# 基于奇异值分解的异常切片挖掘*

遇 辉[1,2+], 马秀莉[1,2], 谭少华[1,2], 唐世渭[1,2], 杨冬青[1]

[1](北京大学 信息科学技术学院,北京 100871)

[2](北京大学 视觉与听觉信息处理国家重点实验室,北京 100871)

## Exceptional Slices Mining Based on Singular Value Decomposition

YU Hui[1,2+], MA Xiu-Li[1,2], TAN Shao-Hua[1,2], TANG Shi-Wei[1,2], YANG Dong-Qing[1]

[1](School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China)

[2](National Laboratory on Machine Perception, Peking University, Beijing 100871, China)

+ Corresponding author: Phn: +86-10-62755745, Fax: +86-10-62754911, E-mail: yuhui@db.pku.edu.cn, http://www.pku.edu.cn

Received 2004-07-16; Accepted 2005-03-11

**Yu H, Ma XL, Tan SH, Tang SW, Yang DQ. Exceptional slices mining based on singular value decomposition.** ***Journal of Software*****, 2005,16(7):1282−1288.** DOI: 10.1360/jos161282

**Abstract**: Slice is one of the major operations in on-line analysis processing, which has played an important role in the application of decision support. In this paper, a method of mining exceptional slices is presented for extracting the distribution feature of the slice data based on the technique of the singular value decomposition, and the exceptional slices can be found by utilizing the distance-based outlier detection technique on the singular value feature. The effectiveness of the approach is experimentally demonstrated on the artificial data and the real slices data.

**Key words**: exceptional slices mining; feature extraction; singular value feature vector; distance-based outlier detection; on-line analytical processing

摘 要: 切片操作是联机分析处理的主要功能之一,在决策支持应用中发挥着重要作用.由于人工的切片过程非常低效,且易忽略重要信息,提出了一种自动、智能的异常切片挖掘方法.该方法基于奇异值分解技术来提取切片的数据分布特征,然后在提取出的奇异值特征之上,利用基于距离的孤立点检测技术发现异常的切片.在人工生成的数据和实际应用的切片数据上所作的实验结果都表明了该方法的高效性和可行性.

关键词: 异常切片挖掘;特征提取;奇异值特征向量;基于距离的孤立点检测;联机分析处理

中图法分类号: TP311 文献标识码: A

联机分析处理(OLAP)是一种数据分析技术,它通过提供多角度、多粒度的查询和展现数据的功能,使得人

* Supported by the National Natural Science Foundation of China under Grant Nos.60473051, 60473072 (国家自然科学基金)

作者简介: 遇辉(1977−),女,黑龙江哈尔滨人,博士生,主要研究领域为数据挖掘,联机分析处理,模式识别;马秀莉(1972−),女,博士后,主要研究领域为数据挖掘,联机分析处理;谭少华(1960−),男,博士,教授,博士生导师,主要研究领域为人工智能,模式识别;唐世渭(1939−),男,教授,博士生导师,CCF 高级会员,主要研究领域为数据库,信息系统;杨冬青(1945−),女,教授,博士生导师,CCF 高级会员,主要研究领域为数据库,信息系统.

们得以灵活地观察和分析数据.

自从 E.F.Codd 于 1993 年提出联机分析处理的概念以来,针对 OLAP 的研究大致分为两个阶段.目前广泛使用的 OLAP 基本操作属于第一阶段,即上钻、下钻、旋转、切片、切块等.近几年,研究者们根据实际应用的需要,进一步提出了 OLAP 的高级操作符,包括智能上钻、智能下钻、用户导航符、异常指示符、对数据立方体进行语义压缩的商方体以及方体间的关联分析等[1−6].

OLAP 的基本操作是用户指导的汇总和比较,功能有限,且由于这种分析依赖于用户的直觉,在大量数据中进行分析时容易迷失.高级操作符的提出弥补了基本操作的部分缺陷,使得 OLAP 中常用的一些分析得以高效、自动地实现.尽管如此,仍有许多分析功能需要更进一步的自动化和智能化,切片操作就在此列.

切片操作是在给定的数据立方体的目标维上选择一个维值,以展现其余的维组合所构成的数据子方.在实际应用中,同一维上不同维值对应的各个切片的数据分布经常会具有某些相关性,因为这些数据并非孤立产生.分析人员之所以进行切片操作,就是为了观察切片内部的数据分布特点以及进行各个切片之间的相互比较,从而发掘各个切片之间数据分布的异同性,找到数据中潜在的信息,为决策支持服务.

通常情况下,如果某一个切片显著地不同于其他切片的数据分布特点,那么它就会引起分析人员的兴趣,因为异常就意味着它可能拥有更大的信息量,具有深入分析的价值.例如,销售管理人员想知道,2003 年哪个月份的销售情况比较特殊?哪个地区与其他地区相比具有独特性?这两个问题分别对应着时间维和地区维的切片分析.

目前,对切片的这种分析都是人工进行的.分析人员首先要对目标维(在上面例子中,目标维分别是时间维和地区维)上的各个维值进行切片展现,然后观察各个切片的数据分布特点,比较各个切片之间是否具有相似性,进而找到某些异常切片.在实际应用中,由于待分析的数据经常是大量的,且数据中存在着不可避免的噪声,使得这种分析方式非常低效.随着数据量的增大和数据维数的增多,人工的探查不仅容易迷失,而且难以发现或忽略有价值的信息.

针对这种状况,本文提出了一种自动的、智能的异常切片挖掘方法.我们以矩阵的视角来看待切片,对切片实施奇异值分解,提取切片数据的基本代数特征,即奇异值特征.我们将奇异值特征映射到空间,每一个向量就对应了空间中的一个点.然后,基于点之间的欧几里德距离进行基于距离的孤立点检测,从而得到孤立点对应的异常切片.

本文第 1 节给出问题的描述.第 2 节详细讨论切片特征的提取方案和异常切片的挖掘方法.第 3 节给出实验结果和比较分析.第 4 节对全文作出总结.

## 1 问题描述

一个切片被称为异常,通常存在几种情况.一种情况是,这个异常切片的数据分布范围明显不同于其他切片.对于这种情况,通过比较切片的一阶矩(均值)和二阶矩(方差),可以很容易地找到异常的切片.另一种情况是,想要寻找的是那些在数据分布特点上比较特殊的切片.对于这种情况,如果所有切片的取值范围相似,均值和方差经常是无能为力的,这时就需要寻找一种能够刻画切片内部数据分布特点的特征提取方法.本文就是要为第 2 类情况提供解决方案.

下面,我们通过一个例子来给出问题的描述.为了描述方便,我们将切片的维数限制在二维,对于多维切片的分析可以很容易地从二维扩展得到.原因是,如果我们以交叉表的方式组织多维切片的展现,只需将多个维进行排列,组合成二维矩阵的行和列即可,因此总可以将多维切片看成是二维矩阵.

假设,数据立方体的三个维分别是时间维、产品维和地区维,度量是销售额.待分析的问题是:某一指定时间段,哪个月份的销售额在产品维和地区维的分布情况比较特殊?在时间维上,以“月”为粒度,对指定时间段内的所有月份进行切片(假设共分析 $t$ 个月的数据).那么,每一个切片都是由产品维和地区维构成的二维数据矩阵,矩阵元素就是度量值,即销售额,如图 1 所示.

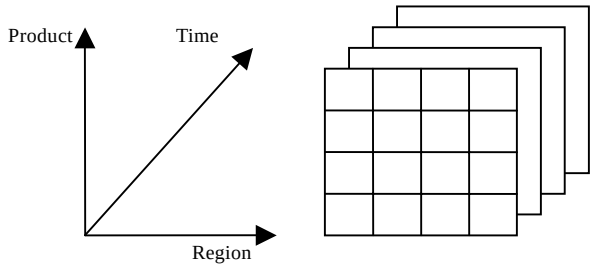



Fig.1 Slices on the target dimension

图 1 目标维上的切片

假设,产品维在某指定粒度下有 $m$ 个取值,地区维在某指定粒度下有 $n$ 个取值,那么,每一个切片就是一个 $m\times n$ 的二维矩阵,$t$ 个切片表示成矩阵形式为$\{c_{ij}^k \mid i=1,\ldots,m,j=1,\ldots,n\},k=1,\ldots,t$.把切片看成数据矩阵之后,对切片之间的分析就是对多个矩阵的数据分布特点进行分析,异常切片的挖掘就是要找到那些数据分布显著不同于其他矩阵的矩阵.我们为每一个切片矩阵提取出 $s$ 个特征$\{f_1,f_2,\ldots,fs\}$,作为刻画该矩阵数据分布情况的 $s$ 维特征向量.$t$ 个切片对应了 $t$ 组特征,$f^k=\{f_1^k,f_2^k,\ldots,f_s^k\},k=1,\ldots,t$.那么,对切片的比较就转换成对 $t$ 个特征向量 $f^1,f^2,\ldots,f^t$ 的比较,找到异常的 $f^i,i\in\{1,2,\ldots,t\}$,就找到了异常的切片.

根据矩阵论,奇异值分解(SVD)是一种有效的代数特征抽取方法,在描述矩阵数据分布特征上具有多项优良特性.它能够捕获矩阵数据的重要的基本结构,可以反映矩阵的代数本质,它在图像压缩、信号处理和模式识别等领域中都有着广泛的应用.因此,我们将采用 SVD 方法抽取出切片矩阵的奇异值特征,然后基于这种奇异值特征向量对切片矩阵数据分布的描述,找到异常的切片.

## 2 基于奇异值分解的异常切片挖掘方法

### 2.1 切片矩阵的奇异值特征向量

引理 **1**(奇异值分解). 对于任一实矩阵 $A_{m\times n}$,秩$(A)=r$,则存在两个标准正交矩阵 $U_{m\times m}$ 和 $V_{n\times n}$ 以及对角阵 $D_{m\times n}$,使得下式成立 $A=UDV^T$,其中, $D_{m\times n}=\begin{bmatrix}\Sigma_{r\times r} & 0\\ 0 & 0\end{bmatrix}$, $\Sigma_{r\times r}=diag(\sigma_1,\sigma_2,...,\sigma_r)$, $U_{m\times m}=(u_1,u_2,...,u_r,u_{r+1},...,u_m)$, $V_{n\times n}=(v_1,v_2,...,v_r,v_{r+1},...,v_n)$ [7].

$T$ 表示矩阵转置, $\sigma_i=\sqrt{\lambda_i}$ $(i=1,2,...,r,...,n)$ 称为矩阵 $A$ 的奇异值,$\lambda_1\geq\lambda_2\geq\ldots\geq\lambda_r\geq 0,\lambda_{r+1}=\lambda_{r+2}=\ldots=\lambda_n=0$ 是矩阵 $A^TA$ 和 $AA^T$ 的特征值. $u_i,v_i$ $(i=1,2,...,r)$ 分别是 $A^TA$ 和 $AA^T$ 对应于非零特征值 $\lambda_i$ 的特征向量,如图 2 所示.

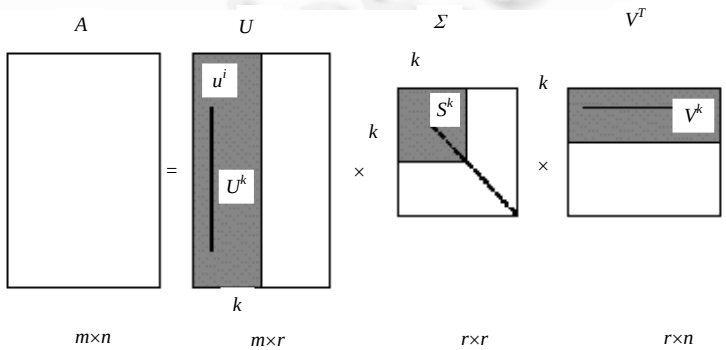



Fig.2 Singular value discomposition

图 2 矩阵的奇异值分解

在$\lambda_1\geq\lambda_2\geq\ldots\geq\lambda_r$ 的限制下,矩阵的奇异值向量 $(\sigma_1,\sigma_2,...,\sigma_r)$ 是唯一的,它刻画了矩阵数据的分布特征.直观上,可以这样理解奇异值分解,将矩阵 $A_{m\times n}$ 看成是一个线形变换,它将 $n$ 维空间的点映射到 $m$ 维空间,而经过对

$A_{m\times n}$ 进行奇异值分解之后,这种映射被分割成 3 个部分,分别是 $U_{m\times m}$,$D_{m\times n}$ 和 $V_{n\times n}$.其中的 $U_{m\times m}$ 和 $V_{n\times n}$ 都是标准正交矩阵,那么它们对应的线形变换就相当于分别对 $m$ 维和 $n$ 维坐标系中坐标轴的旋转变换.而 $D_{m\times n}$ 是对角矩阵,在线形变换中,相当于对各个坐标轴进行伸缩变换.由此可见,对矩阵进行奇异值分解之后,只有奇异值向量构成的对角矩阵保留了矩阵的代数本质.因此,可以将奇异值向量作为切片矩阵的代数特征.

我们以一个简单的例子来说明奇异值向量的代数意义,图 3 中的两个表分别代表两个切片.

| | | |
|---|---|---|
| 1 | 2 | 3 |
| 4 | 5 | 6 |
| 7 | 8 | 9 |

| | | |
|---|---|---|
| 8 | 6 | 1 |
| 5 | 4 | 9 |
| 3 | 2 | 7 |

Fig.3 Two slices

图 3 两个切片

可以看到,它们的均值和方差相同,但数据的分布稍有不同.它们的奇异值特征向量分别是(16.8481,1.0684,0.0000)和(15.3612,6.9968,0.2791).显而易见,这两个切片的奇异值特征捕获了它们在数据分布上的差异,抓住了两个矩阵的数据分布在基本结构上的不同.

矩阵的奇异值特征向量在刻画矩阵数据的代数分布特点的同时,还具有一个重要的优点,即稳定性.稳定性确保了奇异值对矩阵元素的扰动不敏感,因此这种奇异值特征对数据噪声不敏感,保证了对切片的特征提取不会因为一些细微的数据变化而失效.

然而,没有一种特征提取方案是万能的,奇异值特征也有它的适用范围.奇异值特征具有的转置不变性、旋转不变性、位移不变性和镜像变换不变性会导致它在对应情形下失去区分能力.例如,当一个切片的数据是另一个切片的转置时,虽然它们代表了两种不同的数据分布,但是提取出来的奇异值向量却完全相同.因此,对于这种情形,奇异值向量是无能为力的.

综上所述,我们可以总结出将奇异值向量作为切片特征的优点:

- 捕获了切片数据分布的重要的基本结构,反映了切片的内在代数特征,保证了基于奇异值分解分析切片之间数据分布异同的可行性;
- 去除了噪声的影响,保证了特征抽取的稳定性;
- 由于奇异值特征是 $r$ 维的,它使原切片的特征空间由 $m\times n$ 维下降到了 $r$ 维,提高了分析的效率.

### 2.2 异常切片的挖掘方法

原始切片相当于是在 $m\times n$ 维空间中的点,经过对其实施奇异值分解,得到了 $r$ 维的奇异值特征,这相当于是把 $m\times n$ 维空间的点映射到 $r$ 维空间,而这种映射保留了原空间中点(切片)之间的距离[8].也就是说,任意两个切片在 $m\times n$ 维空间中的距离与它们的奇异值特征在降维后的 $r$ 维空间中的距离是等价的.基于这种距离不变的特性,我们可以把对切片之间的相似或相异性比较,通过对奇异值特征的比较来完成.

一组切片对应了 $r$ 维空间中的一组点,这些点在 $r$ 维空间的分布情况反映了原始切片之间的相对关系.因此,我们通过分析这些 $r$ 维空间的点的分布情况来实现对一系列切片的比较分析.数据分布相似的切片,映射到奇异值特征空间后会具有较近的距离,而那些具有显著不同特征的切片,就会以孤立点的形式出现.那么,我们就可以通过挖掘多维空间中的孤立点来找到异常切片.

在数据挖掘领域,孤立点的检测方法分为 3 大类,分别是基于统计的方法、基于距离的方法和基于偏离的方法[9].基于统计的方法在此处是不适用的,因为数据的未知性使我们无法事先给出切片数据分布的统计模型.基于偏离的方法通过检查一组对象的主要特征来确定孤立点,与给出的描述“偏离”的对象被认为是孤立点,这种方法也不适用于检测异常的切片.而基于距离的方法将那些没有“足够多”邻居的对象视为孤立点,这里的邻居是基于距给定对象的距离来定义的,刚好与我们的问题需求相吻合.通过度量多维空间中点之间的距离,找到空间中哪些点所在位置的周围是稀疏的,从而发现孤立的点,即异常的切片.在信号处理和模式识别等领域,大多是采用欧几里德距离公式对提取出的特征进行相似(异)性度量,这是因为欧几里德距离的球状特点符合对特征相似性描述的需求,因此,我们也采用欧几里德距离来度量切片的特征.

基于距离的孤立点检测方法对孤立点作如下定义:如果数据集合 $S$ 中对象至少有 $p$ 部分与对象 $o$ 的距离大于 $d$,则对象 $o$ 是一个带参数 $p$ 和 $d$ 的基于距离的孤立点,即 $DB(p,d)$[10].由于得到的奇异值特征都是高维的,所以不适合采用基于单元(cell-based)的算法.一般情况下,一个维的取值数目不会是非常多的,所以没有必要采用嵌套-循环算法.设 $M$ 是一个孤立点的 $d$-邻域内的最大对象(对象在此处即是多维空间中的点)个数,我们直接计算每个点半径 $d$ 范围内的邻居,当发现多于 $M$ 个点出现时,这个点就不是孤立点.基于奇异值分解的异常切片挖掘算法如下.

算法 **1**(异常切片挖掘).

输入:目标维 $dim$,孤立阈值 $M$,距离 $d$.

输出:异常切片集 $S$.

过程:

(1) 假设目标维 $dim$ 共有 $t$ 个维值,对 $t$ 个维值分别进行切片,得到 $t$ 个切片,用矩阵的形式表示:

$$\{c_{ij}^k | i=1,\ldots,m, j=1,\ldots,n, k=1,\ldots,t\}.$$

(2) 对 $t$ 个切片 $\{c_{ij}^k\}, k=1,\ldots,t$ 进行奇异值分解,得到 $t$ 个奇异值特征向量:

$$(\sigma_1^k, \sigma_2^k, ..., \sigma_{r_k}^k), k=1,\ldots,t.$$

(3) 设 $r_k$ 是第 $k$ 个切片矩阵的秩,取 $r = \max\{r_k\}, k=1,...,t$,得到 $r$ 维空间的 $t$ 个点:

$$(\sigma_1^k, \sigma_2^k, ..., \sigma_r^k), k=1,\ldots,t.$$

(4) 根据欧几里德距离计算每一个点 $d$ 邻域内的点的个数,那些少于 $M$ 个临近点的点即为孤立点,将其对应的切片加入异常切片集 $S$.

## 3 实验与结果分析

### 3.1 实验环境

实验的硬件环境为 P4 1.5Ghz 的 CPU 和 512MB 的内存,软件环境为 Windows 2000(Professional)操作系统,所有代码均用 Visual C++(6.0)实现.

### 3.2 实验结果分析

#### 3.2.1 可行性

为了验证基于奇异值分解的异常切片挖掘方法的可行性,我们对模拟切片数据和实际切片数据均进行了实验.

为了衡量实验结果的正确程度,我们给出准确率的计算公式:

$$\frac{R}{R+W} \times 100\% .$$

其中,$R+W$ 是返回的所有异常切片的总数,$R$ 表示找到的正确的异常切片个数,$W$ 表示返回结果中错误的异常切片个数.

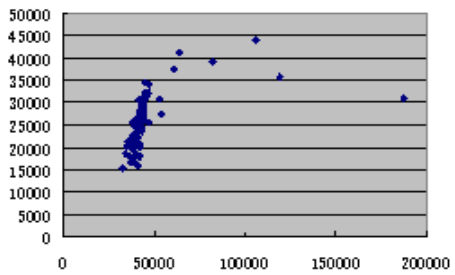
一般情况下,实际应用中的切片数据包括了噪声干扰,人眼很难看出切片内的数据分布状况,也很难进行切片之间的比较.为了精确地验证本文提出的奇异值特征抽取方法的有效性,我们以人工方式生成了多类切片,同类切片具有相似的数据分布特征,而不同类之间切片的数据分布则互不相同.每一类内部的各个切片虽然都具有相似的数据分布特点,但我们在生成每一个切片时加入了随机的扰动,使得类内切片虽然具有相似的形状,但却并不相同.

我们人工生成了 3 种类型的切片,第 1 类切片的数据分布类似于二维正态分布的取值形状,切片矩阵的数据满足中间部分的值大,向四周逐渐减少的特征.第 2 类切片的数据分布呈散射形态,切片矩阵的左下角位置的值最小,从左下角向上、向右,数据取值逐渐增大.第 3 类切片的数据则是随机生成的.生成切片的程序包括如下参数:具有不同数据分布特点的各个类的切片数目、切片的长和宽、切片数据的取值范围、控制各个类内切片数据分布形状的参数.为了验证奇异值特征是否能够刻画切片内数据分布的本质特征,我们使生成的各类切片

具有相同的取值范围和相同的切片大小.

我们进行了多组实验,以第 1 组实验为例,生成第 1 类切片 50 个,第 2 类切片 50 个,第 3 类切片 5 个,其中第 3 类的 5 个切片被认为是异常切片.所有这些切片的数据取值范围都是 1~20,切片的大小均为 20×20.我们为这 105 个切片提取奇异值特征,然后采用基于距离的孤立点检测方法找到一系列异常的切片.我们共进行了 20 组实验,实验结果的准确率均是 100%.虽然我们在生成切片时加入了随机的扰动,但这并不影响挖掘结果的准确性,这也证明了奇异值特征对噪声的稳定性.

此外,我们对实际的多维数据也进行了切片实验.实验数据来自移动网管的数据仓库系统,我们抽取其中 100 个 10×26 大小的切片,进行异常切片的挖掘.通常情况下,矩阵的奇异值从大到小排列呈现迅速的衰减趋势,而前几个最大的奇异值反映了矩阵的主要特征,同时也为了直观地展示奇异值向量对切片数据分布特征的描述能力,我们将前两个最大奇异值映射到二维平面上(如图 4 左所示),将前 3 个奇异值映射到三维空间(如图 4 右所示).



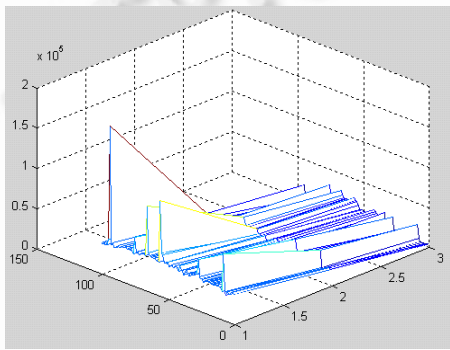




Fig.4 The distinguish ability of the SVD feature

图 4 奇异值特征的区分能力

由图 4 可以看到,有几个切片的奇异值呈现异常情况.我们将这几个异常情况对应的切片与其他切片进行了比较,发现它们的数据分布情况确实存在着差异.由此可见,奇异值特征向量刻画了切片数据的分布特征,对于不同分布形态的切片,具有很强的区分能力.

3.2.2 可扩展性

我们从切片数目和切片大小两个角度来分析算法的效率.首先,我们固定切片的大小,观察随着切片数目的增加,算法执行时间呈现怎样的变化趋势.我们对大小为 50×50 的切片进行实验,当切片数目从 100 个逐步增加到 5000 个时,算法的执行时间如图 5 所示.在切片数目小于 4000 时,虽然算法的孤立点挖掘部分是指数时间算法,但由于孤立点的检测过程在整个异常切片的挖掘过程中占用的时间资源相对较小,故整个算法的执行时间随着切片数目的增加主要呈线性变化.当切片数目达到 4000 左右时,孤立点检测过程所占用的时间资源达到整个算法执行时间的一半.从这一点开始,随着切片数目的增加,系统的时间消耗就主要来自于孤立点检测过程.但在实际应用中分析异常切片时,待分析的切片极少达到这么大的数目,因此依然可以认为算法是可扩展的.

当固定切片的数目、增加切片大小时,由于计算矩阵奇异值的过程首先要计算矩阵的转置与矩阵的乘,这是需要指数时间来完成的,因此,当切片大小增加时,执行时间将呈指数递增.我们分别将切片数目固定为 100 和 200,待分析的切片大小从 50×50 到 200×200 之间发生变化,对它们进行了实验,结果如图 6 所示.算法的执行时间虽然呈指数增加,但对于 200 个大小为 200×200 的切片,执行时间为 500 秒左右,这是数据挖掘过程可以接受的时间范围.

异常切片挖掘算法中有两部分涉及指数级操作,一个是矩阵奇异值的计算过程,算法的时间复杂度是 $O(m\times n^2)$,其中 $m\times n$ 代表矩阵的大小;另一个是孤立点挖掘部分,时间复杂度是 $O(n^2)$.这里,$n$ 表示目标维的维成员总数.由于这两部分是相对独立的过程,因此,为了尽量避免指数级的操作,要分别从如下两个方面入手.一方面可以探索其他的特征抽取方法替代矩阵的奇异值特征;另一方面要研究更高效的孤立点挖掘算法,或是采用其他替代方法找到异常的切片.

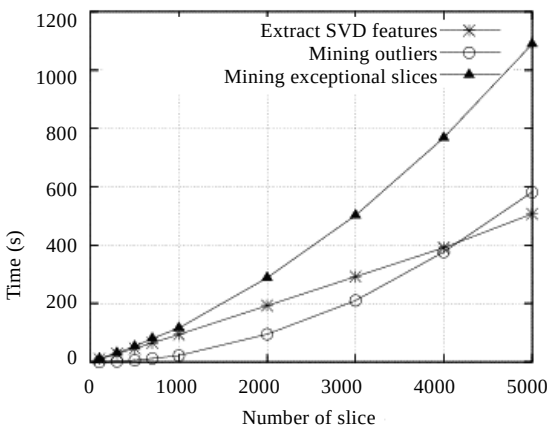



Fig.5 The runtime with respect to the number of slices

图 5 对切片数目的可扩展性

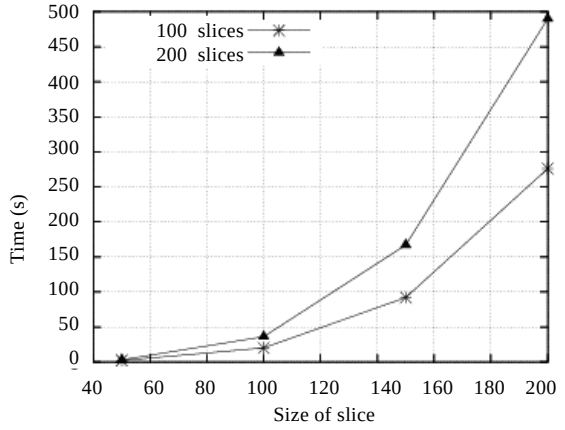



Fig.6 The runtime with respect to the size of slices

图 6 对切片大小的可扩展性

## 4 结 论

本文提出了一种基于奇异值分解的异常切片挖掘方法.我们抽取切片矩阵的奇异值向量作为它的数据分布特征,奇异值特征能够反映切片的内在代数本质,从而能够在此特征的基础上进行异常切片的挖掘.奇异值特征的稳定性保证了这种抽取特征的方法不受数据噪声和细微扰动的影响,可以捕获切片矩阵的基本代数结构.由于奇异值向量保留了原始切片之间的相对距离,因此,经过基于距离的孤立点检测,就可以找到异常的切片.

下一步工作是继续提高该算法的效率,解决其在特征提取和孤立点挖掘过程中的扩展性问题.此外,由于奇异值特征对转置、旋转、位移、镜像变换的不变性导致其在对应情形失去区分能力,我们也将扩展特征抽取的方法以解决这一问题.

致谢 数据库系统实验室的李希婷、梁晓等同学实现了本文部分算法的编程工作,张德辉博士生、姜力争博士生对本文的完成给予了很大帮助,在此一并表示感谢.

**References**:

[1] Imielinski T, Khachiyan L, Abdulghani A. Cubegrades: Generalizing association rules. In: Proc. of the 8th Int'l Conf. on Data Mining and Knowledge Discovery. Edmonton: ACM Press, 2002. 219−257.

[2] Lakshmanan VS, Pei J, Han JW. Quotient cube: How to summarize the semantics of a data cube. In: Proc. of the 28th Int'l Conf. on Very Large Data Bases. Hong Kong: Morgan Kaufmann Publishers, 2002. 778−789.

[3] Sarawagi S, Agrawal R, Megiddo N. Discovery-Driven exploration of OLAP data cubes. In: Proc. of the Int'l Conf. on Extending Database Technology. LNCS 1377, Springer-Verlag, 1998. 168−182.

[4] Sarawagi S. Explaining differences in multidimensional aggregates. In: Proc. of the 25th Int'l Conf. on Very Large Data Bases. Edinburgh: Morgan Kaufmann Publishers, 1999. 42−53.

[5] Sarawagi S. User-Adaptive exploration of multidimensional data. In: Proc. of the 26th Int'l Conf. on Very Large Data Bases. Cairo: Morgan Kaufmann Publishers, 2000. 307−316.

[6] Sathe G, Sarawagi S. Intelligent rollups in multidimensional OLAP data. In: Proc. of the 27th Int'l Conf. on Very Large Data Bases. Roma: Morgan Kaufmann Publishers, 2001. 531−540.

[7] Shi RC. Matrix Analysis. Beijing: Beijing Institute of Technology Press, 1996. 149−153 (in Chinese).

[8] Guha S, Gunopulos D, Koudas N. Correlating synchronous and asynchronous data streams. In: Proc. of the 9th Int'l Conf. on Knowledge Discovery and Data Mining. Washington DC: ACM Press, 2003. 529−534.

[9] Han JW, Kamber M. Data Mining: Concepts and Techniques. Beijing: High Education Press, 2001. 381−388.

[10] Knorr E, Ng R. Algorithms for mining distance-based outliers in large datasets. In: Proc. of the 27th Int'l Conf. on Very Large Data Bases. New York: Morgan Kaufmann Publishers, 1998. 392−403.

附中文参考文献:

[7] 史荣昌.矩阵分析.北京:北京理工大学出版社,1996.149−153.