

文章编号:1001-9081(2009)06-1569-03

基于人工鱼群算法的动态模糊聚类

刘 白,周永权,谢竹诚

(广西民族大学 数学与计算机科学学院,南宁 530006)

(yongquanzhou@126.com)

摘 要:针对传统的模糊 C-均值(FCM)聚类算法的聚类有效性对空间样本分布的依赖性等缺点,提出了一种新的基于人工鱼群算法的动态模糊聚类。通过引入模糊等价矩阵来表示高维样本之间的相似程度,并将高维样本映射到二维平面。然后利用人工鱼群算法不断优化二维样本的坐标值,使样本之间的欧氏距离向样本间的模糊等价矩阵趋近,最终实现模糊聚类。该方法克服了聚类有效性对高维样本空间分布的依赖性并同时提高了效率。仿真实验结果证明了该算法是有效的,具有聚类速度快、精度高等特点。

关键词:动态模糊聚类;人工鱼群算法;模糊相似矩阵;高维样本;模糊等价矩阵

中图分类号: TP181 **文献标志码:** A

Dynamic fuzzy clustering method based on artificial fish swarm algorithm

LIU Bai, ZHOU Yong-quan, XIE Zhu-cheng

(College of Mathematics and Computer Science, Guangxi University for Nationalities, Nanning Guangxi 530006, China)

Abstract: In order to avoid the dependence of the validity of clustering on the space distribution of high dimensional samples of Fuzzy C-Means (FCM), a dynamic fuzzy clustering method based on artificial fish swarm algorithm was proposed. By introducing a fuzzy equivalence matrix to the similar degree among samples, the high dimensional samples were mapped to two dimensional planes. Then the Euclidean distance of the samples was approximated to the fuzzy equivalence matrix gradually by using artificial fish swarm algorithm to optimize the coordinate values. Finally, the fuzzy clustering was obtained. The proposed method, not only avoided the dependence of the validity of clustering on the space distribution of high dimensional samples, but also raised the clustering efficiency. Experiment results show that it is an efficient clustering algorithm with rapid speed and high precision.

Key words: dynamic fuzzy clustering; artificial fish swarm algorithm; fuzzy similarity matrix; high dimension sample; fuzzy equivalence matrix

0 引言

聚类是数据挖掘中一个非常重要的内容,是一种重要的人类行为。简单地讲,聚类分析就是将数据对象分组成为多个类或簇,在同一个簇中的对象之间具有较高的相似度而不同簇中的对象具有较大的差别^[1]。从机器学习观点来看,聚类是一种无指导的学习,与分类不同,聚类和无指导学习不依赖预先定义的类和带类标号的训练实例。传统的聚类方法如模糊 C-均值聚类(Fuzzy C-Means, FCM)^[2]和 C-均值聚类等,它们是直接利用样本进行聚类,都是基于目标函数的聚类^[3],并没有对样本进行相关的预处理。而基于目标函数的聚类对样本的空间分布有较强的依赖性,这样最终的有效性在很大程度上就取决于样本的分布情况^[4]。例如, C-均值聚类对于特征空间呈超球体的情况聚类效果较好,而对于呈任意形状簇分布的情况则聚类效果较差^[5], FCM 模糊聚类对于特征空间呈椭球体结构的情况聚类效果较好^[6],而且 FCM 算法对高维数据聚类时速度较慢^[7]。

因此,为了克服聚类有效性对样本分布的依赖性以及提高聚类的效率,本文提出了一种基于人工鱼群算法的动态模

糊聚类方法,目的是实现分布呈任意形状簇的样本聚类。这样不仅克服了聚类有效性对样本分布的依赖性,又增加了聚类的灵活性和可视化。首先,我们对样本进行降维预处理,通过构造模糊等价矩阵,实现高维样本向二维样本的映射。然后利用人工鱼群算法对初始时随机分布在二维平面内的各聚类样本的坐标值进行全局优化,使样本之间的欧氏距离逐步趋近于它们之间的模糊等价关系,实现动态模糊聚类。

FCM 本质上是一种局部搜索算法,收敛于局部最优,容易陷入局部极小值^[9-10],而人工鱼群算法是一种全局搜索算法,是李晓磊等人^[8,11]模仿鱼类行为方式提出的一种基于动物自治体的优化方法,是集群智能思想的一个具体应用。本文提出的方法在性能上较经典 FCM 聚类方法有一定改进,聚类更准确,收敛时间较快。仿真实验表明该方法能得到较好的聚类结果,并同时验证了其有效性和可行性。

1 基于人工鱼群算法的动态模糊聚类

1.1 建立模糊相似矩阵

模糊相似矩阵用于存储样本之间的相似程度,用 $[0, 1]$ 的数来表示。设样本空间 $X = \{x_1, x_2, \dots, x_n\}$, $\forall x_i \in X$,其

收稿日期:2008-12-08;修回日期:2009-02-20。

基金项目:国家民委科研项目(08GX01);广西自然科学基金资助项目(0832082);广西民族大学创新计划项目(gxun-chx0885)。

作者简介:刘白(1981-),女,河南驻马店人,硕士,主要研究方向:计算智能;周永权(1962-),男,陕西旬邑人,教授,博士,主要研究方向:计算智能、神经网络;谢竹诚(1983-),男,广东韶关人,硕士,主要研究方向:计算智能。

特征矢量为 $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{im})$, \mathbf{x}_{ik} 表示 i 个样本的第 k 个特征属性。记 n 个样本第 k 个特征属性的平均值和标准方差分别为:

$$\mu_k = \frac{1}{n} \sum_{i=1}^n x_{ik} \quad (1)$$

$$s_k = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_{ik} - \mu_k)^2} \quad (2)$$

则原始样本可标准化为:

$$\mathbf{x}_{ik}' = (x_{ik} - \mu_k) / s_k \quad (3)$$

利用极值标准化公式进行归一化,即:

$$\mathbf{x}_{ik} = (\mathbf{x}_{ik}' - \mathbf{x}_{\min k}') / (\mathbf{x}_{\max k}' - \mathbf{x}_{\min k}') \quad (4)$$

其中 $\mathbf{x}_{\min k}'$ 和 $\mathbf{x}_{\max k}'$ 分别为 $\mathbf{x}_{1k}', \mathbf{x}_{2k}', \dots, \mathbf{x}_{nk}'$ 中的最大值和最小值。模糊相似矩阵 $(r_{ij})_{nn}$ 是一个 $n \times n$ 维的对角线元素为 1 的对称矩阵,即:

$$\mathbf{R} = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ r_{31} & r_{32} & \cdots & r_{3n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nn} \end{bmatrix} \quad (5)$$

其中 r_{ij} 代表样本 i 和 j 之间相似性的量化表示。通常,它是一个非负的数值。常采用欧氏距离计算,即:

$$d_{ij} = \sqrt{\sum_{k=1}^m (x_{ik} - x_{jk})^2} \quad (6)$$

$$r_{ij} = 1 - Cd_{ij} \quad (7)$$

其中 C 选取适当的正数,使 r_{ij} 在 $[0, 1]$ 内, m 为样本属性个数。

1.2 建立模糊等价矩阵

用上述方法建立起来的相似关系 \mathbf{R} , 一般只满足反射性和对称性, 不满足传递性, 因而还不是模糊等价关系。为此, 需要将 \mathbf{R} 改造成 \mathbf{R}^* , 在适当的阈值上进行截取, 便可得到所需的分类。将 \mathbf{R} 改造成 \mathbf{R}^* , 可用求传递闭包的方法^[12]。 \mathbf{R} 自乘的思想是按最短距离法原则, 寻求两个向量 \mathbf{x}_i 与 \mathbf{x}_j 的亲密程度。

假设 $\mathbf{R}^2 = (r_{ij}^2)$, 即 $r_{ij}^2 = \bigvee_{k=1}^n (r_{ik} \wedge r_{kj})$, 说明 \mathbf{x}_i 与 \mathbf{x}_j 是通过第三者 k 作为媒介而发生关系, $r_{ik} \wedge r_{kj}$ 表示 \mathbf{x}_i 与 \mathbf{x}_j 的关系密切程度是以 $\min(x_{ik}, x_{kj})$ 为准则, 因 k 是任意的, 故从一切 $r_{ik} \wedge r_{kj}$ 中寻求一个使 \mathbf{x}_i 和 \mathbf{x}_j 关系最密切的通道。 \mathbf{R}^m 随 m 的增加, 允许连接 \mathbf{x}_i 与 \mathbf{x}_j 的链的边就越多。由于从 \mathbf{x}_i 到 \mathbf{x}_j 的一切链中, 一定存在一个使最大边长达到极小的链, 这个边长就是相当于 r_{ij}^* 。

在实际处理过程中, \mathbf{R} 的收敛速度是比较快的。为进一步加快收敛速度, 通常采取如下处理方法:

$$\mathbf{R} \rightarrow \mathbf{R}^2 \rightarrow \mathbf{R}^4 \rightarrow \cdots \rightarrow \mathbf{R}^{2^k} \quad (8)$$

即先将 \mathbf{R} 自乘改造为 \mathbf{R}^2 , 再自乘得 \mathbf{R}^4 , 如此继续下去, 直到某一步出现 $\mathbf{R}^{2^k} = \mathbf{R}^k = \mathbf{R}^*$ 。此时 \mathbf{R}^* 满足了传递性, 于是模糊相似矩阵 (\mathbf{R}) 就被改造成了一个模糊等价关系矩阵 (\mathbf{R}^*) 。

1.3 食物浓度函数的选取

构建模糊等价关系矩阵之后, 可将高维样本映射到二维平面。通过算法对各个样本的坐标值进行迭代优化, 使各样本间的欧氏距离趋近于模糊等价矩阵。因此, 人工鱼群算法的误差函数定义为:

$$E = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n |r_{ij}' - r_{ij}^*| \quad (9)$$

其中: r_{ij}' 表示映射到二维平面的样本 i 和 j 之间的欧氏距离。设样本 i 和 j 的坐标值分别为 (a_i, b_i) 和 (a_j, b_j) , $i = 1, 2, \dots, n, j = 1, 2, \dots, n$, 则 r_{ij}' 为:

$$r_{ij}' = \sqrt{(a_i - a_j)^2 + (b_i - b_j)^2} \quad (10)$$

误差函数值越小, 相似性越大, 个体的食物浓度越高, 因此个体的食物浓度函数定义为:

$$f(x) = \frac{1}{E + 1} \quad (11)$$

1.4 基于人工鱼群算法的动态模糊聚类

基于人工鱼群算法的动态模糊聚类方法描述如下:

- 1) 初始化。将待聚类样本随机分布在二维平面的一定区域内, 即随机赋给每个样本一对坐标值 (a_i, b_i) , 其中 a_i 和 b_i 取值均为 $[0, 1]$, $i = 1, 2, \dots, n$ 。
- 2) 建立模糊相似矩阵。利用式(1) ~ (7) 建立模糊相似矩阵 \mathbf{R} , 利用式(8) 通过模糊相似矩阵 \mathbf{R} 建立模糊等价矩阵 \mathbf{R}^* 。
- 3) 利用式(9) ~ (11) 计算人工鱼的食物浓度的大小。
- 4) 根据当前人工鱼的食物浓度大小进行觅食行为、聚群行为和追尾行为; 若不满足条件, 则执行随机行为。
- 5) 若组内所有人工鱼都结束移动, 则继续向下执行, 否则转3)。
- 6) 终止操作。如果新一代人工鱼的个体的最大食物浓度函数值与上一代最大食物浓度函数值之差小于 ξ (ξ 取值为 0.001) 或者已经达到最大循环次数, 则结束; 否则转3)。
- 7) 对最后得到的二维坐标值应用 FCM 算法, 并把最后的聚类结果对应到原始的高维样本中。

2 仿真实验

为了验证所提出的方法有效性和可行性, 对 UCI 机器学习数据库有关酒的测试数据集进行了仿真实验, 如表 1 所示。

表1 酒类样本数据

编号	属性												
	1	2	3	4	5	6	7	8	9	10	11	12	13
1	14.23	1.71	2.43	15.6	127	2.80	3.06	0.28	2.29	5.64	1.04	3.92	1065
2	13.2	1.78	2.14	11.2	100	2.65	2.76	0.26	1.28	4.38	1.05	3.40	1050
3	13.16	2.36	2.67	18.6	101	2.80	3.24	0.30	2.81	5.68	1.03	3.17	1185
4	14.37	1.95	2.50	16.8	113	3.85	3.49	0.24	2.18	7.80	0.86	3.45	1480
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
177	13.17	2.59	2.37	20.0	120	1.65	0.68	0.53	1.46	9.30	0.60	1.62	840
178	14.13	4.10	2.74	24.5	96	2.05	0.76	0.56	1.35	9.20	0.61	1.60	560

该测试数据集样本个数为 178, 条件属性个数为 13, 聚类类别为 3。类别 1、2 和 3 所包括的样本数分别为 59、71 和 48。初始时, 样本的随机分布情况如图 1 所示。

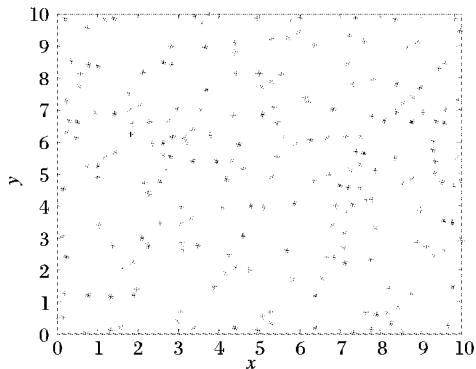


图1 样本在二维平面随机分布

采用提出的新方法, 经过 30 次迭代, 移动步长为 1, 可视域范围为 2, 则动态模糊聚类结果如图 2 所示。

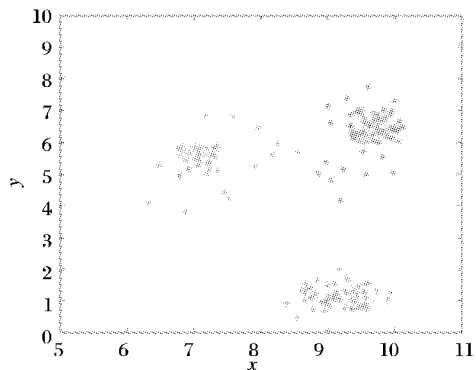


图2 动态模糊聚类结果

与 FCM 算法、文献[13-14]相比, 本文方法其聚类正确率高于 FCM 算法和文献[13-14]的结果, 收敛速度也比 FCM 算法和文献[13-14]快。比较结果如表 2 所示。

表2 几种算法的比较结果

方法	迭代次数	正确率/%		
		类别 1	类别 2	类别 3
FCM	150	90	92	100
文献[13]方法	80	93	98	100
文献[14]方法	80	90	95	100
本文方法	30	95	98	100

3 结语

传统聚类方法的有效性依赖于样本的分布情况, 若样本

界限分明, 则聚类效果好。但是实际情况往往是样本分布呈任意形状簇。对于这类情形, 已有的方法效果不佳, 本文提出的基于人工鱼群算法的动态模糊聚类方法, 通过人工鱼群算法和模糊等价关系矩阵将高维样本映射到二维平面, 迭代优化各样本的坐标值, 使样本之间的欧氏距离逐步趋近于它们之间的模糊等价矩阵, 最终得到全局最优解, 动态实现模糊聚类。仿真结果表明, 该方法在性能上较经典的模糊聚类算法有一定改进, 不依赖于样本特征空间的分布。

参考文献:

- [1] JIA WEI-HAN, KAMBER M. 数据挖掘概念与技术 [M]. 范明, 孟小峰, 译. 北京: 机械工业出版社, 2001.
- [2] 黄凤岗, 宋克欧. 模式识别 [M]. 哈尔滨: 哈尔滨工程大学出版社, 1998.
- [3] ZHANG YUAN-QUAN, RUEDA L. A geometric framework to visualize fuzzy clustered data [C]// Proceedings of the XXV International Conference on the Chilean Computer Science Society. Washington, DC: IEEE Computer Society, 2005: 13-17.
- [4] 张莉, 周伟达, 焦李成. 核聚类算法 [J]. 计算机学报, 2002, 25 (6): 587-590.
- [5] 高新波, 谢维信. 模糊聚类理论发展及应用的研究进展 [J]. 科学通报, 1999, 44(21): 2241-2251.
- [6] HATHAWAY R J, BEZDEK J C. Fuzzy c-means clustering of incomplete data [J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 2001, 31(5): 735-744.
- [7] AGGARWAL C, YU P. Finding generalized projected clusters in high dimensional spaces [C]// SIGMOD'00: Proceedings of ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 2000: 70-81.
- [8] 李晓磊, 邵之江, 钱积新. 一种基于动物自治体的寻优模式: 鱼群算法 [J]. 系统工程理论与实践, 2002, 22(11): 76-82.
- [9] KAMEL S M. New algorithms for solving the fuzzy C-means clustering problem [J]. Pattern Recognition, 1994, 27(4): 14-21.
- [10] 赵艳厂, 谢帆, 宋俊德. 一种新的聚类算法: 等密度线算法 [J]. 北京邮电大学学报, 2002, 25(2): 8-13.
- [11] 李晓磊. 一种新型的智能优化方法——人工鱼群算法 [D]. 杭州: 浙江大学, 2003.
- [12] 高新波. 模糊聚类分析及其应用 [M]. 西安: 西安电子科技大学出版社, 2004.
- [13] 郑岩, 黄荣怀, 战晓苏, 等. 基于遗传算法的动态模糊聚类 [J]. 北京邮电大学学报, 2005, 28 (1): 75-78.
- [14] 张利彪, 周春光, 马铭, 等. 基于粒子群优化算法的模糊 C-均值聚类 [J]. 吉林大学学报: 理学版, 2006, 44(1): 217-222.

(上接第 1565 页)

参考文献:

- [1] AERTS S, LAMBRECHTS D, MAITY S, et al. Gene prioritization through genomic data fusion [J]. Nature Biotechnology, 2006, 24 (5): 537-544.
- [2] de BIE T, TRANCHEVENT L C, van OEFFELLEN L M M, et al. Kernel-based data fusion for gene prioritization [J]. Bioinformatics, 2007, 23(13): 125-132.
- [3] PAVLIDIS P, WESTON J, CAI J S, et al. Learning gene functional classifications from multiple data types [J]. Journal of Computational Biology, 2002, 9(2): 401-411.
- [4] LANCKRIET G R G, DENG M, CRISTIANINI N, et al. Kernel-

based data fusion and its application to protein function prediction in yeast [C]// Pacific Symposium on Biocomputing. Hawaii, USA: World Scientific, 2004: 300-311.

- [5] LANCKRIET G R G, de BIE T, CRISTIANINI N, et al. A statistical framework for genomic data fusion [J]. Bioinformatics, 2004, 20 (16): 2626-2635.
- [6] LANCKRIET G R G, CRISTIANINI N, BARTLETT P, et al. Learning the kernel matrix with semidefinite programming [J]. Journal of Machine Learning Research, 2004, 5: 27-72.
- [7] SCHOLKOPF B, PLATT J C, SHAWE-TAYLOR J, et al. Estimating the support of a high-dimensional distribution [J]. Neural Computation, 2001, 13(7): 1443-1471.