文章编号: 1001-0920 (2012) 11-1711-04

一种改进的粗 k 均值聚类算法

王 莉[1,2], 周献中[1], 沈 捷[2]

(1. 南京大学 工程管理学院，南京 210093；2. 南京工业大学 自动化与电气工程学院，南京 210009)

摘 要: Lingras 提出的粗 k 均值聚类算法易受随机初始聚类中心和离群点的影响, 可能出现一致性和无法收敛的聚类结果. 对此, 提出一种改进的粗 k 均值算法, 选择潜能最大的 k 个对象作为初始的聚类中心, 根据数据对象与聚类中心的相对距离来确定其上下近似归属, 使边界区域的划分更合理. 定义了广义分类正确率, 该指标同时考虑了下近似集和边界区域中的对象, 评价算法性能更准确. 仿真实验结果表明, 该算法分类正确率高, 收敛速度快, 能够克服离群点的不利影响.

关键词: 聚类；粗糙集；粗 k 均值；广义分类正确率

中图分类号: TP18 **文献标志码:** A

An improved rough k-means clustering algorithm

WANG Li[1,2], *ZHOU Xian-zhong*[1], *SHEN Jie*[2]

(1. School of Engineering and Management，Nanjing University，Nanjing 210093，China；2. School of Automation and Electrical Engineering，Nanjing University of Technology，Nanjing 210009，China. Correspondent：WANG Li, E-mail：silyzheda@sina.com)

Abstract: Rough k-means clustering algorithm proposed by Lingras is sensitive to the initial centers of the k cluster and outliers and may result in identical clustering and non-convergence. In this paper, an improved rough k-means clustering algorithm is proposed. The k objects with maximum potentials are chosen as initial centers. The absolute distance between object and center of clusters is considered to decide whether a data object belongs to the lower or upper approximation set of a cluster, so the division of boundary area is more reasonable. General classification accuracy considering the objects in lower approximation set and boundary area is defined for rough k-means clustering algorithm, and it is more appropriate for evaluating rough k means clustering. The simulation results show that, the proposed algorithm has the advantages of high classification accuracy and fast convergence, and can also avoid the bad influence of outlier.

Key words: clustering；rough sets；rough k-means；general classification accuracy

1 引 言

聚类算法是一种数据挖掘的方法, 已广泛地用于各种领域, 如图像识别、文本分类、基因分析等领域[1-6]. 聚类是指按照某种相似性将一组没有类别标记的对象分为若干类别, 使得类内对象距离尽量小, 而类与类之间的距离尽可能大[7].

传统的硬 k 均值聚类算法把每个待辨识的对象严格地划分到某个类中, 具有非此即彼的性质, 这种分类的类别界限是分明的[8]. 模糊聚类是一种软划分算法, 类与类之间的界限是不清晰的, 某个对象在类属上存在中介性, 可以被划分到多个类中, 并用隶属度来描述对象属于各类的程度. Lingras[9-10]将粗糙集思想引入 k 均值聚类算法, 形成粗 k 均值聚类算法[11-12]. 其主要思想是将每个类用上下近似集来描述, 下近似集是上近似集的子集, 上下近似集之差为边界区域. 类下近似集中的对象肯定属于该类, 而边界区域的对象具有不确定性, 根据现有的知识无法确定其明确的归属. 粗 k 均值聚类是一种划分式的算法, 需要预先指定聚类数目和聚类中心, 反复迭代计算, 直至收敛时得到最终的聚类结果. 由于知识的不完备, 粗聚类算法将某些根据已有知识无法确定其明确归属的对象划分到边界区域, 与被错误地划分相比, 这样更为合理. 在更新聚类中心时, 下近似集的对象被

收稿日期: 2011-05-18；**修回日期:** 2011-11-04.

基金项目: 国家自然科学基金项目(70971062)；东南大学复杂工程系统测量与控制教育部重点实验室开放课题(2010A004).

作者简介: 王莉(1979−), 女, 博士生, 从事智能算法、粗糙集理论及应用的研究；周献中(1962−), 男, 教授, 博士生导师, 从事指挥自动化系统理论与技术、智能信息处理与智能系统等研究.

赋予较大的权重, 边界区域的对象被赋予较小的权重. 粗 k 均值算法对初值聚类中心和离群点比较敏感, 而初始聚类中心又是随机选择的, 当初始聚类中心位置不合理时会导致一致性聚类和无法收敛的结果, 同时也不能克服离群点对聚类结果带来的不利影响.

针对上述问题, 本文提出一种改进的粗 k 均值聚类算法, 通过定义潜能来确定初始的聚类中心, 并将 Lingras 粗 k 均值算法用来确定上下近似集的绝对距离改为相对距离. 针对粗聚类算法, 定义了广义分类正确率以验证算法的性能. 最后通过仿真实验验证了所提出的改进算法聚类效果较好, 并能克服离群点的不利影响.

2 改进的 Lingras 粗 k 均值聚类算法

首先为每个数据点定义潜能, 再通过 $k-1$ 次迭代计算, 选择每次迭代计算时潜能最大的对象作为初始聚类中心, 避免了初始聚类中心设置不合理的情况.

定义 1 给定数据集 $X=\{x_1,x_2,\cdots,x_N\}$, 称 $D_{x_i}(k)$ 为第 k 次迭代 x_i 的潜能, 即

$$D_{x_i}(k)=\begin{cases}\displaystyle\sum_{j=1}^{N}\exp\Big(-\frac{\|x_i-x_j\|^2}{\gamma_a^2}\Big), & k=0;\\ D_{x_i}(k-1)-\tilde{D}(k-1)\times & \\ \exp\Big(-\dfrac{\|x_i-\tilde{x}(k-1)\|^2}{\gamma_b^2}\Big), & k\geqslant 1.\end{cases}\tag{1}$$

其中: $\tilde{D}(k-1)$ 为第 $k-1$ 次迭代时的最大潜能, $\tilde{x}(k-1)$ 为第 $k-1$ 次迭代具有最大潜能的对象, 即 $\tilde{D}(k-1)=D_{\tilde{x}(k-1)}(k-1)=\max\limits_{x_i\in X}D_{x_i}(k-1)$; γ_a 为 0 ~ 1 之间的常数; γ_b 为常数, 一般取 $1.25\sim 1.5\gamma_a$.

注 1 数据集 X 已经过归一化处理.

注 2 $D_{x_i}(0)$ 是第 0 次迭代时的潜能, 称为初始潜能, 其值越大, 表明 x_i 周围的对象越多. 离群点附近的对象很少, 因此其初始潜能也很小. $D_{x_i}(0)/\tilde{D}(0)\leqslant\varepsilon,\varepsilon$ 为大于 0 的一个很小的常数, 则判定 x_i 为离群点, 不再参与潜能迭代计算.

注 3 当 $D_{x_i}(k)\leqslant 0$ 时, 数据对象 x_i 不再参与迭代.

定义 2 称 $\tilde{X}$ 为最大潜能对象集, 即

$$\tilde{X}=\{\tilde{x}(m)|m=0,1,\cdots,k-1\}.\tag{2}$$

最大潜能对象集中的每个对象周围分布着较多的数据点, 并且对象之间保持着一定的距离. 选择 $\tilde{X}$ 中的 k 个对象作为初始聚类中心, 每个初始聚类中心分散在样本空间密度较高的地方, 这样可以避免出现初始聚类中心靠得太近或者离群点被选为初始聚类中心的情况. 若离群点参与潜能的迭代计算, 则很容易在第 1 次迭代时被选为聚类中心, 所以必须将其剔除.

本文提出的改进粗 k 均值聚类算法的具体步骤如下.

Step 1: 过滤数据集, 剔除离群点. 根据式 (1) 计算各数据点的初始潜能, 确定新数据集 $\bar{X}=\{x_i\in X|D_{x_i}(0)/\tilde{D}(0)>\varepsilon\}$.

Step 2: 由式 (1) 迭代计算最大潜能对象集 $\tilde{X}$, 确定初始聚类中心 v_j.

Step 3: 确定每个对象的上下近似归属.

1) 由式 (3) 计算对象 x_n 和每个类中心 v_j 的距离

$$d_{n,j}=d(x_n,v_j)=\|x_n-v_j\|.\tag{3}$$

2) 找到与对象 x_n 距离最近的类 h, 即

$$d_{n,h}=\min_{j=1,2,\cdots,k}(d(x_n,v_j)).\tag{4}$$

3) 若 $d_{n,j}/d_{n,h}\leqslant\xi,\forall j\neq h$, 则 $x_n\in\overline{V_h},x_n\in\overline{V_j}$; 否则 $x_n\in\underline{V_h}$. 其中 ξ 取 1 ~ 1.5. $\underline{V_j}$ 表示第 j 类的下近似集, $\overline{V_j}$ 表示第 j 类的上近似集, $\underline{V_j}\subseteq\overline{V_j}$, $|\cdot|$ 表示基数. 若数据对象 x_n 属于某个类的下近似集, 则同时也属于该类的上近似集.

Step 4: 根据式 (5) 更新聚类中心

$$v_j=\begin{cases}\omega_l\dfrac{\displaystyle\sum_{x_n\in\underline{V_j}}x_n}{|\underline{V_j}|}+\omega_u\dfrac{\displaystyle\sum_{x_n\in\bar{V}_j}x_n}{|\bar{V}_j|}, & \underline{V_j}\neq\phi;\\ \dfrac{\displaystyle\sum_{x_n\in\bar{V}_j}x_n}{|\bar{V}_j|}, & \text{otherwise.}\end{cases}\tag{5}$$

其中: ω_l, ω_u 为权系数, 通常取值在 0 ~ 1 之间, $\omega_l+\omega_u=1$.

Step 5: 检查算法是否达到收敛条件. 当聚类中心不再发生变化或达到最大迭代次数时, 计算结束; 否则返回 Step 3.

以往的粗 k 均值聚类算法在计算分类正确率时, 只考虑了下近似集中的对象. 本文为粗聚类算法定义了广义分类正确率, 由于该指标考虑了边界区域中的对象, 作为验证算法性能的指标更加合理. 该指标不仅适用于粗聚类, 对于传统的硬聚类同样适用. 数据集 $X=\{x_1,x_2,\cdots,x_N\}$, 称 $C=\{c_i|c_i=1,2,\cdots,k\}$ 为类别集.

定义 3 第 j 类的原始类中心为

$$v_{oj}=\sum_{x_i\in V_{oj}}x_i\Big/|V_{oj}|,\tag{6}$$

其中 V_{oj} 表示数据集 X 中属于第 j 类的样本集合.

定义 4 称 $\hat{f}$ 为分类映射, 对于数据集 $\hat{X}=\{x_i|x_i\in\underline{V_j},j=1,2,\cdots,k\}$, $\forall x_i\in\hat{X}$, 存在一个确定的 $c_i\in C$ 与之对应, 记为 $\hat{f}:\hat{X}\to C$.

计算聚类中心 v_m 到每个原始类中心 v_{oj} 的距离 $d(v_m,v_{oj})$, 当式 (7) 成立时, $x_i\in\underline{V_m}$ 的类标识为 h, 即

$$\{\hat{f}(x_i)=h|d(v_m,v_{oh})=\min_{j=1,2,\cdots,k}d(v_m,v_{oj})\}. \quad (7)$$

定义 5　称 δ 为广义分类正确率, 有

$$\delta=\frac{\sum_{j=1}^{k}|V_{oj}\bigcap\{x_i|\hat{f}(x_i)=j,x_i\in\hat{X}\}|}{|X|}\times 100\%. \quad (8)$$

粗聚类算法中的边界区域对象由于无法确定属于哪一类, 在计算分类正确率时, 排除在正确划分的对象之外.

3　仿真实验

通过一组人工生成的二维数据集和 3 组 UCI 机器学习数据集对 Lingras 粗 k 均值聚类算法 (算法 1) 和本文提出的改进粗 k 均值聚类算法 (算法 2) 进行了一系列仿真实验. 设定最大迭代次数为 100, 算法 1 的参数为: $\omega_l=0.8,\omega_u=0.2,\varepsilon=0.05$; 算法 2 的参数为: $\gamma_a=0.25,\gamma_b=0.375,\omega_l=0.8,\omega_b=0.2,\xi=1.3$.

3.1　人工数据集仿真实例

首先给出图形中的符号含义: "◯"表示初始聚类中心, "×"和"+"分别表示类 1 和类 2 的下近似集, "△"表示边界区域, "■"表示最终得到的聚类中心, 聚类数目 $k=2$.

3.1.1　一致性聚类情况

当初始聚类中心的位置比较接近时, 算法 1 容易产生一致性聚类. 算法 2 的聚类结果如图 1 所示, 类 1 和类 2 的下近似集数目分别为 35 和 75, 边界区域有 2 个对象. 由于初始聚类中心位置合理, 经过 3 步迭代计算便能收敛, 并且聚类质量较好.

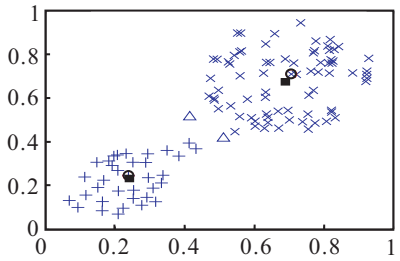

图 1　算法 2 的聚类结果

3.1.2　算法无法收敛情况

当初始聚类中心在图 2 所示的位置时, 算法 1 无法收敛, 最终聚类的结果在图 2 和图 3 之间来回振荡.

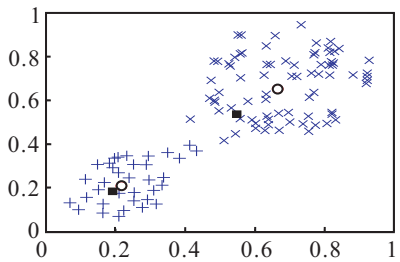

图 2　算法 1 无法收敛时的结果 (1)

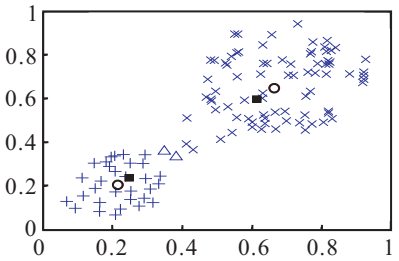

图 3　算法 1 无法收敛时的结果 (2)

这是因为当某次迭代计算得到如图 2 所示的类划分时, 按照算法求出下一步迭代的结果如图 3 所示, 聚类中心与实际情况有偏离. 由于算法 1 在更新聚类中心时, 即使所有对象都属于下近似集, 也需要乘以系数 ω_l, 得到的聚类中心如图 2 所示, 显然是不合理的.

3.1.3　含离群点的情况

为了分析离群点对粗 k 均值算法带来的不利影响, 在数据集中人为加入 1 个离群点 (4.7, 3.8), 这里为了显示清楚, 将图形局部放大. 算法 1 可能出现图 4 的聚类结果, 离群点被单独聚为一类, 其他的数据聚为一类. 这是由于离群点的存在影响了初始聚类中心的位置, 从而改变了最终的聚类结果.

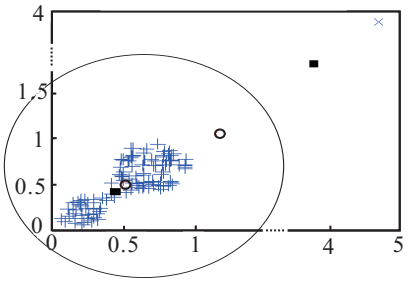

图 4　算法 1 离群点的影响

为了验证本文算法可以克服离群点对聚类结果带来的不利影响, 这里将算法 2 用于同样的数据集, 聚类结果如图 5 所示, 离群点被划分到边界区域. 这是因为通过本文算法找到的初始聚类中心位于数据对象密集的区域, 较为合理, 避免了将离群点选为聚类中心的可能性, 并且以相对距离确定上下近似集, 离群点会被分到边界区域, 对于聚类中心的更新影响也较小.

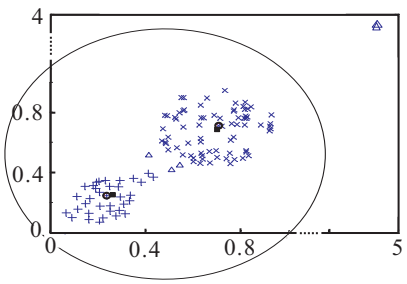

图 5　算法 2 离群点的影响

3.2 UCI数据集仿真实例

从UCI机器学习数据库[13]中选择wdbc, wine和soybean三个分类数据集, 表1对这几个分类数据集进行了描述.

表1 分类数据集

数据集	wdbc	wine	soybean
样本数	569	178	47
条件属性个数	32	13	35
聚类数	2	3	4

变量的量纲不同对聚类的结果有影响, 这里事先对数据集进行归一化处理. 采用算法1对表1中的3组数据集进行30次实验. 由于算法1的初始聚类中心是随机选择的, 聚类的结果呈现多样性. 当初始聚类中心选择不合理时, 会导致算法的边界区域对象数目较多, 甚至会造成某类下近似集为空, 即产生一致性聚类. 由于不能确定类别归属的对象数目太多, 导致相应的分类正确率非常低, 在wdbc数据集上甚至出现了下近似集全为空、分类正确率为零的情形. 算法1在wine和soybean数据集上都出现了无法收敛的情况.

采用算法2对同样的数据集进行聚类, 由于本文算法对ξ值比较敏感, ξ的改变可能引起分类结果发生变化, 这里对ξ取不同的值进行了仿真实验. 对于不同的数据集, 分类正确率最高时对应的ξ值也不同, 并且边界区域对象的数目并不一定随ξ的变大而增加, 在某些情况下, ξ变大, 边界区域对象的数目反而减少.

将算法2和算法1执行30次的平均计算结果进行了比较(见表2), 算法2中ξ取分类正确率最高时对应的值. 由于算法2选择的初始聚类中心位置合理, 只需要对算法执行一次, 且迭代次数较少, 分类正确率较高.

表2 算法2的性能

数据集	ξ	分类正确率/%		边界区域对象百分比/%		迭代次数	
		算法1	算法2	算法1	算法2	算法1	算法2
wdbc	1.01	69.64	94.475	25.69	0.18	8.33	10
wine	1.01	81.54	94.94	9.98	0.56	22.53	9
soybean	1.03	87.16	100	0.21	0	13	8

4 结 论

本文对Lingras的粗k均值算法进行了改进, 为每个数据对象定义潜能, 选择潜能最大的k个对象作为初始聚类中心, 用相对距离代替绝对距离作为评判上下近似归属的标准. 算法获得的初始聚类中心位置合理, 聚类效果较好, 并且可以克服离群点带来的不利影响. 为评判粗聚类算法的性能, 定义了广义分类正确率, 更为恰当.

本文的算法对参数ω_l, ω_u, ξ的选择敏感, 特别是ξ值的选择. ξ值的微小改变都可能引起每个类的上下近似集发生变化, 并且对于不同的数据集最优的ξ值也可能不同. 关于这些参数的选择, 目前主要通过反复实验对比结果来确定, 尚未有理论指导, 这是以后需要进一步研究的工作.

参考文献(References)

[1] Chow C K, Zhu H L, Lacy J, et al. A cooperative feature gene extraction algorithm that combines classification and clustering[C]. IEEE Int Conf on Bioinformatics and Biomedicine Workshop. New York: IEEE Press, 2009: 197-202.

[2] Matsumoto T, Hung E. Fuzzy clustering and relevance ranking of web search results with differentiating clustering label generation[C]. IEEE Int Conf on Fuzzy Systems. New York: IEEE Press, 2010: 1-8.

[3] 韩敏, 范剑超. 单点逼近型加权模糊C均值算法的遥感图像聚类应用[J]. 中国图象图形学报, 2009, 14(11): 2333-2340.
(Han M, Fan J C. A single-point approximation weighted fuzzy C-means clustering method for classifying remote sensing images[J]. J of Image and Graphics, 2009, 14(11): 2333-2340.)

[4] Ukkonen A. Clustering algorithms for chains[J]. Machine Learning Research, 2011, 12: 1389-1423.

[5] 白雪峰, 蒋国栋. 基于改进k-means聚类算法的负荷建模及应用[J]. 电力自动化设备, 2010, 30(7): 80-83.
(Bai X F, Jiang G D. Load modeling based on improved k-means clustering algorithm and its application[J]. Electric Power Automation Equipment, 2010, 30(7): 80-83.)

[6] Frey B J, Dueck D. Clustering by passing messages between data points[J]. Science, 2007, 315: 972-976.

[7] 孙吉贵, 刘杰, 赵连宇. 聚类算法研究[J]. 软件学报, 2008, 19(1): 48-61.
(Sun J G, Liu J, Zhao L Y. Clustering algorithm research[J]. J of Software, 2008, 19(1): 48-61.)

[8] Shamir O, Tishby N. Stability and model selection in k-means clustering[J]. Machine Learning. 2010, 80(2/3): 213-243.

[9] Lingras P, Yan R, West C. Comparison of conventional and rough k-means clustering[C]. Int Conf on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing, Lecture Notes in Artificial Intelligence. Berlin: Springer, 2003: 130-137.

[10] Lingras P, West C. Interval set clustering of web users with rough k-means[J]. J of Intelligent Information Systems, 2004, 23(1): 5-16.

(下转第1719页)