技术与方法

新型数值分类软件 X-Cluster 的开发及应用

黄继翔 惠 明 齐东梅 牛天贵*

(中国农业大学食品科学与营养工程学院 北京 100083)

摘要：针对目前常用数值分类软件的不足，采用多种设计模式开发了新型数值分类软件 X-Cluster。该软件具有界面友好、操作方便、体积小、速度快、功能强大、应用范围广等特点，能够满足大多数情况下数值分类研究工作的需求，并在芽孢杆菌分类研究中得到了验证。

关键词：数值分类，软件开发，芽孢杆菌

中图分类号：Q332 **文献标识码**：A **文章编号**：0253-2654(2006)01-0118-04

Development and Application of X-Cluster: a New Software for Numerical Classification

HUANG Ji-Xiang HUI Ming QI Dong-Mei NIU Tian-Gui*

(*College of Food Science & Nutritional Engineering, China Agriculture University, Beijing* 100083)

Abstract: To remedy the limitations of traditional numerical classification softwares, a new application, X-Cluster, was developed by using various design patterns. X-Cluster had powerful functions to support the researching of numerical classification, and testified by some classify studying about *Bacillus* spp..

Key words: Numerical classification, Software development, *Bacillus* spp.

数值分类是根据微生物分类学信息，应用计算数学原理和技术辅助定义微生物分类单位的方法[1]。目前广泛使用的数值分类软件主要有2类：MINTS[2]、SPSS[3]或SAS[4]，尤其是中国科学院微生物研究所开发的MINTS程序，对我国生态学及微生物分类研究起到了重要作用[5~7]，但二者均存在一些不足之处：MINTS使用QuickBasic编程语言开发，运行于DOS系统，导致操作繁琐、人机界面不友好、数据处理量有限等缺陷；SPSS、SAS作为大型专业数据处理软件，要求使用者具有一定的专业知识，一般用户使用不便，对微生物数值分类研究中的某些功能缺乏支持，而且软件体积庞大、对硬件系统要求较高。针对以上不足，开发一种界面友好、操作方便、体积小、速度快、功能强大的新型数值分类软件对于微生物及相关领域的研究具有重要的现实意义和良好的应用前景。

1 设计思想

1.1 软件架构

为提高软件的灵活性和可扩展性，在宏观架构上将数据容器和数据处理功能相分离，软件则相应分为主程序和插件两个部分，主程序提供数据容器功能，插件提供数据处理功能。主程序和插件均可独立变化，使软件的灵活性得到提高；提供具有更多数据处理功能的插件则体现了软件的可扩展性。在具体架构上，主程序主要包括3个

* 通讯作者 Tel：010-62737045，E-mail：niutiangui@163.com

收稿日期：2005-04-20，修回日期：2005-05-23

部分：视图、视图管理器、插件管理器。其中，视图是数据容器的实现，在接口设计上采取 MVC（模型/视图/控制器，Model/View/Controller）三元组[8]一体化，在其内部保持 MVC 各部分独立性，该设计在保留 MVC 设计模式优点的同时减少了与视图管理器进行交互的复杂性；视图管理器提供视图的新建、打开、保存等功能，视图管理器和视图在结构设计上采取抽象工厂模式[6]，为软件升级与个性化提供支持；插件管理器负责插件的打开、调用、关闭等功能。插件以动态链接库形式存在，在程序运行中可根据数据处理的需求而动态地加载和关闭，将不需要的插件关闭可有效降低系统资源的使用数量，加快运行速度。

1.2　数值分类过程中的设计模式

数值分类过程主要包括数据标准化、分类单元间亲缘关系的计算、分类运算 3 个步骤，每个步骤均可选用多种算法完成[9,10]，对其进行抽象后可归纳为策略模式[8]。在策略模式中，定义一系列算法并将其封装为对象，算法调用者根据情况进行选择，可降低软件的复杂程度。在软件设计中，定义 IWorker 为算法对象基类，IWorkObj 为算法对象管理器基类。对应与数值分类过程的 3 个步骤，IWorker 派生出数据标准化基类 IMatStd、分类单元间亲缘关系计算基类 ICalaRela、分类运算基类 ISysCluster；IWorkObj 也派生出相应的算法管理器类。

1.3　数据显示

在主程序中，数据结构为矩阵，采用表格组件显示。在数值分类过程中，数据结构包括矩阵和树，矩阵的显示与主程序相同，树的显示采用树状视图组件。在树状视图中，一个节点代表一个类群，可以直观的显示分类运算得到的树状图。

2　软件开发及应用

2.1　开发环境

在 Microsoft Windows2000、Borland Delphi6 环境下使用 Object Pascal 编程语言编写了主程序 NatureEdge、数值分类插件 X-Cluster。

2.2　工作流程

软件的工作流程见图 1。

2.3　功能及特点

主程序 NatureEdge 的主要功能为数据的管理和编辑、插件的管理和调用。

数值分类插件 X-Cluster 的主要功能包括：具有用户自定义筛选和自动筛选 2 种数据筛选模式；支持 4 种数据标准化方法、23 种分类单元间距离和相似性计算方法、12 种分类算法；树状图的查看和编辑；以树状图中分类单元顺序查看并输出某类群中原始数据、标准化后数据、类群中分类单元间亲缘关系；计算多个类群的中心分类单元；设置树状图输出绘制参数，以完全、分支、可见等多种方式绘制树状图。

由于在设计和开发过程中采用了多种设计模式和程序编写技巧，使软件具有操作方便、界面友好、应用范围广、功能强大、运行速度快等特点。在 X-Cluster 的计算界面中，用户只需根据需要选择相应算法和参数并点击命令按钮既可完成分类运算，使用极其方便；不仅可用于常规的 0-1 编码二元性状数据的数值分类计算，还可应用于浮点数类型数据，具有广泛的应用范围；在分类结果显示界面（图 2）中，可单独设置树状图中某个节点的颜色、名称、字体风格等，并在树状视图和树状图绘制中体现；树状图的绘制采用矢量绘图的增强型图元格式，具有体积小、缩放下不失真的优点，便于插入 Word 等文字处理软件；在程序设计过程中采用组件自定义绘图、暂停显示更新等技巧加快了软件运行速度。

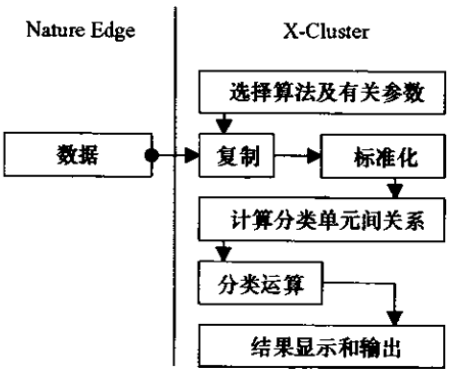

图 1 软件流程

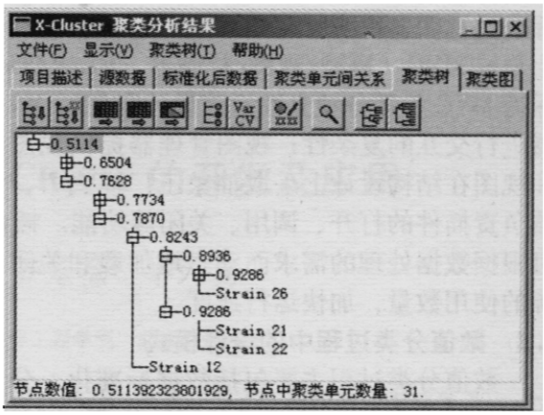

图 2 X-Cluster 聚类分析结果显示界面

2.4 软件的应用

本实验室长期从事芽孢杆菌的资源收集、分类及应用研究工作，已分离纯化 3000 余株芽孢杆菌，对部分菌株的形态、培养条件、生理生化特性进行观察和测试后得到 0-1 编码的二元性状数据，使用简单匹配系数、非加权算术配对平均法（UPGMA）进行数值分类，结果见图 3。图 3 中，测试菌株在 0.65 水平上聚为 3 个表观群，多数菌

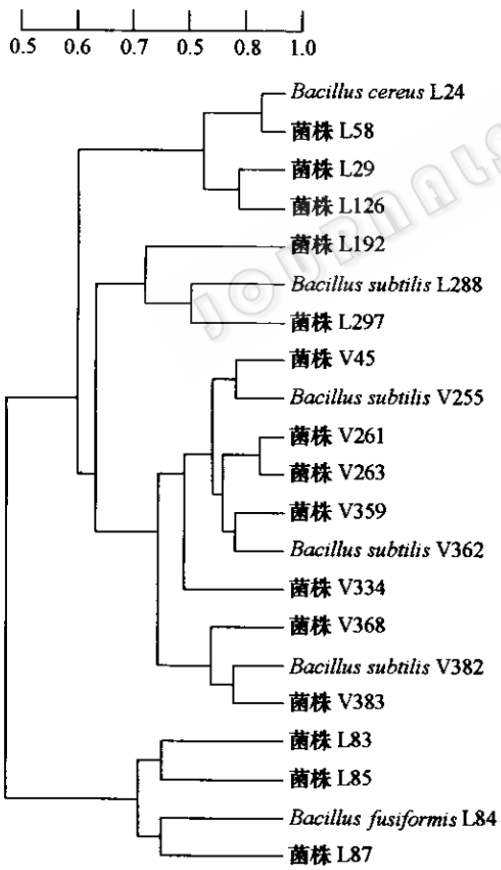

图 3 本实验室部分芽孢杆菌菌株树状图
0－1 编码的表型数据，简单匹配系数
UPGMA 方法

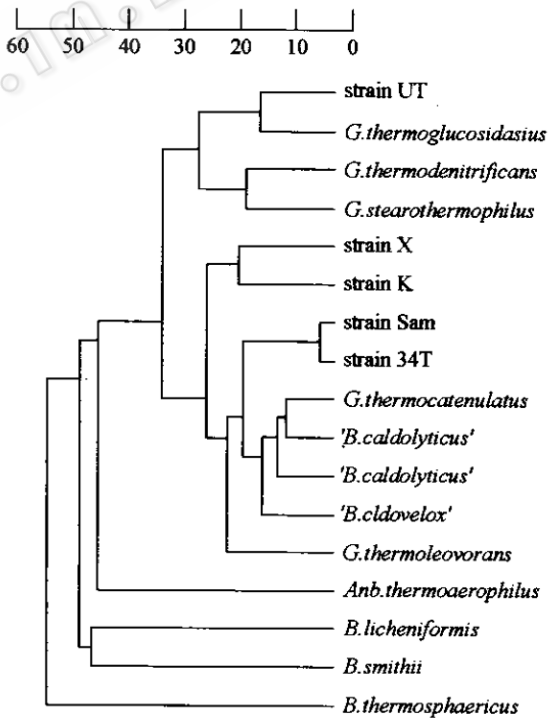

图 4 *Geobacillus* 属中部分菌株树状图
浮点数类型的细胞脂肪酸组分数据，欧氏距离
UPGMA 方法

株聚集在 *Bacillus subtilis* 表观群中并在 0.7 ~ 0.8 水平上形成几个小的亚群，说明 *B. subtilis* 是一个菌株间表型性状差异较大的种，目前 *B. subtilis* 已包括多个亚种，数值分类结果验证了这一观点。*B. cereus* 和 *B. subtilis* 表观群关系较近而与 *B. fusiformis* 表观群相距较远，与 16S rDNA 序列揭示的系统发育关系相符合。

Nazina 对包括 *B. stearothermophilus* 在内的多种嗜热芽孢杆菌进行多相分类研究，建立了 *Geobacillus* 属[11]。对该属中部分菌株的细胞脂肪酸组分数据使用欧式距离、UPGMA 方法进行数值分类研究（图 4），*Geobacillus* 属中菌株均聚集在同一类群中而与其他属中菌株保持距离，验证了该属的划分。

3　结论与展望

针对目前常用数值分类软件的不足，在面向对象的设计思想指导下开发了新型数值分类软件 X-Cluster，具有界面友好、操作方便、体积小、速度快、功能强大、应用范围广等特点。应用于芽孢杆菌的形态、培养条件、生理生化特性测试数据和细胞脂肪酸组分数据的数值分类研究并取得良好效果，表明该软件具有较高的应用价值。

数值分类是在传统的微生物描述分类基础上发展而来的学科，但其应用范围已超越其基础，成为一种具有普遍适用性的工具应用在多种学科和多种类型数据的处理过程中，例如化学分类方法和 DNA 指纹技术方法得到的数据同样可用于数值分类[5,11]。在数值分类研究中一个关键问题是测试性状（特征）的选择，选择合适的测试性状可以提高分类合理性、减少工作量，在测试性状的选择上国外学者已进行了卓有成效的工作[12]，与此有关的软件正在开发中。

参 考 文 献

[1] 刘志恒. 现代微生物学. 北京：科学出版社，2002. 46 ~ 48.
[2] 马俊才，赵玉峰. 微生物学通报，1986，**13** (5)：225 ~ 228.
[3] 张文彤. SPSS 统计分析高级教程. 北京：高等教育出版社，2004. 199 ~ 228.
[4] 洪　楠，侯　军. SAS for Windows (V8) 统计分析系统教程新编. 北京：清华大学出版社，2004. 419 ~ 433.
[5] 陈文峰，陈文新. 应用与环境生物学报. 2003，**9** (1)：53 ~ 58.
[6] 杨亚珍，韦革宏，万晓红，等. 微生物学通报，2004，**31** (2)：20 ~ 25.
[7] 韩　黎，吴桂芝，陈世平，等. 中华微生物学和免疫学杂志，1999，**19** (4)：338 ~ 341.
[8] Erich G，Richard H，Ralph J，*et al*. 李英军等译. 设计模式：可复用面向对象软件的基础. 北京：机械工业出版社，2000.
[9] 徐克学. 生物数学. 北京：科学出版社，1999. 71 ~ 115.
[10] 朱剑英. 智能系统非经典数学方法. 武汉：华中科技大学出版社，2001. 99 ~ 145.
[11] Nazina T N，Tourova T P，Poltaraus A B，*et al*. Int J Syst Evol Microbiol，2001，**51**：433 ~ 445.
[12] Reva O N，Sorokulova I B，Smirnov V V. Int J Syst Evol Microbiol，2001，**51**：1361 ~ 1371.