基于最小聚类单元的聚类算法研究及其在CRM中的应用

张光建　黄贤英
（重庆工学院计算机系　重庆 400050）

摘　要　将聚类分析技术应用于客户关系管理可以改善客户关系，对将来的趋势和行为进行预测，优化营销策略。在综合分析网格聚类算法和K-均值聚类算法的基础上，提出了基于最小聚类单元（Minimum Clustering Cell，简称MCC）的聚类算法，介绍了该算法在CRM中的应用。经证明该算法是一种实用的、速度更快、效率更高的改进聚类算法，它克服了K-均值聚类需要事先给定K值、网格聚类要求数据密集的缺点。

关键词　数据挖掘，聚类，K均值聚类，网格，CRM（客户关系管理）

Study on a New Clustering Algorithm Based on Minimum Clustering Cell and its Application in CRM

ZHANG Guang-Jian　HUANG Xian-Ying
(Chongqing Institute of Technology, Chongqing 400050)

Abstract　The clustering technique of data mining can improve the relationship between enterprise and customers, forecast the trend and behaviors to support people's decision, optimize marketing policy. The advantages and disadvantages of K-means and grid clustering algorithm are given first, then a new clustering algorithm based on Minimum Clustering Cell (MCC) is presented and analyzed, which is proved to be correct, efficient and fast through application in CRM.

Keywords　Data mining, Clustering, k-means clustering, Grid, CRM

1　引言

聚类作为数据挖掘的一种方法被广泛应用。聚类是在预先不知道目标数据对象到底有多少类的情况下，对所有对象进行分类，使得以某种度量为标准的相似性（大多数算法都是基于距离的），在同一类中的对象之间最小化，在不同类中的对象之间最大化。这样，每一类中的成员都可以被统一看待。通过聚类，人们能够识别密集和稀疏的区域，从而发现全局的分布模式以及数据属性间的关系[6]。

K-均值聚类算法是一种不依赖于顺序的基于划分的算法，其时间复杂度与数据集的大小呈线性关系，缺点是必须指定K-值，且对初始中心点的选择比较敏感，如果初始值选择不当，将会收敛成为一个局部最小的准则函数[6]。

网格聚类算法采用多分辨率的网格数据结构，将数据空间量化为有限数目的单元，所有的聚类操作都在网格上进行。优点是处理速度快，处理时间独立于数据对象的数目，只与数据空间量化后的单元数有关，与原始数据对象的个数无关；缺点是要求数据点比较集中[6]。

本文在综合分析K-均值聚类算法和网格聚类算法的优缺点的基础上，提出了一种基于最小聚类单元（Minimum Clustering Cell，简称MCC）的聚类方法，目标是降低算法的时间复杂度。基本思想是：首先对原始数据进行一次网格微聚类，将每个独立的点归入对应的最小网格单元中；其次以网格为单位进行K-均值聚类，大大减少点的数量，从而大幅度地降低了算法的时间复杂度。通过分析证明该算法是一种速度更快、效率更高、聚类质量更好的改进算法，该算法应用于分析某百货公司批发客户的交易数据，可以帮助市场分析人员区分出不同的客户类，以便更加了解客户，进行市场划分，获取新的目标客户，提升现有客户价值，实现营销战略的优化[4]。

2　MCC算法及数据结构描述

2.1　MCC算法中的几个定义

定义1　设G为数据对象的集合，共有m个元素，给定阈值$T>0$，对G中任一个元素g_i，$i=1,2,\cdots,m$，它到中心点的欧几里得距离$d\leqslant T$，则称G为网格。一个网格也可用一个二元组$G=(m,M)$表示，m为网格中的数据对象点的个数，M为m个数据点的中心$M=\frac{1}{m}(\sum_{i=1}^{m}g_i)$。

定义2　网格U_1和U_2的中心点之间的欧几里德距离d称为U_1和U_2间的距离[2]。

定义3　对给定的n个数据对象$(x_1,x_2,\cdots,x_n)$，按照给定的阈值T将n个元素分别归入k个集合中，使得每个集合成为一个网格，称为网格划分[2]。

定义4　若某点不能划分到任意一个网格中，则称该点为异常点。若某点可以划分到多个网格中，则该点称为多个网格的公共点。

定义5　两个网格U_1和U_2，若它们之间有公共点，或是存在另一个网格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。

2.2　MCC算法的数据结构

设原始数据对象可表示为n个P维向量$X_i=(X_{i1},X_{i2},\cdots,X_{ip})$，$i=1,2,\cdots,n$，这些数据构成$n\times p$的数据矩阵DATA，数据点和变量结构的表示形式是：

$$DATA=\begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{pmatrix}$$

张光建　讲师，硕士，主要研究方向：软件工程应用，数据挖掘。

2.3 MCC 算法描述

输入:DATA 矩阵(聚类对象),异常点阈值 T_i,聚类阈值 T_j

输出:类－记录数表(用于存放分类的个数和每个类的记录总和)

方法:

Step1:将原始数据区域应用网格微聚类算法划分为 M 个网格,将每个数据点归入相应的网格中。

Step2:根据异常点阈值 S_i 判断每个网格中的数据对象是否为异常点。

Step3:对上述得到的每个非异常点的网格,计算它与其他网格之间的距离。

Step4:根据网格中心之间的距离是否小于给定的聚类阈值 T_j,合并相邻网格。

Step5:将最后得到的每个网格的中心点作为 K-均值算法的初始随机中心点,计算每个对象到已确定的各个网格中心的距离,比较后将该对象赋给最近的网格。然后重新计算每个网格的中心,重复这个过程,直到数据收敛。

3 MCC 算法在 CRM 中的应用

为了验证 MCC 算法的正确性和有效性,笔者设计并实现了该算法,并应用到某百货公司批发数据库上。在实验中共有 5000 个客户,客户属性包括顾客编号、所在地区、注册资金、经营年限、规模、年销售额、订货次数、购买品种、购买金额、为本企业带来的利润等。实验环境采用 Windows XP 上运行 ORACLE 9i,使用 Visual C^{++} 6.0 编写程序。

3.1 初始类的计算

每个客户作为一个数据对象,将 DATA 矩阵的 n 个数据构成 n 个初始类。为了方便查看测试结果,将原始数据的属性按贡献度大小取一定的权进行处理,使其映射到一定范围(600 * 320 像素)内分布的二维空间数据(X_i,Y_i),整个数据区域划分为 30 * 16 个网格。核查所有的数据点,将其映射后归入相应的网格中,结果如图 1,将网格的中心坐标、序号,以及包含的数据点保存在数据库表中。此步工作可在数据库中用 SQL 语句编写存储过程实现,算法如下:

1. 从数据库中取出记录 i 的所有属性 x_j

2. $\overline{X}_i=(\sum_{j=1}^{kx}\alpha_j\times x_j/\sum_{j=1}^{kx}x_j)\times 600$;

3. $\overline{Y}_i=((\sum_{j=1}^{ky}\beta_j\times x_j)/\sum_{j=1}^{ky}x_j\times 320$;//其中 α_j,β_j 为属性 j 对 $\overline{X}_i,\overline{Y}_i$ 的贡献度权值,kx,ky 为对 $\overline{X}_i,\overline{Y}_i$ 有贡献的属性总数,$\overline{X}_i,\overline{Y}_i$ 为变换后的空间纵坐标。

4. 在屏幕上画点标注该客户。

对屏幕上的点进行网格微聚类,结果如图 2,算法如下:

1. 对每个点的 $\overline{X}_i,\overline{Y}_i$ 坐标与每个网格(共 30 * 16 个网格)范围比较。

2. 若该点落入某个网格,就记录该点信息、网格中心点坐标和网格序号等。

3.2 选择初始的随机中心点

从网格微聚类结果出发,将第一阶段得到的网格中心作为初始类,计算各个网格中心之间的距离,用距离矩阵 D 表示,其中 $d_{ij}(i=1,2,\cdots,m;j=1,2,\cdots,m)$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,因为 $d_{ij}=d_{ji}$,故只需计算下三角元素。

$$D=\begin{pmatrix}0\\ d_{21} & 0\\ \cdots & \cdots & 0\\ d_{m1} & d_{m2} & \cdots & 0\end{pmatrix}$$

再合并距离矩阵中值小于聚类阈值 T_j 的类,得到一个较稳定的网格划分,共 K 个类,实验结果如图 3,有 4 个类。将其作为 K 均值算法的初始随机中心点。

3.3 从确定的初始类开始,运行 K-均值聚类算法

重新计算每个点到已确定的各个类的中心的距离,比较后将它赋给最近的类。重新计算每个类的中心,重复这个过程,直到数据收敛。算法描述如下:

1. Repeat;

2. 计算 K 个初始类 C_i 的中心 $M_i=\frac{1}{T}(\sum_{i=1}^{T}p_i)$,$T$ 为 C_i 中点的数目,p 为 C_i 中的点;

3. 计算每个点与 K 个中心点的欧几里得距离 $E_i=\sum(p-M_i)^2$,从这 K 个距离中找到最小的值,将该点(重新)赋给这个簇,p 为任意的点;

4. Until 各类的中心点不再变化。

运行上述程序后,确定了最后的中心点,得到的聚类结果如图 4。

4 实验结果分析及 MCC 算法评价

4.1 实验结果分析

由于实验采取的是已知数据,程序运行结果与人们的自然分类较吻合,证明算法的正确性。算法中要求输入异常点阈值和聚类阈值,这两个值的设置会影响聚类结果。因此在实际应用中,应根据用户的领域知识、经验及多次实验来设置这两个参数的值。

4.2 MCC 算法时间复杂性分析

假设数据对象数为 n,每个对象的维数为 p,n 个数据对象初始划分的网格数为 t,非异常点网格为 m,最后的聚类个数是 C。算法消耗的时间包括如下几部分:

(1) 将数据点映射到平面,并将其划分到初始网格中,消耗的时间与 $n\times(p+t)$ 成正比;

(2) 计算非异常点网格中心之间的距离,并进行合并,消耗的时间与 m^2 成正比;

(3) 确定最后的聚类,消耗的时间与 $n\times C+C^2$ 成正比;

因此整个算法的时间复杂度 $O(T)$ 约为:

$n\times(p+t)+m^2+n\times C+C^2$

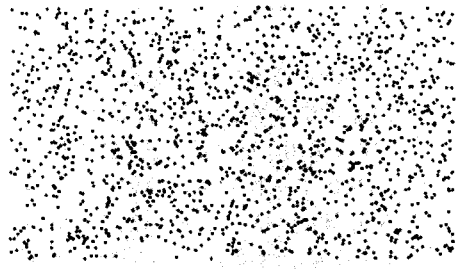

图 1 原始数据点分布

由于 $C\leqslant m\leqslant t$,由上式可看出,当 m 值较大时,整个聚类过程的时间复杂度主要取决于 m 的平方。当初始划分的网

(下转第 203 页)

参 考 文 献

1 Shekhar S, Ravada S, Fetterer A, et al. Spatial Databases: Accomplishments and Research Needs. IEEE Trans Knowledge and Data Eng, 1999,11(1):45～55
2 Nyerges T. Schema Integration Analysis for the Development of CIS Databases. Int J Geographic Information Systems, 1989,3(2):153～183
3 Gueting R H. An Introduction to Spatial Database Systems. The VLDB Journal, 1994,3(4):357～399
4 李德仁，王树良，李德毅，等. 论空间数据挖掘和知识发现的理论与方法. 武汉大学学报(信息科学版)，2002，27(3)：221～233
5 刘大有，王生生，虞强源，等. 基于定性空间推理的多层空间关联规则挖掘算法. 计算机研究与发展，2004，41(4)：565～570
6 李德毅. 知识表示中的不确定性. 中国工程科学，2000,2(10)：73～79
7 刘君强，潘云鹤. 挖掘空间关联规则的前缀树算法设计与实现. 中国图象图形学报，2003,04
8 Tung A K H, Lu Hongjun, Han Jiawei, et al. Efficient Mining of Intertransaction Association Rules. IEEE Transactions on Knowledge and Data Engineering, 2003, 15 (1)
9 裴韬，周成虎，骆剑承，等. 空间数据知识发现研究进展评述. 中国图象图形学报,2001,6(9):854～860
10 Han J, Kamber M. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, 2000
11 Ester M, Kriegel H P, Xu X. Knowledge Discovery in Large Spatial Databases: Focusing Techniques for Efficient Class Identification. In: Proc. 4th Int Symp on Large Spatial Databases, Portland, ME, 1995, Lecture Notes in Computer Science, Springer, 1995, 951:67～82
12 Egenhofer M. A Formal Definition of Binary Topological Relationships. In: Proc. Intl Conf On Foundations of Data Organization and Algorithms (FODO), 1989. 457～472
13 Efenhofer M, Franzosa R D. Point-Set Topological Spatial Relations. International Journal of Geographical Information Systems, 1991,5(2):161～174
14 Theodoridis Y, Stefanakis E, Sellis T K. Efficient Cost Models for Spatial Queries Using R-Trees IEEE Trans. Knowledge and Data Eng, 2000,12(1):19～32
15 Tian Yongqing, Weng Yingjun, Zhu Zhongying. Mining association rules based on cloud model and application in prediction. In: Proceedings of the 4-th World Congress on Intelligent Control and Automation. Shanghai, P. R. China, 2002. 2203～2207
16 Matsakis P, Keller J M, Wendling L, et al. Linguistic Description of Relative Positions in Images. IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics, 2001, 31(4): 573～588
17 Kuok C, Fu A, Wong M. Mining fuzzy association rules in databases. 1998, 27(1): 41～46

(上接第 189 页)

格数 t 越少，则聚类效率越高；当数据真正所占网格(非异常点网格)数 m 越少，则聚类效率越高。

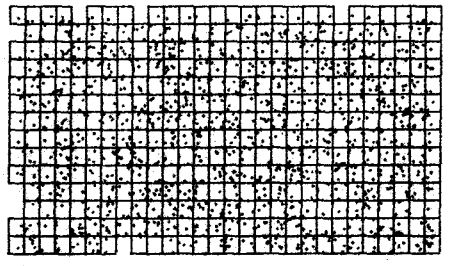

图 2 网格微聚类的结果

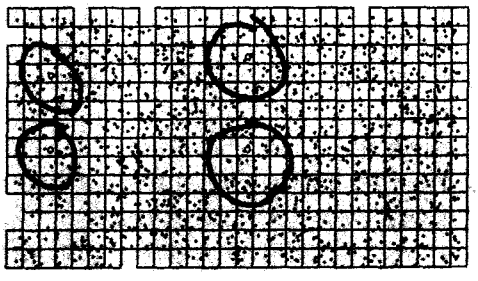

图 3 网格聚类选取的初始随机中心点

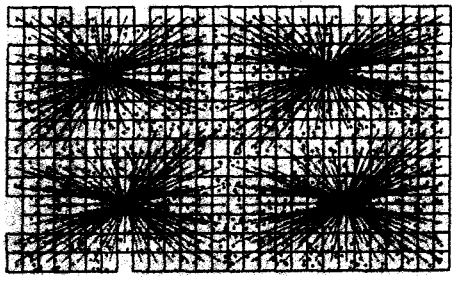

图 4 最后的聚类结果

4.3 MCC 算法评价

MCC 算法具有如下优点：

(1) 该算法克服了 K-均值算法要求事先给定类的个数的缺陷，由算法自动生成；

(2) 不必事先随机选取 K-均值算法的初始点，由算法自动选取初始点，从而避免了因初始点选择不当而导致收敛为一个局部最小准则函数的可能性；

(3) 该算法克服了网格聚类算法要求数据较集中的缺陷，可以用在数据较分散的情况中，而且可以方便有效地发现异常点；

(4) 该算法首先采用了网格聚类，所以算法的时间复杂度与数据对象数只呈一次线性关系，主要由网格的划分及数据的空间分布情况决定。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，时间复杂度低；

(5) 通过网格合并得到的聚类结果，也很容易给出其现实意义描述。

MCC 算法的主要缺点是聚类的结果依赖于异常点阈值和聚类阈值。

参 考 文 献

1 HarPeled S, Mazumdar S. Coresets for k-means and k-median clustering and their applications. In: Proc. 36th Annu. ACM Sympos. Theory Comput. ,2004. 291～300
2 Bach F R, Jordan M I. Learning spectral clustering. Advances in Neural Information Processing Systems (NIPS), 2004,16
3 LI CunHua, SUN ZhiHui. A Mean Approximation Approach to a Class of Grid－Based Clustering Algorithms. Journal of Software, 2003. 1267～1274
4 胡泱，陈刚. 一种有效的基于网格和密度的聚类分析算法. 计算机应用，2003(12)
5 Kanungo T, Mount D M, Netanyahu N, et al. A Local Search Approximation Algorithm for k－Means Clustering, Computational Geometry: Theory and Applications, 2004 (28): 89～112
6 Peng J M, Xia Y. A new theoretical framework for K－means clustering. To appear in Foundation and recent advances in data mining, Eds Chu and Lin, Springer Verlag, 2005
7 田启明，王丽珍，尹群. 基于网格距离的聚类算法的设计、实现和应用. 计算机应用，2005(2)
8 Han Jiawei, Kamber M. 数据挖掘概念与技术[M]. 北京：机械工业出版社，2001(8)
9 Berson A, Smith S, Thearling K. 构建面向 CRM 的数据挖掘应用. 人民邮电出版社，2001(8)