FD-CABOSFV区间变量高维数据聚类

武　森*　张文丽　黄慧敏　叶俞飞
(北京科技大学经济管理学院 北京 100083)

摘　要　针对属性取值为区间变量的高维数据聚类问题，提出基于模糊离散化的CABOSFV改进算法——FD-CABOSFV。针对属性组合利用模糊C均值聚类的思想进行属性取值的离散化，并通过λ水平截取的方式确定各对象对离散化属性的归属，将其转换为二态变量后利用CABOSFV算法进行聚类。采用三组UCI基准数据集将FD-CABOSFV与著名的K-means聚类算法进行比较，实验结果表明FD-CABOSFV更有效。

关键词　CABOSFV算法，属性组合，模糊离散化，区间变量

中图分类号　TP311

1　引言

聚类作为数据挖掘的主要方法之一[1]，越来越引起人们的关注。随着信息技术的迅猛发展，聚类不仅面临着数据量越来越大的问题，还面临着数据的高维度问题[2]。对高维数据的聚类分析，已成为数据挖掘领域的研究重点之一。

目前对高维数据的聚类主要有两种思路：一种是采用主成分分析[3]、流行学习[4,5]等降维方法，将高维数据转化为较低维数据，然后用传统的聚类算法在较低维的空间中进行聚类；另一种是专门针对高维数据的聚类算法[6,7]，不进行数据降维，直接在高维空间中完成聚类过程。其中，CABOSFV[8,9]是一种专门针对高维数据的高效聚类算法。它的思想主要体现在针对高维数据的稀疏特征提出一种新的差异度计算方法，称为集合的“稀疏特征差异度”(Sparse Feature Dissimilarity，SFD)，反映一个集合内所有对象间的总体差异程度，并通过定义的“稀疏特征向量”(Sparse Feature Vector，SFV)实现对数据的有效压缩，使得数据处理量明显减少，且只需一次数据扫描就可以生成聚类结果，是处理高维稀疏数据聚类问题的一种高效的算法。但是该算法的局限性在于它是针对二态变量提出的，不能解决现实中常见的区间变量聚类问题。

为了将CABOSFV算法的思想拓展到求解高维区间变量聚类问题，提出模糊离散化CABOSFV改进算法——FD-CABOSFV(improved CABOSFV based on Fuzzy Discretization for high dimensional data clustering of interval-scaled variables)，引入模糊理论中的模糊C均值聚类[10,11](Fuzzy C-Means clustering，FCM)将区间变量分组离散化为类别变量，并进一步转化为二态变量，然后应用CABOSFV算法对高维数据进行聚类分析。

* 基金项目：国家自然科学基金(70771007)；中央高校基本科研业务费专项资金(FRF-TP-10-006B)。
通信作者：武森，北京科技大学经济管理学院，博士、教授、博士生导师，E-mail：wusen@manage.ustb.edu.cn。

2 FD-CABOSFV 算法概念基础

FD-CABOSFV 算法结合属性组合、模糊离散化和隶属度的概念，来解决区间变量高维数据聚类问题。下面介绍这些概念。

2.1 属性组合

属性的类别特征是否明显决定了对属性进行离散化处理的结果是否理想。对于取值分布具有明显类别特征的单属性可以进行独立离散化；对于数据集中的多个属性之间存在一定关系的属性，如果只进行独立离散化，不考虑连续属性离散化过程中的相互影响，容易产生不合理或多余的离散化划分点。因此需要对其进行整体离散化[12]。由于单属性离散化和整体离散化在区间变量高维数据聚类中不一定能取得很好的效果，因此提出属性组合的概念对高维数据进行分组离散化处理。其基本思想是在进行离散化的过程中将属性分成属性小组，将属性小组作为整体进行离散化。属性组合的定义如下：

定义1(属性组合)：假设数据集 X 有 n 个对象 m 个属性，m 个属性分成 r 个属性小组，则属性小组表示为 $H_j, j \in \{1,2,\cdots,r\}$。其中每个属性小组中包含的属性个数为 $p_i, i \in \{1,2,\cdots,r\}$，且 $\sum_{i=1}^{r} p_i = m$。

现采用直接划分的方法进行属性组合。假设每个属性小组包含 g 维，则一个 m 维的数据集可分为 r 个小组，使用向上取整公式如下：

$$r = \left\lceil \frac{m}{g} \right\rceil \tag{1}$$

其中，第 $1,2,\cdots,r-1$ 个小组都包含 g 维，第 r 个小组的维数为 $m-(r-1)g \in \{1,2,\cdots,g\}$ 对每个属性小组分别进行模糊离散化处理，然后合并每个属性小组得到的离散化结果，得到整个数据集上离散化的最终结果。

2.2 模糊离散化

区间变量是在属性的连续值域范围内任意取值的变量，是一种连续变量，一般取值为线性度量值，例如身高、长度、宽度、重量等都是区间变量。

FD-CABOSFV 算法为了借鉴 CABOSFV 算法提出的针对二态变量的稀疏差异度来度量区间变量对象的集合差异度，需要将属性离散化为不同的类别变量，并最终用特定的二态变量{0,1}来表示。例如描述花的一个区间变量型的属性是“花瓣长度”，将其泛化为花瓣(长)、花瓣(中)、花瓣(短)，三者中只能有一个取值为1，其他取值均为0。

在将区间变量转换成类别变量并用二态值来表示的过程中，FD-CABOSFV 算法未采用非此即彼的思想得出属性取值是1还是0。而是借助模糊 C 均值聚类的思想[10,11]得到模糊离散化后各属性在[0,1]区间的取值。模糊 C 均值聚类的本质是带约束的非线性规划问题。其核心思想是：确定类的个数 c 及每组的聚类中心，使得各对象合并到相应的类以后目标函数最小。

将其用于求解各个对象对离散化属性的隶属程度。具体描述为：对象集 $X=\{x_1,x_2,\cdots,x_n\}$ 被分为 c 类($c>1$ 且 c 为正整数)，模糊 C 均值聚类的结果是各对象对各离散化属性的隶属程度，用模糊矩阵 $\boldsymbol{W}=(w_{ij})(0\leqslant w_{ij}\leqslant 1)$ 表示，其中 w_{ij} 表示第 $i(1\leqslant i\leqslant n)$ 个对象对第 $j(1\leqslant j\leqslant c)$ 个离散化属性的隶属程度。$\boldsymbol{W}$ 具有如下性质：

$$w_{ij} \in [0,1] \tag{2}$$

$$\sum_{j=1}^{c} w_{ij} = 1 \tag{3}$$

$$0 < \sum_{i=1}^{n} w_{ij} < n \tag{4}$$

为了计算各对象对各离散化属性的隶属度，定义模糊 C 均值聚类的目标函数为：

$$J_m(W,Z) = \sum_{i=1}^{n}\sum_{j=1}^{c} w_{ij}^m d_{ij}^2(x_i,z_j), \quad Z = (z_1,z_2,\cdots,z_c) \tag{5}$$

其中，m 为模糊系数，w_{ij} 表示模糊隶属度，z_j 为第 j 个聚类中心。聚类中心表示的是每个类的平均特征，可以认为是这个类的代表点。每个类对应一个离散化属性。

$d_{ij}^2(x_i,z_j) = \| x_i - z_j \|$ 是对象 x_i 到聚类中心 z_j 的欧式距离。模糊 C 均值聚类即是在式(2)到式(4)的约束条件下求目标函数式(5)的最小值，通过对目标函数的迭代优化实现对象集的聚类。

算法的输入是属性组合后的各个属性小组对应的对象数据，输出是聚类中心 Z 以及 $n * c$ 的模糊划分矩阵。根据模糊理论中的最大隶属原则能够确定每个对象归为哪个离散化属性。

2.3 隶属度下限 λ

模糊离散化得到的模糊划分矩阵表示每个对象对聚类中心的隶属程度，该隶属程度用一个在[0,1]区间的数值表示。但是要确定每个对象到底归属于哪一个类别，还需要对隶属度进行水平截取。采用模糊理论中的隶属度下限 λ 对模糊矩阵中的各值进行水平截取。对应的转换函数如下所示：

$$\lambda_{w_{ij}}(x) = \begin{cases} 1, & w_{ij} \geqslant \lambda \\ 0, & w_{ij} < \lambda \end{cases} \tag{6}$$

$\boldsymbol{W}_\lambda = (\lambda_{wij})$ 称为 $\boldsymbol{W}$ 的 λ 截矩阵，是一个布尔矩阵，λ 为置信水平。对象在某类别下取值为 1 时，表示该对象属于此类别；反之，当取值为 0 时，表示该对象不属于此类别。λ 取值不同，会影响对象对类别的归属情况，即代表了对边界值的不同处理方式。λ 值较大时，对象进入某类别的门槛较高；当 λ 值较小时，对象进入某类别的门槛相对较低。

综上所述，对象中的数据进行模糊离散化并用隶属度下限 λ 截取以后，区间变量已转换为类别变量，并表示为二态变量{0,1}，可以进一步利用 CABOSFV 算法解决聚类问题。

3 FD-CABOSFV 算法过程

3.1 算法步骤

假设数据集 X 有 n 个对象 m 个属性，描述第 $i(i \in \{1,2,\cdots,n\})$ 个对象的 m 个属性取值是区间变量值 $x_{i1}, x_{i2}, \cdots, x_{im}$，FD-CABOSFV 算法的步骤如下：

步骤 1：数据标准化。描述同一个对象的各个属性往往有不同的计量单位，为了避免计量单位对差异度计算的影响，需要先对变量进行标准化。采用统计学中的标准化变换方法对数据进行处理，变换公式如下：

$$x_{ij}^* = \frac{x_{ij} - average[x_j]}{sqrt(\mathrm{var}[x_j])} \tag{7}$$

其中，x_{ij} 表示第 i 个对象的第 j 维数据，$average[x_j]$ 表示第 j 维数据的平均值，$sqrt(\mathrm{var}[x_j])$ 表示第 j 维数据的标准差，x_{ij}^* 是标准化后第 i 个对象在第 j 维的取值。这里第 i 个对象的 m 个属性取值仍然用 $x_{i1}, x_{i2}, \cdots, x_{im}$ 来描述。

步骤2：属性组合。将 m 个属性分成 r 个属性小组，根据公式(1)计算出每个属性小组包含的维数。数据集转换为 r 组属性 n 个对象，描述第 i 个对象的 r 组属性为 $y_{i1}, y_{i2}, \cdots, y_{ir}$。数据集表示为 $Y_{n\times r}$。

步骤3：模糊离散化。针对数据集 $Y_{n\times r}$ 中的 r 个属性小组运用 FCM 聚类逐个进行模糊离散化，这里借助 matlab 工具中集成的 FCM 算法实现。具体过程如下：

(1) 读取数据。将第 $k, k\in\{1,2,\cdots,r\}$ 个属性小组对应的对象数据分别读入 matlab 中。

(2) 针对每个属性小组，进行模糊离散化，将每组属性取值划分为 c 类。编写 matlab 代码，利用 matlab 中自带的 fcm 公式[13]，计算模糊隶属度矩阵 $W=[w_{it}]$，其中 $t\in\{1,2,\cdots,c\}$。

(3) matlab 实现迭代过程中满足式(5)中目标函数最小时，输出模糊矩阵 W 和聚类中心 Z。

汇总所有单一属性小组模糊离散化以后的结果，得到的模糊矩阵 W 是 s 维 n 个对象，其中 $s=r\cdot c$。描述第 i 个对象的 s 维属性的隶属度表示为 $w_{i1}, w_{i2}, \cdots, w_{is}$。

步骤4：λ 水平截取。对模糊矩阵 W 的各个隶属度利用式(6)进行 λ 水平截取，得到布尔型矩阵 W_λ。

步骤5：利用 CABOSFV 算法进行聚类。对 n 个对象的数据集 W_λ，其中描述第 i 个对象的 s 维属性已表示为二态变量 $w_{i1}, w_{i2}, \cdots, w_{is}$，设类内对象的集合差异度上限记为 b，利用 CABOSFV 算法[8]进行聚类。

若聚类结果不能满足需求，可以回到对应的步骤中调整参数 r 值、c 值、λ 值以及 b 值，重复上面的计算步骤，直到满足需求为止。

总结上述步骤，FD-CABOSFV 算法的聚类过程模型如下图所示。

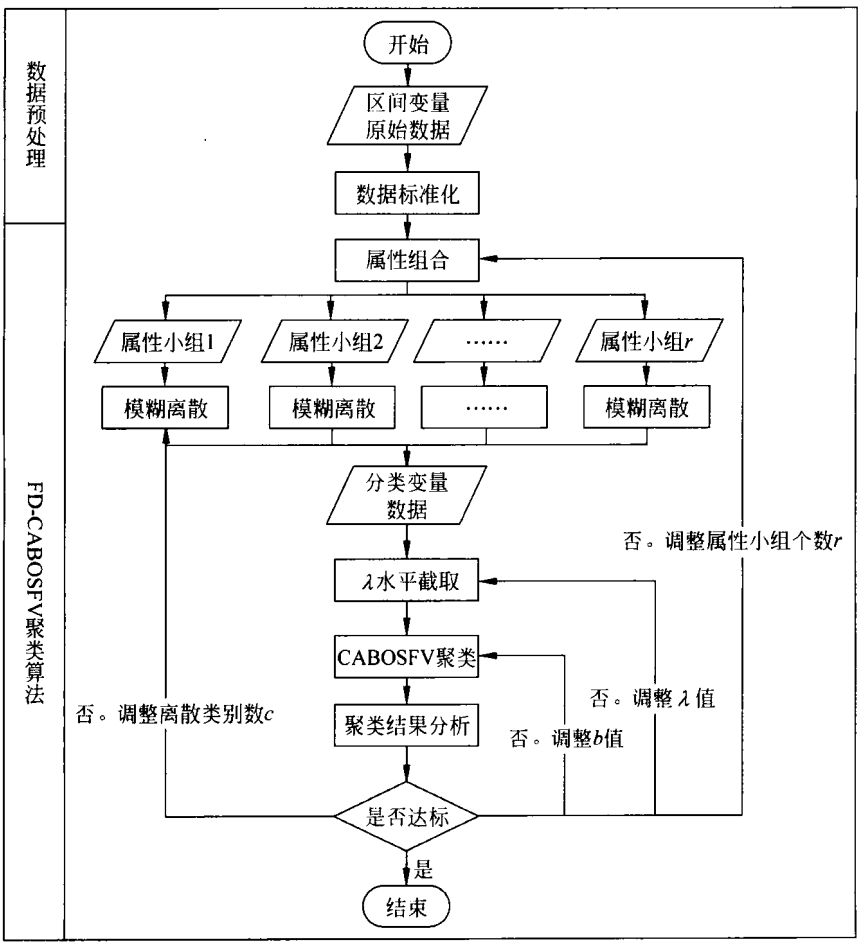

图1 FD-CABOSFV 算法聚类过程模型

3.2 算法示例

设有 10 朵花，花的序号记为{1,2,…,10}，描述每朵花的属性有 4 个，分别为花萼长度、花萼宽度、花瓣长度以及花瓣宽度，其值为区间变量，原始数据如表 1 所示。根据这 10 朵花的属性对花进行聚类，这是一个 10 个对象 4 维的聚类问题。

表 1 原始数据

花朵序号	花萼长	花萼宽	花瓣长	花瓣宽
1	5.1	3.5	1.4	0.2
2	5.0	3.6	1.4	0.2
3	5.0	3.4	1.5	0.2
4	6.9	3.1	4.9	1.5
5	6.1	2.9	4.7	1.4
6	6.0	2.9	4.5	1.5
7	6.3	3.3	6.0	2.5
8	7.2	3.6	6.1	2.5
9	6.0	2.2	5.0	1.5
10	6.7	3.1	5.6	2.4

应用 FD-CABOSFV 算法对上述问题进行聚类。描述 10 朵花 4 个属性的数据集为$[x_{ij}]_{10\times4}$，其中 $i\in\{1,2,\cdots,10\}$，$j\in\{1,2,3,4\}$。x_{ij} 为第 i 朵花在第 j 个属性的取值，FD-CABOSFV 处理步骤如下：

(1) 标准化处理。应用式(7)进行标准化后的数据如表 2 所示。

表 2 标准化后数据

花朵序号	花萼长	花萼宽	花瓣长	花瓣宽
1	−0.177 540	0.124 423	−1.237 300	−2.581 890
2	4.341 902	10.471 620	−0.164 630	−0.207 660
3	−1.145 610	−0.316 570	−2.551 860	−1.171 700
4	0.466 167	−0.683 300	2.121 556	1.771 358
5	−4.495 410	8.770 746	−1.551 730	−0.873 820
6	1.115 338	−0.379 020	0.647 499	−0.102 590
7	−1.924 060	−0.597 910	−3.433 810	−4.374 580
8	2.001 930	18.178 390	1.335 968	0.609 221
9	−2.018 360	−0.149 010	−1.768 180	−1.178 740
10	1.168 590	−0.822 720	1.373 658	1.001 579

(2) 属性组合。以每 2 个属性为一个组合进行属性划分，则该数据集的属性可分为 2 个小组，每小组的数据单独进行离散化处理。

(3) 模糊离散化。设置各属性小组的离散化属性数 $c=3$，使用 FCM 算法针对各属性小组对应的数据分别进行模糊离散化，其离散化结果如表 3 所示，所得的结果用离散化属性 $t,t\in\{1,2,\Lambda,\cdots,c*r\}$ 表示，其中 c 为各属性小组的离散化属性数，r 为属性小组的个数。

表3 模糊隶属度矩阵

花朵序号	离散化属性1	离散化属性2	离散化属性3	离散化属性4	离散化属性5	离散化属性6
1	0.002 668	0.209 050	0.788 280	0.082 196	0.716 280	0.201 520
2	0.771 410	0.107 720	0.120 870	0.493 720	0.448 530	0.057 752
3	0.002 420	0.813 060	0.184 520	0.035 201	0.900 710	0.064 091
4	0.000 902	0.035 356	0.963 740	0.912 750	0.064 294	0.022 961
5	0.368 610	0.358 710	0.272 680	0.013 701	0.976 200	0.010 094
6	0.001 287	0.032 050	0.966 660	0.824 740	0.145 460	0.029 797
7	0.002 177	0.915 110	0.082 711	0.000 444	0.001 688	0.997 870
8	0.925 110	0.037 535	0.037 352	0.996 780	0.002 525	0.000 699
9	0.000 561	0.980 290	0.019 152	0.000 563	0.998 820	0.000 620
10	0.002 698	0.064 086	0.933 220	0.995 150	0.003 761	0.001 089

（4）λ 水平截取。设 $\lambda=0.5$，按式(6)进行水平截取，得到数据集 U，如表4所示。

表4 λ 水平截取数据集

花朵序号	离散化属性1	离散化属性2	离散化属性3	离散化属性4	离散化属性5	离散化属性6
1	0	0	1	0	1	0
2	1	0	0	0	0	0
3	0	1	0	0	1	0
4	0	0	1	1	0	0
5	0	0	0	0	1	0
6	0	0	1	1	0	0
7	0	1	0	0	0	1
8	1	0	0	1	0	0
9	0	1	0	0	1	0
10	0	0	1	1	0	0

（5）CABOSFV 聚类。对数据集 U 进行聚类分析，设一个类内对象的差异度上限 $b=1$，得到FD-CABOSFV 算法聚类的结果如表5所示。

表5 应用 FD-CABOSFV 算法聚类结果

类 别	花朵序号	花朵数目
$U_1^{(1)}$	1,3,5,9	4
$U_2^{(1)}$	4,6,10	3
$U_3^{(1)}$	7	1
$U_4^{(1)}$	2,8	2

类 $U_3^{(1)}$ 仅包含一朵花，为孤立对象类，从形成的类中除去。因此，由 FD-CABOSFV 算法得到的最终聚类结果为3个类，分别为{1,3,5,9}，{4,6,10}以及{2,8}。

4 FD-CABOSFV 算法实验

采用 UCI 机器学习数据库(http://archive.ics.uci.edu/ml)的三组数据类型为区间变量已有对象类标识的真实数据集，对 FD-CABOSFV 进行聚类效果检验，从算法的有效性和参数对算法的影响

两个方面进行实验结果分析。

4.1 实验数据集与实验环境

数据集描述如表 6 所示。其中 Image 数据集中有 1 个属性在所有对象中的取值都相同，所以在实验中去除了该属性，采用了其余的 18 个属性。

表 6 数据集描述

	数 据 集	数据集缩写	样本大小	维度数目	类的数目
1	Iris	Iris	150	4	3
2	Image Segmentation	Image	2 100	19	7
3	Statlog(Landsat Satellite)	Satellite	6 435	36	6

应用 MATLAB 数学工具、SQL SERVER 2005 数据挖掘工具以及 Visual C++开发的 CABOSFV 工具进行实验。用 MATLAB 实现 FCM 算法中模糊矩阵的计算[13]，然后用 CABOSFV 聚类算法实现高维数据的聚类分析得到 FD-CABOSFV 的最终聚类结果，最后利用 SQL Server 2005 数据挖掘工具得到 K-means 聚类结果，并将实验结果与 FD-CABOSFV 进行对比分析。对聚类结果的评判是将实际的聚类结果与原先分类数据集的类别划分进行对比[14]，计算出聚类结果的准确率 。

K-means 是应用最为广泛的聚类方法，其实用性与有效性已得到理论研究和实际应用领域的广泛认可。所以本文在假定 K-means 的聚类结果能够满足需求的情况下，将所提出的 FD-CABOSFV 方法与 K-means 进行对比，实验结果表明 FD-CABOSFV 较 K-means 聚类方法质量更高。

4.2 算法有效性分析

分别采用 FD-CABOSFV 算法和 k-means 算法对 UCI 三个真实数据集进行聚类。在 FD-CABOSFV 算法和 k-means 算法皆进行参数优化的情况下，根据对象排序不同随机进行 20 次实验的平均正确率如表 7 所示。实验结果表明：FD-CABOSFV 聚类效果明显优于 K-means 算法。

表 7 FD-CABOSFV 算法与 K-means 算法平均正确率比较

	数据集缩写	FD-CABOSFVS 算法平均正确率	K-means 算法平均正确率
1	Iris	0.903 000	0.841 333
2	Image	0.734 571	0.462 857
3	Satellite	0.804 167	0.694 167

为了检验聚类效果好是因为所提出的属性组合模糊离散化方法，还是采用的 CABOSFV 算法，将本文所提出的方法与采用其他离散化方法(包括单属性 FCM 模糊离散化方法)和 CABOSFV 聚类算法结合的聚类效果进行比较，20 次随机实验的平均正确率如表 8 所示。可以看出：采用常用离散化方法与 CABOSFV 算法结合后能够有效进行区间变量数据聚类，并且聚类效果优于 K-means，表明进行数据离散化后采用 CABOSFV 算法是有效的；而采用文中所提出的属性组合离散化方法与 CABOSFV 算法结合后的聚类效果优于采用其他常用离散化方法与 CABOSFV 算法结合后的聚类效果，更优于 K-means 算法。它会产生较好聚类效果的主要原因是采用了首先进行属性组合后再进行模糊离散化的方法。如果不进行属性组合而直接进行单个属性的模糊离散化，产生的聚类效果不一定优于传统的离散化方法。

表 8 FD-CABOSFV 算法与其他离散化方法和 CABOSFV 算法结合的平均正确率比较

	Iris	Image	Satellite
K-means 算法	0.841 333	0.462 857	0.694 167
等宽离散化-CABOSFV 算法	0.859 333	0.714 143	0.708 611
等频离散化-CABOSFV 算法	0.896 333	0.648 143	0.724 167
单属性 FCM 模糊离散化-CABOSFV 算法	0.845 333	0.570 429	0.678 333
FD-CABOSFV 算法	0.903 000	0.734 571	0.804 167

FD-CABOSFV 聚类算法针对属性组合进行区间变量数据离散化，并将 FCM 聚类算法应用于离散化过程，适合于对多个属性同时进行离散化后应用 CABOSFV 算法进行聚类。在属性小组的个数 r 取值为数据集的维数时，每个属性组合里只有 1 个属性，FD-CABOSFV 聚类算法就转化为针对单属性进行 FCM 模糊离散化后应用 CABOSFV 算法进行聚类。实验结果表明：针对属性组合进行离散化与针对单个属性进行离散化相比，可以提高 FD-CABOSFV 最终聚类结果的质量。

相关参数有属性小组的个数 r、每个属性小组 FCM 模糊离散化后属性数目 c、对模糊矩阵截取的隶属度下限 λ 以及 CABOSFV 算法的差异度上限 b 值。参数对算法的影响讨论如下。

4.3 属性小组的个数 r 对聚类结果的影响

为了检验属性小组的个数 r 对聚类结果的影响，采用 FD-CABOSFV 算法分别对 UCI 三个真实数据集进行聚类。在其他参数进行优化的情况下，按照对象顺序不同随机进行 20 次实验的平均正确率随属性小组的个数 r 变化的情况如表 9 所示。表中考虑了 r 的所有可能取值，“*”表示平均正确率低于 K-means 算法的情况。从分析表中实验结果可知：在 r 的绝大多数取值情况下，FD-CABOSFV 聚类效果优于 K-means 算法；只有在 Iris 数据集中 $r=1$ 及在 Satellite 数据集中 r 取值为最大的三种情况时，FD-CABOSFV 聚类效果不如 K-means 算法。由于在其他参数不变的情况下，属性小组的个数 r 越小，属性组合模糊离散化后的维数越少，所以在需要降维的情况下建议参数 r 设置为 2 或 3。

表 9 属性小组的个数 r 对聚类结果的影响

Iris		Image		Satellite	
属性小组个数	平均正确率	属性小组个数	平均正确率	属性小组个数	平均正确率
$r=4$	0.845 333	$r=18$	0.570 429	$r=36$	0.678 333*
$r=2$	0.903 000	$r=9$	0.479 571	$r=18$	0.673 333*
$r=1$	0.820 000*	$r=6$	0.734 571	$r=12$	0.690 000*
—	—	$r=5$	0.501 714	$r=9$	0.708 333
—	—	$r=4$	0.486 143	$r=8$	0.698 333
—	—	$r=3$	0.544 286	$r=6$	0.757 222
—	—	$r=2$	0.520 143	$r=5$	0.764 167
—	—	$r=1$	0.465 714	$r=4$	0.774 167
—	—	—	—	$r=3$	0.804 167
—	—	—	—	$r=2$	0.783 611
—	—	—	—	$r=1$	0.711 111

4.4 每个属性小组 FCM 模糊离散化后属性数目 c 对聚类结果的影响

为了检验每个属性小组 FCM 模糊离散化后属性数目 c 对聚类结果的影响，采用 FD-CABOSFV

算法分别对 UCI 三个真实数据集进行聚类。在其他参数进行优化的情况下，按照对象顺序不同随机进行 20 次实验的平均正确率随参数 c 变化的情况如表 10 所示。其中，“#”表示平均正确率最高的情况。实验结果表明：在进行属性组合的情况下，c 的选取在[3,5]区间的某个值时聚类质量达到最高水平；随着 c 的降低，聚类质量可能下降；随着 c 的增加，聚类质量也可能下降。因此 c 的选取一般在[3,5]区间是比较合适的。

表 10 每个属性小组 FCM 模糊离散化后属性数目 c 对聚类结果的影响

	平均正确率					
	$c=2$	$c=3$	$c=4$	$c=5$	$c=6$	$c=7$
Iris	0.713 333	0.903 000#	0.769 333	0.681 334	0.540 000	0.487 667
Image	0.400 000	0.734 571#	0.501 714	0.533714	0.536 714	0.544 286
Satellite	0.644 722	0.656 389	0.744 167	0.804 167#	0.756 111	0.730 556

根据上述分析，参数 r 设置为 2 或 3 及参数 c 设置为[3,5]时一般能取得比较好的聚类结果。在这样的参数设置情况下，FD-CABOSFV 方法对 15 维以上的数据进行聚类时，属性组合模糊离散化的过程能实现降维，而且降维后最多为 15 维，一般维数越高降维比例越高；而对于 15 维以下的数据进行聚类时，属性组合模糊离散化的过程有可能使维数增加，但增加后的维数不会超过 15 维。另外，由于 CABOSFV 算法对二态变量高维数据的处理能力很强，即使离散化的过程没能实现降维甚至使得维数增加，仍然可以在离散化后采用 CABOSFV 算法。

4.5 隶属度下限 λ 对聚类结果的影响

隶属度下限 λ 取值不同，会影响模糊离散化的结果，λ 取值不同代表对边界值的不同处理方式，继而影响到后续的聚类结果。以数据集 Iris 数据集为例，在 $r=2$、$c=3$ 且 λ 在[0.1,0.9]区间间隔 0.1 的九种情况下进行聚类分析，聚类结果随 λ 变化的趋势如图 2 所示。

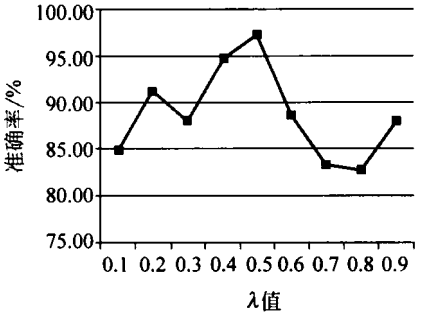

图 2 λ 值对聚类准确率的影响

实验表明，若 λ 值过大，模糊离散化控制过严，将一些原本应该归属到某类别的对象排除在该类别之外。反之，若 λ 值过小，使得边界区域的对象隶属于较多的类，造成类别间差异减弱，影响聚类效果。λ 值设定为 0.5 时一般能得到比较好的聚类结果。

4.6 差异度上限 b 对聚类结果的影响

FD-CABOSFV 算法中的参数 b 为差异度上限值，用来描述集合内部各对象间的总体差异程度上限。以数据集 Image 为例进行试验，在 $r=2$、$c=3$ 且 $\lambda=0.5$ 条件不变的情况下测试 b 的取值对聚类

结果的影响,其实验结果如图3所示。

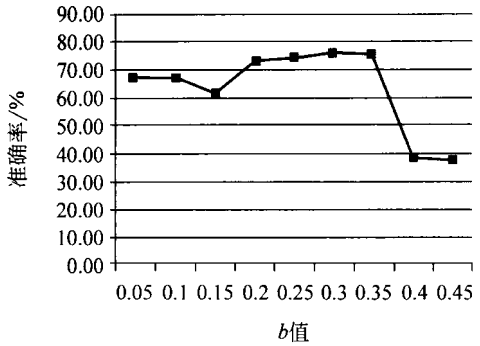

图3 b值对聚类结果准确率的影响

实验表明,如果b值取得较小,形成的聚类结果中集合内的差异度很小,聚类结果相对更精细。但如果b值过小,形成的聚类结果中集合内的对象会较少,甚至使得原数据集中对象各成一类,达不到聚类的目的。本例中当b取为0.03时,聚类个数高达40个。若b值较大,形成的聚类结果中集合内的差异度较大,聚类结果相对粗糙。若b值过大,形成的聚类结果中集合内对象会很多,聚类个数较少,甚至使得原数据集中对象全部聚成一类,达不到聚类的目的。一般可以通过b的取值不同来调整类的规模。参数b是CABOSFV算法本身所需的参数,而不是本文提出的算法引入的,在CABOSFV相关文献中已有充分的讨论。

5 结论

提出一种基于模糊离散化的区间变量高维数据聚类算法FD-CABOSFV,针对属性组合进行区间变量数据离散化,并将模糊C均值聚类算法应用于离散化过程,即利用模糊离散化的概念对区间变量进行离散化处理,得到模糊离散化后的类别变量,进而通过λ水平截取的方式将类别变量转化为二态变量,采用CABOSFV二态变量高维数据聚类算法得到区间变量数据聚类的最终结果。FD-CABOSFV算法能有效解决区间变量高维数据聚类问题,与著名的K-means聚类算法相比具有明显优势。但是该算法受参数的影响,文中给出了进行参数设定的建议。本文在进行属性组合时采用的是简单顺序组合方式,即从第一个属性开始根据属性小组中包含的属性数目顺序选取属性进行组合。是否有更好的属性组合方法使得聚类质量能够进一步高,还有待深入研究。

参考文献

[1] 孙吉贵,刘杰,赵连宇. 聚类算法研究[J]. 软件学报,2008,19(1): 48-61.

[2] 贺玲,蔡益朝,杨征. 高维数据聚类方法综述[J]. 计算机应用研究,2010,27(1): 23-26.

[3] Jolliffe I T. Principal component analysis[M]. New York,USA: Springer-Verlag,1986.

[4] Tenenbaum J B,Silva V,Langford J C. A global geometric framework for nonlinear dimensionality reduction[J]. Science,2000,290(5500): 2319-2323.

[5] Roweis S T,Saul L K. Nonlinear dimensionality reduction by locally linear embedding[J]. Science,2000,290(5500): 2323-2326.

[6] 武森,高学东,M. 巴斯蒂安. 高维稀疏聚类知识发现[M]. 北京: 冶金工业出版社,2003.

[7] 姚忠,魏佳,吴跃. 基于高维稀疏数据聚类的协同过滤推荐算法[J]. 信息系统学报,2008,2(2): 78-96.

[8] Wu S, Gao X D. CABOSFV algorithm for high dimensional sparse data clustering[J]. Journal of USTB, 2004, 11(3): 283-288.
[9] 刘希宋,喻登科,李玥. 基于客户知识的客户 CABOSFV 聚类[J]. 情报杂志,2008,(2): 7-10.
[10] Hathaway R J, Bezdek J C, Tucker W T. An improved convergence theorem for the fuzzy c-means clustering algorithms[C]. In: Bezdek J, ed. Proceedings of the Analysis of Fuzzy Information. Boca Raton, FL: CRC Press, 1987, 3: 123-131.
[11] 刘蕊洁,张金波,刘锐. 模糊 C 均值聚类算法[J]. 重庆工学院学报,2008,22(2): 139-142.
[12] 于金龙,李晓红,孙立新. 连续属性值的整体离散化[J]. 哈尔滨工业大学学报,2000,32(3): 48-53.
[13] 戈国华,肖海波,张敏. 基于 FCM 的数据聚类分析及 Matlab 实现[J]. 福建电脑,2007,(4): 89-90.
[14] 白亮,梁吉业,曹付元,等. 基于粗糙集的改进 K-Modes 聚类算法[J]. 计算机科学,2009,36(1): 162-164 转 176.

FD-CABOSFV High Dimensional Data Clustering for Interval-scaled Variables

WU Sen, ZHANG Wenli, HUANG Huimin & YE Yufei
(School of Economics and Management, University of Science and Technology Beijing, Beijing 100083)

Abstract FD-CABOSFV, an improved algorithm of CABOSFV based on fuzzy discretizaton, is proposed for high-dimensional data clustering of interval-scaled variables. It discretizes the data of each attribute portfolio by using the idea of fuzzy C means clustering, and determines each object's discretized attribute category by λ cut turning the attribute value into binary variables, and then uses CABOSFV algorithm to complete clustering. Three UCI benchmark data sets were used to compare FD-CABOSFV with famous K-means clustering algorithm. The empirical tests show that FD-CABOSFV is more effective.

Key words CABOSFV algorithm, Attribute portfolio, Fuzzy discretization, Interval-scaled variable

作者简介

武森(1971—),女,北京科技大学经济管理学院教授、博士生导师,主要研究方向是数据挖掘;
张文丽(1987—),女,北京科技大学经济管理学院硕士,主要研究方向是数据挖掘;
黄慧敏(1985—),女,北京科技大学经济管理学院硕士,主要研究方向是数据挖掘、企业信息系统;
叶俞飞(1987—),男,北京科技大学经济管理学院硕士研究生,主要研究方向是数据挖掘。