

多属性加权的模糊 c 均值聚类算法

李诗瑾, 李 倩, 徐桂琼
(上海大学 管理学院, 上海 200444)

摘 要: 在多属性模糊 c 均值聚类的基础上, 提出了一种基于属性重要性加权的聚类算法. 为验证新算法的有效性, 在 6 个 UCI 数据集上, 将新算法与结合主成分分析法和基于粗糙集指数加权的聚类方法进行了比较分析. 实验结果表明, 基于属性重要性的聚类算法具有更好的泛用性和稳定性, 且随着平均类间簇心距离的增大而提升聚类有效性.

关键词: 数据挖掘; 模糊 C 均值聚类; 属性重要性

中图分类号: TP391

文献标识码: A

文章编号: 1000—7180(2017)07—0092—04

Fuzzy C-means Clustering Based on the Multivariable Weighted Attributes

LI Shi-jin, LI Qian, XU Gui-qiong

(School of Management, Shanghai University, Shanghai 200444, China)

Abstract: Considering the difference of feature attributes, a new clustering algorithm based on the importance of attributes is proposed according to multivariable fuzzy c-means clustering method. In order to verify the validity of new method, the proposed algorithm is analyzed and compared with principal component comprehensive analysis and attributes weighted by rough set theory on six UCI datasets. The experimental results showed that this method has wider scope of application and stability. Moreover, the clustering efficiency increase when datasets have further average distance among cluster centers.

Key words: data mining; fuzzy c-means algorithm; attributes importance

1 引言

聚类是数据挖掘的一个重要组成部分, 模糊 c 均值聚类(FCM)最早由 Bexdek 提出^[1], 是聚类分析中最流行的算法之一. 在 FCM 算法中, 样本隶属于某一簇的程度是通过欧式距离度量样本与原型之间的相似性, 从而计算隶属度来反映的. 经过长期深入的研究, 许多学者都提出了对 FCM 的改进算法^[2]. 没有考虑不同属性对类别划分的贡献程度是不同的. 因此特征属性约简^[3]或属性赋权对模糊聚类算法有重要意义.

近年来, 许多作者针对聚类算法中属性加权问题展开了研究: 文献[4]提出了在 FCM 算法中考虑样本各个维度对模式分类的不同影响, 提出一种优

化特征加权的模糊聚类算法; Nazari 等^[5]根据 FCM 的划分结果, 提出特征权值随着每次迭代自动地计算; 李振东提出了一种基于粗糙集指数加权的聚类算法, 以减少包含共性信息的属性对聚类的影响^[6].

最近, Pimentel 和 Souza 提出了一种多属性 FCM 算法(MFCM), 它从数据集中提取了更多属性用于聚类的信息, 提高了聚类质量^[7]. 受该研究工作的启发, 本文在 MFCM 算法的基础上, 提出了一种基于多属性重要性加权的模糊 c 均值聚类算法(WAFCM).

2 模糊 C 均值聚类

FCM 的基本思想是将 n 个样本 $X = \{x_1, x_2, \dots, x_n\}$ 分为 c 个簇, 簇心用 $V = \{v_1, v_2, \dots, v_c\}$ 表示. 通过建立表示样本数据点与聚类簇心之间加权

相似性测度的目标函数,并对其进行迭代最小化,最终确定最佳的聚类结果.聚类结果通过模糊隶属度矩阵 $U = [u_{ik}] \in R, i = 1, 2, \dots, c; k = 1, 2, \dots, n$ 表示. FCM 算法的目标函数 J 可由式(1)给出^[8].

$$\min J_1(U, V) = \sum_{i=1}^c \sum_{k=1}^n u_{ik}^m (||x_k - v_i||^2) \quad (1)$$

式中, u_{ik} 表示样本点 x_k 属于第 i 个类的隶属程度,它反映了样本点与簇的相似程度,若接近 1,表示属于此类的程度高;若接近 0,表示属于此类的程度低. m 代表加权指数(模糊指标), $1 < m < +\infty$. 隶属度矩阵中的元素满足如下 3 条限制:

- (1) $u_{ik} \in [0, 1], i = 1, 2, \dots, c; k = 1, 2, \dots, n;$
- (2) $0 < \sum_{k=1}^n u_{ik} < n, i = 1, 2, \dots, c;$
- (3) $\sum_{i=1}^c u_{ik} = 1, k = 1, 2, \dots, n.$

在满足条件(3)的前提下,根据 Lagrange 乘子法,可以得到目标函数取得极小值的必要条件为

$$u_{ik} = [\sum_{j=1}^c (\frac{||x_k - v_i||}{||x_k - v_j||})^{\frac{2}{m-1}}]^{-1}, i \neq j \quad (2)$$

$$v_i = \frac{\sum_{k=1}^n (u_{ik})^m x_k}{\sum_{k=1}^n (u_{ik})^m} \quad (3)$$

FCM 算法的具体步骤如下

输入:数据集、聚类个数 c 、模糊度 m 、迭代阈值 ϵ 、最大迭代次数 T

输出: U, V

步骤 1:随机初始化聚类中心 V ,令 $t = 0$.

步骤 2:根据式(2)更新隶属度矩阵 U ,代入式

(3)更新聚类簇心 V .

步骤 3:若 $t = T$ 或 $||v_{i,t} - v_{i,t-1}|| \leq \epsilon$,则停止,否则重复步骤 2,并且置 $t = t + 1$.

3 基于多属性加权的 FCM 聚类算法

3.1 属性重要性

FCM 算法以样本点与聚类中心之间的欧式距离作为相似性指标,默认每一个属性对聚类作出的贡献相等.图 1 包含 10 个样本点,该图直观地表现出不同属性对于聚类的贡献程度不同:假设样本 1~5 属于第一个簇,样本 6~10 属于第二个簇.通过 FCM 算法得样本 5 隶属于簇 2 的程度却要高于簇 1.分别考虑属性 x 和 y 可以发现,当仅考虑属性 x 时,样本 5 属于簇 1 的程度高;当仅考虑属性 y 时,样本 5 属于簇 2 的程度高.由此可见,不同属性对于聚类有着不同的贡献程度,样本的每个特征属性并不都对聚类结果起决定性作用.

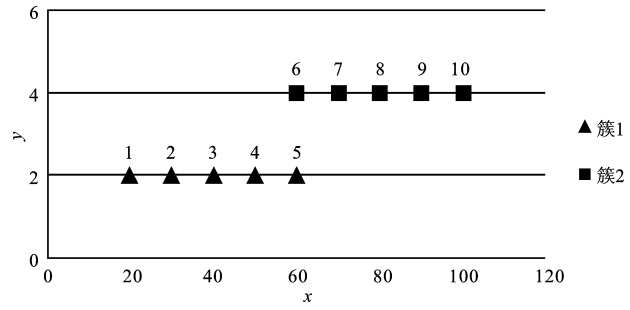


图 1 不同属对聚类的贡献比较

3.2 MFCCM 算法

Pimentel 和 Souza 考虑到 FCM 算法在计算隶属度时,默认样本的每个属性提供的信息相同.因此需要进一步细化数据样本基于不同属性的分簇情况.假设数据集 X 中的每个样本包含 p 个特征属性,样本点可记作 $x_k = (x_{1k}, x_{2k}, \dots, x_{pk}), k \in \{1, 2, \dots, n\}$,簇心记为 $v_i = (v_{i1}, v_{i2}, \dots, v_{ip}), i \in \{1, 2, \dots, c\}$.目标函数 J 则调整为

$$\min J_2(U, V) = \sum_{i=1}^c \sum_{k=1}^n \sum_{j=1}^p u_{ijk}^m (||x_{jk} - v_{ij}||^2) \quad (4)$$

此时的目标函数考虑到每一个属性的累计贡献,通过基于属性 j 计算样本 x_k 与簇心 v_i 的距离.考虑到不同属性对于聚类的贡献程度,基于每一个属性计算对象的隶属度:

$$u_{ijk} = [\sum_{h=1}^c \sum_{l=1}^p (\frac{||x_{jk} - v_{ij}||}{||x_{jk} - v_{hl}||})^{\frac{2}{m-1}}]^{-1}, i \neq h, j \neq l \quad (5)$$

式中, u_{ijk} 代表样本 x_k 在属性 j 上隶属于簇 v_i 的程度,需要满足以下限制条件:

- (1) $u_{ijk} \in [0, 1], i = 1, 2, \dots, c; k = 1, \dots, n, j = 1, \dots, p;$
- (2) $0 < \sum_{j=1}^p \sum_{k=1}^n u_{ijk} < n, i = 1, 2, \dots, c;$
- (3) $\sum_{i=1}^c \sum_{j=1}^p u_{ijk} = 1, k = 1, 2, \dots, n.$

此时,簇心可由式(6)计算得出:

$$v_{ij} = \frac{\sum_{k=1}^n (u_{ijk})^m x_{kj}}{\sum_{k=1}^n (u_{ijk})^m} \quad (6)$$

3.3 基于多属性加权的 FCM 聚类算法

MFCCM 算法虽然计算了样本在不同属性下的聚类结果,但未根据不同属性对聚类结果的影响程度进行赋权,影响了聚类的效率和精度.实际上,包含更多个性信息的属性在聚类中起到的作用更大.由此,本文在 MFCCM 的基础上提出一种根据多属性

重要性加权的聚类算法(WAFCM). 计算每一个样本的不同属性所携带的信息对整体聚类的重要性, 根据其贡献程度对每一个属性自主赋值加权 $\mathbf{W}$, $\mathbf{W}$ 是一个 $p \times p$ 的对角阵, 若属性对样本划分起到积极作用, 则赋予较大的权重, 反之则赋予较小的权重. 目标函数修改为式(7):

$$\min J_3 = \sum_{i=1}^c \sum_{k=1}^n \sum_{j=1}^p u_{ijk}^m ||\omega_j x_{jk} - v_{ij}||^2 \quad (7)$$

当目标函数 J_3 达到最小值时, 就得到了最优的聚类结果. 属性赋权依据每个属性对数据集划分的贡献程度来计算, 属性 j 的赋权如式(8)所示.

$$\omega_j = \frac{\sum_{k=1}^n \sum_{i=1}^c u_{ijk}}{n} \quad (8)$$

相应的簇心更改为式(9):

$$v_{ij} = \frac{\sum_{k=1}^n (u_{ijk})^m x_{kj} \cdot \omega_j}{\sum_{k=1}^n (u_{ijk})^m} \quad (9)$$

固定权重 ω_j 和 v_{ij} , 由式(10) 调整 u_{ijk} .

$$u_{ijk} = [\sum_{h=1}^c \sum_{l=1}^p (\frac{||\omega_j \cdot x_{jk} - v_{ij}||}{||\omega_j \cdot x_{jk} - v_{hl}||})^{\frac{2}{m-1}}]^{-1} \quad (10)$$

WAFCM 算法的具体步骤如下.

输入: 数据集 X 、聚类数量 c 、模糊度参数 m 、迭代阈值 ϵ 、最大迭代次数 T

输出: U 、 V 、 W

步骤 1: 随机初始化隶属度矩阵 $U = [u_{ijk}]$, 满足限制条件 $\sum_{i=1}^c \sum_{j=1}^p u_{ijk} = 1, u_{ijk} \in [0, 1]$ 设置初始属性权重, 假设每个属性重要性相同, 为 $\frac{1}{p}$.

步骤 2: 固定 u_{ijk} 和 ω_j , 根据式(8) 更新原型矩阵 $\mathbf{V} = [v_{ij}]$.

步骤 3: 固定 v_{ij} 和 ω_j , 根据式(9) 更新隶属度矩阵 $\mathbf{U} = [u_{ijk}]$.

步骤 4: 固定 u_{ijk}, v_{ij} , 根据式(10) 更新属性权重 $\mathbf{W} = [\omega_1, \omega_2, \dots, \omega_p]$.

步骤 5: 当 $t = T$ 或 $||J_t - J_{t-1}|| \leq \epsilon$ 时执行步骤 6; 否则重复步骤 2 ~ 4, 并且置 $t = t + 1$.

步骤 6: 累加计算样本基于每个加权属性的隶属度, 得到样本的隶属度用于分簇, 如式(11) 所示.

$$u_{ik} = \sum_{j=1}^p u_{ijk} \quad (11)$$

WAFCM 算法实则是根据 FCM 中的样本不同属性对聚类的贡献程度, 自动计算并迭代更新属性的权重. 本文的 WAFCM 算法选取 RI 指数^[9] 作为有效性指标, 比较计算经过属性加权后的数据集聚

类结果的 RI 值, 具体 WAFCM 方法可参照 4.2 节中的实验, 以数据集 wine 为例介绍属性赋权的过程.

4 实验与结果分析

4.1 实验数据集

为测试算法的有效性和泛用性, 本文选取 UCI 公开数据库中的六个带标签的具有连续属性的数据集即 wine, red wine quality, iris, seeds, pima Indians diabetes, abalone. 数据集分簇后的平均类间中心距离可以表示数据集的分布特点. 六个数据集及聚类特征分布可由表 1 所示.

表 1 实验用数据集

数据集	样本数目	属性个数	类别个数	平均类间中心距离
Iris	150	4	3	0.40
Wine	178	13	3	17.28
Red wine quality	1 599	11	12	21.84
Seeds	210	7	3	0.52
Pima	768	8	2	41.99
Abalone	4 177	8	3	1.87

4.2 确定属性权重实验

实验通过参考 FCM 并使用 Matlab 软件编写 WAFCM 算法, 这里以 wine 数据集为例说明 WAFCM 的聚类算法, 表 2 对 wine 各个属性值对聚类的贡献大小进行了排序, Nonflavanoid phenols, Ash 和 Hue 这三个属性的总贡献度最高, 则通过自主赋权的迭代过程后计算得到的权重也最高, 通过式(8) 计算属性的权重, 当目标函数到达最小值得到 $\mathbf{W}_{\text{wine}}$, 最高的三个属性的权重分别是 20.57%, 12.17%, 11.45%. 利用得到的属性权重, 重新计算隶属度和簇心, 直到使得目标函数达到最小. 本文利用 RI 指数作为聚类有效性指标.

表 2 wine 各个属性对聚类的贡献度

属性	簇 1	簇 2	簇 3	权重/%
1Nonflavanoid phenols	11.29	12.66	12.67	20.57
2 Ash	7.03	6.86	7.78	12.17
3 Hue	7.73	8.32	4.33	11.45
4Alcalinity of ash	9.01	1.01	7.01	9.57
5 Magnesium	5.00	9.01	3.00	9.56
6 Total phenols	3.11	5.68	6.46	8.57
7Proanthocyanins	3.99	3.11	4.94	6.76
8 OD280/OD315 of diluted wines	2.17	4.51	2.09	4.93
9 Alcohol	1.14	2.22	5.30	4.87
10 Color intensity	3.03	2.13	1.64	3.82
11Proline	1.00	4.00	1.00	3.37
12Flavanoids	2.14	1.10	1.11	2.45
13 Malic acid	1.16	1.19	1.08	1.92

4.3 不同聚类算法的对比实验

本文运用 WAFCM 算法与 MFCM,结合主成分分析法和粗糙集指数加权的聚类算法,这两种属性加权的聚类算法进行了比较分析。其中,主成分分析法是一种较为成熟的多元统计分析方法,基于特征值或因子个数提取综合成分;粗糙集指数加权法应用于聚类前的数据处理阶段,减少了包含过多共性信息的属性对聚类的影响。

然而 PCA 分析法与粗糙集指数加权的聚类算法的都有其自身的缺陷,无法绝对客观的对数据集属性进行加权。例如在 Iris、Red wine quality 数据集中,KMO 统计量分别为 0.536,0.431,低于标准值 0.6,不适合作 PCA;在属性值为连续的数据集中,在对属性值离散化的过程中,无法避免地造成数据信息的丢失。WAFCM 方法能够通过计算每个属性在聚类时的贡献度,客观地得到每个属性的权重,不受到方法自身缺陷的约束。

对 MFCM 算法和 PCA 分析法、粗糙集指数加权法及本文利用属性贡献度的加权聚类方法对六个数据集进行聚类,通过 Rand 指数比较聚类结果的有效性,结果如图 2 所示。实验结果表明,对于每一个数据集,相比于其他三种聚类算法,基于 WAFCM 算法的聚类效果均有一定的提升。

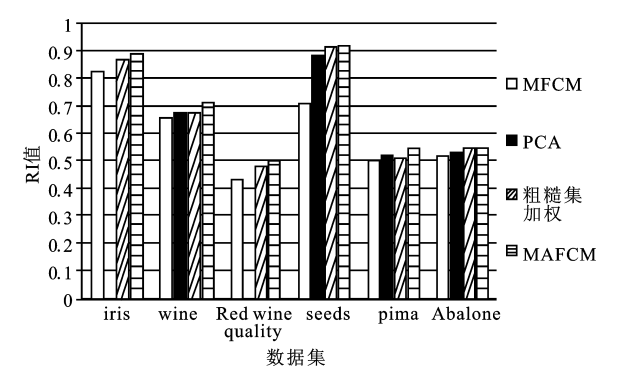


图 2 四种算法聚类有效性比较

4.4 实验结果分析

由于 PCA 分析法、粗糙集指数加权的聚类算法和 WAFCM 聚类算法都是在模糊 c 均值聚类算法基础上基于属性加权的算法改进,故计算这三种算法相较于 FCM 算法在聚类有效性上的提升程度,用 RI 提升率来显示三种方法的改进程度。

从图 3 可以看出,WAFCM 不仅相较于 PCA 分析法、粗糙集指数加权的聚类算法的聚类算法的结果要好,且数据集类间平均中心距离与 WAFCM 算法得到的聚类有效性结果呈正相关,即 RI 提升

率随着数据集簇间平均中心距离的增大而提高,而其他两种算法却是随着 RI 提升率的上升而不断地波动。因此当簇心之间的平均距离较大的时候可以利用本文的算法进行聚类。

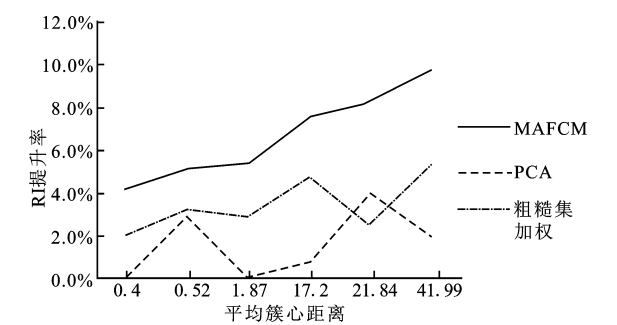


图 3 RI 提升率与平均类间中心距离之间的关系

虽然 PCA 分析法、粗糙集指数加权的聚类算法的聚类算法能对某些数据集进行属性综合加权,但两者均有其局限性。PCA 分析法在当其 KMO 值小于 0.6 时,意味着样本间属性的相关性较弱,数据集不适合作 PCA 分析。同样的,基于粗糙集指数加权的聚类算法的聚类算法只局限于属性值为少数离散值的情况,较难在不失去信息的情况下,通过对连续属性值的离散化处理数据集以扩展该算法。WAFCM 与其他两种算法相比,虽然在聚类有效性上的结果提升程度较小,但是仍然在一定程度上优于其他两种属性加权方法,并且本文的算法无使用的限制条件,更适用于无标签带有连续属性的数据集。

5 结束语

数据挖掘中,FCM 算法是一种经典的聚类分析方法,广泛的应用于诸多领域。为了克服传统 FCM 算法的局限性,本文在多属性模糊 c 均值聚类的基础上,提出了一种基于多属性的模糊 c 均值聚类算法。为验证有效性,在 UCI 的六个数据集上,将新算法与 PCA 分析法、粗糙集指数加权的聚类算法的聚类方法进行了比较分析。实验结果表明,该方法具有更好的泛用性和稳定性。特别是,当数据集的平均类间距离较大的情况下,聚类效果随着类间距离的增大而增大。

参考文献:

[1] Bezdek J C. Pattern recognition with fuzzy objective function algorithms[M]. Plenum Press, 1981.
[2] 肖林云, 陈秀宏, 林喜兰. 特征加权和优化划分的模糊 C 均值聚类算法[J]. 微电子学与计算机, 2016,33 (10):143-146.

配点对数增加 15%以上,准确率相当.

表 2 算法实验结果对比

性能		亮度	缩放	旋转	透视
时间/s	SIFT	4.235	4.398	4.412	4.217
	SURF	1.549	1.621	1.635	1.596
	本文算法	1.386	1.429	1.420	1.387
匹配点对	SIFT	142	193	154	206
	SURF	145	188	157	217
	本文算法	184	219	187	231
准确率/%	SIFT	98.6	98.9	98.7	97.1
	SURF	99.6	98.9	100	100
	本文算法	99.4	99.5	100	100

6 结束语

针对传统 SURF 算法匹配精度低,速度慢的问题,提出一种基于 BBF 搜索策略的改进 SURF 算法.算法利用 Harr-like 特征丰富特征描述符,使图像匹配点对数明显增加,提高了算法描述能力和鉴别力;同时结合一种改进的搜索策略——BBF 算法,在一定程度上提高了算法速度.实验表明,改进的算法在图像亮度变化、缩放、旋转、透视等情况下均具有较好的鲁棒性,因此,该研究为印刷品缺陷检测技术提供了更好的支持.

参考文献:

[1] 李俊平. 数字印刷品的质量控制[J]. 印刷工业, 2016, 11(3):64-65.

[2] 张格森. 一种图像局部特征快速匹配算法[J]. 计算机技术与应用, 2015, 41(11): 124-127.

[3] Lowe D G. Distinctive image features from scale-invariant keypoints[J]. International Journal of Computer

Vision, 2004, 60(2):91-110.

[4] BAY H, TUYTELAARS T, Gool V L, et al. Speeded-up robust features(SURF)[J]. Computer Vision and Image Understanding, 2008, 110(3):346-359.

[5] 尚明姝. 一种基于改进 SURF 的图像配准方法[J]. 微电子学与计算机, 2014, 31(2):125-128.

[6] 王文举,赵萍,陈伟,等. 彩色印刷品缺陷快速精确检测方法研究[J]. 包装工程, 2015, 36(17): 112-130.

[7] Manish I Patel, Vishvjit K Thakar, Shishir K Shah. Image registration of satellite images with varying Illumination level using HOG descriptor based SURF[J]. Procedia Computer Science, 2016, 93:382-388.

[8] Zhang Chunjie, Jing Liu, Chao Liang, et al. Image classification using Harr-like transformation of local features with coding residuals[J]. Signal Processing, 2013, 93(8): 2111-2118.

[9] Nikolett Bereczky, Amalia Duch, Krisztián Németh, et al. Quad-kd trees: A general framework for kd trees and quad trees [J]. Theoretical Computer Science, 2016, 616:126-140.

作者简介:

张培男, (1993-), 硕士研究生. 研究方向为计算机图形学、图像匹配.

郭慧(通讯作者)女, (1964-), 博士, 副教授. 研究方向为计算机图形学、逆向工程. E-mail:ghcad@163.com.

刘芳辉男, (1991-), 硕士. 研究方向图像检索及图像处理.

胡方尚男, (1992-), 硕士. 研究方向计算机图形学.

周绍萍女, (1966-), 博士, 教授. 研究方向状态监测与故障诊断.

(上接第 95 页)

[3] Wang J. Geometric structure of high-dimensional data and dimensionality reduction[M]. Springer Berlin Heidelberg, 2011.

[4] 邢婷, 邢治国, 王凤领. 基于信息熵的 FCM 聚类算法[J]. 计算机工程与设计, 2010, 31(23):5092-5095.

[5] Nazari M, Shanbehzadeh J, Sarrafzadeh A. Fuzzy C-means based on automated variable feature weighting [J]. Lecture Notes in Engineering & Computer Science, 2013, 2202(1):25-29.

[6] 李振东, 李飞. 基于粗糙集指数加权的聚类算法[J]. 统计与决策, 2015, 31(2):83-85.

[7] Pimentel B A, Souza R M C R. A multiWAFCMate fuzzy c-means method[J]. Applied Soft Computing, 2013, 13(4):1592-1607.

[8] Kannan S R, Ramathilagam S, Chung P C. Effective fuzzy c-means clustering algorithms for data clustering problems [J]. Expert Systems with Applications, 2012, 39(7):6292-6300.

[9] 杨燕, 靳蕃, KAMEL Mohamed. 聚类有效性评价综述[J]. 计算机应用研究, 2008, 25(6):1630-1633.

作者简介:

李诗瑾女, (1993-), 硕士研究生. 研究方向为数据挖掘、模糊聚类. E-mail:lilyli@i.shu.edu.cn.

李倩女, (1993-), 硕士研究生. 研究方向为数据挖掘、协同过滤.

徐桂琼女, (1973-), 博士生导师. 研究方向为算法分析与设计、数据挖掘与人工智能.