基于网格和信息熵的多密度聚类算法①

周悦来，谭建豪

(湖南大学 电气与信息工程学院，长沙 410012)

摘 要：虽然现有的很多聚类算法能发现任意形状、任意大小的类，但用于多密度的数据集时却难以取得令人满意的结果。为提高对多密度数据集的聚类效果，提出了一种基于网格和信息熵的多密度聚类算法，它通过不同密度的网格所携带的信息熵，自动计算出密度阈值，找出在多密度数据集中不同的类。实验证明，该算法能有效的去处噪声，发现多密度的类，具有较好的聚类效果。

关键词：聚类；自动阈值；网格；信息熵；多密度

Grid-Based and Information Entropy-Based Clustering Algorithm for Multi-Density

ZHOU Yue-Lai, TAN Jian-Hao

(College of Electrical and Information Engineering, Hunan University, Changsha 410012, China)

Abstract: Although many existing clustering algorithm can find the arbitrary shape and different size clusters, but it is difficult to obtain satisfactory results for multi-density data set. In order to improve the quality and efficiency of clustering algorithm, the paper presents a new improving precision clustering algorithm based on grid and information entropy, which through information entropy which carried by the different densities of grid to automatically calculate the density threshold, and then identify different clusters in the multi-density data set. Experiments show that the algorithm can wipe off the noise effectively and find out the multi-density clusters that have better clustering results.

Key words: clustering; automatic threshold; grid; information entropy; multi-density

所谓聚类，就是根据数据中发现的描述对象及其关系的信息，将相近相似的一组对象划分成类，使类内的对象相似性尽量大，而类间的对象相似性尽量的小[1]。一个好的聚类算法应能识别任意大小和形状的聚类，自动检测并清除孤立点（孤立点是指没有包含在任何聚类内的对象），它对数据点的输入顺序不敏感，输入参数对领域知识的弱依赖性，要有良好的时间效率，当数据增加时具有良好的可伸展性，并且聚类结果具有可解释性和可用性[2]。聚类分析时所使用的数据对象中，各个类的密集度往往并不完全相同，而且有时还有很大的差别，虽然大多数的聚类算法都是致力于怎样发现任意大小和形状的类，但是在处理密度差别较大的数据集时，还是难以取得满意的结果。现在能够处理多密度数据集的聚类算法有Chameleon[3]算法、SNN[4]算法、GDCIC[5]算法、多阶段密度线[6]算法、基于网格梯度[7]算法等。Chameleon算法一种采用动态建模技术的层次算法，它可以用来处理多密度的数据集，但是它的时间复杂度较高，为$O(N^2)$，同时 Chameleon 算法将许多噪声点都纳入类中而没有丢弃，导致聚类的精度不高；SNN 算法采用了一种共享近邻的思想来定义相似度，它在多密度聚类和处理孤立点或噪声点方面精度不高，其时间复杂度也较高，为 $O(N^2)$，并且该算法对参数的敏感的；GDCIC 算法采用了局部网格密度置信区间的方法，能够自动识别稠密区间，但是不能很好的处理某些高孤立点或噪声的区域；多阶段密度线算法是在基于网格的等密度线聚类算法的基础上，采用多阶段的聚类方式对数据集进行聚类，它的缺点是在聚类过程中需要人工参与，对输入参数是敏感的；基于网格梯度算法是根据网格梯度在不同类之间变化的思想找出隐藏在

① 基金项目:湖南省自然科学基金(08JJ3132)
收稿时间:2011-03-03;收到修改稿时间:2011-03-26

多密度数据集中的类，它的缺点是聚类结果对输入的参数是敏感的，且输入参数对领域知识有强依赖性。本文提出一种基于网格和信息熵的多密度聚类算法（Grid-based and Information entropy-based Clustering Algorithm for Multi-density，GICM），它通过不同密度的网格所携带的信息熵，自动计算出密度阈值，然后分离出不同密度区域的核心网格，再使用广度优先的搜索方式和边界处理技术进行聚类，找出在多密度数据集中不同的类。

1　相关定义

定义 1　数据空间

给定一个 d 维数据集 $D=\{D_1,D_2,\cdots D_d\}$，其中属性 D_i 都是有界的，取值区间为 $[l_i,h_i),i=1,2,\cdots,d$，那么 $S=[l_1,h_1)\times[l_2,h_2)\times\cdots\times[l_d,h_d)$ 就是一个 d 维的数据空间。

定义 2　点集

输入的 d 维点集是由 $V=\{V_1,V_2,\cdots,V_m\}$ 组成，其中 $V_i=\{V_{i1},V_{i2},\cdots,V_{id}\}$。$V_i$ 的第 j 个分量 $V_{ij}\in D_j(1\le j\le d)$。

定义 3　网格单元

将 d 维数据空间的每一维划分为 K 个长度相等且不相交的左闭右开区间，从而将整个数据空间划分成 K^d 个不相交且大小相等的超矩形单元，可以用网格单元 $U_i=\{u_{i1},u_{i2},\cdots,u_{id}\}$ 来描述每一个这样的超矩形单元。

定义 4　网格密度

用落入网格单元 U_i 中的数据点的个数 k 来表示 U_i 的网格密度，记为 $den(U_i)$。当 $den(U_i)=0$，称 U_i 为空格单元；当 $den(U_i)>0$ 时，称 U_i 为非空格单元。

定义 5　平均网格密度

平均网格密度 $den(U)_{avg}$ 等于数据点的总数与非空格网格数之比，即 $den(U)_{avg}=\dfrac{Num}{non\_cell}$。其中 Num 为数据点的总数，non_cell 为非空网格的个数。

定义 6　网格信息熵

受信息论[8]、SCI[9]算法中信息熵的启发，GICM 算法提出网格信息熵的概念，但与信息论中稍有不同的是，算法中使用公式(1)来表示网格的信息熵：

$$H(x)=\sum_{i=1}^{x}-k\times P(den(U)=k)\times\log_2 P(den(U)=k) \quad (1)$$

其中 k 是非空网格的密度，x 是 k 能有效取值的个数，$P(den(U)=k)=\dfrac{count(k)}{\sum_{i=1}^{k}i\times count(i)}$。其中 $count(k)$ 网格密度等于 k 的网格数，$\sum_{i=1}^{k}i\times count(i)$ 等于在计算网格信息熵时实际使用的非空网格数。

从信息传播的角度来看，信息熵可以表示信息的价值，这样子我们就有一个衡量信息价值高低的标准。当一个系统越是有序时，信息熵就越低。在 GICM 算法中，是以网格信息熵作为密度阈值的大小，当空间数据分布越均匀，则网格信息熵越小，此时密度阈值也越小。

定义 7　核心网格

以网格信息熵为密度阈值，若网格单元 U_i 的网格密度 $den(U_i)$ 满足不等式(2)：

$$den(U_i)>H(x) \quad (2)$$

则称网格单元 U_i 为核心网格，否则为非核心网格，类的扩展从核心网格开始的。

定义 8　网格邻居

当且仅当网格单元 U_i 和 U_j 有相邻的边界或相邻的点，即 $U_i\cap U_j\ne\phi$ 时，称 U_i 和 U_j 是网格邻居论[10]。那么 d 维数据空间的每个网格单元都有 3^d-1 个网格邻居（处于数据空间边界的网格单元除外）。

定义 9　簇

从任意核心网格开始的所有密度可达网格的集合称为簇。簇是一个类的骨架。

2　一种基于网格和信息熵的多密度聚类算法(GICM算法)

2.1 聚类边界点的处理

簇构成聚类的骨架，而边界点则充实了该骨架[11]，有时一些边界点可能落在骨架之外的网格中，但是这些边界点却又和骨架有着诸多的联系，是对骨架有益的补充，如果只是简单的舍弃掉这些边界点，就会影响到整个聚类的精度，所以需要将聚类的边界点从这些网格中提取出来，以提高聚类的精度。

设 $UL=\{UL_1,UL_2,\cdots UL_k\}$ 为未被聚类的网格单元，$UH=\{UH_1,UH_2,\cdots UH_n\mid n\le 3^{d-1}\}$ 是 UL_i 的网格邻居，计算网格 UL_i 的重心 x_i 和 UH_i 的重心 y_i 之间

的欧式距离 $d(x_i, y_i)$，若 $d(x_i, y_i)$ 满足不等式(3)，则将网格 UL_i 和 UH_i 聚为一类。若都不满足不等式(3)，则将网格 UL_i 作为噪声点。

$$d(x_i, y_i) \le d(x, y) \tag{3}$$

其中 $d(x, y)$ 为网格单元对角线的长度，整个不等式即为两个网格重心的欧式距离小于或等于网格单元对角线的长度。

2.2 GICM 算法描述

算法的主要思想是基于以下的事实：类内的对象相似性大，而类间的对象相似性小。即一个聚类内部的网格密度高于聚类外部的网格密度和边界点的网格密度，所以网格密度从聚类边界到聚类外部，会有一个明显的变化，所以每次聚类都是从核心网格单元开始逐步向外扩展，当遇到边界点后，进行边界点处理，直到所有的网格单元都被处理后，结束聚类。为了便于可视化，算法选择对二维数据进行描述，从基于广义距离确定数据对象相似度上讲，算法分析对多维数据集同样适用。

2.3 GICM 算法步骤

输入：数据集 D

输出：聚类，噪声/孤立点

步骤 1　将数据集 D 的每一维划分为 K 个等长的单元，得到 K^d 个网格单元的数据空间。每一维上区间数目：$K = \sqrt{Num}$，其中 Num 是数据对象的总个数。

步骤 2 将数据点映射到网格，得到数据点在网格中的空间坐标，统计网格单元的密度，找出所有的非空网格单元和统计非空网格的个数，并计算非空网格重心。

步骤 3 统计不同网格密度的个数，及将这些不同的网格密度按升序排列，统计在每一不同网格密度取值下网格的数目。

步骤 4 计算网格信息熵 $H(x)$，根据不等式(2)判断网格是否是核心网格。

步骤 5 计算平均网格密度 $den(U)_{avg}$ 的值，若网格信息熵 $H(x)$ 满足不等式(4)：

$$H(x) > den(U)_{avg} \tag{4}$$

则去掉网格密度最大的网格数，转到步骤 4；若不满足不等式(4)，则转到步骤 5。

步骤 5 从最高密度的任意核心网格开始，按照广度优先的搜索方式将所有密度可达的网格归为一类，遇到边界时，进行边界点处理。直到所有的非空网格都被处理完，输出聚类。

3 实验结果及分析

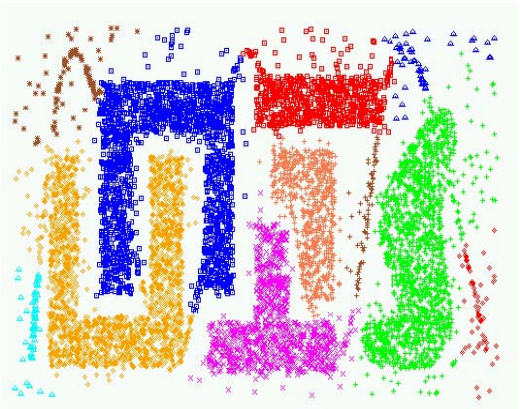
实验环境是英特尔 Celeron(赛扬)E1400 @ 2.00GHz CPU，2 GB 内存，Microsoft Windows XP 专业版。算法的编写和编译是在 VC++6.0 环境下实现的。

3.1 算法复杂度分析

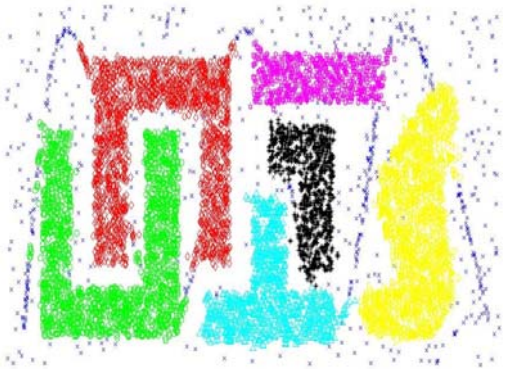
算法只需对数据集进行一次扫描，其时间复杂度为 $O(N)$。从复杂度上可以看出，算法的时间复杂度是数据集大小的线性函数，算法适用对大数据集进行聚类，GICM 算法的时间复杂度低于 SNN 的时间复杂度 $O(N^2)$；低于 Chameleon 算法的时间复杂度 $O(N^2)$。

3.2 聚类的结果的对比

为了验证算法的正确性和精度，笔者做了大量的在实验，在这里仅取其中的两个具有代表性的实验给予具体的说明。图 1、图 2 的数据集都来自于文献[3]，都是带噪声的多密度数据集，其中图 1 样本数为 8000 个，它包含了 6 个不同形状和大小的类以及一些随机分散的孤立点，图 2 样本数为 10000 个，它包含了 9 个不同形状和大小的类以及一些随机分散的孤立点。图 1（a）和图 2（a）是 Chameleon 算法的聚类结果，从图上可以看出虽然它能识别出多密度的类，但是聚类的结果中吸收了附近不少的孤立点，导致聚类精度不高。图 1（b）和图 2（b）是 GICM 算法的聚类结果，可以看出 GICM 有效的识别了多密度的类和孤立点，聚类结果没有吸收孤立点，聚类精度较高。

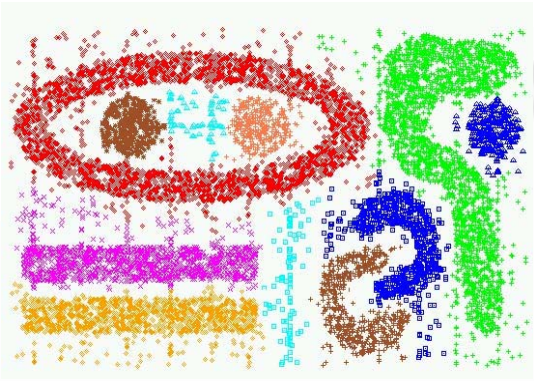


(a) Chameleon 算法

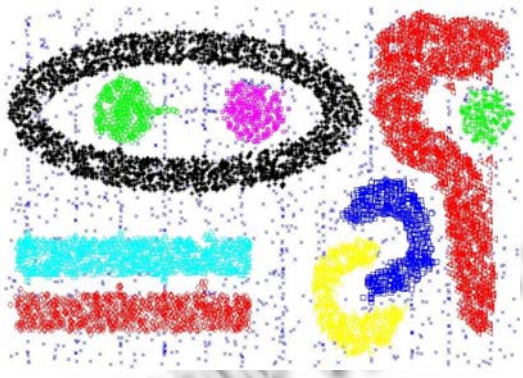


(b)　GICM 算法

图 1　chameleon 算法与 GICM 算法聚类结果比较



(a) Chameleon 算法



(b) GICM 算法

图 2　chameleon 算法与 GICM 算法聚类结果比较

4　结语

本文提出了一种基于网格和信息熵的多密度聚类算法。该算法能识别任意形状和大小的类，提出的边界点处理技术能很好的提取类的边界点，通过对多密度数据集的多次聚类得到了较理想的聚类效果，聚类的精度较高，算法只要求对数据集进行一次扫描，时间复杂度较低，聚类的结果与输入数据的顺序无关，且聚类过程中不需要输入参数，减少了对输入参数的依赖性，实现了参数的自动化。下一步将研究如何改进算法的存储结构，以提高算法对高维的大数据集进行聚类时的效率。

参考文献

1 Han JW, Kamber M.范明,孟小峰译.数据挖掘:概念与技术 第 2 版.北京:机械工业出版社,2007.251–253.

2 Uncu O, Gruver WA, Kotak DB. GRIDBSCAN:Grid density-based spatial clustering of applications with noise. 2006 IEEE International Conference on Systems, Man, and Cybernetics, Taipei, October 8–11, 2006.

3 Karypis G, Han EH, Kumar V. Chameleon: a hierarchical clustering algorithm using dynamic modeling. IEEE Computer, 1999,32(8):68–75.

4 Ertoz L, Steinbach M, Kumar V. Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data. Proc. of the 3rd SIAM International Conference on Data Mining. San Francisco: SIAM Press, 2003. 1–12.

5 Song G, Ying X. Gdcic: a grid-base densityconfidence-interval clustering algorithm for multi-density dataset in large spatial database. Proc. of the 6th International Conference on Intelligent Systems Design and Applications. Washington DC; IEEE Computer, 2006.713–717.

6 赵艳厂,宋梅,采德德,等.用于不同密成聚类的多阶段等密度线算法.北京邮电大学学报,2003,26(2):42–47.

7 夏英,李克非,丰江帆.基于网格梯度的多密度聚类算法.计算机应用研究,2008,25(11):3278–3280.

8 阮吉寿,张华.信息论基础.北京:机械工业出版社,2008.7–11.

9 Hsu CM, Chen MS. Subspace Clustering of High Dimensional Spatial Data with Noises. Heidelberg: Springer, 2004.31–40.

10 Qiu BZ, Li XL, Shen JY. Grid-Based Clustering Algorithm Based on Intersecting Partition and Density Estimation. Proc of PAKDD. Berlin: Springer, 2007. 368–377.

11 程国庆,陈晓云.基于网格相对密度的多密度聚类算法.算机工程与应用,2009,45(1):156–158.