蚁群聚类算法综述

张建华[1,2]　江　贺[1]　张宪超[1]
[1](大连理工大学软件学院,大连　116621)
[2](阜阳师范学院计算机系,安徽阜阳　236032)
E-mail:jianhuazhang2008@163.com

摘　要　数据聚类是重要的数据挖掘技术,在工程和技术等领域具有广泛的应用背景。蚁群算法作为一种新型的优化方法,具有很强的鲁棒性和适应性。文章着重介绍蚁群聚类算法的研究情况,阐述当今流行的蚁群聚类算法的基本原理及其特性,旨在为蚁群聚类算法的发展提供引导作用。

关键词　数据挖掘　蚁群算法　聚类

文章编号　1002-8331-(2006)16-0171-04　**文献标识码**　A　**中图分类号**　TP301

Survey of Ant Colony Clustering Algorithms

Zhang Jianhua[1,2]　Jiang He[1]　Zhang Xianchao[1]
[1](School of Software, Dalian University of Technology, Dalian 116621)
[2](Department of Computer, Fuyang Normal College, Fuyang, Anhui 236032)

Abstract: Clustering is an important technique of data mining.It is widely used in fields of engineering and technology. Ant colony algorithms are robust and adaptable as novel optimization methods.This paper emphatically introduces the research of ant colony clustering algorithms, and describes the basic principle and characteristics of existing popular ant colony clustering algorithms.It affords direction for the future work of ant colony clustering algorithms.

Keywords: data mining, ant colony algorithm, clustering

1　引言

聚类分析是数据挖掘领域中的一个重要分支[1],是人们认识和探索事物之间内在联系的有效手段,它既可以用作独立的数据挖掘工具,来发现数据库中数据分布的一些深入信息,也可以作为其他数据挖掘算法的预处理步骤。所谓聚类(clustering)就是将数据对象分组成为多个类或簇(cluster),在同一个簇中的对象之间具有较高的相似度,而不同簇中的对象差别较大。传统的聚类算法主要分为四类[2,3]:划分方法,层次方法,基于密度方法和基于网格方法。

受生物进化机理的启发,科学家提出许多用以解决复杂优化问题的新方法,如遗传算法、进化策略等。1991 年意大利学者 A.Dorigo 等提出蚁群算法,它是一种新型的优化方法[4]。该算法不依赖于具体问题的数学描述,具有全局优化能力。随后他和其他学者[5~7]提出一系列有关蚁群的算法并应用于复杂的组合优化问题的求解中,如旅行商问题(TSP)、调度问题等,取得显著的成效。后来其他科学家根据自然界真实蚂蚁群堆积尸体及分工行为,提出基于蚂蚁的聚类算法[8,9],利用简单的智能体模仿蚂蚁在给定的环境中随意移动。这些算法的基本原理简单易懂[10],已经应用到电路设计、文本挖掘等领域。本文详细地讨论现有蚁群聚类算法的基本原理与性能,在归纳总结的基础上提出需要完善的地方,以推动蚁群聚类算法在更广阔的领域内得到应用。

2　聚类概念及蚁群聚类算法

一个簇是一组数据对象的集合,在同一个簇中的对象彼此类似,而不同簇中的对象彼此相异。将一组物理或抽象对象分组为类似对象组成的多个簇的过程被称为聚类。它根据数据的内在特性将数据对象划分到不同组(或簇)中。聚类的质量是基于对象相异度来评估的,相异度是根据描述对象的属性值来计算的,距离是经常采用的度量方式。聚类可用数学形式化描述为:设给定数据集 $X=\{x_1, x_2, \cdots, x_n\}$, $\forall i\in\{1, 2, \cdots, n\}$, $x_i=\{x_{i1}, x_{i2}, \cdots, x_{ip}\}$ 是 X 的一个对象, $\forall l\in\{1, 2, \cdots, p\}$, x_{il} 是 x_i 对象的一个属性。根据数据的内在特性将 X 分解成 $C=\{C_1, C_2, \cdots, C_k\}$。其中 $\bigcup_{i=1}^{k} C_i=X$, $\forall i, j\in\{1, 2, \cdots, k\}$, $C_i\neq\phi$, $C_j\neq\phi$, 且 $(C_i\wedge C_j=\phi)$ $(i\neq j)$。$K=\{X, C\}$ 称为一个聚类空间, C_i 称为聚类空间的第类(簇)。

在数据挖掘中,聚类是一个活跃的研究领域[11],涉及的范围从社会学、心理学、生物学到计算机科学。存在多种聚类方法,这些方法不仅算法原理(决定运行时间及可测量性)不同,而且许多基本特性也不相同,例如处理的数据对象,有关簇形状的设想,最终划分的形式或必须提供的参数等。

计算机科学家通过模仿生物行为已经提出一系列解决问题的新颖的成功方法。1991 年 Deneubour 等介绍了基于蚂蚁的聚类和分类[9]方法,当时主要用于机器人作业调度中。后来 Lumer 等[8]修改了这个算法并将之应用于对数字数据分析上。

作者简介:张建华(1978-),男,硕士生,主要研究领域为聚类分析,算法分析与设计。江贺(1980-),男,博士,讲师,主要研究领域为分布式算法设计,无线传感器网络路由,数据挖掘等。张宪超(1971-),男,博士,副教授,主要研究领域为组合优化,算法分析与设计,并行分布式计算等。

后来应用于数据挖掘[12],图像分割[13]和文本挖掘中[14]。2002 年 Labroche 等提出基于蚂蚁化学识别系统的聚类方法。总的说来,基于蚁群算法的聚类方法从原理上可以分为四种:(1)运用蚂蚁觅食的原理,利用信息素来实现聚类[15];(2)利用蚂蚁自我聚集行为聚类;(3)基于蚂蚁堆的形成原理实现数据聚类;(4)运用蚁巢分类模型,利用蚂蚁化学识别系统进行聚类的。

3 算法分析

3.1 基于蚂蚁觅食的聚类算法

蚂蚁的觅食过程可以分为搜索食物和搬运食物两个环节[16]。每个蚂蚁在运动过程中都会在其经过的路径上释放信息素,并能够感知信息素及其强度。经过蚂蚁越多的路径其信息素越强,同时信息素自身也会随着时间的流逝而挥发。蚂蚁倾向于信息素强度高的方向移动,某一路径上走过的蚂蚁越多,后来的蚂蚁选择该路径的概率就越大,整个蚁群的行为表现出信息正反馈现象。基于蚂蚁信息素痕迹的聚类分析基本思想如下:

将数据视为具有不同属性的蚂蚁,聚类中心是蚂蚁所要寻找的“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[17]。假设数据对象为:$X=\{X|X_i=(x_{i1},x_{i2},\cdots,x_{im}),i=1,2,\cdots,N\}$,算法首先进行初始化,将各个路径的信息素置为 0,即 $\tau_{ij}(0)=0$,设置簇的半径为 r,统计误差为 ε 等参数。计算对象 X_i 到 X_j 之间的加权欧氏距离 d_{ij},计算各路径上的信息素 $\tau_{ij}(t)$:

$$\tau_{ij}(t)=\begin{cases}1 & d_{ij}\le r\\ 0 & d_{ij}>r\end{cases} \tag{1}$$

对象 X_i 合并到 X_j 的概率为:

$$p_{ij}(t)=\frac{\tau_{ij}^{\alpha}(t)\eta_{ij}^{\beta}(t)}{\sum_{s\in S}\tau_{sj}^{\alpha}(t)\eta_{sj}^{\beta}(t)} \tag{2}$$

其中 $S=\{X_s|d_{sj}\le r,s=1,2,\cdots,j,j+1,\cdots,N\}$。如果 $p_{ij}(t)$ 大于阈值 p_0,就将 X_i 合并到 X_j 的领域内。这里 η_{ij} 是 d_{ij} 的倒数,称它为能见度。α,β 为调节因子[18],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思想。

该聚类方法中的 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[19]。可以根据情况尝试不同的方法避免算法陷于局部最优。算法虽然不需要预先指定簇的数目,但是由于簇的半径是预置的,所以聚类的规模受到限制。另外在实际计算中,在给定循环次数的条件下很难找到最优解。再者信息素分配策略、路径搜索策略、最优解保留策略等方面均带有经验性和直觉性,导致算法的求解效率不高,收敛性差。

3.2 基于蚂蚁自我聚集行为的聚类算法

蚂蚁能够通过自我聚集行为构建一个树状结构,称之为蚂蚁树(AntTree)[20]。用蚂蚁表示数据并代表该树的节点,初始时蚂蚁放在一个称为支点的固定点上,这个点相当于树根。蚂蚁在这棵树上或已经固定在树上的蚂蚁身上移动,来寻找适合自己的位置。假设蚂蚁能够到达树的任何地方并能粘在该结构的任何位置,不过在结构树形成的过程中受对象间的作用,蚂蚁更趋于固定在树枝的末端[21,22]。树的局部结构及蚂蚁表示的数据之间的相似性引导它的移动,当所有蚂蚁都在树上固定下来后,算法结束,获得对数据集的划分。

为了更好描述算法过程,采用蚂蚁表示的数据代表树中的每个节点,用欧氏距离作为相似度尺度,用 $Sim(i,j)$ 表示。相似度公式为:

$$Sim(i,j)=1-\sqrt{\frac{1}{M}\sum_{k=1}^{M}(v_{ik}-v_{jk})^2} \tag{3}$$

其中 M 表示每个数据对象的属性值,v_{ik} 表示数据对象 i 的第 k 个属性。每对数据 (d_i,d_j),$i\in[1,N],j\in[1,N]$ 的相似度值 $Sim(i,j)$ 在 $[0,1]$ 之间(N 表示数据集内的对象数),意味着 d_i,d_j 完全不同,表示它们相同。AntTree 主要原理如下:

节点 a_0 表示树根,蚂蚁逐步连接到这个初始节点上或连接到固定在该节点的蚂蚁上,直到所有的蚂蚁均连接到结构上(蚂蚁树的停止标准)。移动的蚂蚁 a_i 根据 $Sim(i,j)$ 值和它的局部邻居决定自身的位置。每只蚂蚁 a_i 只有一个父亲结点,最多有 L_{max} 个孩子结点。对每只蚂蚁 a_i 都定义一个相似度阈值 $T_{Sim(a_i)}$ 和相异度阈值 $T_{Dissim(a_i)}$,并且由 a_i 进行局部更新,用来判断蚂蚁 a_i 表示的数据 d_i 与其它蚂蚁表示的数据的相似或相异程度。

蚂蚁的局部行为:第一只蚂蚁直接连接到 a_0 上,对后来的蚂蚁 a_i 要考虑两种情况:第一种情况是 a_i 在支点上。设 a^+ 为支点上且与 a_i 最相似的蚂蚁,如果 a_i 和 a^+ 足够相似 $Sim(a_i,a^+)\ge T_{Sim(a_i)}$,那么 a_i 向 a^+ 移动,使它们能尽可能的聚集在同一子树上,即在同一个簇内。否则(如 a_i 与 a^+ 不相似)如果 a_i 与 a^+ 足够相异 $Sim(a_i,a^+)<T_{Dissim(a_i)}$,那么它就连接在支点上,意味着创建了一棵新子树,该子树上的蚂蚁将尽可能的与以 a_0 为根的其他子树上的蚂蚁不同。如果 a_0 已经有 L_{max} 孩子结点,那么 a_i 向 a^+ 移动。假如 a_i 和 a^+ 既不足够相似也不足够相异,则用 $T_{Sim(a_i)}\leftarrow T_{Sim(a_i)}*\alpha$ 和 $T_{Dissim(a_i)}\leftarrow T_{Dissim(a_i)}+\beta$ 来更新阈值,增加 a_i 下次连接的概率。α,β 为调节因子,实验中通常选择 0.9 与 0.01,因为它能提供更好的结果[11]。由于分布的相似性,相似性阈值的减少速度明显高于相异性阈值的增加速度。

第二种情况 a_i 在蚂蚁 a_{pos} 上移动(a^+ 表示 a_{pos} 上与 a_i 最相似的蚂蚁)。如果 a_{pos} 的孩子结点少于 L_{max} 且 a_i 与 a_{pos} 足够相似($Sim(a_i,a_{pos})\ge T_{Sim(a_i)}$),与 a_{pos} 上其他蚂蚁足够相异(ie.$Sim(a_i,a^+)<T_{Dissim(a_i)}$),那么 a_i 就连接在 a_{pos} 上,否则蚂蚁 a_i 随机地向 a_{pos} 的邻居移动。根据需要按前面的方式更新阈值,寻找合适的位置,当所有蚂蚁都连接好时算法结束。

利用该算法得到的簇更接近于数据的真实分类,并且当蚂蚁连接上以后就不再移动,所以平均执行时间相当低。但是算法的初始化(a_0 的选取)很重要,它影响整个算法的质量。此外对阈值的更新策略也是影响该算法的最重要并且最难确定的因素。

3.3 基于蚂蚁堆形成原理实现聚类

蚁堆聚类方法的基本机制是工蚁堆积蚂蚁尸体过程,小蚁堆不断吸引工蚁堆积更多的死蚂蚁,通过正反馈导致蚁堆逐渐增大。基于这种模型主要有下面几种算法:

3.3.1 基于智能体模型的算法

1991 年 Deneubourg[23]等提出基于智能体(人工蚂蚁)的模

型来模仿蚂蚁行为对数据进行聚类,人工蚂蚁沿着网格单元移动,每个单元只含一个对象。没有搬运数据对象的蚂蚁碰到对象时就会以某个概率拾起它,这个概率依赖对该对象周围的相同对象密度的评估,如果密度高则拾起的概率就低;携带对象的蚂蚁遇到空单元或搬运的对象与邻近的对象相似时就会以某个概率放下它,放下的概率也依赖对周围对象类型密度的评估,如果密度大时放下的概率就高,结果相同类型的对象都被聚集在一起。数据对象在空间的分布状态将影响聚类结果,其基本思想如下。

假设所有的数据对象都随机地分布在二维的与数据集成比例伸缩地网格空间[24]。处于网格中的两个对象 o_i 和 o_j 之间的距离(相似度)$d(o_i,o_j)$ 是它们的欧氏距离,如果 o_i 和 o_j 是同一类对象(即 o_i 和 o_j 相似),则 $d(o_i,o_j)=0$,反之 $d(o_i,o_j)=1$,从而得到二进制的相似度矩阵。设若干个蚂蚁在二维网格上不断运动并反复执行拾起或放下对象操作,蚂蚁某时刻在 r 位置发现对象 o_i,则它的局部密度可以由下面的公式来计算:

$$f(o_i)=\begin{cases}\frac{1}{s^2}\sum_{o_j}[1-\frac{d(o_i,o_j)}{2}] & \text{if } f>0\\ 0 & \text{otherwise}\end{cases} \tag{4}$$

其中 $f(o_i)$ 为相似度密度,$o_j\in Grid_{(s\times s)}(r)$,单元 r 的邻域面积为 $s\times s$,α 为相异度因子。蚂蚁在运动过程中拾起对象的概率 $p_p(o_i)$ 和放下对象的概率 $p_d(o_i)$ 分别为:

$$p_p(o_i)=\left(\frac{k_1}{k_1+f(o_i)}\right)^2 \tag{5}$$

$$p_d(o_i)=\begin{cases}2f(o_i) & \text{if } f(o_i)<k_2\\ 1 & \text{if } f(o_i)\geq k_2\end{cases} \tag{6}$$

其中 k_1,k_2 都是阈值常量。

该算法实际上是一种基于网格和密度的聚类方法[16]。为了便于处理高维数据空间,首先将其映射到某一低维网格空间,映射要确保簇内距离小于簇间距离,同时网格的精细度将会影响聚类质量。蚂蚁拾起或放下对象受局部相似度密度 $f(o_i)$ 的影响,局部相似度密度大,拾起的概率 $p_p(o_i)$ 小,数据不易从该簇中移走,同时放下的概率 $p_d(o_i)$ 大,对象倾向于留在该簇中,反之亦然。该算法不必预先指定簇的数目,并能构造任意形状的簇。

3.3.2 基于蚂蚁的聚类

基于蚂蚁的聚类(ANT-BASED CLUSTERING)[25]也是受蚁卵分类启发的,实际上是对上述算法的改进,主要过程与前面相似。主要修改了几个重要的特性,首先修改了局部相似度密度,此处的邻域密度函数为 $\hat{f}(o_i)$:

$$\hat{f}(o_i)=\begin{cases}\frac{1}{s^2}\sum_{o_j}[1-\frac{d(o_i,o_j)}{2}] & \text{if } t=\text{true}\\ 0 & \text{otherwise}\end{cases} \tag{7}$$

其中:

$$t=\hat{f}(o_i)>0\wedge\forall_{o_j}\left(1-\frac{d(o_i,o_j)}{2}\right)>0 \tag{8}$$

这样拾起或放下的概率公式变为:

$$p_p(o_i)=\begin{cases}1.0 & \text{if } \hat{f}(o_i)\leq 1.0\\ \frac{1}{f(o_i)^2} & \text{otherwise}\end{cases} \tag{9}$$

$$p_d(o_i)=\begin{cases}1.0 & \text{if } \hat{f}(o_i)\geq 1.0\\ \hat{f}(o_i)^4 & \text{otherwise}\end{cases} \tag{10}$$

从(4)和(7)可以看出它对网格单元进行了处理,这样有助于对象形成的簇更加紧凑。此外约束条件 $\forall o_j(1-d(o_i,o_j)/\alpha)>0$ 能处理带有非常高的相异点的邻居,它增大了两个簇的分离空间。

其次,用了一个短期寄存器记下前面放下对象的位置,当拾起一个对象时可以参考寄存器来指导它移动,这样能向最近放下相似对象的位置移动,降低时间复杂度。再次,在计算 $\hat{f}(o_i)$ 时考虑增加感知半径来扩大邻域,把一些小簇合并成一个大簇。最后,对相异度因子 α 作适当的修改。适当的选择相异度因子很重要,选择太小的 α 就不能形成簇,太大时会导致过多的簇合并,极限情况下所有对象聚集在一个簇中。选择适当的 α 主要依靠同一个簇中每对对象间的相异度,因为 α 影响蚂蚁的拾起/放下行为,所以可以通过跟踪蚂蚁的行为获得自动适应的 α,从而调整 α 的值。

3.3.3 混合聚类方法

上面虽然对邻域密度函数进行了修改,但最终聚类的数目往往太高,而且收敛很慢。于是 2000 年 Monmarche 建议一个单元可以放置多个对象,每个拥有物品的单元相当于一个簇。每只蚂蚁 a_i 具有的承载能力为 $c(a_i)$,代替过去一只蚂蚁一次只能搬运一个对象的情况。概率性的从一个堆上最多拾起对象为 $c(a_i)$,并且往堆上丢下对象时要根据堆的特性,如堆中对象间的平均相异度。当蚂蚁决定拾起对象时,挑选与堆中心相异度最大的对象。此时蚂蚁 a_i 的运载能力 $c(a_i)$ 有两种情况 $c(a_i)=1$ 或 $c(a_i)=\infty$,即可以搬运一个对象,也可以是整个堆。Monmarche 建议结合 K-means 方法进行聚类,主要过程如下:

最初同样利用基于蚂蚁算法来形成初始的簇。由于划分时间太长,所以在算法收敛之前就终止了算法,致使创建的划分存在错误划分,所以使用 k-means 算法除去小的分类错误,并分配‘自由’对象,即算法停止时仍然有单独存在单元上的对象或蚂蚁仍在搬运的对象。这虽然能除去分类中的错误,但是 k-means 算法是局部优化算法,不能得到高质量的簇,所以需要再次利用基于蚂蚁的算法。

这次是对对象堆上而不是单个对象运用算法。前面基于蚂蚁的算法同样适用于堆,这些堆可以像单个对象那样再次被拾起或放下,再次构成新的簇。但像前面那样仍然有未分配的堆,因而再次使用 k-means 算法来获得最终的划分。这次因为提供给 k-means 的输入已经很接近最佳了,所以输出的结果质量很高。与这个方法相似的另一种是与模糊 c-means 相结合的方法[24],基于相对简单智能体直接或间接的反馈完成聚类。

3.4 基于蚂蚁化学识别系统的聚类方法

现实中的蚂蚁为了保护自己的巢穴不被敌人或食客攻击破坏,必须具有区别伙伴和敌方的能力,它是靠识别群体间的气味来实现的。当两只蚂蚁相遇时分别检查对方表皮所散发的气味(也叫标签),并与自身的模板比较。模板是蚂蚁在幼年时期获得的,并在成长过程中不断更新。标签是由蚂蚁基因及蚂蚁间不断交换化学物质决定的。同伴间通过不断交换化学物质建立群体气味,该气味可以被每个伙伴识别,不同群体具有不同的气味,同一群体分享相同的气味,这就是所谓的‘群体圈’,也是化学识别系统的基本原理。

2002 年 Labroche 等提出叫做蚂蚁簇(AntClust)的聚类方法,主要是利用化学识别系统原理来聚类的。为了更好地说明这个方法,给每只蚂蚁 a_i 定义如下参数:由蚂蚁巢穴的属性决定的标签 $Label_i$,可以代表巢,开始蚂蚁不受任何巢穴的影响,所以 $Label_i=0$, 随后标签不断变换直到蚂蚁找到最好的巢为止。模板是由蚂蚁的基因 $Genetic_i$ 和接受阀值 $Template_i$ 组成的,前者对应于数据集的对象且在算法过程中不断变化,后者是在初始化阶段获得的,它是蚂蚁与其它蚂蚁相遇期间观察到的最大相似度 $Max(Sim(i,.))$ 和平均相似度 $\overline{Sim(i,.)}$ 的函数,它是动态的,蚂蚁每次和其它蚂蚁相会后对它进行修改。

$$Template_i \leftarrow \frac{\overline{Sim(i,.)}+Max(Sim(i,.))}{2} \quad (11)$$

评价因子 M_i 反映蚂蚁间的相遇情况。相同标签的蚂蚁相遇 M_i 增加,反之 M_i 减少,开始时 $M_i=0$,它反映蚂蚁 a_i 所在巢穴的规模。M_i^+ 表示蚂蚁被接受程度,如果具有相同标签的蚂蚁相遇或两只蚂蚁彼此接受对方,M_i^+ 增加,否则蚂蚁不接受时减少。蚂蚁接受与否可根据下面公式判断, 设 a_i, a_j 分别表示两只蚂蚁,则:

$$Accept\ an\ ce(a_i,a_j) \Leftrightarrow (Sim(a_i,a_j)>Template_i) \wedge (Sim(a_i,a_j)>Template_j) \quad (12)$$

蚂蚁簇算法主要是反复随机选择两只蚂蚁并模仿它们相遇过程:

(1)当两只都没有巢的蚂蚁相遇并彼此接受时将创建一个新的巢(初始簇),并作为‘种子’聚集相似蚂蚁以便产生最终的簇;

(2)当有巢的蚂蚁遇到可以接受没有巢的蚂蚁时,没有巢的蚂蚁加入该巢内,通过加入相似蚂蚁来扩大现存的簇;

(3)属于同一个巢的蚂蚁在接受的情况下增加评价因子和,使巢变得更健壮;

(4)当两个同伴相遇且不能彼此接受,则整体最差的蚂蚁将从巢中被驱除出去,这样通过除去不理想的蚂蚁,使巢变得更完美;

(5)不同巢的蚂蚁相遇且彼此接受时,合并它们的巢,即合并相似的簇,小簇被大簇吸收。算法结束时蚂蚁集中在有限数目的巢内,巢就是期望得到的划分。

反复利用这些过程,能得到最终想要的划分。该算法能处理任意类型的数据,具有很好的鲁棒性和适应性。但是如何确定迭代次数保证算法收敛有待进一步研究。另外巢的删除阀值直接影响到聚类结果的稳定性,目前此阀值的设定尚未有效的解决,尽管有些文献对这作了修改,但没有足够的理论依据,不可靠。

4 总结

本文分析了现在流行的有关蚁群的聚类算法,针对每种方法分别从它的基本思想、聚类原理及主要步骤上进行了论述与分析。它们的不同之处主要在蚂蚁个体间的通信介质不同,有的是根据气味有的完全避开气味。根据气味通信的又分为两种:(1)根据其所经路径上留下的信息素,如基于蚂蚁觅食的聚类方法;(2)根据蚂蚁自身携带的气味,如基于蚂蚁化学识别系统的聚类方法。另外是根据对象的空间分布状态指导蚂蚁间的相互作用完成聚类的,如基于蚂蚁堆的形成原理的聚类方法都是根据对象在网格上的分布情况实现的;其中基于蚂蚁的混合聚类方法,交替使用蚁群聚类方法和 k-means 算法,加速算法收敛并提高了聚类质量。蚁群聚类方法具有许多特有的特性,如灵活性、健壮性、分布性和自组织性等,这些特性使其非常适合本质上是分布、动态及又要交错的问题求解中,能解决无人监督的聚类问题,具有广阔的前景。但仍存在很多问题需要解决,有待进一步研究。(收稿日期:2005 年 9 月)

参考文献

1.Chen MS.Data mining: an overview from a database perspective[J].IEEE Trans on Knowledge and data engineering, 1996; 8(6): 866~883

2.P Berkhin.Survey of clustering data mining techniques[R].Technical report, Accure Software, San Jose, CA, 2002

3.钱卫宁等.从多角度分析聚类算法[J].软件学报, 2002; 13(8)

4.A Dorigo, M Dorigo, V Maniezzo.Distributed optimization by ant colonies[C].In: European Conference on Artificial Life, 1991: 134~142

5.M Dorigo et al.Ant system: optimization by a colony of cooperating agents[J].IEEE Transactions on Systems, Man, and Cybernetics, Part B, 1996; 26(1): 29~41

6.M Dorigo, L M Gambardella.Ant colony system: a cooperative learning approach to the traveling salesman problem[J].IEEE Transactions on Evolutionary Computation, 1997; 1(1): 53~66

7.M Dorigo et al.Guest editorial: special section on ant colony optimization[J].IEEE Transactions on Evolutionary Computation, 2002; 6(4): 317~319

8.E Bonabeau, M Dorigo, G Theraulaz.Swarm Intelligence- From Natural to Artificial System[M].New York, NY: Oxford University Press, 1999

9.J-L Deneubourg, S Goss, N Franks et al.The dynamics of collective sorting: Robot-like ants and ant-like robots[C].In: J-A Meyer, S Wilson eds.Proceedings of the First international Conference on Simulation of Adaptive haviour, From Animals to Animals J, MIT Press, Cambridge MA, 1991: 356~365

10.E Lumer, B Faieta.Diversity and adaptation in populations of clustering ants[C].In: Proceedings of the Third International Conference on Simulation of Adaptive Behavior: From Animals to nimats 3, MIT Press, Cambridge, MA, 1994: 501~508

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

12.E Lumer, B Faieta.Exploratory database analysis via self-organization[M].Unpublished manuscript, 1995

13.P Kuntz, D Snyers, P Layzell.A stochastic heuristic for visualizing graph clusters in a bi-dimensional space prior to partitioning[J].Journal of Heuristics, 1998; 5(3): 327~351

14.J Handl, B Meyer.Improved ant-based clustering and sorting in a document retrieval interface[C].In: Proceedings of the Seventh International Conference on Parallel Problem Solving from Nature (PPSN VII), Springer-Verlag, Berlin, Germany, 2002; 2439: 913~923

15.杨新斌,孙京诰,黄道.一种进化聚类学习新方法[J].计算机工程与应用, 2003; 39(15): 60~62

16.张惟皎,刘春煌,尹晓峰.蚁群算法在数据挖掘中的应用研究[J].计算机工程与应用, 2004; 40(28): 197~193

17.杨欣斌,孙京诰,黄道.基于蚁群聚类算法的离群挖掘方法[J].计算机工程与应用, 2003; 39(9): 12~14

(下转 211 页)

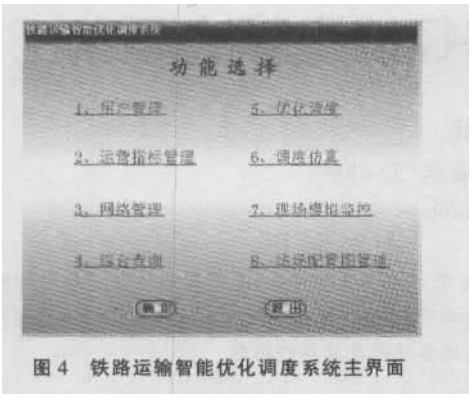

图 4　铁路运输智能优化调度系统主界面

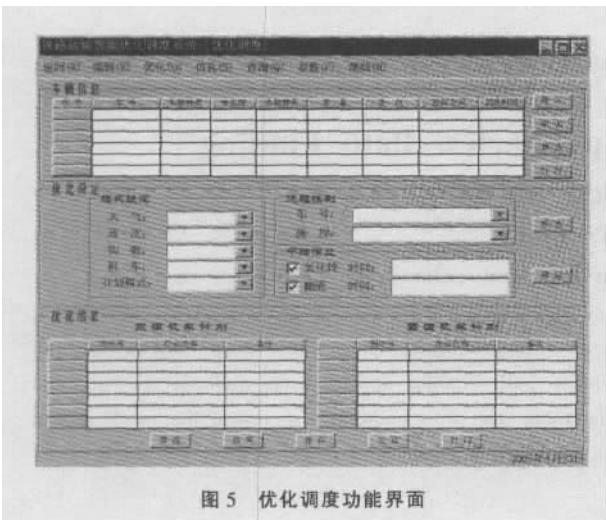
图 5　优化调度功能界面

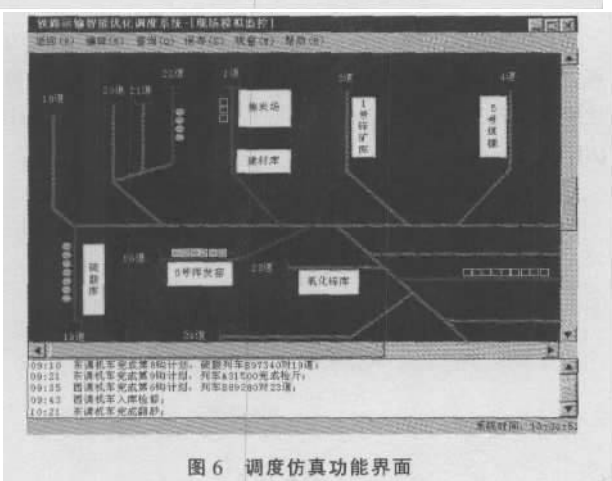
图 6　调度仿真功能界面

路运输管理信息系统导入一定信息和完成若干选项设置后，便可得到该厂东调、西调两台机车的调车钩计划，钩计划可人工修改，系统提供的“调度仿真”功能可对修改后的调车计划进行审查；若选择“现场模拟监控”功能，则调度员可根据现场反馈的信息，通过一步步触发调车计划的模拟来得到站场内停车情况和作业进度的模拟图像（如图 6 所示），增加了调车指挥的直观性。

开发的系统基本满足了企业铁路运输调度的需求，实现了调车计划自动编制和优化，验证了所提出的混合优化策略以及计划模拟方法的可行性，结果令人满意。不过，作为一个实用系统，系统中还存在一定的不足，如铁路站场环境复杂，研究过程中简化了一些工作条件，为此有必要进一步讨论一些复杂的运输调度情况，不断完善调度优化的方法，特别是列车解编组和取送车作业的优化；另外，在系统操作的友好性、调度专家规则的自学习等方面也需作进一步的研究和改进。

（收稿日期：2005 年 12 月）

参考文献

1.田茂勋.冶炼企业铁路运输组织[M].冶炼工业出版社，1987：4~5

2.何世伟等.枢纽编组站智能调度系统的设计与实现[J].北方交通大学学报，2002；26(5)：19~23

3.刘敏.大型钢铁企业编组站调度信息管理系统研究[J].中国铁道科学，2002；23(5)：41~45

4.刘敏.钢铁企业编组站信息与决策系统若干问题探讨[J].计算机工程，2002；28(7)：220~222

5.高四维，毛节铭.编号递推转换法一种编制编组调车作业计划的新算法模型[C].见：西南交通大学峨眉分校教学改革与科学研究论文集，成都西南交通大学出版社，2001：133~138

6.高四维，张殿业.提高调车作业指挥模型系统适应性的研究[J].交通运输系统工程与信息，2003；3(1)：84~88

7.高四维，张殿业.一种新的调车作业原理－“消逆法”[J].铁道学报，2003；25(5)：1~7

8.胡安洲.统筹对口调车法及其原理[C].见：铁道运输与经济论文集，北京：中国铁道出版社，1980：59~83

9.郭富娥编译.列车运行调整支援专家系统[J].铁路运输与经济，1996；(2)：30~32

10.周百川，黄国君.露天矿铁路运输调度专家系统的研究[J].中国矿业，1994；3(1)：77~80

11.郎茂祥，胡思继.车辆路径问题的禁忌搜索算法研究[J].管理工程学报，2004；18(1)：81~84

12.徐俊明.图论及其应用[M].中国科学技术大学出版社，1998：22~28

（上接 174 页）

18.Thomas Stutzle, Marco Dorigo.ACO Algorithms for the Traveling Salesman Problem[M].Evolutionary Algorithms in Engineering and Computer Science, Wiley, 1999：163~183

19.叶志伟，郑肇葆.蚁群算法中参数 α β ρ设置的研究－以 TSP 问题为例[J].武汉大学学报（信息科学版），2004；29(7)：597~601

20.H Azzag, N Monmarche, M Slimance et al.AntTree：a new model for clustering with artificial ants[C].In：IEEE Congress on Evolutionary Computation, Canberra, Australia, 2003：8~12

21.H Azzage, C Guinot, G Venturini.How to use ants for hierarchical clustering[C].In：Fourth international workshop on Ant Colony Optimization and Swarm Intelligence, Brussels, Belgium, LNCS 3172, 2004：350~357

22.H Azzag, N Monmarche, M Slimance et al.A clustering algorithm based on the ants self-assembly behaviour[C].In：Advances in Artificial Life-Proceedings of the 7th European Conference on Artificial Life (ECAL), Dortmund, Germany, 2801：564~571

23.J L Deneubourg, S Goss, N Franks et al.The Dynamics of Collective Sorting：Robot-Like Ants and Ant-Like Robots[C].In：From Animals to Animates：Proceedings of the First International Conference on Simulation of Adaptive Behavior, MIT Press, 1991：356~363

24.M Parag Kanade, O Lawrence Hall.Fuzzy Ants as a Clustering Concept.Dept of Computer Science Engineering[C].In：22nd international conference of the North American fuzzy information processing society, NAFIPS, 227~232

25.J Handl, J Knowles, M Dorigo.Ant-based clustering：a comparative study of its relative performance with respect to k-means, average link and 1d-Som[C].In：Proceedings of the Third International Conference on Hybrid Intelligent Systems, IOS Press, 2003