ISSN 1009-3044
Computer Knowledge and Technology 电脑知识与技术
Vol.13, No16, June 2017
E-mail: jslt@dnzs.net.cn
http://www.dnzs.net.cn
Tel:+86-551-65690963 65690964

基于划分的聚类算法研究与应用

何 宇

(成都信息工程大学,四川 成都 610225)

摘要:随着数学、计算机科学以及统计学、生物学等的快速发展,促进了聚类算法的产生。聚类分析在数据的处理和分析当中有着举足轻重的作用,并且被广泛应用到多个领域,介于此人们发明出了聚类算法。这些算法可以被分为以划分方法为代表的多种多样的处理方法。今天我们着重来探讨一下基于划分的聚类算法的研究与应用。

关键词:划分方法;聚类算法;研究与应用

中图分类号:TP391 文献标识码:A 文章编号:1009-3044(2017)16-0055-02

DOI:10.14004/j.cnki.ckt.2017.1947

随着我国的数学、计算机科学以及经济学学科的快速发展,聚类算法得到广泛使用,加快了数据处理与分析的速度,很大程度上促进了这些学科的发展。而且聚类算法的应用领域已经涉及生活和生产的方方面面,它是将物理或抽象对象的集合分组为由类似的对象组成的多个类的分析过程。这有很多具体应用的实例,比如说在商业方面,聚类分析方法可以帮助销售工作者找到不同的客户群,并且通过聚类分析中特定的模式来展现客户群的差异性。实际出真知,聚类分析方法对于市场的整体分析和数据处理等有着极其重要的作用,而且可以根据对客户群特点的分析准确把握客户的消费心理,这样一来能够促使厂商发现新的商机,开发新型的产业和地区市场,并且能够将这些信息整合起来;在保险行业,聚类分析的应用更是在很大程度上解放的人力,很多数据的收集、处理变得极其方便,主要是根据地区的保险业的平均水平来划分的,以平均值为划分的界限,再结合局部地区的经济发展速度、人均工资水平以及对保险业的态度和购买程度进行分组;再者便是在近些年来最为流行的贸易方式——电子商务。电子商务顾名思义用的就是计算机,其本身在数据处理上就占有一定的优势,利用聚类分析的方法使得电子商务中的交易数据和人群划分更加明显,交易人群特点的掌握有利于电商事业的发展,也为更近一步的商务交流提供了建设性的意见。

1 划分方法的基本概念及其常用的方法

划分方法(PAM: Partitioning method)的定义是首先创建k个划分,k为要创建的划分个数。常用的划分方法有:k-means, k-medics, CLARA(Clustering Large Application), CLARANS(Clustering Large Application based upon Randomized Search). FCM。[1]其中以k-means的使用最为普通,严格来说k—means属于非层次聚类法的一种,下面我们来看一下它的整个执行过程,一共分为两个部分,分别是初始化,循环。所谓初始化就是指选择或是人为指定某些记录作为凝聚点,但是要注意的一点就是按就近原则进行初始化的选择,而且要注意记录中心的数据,最后根据记录数据重新进行这一过程。一直不断地重复这一过程,直到凝聚点位置收敛为止。这种方法一般具有节省运算时间等特点。

2 具体的运算过程

2.1 数据预处理

数据预处理是指我们在对数据进行正式的处理之前,要先对数据的整体进行一下估量,主要从数据的数量、范围、程度和既定标准这几方面入手,进行规划分类和简单的预测分析,然后再就每一个方面对整体数据的影响进行估量式判断,建立起一个预测模式。[2]当然在我们有了明确的数据处理和分析结果时要将这一预测模式清楚,避免结果混淆。

2.2 定义距离函数

聚类的产生是由于多个领域和数据之间存在着相似性,正是由于事物之间相似性的存在,才促生了聚类的算法。但是这些相似性的存在也极容易造成事物之间的混淆。所以给这些数据设置一个定义函数是非常有必要的。函数的设置是为了避免误差,所以在设置相似距离时一定要把握好度量,保持数据点之间的平衡,从而保证整个运算过程的准确性。

2.3 聚类或分组

数据对象的分类要根据数据的特点、适应的环境或是发挥的作用等来进行分类,而且由于分类时采用的方法不同或是人为因素的干扰,总会产生不同的数据分组。划分方法一般从初始划分和最优化一个聚类标准开始。Crisp Clustering,它的每一个数据都属于单独的类;Fuzzy Clustering,它的每个数据可能在任何一个类中,Crisp Clustering和Fuzzy Clustering是划分方法的两个主要技术,划分方法聚类方法具有自身的法则优势,它可以找到在不同的分类组之间的相似性,甚至可以分析出在同一组内分类数据之间存在的差异性,我们常说数据的处理和分析要科学,要辩证的看到事物的两面性。[3]这种方法本身就是一种辩证的方法,所以用它来分析和处理数据最合适不过了。

3 聚类方法的主要应用研究

聚类算法在实际应用过程中涉及多个行业发展。从商业、生物、地理、保险行业、因特网行业以及电子商务行业等都所有涉猎。

收稿日期:2017-05-05

作者简介:何宇(1985—),男,四川仁寿人,硕士,主要研究方向为网络安全及计算机应用。

3.1 商业

在商业市场的发展过程中，往往对于市场未知风险的预测是企业可持续发展的一个重要问题。如果企业能够有效的预测未来的市场风险，探究潜在的消费者动向，那么往往能够取得显著的利益。因此在当前阶段，聚类算法能够为企业研究消费者行为、探究潜在市场发展、选择实验室市场等奠定坚实的理论数据基础。

3.2 生物

在生物行业发展过程中，由于现代化科学技术的发展，在进行生物学的研究过程中，基因数据库的容量大大提升，通过聚类算法能够有效的根据基因数据库的特点进行划分，使人们能够对种群的固有特征有显著的认识。

3.3 保险行业

当前世界保险行业发展速度较快，而不同的行业所需要的保险种类略有区别。聚类算法能够根据不同的行业发展类型，制定相应的保险措施，为保险更好地发挥作用效果奠定基础。

3.4 因特网及电子商务

当前阶段，电子文库的发展规模逐渐扩大，聚类算法电子文库的信息修复以及信息分类上发挥了显著的作用效果。在进行信息特征搜索的过程中，聚类算法能够根据相应的关键词检测整篇文章，大大降低了工作量。

其次当前物联网时代的到来，物联网对人们的影响愈加扩大。而通过聚类算法的数据分析和统计等，能够在最短的时间内根据消费者的消费记录以及浏览行为确定消费者特征，为电子商务的更好更快发展提供有效的保障。

4 总结

通过对聚类算法的研究和分析，我们清楚地了解了划分方法的原理以及其作用机制。加深了对划分方法的理解，也为聚类算法在更多领域的应用提供了完备的理论支持，与此同时也促进了聚类算法自身知识和体系的进一步完善和发展。最后，通过对目前阶段聚类算法的实际应用分析发现，其对于我们生活方式产生了极大的影响，其已经渗透到了人们生活的方方面面。

参考文献：

[1] 李荟娆. K-means聚类方法的改进及其应用[D]. 东北农业大学, 2014.
[2] 刘强, 王艳秋, 张健. 人工免疫聚类算法在交通时段自动划分上的应用[J]. 自动化博览, 2008(Z1).
[3] 陈建娇. 高维数据的K-harmonic Means聚类方法及其应用研究[D]. 上海大学, 2012.

（上接第52页）

$$f_r = \sum_1^m (R_i - R_{\min}) \tag{8}$$

在该公式中R_i表示在第i维处的标准化剩余资源，$R_{\min}$表示标准化剩余资源中的最小只，根据公式(6)计算出利用率约均衡，f_r值越是小，在各维中使用该节点的几率就越不均衡，如果值是不断增大的，那么在各维中该节点的使用率就会越低，则该节点可以作为防止节点。

3) 节点的能耗包括两部分内容，一是运行的能耗，而是基础能耗。因为电源能耗和利用的CPU率之间是存在一定的线性关系，所以，这里可以使用通过CPU的相关使用率计算节点处的能耗。计算节点能耗的相关公式如下：

$$f_p(U_{CPU}) = P_{idle} + \frac{U_{CPU}}{(P_{busy} - P_{idle}) \times U_{CPU}} \times P_{busy} \tag{9}$$

该公式中P_{busy}表示满载电能的相关消耗，P_{idle}是空载电能的有关消耗，，根据公式得出最低耗能的相关节点作为搜索解。

6) 根据不同的目标优化要求进行权值的设置，从而实现多目标的优化。综合的适应度函数表示为：

7) $$f(U_{CPU}, U_{mem}, U_{bw}) = K_1 f_{SLA} + K_2 f_r + K_3 f_p \tag{10}$$

2 虚拟机的迁移过程

2.1 利用负载预测获得物理热点

首先是通过窗口的思想，按照时间相关顺序利用预测指数的算法来确定相应的热点。根据历史数值对某一时刻的CPU使用率进行计算：

$$x_{t+1} = \alpha x_t + \alpha^2 x_{t-1} + \cdots + \alpha^{n+1} x_{t-n} + \alpha_t \tag{11}$$

该公式中α表示平滑指数相关的预测数值（其范围是整数且$\alpha \leqslant 1$），它是对下一个时间段CPU的使用率受窗口影响的预测；α_1是正太随机分布变量，主要是确保预测的值有一定的随机性。

2.2 选择迁移虚拟机

将CPU使用率和虚拟机相关的内存大小策略相结合，同时提高迁移质量，相关函数是：

$$Q = \frac{U_{CPU}}{R_{ram}} \tag{12}$$

该式中R_{ram}是虚拟机的相关内存。如果CPU的使用率高而且内存是小的时候，Q的值就是最大，虚拟机的迁移可以很快的消除热点，如果CPU有关场景数据和内存数据非常少，这时可以减少迁移的时间。

3 结论

本文主要是对云平台有关资源调度的SLA和能耗，以及资源利用与迁移次数等问题进行分析，通过退火思想对虚拟机进行了初始化放置，分析了虚拟机迁移相关的动态变化过程，使能耗和资源利用，以及迁移次数等问题实现了均衡的目的。

参考文献：

[1] 娄建峰,高岳林,李飞,等.基于改进粒子群算法的云计算任务调度算法[J].微电子学与计算机,2016(8):112-116.
[2] 仲伟彪.改进粒子群算法的研究及其云计算资源调度的应用[D].江西理工大学,2015.
[3] 王德文,刘晓萌.基于改进粒子群算法的云计算平台资源调度[J].计算机应用研究,2015(11):3230-3234+3246.
[4] 李超.基于改进粒子群算法的云计算资源调度研究[D].中国矿业大学,2015.
[5] 丁阳,颜惠琴.基于改进粒子群算法的云计算任务调度策略[J].无锡职业技术学院学报,2012(3):66-68+71.