

文章编号:1001-9081(2011)01-0097-04

doi:10.3724/SP.J.1087.2011.00097

粗糙 K-Modes 聚类算法

李仁侃,叶东毅

(福州大学 数学与计算机科学学院, 福州 350108)

(lirenkan@sina.com)

摘要: Michael K. Ng 等人提出了新 K-Modes 聚类算法, 它采用基于相对频率的启发式相异度度量方法, 有效地提高了聚类精度, 但不足的是在计算各类的属性分类值频率时假定类中样本对聚类的贡献相同。为了考虑类中样本对类中心的不同影响, 提出一种粗糙 K-Modes 算法, 通过粗糙集的上、下近似度量数据样本在类内的重要性程度, 不仅可以获得比新 K-Modes 算法更好的聚类效果, 而且可以在保证聚类效果的基础上降低白亮等人提出的基于粗糙集改进的 K-Modes 算法的计算复杂度。对几个 UCI 的数据集的测试实验结果显示出新算法的优良性能。

关键词: 聚类; K-Modes 算法; 粗糙集; 类中心; 聚类精度

中图分类号: TP18 文献标志码:A

Rough K-Modes clustering algorithm

LI Ren-kan, YE Dong-yi

(College of Mathematics and Computer Science, Fuzhou University, Fuzhou Fujian 350002, China)

Abstract: Michael K. Ng et al. proposed the new K-Modes clustering algorithm. It takes the heuristic dissimilarity measure method based on the relative frequency and improves the clustering accuracy. However, when computing the attribute category frequency in each cluster, it assumes each object of the samples plays a uniform contribution to the cluster center. To consider the particular contribution of the different objects, a rough K-Modes algorithm was proposed in this paper. By a new approach based on the upper and lower approximation of rough set to measure the important level of each object in its corresponding cluster, the better clustering results can be achieved than the new K-Modes algorithm, and the computational complexity can be reduced in comparison with the improved K-Modes clustering algorithm based on rough sets of Bai Liang et al. with the equivalent clustering results. The experimental results on several UCI data sets illustrate the effectiveness of the proposed algorithm.

Key words: clustering; K-Modes algorithm; rough set; cluster center; clustering accuracy

0 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, 因此, 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。

在众多聚类算法中, K-means 算法^[1]以其高效而著称, 然而由于局限于处理数值型数据集而使其应用受到限制。K-Modes 算法^[2]就是在字符型数据集范围内, 对 K-means 算法进行扩展的一个有效聚类算法, 由于它继承了 K-means 高效的特点, 且算法简单、易实现, 因此被广泛应用于各个领域。随着聚类分析的模糊性不断增加, Huang 等人^[3]在传统 K-Modes 算法基础上提出了模糊 K-Modes 算法。然而, 无论是传统 K-Modes 算法还是模糊 K-Modes 算法都是由 mode 作为类的代表点, 不同的类可能具有相同的 mode, 并不能充分反映类的特征, 因此应用简单匹配方法度量对象与 mode 的相异度有可能造成信息的丢失, 弱化类内的相似性。

近几年, 很多学者对属性相似性度量方法进行了研究, 提出了几种不同的属性值加权的相似性度量方法^[4-10], 但部分仍不能有效地应用在 K-Modes 算法: 如文献[5-6]提出的方

法为有监督学习而设计; 文献[7]需要先验知识指定距离层次; 文献[9]需要其他属性的取值来衡量属性不同值的相似性, 降低了 K-Modes 的效率; 文献[10]提出了基于频率的加权度量方法, 有效地结合 K-Modes, 同时通过实验说明能取得较优的聚类效果。

Ng 等人^[11]利用基于相对频率的相异度度量对传统的 K-Modes 聚类算法进行了改进, 有效地提高了聚类精度。然而不管是传统的 K-Modes 算法还是 Ng 改进的新 K-Modes 算法^[11], 在计算类中心时都隐含假定类中各数据对象具有一样的重要性, 没有考虑类中数据对象对类中心的不同影响。而在实际应用中, 经常存在可能属于某一类也可能属于另一类的边界数据点, 因此, 在计算类中心时, 有必要给不同的数据点赋予不同的权值。针对这一情况, 利用粗糙集的上、下近似思想划分数据对象, 考虑类中各数据点的不同贡献, 提出一种基于粗糙集改进的 K-Modes 算法。

粗糙集理论^[12]自 1982 年由 Pawlak 教授提出以来, 得到了广泛的应用。Lingras 等人^[13]最早将粗糙集概念应用在聚类分析中, 提出了粗糙 K-means 聚类算法。之后粗糙集广泛应用在聚类分析中, 并获得了较好的效果。Peters 和 Lampart^[14]将粗糙集结合经典的 K-Medoids 算法, 提出粗糙 K-Medoids 聚类算法。文献[15]利用粗糙集改进 K-Modes 算法, 提出了一个基于

收稿日期:2010-06-03 ;修回日期:2010-08-17。

基金项目:国家自然科学基金资助项目(60805042);福建省自然科学基金资助项目(2010J01329)。

作者简介:李仁侃(1986-),男,福建泉州人,硕士研究生,主要研究方向:计算智能; 叶东毅(1964-),男,福建泉州人,教授,博士生导师,博士,主要研究方向:计算智能和最优化算法。

粗糙集的新相异度度量方法并重新定义类中心,有效地提高了聚类精度,但由于它需要度量每个属性下不同属性取值的相似性,而且相似性求解比较复杂,影响了算法的效率。本文提出的粗糙 K-Modes 算法把数据对象划分为各类的上、下近似,考虑了数据对象对所属类的贡献程度,使类中心 mode 对类中分布更具代表性,且各类上下近似的求解简单高效,可以在不降低 K-Modes 算法效率的情况下,提高其聚类精度。

1 相关工作概述

1.1 经典 K-Modes 算法

K-Modes 聚类算法是对 K-means 聚类算法的扩展,使用简单匹配方法度量字符型对象之间的相异度,用 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 mode 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 $U = \{x_1, x_2, \dots, x_n\}$ 是由 n 个字符型数据对象构成的非空有限集合; $A = \{a_1, a_2, \dots, a_m\}$ 是由 m 个字符型属性构成的非空有限集合; $DOM(a_j) = \{a_{j,1}, a_{j,2}, \dots, a_{j,n_j}\}$ 是字符型属性 a_j ($1 \leq j \leq m$) 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:

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

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: $x_i = [x_{i,1}, x_{i,2}, \dots, x_{i,m}]$, $x_j = [x_{j,1}, x_{j,2}, \dots, x_{j,m}]$; 则 x_i 与 x_j 之间简单匹配的相异度量定义^[2] 为:

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

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

经典 K-Modes 算法最优化的目标函数^[2] 为:

$$P(W, Z) = \sum_{l=1}^k \sum_{i=1}^n \omega_{i,l} d(X_i, Z_l) \quad (2)$$

$$\text{s. t. } \sum_{l=1}^k \omega_{i,l} = 1; 1 \leq i \leq n \quad (3)$$

$$\omega_{i,l} \in \{0, 1\}; 1 \leq i \leq n, 1 \leq l \leq k \quad (4)$$

其中: W 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; $\omega_{i,l}$ 表示第 i 个数据对象是否隶属于第 l 个聚类, $Z_l = [z_{l,1}, z_{l,2}, \dots, z_{l,m}]$ ($1 \leq l \leq k$) 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, 直到目标函数收敛为止。

定理 1^[2] 隶属度矩阵 W 元素 $\omega_{i,l}$ 的更新:

$$\omega_{i,l} = \begin{cases} 1, & d(X_i, Z_l) \leq d(X_i, Z_h), 1 \leq h \leq k \\ 0, & \text{其他} \end{cases} \quad (5)$$

定理 2^[2] 设 X 是一个由符号型属性集 $A = \{a_1, a_2, \dots, a_m\}$ 表示的符号型数据集, 并且 $DOM(a_j) = \{a_j^{(1)}, a_j^{(2)}, \dots, a_j^{(n_j)}\}$, 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 $[z_{l,1}, z_{l,2}, \dots, z_{l,m}]$ ($1 \leq l \leq k$) 表示; 则 $\sum_{l=1}^k \sum_{i=1}^n \omega_{i,l} d(X_i, Z_l)$ (目标函数) 最小的充要条件是: $z_{l,j} = a_j^{(r)} \in DOM(a_j)$ 。其中:

$$|\{\omega_{i,l} | x_{i,j} = a_j^{(r)}, \omega_{i,l} = 1\}| \geq$$

$$|\{\omega_{i,l} | x_{i,j} = a_j^{(r)}, \omega_{i,l} = 1\}|;$$

$1 \leq l \leq n_j, 1 \leq j \leq m, |X|$ 表示数据集 X 的元素个数

可以发现经典 K-Modes 算法求出的类中心不唯一,而且使用简单匹配相异度量的值要么是 0 要么是 1, 丢失了数据对象与类中心具有相同分类属性值时隐含的关系,弱化了类内的相似度。

1.2 Ng 的改进 K-Modes 算法

在文献[11]中, Ng 等人提出了新 K-Modes 聚类算法,为了挖掘数据对象与类中心具有相同分类属性时隐含的关系, 新 K-Modes 算法统计了 m 个属性的所有分类值在各类中的出现频率, 各类分别用类中 m 个出现频率最高的属性分类值更新类中心, 并记录它们的出现频率, 以此作为启发式信息来定义数据点与类中心的新的相异度度量, 即使出现相同的类中心, 数据对象分配时也可以通过类中心各属性取值在类中的支配地位(相对频率)来选择不同的类。采用新的度量方法, 强化了类内的相似度。

定义 1^[11] 属性集 $A = \{a_1, a_2, \dots, a_m\}$, 类中心 $Z = \{Z_1, Z_2, \dots, Z_k\}$, $X_i \in U$ ($1 \leq i \leq n$) 和 Z_l ($1 \leq l \leq k$), 新的相异度度量定义如下:

$$d_n(X_i, Z_l) = \sum_{j=1}^m \varphi(x_{i,j}, z_{l,j}) \quad (6)$$

其中: $\varphi(x_{i,j}, z_{l,j}) = \begin{cases} 1, & z_{l,j} \neq x_{i,j} \\ 1 - \frac{|c_{l,j,r}|}{|c_l|}, & \text{其他} \end{cases}$ 。 $|c_l|$ 为第 l 类的对像个数, 即: $|c_l| = |\{i | \omega_{i,l} = 1\}|$; $|c_{l,j,r}|$ 为第 l 类中数据对象在第 j 个属性的取值为 $a_j^{(r)}$ 的个数, 即: $|c_{l,j,r}| = |\{\omega_{s,l} | z_{l,j} = x_{s,j} = a_j^{(r)}, \omega_{s,l} = 1\}|$ 。 $\frac{|c_{l,j,r}|}{|c_l|}$ 为第 j 个属性取值为 $a_j^{(r)}$ 在 l 类中的支配地位(相对频率)。

新 K-Modes 算法仍采用经典 K-Modes 算法的框架, 用新的相异度量方法代替了简单匹配相异度量而已, 即用式(6)的 $d_n(X_i, Z_l)$ 的代替了式(2)、(5) 中出现的 $d(X_i, Z_l)$, 算法步骤、隶属度矩阵的更新方式以及类中心的更新方式都不变。通过分析发现, 在每次的迭代更新过程中, 新 K-Modes 算法只比经典 K-Modes 算法多了 $|c_l|$ 、 $|c_{l,j,r}|$ 的计算和保存, 其中计算 $|c_l|$ 的时间复杂度为 $O(n \times k)$, 计算 $|c_{l,j,r}|$ 的时间复杂度为 $O(n \times k \times M)$ (M 为 m 个属性的分类值个数之和), 而经典 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), 因此新 K-Modes 算法可以在不牺牲效率的基础上提高 K-Modes 算法的聚类精度。

2 本文的粗糙 K-Modes 算法

新 K-Modes 算法提高了在数据点分配时选择所属类的准确度,但在类中心的计算时假定所有数据点对类的贡献是一致的,忽略了现实中每个类存在一些边界点和核心点,本文就是针对这一不足提出了改进。粗糙集理论的核心思想是利用上、下近似来描述不确定的区域,本文的出发点就是利用这一思想将各个类内的边界点和核心点区分出来,并赋予不同的权重,使得求出的类中心更充分地显示类内数据点的分布,提高类中心取值的准确度。

Lingras 的粗糙 K-means 算法采用粗糙集理论的上近似和下近似来描述不确定的事物,该算法具有以下特性^[13]:

- 1) 一个数据对象最多只属于一个类的下近似;
- 2) 如果一个数据对象不属于任何类的下近似,则它属于

两个或者两个以上类的上近似;

3)一个类的下近似是该类上近似的子集。

借鉴以上特性,把既可能属于一类也可能属于另一类的点归入所有可能所属类的上近似,确定属于某一类的点归入所属类的下近似。当一个数据点跟某一类的相异度接近于跟所属类的相异度时,把该数据点划分为该类和所属类的上近似;如果不存在这样的类,则把数据点划分为所属类的下近似。通过上、下近似,可以找出该类的边界点。显然,在求解一个聚类的中心时,边界点的权重应该更小,使类中心更合理表示该类的分布。因此引进下近似和边界点的权重 w_{lower} 和 w_{boundary} (使 $w_{\text{lower}} + w_{\text{boundary}} = 1$),以及控制阈值 ε 。

定义2 属性集 $A = \{a_1, a_2, \dots, a_m\}$,类中心 $Z = \{Z_1, Z_2, \dots, Z_k\}$, $X_i \in U(1 \leq i \leq n)$ 和 $Z_l(1 \leq l \leq k)$,改进后的相异度量公式同定义1的式(6),其中 $|c_l|$ 改为考虑了边界点后第 l 类内所有数据对象的加权比重总和,即:

$$|c_l| = w_{\text{lower}} \times \sum_{x_i \in C_l}^{i=1,2,\dots,n} \omega_{i,l} + w_{\text{boundary}} \times \sum_{x_i \in C_l^B}^{i=1,2,\dots,n} \omega_{i,l}$$

$|c_{l,j,r}|$ 改为考虑了边界点后第 l 类内所有数据对象在第 j 个属性的取值为 $a_j^{(r)}$ 的加权比重总和,即:

$$|c_{l,j,r}| = w_{\text{lower}} \times \sum_{\substack{x_i \in C_l \\ x_i, j = z_i, j = a_j^{(r)}}}^{i=1,2,\dots,n} \omega_{i,l} + w_{\text{boundary}} \times \sum_{\substack{x_i \in C_l^B \\ x_i, j = z_i, j = a_j^{(r)}}}^{i=1,2,\dots,n} \omega_{i,l}$$

粗糙K-Modes算法的最优化目标函数以及隶属度更新方式同新K-Modes算法,其中相异度量方法采用定义2的度量方式,新的聚类中心更新方法采用定理3。

定理3 设 X 是一个由符号型属性集 $A = \{a_1, a_2, \dots, a_m\}$ 表示的符号型数据集,并且 $DOM(a_j) = \{a_j^{(1)}, a_j^{(2)}, \dots, a_j^{(n_j)}\}$,其中 n_j 是属性 $a_j(1 \leq j \leq m)$ 的类别个数,设聚类中心 Z_l 由 $[z_{l,1}, z_{l,2}, \dots, z_{l,m}](1 \leq l \leq k)$ 表示;则 $\sum_{l=1}^k \sum_{i=1}^n \omega_{i,l} d(X_i, Z_l)$ (目标函数)最小的充要条件是: $z_{l,j} = a_j^{(r)} \in DOM(a_j)$ 。其中:

$$\begin{aligned} & w_{\text{lower}} \times \sum_{\substack{x_i \in C_l \\ x_i, j = a_j^{(r)}}}^{i=1,2,\dots,n} \omega_{i,l} + w_{\text{boundary}} \times \sum_{\substack{x_i \in C_l^B \\ x_i, j = a_j^{(r)}}}^{i=1,2,\dots,n} \omega_{i,l} \geq \\ & w_{\text{lower}} \times \sum_{\substack{x_i \in C_l \\ x_i, j = a_j^{(t)}}}^{i=1,2,\dots,n} \omega_{i,l} + w_{\text{boundary}} \times \sum_{\substack{x_i \in C_l^B \\ x_i, j = a_j^{(t)}}}^{i=1,2,\dots,n} \omega_{i,l}; \\ & 1 \leq t \leq n_j, 1 \leq j \leq m \end{aligned}$$

证明 对于给定的隶属度矩阵 W ,且 w_{lower} 和 w_{boundary} 均为正,则目标代价函数 $P(W, Z) = \sum_{l=1}^k \sum_{i=1}^n \omega_{i,l} d_n(X_i, Z_l) = \sum_{l=1}^k \sum_{i=1}^n \sum_{j=1}^m \omega_{i,l} \varphi(x_{i,j}, z_{l,j})$ 是非负的,且假设 m 个属性互相独立,欲使 $P(W, Z)$ 最小,等价于使 $\beta_{l,j} = \sum_{i=1}^n \omega_{i,l} \varphi(x_{i,j}, z_{l,j})$ 最小,当 $z_{l,j} = a_j^{(r)}$ 时:

$$\begin{aligned} \beta_{l,j} &= \sum_{i=1, x_i, j = a_j^{(r)}}^n \omega_{i,l} \times \left(1 - \frac{|c_{l,j,r}|}{|c_l|}\right) + \sum_{i=1, x_i, j \neq a_j^{(r)}}^{i=1,2,\dots,n} \omega_{i,l} \times 1 = \\ & |c_{l,j,r}| \times \left(1 - \frac{|c_{l,j,r}|}{|c_l|}\right) + (|c_l| - |c_{l,j,r}|) = \\ & |c_{l,j,r}| - \frac{|c_{l,j,r}|^2}{|c_l|} + |c_l| - |c_{l,j,r}| = \end{aligned}$$

$$|c_l| - \frac{|c_{l,j,r}|^2}{|c_l|}$$

由于 $|c_l|$ 是固定的,显然 $\beta_{l,j}$ 最小的充要条件是 $|c_{l,j,r}|$ 取

$$\begin{aligned} &\text{得最大值,由定义2可得 } |c_{l,j,r}| = w_{\text{lower}} \times \sum_{\substack{i=1,2,\dots,n \\ X_i \in C_l \\ x_i, j = z_i, j = a_j^{(r)}}} \omega_{i,l} + \\ & w_{\text{boundary}} \times \sum_{\substack{i=1,2,\dots,n \\ X_i \in C_l^B \\ x_i, j = z_i, j = a_j^{(r)}}} \omega_{i,l}, \text{定理3得证。} \end{aligned}$$

粗糙K-Modes聚类算法步骤如下。

输入: 数据集 U , 下近似和边界点权值分别为 w_{lower} , w_{boundary} , 聚类数目 k 以及阈值 ε ;

输出:聚类结果 $\{C_1, C_2, \dots, C_k\}$ 。

步骤1 在数据集中任取 k 个点作为每个类的初始聚类中心点 Z^0 ,并用定义2和式(5)更新隶属度矩阵获得 W^0 ,并且使 $X_i \in \overline{C}_l(\omega_{i,l} = 1)$;

如果 $\exists Z_j(1 \leq j \leq k, j \neq l)$,使得 $d_n(X_i, Z_j) - d_n(X_i, Z_l) \leq \varepsilon$,则使 $X_i \in \overline{C}_j$;否则使 $X_i \in \overline{C}_l$ 。计算 $F(W^0, Z^0)$,设 $t = 0$ 。

步骤2 令 $\hat{W} = W^t$,计算上、下近似并更新各类属性分类值出现频率,使 $X_i \in \overline{C}_l(\omega_{i,l} = 1)$;

如果 $\exists Z_j(1 \leq j \leq k, j \neq l)$,使得 $d_n(X_i, Z_j) - d_n(X_i, Z_k) \leq \varepsilon$,则使 $X_i \in \overline{C}_j$;否则使 $X_i \in \overline{C}_k$ 。并用定理3更新聚类中心获得 $Z^{t+1}(C_l^B = \overline{C}_l - \overline{C}_l)$,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 $F(\hat{W}, Z^t) = F(\hat{W}, Z^{t+1})$ 输出 \hat{W}, Z^t 并停止;否则转到步骤3。

步骤3 令 $\hat{Z} = Z^{t+1}$ 并用定义2和式(5)更新隶属度矩阵获得 W^{t+1} ,如果 $F(W^t, \hat{Z}) = F(W^{t+1}, \hat{Z})$,输出 W^t, \hat{Z} 并停止;否则,令 $t = t + 1$ 并转到步骤2。

粗糙K-Modes算法中,各类上下近似划分的时间复杂度为 $O(n \times k)$,计算 $|c_l|$ 的时间复杂度为 $O(n \times k)$,计算 $|c_{l,j,r}|$ 的时间复杂度为 $O(n \times k \times M)$ (M 为 m 个属性的所有分类值个数)。因此,粗糙K-Modes算法时间复杂度跟新K-Modes算法^[11]的时间复杂度一样。

3 算法实验与分析

为了评价聚类的质量,借助聚类正确度 $FM^{[16]}$,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。

为了测试算法效果,采用UCI数据集(<http://mlearn.ics.uci.edu/databases/>)进行实验。选用的数据集描述如表1。

表1 选取的UCI数据集描述

编号	数据集	样本个数	分类属性个数	类别个数
1	Soybean-small	47	21(剔除只有一个值的属性)	4
2	Zoo	101	16(剔除动物名)	7
3	breast-cancer	286		2
4	Credit	651	9(剔除数值型属性)	2
5	Lenses	24		3
6	Vote	435		2
7	adult	757	8(剔除数值型属性)	2
8	SPECT	267		2

实验采用随机选取类中心初始化,为了体现实验的公平性,在相同的类中心下运行新K-Modes算法^[11]、文献[15]的

粗糙集改进的 K-Modes 算法和本文的算法,为使实验结果更具代表性,实验运行 100 次取平均值分析实验结果如表 2(下划线显示最优的实验结果)。

由表 2 可以看出,本文的粗糙 K-Modes 算法相对于改进前的新 K-Modes 算法聚类精度均得到有效提高,相对于文献[15]粗糙集改进的 K-Modes 各有优劣。

表 2 平均 FM 值比较

数据集	新 K-Modes	文献[15]粗糙集 改进的 K-Modes	粗糙 K-Modes
Soybean-small	0.916 838	0.905 314	<u>0.947 970</u>
Zoo	0.736 516	0.738 116	<u>0.745 406</u>
breast-cancer	0.568 915	<u>0.591 141</u>	0.589 869
Credit	0.623 846	0.610 875	<u>0.630 033</u>
Lenses	0.445 336	<u>0.456 321</u>	0.453 237
Vote	0.773 467	<u>0.784 766</u>	0.774 026
adult	0.609 083	<u>0.637 586</u>	0.614 756
SPECT	0.629 877	0.620 907	<u>0.635 743</u>

为了验证改进后粗糙 K-Modes 聚类算法的效率,表 3(下划线显示最短的运行时间)比较了在相同环境下三个算法 100 次实验总的运行时间,结果表明本文的粗糙 K-Modes 算法和新 K-Modes 算法效率差不多,文献[15]的粗糙集改进的 K-Modes 算法效率最低,由实验结果也可以看出其计算每个属性下不同值相似性的预处理时间占主要部分,远远大于聚类时间。

表 3 100 次实验运行时间比较

数据集	新 K-Modes/s	文献[15]的方法		本文 方法/s
		总时间/s	预处理时间/s	
Soybean-small	202	1 090	900	<u>138</u>
Zoo	<u>495</u>	1 361	700	609
breast-cancer	429	4 209	3 500	<u>339</u>
Credit	574	10 606	9 300	<u>529</u>
Lenses	<u>7</u>	28	20	8
Vote	<u>619</u>	6 013	4 700	626
adult	<u>1 598</u>	72 677	68 900	1 640
SPECT	736	2 436	1 600	<u>715</u>

为了测试下近似权值 w_{lower} 对算法的影响,选取 Soybean-small 数据集在相同条件下取不同的权值进行实验,实验结果如图 1。由图 1 可以看出当 $w_{lower} = 0.5$ 时,粗糙 K-Modes 聚类算法等价于新 K-Modes 聚类算法^[11],一般在 $[0.65, 0.90]$ 范围内效果较优,具体取值多少最优,根据数据集结构各异。

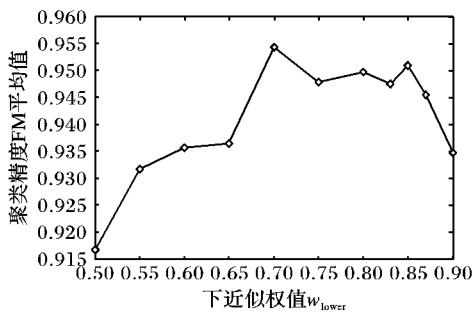


图 1 w_{lower} 取值变化的实验结果

4 结语

本文借鉴粗糙集的思想,提出一种基于数据样本加权的粗糙 K-Modes 算法。实验结果表明,与文献[11]的新 K-Modes

算法相比,该方法可以获得更优的聚类效果;与白亮等人^[15]的粗糙集改进的 K-Modes 算法相比,本文算法计算简捷,复杂度更低,且可以在保证聚类质量。本文算法利用粗糙集理论的上、下近似概念可以区分类内样本对类中心的不同贡献程度,使类中心更充分反映类内样本的分布特征,因此,针对样本分布比较分散及各类边界模糊的数据集,可以取得更好的聚类效果。

参考文献:

- [1] HAN JIAWEI, KAMBER M. Data mining concepts and techniques [M]. San Francisco, USA: Morgan Kaufmann, 2001.
- [2] HUANG ZHUXUE. Extensions to the k-means algorithm for clustering large data sets with categorical values[C]// Data Mining and Knowledge Discovery. Netherlands: Kluwer Academic Publishers, 1998: 283 – 304.
- [3] HUANG ZHUXUE, MICHAEL K NG. A fuzzy k-modes algorithm for clustering categorical data[J]. IEEE Transactions on Fuzzy Systems, 1999, 7(4): 446 – 452.
- [4] PALMER C R, FALOUTSOS C. Electricity based external similarity of categorical attributes[C]// PAKDD '03: Proceedings of the 7th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, LNAI 2637. Berlin: Springer-Verlag, 2003: 486 – 500.
- [5] LE SI QUANG, HO TU BAO. A conditional probability distribution-based dissimilarity measure for categorical data[C]// PAKDD '04: Proceedings of the 8th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, LNAI 3056. Berlin: Springer-Verlag, 2004: 580 – 589.
- [6] CHENG V, LI C-H, KWOK J T, et al. Dissimilarity learning for nominal data[J]. Pattern Recognition, 2004, 37(7): 1471 – 1477.
- [7] LEE S-G, YUN D-K. Clustering categorical and numerical data: a new procedure using multidimensional scaling [J]. International Journal of Information Technology and Decision Making, 2003, 2(1): 135 – 160.
- [8] LI CEN, BISWAS GAUTAM. Unsupervised learning with mixed numeric and nominal data[J]. IEEE Transactions on Knowledge and Data Engineering, 2002, 14(4): 673 – 690.
- [9] AHMAD A, DEY L. A method to compute distance between two categorical values of same attribute in unsupervised learning for categorical data set[J]. Pattern Recognition Letters, 2007, 28(1): 110 – 118.
- [10] HE ZENGYOU, DENG SHENGCHUN, XU XIAOFEI. Improving k-modes algorithm considering frequencies of attribute values in mode[C]// Proceedings of the International Conference on Computational Intelligence and Security, LNCS 3801. Berlin: Springer-Verlag, 2005: 157 – 162.
- [11] NG K N, LI M J, HUANG J Z, et al. On the impact of dissimilarity measure in k-modes clustering algorithm[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(3): 503 – 507.
- [12] PAWLAK Z. Rough sets[J]. International Journal of Information and Computer Sciences, 1982, 11(5): 341 – 356.
- [13] LINGRAS P, WEST C. Interval set clustering of Web users with rough k-means [J]. Journal of Intelligent Information Systems, 2004, 23(1): 5 – 16.
- [14] PETERS G, LAMPART M. A partitive rough clustering algorithm [C]// Rough Sets and Current Trends in Computing, LNCS 4259. Berlin: Springer-Verlag, 2006: 657 – 666.
- [15] 白亮, 梁吉业, 曹付元. 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.
- [16] 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.