计算机研究与发展 DOI:10.7544/issn1000-1239.2016.20150131
Journal of Computer Research and Development 53(5):1018-1028,2016

一种基于信息熵的混合数据属性加权聚类算法

赵兴旺 梁吉业
(山西大学计算机与信息技术学院 太原 030006)
(计算智能与中文信息处理教育部重点实验室(山西大学) 太原 030006)
(zhaoxw84@163.com)

An Attribute Weighted Clustering Algorithm for Mixed Data Based on Information Entropy

Zhao Xingwang and Liang Jiye
(*School of Computer and Information Technology, Shanxi University, Taiyuan* 030006)
(*Key Laboratory of Computational Intelligence and Chinese Information Processing (Shanxi University), Ministry of Education, Taiyuan* 030006)

Abstract In real applications, mixed data sets with both numerical attributes and categorical attributes at the same time are more common. Recently, clustering analysis for mixed data has attracted more and more attention. In order to solve the problem of attribute weighting for high-dimensional mixed data, this paper proposes an attribute weighted clustering algorithm for mixed data based on information entropy. The main work includes: an extended Euclidean distance is defined for mixed data, which can be used to measure the difference between the objects and clusters more accurately and objectively. And a generalized mechanism is presented to uniformly assess the compactness and separation of clusters based on within-cluster entropy and between-cluster entropy. Then a measure of the importance of attributes is given based on this mechanism. Furthermore, an attribute weighted clustering algorithm for mixed data based on information entropy is developed. The effectiveness of the proposed algorithm is demonstrated in comparison with the widely used state-of-the-art clustering algorithms for ten real life datasets from UCI. Finally, statistical test is conducted to show the superiority of the results produced by the proposed algorithm.

Key words clustering analysis; mixed data; attribute weighting; information entropy; dissimilarity measure

摘 要 同时兼具数值型和分类型属性的混合数据在实际应用中普通存在,混合数据的聚类分析越来越受到广泛的关注.为解决高维混合数据聚类中属性加权问题,提出了一种基于信息熵的混合数据属性加权聚类算法,以提升模式发现的效果.工作主要包括:首先为了更加准确客观地度量对象与类之间的差异性,设计了针对混合数据的扩展欧氏距离;然后,在信息熵框架下利用类内信息熵和类间信息熵给出

收稿日期:2015-02-12;**修回日期**:2015-06-09
基金项目:国家自然科学基金项目(61432011,U1435212,61402272);国家"九七三"重点基础研究发展计划基金项目(2013CB329404);山西省自然科学基金项目(2013021018-1)
This work was supported by the National Natural Science Foundation of China (61432011, U1435212, 61402272), the National Basic Research Program of China (973 Program) (2013CB329404), and the Natural Science Foundation of Shanxi Province of China (2013021018-1).
通信作者:梁吉业(ljy@sxu.edu.cn)

了聚类结果中类内抱团性及一个类与其余类分离度的统一度量机制,并基于此给出了一种属性重要性度量方法,进而设计了一种基于信息熵的属性加权混合数据聚类算法.在10个UCI数据集上的实验结果表明,提出的算法在4种聚类评价指标下优于传统的属性未加权聚类算法和已有的属性加权聚类算法,并通过统计显著性检验表明本文提出算法的聚类结果与已有算法聚类结果具有显著差异性.

关键词 聚类分析;混合数据;属性加权;信息熵;相异性度量

中图法分类号 TP391

聚类分析的主要目的在于发现数据中隐含的类结构,将物理或抽象对象划分为不同的类,使得同一类内对象之间相似度较大,而不同类的对象之间相似度较小.作为一种主要的探索性数据分析工具,聚类分析目前已经在机器学习、数据挖掘、模式识别、生物信息学、统计学和社会计算等领域都得到了广泛的研究和应用[1].

半个多世纪以来,研究者针对不同的应用领域已经提出了诸多聚类算法,主要分为层次聚类算法和划分式聚类算法[2-4].其中,划分式聚类算法由于简单、高效、易实现等优点得到了广泛的应用.在传统的划分式聚类过程中,都假定各个属性对聚类的贡献程度相同,即在相似性或相异性度量的计算中所有属性的权重相同.而在大部分实际应用中,用户期望得到的聚类结果,对参与聚类的各个属性的重要程度往往并不相同.特别是在高维数据聚类过程中,样本空间中各属性对聚类效果贡献大小不同成为了一个不可回避的问题[5].因此,识别不同属性在聚类过程中的差异程度,从而提高聚类结果的质量,研究聚类过程中属性自动加权技术具有非常重要的意义.

近年来,为了计算不同属性对聚类贡献程度的差异性,许多学者针对聚类算法中属性加权问题已经开展了一些研究,提出了一系列属性自动加权聚类算法,这些研究大多针对数值型数据[6-8].然而现实生活中遇到的数据既可能是数值型数据,也可能是分类型数据,或同时由数值型和分类型属性共同描述的混合数据.与数值型数据相比,分类型数据的某一属性的取值是有限集合中的某一个值,而且取值之间没有序关系,这些特点使得分类型数据之间相似性度量的定义更为困难,从而使得数值型数据属性加权聚类算法无法直接应用于分类型数据[9-11].在分类型数据属性加权聚类算法方面,文献[10]提出了一种分类型数据属性加权聚类算法,在计算属性权重的过程中同时考虑了类中心出现的频率和类内对象到类中心平均距离.文献[11]提出了一个基于非模的分类型数据属性自动加权聚类方法,依据属性取值的总体分布情况对属性赋予不同权重.文献[12]设计了一种基于类内信息熵的分类型数据软子空间聚类算法,在每一个类中根据属性重要度对不同属性赋予不同的权重.

混合数据由于同时具有数值型属性和分类型属性,在对象之间相似性或相异性度量的定义和不同类型属性加权机制方面更为困难.其中,1998年,文献[13]通过把K-Modes算法和K-Means算法进行简单集成提出了针对混合数据的K-Prototypes算法.在该算法中,对象与类中心之间的相异性度量同时考虑了数值型数据和分类型数据,并通过参数来控制数值型部分和分类型部分贡献的大小.其中,在相异性度量方面,数值型属性之间的相异性度量采用欧氏距离,分类型属性之间的相异性通过简单0-1匹配来度量;在类中心表示方面,数值型数据部分采用均值(means)表示,而分类型数据分别采用众数(modes)表示.K-Prototypes算法由于简单易实现,已经在混合数据聚类中得到了广泛的应用.但是,欧氏距离和简单0-1匹配相异性度量的取值范围不同,而且不同类型数据之间的相异性度量仅仅通过一个参数进行调节,因此在聚类过程中不能客观地体现对象与类之间相异性在数值属性和分类属性各个属性的重要度.针对以上问题,一些学者已经进行了一些探索[14-16].如文献[14]针对分类数据部分,通过考虑同一属性下不同属性值的出现频率给出了一种新的相异性度量方法,并利用倒数比例计算法给出了一种新的类中心更新方式,通过与K-Means算法集成进而提出了一个可以处理混合数据的聚类算法.文献[15]给出了一个针对分类型和数值型属性统一的相似性度量方法,消除了不同类型数据之间度量量纲的差异性,同时也免去了不同度量之间参数的设置.该方法针对分类型数据,利用信息熵对不同分类属性的重要性进行了刻画.但是,在分类型属性加权过程中直接计算各个属性在所有数据上的信息熵,并没有考虑不同类之间的差异性;另外在计算相似性度量过程中把数值型数据的所有属

性当作整体计算得到一个相似度度量的数值,然后与分类型属性部分进行加权,这种计算方法显然削弱了数值型数据部分的权重.

通过以上分析可知,已有的属性加权聚类算法主要存在3个局限性:

1) 大多数加权聚类方法仅仅针对数值型或分类型单一类型数据进行属性加权,这些方法在实际应用中存在一定的局限性;

2) 已有属性加权方法或者依赖于整体数据集属性值的总体分布情况(属性重要性与数据分布的离散程度成反比,数据集中某个属性的取值越集中,该属性的重要性就越高),或者根据对象到类中心的距离定义的分散度来衡量(在某个类中某属性下的分散度越低,则该属性在该类的重要性越高),而并没有考虑不同类之间属性值分布的差异性,这必然导致属性权重计算上的偏差,从而影响聚类质量;

3) 针对现实生活中广泛存在的混合型数据,已有方法在对象与类中心之间相似性或相异性度量方面,通常采用不同的度量机制,存在量纲不同的问题,不能客观地反映混合数据之间的差异性.

针对以上问题,本文基于信息熵理论提出了一个混合数据属性加权聚类算法. 1)为了更加准确客观地度量对象与类之间的差异性,设计了一种针对混合数据的扩展欧氏距离;2)在信息熵框架下利用类内信息熵和类间信息熵给出了聚类结果中类内抱团性及一个类与其余类分离度的统一度量机制,并基于此给出了一种属性加权方法;3)在UCI数据集上实验结果表明,本文提出的属性加权聚类算法在4种评价指标下优于传统的未加权聚类算法和已有加权聚类算法.

1 相关工作

本节主要对混合型数据聚类相关背景知识和*K*-Prototypes聚类算法进行介绍.

设 $X=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N\}$ 表示由 N 个对象组成的待聚类的全体数据集,其中 $\boldsymbol{x}_i(1\leqslant i\leqslant N)$ 表示数据集中第 i 个对象且由 $A_1^{\mathrm{r}},A_2^{\mathrm{r}},\cdots,A_p^{\mathrm{r}},A_{p+1}^{\mathrm{c}},\cdots,A_m^{\mathrm{c}}$ 共 m 个属性进行刻画,其中 $A_1^{\mathrm{r}},A_2^{\mathrm{r}},\cdots,A_p^{\mathrm{r}}$ 为 p 个数值型属性,$A_{p+1}^{\mathrm{c}},A_{p+2}^{\mathrm{c}},\cdots A_m^{\mathrm{c}}$ 为 m-p 个分类型属性. 对于分类型属性 $A_j^{\mathrm{c}}(p+1\leqslant j\leqslant m)$ 的值域为 $D(A_j^{\mathrm{c}})=\{a_j^{(1)},a_j^{(2)},\cdots,a_j^{(n_j)}\}$,其中 n_j 表示属性 A_j^{c} 中值域的个数. 混合数据集中的对象 $\boldsymbol{x}_i\in X$ 可以用一个 m 维的向量来表示,即 $\boldsymbol{x}_i=(\boldsymbol{x}_i^{\mathrm{r}},\boldsymbol{x}_i^{\mathrm{c}})$,其中 $\boldsymbol{x}_i^{\mathrm{r}}=(x_{i,1}^{\mathrm{r}},x_{i,2}^{\mathrm{r}},\cdots,x_{i,p}^{\mathrm{r}})$ 表示数值型数据部分的取值,$\boldsymbol{x}_i^{\mathrm{c}}=(x_{i,p+1}^{\mathrm{c}},x_{i,p+2}^{\mathrm{c}},\cdots,x_{i,m}^{\mathrm{c}})$ 表示分类型数据部分的取值. 假设由数据集 X 划分得到的聚类结果为 $C_1,C_2,\cdots,C_k$ 共 k 个类,即 $C=\{C_1,C_2,\cdots,C_k\}$ 且 $C_i\cap C_j=\varnothing$,$\bigcup_{l=1}^{k}C_l=C(i,j=1,2,\cdots,k;i\neq j)$,其中 $C_l(1\leqslant l\leqslant k)$ 的类中心表示为 $\boldsymbol{z}_l=(\boldsymbol{z}_l^{\mathrm{r}},\boldsymbol{z}_l^{\mathrm{c}})$,其中,数值型数据和分类型数据部分的类中心分别为 $\boldsymbol{z}_l^{\mathrm{r}}=(z_{l,1}^{\mathrm{r}},z_{l,2}^{\mathrm{r}},\cdots,z_{l,p}^{\mathrm{r}})$,$\boldsymbol{z}_l^{\mathrm{c}}=(z_{l,p+1}^{\mathrm{c}},z_{l,p+2}^{\mathrm{c}},\cdots,z_{l,m}^{\mathrm{c}})$.

K-Prototypes算法[13]中,假设数据集 X 在迭代过程中的一个中间类结果为 $C=\{C_1,C_2,\cdots,C_k\}$,$\boldsymbol{z}_l=(\boldsymbol{z}_l^{\mathrm{r}},\boldsymbol{z}_l^{\mathrm{c}})$,其中 $\boldsymbol{z}_l^{\mathrm{r}}=(z_{l,1}^{\mathrm{r}},z_{l,2}^{\mathrm{r}},\cdots,z_{l,p}^{\mathrm{r}})$,$\boldsymbol{z}_l^{\mathrm{c}}=(z_{l,p+1}^{\mathrm{c}},z_{l,p+2}^{\mathrm{c}},\cdots,z_{l,m}^{\mathrm{c}})$ 为类 $C_l\in C$ 的类中心,对象 $\boldsymbol{x}_i\in X$ 与类中心 $\boldsymbol{z}_l$ 的相异性度量同时考虑数值型数据和分类型数据,并通过参数 γ 来控制相异性度量中数值型部分和分类型部分贡献的大小,定义如下:

$$d(\boldsymbol{x}_i,\boldsymbol{z}_l)=d_{\mathrm{r}}(\boldsymbol{x}_i^{\mathrm{r}},\boldsymbol{z}_l^{\mathrm{r}})+\gamma d_{\mathrm{c}}(\boldsymbol{x}_i^{\mathrm{c}},\boldsymbol{z}_l^{\mathrm{c}}),\tag{1}$$

其中,d_{r} 和 d_{c} 分别表示对象与类中心在数值型和分类型属性下的相异性度量,d_{r} 表示欧氏距离 $d_{\mathrm{r}}(\boldsymbol{x}_i^{\mathrm{r}},\boldsymbol{z}_l^{\mathrm{r}})=\sum_{t=1}^{p}(x_{i,t}^{\mathrm{r}}-z_{l,t}^{\mathrm{r}})^2$,$d_{\mathrm{c}}$ 表示0-1简单匹配相异性度量 $d_{\mathrm{c}}(\boldsymbol{x}_i^{\mathrm{c}},\boldsymbol{z}_l^{\mathrm{c}})=\sum_{t=p+1}^{m}\delta(x_{i,t}^{\mathrm{c}},z_{l,t}^{\mathrm{c}})$,其中:

$$\delta(x_{i,t}^{\mathrm{c}},z_{l,t}^{\mathrm{c}})=\begin{cases}1, & x_{i,t}^{\mathrm{c}}\neq z_{l,t}^{\mathrm{c}},\\ 0, & x_{i,t}^{\mathrm{c}}=z_{l,t}^{\mathrm{c}}.\end{cases}$$

K-Prototypes算法最小化以下目标函数:

$$F(W,Z)=\sum_{l=1}^{k}\sum_{i=1}^{N}w_{li}d(\boldsymbol{x}_i,\boldsymbol{z}_l),$$

其中,$w_{li}\in\{0,1\}$,$1\leqslant l\leqslant k$,$1\leqslant i\leqslant n$;

$$\sum_{l=1}^{k}w_{li}=1,\ 1\leqslant i\leqslant n;$$

$$0<\sum_{i=1}^{n}w_{li}<n,\ 1\leqslant l\leqslant k.$$

$w_{li}=1$ 时表示第 i 个对象属于第 l 个类,$w_{li}=0$ 时表示第 i 个对象不属于第 l 个类.

为了使目标函数 F 在给定的约束条件下达到极小值,采用如下算法进行计算.

算法1. *K*-Prototypes聚类算法.

Step1. 从数据集 X 中随机选取 k 个对象作为初始类中心;

Step2. 根据式(1)计算对象与类中心之间的距离,并根据最近原则将每个对象分配到离它最近的类中;

Step3. 更新聚类中心,其中数值属性部分通过计算同一类中对象的平均值得到,分类型属性部分

通过计算类中各属性值出现的频率高低来确定;

Step4. 重复 Step2,Step3,直到目标函数 F 不再发生变化为止.

K-Prototypes 算法由于简单、高效、易实现,已经得到了广泛的应用.但是,在相异性度量的定义中存在数值型数据和分类型数据部分量纲不同、参数 γ 难以确定、不能客观地反映对象与类中心的差异性的缺陷;而且在定义对象与类中心的差异性度量的过程中未考虑各个属性的重要性.本文在 K-Prototypes 算法的框架下,提出了一个针对混合数据迭代式的属性加权聚类算法.

2 基于信息熵的混合数据属性加权聚类算法

在信息论中,信息熵是一种用于度量系统不确定性的度量方法.在一个系统中,某个属性取值的不确定性程度越大,表明系统越混乱,在该属性下系统的信息熵越大,它提供的信息量越小,该属性的重要性也就越小;反之,某个属性取值的不确定性程序越小,表明系统越有序,在该属性下信息熵越小,它提供的信息量越大,该属性的重要性也就越大.作为一种有效的度量机制,信息熵已经在聚类分析、孤立点检测、不确定性度量等领域得到了广泛的应用.本文根据各个属性下取值的不确定程度,利用类内和类间信息熵来度量各个属性在聚类过程中的重要程度.由于描述对象的属性类型不同,信息熵的计算方法也不同,本节将分别进行描述.

在给出数值型和分类型数据基于信息熵度量的属性加权机制之前,首先定义一种新的混合数据相异性度量方法.

2.1 混合数据相异性度量方法

由于 K-Prototypes 算法中分类型数据部分的聚类中心仅仅将某类中当前属性值域中出现频率最高的取值(即 modes)作为类中心,忽略了该属性的其余取值情况,而且往往会出现类中心不唯一的情况.本节首先给出分类型数据一种新的模糊类中心表示方式,并基于此将传统的欧氏距离扩展到混合数据,使得能够在统一的框架下更加客观地度量混合数据中对象与类之间的相异性.

定义 1. 设 C_l 表示数据集 X 在聚类过程中得到的一个类,则分类型属性部分模糊类中心表示为

$$\hat{\boldsymbol{z}}_l^{\mathrm{c}}=(\hat{z}_{l,p+1}^{\mathrm{c}},\hat{z}_{l,p+2}^{\mathrm{c}},\cdots,\hat{z}_{l,m}^{\mathrm{c}}),\tag{2}$$

其中,$\hat{z}_{l,t}^{\mathrm{c}}=\{(a_t^1,f_{l,t}^1),(a_t^2,f_{l,t}^2),\cdots,(a_t^{n_t},f_{l,t}^{n_t})\}$,$p+1\leqslant t\leqslant m$ 表示第 t 个分类型属性 A_t^{c} 下模糊类中心表示形式.n_t 表示属性 A_t^{c} 值域的个数,$f_{l,t}^w(1\leqslant w\leqslant n_t)$ 表示第 t 个分类型属性 A_t^{c} 的值域中取值 a_t^w 在类 C_l 中出现的频率,且 $\sum\limits_{w=1}^{n_t}f_{l,t}^w=1$.

由定义 1 可知,数据集中的单一对象也可以表示为模糊类中心的形式,是模糊类中心表示的一种特殊形式.即针对某一属性下,当前对象的属性值对应的频率为 1,其余值域对应的频率为 0.

基于以上给出的分类型数据新的类中心表示方式,下面给出扩展的欧氏距离度量.

定义 2. 设数据集 X 在聚类过程中,类 C_l 中分类型属性部分的类中心表示为 $\hat{\boldsymbol{z}}_l^{\mathrm{c}}=(\hat{z}_{l,p+1}^{\mathrm{c}},\hat{z}_{l,p+2}^{\mathrm{c}},\cdots,\hat{z}_{l,m}^{\mathrm{c}})$,其中 $\hat{z}_{l,t}^{\mathrm{c}}=\{(a_t^1,f_{l,t}^1),(a_t^2,f_{l,t}^2),\cdots,(a_t^{n_t},f_{l,t}^{n_t})\}$,$p+1\leqslant t\leqslant m$;对象 $\boldsymbol{x}_i\in X$ 中分类型数据部分 $\boldsymbol{x}_i^{\mathrm{c}}$ 对应的模糊类中心表示为 $\hat{\boldsymbol{x}}_i^{\mathrm{c}}=(\hat{x}_{i,p+1}^{\mathrm{c}},\hat{x}_{i,p+2}^{\mathrm{c}},\cdots,\hat{x}_{i,m}^{\mathrm{c}})$,$\hat{x}_{i,t}^{\mathrm{c}}=\{(a_t^1,f_{i,t}^1),(a_t^2,f_{i,t}^2),\cdots,(a_t^{n_t},f_{i,t}^{n_t})\}$,$p+1\leqslant t\leqslant m$,则对象 $\boldsymbol{x}_i^{\mathrm{c}}$ 与类中心 $\hat{\boldsymbol{z}}_l^{\mathrm{c}}$ 之间的欧氏距离定义为

$$\hat{d}_{\mathrm{c}}(\boldsymbol{x}_i^{\mathrm{c}},\hat{\boldsymbol{z}}_l^{\mathrm{c}})=\sum_{t=p+1}^{m}\frac{1}{n_t}\sum_{s=1}^{n_t}(f_{i,t}^s-f_{l,t}^s)^2.\tag{3}$$

由定义 2 可知,由于每一属性下相异度的取值范围在[0,1]之间,则 $0\leqslant\hat{d}_{\mathrm{c}}(\boldsymbol{x}_i^{\mathrm{c}},\hat{\boldsymbol{z}}_l^{\mathrm{c}})\leqslant m-p$.

定义 3. 设 $C_l\subset X$ 是聚类过程中得到的一个类,$\hat{\boldsymbol{z}}_l=(\boldsymbol{z}_l^{\mathrm{r}},\hat{\boldsymbol{z}}_l^{\mathrm{c}})$ 为 C_l 的类中心,其中 $\boldsymbol{z}_l^{\mathrm{r}}$ 为用均值表示的数值型部分的类中心,$\hat{\boldsymbol{z}}_l^{\mathrm{c}}$ 为由定义 1 中给出的分类型数据的模糊类中心,则对象 $\boldsymbol{x}_i\in X$ 与类中心 $\hat{\boldsymbol{z}}_l$ 之间扩展的欧氏距离定义如下:

$$\hat{d}(\boldsymbol{x}_i,\boldsymbol{z}_l)=d_{\mathrm{r}}(\boldsymbol{x}_i^{\mathrm{r}},\boldsymbol{z}_l^{\mathrm{r}})+\hat{d}_{\mathrm{c}}(\boldsymbol{x}_i^{\mathrm{c}},\hat{\boldsymbol{z}}_l^{\mathrm{c}}),\tag{4}$$

其中,$d_{\mathrm{r}}(\boldsymbol{x}_i^{\mathrm{r}},\boldsymbol{z}_l^{\mathrm{r}})$ 表示数值型属性部分的欧氏距离,即 $d_{\mathrm{r}}(\boldsymbol{x}_i^{\mathrm{r}},\boldsymbol{z}_l^{\mathrm{r}})=\sum\limits_{j=1}^{p}(x_{i,j}^{\mathrm{r}}-z_{l,j}^{\mathrm{r}})^2$;$\hat{d}_{\mathrm{c}}(\boldsymbol{x}_i^{\mathrm{c}},\hat{\boldsymbol{z}}_l^{\mathrm{c}})$ 表示由定义 2 给出的分类型数据部分基于模糊类中心表示的扩展的欧氏距离.

2.2 基于信息熵的数值型属性加权机制

针对数值型数据,匈牙利数学家 Renyi[17] 于 1961 年提出了一个可以度量连续型随机变量的信息熵,称作 Renyi 熵.设连续型随机变量 $\boldsymbol{x}$ 的概率密度函数为 $f(\boldsymbol{x})$,则该随机变量的 Renyi 熵定义为

$$H_{\mathrm{R}}(\boldsymbol{x})=\frac{1}{1-\alpha}\ln\int f^{\alpha}(\boldsymbol{x})\mathrm{d}\boldsymbol{x},\ \alpha>0,\alpha\neq1.\tag{5}$$

当 $\alpha=2$,$H_{\mathrm{R}}(\boldsymbol{x})=-\ln\int f^2(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$,记为二阶 Renyi 熵.

二阶 Renyi 熵由于计算方便、具有很好的性质,在实际应用中得到了广泛的应用. Parzen 窗口估计法作为一种非参数估计方法,通常用于利用已知样本对总体样本的概率密度进行估计[18]. 因此二阶 Renyi 熵中的密度函数可以用 Parzen 窗口估计法来进行估计. 设 $X=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N\},\boldsymbol{x}_i\in\mathbb{R}^d$,是一个由独立同分布的 N 个数据对象组成的数据集合,则对于任意随机变量 $\boldsymbol{x}\in X$ 利用 Parzen 窗口估计法估计出的概率密度为

$$\hat{f}(\boldsymbol{x})=\frac{1}{N}\sum_{i=1}^{N}W_{\sigma^2}(\boldsymbol{x},\boldsymbol{x}_i),\tag{6}$$

其中,W_{σ^2} 表示 Parzen 窗函数,σ^2 表示窗宽. 通常选取高斯核函数作为 Parzen 窗函数,即

$$W_{\sigma^2}(\boldsymbol{x},\boldsymbol{x}_i)=\frac{1}{(2\pi)^{\frac{d}{2}}\sigma^d}\exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{x}_i)^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{x}_i)}{2\sigma^2}\right).\tag{7}$$

用 Parzen 窗口估计法得到的 $\hat{f}(\boldsymbol{x})$代替二阶 Renyi 熵中的 $f(\boldsymbol{x})$即可得到数据集 X 的 Renyi 熵[19]:

$$H_{\mathrm{R}}(\boldsymbol{x})=-\ln\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{d}W_{2\sigma^2}(x_{i,t},x_{j,t}),$$

其中,$W_{2\sigma^2}(x_{i,t},x_{j,t})=\dfrac{1}{(2\pi)^d\sigma^{2d}}\exp\left(-\dfrac{(x_{i,t}-x_{j,t})^2}{2\sigma^2}\right)$.

通过分析可知,基于 Parzen 窗口估计法得到的 Renyi 熵可能出现负值的情况. 为了保证熵值为正及计算的方便,本文在 $W_{2\sigma^2}(x_{i,t},x_{j,t})$之前乘以一个相对小的正数,即:

$$W'_{2\sigma^2}(x_{i,t},x_{j,t})=\exp\left(-\frac{(x_{i,t}-x_{j,t})^2}{2\sigma^2}\right).$$

基于以上给出的 Renyi 熵,数值型数据在某一属性下类内熵、类间熵分别定义如下[20].

定义 4. 设 $X^{\mathrm{r}}=\{\boldsymbol{x}_1^{\mathrm{r}},\boldsymbol{x}_2^{\mathrm{r}},\cdots,\boldsymbol{x}_N^{\mathrm{r}}\}$,其中 $\boldsymbol{x}_i^{\mathrm{r}}=(x_{i,1}^{\mathrm{r}},x_{i,2}^{\mathrm{r}},\cdots,x_{i,p}^{\mathrm{r}})(1\leqslant i\leqslant N)$是一个由 p 个属性描述的数值型数据集,在聚类过程中被划分为 k 个类 $C^k=\{C_1^{\mathrm{r}},C_2^{\mathrm{r}},\cdots,C_k^{\mathrm{r}}\}$. 在属性 $A_t^{\mathrm{r}}(1\leqslant t\leqslant p)$下,任意一个类 $C_{k'}^{\mathrm{r}}\in C^k$ 的类内熵 $WEN(C_{k'}^{\mathrm{r}},A_t^{\mathrm{r}})$定义为

$$WEN(C_{k'}^{\mathrm{r}},A_t^{\mathrm{r}})=-\ln\frac{1}{N_{k'}^2}\sum_{\boldsymbol{x}_i\in C_{k'}^{\mathrm{r}}}\sum_{\boldsymbol{x}_j\in C_{k'}^{\mathrm{r}}}W'_{2\sigma^2}(x_{i,t},x_{j,t}),\tag{8}$$

其中,$N_{k'}=|C_{k'}^{\mathrm{r}}|$表示类 $C_{k'}^{\mathrm{r}}$ 中对象的个数.

上述定义给出的类内熵反映了聚类划分结果中某一类在不同的属性下数据分布的不确定程度. 即在一个类中如果某个属性下的类内熵越小,则该属性在该类中的不确定性越小,聚类过程中该属性的权重就越大.

定义 5. 设 $X^{\mathrm{r}}=\{\boldsymbol{x}_1^{\mathrm{r}},\boldsymbol{x}_2^{\mathrm{r}},\cdots,\boldsymbol{x}_N^{\mathrm{r}}\}$,其中 $\boldsymbol{x}_i^{\mathrm{r}}=(x_{i,1}^{\mathrm{r}},x_{i,2}^{\mathrm{r}},\cdots,x_{i,p}^{\mathrm{r}})(1\leqslant i\leqslant N)$是一个由 p 个属性描述的数值型数据集,在聚类过程中被划分为 k 个类 $C^k=\{C_1^{\mathrm{r}},C_2^{\mathrm{r}},\cdots,C_k^{\mathrm{r}}\}$. 在属性 $A_t^{\mathrm{r}}(1\leqslant t\leqslant p)$下,任意一个类 $C_{k'}^{\mathrm{r}}\in C^k$ 与其余类之间的平均类间熵定义为

$$BEN(C_{k'}^{\mathrm{r}},A_t^{\mathrm{r}})=\frac{1}{k-1}\sum_{C_s^{\mathrm{r}}\in C^k,s\neq k'}-\ln\frac{1}{N_{k'}N_s}\sum_{\boldsymbol{x}_i\in C_{k'}^{\mathrm{r}}}\sum_{\boldsymbol{x}_j\in C_s^{\mathrm{r}}}W'_{2\sigma^2}(x_{i,t},x_{j,t}),\tag{9}$$

其中,$N_{k'}=|C_{k'}^{\mathrm{r}}|$和 $N_s=|C_s^{\mathrm{r}}|$分别表示类 $C_{k'}^{\mathrm{r}}$ 和 C_s^{r} 中对象的个数.

显然,如果 $BEN(C_{k'}^{\mathrm{r}},A_t)$的值越大表明在属性 A_t^{r} 下类 $C_{k'}^{\mathrm{r}}$ 与其余类的分离程度越大,则聚类过程中属性 A_t^{r} 对类 $C_{k'}^{\mathrm{r}}$ 的权重也越大.

基于定义 4 和定义 5,下面给出数值型数据属性加权的定义.

定义 6. 设 $X^{\mathrm{r}}=\{\boldsymbol{x}_1^{\mathrm{r}},\boldsymbol{x}_2^{\mathrm{r}},\cdots,\boldsymbol{x}_N^{\mathrm{r}}\}$,其中 $\boldsymbol{x}_i^{\mathrm{r}}=(x_{i,1}^{\mathrm{r}},x_{i,2}^{\mathrm{r}},\cdots,x_{i,p}^{\mathrm{r}})$是一个由 p 个属性描述的数值型数据集,$1\leqslant i\leqslant N$,在聚类过程中被划分为 k 个类 $C^k=\{C_1^{\mathrm{r}},C_2^{\mathrm{r}},\cdots,C_k^{\mathrm{r}}\}$. 对于任意一个类 $C_{k'}^{\mathrm{r}}\in C^k$,属性 $A_t^{\mathrm{r}}(1\leqslant t\leqslant p)$的权重定义为

$$WN(C_{k'}^{\mathrm{r}},A_t^{\mathrm{r}})=\exp(-WEN(C_{k'}^{\mathrm{r}},A_t^{\mathrm{r}}))\times(1-\exp(-BEN(C_{k'}^{\mathrm{r}},A_t^{\mathrm{r}}))).\tag{10}$$

由定义 6 可知,属性权重与类内熵成反比,与类间熵成正比,而且类内熵、类间熵都已经归一化到[0,1]之间,因此,各个属性权重的范围为[0,1]. 由定义 5 可知,如果聚类个数为 2 时,类与类之间的平均类间熵则相等,此时属性权重只与类内熵有关.

2.3 基于信息熵的分类型属性加权机制

针对分类型数据,文献[21]中提出了一个既可以用来度量随机性,又可以度量模糊性的互补信息熵. 目前,该信息熵已经在不确定性度量、特征选择、聚类分析、孤立点检测等领域得到了广泛的应用[12,22-26].

基于互补熵,分类型数据的类内熵、类间熵分别定义如下[25]:

定义 7. 设 $X^{\mathrm{c}}=\{\boldsymbol{x}_1^{\mathrm{c}},\boldsymbol{x}_2^{\mathrm{c}},\cdots,\boldsymbol{x}_N^{\mathrm{c}}\}$,其中 $\boldsymbol{x}_i^{\mathrm{c}}=(x_{i,p+1}^{\mathrm{c}},x_{i,p+2}^{\mathrm{c}},\cdots,x_{i,m}^{\mathrm{c}})$是一个由 $m-p$ 个属性描述的分类型数据集,$1\leqslant i\leqslant N$,在聚类过程中被划分为 k 个类 $C^k=\{C_1^{\mathrm{c}},C_2^{\mathrm{c}},\cdots,C_k^{\mathrm{c}}\}$. 在属性 $A_t^{\mathrm{c}}(p+1\leqslant t\leqslant m)$下,任意一个类 $C_{k'}^{\mathrm{c}}\in C^k$ 的类内熵 $WEC(C_{k'}^{\mathrm{c}},A_t^{\mathrm{c}})$定义为

$$WEC(C_{k'}^{\mathrm{c}},A_t^{\mathrm{c}})=\sum_{s=1}^{n_t}\frac{|Y_s|}{|C_{k'}^{\mathrm{c}}|}\left(1-\frac{|Y_s|}{|C_{k'}^{\mathrm{c}}|}\right),\tag{11}$$

其中,n^t 表示分类型属性 A_t^{c} 值域的个数,$|Y_s|$表示

分类型属性 A_t^c 的值域中第 s 个取值 Y_s 在类 $C_{k'}^c$ 中出现的次数,$|C_{k'}^c|$ 表示类 $C_{k'}^c$ 中对象的个数.

通过分析可知,分类型数据在属性 $A_t^c(p+1\leqslant t\leqslant m)$ 下,任意一个类 $C_{k'}^c\in C^k$ 的类内熵 $(C_{k'}^c,A_t^c)$ 与该类内部任意对象两两之间的简单 0-1 匹配相异性度量具有如下关系[25]:

$$WEC(C_{k'}^c,A_t^c)=\frac{1}{|C_{k'}^c|}\sum_{x_i\in C_{k'}^c}\sum_{x_j\in C_{k'}^c}\delta(x_{i,t},x_{j,t}),\quad(12)$$

其中,$\delta(x_{i,t},x_{j,t})=\begin{cases}1, & x_{i,t}\neq x_{j,t}.\\ 0, & x_{i,t}=x_{j,t}.\end{cases}$

式(12)表明分类型数据类内熵与类内平均距离是等价的.因此,可以从距离的角度来定义分类型数据一个类与其余类之间的平均类间熵.

定义 8. 设 $X^c=\{\boldsymbol{x}_1^c,\boldsymbol{x}_2^c,\cdots,\boldsymbol{x}_N^c\}$,其中 $\boldsymbol{x}_i^c=(x_{i,p+1}^c,x_{i,p+2}^c,\cdots,x_{i,m}^c)$ 是一个由 $m-p$ 个属性描述的分类型数据集,$1\leqslant i\leqslant N$,在聚类过程中被划分为 k 个类 $C^k=\{C_1^c,C_2^c,\cdots,C_k^c\}$.在属性 $A_t^c(p+1\leqslant t\leqslant m)$ 下,任意一个类 $C_{k'}^c\in C^k$ 与其余类之间的平均类间熵定义为

$$BEC(C_{k'}^c,A_t^c)=\frac{1}{k-1}\sum_{C_s^c\in C^k,s\neq k'}\times\frac{1}{|C_{k'}^c||C_s^c|}\sum_{\boldsymbol{x}_i\in C_{k'}^c}\sum_{\boldsymbol{x}_j\in C_s^c}\delta(x_{i,t},x_{j,t}),\quad(13)$$

其中,$|C_{k'}^c|$ 和 $|C_s^c|$ 分别表示类 $C_{k'}^c$ 和 C_s^c 中对象的个数;$\delta(x_{i,t},x_{j,t})=\begin{cases}1, & x_{i,t}\neq x_{j,t}.\\ 0, & x_{i,t}=x_{j,t}.\end{cases}$

与数值型数据的信息熵类似,定义 7 和定义 8 给出的分类型数据的类内熵、类间熵分别表示在一个聚类划分结果中一个类在不同的属性下类内数据分布的不确定程度和与其余类的分离程度.

基于定义 7 和定义 8,下面给出分类型数据属性加权的定义.

定义 9. 设 $X^c=\{\boldsymbol{x}_1^c,\boldsymbol{x}_2^c,\cdots,\boldsymbol{x}_N^c\}$,其中 $\boldsymbol{x}_i^c=(x_{i,p+1}^c,x_{i,p+2}^c,\cdots,x_{i,m}^c)$ 是一个由 $m-p$ 个属性描述的分类型数据集,$1\leqslant i\leqslant N$,在聚类过程中被划分为 k 个类 $C^k=\{C_1^c,C_2^c,\cdots,C_k^c\}$.对于任意一个类 $C_{k'}^c\in C^k$,属性 $A_t^c(p+1\leqslant t\leqslant m)$ 的权重定义为

$$WC(C_{k'}^c,A_t^c)=\exp(-WEC(C_{k'}^c,A_t^c))\times(1-\exp(-BEC(C_{k'}^c,A_t^c))).\quad(14)$$

2.4 基于信息熵的混合数据属性加权聚类算法

基于式(4)(10)(14),任一对象 $\boldsymbol{x}_i\in X$ 和类 C_l(类中心表示为 $\boldsymbol{z}_l$)的加权相异度度量定义为

$$WD(\boldsymbol{x}_i,\boldsymbol{z}_l)=\sum_{t=1}^{p}WN(C_l^r,A_t^r)(x_{i,t}^r-z_{l,t}^r)^2+\sum_{t=p+1}^{m}WC(C_l^c,A_t^c)\frac{1}{n_t}\sum_{s=1}^{n_t}(f_{i,t}^s-f_{l,t}^s)^2.\quad(15)$$

由式(15)可知,在信息熵机制下结合类内信息熵和类间信息熵给出了针对数值型和分类型数据统一的加权方法,而且不同类型属性下的相异性度量的范围都在[0,1]之间,避免了量纲不同的问题,能够更加客观地反映混合数据中对象与类之间的差异性.

设 $X=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N\}$ 是一个混合型数据集,将上述给出的加权相异性度量应用到 K-Prototypes 算法框架中,算法的目标函数定义为

$$F'(\boldsymbol{W},Z)=\sum_{l=1}^{k}\sum_{i=1}^{N}w_{li}\times WD(\boldsymbol{x}_i,\boldsymbol{z}_l),$$

其中:

$$w_{li}=\begin{cases}1, & WD(\boldsymbol{x}_i,\boldsymbol{z}_l)\leqslant WD(\boldsymbol{x}_i,\boldsymbol{z}_r),\forall 1\leqslant r\leqslant k.\\ 0, & \text{否则.}\end{cases}$$

$$1\leqslant l\leqslant k,1\leqslant i\leqslant n;$$

$$\sum_{l=1}^{k}w_{li}=1,\ 1\leqslant i\leqslant n;$$

$$0<\sum_{i=1}^{n}w_{li}<n,\ 1\leqslant l\leqslant k.$$

以上优化问题是一个非常复杂的非线性规划问题,和 K-Prototypes 类型算法类似,采用逐步迭代优化的策略,即首先固定聚类中心 Z,最小化目标函数 F' 得到隶属矩阵 $\boldsymbol{W}$;然后固定隶属矩阵 $\boldsymbol{W}$,最小化目标函数 F' 得到新的聚类中心 Z;如此迭代,直到目标函数 F' 收敛为止.算法描述如下.

算法 2. 基于信息熵的混合数据加权聚类算法.

输入:数据集 $X=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N\}$、类个数 k;

输出:聚类结果.

Step1. 从数据集 X 中随机选取 k 个不同的对象作为初始聚类中心;

Step2. 把 k 类中每一个属性的权重初始化为相同值,即任意一个类 $C_{k'}(1\leqslant k'\leqslant k)$ 在属性 $A_t(1\leqslant t\leqslant m)$ 的权重都为 $\frac{1}{m}$;

Step3. 根据式(15)计算对象与类中心之间的相异度,按照最近邻原则将数据对象划分到离它最近的聚类中心所代表的类中;

Step4. 更新聚类中心,其中数值属性部分通过计算同一类中对象取值的平均值得到,分类型属性部分根据定义 1 计算模糊类中心;

Step5. 根据定义 6 和定义 9 分别计算各个类在数值型和分类型数据部分各个属性的权重;

Step6. 重复 Step3～Step5,直到目标函数 F' 不再发生变化为止.

算法 2 的时间复杂度分析如下:在算法 2 的每一次迭代过程中需要更新各个类的属性权重,因此需要计算类内熵和类间熵,该步骤的时间复杂度为 $O(\sum_{i=1}^{k}\sum_{j=1}^{k}|C_i||C_j|)$,其中 k 表示聚类个数,$|C_i|$ 和 $|C_j|$ 分别表示类 C_i 和 C_j 中对象的个数.因此,算法的时间复杂度为 $O(Nkt\sum_{i=1}^{k}\sum_{j=1}^{k}|C_i||C_j|)$,其中 N 表示数据集对象个数,t 表示算法的迭代次数.

3 实验结果与分析

为了测试本文提出算法的有效性,我们从 UCI 真实数据集中分别选取了数值型、分类型和混合型 3 种不同类型的数据集进行了测试,并将本文提出的算法与 K-Prototypes 算法[13]、K-Centers 算法[14]和基于属性加权的 OCIL 算法[15]、改进的 K-Prototypes 算法[16]进行了比较.10 组数据集信息描述如表 1 所示,其中包括 3 个数值型数据、3 个分类型数据和 4 个混合型数据.

Table 1 The Summary of Data Sets' Characteristics

表 1 数据集信息描述

Data Sets	# Objects	# Numerical Attributes	# Categorical Attributes	# Classes
Segment	2 310	19	0	7
Waveform	5 000	22	0	3
Waveform+Noise	5 000	41	0	3
Promoters	106	0	58	2
DNA	3 190	0	60	3
Chess	3 196	0	36	2
Flag	194	10	18	8
Dermatology	366	1	33	6
German Credit	1 000	7	13	2
CMC	1473	2	7	3

为了对聚类结果的有效性进行评价,本文采用 3 个外部有效性评价指标和 1 个内部有效性评价指标对聚类结果评价.外部有效性评价指标包括聚类精度(clustering accuracy, CA)[14]、标准互信息(normalized mutual information, NMI)[27]和调整的兰特指数(adjusted rand index, ARI)[25];内部有效性指标包括混合数据分类效用函数(category utility function for mixed data, CUM)[25].

实验中,由于本文提出的新算法和被比较算法的聚类结果均受初始类中心选择的影响,不同的初始类中心可能有不同的聚类结果.因此,以下实验结果均为算法随机运行 50 次评价指标的平均值和方差.K-Prototypes 算法[13]、K-Centers 算法[15]中的权重参数 γ 根据作者建议分别设置为 $\gamma=1.5$ 和 $\gamma=0.5$.改进的 K-Prototypes 算法[16]中参数 λ 根据作者建议设置为 $\lambda=8$.另外,为了避免数值型不同属性间量纲的影响,在聚类之前对数值型数据进行了标准化处理.

3.1 数值型数据聚类结果分析

在不同评价指标下,本文提出的新算法和其他 4 种聚类算法在数值型数据上聚类结果如表 2～5 所示.由于 K-Prototypes 算法、K-Centers 算法和基于属性加权的 OCIL 算法在针对纯数值型数据聚类时,都退化为经典 K-Means 算法,因此本实验中针对数值型数据只将本文提出算法与 K-Means 算法、改进的 K-Prototypes 算法进行了比较.

Table 2 Comparison of CA Values (means±std) of Different Algorithms on Numerical Data

表 2 数值型数据聚类结果比较:CA 值(均值±方差)

Data Sets	K-Means	Improved K-Prototypes	Proposed Algorithm
Segment	0.5308±0.0003	**0.6025±0.0373**	0.5612±0.0042
Waveform	0.5199±0.0004	0.5327±0.0001	**0.5911±0.0007**
Waveform+Noise	0.6431±0.0313	0.5242±0.0000	**0.6978±0.0381**

Table 3 Comparison of NMI Values (means±std) of Different Algorithms on Numerical Data

表 3 数值型数据聚类结果比较:NMI 值(均值±方差)

Data Sets	K-Means	Improved K-Prototypes	Proposed Algorithm
Segment	0.3841±0.0000	**0.6200±0.0383**	0.4671±0.0049
Waveform	0.3644±0.0009	0.3568±0.0003	**0.4122±0.0003**
Waveform+Noise	0.6113±0.0153	0.3563±0.0000	**0.6406±0.0369**

Table 4 Comparison of ARI Values (means±std) of Different Algorithms on Numerical Data

表 4 数值型数据聚类结果比较:ARI 值(均值±方差)

Data Sets	K-Means	Improved K-Prototypes	Proposed Algorithm
Segment	0.2535±0.0001	0.2483±0.0517	**0.2789±0.0401**
Waveform	0.2515±0.0003	0.2515±0.0007	**0.2890±0.0005**
Waveform+Noise	0.4780±0.0388	0.2501±0.0001	**0.5139±0.0507**

Table 5 Comparison of CUM Values (means±std) of Different Algorithms on Numerical Data

表 5 数值型数据聚类结果比较:CUM 值(均值±方差)

Data Sets	*K*-Means	Improved *K*-Prototypes	Proposed Algorithm
Segment	**2.7943±0.0000**	1.5802±0.1308	2.7879±0.0317
Waveform	2.8066±0.0001	2.7923±0.0002	**2.9043±0.0021**
Waveform+Noise	1.4703±0.0651	**1.7951±0.0005**	1.6404±0.0820

由表 2～5 可知,从 CA,NMI,CUM 指标看,本文提出的算法在 Segment 数据集上获得的聚类结果劣于其他算法,在 CUM 指标下本文算法在 Waveform+Noise 数据集上获得的聚类结果劣于改进的 *K*-Prototypes 算法.除此之外,在不同指标下本文提出的聚类算法在其他数据集上的聚类结果均优于其他算法.

3.2 分类型数据聚类结果分析

在不同评价指标下,本文提出的新算法和其他 4 种聚类算法在分类型数据上聚类结果如表 6～9 所示:

Table 6 Comparison of CA Values (means±std) of Different Algorithms on Categorical Data

表 6 分类型数据聚类结果比较:CA 值(均值±方差)

Data Sets	*K*-Prototypes	*K*-Centers	OCIL	Improved *K*-Prototypes	Proposed Algorithm
Promoters	0.5988±0.0753	0.6281±0.0676	0.6292±0.0791	0.6146±0.0701	**0.6468±0.0794**
DNA	0.5188±0.0000	0.5217±0.0113	0.5702±0.0367	0.5211±0.0077	**0.6267±0.0452**
Chess	0.5557±0.0429	0.5531±0.0381	0.5581±0.0384	**0.8883±0.0000**	0.5701±0.0383

Table 7 Comparison of NMI Values (means±std) of Different Algorithms on Categorical Data

表 7 分类型数据聚类结果比较:NMI 值(均值±方差)

Data Sets	*K*-Prototypes	*K*-Centers	OCIL	Improved *K*-Prototypes	Proposed Algorithm
Promoters	0.0491±0.0636	0.0672±0.0672	0.0770±0.0784	0.0581±0.0634	**0.0906±0.0835**
DNA	0.0321±0.0191	0.0711±0.0503	0.1807±0.0741	0.0395±0.0227	**0.1877±0.0117**
Chess	0.0162±0.0261	0.0127±0.0165	0.0141±0.0182	0.0347±0.0339	**0.0439±0.0188**

Table 8 Comparison of ARI Values (means±std) of Different Algorithms on Categorical Data

表 8 分类型数据聚类结果比较:ARI 值(均值±方差)

Data Sets	*K*-Prototypes	*K*-Centers	OCIL	Improved *K*-Prototypes	Proposed Algorithm
Promoters	0.0525±0.0773	0.0750±0.0710	0.0829±0.0891	0.0629±0.0802	**0.1027±0.0982**
DNA	0.0167±0.0110	0.0402±0.0384	0.1103±0.0698	0.0230±0.0187	**0.1327±0.0120**
Chess	0.0185±0.0341	0.0159±0.0224	0.0182±0.0243	0.0373±0.0466	**0.0487±0.0219**

Table 9 Comparison of CUM Values (means±std) of Different Algorithms on Categorical Data

表 9 分类型数据聚类结果比较:CUM 值(均值±方差)

Data Sets	*K*-Prototypes	*K*-Centers	OCIL	Improved *K*-Prototypes	Proposed Algorithm
Promoters	0.4808±0.0525	0.4557±0.0570	0.4692±0.0578	0.3557±0.0484	**0.5129±0.0710**
DNA	0.2589±0.0507	0.1523±0.0426	0.2688±0.0510	0.1559±0.0184	**0.2972±0.0527**
Chess	0.3358±0.0709	0.4010±0.0807	0.4426±0.0913	0.2788±0.0668	**0.5156±0.1077**

由表 6～9 可知,本文提出的算法除了在 Chess 数据集上聚类结果的 CA 值劣于改进的 *K*-Prototypes 算法之外,其余聚类结果均均优于其他算法.

3.3 混合型数据聚类结果分析

在不同评价指标下,本文提出的新算法和其他 4 种聚类算法在混合数据上聚类结果如表 10～13 所示.

由表 10～13 可知,从 CA 和 ARI 评价指标看,K-Centers 算法在 Flag 和 German Credit 数据集上取得了最优的聚类结果.除此之外,本文提出的算法在其余数据集上均取得了最优的聚类结果.

Table 10 Comparison of CA Values (means±std) of Different Algorithms on Mixed Data

表 10 混合型数据聚类结果比较:CA 值(均值±方差)

Data Sets	*K*-Prototypes	*K*-Centers	OCIL	Improved *K*-Prototypes	Proposed Algorithm
Flag	0.4671±0.0254	**0.5019±0.0316**	0.4638±0.0203	0.4384±0.0379	0.4830±0.0191
Dermatology	0.6825±0.0825	0.6960±0.0512	0.7874±0.0599	0.6029±0.0779	**0.7955±0.0687**
German Credit	0.7100±0.0000	0.7100±0.0000	0.7100±0.0000	0.7002±0.0000	**0.7843±0.0000**
CMC	0.4276±0.0020	0.4285±0.0025	0.4290±0.0026	0.4302±0.0087	**0.4391±0.0112**

Table 11 Comparison of NMI Values (means±std) of Different Algorithms on Mixed Data

表 11 混合型数据聚类结果比较:NMI 值(均值±方差)

Data Sets	*K*-Prototypes	*K*-Centers	OCIL	Improved *K*-Prototypes	Proposed Algorithm
Flag	0.2224±0.0201	0.2592±0.0235	0.2138±0.0206	0.2000±0.0291	**0.2651±0.0193**
Dermatology	0.5659±0.0992	0.5513±0.0592	0.7310±0.0537	0.4677±0.0765	**0.7464±0.0658**
German Credit	0.0086±0.0099	0.0106±0.0066	0.0102±0.0187	0.0109±0.0121	**0.0368±0.0098**
CMC	0.0253±0.0061	0.0363±0.0066	0.0311±0.0072	0.0203±0.0104	**0.0433±0.0085**

Table 12 Comparison of ARI Values (means±std) of Different Algorithms on Mixed Data

表 12 混合型数据聚类结果比较:ARI 值(均值±方差)

Data Sets	*K*-Prototypes	*K*-Centers	OCIL	Improved *K*-Prototypes	Proposed Algorithm
Flag	0.0928±0.0215	**0.1214±0.0254**	0.0968±0.0157	0.0804±0.0314	0.0618±0.0159
Dermatology	0.4466±0.1423	0.4319±0.0870	0.6183±0.1068	0.3054±0.1226	**0.6665±0.1215**
German Credit	0.0142±0.0186	**0.0337±0.0184**	0.0100±0.0285	0.0155±0.0288	0.0160±0.0147
CMC	0.0157±0.0062	0.0178±0.0075	0.0127±0.0052	0.0061±0.0089	**0.0251±0.0132**

Table 13 Comparison of CUM Values of (means±std) Algorithms on Mixed Data

表 13 混合型数据聚类结果比较:CUM 值(均值±方差)

Data Sets	*K*-Prototypes	*K*-Centers	OCIL	Improved *K*-Prototypes	Proposed Algorithm
Flag	0.2838±0.0141	0.2659±0.0105	0.2846±0.0088	0.2390±0.0374	**0.2859±0.0188**
Dermatology	0.6301±0.0972	0.6296±0.0533	0.6754±0.0624	0.5062±0.0925	**0.7212±0.0374**
German Credit	0.1292±0.0272	0.0844±0.0169	0.1707±0.0383	0.1081±0.0259	**0.1782±0.0311**
CMC	0.1745±0.0172	0.1423±0.0154	0.1981±0.0111	0.1371±0.0250	**0.2132±0.0219**

3.4 聚类结果统计显著性分析

针对本文提出算法的聚类结果与其他算法聚类结果差异是否显著的问题,本节利用无参数的 Wilcoxon 秩和检验方法进行了统计显著性检验.在不同评价指标下,将本文提出算法实验结果的均值分别与其余算法实验结果的均值进行了统计显著性检验.其中,原假设为在当前指标下本文算法与已有算法的聚类结果没有显著性差异,备择假设为不同聚类算法结果具有显著性差异.

置信区间为 95% 的 Wilcoxon 秩和检验结果 $h(p)$ 如表 14 所示.其中数据分别表示假设检验结果 h 和 p 值.$h=1$ 表示在置信区间为 95%时拒绝原假设接受备择假设,即表示本文提出算法的聚类结果与已有算法聚类结果具有显著差异性;$h=0$ 表示本文提出算法的聚类结果与已有算法聚类结果无显著

Table 14 Results $h(p)$ of Wilcoxon Signed-ranks Test for the Proposed Algorithm versus Other Algorithms

表 14 本文算法与其余算法 Wilcoxon 秩和检验结果 $h(p)$

Indices	Proposed Algorithm vs *K*-Prototypes	Proposed Algorithm vs *K*-Centers	Proposed Algorithm vs OCIL	Proposed Algorithm vs Improved *K*-Prototypes
CA	**1(0.0020)**	**1(0.0137)**	**1(0.0020)**	0(0.1602)
NMI	**1(0.0020)**	**1(0.0020)**	**1(0.0020)**	**1(0.0488)**
ARI	**1(0.0195)**	0(0.0840)	**1(0.0371)**	**1(0.0098)**
CUM	**1(0.0059)**	**1(0.0039)**	**1(0.0059)**	**1(0.0273)**

差异性.由表14可知,本文提出算法在10个数据集获得的聚类结果在87.5%的情况下与其余聚类算法获得的聚类结果具有显著差异性.

4 结束语

本文首先基于分类型数据类中心的模糊表示形式,给出了一种针对混合数据扩展的欧氏距离,能够更加客观准确地度量对象与类之间的差异性.其次,分别基于Renyi熵和互补熵定义了数值型数据和分类型数据的类内熵、类间熵,给出了属性加权机制,进而设计了一个基于信息熵的混合数据属性加权聚类算法.新提出的算法克服了主流聚类算法仅依据数据集总体分布或类内分散度进行属性加权的缺陷.在多个真实数据集上进行了实验验证,与其他混合数据聚类算法相比,本文提出的算法可以获得较高的聚类质量,而且与其他聚类结果具有显著差异性.

参 考 文 献

[1] Han Jiawei, Kamber M, Pei Jian. Data Mining: Concepts and Techniques [M]. 3rd ed. San Francisco: Morgan Kaufmann, 2011

[2] Jain A K. Data clustering: 50 years beyond *K*-means [J]. Pattern Recognition Letters, 2010, 31(8): 651-666

[3] Xu Rui, Wunsch D. Survey of clustering algorithm [J]. IEEE Trans on Neural Networks, 2005, 16(3): 645-678

[4] Sun Jigui, Liu Jie, Zhao Lianyu. Clustering algorithms research [J]. Journal of Software, 2008, 19(1): 48-61 (in Chinese)
(孙吉贵, 刘杰, 赵连宇. 聚类算法研究[J]. 软件学报, 2008, 19(1): 48-61)

[5] Kriegel H P, Kröger P, Zimek A. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering [J]. ACM Trans on Knowledge Discovery from Data, 2009, 3(1): 1-58

[6] Chan E Y, Ching W K, Ng M K, et al. An optimization algorithm for clustering using weighted dissimilarity measures [J]. Pattern Recognition, 2004, 37(5): 943-952

[7] Huang J Z, Ng M K, Rong Hongqiang, et al. Automated variable weighting in *k*-means type clustering [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2005, 27(5): 657-668

[8] Jing Liping, Ng M K, Huang J Z. An entropy weighting *k*-means algorithm for subspace clustering of high-dimensional sparse data [J]. IEEE Trans on Knowledge and Data Engineering, 2007, 19(8): 1026-1041

[9] Liang Jiye, Bai Liang, Cao Fuyuan. *K*-modes clustering algorithm based on a new distance measure [J]. Journal of Computer Research and Development, 2010, 47(10): 1749-1755 (in Chinese)
(梁吉业, 白亮, 曹付元. 基于新的距离度量的*K*-modes聚类算法[J]. 计算机研究与发展, 2010, 47(10): 1749-1755)

[10] Bai Liang, Liang Jiye, Dang Chuangyin, et al. A novel attribute weighting algorithm for clustering high-dimensional categorical data [J]. Pattern Recognition, 2011, 44(12): 2843-2861

[11] Chen Lifei, Guo Gongde. Non-mode clustering of categorical data with attributes weighting [J]. Journal of Software, 2013, 24(11): 2628-2641 (in Chinese)
(陈黎飞, 郭躬德. 属性加权的类属型数据非模聚类[J]. 软件学报, 2013, 24(11): 2628-2641)

[12] Cao Fuyuan, Liang Jiye, Li Deyu, et al. A weighting *k*-modes algorithm for subspace clustering of categorical data [J]. Neurocomputing, 2013, 108: 23-30

[13] Huang Zhexue. Extensions to the *k*-means algorithm for clustering large data sets with categorical values [J]. Data Mining and Knowledge Discovery, 1998, 2 (3): 283-304

[14] Zhao Weidong, Dai Weihui, Tang Chunbin. *K*-centers algorithm for clustering mixed type data [C] //Proc of the 11th Pacific-Asia Conf on Knowledge Discovery and Data Mining. Berlin: Springer, 2007: 1140-1147

[15] Cheung Y, Jia Hong. Categorical-and-numerical-attribute data clustering based on a unified similarity metric without knowing cluster number [J]. Pattern Recognition, 2013, 46 (8): 2228-2238

[16] Ji Jinchao, Bai Tian, Zhou Chunguang, et al. An improved *k*-prototypes clustering algorithm for mixed numeric and categorical data [J]. Neurocomputing, 2013, 120: 590-596

[17] Renyi A. On measures of entropy and information [C] //Proc of the 4th Berkley Symp on Mathematics of Statistics and Probability. Berkeley: University of California Press, 1961: 547-561

[18] Parzen E. On the estimation of a probability density function and the mode [J]. Annals of Mathematical Statistics, 1962, 33(3): 1065-1076

[19] Jenssen R, Eltoft T, Erdogmus D, et al. Some equivalences between kernel methods and information theoretic methods [J]. Journal of VLSI Signal Processing Systems, 2006, 45 (1/2): 49-65

[20] Gokcay E, Principe J C. Information theoretic clustering [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2002, 24(2): 158-171

[21] Liang Jiye, Chin K S, Dang Chuangyin, et al. A new method for measuring uncertainty and fuzziness in rough set theory [J]. International Journal of General Systems, 2002, 31(4): 331-342

[22] Xu Weihua, Zhang Xiaoyan, Zhang Wenxiu. Knowledge granulation, knowledge entropy and knowledge uncertainty measure in ordered information systems [J]. Applied Soft Computing, 2009, 9(4): 1244-1251

[23] Wang Junhong, Liang Jiye, Qian Yuhua. Uncertainty measure of rough sets based on a knowledge granulation of incomplete information systems [J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2008, 16(2): 233-244

[24] Qian Yuhua, Liang Jiye, Pedrycz W, et al. Positive approximation: An accelerator for attribute reduction in rough set theory [J]. Artificial Intelligence, 2010, 174(9/10): 597-618

[25] Liang Jiye, Zhao Xingwang, Li Deyu, et al. Determining the number of clusters using information entropy for mixed data [J]. Pattern Recognition, 2012, 45(6): 2251-2265

[26] Zhao Xingwang, Liang Jiye, Cao Fuyuan. A simple and effective outlier detection algorithm for categorical data [J]. International Journal of Machine Learning and Cybernetics, 2014, 5(3): 469-477

[27] Ienco D, Pensa R G, Meo R. From context to distance: Learning dissimilarity for categorical data clustering [J]. ACM Trans on Knowledge Discovery from Data, 2012, 6(1): 1-25

Zhao Xingwang, born in 1984. PhD candidate. Member of China Computer Federation. His main research interests include data mining and machine learning.

Liang Jiye, born in 1962. Professor and PhD supervisor. Distinguished member of China Computer Federation. His main research interests include granular computing, data mining and machine learning.

2014 年《计算机研究与发展》高被引论文 TOP10

排名	论文信息
1	刘建伟，刘媛，罗雄麟. **玻尔兹曼机研究进展**[J]. 计算机研究与发展，2014，51(1)：1-16 Liu Jianwei, Liu Yuan, and Luo Xionglin. Research and Development on Boltzmann Machine [J]. Journal of Computer Research and Development, 2014, 51(1): 1-16
2	丁兆云，贾焰，周斌. **微博数据挖掘研究综述**[J]. 计算机研究与发展，2014，51(4)：691-706 Ding Zhaoyun, Jia Yan, and Zhou Bin. Survey of Data Mining for Microblogs [J]. Journal of Computer Research and Development, 2014, 51(4): 691-706
3	王静远，李超，熊璋，单志广. **以数据为中心的智慧城市研究综述**[J]. 计算机研究与发展，2014，51(2)：239-259 Wang Jingyuan, Li Chao, Xiong Zhang, and Shan Zhiguang. Survey of Data-Centric Smart City [J]. Journal of Computer Research and Development, 2014, 51(2): 239-259
4	黄冬梅，杜艳玲，贺琪. **混合云存储中海洋大数据迁移算法的研究**[J]. 计算机研究与发展，2014，51(1)：199-205 Huang Dongmei, Du Yanling, and He Qi. Migration Algorithm for Big Marine Data in Hybrid Cloud Storage [J]. Journal of Computer Research and Development, 2014, 51(1): 199-205
5	李晖，孙文海，李凤华，王博洋. **公共云存储服务数据安全及隐私保护技术综述**[J]. 计算机研究与发展，2014，51(7)：1397-1409 Li Hui1, Sun Wenhai, Li Fenghua, and Wang Boyang. Secure and Privacy-Preserving Data Storage Service in Public Cloud [J]. Journal of Computer Research and Development, 2014, 51(7): 1397-1409
6	林闯，董扬威，单志广. **基于 DTN 的空间网络互联服务研究综述**[J]. 计算机研究与发展，2014，51(5)：931-943 Lin Chuang, Dong Yangwei, and Shan Zhiguang. Research on Space Internetworking Service Based on DTN [J]. Journal of Computer Research and Development, 2014, 51(5): 931-943
7	张玉清，王凯，杨欢，方喆君，王志强，曹琛. **Android 安全综述**[J]. 计算机研究与发展，2014，51(7)：1385-1396 Zhang Yuqing, Wang Kai, Yang Huan, Fang Zhejun, Wang Zhiqiang, and Cao Chen. Survey of Android OS Security [J]. Journal of Computer Research and Development, 2014, 51(7): 1385-1396
8	刘雅辉，张铁赢，靳小龙，程学旗. **大数据时代的个人隐私保护**[J]. 计算机研究与发展，2015，52(1)：229-247 Liu Yahui, Zhang Tieying, Jin Xiaolong, and Cheng Xueqi. Personal Privacy Protection in the Era of Big Data [J]. Journal of Computer Research and Development, 2015, 52(1): 229-247
9	蒋卓轩，张岩，李晓明. **基于 MOOC 数据的学习行为分析与预测**[J]. 计算机研究与发展，2015，52(3)：614-628 Jiang Zhuoxuan, Zhang Yan, and Li Xiaoming. Learning Behavior Analysis and Prediction Based on MOOC Data [J]. Journal of Computer Research and Development, 2015, 52(3): 614-628
10	周江，王伟平，孟丹，马灿，古晓艳，蒋杰. **面向大数据分析的分布式文件系统关键技术**[J]. 计算机研究与发展，2014，51(2)：382-394 Zhou Jiang, Wang Weiping, Meng Dan, Ma Can, Gu Xiaoyan, and Jiang Jie. Key Technology in Distributed File System Towards Big Data Analysis [J]. Journal of Computer Research and Development, 2014, 51(2): 382-394

数据来源：中国知网，CSCD；统计日期：2016-02-16