一种面向非规则非致密空间分布数据的聚类方法

唐贤伦[1] 仇国庆[1] 庄 陵[2]
(重庆邮电大学网络化控制与智能仪器仪表教育部重点实验室 重庆 400065)[1]
(重庆邮电大学通信与信息工程学院 重庆 400065)[2]

摘 要 针对目前很少关注非致密非规则数据聚类的情况,利用蚁群算法具有的组合优化方面的优势,引入近邻函数准则,提出了基于蚁群算法和近邻函数准则的聚类算法,来求解非规则非致密数据聚类问题。实验表明,对于非规则非致密分布数据的聚类问题,该聚类算法可根据连接关系合理地进行聚类,相比K均值算法等其他采用样本距离作为分类指标的聚类方法,可有效降低错聚率,一定程度上较好地解决了这类问题。

关键词 蚁群算法,近邻函数准则,聚类分析,K均值聚类

Clustering Method for Irregular and Uncompact Data

TANG Xian-lun[1] QIU Guo-qing[1] ZHUANG Ling[2]
(Key Laboratory of Network Control & Intelligent Instrument, Ministry of Education, Chongqing University of Posts and Telecommunications, Chongqing 400065, China)[1]
(College of Communications and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China)[2]

Abstract Taking advantage of ant colony algorithm's superiority on combinatorial optimization problems, a new clustering algorithm based on ant colony algorithm and neighbor function criterion was presented for uncompact and irregular data. Result shows that this new clustering algorithm can obtain better clustering results than K-means clustering algorithm and can commendably solve the problem about the uncompact and irregular distribution to a certain extent.

Keywords Ant colony algorithm, Neighbor function criterion, Clustering analysis, K-means clustering

1 引言

聚类分析是多元统计分析的一种,作为一种无监督分类方法,聚类分析已经被广泛地应用于模式识别、数据挖掘、计算机视觉和模糊控制等诸多领域。传统的聚类算法主要分为4类[1]:划分方法、层次方法、基于密度方法和基于网格方法。后来科学家提出了许多新的聚类方法,比如基于遗传算法的聚类方法、基于蚁群算法的聚类算法等[2-4],这些算法已经应用到电路设计、文本挖掘等领域。

然而传统的聚类算法以及后来提出的众多新的聚类方法几乎都采用样本距离作为分类指标,对于球形或椭球形等规则且致密目标簇的聚类效果比较好,而对于非规则非致密形状的聚类相对较差[5,6]。因为具有非规则非致密特性的数据如果单纯地根据样本距离进行分类,往往容易将核函数不确定和不能用简单函数表示的样本错误地划分,造成数据的误判,且不容易获得全局最优的聚类中心。本文结合蚁群算法和近邻函数准则提出一种新的聚类方法,用于求解此类情况下的数据聚类问题。实验结果表明,对于非规则非致密分布数据,该方法具有良好的聚类效果。

2 蚁群算法

蚁群算法是模拟自然界中蚁群从蚁巢到目的地之间寻找最短路径时的交互过程的一种随机搜索算法[7]。Marco Dorigo等人提出该算法时,充分利用了蚁群搜索食物的过程与著名的旅行商(TSP)问题之间的相似性,通过人工模拟蚂蚁搜索食物的过程来求解TSP。

设m是蚁群中蚂蚁的数量,$b_i(t)$表示t时刻位于城市i的蚂蚁数,$\tau_{ij}(t)$表示t时刻在ij连线上的信息量。在算法的初始时刻,将m只蚂蚁随机地放到n座城市上,此时各路径上的信息量相等,设$\tau_{ij}(0)=C(C$为常数)。每只蚂蚁根据路径上保留的信息量独立地选择下一个城市。在时刻t,蚂蚁k从城市i转移到城市j的概率为

$$P_{ij}^{k}=\begin{cases}\dfrac{\tau_{ij}^{\alpha}(t)\eta_{ij}^{\beta}(t)}{\sum_{l\in allowed_k}\tau_{il}^{\alpha}(t)\eta_{il}^{\beta}(t)}, & \text{if } j\in allowed_k\\ 0 & \text{otherwise}\end{cases} \tag{1}$$

其中,$allowed_k=\{0,1,2,\cdots,n-1\}-tabu_k$表示蚂蚁$k$下一步允许选择的所有城市,列表$tabu_k$记录了当前蚂蚁$k$所走过的城市。当所有$n$个城市都加入到$tabu_k$中时,蚂蚁$k$便完成了一次循环,此时蚂蚁$k$所走过的路径便是问题的一个解。$\eta_{ij}$是一个启发式因子,表示蚂蚁从城市$i$转移到城市$j$的期望程度,在蚂蚁算法中,$\eta_{ij}$通常取城市$ij$之间距离的倒数。$\alpha$和$\beta$分别表示路径上信息量和启发因子的重要程度。

当蚂蚁完成了一次循环之后,相应边上的信息素浓度根据式(2)进行更新处理:

到稿日期:2008-06-21 本文受国家自然科学基金(60506055),重庆市教委项目(KJ070509)资助。
唐贤伦 博士,副教授,主要研究方向为群体智能、智能信息处理,E-mail:tangxlun@hotmail.com;仇国庆 副教授,主要研究方向为智能信息处理;庄 陵 博士研究生,讲师,主要研究方向为人工智能与信息处理。

$$\tau_{ij}(t+1)=\rho\tau_{ij}(t)+\Delta\tau_{ij}\quad \rho\in(0,1)$$

$$\Delta\tau_{ij}=\sum_{k=1}^{m}\Delta\tau_{ij}^{k} \tag{2}$$

其中 ρ 为一取值范围在 0 到 1 之间的常数系数，表示残留信息的保留部分，$1-\rho$ 表示信息素的挥发程度。$\Delta\tau_{ij}$ 表示本次循环路径上信息的增量。$\Delta\tau_{ij}^{k}$ 表示第 k 只蚂蚁在本次循环中留在路径 ij 上的信息量，如果蚂蚁 k 没有经过路径 ij，则 $\Delta\tau_{ij}^{k}$ 的值为零，$\Delta\tau_{ij}^{k}$ 表示为：

$$\Delta\tau_{ij}^{k}=\begin{cases}\dfrac{Q}{L_k}, & \text{若第 } k \text{ 只蚂蚁经过路径 } ij\\ 0 & \text{否则}\end{cases} \tag{3}$$

其中，Q 为常数，L_k 表示第 k 只蚂蚁在本次循环中所走过的路径的长度。

3 近邻函数准则算法

在某些情况下，采用距离度量方式求近似性的方法不一定能达到合理效果。图 1 所表示的就是几种不同形状的不规则数据构造的例子。为解决类似情况下的聚类问题，可采用近邻函数准则算法，根据连接关系来合理地聚类成真实的几何形状。

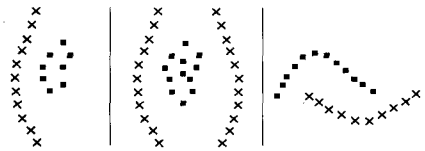

图 1 不规则空间数据分布

本文采用文献[8]提出的思想，基于近邻函数给出评价聚类结果质量的准则函数，其中给出了几个近邻函数准则用到的定义。

定义 1(近邻系数) 对于数据集中的任何两个样本 y_i 和 y_j，若 y_j 是 y_i 的第 I 个近邻，则称 y_j 对 y_i 的近邻系数为 I；类似地，如果 y_i 是 y_j 的第 K 个近邻，那么 y_i 对 y_j 的近邻系数为 K。

定义 2(近邻函数值) 对于两个聚类样本 y_i 和 y_j，如果 y_i 对 y_j 的近邻系数为 K，y_j 对 y_i 的近邻系数为 I，那么这两个样本的近邻函数值(用 a_{ij} 表示)定义为 $a_{ij}=I+K-2$。如果 y_i 和 y_j 互为最近邻，这时的近邻函数值最小，等于零。

定义 3(连接损失) 对于数据集中的任何两个样本 y_i，y_j，若 y_j 是 y_i 的第 I 个近邻(按距离排序)，则称 y_i 对 y_j 的正向连接损失为 I。同样，若 y_i 是 y_j 的第 K 个近邻，则称 y_i 对 y_j 的逆向连接损失为 K。由此就可与定义 2 中表述的近邻函数值建立关系，即定义 y_i 和 y_j 之间的连接损失(也称作近邻函数值)为 $a_{ij}=(K+I-2)$。如果 y_i 和 y_j 互为最近邻，这时的连接损失最小，$a_{ij}=0$。这样规定“连接”损失的好处是使密度较为接近的点容易聚成一类。

当规定了样本间的连接损失之后，就可以规定类内损失和类间损失。

定义 4(类内损失) 该类中所有样本之间的连接损失之和，即

$$L_{IA}=\sum_{i=1}^{N}\sum_{j=1}^{N}a_{ij} \tag{4}$$

为要计算类间损失，首先计算 c_i 和 c_j 两类之间的最小近邻函数值 γ_{ij}。即计算 c_i 类中的每一点与 c_j 类中所有各点的近邻函数值，并取其中的最小值为 γ_{ij}。显然 c_i 类与其余的 $c-1$ 类的最小近邻函数值为

$$\gamma_i=\min_{\substack{j\in\theta\\ j\neq i}}\gamma_{ij},\ \theta=\{1,2,\cdots,c\} \tag{5}$$

定义 5(类间损失) 给定一个聚类结果，对于其中的任意一个分类 c_i，假设它对应的最小近邻函数值为 γ_i。用 $\alpha_{i\max}$ 和 $\alpha_{k\max}$ 分别表示两个分类 c_i 和 c_k 中具有最大连接损失的两个样本的近邻函数值。β_i 表示类 c_i 的类间损失：

$$\beta_i=\begin{cases}-[(\gamma_i-\alpha_{i\max})+(\gamma_i-\alpha_{k\max})], & \text{若}\begin{cases}\gamma_i>\alpha_{i\max}\\ \gamma_i>\alpha_{k\max}\end{cases}\\ \gamma_i+\alpha_{i\max}, & \text{若}\begin{cases}\gamma_i\leqslant\alpha_{i\max}\\ \gamma_i>\alpha_{k\max}\end{cases}\\ \gamma_i+\alpha_{k\max}, & \text{若}\begin{cases}\gamma_i>\alpha_{i\max}\\ \gamma_i\leqslant\alpha_{k\max}\end{cases}\\ \gamma_i+\alpha_{i\max}+\alpha_{k\max}, & \text{若}\begin{cases}\gamma_i\leqslant\alpha_{i\max}\\ \gamma_i\leqslant\alpha_{k\max}\end{cases}\end{cases} \tag{6}$$

总类间损失

$$L_{IR}=\sum_{i=1}^{c}\beta_i \tag{7}$$

从上面的讨论中可以看到，定义

$$J_{NN}=L_{IA}+L_{IR} \tag{8}$$

为本算法的准则函数显然是合理的，聚类的结果应使准则函数 J_{NN} 取最小值。

4 非规则非致密数据聚类算法

基于样本的聚类与蚂蚁觅食及旅行商问题的相似性，将每一个样本数据看成一只蚂蚁所需访问的地点。每一只蚂蚁根据样本数据的近邻函数及样本之间的信息素的大小，以一定的概率选择下一个访问的地点，并将其加入自身的数据列表里。直到每一只蚂蚁对全部的数据样本都进行了一次唯一的访问，从而形成了全部数据样本的一个有序的连接。在每个数据列表里，每个样本数据都只出现一次。形成全部样本的一个有序连接后，对每只蚂蚁的数据列表根据其样本之间的近邻函数值的大小，断开其中值最大的两个样本数据之间的连接，从而形成了最初的两个类别。对于已经进行了聚类的样本数据，计算每一类之间的连接损失。依此类推，直到样本数据之间的连接损失最小为止，最终形成该蚂蚁的聚类结果。对每一只蚂蚁的聚类结果进行比较，选取当前运行的聚类连接损失最小的蚂蚁的结果为本次循环的结果。当算法达到规定的最大运行次数或者样本数据的类别之间的连接损失达到给定最小值时，算法结束，此时的运行结果为最终解。

基于蚁群算法和近邻函数准则求解非规则非致密数据聚类问题的算法具体流程如下。

Step 1 计算样本数据的距离矩阵 Δ，使其元素 $\Delta_{ij}=\Delta(y_i,y_j)$ 表示样本 y_i 和 y_j 间的欧氏距离。

$$\Delta_{ij}=\|D(Y_i-Y_j)\|_2=\sqrt{\sum_{k=1}^{m}(y_{ik}-y_{jk})^2}$$

Step 2 根据矩阵 Δ，得到样本的近邻矩阵 M，其元素 M_{ij} 为样本 y_i 对 y_j 的近邻函数值。一般 M 为正定矩阵。由于样本点的近邻序数只能是 $1,2,\cdots,N$，因此 M 矩阵中各元素均为整数。

Step 3 根据 $L_{ij}=M_{ij}+M_{ji}-2$ 得到近邻函数矩阵 L，如果 y_i 和 y_j 间有“连接”关系，则 L_{ij} 给出了它们之间的近邻函数值。置 L 矩阵的对角元素 L_{ii} 的值为 $2N$ 或更大。

Step 4 把样本数据看作蚂蚁所需访问的地点，每只蚂蚁根据样本数据的近邻函数矩阵 L，由下式决定下一个将加入其数据列表的不在禁忌列表里的样本数据：

$$P_{ij}^{k}=\begin{cases}\dfrac{\tau_{ij}^{\alpha}(t)\eta_{ij}^{\beta}(t)}{\sum_{l\in allowed_k}\tau_{il}^{\alpha}(t)\eta_{il}^{\beta}(t)}, & 若\ j\in allowed_k\\ 0 & 若\ j\notin allowed_k\end{cases}$$

$$\eta_{ij}=\frac{1}{(L_{ij}+1)}$$

其中 L_{ij} 为样本 i,j 之间的近邻函数值。

Step 5　将加入到蚂蚁数据列表的样本数据加入其禁忌列表 $tabu_k(k=1,2,\cdots,m)$。

Step 6　判断是否所有的蚂蚁都对全部样本数据完成了一次访问。否则返回 Step 4。

Step 7　由每一只蚂蚁的数据列表中的样本数据的连接情况，根据样本之间的近邻函数值的大小，断开当前有连接情况的近邻函数值最大的样本之间的连接，从而形成不同的类别。

Step 8　由式(4)一式(8)计算所有类别的类内损失和类间损失及总的代价。

Step 9　若每一只蚂蚁的 $J_{NN}(k)$ 达到最小，则完成一次聚类，否则转到 Step 7 重复执行。

Step 10　取所有蚂蚁中 $J_{NN}(k)$ 最小的蚂蚁的聚类结果作为本次循环的求解结果。

Step 11　根据式(3)，调整样本数据连接之间的信息素大小：

$$\Delta\tau_{ij}^{k}=\begin{cases}\dfrac{Q}{L_k}, & 若第\ k\ 只蚂蚁在本次循环使样本\ ij\ 有连接\\ 0 & 否则\end{cases}$$

其中 L_k 为总连接损失。

Step 12　判断循环次数是否达到最大步数的给定值，$J_{NN}=L_{IA}+L_{IR}$ 是否达到给定的最小值：是，则中止运行，给出最终解；否则，返回 Step 4。

本算法的优点如下：1)减少了样本相互之间反复计算近邻函数值的过程，提高了运算效率；2)使样本点能根据原自身固有的在样本空间中的分布得到优化，而不会像在 K 均值或模糊均值基础上求解近邻函数值而获得的球形或椭球形分布，在空间中改变了原有形状。

5　实验与结果分析

为了验证新算法的有效性，本文选用模式分类中常用的 Iris 数据集进行测试，采用分类正确率度量聚类的质量。该数据集以花瓣长度、花瓣宽度、萼片长度和萼片宽度 4 个参数作为特征，共 150 个样本，数据为 150×4 维矩阵，其中 1～50 记录属于第一类 Iris-setosa，51～100 记录属于第二类 Iris-versicolor，101～150 记录属于第三类 Iris-virginica。

图 2 显示了 Iris 数据集中各个样本的前 3 个属性值的三维空间分布。可以看出部分样本数据的致密和规则特性不是很好，所以可采用本文所提算法进行验证。

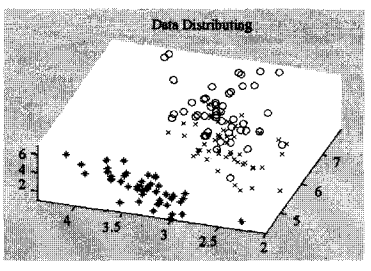

图 2　Iris 前 3 个属性数据分布

用本文提出的基于蚁群算法和近邻函数准则的聚类算法、近邻函数准则算法以及 K 均值算法分别做了 3 次实验，以算法的错聚率作为实验结果，如表 1 所列。

表 1　3 种算法的实验结果

算法错聚率	K 均值算法	近邻函数准则	本文算法
1	8.7%	4.7%	2.7%
2	8.0%	5.3%	3.3%
3	10%	5.3%	3.3%

从上述实验结果可看出，K 均值算法由于根据样本距离进行分类，从数据致密程度考虑问题，把簇居在一起呈规则形状分布的数据做整体对待，且 K 均值算法是建立在局部最优基础上，不容易获得全局最优的聚类中心，因此对这种样本数据的致密和规则特性不是很好的情况聚类效果不很理想。近邻函数准则根据连接关系进行聚类，因此较之 K 均值算法效果更好。蚁群算法具有组合优化方面的优势，利用其求解问题的全局优化特性结合近邻函数准则来求解非规则非致密聚类问题，结果表明相比前两种方法可以得到更优的聚类效果。

从时间效能上看，考虑到蚁群算法群体计算的时间损耗，整体效率较低些。在聚类效果上，由于考虑了样本数据间的空间分布(指空间近邻特性)，因而对于远离聚类中心的“噪声数据”的正确归类有很大的帮助。在遇到非致密空间数据问题时，尤其要考虑到数据在空间构造的非规则特性的时候，新算法明显会有更好和更为准确的聚类效果。当然，遇到诸如数据样本致密分布特性的问题时，新算法所表现出来的性能有所下降。

所以，在运用不同算法解决实际问题的时候，需依据数据分布的空间几何特性来确定：对于规则、致密空间分布的数据样本可采用 K 均值算法和模糊均值算法，对于非致密、非规则几何空间分布的数据样本可采用本文所提算法。

结束语　本文提出一种新的基于蚁群算法和近邻函数准则算法的聚类方法来求解非致密非规则数据聚类问题。实验表明，对于非规则非致密空间数据而言，与 K 均值聚类算法和近邻函数准则相比，本文提出的算法聚类效果更好，具有一定的理论和实际意义。

参 考 文 献

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

[2] 陈传波，罗增琦. 一种基于蚁群聚类的快速算法[J]. 计算机工程，2007，33(6)：206-207，210

[3] 王宝文，等. 基于遗传算法的高维数据模糊聚类[J]. 计算机工程与应用，2007，43(16)：191-192，221

[4] Azzage H, Guinot C, Venturini G. How to use ants for hierarchical clustering[C]// Fourth International Workshop on Ant Colony Optimization and Swarm Intelligence. Brussels, Belgium, LNCS 3172. 2004：350-357

[5] 高新波. 模糊聚类分析及其应用[M]. 西安：西安电子科技大学出版社，2004

[6] Kantardzic M. 数据挖掘—概念、模型、方法和算法[M]. 北京：清华大学出版社，2003

[7] Dorigo M, Maniezzo V, Colorni A. The ant system: Optimization by a colony of cooperating agents[J]. IEEE Transactions on Systems, Man, and Cybernetics Part B, 1996, 26(1)：29-41

[8] 边肇祺，张学工，等. 模式识别[M]. 北京：清华大学出版社，2001