基于模糊质心的混合属性数据模糊加权聚类算法

冀进朝[1,2] 赵晓威[1,2] 何 飞[1,2] 胡英慧[1] 白 天[3] 李在荣[4]

(东北师范大学信息科学与技术学院 长春 130117)[1] (东北师范大学计算生物研究所 长春 130117)[2]

(吉林大学计算机科学与技术学院 长春 130012)[3] (东北师范大学传媒科学学院 长春 130117)[4]

摘 要 在模糊聚类算法中,模糊系数被用来控制簇可能重叠的程度,其负面影响是所有的数据对象会影响所有的簇。为解决该问题,Klawonn和Höppner使用模糊函数替换模糊系数(KH算法),但该方法是针对数值属性数据而设计的。然而,在许多真实的应用中,数据对象通常同时由数值属性和分类属性描述。面向混合属性数据,文中提出了一种新的基于模糊质心的模糊加权聚类算法。首先结合模糊质心和均值来表示混合属性条件下的簇中心,然后使用能够评估不同属性在聚类过程中作用的度量来评估数据对象和簇中心之间的相异度,最后给出算法框架。在3个混合属性数据集上对新算法进行了一系列的测试,实验结果表明新算法的性能优于传统算法。

关键词 模糊聚类,数据挖掘,混合数据,相异性度量

中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.02.019

Fuzzy Weighted Clustering Algorithm with Fuzzy Centroid for Mixed Data

JI Jin-chao[1,2] ZHAO Xiao-wei[1,2] HE Fei[1,2] HU Ying-hui[1] BAI Tian[3] LI Zai-rong[4]

(School of Information Science and Technology, Northeast Normal University, Changchun 130117, China)[1]

(Institute of Computational Biology, Northeast Normal University, Changchun 130117, China)[2]

(College of Computer Science and Technology, Jilin University, Changchun 130012, China)[3]

(School of Media Science, Northeast Normal University, Changchun 130117, China)[4]

Abstract In fuzzy c-means type algorithms, fuzy parameters are used to control the degree of possible overlap, but it also has the negative effects that all data objects tend to influence all clusters. To solve this issue, Klawonn and Höppner proposed a fuzzy function for replacing the fuzzier. However, this method is only designed for numeric data. In many real-world applications, data objects are usually described by both numeric and categorical attributes. In this paper, a novel weighted fuzzy clustering algorithm based on fuzzy centroid (FWFC) was proposed for the data with both numeric and categorical attributes, i.e. mixed data. In this method, the mean is first integrated with fuzzy centroid to represent the cluster centers. Then, a measure which can evaluate the influence of different attributes in the process of clustering is used to evaluate the dissimilarity between data objects and cluster centers. Finally, the algorithm is presented for clustering the data with mixed attributes. The proposed algorithm was tested by a series of experiments on three mixed datasets. Experimental results show that the proposed algorithm outperforms traditional clustering algorithms.

Keywords Fuzzy clustering, Data mining, Mixed data, Dissimilarity measure

1 引言

聚类分析是数据挖掘[1]中的重要技术之一。聚类被广泛应用于许多领域,如信息检索[2]、模式识别[3]、图像处理[4]和生物信息学[5]等。聚类的目标是把数据集中的数据对象划分到不同的簇中,使得同一个簇中的数据对象之间的相似程度高,不同簇中的数据对象之间的相似程度低[6]。在许多真实的应用中,数据通常同时包含数值属性和分类属性[7]。黄哲学最早探索了这类数据的聚类问题[8],之后不断有研究者对这个问题进行进一步的探索[9-13],如Li和Biswas提出了层次聚类算法[9],Foss等人提出了一种半参数方法[10],Pang和Zhao基于先验信息提出了簇数的确定方法[13]。

模糊c均值算法是聚类分析中最具影响力的方法之一,这类算法大多是针对数值属性数据而设计的。Kim等人提出

到稿日期:2017-03-16 返修日期:2017-05-26 本文受国家自然科学基金项目(61502093,61403077),吉林省教育厅科研项目(2016504),吉林省科技发展计划资助项目(20170520058JH),中央高校基本科研业务费专项基金(13QN002)资助。

冀进朝(1982-),男,博士,讲师,CCF会员,主要研究方向为数据挖掘;**赵晓威**(1984-),女,博士,讲师,主要研究方向为生物信息学;**何 飞**(1985-),男,博士,讲师,主要研究方向为深度学习;**胡英慧**(1992-),女,硕士生,主要研究方向为数字化教育;**白 天**(1983-),男,博士,讲师,主要研究方向为生物信息学;**李在荣**(1982-),男,博士,讲师,主要研究方向为数字媒体,E-mail:zairong0431@163.com(通信作者)。

了模糊 k 质心方法来处理分类属性数据[14]，该方法使用模糊质心表示分类属性的簇中心。最近，有研究者针对混合属性数据提出了几种模糊 c 均值算法[15-17]。所有的这些模糊 c 均值算法都涉及到一个参数——模糊系数，该参数控制着簇的重叠程度。模糊系数的引入尽管使得聚类算法能够处理数据对象隶属于不同簇的情况，但也带来了意想不到的效果：无论数据对象离簇有多远，都会影响到所有的簇[18]，从而导致不准确的聚类结果。为解决该问题，Klawonn 和 Höppner 提出了一种新的模糊 c 均值算法（简称 KH 算法），然而该方法仅考虑了数值属性数据。

考虑到混合属性数据的普遍存在性，在 KH 算法框架的基础上，面向混合属性数据提出了一种基于模糊质心的模糊加权聚类算法（FCFW）。FCFW 结合模糊质心和均值表示混合属性数据簇的簇中心，采用能够评估不同属性在聚类中作用的度量来估算数据对象和簇中心之间的相异度。最后，在数据集上测试该方法。

本文的主要贡献如下：

1）面向混合属性数据，提出了一种基于模糊质心的模糊加权聚类算法；

2）结合模糊质心和均值，设计了混合属性条件下的簇中心表示方式；

3）提出一种考虑不同属性在聚类过程中作用的相异性度量，并引入评估机制，使得聚类算法能够实时评估不同属性在聚类过程中的作用。

2 相关符号

$X=\{x_1,x_2,\cdots,x_n\}$ 是数据集，包含 n 个数据对象，$x_i(1\leqslant i\leqslant n)$ 是数据对象，由 m 个属性 $A_1,A_2,\cdots,A_m$ 描述，每个属性 A_j 有一个值域。混合属性数据集的属性域通常分为两类：数值和分类。数值域由实数表示，分类域由无自然序的分类值（如颜色、职业的相关值）的有限集合表示。分类属性域通常表示为 $Dom(A_j)=\{a_j^1,a_j^2,\cdots,a_j^t\}$，其中 t 表示属性 j 中不同属性值的数目。数据对象 X_i 在逻辑上表示为属性值对的合集：

$$[A_1=x_{i1}]\wedge[A_2=x_{i2}]\wedge\cdots\wedge[A_m=x_{im}]$$

其中，$x_{ij}\in Dom(A_j)$，$1\leqslant j\leqslant m$。在不混淆的情况下，X_i 也可表示为向量 $[x_{i1}^r,x_{i2}^r,\cdots,x_{ip}^r,x_{ip+1}^c,x_{ip+2}^c,\cdots,x_{im}^c]$，具有上标 r 的前 p 个元素是数值属性值，剩下的上标为 c 的元素是分类属性值。

3 KH 算法

KH 算法[18]的目标是把规模为 n 的数据集 X 划分到 c 个簇中。该目标的实现需最小化目标成本函数：

$$E=\sum_{l=1}^{c}\sum_{i=1}^{n}g(u_{il})d_{il} \tag{1}$$

其中，u_{il} 是划分矩阵 $U_{n\times c}$ 的元素；$g(u_{il})=\alpha u_{il}^2+(1-\alpha)u_{il}$，$\alpha(0\leqslant\alpha\leqslant1)$ 是模糊系数（通常具有 u_{il}^{β} 的形式，$\beta>1$）的替代函数；d_{il} 是相异度，定义如下：

$$d_{il}=(x_i-v_l)^2 \tag{2}$$

其中，x_i 表示数据对象 i，v_l 表示簇 l 的中心。

KH 算法的流程与模糊 c 类型的算法类似，这里不再赘述。

4 基于模糊质心的模糊加权聚类算法 FWFC

新算法考虑了两方面：1）在决策数据对象的簇归属之前尽可能长时间地保留数据集内在的不确定性；2）考虑不同属性在聚类过程中的作用。为此，本文引入模糊质心[14]的概念和属性影响力的计算策略。

4.1 模糊质心

模糊质心[14]的概念由 Kim 等人提出，用于表示分类属性的簇中心。在模糊质心中，每个分类属性值用一个模糊分类值来描述其在簇中的信息分布。若分类属性 A_j 的属性域为 $Dom(A_j)=\{a_j^1,a_j^2,\cdots,a_j^t\}$，则第 l 个模糊质心 $\widetilde{V}_l$ 被定义为：

$$\widetilde{V}_l=[\widetilde{v}_{l1},\cdots,\widetilde{v}_{lj},\cdots,\widetilde{v}_{lm}] \tag{3}$$

其中：

$$\widetilde{v}_{lj}=\{\{a_{lj}^1,\omega_{lj}^1\},\{a_{lj}^2,\omega_{lj}^2\},\cdots,\{a_{lj}^{\kappa},\omega_{lj}^{\kappa}\},\cdots,\{a_{lj}^t,\omega_{lj}^t\}\} \tag{4}$$

ω_{lj}^{κ} 满足两个条件：

$$0\leqslant\omega_{lj}^{\kappa}\leqslant1,1\leqslant\kappa\leqslant t \tag{5}$$

$$\sum_{\kappa=1}^{t}\omega_{lj}^{\kappa}=1,1\leqslant j\leqslant m \tag{6}$$

每个属性 $\widetilde{v}_{lj}\in\widetilde{V}_l$ 表示模糊集的模糊属性值 $\{a_j^{\kappa},\omega_{lj}^{\kappa}\}$ $(1\leqslant\kappa\leqslant t)$，该值由属性 A_j 的属性值在簇 l 中的信息分布情况决定。

4.2 属性加权策略——距离和影响力

该策略由 Ahmad 和 Dey 提出[19]。假设 X 为具有 m 个属性的分类属性数据集；A_i 为第 i 个分类属性，拥有两个值 a 和 b；A_j 表示第 j 个分类属性；s 为第 j 个属性值的子集，$\bar{s}$ 是 s 的补集。$p_i(s/a)$ 表示数据对象在第 i 个属性值为 a 的条件下，第 j 个属性值属于集合 s 的条件概率；$p_i(\bar{s}/b)$ 表示数据对象在第 i 个属性值为 b 的条件下，第 j 个属性值属于 $\bar{s}$ 的条件概率。

定义 1 设 a 和 b 是第 i 个分类属性的属性值，则 a 和 b 相对第 j 个属性的距离定义为：

$$\delta^{ij}(a,b)=p_i(\vartheta/a)+p_i(\bar{\vartheta}/b)-1.0 \tag{7}$$

其中，$\vartheta=\arg\max_{s}\{p_i(s/a)+p_i(\bar{s}/b)\}$。

属性值 a 和 b 相对其他属性的距离可用类似方法计算得到。在混合属性条件下，a 和 b 相对数值属性的距离则通过离散化数值属性得到，a 和 b 之间的绝对距离则是所有这些值的均值。

定义 2 设 X 是具有 m 个属性的混合属性数据集，数值属性已离散化处理，则第 i 个分类属性的不同属性值 a 和 b 之间的距离为：

$$\delta(a,b)=\frac{\sum_{j=1,i\neq j}^{m}\delta^{ij}(a,b)}{m-1} \tag{8}$$

其中，$\delta(a,b)$ 具有如下属性：

1）$0\leqslant\delta(a,b)\leqslant1$；

2）$\delta(a,b)=\delta(b,a)$；

3）$\delta(a,a)=0$。

属性的影响力表示属性在聚类过程中的作用。研究表明,在一个属性中,如果共同出现的属性值在其他属性对应的组中表现出更好的分割性,则该属性具有更大的影响力[19]。这意味着,如果属性的值对之间的整体距离值更大,则该属性在聚类过程中的影响力更大。由于已考虑不同属性值的分组,因此分类属性的属性值之间的距离暗含其在聚类过程中的影响力。

数值属性 A_i 的属性值 a 和 b 之间的距离为 $d(a,b)=(s_i(a-b))^2$,s_i 表示 A_i 的影响力。。

在离散化数值属性时,需指定离散区间数,每个区间对应一个分类属性值 $u[1],u[2],\cdots,u[T]$。离散化之后,采用式(8)计算属性值对之间的距离 $\delta(u[r],u[e])$。数值属性的影响力是其所有属性值对距离的平均值。

定义 3　数值属性 A_i 的影响力 s_i 的计算公式为:

$$s_i=\frac{2\sum_{k=1}^{T}\sum_{j>k}^{T}\delta(u[k],u[j])}{T(T-1)} \tag{9}$$

其中,T 是离散区间数。

4.3　新算法 FWFC

在 KH 算法框架下,结合 4.1 节－4.2 节的思想,面向混合属性数据提出了新的聚类算法。如前文所述,混合属性条件下的数据对象 X_i 表示为向量 $[x_{i1}^r,x_{i2}^r,\cdots,x_{ip}^r,x_{ip+1}^c,x_{ip+2}^c,\cdots,x_{im}^c]$,带有上标 r 的前 p 个元素是数值属性值,剩余的带有上标 c 的元素是分类属性值。混合属性条件下的簇中心分为两部分:前 p 个质心使用均值;剩余的使用模糊质心。聚类混合属性数据的目标成本函数为:

$$E(U,Q)=\sum_{j=1}^{k}\sum_{i=1}^{n}g(u_{ij})d(x_i,Q_j) \tag{10}$$

其中,n 和 k 分别表示数据集的规模和簇数;$U=(u_{ij})_{n\times k}$ 是模糊划分矩阵,满足条件 $0\leqslant u_{ij}\leqslant 1$ 和 $\sum_{j=1}^{k}u_{ij}=1$;$Q_j=[q_{j1}^r,q_{j2}^r,\cdots,q_{jp}^r,\tilde{v}_{jp+1}^c,\cdots,\tilde{v}_{jm}^c]$ 是簇 j 的簇中心;$g(u_{ij})=\alpha u_{ij}^2+(1-\alpha)u_{ij}$ $(0\leqslant\alpha\leqslant 1)$ 是模糊系数的替代函数;$d(x_i,Q_j)$ 是数据对象 x_i 和簇中心 Q_j 之间的相异性度量,定义如下:

$$d(x_i,Q_j)=\sum_{l=1}^{p}(s_l(x_{il}^r-q_{jl}^r))^2+\sum_{l=p+1}^{m}\varphi(x_{il}^c,\tilde{v}_{jl}^c)^2 \tag{11}$$

至此,可将目标成本函数重写为:

$$E(U,Q)=\sum_{j=1}^{k}\sum_{i=1}^{n}g(u_{ij})\Big(\sum_{l=1}^{p}(s_l(x_{il}^r-q_{jl}^r))^2+\sum_{l=p+1}^{m}\varphi(x_{il}^c,\tilde{v}_{jl}^c)^2\Big) \tag{12}$$

其中,

$$u_{ij}=\frac{2\alpha}{1+\alpha}\Bigg(\frac{1+(\hat{c}-1)\left(\frac{1-\alpha}{1+\alpha}\right)}{\sum_{z=1,d(x_i,Q_z)\neq 0}^{k}\left(\frac{d(x_i,Q_j)}{d(x_i,Q_z)}\right)}-\left(\frac{1-\alpha}{1+\alpha}\right)\Bigg) \tag{13}$$

其中,$\hat{c}$ 是数据对象 x_i 具有的非零成员隶属度的簇数。寻找值为零的 u_{ij} 的方法如下:对固定的 i,把相异性度量值 $d_{ij}(x_i,Q_j)$ 以降序排序,不失一般性,假设 $d_{i1}(x_i,Q_1)\geqslant d_{i2}(x_i,Q_2)\geqslant\cdots\geqslant d_{ik}(x_i,Q_k)$,如果没有成员隶属度为零的元素,为最小化目标成本函数,则具有更大相异度的 u_{ij} 必须为零。为此,找到使式(13)产生正值的最小索引值 j_o。当 $j<j_o$ 时,成员隶属度 $u_{ij}=0$;当 $j\geqslant j_o$ 时,按照式(13)计算 u_{ij},其中 $\hat{c}=k-j_o+1$。q_{jl}^r 是数值属性的质心,其形式为:

$$q_{jl}^r=\frac{\sum_{i=1}^{n}g(u_{ij})\cdot x_{il}^r}{\sum_{i=1}^{n}g(u_{ij})} \tag{14}$$

$\tilde{v}_{jl}^c$ 是分类属性的质心,其形式为:

$$\tilde{v}_{jl}^c=\{\{a_{jl}^1,\omega_{jl}^1\},\{a_{jl}^2,\omega_{jl}^2\},\cdots,\{a_{jl}^\kappa,\omega_{jl}^\kappa\},\cdots,\{a_{jl}^t,\omega_{jl}^t\}\} \tag{15}$$

式(12)中,$\varphi(x_{il}^c,\tilde{v}_{jl}^c)$ 的定义如下:

$$\varphi(x_{il}^c,\tilde{v}_{jl}^c)=\sum_{\kappa=1}^{t}\tau(x_{il}^c,a_{jl}^\kappa)\times\delta(x_{il}^c,a_{jl}^\kappa) \tag{16}$$

其中,$\delta(x_{il}^c,a_{jl}^\kappa)$ 由式(8)计算,$\tau(x_{il}^c,a_{jl}^\kappa)$ 的形式为:

$$\tau(x_{il}^c,a_{jl}^\kappa)=\begin{cases}0, & x_{il}^c=a_{jl}^\kappa\\ \omega_{jl}^\kappa, & x_{il}^c\neq a_{jl}^\kappa\end{cases} \tag{17}$$

式(15)中:

$$\omega_{jl}^\kappa=\sum_{i=1}^{n}\gamma(a_{jl}^\kappa) \tag{18}$$

其中,

$$\gamma(a_{jl}^\kappa)=\begin{cases}\dfrac{g(u_{ij})}{\sum_{i=1}^{n}g(u_{ij})}, & a_{jl}^\kappa=x_{il}^c\\ 0, & a_{jl}^\kappa\neq x_{il}^c\end{cases} \tag{19}$$

至此,相关变量的计算公式已全部列出,下面给出算法的具体步骤。

步骤 1　给定最大迭代次数 $maxIte$、簇数 k、参数 α、离散区间数 T 和阈值 ε 的值,计算分类属性值对之间的距离和每个数值属性的影响;从数据集中随机选择 k 个不同的数据对象,把它们转变为初始的簇中心 $Q=(Q_1,Q_2,\cdots,Q_k)$。

步骤 2　按照式(13)计算模糊划 U。

步骤 3　计算簇中心 Q;对簇中心的数值属性部分使用式(14)进行计算,对簇中心的分类属性部分使用式(15)、式(18)和式(19)进行计算。

步骤 4　如果相邻的目标成本函数值之间的差小于阈值 ε 或 $maxIte$ 等于 0,则算法终止;否则,设置 $maxIte\leftarrow maxIte-1$,转到步骤 2。

步骤 1 中的转换规则为:若第 j 个属性是数值属性,则质心 $q_{lj}\in Q_l$ 是数值属性的值。若第 j 个属性是分类属性,则质心 $\tilde{v}_{lj}\in Q_l$ 的值的分配规则为:若 $x_{lj}=a_{lj}^\kappa$,则 ω_{lj}^κ 的值为 1;若 $x_{lj}\neq a_{lj}^\kappa$,则 ω_{lj}^κ 的值为 0,其中 $1\leqslant\kappa\leqslant t$。

4.4　算法的复杂性分析

本节考虑算法 FWFC 的时间复杂性和空间复杂性。算法的时间复杂性由算法执行过程中簇中心和划分矩阵的迭代更新、属性值对之间距离的计算成本构成,分别为 $O(k(p+Nm-Np)n)$,$O(kmn)$ 和 $O(m^2n+m^2T^3)$。其中,k 表示簇数,p 表示数值属性的个数,m 表示所有属性的数目,$N=\max(t)$ 表示 $(m-p)$ 分类属性中单个属性包含的不同属性值的最大数目,n 表示数据集的规模,T 表示分类属性包含的不同属性值的平均数目。因此,算法的整体时间复杂性为 $O(m^2n+m^2T^3+k(m+p+Nm-Np)ns)$,$s$ 是算法收敛所需的迭代次数。该复杂性与数据对象数呈线性关系。在空间复杂性方面,算法需要 $O(k(p+mN-pN)+mn)$ 的空间存储簇中心和数据集 X 以及 $O(kn)$ 的空间存储划分矩阵。因此,整体的空间复

杂性为 $O(k(p+n+mN-pN)+mn)$。

5 实验与结果

为在相同的尺度上比较不同的数值属性，采用 Witten 和 Frank 所提的规格化框架[20]，其形式为：

$$n_{ij}=\frac{x_{ij}-v_{j\min}}{v_{j\max}-v_{j\min}} \tag{20}$$

其中，$v_{j\min}$ 和 $v_{j\max}$ 分别是第 j 个数值属性的最小值和最大值；x_{ij} 和 n_{ij} 分别是原始值和规格化后的值。为简化计算，在评估分类属性值对之间的距离和数值属性的影响时，采用等宽离散化。离散间隔数 T 的值为数据集中分类属性值拥有的不同属性值的平均数目。为评估算法的有效性，使用新算法在 3 个混合属性数据集上进行实验。这 3 个数据集分别为 Zoo，Heart Disease 和 Credit Approval，它们来自 UCI 机器学习库[1]。在聚类分析中，聚类准确率 r[21] 是评估聚类结果质量的常用度量之一，也常用于验证聚类算法的性能。本文采用该度量来评估所得的聚类结果，聚类准确性度量 r 的定义为：

$$r=\frac{\sum_{i=1}^{k}a_i}{n} \tag{21}$$

其中，a_i 是出现在第 i 个簇和相应的真实类中的数据对象的数目，n 是数据集的规模。按照该准确率度量的定义，r 的值越大，聚类结果越好。鉴于初始簇中心是随机产生的，将新算法运行 100 次并取平均聚类准确率。在所有的实验中，设置 $maxIte=100$，$\varepsilon=1.1102\times10^{-21}$。

Zoo 数据集是人工混合数据集，有 101 个数据对象，每个数据对象由 16 个分类属性和 1 个数值属性描述。最后 1 个分类属性是类属性，具有 7 个值。由于每个分类属性拥有的不同属性值的数目平均为 2，因此式(9)中 $T=2$。评估新算法 FWFC 和 3 个著名的面向混合属性数据的聚类算法 SBAC[9]，KL-FCM-GM[17] 和 K-prototypes[6] 的聚类准确率，结果如表 1 所列。K-prototypes，SBAC 和 KL-FCM-GM 的聚类准确率分别为 0.806，0.426 和 0.864。而新算法 FWFC 在 $\alpha=0.9$ 时得到了更高的聚类准确率 0.901。在这种情况下，FWFC 算法的聚类准确率比 K-prototypes，SBAC 和 KL-FCM-GM 算法分别高出了 9.5%，47.5% 和 3.7%。

表 1 不同算法在 Zoo 数据集上的准确率

Table 1 Accuracy of different algorithms on Zoo dataset

Algorithm	Clustering accuracy(r)
K-prototypes	0.806
SBAC	0.426
KL-FCM-GM	0.864 ($\beta=1.3$)
FWFC	0.901 ($\alpha=0.9$)

真实数据集 Heart Disease 包含 303 个病人实例，每个实例由 6 个数值属性和 9 个分类属性描述。最后两个分类属性是类属性，当使用第 15 个属性作为类属性时，数据集有 5 个类(s_1，s_2，s_3，s_4 和 H)，由 14 个属性描述；当使用第 14 个属性作为类属性时，数据集分为两个类(buff，sick)，数据对象由 13 个属性描述。对于第一种情况，由于分类属性拥有不同属性值的平均数目为 3，因此在式(9)中 $T=3$。表 2 列出了 K-prototypes，SBAC，KL-FCM-GM 和 FWFC 算法的聚类准确率。从表 2 可以看出，在这种情况下，FWFC 算法的聚类准确率比 K-prototypes，SBAC 和 KL-FCM-GM 算法分别高出了 16.4%，16.1% 和 5.3%。

表 2 不同算法在 Heart Disease 数据集(5 个类，14 个属性)上的准确率

Table 2 Accuracy of different algorithms on Heart Disease dataset (5 classes，14 attributes)

Algorithm	Clustering accuracy (r)
K-prototypes	0.542
SBAC	0.545
KL-FCM-GM	0.653 ($\beta=1.3$)
FWFC	0.706 ($\alpha=0.6$)

对于第二种情况，Heart Disease 数据集中的数据对象由 13 个属性描述，使用第 14 个属性作为数据集的类属性。由于分类属性拥有不同属性值的平均数目为 3，因此在式(9)中 $T=3$。表 3 列出了 K-prototypes，SBAC，KL-FCM-GM 和 FWFC 算法的聚类准确率。表 3 中的聚类结果表明，FWFC 算法的聚类准确率比 K-prototypes，SBAC 和 KL-FCM-GM 算法分别高出了 25.1%，7.6% 和 7.0%。

表 3 不同算法在 Heart Disease 数据集(2 个类，13 个属性)上的准确率

Table 3 Accuracy of different algorithms on Heart Disease dataset (2 classes，13 attributes)

Algorithm	Clustering accuracy (r)
K-prototypes	0.577
SBAC	0.752
KL-FCM-GM	0.758 ($\beta=1.7$)
FWFC	0.828 ($\alpha=0.7$)

Credit Approval 数据集也是真实数据集，由 10 个分类属性和 6 个数值属性描述，最后一个属性为类属性。该数据集包含来自信用卡公司的 690 个数据对象，分为两个类：正类和负类。由于该数据集中分类属性拥有不同属性值的平均数目为 5，因此在式(9)中 $T=5$。表 4 列出了 K-prototypes，SBAC，KL-FCM-GM 和 FWFC 算法的聚类准确率。从表 4 可以看出，新算法 FWFC 的聚类准确率比 K-prototypes，SBAC 和 KL-FCM-GM 算法分别高出了 21.9%，22.6% 和 19.7%。

表 4 不同算法在 Credit Approval 数据集上的准确率

Table 4 Accuracy of different algorithm on credit approval dataset

Algorithm	Clustering accuracy (r)
K-prototypes	0.562
SBAC	0.555
KL-FCM-GM	0.584 ($\beta=2.3$)
FWFC	0.781 ($\alpha=0.4$)

表 1－表 4 中的实验结果表明，在聚类准确率方面，所提算法 FWFC 取得了比其他算法更好的聚类结果，原因如下：

1)通过引入 KH 算法框架，FWFC 避免了面向混合属性数据的大多数模糊聚类算法所面临的问题，即不论数据对象都离簇多远，所有的数据对象都趋向于影响所有的簇；

[1] http://archive.ics.uci.edu/ml/datasets.html

2)通过使用模糊质心和均值表示混合属性条件下的簇中心,FWFC 可以在最终决策簇归属前更长时间地保留数据集的不确定性,因此不易陷入局部最优;

3)算法考虑了不同属性在聚类过程中的影响力。

表 5 列出了 FWFC 算法在不同数据集上的运行时间(取 100 次运行时间的平均值)。实验平台是联想 ThinkCentre M8600t-N000 台式电脑,处理器为英特尔 Core i7-6700 @ 3.40GHz 四核,内存为 16GB。

表 5　FWFC 算法在不同数据集上的运行时间

Table 5　Running time of FWFC on different datasets

数据集	运行时间 /s
Zoo	1.422
Heart Disease(5 个簇)	4.303
Heart Disease(2 个簇)	0.930
Credit Approval	17.356

由表 5 可知,FWFC 在各个数据集上的运行时间与算法的时间复杂性分析结果是一致的。

结束语　混合属性数据在实际应用中很常见。本文面向该类型数据,提出了一种新的模糊加权聚类算法 FWFC。该方法首先结合模糊质心和均值来表示簇中心,接着使用能够评估不同属性在聚类过程中的影响力的度量来评估数据对象和簇中心之间的相异度。这些因素使得 FWFC 算法不仅具有利用模糊集区分位于簇边界区域的模糊数据对象的能力,而且还避免了针对混合属性数据的大多数聚类算法中不管数据对象离簇的距离有多远,数据对象都趋向于影响所有簇的不足。

然而,与大多数(模糊)c 均值类型聚类算法一样,本文的方法也需要确定簇数。实验中,簇数是根据数据集中类的数目确定的。在实际应用中,类信息通常是未知的。在这种情况下,如何确定合适的簇数是一个重要的问题,这将是下一步研究工作的主题。

参 考 文 献

[1] CELEBI M E,KINGRAVI H A,VELA P A. A comparative study of efficient initialization methods for the k-means clustering algorithm[J]. Expert Systems with Applications,2013,40(1):200-210.

[2] BORDOGNA G,PASI G. A quality driven hierarchical data divisive soft clustering for information retrieval[J]. Knowledge-Based Systems,2012,26:9-19.

[3] LI T,CORCHADO J M,SUN S,et al. Clustering for filtering: Multi-object detection and estimation using multiple/massive sensors[J]. Information Sciences,2017(388-389):172-190.

[4] VERMA H,AGRAWAL R K,SHARAN A. An improved intuitionistic fuzzy c-means clustering algorithm incorporating local information for brain image segmentation[J]. Applied Soft Computing,2016,46:543-557.

[5] SAEED F,SALIM N,ABDO A. Information theory and voting based consensus clustering for combining multiple clusterings of chemical structures[J]. Molecular Informatics,2013,32(7):591-598.

[6] HUANG Z. 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.

[7] ZHANG X,MEI C,CHEN D,et al. Feature selection in mixed data:A method using a novel fuzzy rough set-based information entropy[J]. Pattern Recognition,2016,56(1):1-15.

[8] HUANG Z. Clustering large data sets with mixed numeric and categorical values[C]// Proceedings of the first Pacific-Asia Conference on Knowledge Discovery and Data Mining. 1997:21-34.

[9] LI C,BISWAS G. Unsupervised learning with mixed numeric and nominal data[J]. IEEE Transactions on Knowledge and Data Engineering,2002,14(4):673-690.

[10] FOSS A,MARKATOU M,RAY B,et al. A semiparametric method for clustering mixed data[J]. Machine Learning,2016,105(3):419-458.

[11] BAI L,LIANG J Y,DANG C,et al. A cluster centers initialization method for clustering categorical data[J]. Expert Systems with Applications,2012,39(9):8022-8029.

[12] PANG T J,LIANG J Y. Clustering Ensemble Algorithm for Large-scale Mixed Data Based on Sampling[J]. Computer Science,2016,43(9):209-212. (in Chinese)
庞天杰,梁吉业. 一种基于抽样的大规范混合数据聚类集成算法[J]. 计算机科学,2016,43(9):209-212.

[13] PANG T J,ZHAO X W. Algorithm to Determine Number of Clusters for Mixed Data Based on Prior Information[J]. Computer Science,2016,43(2):101-104. (in Chinese)
庞天杰,赵兴旺. 一种基于先验信息的混合数据聚类个数确定算法[J]. 计算机科学 ,2016,43(2):101-104.

[14] KIM D W,LEE K H,LEE D. Fuzzy clustering of categorical data using fuzzy centroids[J]. Pattern Recognition Letters,2004,25(11):1263-1271.

[15] AHMAD A,DEY L. Algorithm for fuzzy clustering of mixed data with numeric and categorical attributes[M]// Distributed Computing and Internet Technology. Berlin: Springer Berlin Heidelberg,2005:561-572.

[16] LEE M,PEDRYCZ W. The fuzzy c-means algorithm with fuzzy p-mode prototypes for clustering objects having mixed features[J]. Fuzzy Sets and Systems,2009,160(24):3590-3600.

[17] CHATZIS S P. A fuzzy c-means-type algorithm for clustering of data with mixed numeric and categorical attributes employing a probabilistic dissimilarity functional[J]. Expert Systems with Applications,2011,38(7):8684-8689.

[18] KLAWONN F,H PPNER F. What Is Fuzzy about Fuzzy Clustering? Understanding and Improving the Concept of the Fuzzifier[M]// Advances in Intelligent Data Analysis V. Berlin: Springer Berlin Heidelberg,2003:254-264.

[19] AHMAD A,DEY L. A k-mean clustering algorithm for mixed numeric and categorical data[J]. Data & Knowledge Engineering,2007,63(2):503-527.

[20] WITTEN I H,FRANK E. Data Mining Practical Machine Learning Tools and Techniques with Java Implementation[M]. San Fransisco:Morgon Kaufmann Publishers,1999.

[21] HUANG Z X,NG M K. A fuzzy k-modes algorithm for clustering categorical data[J]. IEEE Transactions on Fuzzy Systems,1999,7(4):446-452.