

双重标准差法在昆虫科阶元分类学上的应用

杜瑞卿, 王庆林, 庞发虎, 王明伟

(南阳师范学院生命科学系, 河南南阳 473061)

摘要: 在科分类阶元上对半翅目、鳞翅目和鞘翅目 8 个科的 23 种昆虫图像中提取的昆虫面积、周长等 11 项数学形态特征进行了双重标准差法分析, 以评估该方法在昆虫科阶元分类上的应用有效性。结果表明, 在科的阶元上 11 项特征可靠性大小依次为: (似圆度、偏心率、圆形性) > (横轴长、形状参数、叶状性) > (面积、周长、球状性) > (纵轴长、亮斑数)。科的亲缘关系远近结果显示: 螻蛄科和缘螻蛄科关系较近 > 丽金龟科、天牛科与鳃金龟科关系较近 > 夜蛾科、大蚕蛾科和粉蝶科关系较近。所得结果与统计假设检验分析所得结果非常相似。

关键词: 昆虫分类; 科阶元; 数学形态特征; 双重标准差法

中图分类号: Q69 文献标识码: A 文章编号: 0454-6296(2006)06-0882-05

Use of double standard deviation at family level of insect taxonomy

DU Rui-Qing, WANG Qing-Lin, PANG Fa-Hu, WANG Ming-Wei (Department of Biology, Nanyang Normal University, Nanyang, Henan 473061, China)

Abstract: Double standard deviation analysis with 11 math-morphological features (MMFs) (such as area, perimeter, etc.) from the images of 23 species of insects of the Pentatomidae, Coreidae, Noctuoidea, Saturniidae, Pieridae, Melolonthidae, Rulelidae and Cerambycidae families was made to assess its use in insect taxonomy at family level. The results indicate that the ranked reliability of MMF in the identification of insect families is, from high to low: roundness-likelihood, eccentricity, circularity > X-length, form parameter, lobation > area, perimeter, sphericity > Y-length, hole number. From the perspective of mathematical morphology, the kinship can be ranked as follows: Pentatomidae and Coreidae > Rulelidae, Cerambycidae and Melolonthidae > Noctuoidea, Saturniidae and Pieridae. The results are similar to those of statistical hypothesis testing analysis.

Key words: Insect taxonomy; family level; math-morphological features; double standard deviation

在前文“粗糙集模糊聚类分析法在昆虫分类学上的应用”(杜瑞卿等, 2006)和“粗糙集神经网络分析法在昆虫总科阶元分类学上的应用”(杜瑞卿等, 待发表)中, 论述了昆虫数学形态特征在目和总科阶元上的应用可行性, 并与赵汗青等(2003a, 2003b)两篇论文的统计分析结果进行了比较, 获得较为满意的结论。而科阶元作为昆虫分类上的一个重要阶元, 探讨数学形态特征在该阶元上的应用亦是很重要的。但是从目、总科到科, 随分类级别降低所包括分类单元越来越多, 分析其特征属性的重要性以及亲缘关系的远近也就越来越复杂, 计算量也不断增大。作为统计假设检验法, 一方面检验总体的样本越来越少, 检验可靠性减小; 另一方面, 检验总体增

多, 计算量增大。此时选择统计假设检验对各特征属性在各科的显著性检验已不是理想方法。粗糙集理论在科阶元分类上, 虽然可以获得约简属性, 但得出的相对约简较多, 无法确定属性指标在科阶元分类上的重要性。作为好的分类指标(特征属性), 应满足两个基本要求: 一是指标值在同一类内分布均衡稳定, 体现类的数值特征; 二是指标值在不同类上的分布差异较大, 体现指标的敏感性。本文作者正是从这样的思想出发, 根据昆虫在阶元分类上的特点, 提出一种“双重标准差法”, 对昆虫在科阶元分类上各特征属性的重要性进行分析, 并确定科之间的亲缘关系。

基金项目: 南阳师范学院科研资助项目(NYCT2004K01)

作者简介: 杜瑞卿, 男, 1968年生, 硕士, 讲师, 主要从事生物医学工程学研究, E-mail: duruiqing8@163.com

收稿日期 Received: 2006-01-04; 接受日期 Accepted: 2006-08-28

1 材料与方法

翅目、鳞翅目、鞘翅目的 3 目 8 科的 23 种昆虫, 每种昆虫各取 50 个左右的成虫标本用数码相机获取图像。昆虫名录和原始数据见表 1。

1.1 数据来源

原始数据选自赵汗青等(2003a), 涉及隶属于半

表 1 23 种昆虫的 11 项数学形态特征提取值

Table 1 Eleven math-morphological characters extracted from 23 species of insects

昆虫名称 Insect name	面积 Area (A)	周长 Perimeter (P)	横轴长 X-length (XL)	纵轴长 Y-length (YL)	形状参数 Shape parameter (F)	叶状性 Lobation (B)	球状性 Sphericity (S)	圆形性 Circularity (C)	似圆度 Roundness (R)	偏心率 Eccentricity (E)	亮斑数 Hole number (H)
蜻科 Pentalonidae											
碧蜻 <i>Palmenia angulosa</i> Motschulsky	2 718	215.9	48.3	74.6	1.3685	0.4080	0.4813	4.4155	1.4840	1.5470	1.00
麻皮蜻 <i>Erthesina fullo</i> (Thunberg)	4 725.7	302.7	60.4	104.5	1.4587	0.3560	0.3292	5.3520	1.6450	1.7310	1.00
缘蜻科 Coreidae											
褐奇缘蜻 <i>Derepteryx fuliginosa</i> (Uhler)	5 216.5	378.3	62.2	117.6	2.2014	0.2760	0.2210	4.0697	1.7350	1.9050	1.17
波原缘蜻 <i>Coreus potanini</i> (Jakovlev)	2 085.4	231.0	41.3	75.1	2.2619	0.3370	0.2964	4.0173	1.5620	1.8230	1.07
夜蛾科 Noctuidae											
小地老虎 <i>Agrotis ypsilon</i> (Rotttemberg)	20 365.0	932.9	273.5	117.8	3.315	0.089	0.156	5.900	0.335	0.431	10.12
棉铃虫 <i>Helicoverpa armigera</i> (Hübner)	10 394.0	768.8	188.2	86.0	4.686	0.085	0.151	4.909	0.372	0.441	26.17
白点雍夜蛾 <i>Oederernia esox</i> Draudt	13 841.0	717.8	228.4	96.0	2.980	0.076	0.139	5.874	0.338	0.409	1.84
大蚕蛾科 Saturniidae											
大蚕蛾 <i>Rhodinia jankowskii hattoriae</i> Inoue	72 155.0	1 685	485.1	124.7	3.142	0.048	0.084	7.750	0.442	0.254	5.20
丁目大蚕蛾 <i>Aglia tau</i> L.	82 879.0	1 821	503.4	147.8	3.201	0.069	0.087	9.366	0.426	0.279	3.40
黄目大蚕蛾 <i>Caligula anna</i> Moore	103 368.0	2 143	561.0	156.4	3.545	0.041	0.072	9.243	0.420	0.270	53.17
猫目大蚕蛾 <i>Salassa thespis</i> Leech	149 687.0	2 486	640.9	178.4	3.297	0.050	0.086	9.758	0.464	0.261	10.63
菜粉蝶科 Papilionidae											
菜粉蝶 <i>Pieris rapae</i> L.	37 948.0	1 206	298.0	92.7	2.901	0.086	0.145	6.923	0.551	0.316	7.73
黄粉蝶 <i>Colias hyale</i> (L.)	38 479.0	1 105	305.9	105.1	2.538	0.110	0.189	6.886	0.526	0.344	1.22
山楂粉蝶 <i>Aporia crataegi dilata</i> Verity	54 812.0	494	375.5	117.5	3.278	0.064	0.109	8.275	0.495	0.264	1.39
尖钩粉蝶 <i>Gonepteryx mahaguru aspasia</i> Ménétrière	60 053.0	1 344	377.2	125.4	2.404	0.114	0.198	8.521	0.542	0.321	1.76
鳃金龟科 Melolonthidae											
棕色鳃金龟 <i>Holotrichia titanis</i> Reitter	5 331.4	303.0	60.7	113.6	1.373	0.420	0.424	5.554	1.844	1.875	1.63
华北大黑鳃金龟 <i>Holotrichia oblita</i> (Faldemann)	5 252.0	299.6	60.7	113.0	1.362	0.422	0.421	5.564	1.828	1.870	1.57
丽金龟科 Rutelidae											
中华弧绿丽金龟 <i>Popillia quadriguttata</i> Fabr.	4 350.3	266.8	59.3	98.0	1.306	0.436	0.492	5.163	1.571	1.652	6.74
铜绿丽金龟 <i>Anomala corpulenta</i> Motschulsky	4 606.4	275.8	61.3	101.0	1.319	0.445	0.514	5.342	1.558	1.648	2.77
天牛科 Cerambycidae											
黄斑星天牛 <i>Anoplophora nobilis</i> (Ganglbauer)	8 011.5	488.4	60.8	167.3	2.389	0.349	0.222	5.011	2.743	2.750	16.20
松幽天牛 <i>Asemum amurense</i> Kraatz	2 402.9	234.4	32.0	91.2	1.860	0.363	0.226	4.139	2.991	2.853	1.37
榆绿天牛 <i>Chelidonium provosti</i> (Fairmaire)	1 871.9	220.0	27.3	88.7	2.067	0.409	0.223	3.682	3.214	3.265	1.00
绿翅楔天牛 <i>Saperda viridipennis</i> Gressitt	2 700.8	262.0	32.9	98.9	2.032	0.411	0.249	3.919	3.176	3.013	1.35

1.2 研究方法

1.2.1 特征值的提取：方法见赵汗青等 (2003a, 2003b)。提取的特征包括图像中昆虫的面积、周长、横轴长、纵轴长、形状参数、叶状性、球状性、圆形性、偏心率、似圆度、亮斑数等 11 项指标 (表 1)。

1.2.2 双重标准差法：

(1) 数据标准化：由于各特征属性值 (表 1) 的单位不同，数据大小不同。为了消除各特征属性单位的影响，进行标准化处理。把表 1 可以看成是 23 × 11 矩阵表，表中元素为 X_{ij} ($i = 1, 2, \dots, 23; j = 1, 2, \dots, 11; i$ 为行 j 为列)。

$$T_{ij} = \frac{x_{ij}}{x_j} \quad (i = 1, 2, \dots, n_t; j = 1, 2, \dots, 11, t = 1, 2, \dots, 8, t \text{ 为科数 } n_1 + n_2 + \dots + n_8 = 23)$$

n_t 为 t 科样本数 \bar{x}_{ij} 为 t 科 j 指标 (特征属性) 的平均值。这里与其他算法不同，没有用 j 指标的所有对象的平均值进行标准化处理，目的就是减少各科之间的平均化，突出各科的特征属性的特性。

(2) 计算科内标准差：计算 t 科内 j 指标的标准差：

$$S_{ij} = \sqrt{\frac{\sum_{i=1}^{n_t} (T_{ij} - \bar{T}_{ij})^2}{n_t}} \quad \bar{T}_{ij} \text{ 为 } T_{ij} \text{ 在 } t \text{ 科内的平均值}$$

$$\bar{T}_{ij} = \frac{\sum_{i=1}^{n_t} T_{ij}}{n_t}$$

S_{ij} 越小，反映了 j 指标在 t 科上分布越均衡稳定，代表性就越好。利用表 1，通过以上运算形成表 2。

表 2 8 科各指标科内标准差 (S_{ij})
Table 2 Standard deviation (S_{ij}) of 8 families of insects

科编号 Family no. (t)	1 (A)	2 (P)	3 (XL)	4 (YL)	5 (F)	6 (B)	7 (S)	8 (C)	9 (R)	10 (E)	11 (H)
1	0.27	0.167	0.111	0.167	0.032	0.068	0.188	0.096	0.052	0.056	0
2	0.429	0.242	0.202	0.221	0.014	0.1	0.146	0.0065	0.052	0.022	0.045
3	0.28	0.114	0.152	0.961	0.202	0.066	0.049	0.083	0.048	0.031	0.795
4	0.291	0.152	0.111	0.127	0.047	0.200	0.072	0.085	0.039	0.035	1.129
5	0.205	0.114	0.109	0.113	0.122	0.219	0.224	0.098	0.040	0.094	0.900
6	0.008	0.006	0	0.002	0.004	0	0.003	0.001	0.004	0.001	0.019
7	0.029	0.017	0.017	0.015	0.005	0.011	0.021	0.017	0.004	0.001	0.417
8	0.662	0.363	0.345	0.291	0.009	0.072	0.049	0.130	0.059	0.065	1.301

(3) 计算科之间的标准差： S_{ij} 只反映了科内总体分布情况，并没有反映科之间的差异。由于标准化的处理，使得各科之间的特征值的绝对值大小被掩盖，为了反映指标绝对值在各科上的差异，有必要计算各科均值系数：

$$k_{ij} = \frac{\bar{x}_{ij}}{x_j} \quad \bar{x}_j = \frac{\sum_{i=1}^8 \bar{x}_{ij}}{8}$$

$$\bar{x}_{ij} = \frac{\sum_{i=1}^{n_t} \bar{x}_{ij}}{n_t} \quad (i = 1, 2, \dots, n_t; j = 1, 2, \dots, 11,$$

$t = 1, 2, \dots, 8, t$ 为科数 $n_1 + n_2 + \dots + n_8 = 23$)
 \bar{x}_{ij} 为 t 科 j 指标值的均值。 k_{ij} 为均值系数。我们希望 S_{ij} 越小越好，但同时又希望 k_{ij} 之间差异越大越好，因此，我们采用科内标准差与均值系数标准差比值法评价指标的优劣。

$$d_j = \sqrt{\frac{\sum_{t=1}^8 (k_{ij} - \bar{k}_j)^2}{8}} \quad \bar{k}_j = \frac{\sum_{t=1}^8 k_{ij}}{8}$$

d_j 越大越好，反映了 j 指标在各科上的差异越大， j 指标作为分类指标灵敏性高。

我们用综合评价指标 $h_j = \frac{\sum_{t=1}^8 S_{ij}}{d_j}$ 来评价 j 指标， h_j 越小说明 j 指标在科上越均匀，在科之间分布越有差异。

(4) 计算指标之间的标准差：

$$p_t = \sqrt{\frac{\sum_{j=1}^{11} (k_{ij} S_{ij} - \bar{S}_t)^2}{8}} \quad \bar{S}_t = \frac{\sum_{j=1}^{11} S_{ij} k_{ij}}{11}$$

p_t 越小越好，反映了各指标值在 t 科上分布均衡稳定，代表性越好。

2 结果与分析

利用上述方法 , 可获得科内标准差(S_j)表(表

表 3 8 科各指标均值系数(k_j)

Table 3 Average coefficients(k_j) of indices of 8 families of insects

科编号 Family no. (t)	指标编号 Index no. (j)										
	1 (A)	2 (P)	3 (XL)	4 (YL)	5 (F)	6 (B)	7 (S)	8 (C)	9 (R)	10 (E)	11 (H)
1	0.16	0.373	0.315	0.821	0.623	1.415	1.467	0.169	1.311	1.128	0.846
2	0.157	0.438	0.300	0.884	0.984	1.135	0.937	0.189	1.491	1.188	0.701
3	0.641	1.159	1.331	0.917	1.613	0.307	0.540	2.150	0.342	0.251	0.964
4	4.397	2.923	3.170	1.393	1.453	0.193	0.297	3.062	0.213	0.316	1.565
5	2.061	1.850	1.963	1.011	1.225	0.346	0.580	0.512	0.249	0.381	1.326
6	0.228	0.433	0.351	1.039	0.603	1.556	1.530	0.271	1.498	1.323	0.963
7	0.193	0.390	0.349	0.913	0.579	1.633	1.822	0.804	1.320	1.128	0.910
8	0.161	0.433	0.221	1.023	0.920	1.419	0.833	0.842	2.376	2.289	0.725

表 4 8 科 11 指标科内标准差与均值系数标准差比值(h_j)

Table 4 Ratio of standard deviation(h_j) and averaging coefficients of 11 indices among 8 families of insects

	指标编号 Index no. (j)										
	1 (A)	2 (P)	3 (XL)	4 (YL)	5 (F)	6 (B)	7 (S)	8 (C)	9 (R)	10 (E)	11 (H)
h_j	1.530	1.338	1.037	11.495	1.166	1.280	1.472	0.524	0.416	0.498	16.45

表 5 8 科 11 指标之间标准差(p_t)

Table 5 Standard deviation(p_t) of 11 indices among 8 families of insects

	科编号 Family no. (t)							
	1	2	3	4	5	6	7	8
p_t	0.079	0.056	0.288	0.556	0.321	0.095	0.106	0.247

从表 4 可以看出 , 9 号(似圆度)、10 号(偏心率)、8 号(圆形性)指标最小 ; 3 号(横轴长)、5 号(形状参数)、6 号(叶状性)指标次之 ; 1 号(面积)、2 号(周长)、7 号(球状性)指标再次之 ; 4 号(纵轴长)、11 号(亮斑数)指标最大。即 9 号(似圆度)、10 号(偏心率)、8 号(圆形性)指标是主要的。

从表 5 可以看出 , 1 科(蜻科)、2 科(缘蜻科)值相近且最小 ; 6 科(鳃金龟科)、7 科(丽金龟科)、8 科(天牛科)值相近且次之 ; 3 科(夜蛾科)、4 科(大蚕蛾科)、5 科(粉蝶科)值相近且最大。

3 结论与讨论

通过上面的分析得出如下结论 , 即在昆虫科阶元上分类时其指标的重要性依次是 : (似圆度、偏心率、圆形性) > (横轴长、形状参数、叶状性) > (面

积、周长、球状性) > (纵轴长、亮斑数)。科的亲缘关系远近结果显示 : 蜻科和缘蜻科关系较近 > 丽金龟科、天牛科与鳃金龟科关系较近 > 夜蛾科、大蚕蛾科和粉蝶科关系较近。

沈佐锐等(2003)通过对各科各指标的假设检验法所得结论是 : (似圆度、偏心率) > (面积、周长、横轴长、球状性) > (纵轴长、圆形性) > (形状参数、叶状性) > 亮斑数。夜蛾科等 3 个科的亲缘关系远近为夜蛾科与粉蝶科 > 大蚕蛾科与粉蝶科 > 夜蛾科与大蚕蛾科 ; 鳃金龟科等 3 科的亲缘关系远近为鳃金龟科与天牛科 > 丽金龟科与天牛科 > 鳃金龟科与丽金龟科。

可以看出 , 本文结论与沈佐锐等人的指标重要性的结论基本相似 , 科亲缘关系上有所不同。双重标准差法作为一种创新研究的方法 , 既有其原理的合理性 , 又有其计算的简便性 , 而且还包含了研究对

象信息的综合性。实验的结果证明,双重标准差法不失为一种有效的方法,但有待于进一步的完善。

参 考 文 献 (References)

- Du RQ, Zhang ZT, Liu GL, 2006. Application of rough-set theory and fuzzy clustering analysis in insect taxonomy. *Acta Entomol. Sin.*, 49(1): 106 - 111. [杜瑞卿, 张征天, 刘光亮, 2006. 粗糙集模糊聚类分析法在昆虫分类学上的应用. 昆虫学报, 49(1): 106 - 111]
- Shen ZR, Zhao HQ, YU XW, 2003. Use of math morphology features in insect taxonomy. III. At the family level. *Acta Entomol. Sin.*, 46(3): 339 - 344. [沈佐锐, 赵汗青, 于新文, 2003. 数学形态学在昆虫分类学上的应用研究. III. 在科级阶元上的应用研究. 昆虫

学报, 46(3): 339 - 344]

- Zhao HQ, Shen ZR, YU XW, 2003a. Use of math morphology features in insect taxonomy. I. At the order level. *Acta Entomol. Sin.*, 46(1): 45 - 50. [赵汗青, 沈佐锐, 于新文, 2003a. 数学形态学在昆虫分类学上的应用研究. I. 在目级阶元上的应用研究. 昆虫学报, 46(1): 45 - 50]

- Zhao HQ, Shen ZR, YU XW, 2003b. Use of math morphology features in insect taxonomy. II. At the superfamily level. *Acta Entomol. Sin.*, 46(2): 201 - 208. [赵汗青, 沈佐锐, 于新文, 2003b. 数学形态学在昆虫分类学上的应用研究. II. 在总科级阶元上的应用研究. 昆虫学报, 46(2): 201 - 208]

(责任编辑:袁德成)