仿高阶矩的结点不变量及其组成的图不变量

江顺亮 葛 芸 唐祎玲 徐少平 叶发茂
(南昌大学信息工程学院 南昌 330031)

摘 要 借鉴高阶矩的方法,采用层序的计算框架,依据结点的连接距离和层序信息定义了20种结点不变量。这些结点不变量体现图整体的上下偏分布特性、整体不均匀性和整体平滑性,结点不变量中的每层结点度数平方之和反映了层内结点度数的分布情况。通过比较这些结点不变量的可区分结点数,发现每层结点度数平方之和明显改善了结点不变量的细分能力。把排序后的结点不变量组成一个矢量后作为图的不变量。计算结果表明,共有9种图不变量可以区分所有结点数 $N<25$ 的非同构树和 $N<34$ 的非同构同胚不可约树(没有度数为2的树),对于更多结点的树,还没有发现非同构树有相同图不变量的例子;把这些图不变量应用到非同构图($N<10$),区分结果好于文献[8]中列出的22种图不变量的19种,而且文中9种图不变量的简并度不大,提高了随机图的同构测试性能。

关键词 图同构,结点不变量,图不变量,树不变量
中图法分类号 TP301.6 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.08.054

Node Invariants by Imitating High-order Moments and Their Graph Invariants

JIANG Shun-liang GE Yun TANG Yi-ling XU Shao-ping YE Fa-mao
(School of Information Engineering, Nanchang University, Nanchang 330031, China)

Abstract By the way of high-order moments and the level-order computing framework, twenty kinds of node invariants were defined with the connection distance and level-order information of nodes. These node invariants reflect overall bias distribution characteristics, non-uniformity and smoothness, while the sum of squares of nodal degree reflects the distribution of the nodal degrees in the level. By comparing the number of distinguishable nodes, it is found that the sum of squares of nodal degrees obviously improves the refining ability of node invariants. The node invariants are ordered to form a vector as the graph invariant. The calculation results show that there are nine graph invariants that can distinguish between all $N<25$ non-isomorphic tree and $N<34$ non-isomorphic irreducible tree (no 2-degree tree) without instance being found so far for non-isomorphic tree with the same graph invariants for trees with more nodes. The distinguished ability of nine graph invariants to non-isomorphic graphs ($N<10$) exceeds the 19 results of 22 graph invariants in reference [8], and the degeneracy of nine graph invariants is small, thus improving the performance of random graph isomorphism testing.

Keywords Graph isomorphism, Node invariants, Graph invariants, Tree invariants

1 引言

图同构问题是计算机科学复杂性理论中的重大问题[1]。芝加哥大学的László Babai教授于2015年宣称发现了一个拟多项式的图同构算法[2],引起了计算机学术界的广泛关注。但有学者怀疑这种算法在实际应用中的价值,真正的应用价值或许在未来才能得以体现[1]。目前获得广泛认可的图同构方法是Nauty[3],该方法充分利用结点细分、结点不变量和群变换理论,在算法实施方面进行了精致的处理,从而获得了良好的实际效果[3]。结点不变量在各种图同构方法中显得非常重要,它不仅可以对结点进行细分从而提高图同构算法的效率[3-4],还可组合形成图的不变量从而进行图同构的判断[5-8]。寻找更好、更充分的图不变量仍是一个开放的研究问题[7]。Dehmer等[8]比较了22种图不变量的区分力,并把其中一些不变量组合成了一个超级指数(Super Indices),对于结点数 N 小于10的图获得了良好的效果。寻找更好、更充分的图不变量也可以针对某种类型图进行,因为对所有非同构图都能区分的完备图不变量也许不存在,或者虽存在但因复杂度过

到稿日期:2017-07-24 返修日期:2017-09-15 本文受2017中国国家自然科学基金(61662044)资助。
江顺亮(1965-),男,博士,教授,博士生导师,主要研究方向为算法设计与分析、计算机模拟与仿真、人工智能,E-mail:jiangshunliang@ncu.edu.cn;**葛 芸**(1983-),女,博士生,讲师,主要研究方向为人工智能、机器视觉;**唐祎玲**(1977-),女,博士生,讲师,主要研究方向为人工智能、图像处理,E-mail:tangyiling@ncu.edu.cn(通信作者);**徐少平**(1976-),男,博士,教授,博士生导师,主要研究方向为机器视觉、计算机图形学;**叶发茂**(1978-)男,博士,副教授,主要研究方向为数字图像处理、计算机图形学。

高而失去了实用价值。树是图的一种，目前已经有了高效的树同构算法[9-10]，但是利用树不变量来进行树同构测试还未实现[5,11-12]，而树的不变量和非同构问题在分子结构、汉字识别等领域有着重要的意义[4,11-13]。

侯爱民对几种常见的结点不变量及其“异或距离”进行了初步比较[4]，基于几个例子分析了结点不变量的细分能力，但这种分析还不够系统。针对很多结点不变量，系统地进行结点不变量细分能力的比较和分析的研究还较少。

本文仿照高阶矩概念提出了一系列结点不变量，并把排序后的结点不变量组成一个矢量作为图的不变量。针对非同构自由树数据，利用可区分结点数比较了这些结点不变量的细分能力。已有计算结果表明，共有 9 种图不变量对非同构树的区分没有出现反例，而且对非同构图的区分性能较好。

2　系列结点不变量

图结点不变量主要有结点的度数、最短距离、路径数目、简单回路数目、包含于三角形的数目等[4]，以及由这些不变量衍生出来的各种不变量[5-8]。这些不变量中，有的反映了图的局部特征，有的反映了图的某些拓扑特性，但少有能反映图的整体特性的不变量。真实物体的高阶矩能够反映物体的整体特性，比如一阶矩反映了重心、二次矩是旋转惯量。离散刚体系统的 k 阶矩公式如下：

$$M_k=\sum m_i x_i^k \tag{1}$$

其中，m_i 是离散刚体系统的第 i 组成部件的质量，x_i 是其与原点之间的距离。本文借鉴高阶矩的方法，从一个结点出发，按层序，依据结点间的连接距离来计算各种高阶矩，m_i 取为结点的各种局部信息。

结点不变量的计算是从第一个结点 Start 出发，依据连接层序，逐层累加，如图 1 所示。图 1 中包括当前扩展层 FRONT、连接距离 h（第一个结点的连接距离为 1）、当前扩展层的结点数 w_1（可以看成局部宽度）、当前扩展层的结点度数平方之和 S_1。新扩展获得的结点属于 NEW_FRONT 层，结点数为 w_2，结点度数平方之和为 S_2。将每层结点度数的平方之和作为主要变量，因为它能体现结点度数的分布。每层数据累加时，将距离 h，h^2，h^3 作为权，20 个不变量的定义如表 1 所列，表 1 中用到的变量在图 1 中有示例。

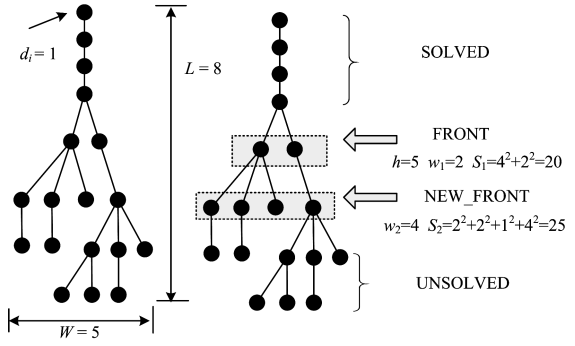

图 1　结点不变量层序计算示意图

Fig. 1　Diagram of calculating sequence nodal invariants

表 1　20 种不变量的定义

Table 1　Definition of 20 invariants

不变量 /代号	不变量定义	不变量 /代号	不变量定义
I_1/D	d_i	I_{11}/w_1w_2hhh	$\sum w_1w_2h^3$
I_2/W	W	I_{12}/w_2w_2hhh	$\sum w_2w_2h^3$
I_3/L	L	I_{13}/w_1S_1h	$\sum w_1S_1h$
I_4/w_1w_1	$\sum w_1w_1$	I_{14}/w_1S_2h	$\sum w_1S_2h$
I_5/w_1h	$\sum w_1h$	I_{15}/w_2S_1h	$\sum w_2S_1h$
I_6/w_1hh	$\sum w_1h^2$	I_{16}/w_2S_2h	$\sum w_2S_2h$
I_7/w_1w_1hh	$\sum w_1w_1h^2$	I_{17}/w_1S_1hh	$\sum w_1S_1h^2$
I_8/w_1w_2hh	$\sum w_1w_2h^2$	I_{18}/w_1S_2hh	$\sum w_1S_2h^2$
I_9/w_2w_2hh	$\sum w_2w_2h^2$	I_{19}/w_2S_1hh	$\sum w_2S_1h^2$
I_{10}/w_1w_1hhh	$\sum w_1w_1h^3$	I_{20}/w_2S_2hh	$\sum w_2S_2h^2$

表 1 中的累加符号是指按层累加。设计 20 种不变量的原因在于：1）方便计算，所有结点不变量可以用一个算法进行一次性计算；2）不同不变量能不同程度地反映图的整体连接特性。第一个结点不变量只是结点的局部信息；第二个和第三个结点不变量是从结点出发层序遍历图的长度和宽度，它们只是图的整体初级信息；后面 17 个结点不变量用到了 3 种信息：层的宽度 w_1 和 w_2、层中结点度数平方之和 S_1 和 S_2 以及结点的距离 h。距离 h，h^2，h^3 可以认为是层的权，高阶权的不变量能体现图形整体上下的偏分布特性，结点越集中在图形的远端，不变量越大。结点不变量公式中如果只有一个层的宽度，则实际是对所有结点进行累加，因此第五个结点不变量（代号 w_1h）可以表示图形的重心，第六个不变量（代号 w_1hh）是图形的二次矩（与物理意义上的旋转惯量对应）。相同层的同一个参数相乘（如 w_1*w_1）或不同参数相乘（如 w_1*S_1）可以反映图形的整体不均匀性，比如层宽为 2/2/2 的计算结点不变量会小于层宽为 1/3/2 的。不同层之间的两个变量相乘可以反映图形层间变化的平滑性，平滑性越好，不变量越大，这是因为如果顺序多个数的和相同，把两两相邻的数相乘然后累加，当大数与大数相邻、小数与小数相邻时，这个累加值会更大，比如 4 个数(1,1,4,4)的 $1*1+1*4+4*4$ 大于(1,4,1,4)的 $1*4+4*1+1*4$。

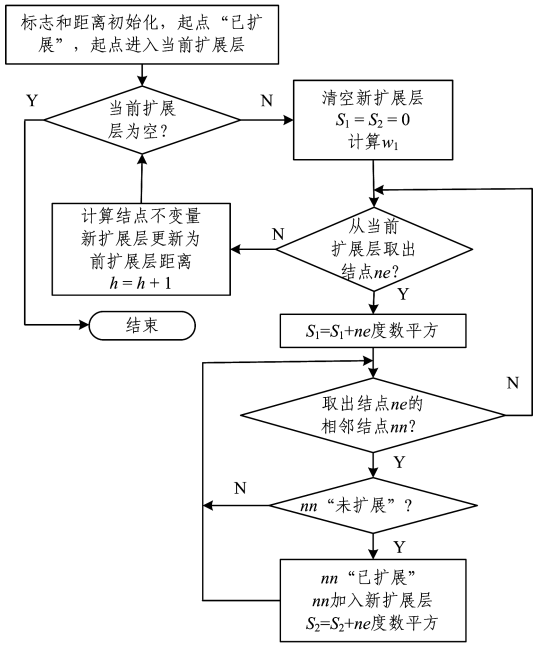

图 2　结点不变量的计算流程图

Fig. 2　Flowchart of calculating nodal invariants

综上所述，如果结点不变量包括 h 及 FRONT 和 NEW_FRONT 两层的数据，这些不变量就可以体现图形整体的上下偏分布特性、整体不均匀性和整体平滑性，尤其是 S_1 或 S_2 还可以反映层内结点度数的分布情况。有些结点不变量的差别不大，但由于要与其他不变量进行组合，为了选出更好的组合，保留了一些差别不大的不变量。

对于结点不变量，采用单源最短路径的计算框架，并用广度优先的层序计算方式进行计算，具体算法流程如图 2 所示。

3 结点不变量的结点细分能力比较

本文所指的“树”是自由树，即树的根结点未指定而且子树没有顺序的树，所说的“图”是指简单连通无向图。

只针对树进行结点不变量的结点细分能力比较，是基于以下原因：1)非同构图的数量巨大且不够完整，目前只有结点数 $N\leqslant 10$ 的所有非同构简单图数据，且 $N=10$ 的所有非同构简单图有 11716571 种，虽然可以产生 $N=11$ 的所有非同构简单图数据[3]，但其数量太大，共拥有 1006700565 种；2)由于图同构问题的复杂性，如何一次性地快速获取系列结点不变量是亟需解决的问题；3)非同构树的数据比较完备，互联网上比较容易获取结点数 $N\leqslant 22$ 的所有非同构树数据[14]，$N=22$ 时共有 5623756 种非同构树，此网站上还有 $N\leqslant 30$ 的所有非同构同胚不可约树(Homeomorphically Irreducible Trees)的数据；4)虽然对于树同构问题已经有高效的算法[9-10]，但还没有较好的结点不变量可以用来进行树同构的测试，因此针对树同构，寻找更好的乃至完备的结点不变量很有研究价值。

一种结点不变量对结点的细分能力一般是有限的，利用多种结点不变量来区分结点是比较常用的方法[3,5-6,8]，因此比较某种结点不变量时不仅需要考虑这种结点不变量对结点的细分能力，还需要考虑它与其他结点不变量结合时的结点细分能力。一个结点的多种结点不变量组成一个矢量，所有结点的多种结点不变量组成矢量的矢量，对这个矢量进行矢量排序后可以看成图的特征矢量，排序时将一个结点的矢量看作是一个固定的量，排序后结点的顺序发生了改变，而结点的多种结点不变量的相对顺序未改变。

生成图的不变量后，可以对图的结点进行划分。划分时，如果某个结点的不变量矢量与前后结点的不变量矢量不等，划分的结果就是该结点单独形成一个子集，意味着它是已经被区分的结点，此类结点被称为“可区分结点”。属于多个结点子集的结点就是不可区分的结点，即子集内所有结点的不变量相同。针对低阶的所有非同构图，统计所有可区分结点的数量，并将其作为该类特征矢量区分度的比较依据，可区分结点的数量越多时区分度越好。

用结点数 $N\leqslant 17$ 的所有非同构树(0 个和 1 个结点的树除外)进行结点不变量的细分能力比较，共有 205003 种非同构树，总计 3553678 个结点。比较依据的原始数据是区分结点总数，折算得分最好为 100，最差为 0，其他按线性比例计算，具体得分如表 2 所列。表 2 中的“组合 2”指某个结点不变量与其他一个结点不变量组合进行结点划分，把所有组合计算获得的区分结点数累加，然后用这个累加数进行折算得分；同理，“组合 3”指某个结点不变量与其他两个结点不变量组合进行结点划分。“Top15%次数统计”指找出所有组合的 Top 15%的结果，统计某个结点不变量出现的次数，依据最好为 100、最差为 0、其他按线性比例的方法折算得分。将所有得分进行平均后作为该结点不变量的细分能力比较依据，数据越大越好，这个数据不是绝对指标，而是与其他结点不变量进行比较后的相对指标。

表 2 利用树数据的结点不变量比较结果

Table 2 Comparison results of nodal invariants by tree data

结点不变量代号	区分结点总数累计折算			Top15%次数统计折算		平均
	组合 3	组合 2	单个	组合 3	组合 2	
D	17.1	14.7	6.9	0.0	0.0	7.7
W	0.0	0.0	0.8	0.0	0.0	0.2
L	2.1	9.2	0.0	0.0	0.0	2.3
w_1w_1	17.6	45.1	48.1	0.0	0.0	22.2
w_1h	71.6	82.6	78.0	9.6	0.0	48.4
w_1hh	83.9	96.0	96.9	19.2	14.3	62.1
w_1w_1hh	87.2	98.1	98.9	19.2	14.3	63.5
w_1w_2hh	86.9	97.7	98.3	17.8	10.7	62.3
w_2w_2hh	87.1	98.0	98.7	19.2	10.7	62.8
w_1w_1hhh	88.1	98.8	99.9	41.1	17.9	69.2
w_1w_2hhh	88.2	98.8	99.8	39.7	17.9	68.9
w_2w_2hhh	88.1	98.8	99.8	41.1	17.9	69.1
w_1S_1h	98.5	98.7	97.6	41.1	14.3	70.0
w_1S_2h	98.1	97.9	96.7	80.8	35.7	81.8
w_2S_1h	98.1	99.0	98.4	46.6	14.3	71.3
w_2S_2h	98.8	98.7	97.7	45.2	14.3	70.9
w_1S_1hh	100.0	100.0	99.9	78.1	82.1	92.0
w_1S_2hh	99.8	99.8	99.8	100.0	92.9	98.5
w_2S_1hh	100.0	100.0	100.0	57.5	89.3	89.4
w_2S_2hh	99.9	99.9	99.9	46.6	100.0	89.3

依据 20 个结点不变量的比较结果，可以把这 20 个结点不变量大致分为 4 档：差、及格、中、良。前 5 个结点不变量为差，I_6 到 I_9 为及格，I_{10} 到 I_{16} 为中，后 4 个为良。从数据的整体趋势看，结点不变量的细分能力与阶次有关，具体是 0 阶和 1 阶的结点不变量细分能力较差。从 I_6 到 I_9 的比较结果来看，提高局部宽度 w_1 的阶次并没有改善结点不变量的性能。从整体来看，S_1/S_2 的加入明显改善了结点不变量的性能。

4 图不变量在树同构测试中的结果分析

本文的目标是寻找一种结点不变量，由它组成的图不变量对非同构树的区分是完备的，或接近完备的，而对非同构图的区分性能较好。把一棵树所有的结点不变量排序后组成一个矢量，这个矢量可以看成树的不变量，用这个树不变量进行同构测试。

对结点数 $N\leqslant 22$ 的所有非同构树进行同构测试，令 G 为非同构树的集合，计算 G 集合中的所有非同构树的树不变量，这些树不变量构成集合 VA。令 $|G|$ 表示集合 G 的元素个数，如果 G 中的所有不同构树可以获得不同的树不变量，则 $|G|=|VA|$；如果有非同构树获得相同的树不变量，则 $|G|>|VA|$，而且相同树不变量的非同构树越多，两者的差值就越大，因此用 $nd=|G|-|VA|$ 表征该不变量的同构测试性能(本文称其为不变量欠量)，$nd=0$ 意味获得最佳的测试性能，nd 越大则性能越差。结果如表 3 所列，其中 I_i^g 是由结点不变量 I_i 排序后组成的图不变量，它们的代号相同。

表 3　不同图不变量计算获得的不变量欠量

Table 3　Number of deficiency calculated from different graph invariants

N	I_1^g	I_2^g	I_3^g	I_4^g	I_5^g	I_6^g	I_7^g，I_9^g	I_8^g	I_{10}^g，I_{12}^g	$I_{13}^g-I_{20}^g$
6	1	1	1	0	0	0	0	0	0	0
7	4	2	3	0	0	0	0	0	0	0
8	12	8	10	0	0	0	0	0	0	0
9	32	21	26	0	0	0	0	0	0	0
10	84	58	72	1	1	0	0	0	0	0
11	205	146	180	3	1	0	0	0	0	0
12	509	379	462	8	14	0	0	0	0	0
13	1245	991	1157	13	11	0	0	0	0	0
14	3082	2577	2926	30	133	0	0	0	0	0
15	7640	6656	7364	53	132	0	0	0	0	0
16	19185	17323	18710	120	1161	4	0	0	0	0
17	48453	44961	47642	259	1142	2	0	0	0	0
18	123636	117015	122270	569	9241	43	2	2	2	0
19	317658	305296	315371	1268	9702	37	2	1	1	0
20	822680	799807	818884	2793	72681	359	13	12	12	0
21	2144015	2101655	2137740	6622	80006	357	11	8	8	0
22	5623129	5544739	5612810	15675	550347	2653	99	88	86	0

表 3 中前 5 个结点不变量（代号分别为 D，W，L，w_1w_1，w_1h）的数据表明：树的单一简单特性对树的区分度不好，但图不变量 $I_4^g(w_1w_1)$明显好于其他 4 个图不变量，它能反映树的整体不均匀性，结点在某层越集中，该不变量越大；图不变量 $I_5^g(w_1h)$反映了树的重心，虽然对树区分度的性能不如图不变量 $I_4^g(w_1w_1)$，但相比前 3 个图不变量（结点度数、树长度、树宽度）有实质提高。

图不变量 $I_6^g(w_1hh)$是树的二次矩，相当于旋转惯量，可以反映树的上下偏分布，性能优于前 5 个图不变量，次于后 14 个图不变量。由于结点不变量是按层累加的，结点不变量公式中一个 w_1 或 w_2 可以看作是对所有结点的累加，其余的 w_1 或 w_2 可以看成是 1 阶的，S_1/S_2 由于是层中结点度数的平方之和，可以看成是 2 阶的，因此后 14 个图不变量是 3 阶或 4 阶的，性能好于前 6 个图不变量是合理的。层中结点度数的平方之和 S_1/S_2 可以反映不同度数结点在层中的分布，度数越大的结点在某一层越集中，不变量越大，因此后 8 个图不变量对于结点数 $N\leqslant22$ 的所有非同构树具有完全的区分度。

根据表中的数据，可以获得不同图不变量对树区分度的性能排序：$(D,W,L)<w_1h<w_1w_1<w_1hh<(w_iw_jhh,w_iw_jhhh)<(w_iS_jh,w_iS_jhh)$，其中 i 和 j 取 1 或 2。现有文献中这方面的数据较少，可以从文献[5]中获得树的不变量欠量 $nd=|G|-|VA|$，当 $N=14,15,16,17$ 时文献[5]的结果分别为 1，3，13，35；本文后 15 个结点不变量的结果明显好于文献[5]的结果。侯爱民分析了多种不变量对树的区分情况[4]，认为“最短距离”不变量强于“异或距离”不变量、“同或距离”不变量、“顶点的邻接点度数序列”不变量等的细分能力。实际上，与本文的其他不变量相比，“最短距离”不变量（本文的 I_3/L）的细分能力较弱。

为了进一步测试 I_{13}^g 到 I_{20}^g 这后 8 个图不变量 $JP2$，首先生成 $N=23$ 和 $N=24$ 的所有非同构树，生成方法是对 $N=22$ 的每一棵树在每个结点上增加一条边和一个结点形成一棵 $N=23$ 的树，每生成一棵新的 $N=23$ 的树，就判断在已有的 $N=23$ 的树中是否存在与之同构的树，如果没有则加入到 $N=23$ 的树中，如果有则抛弃。同构测试只需用图不变量 I_{18}^g (w_1S_2hh)排序后形成的树不变量进行比较，实践证明该树不变量可以区分所有结点数 $N<25$ 的非同构树和 $N<34$ 的非同构同胚不可约树（Homeomorphically Irreducible Trees）。

生成所有 $N=23$ 的树之后，同理可以生成 $N=24$ 的树。由于 $N=23$ 和 $N=24$ 的树数量较大（14828074 种和 39299897 种），对计算机内存和计算时间需求较大，因此采用分类生成的方法，$N=23$ 的树采用结点最大度数和叶子数进行分类，$N=24$ 的树采用结点最大度数、叶子数和度数为 2 的结点数进行分类。采用类似方法可以正确生成 $N=31$，32 和 33 的所有非同构同胚不可约树。生成 $N=31$ 的所有非同构同胚不可约树要用到 $N=29$ 和 30 的所有非同构同胚不可约树，方法是对每个 $N=29$ 的同胚不可约树在叶子结点上增加两个结点和两条边形成一个 $N=31$ 的同胚不可约树，对每个 $N=30$ 的同胚不可约树在非叶子结点上增加一个结点、一条边形成一个 $N=31$ 的同胚不可约树。$N=33$ 的同胚不可约树采用结点最大度数、最大度数的结点个数和度数为 3 的结点数进行分类计算。

有了 $N=23$ 和 $N=24$ 的非同构树和 $N=31$，32 和 33 的非同构同胚不可约树的数据后，可以计算图不变量 $I_{13}^g-I_{20}^g$ 的不变量欠量 $nd=|G|-|VA|$，结果发现这 8 个图不变量计算的不变量欠量全部为零（$nd=0$），即它们对这些非同构树具有完全的区分度。但这 8 个图不变量对应的结点不变量的细分能力是有明显区别的，主要原因应该是细分是比较同一个图中的不同结点的不变量，而两树之间的区分是比较不同树对应结点的不变量，只要存在一对对应结点不变量不相等，两个棵树就被区分开了。因此，树不变量区分非同构树的优良性能并不要求对应的结点不变量拥有极强的细分能力，只需较好的细分能力即可。这样的结论是否可以推广到图还有待研究。

为了进一步寻找最小完备的树不变量（此处“完备”是指对非同构树具有完全的区分度），限定图不变量 I_{13}^g 到 I_{20}^g 定义中的部分变量。分 3 种情况：1）消去层宽度，令 $w_1=w_2=1$；

2)降低 S_1 和 S_2 的阶数,令 S_1 为 FRONT 层的结点度数之和,而不是结点度数平方之和,同理令 S_2 为 NEW_FRONT 层的结点度数之和;3)消去距离,令 $h=1$。3 种情况下的计算情况如表 4 所列。

表 4 部分变量限定情况下计算获得的不变量欠量 nd

Table 4 Number of deficiency cacluated under variable constraint

N	$w_1=w_2=1$			$S_1,S_2=\text{sum } d_i$		$h=1$
	I^g_{13},I^g_{16}	I^g_{14}	$I^g_{15},I^g_{17}-I^g_{20}$	$I^g_{13}-I^g_{16}$	$I^g_{17}-I^g_{20}$	$I^g_{13}-I^g_{20}$
14	0	0	0	0	0	0
15	0	0	0	2	0	0
16	0	0	0	13	0	0
17	0	0	0	42	0	0
18	2	2	2	156	0	0
19	1	2	1	469	0	0
20	14	13	12	1492	1	0
21	8	8	8	4530	0	0
22	88	89	86	13614	5	0

情况 1)和情况 2)存在不变量欠量 $nd>0$ 的情况,即存在非同构树有相同不变量的情况,这些树不变量是不完备的。对于情况 3),消去距离,令 $h=1$,从 8 种树不变量($I^g_{13}-I^g_{20}$)中可以获得 4 种结点不变量(依表 1 中"代号"定义的方式,代号分别为 w_1S_1,w_1S_2,w_2S_1,w_2S_2)。

进一步计算 $N=23$ 的非同构树,获得树不变量 w_1S_1,w_1S_2,w_2S_1,w_2S_2 的 nd 分别为 2,0,1,2。因此,树不变量 w_1S_1,w_2S_1,w_2S_2 是不完备的。然后,对树不变量 w_1S_2 进行深入计算,它可以区分所有结点数 $N<25$ 的非同构树和 $N<34$ 的非同构同胚不可约树。由于与树不变量 w_1S_2 非常类似的树不变量 w_1S_1,w_2S_1 和 w_2S_2 是不完备的,本文对"树不变量 w_1S_2 是完备的"持谨慎态度。与之对应,本文对"树不变量 $I^g_{14}(w_1S_2h)$是完备的"持乐观态度,而且针对非同构树的计算表明树不变量 I^g_{14} 在 8 种树不变量($I^g_{13}-I^g_{20}$)中具有最小的数值。对更多结点数的树进行随机测试。取 $N=100$ 的 10 组 10 万棵随机非同构树,数据是用 Nauty 2.6 中的 genrang 生成并用 shortg 确保了 10 万棵树是相互不同构的。树不变量($I^g_{13}-I^g_{20}$)和树不变量 w_1S_2 对这 10 组数据都具有完全的区分度。

依据现有的测试结果,本文猜想树不变量 $I^g_{14}(w_1S_2h)$是最小完备的树不变量,若要从数学上严格证明这个猜想,还需要进行大量的研究工作。

5 结点不变量在图同构测试的结果分析

树是一种特殊的连通图,本文已经找到 9 种图不变量,它们在所有测试的树数据上都获得了完全区分度。这些不变量包含结点度数、距离、连接层序的信息,但并不包含一般图具有而树不具有的信息,比如路径数、回路数、同层之间的连接数等,而这些信息对图同构测试非常重要。文献[8]列举了 22 种图不变量对结点数 $N\in[5,9]$的所有非同构图的测试结果。文献[8]的数据是不可区分图的数量,不可区分图是指与其他不同构的图具有相同图不变量的图,本文的计算结果与文献[8]中 22 个结果中最好的 3 个结果如表 5 所列。

表 5 用图不变量计算的不可区分图数量

Table 5 Number of indistinguishable graphs calculated with graph invariants

N	Top 3 in 22 Results[8]			w_1S_1h	w_1S_2h	w_1S_1h+cm	w_1S_2h+cm
	Top 1	Top 2	Top 3				
5	6	2	2	0	0	0	0
6	16	5	23	8	4	0	0
7	34	7	96	71	77	12	12
8	385	482	1701	1938	1952	497	529
9	6009	27030	58202	82594	83067	32203	33202

表 5 中的代号"w_1S_1h"和"w_1S_2h"与表 1 中的相应结点不变量的代号相同,表中的"$+cm$"是指用图的补图(Complementary Graph)计算图不变量,然后联合原图的不变量对图进行同构测试。本文计算结果的所有数据优于文献[8]中其他 19 个结果,因此这 19 个结果没有在表 5 中呈现,这 19 个结果的平均值($N=5,6,7,8,9$)分别为 9,64,584,8983,231809。

9 种图不变量($I^g_{13}-I^g_{20}$和 w_1S_2)计算获得的不可区分图数量的差别较小,而且用不同的不变量组合成的矢量作为图的不变量,计算获得的不可区分图数量与表 5 中的结果的区别也较小,比如 3 个不变量(w_2w_2hhh,w_1S_2h,w_1S_1hh)包含 2 层 5 种信息(w_1,w_2,S_1,S_2,h),用它们组成的图不变量计算获得的结果($N=5,6,7,8,9$)分别为 0,4,71,1936,82594。但是,用原图与补图的结点不变量联合组成原图的不变量进行计算时获得的结果有较大的改进,具体见表 5 中的"w_1S_1h+cm"和"w_1S_2h+cm"的结果。主要原因是这 9 种图不变量较好地反映了类似树的图特征,而对于边数较多的图,它们的补图也类似于树或森林,因此把补图的不变量加入到同构测试中会明显改善结果。

为了探讨不变量对不同密度图(边数的疏密度)的区分情况,对不同边数的结点数 $N=10$ 的图进行了同构测试。$N=10$ 的非同构图共有 11716571 种,边数 $E=20\sim25$ 的非同构图都超过 100 万种,其中 $E=23$ 的非同构图最多,共有 1348674 种。用不变量 w_1S_2h 计算的不变量欠量总数是 3999232,百分比为 34.13%,其中 $E=24$ 的最多,数值为 554612;如果按百分比计算,则是 $E=27$ 的最多,数值为 49.69%。把补图的不变量加入到同构测试中,极大地改善了边数较多的结果,同时总体不变量欠量百分比从 34.13%降为 16.67%,数值和百分比最多的都在 $E=23$ 处。因此,当图及其补图的边数都较多时,9 种图不变量计算的结果都不理想。

对结点数 $N=15$ 的图进行随机测试。取边数 $E=40\sim49$ 的 10 组 100 万个随机非同构图,数据是用 Nauty 2.6 中的 genrang 生成并用 shortg 确保了 100 万个图是相互不同构的。用 5 种图不变量(代号分别是 L,w_1hh,w_1w_1hh,w_1S_2,w_1S_2h)对这 10 组数据进行了不变量欠量的计算,结果如表 6 所列。其中,图不变量 w_1S_2 和 w_1S_2h 对总共 1000 万个图只有不变量欠量各为 1,这意味着各存在一对图是不可区分的。对于性能良好的图不变量,存在比较少的不可区分图是图同构的特点[15],而且找出这样的图例也是非困难的[15]。针对这两个不变量的具体情况,原因可能是大部分不可区分的图

是成对、成三或以非常小图群的形式出现。为了分析这个原因，计算了图不变量的简并度(Degeneracy)，结果如表 7 所列。简并度指一个不变量对应非同构图的数量，对应 2 个非同构图时简并度为 2，对应 3 个非同构图时简并度为 3。

表 6　$N=15$ 时不同边数的不变量欠量

Table 6　Number of deficiency from graphs with $N=15$

E	L	w_1hh	w_1w_1hh	w_1S_2	w_1S_2h
40～42	999892	＞192000	＞192000	0	0
43～45	999919	＞406000	＞407000	0	0
46～48	999947	＞612000	＞612000	0	0
49	999962	761027	761183	1	1

表 7 中的数据不包括只对应一个图的不变量。不变量 w_1S_2 的简并度与不变量 w_1S_2h 的简并度相差较小，因此表 6 中没有列出不变量 w_1S_2 的简并度数据。从表 6 中可以看出，如果一个不变量 w_1S_2h 对应几个非同构图，一般对应 2 个或 3 个非同构图，也许随着结点数的增加该数据会相应增加，但与表中其他不变量的数据相比，这个数据应该是比较小的。另一方面，非同构图数量巨大，比如 $N=15$，$E=40$ 的非同构图有 124262478451794760 之巨，远超 100 万，因此在很多个非常小的图群中出现 2 个图在 1 个图群的概率较小。但是，不变量 L，w_1hh，w_1w_1hh 的简并度比 w_1S_2h 的简并度更大，增长很快，而且这种图的总数量比例较高，因此在 100 万个非同构图中会出现很多具有相同不变量的非同构图，故计算的不变量欠量比较大，不变量 L 的不变量欠量占比几乎为 100%，不变量 w_1hh 和 w_1w_1hh 的不变量欠量占比为 19%到 76%不等。这也说明，9 种图不变量($I^g_{13}-I^g_{20}$ 和 w_1S_2)对非同构图的区分性能远远超过本文的其他不变量，而且对随机图的区分性能非常好。

表 7　简并度大于 1 的简并度统计

Table 7　Statistic of degeneracy(＞1)

N	Invariant Code	$\|G\|$	$\|VA\|$	Degeneracy	
				Average	St. Dev.
8	w_1S_2h	1952	827	2.4	0.8
	w_1hh	8296	1457	5.7	8.9
	w_1w_1hh	8526	1429	6.0	9.0
	L	11113	36	308.7	658.3
9	w_1S_2h	83067	30438	2.7	3.6
	w_1hh	235042	19446	12.1	48.8
	w_1w_1hh	237634	17439	13.6	53.0
	L	261076	60	4351.3	10770.9

结束语　仿照高阶矩概念，依据结点的连接距离和层序信息定义了 20 种结点不变量，这些结点不变量可以采用单源最短距离的层序计算框架进行一次计算。这些结点不变量体现了图整体的上下偏分布特性、整体不均匀性和整体平滑性。利用可区分结点数比较了这些结点不变量的细分能力，发现结点不变量包含每层结点度数的平方之和，明显改善了它的细分能力。把排序后的结点不变量组成一个矢量作为图的不变量。计算结果表明，共有 9 种图不变量可以区分所有已经测试的非同构树，至今没有出现反例；把这些图不变量应用到非同构图($N<10$)，区分结果好于文献[8]中列出的 22 种图不变量的 19 种，而且由于这些图不变量的简并度较小，因而提高了随机图的同构测试性能。

参 考 文 献

[1] SAVAGE N. Graph matching in theory and practice[J]. Communications of the ACM, 2016, 59(7): 12-14.

[2] BABAI L. Graph Isomorphism in Quasipolynomial Time[OL]. http://arxiv.org/abs/1512.03547.

[3] MCKAY B D, PIPERNO A. Practical graph isomorphism, II[J]. Journal of Symbolic Computation, 2014, 60(1): 94-112

[4] HOU A M. Theory Research and Algorithm Design On Halmiltonian Cycle and Graph Isomorphism Problems[D]. Guangzhou: South China University of Technology, 2013. (in Chinese)
侯爱民. 哈密顿环与图同构问题的理论研究及算法设计[D]. 广州：华南理工大学，2013.

[5] PIEC S M, MALARZ K, KULAKOWSKI K, et al. How to count trees[J]. International Journal of Modern Physics C, 2005, 16(10): 1527-1534.

[6] ZOU X X, DAI Q. A Vertex Refinement Method for Graph Isomorphism[J]. Journal of Software, 2007, 18(2): 213-219. (in Chinese)
邹潇湘，戴琼. 图同构中的一类顶点细分方法[J]. 软件学报，2007，18(2)：213-219.

[7] GAMKRELIDZE A, VARAMASHVILI L, HOTZ G. New Invariants for the Graph Isomorphism Problem[J]. Journal of Mathematical Sciences, 2016, 218(6): 754-762.

[8] DEHMER M, GRABNER M, MOWSHOWITZ A, et al. An efficient heuristic approach to detecting graph isomorphism based on combinations of highly discriminating invariants[J]. Advances in Computational Mathematics, 2013, 39(2): 311-325.

[9] BUSS S R. Alogtime algorithms for tree isomorphism, comparison, and canonization[M]// Computational Logic and Preef Theory. Springer Berlin Heidelberg, 1997: 18-33.

[10] ZHANG B, TANG Y, WU J, et al. LD: A Polynomial Time Algorithm for Tree Isomorphism[J]. Procedia Engineering, 2011, 15(8): 2015-2020.

[11] ZHANG N. Research the Longest Path Length Based On the Degree Sequence of Non-isomorphic Undirected Tree[D]. Huhehot: Inner Mongolia Normal University, 2013. (in Chinese)
张楠. 基于度序列的非同构无向树的最长路径长度的研究[D]. 呼和浩特：内蒙古师范大学，2013.

[12] SCHMUCK N S, WAGNER S G, HUA W. Greedy trees, caterpillars, and Wiener-type graph invariants[J]. Match Communications in Mathematical and in Computer Chemistry, 2012, 68(68): 273-292.

[13] LOHREY M, MANETH S, PETERNEK F. Compressed Tree Canonization[C]// International Colloquium on Automata, Languages, and Programming. Springer Berlin Heidelberg, 2015: 337-349.

[14] BRINKMANN G, COOLSAET K, GOEDGEBEUR J. House of Graphs: A database of interesting graphs[J]. Discrete Applied Mathematics, 2013, 161(1/2): 311-314.

[15] BABAI L, ERDOS P, SELKOW S M. Random Graph Isomorphism[J]. Siam Journal on Computing, 1980, 9(3): 628-635.