

基于自适应的粗糙 C -均值聚类算法

严静静,张腾飞

(南京邮电大学 自动化学院,江苏 南京 210023)

摘要:粗糙 C -均值的提出,首次将粗糙集与聚类算法结合起来。随后,众多学者对其进行了广泛研究。然而,绝大多数算法在研究簇的下近似、边界对象时,使用统一的权重,忽略了这些对象本身的差异性以及对所在簇的贡献。针对此问题,文中提出一种改进的聚类方法。通过样本对象偏移其所在簇心的程度,设定不同的簇偏移量,距离簇心越近的样本对象其簇偏移量越大,反之越小。通过此举以客观描述这些样本对象对其所在簇的贡献,使得最终聚类结果更加精确、簇内更加紧密、簇间更加稀疏。实例计算结果以及通过 MATLAB 对数据库中 IRIS 的数据集进行仿真验证,表明提出的改进算法具有一定的可行性。

关键词:聚类;粗糙集;粗糙 C -均值;簇偏移量

中图分类号:TP301.6

文献标识码:A

文章编号:1673-629X(2016)03-0067-04

doi:10.3969/j.issn.1673-629X.2016.03.016

Rough C -means Clustering Algorithm Based on Self-adaption

YAN Jing-jing,ZHANG Teng-fei

(College of Automation,Nanjing University of Posts and Telecommunications,Nanjing 210023,China)

Abstract: Rough C -means is proposed to combine the rough set with clustering algorithm first. In the following, many scholars have been doing extensive research. However, for the objects in the low approximation or boundary, the most of algorithms use unified weights, ignoring the difference of the objects themselves and the contribution to the classes. Aiming at this problem, an improved clustering method is put forward. Based on degree of objects deviated centroid of clusters, it sets different offsets to highlight these objects on contribution to the classes in this paper, making the result of clustering more precise, intra-classes more close, and inter-classes more sparse. The experimental results and simulation verification on IRIS by MATLAB shows the method is feasible.

Key words: clustering; rough set; rough C -means; offsets of classes

1 概述

聚类是满足同类对象相似而不同类对象差异的一种数据分类过程。通过优化对象函数,分组 N 个样本对象为 c 个可能的簇,实现簇内对象具有较高的相似性,簇间对象具有较低的相似性^[1]。

传统的聚类技术强制划分数据对象到某个簇中,然而在许多数据挖掘应用中,这种要求过于局限。在某些情况下,数据对象可能属于两个或两个以上的簇。由此可知,类边界严重交叉重叠^[2]。Bezdek 提出模糊聚类如模糊 C 均值(FCM),定义 $0 \sim 1$ 之间的模糊隶属度,使某个对象属于多个簇成为可能。2002 年,Lingras^[3-4]在 Web 数据挖掘时,把粗糙集理论融入到 K -均值算法中,定义三种集合:上近似、下近似、边界。下近似中的对象一定属于所在的簇,上近似中的对象可

能属于所在或者其他的簇,在解决边界重叠问题上具有一定的可行性。然而,Lingras 在计算簇心时,只考虑了边界为空和边界非空的情况。当下近似为空时,算法会出现数值不稳定;计算对象是否属于上下近似时,使用绝对距离公式,忽视了存在的离散数据点。Peters^[5]针对上(下)近似为空时,对相应的下(上)近似权重赋值为 1;考虑到离散点的干扰,用相对距离代替绝对距离,并通过实例,证明其优势。文献[6-7]提出粗糙模糊算法和模糊粗糙算法,通过加入模糊隶属度,每个类有一个模糊下近似和模糊边界,下近似中的对象有明确的权重值,而边界对象有模糊的权重值,有效提高了边界精度。

上述粗糙 C -均值算法在计算簇心时,上、下近似使用统一的权重值,忽视了样本对象内部的差异性。

收稿日期:2015-06-17

修回日期:2015-09-23

网络出版时间:2016-02-18

基金项目:江苏省普通高校研究生科研创新计划项目(46888LX14819)

作者简介:严静静(1990-),女,硕士生,研究方向为模式识别与智能系统;张腾飞,副教授,博士,研究方向为智能信息处理、智能控制等。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20160218.1634.054.html>

根据样本对象对其所在簇的贡献程度,无法使用一致的度量值来衡量,否则必将导致某些点的错误分类。

文中通过计算样本对象偏移簇心的程度,每个对象分别赋予不同的权重值,越是靠近簇心的样本对象其所在簇的权值越大,表明此对象对样本贡献最重。通过实验数据,对比不同的聚类算法,证明了文中提出算法的可行性。

2 粗糙 C-均值聚类算法

2.1 粗糙 C-均值聚类算法的基础

粗糙集在不精确和不完备信息建模下取得了不错的成绩^[7]。这种方法的基本思想是根据某些属性,讨论不可分辨的样本对象分类问题。粗糙 C-均值算法把每个样本对象视为某个区间或者粗糙的集合,是 C-均值算法的扩展。一个粗糙集 $X \in U$, \bar{BU} 、 \underline{BU} 分别为其上近似、下近似集合,它的属性为:

(1) 对某个对象 x_k , 如果 $x_k \in \underline{BU}_i$, 那么 $x_k \notin \underline{BU}_j, i \neq j$;

(2) 如果对象 $x_k \in \underline{BU}_i$, 那么 $x_k \in \bar{BU}_i$;

(3) 如果对象 $x_k \notin \underline{BU}_i$, 那么 $x_k \in \bar{BU}_i, \bar{BU}_j, i \neq j$ 。

粗糙 C-均值聚类算法流程如下:

步骤 1: 初始化参数, 设置聚类个数 c , 任意选取 c 个样本中心 $C_i (i = 1, 2, \dots, c)$, 距离阈值 ε , 权重值 w_l, w_u 。

步骤 2: 计算每个对象 x_k 到簇心 $C_i (i = 1, 2, \dots, c)$ 的欧氏距离。 d_{ik}, d_{jk} 为 x_k 到簇心 c_i, c_j 的欧氏距离。

步骤 3: 选择 d_{ik} 为最小值, d_{jk} 为最大值。如果 $d_{jk} - d_{ik} \leq \varepsilon$, 那么 $x_k \in \bar{BU}_i, x_k \in \bar{BU}_j, x_k$ 不可能属于任何一个簇的下近似(性质 3); 否则, $x_k \in \underline{BU}_i, d_{ik}$ 是 c 个聚类中的最小值(性质 1)。

步骤 4: 更新簇心公式。

$$\vec{m}_k = \begin{cases} w_l \sum_{\vec{x}_n \in \underline{C}_i} \frac{\vec{x}_n}{|\underline{C}_i|} + w_u \sum_{\vec{x}_n \in \underline{C}_i^B} \frac{\vec{x}_n}{|\underline{C}_i^B|}, & \text{如果 } \underline{C}_i \neq \emptyset \cap \underline{C}_i^B \neq \emptyset \\ \sum_{\vec{x}_n \in \underline{C}_i} \frac{\vec{x}_n}{|\underline{C}_i|}, & \text{如果 } \underline{C}_i \neq \emptyset \cap \underline{C}_i^B = \emptyset \\ \sum_{\vec{x}_n \in \underline{C}_i^B} \frac{\vec{x}_n}{|\underline{C}_i^B|}, & \text{如果 } \underline{C}_i = \emptyset \cap \underline{C}_i^B \neq \emptyset \end{cases} \quad (1)$$

步骤 5: 重复步骤 2~4, 直到没有新的划分对象。

2.2 粗糙 C-均值聚类算法的改进

文献[7]将模糊集的模糊隶属度加入到 C-均值

算法可以处理重叠问题, 将粗糙集的上、下近似加入到 C-均值算法可以定义不确定的、模糊的、不完备的类, 改进后的算法能高效选取聚类原型。文献[8-13]提出粗糙模糊 C-均值融合算法。该算法通过粗糙集上、下近似的引入改变了模糊 C-均值算法中隶属度函数的分布情况, 修正了模糊隶属度计算公式(2)和簇心的更新公式(3)。

模糊隶属度 u_{ik} 表明对象 x_k 与类 U_i 的关联程度, 隶属度越大表明对象对其类关联程度越高、簇作用越大。

$$u_{ik} = \frac{1}{\sum_{j=1}^c \left(\frac{d_{ik}}{d_{jk}} \right)^{\frac{2}{m-1}}} \quad (2)$$

将模糊隶属度加入粗糙集质心迭代公式:

$$V_i = \begin{cases} w_l \frac{\sum_{x_j \in \underline{BU}_i} u_{ij}^m x_j}{\sum_{x_j \in \underline{BU}_i} u_{ij}^m} + w_u \frac{\sum_{x_j \in \underline{BU}_i - \underline{BU}_i} u_{ij}^m x_j}{\sum_{x_j \in \underline{BU}_i - \underline{BU}_i} u_{ij}^m}, & \bar{BU}_i - \underline{BU}_i \neq \emptyset, \underline{BU}_i \neq \emptyset \\ \frac{\sum_{x_j \in \underline{BU}_i} u_{ij}^m x_j}{\sum_{x_j \in \underline{BU}_i} u_{ij}^m}, & \bar{BU}_i - \underline{BU}_i = \emptyset, \underline{BU}_i \neq \emptyset \\ \frac{\sum_{x_j \in \underline{BU}_i - \underline{BU}_i} u_{ij}^m x_j}{\sum_{x_j \in \underline{BU}_i - \underline{BU}_i} u_{ij}^m}, & \bar{BU}_i - \underline{BU}_i \neq \emptyset, \underline{BU}_i = \emptyset \end{cases} \quad (3)$$

文献[6,8]将下近似中的对象的权重值设为 1, 认为下近似中的对象肯定是属于所在的类, 对其类关联程度最高, 簇心迭代公式更新为:

$$V_i = \begin{cases} w_l \frac{\sum_{x_j \in \underline{BU}_i} x_j}{\sum_{x_j \in \underline{BU}_i} |\underline{BU}_i|} + w_u \frac{\sum_{x_j \in \underline{BU}_i - \underline{BU}_i} u_{ij}^m x_j}{\sum_{x_j \in \underline{BU}_i - \underline{BU}_i} u_{ij}^m}, & \bar{BU}_i - \underline{BU}_i \neq \emptyset, \underline{BU}_i \neq \emptyset \\ \frac{\sum_{x_j \in \underline{BU}_i} x_j}{\sum_{x_j \in \underline{BU}_i} |\underline{BU}_i|}, & \bar{BU}_i - \underline{BU}_i = \emptyset, \underline{BU}_i \neq \emptyset \\ \frac{\sum_{x_j \in \underline{BU}_i - \underline{BU}_i} u_{ij}^m x_j}{\sum_{x_j \in \underline{BU}_i - \underline{BU}_i} u_{ij}^m}, & \bar{BU}_i - \underline{BU}_i \neq \emptyset, \underline{BU}_i = \emptyset \end{cases} \quad (4)$$

3 基于自适应的粗糙 K-均值聚类算法

3.1 对象偏移质心程度的偏移量

文中提出的算法(IMP-RCM)基于对象偏移簇心

程度的偏移量,在进行公式迭代时,每个样本对象的权重值设为:

$$A_{ij} = \frac{1}{\sigma_{ij}^2 \sum_{j=1}^{m_i} \frac{1}{\sigma_{ij}^2}}$$

(5)

约束条件:

$$\sum_{j=1}^{m_i} A_{ij} = 1$$

其中: σ_{ij} 为样本对象到簇心的标准差; m_i 为类 i 中样本数。

从式中可以看出类内对象离簇心越近, σ_{ij} 值越小,其对所在类的贡献越大,分配的权重值越大。反之,如果偏移度越大, σ_{ij} 样本对象到簇心距离越远,对其所在类的贡献越低,相应的样本对象获得的权重值越低。

对于 w_l 与 w_u 值的公式,文中不再使用默认参数值,而是参照样本中上、下近似对象的个数, w_l (w_u) 是样本下(上)近似对象的个数比样本数。

$$w_l = \frac{\sum_{x_j \in \underline{BU}} 1}{\sum_{x_j \in \underline{BU}} 1 + \sum_{x_j \in \underline{BU}} 1}, w_u = \frac{\sum_{x_j \in \underline{BU}} 1}{\sum_{x_j \in \underline{BU}} 1 + \sum_{x_j \in \underline{BU}} 1}$$

(6)

3.2 基于自适应的粗糙 C-均值聚类算法

文中提出的算法(IMP-RCM)是在式(4)的基础上进行改进。式(4)中簇内样本对象权重是所有样本总数的倒数,忽视样本个体对类的贡献差异。文中将式(5)引入到簇心迭代公式中。同时式(4)中 w_l 、 w_u 受初始参数影响较大,提出式(6),考虑上、下近似样本的对象个数。

步骤 1:初始化参数,设置聚类个数 c ,任意选取 c 个样本中心 $C_i(i = 1, 2, \dots, c)$,距离阈值 ε 。

步骤 2:计算权重值 w_l 、 w_u ,以及每个对象 x_k 到样本中心 $C_i(i = 1, 2, \dots, c)$ 的欧氏距离。 d_{ik} 、 d_{jk} 为 x_k 到质心 c_i 、 c_j 的欧氏距离。

步骤 3:选择 d_{ik} 为最小值, d_{jk} 为最大值。如果 $d_{jk} - d_{ik} \leq \varepsilon$,那么 $x_k \in \underline{BU}_i$, $x_k \in \underline{BU}_j$, x_k 不可能属于任何一个类的下近似(性质 3);否则, $x_k \in \underline{BU}_i$, d_{ik} 是 c 个聚类中最小值(性质 1)。

步骤 4:更新簇心公式。

$$V_i = \begin{cases} w_l \sum_{x_j \in \underline{BU}_i} A_{ij} x_j + w_u \sum_{x_j \in \underline{BU}_i - \underline{BU}_i} A_{ij} x_j, \underline{BU}_i - \underline{BU}_i \neq \emptyset, \underline{BU}_i \neq \emptyset \\ \sum_{x_j \in \underline{BU}_i} A_{ij} x_j, \underline{BU}_i - \underline{BU}_i = \emptyset, \underline{BU}_i \neq \emptyset \\ \sum_{x_j \in \underline{BU}_i - \underline{BU}_i} A_{ij} x_j, \underline{BU}_i - \underline{BU}_i \neq \emptyset, \underline{BU}_i = \emptyset \end{cases}$$

(7)

步骤 5:重复步骤 2~4,直到没有新的划分对象。
在步骤 4 中,不再是传统意义上对象权重值都是固定值,在每次迭代过程中,簇心越来越收敛,趋于稳定。

4 实验仿真分析

为了验证算法的处理效果,文中对 UCI 数据库中的 IRIS 数据集进行 MATLAB 仿真分析。IRIS 数据特征见表 1。

表 1 IRIS 数据集特征

数据名称	分类个数	数据个数	特征数
IRIS	3	150	4

先通过主成分分析法将 IRIS 数据集进行降维处理,随后采用 HCM 算法、粗糙 RCM 算法、文中改进的 RCM 算法对降维后的数据集进行聚类分析。聚类分布图分别为图 1~3 所示。其中, ∇ 、 Δ 、 \bigcirc 分别表示 IRIS 的三个类对象,即簇 1、簇 2、簇 3;黑圆 \times 表示簇心。簇 1 相距簇 2、簇 3 较远,而且簇 2 与簇 3 边界交叉重叠。

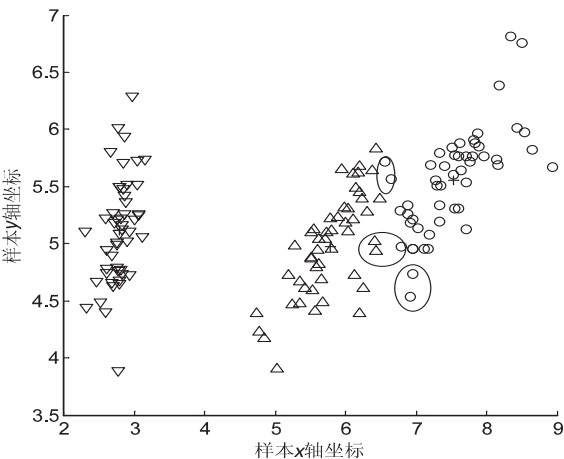


图 1 HCM 算法 IRIS 数据点分布

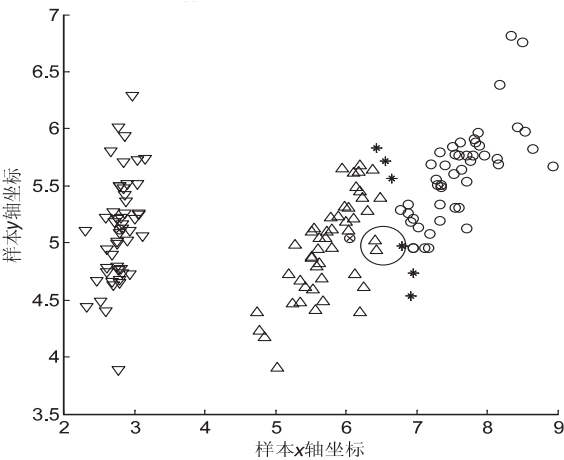


图 2 RCM 算法 IRIS 数据点分布

HCM 算法强制认为非此即彼的划分,没有考虑边

界严重交织、模糊的划分。图 2 标出有些点属于某个簇是不科学的;粗糙 RCM 的权重 $w_l = 0.7$, $w_u = 0.3$ 。文中改进的权重值属于自适应值,从图 2 看出,出现划分的错误点,而改进算法得到的簇心(见图 3)较图 2 更科学,与所在簇更紧密。

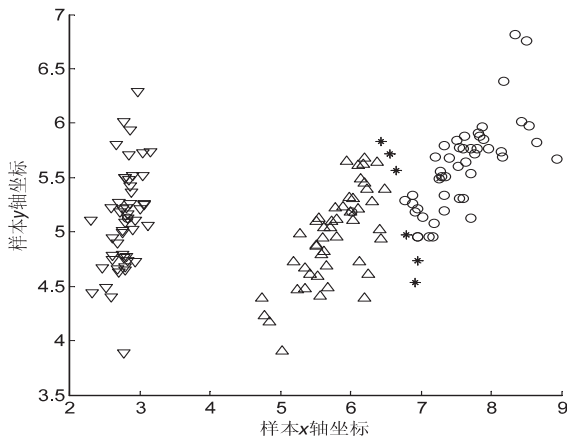


图 3 文中改进的 RCM 算法 IRIS 数据点分布

5 结束语

文中主要针对绝大多数算法在研究簇的下近似、边界对象时,使用统一的权重,忽略了这些对象本身的差异性以及对所在簇的贡献差异,提出一种新的改进方法。通过对样本对象偏移其所在簇心的程度,设定不同的簇偏移量,以客观描述这些样本对象对其所在簇的贡献,使得最终聚类结果更加精确,簇内更加紧密、簇间更加稀疏。通过 MATLAB 仿真,对比以往的聚类,表明基于自适应的粗糙 C-均值聚类算法的聚类效果更优。

参考文献:

- [1] Mitra S, Banka H, Pedrycz W. Rough-fuzzy collaborative clustering[J]. IEEE Trans on Systems, Man and Cybernetics, Part B: Cybernetics, 2006, 36(4): 795-805.
- [2] Lingras P, Peters G. Rough clustering[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2011, 1

(1): 64-72.

- [3] 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.
- [4] Pawan L, Rui Y, Chad W. Comparison of conventional and rough k -means clustering[C]//Proc of 9th international conference on rough sets, fuzzy sets, data mining, and granular computing. Berlin: Springer, 2003: 130-137.
- [5] Georg P. Some refinements of rough k -means clustering[J]. Pattern Recognition, 2006, 39(8): 1481-1491.
- [6] Hu Qinghua, Yu Daren. An improved clustering algorithm for information granulation[C]//Proc of 2nd international conference on FSKD. Berlin: Springer, 2005: 494-504.
- [7] Maji P, Pal S K. RFCM: A hybrid clustering algorithm using rough and fuzzy sets[J]. Fundamenta Informaticae, 2007, 80(4): 475-496.
- [8] 王丹, 吴孟达. 粗糙模糊 C 均值融合聚类[J]. 国防科技大学学报, 2011, 33(3): 145-150.
- [9] Maji P, Paul S. Robust rough-fuzzy c-means algorithm: design and applications in coding and non-coding RNA expression data clustering[J]. Fundamenta Informaticae, 2013, 124: 153-174.
- [10] Lai J Z C, Juan E Y T, Lai F J C. Rough clustering using generalized fuzzy clustering algorithm[J]. Pattern Recognition, 2013, 46(9): 2538-2547.
- [11] Peters G, Crespo F, Lingras P, et al. Soft clustering - fuzzy and rough approaches and their extensions and derivatives[J]. International Journal of Approximate Reasoning, 2013, 54(2): 307-322.
- [12] Scitovski R, Sabo K. Analysis of the k -means algorithm in the case of data points occurring on the border of two or more clusters[J]. Knowledge-based Systems, 2014, 57(2): 1-7.
- [13] Velmurugan T. Performance based analysis between k -means and fuzzy c-means clustering algorithms for connection oriented telecommunication data[J]. Applied Soft Computing, 2014, 19(6): 134-146.