·图形图像处理· 文章编号：1000-3428(2016)01-0231-06 文献标识码：A 中图分类号：TP393

基于最小描述长度的图分割结构检测改进算法

魏长宝，姚汝贤

（黄淮学院信息工程学院，河南 驻马店 463000）

摘 要： 针对现有图分割变化检测（GPCD）算法中易出现重复分割及忽略图形变化成本的不足，利用概率树表示图分割结构的概率模型。将 GPCD 问题转化为基于最小描述长度的树变化检测问题，利用树算法来求解 GPCD 问题。实验结果表明，在考虑变化成本的情况下，与 GraphScope 基准算法相比，TREE 算法具有较低的虚警率和较高的检测精度。

关键词： 图分割变化检测；最小描述长度；概率树；变化成本；虚警率

中文引用格式： 魏长宝，姚汝贤. 基于最小描述长度的图分割结构检测改进算法[J]. 计算机工程，2016，42(1)：231-236，242.

英文引用格式： Wei Changbao，Yao Ruxian. Improved Algorithm of Graph Partitioning Structure Detection Based on Minimum Description Length[J]. Computer Engineering，2016，42(1)：231-236，242.

Improved Algorithm of Graph Partitioning Structure Detection Based on Minimum Description Length

WEI Changbao，YAO Ruxian

（School of Information Engineering，Huanghuai University，Zhumadian，Henan 463000，China）

【Abstract】 Aiming at the disadvantages of the existing Graph Partitioning Change Detection（GPCD） algorithm like repeated segmentation and ignoring change cost of images，it employs probabilistic trees to represent probabilistic models of graph partitioning structures. Then reduce GPCD into the issue of detecting changes of trees on the basis of the Minimum Description Length（MDL） principle. It proposes TREE algorithm for solving the GPCD problem. Simulation experimental results show that，by taking the cost of changes into consideration，TREE realizes significantly less False Alarm Rate（FAR） for change detection than the baseline method called GraphScope. And it is able to detect changes more accurately than GraphScope.

【Key words】 Graph Partitioning Change Detection（GPCD）；Minimum Description Length（MDL）；probabilistic tree；cost of change；False Alarm Rate（FAR）

DOI： 10.3969/j.issn.1000-3428.2016.01.041

1 概述

图分割技术[1-3]（也称图聚合问题）是一种按照某种标准将图分成若干子模块的方法，可以使得分割后各子模块中节点联系更紧密，模块间联系更低。该技术在蛋白质网络、网络数据挖掘等许多领域都有广泛应用。将二分图时间序列的图分割结构检测问题称为图形分割变化检测（Graph Partitioning Change Detection，GPCD）问题。根据链路关系，图分割结构可看成图节点的聚类结构，这一图分割过程将会产生多个网络社区。因此，它有助于社区结构变化检测及图分割结构变化检测，而社区结构变化往往对应于真实世界中的重要世界，因此研究 GPCD 问题具有重要意义[4-5]。

文献[6-7]基于最小描述长度（Minimum Description Length，MDL）原则，从无损数据压缩角度为选择最优分割提供了一种准则。该准则认为，如果有种分割策略，进行图形编码时所需总码长及相对数据量最小，则该分割策略即为最优分割。为此，文献[8-9]提出基于 MDL 的静态二分图分割方法和普通图分割方法。对于动态图分割，文献[10]提出 GraphScope 图形变化检测方法。该方法根据 MDL 准则将二分图分割为一组子图，以便使数据的码长及图形分割码长之和最小，通过进行分割是否发生

基金项目： 河南省科技攻关计划基金资助项目（122102210430）。

作者简介： 魏长宝（1972－），男，副教授、硕士，主研方向为图像处理、智能信息处理；姚汝贤，副教授、硕士。

收稿日期： 2014-12-24 **修回日期：** 2015-02-07 **E-mail：** 2525100689@qq.com

变化的假设性检验实现图形分割结构的变化检测。GraphScope 是 GPCD 问题的有效求解算法,但是存在如下缺陷:(1)初始图必须是节点分割的直积,图形分割结构的表示非常有限,本文将这一现象称为基于直积的分割策略。当描述具有分层特征的图分割结构时会出现大量参数,比如进行一次分割后,对每个被分割的子图还需再一次分割,依次类推。(2)没有考虑图形分割结构的动态特点。实际上,该方法也没有考虑分割转换概率模型。因此,图形变化的成本被忽略,导致出现虚警。

为此,本文提出 GPCD 问题的一种新的求解算法——树(TREE)算法,以克服 GraphScope 算法的上述缺陷。主要工作如下:(1)基于树进行图分割。TREE 算法利用基于树的概率模型来表示图形分割结构,本文称其为基于树的分割方法,GPCD 问题可转化为树结构的变化检测问题。与基于直积的分割方法相比,基于树的分割问题可使本文使用较少参数来表示分层分割图。(2)GPCD 问题的动态模型选择。通过引入图分割之间的转换概率,以考虑基于树的图分割的动态特性。图分割变化成本可定义为转换概率的码长。将动态模型选择(DMS)理论[11]应用于图分割序列选择过程中,即通过选择一组图分割,可使数据的码长与变化成本之和最小。根据数据拟合和变化复杂度间的平衡情况确定最优序列。

2 图形分割变化检测

假设图形序列为:

$$\Psi=\{G_1,G_2,\cdots,G_t,\cdots\}$$

其中,每个二分图 $G_t(t=1,2,\cdots)$ 有 m 个发送节点和 n 个接收节点。更准确地说,每个 G_t 可表示为一个 $m\times n$ 矩阵,第 (i,j) 个元素 g_{ij} 表示第 i 个发送方至第 j 个接收方的链接。本文假设 m 和 n 固定。将图形序列 Ψ 分割为如下一组图形子序列:

$$\Psi=\{\Psi^1,\Psi^2,\cdots,\Psi^s,\cdots\}$$

其中,每个 Ψ^s 表示从 t_s 至 $t_{s+1}-1$ 的一个图形序列:

$$\Psi^s=\{G_{t_s},G_{t_s+1},\cdots,G_{t_{s+1}-1}\}$$

其中,t_s 称为变化点;Ψ^s 称为一个分段($s=1,2,\cdots$)。将图 G 的分割看成图 G 分解为一组连接子图 $\{G(u)\}$ 的过程,$G=\cup_u G(u)$。可得 GPCD 问题定义如下:已知一个图形序列 $\Psi=G_1,G_2,\cdots$,根据 $G_1,G_2,\cdots$的分割序列检测出 Ψ 中的变化点及其相应变化。

3 基于树的 GPCD 问题

3.1 基于树分割的概率模型

利用二分树实现基于树的分割表示。在二分树中,每个树节点关联一组发送节点和接收节点。为了区分树中节点和图中节点,将前者称为树节点,后者称为节点。对每个树节点,其子树节点的分配有 2 种情况:(1)2 个子树节点的发送节点集合相同,但是接收节点集合分离。(2)2 个子树节点的接收节点集合相同,但是发送节点集合分离。

例如,如果图 G 的矩阵表示如式(1)所示,则图 G 可由图 1 中的二分树确定。

$$G=\left(\begin{array}{cc}\begin{bmatrix}1&2\\2&1\end{bmatrix} & \begin{bmatrix}5&3\\4&7\end{bmatrix}\\ \begin{bmatrix}9&8\\12&9\end{bmatrix} & \begin{bmatrix}6&3\\7&4\end{bmatrix}\end{array}\right) \tag{1}$$

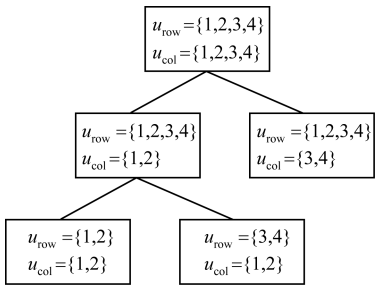

图 1 G 基于树的分割过程

接收方节点集合可分为{1,2}和{3,4},然后对于具有前一集合的子树节点,接收节点集合可分为{1,2}和{3,4}。

已知树 M,设 $u=(u_{row},u_{col})$ 表示树 M 中叶子 u 的行节点和列节点组成的元组。设 $G(u)$ 表示 u 的相应子图。此时,可以获得一种基于数的分割 $G=\cup_u G(u)$。

设 M_t 表示 G_t 的对应树,二分树序列可表示如下:

$$\Theta=\{M_1,M_2,\cdots,M_t,\cdots\}$$

Ψ 的分割问题转化为 Θ 的分割问题,如图 2 所示,将 Θ 的分割表示为 $\Theta=\{\Theta^1,\Theta^2,\cdots\}$。对每个 s,任意 $M,M'\in\Theta^s$ 有 $M=M'$。如果 $G_t\in\Psi^s$,则 $M_t\in\Theta^s$。也就是说,一个分段只对应于一个树。

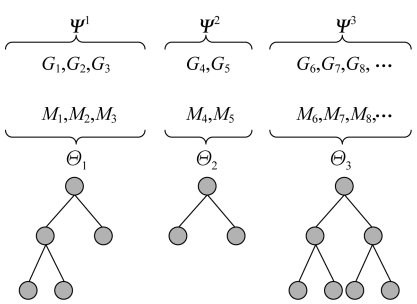

图 2 基于树的 GPCD

设 M 表示图 G 相应的二分树。M 的概率分布定义可表示如下:对树 M 中的叶子 u,$G_{r,c}(u)$ 表示 u 相应子图中第 r 行第 c 列的元素。假设根据参数为 $\lambda(u)$ 的泊松分布生成 $G_{r,c}(u)$,并将其表示为

$f(G_{r,c}(u)\mid\lambda(u))$,于是有:

$$f(G_{r,c}(u)\mid\lambda(u))\overset{\text{def}}{=}\mathrm{e}^{-\lambda(u)}\frac{\lambda(u)G_{r,c}(u)}{G_{r,c}(u)!}\quad(2)$$

3.2　动态模型选择

假设已知图形序列 $\Psi=G_1,G_2,\cdots,G_T$(T:数据规模)及基于树的分割序列 $\Theta=M_1,M_2,\cdots,M_T$,其中 M_i 会随着时间而变化。假设每个码字不是另一个码字的前缀,在此前提下考虑 ψ 和 Θ 的编码问题。本文引入动态模型选择(Dynamic Model Selection,DMS)准则[8]来衡量已知 ψ 时 Θ 的质量。将其定义为对 ψ 和 Θ 进行编码所需要的总码长 $L(\Psi;\Theta)$。

设 $P(G_t\mid M_t)$表示已知 M_t 时 G_t 的概率。引入从 $M^{t-1}=M_1M_2\cdots M_{t-1}$到 M_t 的转换概率 $P(M_t\mid M^{t-1})$。于是,DMS 准则 $L(\Psi;\Theta)$表述如下:

$$L(\psi;\Theta)\overset{\text{def}}{=}\sum_{t=1}^{T}(-\mathrm{lb}P(G_t\mid M_t))+\sum_{t=1}^{T}(-\mathrm{lb}P(M_t\mid M^{t-1}))\quad(3)$$

其中,M_0 已知;式(3)右侧第 1 项表示 ψ 相对于 Θ 的码长;第 2 项表示 Θ 本身码长。DMS 可看成是将传统的基于 MDL 的静态模型选择拓展为一种模型序列选择[11]。

3.3　基于树的图形分割编码方法

为计算式(3)右侧第 1 项和第 2 项,M_t 已知时 G_t 的码长可计算为归一化最大似然(Normalized Maximum Likelihood,NML)码长,该码长定义为归一化最大似然的负对数,如下式所示:

$$-\mathrm{lb}P(G_t\mid M_t)=\sum_{u\in leaf(M_t)}(-\mathrm{lb}P_{\mathrm{NML}}(G_t(u)\mid M_t))\quad(4)$$

其中,$leaf(M_t)$表示 M_t 的一组叶节点;$G_t(u)$表示第 u 个叶节点的子图。设 $G_{t;r,c}(u)$表示 $G_t(u)$的第(r,c)个元素。u_{row}和 u_{col}分别表示树节点 u 的行节点集合和列节点集合。于是,P_{NML}表示归一化最大似然分布,定义如下:

$$P_{\mathrm{NML}}(G_t(u)\mid M_t)\overset{\text{def}}{=}\frac{\prod_{r\in u_{\mathrm{row}}}\prod_{c\in u_{\mathrm{col}}}f(G_{t;r,c}(u)\mid\hat{\lambda}(u))}{\sum_{G_t(u)}\prod_{r\in u_{\mathrm{row}}}\prod_{c\in u_{\mathrm{col}}}f(G_{t;r,c}(u)\mid\hat{\lambda}(u))}\quad(5)$$

其中,$\hat{\lambda}(u)$表示 λ 基于 $G_t(u)$的最大似然估计;u 固定时对 $G_t(u)$所有可能值求出分母之和。文献[6]指出,因为 NML 码长实现了 Shtarkov 极小极大遗憾准则,所以 NML 码长最优。本文将 NML 码长 $-\mathrm{lb}P(G_t\mid M_t)$称为 G_t 的随机复杂性,并将其表示为 SC_t。于是有:

$$SC_t\overset{\text{def}}{=}-\mathrm{lb}P(G_t\mid M_t)$$

式(5)右侧的分母难以进行解析计算,根据 Rissanen 方程[12]进行近似计算:设 $I(\lambda)$表示定义为 $I(\lambda)\overset{def}{=}E[-\partial^2\mathrm{lb}P(G\mid\lambda)/\partial^2\lambda]$的 Fisher 信息矩阵,$G_t$ 的随机复杂性有如下等式:

$$\begin{aligned}-\mathrm{lb}P_{\mathrm{NML}}(G_t(u)\mid M_t)&=-\mathrm{lb}f(G_t(u)\mid\hat{\lambda}(u))+\frac{1}{2}\mathrm{lb}\frac{|G_t(u)|}{2\pi}+\mathrm{lb}\int\sqrt{|I(\lambda)|}\,\mathrm{d}\lambda\\&=\sum_{r\in u_{\mathrm{row}}}\sum_{c\in u_{\mathrm{col}}}(-\mathrm{lb}f(G_{t;r,c}(u)\mid\hat{\lambda}(u)))+\frac{1}{2}\mathrm{lb}\frac{|G_t(u)|}{2\pi}+(1+\frac{a}{2})\mathrm{lb}2+\mathrm{lb}*a\end{aligned}\quad(6)$$

其中,a 表示最小整数;$\hat{\lambda}\in[0,2^a]$且 $\mathrm{lb}*a=\mathrm{lb}a+\mathrm{lblb}a+\cdots+\mathrm{lb}b$ 表示对 a 编码时需要的码长,此时对所有正项求和,且 $\mathrm{lb}b\approx2.865\ 064$。如果数据范围无限,则 Fisher 信息发散,因此本文将其限制在范围$[0,2^a]$内以便使整数有限。将式(6)代入式(4),有:

$$\begin{aligned}SC_t&=-\mathrm{lb}P(G_t\mid M_t)\\&=\sum_{u\in leaf(M_t)}\{\sum_{r\in u_{\mathrm{row}}}\sum_{c\in u_{\mathrm{col}}}(-\mathrm{lb}f(G_{t;r,c}(u)\mid\hat{\lambda}(u)))+\frac{1}{2}\mathrm{lb}\frac{|G_t(u)|}{2\pi}+(1+\frac{a}{2})\mathrm{lb}2+\mathrm{lb}*a\}\end{aligned}\quad(7)$$

3.4　树转换的编码方法

下文给出式(3)中第 2 项的计算方法。设 $\alpha>0$ 表示一维参数,$N_o(M_t)$表示 M_t 中叶节点的数量,$N_1(M_t)$表示 M_t 中内部树节点的数量,且:

$$P_0\overset{\text{def}}{=}\frac{N_o(M_t)}{N_o(M_t)+N_1(M_t)},P_1\overset{\text{def}}{=}\frac{N_1(M_t)}{N_o(M_t)+N_1(M_t)}$$

用 $int(M_t)$表示 M_t 的内部树节点集合,m_u 表示在树节点 u 处可被分割的树节点数量。转换概率定义如下:对 $t\geqslant1$,有

$$P(M_t\mid M^{t-1};\alpha)\overset{\text{def}}{=}\begin{cases}\alpha P_1^{N_1(M_t)}P_0^{N_0(M_t)}\prod_{u\in int(M_t)}\dfrac{1}{2m_u\cdot m_u!} & M_t\neq M_{t-1}\\1-\alpha & M_t=M_{t-1}\end{cases}\quad(8)$$

该转换定义表明,树节点不发生变化的概率为 $1-\alpha$,发生变化的概率为 α。在后一种情况下,根据概率 $P_1^{N_1(M_t)}P_0^{N_0(M_t)}\prod_{u\in int(M_t)}\dfrac{1}{2m_u\cdot m_u!}$选择 M_t。因此,树转换的码长为:

$$-\mathrm{lb}P(M_t\mid M^{t-1};\hat{\alpha})\quad(9)$$

此时,每次转换时树编码方法与 Rissanen 方法[12]吻合,于是有:

$$\mathscr{L}(u)\overset{\text{def}}{=}\mathrm{lb}2+\mathrm{lb}m+\mathrm{lb}m!\quad(10)$$

GraphScope 算法没有考虑 $\hat{\alpha}$ 的转换成本。本文

考虑该成本。在式(9)中,$\hat{\alpha}$ 是 α 基于 $M^{t-1}=M_1M_2\cdots M_{t-1}$ 的估计。利用文献[13]中 Krichevsky 和 Trofimov 的估计值来估计 α。KT 估计值为:

$$\hat{\alpha}=(n(M^{t-1})+1/2)/t \tag{11}$$

其中,$n(M^{t-1})$表示 M^{t-1}中树的变化总量。

3.5 基于树的 GPCD 问题

本文提出的 TREE 算法总体流程如下:

(1)通过分割和融合操作构建一个树序列。本文利用图序列构建树序列。树序列包括多个子序列,每个子序列称为一个段。假设在同一分段内的所有树相同,且对不同的时间分段,一个段中的树与相邻段中的树不同。为了构建每个段中的树,对节点进行分配,对树进行分割或修剪操作以使随机复杂性最小。

算法 1 节点分配

已知:图形子序列 Ψ^s,对叶节点 u,有一组行节点 $u_{1\text{row}},u_{2\text{row}}$

步骤 1 设置 $R=u_{1\text{row}}\cup u_{2\text{row}}$。

步骤 2 设置 $x(\in R)$为一个列节点,设置 $R\leftarrow R-\{x\}$。

步骤 3 计算:

$$\lambda_x=\frac{\sum_{t=t_s}^{t_{s+1}-1}\sum_{c\in u_{\text{col}}}G_{x,c}^{(t)}}{(t_{s+1}-t_s)\,|u_{\text{col}}|}$$

$$\lambda_1=\frac{\sum_{t=t_s}^{t_{s+1}-1}\sum_{c\in u_{\text{col}}}\sum_{r\in u_{1\text{row}}}G_{r,c}^{(t)}}{(t_{s+1}-t_s)\,|u_{\text{col}}|\,|u_{1\text{row}}|}$$

$$\lambda_2=\frac{\sum_{t=t_s}^{t_{s+1}-1}\sum_{c\in u_{\text{col}}}\sum_{r\in u_{2\text{row}}}G_{r,c}^{(t)}}{(t_{s+1}-t_s)\,|u_{\text{col}}|\,|u_{2\text{row}}|}$$

步骤 4 计算 $KL(\lambda_x,\lambda_1)$,$KL(\lambda_x,\lambda_2)$,其中,$KL(\lambda_1,\lambda_2)=\lambda_1\log\frac{\lambda_2}{\lambda_1}+\lambda_1-\lambda_2$ 表示参数 λ_1 和 λ_2 时泊松分布间的 Kullback-Leibler 散度。

步骤 5 如果 $KL(\lambda_x,\lambda_1)<KL(\lambda_x,\lambda_2)$且 $x\in u_{2\text{row}}$,则 $u_{1\text{row}}\leftarrow u_{1\text{row}}\cup\{x\}$,$u_{2\text{row}}\leftarrow u_{2\text{row}}-\{x\}$,否则如果 $KL(\lambda_x,\lambda_2)<KL(\lambda_x,\lambda_1)$且 $x\in u_{1\text{row}}$,则 $u_{2\text{row}}\leftarrow u_{2\text{row}}\cup\{x\}$,$u_{1\text{row}}\leftarrow u_{1\text{row}}-\{x\}$,否则保持不变。

步骤 6 如果 $R=\varnothing$,则终止;否则执行步骤 2。

(2)对树序列进行最优分段。为了使总码长最小,根据式(3)DMS 准则对树序列进行分段。按照如下方法确定一个变化点序列。对已知分段 Ψ,Ψ 的随机复杂度定义为:

$$SC=\sum_{G_t\in\psi}(-\text{lb}P(G_t\mid M_t))$$

其中,M_t 表示 G_t 的树。设 SC_s 表示最后一个分段 Ψ^s 的随机复杂性;SC_{s+t}表示 $\Psi^s\cup G_t$ 的随机复杂性;SC_t 表示 G_t 的随机复杂性;$\hat{\alpha}$ 表示从一个段到另一个段的转换概率估计。将式(3)DMS 准则应用到树序列的分割中。于是发现,如果:

$$SC_s+SC_t-\text{lb}\hat{\alpha}+L(M_t)<SC_{s+t}-\text{lb}(1-\hat{\alpha}) \tag{12}$$

则可将时间 t 看成是一个变化点,否则不将其看成变化点。此时:

$$L(M_t)\overset{\text{def}}{=}-N_1(M_t)\text{lb}P_1-N_0(M_t)\text{lb}P_0+\sum_{u\in int(M_t)}\mathscr{L}(u)$$

鉴于两者间的时间差异,通过式(3)DMS 准则和式(9)可推出式(12)。此外,G_t 的变化点指数可定义为:

$$score(t)\overset{\text{def}}{=}SC_{s+t}-\text{lb}(1-\hat{\alpha})-(SC_s+SC_t-\text{lb}\hat{\alpha}+L(M_t)) \tag{13}$$

该指数可衡量在时间 t 进行分割时码长的下降情况。本文定义,只要下式成立则 t 即是一个变化点。

$$score(t)>0 \tag{14}$$

TREE 的时间复杂度为 $O(mn(\text{lb}mn)T)$。其中,m 表示列的数量;n 表示行的数量;T 表示数据规模。这是因为每个树进行分割和融合的计算复杂度为 $O(mn)$,树的最大规模为 $O(\text{lb}mn)$,TREE 算法运行时与 T 呈线性关系。

算法 2 树分割

已知:u 为双支树的一个叶节点

步骤 1 对叶节点 u,生成新的叶节点 u_1,u_2,并设置 $u_{1\text{row}}=u_{\text{row}}$,$u_{2\text{row}}=\varnothing$。

步骤 2 $SC_1=SC(u_1)+SC(u_2)$。

步骤 3 对 $x\in u_{1\text{row}}$,计算将 x 从 u_1 移动到 u_2 时所生成的树的 SC,并将其记为 SC_2。

步骤 4 如果 $SC_2<SC_1$,则 $u_{1\text{row}}\leftarrow u_{1\text{row}}-\{x\}$,且 $u_{2\text{row}}\leftarrow u_{2\text{row}}\cup\{x\}$。

步骤 5 对所有 $x(\in u_{1\text{row}})$,重复步骤 2~步骤 4。

步骤 6 对 u_1,u_2,运行算法 1。

步骤 7 如果 $SC(u_1)+SC(u_2)-\text{lb}P_0+\mathscr{L}(u)<SC(u)-\text{lb}P_1$,则分割列节点。

算法 3 树融合

已知:u 为树的根;q 为队列

步骤 1 $q=\{u\}$。

步骤 2 $u'\leftarrow q.dequeue$,d:u'的深度。

步骤 3 设 u_1,u_2 是 u'的子树节点。计算:

$SC_1=SC(u')-\text{lb}P_1$

$SC_2=SC(u_1)+SC(u_2)-\text{lb}P_0+\mathscr{L}(u)$

步骤 4 如果 $SC_1<SC_2$,则修剪 u_1,u_2,否则 $q.enqueue(u_1,u_2)$。

步骤 5　重复步骤 2 ~ 步骤 4,直到 $q=\varnothing$。

4　基于直积的 GPCD

GraphScope 算法是一种典型的基于硬聚类的 GPCD 算法。首先分别对接收节点和发送节点进行聚类,然后生成一个图形分割作为发送节点和接收节点聚类的直积。文中将这种分割称为基于直积的分割。例如,假设发送节点集合为 $\{A,B,C,D\}$,接收节点的集合为 $\{1,2,3,4\}$。当为接收节点生成 $\{A,B\}$ 和 $\{C,D\}$ 2 个聚类,并为接收节点生成 $\{1,2\}$ 和 $\{3,4\}$ 2 个聚类时,生成的分割策略有 4 个区域分割成为它们的直积,如图 3 所示。

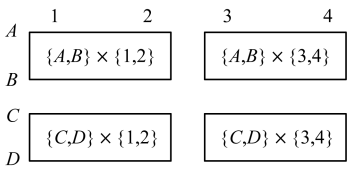

图 3　基于直积的分割

GraphScope 算法使用多种启发式策略构建基于直积的分割,此时总码长在所有可能的分割中局部最小。GraphScope 的计算复杂度为 $O(mnT(k*+\mathscr{L}*))$,其中,$k*$ 表示列聚类数量;$\mathscr{L}*$ 表示行聚类数量。

5　仿真实验与结果分析

5.1　基于人工数据集的实验

利用人工数据集比较 TREE 和 GraphScope 算法。设置接收节点和发送节点数量均为 20 个,并准备 4 个数据集,每个数据集有 30 个数据。根据 Θ_1 生成 $t=1\sim10$ 之间的数据,根据 Θ_2 生成 $t=11\sim20$ 间的数据,根据 Θ_3 生成 $t=21\sim30$ 间的数据。变化点为 $t_c=11$ 和 21。

生成数据集 1 ~ 数据集 4,其分割结构变化如图 4 ~ 图 7 所示。矩阵中的数据表明各分段的数量。使用效益和虚警率(False Alarm Rate,FAR)2 个指标衡量 GPCD 的性能。

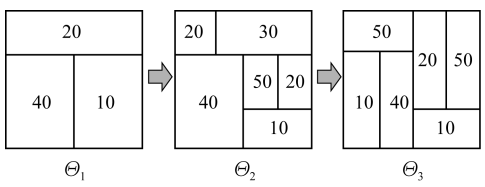

图 4　数据集 1

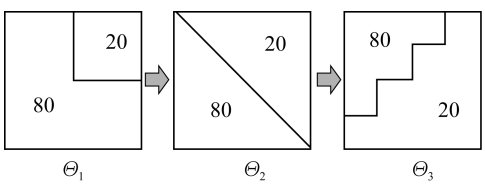

图 5　数据集 2

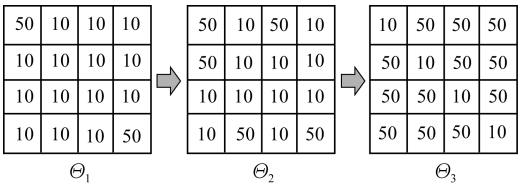

图 6　数据集 3

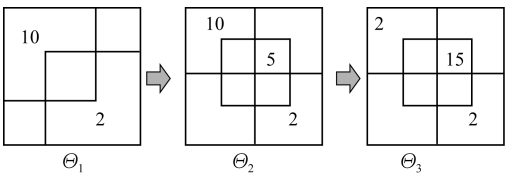

图 7　数据集 4

假设 x 表示被检测出来的变化点,t_c 表示真实点,x 的效益定义为:

$$benefit(x)=\max\{0,1-|x-t_c|/T\}$$

其中,T 表示已知正数。当且仅当 x 与 t_c 吻合时值为 1,当 $|x-t_c|$ 增加时线性降为 0。如果警报点大于 T 且离真实点很远,则认为警报丢失。FAR 表示所有警报中不是变化点的警报比例。进行 50 次实验取均值。

表 1 给出了不同算法在 4 个数据集上的效益和虚警率比较结果。本文在计算效益时设置 $T=3$。对数据集 1,2 种算法的效益值均较高,TREE 的 FAR 值远低于 GraphScope。这是因为直积方法无法有效表示真实的分割,导致 GraphScope 的 FAR 高于 TREE。

表 1　数据集 1 ~ 数据集 4 的效益和 FAR

算法	数据集 1		数据集 2		数据集 3		数据集 4	
	效益	FAR	效益	FAR	效益	FAR	效益	FAR
TREE	1.00	0.03	1.00	0.02	1.00	0.28	1.00	0.05
GraphScope	1.00	0.82	1.00	0.82	1.00	0.13	0.92	0.39

对数据集 2,由于与数据集 1 同样的原因,GraphScope 算法的 FAR 较高。然而,即使对数据集 2,基于树的方法无法有效表示真实分割,TREE 的 FAR 仍然远低于 GraphScope。对数据集 3,Graph-Scope 的 FAR 低于 TREE。这是因为直积方法非常有效地表示了真实分割,TREE 方法需要更多的参数才能表示真实分割。在实践中这种情况比较罕见。从以上结果可以看出,除了直积可以有效表示真实分割的少部分情况外,TREE 的性能优于 GraphScope。

表 2 给出了不同算法在数据集 1 ~ 数据集 4 上的内存占用量比较结果。实验环境为:CPU 为酷睿 3.4 GHz,内存容量为 4 GB,Windows 7 操作系统(64 位)。从表 2 可以看到,TREE 算法在 4 种数据集上的表现都要优于 GraphScope 算法。仔细分析其原因可知,这主要是因为 GraphScope 算法在描述

具有分层特征的图分割结构时会出现大量参数,需要多次分割以保证检测的准确性,占用了较多的系统资源。而本文方法利用基于树的概率模型表示图分割结构,并将动态模型选择(DMS)理论应用于图形分割序列选择过程中,与 GraphScope 算法相比,可以用较少的参数表示分层分割图,因此取得了更好的系统性能。

表 2 不同算法的内存占用量比较 MB

数据集	GraphScope	TREE
数据集 1	123	74
数据集 2	108	59
数据集 3	115	68
数据集 4	88	49

5.2 基于真实数据集的实验

利用日本内务交通省、统计局、政策规划和统计研究及培训学院总干事提供的人口流动数据[14]进行实验。这一数据对公众开放,给出了 2005 年 4 月-2011 年11 月间日本所有县市每月人口流动情况。发送节点是人口流出的县市,接收节点是人口流入的县市。从发送节点到接收节点之间链路的数值表示一个月内从发送节点到接收节点间的流动人口数量。时间点总数为 81,发送节点和接收节点总数均为 47 个。

图 8 和图 9 给出了 2 种算法的运行结果。每个图中的横坐标显示了 2005 年 4 月开始的时间。在 2 种情况下,当纵坐标数值超过 0 时即可检测出变化点。2 种算法均可检测出 $t=71$ 时的变化(从竖直虚线可以看出)。该变化对应于 2011 年日本远东大地震导致的人口迁移。2011 年 5 月检测出来的变化对应于日本政府宣布日本部分地区辐射水平显著升高时。2 种算法还检测出了 2011 年6 月-2011 年 8 月的变化,这 2 次变化对应于日本政府宣布 2011 年 6 月-2011 年 8 月间的人口流动。虽然 TREE 和 GraphScope 可以成功检测出重要真实事件的变化,但是 GraphScope 发布的与任何事件均无关联的虚警数量高于 TREE。这表明 TREE 的稳定性优于 GraphScope。

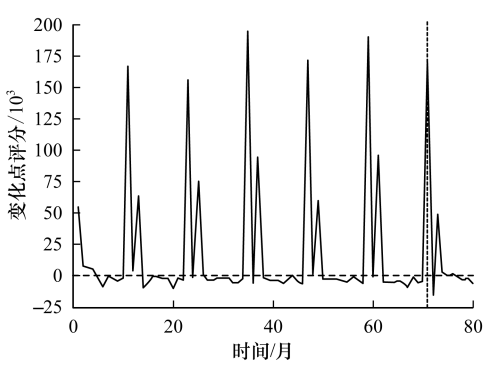

图 8 TREE 算法统计的人口流动数据

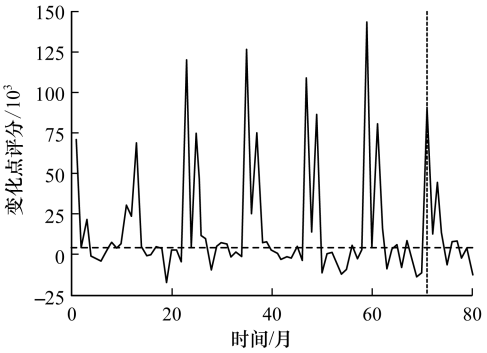

图 9 GraphScope 算法统计的人口流动数据

6 结束语

本文研究了图分割变化检测问题,提出一种基于树聚类的 GPCD 求解算法 TREE。该算法将传统的基于直积的方法(GraphScope)拓展至分割结构具有分层特点这一情况。根据 MDL 准则,从动态模型选择角度设计本文算法,比较了 TREE 和 GraphScope 算法的性能,结果表明对多种图形分割结构,TREE 算法的 GPCD 性能优于 GraphScope 算法。下一步研究工作的重点是针对有权图分割时不能很好解决子图内部耦合度不高的问题,使用可以同时优化子图内部顶点耦合度和子图之间顶点耦合度的 Ncut 准则,拟提出基于散列技术的图分割改进算法。

参考文献

[1] 文政颖,于海鹏. 基于多 Gamma 分布模型的 SAR 图像直方图分割算法[J]. 计算机工程与设计,2014,35(6):2104-2108.

[2] 汪云飞,毕笃彦,孙 毅,等. 一种采用双势阱策略的小直径图分割方法[J]. 计算机应用与软件,2013,30(4):275-278.

[3] Stanton I, Kliot G. Streaming Graph Partitioning for Large Distributed Graphs[C]//Proceedings of the 18th ACM International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2012: 1222-1230.

[4] Bansal N, Feige U, Krauthgamer R, et al. Min-max Graph Partitioning and Small Set Expansion[J]. SIAM Journal on Computing, 2014, 43(2): 872-904.

[5] Nishimura J, Ugander J. Restreaming Graph Partitioning: Simple Versatile Algorithms for Advanced Balancing[C]//Proceedings of the 19th ACM International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2013: 1106-1114.

[6] López-Ruiz R, Sañudo J, Romera E, et al. Statistical Complexity and Fisher-Shannon Information: Applications[M]. Berlin, Germany: Springer, 2011.

[7] Silva T C, Zhao Liang. Stochastic Competitive Learning in Complex Networks[J]. IEEE Transactions on Neural Networks and Learning Systems, 2012, 23(3): 385-398.

(下转第 242 页)

表 5 本文算法在图像 JPEG 压缩后的检测率

质量因子	图像数	检测数	检测率/%
30	200	185	92.5
40	200	183	91.5
50	200	180	90.0
60	200	177	88.5
70	200	176	88.0

从表 4、表 5 可以看出,本文算法具有一定的鲁棒性,能够有效地抵抗图像经过不同角度旋转或者不同质量因子 JPEG 压缩的操作。

4 结束语

模糊伪造图像定位检测是数字图像取证技术的重要内容之一。本文利用色彩一致性对图像中的模糊篡改操作进行检测,通过提取色调变化率、异常色调率的相对标准差和均值 4 个特征来量化描述模糊操作对图像色调一致性的破坏程度。将计算得到的 4 个统计特征输入 SVM 进行训练和分类,通过对原始像素点和模糊像素点的分类定位模糊篡改的区域。实验结果表明,本文方法可以准确地定位出图像中经过模糊篡改的区域,且对旋转和 JPEG 压缩操作具有一定的鲁棒性,在今后工作中将对适用于多种篡改操作的检测算法做进一步研究。

参考文献

[1] Shen Xuanjing, Tang Bohao, Li Xiaofei. A Blur Image Blind Identify Algorithm Based on the Edge Feature[C]//Proceedings of the 3th International Conference on Multimedia Information Networking and Security. Washington D. C., USA: IEEE Press, 2011: 309-313.

[2] 和 平,李 峰,向凌云. 融合 LWT 纹理特征的图像复制篡改检测算法[J]. 计算机工程,2013,39(10):267-270.

[3] Fan Shaosheng, Wang Hainan. Multi-direction Fuzzy Morphology Algorithm for Image Edge Detection[J]. Journal of Networks, 2011, 6(6): 95-898.

[4] 王 波,孙璐璐,孔维祥. 图像伪造中模糊操作的异常色调率取证技术[J]. 电子学报,2006,34(12A):2451-2454.

[5] Peng Fei, Wang Xilan. Digital Image Forgery Forensics by Using Blur Estimation and Abnormal Hue Detection[C]//Proceedings of 2010 Symposium on Photonics and Optoelectronic. Washington D. C., USA: IEEE Press, 2010: 1-4.

[6] 王 波,孔祥维,尤新刚. 利用色彩一致性的数字伪造图像取证方法[C]//全国计算机安全学术交流会论文集. 上海:[出版者不详],2008:118-123.

[7] Yang Benjuan, Zuo Juxian, Liu Benyong, et al. Blur Detection in Image Forensics Using Linear Correlation of Pixels[C]//Proceedings of 2010 Chinese Conference on Pattern Recognition. Washington D. C., USA: IEEE Press, 2010: 1-5.

[8] 刘 凯,扈文斌. 动态阈值模糊检测在篡改图像检测中的应用[J]. 上海大学学报:自然科学版,2011,17(5):586-590.

[9] 孙堡垒,周琳娜,张 茹. 基于 Benford 定律的高斯模糊篡改取证[J]. 计算机研究与发展,2009,46(Suppl):211-216.

[10] Sutcu Y, Coskun B, Sencar H T. Tamper Detection Based on Regularity of Wavelet Transforms Coefficients[C]//Proceedings of International Conference on Multimedia and Explore. Washington D. C., USA: IEEE Press, 2007: 397-400.

[11] Wang Xin, Xuan Bo, Peng Silong. Digital Image Forgery Detection Based on the Consistency of Defocus Blur[C]//Proceedings of International Conference on Intelligent Information Hiding and Multimedia Signal Processing. Washington D. C., USA: IEEE Press, 2008: 192-195.

[12] 郑美珠,赵景秀. 基于区域一致性测度的彩色图像边缘检测[J]. 计算机应用,2011,31(9):2485-2492.

[13] 吴德刚,张宇波,曹立波. 一种改进的模糊边缘检测算法[J]. 激光与红外,2010,40(12):1374-1377.

[14] Riess C, Angelopoulou E. Scene Illumination as an Indicator of Image Manipulation[M]//Böhme R, Fong P W L. Information Hiding. Berlin, Germany: Springer, 2010: 66-80.

编辑 陆燕菲

(上接第 236 页)

[8] Chakrabarti D, Papadimitriou S, Modha D S, et al. FullyAutomatic Cross-associations[C]//Proceedings of the 10th ACM International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2012: 79-88.

[9] Chakrabarti D. Autopart: Parameter-free Graph Partitioning and Outlier Detection[M]. Berlin, Germany: Springer, 2004.

[10] Sun Jimeng, Faloutsos C, Papadimitriou S, et al. Graphscope: Parameter-free Mining of Large Time-evolving Graphs[C]//Proceedings of the 13th ACM International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2007: 687-696.

[11] Yamanishi K, Maruyama Y. Dynamic Model Selection with Its Applications to Novelty Detection[J]. IEEE Transactions on Information Theory, 2013, 53(6): 2180-2189.

[12] Rissanen J. FisherInformation and Stochastic Complexity[J]. IEEE Transactions on Information Theory, 2012, 42(1): 40-47.

[13] Haldar J P, Hernando D, Liang Z P. Compressed-sensing MRI with Random Encoding[J]. IEEE Transactions on Medical Imaging, 2011, 30(4): 893-903.

[14] Jonsen I D, Flemming J M, Myers R A. Robust State-space Modeling of Animal Movement Data[J]. Ecology, 2005, 86(11): 2874-2880.

编辑 顾逸斐