11]相比,KSCC 由于进行特征选择识别出重要的属性 3 与属性 4,聚类结果明显较好.从 DataSet2 和 DataSet3 中可以看出,随着属性相关的数目增多,KSCC 的聚类精度明显高于 WKM 和 MWKM.这是由于 KSCC 进行了“核化”的操作,考虑了属性之间的关系.

下面通过一个实例说明 KSCC 在挖掘属性间关系的效果.以合成数据集 DataSet1 为例,属性 1 与属性 2 相关,图 4 绘制了在这两个属性的二维子空间上样本的分布情况.原空间中两个属性相关,样本呈现环形分布.经过核变换后,样本由原空间中的环形分布在非线性空间中拉伸成了近似线性分布,如图 5.图 5 表明,KSCC 通过核变换将样本投影到了一个高维空间,在核空间中挖掘出了属性间非线性的组合形式,增大了数据线性可聚的概率.

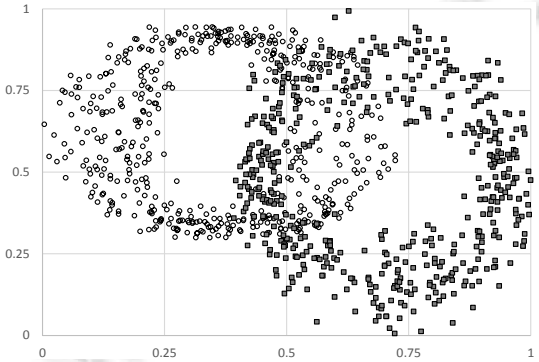

Fig.4 Distribution of samples in the original space

图 4 原空间中样本分布

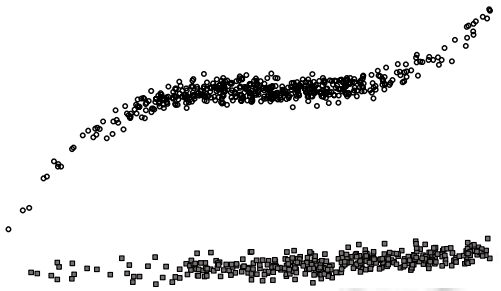

Fig.5 Distribution of samples in non-linear space

图 5 非线性空间中样本分布

4.3 真实数据及结果分析

为了能有效测试 KSCC 算法的聚类性能,本文同样在真实数据集上进行了测试,实验采用来自 UCI 的 6 个真实数据集,其详细信息见表 3.

Table 3 Summary of the parameters for the real-world datasets

表 3 真实数据集的有关信息

UCI 数据集	属性数目 *D*	簇数目 *K*	样本数目 *N*
Breastcancer	9	2	699
Vote	16	2	435
Mushroom	21	2	8 124
Soybeansmall	35	4	47
Dermatology	33	6	366
Zoo	15	7	101

数据集 Breastcancer 是乳腺癌数据;Vote 来自美国国会投票记录;蘑菇数据集 Mushroom 包含的样本较多,并且由于其中的 veil-type 属性因取值唯一在实验中剔除;Soybean(Small)是著名的大豆疾病数据;Dermatology 数据集用于医疗领域皮肤病诊断;Zoo 是动物数据,由于动物名称属性取值皆不相同,与动物种类无关,因此在实

验中剔除.

根据第 4.2 节的方法对各真实数据集确定聚类簇数 K 和权重参数 θ.图 6 显示了 θ=5 时,每个真实数据集在不同的 K 上运行 30 次的平均 V_{KC} 结果.结果表明,Breastcancer、Vote 等这 6 个数据集的最佳簇数目分别为 2,2,2,4,7,7.除了数据集 Dermatology 以外,这些簇数目恰恰是真实数据集中的簇数目,再次验证了 V_{KC} 的有效性.分析 Dermatology 数据集可知,每个类别的样本数呈现分布不均衡的特点,所以影响了 V_{KC},而真实簇数目对应的 V_{KC}(K=6)=3.5012 与 V_{KC} 最小值(V_{KC}(K=7)=3.4919)之间误差极小.图 7 显示了 6 个数据集上固定簇数目 K,不同 θ 取值下的 V_{KC}.由于数据集 Dermatology 与 Zoo 的 V_{KC} 与其他相差过大,图 7 对其进行适当放缩.从图 7 可知,当 θ 为 1.5(Breastcancer),1.5(Vote),2(Mushroom),2.5(Soybeansmall),3.5(Dermatology),1.5(Zoo)时,聚类质量最好.表 4 列出了 5 种算法在真实数据集上获得的聚类结果.从表 4 可以看出,与其他 3 种对比算法相比,KSCC 算法在大部分真实数据集上均获得较高的聚类结果,尤其在样本数较多的 Mushroom、类别数最多的 Zoo 和相对高维的 Soybeansmall 数据集上,说明新算法对类属型数据集具有良好的适应性.在 Dermatology(医疗领域皮肤病诊断)数据集中,红斑等皮肤病属性与是否发痒等属性有着明显的关系,WKM 算法与 MWKM 算法基于特征相互独立假设,当数据具有较多的属性相关时,这类方法严重影响了聚类结果;而 KSCC 算法通过核方法以黑盒的方式考虑了属性间的关系,提高了聚类质量.Mushroom 数据集不仅属性间相关,这 21 个属性的统计特性还存在明显的差异.与 KKM[11]算法相比,KSCC 算法通过特征加权区分出 bruises(第 4 个属性)和 veil-color(第 16 个属性)等重要的属性,赋予它们较大的权重,进一步说明了核子空间聚类的优势.

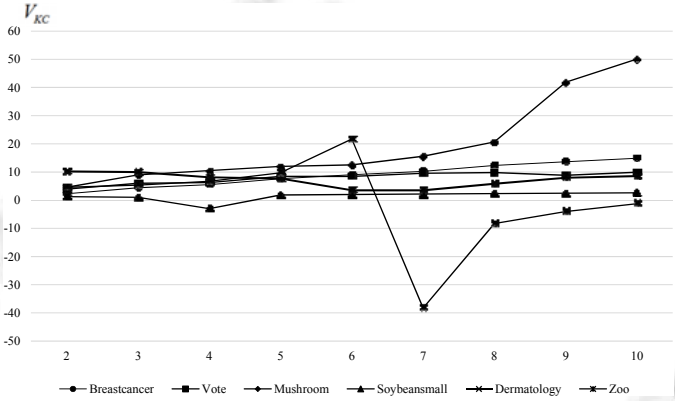

Fig.6 Change in the cluster validity index V_{KC} with various K on the real-world datasets

图 6 真实数据集上,不同 K 值对应的 V_{KC} 变化

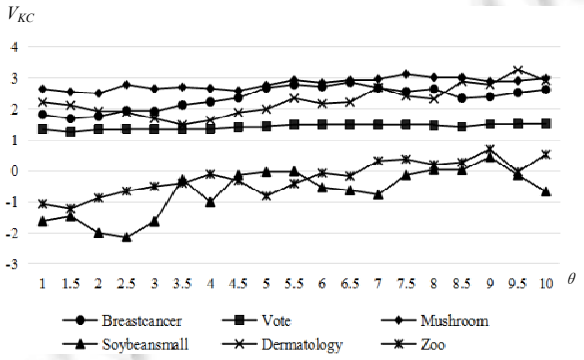

Fig.7 Change in the V_{KC} with various θ on the real-world datasets

图 7 真实数据集上,不同 θ 值对应的 V_{KC} 变化

Table 4 Comparison of *F*-score and Accuracy by different algorithms, on the real-world datasets

表 4 真实类属型数据集上不同算法的 *F*-score 和 Accuracy 指标对比

指标	数据集	KSCC	WKM	MWKM	KKM
F-score	Breastcancer	**0.9659±0.00**	0.7713±0.06	0.8514±0.06	0.9125±0.02
	Vote	**0.8841±0.00**	0.8223±0.06	0.8623±0.06	0.8438±0.04
	Mushroom	**0.7733±0.13**	0.6746±0.08	0.7136±0.08	0.7014±0.02
	Soybeansmall	**0.8975±0.04**	0.7538±0.13	0.7938±0.13	0.8126±0.07
	Dermatology	**0.7241±0.02**	0.6442±0.11	0.6542±0.11	0.6717±0.02
	Zoo	**0.7603±0.05**	0.7425±0.03	**0.7625±0.03**	0.7625±0.05
Accuracy	Breastcancer	**0.9654±0.00**	0.8103±0.03	0.8631±0.03	0.9154±0.03
	Vote	0.8805±0.00	0.8324±0.06	**0.8824±0.06**	0.8562±0.04
	Mushroom	**0.7856±0.08**	0.6862±0.12	0.7195±0.12	0.7326±0.03
	Soybeansmall	**0.9326±0.03**	0.7869±0.12	0.8069±0.12	0.8155±0.07
	Dermatology	**0.8678±0.04**	0.6821±0.08	0.6821±0.08	0.6959±0.02
	Zoo	0.7732±0.03	0.7726±0.04	**0.8126±0.04**	0.7789±0.02

以 Breastcancer 为例,图 8 给出了各类算法运行 100 次的聚类精度分布,横坐标代表各算法运行的次数,纵坐标是以 *F*-score 指标衡量每次聚类获得的聚类结果.如图 8 所示,KSCC 算法与其他算法相比,波动最小.由于 *k*-modes[15]型算法在聚类过程中仅考虑模而易陷入局部最优以及初始簇中心为 *K* 个随机选择的对象,所以导致聚类结果反差很大(体现在平均精度的标准差上).而 KSCC 算法将模向量化,避免了以上问题,具有比其他算法更稳定的性能.

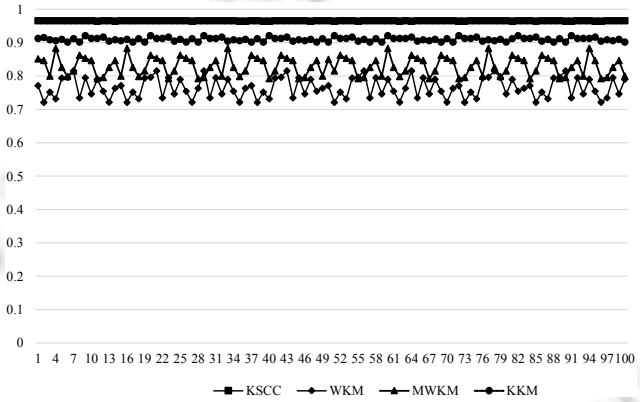

Fig.8 Comparison of *F*-score with different algorithms on Breastcancer

图 8 不同算法在 Breastcancer 上的 *F*-score 指标对比

图 9 给出了各算法分别在 6 个真实数据集上独立运行 100 次聚类花费的平均时间对比.

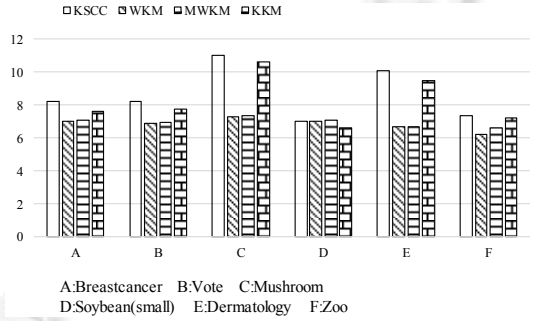

Fig.9 Comparison of different algorithms running average time

图 9 不同算法运行的平均时间对比

由于低维数据聚类和维度较高的数据聚类花费的平均时间跨度较大,直接用实际平均时间做比较,图示效果不佳.因此,本文统一对聚类花费的平均时间取对数.其中,纵坐标代表各算法在真实数据集上运行的平均时间(ms).

从图 9 可以看出,WKM[12]和 MWKM[20]这两种算法具有较高的聚类效率,这正是基于模的聚类算法的一个优势所在,只需考虑类属属性的模,而忽略其余类属符号的统计信息,这大大降低了算法时间;KKM[11]算法与 KSCC 相比没有属性加权的过程,时间也较低.

5 结 论

针对现有类属型数据子空间聚类方法线性度量对象间相似性的问题,本文提出了一种新的类属型数据核子空间聚类方法,用于类属型数据无监督的统计学习.利用核的思想定义了一个新的相似性度量;然后基于该相似性度量,提出一种核子空间聚类算法 KSCC 来优化目标函数.新方法不仅克服了线性相似性度量的不足,还能够进行自动的属性加权.经过合成数据和真实数据集的实验验证,与现有其他类属型数据子空间聚类算法相比,KSCC 算法在实验数据上的聚类质量获得较为明显的改善.

下一步的工作可分为两方面:一方面,对于类属型数据我们无法判断其分布形式,下一步寻找一种能自适应学习出核函数形式(即自适应学习出核矩阵)的方法;另一方面,高维数据的聚类分析是数据挖掘的一类难点问题,类属型数据中存在较多高维数据,比如序列数据,下一步对高维类属型数据进行聚类分析.

References:

[1] Han JW, Kamber M, Pei J, Worte; Fan M, Meng XF, Trans. Data Mining: Concepts and Techniques. 3rd ed., Beijing: China Machine Press, 2012 (in Chinese). [doi: 10.3969/j.issn.1674-6511.2008.03.043]

[2] Chen LF, Wu T. Feature Reduction in Data Mining. Beijing: Science Press, 2016 (in Chinese).

[3] Cai XY, Dai GZ, Yang LB. Survey on spectral clustering algorithms. Computer Science, 2008,35(7):14−18 (in Chinese with English abstract). [doi: 10.3969/j.issn.1002-137X.2008.07.004]

[4] Jain AK, Murty MN, Flynn PJ. Data clustering: A review. ACM Computing Surveys, 1999,31(3):264−323.

[5] Perona P, Freeman W. A factorization approach to grouping. In: Proc. of the European Conf. on Computer Vision. 1998. 655−670.

[6] Huang JZ, Ng MK, Rong H, *et al*. Automated variable weighting in *k*-means type clustering. IEEE Trans. on Pattern Analysis & Machine Intelligence, 2005,27(5):657−668. [doi: 10.1109/TPAMI.2005.95]

[7] Chen LF, Guo GD, Jiang QS. Adaptive algorithm for soft subspace clustering. Ruan Jian Xue Bao/Journal of Software, 2010,21(10): 2513−2523 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3763.htm [doi: 10.3724/SP.J.1001.2010.03763]

[8] Ng MK, Li MJ, Huang JZ, *et al*. On the impact of dissimilarity measure in *k*-modes clustering algorithm. IEEE Trans. on Pattern Analysis & Machine Intelligence, 2007,29(3):503−507. [doi: 10.1109/TPAMI.2007.53]

[9] Boriah S, Chandola V, Kumar V. Similarity measures for categorical data: A comparative evaluation. In: Proc. of the 2008 SIAM Int'l Conf. on Data Mining. 2008. 243−254. [doi: 10.1137/1.9781611972788.22]

[10] Knippenberg RW. Orthogonalization of categorical data: How to fix a measurement problem in statistical distance metrics. Ssrn Electronic Journal, 2013. [doi: 10.2139/ssrn.2357607]

[11] Kong R, Zhang GX, Shi ZS, *et al*. Kernel-based *K*-means clustering. Computer Engineering, 2004,30(11):12−13,80 (in Chinese with English abstract). [doi: 10.3969/j.issn.1000-3428.2004.11.005]

[12] Chan E, Ching W, Ng M, *et al*. An optimization algorithm for clustering using weighted dissimilarity measures. Pattern Recognition, 2004,37(5):943−952. [doi: 10.1016/j.patcog.2003.11.003]

[13] Cao F, Liang J, Li D, *et al*. A weighting *k*-modes algorithm for subspace clustering of categorical data. Neurocomputing, 2013, 108(5):23−30. [doi: 10.1016/j.neucom.2012.11.009]

[14] Chen L, Wang S, Wang K, *et al*. Soft subspace clustering of categorical data with probabilistic distance. Pattern Recognition, 2016, 51(C):322−332. [doi: 10.1016/j.patcog.2015.09.027]

[15] Huang Z, Ng MK. A note on *K*-modes clustering. Journal of Classification, 2003,20(2):257−261. [doi: 10.1007/s00357-003-0014-4]

[16] Hartigan J A, Wong MA. A *K*-means clustering algorithm. Applied Statistics, 1979,28(1):100−108. [doi: 10.2307/2346830]

[17] Sun H, Wang S, Jiang Q. FCM-based model selection algorithms for determining the number of clusters. Pattern Recognition, 2004, 37(10):2027−2037. [doi: 10.1016/j.patcog.2004.03.012]

[18] Burnham KP, Anderson DR. Model Selection and Multimodel Inference: A Practical Information-theoretic Approach. Springer-Verlag, 2002. [doi: 10.1198/tech.2003.s146]

[19] Pelleg D, Moore AW. *X*-means: Extending *K*-means with efficient estimation of the number of clusters. In: Proc. of the 17th Int'l Conf. on Machine Learning. 2000. [doi: 10.1007/3-540-44491-2_3]

[20] Bai L, Liang J, Dang C, *et al*. A novel attribute weighting algorithm for clustering high-dimensional categorical data. Pattern Recognition, 2011,44(12):2843−2861. [doi: 10.1016/j.patcog.2011.04.024]

[21] Chen L, Jiang Q, Wang S. A probability model for projective clustering on high dimensional data. In: Proc. of the IEEE Int'l Conf. on Data Mining. 2008. 755−760. [doi: 10.1109/ICDM.2008.15]

[22] Chen LF. A probabilistic framework for optimizing projected clusters with categorical attributes. Science China Information Sciences, 2015,58(7):1−15. [doi: 10.1007/s11432-014-5267-5]

[23] Bezdek JC. Pattern Analysis in Handbook of Fuzzy Computation. IOP Publishing Ltd., 1998. [doi: 10.1887/0750304278]

附中文参考文献:

[1] 韩家炜,Kamber M,裴健,著;范明,孟小峰,译.数据挖掘:概念与技术.第 3 版,北京:机械工业出版社,2012. [doi: 10.3969/j.issn.1674-6511.2008.03.043]

[2] 陈黎飞,吴涛.数据挖掘中的特征约简.北京:科学出版社,2016.

[3] 蔡晓妍,戴冠中,杨黎斌.谱聚类算法综述.计算机科学,2008,35(7):14−18. [doi: 10.3969/j.issn.1002-137X.2008.07.004]

[7] 陈黎飞,郭躬德,姜青山.自适应的软子空间聚类算法.软件学报,2010,21(10):2513−2523. http://www.jos.org.cn/1000-9825/3763.htm [doi: 10.3724/SP.J.1001.2010.03763]

[11] 孔锐,张国宣,施泽生,等.基于核的 *K*-均值聚类.计算机工程,2004,30(11):12−13,80. [doi: 10.3969/j.issn.1000-3428.2004.11.005]

徐鲲鹏(1994－),男,硕士,主要研究领域为数据挖掘,模式识别,机器学习.

孙浩军(1963－),男,博士,教授,CCF 专业会员,主要研究领域为数据挖掘,模式识别,信息系统.

陈黎飞(1972－),男,博士,教授,博士生导师,主要研究领域为数据挖掘,模式识别,机器学习.

王备战(1965－),男,博士,教授,博士生导师,CCF 专业会员,主要研究领域为数据挖掘,数据库与数据仓库,软件体系结构.