doi:10.3969/j.issn.1000-436x.2012.z2.042

基于信息熵的蚁群聚类 DBSCAN 改进算法

张拥华，杜飞明，吴代文
（湖南工业职业技术学院 经济管理系，湖南 长沙 410208）

摘 要：针对 DBSCAN 算法对数据分布不均匀和大规模数据处理问题上的不足，提出了一种新的整合算法，算法使用信息熵和蚁群聚类技术对聚类数据集进行代表性子集选择，在子集基础上进行 DBSCAN 聚类，实验证明这一算法能显著降低 I/O 耗费和内存需求，有效地解决含有分类属性的高维大规模数据集的聚类问题。
关键词：信息熵；聚类；DBSCAN；蚁群算法
中图分类号：TP18 **文献标识码**：A **文章编号**：1000-436X(2012)Z2-0290-04

Improved density-dased clustering algorithm based on information entropy and ant colony optimization abstract

ZHANG Yong-hua, DU Fei-ming, WU Dai-wen
(Department of Economic Management, Hunan Industry Polytechnic, Changsha 410208, China)

Abstract: An integration of clustering algorithm was proposed for the shortage of the DBSCAN algorithm in inhomogeneous distribution and large-scale data processing. The algorithm extracted representative data from the original data set using information entropy and ant colony clustering technology, and did DBSCAN clustering based on the representative data subset. The experiment show that this algorithm is effective to reduce the I/O-consuming and memory requirements, and resolve the cluster problem of large-scale data sets containing categories property.
Key words: information entropy; clustering; DBSCAN; ant colony optimization

1 引言

DBSCAN（density-based spatial clustering of applications with noise）是一个基于密度的聚类算法。该算法将具有足够高密度的区域划分为簇，并可以在带有“噪声”的空间数据库中发现任意形状的聚类。DBSCAN 算法作为一种有代表性的基于密度的聚类算法，从提出至今，已经成功应用于城市规划、城市建设、选址等多个研究领域，并在发展过程中产生了几种有效的改进算法[1~5]。为了解决含有分类属性的数据集聚类问题，人们探索了许多途径[6,7]，文献[8]所提出的将信息熵理论引入聚类算法是一种比较好的方法。但 DBSCAN 算法对数据集不进行任何预处理，在进行聚类操作前，直接构建 R^*树，并绘出 k-dist 图，因此时间耗费巨大，并且在聚类过程中 DBSCAN 算法的全局变量不断优化导致算法需要更大的内存和 I/O 资源进行支撑。当数据分布不均匀和大规模数据处理时，DBSCAN 算法就出现严重不足。文献[9]提出了 OR-DBSCAN 算法，该算法基于一种代表性子集选择技术—最优 K 相异性算法(optisim)来扩展 DBSCAN 算法，使之能够有效地对大规模数据集进行聚类分析，但是它不适用于包含分类属性的数据集。本文在信息熵理论和蚁群算法的基础上，设计一种新的融合算法—

收稿日期：2012-10-23
基金项目：湖南省教育厅科研基金资助项目（11C0451）
Foundation Item: The Scientific Research Foundation of Education Department of Hunan Province (11C0451)

基于信息熵的蚁群聚类，应用到有代表性子集选择中，从而能有效地解决含有分类属性的高维大规模数据集聚类问题。

2　基于信息熵的聚类算法

1948 年，Shannon 引入了信源的信息熵：

$$H(R) = -\sum_{x \in S(X)} p(x)\log(p(x)) \tag{1}$$

其中，X 是一个随机变量，$S(X)$ 是 X 可能取的值的集合，$p(x)$ 是 X 的概率函数。

基于信息熵的聚类问题，就是要将集合 $R=\{X_1, X_2, \mathrm{L}, X_n\}$ 中的 n 个记录分配到 k 个聚类中，并且使得整个分配的总信息熵最小。由以上定义可知，其目标是使期望信息熵最小，那么则可以考虑使用启发式函数来进行。

$$\mathrm{E}'\hat{C} = \sum_{i=1}^{k} \frac{|\mathrm{P}(C_i)|}{n} \mathrm{E}(C_i) \tag{2}$$

其中，$|\mathrm{P}(C_i)|$ 是分配给聚类 C_i 的数据个数，$\mathrm{E}(C_i)$ 表示聚类 C_i 的信息熵，并且 $C_i \mathrm{I}\ C_j = \varnothing,\ i, j = 1, 2, \mathrm{L}, k, i \neq j$，即不同聚类中的数据将不会重复，最后目标是使得聚类结果 $\hat{C}$ 的期望信息熵 $\mathrm{E}'\hat{C}$ 最小。

算法描述如下。

输入：包含 n 个对象的数据库，半径，最少数目 MinPts。

输出：所有生成的簇，到达密度要求。

1) 在待聚类数据集中随机选取 k 个数据，分别分配在 k 个不同的聚类 C_i；

Repeat

2) 从第 k+1 个数据起，遍历分配给每个聚类 C_i 并计算其信息熵；

3) 从 K 个不同聚类 C_i 中，找出信息熵最小的聚类 $C_{\min}$；

4) 将第 k+1 个数据正式分配给该聚类 $C_{\min}$。

Until 从第 k+1 到第 n 个对象均被处理。

算法存在的问题：1) 当数据分布不均匀时，随机选取的 k 个数据可能十分相似，由此初始聚类相似度高，因而造成聚类的失败。2) 对数据进行聚类时，随着聚类集合中的数据逐渐增加，此前分配到该聚类集合中的数据可能不再与大多数数据相似，这样对最后获得的聚类质量产生很大影响。

3　基于信息熵的蚁群聚类 DBSCAN 改进算法

3.1　蚁群算法初始化

蚁群算法（ACO, ant colony optimization）是一种新型的模拟进化算法，在求解复杂组合优化问题方面具有并行性、正反馈、顽健性强等先天优越性，但在蚁群算法初始时，却存在蚂蚁路径上信息素缺乏的缺点，使得算法在开始时运行效果不好。本文采用文献[8]中所描述的初始化方法，为蚂蚁路径预先留下信息素：设待聚类的数据集合中有 n 个数据，随机选取 m 个数据（$m << n$）作为样本空间 M。

For i=1 to m do

计算任意 2 个数据之间的信息熵 E($\{R_i, R_j\}$)；

For i=1 to m do

求出每个数据与其他数据之间的最小信息熵 $g_i = \min(\mathrm{E}(\{R_i, R_j\}))$；

For i=1 to m do

对最小信息熵 g_i 按照降序产生队列 G；

由队列将对应数据分配给不同的聚类；

Until k 个聚类中都有分配数据，从而得到初始化的集合

$$\mathscr{C} = (C_1, \mathrm{L}, C_k)$$

3.2　蚁群算法在聚类问题中的使用

为了将蚁群算法引入聚类问题，现将蚁群算法修改为

$$p_{ij}^{m} = \begin{cases} \dfrac{[\tau_{1j}(t) + \mathrm{L} + \tau_{ij}(t)]^{\alpha} \left[\dfrac{1}{d_{1j} + \mathrm{L} + d_{ij}}\right]^{\beta}}{\displaystyle\sum_{s \notin \mathrm{tad}u_m} [\tau_{1s}(t) + \mathrm{L} + \tau_{is}(t)]^{\alpha} \left[\frac{1}{d_{1s} + \mathrm{L} + d_{is}}\right]^{\beta}}, & j \notin \mathrm{tad}u_m \\ 0, \text{其他} \end{cases} \tag{3}$$

其中，d_{ij} 为数据 x_i 与数据 x_j 之间的距离，即通信代价；$\tau_{ij}(t)$ 为数据 x_i 与数据 x_j 之间的熟悉度；$\mathrm{tad}u_m$ ($m = 1, 2, \mathrm{L}, k$)表示第 m 只蚂蚁的禁忌表，即它不允许访问数据的集合；参数 α, β 表示轨迹的相对重要性和能见度的相对重要性；p_{ij}^{m} 则表示第 m 只蚂蚁选择数据 x_j 加入联盟的概率。

当蚂蚁完成一次循环后，各数据之间的 τ_{ij} 将按照下式进行调整：

$$\tau_{ij}(t+1) \leftarrow \rho\tau_{ij}(t) + \Delta\tau_{ij} \tag{4}$$

$$\Delta\tau_{ij} = \frac{Q}{d_{ij}} \tag{5}$$

其中，ρ 表示轨迹的相对持久性，$1-\rho$ 表示轨迹衰减度随时间推移，以前留下的信息逐渐消失，$\Delta\tau_{ij}$ 表示一次循环后数据 x_i 与数据 x_j 之间信息素的增量。

3.3　基于信息熵的蚁群聚类 DBSCAN 改进算法描述

将基于信息熵的蚁群聚类的 DBSCAN 算法，简记为 HACO-DBSCAN。

HACO-DBSCAN 算法的步骤如下。

1) 对数据集 DS 使用基于信息熵的蚁群聚类技术进行代表性子集选择，产生代表性数据子集 HACO-DS。

2) 为代表性数据子集 HACO-DS 构造 R*树。

3) 使用 DBSCAN 对代表性数据子集 HACO-DS 进行聚类。

4) 对代表性数据子集 HACO-DS 中的每一个核心对象 P 执行：

① 在 HACO-DS 中根据半径 Eps 对 P 进行区域查询，得到其邻域 $Eps-P$；

② 对 $Eps-P$ 中的点及其同一簇中其他未被选为代表对象的点都按 P 的类别进行标记。

在这里，步骤 1)使用了在第 2 节中提到的基于信息熵的蚁群聚类算法，从数据集中均衡地提取数据样本。步骤 2)为代表性数据子集 HACO-DS 建立 R*树，以提高聚类的效率。步骤 3)和步骤 4)分别进行聚类和标记。

4　实证分析

在 DBSCAN 算法基础上使用 MATLAB 编写程序，实现了 HACO-DBSCAN 改进算法。实验中计算机（PC 机）环境为：P4CPU(1.5GMHz)、1G 内存和 80G 硬盘。数据集为 SEQUOIA 2 000 基准数据库中的点数据集。

图 1 展现了 DBSCAN 和 HACO-DBSCAN 算法的性能比较结果。实证中使用的数据集是 D_1、D_2、D_3、D_4 和 D，其中，D 是 SEQUOIA 2 000 基准数据库中的点数据集，而 D_1、D_2、D_3、D_4 分别是它的 10%、20%、30%、50%选择性代表子集。实证参数设置为：DBSCAN 算法($e=20$，$MinPts=4$)，HACO-DBSCAN 算法($m=100$，$k=10$，$\alpha=4$，$\beta=1, \rho=0.9$，循环 1 000 次)。图 1 中的曲线表明，HACO-DBSCAN 算法明显快于 DBSCAN 算法，且 HACO-DBSCAN 算法比 DBSCAN 算法对数据集的大小具有更好的扩展性。

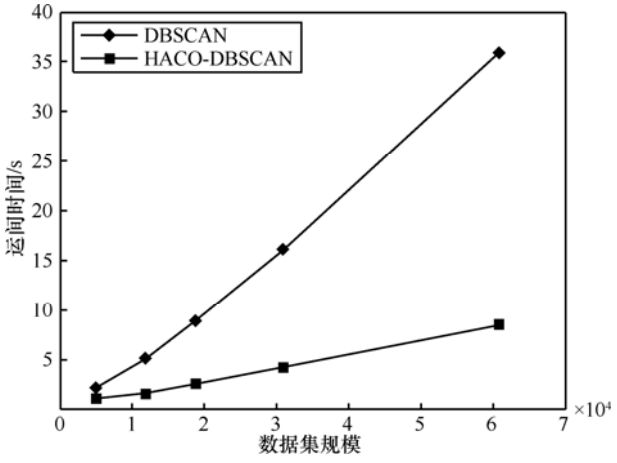

图 1　DBSCAN 与 HACO-DBSCAN 性能对比

与 DBSCAN 算法相比，HACO-DBSCAN 算法中的区域查询操作频率大幅度削减。若在 DBSCAN 算法中区域查询操作的数目是 n，则在 HACO-DBSCAN 算法中这个数值至多是 ki (k 是基于信息熵的蚁群聚类的簇数，i 是从每一簇中选择出来作为代表对象的数目)。HACO-DBSCAN 算法的聚类质量依赖于每一簇中代表对象的选择比率和选择方法，由于基于信息熵的蚁群聚类代表性子集选择技术较为均衡，当选择比率适当，则 HACO-DBSCAN 算法的聚类质量可以得到保证。

5　结束语

HACO-DBSCAN 算法通过在待聚类的数据集中使用基于信息熵的蚁群聚类技术进行代表性子集选择，产生代表性数据子集 HACO-DS，并为该子集构造 R*树，再使用 DBSCAN 算法对该子集进行聚类，且并行地聚类整个数据集。这样，I/O 和内存的资源需求得到显著的降低，且聚类运行时间得到大量的减少；另外，由于信息熵的引入，该算法也能够处理含有分类属性的大规模数据集。

参考文献：

[1] 周水庚, 周傲英, 曹晶等. 一种基于密度的快速聚类算法[J]. 计算机研究与发展, 2000, 37(11):1187-1192.
ZHOU S G, ZHOU A Y, CAO J, *et al*. A fast density based clustering algorithm[J].Journal of Computer Research and Development, 2000,

37(11):1187-1192.

[2] 冯少荣，肖文俊．一种提高 DBSCAN 聚类算法质量的新方法[J]. 西安电子科技大学学报, 2008, 35(3):523-529.
FENG S R, XIAO W J. New method to improve DBSCAN clustering algorithm quality[J]. Journal of Xidian University, 2008, 35(3): 523-529.

[3] 马帅，王腾蛟，唐世渭等．一种基于参考点和密度的快速聚类算法[J]. 软件学报,2003, 14(6):1089-1095.
MA S, WANG T J, TANG S W, *et al*. A fast clustering algorithm based on reference and density[J]. Journal of Software, 2003, 14(6): 1089-1095.

[4] 崔尚卿，马秀莉，唐世渭等．基于不均匀密度的自动聚类算法[J]. 计算机工程, 2008, 34(23):86-88.
CUI S Q, MA X L, TANG S W, *et al*. Auto-clustering algorithm based on non-uniform density[J]. Computer Engineering, 2008, 34(23):86-88.

[5] LIU D Q, SOURINA O. Free-parameters clustering of spatial data with non-uniform density[A]. Proc of the 2004 IEEE Conf on Cybernetics and Intelligent Systems[C]. Singapore, 2004. 387-392.

[6] GUHA S, RASTOGI R, SHIM K. ROCK: a robust clustering algorithm for categorical attributes[J]. Information Systems, 2000, 25(5): 345-366.

[7] BARBARA D, COUTO J, LI Y. COOLCAT: An entropy-based algorithm for categorical clustering[D]. George Mason University, Information and Software Engineering Department, Fairfax: George Mason University, 2001.

[8] 熊家军，李庆华．信息熵理论与入侵检测聚类问题研究[J].小型微型计算机系统,2005, 26(7):1163-1166.
XIONG J J, LI Q H. Study on clustering problem for intrusion detection with information entropy[J]. Mini-micro Systems, 2005, 26(7):1163-1166.

[9] 胡文瑜，孙志挥，周晓云．基于相异性选择的密度聚类算法研究[J]. 小型微型计算机系统, 2006, 27(9):1601-1604.
HU W Y, SUN Z H, ZHOU X Y. Research of density-based clustering algorithm based on dissimilarity selection[J]. Journal of Chinese Computer Systems, 2006, 27(9):1601-1604.

[10] ESTER M, KRIEGEL H P, SANDER J, *et al*. A density-based algorithm for discovering clusters in large spatial databases with noise[A]. Proceedings of 2nd International Conference on Knowledge Discovering in Databases and Data Mining(KDD-96)[C]. Portland, Oregon, USA, 1996. 226-231.

作者简介：

张拥华（1977-），女，湖南宁乡人，硕士，湖南工业职业技术学院讲师，主要研究方向为机器学习、数据挖掘。

杜飞明（1963-），男，湖南益阳人，硕士，湖南工业职业技术学院副教授，主要研究方向为自动化控制。

吴代文（1980-），男，湖南邵阳人，硕士，湖南工业职业技术学院讲师，主要研究方向为数据挖掘。

ISSN 1000-436X

发行代号：国内2-676 / 国外M395　　2012 年 11 月 30 日出版　定价：58.00 元