文章编号：1004-4353(2015)03-0254-03

自动连结链聚类算法

李隘优
（闽西职业技术学院 计算机系，福建 龙岩 364021）

摘要：针对传统聚类算法存在时间性能低效且需要输入参数的缺点，本文提出了一种自动连结链聚类新算法.该算法在确立数据的基础上，通过计算数据点与各顶点的距离并加以排序形成不同群组，然后快速搜寻出它们的相邻点形成连结链网络，再根据连结链的平均距离删除过长的连结链，从而达到聚类的目的.实验结果表明，本文算法与 DBSCAN 及 Single-Link 算法具有相同的聚类效果，但执行时间约仅为这两种算法的 10%.

关键词：自动连结链；聚类算法；象限；网络
中图分类号：TP309.3　　**文献标识码**：A

Automatic link-chain clustering algorithm

LI Aiyou
(*Department of Computer, Minxi Vocational & Technical College, Longyan* 364021, *China*)

Abstract: Time performance for shortcomings and inefficiencies of traditional clustering algorithms require input parameters, this paper proposes a new algorithm. The algorithm on the basis of the data side established by the distance calculation of data points and the vertices and be able to sort the formation of different groups, and then quickly find out their adjacent points form a network link chain, according to the average distance is too long and then delete the link chain link chain, which serve the purpose of clustering. Experimental results show that the execution time of the automatic link chain clustering algorithm accounts for about 10% of the common algorithm.

Key words: automatic link-chain; clustering algorithm; quadrant; network

0 引言

群集分析技术[1]由于能够明显突出群体间的差异性，因此被广泛应用于图像识别、数据压缩、影像处理、空间分析和生物信息特征分析等领域.但目前大多群集分析技术算法需要事先给出(或输入)一个或多个参数，而确定适当的这些参数本身就不是一件易事，这不仅加大了聚类分析过程的复杂度，有时也影响了聚类结果[2].例如 *K*-Means 聚类算法[3]中，必需代入参数 k 以确立所要聚类的群体数，并需要反复尝试及验算，才能得到较好的聚类结果，计算量非常大.这类代入参数的聚类算法需建立一套参数范围估算的验算公式，才能有效地执行群集分析.鉴于此，本文引入自动连结链聚类算法(automatic link-chain clustering algorithm, ALC Algorithm)，它无需输入参数，也无需反复针对聚类结果加以尝试及验算.实验表明，该算法既保证了聚类的准确率，又提高了聚类的速度.

1 自动连结链聚类(ALC)算法

自动连结链聚类算法是近年来兴起的一个简

收稿日期：2015-06-28
作者简介：李隘优(1980—)，男，讲师，研究方向为算法分析与设计.

单聚类方式，它有别于基于阶层式聚类法[4]、基于密度聚类法[5]、基于网格聚类法[6]及基于模型式聚类法[7]，但与分割式聚类法[8]接近. ALC 算法在分割群体之前，每一个数据点必需找到各个象限中与它最接近的数据点，然后连结各点形成连结链网络. 因此，从网络中很容易就能发现数据点间的分布关系，通过删除过长的连结链，就能达到快速分割群体的目的. 本文重点讨论如何在这些数据群中，不输入任何参数且只需使用少量的计算就能建立起连结链网络结构.

ALC 算法的具体步骤如下：

1) 找出数据点边界顶点. 寻找出数据点分布的边界，以边界的顶点作为排序基准点. 如对于图 1 所描绘的数据分布，根据数据点分布的情况寻找数据点的边界，并以 A、B、C、D 点作为数据的边界顶点.

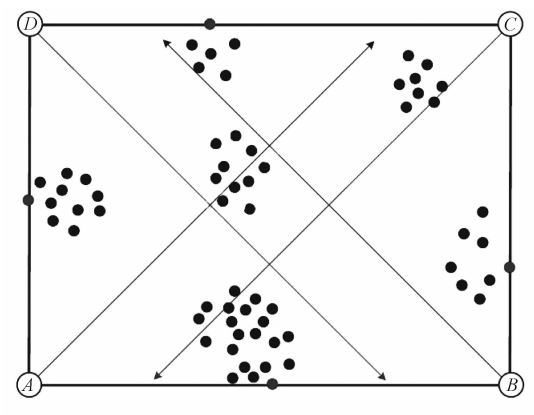

图 1　样本分布边界点

2) 数据点对边界顶点排序，形成不同群组. 如将图 1 中数据集合中的数据点分别对边界顶点 A、B、C、D 进行排序，即计算所有的数据点与 A、B、C、D 点的欧氏距离(Euclidean distance)：

$$d(x_i,x_j)=\|x_i-x_j\|_2=\sqrt{\sum_{d=1}^{k}|x_{id}-x_{jd}|^2},$$

其中 k 为维度；用计算出的距离加以排序形成不同群组，群组排序结果以 Sort A、Sort B、Sort C、Sort D 表示.

3) 快速搜寻邻点，形成连结链网络. 要存储一个连结链网络数据就必须定义一个数据结构，以便记录连结链网络信息. 从任一个数据点开始，根据上面所获得的排序群组来搜寻每个象限中最接近(相邻)的点并计算出它们之间的距离，如果该象限中没有找到相邻的点则以 NULL 表示；重复上述步骤，直到每个数据点都在每个象限中找到它的相邻的点. 最终生成的连结链网络结构如图 2 所示.

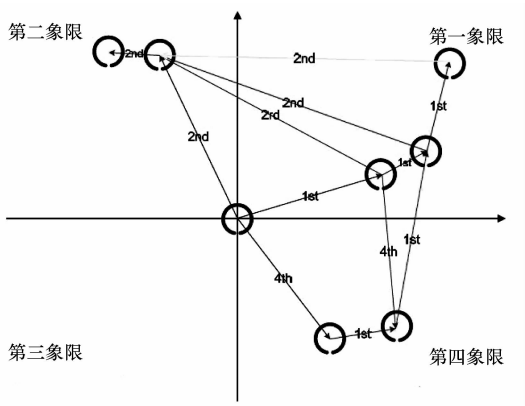

图 2　连结链网络生成图

快速搜寻相邻点的方法，如图 3 所示，以 P 点搜寻第一象限相邻的点为例，在排序群 C 序列中找到与 P 相邻的排序点 $P2$，以 $P2$ 在排序群 C 序列中的位置作为搜寻范围的起点；在排序群 A 序列中找到与 P 相邻的排序点 $P1$，以 $P1$ 在排序群 C 序列中的对应位置作为搜寻范围的终点；在排序群 C 中，由 $P2$(起点) 到 $P1$(终点) 间的范围为搜寻范围(图 3 中虚线所绘区域)，将大幅度地减少寻找相邻点所需要比对数据点的次数，可大大缩短计算机的运算时间.

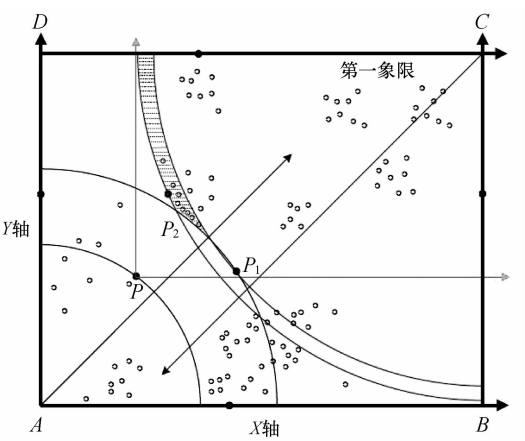

图 3　P 点搜寻第一象限的最近点

4) 计算连结链的平均距离. 生成连结链网络后,必须进行数据分析才可进一步进行聚类. 根据所有的连结链的距离(即数据点间的差异性)计算出所有连结链的平均值(D_{AVG}),并以此作为聚类的标准.

$$D_{AVG}=\sum_{1}^{N} Des/N,$$

其中 $Des \in (Des\,A, Des\,B, Des\,C, Des\,D)$, N 为有效的连结链总数.

5) 删除过长的连结链. 根据聚类的评估标准值 D_{AVG} 进行聚类,删除连结链网络中所有连结链距离大于 D_{AVG} 的数据,被删除的数据点归到不同的群组中,存在连结链关系的数据点归为同一群组,从而使数据产生群聚,形成聚类效果.

6) 群集分析统计及群集聚类定义. 分别计算出每一群组的数据点数,然后求出群组的平均数据点数量 G_{AVG}(群平均数): $G_{AVG}=\sum_{1}^{GN} G_{num}/GN$, 其中 GN 表示群组数量, G_{num} 表示每一个群组的数据点数. 以 G_{AVG} 值作为群集分析的依据, $G_{num} \geqslant G_{AVG}$ 的为真实群, G_{num} 界于 $\frac{1}{2}G_{AVG}$ 与 G_{AVG} 之间的为模糊群, $G_{num} < \frac{1}{2}G_{AVG}$ 的为噪声群,由此达到聚类的结果.

2 实验与分析

为分析 ALC 算法的性能,与传统 Single-Link 算法、DBSCAN 算法进行了对比分析. 聚类算法执行时间的实验设计:分别输入 500、1 000、1 500、2 000、2 500、3 000、3 500、4 000、4 500、5 000个随机样本点,数据集合为二维坐标点,数据数值范围为±200,分别使用 ALC、DBSCAN、Single-Link 聚类算法进行群集分析,其中 DBSCAN、Single-Link 聚类算法分别参考文献[5]与文献[4]. 由 ALC 进行群集分析之后,将所得到的 D_{AVG} 作为 DBSCAN 算法的 Eps 及 Single-Link 算法的 Threshold 参数值. 不同算法执行时间的结果如表 1 所示.

表 1 不同聚类算法执行时间的比较

样本数量	DBSCAN 及 Single-Link 算法的参数值			执行时间/s		
	Eps	MinPts	Threshold	ALC 算法	DBSCAN 算法	Single-Link 算法
500	11.94	5	11.94	0.001	0.045	0.018
1 000	11.21	5	11.21	0.017	0.181	0.074
1 500	9.33	5	9.33	0.016	0.412	0.139
2 000	8.14	5	8.14	0.032	0.722	0.253
2 500	8.27	5	8.27	0.034	1.068	0.398
3 000	7.48	5	7.48	0.050	1.539	0.059 9
3 500	7.52	5	7.52	0.070	2.108	0.816
4 000	7.23	5	7.23	0.077	2.720	1.102
4 500	6.56	5	6.56	0.085	3.471	1.330
5 000	6.28	5	6.28	0.096	4.460	1.766

由表 1 可知,3 种聚类法都能将非固定形状样本数据分出相同的群组数量,但 DBSCAN 及 Single-Link 算法都需反复调整 Eps 及 Threshold 参数值才能得到较佳的聚类效果. 3种聚类算法的效果虽然相同,但执行效率上 ALC 算法所消耗的时间远小于 DBSCAN 和 Single-Link 算法,如图 4 所示.

(下转第 266 页)

窗的研究还较为不足,今后将对此做进一步地研究,使之更加贴近实际情况.

参考文献:

[1] 张晓龙.电子商务下现代物流企业配送系统优化研究[J].物流技术,2011,30(6):135-138.
[2] 柳林,朱建荣.基于遗传算法的物流配送路径优化问题的研究[J].计算机工程与应用,2005(27):227-229.
[3] 胡大伟,朱志强,胡勇.车辆路径问题的模拟退火算法[J].中国公路学报,2006,19(4):123-126.
[4] 王雪莲,汪波,钟石泉.一类半开放式车辆路径问题及其晋江算法研究[J].机系统仿真学报,2008,20(8):1969-1972.
[5] 杨从平.基于蚁群算法的快递物流配送路径优化[J].物流工程与管理,2014,36(4):27,65-67.
[6] 蒋国清,潘勇,胡飞跃.两阶段式的物流配送路径优化方法[J].计算机工程与应用,2015,51(2):255-258.
[7] 任璐.基于遗传算法的建立与求职岗位匹配研究[D].广州:暨南大学,2009:22.
[8] 宋娟,崔艳.基于改进遗传算法的同城快递配送模型[J].电子技术应用,2014,40(12):136-139.

(上接第 256 页)

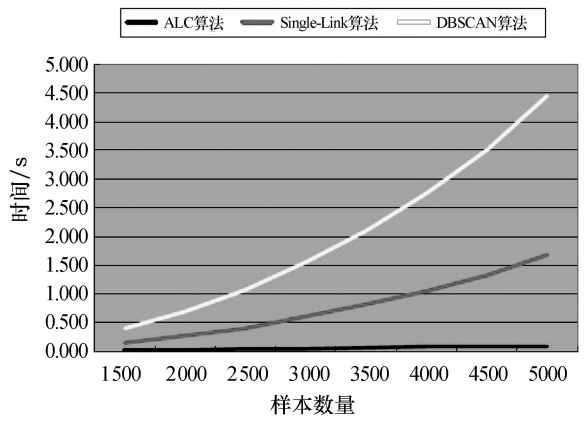

图 4 不同算法执行时间的比较

3 结论

本文提出的自动连结链聚类算法克服了其他聚类算法使用参数而对聚类结果及效率产生不良影响的问题,在不需要输入参数的情况下,通过引入自动连结链的方法,增加了数据点的线性可分概率,即扩大了数据类之间的差异,从而提高了聚类的质量.实验表明,该算法与 Single-Link 聚类算法及 DBSCAN 聚类算法具有相同的聚类效果,但其执行时间仅为这两种算法的 10%,大大节省了运算时间.由于目前的研究只针对了二维数据样本上的聚类实验,高维度以及其他形态的数据集有待今后进一步地研究.

参考文献:

[1] 张晓伟.聚类算法及在搜索引擎系统中的应用[D].哈尔滨:哈尔滨理工大学,2014.
[2] Mali U, Bandyopadhyay S. Genetic algorithm-based clustering technique [J]. Pattern Recognition, 2000,33(9):1455-1465.
[3] 王辉,张望,范明.基于集群环境的 *K*-Means 聚类算法的并行化[J].河南科技大学学报(自然科学版),2008,29(4):42-45.
[4] 赵玉艳,郭景峰,郑丽珍,等.一种改进的 BIRCH 分层聚类算法[J].计算机科学,2008,35(3):180-183.
[5] 熊忠阳,吴林敏,张玉芳.针对非均匀数据集的 DBSCAN 过滤式改进算法[J].计算机应用研究,2009,26(10):3721-3723.
[6] Song B C, Ra J B. A fast search algorithm for vector quantization using L2-norm pyramid of codewords[J]. IEEE Transactions Image Processing, 2002,59(11):10-15.
[7] Jim Z C Lai, Yi-Ching Liaw. Fast-searching algorithm for vector quantization using projection and triangular inequality[J]. IEEE Transactions Image Processing, 2004,13(12):1554-1562.
[8] 吴昌友,王福林,马力.采用链路聚类的动态网络社团发现算法[J].西安交通大学学报,2014,48(8):73-79.