

一种基于引力的分层聚类算法

贾瑞玉, 查 丰, 耿锦威, 宁再早

(安徽大学 计算机科学与技术学院, 安徽 合肥 230039)

摘 要:传统的分层聚类算法在聚类过程中,仅使用样本间的距离作为相似度的唯一标准,其描述过于单一。考虑到宇宙中星系的形成过程本质也是一种聚类过程,星系之间吸引力是靠万有引力作用。将万有引力思想引入分层聚类中,提出一种基于引力的层次聚类算法 HCBG(Hierarchical Clustering Base Gravity),从样本间的距离和类簇的大小两个方面更加精确地刻画相似度。把分层聚类的过程看成样本点之间依据“万有引力”自发吸引的过程。采用 UCI 机器学习数据库的 Iris, Wine 和 Glass 数据集,实验结果表明,提出的 HCBG 算法的聚类结果比经典的基于距离的层次聚类 HC(Hierarchical Clustering)提高 5%~10% 左右。

关键词:引力;分层聚类;相似度

中图分类号:TP301.6

文献标识码:A

文章编号:1673-629X(2011)03-0076-03

A Hierarchical Clustering Algorithm Based on Gravity

JIA Rui-yu, ZHA Feng, GENG Jin-wei, NING Zai-zao

(School of Computer Science and Technology, Anhui University, Hefei 230039, China)

Abstract: The traditional hierarchical clustering algorithm for clustering process, only uses the distance between samples as the sole criterion for similarity, this description is too simple. Associated with the formation of galaxies in the universe is essentially a clustering process by gravitational attraction between galaxies role. Introduce the idea of hierarchical gravitational clustering, propose a hierarchical clustering algorithm based on gravitational HCBG (Hierarchical Clustering Base Gravity), from two aspects of the distance between the samples and the cluster size classes more accurately depicts the similarity. The hierarchical clustering process is regarded as the sample points based on "gravity" to attract spontaneous process. Use UCI machine learning database: Iris, Wine and Glass as data sets, experimental results show that the proposed algorithm HCBG clustering results than classical hierarchical clustering based on distance HC (Hierarchical Clustering) increase 5%~10% or so.

Key words: gravity; hierarchical clustering; similarity

0 引 言

聚类分析是数据挖掘的一个非常活跃的研究分支,广泛应用于模式识别、图像处理、市场研究、决策支持等领域。目前已提出的聚类算法有很多,这些算法可以被分为基于划分方法、基于层次方法、基于密度方法、基于网格方法和基于模型方法^[1]。层次聚类分析由于其强壮性和对于输入记录的顺序不敏感及简单易用而成为聚类分析中的一个广泛应用的算法^[2,3]。

蒋盛益在文献[4]中提出的引力聚类方法,首次将类的大小,即类簇包含的样本点数,作为衡量相似度的一部分;石剑飞等在文献[5]中介绍了凝聚型层次聚类的基本算法和优缺点;梁斌梅等在文献[6]中提

出基于层次聚类的孤立点检测,对距离矩阵按簇间距离从大到小检测孤立点,可检测出指定离群程度的孤立点。文中将万有引力的思想加入分层聚类中,提出一种基于引力的层次聚类算法 HCBG(Hierarchical Clustering Base Gravity),该算法用引力作为相似度度量的标准,该方法的特点是对相似度的刻画更精确,将相似度度量标准加入类的大小等因素,可以准确地度量类之间的相似性。

1 基本概念

1.1 万有引力的基本原理

宇宙中各星系的形成和笔者对数据集的聚类分析极为类似,每个星系可以看作聚类分析中的一个类。在宇宙形成初期,各种物质杂乱无章地分布在宇宙中各个角落,由于万有引力的存在,使得两个引力大的物质聚集在一起进而演化成星系。

万有引力公式如下:

收稿日期:2010-06-19;修回日期:2010-09-25

基金项目:安徽省自然科学基金项目(KJ2008B092)

作者简介:贾瑞玉(1965-),女,副教授,研究方向为数据挖掘、人工智能、计算机图形学。

$$g = G \frac{M \bullet m}{r^2} \quad (1)$$

其中:

M, m 表示两个星系的质量;

r 表示两个星系的距离;

G 表示引力常数。

可以看出,在宇宙星系形成过程中,不仅仅是距离在起作用,质量(即聚类的大小)也同样影响到聚类的结果。

1.2 层次聚类的基本原理

基于层次聚类的方法可以分为自顶向下的层次聚类(分裂层次聚类)和自底向上的层次聚类(凝聚层次聚类)。凝聚层次聚类的策略是首先将每个样本点都看作一个类,然后合并这些原子类为越来越大的类,直到所有的样本点都聚为一类,或满足某终止条件。大部分层次聚类使用这种方法。不同的层次聚类算法在每一层上合并类的方式也不同^[7-9]。

1)最小距离,又称单连接或最近邻方法。取两个聚类间样本的最近距离作为两个类间距离。这种方法在大数据样本的情况下,可能会产生链式现象,当两个类之间有彼此很靠近的点时,这两个类就被合并为一类。

2)最大距离,又称全连接或最远邻方法。取两个聚类间样本的最远距离作为两个类间距离。这种算法易找出紧凑的类簇。

3)平均距离,又称平均连接方法。取两个聚类间样本的平均距离作为这两个类间距离。这种方法考虑了类簇的结构,产生的聚类结果有相对的鲁棒性^[10-12]。

层次聚类算法(HC算法)的主要步骤^[13]如下:

Step1:输入最终的聚类数。将每个样本点作为一个类簇,初始化距离矩阵。

Step2:如果类簇 C_i 和 C_j 之间的平均距离最小,则合并 C_i 和 C_j ,并更新距离矩阵。

Step3:对合并后的类簇,计算平均距离,更新距离矩阵。

Step4:如果聚类数达到预先指定的个数,则聚类结束;否则转至 Step2。

2 基于引力的层次聚类算法(HCBG 算法)

2.1 类间的引力公式

定义类 C_1 与 C_2 之间的引力为:

$$g = G \frac{N_{C1} \bullet N_{C2}}{r^2} \quad (2)$$

由于在该算法比较时,不需要计算准确的引力大小,只是需要比较两个引力相对大小。所以(2)式中 G

可以去掉。则引力公式变为:

$$g = \frac{N_{C1} \bullet N_{C2}}{r^2} \quad (3)$$

其中:

N_{C1}, N_{C2} 表示聚类 C_1 和聚类 C_2 所含样本数;

r 表示两个聚类之间的距离(采用欧式距离计算)。

但是当有部分区域非常密集时,会出现类似黑洞的现象。也就是说,某一个类簇包含的样本非常多,从而它对其他类簇的引力巨大,把其他本不属于它的类簇吞并了。为了解决这种缺陷,改进引力公式为:

$$g = \frac{\ln(\ln(N_{C1})) * \ln(\ln(N_{C2}))}{r^2} \quad (4)$$

2.2 HCBG 算法

算法步骤如下:

Step1:输入最终想要得到的聚类个数 c ;

Step2:将每一个样本点作为一类,计算两两之间的引力 $g(i, j)$, $i, j = 0, 1, \dots, n$, 得到初始化矩阵 I ;

Step3:将引力最大的两个类合并成一个类;

Step4:计算合并成的新类的质心;

Step5:重新计算新的类与所有类之间的引力,更新引力矩阵;

Step6:重复 Step3 ~ Step5,直到合并成想要得到的聚类个数为止。

在本算法中,类簇之间引力是用其质心之间的引力计算。

2.3 HCBG 算法流程图

HCBG 算法流程图见图 1。

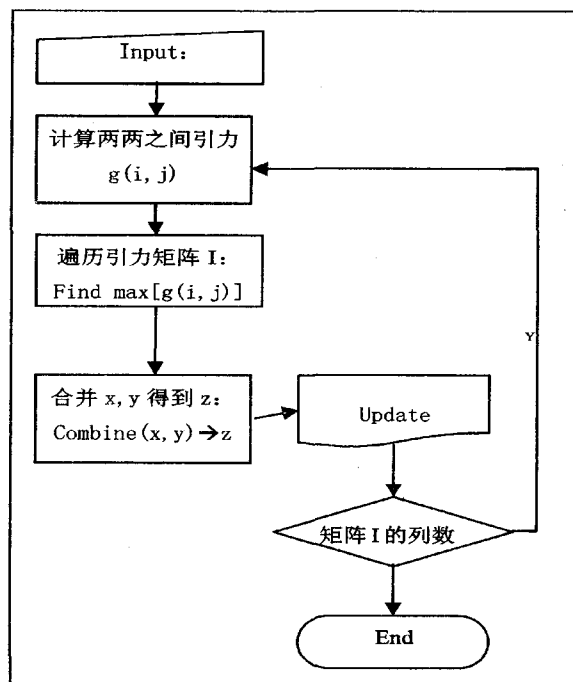


图 1 HCBG 算法流程图

3 实验结果及分析

文中选用的测试数据集是来自 UCI 机器学习数据库的 Iris, Wine 和 Glass(见表 1), 测试数据集分别用 HCBG, HC 和 k-means 算法进行测试。所有实验都在 CPU 为 Intel Pentium4 3.0GHz, 1.00GB 内存, 编程环境为 JAVA 下完成的。

表 1 实验数据库

数据库	样本数	样本特征数	类数
Iris	150	4	3
Wine	178	13	3
Glass	214	9	6

表 2 中 HCBG 算法采用的是平均引力大小来表示类与类之间的相似度。如果采用最小引力或最大引力作为相似度度量, 结果不理想。这是由于: 一方面, 这和宇宙中万有引力计算公式的前提条件违背, 宇宙是浩瀚无穷的, 其中的每个星系可以看作一个质点, 而星系的直径比起宇宙直径微乎其微。而在测试的数据集中, 每个类的直径和总的数据集直径相比, 不能忽略不计, 所有采用平均引力来计算。另一方面, 经过实验验证, 采用最小引力或最大引力将形成类似黑洞的效应, 也就是说某一类包含的样本非常多, 而其他类包含的样本稀少。这是由于某些类的边缘数据十分靠近, 根据引力公式, 这些聚类间的引力比较大, 从而会聚成更大的类, 如此下去其中某一类将会变成“黑洞”, 把所有的类都会吸到其中。

表 2 HCBG 算法和 HC 算法的测试结果(正确率)

数据库	HCBG 算法	HC 算法	k-means 算法
Iris	94.6%	89.2%	89.3%
Wine	87.0%	68.3%	69.4%
Glass	73.9%	52.4%	<50%

从表 2 可以看出, HCBG 算法比传统的 HC 算法和经典的 k-means 算法聚类精度要高, 这表明 HCBG 算法是有效的。从实验结果也可看出, 将相似度度量标准加入人类的大小等因素, 可以准确地度量类之间的相似性。

4 结束语

文中提出的基于引力的层次聚类算法 HCBG 能够聚成不同形状类簇, 通过在三个数据集进行对比试验, 聚类效果明显高于经典 HC 和 k-means 算法。但是本算法也存在一些局限性, 如: 算法的空间和时间复杂度较高等等。

参考文献:

- [1] Han Jiawei, Kamber M. 数据挖掘: 概念与技术[M]. 第 2 版. 范明, 孟小峰, 译. 北京: 机械工业出版社, 2007: 251-252.
- [2] Jiang Shengyi, Li Xia. A Hybrid Clustering Algorithm[J]. Fuzzy Systems and Knowledge Discovery, 2009(1): 366-370.
- [3] Fisher R A. UCI repository of machine learning databases. iris [EB/OL]. 1998. <http://archive.ics.uci.edu/mYdatasets.html>.
- [4] 蒋盛益, 李庆华. 一种基于引力的聚类方法[J]. 计算机应用, 2005, 25(2): 286-300.
- [5] 石剑飞, 闫怀志, 牛占云. 基于凝聚的层次聚类算法的改进[J]. 北京理工大学学报, 2008, 28(1): 66-69.
- [6] 梁斌梅. 基于层次聚类的孤立点检测方法[J]. 计算机工程与应用, 2009, 45(32): 117-119.
- [7] 王鑫, 王洪国, 王珏, 等. 数据挖掘中聚类方法比较研究[J]. 计算机技术与发展, 2006, 16(10): 20-25.
- [8] 段明秀, 杨路明. 对层次聚类算法的改进[J]. 湖南理工学院学报(自然科学版), 2008, 21(2): 28-36.
- [9] 游芳, 姜建国, 张坤. 基于二维属性的高维数据聚类算法研究[J]. 计算机技术与发展, 2009, 19(5): 111-113.
- [10] 李光强, 邓敏. 一种双距离空间的聚类算法[J]. 测绘学报, 2008, 37(4): 482-487.
- [11] 李泽文. 基于 WEB 的数据挖掘技术[J]. 现代计算机, 2004(11): 29-32.
- [12] Han Jiawei, Kamber M. 数据挖掘概念与技术[M]. 范明, 孟小峰, 译. 北京: 机械工业出版社, 2006.
- [13] 陈学进. 数据挖掘中聚类分析的研究[J]. 计算机技术与发展, 2006, 16(9): 44-49.

(上接第 75 页)

- [6] 黄智维, 倪子伟. 网格计算环境下资源管理的研究[J]. 计算机技术与发展, 2009, 19(3): 200-203.
- [7] Hsien-Po Shiang, van der Schaar M. Distributed Resource Management in Multihop Cognitive Radio Networks for Delay-Sensitive Transmission[J]. IEEE Transactions on Vehicular Technology, 2009, 58(2): 941-953.
- [8] 李育强, 罗光春. Web2.0 构建网格资源平台技术研究[J]. 电子科技大学学报, 2007, 36(6): 1389-1392.
- [9] 何清林, 杨森, 徐泽同. 基于元数据和 Web Service 中间件的分布式资源库集成[J]. 计算机工程与设计, 2009, 30

(9): 2201-2204.

- [10] 陈东毅. 面向服务的分布式技术在网络教学平台中的应用研究[J]. 软件导刊, 2009, 8(5): 141-142.
- [11] 唐爱国, 罗新密, 杭志. 基于 J2EE 网络教学平台的研究与应用[J]. 计算机技术与发展, 2009, 19(6): 236-239.
- [12] 付长青. 公共计算机课程网络教学平台的设计与实现[D]. 北京: 北京工业大学, 2009.
- [13] Gibb G, Lockwood J W, Naous J, et al. Design and Realization of Network Teaching Platform Based on E-learning[J]. IEEE Transactions on Education, 2008, 51(3): 364-369.