计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(3):213−217 [doi: 10.15888/j.cnki.csa.007342]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 PCA 优化的 PSO-FCM 聚类算法①

陈　诚, 刘振宇

(南华大学 计算机科学与技术学院, 衡阳 421001)
通讯作者: 陈　诚, E-mail: 610396646@qq.com

摘　要: 为解决 PSO-FCM 聚类算法针对多聚类问题, 性能不足, 容易陷入局部最优解, 影响多聚类结果的准确度. 提出一种基于 PCA 优化的 PSO-FCM 聚类算法, 通过引入 PCA 分析方法, 在粒子的各维度上设定不同的移动权重, 降低粒子的敏感度, 合理的控制粒子各维度上移动的速度, 有效的降低粒子各维度上粒子无约束, 位于多个聚类群交界处的粒子过分敏感, 移动到错误的聚类的可能性增加. 本文简要介绍了 PSO-FCM 算法的相关情况, 详细介绍了本文的优化算法, 最后通过实验证明, 本文提出的优化算法在多个数据集上结果总体优于其他算法.

关键词: 模糊聚类; 主成分分析; 粒子群模糊聚类

引用格式: 陈诚,刘振宇.基于 PCA 优化的 PSO-FCM 聚类算法.计算机系统应用,2020,29(3):213–217. http://www.c-s-a.org.cn/1003-3254/7342.html

Optimized PSO-FCM Cluster Algorithm Based on Principal Component Analysis

CHEN Cheng, LIU Zhen-Yu

(School of Computer Science and Technology, University of South China, Hengyang 421001, China)

Abstract: For multi-cluster problems, PSO-FCM cluster algorithm is lack of performance and easily leads to local optimum, which affects the accuracy of multi-cluster result. To tackle these issues, an optimized PSO-FCM cluster algorithm based on PCA is put forward. By introducing PCA processing method, setting different movement weight on each dimension of particle and reducing particle sensitivity, reasonably controlling movement speed of particles on each dimension and effectively decreasing unconstrained particles on each dimension, possibility of moving into false cluster is increased due to over-sensitive particles on interface of multi-cluster groups. This paper introduces related conditions of PSO-FCM algorithm briefly and the proposed optimized algorithm in detail. Finally, this paper presents the experiment results, i.e., the optimized algorithm proposed in this study is totally better than other algorithms in many data sets.

Key words: fuzzy clustering; principal component analysis; particle swarm fuzzy clustering

作为传统聚类算法模糊 C-均值聚类算法 (Fuzzy C- Mean clustering algorithm, FCM) 的一种优化算法, 引入了粒子群优化算法 (Particle Swam Optimization, PSO), 粒子群模糊聚类算法 (Particle Swarm-based Fuzzy Clustering algorithms, PSO-FCM), 通过 PSO 算法的收敛速度快, 粒子收敛由自身最优位置和群体最优位置相结合, 在一定程度上解决了 FCM 对初始值敏感, 对噪声数据敏感, 容易陷入局部最优解的缺点. 如今, 随着数据量多样化, 复杂化, 多类别化, PSO-FCM 只是单一优化初始聚类中心选取问题, 没有合理

① 基金项目: 国家自然科学基金 (61402220, 61502221); 湖南省哲学社会科学基金 (16YBA323); 湖南省自然科学基金 (2015JJ3015); 湖南省教育厅青年项目 (15B207, 18B279); 南华大学科研创新项目 (193YXC015)
Foundation item: National Natural Science Foundation of China (61402220, 61502221); Philosophic Social Science Fund of Hunan Province (16YBA323); Natural Science Foundation of Hunan Province (2015JJ3015); Youth Program of Education Bureau, Hunan Provinve (15B207, 18B279); Innovative Research Program of University of South China (193YXC015)
收稿时间: 2019-07-29; 修改时间: 2019-09-05, 2019-09-18; 采用时间: 2019-09-29; csa 在线出版时间: 2020-02-28

的限制粒子的移动, 并不能更好优化好 FCM 算法面对多聚类问题时[1–7].

为了解决上述问题, 引入主成分分析 (Principal Component Analysis, PCA), 本文提出基于 PCA 优化的粒子群模糊聚类算法 (PCA-PSO-FCM), 通过 PCA 对数据各维度的分析和评定综合给出一个权重值, 粒子各维度会根据该调整权重速度和方向. 本文详细介绍了 PCA-PSO-FCM, 并且与 FCM 和 PSO-FCM 进行了实验结果的比对, 从实验上来看, 本文的算法在多种群聚类问题上性能更好, 是一种很有潜力的聚类算法.

本文结构如下: 第 1 部分主要对已有的算法的研究成果进行简要分析总结; 第 2 部分对于本文的优化算法进行详细说明; 第 3 部分说明实验过程相关细节, 设定参数以及实验结果的分析; 第 4 部分总结全文.

1 算法介绍

PSO-FCM 算法是模糊均值聚类算法基础上的优化算法, 传统的模糊 C 均值算法的结果精度, 对初始中心的选取有很严格的要求, 并且容易陷入局部最优解. 为了解决这个问题, 国内许多学者, 利用具有集体智能的粒子群优化算法, 与传统模糊 C 均值算法结合. 利用 PSO 算法求解初始聚类中心, 进而优化了 FCM 依赖初始中心的问题; 利用 PSO 算法中, 粒子个体与粒子群体之间关系, 粒子整体移动的速度可以调节, 进而降低了 FCM 容易陷入最优解的可能性.

1.1 PSO-FCM 算法

PSO-FCM 算法是基于数据样本之间的模隶属矩阵建立的聚类算法. 算法的核心思想是: n 个文本样本为$X=(x_1,x_2,\cdots,x_n)$, 划分为$\mathrm{C}=(c_1,c_2,\cdots,c_n)$, p 个聚类中心, 计算出每个文本的隶属度μ_{ij}, μ_{ij}表示第 j 个样本隶属于第 i 个样本的隶属度.

$$\sum_{i=1}^{p}\mu_{ij}=1,\mu_{ij}\in[0,1] \tag{1}$$

$$\mu_{ij}=\begin{cases}1/\sum_{k=1}^{p}\left(d_{ij}/d_{kj}\right)^{\frac{2}{m-1}}, & d_{kj}\neq 0\\ 0, & d_{kj}=0\end{cases} \tag{2}$$

根据每个样本的隶属度值计算出适应度函数值:

$$J_m=\sum_{j=1}^{n}\sum_{i=1}^{p}(\mu_{ij})^m(x_j-c_i)^2 \tag{3}$$

$$c_i=\frac{\sum_{j=1}^{n}(\mu_{ij})^m x_j}{\sum_{j=1}^{n}(\mu_{ij})^m} \tag{4}$$

$$v_i^{k+1}=\begin{cases}\varpi v_i^k+c_1r_1(\rho_i-x_i^k)+c_2r_2(\sigma_i-x_i^k), & k>1\\ c_2r_2(\sigma_i-x_i^k), & k=1\end{cases} \tag{5}$$

$$x_i^{k+1}=x_i^k+v_i^{k+1} \tag{6}$$

式中, m 是加权指标, $m>1$, x_j-v_i表示样本 x_j到第 i 个样本中心的聚类, PSO-FCM 算法适应度函数J_m值越小说明性能越好; ρ_i是粒子最优适应度的位置, σ_i是群体最优适应度函数, c_1和c_2是学习因子; r_1和r_2是[0,1]之间的随机因子数, ϖ是惯性权重.

2 算法优化

2.1 PCA 算法

随着数据量的爆发和激增, 数据类型的增多, 数据复杂程度的加深, PSO-FCM 算法的性能无法完全发挥. 于是近年来有学者对该算法进行了再度优化, 陈寿文[8]提出利用混沌粒子融合粒子群模糊聚类算法 (CCPSOFCM), 余晓东等[9]利用直觉模糊核优化粒子群模糊聚类算法. 雷浩辖等[10]利用遗传算法 (GA) 与 PSO 混合优化的遗传粒子群模糊聚类 (GA-PSO-FCM). 这些学者都是针对于 PSO-FCM 算法依赖初始解这个问题上进行的优化. 算法核心是通过比较隶属度, 移动该粒子并决定属于哪一类, 但是在各维度上面的移动上并没有一个主次之分, 在各维度上的移动全部是随机因子数决定. 随着聚类中心数量的增加, 隶属度矩阵上, 各聚类中心隶属度值接近, 粒子各维度移动不受限, 这样导致部分粒子可能会被分入, 与正确聚类中心隶属度值接近的错误聚类中心中的问题. 在维度增加, 聚类中数量增加, 这个问题会越来越频繁出现.

为了在一定程度降低上面的问题出现的可能性, 本文引入了 PCA[11–13]算法对原算法进行优化, PCA 是一种统计分析的方法, 通过正交变换将具有一定相关性的向量转为彼此正交, 且互相独立的一维新向量 (即主成分). 每个主成分都是初始变量的线性组合, 没有冗余信息, 构成空间的正交基. 主成分分析法可以简化统计数据, 揭示特征变量之间的关系. 在本文优化中并没直接对数据进行降维, 根据 PCA 中主成分贡献率公式: $\eta_k=\frac{x_k}{\sum_{i=1}^{m}x_i}$, 计算出样本空间各维度之间的贡献率$\eta=(\eta_1,\eta_2,\cdots,\eta_n)$, 进一步优化 PSO-FCM 算法中速

度的v_i^{k+1}的迭代公式:

$$\begin{cases}\gamma_1=\eta_1 r_1,\ \gamma_2=\eta_2 r_2\\ v_i^{k+1}=\begin{cases}\varpi v_i^k+c_1\gamma_1(\rho_i-x_i^k)+c_2\gamma_2(\sigma_i-x_i^k), & k>1\\ c_2 r_2(\sigma_i-x_i^k), & k=1\end{cases}\end{cases} \tag{7}$$

2.2 优化算法实现

输入: 短文本数据集$X=(x_1,x_2,\cdots,x_n)$

初始化: 加权指标 m; 学习因子c_1和c_2; 惯性权重ϖ; 训练轮数 T.
随机初始化: 初始聚类中心$C_0=(c_1^{(0)},c_2^{(0)},\cdots,c_p^{(0)})$

1. $C^{\text{best}}\leftarrow C_0$
2. $J_m^{(0)}\leftarrow\sum_{j=1}^{n}\sum_{i=1}^{p}(\mu_{ij})^m\left(x_j-c_i^{(0)}\right)^2$
3. $J_m^{\text{best}}\leftarrow J_m^{(0)}$
4. for t=1,2,$\cdots$,T do
5. $\rho_{\text{best}}\leftarrow x_i^t$
6. $pindex\leftarrow\arg\max(\mu_{ij},j=1,2,\cdots,p)$
7. $\sigma_{\text{best}}\leftarrow c_{pindex}^{(0)}$
8. for i=1,2,$\cdots$,I do
9. $v_i^{t+1}\leftarrow c_2\gamma_2(\sigma_2-x_i^t)$
10. $x_i^{t+1}\leftarrow x_i^t+v_i^{t+1}$
11. if $J_m^{\text{best}}>J_m^{(t)}$ then
12. $J_m^{\text{best}}\leftarrow J_m^{(t)}$
13. $J_m=\sum_{j=1}^{n}\sum_{i=1}^{p}(\mu_{ij})^m(x_j-c_i)^2$
14. $C^{\text{best}}=\frac{\sum_{j=1}^{n}(\mu_{ij})^m x_j}{\sum_{j=1}^{n}(\mu_{ij})^m}$
15. else continue
16. end for
17. 输出中心C^{best}, J_m^{best}

3 实验分析

3.1 实验数据处理

在测试算法的性能, 本文选择 UCI 机器学习数据库中, Wine, Breast Tissue, Dermatology, 以及 Glass Identification, 每一组数据都进行了清洗, 并且都做了使用线性函数归一化将数据集进行标准化处理. 各维度的权重是通过主成分分析得出各维度贡献率, 数据集参见表 1 数据集表.

表 1 实验数据集表

数据集名称	数据维度	聚类中心数	各维度的权重
Wine	13	3	(0.4075, 0.1897, 0.0856, 0.0743, 0.0557, 0.0466, 0.0366, 0.0241, 0.0227, 0.0225, 0.0138, 0.0127, 0.0082)
Dermatology	33	6	(0.3222, 0.1739, 0.0712, 0.0464, 0.0452, 0.0342, 0.0322, 0.025, 0.0236, 0.0217, 0.0198, 0.0187, 0.0168, 0.0159, 0.0136, 0.0124, 0.0113, 0.0107, 0.0101, 0.0093, 0.009, 0.0086, 0.0079, 0.0072, 0.006, 0.0041, 0.004, 0.0037, 0.0035, 0.0029, 0.0026, 0.0025, 0.002, 0.0014, 0.0005)
Breast Tissue	10	6	(0.6682, 0.1819, 0.078, 0.0361, 0.0234, 0.0088, 0.0031, 0.0003, 0.0002)
Glass Identification	10	7	(0.4543, 0.1799, 0.1265, 0.098, 0.0686, 0.0421, 0.0261, 0.0043, 0.0001)

3.2 模型评价指标

通过对比本算法与 K-近邻 (KNN), FCM, PSO-FCM 在数据集训练的结果. 本文采取的评价算法性能的指标: 调整互信息 (Adjusted Mutual Information based scores, AMI); 调整兰德系数 (Adjusted Rand Index, ARI); FM 指数 (Fowlkes and Mallows Index, FMI). 3 个指标都是评价聚类算法性能的外部指标, 通过聚类结果与参考数据集的标签比较而获得, 这些外部指标度量的结果都在[0, 1]之间, 指标值越接近 1 说明聚类的结果越好.

3.3 检验模型性能

图 1 和图 2 根据 Breast Tissue 数据集的主成分贡献率所选择的平面图, 图 1 是本文算法在数据集上, 两个高贡献率维度的图像, 图 2 是 PSO-FCM 算法, 从图中可以明显的对比出来, 在相同数据集, 相同维度下的本文算法聚类的结果明显优于 PSO-FCM, PSO-FCM 算法在数据比较集中的区域, 对于多个聚类中心的交界处的数据敏感程度低, 无法有效的给出数据的准确的聚类中心, 相反本文算法面对这类粒子, 敏感度高, 能够更加有效的且准确的给出聚类中心. 粒子各维度之间无差别移动, 在多个聚类中心的粒子会被错误的移动到不正确的聚类中心中: 本算法对于不同贡献率的空间中, 采取相对应的移动权重的能够较低粒子错误移动的概率, 说明该策略效果是显著的.

由表 2 和表 3 中可以看出, 本文算法只是在 Dermatology 数据集上的 AMI 这一个指标上落后 KNN, 这是因为作为硬聚类算法, 随着聚类中心数目的增加, 每一个数据只能存在单一的一个聚类结果, 不会存在多种可能性, 聚类的结果纯度更高. KNN 算法性能很稳定, 在随着聚类中心增多, 性能反超 FCM, PSO-FCM 两个算法, 但是综合指标上, 本文的算法总体仍是优于 FCM, PSO-FCM, KNN 这 3 个算法. FCM 采用随机初始的中

心, 指标随着聚类中心的增多, 算法性能下降明显. PSO-FCM 采取使用 PSO 算法得出的初始中心, 明显的发现, 综合性能上面性能上优于 FCM, 但是算法精度提升不高.

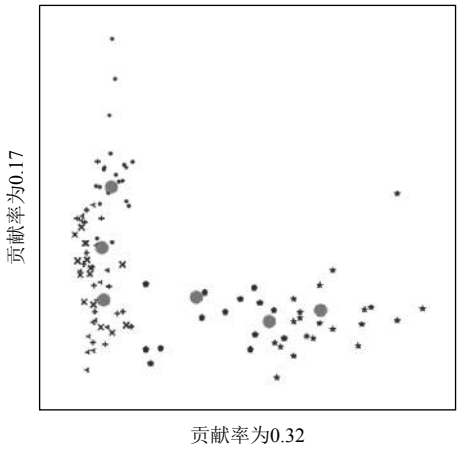

图 1 PCA-PSO-FCM 高贡献率图

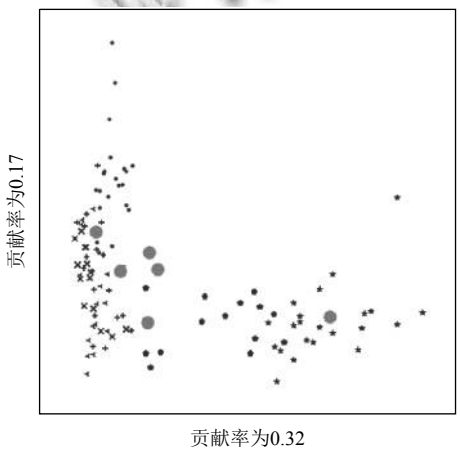

图 2 PSO-FCM 高贡献率图

随着各数据集的聚类中心的增加, 聚类的问题的复杂化, 从表中各指标上, 侧面体现本算法面对多个聚类中心的之间的粒子敏感度更高, 分辨能力更强. 总体上指标上来看, 本文算法性能更强, 鲁棒性更高, 适用面更广.

4 结论与展望

采取 PCA 优化的 PSO-FCM 算法, 通过主贡献率加权的限制, 控制粒子各维度上的移动, 降低多聚类群交界粒子的敏感性, 增强了粒子的搜索能力, 降低粒子被不正确粒子群吸入, 能够一定程度上, 跳出局部最优, 有效的弥补了传统 PSO-FCM 性能上的不足, 增加算法精度, 增强算法的鲁棒性, 相对于其他算法, 在综合指标上面更优, 部分指标上有着更好的精度, 适用面更广, 鲁棒性更强. 接下来的工作会将优化算法应用到更多领域.

表 2 算法性能表 1

数据集	算法	FM(%)	AMI(%)
Wine	KNN	61.25	46.59
	FCM	78.28	67.89
	PSO-FCM	82.27	74.39
	本文算法	**83.85**	**75.01**
Dermatology	KNN	73.77	**76.42**
	FCM	69.62	60.96
	PSO-FCM	70.06	65.18
	本文算法	**78.02**	70.86
Breast Tissue	KNN	40.72	46.20
	FCM	42.14	46.39
	PSO-FCM	42.40	46.86
	本文算法	**47.59**	**47.27**
Glass Identification	KNN	39.18	30.31
	FCM	32.74	23.56
	PSO-FCM	33.27	24.15
	本文算法	**40.49**	**32.52**

表 3 算法性能表 2

数据集	算法	ARI(%)	综合指标 (%)
Wine	KNN	45.11	50.98
	FCM	67.03	71.07
	PSO-FCM	69.65	75.44
	本文算法	**75.56**	**78.14**
Dermatology	KNN	67.27	72.49
	FCM	58.47	63.02
	PSO-FCM	61.28	65.51
	本文算法	**70.65**	**73.18**
Breast Tissue	KNN	28.34	38.42
	FCM	29.25	39.26
	PSO-FCM	29.67	39.64
	本文算法	**33.36**	**42.74**
Glass Identification	KNN	17.09	28.86
	FCM	14.63	23.64
	PSO-FCM	15.02	24.15
	本文算法	**22.14**	**31.72**

参考文献

1 许磊, 张凤鸣. 基于 PSO 的模糊聚类算法. 计算机工程与设计, 2006, 27(21): 4128–4129. [doi: 10.3969/j.issn.1000-7024.2006.21.054]

2 Mekhmoukh A, Mokrani K. Improved fuzzy C-means based Particle Swarm Optimization (PSO) initialization and outlier rejection with level set methods for MR brain image segmentation. Computer Methods and Programs in Biomedicine, 2015, 122(2): 266–281. [doi: 10.1016/j.cmpb.

2015.08.001]

3 Filho TMS, Pimentel BA, Souza RMCR, *et al*. Hybrid methods for fuzzy clustering based on fuzzy c-means and improved particle swarm optimization. Expert Systems with Applications, 2015, 42(17–18): 6315–6328. [doi: 10.1016/j.eswa.2015.04.032]

4 Jie LL, Liu WD, Sun Z, *et al*. Hybrid fuzzy clustering methods based on improved self-adaptive cellular genetic algorithm and optimal-selection-based fuzzy c-means. Neurocomputing, 2017, 249: 140–156. [doi: 10.1016/j.neucom.2017.03.068]

5 Benaichouche AN, Oulhadj H, Siarry P. Improved spatial fuzzy c-means clustering for image segmentation using PSO initialization, Mahalanobis distance and post-segmentation correction. Digital Signal Processing, 2013, 23(5): 1390–1400. [doi: 10.1016/j.dsp.2013.07.005]

6 邱宁佳, 李娜, 胡小娟, 等. 基于粒子群优化的朴素贝叶斯改进算法. 计算机工程, 2018, 44(11): 27–32, 39.

7 李锋. 粒子群模糊聚类算法在入侵检测中的研究. 计算机技术与发展, 2014, 24(12): 138–141.

8 陈寿文. 基于 Chebyshev 映射的混沌粒子群融合 FCM 聚类算法. 计算机应用与软件, 2015, 32(7): 255–258. [doi: 10.3969/j.issn.1000-386x.2015.07.062]

9 余晓东, 雷英杰, 岳韶华, 等. 基于粒子群优化的直觉模糊核聚类算法研究. 通信学报, 2015, 36(5): 78–84.

10 雷浩辖, 刘念, 崔东君, 等. 基于 GA 与 PSO 混合优化 FCM 聚类的变压器故障诊断. 电力系统保护与控制, 2011, 39(22): 52–56. [doi: 10.7667/j.issn.1674-3415.2011.22.010]

11 李元, 白岩松. 改进主成分分析的 KNN 故障检测研究. 沈阳化工大学学报, 2018, 32(4): 366–371. [doi: 10.3969/j.issn.2095-2198.2018.04.014]

12 王帅, 黄海鸿, 韩刚, 等. 基于 PCA 与 GA-BP 神经网络的磁记忆信号定量评价. 电子测量与仪器学报, 2018, 32(10): 190–196.

13 徐明月, 林泽轩, 顾彦. 基于 PCA-SVM 模型的红斑鳞屑性皮肤病识别研究. 杭州电子科技大学学报 (自然科学版), 2018, 38(6): 35–40.