聚类问题的蚁群算法

高　尚[1,2]　杨静宇[1]　吴小俊[2]
[1]南京理工大学计算机系,南京 210094)
[2]华东船舶工业学院电子与信息系,镇江 212003)
E-mail:gao_shang@hotmail.com

摘　要　文章建立了聚类分析问题模型,分析了 K-均值算法、模拟退火算法和蚁群算法的优缺点,结果表明蚁群算法比较有效。

关键词　聚类分析　蚁群算法　K-均值算法　模拟退火算法

文章编号 1002-8331-(2004)08-0090-02　**文献标识码** A　**中图分类号** O235;TP301.6

An Ant Colony Algorithm for Clustering Problem

Gao Shang[1,2]　Yang Jingyu[1]　Wu Xiaojun[2]
[1](Department of Computer,Nanjing University of Science and Technology,Nanjing 210094)
[2](Department of Electronics and Information,East China Shipbuilding Institute,Zhenjiang 212003)

Abstract: An optimization model of clustering problem is given in this paper.The advantages and shortages of K-Means algorithm,simulated annealing algorithm and ant colony algorithm are analyzed,and the effectiveness of ant colony algorithm is illustrated through result.

Keywords: Clustering problem,Ant colony algorithm,K-Means algorithm,Simulated annealing algorithm

1　引言

聚类分析是按不同对象之间的差异,根据特定的准则做模式分类,其应用面相当广泛。大致分“分类数目未知”和“分类数目已知”两类问题。聚类分析方法比较多,比如对于“分类数目已知”聚类算法有 K-均值算法,ISODATA 算法,修正的 ISODATA 算法[1,2]等。该文用蚁群算法来解决此问题,并与 K-均值算法、模拟退火算法做了比较。

2　数学模型

已知模式样本集$\{X\}$有 n 个样本和 K 个模式分类$\{S_j,j=1,2,\cdots,K\}$,以每个模式样本到聚类中心的距离之和达到最小为准则,其数学模型为:

$$\min\sum_{j=1}^{K}\sum_{X\in S_j}\|X-m_j\| \qquad (1)$$

式中 K 为聚类数目,m_j 为 j 类样本的均值向量。若模式样本 i 分配第 j 聚类中心,则令 $y_{ij}=1$,否则令 $y_{ij}=0$。

$m_j=\dfrac{1}{\sum_{i=1}^{n}y_{ij}}\sum_{i=1}^{n}y_{ij}X_i$, $\sum_{j=1}^{n}y_{ij}=1$ 表示模式样本 i 只能分配到一个聚类中心上,因此聚类问题的数学模型为:

$$\min\sum_{i=1}^{n}\sum_{j=1}^{K}(y_{ij}\|X_i-m_j\|)$$

$$\text{s.t.}\sum_{j=1}^{K}y_{ij}=1 \quad (i=1,2,\cdots,n)$$

$$m_j=\frac{1}{\sum_{i=1}^{n}y_{ij}}\sum_{j=1}^{n}y_{ij}X_i \quad (j=1,2,\cdots,K) \qquad (2)$$

$y_{ij}=0,1$

此问题是一个非线性规划问题,目前没有有效的算法解此问题。

3　K 均值算法

K-均值算法的步骤如下[1,2]:

(1)任选 K 个初始聚类中心:$z_1,z_2,\cdots,z_K$;

(2)逐个将样本集$\{X\}$中各个样本按最小距离原则分配给 K 个聚类中心的某一个 z_j;

(3)计算新的聚类中心 z'_j $(j=1,2,\cdots,K)$,即 $z'_j=\dfrac{1}{N_j}\sum_{X\in S_j}X$,其中 N_j 为第 j 个聚类域 S_j 包含的个数;

(4)若 $z'_j\neq z_j$ $(j=1,2,\cdots,K)$,转步(2);否则算法收敛,计算结束。

4　模拟退火算法

模拟退火算法用于优化问题的出发点是基于物理中固体物质的退火过程与一般优化问题的相似性。算法的基本思想是从一给定解开始的,从邻域中随机产生另一个解,接受准则允许目标函数在有限范围内变坏,它由一控制参数 t 决定,其作用类似于物理过程中的温度 T,对于控制参数 t 的每一取值,算法持续进行“产生新解-判断-接受或舍弃”的迭代过程,对应着固体在某一恒定温度下趋于热平衡的过程。经过大量的解变换后,可以求得给定控制参数 t 值时优化问题的相对最优解。然后减小控制参数 t 的值,重复执行上述迭代过程。当控制参数逐渐减小并趋于零时,系统亦越来越趋于平衡状态,最后系统

作者简介:高尚(1972-),男,江苏镇江市人,硕士,讲师,从事系统工程理论与优化研究。

状态对应于优化问题的整体最优解，该过程也称冷却过程。

其模拟退火算法的步骤为：

(1)给定起、止"温度"T=100000、T_0=1和退火速度 α=0.9，随机产生一个聚类方案 $Y_0=(x_{ij})_{n\times K}$，计算每个模式样本到聚类中心的距离；

(2)若 $T>T_0$，转步(3)，否则算法停止，输出 Y_0；

(3)随机产生模式样本 i 和随机产生聚类中心 j，令 y_{ij}=1，其它 y_{im}=0 (m=1,2,…,K，$m\neq j$)，此时变量记为 Y_1；

(4)计算每个模式样本到新的聚类中心的距离之和 f_1，$\Delta E=f_1-f_0$，若 $\Delta E\leq 0$，接受新值，$Y_0\leftarrow Y_1$，$T\leftarrow\alpha T$，转步(2)；否则若 $\exp(-\Delta E/T)>rand(0,1)$，也接受新值，$Y_0\leftarrow Y_1$，$T\leftarrow\alpha T$，转步(2)；否则转步(3)。

5 蚁群算法

上世纪50年代中期创立了仿生学，人们从生物进化的机理中受到启发，提出了许多用以解决复杂优化问题的新方法，如遗传算法、进化规划、进化策略等，蚁群算法是最近几年才提出的一种新型的模拟进化算法，由意大利学者 M.Dorigo 等人首先提出来[3]，用蚁群在搜索食物源的过程中所体现出来的寻优能力来解决一些离散系统优化中困难问题。已经用该方法求解了旅行商问题、指派问题、调度问题等，取得了一系列较好的实验结果[4,5]。

人们经过大量研究发现，蚂蚁个体之间是通过一种称之为外激素(pheromone)的物质进行信息传递，从而能相互协作，完成复杂的任务。蚂蚁在运动过程中，能够在它所经过的路径上留下该种物质，而且蚂蚁在运动过程中能够感知这种物质的存在及其强度，并以此指导自己的运动方向，蚂蚁倾向于朝着该物质强度高的方向移动。因此，由大量蚂蚁组成的蚁群的集体行为便表现出一种信息正反馈现象：某一路径上走过的蚂蚁越多，则后来者选择该路径的概率就越大。蚂蚁个体之间就是通过这种信息的交流达到搜索食物的目的。

聚类问题的蚁群算法思路如下：在第 i 模式样本处分别设置1个蚂蚁，模式样本分配给第 j 个聚类中心 z_j (j=1,2,…,K)，蚂蚁就在模式样本 i 到聚类中心 z_j 的路径上留下外激素 τ_{ij}，第 i 个蚂蚁选择聚类中心 z_j 概率为：

$$p_{ij}=\frac{\tau_{ij}}{\sum_{j=1}^{n}\tau_{ij}} \qquad (3)$$

更新方程为：

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

d_{ij} 为模式样本 i 到聚类中心 z_j 的距离，ρ 表示强度的持久性系数，一般取0.5~0.9左右，Q 为一正常数。

解聚类问题的蚁群算法如下：

(1)$nc\leftarrow 0$ (nc 为循环次数)，给 τ_{ij} (i,j=1,2,…,n)赋相同的数值，给出 Q、ρ 的值，随机给出一个分配方案；

(2)对每个蚂蚁按转移概率 p_{ij} 选择下一个节点，按更新方程修信息强度；

(3)计算新的聚类中心，计算每个模式样本到新的聚类中心的距离 d_{ij}；

(4)$nc\leftarrow nc+1$，若 nc>规定的次数 NC，停止运行，根据外激素输出最好的解，否则转步(2)。

6 算法比较

例如对于14个2维样本[6]蝶形数据，最优聚类结果如图1所示。采用K-均值算法，此方法虽然简单，但其结果与初始聚类中心有关，假如初始聚类中心取14个样本中的任意两个，共有 C_{14}^{2}=91种情况，经过验证有12种情况：$\{x_1,x_2\}$，$\{x_1,x_3\}$，$\{x_2,x_3\}$，$\{x_4,x_5\}$，$\{x_4,x_6\}$，$\{x_5,x_6\}$，$\{x_9,x_{10}\}$，$\{x_9,x_{11}\}$，$\{x_{10},x_{11}\}$，$\{x_{12},x_{13}\}$，$\{x_{12},x_{14}\}$，$\{x_{13},x_{14}\}$收敛不到最优聚类结果。如初始聚类中心取$\{x_1,x_2\}$时，聚类结果如图2所示。因此对K-均值算法改进的方法是，随机产生一个聚类方案，计算其聚类中心作初始聚类中心。其效果如表1所示。

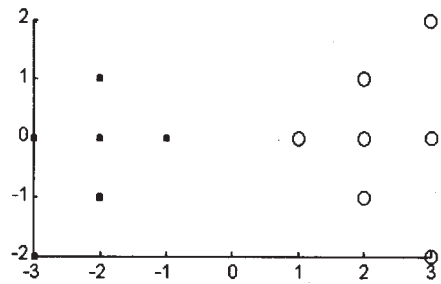

图1 最优聚类结果

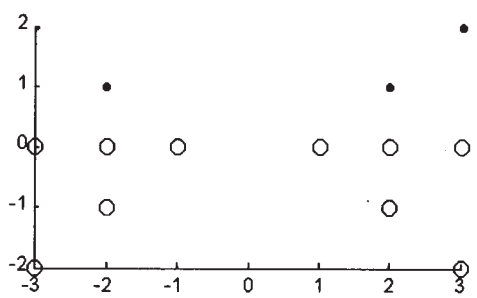

图2 初始聚类中心取$\{x_1,x_2\}$时聚类结果

采用模拟退火算法，参数设置如下：起始温度 T=1000，终止温度 T_0=1，退火速度 α=0.95。模拟退火算法的结果与初始聚类中心关系不大，但由于固体退火必须缓慢降温，才能使固体在每一温度下都达到热平衡，最终趋于平衡状态，因此，控制参数的值必须缓慢衰减，才能确保模拟退火算法最终趋于优化问题的整体最优解，因此迭代次数一般较长，且其正确率也不高。

采用蚁群算法参数设置如下：ρ=0.9，Q=100，NC=100。蚁群算法的结果与初始聚类中心关系也不大。与模拟退火算法类似，也采用概率改变变量，模拟退火算法在解的附近随机地找下一个解，以概率方式选取下一个解，而蚁群算法考虑到了附近解"外激素"，好的解附近的"外激素"较多，它被选取的概率就大，蚁群算法的迭代次数一般比较少，其方法相对比较有效。

表1 各种算法测试结果 后3种算法各运行100次)

算法	K-均值算法(以任意两个样本作聚类中心)	K-均值算法(随机产生一个聚类方案)	模拟退火算法	蚁群算法
正确率	79/91=86.8%	90%	78%	92%

7 结束语

蚁群算法的严格理论基础尚未奠定，有许多问题有待进一步研究，如算法的收敛性、理论依据等。但从当前的应用效果来

(下转232页)

洪灾损失指标来表示。如：亩均损失值，单位面积损失值，人均损失值等指标。综合洪灾损失率为：

$$\eta_{Colligate}=\frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} W_i}$$

其中 w_i 表示各经济部门的洪灾损失值，W_i 则表示 i 类经济部门的财产损失值。

5 结束语

鄱阳湖区洪涝灾害遥感动态监测系统能及时接收卫星遥感信息，并对其加以分析提取出水体淹没的范围和面积，及时反映淹没的程度及地理位置；通过所建立的降水与水位关系模型（通过了检验）预测某天的水位及淹没面积；预测和评估湖区洪涝灾害损失，使相关部门做好必要的防范措施，以减少灾害的损失。另外，随着 3S 技术和数据挖掘技术的迅速发展和不断成熟，可以考虑利用这些技术对此系统做进一步的完善。通过数据挖掘技术的利用，可以预先挖掘到一些有关降水、灾害等信息，对有关人员做出决策有很重要的作用。

（收稿日期：2003 年 7 月）

参考文献

1.[美]Michael N DeMers 著.武法东，付宗堂，王小牛等译.地理信息系统基本原理[M].第二版，北京：电子工业出版社，2001-10

2.Draper N，Smith R H.Applied Regression Analysis[M].2nd Ed，John Wiley，1981

3.上海师范大学数学系概率统计教研组.回归分析及其试验设计[M].上海教育出版社，1987

4.周复恭，黄运成.应用线性回归分析[M].北京：中国人民大学出版社，1989

5.张智星.MATLAB 程序设计与应用[M].北京：清华大学出版社，2002

6.陈桂明，戚红雨，潘伟.MATLAB 数理统计（6.x）[M].科学出版社，2002

7.江西省发展计划委员会，江西省气象局.鄱阳湖区生态环境遥感调查报告：江西省国土资源遥感综合调查之十一课题[R].2002-04

8.万庆等.洪水灾害系统分析与评估[M].科学出版社，1999

9.李博轩.Visual C++数据库开发指南[M].清华大学出版社

10.陈秀万.洪水灾害损失评估系统-遥感与 GIS 技术应用研究[M].中国水利水电出版社，1999

11.李于剑.Visual C++ 实践与提高（图形图像编程篇）[M].中国铁道出版社，2001

12.万中英，钟茂生等.鄱阳湖水位动态预测模型[J].江西师大学报（自然科学版），2003；(1)

（上接 38 页）

networks for image analysis and compression[J].Acoustics，Speech and Signal Processing，IEEE Transactions on，1988；36（7）：1169~1179

7.M Porat，Y Y Zeevi.The generalized Gabor scheme of image representation in biological and machine vision[J].Pattern Analysis and Machine Intelligence，IEEE Transactions on，1988；10（4）：452~468

8.T Ebrahimi，M Kunt.Image compression by Gabor expansion[J].Opt Eng，1991；30：873~880

9.A C Bovik，M Clark，W S Geisler.Multichannel texture analysis using localized spatial filters[J].Pattern Analysis and Machine Intelligence，IEEE Transactions on，1990；12（1）：55~73

10.T Saito，H Kudo，S Suzuki.Texture image segmentation by optimal Gabor filters[C].In：Signal Processing，1996，3rd International Conference on，1996：380~383

11.I J Cox，J Kilian，F T Leighton et al.Secure spread spectrum watermarking for multimedia[J].Image Processing，IEEE Transactions on，1997；6（2）：1673~1687

12.M J Bastiaans.A sampling theorem for the complex spectrum，and Gabor's expansion of a signal in Gaussian elementary signals[J].Opt Eng，1981；20：594597

13.S Qian，D Chen，Discrete Gabor Transform[J].Signal Processing，IEEE Transactions on，1993；41（7）：2429~2438

（上接 91 页）

看这种模仿自然生物的新型系统寻优思想无疑具有十分光明的前景，更多深入细致的工作还有待于进一步展开。聚类问题是一个线性整数规划问题，特别当模式样本和分类数很大时，用蚁群算法解决多聚类问题确实有效。对该算法稍加修改，可解决类似的整数规划问题。对于“分类数目未知”的情况，如何用蚁群算法来解，可以做进一步的研究。（收稿日期：2003 年 4 月）

参考文献

1.黄振华，吴诚一.模式识别[M].杭州：浙江大学出版社，1991：40~62

2.蔡元龙.模式识别[M].西安：西北电讯工业出版社，1986：17~32

3.Colorni A，Dorigo M，Maniezzo V.An investigation of some properties of an ant algorithm[C].In：Proc of the Parallel Problem Solving from Nature Conference（PPSN'92），Brussels，Belgium：Elsevier Publishing，1992：509~520

4.张纪会，徐心和.具有变异特征的蚁群算法[J].计算机研究与发展，1999；36（10）：1240~1245

5.马良，项培军.蚂蚁算法在组合优化中的应用[J].管理科学学报，2001；4（1）：32~37

6.谢维信.工程模糊数学方法[M].西安电子科技大学出版社，1991：142~160

（上接 96 页）

参考文献

1.Rajeev Balasubramonian，David Albonest，Alper Buyuktosunoglu et al.Memory Hierarchy Reconfiguration for Energy and Performance in General-Purpose Processor Architectures[J].2000IEEE 0-0765-0924-X/100，2000：245~257

2.G McFarland.CMOS Technology Scaling and its Impact on Cache Delay[D].PhD Thesis.Stanford University，1997-06

3.郑纬民，汤志忠.计算机系统结构[M].第二版，北京：清华大学出版社，1998