统计显著性标记的聚类分析算法与网络实现*1)

张文军　冯永军　古德祥

(中山大学昆虫学研究所与生物防治国家重点实验室　广州 510275)

摘　要

聚类分析方法应用广泛，但过程及结果缺乏可靠的统计学检验，数学上不严格．另外，用于聚类分析的数据分布类型复杂多样，往往无法确定，而经典统计检验方法设定了各种统计前提和假设，应用依据不足．鉴于此，本研究用随机化方法对分类进行统计显著性检验，建立了具有统计显著性标记的聚类分析算法，用于对若干个样品进行有显著性标记的聚类分析．该算法包括数据加权与规范化，计算距离测度，系统聚类，及随机化统计检验等过程．在该算法中，有 14 种距离测度、5 种系统聚类方法、3 种数据规范化方法及指标加权与否可供选择．随机化检验不需统计前提和假设，适用于各种统计问题．算法用 Java 语言网络化实现，包含 6 个类和一个 HTML 文件．可通过网络在多种 Java 兼容的浏览器上实现算法共享．以水稻田无脊椎动物多样性的调查数据，对该算法进行了对比分析，给出了选择距离测度的一些原则．

关键词：聚类分析，统计显著性，随机化检验，距离测度，算法与实现

ALGORITHM AND IMPLEMENTATION OF A CLUSTER ANALYSIS WITH STATISTIC SIGNIFICANCE

Zhang Wenjun　Feng Yongjun　Gu Dexiang

(*Research Institute of Entomology and State Key Laboratory for Biological Control, Zhongshan University, Guangzhou* 510275)

Abstract

Algorithms of cluster analysis are not always statistically tested. An algorithm of cluster analysis with statistic significance test to classification was developed in this paper. It was made of four parts, i.e., data weighting and standardization, calculation of distance measures, hierarchical clustering, and randomization statistic test. Fourteen distance measures and five methods of hierarchical clustering were provided in the algorithm to be choosed. The algorithm was implemented as the network program with Java language, which made of 6 Java classes and a HTML file and can be loaded and run on Java-enabled web browsers. The algorithm was

* 2003 年 12 月 26 日收到.

1) 国家自然科学基金 "农田有害生物可持续管理的生物多样性计算方法及软件"(30170184)，教育部留学回国人员科研基金 "农田有害生物管理的生物多样性定量算法及计算软件研制"(2000).

tested with investigation data of rice invertebrate diversity. Principles for choosing distance measures was explained.

Key words: cluster analysis, distance measures, randomization statistic test, algorithm, implementation

聚类分析方法应用极为广泛，在计算机领域，用于模式识别与分类；在地质学领域，用于地层分类，矿产资源分类等；在地理学领域，用于地理气候区划；在农业领域，用于农业区划，品种分类等；在生物学领域，用于同源性分析，生物进化研究，形态学分类，群落分析等等. 聚类分析的应用已深入到各个角落. 根据相似性进行聚类是人类最早认识和应用的科学思想，是人类认识自然的基本工具之一. 迄今为止，已提出了众多聚类分析方法，并得到了广泛应用. 但在数学原理上没有严格性 (张尧庭，方开泰，1982). 主要原因是，聚类过程及其结果缺乏可靠的统计学标准，无法在统计学上确定分类是否可信. 另一方面，经典统计学方法依赖的统计前提和假设在绝大多数聚类分析数据中并不成立，也不易确定. 随机化检验不需统计前提和假设，适用于各种统计问题，已得到了广泛应用 (Solow, 1993, Manly, 1997, Zhang and Schoenly, 1999, Zhang et al., 2001, 2002). 本文用随机化方法对分类进行统计显著性检验，建立了具有统计显著性标记的聚类分析算法，用于对若干个样品进行有显著性标记的聚类分析. 算法用 Java 语言网络化实现，为具有统计显著性标记的聚类分析提供一种在线计算工具.

§1. 统计显著性标记的聚类分析算法

该算法的目的是对若干个样品进行聚类分析，并对分类合理性给出统计显著性标记.

算法主要包括四部分内容：数据加权与规范化，计算距离测度，系统聚类，随机化统计检验.

1.1 数据加权与规范化

设有 m 个指标 (分类单元), n 个样品，结果得一 $n \times m$ 取值矩阵 (y_{ij}).

若需对指标给定权重，各指标的权重为 $w_1, w_2, \ldots, w_m$, 且有 $\sum_{i=1}^{m} w_i$=1, 则加权后的取值矩阵为 $y_{ij} = w_j \times y_{ij}$, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m$.

若选择定量的距离测度，如 Minkowski 距离和相关系数距离，则可对数据进行规范化. 有 3 种规范化方法，分别为：

标准差法

$$y_{ij} = (y_{ij} - y_{bj})/s_j,$$

其中 $y_{bj} = \sum_{i=1}^{n} y_{ij}/n$, $s_j = (\sum_{i=1}^{n}(y_{ij} - y_{bj})^2/(n-1))^{1/2}$, $j = 1, 2, \ldots, m$. 经过规范化，各指标均值为 0, 标准差为 1.

最大最小法

$$y_{ij} = (y_{ij} - y_j^{min})/(y_j^{max} - y_j^{min}),$$

其中 $y_j^{max} = \max y_{ij}$, $y_j^{min} = \min y_{ij}$, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m$. 经过规范化后，就有 $0 \leq y_{ij} \leq 1$, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m$.

比率法

$$y_{ij} = y_{ij}/y_j^{max}.$$

规范化后的各指标最大值为 1.

1.2 距离测度计算

距离测度用以表示样品之间的相似性程度. 距离值小, 则样品之间相似性大. 本算法中有四类, 共 14 种距离测度可供选择 (张尧庭, 方开泰, 1982, Krebs, 1989, 齐艳红等, 2003).

第一类是 Minkowski 距离, 包括欧氏距离 (Euclidean Distance), 曼氏距离 (Manhattan Distance) 和切比雪夫距离 (Chebyshov Distance),

$$\begin{aligned} d_{ij} &= [\textstyle\sum_{k=1}^{m}(y_{ik}-y_{jk})^2/m]^{1/2}, && i,j=1,2,\ldots,n,\\ d_{ij} &= \textstyle\sum_{k=1}^{m}|y_{ik}-y_{jk}|/m, && i,j=1,2,\ldots,n,\\ d_{ij} &= max_k|y_{ik}-y_{jk}|, && i,j=1,2,\ldots,n. \end{aligned}$$

第二类是相关系数距离, 包括以简单相关系数 (Correlation Coefficient) 和夹角余弦 (Angular Cosine) 为基础的距离

$$d_{ij} = 1-\textstyle\sum_{k=1}^{m}[(y_{ik}-y_{ib})(y_{jk}-y_{jb})]/[\sum_{k=1}^{m}(y_{ik}-y_{ib})^2\sum_{k=1}^{m}(y_{jk}-y_{jb})^2]^{1/2},\ i,j=1,2,\ldots,n.$$

$$d_{ij} = 1-\textstyle\sum_{k=1}^{m}(y_{ik}y_{jk})/[\sum_{k=1}^{m}y_{ik}^2\sum_{k=1}^{m}y_{jk}^2]^{1/2},\ i,j=1,2,\ldots,n.$$

前两类属于定量指标, 即指标取值连续的距离测度, 如气温, 降雨量, 植被盖度, 等等.

第三类是多值定性距离, 包括以联列系数 (Link Coefficient) 和三种连关系数 (Colink Coefficient 1, Colink Coefficient 2, Colink Coefficient 3) 为基础的距离

$$\begin{aligned} d_{ij} &= 1-[x^2/(x^2+n_{..})]^{1/2},\\ d_{ij} &= 1-[x^2/(n_{..}max(p-1,q-1))]^{1/2},\\ d_{ij} &= 1-[x^2/(n_{..}min(p-1,q-1))]^{1/2},\\ d_{ij} &= 1-\{x^2/[n_{..}\sqrt{((p-1)(q-1))}]\}^{1/2}, \end{aligned}$$

其中 $x^2 = n_{..}[\sum_{i=1}^{p}\sum_{j=1}^{q}n_{ij}^2/(n_{i.}n_{.j})-1]$, $n_{..}=\sum_{i=1}^{p}n_{i.}$, $n_{i}=\sum_{j=1}^{q}n_{ij}$, $n_{.j}=\sum_{i=1}^{p}n_{ij}$. 样品 i 取值为 $t_1,t_2,\ldots,t_p$, 样品 j 取值为 $r_1, r_2, \ldots, r_q$. n_{kl} 为 i 取 t_k, j 取 r_l 的指标数.

第四类是二值定性距离, 包括以点相关系数 (Point Correlation Coefficient) 、四分相关系数 (Quadratic Correlation Coefficient) 、两种夹角余弦 (Angular Cosine 1, Angular Cosine 2) 为基础的距离, 以及 Jaccard 距离

$$\begin{aligned} d_{ij} &= 1-(ad-bc)/[(a+b)(c+d)(a+c)(b+d)]^{1/2},\\ d_{ij} &= 1-\sin[(a+d-(b+c))/(a+b+c+d)*3.1415926/2],\\ d_{ij} &= 1-[a*a/((a+b)(a+c))]^{1/2}, && i,j=1,2,\ldots,n,\\ d_{ij} &= 1-[a*a*d*d/((a+b)(a+c)(b+d)(c+d))]^{1/2},\\ d_{ij} &= (b+c)/(b+c+d), \end{aligned}$$

其中 a 为 m 个指标中样品 i 和 j 同取 0 的个数， d 为样品 i 和 j 同取 1 的个数， b 为样品 i 取 0 且样品 j 取 1 的个数， c 为样品 i 取 1 且样品 j 取 0 的个数.

第三类和第四类距离测度适用于指标为定性指标的情形，其中多值定性距离用于多值定性指标，后者如 DNA 序列碱基类型，蛋白质肽链氨基酸类型，等等. 二值定性距离用于二值定性指标，二值定性指标的例子包括二进制序列 0-1 信号，有 - 无，高 - 低，好 - 坏等等.

1.3 系统聚类

选用 5 种聚类方法：最短距离法，最长距离法，类平均法，重心法，离差平方和法 (张尧庭，方开泰， 1982, Krebs, 1989). 设有类 A 和类 B, 则对最短距离法、最长距离法、类平均法、重心法及离差平方和法. 类 A 和类 B 之间的距离分别为

$$
\begin{aligned}
&D_{AB} = mind_{ij} && i \in A,\ j \in B,\\
&D_{AB} = maxd_{ij} && i \in A,\ j \in B,\\
&D_{AB} = [(\textstyle\sum_i \sum_j d_{ij}^2)/(n_A n_B)]^{1/2}, && i \in A,\ j \in B,\\
&D_{AB} = d_{AB}(\textstyle\sum\sum d_{ij}/n_A,\ \sum\sum d_{ij}/n_B),\\
&D_{AB} = r - p - q,
\end{aligned}
$$

其中 n_A 和 n_B 分别为类 A 和类 B 中的样品个数. p, q, r 为类 A, 类 B 及其合并类的离差平方和. 聚类开始时， n 个样品各自成一类. 在类集合中，选距离值 D_{AB} 最小的两类进行聚合，得一新类， ...，如此类推，直到 n 个样品聚为同一类.

1.4 随机化统计检验

聚类过程的对偶是分类过程. 设某父类已划分为两个子类，类 A 和类 B. 随机化检验过程是 (Manly, 1997, Zhang et al., 2001,2002), 将父类样品随机划分为各有 n_A 和 n_B 个样品的两类，虚类 A 和虚类 B. 计算虚类 A 和虚类 B 之间的距离值，比较是否不小于实类间的距离值. 重复该过程 w 次，设计算值不小于实际值的次数为 v, 令统计检验值 $p = v/w$. 若 p 小于临界值，如 0.01, 或 0.05, 或 0.1 等，则认为分类有统计显著性，该父类划分为类 A 和类 B 是统计上可接受的. 反之，若 p 大于临界值，则认为分类无统计显著性.

本算法的特点是，用随机化方法对分类合理性进行严格的统计显著性检验. 随机化检验不需统计前提和假设，适用于各种统计问题.

§2. 算法实现

本算法用 Java 语言网络化实现，包含 6 个类和一个 HTML 文件.

(1) StatTestCluster 类：在该类中进行计算，并调用其它类完成有关任务. 该类的 Applet 被载入 Java 兼容浏览器后，显示输入窗口. 内容包括：选择距离测度，选择聚类方法，选择指标加权与否，选择数据规范化方法，输入样品数，指标数，随机化次数 (如 100, 1000 等)，临界 p 检验值 (如 0.1, 0.05), 打开聚类数据文件 (图 1).

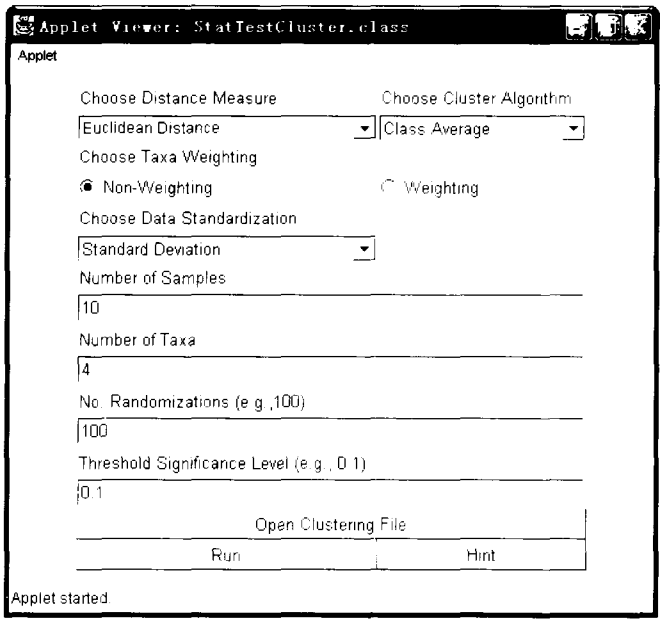

图 1 参数输入窗口 (StatTestCluster 类)

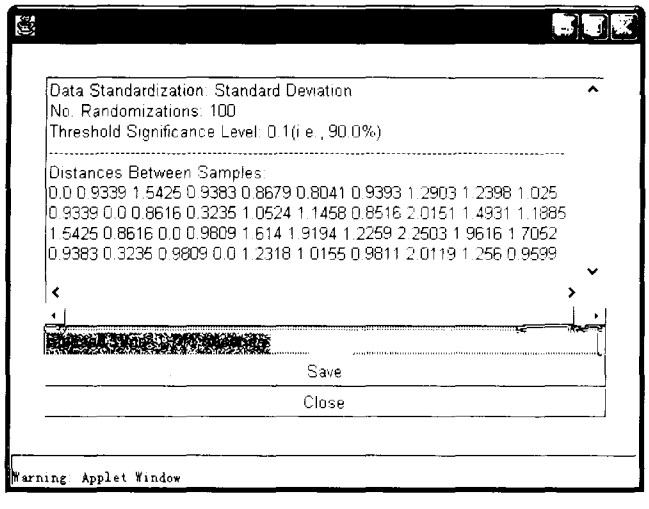

图 2 结果输出窗口 (ResultShow 类)

(2) ResultShow 类 (图 2), Hint 类, GraphicsFrame 类, 以及 WarningShow 类: 见文献所述 (齐艳红等, 2002, 2003, 张文军等, 2002).

(3) ClusterGraphics 类: 该类输出样品的聚类树状图. 通过 ClusterGraphics 类窗口上的 "Rotate" 按钮, 可对聚类树状图进行四个方向的旋转 (图 3).

(4) StatTestCluster.html 文件: 该文件将 StatTestCluster 类载入 Web 浏览器, 并传入窗口大小及组件大小参数.

在聚类数据文件中，第一行为各样品的编号. 以后每行第一个值为指标编号，若选择给定指标权重，则每行第一个值为各指标的权重值. 该行中其余值为该指标下各样品的取值. 原始数据文件为 MS-DOS 文本文件 (.txt). 可在 MS-DOS 的文本编辑器中编辑，或在 Windows 中选开始 → 程序 → 附件 → 记事本，在记事本中编辑文件.

算法运行后输出聚类结果，每次分类过程下各分类的显著性水平，样品的聚类树状图.

§3. 聚类分析数据

于 2003 年 6 月 12 日在清新县三坑镇水稻田取样调查无脊椎动物多样性. 用汽油发动机吸虫器收集稻田取样点内的无脊椎动物，取 10 个样点. 带回室内鉴定计数各样点的无脊椎动物种类和数量. 最后，合并为捕食性无脊椎动物，寄生性无脊椎动物，植食性无脊椎动物，及中性无脊椎动物在各样点的数量，得到聚类原始数据文件见表 1.

表 1　水稻田无脊椎动物数量取样调查结果 (清新，2003.6.12)

	1	2	3	4	5	6	7	8	9	10
捕食性无脊椎动物	9	12	9	13	9	14	9	2	13	13
寄生性无脊椎动物	10	15	20	15	8	10	16	11	9	10
植食性无脊椎动物	79	47	51	75	17	91	38	122	168	141
中性无脊椎动物	679	442	221	458	375	895	862	1083	333	349

§4. 结果分析

4.1　统计显著性标记对分类结果选择的影响

聚类原始数据文件中共有 10 个样品，4 个无脊椎动物分类单元 (指标). 以本算法进行聚类分析. 取欧氏距离，类平均法，分类单元不加权，数据以标准差方法规范化，随机化 100 次，临界 p 检验值为 0.1(图 1). 正如下述结果所示，这些分类是统计上显著的，可信的：(1 2 4 5 6 7 9 10)→(1 2 4 5 6 7)+(9 10), (2 4 7)→(2 4)+(7), (1 6)→(1)+(6), (2 4)→(2)+(4), (9 10) →(9)+(10).

Clustering Distance=0.0
(1) (6) (2) (4) (7) (5) (9) (10) (3) (8)
p=0.0*
Clustering Distance=0.3094
(1) (6) (2) (4) (7) (5) (9 10) (3) (8)
p=0.0*
Clustering Distance=0.3235
(1) (6) (2 4) (7) (5) (9 10) (3) (8)
p=0.0*
Clustering Distance=0.8041
(1 6) (2 4) (7) (5) (9 10) (3) (8)
p=0.0*

Clustering Distance=0.9186
(1 6) (2 4 7) (5) (9 10) (3) (8)
p=0.88
Clustering Distance=1.0311
(1 2 4 6 7) (5) (9 10) (3) (8)
p=0.83
Clustering Distance=1.1957
(1 2 4 5 6 7) (9 10) (3) (8)
p=0.05*
Clustering Distance=1.3828
(1 2 4 5 6 7 9 10) (3) (8)
p=0.67
Clustering Distance=1.5261
(1 2 3 4 5 6 7 9 10) (8)
p=0.92
Clustering Distance=1.8961
(1 2 3 4 5 6 7 8 9 10)

显然，具有统计显著性标记的聚类分析与传统聚类分析的可用结果有所不同. 前者包含了更多的可用信息，使结果的可选性和可信性增强.

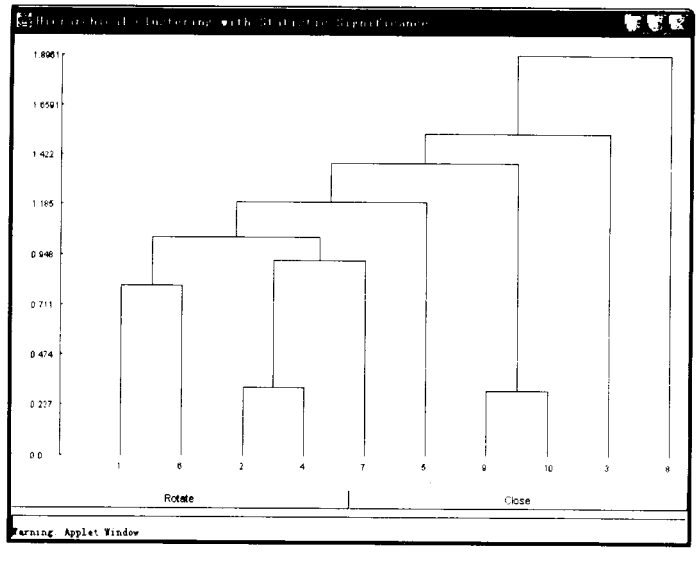

图 3 水稻田无脊椎动物多样性数据 (10 个样品， 4 个无脊椎动物分类单元) 的聚类树状图.
取欧氏距离，类平均法，分类单元不加权，数据以标准差方法规范化，随机化 100 次，临界 p 检验值为 0.1

4.2 不同聚类方法对分类结果选择的影响

取重心法，其它与 4.1 中相同，则如下分类是统计上显著的，可信的： (1 2 3 4 5 6 7 9 10)→(1 2 3 4 5 6 7)+(9 10), (1 2 3 4 5 6 7)→(1 5 6 7)+(2 3 4), (1 6 7)→(1 6)+(7).

(1 6)→(1)+(6), (9 10)→(9)+(10); 取最短距离法，其它条件相同，则如下分类是统计上显著的： (1 2 4 5 6 7 9 10)→(1 2 4 5 6 7)+(9 10), (2 4 7)→(2 4)+(7), (1 6)→(1)+(6). (2 4)→(2)+(4), (9 10)→(9)+(10).

显而易见，这些结果并不相同. 5 种聚类方法的比较说明，随聚类方法选择的不同，分类结果选择将会有程度不同的差异.

4.3 不同距离测度对分类结果选择的影响

取 Manhattan 距离，其它与 4.1 中相同，则如下分类是统计上显著的： (1 2 3 4 5 6 7 8 9 10)→(1 2 3 5 7)+(4 6 8 9 10), (1 2 3 5 7)→(1 3 5 7)+(2), (4 6 8 9 10)→(4 6 9 10)+(8), (1 3 5 7)→(1 3 5)+(7), (4 6 9 10)→(4 9 10)+(6), (1 3 5)→(1 3)+(5). (4 9 10)→(4)+(9 10), (1 3)→(1)+(3), (9 10)→(9)+(10), 即所有分类在统计上显著. 计算表明，不同距离测度的选择对分类结果选择会产生重大影响.

4.4 数据加权与否对聚类结果的影响

选择分类单元加权 (按重要性大小，捕食性无脊椎动物、寄生性无脊椎动物、植食性无脊椎动物及中性无脊椎动物的权重分别为 0.2,0.2,0.5,0.1), 其它与 4.1 中相同，则所有分类在统计上显著. 指标的相对重要性不同，将对分类结果选择产生显著影响.

4.5 数据规范化对聚类结果的影响

不对数据进行规范化，其它与 4.1 中相同，则所有分类在统计上显著，与 4.1 中结果不同. 数据规范化与否及规范化方法对分类结果选择会产生显著影响.

§5. 讨　论

5.1 选择距离测度的原则

在原始数据中，若所有的指标均为同一种类型，即全为定量指标，或全为多值定性指标，或全为二值定性指标，则可按照前面所述，选定相应的距离测度. 定量指标 (或多值定性指标) 和二值定性指标可同时使用. 此时，若选二值定性距离测度，则定量指标 (或多值定性指标) 被转化为二值定性指标进行运算，即其非零值转化为 1，而 0 值保持不变. 若选定量指标 (或多值定性指标) 的距离测度，则定量指标 (或多值定性指标) 和二值定性指标的各样品值直接进行运算. 定量指标和多值定性指标同时使用时，其各值不变并参加运算. 不同类型的指标同时使用，将造成程度不同的信息损失.

原始数据中的非零值过多或过少时，使用二值定性距离测度的效果较差.

相关系数距离测度只考虑两样品之间的相关性，而不考虑各指标的差值大小. Minkowski 距离测度考虑了两样品之间各指标的差值大小，差值大，则距离值就大. 但没有反映两样品之间的相关性.

在聚类分析中需根据指标类型和样品相似性的含义选择合适的距离测度.

5.2 随机化与结果差异

不同系统产生随机数的机制有所差异，本文聚类分析中统计显著性的结果在不同系统下可能会有微细差异. 增大随机化次数，可提高结果的准确性.

参 考 文 献

[1] 齐艳红，张文军，有害生物侵扰在多样化生境中的一种随机扩散过程及网络计算软件，现代计算机， 133(2002) 16-19.

[2] 齐艳红，图书期刊评价分析的混合优序图及网络计算软件研究，现代计算机， 151(2002) 14-16,56.

[3] 齐艳红，张文军，CorreDetector: 一种用于信息资料相关性分析的网络共享软件，情报学报，22(Suppl.)(2003) 266-268

[4] 齐艳红，网络计量学的一种 Internet 分布式聚类分析软件，情报科学， 21:10(2003) 1069-1071,1079.

[5] 张文军，齐艳红，生物多样性和均匀度显著性的统计检验及网络计算软件，现代计算机， 145(2002)，6-9.

[6] 张尧庭，方开泰，多元统计分析引论，科学出版社，北京， 1982.

[7] Krebs. C.J., Ecological Methodology. Happer Collins Publishers, Inc. New York (1989).

[8] Manly, B.F.J., Randomization, Bootstrap and Monte Carlo Methods in Biology (Second Edition). Chapman & Hall, London, UK, (1997).

[9] Zhang, W.J., Schoenly, K.G. 1999. IRRI Biodiversity Software Series. II. COLLECT1 and COLLECT2: programs for calculating statistics of collectors' curves. IRRI Technical Bulletin No.2. Manila (Philippines): International Rice Research Institute.

[10] Zhang, W.J., Schoenly, K.G. 1999. IRRI Biodiversity Software Series. III. BOUNDARY: a program for detecting boundaries in ecological landscapes. IRRI Technical Bulletin No.3. Manila (Philippines): International Rice Research Institute.

[11] Zhang, W.J., Schoenly, K.G. 1999. IRRI Biodiversity Software Series. IV. EXTSPP1 and EXTSPP2: programs for comparing and performance-testing eight extrapolation-based estimators of total taxonomic richness. IRRI Technical Bulletin No.4. Manila (Philippines): International Rice Research Institute.

[12] Zhang, W.J., K.G. Schoenly, A randomization test and software to compare ecological communities. International Rice Research Notes, 26:2(2001) 48-49.

[13] Zhang, W.J., Y.H. Qi, K.G. Schoenly, Randomization tests and computational software on significance of community biodiversity and evenness. Biodiversity Science, 10:4(2002) 431-437.