网格和密度的聚类算法在CRM中的应用

朵春红，王翠茹

(华北电力大学计算机学院 河北 保定 071003)

【摘要】聚类分析是数据挖掘领域中一种非常有用的技术，它用于从大量数据中寻找隐含的数据分布模式，主要有分割法、层次法、密度法、网格法和模型法等。该文主要讨论数据挖掘中一种基于密度和网格的聚类分析算法及其在客户关系管理中的应用。该算法具有较高的聚类效率而且容易实现，可以发现任意形状的聚类，时间复杂度低，聚类精度高，适用于数据的批量更新。该文还提出增量式聚类技术，它不仅能够利用前期聚类的结果，充分提高聚类分析的效率，而且可以降低维护知识库所带来的巨大开销。实验证明了算法的有效性。

关 键 词 聚类分析; 客户关系管理; 数据挖掘; 密度; 网格
中图分类号 TP311.13 文献标识码 A

Application of a Clustering Algorithm Based on Density and Grid in CRM

DUO Chun-hong，WANG Cui-ru

(Institute of Computer, North China Electric Power University Baoding Hebei 071003)

Abstract Clustering analysis is a very useful tool in the domain of data mining for searching distributing mode from a great deal of data. Its main algorithms are partition-based algorithm, hierarchy-based algorithm, density-based algorithm, grid-based algorithm, and model-based algorithm. The paper mainly discusses a clustering algorithm based on density and grid in data mining, which has high clustering efficiency and low time complexity. It is efficient and effective for multi-density and uniformity density data sets with noise and suitable for batch update. After that an incremental clustering technique is presented. This technique not only makes best use of the former clustering results and improves the efficiency of clustering analysis, but also brings to the reduction of enormous expenditure on knowledge base maintenance. At last an application of the algorithm in Customer Relationship Management (CRM) is gien.

Key words clustering analysis; customer relationship management; data mining; density; grid

聚类分析是数据挖掘领域研究的重要课题[1]，其基本思想是：按照数据的相似性和差异性，将数据划分为若干组，同组的数据尽量相似，不同组的数据尽量相异[2-3]。迄今为止，人们已经提出了许多聚类算法，主要有分割法、层次法、密度法、网格法和模型法等[4-6]。基于网格和密度的聚类算法由于易于增量实现和高维数据挖掘而被广泛地应用于聚类算法中。基于网格的方法在聚类过程中将网格中的点作为一个整体处理，而不是考虑单元中的每一个点，基于这一特性，该方法在所有的聚类方法中效率最高。其优点是聚类的结果与输入数据的顺序无关，算法的时间复杂度是数据点个数的线性函数，速度快、可扩展性好，能识别不同形状的聚类。

本文给出一种基于密度和网格的聚类算法，它是一个基于密度的算法，既保留了基于网格算法运行速度快的特点，又通过细化技术弥补了该类算法精度不好的弱点。

1 算法分析

在基于密度的算法中，一个聚类就是一个比周围区域有更高数据点密度的区域[7]。为识别数据点的密度，将数据空间进行划分并找出每个单元中数据点的数目。为使计算点的密度的方法简单一些，将数据空间分割成网格状，把数据空间中的每一维划分成相同的区间数，每一个单元具有相同的“体积”[8-9]。单元中点的密度的计算可以转换成简单的点计数，然后把落到某个单元中的点的个数作为该单元的密度。这时可以指定一个阈值r，当某单元格中点的个数大于该阈值时，就称该单元格是密集的，聚类也就是所有相邻近的密集单元格的集合。

收稿日期：2007－09－14

作者简介：朵春红(1982－)，女，硕士生，主要从事数据库技术与信息处理、决策支持系统方面的研究；王翠茹(1954－)，女，教授，主要从事数据库技术、决策支持系统、计算机在电力系统的应用等方面的研究.

1.1 相关概念

设$A=\{D_1,D_2,\cdots,D_n\}$是n个有界定义域，那么$S=D_1\times D_2\times\cdots\times D_n$就是一个$n$维空间。将$D_1,D_2,\cdots,D_n$看成是$S$的维(属性、字段)，算法的输入是一个$n$维空间中的点集，设为$D=\{p_1,p_2,\cdots,p_n\}$，其中$p_i=\{p_{i1},p_{i2},\cdots,p_{in}\}$。$p_i$的第$j$个分量$p_{ij}\in D_j$。通过一个输入参数$m$，可以将空间$S$的每一维分成相同的$m$个区间，从而将整个空间分成有限个不相交的类矩形单元。每一个这样的矩形单元可以描述为$\{u_1,u_2,\cdots,u_n\}$，其中$u_i=[l_i, h_i)$是一个左闭右开区间。当单元$v=\{v_1,v_2,\cdots,v_n\}$落入一个区间$u=\{u_1,u_2,\cdots,u_n\}$时，当且仅当对于每一个v_i都有$l_i<=v_i<h_i$成立。

定义 1 单元格u的选择率selectivity(u)=单元格中点的个数/总的点数。对于用户的输入参数r，当且仅当selectivity(u)>r时，称数据单元u是密集的。

定义 2 单元中心点：一个单元U的中心点是一个n维向量$(u_{c1},u_{c2},\cdots,u_{cn})$，其中$u_{ci}=(l_i+h_i)/2$，$l_i$和$h_i$分别为该区间的最小值和最大值。

定义 3 单元重心点：假设单元U包含k个数据点$p_1,p_2,\cdots,p_k$，则单元U的重心点是一个n维向量$(p_{u1},p_{u2},\cdots,p_{un})$，其中$p_{ui}=(p_{1i}+p_{2i}+\cdots+p_{ki})/k$。

定义 4 直接关联单元：如果单元$u_1=\{r_1, r_2, \cdots, r_n\}$和单元$u_2=\{r_1',r_2',\cdots,r_n'\}$存在$n-1$维，假设为$D_1,D_2,\cdots,D_{n-1}$，使得$r_j=r_j'$，$j$=1,2,$\cdots$,$n$–1，并且$h_n=l_n'$或$h_n'=l_n$，那么$u_1$与$u_2$有一个共同的面。如果$u_1$和$u_2$有一个共同的面或至少在一个维上有交集，那么$u_1$和$u_2$是直接相关联的。

定义 5 关联单元：如果u_1和u_2直接关联，或存在另一个单元u_3，u_3分别与C_1和C_2直接相关联，那么u_1和u_2是相关联的。

定义 6 细化技术：单元的中心与单元的重心不在同一个位置，其重心偏向于邻近的密集单元，可以以该单元的重心作为新的中心点重新画一个单元，使得其中的数据点分布尽可能均匀。从本质上来讲，新的单元相当于原来单元向密集单元移动，最终的目标是使得包含属于同一个cluster的数据点的单元靠得更近一些[10]。

1.2 基于密度和网格的聚类算法

首先把整个数据空间划分成多个单元，数据集合中的每个点都应该映射到其中的一个数据单元中。单元是算法的基本处理单位，不必单独处理单元中的每个数据点，因此可以加速聚类的过程。

算法1：基于密度和网格的聚类算法。该算法步骤如图1所示，具体为：(1) 把d维空间的每一维均匀划分成m个区间，即把数据集合S对应的数据空间划分成md个互不相交的单元。(2) 通过细化技术来发现新的密集单元，其基本思想是把非密集单元向密集单元移动，从而获得更好的聚类效果。首先创建新的数据单元，然后计算映射到该数据单元的数据点，在这个过程中不需要扫描整个数据集合，只需要查看邻近的数据单元。这样，整个数据空间就由最初的密集单元和新增的单元构成。(3) 生成cluster，并对cluster中的每个数据点标注一个cluster编号clusterID。在最大相关联的密集单元中的数据点构成一个c1uster，使用深度优先的算法来查找相关联的密集单元。

算法1可以处理各种形状的cluster，受噪音数据的影响小，与数据的输入顺序无关并且可以并行化执行。

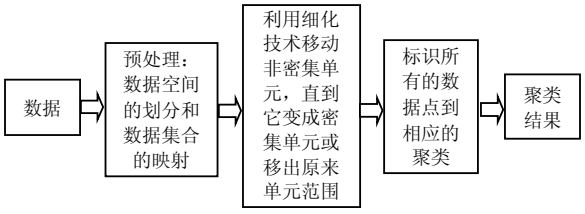

图1 一般步骤

1.3 基于密度和网格的增量式聚类算法

当对一个新增数据进行聚类时，并不需要知道它和全部数据的关系，只需要知道其相邻近的那些数据的聚类情况。因为对新增数据的聚类是通过确定和它邻近的数据所属的聚类以及和它邻近的数据的个数来进行的，与新增数据邻近的数据往往都在包含这个新增数据在内的一个局部区域内。

图2所示是和新增数据聚类有关的数据图，图中，“•”表示已聚成类的数据，“*”表示新增的数据。要对新增数据进行聚类，需要考虑和新增数据邻近的数据。

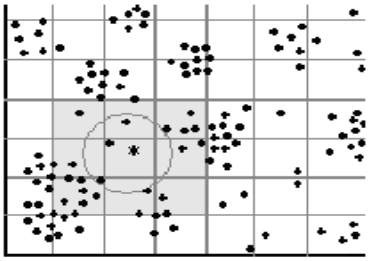

图2 和新增数据聚类有关的数据

对新增数据聚类时，首先确定新增数据所在的网格，找出所有包围这个网格的网格，因为只有这些网格中的数据才可能和新增数据邻近。然后逐个考察新增数据所在的网格以及包围这个网格的全部

网格中的已聚类数据，对新增数据进行聚类分析。

算法2：基于密度和网格的增量式聚类算法。该算法步骤如下：

1) 将聚类空间划分为网格，对每个网格建立它所包含的已聚类数据的索引。2) 对每一个新增数据，确定新增数据所在的网格以及包围这个网格的全部网格，也就是邻近网格。3) 将新增数据与全部邻近的网格区域中的数据记为邻近的数据。4) 根据和新增数据邻近的数据的个数，进行如下操作：(1) 如果没有和新增数据贴近的数据，则新增数据形成新的聚类。(2) 如果仅有一个数据和新增数据贴近，则新增数据归入和这个数据同一聚类的类。(3) 如果存在多个和新增数据贴近的数据，则合并这些贴近的数据所在的聚类为一个新的聚类，并将新增数据归入。5) 建立新增数据所属网格的索引。6) 如果所有新增数据都已聚类，算法结束，否则转步骤2。

基于密度和网格的增量式聚类算法与聚类的数据规模无关，仅与其邻近的网格数据量有关，因此它产生聚类的时间复杂度为新增数据的线性函数。

1.4 算法分析

假设N是数据集合中数据点的个数，则算法映射阶段的时间复杂度为$O(N)$，假设数据集合中的数据点是d维数据，每一维上划分成m个区间，则共有$K=md$个单元。移动一个非密集单元需要扫描2^d个相邻近单元。每个单元平均包含N/K个数据点，因此，移动一个非密集单元的时间复杂度为$O(2^dN/K)$。假设非密集单元的百分比为c，则需要移动$k=cK$个非密集单元，总的时间复杂度为$O(cK*2^dN/K)= O(c*2^dN)$。如果密集单元的总个数为M，则总的存取数据结构的个数为2^dM。所以总时间复杂度为$O(N)+O(C*2^dN)+O(2^dM)$。

对于增量式聚类算法，增量数据越大、分布越广、受影响的单元越多即需要更新的聚类单元越多，聚类所需时间越多。

图3为DIGCA算法与DBSCAN算法的聚类性能比较，由对比可以看出DIGCA算法的时间性能优于DBSCAN算法。

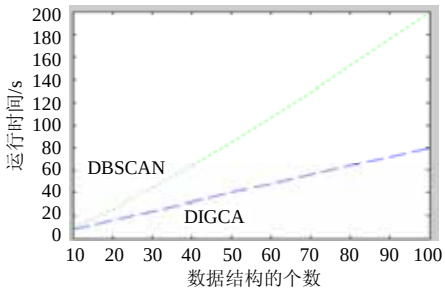

图3　DIGCA与DBSCAN的时间复杂度分析

2　DIGCA算法在CRM中的应用

客户关系管理是通过向企业的销售、市场和服务等部门和人员提供全面、个性化的客户资料，利用数据挖掘技术，发掘客户数据中蕴涵的知识，强化跟踪服务和信息分析能力，使企业能够建立和维护一系列与客户及生意伙伴间卓有成效的“一对一关系”。随着各行业业务操作流程的自动化，企业内会产生大量的业务数据，分析这些数据是为商业决策提供真正有价值的信息，进而获得利润[11]。

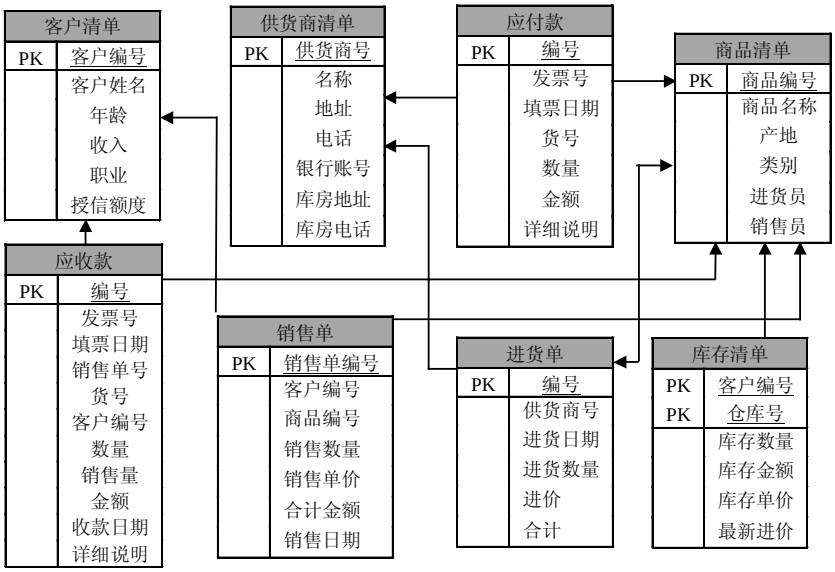

图4　模拟数据库结构

(下转第1314页)

5 结 论

本文提出了一种基于互信息计算的图像检索方法，对图像熵的定义进行推广，引进了图像分块颜色熵、形状熵和纹理熵的新概念。在进行图像检索时，考虑了图像的形状、纹理和颜色三个最主要的低层特征。试验表明，本文的方法适用于很多类型的图像检索，能够取得很好的效果。所以本文的方法有很好的应用前景。但本文的方法在进行纹理互信息计算时，时间和空间的开销较大，这是下一步需要解决的问题。

本文研究工作得到华东交通大学校立科研基金(06ZKJC03)的资助，在此表示感谢！

参 考 文 献

[1] SWAIN M, BALLARD D. Color indexing[J]. International Journal of Computer Vision, 1991, 7(1):11-32.
[2] PASS G, ZABIH R, MILLER J. Comparing images using color coherence vectors[C]//In: Proceedings of ACM Intern Conf Multimedia. Boston: [s.n.], 1996.
[3] STRICKER M, ORENGO M. Similarity of color images[C]//In: Proceedings of SPIE Storage and Retrieval for Image and Video Databases. San Jose: [s.n.], 1995.
[4] 范自柱, 蒋先刚. 基于包围盒的图像检索技术[J]. 计算机研究与发展, 2005, 42(增刊): 260-263.
[5] KLASSEN E, SRIVASTAVA A, WASHINGTON M, et al. Analysis of planar shapes using geodesic paths on shape spaces[J]. Transactions on Pattern Analysis And Machine Intelligence, 2004, 26(3): 372-383.
[6] FENG Jing, LI Ming-jing, ZHANG Hong-Jiang, et al. An efficient and effective region-based image retrieval framework[J]. Transactions on Image Processing, 2004, 13(5): 699-708.
[7] 时永刚, 邹谋炎. 图像配准中统计型相似性测度的比较与分析[J]. 计算机学报. 2004, 27(9): 1278-1283.
[8] 夏良正. 数字图像处理[M]. 修订版. 南京: 东南大学出版社, 1999.
[9] FARAGAND A, DELP E J. Edge linking by sequential search[J]. Pattern Recognition, 1995, 28(5): 611-633.

编 辑 熊思亮

(上接第1291页)

图4是超市管理系统的模拟数据库，由八个表组成。其中商品清单、供货商清单和客户清单是关键表，用于存放基本的数据信息、记录客户购买商品的情况以及供应商供货的情况。

根据表中信息，利用本文给出的聚类算法，可以对客户进行分析，得出具有相似购买能力的客户群，从而可以使超市提供更快捷和周到的优质服务，提高客户满意度，吸引和保持更多的客户，增加营业额，并通过信息共享和优化商业流程有效地降低经营成本。最具有意义的聚类是那些包含最多事例的聚类，因为较小的聚类由于其一致特性只能提供较少的含义。由于不同特性的具体综合不能提供足够的信息以做出概括性的区别，所以有一些类可能被忽略。

3 结 论

本文结合基于密度和基于网格方法优势，给出了一种基于密度和网格的增量式聚类算法，在模拟数据库中进行实验，验证了算法的有效性。但是算法的聚类质量依赖于网格结构的最低层细度。若细度非常高，那么处理开销将会增加许多；若网格结构的最低层太粗，那就会降低聚类分析的质量。对于网格结构的细度以及对新增数据聚类的处理，还有待进一步的研究。

参 考 文 献

[1] SUN Zhi-wei, ZHAO Zheng, WANG Hong-mei. A clustering algorithm based on grid and density with random sampling[J]. Journal of Tianjin University, 2006, 39(5): 621-626.
[2] CHEN Ning, CHEN An, ZHOU Long-xiong. An incremental grid density-based clustering algorithm[J]. Journal of software, 2002, 13(1): 1-7.
[3] YAN Xin, ZHOU Li-hua, HEN Ke-ping, et al. Improved clustering algorithm based on density and grid in the presence of obstacles[J]. Computer Applications, 2005, 25(8): 1818-1823.
[4] LIU Jun-ling, SUN Huan-ling, WANG Da-ling, et al. Optimized cell-based clustering algorithm[J]. Mini-Micro Systems, 2006, 27(10): 1927-1930.
[5] LAI Jiang-zhang, NI Zhi-wei, LIU Zhi-wei. A grid fast clustering algorithm based on density-tree[J]. Computer Engineering, 2006, 31(17): 69-70.
[6] MA Guang-zhi, NI Guo-yuan. An increasable fuzzy clustering algorithm[J]. Microcomputer Applications, 2005, 26 (1): 5-7.
[7] QIU Bao-zhi, SHEN Jun-yi. Grid-based and extend-based clustering algorithm for multidensity[J]. Control and Decision, 2006, 21(9): 1011-1014.
[8] HU Yang, CHEN Gang. An effective cluster analysis algorithm based on grid and intensity[J]. Computer Application, 2003, 23(12): 64-67.
[9] ZHANG Guang-Jian, HUANG Xian-Ying. Study on a new clustering algorithm based on minimum clustering cell and its application in CRM[J]. Computer Technology, 2006, 33(7): 188-189.
[10] 陈梅兰. 基于网格和密度聚类算法研究[J]. 计算机与现代化, 2005, 2: 1-6.
[11] YAO Xiao-yun, LI Xuan, SU Qiang. Study on the customer relationship management and its application in Chinese hospital[C]// ICSSSMO5. chongqing: IEEE, 2005: 188-192.

编 辑 漆 蓉