

基于蚁群和模拟退火算法的聚类新方法

陶 红, 史小伍, 李 莎, 高 尚

(江苏科技大学 计算机科学与工程学院, 江苏 镇江 212003)

摘 要: 模拟退火算法具有良好的全局搜索能力, 而蚁群算法具有良好的分布式并行性和正反馈能力. 针对样本维数大、数目多时聚类效果不满意的问题, 提出了混合的蚁群模拟退火算法, 思路是利用 K -均值算法的结果作为初值, 再使用蚁群算法和模拟退火算法对初值进行调整聚类, 结果表明这种算法比较有效.

关键词: 聚类分析; K -均值算法; 蚁群算法; 模拟退火算法

中图分类号: TP301.6

文献标识码: A

文章编号: 1000-7180(2011)12-0096-03

Clustering Algorithm Based on Hybrid Intelligent Algorithm

TAO Hong, SHI Xiao-wu, LI Sha, GAO Shang

(School of Computer Science and Engineering, Jiangsu University of Science and Technology, Zhenjiang 212003, China)

Abstract: Simulated annealing algorithm has the ability of doing a global stochastically, and ant colony algorithm has the ability of distributed parallel processing and good feedback. Due to the problem that when the dimension and the number of the sample are large, the cluster result will be unsatisfied, a hybrid intelligent algorithm is proposed. The algorithm is extend to use K -Means clustering to seed the initial solution and the ant colony algorithm and simulated annealing algorithm to adjust the initial cluster. Through the result, the more effectiveness of this algorithm is illustrated.

Key words: clustering analysis; K -means algorithm; ant colony algorithm; simulated annealing algorithm

1 引言

聚类就是将数据对象按照它们之间的差异进行分类, 使得同组内的对象差异尽可能小, 而不同组间的对象差异尽可能大. 聚类方法很多, 比较典型的有: 基于划分的聚类方法 K -Means 和 K -Medoies, 基于层次的聚类方法 AGENS (AGlomerative NESTing) 和 DIANA (DIvisive ANALysis), 基于密度的聚类方法 DBSCAN (Density-Based Spatial Clustering of Applications with Noise), 基于蚁群的聚类方法等等. 蚁群算法 (Ant Colony Optimization, ACO)^[1] 是一种新型的模拟蚂蚁群体智能行为的仿生优化算法, 由意大利学者 M. Dorigo 等人于 1991 年率先提出来, 具有良好的正反馈和分布式计

算的特性, 同时也存在易陷入局部收敛的缺陷^[2]. 蚁群算法常用来解决离散域中许多优化问题, 典型的应用有旅行商问题、调度问题、车辆绕径问题等^[3-5]. 模拟退火算法 (Simulated Annealing, SA) 最早由 Metropolis 等人于 1953 年提出, 1983 年 Kirpatrick 等人将其应用于组合优化问题中. 它具有渐进收敛性, 在理论上已经被证明是一种以概率 1 收敛于全局最优解的全局优化算法^[6], 它的初始解的设定对算法的收敛速度影响很大.

2 蚁群算法和模拟退火算法的融合

2.1 蚁群算法和模拟退火算法的融合的基本思想

蚁群算法具有收敛速度快的优点, 但容易陷于局部最优解从而导致算法停滞. 模拟退火算法具有

全局搜索能力,但搜索时间较长.利用它们各自的优势,提出一种融合蚁群算法和模拟退火算法的算法,基本思想是,首先利用 K-Means 算法对数据对象进行快速聚类,然后利用蚁群算法进行迭代,对每次迭代的最优结果进行模拟退火,由于模拟退火算法容许解以一定的概率变坏,这样既可以避免蚁群算法陷入局部最优,也可以提高模拟退火的速度.

2.2 蚁群算法和模拟退火算法的融合

2.2.1 问题描述

基于蚁群的聚类方法可以分为两种,一种是基于蚁堆形成原理来实现聚类的方法,另一种是基于蚁群觅食原理来实现聚类的方法^[7].本文采用的是第二种蚁群聚类方法,蚁群觅食原理的思想是,将数据对象视为具有不同属性的蚂蚁,而将聚类中心看做是蚂蚁要寻找的食物源,那么聚类的过程就可以看成是蚂蚁寻找食物源的过程.

已知数据样本集 $\{X\}$ 有 N 个样本和 K 个样本分类 $\{S_j\}$, $j=1,2,\dots,K$,以每个数据样本到对应的聚类中心的距离之和最小为聚类准则,它的数学模型为

$$\min \sum_{j=1}^K \sum_{x \in S_j} \|X - M_j\| \quad (1)$$

式中, K 为聚类数目, $M_j = \frac{1}{N_j} \sum_{x \in S_j} X$,为 S_j 中数据样本的均值向量.

2.2.2 初始化

初始的数据样本由 K-Means 算法进行快速分类,蚁群算法根据分类的启发式信息,初始化信息素浓度,若数据样本 X_i 分配给聚类中心 S_j , $j=1,2,\dots,K$,则数据样本 X_i 到 S_j 的路径上的信息素浓度 τ_{ij} 为

$$\tau_{ij} = \tau_{ij}(0) + \frac{Q}{d_{ij}} \quad (2)$$

式中, $\tau_{ij}(0)$ ($i,j=1,2,\dots,K$)为初始的信息素浓度值,均为相同的数值, Q 为一正常数, d_{ij} 为数据样本 X_i 到 S_j 聚类中心的距离.

2.2.3 移动

第 i 只蚂蚁选择聚类中心 S_j 的概率为

$$P_{ij} = \frac{\tau_{ij}}{\sum_{j=1}^K \tau_{ij}} \quad (3)$$

信息素更新方程为

$$\tau_{ij}^{\text{new}} = \rho \tau_{ij}^{\text{old}} + \frac{Q}{d_{ij}} \quad (4)$$

ρ 表示信息素残留系数,一般取 $0.5 \sim 0.9$

左右^[8].

2.2.4 模拟退火局部搜索

为快速获取最优解,模拟退火邻域搜索从一次蚂蚁搜索获得的最优解 W_0 开始随机扰动产生另一个解,定义为 W_1 .然后,根据前述的目标函数式(1),计算最优解和局部搜索解的目标函数值,分别为 $f(W_0)$ 和 $f(W_1)$,并以概率式(5)的方式接受新解为当前最优解,其中 T 为当前温度.

$$P = e^{-(f(W_1) - f(W_0))/T} \quad (5)$$

2.3 融合的蚁群和模拟退火算法的求解步骤

步骤1:参数的初始化;

步骤2:利用 K-Means 算法进行快速分类,获得当前全局最优解,并按式(2)初始化蚁群的信息素浓度;

步骤3:For 每次迭代 do

(1)For 每只蚂蚁 do

(2)按照式(3)计算选择概率 p_{ij} ,并以该概率选择聚类中心;

(3)按照式(4)局部更新信息素浓度;

End For

(4)按照式(1)计算目标函数值,并计比较得出本次全局搜索中的最优解;

(5)设定初始温度 T ,终止温度 T_0 ,使用当前最优解作为初始解 W_0 ;

(6)对当前最优解 W_0 进行随机扰动,产生变异解 W_1 ;

(7)以式(5)的概率接受新解 W_1 为当前最优解,存为当前最优解;

(8)检查退火终止条件,如果满足则转向步骤4,否则转向(9);

(9)按照降温公式计算下次迭代温度,转向(6);

步骤4:将全局最优解和当前最优解进行比较,更新全局最优解;

End For

步骤5:结束搜索过程,输出最终结果.

3 测试

采用14个2维样本蝶形数据和著名的 Iris 植物样本数据进行测试.算法参数为: $\rho=0.9$, $Q=80$,迭代次数 $NC=1\ 000$,蚂蚁数目 $R=30$,初始温度为 $T=1\ 000$, $T_0=0.01$,降温速度为 $\alpha=0.95$,算法测试50次.结合文献[8]提供的数据进行了比较,文献[8]的结果如表1和表2所示.基于蚁群和模拟

退火算法的数据结果如表 3 所示. 从表 1、表 2 和表 3 可以看出,随着样本维数和数量的增大,算法的聚类效果均有所改变. 由于融合的蚁群和模拟退火算法,每次迭代都是在当前最优解的基础上进行模拟退火,相比单纯的使用蚁群算法和模拟退火算法,聚类效果比较好.

表 1 蝶形数据的结果比较

算法	平均值	最好解	最差解	最好解的次数	正确率
模拟退火算法	29.125 5	26.857 1	42.547 1	39	78%
K-Means 均值算法(91 次)	34.840 8	26.857 1	87.400 0	79	79/91=86.81%
基本蚁群算法	27.170 8	26.857 1	42.541 7	49	98%
K-Means 均值与蚁群算法结合	26.857 1	26.857 1	26.857 1	50	100%

表 2 Iris 植物样本数据的结果比较

算法	平均值	最好解	最差解	最好解的次数	正确率
模拟退火算法	113.363 7	76.658 1	562.352 6	5	10%
K-Means 均值算法(551 300 次)	92.189 4	78.651 0	682.838 5	221 234	40.13%
基本蚁群算法	91.941 7	78.651 0	143.272 5	21	42%
K-Means 均值与蚁群算法结合	80.169 2	78.651 0	110.971 1	32	64%

表 3 基于蚁群和模拟退火算法的结果比较

数据	平均值	最好解	最差解	最好解的次数	正确率
蝶形数据	26.857 1	26.857 1	26.857 1	50	100%
Iris 植物样本数据	78.68	78.651 0	80.102 4	49	98%

4 结束语

本文针对这一问题,结合蚁群算法与模拟退火算法的特点,将两者融合,并利用 K-Means 算法做快速分类,不仅有利于提高之后的分类速度,而且也给蚁群算法提供了启发式信息,聚类效果有了很大的改善. 针对大容量样本的聚类方法,还有待进一步研究.

参考文献:

[1] Dorigo M, M aniezzo V, Colorni A. Ant system: optimization by a colony of cooperating agents[J]. IEEE Transaction on Systems, Man, and Cybernetics—Part B, 1996,26(1):1—13.

[2] 段海滨,王道波,朱家强,等. 蚁群算法理论及应用研究的进展[J]. 控制与决策,2004,19(12): 1321—1340.

[3] 吴启迪,汪镭. 智能蚁群算法及其应用[M]. 上海:上海科技出版社,2004.

[4] 陈欢欢,刘宴兵. 基于自适应调度的网格容错机制研究[J]. 微电子学与计算机,2010,35(2):145—148.

[5] 杨慧,王直. 基于蚁群退火算法的最优航线设计[J]. 现代电子技术,2009,19(11):188—190.

[6] 刘超,戎蒙恬. 模拟退火算法的 WLAN 信道分配中的应用[J]. 微电子学与计算机,2010,6(6):17—19.

[7] 郑耿忠,刘秋梅. 蚁群算法在分布式智能答疑系统中的应用研究[J]. 微电子学与计算机,2010,15(6):50—53.

[8] 高尚,汤可宗,杨靖宇. 一种新的基于混合蚁群算法的聚类方法[J]. 微电子学与计算机,2006,23(12):38—40.

作者简介:

陶 红 女,(1988—),硕士研究生. 研究方向为智能计算与技术.

史小伍 男,(1987—),硕士研究生. 研究方向为个人信用评价、模式识别与人工智能.