蚁群聚类组合方法的研究

邢洁清[1,2],朱庆生[1],郭 平[1]
XING Jie-qing[1,2],ZHU Qing-sheng[1],GUO Ping[1]

1.重庆大学 计算机学院,重庆 400044
2.海南省琼台师范高等专科学校 现代教育技术系,海口 571100
1.Department of Computer Science,Chongqing University,Chongqing 400044,China
2.Department of Modern Education Technology,Qiongtai Teachers College,Haikou 571100,China
E-mail:qqxjq@21cn.com

XING Jie-qing,ZHU Qing-sheng,GUO Ping.Research on ant colony clustering combination method.Computer Engineering and Applications,2009,45(18):146-148.

Abstract: The ant-based clustering algorithm is applicated in the data mining community.Due to the disadvantage of the classical algorithm,this paper presents an improved ant colony clustering combination method.The paper introduces *K*-means to take the ant colony algorithm the pre-computation process.Through *K*-means,it definites cluster center fastly and sketchily,and takes the starting value using the *K*-means method result,again executes the ant colony algorithm cluster. It solves the ant colony algorithm for early slow convergence effectively.

Key words: clustering;ant colony algorithm;pheromone;clustering combination

摘 要:基于蚁群算法的聚类算法已经在当前的数据挖掘研究中得到应用。针对蚁群聚类算法早期出现的缺点,提出一种蚁群聚类组合方法使其得以改进。改进思路是引入 *K*-means 作为蚁群算法的预处理过程。通过 *K*-means 快速、粗略地确定聚类中心,利用 *K*-means 方法的结果作为初值,再进行蚁群算法聚类。有效地解决了蚁群算法早期收敛过慢等问题。

关键词:聚类;蚁群算法;信息素;聚类组合
DOI:10.3778/j.issn.1002-8331.2009.18.044 **文章编号:**1002-8331(2009)18-0146-03 **文献标识码:**A **中图分类号:**TP311

1 引言

蚁群算法是在 20 世纪 90 年代初由意大利学者 M.Dorigo,Maniezzo 等人首先提出来的一种模拟进化的算法[1],其在数据挖掘中的应用正逐步得到人们的关注。蚂蚁等群居类昆虫具有分布式、自组织、信息素通信、合作等性能,模拟这种智能行为的蚁群算法能够解决许多复杂的问题,还可以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的困难问题。蚁群算法在旅行商问题(TSP)、指派问题、调度问题等的求解中,取得了一系列较好的实验结果[2]。此外在一些实际问题的解决中也取得一定进展,如大规模集成电路综合布线以及网络数据包的路由。目前,人工蚁群在知识发现的过程中主要用于挖掘聚类模型和分类模型。

聚类是指将物理的或抽象的集合分成由相似对象所组成的多个类别的过程。聚类是一种无监督的学习,它使得类内对象相似性大,类间对象相似性尽可能小[3]。从实际应用的观点来看,聚类在科学数据探测、图像处理、模式识别、医疗诊断、计算生物学、文档检索、Web 分析等领域起着非常重要的作用,它已经成为当前数据挖掘研究领域中一个非常活跃的研究课题[4]。聚类是一个具有很强挑战性的领域,对聚类的研究始于 20 世纪 60 年代初。经典聚类方法包括分层算法,划分方法如 *K* 均值算法、模糊 *C* 均值算法,图论聚类法,神经网络法,以及基于统计的方法等[5]。近来随着数据挖掘研究的深入,涌现了大量新的聚类算法,如蚁群聚类算法等。目前基于蚁群算法的聚类方法从原理上可以分为 4 种[6]:运用蚂蚁觅食的原理利用信息素来实现聚类、利用蚂蚁自我聚集行为聚类、基于蚂蚁堆的形成原理实现数据聚类以及运用蚁巢分类模型利用蚂蚁化学识别系统进行聚类。

蚁群聚类算法虽具有许多优点,如自治性(聚类不再是根据所要求的对数据进行原始分割和分类[7],而是通过蚁群搜索行为自然地形成)、灵活性(为了避免局部最优不再采用决定性搜索,而是采用随机搜索)、并行性(代理操作是固有的并行),但仍存在一些缺陷,比如算法收敛性差和较长时间的花费。特别是运用蚂蚁觅食的原理利用信息素来实现聚类的蚁群聚类方法,如其信息素的值从 0 或相等值开始,各条路径上的信息素要想明显区别开,一般需要很长时间。

针对其进行改进,来弥补这些缺陷。下面就蚁群聚类算法、

基金项目:国家科技支撑计划项目(the National Science and Technology Support Plan Project of China under Grant No.2007BAH08B04)。
作者简介:邢洁清(1977-),男,讲师,主要研究领域为软件应用、人工智能等;朱庆生(1956-),男,教授,博士生导师,主要研究领域为数据挖掘、虚拟植物等;郭平(1963-),男,副教授,硕士生导师,主要研究领域为数据仓库与数据挖掘、智能信息系统、GIS 等。
收稿日期:2008-04-10 **修回日期:**2008-07-14

K-means 算法及蚁群聚类组合方法进行比较和讨论并进行实验分析。

2　蚁群聚类算法

蚁群聚类算法是一种全局优化的启发式算法[8]，能根据聚类中心的信息量把周围数据归并到一起，从而得到聚类。在将数据视为具有不同属性的蚂蚁，聚类中心是蚂蚁所寻找的“食物源”，那么数据聚类过程就可以看作蚂蚁寻找食物源的过程。

假设数据对象为：$X=\{X|X_i=(x_{i1},x_{i2},\cdots,x_{im})\}$，$i=1,2,\cdots,N\}$，有 N 个输入样本数据。算法首先进行初始化，将各个路径的信息素置为 C，即 $\tau_{ij}(0)=C$(C 可为零)。确定聚类中心的过程就是蚁群从蚁穴出发去寻找食物的过程，蚂蚁在搜索时，不同的蚂蚁选择某个数据元素是相互独立的。d_{ij} 表示 X_i 到 X_j 之间的加权欧氏距离。令 ε 为本次聚类中心与上次的聚类中心的误差，$\tau_{ij}(t)$是 t 时刻数据 X_i 到 X_j 路径上残留的信息量。在路径上的信息量为：

$$\tau_{ij}(t+n)=\rho\cdot\tau_{ij}(t)+\Delta\tau_{ij} \tag{1}$$

信息素的增量为：

$$\Delta\tau_{ij}=\begin{cases}\dfrac{Q}{L_k} & \text{若第 } k \text{ 只蚂蚁在本次循环经过 } ij\\ 0 & \text{否则}\end{cases} \tag{2}$$

式中 Q 是一个常量，用来表示蚂蚁完成一次完整的路径搜索后，所释放的信息素总量；L_k 是第 k 个蚂蚁环流一周的路径长度。如果长度值越高，其在单位路径上所释入的信息素浓度越低。

X_i 是否归并到 X_j 由转移概率给出：

$$p_{ij}(t)=\frac{\tau_{ij}^{\alpha}(t)\eta_{ij}^{\beta}(t)}{\sum_{s\in S}\tau_{sj}^{\alpha}(t)\eta_{sj}^{\beta}(t)} \tag{3}$$

其中 S 是蚂蚁 X_s 下一步可以选择的路径集合。如果 $p_{ij}(t)$大于阈值 p_0，就将 X_i 合到 X_j 的领域内。这里 η_{ij} 称为边弧(i,j)的能见度。α,β 为调节因子，起到既防止所有蚂蚁均沿相同路径得到相同结果所产生的停滞搜索，又再现了经典的贪心算法。令 C_j 表示所有归并到 X_j 领域的数据集合。求出理想的聚类中心：

$$\overline{C_j}=\frac{1}{J}\sum_{k=1}^{J}X_k,X_k\in C_j \tag{4}$$

将 C_j 加入数据样本空间 X，初始化蚁群的初始位置进行重新的迭代搜索，直到达到规定的迭代次数或是本次聚类中心与上次的聚类中心的误差值达到规定的范围 ε。

蚁群聚类算法存在的问题：

(1)算法效率：蚁群聚算法的收敛过程比较缓慢。特别是在迭代初期，由于系统参数改变很慢，导致整个计算过程非常耗时。在基于蚂蚁觅食原理的聚类分析中，对各路径上的信息素的初始化，为简化操作，一般全都赋为相等的常量 C(通常为0)。因此，信息素的值从相等常量 C 开始，各条路径上的信息素要想明显区别开，一般需要很长时间。

(2)初始值的选择：初值的选择对聚类的最终结果影响很大。而在经典蚁群算法中，确定初始参数时，一般缺乏已知的指导经验。初始参数的确定带有很大的盲目性。该聚类方法中 α、β 的选择对算法运行效率和聚类结果影响较大，选择不当将影响算法执行效率和效果，所需时间增长等缺点。可以根据情况尝试不同的方法避免算法陷于局部最优。

3　K-means 聚类算法

K-means 算法是基于划分的聚类方法，也是最常用的聚类算法。该算法不断计算每个聚类的中心，也就是聚类中对象的平均值，作为新的聚类种子。

K-means 算法具体描述如下[9]：

(1)随机选择 k 个对象作为初始的聚类种子；

(2)根据聚类种子的值，将每个对象重新赋给最相似的簇；

(3)更新聚类种子，即重新计算每个簇中对象的平均值，用对象均值点作为新的聚类种子；

(4)重复执行(2)和(3)两步，直到各个簇不再发生变化。

K-means 算法试图找出使平方误差函数值最小的 k 个划分。当结果簇密集并且各簇之间的区别明显时，它的效果较好。处理大数据集时，K-means 算法具有较好的可伸缩性和高效率。

K-means 聚类算法存在的问题是：当结果簇密集并且各簇之间的区别明显时，它的效果较好。但区别不明显时则效果较差。该算法的缺点是必须事先给出要生成的聚类数目。

4　基于 K-means 的蚁群聚类算法

通过实验验证可知，K-means 算法收敛速度比蚁群聚类算法快，因而引入 K-means 作为蚁群算法的预计算过程。通过 K-means 快速、粗略地确定聚类中心，即“食物源”。K-means 预处理过程结束后，根据现有的聚类结果，获得每只蚂蚁的标签情况，确定各路径的信息素。对于样本数据 X_i，若经过 K-means 后，其属于聚类中心 C_j，则 X_i 到 C_j 的最短路径上信息素相对分配较多。

根据 K-means 作为蚁群算法的预计算得到的聚类结果确定各路径上的信息素，使得在下一步使用蚁群算法时蚂蚁 X_i 其信息素初值取值是不同的。

因此引入 K-means 作为预计算求解聚类问题的基本蚁群算法(AOC)做为一种蚁群聚类组合方法(KMAOC)如下：

步骤 1　任选 k 个初始聚类中心：$C_1,C_2,C_3,\cdots,C_K$；

步骤 2　逐个将数据集$\{X\}$中各个数据对象按最小距离原则分配给 k 个聚类中心的某一个 C_j；

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

步骤 4　若 $C_j'\neq C_j(j=1,2,\cdots,k)$且未快速分类到设定聚类效果阈值 γ 或是指定次数时转步骤 2；

步骤 5　$nc\leftarrow 0$(nc 为循环次数)，由 K-means 算法分类结果计算出的聚类中心 $C_j(j=1,2,\cdots,k)$，计算每个样本数据 X_i 相对应的到不同聚类中心 C_j 的初值 $\tau_{ij}(0)(i,j=1,2,\cdots,N)$。给出 Q、ρ(信息素持久)、n(蚂蚁数)的值，随机给出分配方案；

步骤 6　对每个蚂蚁按转移概率 $p_{ij}(t)$选择下一个节点；

步骤 7　计算新的聚类中心，计算每个样本数据到新的聚类中心的距离。由蚂群聚类公式修改信息素强度 $\tau_{ij}(t)$；

步骤 8　若 $nc<$预定的迭代次数且无退化行为(即找到的都是相同的解)，则输出最好的解；否则转步骤 6。

5　算法测试

实验数据取于 UCI 机器学习数据库的 Iris 和 Wine 及 Balance 数据集。这些数据库有自己的分类，可用于聚类性能的评价。

表 1 数据库描述

数据库名称	Iris	Wine	Balance
数据大小	150	178	625
属性个数	4	13	4
分类数目	3	3	3

对于聚类算法的性能评价通常采用外部和内部两种,其依据是有无关于数据集的先验知识。本文采用外部评价 F-measure 方法[10]以及总的运行时间对提出的聚类算法与 K 均值算法和标准蚁群算法进行评价和比较。F-measure 的取值在[0,1]之间,取值越接近 1 越好。

表 2 给出了 3 种算法 3 个数据集的 F-measure 和算法运行时间的比较结果(取 100 次实验的平均值)。算法参数如下:α=1,β=5,ρ=0.99,Q=80,蚂蚁数 m=60。实验结果表明:对于聚类数目比较明显的数据,3 种算法都可以达到较好的 F-measure。相比而言,聚类组合法的 F-measure 略高于标准蚁群算法,K 均值算法最次。而由于聚类组合法前期收敛性较标准蚁群算法前期有所改进,因而在总的时间花费上有所减少。

表 2 3 种算法的 F-measure 与算法运行时间

算法	Iris		Wine		Balance	
	F-measure	Runtime①	F-measure	Runtime①	F-measure	Runtime①
K-means 算法	0.897	1.116	0.685	1.089	0.548	1.072
标准蚁群算法	0.911	1.375	0.706	1.138	0.570	1.104
KMAOC 算法	0.913	1.000	0.719	1.000	0.595	1.000

注①:在同一台计算机上运行以 KMAOC 算法为标准时间,取值为 1。其他算法的 Runtime 取值为其他算法实际运行时间/KMAOC 算法实际运行时间得出的比值。

在实验中发现了改进后算法的缺点:在 K-means 阶段限制了聚类的个数。只适用于聚类个数明确的聚类问题。故对聚类个数不明确的聚类问题,作如下改进:根据实际问题的应用背景,确定多个 K,比较各中情况下的聚类效果,选择聚类效果最好的。

6 结束语

提出了一种引入 K-means 作为预处理过程的蚁群算法(KMAOC),该方法通过 K-means 快速、粗略地确定聚类中心,即"食物源"。根据现有的聚类结果,即每只蚂蚁的标签情况,确定各路径的信息素。改进后算法算法效率提高。初始阶段使用 K-means,快速、粗略地确定了初始参数,信息素不必从 0 及相等常量开始训练。避免了经典蚁群算法初始阶段学习缓慢的缺点。使得初始值的选择具有更多的可参考的指导经验,减小了确定初始参数的盲目性。

参考文献:

[1] Russo F,Ramponi G.A fuzzy filter for images corrupted by impulse noise[J].IEEE Signal Processing Lett,1996,3(6):168-170.

[2] 赵宝江,李士勇,金俊.基于自适应路径选择和信息素更新的蚁群算法[J].计算机工程与应用,2007,43(3):12-15.

[3] Chen M S.Data mining:An overview from a database perspective[J].IEEE Trans on Knowledge and Data Engineering,1996,8(6):866-833.

[4] Handl J,Knowles J,Dorigo M.On the performance of ant-based clustering[C]//Proc of the 3rd Int Conf on Hybrid Intelligent Systems.Australia:IOS Press,2003-12.

[5] Han Jia-wei,Kamber M.数据挖掘:概念与技术[M].北京:机械工业出版社,2001:223-251.

[6] 张建华,江贺,张宪超.蚁群聚类算法综述[J].计算机工程与应用,2006,42(16):171-174.

[7] Topchy A,Jain A K,Punch W.A mixture model of clustering ensembles[C]//SIAM Intl Conf on Data Mining.Orlando: ACM Press,2004:379-390.

[8] 刘念涛,刘希玉.基于改进的启发式蚁群算法的聚类问题的研究[J].计算机技术与发展,2007(8):37-39.

[9] 张群,熊英,黄庆炬.基于蚁群算法的数据挖掘方法研究[J].湖北工业大学学报,2007,22(2):5-9.

[10] Rijsbergen C.Information retrieval[M].2nd ed.Butterworths,London,UK:[s.n.],1979:99.

(上接 111 页)

首数目,从而大大降低了节点的总能耗,并且在一定程度上使节点的能耗更加均匀,进一步延长了网络的生命周期。

参考文献:

[1] 任丰原,黄海宁,林闯.无线传感器网络[J].软件学报,2003,14(7):1282-1291.

[2] Heinzelman W B.An application-specific protocol achritecture for wireless microsensor networks[J].IEEE Transactions on Wireless Communicaions,2002,1(4):660-670.

[3] Younis O,Fahmy S.HEED:A hybrid energy-efficient distributed clustering approach for ad hoc sensor networks[J].IEEE Trans on Mobile Computing,2004,3(4):660-669.

[4] Manjeshwar A,Agrawal D P.TEEN:A routing protocol for enhanced efficiency in wireless sensor networks[C]//Proceedings of 15th Parallel and Distributed Processing Symposium.[S.l.]:IEEE Computer Society,2001:2009-2015.

[5] Raicu I,Schwiebert L.e3D:An energy-efficient routing algorithm for wireless sensor networks[C]//The International Conference on Intelligent Sensors,Sensors Networks and Information Processing,ISSNIP 2004,Melbourne,Australia,December 2004.

[6] Heinzelman W,Chandrakasan A,Balakrishnan H.Energy-efficient communication ptotocol for wireless microsensor networks[C]//Proceedings of the 33th Hawaii International Conference on System Science,2000.

[7] Youni M,Youssef M,Arisha K.Energy-aware routing in cluster-based sensor networks[C]//Proceedings of the 10th EEE/ACM Int'l Symp on Modeling,Analysis and Simulation of Computer and Telecommunication Systems,2002:129-136.

[8] Heinzelman W R,Chandrakasan A,Balakrishnan H.Energy-efficient communication protocol for wireless microsensor networks[C]//Proc of the Hawaii International Conference on System Sciences,Maui,Hawaii,2000.

[9] Lindsey S,Raghavendra C S.PEGASIS:Power-efficient gathering in sensor information systems[C]//Aerospace Conference Proceedings,IEEE,2002,3.