基于 Squeezer 算法的大规模矩阵聚类分析

李 岩 王惠文 叶 明 刘 丹

(北京航空航天大学 经济管理学院，北京 100191)

摘 要：对 Squeezer 算法进行分析研究，在定义 2 个矩阵之间距离的基础上，提出了一种改进的 Squeezer 算法，用于对维数相同的大规模矩阵进行聚类分析. 改进的算法在设定距离阈值的基础上，对类别的半径设定阈值来控制分类精度，给出具体的算法步骤来实现针对大量矩阵的聚类分析. 对聚类后所得矩阵集合，给出集合质心和半径的定义，来描述矩阵集合的特性. 所提算法能使聚类结果避免受到链条效应的影响而使类不断扩容，从而导致聚类精度下降的问题. 仿真实验分析验证了所提算法具有良好的聚类效果和适用性.

关 键 词：矩阵；聚类分析；Squeezer 算法；阈值

中图分类号：TP 311

文献标识码：A **文 章 编 号**：1001-5965(2009)12-1499-04

Modifiable Squeezer cluster algorithm used in large-scale matrix

Li Yan Wang Huiwen Ye Ming Liu Dan

(School of Economics and Management, Beijing University of Aeronautics and Astronautics, Beijing 100191, China)

Abstract: To solve the clustering method to the large-scale matrixes in the same dimension, the modifiable Squeezer cluster algorithm was proposed, based on the analysis of Squeezer cluster algorithm and the definition of the distance between the matrixes. The modifiable algorithm set a distance threshold, put forward a threshold of radius to control the accuracy of classification, and gave the detailed algorithm steps to realize cluster analysis for a large number of matrices. When the matrix cluster set was obtained, the modifiable algorithm provided the definition of center and radius to describe the properties of the matrix set. The proposed method could control the accuracy of classification in order to prevent chain effect in the course of clustering. The simulation experiment was addressed to validate the rationality and effectiveness of the modifiable algorithm.

Key words: matrix; cluster analysis; Squeezer; threshold

随着计算机技术、通讯技术以及网络技术的飞速发展，在许多领域出现了大量的数据记录. 为帮助用户从这些数据中分析出其所蕴涵的有价值的知识，数据挖掘技术应运而生. 聚类分析是数据挖掘研究领域中一个非常活跃的研究课题. 聚类分析是一个将数据集划分为若干个组或类的过程. 通过分类使得同一类的数据对象具有较高的相似度，而不同类间的数据对象差异较大，其中相似与否的度量是基于数据对象描述属性的取值来决定的. 聚类方法是通过比较数据的相似性和差异性，发现数据内在特征的分布规律，从而获得对数据更深刻的理解与认识. 现有的聚类技术大致可以分为如下几类[1]：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)和基于模型的方法(model-based method).

矩阵是很多科学和工程计算问题中研究的数学对象，大量的数据记录经过数学模型处理后，通常可以采用矩阵的形式进行表示. 本文拟讨论大

收稿日期：2009-02-15

基金项目：国家自然科学基金资助项目(70771004,70521001,70531010)

作者简介：李 岩(1979—)，男，河南正阳人，博士生，buaaly@126.com.

规模维数相同矩阵的聚类分析方法.在给出2个矩阵之间距离定义的基础上,通过对Squeezer聚类分析算法[2]进行研究和改进,实现了对大量矩阵的聚类分析.在计算过程中,还根据一类矩阵的质心与半径的定义[3],来设定距离的阈值,从而有效控制聚类的效果和精度.

1 矩阵聚类分析中的基本概念

矩阵聚类(martrix cluster)最初是为客户关系管理(CRM,Customer Relationship Management)提出的.矩阵聚类是一种新的数据挖掘算法,通过定义置信度和支持度的基础上,找出面积最大或密度最大的子矩阵,进行聚类操作.矩阵聚类算法由文献[4]提出,应用它同时对用户和资源项进行聚类,它从稀疏矩阵中抽取密度较大的区域来进行聚类.与该矩阵聚类思想不同,本文所提出的矩阵聚类算法是针对大规模矩阵集合进行的聚类研究,其中每个矩阵并非稀疏矩阵,而是具有相同$m\times n$阶实数所组成的矩阵.

矩阵可以看成一组数据的二维结构形式,也可以看成某类信息或某种特征集的二维有序排列.本节首先从矩阵范数[5]定义出发,提出2个矩阵之间的距离定义,并用它来衡量矩阵之间的相似程度.然后根据这一概念,讨论大规模矩阵的聚类分析算法.

定义1 对任一矩阵$\boldsymbol{A}=(a_{ij})\in\mathbf{R}^{m\times n}$,定义它的$\|\boldsymbol{A}\|$如式(1)所示:

$$\|\boldsymbol{A}\|\xlongequal{\text{def}}\operatorname{tr}\boldsymbol{A}'\boldsymbol{A} \tag{1}$$

由公式(1)所定义的范数又称为F-范数.根据矩阵范数的定义可以给出2个矩阵之间的距离定义.

定义2 设有2个$m\times n$维矩阵$\boldsymbol{A},\boldsymbol{B}\in\mathbf{R}^{m\times n}$,则$\boldsymbol{A},\boldsymbol{B}$之间的距离$d(\boldsymbol{A}|\boldsymbol{B})$定义为

$$d(\boldsymbol{A}\mid\boldsymbol{B})=\|\boldsymbol{A}-\boldsymbol{B}\|=\operatorname{tr}(\boldsymbol{A}-\boldsymbol{B})'(\boldsymbol{A}-\boldsymbol{B}) \tag{2}$$

选用该距离定义可以计算两个维数相同矩阵$\boldsymbol{A}$与$\boldsymbol{B}$中的元素差异的总量,因此可以作为衡量2个矩阵$\boldsymbol{A}$和$\boldsymbol{B}$相似程度的一个指标.

在矩阵聚类的过程中,为了衡量一个矩阵到某个矩阵集合的距离,还需要给出矩阵集合的质心概念.本文采用在矩阵集合中与其它矩阵距离总和最小的矩阵来表示该矩阵集合的质心,下面给出矩阵集合质心的定义[3].

定义3 记某一矩阵集合$\Lambda=\{\boldsymbol{M}_1,\boldsymbol{M}_2,\cdots,\boldsymbol{M}_n\}$,定义该矩阵集合的质心$\boldsymbol{M}$为一个满足公式(3)的矩阵:

$$\boldsymbol{M}=\min_{i=1,2,\cdots,n}\sum_{j=1}^{n}d(\boldsymbol{M}\mid\boldsymbol{M}_j) \tag{3}$$

为了使聚类的结果达到一定的精度,应把矩阵集合中的所有矩阵控制在一定的精度范围之内.下面给出矩阵集合半径的定义[3].利用该定义,可以设定聚类分析过程中的半径阈值,从而有效控制聚类分析的效果和精度.

定义4 记某一矩阵集合$\Lambda=\{\boldsymbol{M}_1,\boldsymbol{M}_2,\cdots,\boldsymbol{M}_n\}$,质心为$\boldsymbol{M}$,则该矩阵集合的半径$\boldsymbol{R}$定义为

$$\boldsymbol{R}=\max_{i=1,2,\cdots,n}d(\boldsymbol{M}_i\mid\boldsymbol{M}) \tag{4}$$

2 基于改进Squeezer算法矩阵聚类

Squeezer算法[2]是一种应用于大规模数据集的聚类算法.Squeezer算法不需要先验知识,基于数据间的最大相似度或最小距离,只需扫描数据集一遍即可将数据集分割成几个不同的超球体,对数据集进行聚类.

Squeezer算法采用不同属性值的取值频度来表示类.算法的基本思想如下:

1) 初始时,聚类集合为空,读入一个新的对象;

2) 以这个对象构造一个新的类;

3) 若已到数据库末尾,则转6),否则读入新对象,利用给定的距离定义,计算它与每个已有类间的距离,并选择最小的距离;

4) 若最小距离超过给定的阈值,转2);

5) 否则将该对象并入具有最小距离的类中,转3);

6) 结束.

Squeezer算法的优点是能够实现对大规模的数据集进行快速有效的聚类,然而该方法却容易受到"链条效应"的影响,也就是说在样本点之间的距离关系出现链条形式的连接时,往往会导致类质心的不断偏倚进而使类不断扩容,因此无法有效控制聚类结果的半径长度.而一个类别的半径如果过大,则类质心对该类中的全部元素的代表性就会很差.

为了克服Squeezer算法中的困难,本文提出了一种改进的Squeezer算法.为使聚类结果避免受到链条数据[6]的影响而使类不断扩容,从而导致聚类精度下降这一问题,算法在设定距离阈值的基础上,对类别的半径设定阈值.如果新的数据达到,在距离阈值许可范围内聚到某一类别中后,导致该类的半径大于给定的半径阈值,则新的数

据不能划分到该类别中,应构造一个新的类.本文采用该思想,并根据上节介绍的距离、质心以及半径的定义,设定矩阵集合的距离与半径阈值,进而对大规模矩阵进行聚类,算法描述如下.

对于大规模矩阵的集合 $\Lambda=\{\boldsymbol{M}_1,\boldsymbol{M}_2,\cdots,\boldsymbol{M}_n\}$,设定矩阵距离阈值为 r;矩阵集合的半径阈值为 δ.

1) 初始时,聚类集合为空,读入初始矩阵 $\boldsymbol{M}_1$;

2) 以这个矩阵构造一个新的类;

3) 若已到集合 Λ 的末尾,则转 9);

4) 否则读入新的矩阵,计算它与每个已有类间的距离 d,并选择最小的距离;

5) 若最小距离超过给定的距离阈值 r,转 2);

6) 否则先假设该矩阵属于具有最小距离的类中,并算出该类的质心 $\boldsymbol{M}$ 和半径 $\boldsymbol{R}$;

7) 若该类的半径超过给定的半径阈值 δ,转 2);

8) 若该类的半径在给定的半径阈值 δ 范围内,则将该矩阵正式并入该类中;

9) 结束,算出所有类的质心 $\boldsymbol{M}$ 和半径 $\boldsymbol{R}$.

3　仿真实验分析

本节将通过仿真实验来验证本文所提出的聚类算法的有效性.此外,为了说明避免链条效应在聚类分析中的意义,下面分别采用 Squeezer 方法与改进 Squeezer 方法进行聚类分析,并讨论 2 个算法计算结果之间的差异性.

在仿真实验中,采用随机数发生器,按照 $\boldsymbol{M}_1\sim N(4,1)$,$\boldsymbol{M}_2\sim N(5,1)$,$\boldsymbol{M}_3\sim N(6,1)$生成 3 个 6×5 的基本矩阵:

$$\boldsymbol{M}_1=\begin{bmatrix}5.475\,041 & 3.580\,548 & 3.347\,367 & 3.382\,558 & 4.967\,930\\ 2.780\,684 & 4.115\,887 & 3.681\,874 & 3.050\,722 & 3.704\,646\\ 2.668\,391 & 6.269\,226 & 4.176\,437 & 4.034\,393 & 4.254\,595\\ 2.896\,719 & 1.128\,834 & 5.434\,273 & 3.423\,515 & 3.327\,933\\ 3.927\,454 & 3.050\,722 & 3.382\,002 & 2.275\,725 & 2.367\,639\\ 5.650\,305 & 3.214\,990 & 4.856\,032 & 4.154\,644 & 3.161\,928\end{bmatrix}$$

$$\boldsymbol{M}_2=\begin{bmatrix}4.717\,567 & 4.734\,644 & 4.449\,873 & 2.572\,377 & 5.515\,913\\ 4.902\,041 & 3.850\,558 & 4.763\,617 & 5.893\,037 & 3.802\,434\\ 4.483\,562 & 6.962\,671 & 4.255\,169 & 4.300\,837 & 5.331\,999\\ 4.818\,743 & 6.255\,166 & 4.058\,257 & 2.721\,643 & 6.034\,871\\ 5.441\,103 & 5.048\,404 & 5.215\,117 & 4.774\,930 & 5.775\,055\\ 6.160\,022 & 4.257\,388 & 7.059\,032 & 5.307\,361 & 5.170\,691\end{bmatrix}$$

$$\boldsymbol{M}_3=\begin{bmatrix}7.023\,179 & 6.469\,017 & 6.947\,839 & 6.269\,400 & 6.268\,052\\ 4.804\,155 & 4.215\,126 & 5.958\,948 & 5.520\,110 & 5.742\,638\\ 6.341\,471 & 5.523\,541 & 4.612\,898 & 7.240\,669 & 8.014\,021\\ 5.023\,971 & 6.890\,759 & 4.878\,254 & 5.211\,027 & 7.098\,792\\ 6.466\,883 & 4.314\,020 & 6.538\,064 & 5.656\,907 & 5.297\,415\\ 5.248\,896 & 5.683\,563 & 5.462\,555 & 7.187\,907 & 5.265\,729\end{bmatrix}$$

然后,对基本矩阵的每一项 m_{ij},$i=1,2,\cdots,6$,$j=1,2,\cdots,5$,加上线性趋势函数 $y=0.01x+\varepsilon$,其中 $x=1,2,\cdots,199$,$\varepsilon\sim N(0,0.001)$为一随机干扰项.在对每个基本矩阵加上 199 个链式数据后,共生成 600 个矩阵,作为初始数据源,然后利用改进 Squeezer 对 600 个矩阵集合进行聚类.

本文采用的实验环境如下:Intel Pentium 1.5 GHz处理器,758 M 内存,OS 为 Windows XP,以 Visual C# 为工具实现算法编程.实验中,矩阵总数$N=600$,矩阵距离阈值 r 和矩阵集合半径阈值 δ 可调.

采用 Squeezer 算法的思想对矩阵进行聚类,取矩阵距离阈值 $r=4$,得到聚类结果如表 1 所示.采用改进 Squeezer 算法的思想对矩阵进行聚类,取矩阵距离阈值 $r=4$,矩阵集合半径阈值 $\delta=2.8$,得到聚类结果如表 2 所示.两种算法的聚类效果比较如图 1 所示.

表 1　Squeezer 算法聚类结果

类别	类中元素个数	类的半径值
1	146	3.998 175 531 963 60
2	54	1.475 186 468 108 76
3	146	3.998 437 823 577 47
4	54	1.479 126 911 438 31
5	147	3.999 683 731 358 63
6	53	1.425 238 122 435 37

由表1、表2和图1可以看出，改进Squeezer算法在Squeezer算法聚类结果上进行了进一步的细分. Squeezer算法在一定距离阈值许可范围内对数据进行聚类，当某些具有一定趋势的数据聚到某一类中，能使类的边界不断放大，质心也在不断滑动，从而对聚类的效果和精度有一定程度的影响. 改进Squeezer算法设定类的半径阈值来控制类的大小，当新的数据到来时，在距离阈值的范围内并也在类的半径范围内时，则归为该类；如果超过半径阈值的许可范围，则重新构成一个新的类. 因此，采用改进Squeezer算法能达到更好的聚类效果，并且能有效控制分类的精度.

表2 改进Squeezer算法聚类结果

类别	类中元素个数	类的半径值
1	103	2.794 600 066 900 38
2	97	2.629 924 038 649 45
3	103	2.792 924 411 735 35
4	97	2.630 544 558 480 29
5	102	2.792 584 757 149 63
6	98	2.683 460 662 534 04

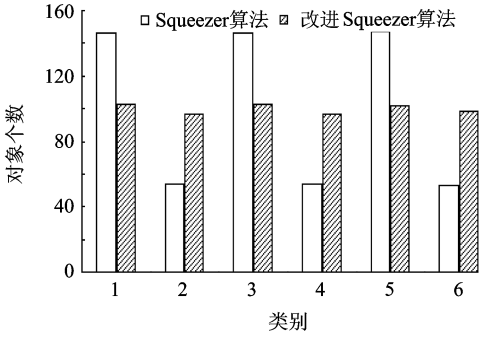


图1 算法聚类效果比较

4 结束语

本文通过F-范数给出2个矩阵之间的距离定义. 在此基础上，提出一种改进的Squeezer方法，并用于矩阵集合的聚类分析. 通过仿真分析，说明了所提方法可以有效地控制矩阵集合的聚类分析精度. 文中通过仿真实验验证了所提算法具有良好的聚类效果和适用性. 此外，还通过对Squeezer方法与改进Squeezer方法进行比较分析，讨论2种计算方法的差异性，说明改进的Squeezer方法可以明显地控制矩阵聚类过程中的链条效应.

参考文献（References）

[1] 胡庆林，叶念渝，朱明富. 数据挖掘中聚类算法的综述[J]. 计算机与数字工程，2007，35(2)：17－20
Hu Qinglin, Ye Nianyu, Zhu Mingfu. Survey of cluster analysis in data mining[J]. Computer & Digital Engineering, 2007, 35(2): 17－20(in Chinese)

[2] He Zengyou, Xu Xiaofei, Deng Shengchun. Squeezer: An efficient algorithm for clustering categorical data[J]. Journal of Computer Science and Technology, 2002, 17(5): 611－624

[3] Ye Ming, Wang Huiwen, Wang Lanhui. Application of improved hierarchical clustering method to classification of curves[C]// The 9th International Conference on Industrial Management. Beijing: China Aviation Industry Press, 2008: 325－330

[4] Oyanagi S, Kubota K, Nakase A. Application of matrix clustering to web log analysis and access prediction[C]// Proceedings of the ACM Web KDD Workshop on Mining Log Data across all Customer Touch Points. Berlin: Springer-Verlag, 2001

[5] 陈祖民，周家胜. 矩阵论引论[M]. 北京：北京航空航天大学出版社，1998：281－288
Chen Zumin, Zhou Jiasheng. Introduction of matrix theroy [M]. Beijing: Beijing University of Aeronautics and Astronautics Press, 1998: 281－288 (in Chinese)

[6] Li Yan, Ye Ming, Wang Huiwen, et al. A data streams clustering algorithm based on interval data[C]//Beijing: The 38th International Conference on Computers and Industrial Engineering. Beijing: Publishing House of Electronics Industry, 2008: 2775－2778

（编 辑：文丽芳）