

约束蚂蚁聚类算法

徐晓华, 潘舟金, 陆 林, 陈 峻

(扬州大学 信息工程学院, 江苏 扬州 225009)

摘 要: 通过对真实世界蚁群的模拟仿真, 提出一种基于随机游走的约束蚂蚁聚类算法来处理以 must-link 和 cannot-link 形式出现的约束聚类问题. 在人工数据集和 UCI 标准数据集上的实验结果表明我们的算法优于无监督的蚁群聚类算法和 COP-Kmeans 算法.

关键词: 约束聚类; 蚂蚁聚类; 随机游走

中图分类号: TP181

文献标识码: A

文章编号: 1000-7180(2012)04-0169-04

Constrained Ant Clustering Algorithm

XU Xiao-hua¹, PAN Zhou-jin¹, LU Lin¹, CHEN Ling¹

(College of Information Engineering, Yangzhou University, Yangzhou 225009, China)

Abstract: By simulating the clustering behaviors of the real-world ant colonies, we propose in this paper a constrained ant clustering algorithm based on random walk to deal with the constrained clustering problems with pairwise must-link and cannot-link constraints. Experimental results show that our approach is more effective on both synthetic datasets and UCI datasets compared with the unsupervised random walk ant-based clustering algorithm and the COP-Kmeans algorithm.

Key words: constrained clustering; ant clustering; random walk

1 引言

聚类分析是数据挖掘领域中一种十分重要的数据分析方法. 在许多情况下我们可以获得数据集的少量先验信息或者是领域知识. 如何利用这些少量但又重要的信息来改善聚类算法的效果变的非常重要^[1-2]. 半监督聚类的提出就是针对这类问题的, 即如何利用已有的先验信息来提高无监督聚类算法的性能.

群居的社会性昆虫具有高度的群体智能^[3], 人们受蚂蚁等社会性昆虫生物繁衍后代、寻找食物、构造巢穴、清除垃圾、保卫领地等行为的启发, 设计了一系列算法, 并成功应用于函数优化、组合优化、机器人技术领域进行问题求解.

将群体智能运用到聚类分析领域的发展过程

中, 最早的成功算法模型来自 Deneubourg^[4]提出的聚类模型, 通常称之为基本模型(Basic Model, BM). 这一模型用来解释蚂蚁尸体堆积成墓穴的行为. Lumer 和 Faieta 对 BM 模型进行了改造^[4], 一般化为用实值元素来聚类数据向量, 给出了数据对象的相似性度量表达式, 设计了通常称之为 LF 的聚类模型. 最近, Xu 等人提出了一种结合细胞自动机的蚂蚁聚类模型 ASM^[5]改善了蚂蚁聚类算法的效率; He 和 Hui 用蚂蚁聚类算法(Ant-C)^[6]分析基因表达数据; El-Feghi 等提出引入了信息素聚合的性质和对环境感知的能力来提高算法的收敛效果^[7]. Abul Hasan 等综述了基于群体智能的聚类算法的巨大研究进展^[8]; Lutz Herrmann 和 Alfred Ultsch 基于 ESM^[9-10]提出了人工生命系统(Artificial Life System)^[11]来处理聚类问题.

收稿日期: 2011-10-09; **修回日期:** 2011-12-25

基金项目: 国家自然科学基金项目(61003180, 61070047); 江苏省科技厅自然科学基金项目(BK2010318); 江苏省教育厅自然科学基金项目(09KJB20013B)

在我们先前的工作中,我们提出了新的蚂蚁聚类框架 RWAC^[6]. RWAC 与传统的蚂蚁聚类算法的不同之处在于,每一个数据点本身就是一只可以游走的蚂蚁,这个想法更加简单直接.在本文中我们基于 RWAC 提出了约束蚂蚁聚类框架,利用了先验的约束对信息对适应度函数进行了修正使得算法简单且更有效.

2 约束聚类问题描述

聚类问题可概述为用户希望得到数据集 X 的 k 个划分,记为 $\Pi_X (\Pi_X = \pi_1 \cup \pi_2 \cdots \cup \pi_k, \cap \pi_i = \Phi)$,每个划分间相似度尽量大,类内相似度尽量小.在现实的聚类问题中,可以获得关于数据点少量的领域先验信息,对于少量数据点是否该被划分为一类或者不能划分为同一类的先验指导信息更加容易获得,最大化这些约束信息的指导价值能够很大程度上改善无监督算法的性能.一般意义上,我们考虑两种类型的约束关系.

(1) Must-link 约束关系 $c = (i, j)$,被这类约束限制的两个数据点在聚类后必须被分到同一个簇中.

(2) Cannot-link 约束关系 $c \neq (i, j)$,与 must-link 约束关系相反,受这类约束限制的两个数据点在聚类后不能被放在同一个簇中.

约束聚类算法的任务是在先验约束信息的指导下对数据集进行划分,在不违背约束的情况下取得的簇内数据点尽量相似,簇间相似度尽量相异.

3 RWAC 聚类框架

RWAC 模拟了蚂蚁在其生存环境中寻找安全的位置睡眠的行为.对于像蚂蚁一类比较弱小的个体为了安全起见便会产生同类相聚、异类相斥的行为.蚂蚁寻找一个安装的位置来休眠的行为很简单:在游走过程中,蚂蚁对周围环境进行判断,当对环境的适应度比较高即比较安全时就停止游走转而睡眠,当对环境的适应度比较低时就苏醒过来离开这个地方,在没找到比较安全的位置前不停随机游走.

在 RWAC 算法中,适应度函数如下与 BM 中的相似:

$$f(i) = \max\left(0, \frac{1}{S} \sum_{j \in N(i)} \left[1 - \frac{d(i, j)}{\alpha}\right]\right) \quad (1)$$

式中, $d(i, j)$ 是 i 蚂蚁和 j 蚂蚁某种距离尺度,一般是欧氏距离.

蚂蚁从睡眠状态到苏醒状态的概率为

$$p_{\text{active}}(i) = \frac{\beta^i}{\beta^\lambda + f(i)^\lambda} \quad (2)$$

式中, β 为适应度阈值, λ 为激活压力,一般取 $\lambda=2$.

基于前文的说明和算法执行过程的描述,我们给出 RWAC 算法描述如表 1 所示.

表 1 RWAC 算法

Algorithm RWAC
Data: ants with affinity matrix
Result: clustering information of all ants
Begin:
Parameters initialization
For each ant do:
Randomly place the ant in different grid of G and initialize its class label
Endfor
While not reach the termination do:
For each ant do:
Calculate ant's fitness $f(\text{ant})$
Calculate the probability of awoken $p_a(\text{ant})$
Generated random number $r = \text{random}([0, 1])$
If $r \leq p_a(\text{ant})$ then:
Active the ant and let it walk into a free location
Else:
Stay in sleep in the original grid
Endif:
Update the label vector according to the local random walk theory δ
Endif
Endfor
Update the parameters, $t = t + 1$
Endwhile
End

4 约束蚂蚁聚类框架

本节我们提出了基于 RWAC 的约束聚类算法:对适应度函数进行了改造,引入高斯核函数作为相似度的度量,从而借鉴的谱方法而达到忽略分离时的度量参数 α ;对在通过对相似度的增强和减弱来实现对应的 must-link 和 cannot-link 两类约束对聚类的指导作用.我们给出了最大最小修正.

4.1 改进的适应度函数

无论是 BM 模型还是 RWAC 模型,适应度函数中的分离度量参数决定了两个蚂蚁在什么情况下应该分离,在程序执行过程中,自适应的减小它的数值大小,某种意义上讲,该参数影响着聚类的数目.我们通过引入高斯核函数作为相似度的度量,对适应度函数进行了改造,使算法变简单直接的同时提高了聚类精度.

我们先以高斯核函数作为相似度得度量方式:

$$w_{ij} = e^{-\frac{d_{ij}^2}{2\sigma_i^2}} \quad (3)$$

式中, d_{ij} 为数据点属性空间的距离, σ_i 的值是与当前数据点的第 i 个近邻的属性空间距离.于是我们得到新的适应度函数:

$$f(i) = \frac{1}{S} \sum_{j \in N(i)} (w_{ij} - \beta e^{-\frac{1}{2}}) \quad (4)$$

在我们新的适应度函数中, S 是局部领域的大小, 参数 β 将从 0 逐步增加到 1:

$$\beta = \beta^{\Delta t} + k(f^t(i) - f^{\Delta t}(i)) \quad (5)$$

式中, t 是当前的迭代次数, Δt 是步长, 用来确定 β 的增加速率. k 是某个正常数.

4.2 最大最小修正

受约束的蚂蚁在网格上随机游走的过程中, 如果得知当前所处的位置的领域中存在与之成对的另外一只蚂蚁时, 用最大相似度或是最小相似度来代替原本两蚂蚁的相似度来实现 must-link 和 cannot-link 约束的指导作用. 公式表示为:

$$\tilde{\omega}_{ij} = \begin{cases} e^{\frac{d_{\min}^2}{2\sigma^2}}, & \text{if must-linked}(i, j) \\ e^{\frac{d_{\max}^2}{2\sigma^2}}, & \text{if cannot-linked}(i, j) \\ e^{\frac{d_{ij}^2}{2\sigma^2}}, & \text{otherwise} \end{cases} \quad (6)$$

式中, $d_{\min}$ 是所有数据点中与 i 数据点距离最近的点, 相对的 $d_{\max}$ 是所有数据点中与 i 数据点距离最远的点.

5 实验分析

本节我们将对以上改进的约束聚类 CAC 算法, RWAC 算法与 COP-Kmeans[1] 算法进行聚类性能的比较. 在 UCI 上挑选了四个数据集: Iris, TAE, Hayes-roth, Glass 以及人工数据集上行测试和比较.

5.1 实验设计

我们使用 C 语言实现了 CAC 算法, RWAC 算法以及 COP-Kmeans 算法. 在上述的 UCI 数据集上进行实验. 针对每个数据集我们随机生成了不同数目的约束对. 在不同的数据集和不同约束对个数的情况下, 采用两个第三方聚类评价标准来衡量三个算法的优劣.

5.2 结果评价方式

本文采用了比较通用的两个第三方评价标准来评价聚类的质量. 纯度(purity)是比较简单明了的评价方法. 在计算纯度时, 形式化为

$$\text{purity}(\Omega, C) = \frac{1}{N} \sum_k \max_j |\omega_k \cap c_j| \quad (7)$$

式中, $\Omega = \{\omega_1, \omega_2, \dots, \omega_k\}$ 是聚类结束后得到了簇集, $C = \{c_1, c_2, \dots, c_k\}$ 是数据集本来的分类. 如果聚类数目很多的话, 会使纯度值偏高. 在使用纯度作为评价标准时就不考虑聚类的个数.

F measure(F score)为这两类错误添加了不同的权重.

$$F = \frac{2PR}{P+R} \quad (8)$$

在这里, $P = TP/(TP + FP)$, $R = TP/(TP + FN)$.

相对而言 F measure 作为第三方评价标准是最合理的.

5.3 数据集上的测试和比较

表 2 总结了用于比较 COP-Kmeans, CAC, RWAC 三种算法的该 4 个测试数据集. 最后根据实验结果进行讨论和总结.

表 2 测试数据集

Data sets	# cases	# attributes	# classes
Iris	150	4	3
TAE	151	5	3
Hayes-Roth	132	5	3
Glass	214	9	6

我们为每个数据集随机生成 1%, 2%, 4%, 7%, 11%, 16%, 20% 的受约束的数据点. 从 purity, F measure 两个方面来与 RWAC, COP-Kmeans 进行比较. 图 1~图 5 为三种算法在五个数据集上的性能比较图.

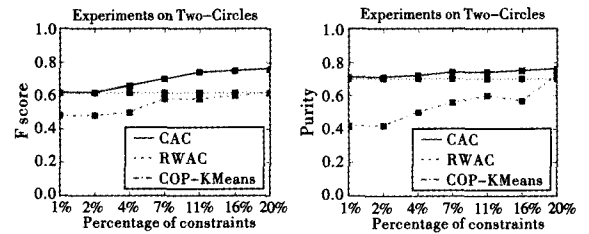


图 1 F-score and Purity on Two-circles with different percentage of constraints

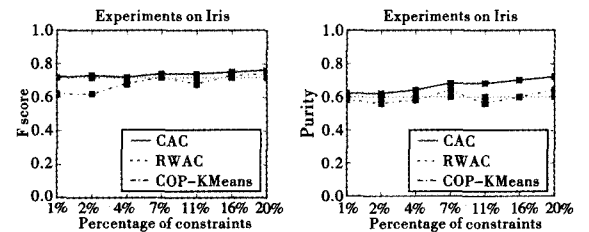


图 2 F-score and Purity on Iris with different percentage of constraints

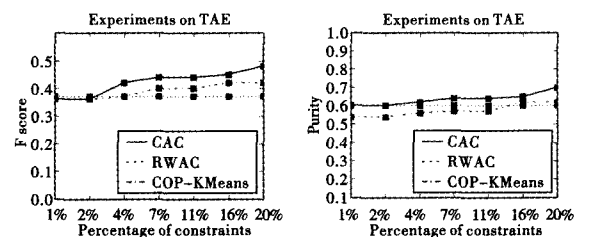


图 3 F-score and Purity on TAE with different percentage of constraints

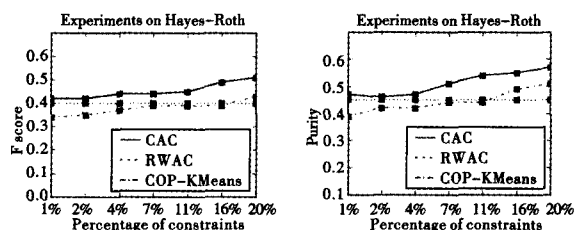


图4 F-score和Purity on hayes-roth with different percentage of constraints

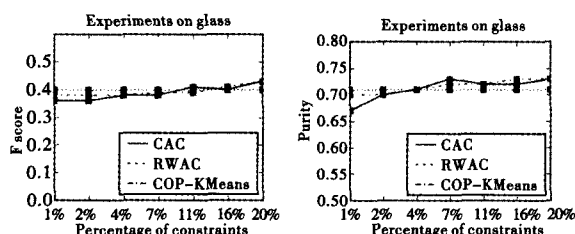


图5 F-score和Purity on glass with different percentage of constraints

6 结束语

本文我们基于 RWAC 算法,引入的自调节的高斯核函数来改进适应度函数,提出了约束的蚂蚁聚类算法框架来处理以约束对形式出现的先验指导信息.在人工生成的数据集和在 UCI 标准的数据机上的实验结果表明我们的 CAC 算法要优于 RWAC 和 COP-Kmeans 算法.

参考文献:

- [1] Zhu X. Semi-supervised learning with graphs [D]. USA: Carnegie Mellon University. 2005.
- [2] Wagstaff K, Cardie C, Rogers S, et al. Constrained k-means clustering with background knowledge[C]// Proceedings of the Eighteenth International Conference on Machine Learning, San Francisco, CA, USA: ACM, 2001:577-584.
- [3] Eric Bonabeau, Marco Dorigo, Guy Theraulaz. Swarm intelligence: from natural to artificial systems (santa fe institute studies in the sciences of complexity proceedings)[M]. USA: Oxford University Press, 1999.
- [4] Dorigo M, Bonabeau E, raulaz G. Ant algorithms and stigmergy[J]. Future Generation Computer Systems, 2000, 16(8):851-871.
- [5] Xiaohua Xu, Lin Chen, Ping He. A novel ant clustering algorithm based on cellular automata[J]. Web Intelligence and Agent Systems, 2007, 5(1):1-14.
- [6] He Y, Hui SC. Exploring ant-based algorithms for gene expression data analysis[J]. Artifl Intell Med. 2009, 47(2):105-119.
- [7] El-Feghi I, Errateeb M, Ahmadi M. Sid-Ahmed MA (2009) an adaptive ant-based clustering algorithm with improved environment perception[C]// Proceedings of the 2009 IEEE international conference on systems, man, and cybernetics. San Antonio:IEEE,2009.
- [8] Abul Hasan, Mohamed, Ramakrishnan, et al. A survey: hybrid evolutionary algorithms for cluster analysis[J]. Artificial Intelligence Review, 2011:1~26.
- [9] Ultsch A, Herrmann L. Automatic Clustering with U * C[R]. Technical Report, Dept. of Mathematics and Computer Science, University of Marburg. 2006.
- [10] Ultsch A. Emergence in Self-Organizing Feature Maps [C] // Proc Workshop on Self-Organizing Maps (WSOM 2007). Bielefeld, Germany:[s. n.],2007.

作者简介:



徐晓华 男,(1979—),博士.研究方向为数据挖掘与机器学习.