

# 自适应快速最大信息系数算法实现

陈凯俊

(中国科学院 空间应用工程与技术中心, 北京 100094)

**摘 要:** 最大信息系数算法其本身的时间复杂度很高, 再加上处理对象的数据量很庞大, 导致其运算速率很慢, 为了解决这个问题, 提出了自适应的快速最大信息系数算法. 该算法是将原有的最大信息系数算法用多线程实现. 对于变量对的数量多少, 又提出了不同的并行策略. 并针对不同的计算机 CPU 的核数自适应地产生相应的线程数, 从而大大提高了算法的处理效率.

**关键词:** 最大信息系数; 自适应; 多线程; 快速; 大数据

**中图分类号:** TP391

**文献标识码:** A

**文章编号:** 1000-7180(2016)09-0070-04

## An Adaptive and Fast Algorithm Based on Maximal Information Coefficient

CHEN Kai-jun

(Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences, Beijing 100094, China)

**Abstract:** The algorithm based on Maximal Information Coefficient is a frontier algorithm in Big Data field. Since this algorithm has a very high complexity of time, it is very slow to process the big data. To solve this problem, the adaptive and fast algorithm based on Maximal Information Coefficient is proposed. This new algorithm which is realized through multithreading improves the speed of the data processing. For the different variables of quantity, a different parallel strategy was put forward. And the number of thread is based on the core of CPU through a tactic. As a result, this algorithm improve the efficiency of the original algorithm based on Maximal Information Coefficient dramatically.

**Key words:** maximal information coefficient; adaptation; multi-thread; fast; big data

### 1 引言

随着社会各领域的发展, 我们已经获得了庞大的不断增长的数据. 在这些庞大的数据量中寻找不同类型数据间的相关关系是一件很有意义的事情<sup>[1]</sup>. 自然界中的许多数据都是非线性的. 度量这些非线性数据相关性的强弱是一个重要的课题. 基于阈值相关<sup>[2]</sup>、互信息<sup>[3]</sup>、相位同步<sup>[4]</sup>等度量方法先后被提出.

现实中, 关系数据集中存在许多不同类型的关联. 而传统的方法只能通过一种模式观察一种信号, 在应用方面局限性比较大. 最大信息系数算法 (MIC) 是基于互信息技术提出的一种算法. 如果想

挖掘出两个数据项间的关系, 只要数据项是一个个数据集, 那么最大信息系数算法就能算出数据项间的关系强弱, 这为进一步的数据挖掘提供了很大的参考价值. 而且目前, 基于最大信息系数算法提供一种称为 MINE 的工具 (Maximal Information-based Nonparametric Exploration), 它能够发现不同类型数据之间的关系<sup>[5]</sup>. 图 1 展示了 MIC 捕获非线性关系的有效性. 随着噪声程度增加其所表现的健壮性. 图中颜色越深, 表示相关性强度越强. 该实验的测试用例包含了线性的以及非线性的典型图形, 并且每一类图形都有四张噪声影响逐渐递增的图形. 从图 1 中可以看出, 对于所有的这些图形, MIC 算法都能正确地算出它们的相关性强弱, 并且有较强

的抗噪声能力。

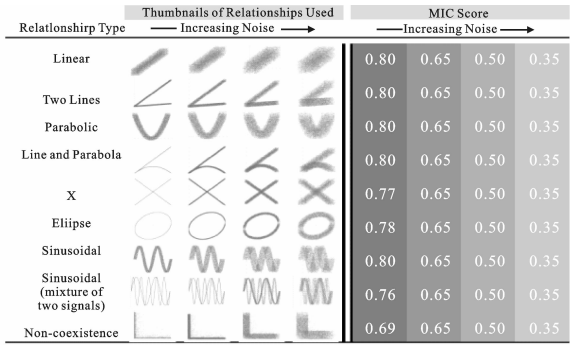


图 1 MIC 对不同关系给出的评价价值及 MIC 健壮性实验<sup>[6]</sup>

## 2 最大信息系数算法原理介绍

### 2.1 互信息的概念

我们用 shannon 熵来度量随机变量中所含的信息量, shannon 熵的定义如下:  $H(p) = -\sum_{i=1}^n p_i \log p_i$ , 其中  $p_i$  是事件  $i$  的发生概率,  $\sum_{i=1}^n p_i = 1$ . 两个随机变量之间的互信息为  $I(X, Y) = H(X) + H(Y) - H(X, Y)$ .

### 2.2 最大信息系数

**定义 1** 给定一个有限的有序数据集  $D : (x, y)$ , 把  $x$ - $y$  坐标轴平面划分成若干个小网格使得数据集  $D$  中的所有的点都落入网格  $G$  ( $G$  是将  $X$  轴划分为  $x$  列, 将  $Y$  轴划分为  $y$  行的网格) 中, 这里允许某几个网格内没有落入数据集  $D$  中的点. 令  $D|G$  为数据集  $D$  在网格  $G$  上的分布, 则基于数据集  $D$  的互信息为  $I(D, x, y) = \max\{I(D|G)\}$ .

其中,  $I(D|G)$  表示  $D|G$  的互信息. 这里是对所有可能的划分  $G$  取互信息的最大值.

**定义 2** 一个双变量的数据集的特征矩阵为  $M(D)_{x,y} = I(D, x, y) \cdot \log^{-1}(\min(x, y))$

**定义 3** 给定一个长度为  $n$  的双变量的数据集, 假设网格数的大小  $x * y$  小于  $B(n)$ , 那么这个数据集的最大信息系数 (MIC) 定义为

$MIC(D) = \max_{x,y < B(n)} M(D)_{x,y}$ , 式中,  $B(n) = n^c$ ;  $n$  为采样点的总数. 根据经验值,  $c$  取 0.6,  $c$  的数值也可以根据实际情况进行微调.

最大信息系数算法的关键是找出一种网格分割  $G$ , 使得落在坐标平面上的采样点在该网格下的分布能取到最大的特征矩阵  $M(x, y)$ .

### 2.3 MIC 的性质

(1) 对于所有无噪声并且非常数的函数关系, MIC 的值都为 1;

(2) 对于两个有确定函数关系的随机变量, 不论这个函数关系是什么样的, MIC 的值都是 1;

(3) 对于两个统计独立的随机变量 MIC 的值接近于 0;

### 2.4 MIC 算法

#### 2.4.1 MIC 算法原理

(1) 因为网格的划分有无数种, 实际情况中不可能针对每一种网格划分来计算互信息的特征矩阵. 因此该算法对网格划分进行简化. 首先均分纵坐标轴, 然后以这些均分线与数据集构成的曲线的交点做横坐标轴的垂直线, 最后从这些垂直线挑选出我们所要的横坐标轴的划分线.

(2) MIC 的算法是一种动态规划算法.  $I(X, Y) = H(X) + H(Y) - H(X, Y) = H(X) - H(X|Y) + H(Y)$ . 因为纵坐标的划分是固定的, 因此  $H(Y)$  是一个定值. 这样, 计算  $I(X, Y)$  的最大值也就是计算  $H(X) - H(X|Y)$  的最大值. 令  $F(m, l) = \max\{H(P) - H(P, Q)\}$ , 则该算法的状态转移方程为

$$F(m, l) = \max_{1 \leq i < m} \left( \frac{i}{m} \times F(i, l-1) - \frac{m-i}{m} \times H((i, m), Q) \right),$$

式中,  $m$  表示数据集的点数,  $i$  表示第  $i$  条垂直线前的点数;  $l$  表示分割线的数目.

边界条件是: 网格在横坐标轴上分割为两部分时的  $F(m, 2)$  值.

#### 2.4.2 MIC 算法步骤

(1) 对于输入的若干的数组 (长度相同且按时间顺序排序), 两两为一组构成一组数据集, 作为算法的输入;

(2) 将纵坐标均分为两部分, 在分割线与数据集构成的曲线 (由离散点组成的近似曲线) 的交点上做与横坐标垂直的垂线;

(3) 逐个选取这些垂线作为横坐标轴的分割线, 将横坐标分为两部分, 计算出对应的互信息特征矩阵, 并找到使特征矩阵最大的分割;

(4) 以步骤 (3) 的结果作为边界条件, 代入转移方程迭代, 计算出纵坐标分割数为 2 的时候对应网格的互信息系数的特征矩阵;

(5) 按照步骤 (3) 和 (4) 计算出均分纵坐标分别为 3 到  $B(m)/2$  个部分时对应的互信息特征矩阵;

(6) 比较所有计算得到的互信息特征矩阵, 找到最大的那个特征矩阵作为备选;

(7)将纵坐标与横坐标对调,并按照步骤(2)到(5),得到最大的特征矩阵.并与步骤(5)中得到的特征矩阵作比较,相对较大的特征矩阵即为我们要求的最大互信息系数.

### 3 自适应最大信息系数(MIC)算法多线程实现

#### 3.1 对于只求一个变量对的 MIC 值的并行策略

对于不同的纵坐标划分,算法都会计算出对应的互信息特征矩阵,而且这些互信息特征矩阵的计算是独立.因此可以并行计算每一种或若干种纵坐标划分的互信息特征矩阵.

因为纵坐标的划分有  $B(n)/2-1$  种( $n$  为数据集大小),当数据量很大时,会同时启动很多线程,反而会降低算法效率.这时候需要用到自适应策略,同时也可以让一个线程计算若干种纵坐标划分的互信息特征矩阵.

#### 3.2 对于多个变量两两间的 MIC 值的并行策略

MIC 算法在计算每个变量对之间的 MIC 值时,首先会均分  $Y$  轴为  $y$  行,  $1 < y < n^c/2$ , 然后以动态规划的思想寻找  $X$  轴上的最佳分割,求出在该分割下的最大特征矩阵;接着会均分  $X$  轴为  $x$  列,  $1 < x < n^c/2$ , 然后以同样的方法寻找  $Y$  轴上的最佳分割,求出在该分割下的最大特征矩阵.最后,这两个最大特征矩阵中最大的值就是 MIC.

对于多个变量两两间的 MIC 值的并行策略是把均分  $Y$  轴时的处理过程与均分  $X$  轴时的处理过程分为两个线程并行计算,同时,不同变量对之间的运算也是放在不同的线程进行并行计算.按照该方案,需要开启的线程数是固定的.

#### 3.3 自适应策略

在实际的多线程程序运行中,并不是线程数越多程序处理数据的效率就越高<sup>[7]</sup>.相反,过多的线程数会造成 CPU 在不同线程中频繁地切换,造成数据处理效率降低.基于以上考虑,本文在设计最大信息系数算法多线程实现时,采用了自适应计算机的 CPU 核数的方法,使进程运行时最大的线程数不超过 CPU 的核数.这样,算法在不同配置的计算机上运行,都会自适应地启动合适的线程数,以达到最优的运算效率.

程序读入数据后,会根据 CPU 的核数开启不超过核数个线程.而主线程则处在阻塞状态.一旦有子线程结束,主线程会扫描已完成的线程数,然后会开启相应数目的线程,直到开启完所有线程.这样

能够保证程序充分的利用 CPU 资源进行计算.自适应创建子线程的流程图如图 2 所示.

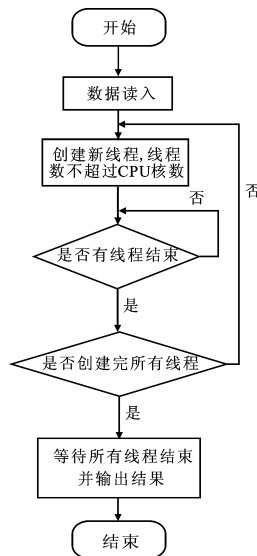


图2 主线程自适应分配子线程流程图

### 4 自适应快速最大信息系数算法实现的速率对比

将最大系数算法进行多线程并行化处理后,速率有了显著的提升.

#### 4.1 单个变量对并行化处理效果对比

测试样例的数据量共有八类,分别是1 000, 1 500, 2 000, 2 500, 3 000, 3 500, 4 000, 8 000 个.对于不同测试样例,分别记录普通 MIC 算法和多线程实现的快速 MIC 算法的数据处理时间.如图 3 所示,坐标图上方的曲线代表单线程的数据处理时间,坐标图下方的曲线代表多线程的数据处理时间.从图 3 中可以明显看出多线程实现的 MIC 算法比单线程的算法快很多.

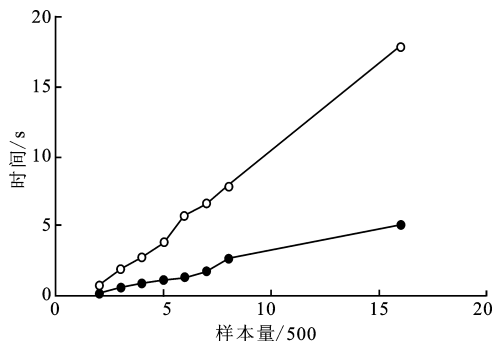


图3 单个变量对,变量数据量递增时数据处理速率对比

#### 4.2 多个变量对并行化处理效果对比

以下是针对两类不同情况进行的测试结果

图. 图4的测试样例为5个变量,每个变量初始为500个数据(记为 $5 \times 500$ ),之后数据量依次递增为 $5 \times 1000$ ,  $5 \times 1500$ ,  $5 \times 2000$ ,  $5 \times 2500$ ,  $5 \times 3000$ ,  $5 \times 3500$ ,  $5 \times 4000$ ,  $5 \times 8000$ . 对于不同测试样例分别记录普通MIC算法和多线程实现的快速MIC算法的数据处理时间. 如图4所示,位于上方的折线为普通MIC算法的数据处理时间,位于下方的为多线程实现的快速MIC算法的数据处理时间. 可以看出,多线程实现的快速MIC算法大大缩短了数据处理时间.

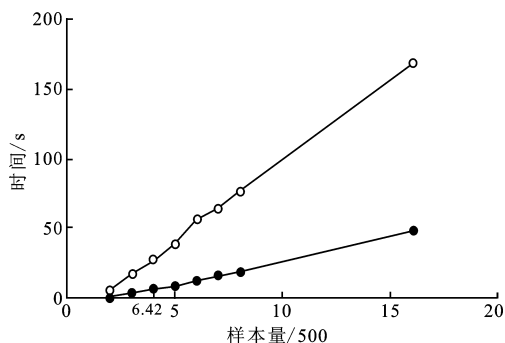


图4 变量个数不变,变量数据量递增时数据处理速率对比

图5的测试样例初始为5个变量,每个变量均包含1000个数据,之后变量个数依次递增为10、15、20、25、30. 如图5所示,位于上方的折线为普通MIC算法的数据处理时间,位于下方的为多线程实现的快速MIC算法的数据处理时间. 可以看出,多线程实现的快速MIC算法大大缩短了数据处理时间.

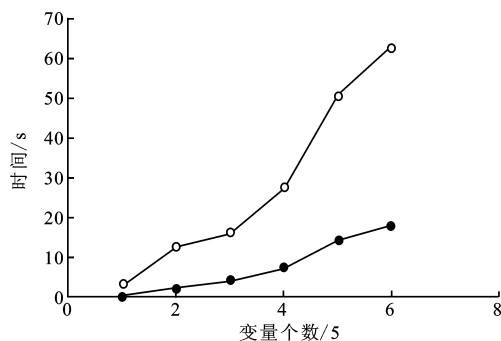


图5 变量数据量不变,变量个数递增时数据处理速率对比

## 5 结束语

MIC在关联性分析中有着较大的应用前景. 比如: 在生物医学方面,可以用来研究基因序列;在经济方面可以研究股票数据以及CPI数据与其他社会因素的关系;在航天研究方面,可以用来分析大量的空间数据,用来判断元部件是否异常;在科学研究方面,尤其是数学建模方面,可以通过计算关联性来筛选建模中需要的参数类型等. MIC算法还有着巨大的应用空间. 本文介绍的自适应快速最大信息系数算法的多线程实现方法,显著提高了MIC的计算速率. 这利于MIC算法在大数据领域上的应用推广.

在MIC算法的运算过程中会有大量的矩阵运算,例如熵的计算包含矩阵除法. 这种特点使得该算法能利用GPU进行协处理,从而进一步提升运算速率. 在大数据技术发展日新月异的今天,也可以尝试利用分布式处理的方法,例如spark计算框架来处理MIC算法.

## 参考文献:

- [1] David N R, Yakir A R, Hilary K F. Detecting novel associations in large datasets[J]. Science, 2011, 334(6062): 1518-1524.
- [2] Mokhtar B A. On a robust correlation coefficient[J]. The Statistician, 1990, 39(4):455-460.
- [3] 詹曙,李敏,徐甲甲,等. 局域化互信息度量的ACM下医学图像的分割[J]. 电子测量与仪器学报,2013,27(4):340-346.
- [4] Ernesto P, Rodrigo Q Q, Joydeep B. Nonlinear multivariate analysis of neurophysiological signals[J]. Progress in Neurobiology, 2005, 77(1):1-37.
- [5] 战泉茹. 基于最大信息系数的人脸特征选择[D]. 长春:东北师范大学,2013:7-13.
- [6] Reshef D N, Reshef Y A, Finucane H K, et al. Detecting novel associations in large data sets[J]. Science, 2011, 334(6062): 1518-1524.
- [7] 郭辉. 多线程的效率[J]. 计算机应用, 2008, 28(s2): 141-143.

## 作者简介:

陈凯俊 男,(1989-),硕士. 研究方向为大数据关联性挖掘.  
E-mail:736000798@qq.com.