基于自适应权重的粗糙K均值聚类算法

周 杨 苗夺谦 岳晓冬
(同济大学电子与信息工程学院 上海 201804)
(同济大学嵌入式系统与服务计算教育部重点实验室 上海 201804)
(国家高性能计算机工程中心同济分中心 上海 201804)

摘 要 原有Rough K-means算法中类的上、下近似采用固定经验权重,其科学性值得商榷,针对这一问题,设计了一种基于自适应权重的粗糙K均值聚类算法。基于自适应权重的粗糙聚类算法在每一次迭代过程中,根据当前的数据划分状态,动态计算每个样本对于类的权重,降低了原有算法对初始权重的依赖。此外,该算法采用近似集合中的高斯距离比例来表现样本权重,从而可以在多种数据分布上得到更精确的聚类结果。实验结果表明,基于自适应权重的粗糙K均值算法是一种较优的聚类算法。

关键词 聚类,粗糙集,粗糙K均值,自适应权重
中图法分类号 TP391 **文献标识码** A

Rough K-means Clustering Based on Self-adaptive Weights

ZHOU Yang MIAO Duo-qian YUE Xiao-dong
(The School of Electronics and Information Engineering, Tongji University, Shanghai 201804, China)
(The Key Laboratory of Embedded System and Service Computing Ministry of Education, Shanghai 201804, China)
(Tongji Branch, National Engineering & Technology Center of High Performance Computer, Shanghai 201804, China)

Abstract The fixed weights are adopted in the traditional rough K-means algorithm to represent the different approximations of the clusters, but it is always difficult to predefine the optimal weights with little priori knowledge before clustering. Therefore, an improved rough K-means algorithm based on self-adaptive weights was proposed in this paper. The new method computes the weights for every data according to the current clustering state and no more does rely on the initial weights. Furthermore, the self-adaptive weights are obtained from the Gaussian distance ration in cluster approximation, which can lead to the more accurate clustering results. The experiments indicate that the rough K-means based on self-adaptive weights is an effective rough clustering algorithm.

Keywords Clustering, Rough sets, Rough K-means, Self-adaptive weight

1 引言

聚类分析是数据挖掘领域的一个重要分支。聚类就是一个将数据集划分为若干组或类的过程,通过聚类使得同一组内的数据样本具有较高的相似度,而不同组中的数据样本则是不相似的[1]。现有的聚类方法主要可分为以K-means[2],Fuzzy K-means[3]为代表的划分型聚类,以Cure[4],Agnes[5],Dlana[5],Rock[6],Chamelon[7],Birch[8]为代表的层次聚类、混合聚类和以Dbscan[9],Optics[10]为代表的基于密度的聚类等类型。

粗糙集[11]是Z. Pawlak于1982年提出的一种数据分析理论,是一种可以有效处理不精确、不完备、不确定性数据的软计算工具。Lingras将粗糙集理论引入聚类分析,提出了粗糙聚类(Rough Clustering)并设计了相应的Rough K-means算法[12],该算法根据样本到类心的距离,以上、下近似划分可能隶属于不同类别的数据,进而在多个层次上对数据进行聚类。粗糙聚类的基本思想在于将确定属于某类的样本点放到该类的下近似中,而将可能属于某类的样本点放到该类的上近似中,它能较好地处理无法精确分类的样本数据,因此已被广泛应用于Web数据挖掘、基因数据分析等领域,并取得了良好的应用效果。

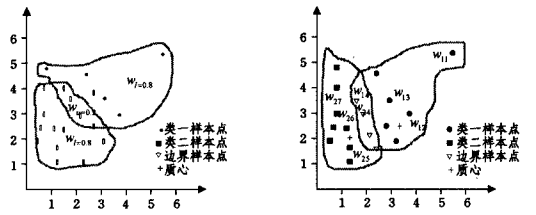
然而基本的Rough K-means算法仍然存在不足,如当下近似为空时导致目标函数无法计算,阈值选取不合理而导致聚类结果波动较大,近似表示的固定权重难以精确选定等。国内外相关文献[14-18]针对此进行了深入的研究:George Peters通过对基本Rough K-means算法的目标函数、数据稳定性和聚类稳定性进行分析,提出了针对粗糙聚类的多种改进策略[14];考虑到数据分布密度的差异性,P. Viswanath和V. Suresh Babu将近似表示引入DBSCAN算法[15],以改进聚类精度;针对传统聚类中采用的基于欧氏距离的划分方法,文献

到稿日期:2010-07-21 返修日期:2010-10-21 本文受国家自然科学基金(60475019,60970061)资助。
周 杨(1986—),男,硕士生,主要研究方向为模式识别、粗糙集理论等,E-mail:zhouyangsky109@163.com;苗夺谦(1961—),男,教授,博士生导师,主要研究方向为人工智能、模式识别、知识发现、粗糙集理论等;岳晓冬(1980—),男,博士生,主要研究方向为人工智能、模式识别等。

[16]将数据场理论引入 Rough K-means 和 Rough K-medoids 中，以改进聚类的收敛速度；文献[17]将谱聚类与粗糙聚类相结合，通过降低数据维数，使数据在子空间中的分布结构更加明显，由此改进聚类算法效率；文献[18]则针对基本 Rough K-means 算法中类均值计算问题提出了基于密度加权的粗糙聚类改进算法。

在基本的 Rough K-means 算法之中，聚类的上、下近似体现于初始步中设定的经验权重，但是固定的权重忽略了数据分布的差异性和迭代过程中类心位置的变化，通常难以精确选定，严重地影响了聚类算法的性能。因此本文将根据数据的具体分布情况，设计自适应的权重计算策略以改进聚类效果：基于自适应权重的粗糙聚类认为在聚类过程中，同一个类中的样本对类心计算的影响并不相同，因此将针对每个样本计算其所属类别的权重；自适应权重聚类将在每一次迭代过程中根据当前的数据划分状态，动态计算样本权重，降低了基本聚类方法对初始权重的依赖；此外，在计算权重的过程中，算法采用近似集合中的高斯距离比例来表现样本权重，从而可以有效减弱数据密度变化对于聚类效果的影响。

本文提出的基于自适应权重的 Rough K-means 算法，在每次迭代过程中动态计算每个样本所属不同类别的权重，进而根据样本权重和当前划分状态更新类中心，直至形成稳定的聚类结果。如图 1 所示，基本的 Rough K-means 在整个聚类过程中，对下近似和类边界中的样本设定固定权重，而基于自适应权重的方法则对于每一个样本，动态计算其所属类的权重。



(a) 基本的 Rough K-means　　(b) 自适应权重的 Rough K-means

图 1　基本 Rough K-means 和自适应权重 Rough K-means 的聚类图示

图 1(a)显示了基本 Rough K-means 算法在聚类过程中数据的上、下近似-边界区域以及对应的权重设定。显然，每个类的下近似由确定属于该类的样本组成，而边界中的样本则可能同属于多个不同类别，类的下近似及其边界构成了对应的上近似。基本的 Rough K-means 算法认为处于同一下近似或边界区域中的样本对于计算类心的作用相同，因此设定下近似的固定权重为 0.8，而边界区域的固定权重为 0.2。图 1(b)则展示了基于自适应权重的 Rough K-means 聚类算法形成的数据近似和样本权重，算法在每一次迭代中，依据当前数据划分状态，动态计算每个样本所属类别的权重。对于类下近似中的样本，只具有唯一属于该类的权重，如 $w_{1,2}$ 表示样本 2 确定属于类别 1 的权重。而对于边界区域中的样本，则具有多个可能所属类的权重，如 $w_{1,4}$ 和 $w_{2,4}$ 分别表示样本 4 属于类别 1 和属于类别 2 的权重。

本文第 2 节介绍研究背景，包括相关理论的基本概念以及传统方法存在的不足；第 3 节介绍自适应权重的计算方法，并据此给出基于自适应权重的 Rough K-means 聚类算法；第 4 节验证改进算法的有效性，实验包括人工数据实验和 UCI 数据集实验，并采用 Rough Davies- Bouldin Index [22] 和 Rand 准确率评价聚类效果；最后进行工作总结。

2　研究背景

传统的聚类算法倾向于将每个样本划分至一个确定的类中，由此形成了数据的精确划分。但是现实中需要处理的数据通常具有不确定性和不精确性，因此传统聚类算法在一些应用中就无法取得令人满意的效果。近年来，粗糙集作为一种软计算方法被广泛应用于数据挖掘领域以有效处理数据具有的多种类型的不确定性，Lingras 据此将粗糙集引入聚类设计了 Rough K-means 算法。与传统聚类方法不同，粗糙聚类算法可以在数据空间中形成覆盖，从而将无法精确分类的样本置于边界之中进行分析，由此优化聚类效果。下面简要介绍与本文方法相关的粗糙集理论基础以及基本的 Rough K-means 聚类算法。

2.1　粗糙集基本概念

设非空样本集合 U 为论域。如果在 U 上定义一个概念，那么它可以由 U 的子集 X 表示。集合的近似表示是粗糙集理论中的重要组成部分：任何在 U 上的集合概念 X 都能用它的上、下近似集合表示。

定义 1　设 R 为集合 U 上的二元关系，如果 R 是自反、对称和传递的，则 R 为 U 上的等价关系。

定义 2　设 R 为集合 U 上的等价关系，对任何 $x\in U$，集合 $[x]_R=\{y|y\in U, xRy\}$ 称为元素 x 形成的 R 等价类。

定义 3　对于每个概念 $X\subseteq U$，定义其下近似为 $\underline{R}X=\bigcup\{x\in U|[x]_R\subseteq X\}$，上近似为 $\overline{R}X=\bigcup\{x\in U|[x]_R\cap X\neq\emptyset\}$，其中 $\underline{R}X$ 表示论域中确定属于 X 的元素集合，而 $\overline{R}X$ 表示可能属于 X 的论域元素集合。

2.2　基本 Rough K-means 算法

经典 K-means 聚类算法可以简要表述为：给定 k 个类，随机挑选 k 个样本为初始类心，根据距离最近原则，将其余数据集样本点分到 k 个类中去。算法采用迭代更新的方法，通过判定给定的聚类目标函数，使每一次迭代都向目标函数值减少的方向进行。在迭代过程中，依据上步的 k 个参照点划分样本，进而更新聚类中心，直至产生稳定的聚类结果。

Lingras 将粗糙集引入经典 K-means 算法，提出了 Rough K-means 聚类算法，方法主要涉及以下性质：

性质 1　一个数据样本至多属于一个类的下近似。

性质 2　若一个数据样本属于一个类的下近似，那么它一定属于一个类的上近似。

性质 3　若一个数据样本不属于任何一个类的下近似，那么它一定同时属于两个以上类的上近似。

在 Rough K-means 算法的迭代过程中，类心的计算公式如下：

$$m_k=\begin{cases} w_l\sum\limits_{X_i\in\underline{C}_k}\dfrac{X_i}{|\underline{C}_k|}+w_u\sum\limits_{X_i\in\overline{C}_k-\underline{C}_k}\dfrac{X_i}{|\overline{C}_k-\underline{C}_k|}, & (\overline{C}_k-\underline{C}_k\neq\emptyset) \\ w_l\sum\limits_{X_i\in\underline{C}_k}\dfrac{X_i}{|\underline{C}_k|}, & (\overline{C}_k-\underline{C}_k=\emptyset) \end{cases} \tag{1}$$

设数据集 $U=\{X_i, i=1,\cdots,N\}$，上式中 $\overline{C}_k$ 和 $\underline{C}_k$ 分别表示类 C_k 的上、下近似，由此可得边界区域 $C_k^B=\overline{C}_k-\underline{C}_k$，$w_l+w_u=1$，其中 w_l 为下近似权重，w_u 为上近似权重。

2.3 存在的问题和改进策略

Lingras 提出的基本 Rough K-means 算法仍然存在问题，如当下近似为空时无法计算目标函数，不合理的阈值选取将导致波动较大的聚类结果，近似表示的固定权重难以精确选定。本文将深入分析基本 Rough K-means 算法中采用固定经验权重带来的问题，并据此提出改进策略。

(1) 在基本 Rough K-means 算法中，通常很难在初始步中依据经验直接定义准确的 w_l 和 w_u，而基于自适应权重的聚类将在迭代过程中根据当前的数据划分状态，动态计算样本权重，降低了基本方法对初始权重的依赖。

(2) 基本 Rough K-means 算法认为同一上近似或下近似中的样本对类心计算的影响是相同的，而自适应权重策略认为在聚类过程中，即使是处于同一近似中的样本在类心计算的作用也并不相同，因此将针对每个样本计算其所属类别的权重。

(3) 基本 Rough K-means 算法的固定近似权重适用于某些特定的数据分布，但是对于更复杂、非均匀分布的数据，固定权重的普适性就存在局限。因此自适应权重将采用样本与类心的高斯距离在近似集合中所占的比例来计算样本权重，从而有效减弱数据密度变化对于聚类效果的影响。

基于以上问题和拟定的改进策略，本文提出一种新的粗糙聚类算法：基于自适应权重的粗糙 K 均值聚类算法。

3 基于自适应权重的粗糙 K 均值聚类

下面将介绍自适应权重的计算策略以及基于自适应权重的 Rough K-means 聚类算法。此外，本节也将对算法的收敛控制进行简要说明。

3.1 自适应权重计算

给定 N 个样本的数据集 $U=\{X_i, i=1,\cdots,N\}$，利用高斯距离计算样本对于聚类的影响，样本的作用随着与类心距离的不断增大而平滑降低。样本 X_i 在聚类过程中对于第 k 类的影响 $f_{i,k}$ 计算如下：

$$f_{i,k}=e^{-\frac{\|X_i-m_k\|^2}{\delta_k^2}} \tag{2}$$

式中，m_k 表示第 k 类的类心，$||\cdot||$ 为欧式距离运算，δ_k 表示第 k 类的有效半径，可依据第 k 类中所有样本与类心的平均欧式距离来定义类半径：

$$\delta_k=\frac{1}{|C_k|}\sum_{X_j\in C_k}||X_j-m_k|| \tag{3}$$

式中，C_k 表示属于第 k 类的样本集，$|\cdot|$ 为集合的势。

根据单个样本对于类心的影响，可以计算近似集和边界集样本对于聚类的影响：

$$\begin{aligned} f_{up,k}&=f_{low,k}+f_{B,k}\\ f_{low,k}&=\sum_{X_i\in\underline{C}_k}f_{i,k}\\ f_{B,k}&=\sum_{X_i\in C_k^B}f_{i,k}\end{aligned} \tag{4}$$

式中，C_k^B 为第 k 类的边界集，$f_{up,k}$ 表示第 k 类上近似中所有样本对其类心计算的影响，$f_{low,k}$ 和 $f_{B,k}$ 分别表示 k 类下近似和边界中所有样本对其类心的影响。

根据样本所处的区域，归一化单个样本的影响，就可以得到样本在聚类过程中对于 k 类的权重：

$$W_{X_i,k}=\begin{cases}\dfrac{f_{i,k}}{f_{low,k}}\cdot\dfrac{f_{low,k}}{f_{up,k}}=\dfrac{f_{i,k}}{f_{up,k}}, & X_i\in\underline{C}_k,k\in\{1,2,\cdots,K\}\\ \dfrac{f_{i,k}}{f_{B,k}}\cdot\dfrac{f_{B,k}}{f_{up,k}}=\dfrac{f_{i,k}}{f_{up,k}}, & X_i\in C_k^B,k\in\{1,2,\cdots,K\}\end{cases} \tag{5}$$

式中，$W_{X_i,k}$ 表示样本 X_i 对于第 k 类的权重，归一化的处理不仅确保属于某一类的所有样本权重总和为 1，而且可以在一定程度上减弱不均匀的密度分布对于聚类的干扰。由获得的样本权重，可以按照如下公式计算并更新类心：

$$m_k=\begin{cases}\sum\limits_{X_i\in\underline{C}_k}W_{X_i,k}X_i+\sum\limits_{X_i\in C_k^B}W_{X_i,k}X_i, & (C_k^B\neq\varnothing)\\ \sum\limits_{X_i\in\underline{C}_k}W_{X_i,k}X_i, & (C_k^B=\varnothing)\end{cases} \tag{6}$$

与基本的 Rough K-means 算法相似，在聚类过程中，需要根据样本到类心的距离进一步确定类的上、下近似。对于 $\forall X_i\in U$，首先判别距 X_i 最近的类心 m_k，即 $d(X_i,m_k)=\min\{d(X_i,m_j)\}, j\in\{1,\cdots,K\}$。

设集合 $T=\{r|d(X_i,m_r)-d(X_i,m_k)<=\xi, r\in\{1,\cdots,K\}\wedge r\neq k\}$，若 $T=\varnothing$，则样本属于 k 类的下近似，$X_i\in\underline{C}_k$，否则将样本划入该类的边界 $X_i\in C_k^B$。类似于粗糙集中的近似表示，X_i 将根据距离阈值 ξ 被划分为确定属于某类或可能属于某几类的样本，阈值 ξ 对聚类效果的影响及其选定方法，将在实验部分中进行详细说明。

3.2 算法步骤描述

根据以上所述的自适应权重计算方法以及类近似表示，可以给出基于自适应权重的粗糙聚类算法的具体步骤。

算法 1 基于自适应权重粗糙 K 均值算法

输入：数据集 U，聚类数目 K，阈值 ξ 用于设定边界区域；

输出：聚类的结果 $\{\underline{C}_1, C_1^B, \underline{C}_2, C_2^B, \cdots, \underline{C}_k, C_k^B\}$；

Step1 在数据集中随机选取 K 个样本作为初始化类中心；

Step2 根据初始化类中心，对 U 中的样本进行划分，对于任意 $X_i\in U$，首先寻找与其距离最近的类心 m_k，然后根据阈值 ξ 判定 X_i 属于第 k 类的下近似或边界区域；

Step3 计算各类的半径参数 δ，然后根据当前的数据划分状态计算每一个样本 X_i 对应其各可能所属类的影响 $f_{i,k}$；

Step4 基于各样本 X_i 对其可能所属类的影响，依据式(5)计算样本权重，再依据式(6)更新对应的类中心；

Step5 若聚类趋于稳定或迭代次数超过阈值，则算法结束输出结果，否则转至 Step2。

3.3 迭代控制

算法 1 主要通过判定样本集与各类的均方差是否趋于稳定来停止迭代，可以依据如下公式来度量相邻迭代步中的均方差变化：

$$Dev=\frac{S_{\text{sum}}^t-S_{\text{sum}}^{t-1}}{S_{\text{sum}}^{t-1}} \tag{7}$$

式中，S_{sum}^t 表示迭代步 t 得到的聚类结果中样本集与各类的均方差，相似地，S_{sum}^{t-1} 为 $t-1$ 迭代步中聚类结果的均方差，设 m_k^t 和 C_k^t 为迭代步 t 得到的第 k 类的类心以及样本集，则有：

$$\begin{aligned} S_{\text{sum}}^t&=\sum_{k=1}^{K}S_{C_k}^t\\ S_{C_k}^t&=\frac{1}{|C_k|-1}\sum_{X_i\in C_k}\sqrt{(X_i-m_k)^2}\end{aligned} \tag{8}$$

在具体实验中，设定当 $Dev<10^{-4}$ 时认为聚类结果稳定，迭代停止。此外，与基本的 Rough K-means 算法相似，由于尚未从理论上证明算法的收敛性，算法的具体实现中，也采用了迭代次数的上限控制以避免可能存在的均方差波动对算

法有效性的影响。

4 实验分析

4.1 评价指标

实验将分别从非监督和监督学习的不同视角采用两类标准来评价算法的数据划分精度。对于类别信息缺失的数据，采用改进的 DBI 指标评价聚类效果。

(1)DBI(Davies-Bouldin Index)为 Davies -Bouldin 聚类有效性度量指标函数[19]，由于 DBI 指标独立于初始类数 K 的设定，因此可以采用 DBI 来评价数据划分的有效性。对于粗糙聚类算法，处于类上、下近似中的样本对类内聚的影响不同，因此 Mitra 等[20]中对 DBI 指标函数进行了改进，提出了 *RDB* 评价指标，*RDB* 指标越小，聚类效果越好。

$$RDB=\frac{1}{K}\sum_{k=1}^{K}\max_{l\neq k}\left\{\frac{S_r(C_k)+S_r(C_l)}{d(C_k,C_l)}\right\} \tag{9}$$

式中，$S_r(C_k)$为第 k 类的近似与边界区域中的样本与类心的加权均方差，用以表征该类的内聚程度。对于具有类别信息的数据，在非监督的聚类评价指标之外，还可以采用分类准确率来评价聚类的数据划分精度。

(2) 准确率指标 *Rand*，利用已知样本的类标签评价聚类结果对数据划分的准确率，对于具有 N 个样本的数据集：

$$Rand=\left(\sum_{k=1}^{K}|R_k|/N\right)\cdot 100\% \tag{10}$$

式中，R_k 为正确划入第 k 类的样本集合，对于聚类划分(或覆盖)结果中的任一子集，若其中含有 k 类别的样本数目最多，则认为该集合表征第 k 类数据的分布。

4.2 实验分析

分别采用人工数据以及 UCI 标准数据集对改进的聚类算法进行测试，依据 RDB 指标和 Rand 准确率比较分析基本 Rough K-means 算法和基于自适应权重的粗糙 K 均值算法的聚类效果。

4.2.1 人工数据实验

人工数据由 3 个高斯分布函数生成的 120 个 2 维样本组成，3 个近似高斯分布的数据集在边界处互相交叠，数据分布如图 2 所示。

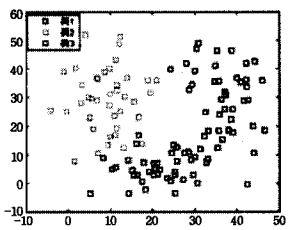
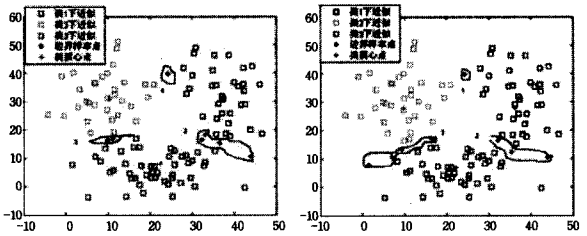


图 2 人工数据集图示

图 3 展示了基本 Rough K-means 算法和基于自适应权重的 Rough K-means 算法对人工数据集的聚类效果，参照图 2，不同的颜色表示不同的类别，曲线包围的样本为类之间的区域样本集。实验将随机选取的样本作为初始化类均值，设置阈值 $\xi=0.3$ 来确定类的近似和边界集。综合 20 次实验结果取得的聚类效果如图 3 所示，从中可见基于自适应权重的粗糙聚类算法比基本 Rough K-means 算法生成的聚类结果更接近于原有的数据类别分布。相较于原有方法，基于自适应权重的算法形成了更合理的边界区域，将不同类别样本混杂的区域判别为边界，从而降低了错误划分类别的样本数量。



(a) 基本的 Rough K-means　(b) 基于自适应权重的 Rough K-means

图 3 基本 Rough K-means 和自适应权重 Rough K-means 图示

在评价指标方面，两种算法对人工数据进行聚类，得到的 RDB 值分别为 0.4925(图 3(a))和 0.4135(图 3(b))，准确率分别为 90%(图 3(a))和 92.5%(图 3(b))。实验对多个生成的人工数据集进行测试，可知相较于基本的 Rough K-means 算法，基于自适应权重的 K 均值算法可以获得更优的聚类效果。

4.2.2 UCI 数据集实验

在人工数据之外，实验也采用 UCI 数据对改进算法的聚类效果进行测试，表 1 展示了数据集 Iris，Balance-scale 和 Wine 的基本信息。

表 1 UCI 数据集基本信息

数据集	样本点个数	维数	类数
Iris	150	4	3
Balance-Scale	625	4	3
Wine	178	13	3

实验将应用基本的 Rough K-means 和基于自适应权重的粗糙 K 均值算法对每个 UCI 数据集进行 20 次聚类，并综合各次聚类结果对算法进行比较分析。实验设定 $K=3$，由于 $\xi>1$ 时可能出现下近似为空的情况，因此设定 $\xi\in[0,1]$。两种算法在各数据集上聚类结果的 RDB 值和 Rand 准确率如图 4、图 5 所示。

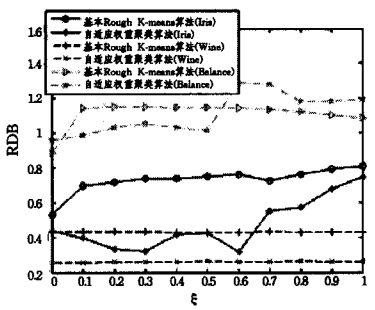


图 4 UCI 数据集 DBI 系数比较

参照非监督的 RDB 标准，由图 4 可见，对于数据集 Iris 和 Wine，基于自适应权重的算法相较于基本 Rough K-means 算法在阈值 ξ 的各个取值上均取得了更优的聚类效果。对于数据集 Balance，当 $\xi\in[0,0.5]$时，改进的权重算法取得了更好的聚类效果，而当 ξ 大于 0.5 时，RDB 值出现了较大的波动，Iris 数据集的测试结果也出现了相似的现象。造成这种现象的原因在于，对于某些数据分布，当 ξ 增大至一定程度时，会将过多的样本划入聚类的边界区域之中，从而严重地影响了各类心的更新计算。由于改进的聚类算法将根据当前数据划分状态，针对每个样本计算权重进而更新类心，因此算法对于边界区域规模的变化更加敏感，聚类结果就随着阈值的增长而出现了较大的波动。鉴于这个问题，实验选定阈值

ξ=0.3，在此取值范围内，改进算法对于大多数数据集都可以取得更优的聚类效果。此外也应注意到，当ξ为0时，边界区域为空，所有样本都被确定划入某类的下近似，这时基本Rough K-means以及改进的聚类算法都退化成为传统的K均值聚类算法。

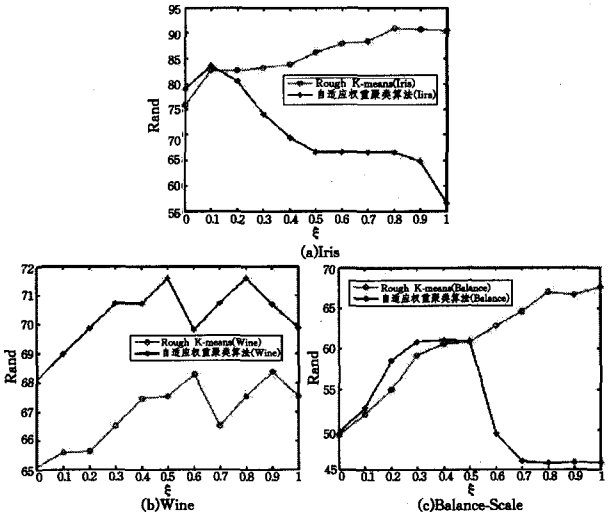


图5 UCI数据集Rand准确率比较

在准确率方面，由图5所示，在ξ处于较小的值区间时，基于自适应权重的算法相较于基本算法可以取得更高的准确率，但是对于某些特定的数据分布，当ξ超越一定范围时，改进算法的准确率会出现大幅降低。这是因为，当阈值ξ超越一定范围不断增大时，改进算法会将过多样本划入可能所属的各类，形成大规模的边界区域，又由于计算准确率前提需要判别数据子集中的多数样本类别，因此过多边界样本的加入会改变数据子集所代表的类别，由此造成准确率的大幅下降。故实验设定ξ为一个较小值(0.3)以保持适度边界，避免过大阈值导致的不良聚类效果。

综上所述，通过实验比较分析基本Rough K-means算法以及本文提出的基于自适应权重的粗糙K均值算法，总结改进聚类算法的优点和不足如下：改进算法的优势在于可以根据具体数据分布状态自适应计算样本权重，避免了基本方法中固定经验权重的设置，此外改进算法可以在ξ取值较小、形成适度边界时取得相较于基本算法更优的聚类效果；改进算法的不足在于改进的聚类算法对于边界的变化过于敏感，对于某些数据分布，在阈值ξ取值较大时，聚类结果会出现大幅度的波动，从而导致不良的数据划分精度。

结束语 本文针对原始Rough K-means算法中类的上、下近似采用固定经验权重引起的问题，提出基于自适应权重的粗糙K均值聚类算法。新聚类算法根据每次迭代的聚类结果动态地确定下一次迭代聚类中样本点的权重，并以此来改进原始K-means算法对经验权重的依赖，从而使权重的选取更加的科学，聚类结果更加精确。下一步的研究工作包括：针对边界阈值的敏感性进一步优化算法；降低算法的复杂度。

参考文献

[1] Xu Rui. Donald Wunsch II. Survey of clustering algorithm [J]. IEEE transaction on neural networks (S1045- 9227), 2005, 10 (3): 645-678

[2] Mac Queen J. Some methods for classification and analysis of multivariate observations [C]// LeCam L M, Neyman J, eds. Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probabilit. Berkeley: University of California Press, 1967: 281-297

[3] Kim T, Bezbek J C. Optimal tests for the fixed points of the fuzzy C-means algorithms [J]. Pattern Recognition (S0031-3203), 1988, 31: 651-663

[4] Guha S, Rastogi R, Shim K. CURE: An efficient clustering algorithm for large databases [C]// Proc. ACM SIGMOD Int. Conf. Management of Data. Seattle, Washington: ACM Press, 1998: 73-84

[5] Kaufmann L, Rousseeuw P J. Finding Groups in Data: An Introduction to Cluster Analysis [M]. New York: Lohn Wiley & Sons, 1990: 67-89

[6] Guha S, Rastogi R, Shim K. ROCK: A robust clustering algorithm for categorical attributes [J]. Informatic Systems (S1746-0980), 2000, 25(5): 345-366

[7] Karypis G, Han E, Kumar V. Chameleon: Hierarchical cluster in gusing dynamic modeling [J]. IEEE Computer (S0018-9162), 1999, 32(8): 68-75

[8] Zhang T, Ramakrishnan R, Livny M. BIRCH: An efficient data clustering method for very large databases [C]// Proc. of the 15th ACM SIGMOD Int'l Conf. on Management of Data. Montrel: ACM Press, 1996: 103-114

[9] Ester M, Kriegel H P, Sander J, et al. A density- based algorithm for discovering clusters in large spatial databases with noise [C]// Simoudis E, Han J W, Fayyad U, eds. Proc of the 2nd Int'l Conf on Knowledge Discovery and Data Mining, KDD96. Menlo Park: AAAI Press, 1996: 226-231

[10] Ankerst M, Breuing M, Kriegel H P, et al. OPTICS: Ordering points to identify the clustering structure [C]// Delis A, Faloutsos C, Ghandeharizadeh S, eds. Proc of the 1999 ACM SIGMOD Int'l Conf on Management of Data, 1999 ACM SIGMOD. New York: ACM Press, 1999: 46-60

[11] Pawlak Z. Rough sets[J]. International Journal of Information and Computer Sciences, 1982, 11: 145 -172

[12] Lingras P, West C. Interval set clustering of web users with rough k-means[J]. Journal of Intelligent Information Systems, 2004, 23(1): 5-1643

[13] Wang Rui-zhi, Miao Duo-qian, Li Gang, et al. Rough Overlapping Biclustering of Gene Expression Data[C]// Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering. 2007: 828-834

[14] Peters G. Some refinements of rough k-means clustering [J]. Pattern Recognition, 2006, 39(8): 1481-1491

[15] Viswanath P, Suresh Babu V. Rough-DBSCAN: A fast hybrid density based clustering method for large data sets [J]. Pattern Recognition Letters, 2009, 30(16): 1477-1488

[16] 李学，苗夺谦，冯琴荣. 基于数据场的粗糙聚类算法[J]. 计算机科学，2009，36(2)：203-206

[17] 胡云，苗夺谦，王睿智，等. 一种基于粗糙k均值的双聚类算法[J]. 计算机科学，2007，34(11)：174-177

[18] 郑超，苗夺谦，王睿智. 基于密度加权的粗糙k均值聚类改进算法[J]. 计算机科学，2009，36(3)：220-222

[19] Davies D L, Bouldin D W. A cluster separation measure [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1979, 1: 224-227

[20] Mitra S, Banka H, Pedrycz W. Collaborative Rough Clustering [J]. Lecture Notes in Computer Science, 2005, 3776: 768-773