

文章编号:1001-9081(2005)02-0294-03

基于网格距离的聚类算法的设计、实现和应用

田启明^{1,2}, 王丽珍¹, 尹群¹

(1. 云南大学信息学院, 云南昆明 650091; 2. 温州职业技术学院 计算机系, 浙江温州 325035)
(slm78_8@hotmail.com)

摘要:提出了一种新的基于网格距离的聚类算法。该算法不仅克服了K-代表点算法中必须先给定K值、难以确定初始代表点、聚类结果的现实意义难以描述等缺点,而且克服了基于网格的聚类算法中要求数据必须在空间密集的缺陷。通过实验验证了新算法的正确性和有效性。

关键词:数据挖掘; 聚类; 网格; K-均值聚类; 相似度量; 内涵知识

中图分类号: TP311 **文献标识码:** A

Design, realization and application of clustering algorithm based on the distance between grids

TIAN Qi-ming^{1,2}, WANG Li-zhen¹, YIN Qun¹

(1. College of Information Engineering, Yunnan University, Kunming Yunnan 650091, China;
2. Department of Computer Science, Wenzhou Vocational and Technological College, Zhejiang Wenzhou 325035, China)

Abstract: This paper presented a new clustering algorithm based on the distance between grids. The new algorithm not only overcame the shortcoming of K-Medoid algorithm which has much difficulties to suppose K in advance, confirm the initialized points and explain realistically, but also overcame the shortcoming of clustering algorithm based on grids which requests dense data in spaces. The new algorithm was proved to be correct and efficient by the results of experiments.

Key words: data mining; Clustering; grid; K-means Clustering; similarity metrics; intentional knowledge

0 引言

聚类挖掘是数据挖掘中的一个重要研究领域。聚类需要解决的问题是将已给定的若干无标记的模式聚集起来使之成为有意义的类^[1]。现存的聚类分析算法可以分为以下几类:基于划分的方法^[2]、基于层次的方法^[3]、基于密度的方法^[4]、基于网格的方法^[5-7]和基于模型的方法^[8]等。其中基于划分的方法是最常用的聚类分析方法之一,它最典型的代表算法是基于K-平均值和基于K-代表点的聚类算法。

最早由MacQueen^[9]提出了K-平均值算法。在这个算法中,每个类用该类中现有对象的平均值表示。由于该算法对于脏数据非常敏感,于是由Kaufman和Rousseeuw^[2]提出了基于K-代表点的PAM和CLARA算法。这些算法与K-平均值算法的主要区别在于,每个类不是用样本的平均值代表,而是用接近聚类中心的一个称为medoid的对象(中心点)来表示。这种算法对于脏数据和异常数据不敏感,但计算量显然要比K均值要大,一般只适合小数据量。

无论是在K-平均值算法中,还是在K-代表点算法中,都存在着如下缺陷:1)必须先假定类的个数,即K值。但在进行真实数据的聚类挖掘时,K值很难事先确定。2)初始聚类中心的随机选取,可能会陷入局部最优解,而难以获得全局最优解。3)难以消除噪声的影响。4)聚类结果的现实意义难以描述。

本文提出了一种新的基于网格距离的聚类算法,在新算

法中一方面采用基于K增量的代表点聚类算法弥补了K-平均值算法和K-代表点算法的前面个缺陷,另一方面网格工具的引入不仅有效地解决了K-平均值算法和K-代表点算法的后面个问题,而且提高了算法的运行效率。

1 基于网格距离的聚类算法

1.1 三个定义

定义1 类

设 G 为元素的集合,它共有 m 个元素,记为 $g_i, i=1,2,\dots,m$,另外给定一个阈值 $T>0$,对 G 中任意一个元素 g_i ,总存在另一个元素 g_j ,它们的距离不大于阈值 T ,即有 $d_{ij} \leq T$,则称 G 为类。

定义2 相邻网格

两个 K 维网格 u_1 和 u_2 ,若它们之间有一个公共面,或是存在另一个 K 维网格 u_3 , u_3 与 u_1 相邻, u_3 也与 u_2 相邻,则 u_1 和 u_2 为相邻网格。

定义3 网格间距

两个 K 维网格间距用网格中心点之间的欧几里得距离表示。

定义4 异常点。异常点包括噪声和例外两种情况。

1.2 基于网格距离的聚类算法描述

1.2.1 基于网格距离的聚类算法

输入:inFormation表(聚类对象),异常点阈值 D_t ,聚类阈

收稿日期:2004-07-15;修订日期:2004-10-19 基金项目:云南省自然科学基金资助项目(2002F0013M)

作者简介:田启明(1974-),女,江西赣州人,讲师,硕士研究生,主要研究方向:数据挖掘;王丽珍(1962-),女(纳西族),云南丽江人,教授,主要研究方向:数据挖掘;尹群(1970-),男,浙江宁波人,高级工程师,硕士研究生,主要研究方向:数据挖掘。

值 D_i

输出:区域表、类_记录数表(用于存放分类的个数和每类的记录总和)

方法:

1) 将每一维划分区间,实现多维空间的网格化。对网格进行编码并统计每个网格中的记录个数。每个网格用网格中心点作为代表点。

2) 判断每个网格中的记录是否小于异常点阈值 D_e ,若小于,则标记该网格为异常点。

3) 对每个非异常点的网格,计算与其他网格之间的距离存入数组 $Comp$ 中($Comp(i,j)$ 则表示第 i 个非异常点网格与第 j 个非异常点网格之间的距离)。

4) 找出距离最远的两个非异常点网格分别代表两个初始类。

5) 判断其他的未被分类的非异常点网格与距离最近的现有类的代表网格之间的距离是否小于聚类阈值 D_i ,若小于,则将该网格分配到对应类中。否则将该网格标记为一个新类。一直迭代直至所有网格分类完毕,并生成类_记录数表。

6) 将同一类中的相邻网格进行合并,生成区域表。

1.2.2 算法的几点说明

网格编码:网格是通过其每一维所属区域的流水号进行编码的。如在图1中所示的二维网格中,网格 c 的编码为13。

相邻网格的判断:网格的相邻包括直接相邻和间接相邻两种情况。两个网格直接相邻指的是两个网格只有一个公共面,其判断条件是两个网格的编码只有一维值不同且只相差1,而其他维的编码均相同。如下表1中,网格 a 的编码为23,网格 c 的编码为13,网格 b 的编码为22,网格 d 的编码为24,网格 e 的编码为33。与网格 a 直接相邻的网格有四个。而间接相邻是指与同一个网格的相邻的所有网格均相邻。

1				
2		b		
3	c	a	e	
4		d		
	1	2	3	4

图1 数据分布及网格划分

1.3 数据存储结构

1) 区域表:用于存放聚类结果的现实描述

表结构如下:

类	区域号	字段号	左边界	右边界	方格个数	记录个数
---	-----	-----	-----	-----	------	------

说明:网格的合并是将某个聚类中相邻的网格合并为一个区域。通过每个区域的每维(字段)上的区间对聚类的结果进行现实描述。方格个数存放该类该区域中的网格个数。如:

3	4	3	0.35	0.7	23	148
---	---	---	------	-----	----	-----

上面这条记录的含义是:第3类中的第4个区域的第3个字段值在 $[0.35, 0.7)$ 之间,该区域中的网格个数是23个,该区域内的记录条数为148条。

2) 类_记录数表:用于存放聚类的个数和每类的记录总个数。表结构如下:

类编号	记录总数
-----	------

2 实验分析

为了验证基于网格距离的聚类算法的正确性和有效性,笔者实现了该算法,并在合成数据上进行了测试。实验环境如下:

硬件配置:毒龙 750MHZ/128M Personal Computer

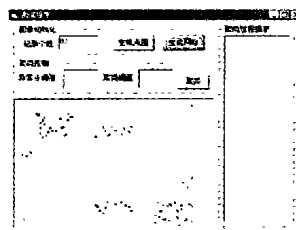
软件配置:Microsoft Windows 98

Visual Basic 6.0

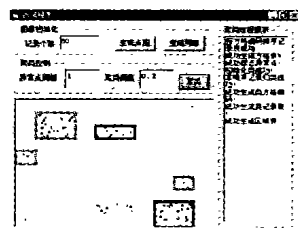
Microsoft Access 2000

2.1 实验结果

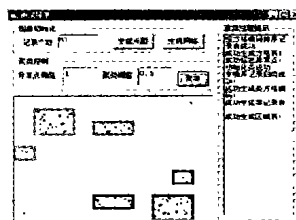
为了方便查看测试结果,测试数据采用在二维空间中的点数据,并且每一维的数据值在 $[0, 1]$ 区间内。将每一维按照0.1的宽度进行网格划分,共生成100个网格。聚类后的结果根据生成的区域表中的区域用深浅不同颜色的方框标出不同的类(同一类中的不同区域用同一颜色标注)。数据分布及网格划分效果如图1所示。调整不同的异常点阈值和聚类阈值,将得到不同的聚类结果。如图2所示。



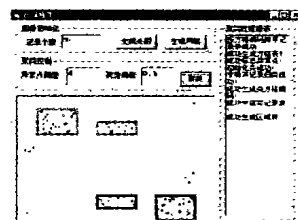
(a) 将实验数据进行网格划分



(b) 异常点阈值=1,聚类阈值=0.2



(c) 异常点阈值=1,聚类阈值=0.5



(d) 异常点阈值=4,聚类阈值=0.5

图2 聚类结果

当异常点阈值=1,聚类阈值=0.5时的聚类结果如图3所示。从图可得出,在该种情况下共聚为两大类,每类均包括3个区域。其中第1类的第1个区域为字段1在 $0.1 \sim 0.3$ 之间,同时字段2在 $0.7 \sim 0.9$ 之间。其他的区域的描述可以此类推。

类	区域号	字段号	左边界	右边界	方格个数	记录个数
1	1	1	0.1	0.3	4	24
1	1	2	0.7	0.9	4	24
1	2	1	0	1	1	3
1	2	2	0.5	0.6	1	3
1	3	1	0.4	0.6	2	11
1	3	2	0.7	0.8	2	11
2	1	1	0.7	0.9	4	25
2	1	2	0	0.2	4	25
2	2	1	0.4	0.6	2	11
2	2	2	0.7	0.8	2	11
2	3	1	0.8	0.9	1	3
2	3	2	0.3	0.4	1	3

图3 异常点阈值=1,聚类阈值=0.5时的区域表

2.2 实验分析

由于采用的是合成数据,预先知道测试用例中数据的分布,因此实验结果可以证明算法可以正确找出数据的聚类。

实验的输入参数:异常点阈值和聚类阈值的设置会影响数据挖掘的结果。因此在实际聚类挖掘中,这两个参数应根据用户对挖掘对象的专业知识、经验及反复实验来设置。

3 算法评价

3.1 时间复杂性分析

假设挖掘对象共 X 条记录, m 维, 每维划分为 n 个区间, X 条记录实际所占网格数为 k 个, 由该假设可得出: $k \leq n^m$ 。设所有 k 个网格中, 非异常点网格有 p 个, 则 $p \leq k$ 。设最后的聚类个数是 $ClassSum$ 。整个算法的时间复杂度近似为:

$$X \times m \times n + k + [p \times (p-1) \times m] / 2 + (p-2) \times (ClassSum + 2) / 2 + (p^2 - p) \times m / 2 \quad (1)$$

由此可看出, 当 p 值较大时, 整个聚类过程的时间复杂度主要取决于 p 的平方级。而由于 $p \leq k \leq n^m$, 故当每维的区间数 n 越少, 则网格数 k 越少, 则聚类效率越高; 当数据真正所占网格数 p 越少, 则聚类效率越高。聚类过程的时间复杂度虽然与每一维的划分的区间数 n 有间接的指数关系, 但由于区间数 n 并不是由客观数据决定, 因此在用此算法进行实际数据的挖掘应用时, 对于高维数据可减少区间数 n , 则可以达到有效地降低时间复杂度的目的。

3.2 算法评价

基于网格的聚类算法的优点主要有: 1) 可以适用于比较分散、并不密集的空间多维数据的挖掘。这正弥补了基于密度的聚类算法的缺陷。2) 不必事先假设类的个数, 类的个数由算法自动生成。这正弥补了 K-平均值和 K-代表点聚类算法的缺陷。3) 不必事先随机选取初始点, 由算法自动选取距离最远的非异点数据为初始点, 从而避免了因初始点选择不当而导致局部最优解的错误。4) 通过网格合并而生成的区域表可以方便地给出聚类结果的现实意义描述。5) 可以方便而有效地发现异常点。6) 运行效率高, 时间复杂度低。由于算法采用了网格技术, 所以算法的处理时间与数据对象的数目只呈一次线性关系, 而主要由网格的划分及数据的空间分布情况决定。当挖掘对象进行网格划分时每维上的划分区间数越少, 聚类效率越高。如果数据的空间分布越集中, 则实际所占的网格数越少, 则聚类效率越高。

本算法的主要不足在于聚类的结果还依赖于异常点阈值和聚类阈值的输入。而对于异常点阈值的依赖性继承于基于网格的算法, 而对于聚类阈值的依赖性主要是继承于基于距离的算法。

4 应用

本实验的研究目的就是用数据挖掘中的聚类方法对某高校已就业的毕业生数据进行挖掘, 试图发现已就业毕业生的共同点, 为高校如何培养容易就业的学生提供宝贵建议。

在该高校中的就业信息库中随机抽取了 5881 条记录, 经过前期整理后, 共有 8 维。采用本文提出的算法对其进行聚类挖掘后结果如表 1 所示。

表 1 聚类结果的比较

异常点阈值	聚类阈值	类的个数	类内距	异常点个数
5	0.6	8	82.725	1358
1	0.6	22	1493.874	607
2	0.8	12	838.1641	935
3	0.6	10	214.7241	1139

在表 1 中, 当异常点阈值取值为 5 时, 所得到的异常点个数已经达到 1358 个。这个实验结果说明该数据库在高维空间中分布是非常分散的。所以该数据库无法采用传统的基于网格的聚类算法, 这也正是提出基于网格距离的聚类算法的原因之一。

5 结语

本文设计并实现了一种基于网格距离的聚类算法。通过实验验证了基于网格距离的聚类算法的正确性和有效性。

在今后的研究中, 可以基于现有的工作从以下几个方面进行拓展:

1) 对异常点进一步研究。在本算法中只是标志出了异常点, 但对异常点究竟是噪声还是例外还有待于进一步研究。尤其是其中的例外挖掘将是未来长期的研究目标。

2) 网格合并算法的进一步完善。由于网格的合并很容易扩大数据真正所属的区间, 因此设计出尽量减少合并后生成区间的误差的网格合并算法将会增强基于网格距离的聚类算法的有效性。

3) 完善聚类结果内涵知识获取方面的工作。虽然本聚类算法最后生成了用于描述聚类结果现实意义的区域表, 但当面对区域表中众多类的众多区间时, 内涵知识的获取依然困难。因此聚类结果内涵知识的获取依然是未来长期的工作目标。

参考文献:

- [1] 韩家炜, KAMBER M. 数据挖掘概念与技术[M]. 北京: 机械工业出版社, 2001.
- [2] KAUFMAN L, ROUSSEEUW PJ. Finding Groups in Data: An Introduction to Cluster Analysis[Z]. New York: John Wiley & Sons, 1990.
- [3] ESTER M, KRIEGLER H-P, SANDER J, et al. A density-based algorithm For discovering clusters in large spatial databases[J]. In Proc1996 Int Conf Knowledge Discovery and Data Mining (KDD96), 1996, 8: 226 - 231.
- [4] ANKERST M, BREUNIG M, KRIEGLER H-P, et al. OPTICS: Ordering points to identify the clustering structure[A]. In Proc1999 ACM-SIGMOD Int Conf Management of Data (SIGMOD99) [C]. Philadelphia, PA, 1999. 49 - 60.
- [5] WANG W, YANG J, MUNTZ R. STING: A statistical inFormation grid approach to apatial data mining[A]. In Proc 1997 Int Conf Very Large Data Bases (VLDB97) [C]. Athens, Greece, 1997. 186 - 195.
- [6] SHEIKHOLESLAMI G, CHATTERJEE S, ZHANG A. WaveCluster: A multi-resolution clustering approach For very large spatial databases [A]. In Proc 1998 Int Conf Very Large Data Bases (VLDB98) [C]. New York, 1998. 428 - 439.
- [7] AGRAWAL R, GEHRKE J, GUNOPULOS D, et al. Automatic subspace clustering of high dimensional data For data mining applications[A]. In Proc 1998 ACM-SIGMOD Int Conf Management of Data (SIGMOD98) [C]. Seattle, WA, 1998. 94 - 105.
- [8] FISHER D. Improving inference through conceptual clustering[A]. In Proc 1987 AAAI Conf [C]. Seattle, WA, 1987. 461 - 465.
- [9] MACQUEEN J. Some methods For classification and analysis of multivariate observations [J]. Proc 5th Berkeley Symp Math Statist, Prob, 1967, 1: 281 - 297.