ISSN 1673-9418 CODEN JKYTA8
Journal of Frontiers of Computer Science and Technology
1673-9418/2012/06(01)-0090-07
DOI: 10.3778/j.issn.1673-9418.2012.01.007

E-mail: fcst@vip.163.com
http://www.ceaj.org
Tel: +86-10-51616056

属性赋权的 *K*-Modes 算法优化*

李仁侃+, 叶东毅
福州大学 数学与计算机科学学院, 福州 350108

Optimization of *K*-Modes Algorithm with Feature Weights*

LI Renkan+, YE Dongyi
College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350108, China
+ Corresponding author: E-mail: lirenkan@sina.com

LI Renkan, YE Dongyi. Optimization of *K*-Modes algorithm with feature weights. Journal of Frontiers of Computer Science and Technology, 2012, 6(1): 90–96.

Abstract: One major problem of the traditional *K*-Modes algorithm is the selection of features. The *K*-Modes clustering algorithm treats all features equally in the clustering process. But in practice, there are only a few important features in many data sets. To consider the particular contribution of different attributes, this paper proposes an improved algorithm called FW-*K*-Modes algorithm, which incorporates the *K*-Modes clustering algorithm with feature weight optimization. The proposed algorithm can not only improve the clustering precision in comparison with the traditional *K*-Modes clustering algorithm, but also analyze the important level of each feature in the clustering process and implement the selection of key features. The experimental results on several UCI machine learning data sets validate the effectiveness of the proposed algorithm.

Key words: *K*-Modes clustering; feature selection; automated feature weighting

摘 要：传统 *K*-Modes 算法的一个主要问题是属性选择问题。*K*-Modes 算法在聚类过程中对每一个属性都同等看待，而在实际应用中，很多数据集仅有几个重要属性对聚类起作用。为了考虑不同属性对聚类的不同影响，将 *K*-Modes 聚类算法与属性权重的最优化结合起来，提出一种属性自动赋权的 FW-*K*-Modes 算法。该算法不仅可以提高传统 *K*-Modes 聚类算法的聚类精度，还能分析各维属性对聚类的贡献程度，实现关键

*The Natural Science Foundation of Fujian Province of China under Grant No. 2010J01329 (福建省自然科学基金); the Key Science and Technology Project of Fujian Province of China under Grant Nos. 2010H6012, 2009J1007 (福建省科技重点项目).
Received 2011-03, Accepted 2011-05.

属性的选择。对多个 UCI 数据集进行了实验，验证了该算法的优良特性。

关键词：*K*-Modes 聚类；属性选择；自动属性赋权
文献标识码：A　　中图分类号：TP18

1　引言

聚类分析[1]是数据挖掘领域的一个重要分支，由于它不需要数据集结构的任何先验知识，在机器学习和智能技术等领域，通常被称为无监督的学习方法。聚类的目标是将一个数据集划分为若干个子类，使得类内对象尽可能相似，而类间对象尽可能相异。

黄哲学教授提出的 *K*-Modes 算法[2]是针对类属型数据的一种有效的聚类分析方法，在机器学习、模式识别等领域有着极为广泛的应用。*K*-Modes 算法计算相异度时，假定待分析样本的各个属性对聚类结果贡献均匀。而在实际应用中，很多数据集仅有几个重要属性对分类起作用，部分属性作用次要甚至可以忽略，此时若同等重要地依赖所有属性计算相异度，会引起“维数陷阱”[3]。例如：一个数据集由 20 个属性组成，其中仅两个属性对分类起作用，即两个数据对象是否相似仅由这两个属性决定，两个同类的对象就有可能因为不相关的 18 个属性不同导致距离很远，此时，等同作用计算相异度则引起了误导[3]。

为了考虑特征矢量中各维特征对聚类的不同贡献，一些研究者采用的特征权值学习方法主要有如下几类：

(1) 通过评价函数间接学习特征权值的方法。典型的代表有王熙照等人通过判据函数 $E(w)$[4−5]、交叉熵判据函数 $CFuzziness(w)$[6]间接学习特征权值。文献[4–6]通过实验表明，该类方法均能有效改善聚类效果，但它们局限于数值型数据的聚类，且评价函数的计算复杂度较高。

(2) 将有监督学习的特征评价方法，如信息增益方法、Relief/ReliefF 方法等，应用在无监督的聚类算法上。典型的代表有李洁等人[7]利用 ReliefF 技术对特征进行加权选择，并进行模糊聚类。文献[7]通过实验验证，其特征加权方法能使聚类效果优于传统聚类方法，但它使用有监督的学习方法进行特征加权，学习效果依赖于第一次聚类的结果，由于采用随机初始化方法，运行结果不稳定，有可能得到较差的聚类效果。

(3) 将 *K*-Means 聚类算法与特征权重优化相结合的方法。典型的代表有 Huang 等人[8]提出的一个改进的 *K*-Means 算法——基于 *K*-Means 的变量自动加权聚类算法(weighting in *K*-Means, W-*K*-Means)。该算法不牺牲算法的效率，具有良好的扩展性。然而，它需要用户指定一个属性权重指数 β，而且用户很难确定一个合适的值来获得好的聚类效果。

本文借鉴 W-*K*-Means[8]算法，引进一个属性权值矩阵($\boldsymbol{W}=\{w_1,w_2,\ldots,w_m\}$)，考虑各个属性对聚类的不同影响，提出一种基于经典 *K*-Modes[2]的属性自动赋权算法(feature weighting in *K*-Modes, FW-*K*-Modes)。新算法采用一种变量自动赋权的机制，不改变原有算法的框架，增加迭代优化属性权值矩阵的步骤，不牺牲算法的效率。实验结果表明，本文算法不仅可以提高聚类精度，还能分析各维属性对聚类的贡献程度，实现关键属性的选择。

2　经典 ***K*-Modes** 算法

K-Modes 聚类算法[2]是对 *K*-Means 聚类算法的扩展，使用简单的匹配方法度量字符型对象之间的相异度，用众数(mode)代替 *K*-Means 算法中的均值，通过基于频率的方法在聚类过程中不断更新众数，使目标函数最小化。*K*-Modes 可以应用在字符型数据集，并将聚类过程转换为带约束的使得目标函数最小的最优化问题。

字符型数据描述为：设 $U=\{x_1,x_2,\ldots,x_n\}$是由 n 个字符型数据对象构成的非空有限集合；$A=\{a_1,a_2,\ldots,a_m\}$是由 m 个字符型属性构成的非空有限集合；$DOM(a_j)=\{a_{j,1},a_{j,2},...,a_{j,n_j}\}$ 是字符型属性 $a_j(1\le j\le m)$的值域，其中 n_j 是属性 a_j 所包含的不同属性值的个数；$DOM(a_j)$是一个非空有限无序的

集合。$x_i \in U(1 \leq i \leq n)$被 A 描述为:

$$x_i = \{x_{i,a_1}, x_{i,a_2}, ..., x_{i,a_m}\}$$

其中，$x_{i,a_j} \in DOM(a_j)$ $(1 \leq j \leq m)$。

令 x_i, $x_j \in U$, 分别描述为 $x_i = [x_{i,1}, x_{i,2}, \ldots, x_{i,m}]$, $x_j = [x_{j,1}, x_{j,2}, \ldots, x_{j,m}]$, 则 x_i 与 x_j 之间简单匹配的相异度量定义[2]为：

$$d(x_i, x_j) = \sum_{l=1}^{m} \delta(x_{i,l}, x_{j,l}) \tag{1}$$

$$\delta(x_{i,l}, x_{j,l}) = \begin{cases} 1, & \text{if } x_{j,l} \neq x_{i,l} \\ 0, & \text{if } x_{j,l} = x_{i,l} \end{cases}$$

经典 K-Modes 算法最优化的目标函数[2]为：

$$P(\boldsymbol{\mu}, \boldsymbol{Z}) = \sum_{l=1}^{k} \sum_{i=1}^{n} \mu_{i,l} d(X_i, Z_l) \tag{2}$$

$$\text{s.t.} \quad \sum_{l=1}^{k} \mu_{i,l} = 1 \qquad 1 \leq i \leq n \tag{3}$$

$$\mu_{i,l} \in \{0,1\} \qquad 1 \leq i \leq n,\ 1 \leq l \leq k \tag{4}$$

其中, $\boldsymbol{\mu}$ 是一个 $n \times k$ 的隶属度矩阵, n 表示数据集 U 中包含的数据对象个数, k 表示聚类的个数; $\mu_{i,l}$ 表示第 i 个数据对象是否隶属于第 l 个聚类; $Z_l = [z_{l,1}, z_{l,2}, \ldots, z_{l,m}]$ $(1 \leq l \leq k)$是第 l 类的类中心。

为了使目标函数在满足约束下达到极小化, 经典 K-Modes 算法交替使用如下两个定理迭代优化, 直到目标函数收敛。

定理 **1**[2] 隶属度矩阵 $\boldsymbol{\mu}$ 的元素 $\mu_{i,l}$ 的更新：

$$\mu_{i,l} = \begin{cases} 1, & \text{if } d(X_i, Z_l) \leqslant d(X_i, Z_h),\ 1 \leqslant h \leqslant k \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

定理 **2**[2] 设 X 是一个由符号型属性集 $A = \{a_1, a_2, \ldots, a_m\}$ 表示的符号型数据集, 并且 $DOM(a_j) = \{a_j^{(1)}, a_j^{(2)}, ..., a_j^{(n_j)}\}$, 其中 n_j 是属性 a_j $(1 \leq j \leq m)$的类别个数, 设聚类中心(mode)Z_l 由 $[z_{l,1}, z_{l,2}, \ldots, z_{l,m}]$ $(1 \leq l \leq k)$表示, 则 $\sum_{l=1}^{k} \sum_{i=1}^{n} \mu_{i,l} d(X_i, Z_l)$ (目标函数)最小的充要条件是: $z_{l,j} = a_j^{(r)} \in DOM(a_j)$。其中,

$$\left| \{\mu_{i,l} \mid x_{i,j} = a_j^{(r)}, \mu_{i,l} = 1\} \right| \geqslant \left| \{\mu_{i,l} \mid x_{i,j} = a_j^{(t)}, \mu_{i,l} = 1\} \right|,$$

$$1 \leqslant t \leqslant n_j,\ 1 \leqslant j \leqslant m$$

($|X|$表示数据集 X 的元素个数)

可以发现经典 K-Modes 算法[2]基于简单匹配的相异度定义。该定义假设各维属性在聚类过程中的重要性相同, 而实际情况中, 有些属性在聚类中发挥重要作用, 有些属性作用次要甚至是无关属性, 在这种情况下, 经典 K-Modes 算法受噪声属性误导, 得不到较好的聚类结果。

3 FW-*K*-Modes 算法

本文基于经典 K-Modes 算法[2]提出一种属性自动赋权的方法——FW-K-Modes 算法, 引进一个属性权值矩阵($\boldsymbol{W} = \{w_1, w_2, \ldots, w_m\}$), 考虑各个属性对聚类的不同影响, 基于类间距离与类内距离比值迭代优化, 动态调整属性权值矩阵, 以获得更好的聚类质量。

FW-K-Modes 算法度量数据对象之间相异度的方法是基于经典 K-Modes 算法的简单匹配法进行属性加权。

定义 **1** 令 $x_i, x_j \in U$, 分别描述为 $x_i = [x_{i,1}, x_{i,2}, \ldots, x_{i,m}]$, $x_j = [x_{j,1}, x_{j,2}, \ldots, x_{j,m}]$, 则 x_i 与 x_j 之间属性加权的相异度定义为：

$$d_w(x_i, x_j) = \sum_{l=1}^{m} w_j \delta(x_{i,l}, x_{j,l}) \tag{6}$$

$$\delta(x_{i,l}, x_{j,l}) = \begin{cases} 1, & \text{if } x_{j,l} \neq x_{i,l} \\ 0, & \text{if } x_{j,l} = x_{i,l} \end{cases}$$

FW-K-Modes 算法最优化的目标函数为：

$$F(\boldsymbol{\mu}, \boldsymbol{Z}, \boldsymbol{W}) = \sum_{l=1}^{k} \sum_{i=1}^{n} \mu_{i,l} d_w(X_i, Z_l) = \sum_{l=1}^{k} \sum_{i=1}^{n} \sum_{j=1}^{m} \mu_{i,l} w_j \delta(x_{i,j}, z_{l,j}) \tag{7}$$

$$\text{s.t.} \quad \sum_{l=1}^{k} \mu_{i,l} = 1,\ 1 \leqslant i \leqslant n$$

$$\mu_{i,l} \in \{0,1\},\ 1 \leqslant i \leqslant n, 1 \leqslant l \leqslant k$$

$$\sum_{j=1}^{m} w_j = 1,\ 0 \leqslant w_j \leqslant 1$$

其中, $\boldsymbol{\mu}$ 是 $n \times k$ 隶属度矩阵, n 表示数据集 U 中包含的数据对象个数, k 表示聚类的个数, 矩阵元素 $\mu_{i,l}$ 表示第 i 个数据对象是否隶属于第 l 个聚类; $\boldsymbol{Z} = [Z_1, Z_2, \ldots, Z_k] \in R^{m \times k}$, 是类中心(mode)矩阵, m 表示属性的个数; $\boldsymbol{W}$ 是 $m \times 1$ 的属性权值矩阵; $d_w(X_i, Z_l)$ 描述数据对象 X_i 与类中心 Z_l 的加权相异度; w_j 表示第

j 个属性的权重。

类似 W-*K*-Means 聚类算法[8]，FW-*K*-Modes 算法最优化目标代价函数 F 的问题，可以通过迭代解决以下三个子问题进行求解：

子问题 F1 固定 $\boldsymbol{Z}=\hat{\boldsymbol{Z}}$ 和 $\boldsymbol{W}=\hat{\boldsymbol{W}}$，求解最小化问题 $F(\boldsymbol{\mu},\hat{\boldsymbol{Z}},\hat{\boldsymbol{W}})$。

子问题 F2 固定 $\boldsymbol{\mu}=\hat{\boldsymbol{\mu}}$ 和 $\boldsymbol{W}=\hat{\boldsymbol{W}}$，求解最小化问题 $F(\hat{\boldsymbol{\mu}},\boldsymbol{Z},\hat{\boldsymbol{W}})$。

子问题 F3 固定 $\boldsymbol{\mu}=\hat{\boldsymbol{\mu}}$ 和 $\boldsymbol{Z}=\hat{\boldsymbol{Z}}$，求解最小化问题 $F(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{Z}},\boldsymbol{W})$。

子问题 F1 求解如下：

$$\mu_{i,l}=\begin{cases}1, & \text{if } \sum_{j=1}^{m}w_j\delta(x_{i,j},z_{l,j})\leqslant\sum_{j=1}^{m}w_j\delta(x_{i,j},z_{t,j})\\ & 1\leqslant t\leqslant k\\ 0, & t\neq l\end{cases}\tag{8}$$

子问题 F2 的求解同经典 *K*-Modes 算法类中心的求解，详见定理 2。

聚类的目标是对给定的数据集确定一种划分，使得类内的数据对象尽可能相似，类间的数据对象尽可能相异。同理，类间距离越大类内距离越小，聚类效果越好。因此，属性权值矩阵的调整可以以极大化类间距离与类内距离的比值为目标，优化聚类的效果。每个类由类中心代表，类间距离用类中心之间的距离来衡量，类内距离为每个数据对象与所属类别的类中心距离之和，通过最大化如下目标函数(类间距离与类内距离的比值)获得最优的属性权值矩阵：

$$P(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{Z}},\boldsymbol{W})=\frac{B(\hat{\boldsymbol{Z}})}{I(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{Z}},\boldsymbol{W})}=\frac{\sum_{i=1}^{k}\sum_{l=i+1}^{k}d_w(Z_i,Z_l)}{\sum_{i=1}^{n}\sum_{l=1}^{k}\mu_{i,l}d_w(X_i,Z_l)}=\frac{\sum_{j=1}^{m}w_j\sum_{i=1}^{k}\sum_{l=i+1}^{k}\delta(z_{i,j},z_{l,j})}{\sum_{j=1}^{m}w_j\sum_{i=1}^{n}\sum_{l=1}^{k}\mu_{i,l}\delta(x_{i,j},z_{l,j})}\tag{9}$$

$$\text{s.t. }\sum_{j=1}^{m}w_j=1,\ 0\leqslant w_j\leqslant 1$$

其中，$\hat{\boldsymbol{\mu}}$ 是该次迭代固定的隶属度矩阵；$\hat{\boldsymbol{Z}}$ 为该次迭代固定的类中心(mode)矩阵；$B(\hat{\boldsymbol{Z}})$ 为类间距离，$I(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{Z}},\boldsymbol{W})$ 为类内距离；n 表示数据集 U 中包含的数据对象个数；k 表示聚类的个数；m 表示属性的个数；$d_w(*,*)$ 描述两个对象的加权相异度；w_j 表示第 j 个属性的权重。

不妨设第 j 个属性的类间距离为：

$$B_j=\sum_{i=1}^{k}\sum_{l=i+1}^{k}\delta(z_{i,j},z_{l,j})$$

第 j 个属性的类内距离为：

$$I_j=\sum_{i=1}^{n}\sum_{l=1}^{k}\mu_{i,l}\delta(x_{i,j},z_{l,j})$$

则 $$P(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{Z}},\boldsymbol{W})=\frac{\sum_{j=1}^{m}w_jB_j}{\sum_{j=1}^{m}w_jI_j}=\frac{\sum_{j=1}^{m}w_j\sum_{i=1}^{k}\sum_{l=i+1}^{k}\delta(z_{i,j},z_{l,j})}{\sum_{j=1}^{m}w_j\sum_{i=1}^{n}\sum_{l=1}^{k}\mu_{i,l}\delta(x_{i,j},z_{l,j})}$$

直接求出最大化函数 P 的解如下：

$$w_j=\begin{cases}1, & \frac{B_j}{I_j}\geqslant\frac{B_t}{I_t},\ t=1,2,...,m\\ 0, & \text{otherwise}\end{cases}\tag{10}$$

由式(10)可以看出最后的解是第 j 个属性的权值为 1，其他属性的权值为 0，即剔除了所有其他属性，即使它们对聚类也有贡献，只是贡献得没有第 j 个属性多。显然这样的解是不合理的，对聚类也是没有意义的。

为了避免这种不合理的现象出现，参考最优解的结果式(10)，逐步调整属性权值矩阵。令 $\boldsymbol{W}^{(s)}=\{w_1^{(s)},w_2^{(s)},...,w_m^{(s)}\}$ 为第 s 次迭代优化的属性权值矩阵，每次更新属性权值矩阵的量以该属性对聚类的贡献程度为标准，则第 s+1 次迭代优化的属性权值更新为：

$$w_j^{(s+1)}=w_j^{(s)}+\Delta w_j^{(s)}\tag{11}$$

$$\Delta w_j^{(s)}=\frac{B_j^{(s)}/I_j^{(s)}}{\sum_{j=1}^{m}B_j^{(s)}/I_j^{(s)}}\tag{12}$$

属性矩阵按照式(11)、(12)更新后，可能不满足 $\sum_{j=1}^{m}w_j=1$ 的约束，因此需要对更新后的属性权值进行如下归一化处理：

$$w_j^{(s+1)}=\frac{w_j^{(s+1)}}{\sum_{j=1}^{m}w_j^{(s+1)}}\tag{13}$$

综上所述, 子问题 F3 可由式(11)、(12)、(13)联合求解。

下面给出 FW-*K*-Modes 算法的具体步骤：

输入：数据集 U, 聚类数目 k。

输出：聚类结果$\{C_1,C_2,\ldots,C_k\}$。

步骤 **1** 在数据集中任取 k 个数据对象作为每个类的初始聚类中心点 Z^0, 初始化属性权值矩阵 $\boldsymbol{W}^{(0)}=\{w_1^{(0)},w_2^{(0)},\ldots,w_m^{(0)}\}=\{\frac{1}{m},\frac{1}{m},\ldots,\frac{1}{m}\}$, 并用子问题 F1 的求解方法式(8)更新隶属度矩阵获得 $\boldsymbol{\mu}^0$。设 t=0。

步骤 **2** 令 $\hat{\boldsymbol{W}}=\boldsymbol{W}^t$ 和 $\boldsymbol{\mu}=\boldsymbol{\mu}^t$, 用子问题 F2 的求解方法(定理2)更新聚类中心获得 $\boldsymbol{Z}^{t+1}$, 并判断是否出现空类, 若出现, 则新的类中心用初始的类中心代替, 如果 $F(\hat{\boldsymbol{\mu}},\boldsymbol{Z}^t,\hat{\boldsymbol{W}})=F(\hat{\boldsymbol{\mu}},\boldsymbol{Z}^{t+1},\hat{\boldsymbol{W}})$, 输出 $(\hat{\boldsymbol{\mu}},\boldsymbol{Z}^t,\hat{\boldsymbol{W}})$ 并停止; 否则, 转到步骤 3。

步骤**3** 令 $\hat{\boldsymbol{\mu}}=\boldsymbol{\mu}^t$ 和 $\hat{\boldsymbol{Z}}=\boldsymbol{Z}^{t+1}$, 用子问题F3的求解方法, 即式(11)、(12)、(13)联合求解, 更新属性权值矩阵, 获得 $\boldsymbol{W}^{t+1}$, 如果 $F(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{Z}},\boldsymbol{W}^t)=F(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{Z}},\boldsymbol{W}^{t+1})$, 输出$(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{Z}},\boldsymbol{W}^t)$ 并停止; 否则, 转到步骤 4。

步骤**4** 令 $\hat{\boldsymbol{W}}=\boldsymbol{W}^{t+1}$ 和 $\hat{\boldsymbol{Z}}=\boldsymbol{Z}^{t+1}$, 用子问题F1的求解方法(式(8))更新隶属度矩阵, 获得 $\boldsymbol{\mu}^{t+1}$, 如果 $F(\boldsymbol{\mu}^t,\hat{\boldsymbol{Z}},\hat{\boldsymbol{W}})=F(\boldsymbol{\mu}^{t+1},\hat{\boldsymbol{Z}},\hat{\boldsymbol{W}})$, 输出 $(\boldsymbol{\mu}^t,\hat{\boldsymbol{Z}},\hat{\boldsymbol{W}})$ 并停止; 否则, 令 $t=t+1$, 并转到步骤 2。

FW-*K*-Modes 算法是在经典 *K*-Modes 算法[2]的框架上增加更新属性权值矩阵的步骤。计算第 j 个属性的类间距离 $B_j^{(s)}$的时间复杂度为 $O(k\times k)$, 计算第 j 个属性的类内距离 $I_j^{(s)}$的时间复杂度为 $O(n\times k)$, 则每次迭代属性权值矩阵更新的时间复杂度为 $O(m\times n\times k)$。经典 *K*-Modes 算法的时间复杂度为 $O(t\times n\times k\times m)$($t$ 为算法迭代次数)。因此 FW-*K*-Modes 算法时间复杂度与经典 *K*-Modes 算法[2]的时间复杂度一样为 $O(t\times n\times k\times m)$($t$ 为算法迭代次数)。

4 算法实验与分析

实验环境是一台 PC 机(英特尔奔腾双核 E2200, 1 GB 内存, Windows XP 环境, VC++6.0 开发平台)。为了评价聚类的质量, 借助聚类正确度 FM(Folkes and Mallows index)[9], 将聚类结果和数据集内在分类结构进行对比, FM 值越大说明聚类结果的结构和数据集内在结构越相似, 从而直观地得出聚类的准确度来评价聚类质量。

第一个实验使用的数据集是UCI的大豆疾病数据集(Soybean-small), 共有 47 条记录, 每条记录由 35 个符号型属性描述, 且都被标记为四种疾病中的一种：Diaporthe StemCanker、Charcoal Rot、Rhizoctonia Root Rot 和 Phytophthora Rot。除了 Phytophthora Rot 有 17 条记录外, 其他的每种疾病都有 10 条记录。

注：35 个属性中有 14 个属性只有 1 个属性取值, 其属性编号分别为 11, 13, 14, 15, 16, 17, 18, 19, 29, 30, 31, 32, 33, 34。

为了验证本文 FW-*K*-Modes 算法的优良特性,在相同条件下分别用传统 *K*-Modes 算法[2]和 FW-*K*-Modes 算法对大豆疾病数据集进行聚类。随机选取 k 个数据样本为初始类中心, 为了体现实验的公平性, 在相同的类中心下运行传统 *K*-Modes 算法[2]和 FW-*K*-Modes 算法, 为使实验结果更具代表性, 实验运行 100 次, 取 *FM* 平均值。实验结果如表 1。

Table 1 Experimental results of Soybean-small

表 **1** **Soybean-small** 实验结果

结果	传统 *K*-Modes 算法	FW-*K*-Modes 算法
100 次平均精确度(平均 *FM*)	0.786 906	0.804 814
100 次总运行时间/s	58	61

表 1 比较了本文 FW-*K*-Modes 算法与传统 *K*-Modes 算法[2]聚类 100 次的平均精确度及总运行时间, 可以看出 FW-*K*-Modes 算法与传统 *K*-Modes 算法计算时间差不多, 平均聚类精确度更高。表明 FW-*K*-Modes 算法在几乎不损失算法效率的前提下, 提高了聚类精度。

表 2 给出了在相同条件下 FW-*K*-Modes 算法与传统 *K*-Modes 算法分别运行 100 次相应聚类精度等级次数对比。由表中可以看出 FW-*K*-Modes 算法 28 次聚类完全正确, 而传统 *K*-Modes 算法只有 20 次; 如果以 *FM*>0.7 定义为较好的聚类效果, 则 FW-*K*-Modes 算法 71 次获得较好的聚类效果, 而传

统 *K*-Modes 算法只有 60 次。由以上数据比较可得，FW-*K*-Modes 算法能从整体上显著地改善传统 *K*-Modes 算法[2]的聚类效果。

Table 2 Comparison of the clustering number

表 **2** 聚类次数对比

FM 值	聚类次数	
	传统 *K*-Modes 算法	FW-*K*-Modes 算法
1	20	28
0.9~1	15	5
0.8~0.9	2	4
0.7~0.8	25	36
0.6~0.7	32	23
0.5~0.6	5	3
≤0.5	1	1

从 FW-*K*-Modes 算法的 100 组实验结果中随机取出 5 组聚类精度(*FM*)为 1 的实验数据，用属性权值矩阵的值作图，横坐标为各维属性编号，纵坐标为 FW-*K*-Modes 算法聚类后各维属性对应的属性权值。如图 1 所示，5 组数据分别用菱形、正方形、三角形、交叉、米字标识画出，连成 5 条曲线。

由图 1 可以看出，5 条曲线走向一致，第 11 维属性与第 13~19 以及 29~34 维属性的权值均为最小，接近于 0，说明这 14 维属性对聚类不起作用。对照原始数据，发现在这 14 维属性上，所有样本的属性值均相同，即对分类不起任何作用，为无关属性，这验证了本文 FW-*K*-Modes 算法可以识别无关属性的优良特性。同时，也可以看出第 25 维属性对聚类也几乎不起作用，第 2、12、23、28、35 维属性权值均较大，应为该数据集的关键属性。图 1 实验结果图可以用于分析各维属性对聚类的贡献程度。

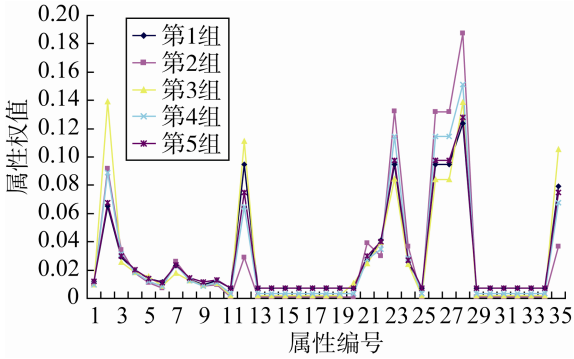

Fig.1 Feature weight curve with clustering precision (*FM*) for 1

图 **1** 聚类精度(***FM***)为 **1** 的属性权值曲线

为了进一步验证本文 FW-*K*-Modes 算法的优越性，再选取几个 UCI 数据集(见表 3)进行测试，测试结果如表4(粗体字显示最优的聚类效果)。由表 4 结果可以看出，本文 FW-*K*-Modes 算法除了在数据集 Lenses 上与传统的 *K*-Modes 算法[2]聚类结果一样，其他均获得更高的聚类精度，再次验证了 FW-*K*-Modes 算法属性加权方法的有效性。

Table 4 Comparison of the average *FM* value

表 **4** 平均 ***FM*** 值比较

数据集名	平均 *FM* 值	
	传统 *K*-Modes 算法	FW-*K*-Modes 算法
Soybean-small	0.786 906	**0.804 814**
Zoo	0.722 715	**0.731 352**
Breast-cancer	0.564 006	**0.577 253**
Credit	0.613 846	**0.634 969**
Lenses	0.445 336	0.445 336
Vote	0.767 308	**0.776 294**
Adult	0.615 351	**0.623 241**
SPECT	0.606 652	**0.617 437**

Table 3 Description of the selected UCI data sets

表 **3** 选取的 **UCI** 数据集描述

编号	数据集名	样本个数	分类属性个数	类别个数
1	Soybean-small	47	35	4
2	Zoo	101	16(剔除动物名属性)	7
3	Breast-cancer	286	9	2
4	Credit	651	9(剔除数值型属性)	2
5	Lenses	24	4	3
6	Vote	435	16	2
7	Adult	757	8(剔除数值型属性)	2
8	SPECT	267	22	2

5 结语

本文在 *K*-Modes 聚类算法框架上，提出一种自动属性赋权的 FW-*K*-Modes 算法。该算法以最大化类间距离与类内距离的比值为目标，迭代优化属性权值，保证了算法的效率。实验结果表明，FW-*K*-Modes 算法不仅能改善经典 *K*-Modes 算法的聚类效果，而且能识别关键属性和噪声属性，这对存在噪声数据的大规模高维数据的聚类具有重要意义。

References:

[1] Han Jiawei, Kamber M. Data mining concepts and techniques[M]. San Francisco, USA: Morgan Kaufmann, 2001.

[2] 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.

[3] Mitchell T M. Machine learning[M]. New York: McGraw-Hill Companies Inc, 1997: 230–247.

[4] Wang Xizhao, Wang Yadong, Zhan Yan, et al. Optimization of *K*-means clustering by feature weight learning[J]. Journal of Computer Research and Development, 2003, 40(6): 869–873.

[5] Wang Xizhao, Wang Yadong, Wang Lijuan. Improving fuzzy C-means clustering based on feature-weight learn-ing[J]. Pattern Recognition Letters, 2004, 25(10): 1123– 1132.

[6] Wang Lijuan, Guan Shouyi, Wang Xiaolong, et al. Fuzzy C mean algorithm based on feature weights[J]. Chinese Journal of Computers, 2006, 29(10): 1797–1802.

[7] Li Jie, Gao Xinbo, Jiao Licheng. A new feature weighted fuzzy clustering algorithm[J]. Acta Electronica Sinica, 2006, 36(1): 89–92.

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

[9] Zhao Heng, Yang Wanhai. Analysis of fuzzy *K*-Modes clustering accuracy[J]. Computer Engineering, 2003, 29(12): 27–28.

附中文参考文献：

[4] 王熙照，王亚东，湛燕，等. 学习特征值对 *K*-均值聚类算法的优化[J]. 计算机研究与发展，2003，40(6)：869–873.

[6] 王丽娟，关守义，王晓龙，等. 基于属性权重的 Fuzzy C Mean 算法[J]. 计算机学报，2006，29(10)：1797–1802.

[7] 李洁，高新波，焦李成. 基于特征加权的模糊聚类新算法[J]. 电子学报，2006，36(1)：89–92.

[9] 赵恒，杨万海. 模糊 *K*-Modes 聚类精确度分析[J]. 计算机工程，2003，29(12)：27–28.

LI Renkan was born in 1986. He is a master candidate at College of Mathematics and Computer Science, Fuzhou University. His research interests include data mining and computational intelligence, etc.

李仁侃(1986—)，男，福建泉州人，福州大学数学与计算机科学学院硕士研究生，主要研究领域为数据挖掘，计算智能等。

YE Dongyi was born in 1964. He received his Ph.D. degree in applied mathematics from University of Toulouse (France) in 1992. Now he is a professor and Ph.D. supervisor at College of Mathematics and Computer Science, Fuzhou University. His research interests include computational intelligence, rough set theory and optimization method, etc. He has published more than 90 papers in domestic and international journals and conferences.

叶东毅(1964—)，男，福建泉州人，1992 年于法国图卢兹大学应用数学专业获得博士学位，现为福州大学数学与计算机科学学院教授、博士生导师，主要研究领域为计算智能，粗糙集理论，最优化方法等。发表论文 90 余篇。