第 33 卷 第 5 期 计 算 机 工 程 2007 年 3 月
Vol.33 No.5 Computer Engineering March 2007
·博士论文· 文章编号：1000—3428(2007)05—0013—03 文献标识码：A 中图分类号：TP301.6

基于两阶段模糊蚂蚁聚类陆战旅待机地域的选取

傅调平，刘玉树
(北京理工大学计算机科学技术学院，北京 100081)

摘 要：根据“隐蔽疏散配置”选取原则，提出了一种计算机辅助生成陆战旅待机地域选取方案的新方法。为克服传统蚂蚁聚类算法运行时间长、仅能处理结构化数据等不足，给出了一种两阶段模糊蚂蚁聚类算法。对第 1 阶段聚类后数据进行融合操作，减少了第 2 阶段聚类的数据量、数据分布空间和迭代次数。实验证明，该算法是一种高效率、鲁棒性好的算法。该选取方法实现了陆战旅待机地域选取方案的自动、准确、快速计算机辅助生成。

关键词：两阶段聚类；模糊蚂蚁聚类算法；待机地域选取

Marine Brigade Deployment Region Choice Based on Two Phase Fuzzy-ant-clustering Algorithm

FU Tiaoping, LIU Yushu
(School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081)

【Abstract】 A new method is developed for auxiliary making marine brigade deployment plan by computer, which is based on the “concealed and scattered deployment” principle. Aiming at improving the shortage of traditional ant clustering algorithm, such as long runtime and merely handling the constructive data object, a new two phase fuzzy-ant-clustering algorithm is proposed. The first phase merges the data after clustering, so makes number of data, data distribution space and iteration time of the second phase reduced. The results of the experiment demonstrate that the algorithm is a high efficiency and good robustness algorithm. The new method realizes auxiliary choosing marine brigade deployment region by computer automatically, precisely and fast.

【Key words】 Two phase clustering; Fuzzy-ant-clustering algorithm; Deployment region choice

随着信息技术的迅猛发展，传统靠查阅资料和用沙盘、地形图来形成决策方案的指挥方式已无法适应现代化登陆作战的需要，待机地域选取方案的计算机辅助生成是现代战争的迫切需要。本文从“隐蔽疏散配置”这一登陆作战待机地域选取原则[1]和有效提高蚂蚁聚类效率出发，建立了一种陆战旅待机地域选取辅助决策的新方法，提出了一种两阶段模糊蚂蚁聚类算法(Two Phase Fuzzy-ant-clustering Algorithm, TPFAC)。实验证明，采用该算法与其它直接聚类蚂蚁算法相比，能显著提高聚类的效率和准确率，并更符合陆战旅待机地域选取问题的实际需要。

1 蚂蚁聚类和陆战旅待机地域选取问题分析

1.1 两阶段蚂蚁聚类的特点

蚂蚁聚类算法取得了很多研究成果和应用。Lumer等首先改进了Deneubourg算法，将数据对象之间相似度的度量引入到Deneubourg分类模型中，提出了LF聚类算法[2]。文献[3]提出了一种基于密度的启发性蚂蚁聚类算法，有效地提高了蚂蚁聚类算法的效率。

然而，LF算法及其后的一些改进算法存在的一个突出的问题是计算效率问题。蚂蚁在搜索过程中，其运动范围覆盖整个 2D的数据分布空间，也就是说，其搜索空间面积为m^2，其中m为正方形二维空间的边长。在算法的整个求解过程中，整个数据空间的面积不发生任何改变，这就造成了大量的冗余计算。因此在提高算法效率的方案中，首先可考虑利用两阶段聚类的思想，通过分阶段融合操作将邻近的数据融合为小簇，减少数据量，从而相应地减小数据分布空间的面积，以达到迅速提高聚类效率的目的。

1.2 陆战旅待机地域选取问题

待机区域，是登陆部队实施登陆作战的出发基地。在选择待机地域时，必须考虑既具备隐蔽作战企图，又具备便于出海的地形条件。

基于陆战旅登陆作战的特点，所属部队待机地域通常成建制分别配置。营与营之间和营内部的连之间根据其特点都有相应的配置要求。根据陆战旅“隐蔽疏散配置”这一待机地域选取基本原则，着眼于增强陆战旅自身安全性，本文陆战旅待机地域选取的思路为：

(1)根据陆战营选取的原则，在陆战营的相应地幅大小内根据连与连之间配置的要求得到满足陆战营待机地域要求的聚类小簇；

(2)将满足陆战营小簇的聚类中心模式作为输入数据对象，根据营与营之间配置的要求再次聚类分析，得出满足陆战旅待机地域要求的聚类簇，完成陆战旅待机地域的选取。

基于上述观点，将隐蔽疏散配置登陆部队归结为一种两阶段蚂蚁聚类。采用这种两阶段的陆战旅待机地域选取蚂蚁聚类算法和其它直接聚类蚂蚁算法相比，更加符合陆战旅待

基金项目：国家部委预研基金资助项目(10504033)
作者简介：傅调平(1975－)，男，博士生，主研方向：数据挖掘，智能辅助决策；刘玉树，教授、博导
收稿日期：2006-04-17 **E-mail：**ftpjh1121@sina.com

机地域选取的实际需要。

1.3 选取陆战旅待机地域的方法

本文提出的陆战旅待机地域选取方法在 GIS 上分两大步进行：第 1 步是基于空间数据挖掘的登陆部队集结上船地域地形分析，利用空间数据挖掘技术对指定区域作地形分析，遴选出符合待机地域地形条件的栅格，可视为数据预处理阶段；第 2 步是阵地栅格聚类分析，按照陆战旅待机地域选取的原则对阵地栅格作聚类分析，得到陆战旅待机阵地，把最终结果在地理信息系统上重现，该阶段是数据分析阶段。整体方法详细描述如下：

Step1 人机交互在作战地域内划定一个区域作为集结上船地域，该地域界定了陆战旅待机地域的选择范围。

Step2 将配置地域进行栅格化，栅格大小满足连阵地幅员；分析所有栅格的地形情况(包括基本地形条件、离海岸距离、可通行性、居民地、隐蔽度、平坦度等)，从中选取满足地形条件的n个栅格构成空间数据集$S=\{o_1,o_2,...,o_n\}$；

Step3 采用蚂蚁聚类算法对数据集 S 进行第 1 阶段聚类，去除阵地栅格少于指定阈值的小簇，选出符合营待机地域要求的 a 个小簇；

Step4 对每个小簇内阵地栅格的平均值进行第 2 阶段聚类，得到符合陆战旅待机地域要求的 b 个大簇；

Step5 分别围绕每个大簇选取符合陆战旅配置要求的 j 个营阵地小簇，再分别围绕每个小簇选取符合陆战营配置要求的 i 个连阵地栅格。

Step3 和 Step4 是难点也是本文要详述的重点。

基于模糊聚类适合描述半结构化数据的特点和蚂蚁聚类不必预设聚类中心数目、可视化、鲁棒性好等优点，本文提出一种改进的模糊蚂蚁聚类算法 TPFAC，该算法能够较好地解决陆战旅待机地域选取问题。

2 两阶段模糊蚂蚁聚类算法的设计

针对蚂蚁聚类算法存在的问题和陆战旅待机地域选取问题的特点，对蚂蚁聚类算法进行了改进。该算法建立在LF模型[2]及其后的一些改进算法[3,4]基础上。

2.1 两阶段蚂蚁聚类算法分析

两阶段蚂蚁聚类算法与蚂蚁直接聚类算法的最大不同是对第 1 阶段聚类后数据的融合操作。原始数据经过若干次迭代后，其中由更相似数据形成的较小规模的簇已基本出现。在这种条件下，通过有效的融合操作能够将相邻的近似数据点聚合为一个新的数据点，从而减少数据量。融合时，以一个数据点为中心对其周围相邻的若干网格上存在的数据进行融合条件的判断，根据融合判据确定新数据点的生成。其中，融合判据的基本条件设置为

$$d_{ij} \le d_{limit}, d_{ij} = |O_i - O_j|, d_{limit} = \sqrt{S}/2 \tag{1}$$

i 为融合中心点，j 为相邻数据点，S 为营待机地域所需面积。

根据以上判据，可形成由多个数据点聚合成的新数据点。新数据点的属性按照均值方式确定。属性大小为

$$x_{new} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad y_{new} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{2}$$

其中x_{new}，y_{new}为相邻区域内满足相似性度量标准的近似数据点融合成的新数据点的属性值，x_i，y_i为纳入同一新数据点内部的各数据点的属性值。对于更高维属性的数据，其新数据点的属性确定可参照此方法进行。

按照融合操作的规则生成新的数据集之后，即可重新设定数据分布空间。经过融合之后，数据量已大幅压缩，与之对应的 2D分布空间同样需要进行适当的规模压缩，以减少冗余网格的数量，提高算法效率。有些研究[5]中指出，二维分布空间边长和数据量的关系遵循如下规则较好，即

$$m = 2*\sqrt{n} \tag{3}$$

m 是二维分布空间边长，n 是数据数量。

由于空间的压缩和数据量的减少，蚂蚁在空间的搜索效率得以提高，因此所需的算法迭代次数也可相应地减少。其调整方法可根据问题性质灵活设置，并通过实验方法逐步优化。

2.2 基于模糊隶属度距离的相似度函数

经典聚类算法通常用于处理结构化数据对象，对象之间相似度基于属性空间的几何距离，距离越近则相似度越高。然而，阵地栅格数据并不是一般意义上的空间数据，虽然每个阵地栅格都可以用一对两维坐标来表示，但栅格间距离受“最近配置间距”和“最远配置间距”的约束，数据对象之间的相似性无法用确定的模型来表述，因此属于半结构化数据。本文引入了模糊数学中的隶属度概念，采用正态型隶属度函数来取代欧式距离作为相似性的度量，构造出基于模糊隶属度的距离模型：

$$u_{ij} = \begin{cases} 0 & d_{ij} < d_{\min} \\ e^{-k(d_{ij}-\alpha)^2} & d_{\min} \le d_{ij} \le d_{\max} \\ 0 & d_{ij} > d_{\max} \end{cases} \tag{4}$$

其中d_{ij}是阵地栅格数据集S中数据对象O_i和O_j的属性空间欧式距离，d_{min}和d_{max}分别为最近配置间距和最远配置间距，α为最优间距。

d_{min}、d_{max}和α 在两个聚类阶段根据陆战旅待机阵地和营待机阵地的特点分别采用不同的数值。

基于模糊隶属度距离的相似度函数为

$$f(o_i) = \frac{1}{\pi r^2}\sum_{o_j \in Neigh(r)} u_{ij} \tag{5}$$

$Neigh(r)$ 表示对象O_i的半径为r的邻域。

2.3 设置蚂蚁搜索禁忌表保证正确归类

为使蚂蚁能够记住数据对象，在同一路径的搜索中不重复搜索同一个样本点，给每只蚂蚁设置一个禁忌表 $tab_k(N)$。规定：如果 $tab_k(j)$ 的值为 true，则数据 j 是可以选择的搜索样本点；当蚂蚁 k 选择了数据 j，就将 $tab_k(j)$ 置为 false，此后蚂蚁就不能选择数据 j。

设置禁忌表时存放的是数据对象的索引号，因此，增加设置禁忌表并不会显著增大系统的空间开销。而给每只蚂蚁设置一个禁忌表既保证了每个蚂蚁都可以对所有数据对象遍历一遍，避免未指派数据对象的存在；又因为多个蚂蚁的存在，使得数据对象都有被考察多次的可能，保证了正确归类，减小了分类错误率。

2.4 调整 C 值控制后期收敛

概率转换函数是以群体相似度为变量的函数，将数据对象的平均相似性转化为拾起概率或放下概率。本文选取对称式 *Sigmoid* 函数作为概率转换函数，对 LF 聚类算法进行改进。*Sigmoid* 是 $f(o_i)$的一个函数，*Sigmoid* 函数具有非线性，运算中只需调整一个参数，比 LF 算法中的二次函数有更快的收敛性。

一个未负载的随机运动的蚂蚁拾起一个对象的拾起概率

定义为

$$P_p = 1 - Sigmoid(f(o_i)) \quad (6)$$

同样，一个负载的随机运动的蚂蚁放下一个对象的放下概率定义为

$$P_d = Sigmoid(f(o_i)) \quad (7)$$

$Sigmoid(x) = (1-e^{-Cx})/(1+e^{-Cx})$ 为自然指数形式，参数 C 越大，曲线饱和越快，算法收敛速度也越快。

在聚类的过程中，有一些稀疏区域的对象与其它对象相似度较低，蚂蚁拾起它们后，难以尽快放下它们，以至影响算法的收敛速度。为了进一步提高蚂蚁聚类算法的效率，本文采取在算法后期逐渐增大 C 值、尽快放下稀疏区域对象的策略。

2.5 TPFAC 算法描述

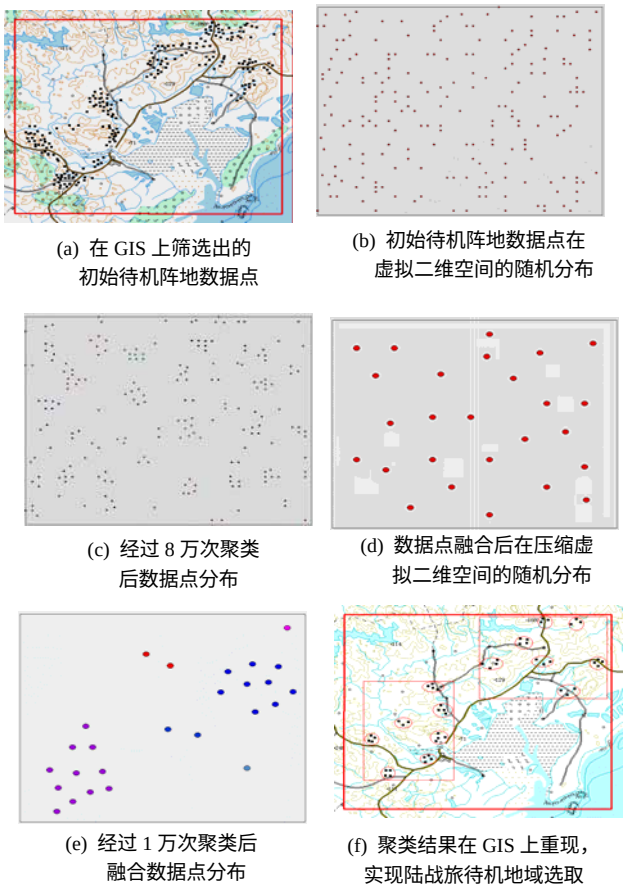
设定数据集大小为 n_{item}，蚂蚁总数为 n_{ant}，算法的循环次数为 $n_{iterate}$，算法蚂蚁聚类部分的时间复杂度大致推算为 $O(n_{item}n_{ant}n_{iterate})$。

根据上述设计思想，两阶段模糊蚂蚁聚类算法的伪代码描述如下。

```
/*Initialization*/
For every Oi do
  将Oi随机放置到二维网格上
End For
For every ant do
  放置在随机位置
End For
/*Main Loop*/
For t =1 to tmax
    For every ant do
      If (ant 空载) and (当前位置有Oi && tab_antk(Oi)=true) then
        计算 f(Oi) 和 pp(Oi)，并在[0,1]间产生随机概率 pr ；
        If ( pp(Oi) ≥ pr )) then
          Pick up Oi
          tab_antk(Oi)=false
        End If
      Else If (ant 携带Oi) and (当前位置空) then
        计算 f(Oi)和 pd(Oi)，并在[0,1]间产生随机概率 pr ；
        If ( pd(Oi) ≥ pr ) then
          Drop Oi
        End If
      End If
      随机选择未被占据的网格为下一目标；
    End For
    If 迭代数 ≥ 收敛阈值
      C = C + k1C ；
    End If
    If 迭代数 = 融合阈值
      进行数据融合操作；
      随机生成融合后数据的二维分布空间；
    End If
    If 满足算法终止条件
      结束算法
    End If
End For
```

3 TPFAC 算法测试与比较分析

为了验证方法的可行性，编制相应程序在P4 2.8GHz/512MB微机上进行实验。阵地栅格大小为 0.04km^2(满足连最小阵地幅员)，在指挥员标定区域内，经阵地地形分析后筛选出 219 个初始陆战连阵地栅格，如图 1(a)所示。图中方框区域代表指挥员标定的登陆待机地域，方框内的黑色栅格点表示数据预处理遴选出的连阵地栅格。聚类算法参数：蚂蚁数 N_{ant} =5，蚂蚁观察邻域半径r=9，分类半径dist=2，融合前数据最优间距 α =400 米，融合后数据最优间距 α =2 500 米。图 1(b)为连阵地栅格数据在虚拟二维平面的初始随机均匀散布，图 1(d)是融合后数据在压缩后虚拟二维平面的初始随机均匀散布，图 1(c)、图 1(e)分别是数据对象在虚拟二维平面上经过 80 000 次和 10 000 次聚类后的结果，耗时分别为 27s 和 2s。图 1(f)是聚类结果经步骤 5 处理后的最终阵地栅格在地理信息系统上的重现，选出了 2 个旅阵地(如图中方框区域)，每个旅阵地选定了 7 个营阵地(如图中椭圆区域)，每个营阵地区域内标定了若干个连阵地。



(a) 在 GIS 上筛选出的初始待机阵地数据点　(b) 初始待机阵地数据点在虚拟二维空间的随机分布

(c) 经过 8 万次聚类后数据点分布　(d) 数据点融合后在压缩虚拟二维空间的随机分布

(e) 经过 1 万次聚类后融合数据点分布　(f) 聚类结果在 GIS 上重现，实现陆战旅待机地域选取

图 1 基于两阶段模糊蚂蚁聚类选取待机地域的过程

为了测试 TPFAC 算法的性能，本文还选取了不同大小的地形数据集与 LF 算法进行实验比较。不同大小数据集下两种算法的运行时间和分类错误率如图 2 所示。

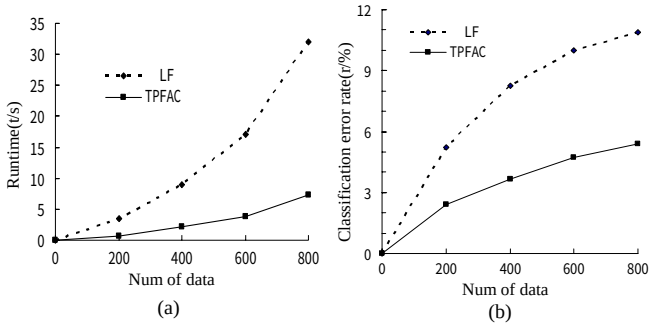


图 2 两种算法的运行时间和分类错误率比较

（下转第 19 页）