网格聚类算法

赵 慧，刘希玉，崔海青
（山东师范大学 管理与经济学院，山东 济南 250014）

摘 要：聚类分析有广泛的应用，是数据挖掘中非常重要的方法。聚类分析算法有多种分类，每种方法在不同领域发挥了不同的作用。以研究网格聚类算法为目的，介绍了聚类分析算法的要求以及常见的聚类算法；针对基于网格方法的聚类算法进行专门研究，比较分析了传统的和改进的基于网格方法的聚类算法。介绍的各种网格聚类算法都有自身的优点和不足。通过对这些网格聚类算法的学习便于深入研究网格聚类算法，以便将其与实际问题相结合，设计更好的算法。

关键词：聚类分析；聚类算法；网格；基于网格的聚类算法
中图分类号：TP311 **文献标识码**：A **文章编号**：1673－629X(2010)09－0083－03

Grid－Based Clustering Algorithm

ZHAO Hui, LIU Xi-yu, CUI Hai-qing
(College of Management and Economic, Shandong Normal University, Jinan 250014, China)

Abstract: Cluster analysis is a very important method in data mining and is widely used. Clustering algorithms have a variety of categories, and each method plays a different role in different areas. For the purpose of research gird－based clustering algorithm, describe the requirements of the clustering algorithm and a variety of clustering algorithms; Make special research for gird－based clustering algorithm and comparatively analysis's of gird－based clustering algorithm . Each gird－based clustering algorithm has its own advantage and disadvantage. Based on learning these gird－based clustering algorithms facilitate further research on clustering algorithm, so as to use gird－based clustering algorithm in practical problems and design better algorithm.

Key words: cluster analysis; clustering algorithms; grid; gird－based clustering algorithm

0 引 言

聚类分析有广泛的应用，在数据挖掘中非常重要。聚类是指将物理对象或抽象对象的集合分组成为由类似对象组成的多个类的过程[1]。在聚类的分组结果中，同一簇内的数据对象之间应具有较高的相似度，而不同簇内的对象之间相似度低。其中，两个对象之间的相似度通常由它们之间的距离来度量。聚类分析广泛应用于模式识别、图像处理、信息检索、生物学、医学、考古学、地质学、地理学、市场学等多个学科[2]。

1 聚类分析算法

1.1 聚类分析算法的要求

聚类分析是一个富有挑战性的研究领域，它的潜在应用提出了各种特殊的要求[1]：①可伸缩性。②处理不同类型属性的能力。③发现任意形状的聚类。④用于决定输入参数的领域知识最小化。⑤对于输入记录的顺序不敏感。⑥处理噪声数据的能力。⑦高维度数据处理能力。⑧基于约束的聚类。在实际应用中常需要在各种约束条件下进行聚类。⑨可解释性和可用性。上述的要求使人们围绕着提高聚类算法对大型数据库的可伸缩性、能够识别复杂形状的簇、处理高维数据等目标进行聚类算法的研究改进[3]。

1.2 常见聚类算法

大体上，主要的聚类算法可以划分为如下几类：划分方法（partitioning method）、层次方法（hierarchical method）、基于密度的方法（density－based method）、基于网格的方法（grid－based method）以及基于模型的方法（model－based method）。上述方法各有特点，在不同的领域以及数据特点下发挥了不同的作用，实现了数据的有效聚类。

2 基于网格方法的聚类算法

基于网格的聚类方法采用一个多分辨率的网格数

收稿日期：2010－01－25；**修回日期**：2010－04－16
基金项目：国家自然科学基金（60873058）；山东省自然科学基金重点项目（Z2007G03）；泰山学者建设工程资助项目
作者简介：赵 慧（1986－），女，硕士研究生，研究方向为聚类分析；刘希玉，“泰山学者”，教授，博士生导师，研究方向为数据挖掘与人工智能。

据结构。它将空间量化为有限数目的单元,这些单元形成网格结构,所有的聚类操作都在网格上进行,网格中的数据压缩质量就决定了算法的聚类质量。

2.1 传统的网格聚类算法

(1)STING 算法。

STING (Statistical Information Grid)[4]将空间区域划分为矩形单元,它是一种基于网格的多分辨率聚类技术。STING 算法网格的计算独立于查询;网格结构利于并行处理和增量更新;效率很高:时间复杂度是 $O(n)$,其中 n 是对象的数目。网格结构的最低层的粒度决定了 STING 算法聚类的质量。该算法处理速度较快,但簇的质量和精确性有可能会降低。

(2) WaveCluster 算法。

WaveCluster[5] 首先通过在数据空间上强加一个多维网格结构来汇总数据,再采用一种小波变换来变换原特征空间,然后在变换后的空间中找到密集区域,它是一种多分辨率的聚类算法。基于小波变换的计算复杂度是 $O(n)$,n 是数据库中对象的数目。这个算法聚类速度很快,它的实现可以并行化。

(3)CLIQUE 算法。

CLIQUE(Clustering In Quest)[6]对于大型数据库中的高维数据的聚类非常有效,它是综合了基于密度和基于网格的聚类算法。但是,由于 CLIQUE 方法简化,聚类结果的精确性可能会降低。

2.2 改进的网格聚类算法

国内外改进的网格聚类算法很多,比如国外的改进基于网格的聚类算法有:SCI[7]算法、MAFIA[8]算法、ENCLUS[9]聚类算法、DCLUST[10]聚类算法等等。

国内改进的网格聚类算法有:

(1)IGDCA 算法。

IGDCA[11]聚类算法是在基于密度的网格聚类算法 GDCA 的基础上提出来的。GDCA 适用于发现大规模空间数据库中任意形状的聚类,该算法首先将数据空间划分成若干个体积相同的单元,然后对单元进行聚类。只有密度不小于给定阈值的单元才得到扩展,从而大大降低了时间复杂性。在 GDCA 的基础上,提出了增量式聚类算法 IGDCA,该算法能处理任意形状的簇,适用于数据的批量更新。

(2)基于网格的层次聚类。

基于网格的层次聚类算法[12]是针对传统的凝聚层次聚类算法因时间复杂度太高无法应用到大型数据集的问题而改进的。基于网格的层次聚类算法,先用基于网格的方法进行一次微聚类,然后再用凝聚的层次聚类算法进行聚类。而且在进行凝聚的层次聚类时,采用一种新的簇间聚类度量方法,方法中采用簇中权值最高的代表点的最小距离作为簇间的距离。该算法的空间复杂度只和所创建的网格单元个数相关,算法的时间复杂度也只和所创建的网格单元个数相关,而和数据集中元素个数无关。

(3)基于测度的网格聚类算法。

基于测度的网格聚类方法[13]的基本思想是:在数据空间上定义计数测度,而且划分数据空间质量的评价标准是计数测度即空间内数据的个数的方差。数据空间划分时总希望可以得到每个集合有较大的差异,这样便于对集合分类。当集合的计数测度的方差作为衡量标准时,方差越大划分越利于聚类。所以当测度的方差最大时,对应的划分是最优划分,对应的分辨率是最优分辨率。将数据空间按照最优分辨率划分,再根据连通性将数据聚类。该算法中没有对参数值进行人为设定,可以实现提高准确性的目的。

(4)自动化网格聚类探究。

自动化网格聚类算法(GCA)[14]主要采用密度阈值技术提取不同的类,使用边界点处理技术提高聚类精度,只要求对数据集进行一遍扫描。实验表明,GCA 算法可扩展性好,能处理任意形状和大小的聚类,能够很好地识别出孤立点和噪声,在处理多密度聚类方面有很好的精度。

(5)新型的基于密度 - 网格的自适应免疫聚类算法。

新型自适应免疫聚类算法(AICDG)[15]保留了基于密度网格聚类算法快速搜索包含聚类子空间的优点,同时根据免疫原理对抗体按密度进行自适应抑制和克隆,从而利用免疫网络改进聚类生成过程。AICDG 能够发现任意形状的聚类,有效地处理高维数据和一些特殊的聚类,有良好的数据规模可扩展性。该算法还具有用户定义参数少、保存数据原型密度信息的特征。

(6)数据流的网格密度聚类。

基于网格密度的实时数据流聚类算法(RTCS)[16]首先在在线层不断地读取每一个新来的数据项,将其放入相应的空间单元格,同时更新该单元格的特征向量。其次,算法定义一个时间间隔 gap。在离线层每经过 gap 时间后做一次聚类。算法在第一个 gap 时间后,调用算法计算各个单元格的特征向量,找出稠密单元格、过渡单元格,并将它们聚集为单元格簇或类,给出相应类号,作为初始聚类。接着,算法在每经过 gap 时间后都要进行聚类调整。调整时首先检查由于新数据的到达而引起的单元格密度变化,再根据孤立点检查策略去删除那些仅有孤立点的单元格,以减少算法的工作量。然后调用算法根据密度的变化对所有在此

gap周期内调整过特征向量的单元格作相关的类别调整,将已变为稀疏性的单元格从类中删除,同时也将转变为稠密或过渡性的单元格合并到最佳的邻居类中。RTCS算法能够根据密度的动态变化区分出真正的孤立点并剔除它,而这种剔除对后面的聚类结果没有影响。该算法有较快的处理速度,对数据维数和规模有更好的可扩展性。

(7)基于网格的增量聚类算法。

一种基于网格的增量聚类算法(IGrid)[17]具有传统网格聚类算法的高效性,且通过维度半径对网格空间进行了动态增量划分以提高聚类的质量。IGrid算法分为两步,首先对数据集中的数据点进行访问,并通过维度半径动态增量式地逐步形成网格结构,然后再利用网格聚类方法进行聚类 。IGrid算法在聚类准确度以及效率上要高于传统的网格聚类算法。

(8)基于网格距离的高精度聚类算法。

一种新的基于网格距离的高精度聚类算法[18]是为了提高基于网格聚类技术的聚类精度和效率而提出的。该算法一方面通过参考网格在逻辑空间的相对距离进行聚类,从而弥补了大多数计算网格之间距离的算法中需要大量数学运算的不足,另一方面,采用一种新的边界点处理技术,该技术能够有效地提取有意义的边界点,运行速度快、聚类精度高。

(9)基于网格密度和距离信息特征的聚类算法。

基于网格密度和距离信息特征的聚类算法(GDD)[19]将数据空间划分成网格单元,并构建基于簇中心距离信息的跃迁函数,通过考察局域范围内网格单元的密度跃迁比,并比对计算出的当前网格单元的跃迁函数值,以决定是否继续扩展和增长聚类簇规模。GDD算法能够发现任意形状的簇,对噪音数据不敏感,且具有线性于网格数目的时间复杂性,适合对大规模真实数据集的聚类。

(10)基于网格方法的高维数据流子空间聚类算法。

基于网格方法的高维数据流子空间聚类算法(GSCDS)[20]通过利用由底向上网格方法对数据的压缩能力和自顶向下网格方法处理高维数据的能力,算法能基于对数据流的一次扫描,快速识别数据中位于不同子空间内的簇。理论分析以及在多个数据集上的实验表明算法具有较高的计算精度与计算效率。

(11)基于相似度的网格聚类。

基于相似度的网格聚类算法(SGCA)[21]主要利用网格技术去除数据集中的部分孤立点或噪声,使用边界点阈值函数提取类的边界点,最后利用相似度方法进行聚类。SGCA算法只要求对数据集进行一遍扫描。该算法可扩展性好,能处理任意形状和大小的聚类,能够很好地识别出孤立点或噪声,适用于综合数据集以及高维数据集。

(12)基于网格和密度的微粒群混合聚类方法。

基于网格和密度的微粒群混合聚类方法(CGDP)[22]在分析现有的基于网格和密度的聚类方法的基础上,借鉴密度函数的思想提出了一种新的网格单元密度的计算方法,在此基础上,将计算方法与微粒群算法相结合而设计的算法。CGDP有效地解决了网格空间下数据对象对周围空间影响信息的丢失问题,增强了区分噪音数据的能力,增强了用户对聚类过程的可控性。

3　结束语

网格聚类算法具有网格数量独立于数据对象的数量,对数据的输入顺序不敏感,处理速度独立于数据集的大小,处理速度快,可伸缩性强等特点。网格聚类算法有其自身的优缺点,如何将网格聚类算法进行研究改进,如何将网格聚类算法与实际问题相结合,从而有效地应用于实践,也是现在国内外的研究热点。

参考文献:

[1] 韩家炜.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,2001.

[2] Jain A K, Murty M N, Flynn P J. Data Clustering: A Review[J]. ACM Computing Surveys, 1999, 31(3): 264-323.

[3] 曾蒙福. 基于自适应网格的聚类算法及在信息提取中的应用研究[D]. 福州: 福州大学, 2005.

[4] Wang W, Yang J, Muntz R. STING: A Statistical Information Grid Approach to Spatial Data Mining[C]// In: Proceedings of the 23rd VLDB Conference. Athens, Greece: [s.n.], 1997:186-195.

[5] Sheikholeslami G, Chatterjee S, Zhang A. WaveCluster: A Multi-Resolution Clustering Approach for Very Large Spatial Databases[C]// In: Proceedings of the 24th VLDB Conference. New York, USA: [s.n.], 1998:428-439.

[6] Agrawal R, Gehrke J, Gunopulos D, et al. Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications[C]// In: Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data. [s.l.]: [s.n.], 1998:94-105.

[7] Hsu Chih-Ming, Chen Ming-Syan. Subspace clustering of high dimensional spatial data with noises[C]// PAKDD 2004. [s.l.]: [s.n.], 2004:31-40.

[8] Agrawal R, Gehrke J, GunoPulos D, et al. Automatic Sub-

(下转第89页)

3 个隐层神经元和一个输出层神经元情况下 BP 算法能够达到目标误差要求。网络拓扑结构如图 6 所示。训练后对每个样本的识别误差如下:

0.1151 0.0306 0.0099 0.0696 0.1124 0.0280 0.0162 0.0722 0.0362 0.0002 0.0286 0.0124 0.0975 0.0176 0.0069 0.1366

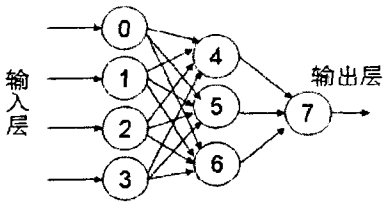

图 6　标准 BP 网络解决逻辑运算问题的网络结构

可见,自进化系统能够找到更优的解决逻辑运算问题的网络结构。

4　结束语

通过自进化系统寻找解决逻辑运算问题的神经元网络结构,并与标准 BP 网络解决逻辑运算问题的网络结构进行比较。实验结果表明,自进化系统能够找到更优的解决问题所需的网络结构。

参考文献:

[1] 卢格尔. 人工智能复杂问题求解的结构和策略[M]. 北京:机械工业出版社,2006.

[2] Tsai Huai-Kuang, Yang Jinn-Moon, Tsai Yuan-Fang, et al. An Evolutionary Algorithm for Large Traveling Salesman Problems[J]. IEEE Trans on Systems, Man and Cybernetics-Part B: Cybernetics, 2004, 34(4): 1718-1729.

[3] 阎平凡,张长水. 人工神经网络与模拟进化计算[M]. 北京:清华大学出版社,2005.

[4] Yao X. Evolutionary Artificial Neural Networks[J]. International Journal of Neural Systems, 1993, 4(3): 203-222.

[5] Yao X. A New Evolving System for Evolving Artificial Neural Networks[J]. IEEE Trans NN, 1997, 8(2): 694-713.

[6] Yao X. Evolving Artificial Neural Networks[J]. Proc IEEE, 1999, 87(5): 1423-1447.

[7] 张秉森,王　莹,李　莉. 遗传算法改进 BP 网络对织物配色的优化研究[J]. 计算机工程与设计,2008,29(19):5033-5036.

[8] 王宏刚,钱　锋. 基于遗传算法的前向神经网络结构优化[J]. 控制工程,2007,14(4):387-390.

[9] 李智勇,童调生. 基于多物种进化遗传算法的神经网络结构学习方法[J]. 计算机工程与设计,2003,39(22):87-90.

[10] 高　坚,贺秉庚. 网络结构拓扑扩展的混合遗传算法[J]. 计算机工程与科学,2002,24(3):3-4.

[11] 黄　浩,宋瀚涛,陆玉昌. 基于小生境遗传算法的贝叶斯网络结构学习算法研究[J]. 计算机应用研究,2007,24(4):100-103.

[12] 方建安. 采用神经元网络学习的网络控制器[J]. 控制与决策,1993,3(3):208-212.

[13] 拉马克. 动物哲学[M]. 北京:商务印书馆,1936.

[14] 张良均,曹　晶,蒋世忠. 神经网络实用教程[M]. 北京:机械工业出版社,2008.

(上接第 85 页)

space Clustering of High Dimensional Data for Data Mining Applications[C]// Proc. of the ACM SIGMOD Int'l Conference on Management of Data. Seattle, Washington: [s. n.], 1998: 94-105.

[9] Cheng C-H, Fu A W, Zhang Y. Entropy-based Subspace Clustering for Mining Numerical Data[C]// Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. [s. l.]: ACM Press, 1999: 84-93.

[10] Zhang Ji, Hsu W, Lee Mong Li. Clustering in Dynamic Spatial Databases[J]. Journal of Intelligent Information Systems, 2005, 24(1): 5-27.

[11] 陈　宁,陈　安,周龙骤. 基于密度的增量式网格聚类算法[J]. 软件学报,2002(1):1-5.

[12] 姚玉钦,李金广. 基于网格的层次聚类算法[J]. 河南师范大学学报:自然科学版,2009(4):42-44.

[13] 白　鹭,马　骥. 基于测度的网格聚类算法[J]. 沈阳大学学报,2009(4):61-63.

[14] 刘敏娟,李勇军. 自动化网格聚类探究[J]. 软件导刊,2009(8):144-145.

[15] 黄柳萍,冯朝一,周　明. 一种新型的基于密度-网格的自适应免疫聚类算法[J]. 福建电脑,2009(8):83-84.

[16] 屠　莉,陈　峻,邹凌君. 数据流的网格密度聚类算法[J]. 小型微型计算机系统,2009(7):1376-1382.

[17] 印桂生,于　翔,宁　慧. 一种基于网格的增量聚类算法[J]. 计算机应用研究,2009(6):2038-2040.

[18] 孟建良,程伟想,牛为华. 基于网格距离的高精度聚类算法[J]. 计算机应用与软件,2009(6):262-264.

[19] 戴维迪,张　璐,王文俊,等. 基于网格密度和距离信息特征的聚类算法[J]. 华南理工大学学报:自然科学版,2009(4):18-23.

[20] 孙玉芬,卢炎生. 一种基于网格方法的高维数据流子空间聚类算法[J]. 计算机科学,2007(4):199-203.

[21] 刘敏娟,柴玉梅,张西芝. 基于相似度的网格聚类[J]. 计算机工程与应用,2007(7):198-201.

[22] 单世民. 一种基于网格和密度的微粒群混合聚类算法[J]. 计算机科学,2006,33(11):164-165.