

文章编号:1001-9081(2006)04-0870-02

基于动态矩形的聚类方法的设计与实现

高原,耿国华,王怡

(西北大学 信息科学与技术学院,陕西 西安 710068)

(tutuyuan@sina.com.cn)

摘要:提出了一种新的基于动态矩形的聚类方法 DRCA。该方法减少了参与聚类计算的数据元素的数量,在每一次基本聚类过程中,采用数据之间空间位置比较取代复杂的聚类距离函数计算,使得算法复杂度与数据量具有近似线性时间关系。试验结果表明了 DRCA 的正确性和有效性。

关键词:数据挖掘;聚类;距离函数

中图分类号: TP311.13 **文献标识码:** A

Design and realization of dynamic rectangle-based clustering approach

GAO Yuan, GENG Guo-hua, WANG Yi

(Institute of Information Science and Technology, Northwest University, Xi'an Shaanxi 710068, China)

Abstract: A dynamic rectangle-based clustering approach(DRCA) was presented. The number of data that needed to be examined for clustering was reduced, and in each procedure of clustering, the position comparison between numbers was used without any distance comparison, which made DRCA had the nearly linear time complexity with the size of dataset. The experiment results show that the DRCA is correct and efficient.

Key words: data mining; clustering; distance function

0 引言

数据挖掘是发现隐藏在大型数据库中有意义的、潜在的信息模式的过程,聚类分析是数据挖掘领域的一个非常活跃的研究分支。聚类是一个在数据库中发现“群”或“簇”的过程,根据定义聚类时所采用的技术,可以分为:层次的、分区的、基于密度的和基于网格的聚类算法^[1]。其中,分区的方法是最常用的聚类分析方法之一,算法以每个模式作为单一聚类开始,迭代重新分配数据点到每个聚类,直至满足某个终止的标准为止。

k-均值是最常用的分区算法,被广泛应用于机器学习^[2,3],模式识别^[4,5]以及统计学中^[6]。但是 k-均值算法在进行超大型真实数据集上的聚类挖掘中计算量大,效率低;并且由于初始聚类中心的随机选取,聚类过程的结果容易陷入局部最优解^[7]。

本文设计并实现了一种基于动态矩形的聚类算法 DRCA (Dynamic Rectangle-based Clustering Approach),算法效率有所提高,并通过试验证明了算法的正确性。

1 动态矩形的聚类算法

1.1 基本概念

为叙述方便,先引入以下定义:

定义 1 设数据集 D 中的属性集由分类属性集 C 和数值属性集 S 组成,则 D 中任一元素 d_i 由 k 个属性组成: $d_i = \{d_{i1}, d_{i2}, \dots, d_{in}, \dots, d_{ik}\}$,称 d_{in} 为 d_i 的一个属性分量, $d_{in} \in C \cup S$,分类属性分量个数 C_n 与数值属性分量个数 S_n 之和 $(C_n + S_n) \geq k$ 。

定义 2 设集合 I 为数据集 D 中的 n 个元素组成的集合, $I = \{(i_1, \dots, i_j, \dots, i_n) \mid i_j \in D, 1 \leq j \leq n\}$, 给定一个阈值 T ($T > 0$), 对于 I 中的任意一个元素 i_j , 总存在元素 $i_k \in I$, 使得 $q(i_j, i_k) = \|i_j - i_k\|^2 < T$, 则称 I 为一个类。

定义 3 元素相关矩形。设 d_i 为集合 D 中任一元素, 存在以 d_i 为顶点的矩形 R_i , R_i 中有以 d_i 为顶点的边 L_i , 则称以 d_i 为原点, L_i 为轴旋转至 $0^\circ, 90^\circ, 180^\circ, 270^\circ$ 位置上获得的四个矩形为 d_i 元素相关矩形。

1.2 基于动态矩形的聚类算法描述

目前有众多的 k-均值改进算法, 这些优化方法主要通过两种途径改进 k-均值的性能:

- 1) 优化初始聚类的选取方法;
- 2) 改变距离函数的度量和降低距离函数的比较次数的方法。

本文基于第二种优化思想, 提出基于动态矩形的聚类算法, 算法描述如下:

输入: 数据集 D , 异常点阈值 D_y , 聚类阈值 D_j

输出: 分类个数及类中元素的集合

步骤 1: 读入一个新的元素 d_i ;

步骤 2: 统计 d_i 的相关矩形中包含的未被访问元素个数;

步骤 3: 选取元素个数最多的矩形, 判断其元素的个数是否小于 D_y , 若是, 则将其标注为临时聚类, 转步骤 7; 否则, 以该相关矩形为基础构造一个新的类;

步骤 4: 沿着水平和垂直的方向移动矩形, 直到移动矩形不会增加元素的数量;

收稿日期: 2005-10-08 基金项目: 国家自然科学基金资助项目(60271032)

作者简介: 高原(1975-), 女, 陕西清涧人, 博士研究生, 主要研究方向: 信息系统与人工智能; 耿国华(1955-), 女, 山东人, 教授, 博士生导师, 主要研究方向: 信息系统与人工智能; 王怡(1971-), 男, 天津人, 博士研究生, 主要研究方向: 计算机图形图像。

步骤5:将矩形在一定范围内进行等值放大,使该类中包含尽量多的元素,直到矩形范围的增长不会明显增加元素的数量;

步骤6:更新该类中元素的各项分类属性值的统计频度及数值属性的质心;

步骤7:判断 D 中元素是否均被访问,如果不存在未被访问的元素,转步骤8;否则,选取离当前类距离最远的元素 d_i ,转步骤2;

步骤8:判断临时聚类与距离最近的现有类之间的距离是否小于阈值 D_j ,若是,则将该临时聚类合并到对应类中;否则,标注为异常点;

步骤9:保存并输出每个类的信息,结束。

1.3 阈值的选取

聚类阈值 D_j 的选取直接影响算法执行的结果及时间复杂度,当 D_j 大到一定程度时,只能得到少数的类甚至一个类,而当 D_j 的取值太小时,得到众多小的聚类,实践中 D_j 太大或太小都不能得到有意义的聚类。为了确定聚类 D_j 的大小,本文采用了文献[11]中提出的确定聚类阈值的抽样技术,并对其进行改进以便适应 DRCA 算法,从而获得聚类阈值的初始值,这种通过抽样技术确定聚类阈值的方法特别适合大数据集的情况。算法描述如下:

1) 在数据集 D 中随机选择 N 个对象;

2) 计算 N 个对象中两两之间的距离:

$$q(i_j, i_k) = \|i_j - i_k\|^2;$$

3) 计算2)中距离的平均值 EX ;

4) 取 D_j 的值介于 $EX/2$ 与 $EX/3$ 之间。

2 实验结果

2.1 实验环境及数据集

为了验证算法的正确性和有效性,本文在UCI^[12]中的多个数据集上进行测试,实验硬件环境为联想2.5 GHz/256M个人计算机,实验所涉及的主要软件环境为:Microsoft Windows XP, Oracle 9i, Visual C++ 6.0。

实验中涉及到的数据集见表1。

表1 实验中涉及到的数据集

数据集名称	记录个数	分类属性	类别
iris	150	4	3
soybean-large	683	35	19
new-thyroid	215	5	3

2.2 性能测试

将本文提出的 DRCA 算法与 k 均值算法进行运行时间比较测试实验,实验结果如表2所示。

表2 k 均值算法与 DRCA 算法性能比较测试

数据集名称	处理方法	处理时间/s	准确率(%)
iris	k-means	9.4	95.72
	DRCA	7.2	96.9
soybean-large	k-means	18.2	91.86
	DRCA	10.5	95.54
new-thyroid	k-means	15.7	91.79
	DRCA	8.6	94.36

2.3 算法评价

实验表明,采用本文提出的 RCA 算法进行聚类分析,由于其核心算法采用简单的数据分布的位置坐标比较实现,算法时间复杂度与数据对象的数目呈近似线性关系,其整体效率优于传统的 k 均值聚类算法。

传统的聚类分析算法由于采用基于欧几里德距离的相似性度量方法,发现的聚类通常是一些球状的、大小和密度相近的类,而 DRCA 可以发现具有任意形状的聚类,聚类质量高,适用范围广。

3 结语

本文设计并实现了一种有效的基于动态矩形的聚类方法 DRCA,通过降低距离函数的比较次数和度量方法的复杂度改善聚类性能,实验数据表明,DRCA 聚类算法可以在保证聚类质量的前提下,提高聚类的速度。然而算法也存在着不足,首先,每一次迭代过程中数据点的搜索算法是下一步需要改进和完善的工作,并且,聚类结果的内涵知识获取是需要长期进行的研究工作。

参考文献:

- [1] 韩家, KAMBER M. 数据挖掘概念与技术[M]. 北京: 机械工业出版社, 2001
- [2] MOODY J, DARKEN C. Fast learning in networks of locally-tuned processing units[J]. Neural Computation, 1989, 1(2): 281-249.
- [3] ESKIN E, ARNOLD A, PRERAU M, et al. A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data[A]. In data mining for security applications[C]. 2002.
- [4] LLOYD SP. Least squares quantization in PCM[J]. IEEE Trans on Information Theory, 1982, IT-28(2): 129-137.
- [5] 王熙照, 王亚东, 湛燕, 等. 学习特征权值对 k -均值聚类算法的优化[J]. 计算机研究与发展, 2003, 40(6): 869-873
- [6] KRISHNAPURAM R, KELLER JM. A possibilistic approach to clustering[J]. IEEE Transactions on Fuzzy Systems, 1993, 1(2): 98-110.
- [7] SELIM SZ, ISMAIL MA. K-means type algorithms: a generalized convergence theorem and characterization of local optimality[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1984, 6(1): 81-87.
- [8] KHAN SS, AHMAD A. Cluster center initialization algorithm for k-means clustering[J]. Elsevier science inc, 2004, 25(11): 1293-1302.
- [9] HUANG Z. Extensions to the K-means algorithm for clustering large data sets with categorical values[J]. Data Mining knowledge discovery, 1998, 2(3): 283-304.
- [10] 闫德勤, 迟忠先. 一种新的聚类算法[J]. 小型微型计算机系统, 2004, 25(11): 1984-1985.
- [11] 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): 286-288.
- [12] HETTICH S, BLAKE CL, MERZ CJ. UCI Repository of machine learning databases[EB/OL]. <http://www.ics.uci.edu/~mlearn/MLRepository.html>, 1998.