doi:10.11887/j.cn.201905011 http://journal.nudt.edu.cn

卫星遥测数据相关性知识发现方法*

杨甲森[1,2],孟 新[1],王春梅[1]
(1. 中国科学院 国家空间科学中心 复杂航天系统电子信息技术重点实验室,北京 100190;
2. 中国科学院大学,北京 100049)

摘 要:为快速发现海量遥测数据中的相关关系,提出一种基于改进最大信息系数(Maximal Information Coefficient, MIC)的遥测数据相关性知识发现方法。以Mini Batch *K*-Means聚类算法为前驱过程对数据进行网格划分;计算该网格划分下的互信息,并以信息熵代替原有最大熵对互信息进行归一化矫正得到信息系数;选择不同网格划分下MIC作为变量相关性的测度。采用量子卫星遥测数据进行试验,结果表明:与基于动态规划算法的MIC方法相比,所提方法可有效解决MIC测度偏向多值变量的问题,时间复杂度从$O(n^{2.4})$下降为$O(n^{1.6})$,是一种适用于大规模遥测数据相关性分析的有效方法。

关键词:Mini Batch *K*-Means;信息熵;最大信息系数;遥测数据;相关性;量子卫星
中图分类号:V557.3 **文献标志码**:A **文章编号**:1001-2486(2019)05-071-08

Correlation knowledge discovery method for satellite telemetry data

YANG Jiasen[1,2], *MENG Xin*[1], *WANG Chunmei*[1]
(1. Key Laboratory of Electronics and Information Technology for Space Systems, National Space Science Center, Chinese Academy of Sciences, Beijing 100190, China; 2. University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: To discover correlations in massive telemetry data efficiently, a novel correlation knowledge discovery method based on the improved MIC (maximal information coefficient) was proposed. The Mini Batch *K*-Means clustering algorithm was used to discretize data in the precursor process; the mutual information between two variables under this partition was calculated and normalized by information entropy instead of maximal entropy to obtain the information coefficient; the MIC was selected as the measure of variable correlation. Afterwards, the method was applied to the correlation analysis of the quantum satellite telemetry data, and the results show that the proposed method can effectively solve the problem of MIC measure bias to multi-valued variables compared with the method based on dynamic programming algorithm, the time complexity dropped from $O(n^{2.4})$ to $O(n^{1.6})$, and it is an effective method for large-scale telemetry data correlation analysis.

Keywords: Mini Batch *K*-Means; information entropy; maximal information coefficient; telemetry data; correlation; quantum satellite

航天器遥测数据是地面运管系统判断其在轨运行状态的唯一依据[1]。作为典型复杂系统,航天器所包含的供配电、姿控、轨控、热控、有效载荷等分系统[2]之间、分系统内部各模块之间,均存在着大量电气、数据、热控接口以及复杂的系统交互,这就决定了反映星载设备状态的遥测数据之间普遍存在着相关关系。掌握这些相关关系对于实现多元遥测数据的关联检测、航天器异常事件原因的深层次挖掘以及特征选择和数据降维等都具有十分重要的意义。

围绕变量相关性,国内外学者提出了多种有效的相关性度量方法。其中Pearson相关系数[3]、最大信息压缩指数[4]、最小平方回归误差等被广泛用于度量变量间的线性相关,但难以刻画遥测变量间普遍存在的非线性相关;互信息[5]、信息增益比等测度虽然能够同时对线性相关和非线性相关进行度量,但这些测度依赖的概率密度函数的估计较为困难[5-6],且不具有普适性和等价性[7]。2011年,Reshef等[7]通过基于网格划分的互信息估计思想提出的最大信息系数(Maximal Information Coefficient, MIC),不但能够同时刻画变量间线性相关和非线性相关,而且对函数、超函数以及分段函数关系的度量都十分有效,具有较好的普适性和等价性,被广泛应用于复杂装备监测数据[8]、航天器遥测数据[9]的相关性分析。然而,在将MIC方法应用于高维度、大规

* **收稿日期**:2018-06-12
基金项目:中国科学院空间科学战略性先导专项资助项目(XDA04080201);中国科学院复杂航天系统电子信息技术重点实验室开放基金资助项目(N201708)
作者简介:杨甲森(1979—),男,山东冠县人,研究员,博士研究生,E-mail: jsy@nssc.ac.cn

模卫星遥测数据的相关性分析时，存在以下问题：①算法时间复杂度高，即使基于动态规划的近似算法，时间复杂度也达到 $O(n^{2.4})$[10]；②MIC 测度偏向于多值变量，取值较少变量的得分往往偏小。提高 MIC 算法的精度和性能历来是相关性研究的两大热点。文献[11]基于二次优化过程提高了 MIC 精度但却再度提高了复杂度；文献[12]采用图形处理单元和现场可编程门阵列组成的异构加速器提高了 MIC 计算的性能；文献[13]提供了一种基于并行处理的 MIC 快速计算跨平台工具；文献[14]基于 MapReduce 模型对 MIC 算法进行了并行化设计；文献[15]采用 C 语言重新实现了 MINE 套件。上述文献针对 MIC 算法性能提升的研究，多是围绕硬件环境、并行运算等算法实现的方式展开，并未改变其等深划分后基于动态规划算法寻优的本质。此外，已有文献尚无关于 MIC 测度偏向多值变量问题的讨论。

1 最大信息系数

1.1 MIC 算法原理

MIC 是以互信息为基础的，其主要思想是基于一种认识：如果两个变量存在相关关系，那么可以在两个变量构成的散点图上绘制网格，数据在网格中的分布情况可以反映二者之间的相关关系。网格绘制中，MIC 方法考虑两个因素：①网格划分的数量；②网格划分的方法。其原理可通过以下定义[7]来说明。

定义 1（特定划分下最大信息系数 $I(D,s,t)$） 对二维有序对数据集 $D(X,Y)$，分别在 X、Y 方向上进行 s、t 段划分。定义该划分对应的网格为 G，G 划分下 D 的概率分布为 $D|_G$、互信息为 $I(D|_G)$，那么此划分下的最大信息系数为：

$$I(D,s,t)=\max I(D|_G) \tag{1}$$

式中 max 的含义是指：将二维变量划分为 s、t 段的方法有多种，如等宽、等深划分[16]，每一种方法对应不同的互信息值，$I(D,s,t)$ 为其中的最大值。$I(D|_G)$ 计算方法如下：

$$I(D|_G)=\sum_{s\in X,t\in Y} p(s,t)\log_2\frac{p(s,t)}{p(s)p(t)} \tag{2}$$

式中，$p(s,t)$ 为联合概率密度，$p(s)$、$p(t)$ 为边缘概率密度。根据大数定理，当观测数据量足够大时，$p(s,t)$ 的计算可用落入网格中的样本数量占比来近似，$p(s)$、$p(t)$ 分别用落入 $(k,k+1)$ 和 $(l,l+1)$ 网格的样本数量占比来近似，其中 $k\in[0,s-1]$，$l\in[0,t-1]$。

定义 2（特征矩阵 $\boldsymbol{M}(D)$） 特征矩阵 $\boldsymbol{M}(D)$ 第 s 行、t 列元素的定义为：在 s，t 段网格划分下的最大信息系数（定义 1）的归一化矫正，如式(3)。s，t 为任意正整数，即 $\boldsymbol{M}(D)$ 是无限维矩阵。

$$\boldsymbol{M}(D)_{s,t}=\frac{I(D,s,t)}{\log_2(\min(s,t))} \tag{3}$$

定义 3（最大信息系数 $MIC(D)$） 设 D 的样本容量为 n，那么：

$$MIC(D)=\max_{st\leqslant B(n)}\{\boldsymbol{M}(D)_{s,t}\} \tag{4}$$

式中，$B(n)$ 为网格划分数量的上限，它将特征矩阵 $\boldsymbol{M}(D)$ 限定为有限维，文献[7]给出了 $B(n)$ 的推荐值为 $n^{0.6}$。

1.2 基于动态规划的 MIC 近似算法

在计算 $I(D,s,t)$ 时，为了避免网格穷举切割，进行遍历寻优，降低 MIC 方法计算的复杂度，文献[7]提出了一种基于动态规划算法来近似求解 MIC 的方法。详细步骤如下：

步骤 1：在 X 方向，以等值样本落入同一网格为原则，确定 X 方向 $s=2$ 段的等深划分 Q，如图 1 所示。等深的含义是指，划分后每一个分段的样本点数相差不多。

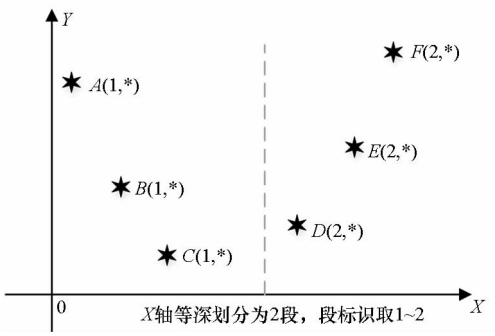


图 1 X 方向等深划分

Fig. 1 X direction equal partition

步骤 2：在 Y 方向，以等值样本落入同一网格为原则，确定 Y 方向候选划分 P，如图 2 所示。其方法是以 X 方向分割线与数据曲线（按 Y 值升序

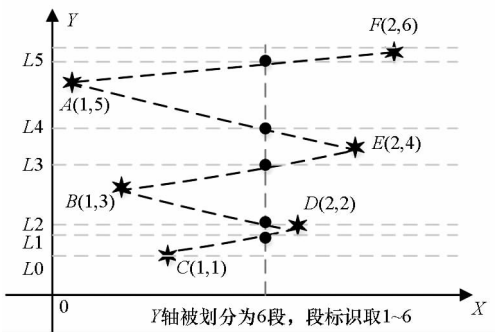


图 2 Y 方向候选划分

Fig. 2 Y direction candidate partition

顺序,逐点连接而成的线)的交点作 Y 轴的垂线,Y 方向等值样本若不在同一 X 划分 Q 中,则需增加一个 Y 方向的候选划分。

步骤 3:设 Y 方向候选划分 $P=\langle 0=c_0,\cdots,c_k=n\rangle$,其中 $c_0\leqslant c_1\leqslant\cdots\leqslant c_k$,第 j 行样本数量为 c_j-c_{j-1},k 为分段数量(为降低复杂度,设定 $k\leqslant\frac{cB}{s}$,c 为常量,如取 5 或者 15 等)。计算将前 c_r 个样本划分为 2 段的 $F(c_r,2)$ 值。F 定义如式(5),$H(\cdot)$ 为香农熵,$r\in[2,k]$。

$$F(c_r,2)=\max_{1\leqslant i<r}\{H(c_0,c_i,c_r)-H(Q,\langle c_0,c_i,c_r\rangle)\}\quad(5)$$

步骤 4:以 $F(c_r,2)$ 作为初始条件,代入动态规划算法的状态转移方程式(6)进行迭代,计算 $F(c_r,l)$ 的值,其中 $l\in[3,t]$,$r\in[l,k]$,$t\leqslant\frac{B}{s}$。

$$F(c_r,l)=\max_{c_{l-1}\leqslant c_i<c_r}\left\{\frac{c_i}{c_r}F(c_i,l-1)-\frac{c_r-c_i}{c_r}H(Q,\langle c_i,c_r\rangle)\right\}\quad(6)$$

$F(c_k,l)$ 是所有样本 l 分段下,$H(P)-H(Q,P)$ 的最大值。由于互信息 $I(Q;P)=H(Q)+H(P)-H(Q,P)$,在 X 方向 Q 划分不变的情况下,$H(Q)$ 为定值,因此 $F(c_k,l)+H(Q)$ 即为所有样本 l 分段下的最大互信息值。

步骤 5:计算 X 方向 $s=2$ 段等深划分下的特征矩阵 $\boldsymbol{M}_X(D)_{s,l}=\frac{[F(c_k,l)+H(Q)]}{\log_2(\min(s,l))}$,$l\in[2,t]$。

步骤 6:重复步骤 1 ~ 5,计算 $s\in\left[3,\frac{B}{2}\right]$ 段等深划分下的信息系数。

步骤 7:将纵坐标与横坐标交换,重复步骤 1 ~ 6,求得 Y 方向等深划分下的特征矩阵 $\boldsymbol{M}_Y(D)$。

步骤 8:以两个特征矩阵 $\boldsymbol{M}_X(D)$、$\boldsymbol{M}_Y(D)$ 中的最大元素作为两个变量的 MIC 值。

近似算法计算量主要集中在步骤 4,即动态规划算法求 F 值部分,该部分时间复杂度为 $O(k^2st)$[7],等深划分循环次数为 $B/2$,因此其整体复杂度为 $O(k^2stB)=O(k^2B^2)=O(n^{2.4})$。

2　Mini Batch *K*-Means 算法

Mini Batch K-Means 算法是 K-Means 算法的变种。与传统 K-Means 方法相比,Mini Batch K-Means 在每次迭代更新质心的过程中,采用随机抽样获得的数据子集进行更新。实践证明 Mini Batch K-Means 在数据量较大时,可以有效地减少算法收敛时间,但其准确度稍有下降[17]。

给定样本集 $Z=\{z_1,\cdots,z_n\}$,n 为样本容量。Z 将被划分为 k 个簇,簇的中心为 $C=\{c_1,\cdots,c_k\}$。Mini Batch K-Means 聚类的迭代步骤如下。

步骤 1:从 Z 中随机选择 k 个样本作为初始中心 C。

步骤 2:从 Z 中随机抽取容量为 b 的样本子集 $L=\{l_1,\cdots,l_j,\cdots,l_b\}$,组成一个 Batch。

步骤 3:对 L 中每一个样本点 l_j,计算其与 k 个簇类中心的相似度,将样本点 l_j 划入相似度最大的簇。

步骤 4:L 中所有样本经过步骤 3 后,根据各样本的簇标号重新计算聚类中心。

步骤 5:判断是否满足聚类结束的条件(如达到迭代次数 t),若未满足,回到步骤 2,否则进入步骤 6。

步骤 6:对 Z 中的每个样本点,根据其与 k 个聚类中心的相似程度,将其划分给相似度最大的簇。

Mini Batch K-Means 算法的主要计算量集中在步骤 3,即计算每个样本点的相似度并确定归属,其时间复杂度为 $O(knt)=O(n)$。

3　改进算法

3.1　改进的归一化因子

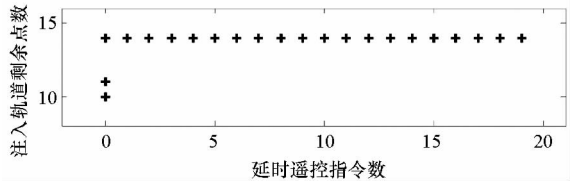
式(3)中,MIC 采用最大熵 $\log_2(\min(s,t))$ 对互信息进行归一化矫正。当变量取值较少,其数据分布只集中在少量网格中时,式(3)计算的 MIC 测度往往偏小,而这种取值较少的变量,在航天器遥测领域普遍存在。以量子卫星遥测参数“延时遥控指令数”“注入轨道剩余点数”为例,其取值如图 3 所示,二者采用式(3)方法计算的 MIC 值为 0.41,而皮尔逊相关系数达到了 0.95。

分析 MIC 取值偏小的原因为:归一化因子采用的最大熵,对应的是特殊的均匀分布,而不是变量的实际分布。本文以信息熵代替原有最大熵,作为归一化因子对互信息进行矫正,如式(7)所示,式中变量分布对分子、分母具有等同的贡献而被相互约减,可有效降低变量分布对 MIC 测度的影响。其中 Q 为 X 方向的 s 段划分,P 为 Y 方向的 t 段划分。

$$\boldsymbol{M}(D)_{s,t}=\frac{I(D,s,t)}{\min(H(Q),H(P))}\quad(7)$$

(a) 遥测时间序列

(a) Time series of telemetry



(b) 二维遥测分布

(b) Two-dimensional telemetry distribution

图 3 数据分布集中在少量网格

Fig. 3 Data is concentrated in a small number of grids

3.2 改进的 MIC 算法

3.2.1 相关定义

传统 Mini Batch *K*-Means 算法的初始聚类中心是随机选择的,聚类结果具有不可预见性[18]。本文定义 4、5,提出一种依据样本分布选择初始聚类中心的方法。

定义 4(一维序列统计信息) 给定遥测时间序列 $X=\{x_k\mid k=1,\cdots,n\}$,$X'=\{x'_k\mid k=1,\cdots,n\}$ 是 X 的升序排列,按索引顺序逐一访问 X' 中的元素,对取值相等元素的计数进行统计,可得一维序列统计信息 $S=\{\langle Value_i,Count_i\rangle\mid i=1,\cdots,m\}$,$\sum_{i=1}^{m}Count_i=n$。即:$X$ 包含 m 个不同取值,$Value_i$ 是其第 i 个取值,该值出现了 $Count_i$ 次。

定义 5(初始 k 个聚类中心) $S'=\{\langle Value'_i,Count'_i\rangle\mid i=1,\cdots,m\}$ 是一维序列统计信息 S 按 *Count* 的降序排列,则初始 k 个聚类中心 $C=\{c_1,\cdots,c_k\}=\{Value'_1,\cdots,Value'_k\}$。

3.2.2 小批量 *K* 均值 MIC 改进算法(MBKM_MIC)

MBKM_MIC 由两个阶段组成,如算法 1 所示,第一阶段(行 1 ~ 2)获取一维序列的统计信息,并按 *Count* 字段降序排列,最频繁的取值将被作为第二阶段初始的聚类中心;第二阶段(行 3 ~ 8)进行 MIC 值计算,主要包括 3 个步骤:①利用 Mini Batch *K*-Means 算法将 X 轴划分为 s 段、Y 轴划分为 t 段;②计算在 s、t 划分下的互信息值并进行归一化得到信息系数;③循环①、②遍历 $st\leqslant B$ 条件下所有划分的信息系数,选择极大值作为 MIC 值。

算法 1 MBKM_MIC 算法

Alg. 1 MBKM_MIC algorithm

输入: $D(X,Y)$ 为二维有序对样本集合,n 为样本容量,*bsize*、*time* 分别为 Mini Batch *K*-Means 算法随机抽样的 Batch 大小、迭代次数

输出: $MIC(D)$

//获取 X、Y 的一维序列统计信息并按 *Count* 字段降序排列

1. $S'_x\leftarrow$GetStatisticInfo(X,n)
2. $S'_y\leftarrow$GetStatisticInfo(Y,n)

// MIC 计算

3. While $st\leqslant B$
4. $P=\{x_1,\cdots,x_s\}\leftarrow$MBKM$(X,n,s,bsize,time,S'_x)$
5. $Q=\{y_1,\cdots,y_t\}\leftarrow$MBKM$(Y,n,t,bsize,time,S'_y)$
6. $\boldsymbol{M}(D)_{s,t}\leftarrow MI(D,P,Q,s,t)$
7. end while
8. return $MIC(D)=\max_{st\leqslant B}(\boldsymbol{M}(D)_{s,t})$

算法 1 调用了算法 2 ~ 4,描述如下。

算法 2 GetStatisticInfo 算法

Alg. 2 GetStatisticInfo algorithm

输入: X 为一维样本集合,n 为样本容量

输出: X 一维序列统计信息 S'(按 *Count* 字段降序排列)

1. $X'\leftarrow X$ 升序排列
2. $S\leftarrow\varnothing$
3. for each $i\in[1,n]$
4. $Count\leftarrow 1$
5. while$(x'_i=x'_{i+1}\,\&\&\,i+1\leqslant n)$
6. $Count\leftarrow Count+1$
7. $i\leftarrow i+1$
8. end while
9. $S\leftarrow S\cup\langle x'_i,Count\rangle$
10. $i\leftarrow i+1$
11. end for
12. $S'\leftarrow S$ 按 *Count* 字段降序排列
13. return S'

3.2.3 时间复杂度分析

第一阶段完成 4 个快速排序,时间复杂度为 $O(4n\log_2 n)$;第二阶段中,X、Y 方向的划分为 Mini Batch *K*-Means 算法,其复杂度为 $O(n)$,互信息计算需要遍历所有数据,其复杂度亦为 $O(n)$,故第二阶段复杂度为 $O(3nB)=O(n^{1.6})$。算法总复杂度为 $O(4n\log_2 n)+O(n^{1.6})=O(n^{1.6})$。

算法 3　MBKM 算法

Alg. 3　MBKM algorithm

输入: X 为一维样本集合,n 为样本容量,m 为划分的分段数,*bsize*、*time* 分别为 Mini Batch *K*-Means 算法随机抽样的 Batch 大小、迭代次数,S' 为 X 一维序列统计信息 S 按 *Count* 字段的降序排列

输出: X 的 m 段划分 P

$C=\{c_1,\cdots,c_m\}\leftarrow S'$初始 m 个聚类中心
for $i=1$ to *time* do//循环 *time* 次
　$CS_j\leftarrow\varnothing(1\leqslant j\leqslant m)$ //簇置空
　$CH=\{ch_1,\cdots,ch_{bsize}\}\leftarrow X$ 中随机抽取样本子集
　for $j=1$ to *bsize* do //Batch 样本簇归属
　　计算 ch_j 与聚类中心 $c_k(1\leqslant k\leqslant m)$ 的距离 d_{jk}
　　确定 ch_j 的簇标记:$\lambda_j\leftarrow\arg\min_{k\in[1,m]}d_{jk}$
　　将样本 ch_j 划入相应簇 $CS_{\lambda_j}\leftarrow CS_{\lambda_j}\cup ch_j$
　end for
　for each $CS_j(1\leqslant j\leqslant m)$//更新质心
　　$c_j\leftarrow\frac{1}{|CS_j|}\sum_{ch\in CS_j}ch$
　end for
end for
//确定所有样本簇归属
$CS_j\leftarrow\varnothing(1\leqslant j\leqslant m)$ //簇置空
for each $x_j\in X$
　计算 x_j 与聚类中心 $c_k(1\leqslant k\leqslant m)$ 的距离 d_{jk}
　确定 x_j 的簇标记:$\lambda_j\leftarrow\arg\min_{k\in[1,m]}d_{jk}$
　将样本 x_j 划入相应簇 $CS_{\lambda_j}\leftarrow CS_{\lambda_j}\cup x_j$
end for
根据 *CS* 获得 m 划分 $P=\{x_1,\cdots,x_m\}$
return P

算法 4　MI 算法

Alg. 4　MI algorithm

输入: $D(X,Y)$ 为二维有序对样本集合,P、Q 分别为 X、Y 的 s、t 段划分

输出: s、t 划分下的信息系数 $I(D,s,t)$

边缘及联合概率分布 $p(x_i)$、$p(y_j)$、$p(x_i,y_j)$ $(i\in[1,s],j\in[1,t])$
信息熵 $H(P)\leftarrow-\sum_{i=1}^{s}p(x_i)\log_2p(x_i)$
信息熵 $H(Q)\leftarrow-\sum_{j=1}^{t}p(y_j)\log_2p(y_j)$
联合熵 $H(P,Q)\leftarrow-\sum_{i=1}^{s}\sum_{j=1}^{t}p(x_i,y_j)\log_2p(x_i,y_j)$
互信息 $I(D,s,t)\leftarrow H(P)+H(Q)-H(P,Q)$
return $I(D,s,t)\leftarrow I(D,s,t)/\min(H(P),H(Q))$

4　实验结果与分析

为验证所提方法的有效性,进行了三组实验。第一组实验采用最大熵、信息熵两种归一化因子对遥测数据进行相关性分析,用于验证信息熵因子对取值较少变量的适用性;第二组实验用于验证 MBKM_MIC 方法的效能,分析该方法得分与动态规划方法得分的差异,评估所提方法的工程可用性;第三组实验,用于验证 MBKM_MIC 方法的处理效率,确认该方法对大规模数据相关性分析的适用性。

4.1　改进的归一化因子实验验证

采用量子卫星遥测数据,对改进前后归一化因子的应用效果进行试验。其中最大熵因子算法采用基于动态规划的 ApproxMIC[7] 算法,信息熵因子算法是在该算法 C 语言代码基础上进行的改进。

实验选用 3 组遥测数据,如表 1 所示(变量维数为所选时段方差不为 0 的遥测变量个数)。对每组数据中两两变量的 MIC 值进行计算,以阈值 0.8 作为相关性筛选条件。实验中发现的相关关系数量对比如图 4 所示。

表 1　三组量子卫星数据

Tab. 1　Three sets of quantum data

遥测数据	覆盖卫星时	变量维数	记录数
轨道实时包	2017 年 10 月 01 日 00 时—2017 年 10 月 03 日 00 时	23	10 780
姿控慢速包	2017 年 10 月 01 日 00 时—2017 年 10 月 01 日 08 时	46	14 398
平台实时包	2017 年 10 月 01 日 0 时—2017 年 10 月 01 日 16 时	40	14 398

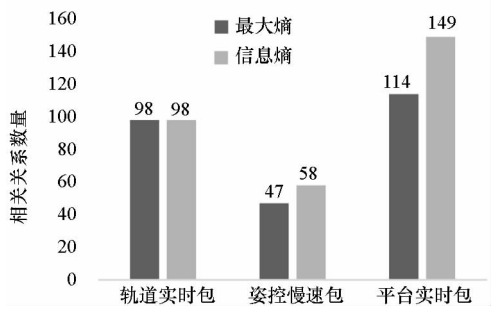


图 4　两种归一化因子发现的相关关系数量

Fig. 4　Number of correlation relationships found by two normalization factors

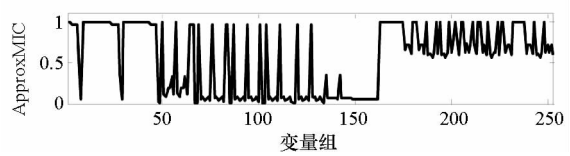
图 4 中,采用最大熵和信息熵归一化因子发现的相关关系总计分别为 259、305 组。经核验:①采用最大熵因子发现的相关关系是信息熵发现关系的子集;②同组变量的相关关系,信息熵因子计算的 MIC 值均高于最大熵因子;③多出的 46 组相关关系均为变量取值偏少导致最大熵因子下得分较小、信息熵因子下得分较大的案例,以第 3.1 节中“延时遥控指令数”“注入轨道剩余点数”为例,其信息熵归一化 MIC 取值达到了 0.94。因此,较之最大熵,采用信息熵归一化因子能够发现遥测数据中更为广泛的相关关系,对于取值较少变量的相关性分析,信息熵归一化因子仍具有良好的适用性。

4.2　MBKM_MIC 算法效能实验验证

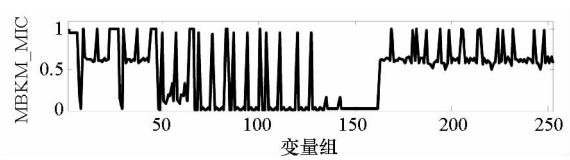
分别采用 MBKM_MIC、ApproxMIC 方法,对表 1 中量子卫星遥测数据进行了处理。MBKM_MIC 方法的实验参数设置为 $B = n^{0.6}$, $time = 3$, $bsize = 100$; ApproxMIC 参数设置为 $B = n^{0.6}$, $c = 15$。

三组数据的处理结果如图 5 ~ 7 所示,图中横轴为三组量子数据中两两组合的变量组索引,图(a)为 ApproxMIC 计算结果,图(b)为 MBKM_MIC 方法计算结果,图(c)为两个测度取值的偏差。三组数据两个测度取值的平均绝对误差(假设 ApproxMIC 方法取值为标准值)分别为 0.103、0.035、0.034,两种方法取值趋势基本一致,误差也非常小。

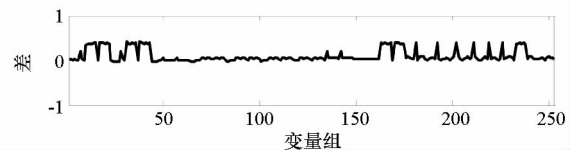
从图 5 ~ 7 可见,部分变量组的 MBKM_MIC 方法测度取值低于 ApproxMIC 方法,这是由 MBKM_MIC 直接网格划分获得测度取值,而 ApproxMIC 经历了动态规划寻优过程导致的;另有部分变量组的 MBKM_MIC 取值高于 ApproxMIC 方法,分析是由二者分别采用信息熵和最大熵作为归一化因子引起的。



(a) ApproxMIC



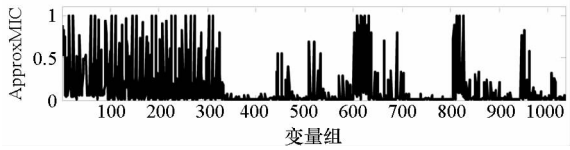
(b) MBKM_MIC



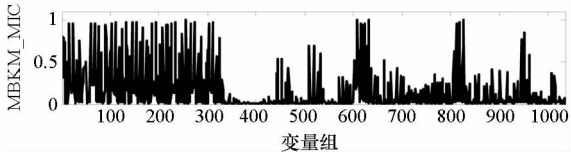
(c) ApproxMIC – MBKM_MIC

图 5　轨道实时包处理结果比较

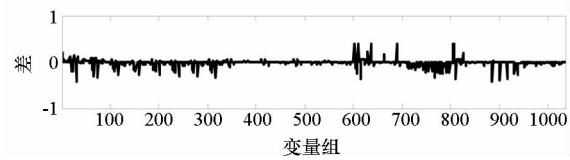
Fig. 5　Comparison of real-time orbit package processing results



(a) ApproxMIC



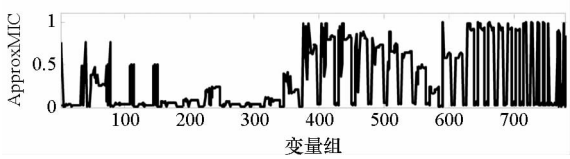
(b) MBKM_MIC



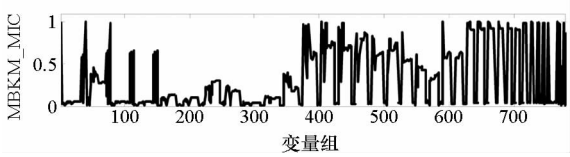
(c) ApproxMIC – MBKM_MIC

图 6　姿控慢速包处理结果比较

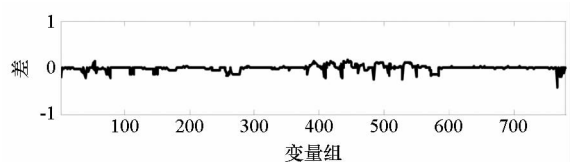
Fig. 6　Comparison of slow attitude package processing results



(a) ApproxMIC



(b) MBKM_MIC



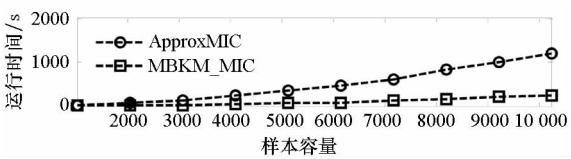
(c) ApproxMIC – MBKM_MIC

图 7　平台实时包处理结果比较

Fig. 7　Comparison of real-time platform package processing results

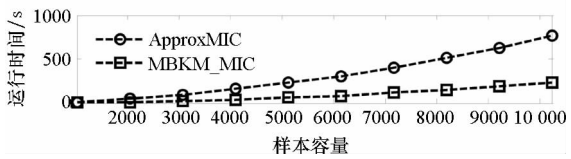
4.3　MBKM_MIC 算法效率实验验证

采用与第 4.2 节实验相同的参数设置,对三组量子数据的处理时间进行统计,实验共进行了 10 次,分别对应样本容量 n 为 1024 ~ 10 240 ,间隔为 1024,性能测试结果如图 8 所示。



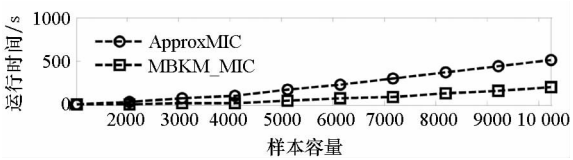
(a) 轨道实时包

(a) Real-time orbit package



(b) 姿控慢速包

(b) Slow attitude package



(c) 平台实时包

(c) Real-time platform package

图 8　处理性能比较

Fig. 8　Comparison of processing performance

由图 8 可知,对于遥测数据的相关性分析,MBMK_MIC 方法的性能普遍优于 ApproxMIC,且样本规模越大,性能优势越明显。

5　结论

遥测数据相关性发现在卫星数据分析、故障诊断中具有关键的“导航”作用。从高维的航天器遥测数据中挖掘出具有可解释性的相关性知识,可以快捷、高效地发现星载设备间内在的关联关系。本文针对 MIC 方法应用于大规模遥测数据相关性分析过程中,处理性能较低、偏向于多值变量的问题,提出了改进归一化因子、以 Mini Batch K-Means 聚类为前驱过程的改进 MIC 方法。试验结果表明,改进的归一化因子方法对取值较少的变量也具有良好的适用性;改进的 MIC 方法,在计算结果与 ApproxMIC 测度偏差可承受的前提下,显著提升了数据处理的性能,是一种适用于大规模遥测数据相关性分析的有效方法。

参考文献(References)

[1] 彭喜元, 庞景月, 彭宇, 等. 航天器遥测数据异常检测综述[J]. 仪器仪表学报, 2016, 37(9): 1929 – 1945.
PENG Xiyuan, PANG Jingyue, PENG Yu, et al. Review on anomaly detection of spacecraft telemetry data [J]. Chinese Journal of Scientific Instruments, 2016, 37(9): 1929 – 1945. (in Chinese)

[2] 谭春林, 胡太彬, 王大鹏, 等. 国外航天器在轨故障统计与分析[J]. 航天器工程, 2011, 20(4): 130 – 136.
TAN Chunlin, HU Taibin, WANG Dapeng, et al. Analysis on foreign spacecraft in-orbit failure [J]. Spacecraft Engineering, 2011, 20(4): 130 – 136. (in Chinese)

[3] Pearson K. Notes on the history of correlation [J]. Biometrika, 1920, 13(1): 25 – 45.

[4] Mitra P, Murthy C A, Pal S K. Unsupervised feature selection using feature similarity[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(3): 301 – 312.

[5] Moon Y I, Rajagopalan B, Lall U. Estimation of mutual information using kernel density estimators [J]. Physical Review E, 1995, 52(3): 2318 – 2321.

[6] Kraskov A, Stogbauer H, Grassberger P. Estimating mutual information [J]. Physical Review E, 2004, 69(6): 066138.

[7] 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.

[8] 曾令男, 丁建伟, 赵炯, 等. 基于互信息的复杂装备高维状态监测数据相关性发现与建模[J]. 计算机集成制造系统, 2013, 19(12): 3017 – 3025.
ZENG Lingnan, DING Jianwei, ZHAO Jiong, et al. Detecting and modeling for associations between high-dimension condition monitoring data of complex equipment based on mutual information [J]. Computer Integrated Manufacturing System, 2013, 19(12): 3017 – 3025. (in Chinese)

[9] 王鹏, 张善从. 基于最大信息系数的时延数据相关性分析方法[J]. 电子测量技术, 2015, 38(9): 112 – 115.
WANG Peng, ZHANG Shancong. Method for the correlation analysis of data with time delay based on maximal information coefficient [J]. Electronic Measurement Technology, 2015, 38(9): 112 – 115. (in Chinese)

[10] 邵福波. 基于最大信息系数改进算法及其在铁路事故分析中的应用[D]. 北京: 北京交通大学, 2016: 36.
SHAO Fubo. The improved algorithm of maximal information coefficient and its applications in railway accident analysis [D]. Beijing: Beijing Jiaotong University, 2016: 36. (in Chinese)

[11] Wang S L, Zhao Y P, Shu Y, et al. Improved approximation algorithm for maximal information coefficient [J]. International Journal of Data Warehousing and Mining, 2017, 13(1): 76 – 93.

[12] Wang C, Li X, Wang A L, et al. Brief announcement: MIC + +: accelerating maximal information coefficient calculation with GPUs and FPGAs [C]// Proceedings of the 28th ACM Symposium on Parallelism in Algorithms and Architectures, 2016: 287 – 288.

[13] Tang D P, Wang M W, Zheng W F, et al. RapidMic: rapid computation of the maximal information coefficient [J].

Evolutionary Bioinformatics, 2014, 10(10): 11 – 16.

[14] 吕瑞, 蔡国永, 裴广战. 基于MapReduce的MIC算法并行化[J]. 计算机科学, 2015, 42(11): 80 – 83.
LYU Rui, CAI Guoyong, PEI Guangzhan. Parallelization of MIC algorithm based on MapReduce [J]. Computer Science, 2015, 42(11): 80 – 83. (in Chinese)

[15] Albanese D, Filosi M, Visintainer R, et al. Minerva and minepy: a C-engine for the MINE suite and its R, Python and MATLAB wrappers [J]. Bioinformatics, 2013, 29(3): 407 – 408.

[16] 郑西西. 基于关联规则的火电厂优化目标值确定的研究[D]. 北京: 华北电力大学, 2011: 21.
ZHENG Xixi. Research on the thermal power plant operation optimization value determining base on association rule [D]. Beijing: North China Electric Power University, 2011: 21. (in Chinese)

[17] 陈涛, 云利军, 程飞燕, 等. 复杂背景下给予YCbCr颜色空间和Mini-Batch聚类的肤色检测[J]. 云南师范大学学报(自然科学版), 2017, 37(5): 27 – 33.
CHEN Tao, YUN Lijun, CHENG Feiyan, et al. Skin color detection based on YCbCr color space and Mini-Batch clustering in complex background [J]. Journal of Yunan Normal University(Natural Sciences Edition), 2017, 37(5): 27 – 33. (in Chinese)

[18] Dong X, Pi D C. An effective method for mining quantitative association rules with clustering partition in satellite telemetry data [C]//Proceedings of Second International Conference on Advanced Cloud and Big Data, 2014: 26 – 33.