网上聚类分析的可视化设计与开发①

Design and Development on Visualizing Internet – Based Cluster Analysis

石彤菊　马新顺　（华北电力大学 数理系 河北 保定 071003）

摘　要： 利用 Microsoft Visual J + + 6.0 作为实验开发平台，采用 Java 语言及 Applet 嵌入网页技术，实现了互联网环境下的聚类分析可视化。所设计的可视化窗口上，能够通过范例演示聚类分析全过程，并可通过选择快速聚类法、最短距离法、最长距离法对录入数据进行聚类分析，具有数据的录入、修改、输出分类结果及显示分类谱系图等功能。

关键词： 数学实验 聚类分析 Microsoft visual J + + 6.0 Applet 小程序

1　引言

数学实验是大学数学课程的一种新的教学模式。该课程将数学知识、数学建模与计算机应用相结合，让同学们自己做实验，体验解决问题的过程，从实验中去学习，从成功和失败中去获得真知。本文的主要内容是我校基于 Visual J + + 6.0 开发的数学实验的一部分，实现聚类分析可视化设计与编程。

聚类分析是对被评价对象进行定量分类的一种多元统计分析方法。聚类的目的是挖掘数据潜在的自然分组结构和关系。聚类分析方法已经在众多的不同领域中得到应用，有效地解决了科学研究中多因素、多指标的分类问题。在本设计实现分类的过程中，主要是把原始资料列为一个矩阵，通过对矩阵进行求最大最小值、去掉相关性最小的行或列等运算来实现。通过本实验，能使同学们更好的了解聚类分析的原理、应用，并能够用它解决实际中的聚类问题。

2　聚类分析理论

2.1 原始资料矩阵

设有 n 个样品，每个样品测得 m 项指标，原始资料可用 n * m 的矩阵表示为：

$$X=\begin{pmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{pmatrix}$$

其中 $x_{ij}(i=1,\cdots\cdots n;j=1,\cdots\cdots,m)$ 为第 i 个样品的第 j 个指标的观测数据。第 i 个样品 X_i 为矩阵 X 的第 i 行所描述，所以任何两个样品 X_K 与 X_L 之间的相似性，可以通过矩阵 X 中的第 K 行和第 L 行的相似程度来刻划，任何两个变量 x_k 与 x_l 之间的相似性，可以通过第 K 列和第 L 列的相似程度来刻划。

2.2 距离

如果把 n 个样品（X 中的 n 行）看成 m 维空间（m 个变量描述）中 n 个点，则两个样品间相似程度可用 m 维空间中两点的距离来度量。令 d_{ij} 表示样品 X_i 和 X_j 的距离。我们取明氏距离 $d_{ij}(q)=(\sum_{\alpha=1}^{m}|x_{i\alpha}-x_{j\alpha}|^q)^{1/q}$ 中的 q = 2 也就是用欧氏距离来定义样本间的距离[1]。

2.3 聚类方法

2.3.1 最短距离法

定义类 G_i 和 G_j 之间的距离为两类最近样品的距离 $D_{ij}=\min_{X_i\in G_i,X_j\in G_j} d_{ij}$. 设类 G_p 与 G_q 合并为一个新类记为 G_r，则任一类 G_k 与 G_r 的距离：

$$D_{ij}=\min\{\min_{X_i\in G_k,X_j\in G_p} d_{ij},\min_{X_i\in G_k,X_j\in G_q} d_{ij}\}=\min\{D_{kp},D_{kq}\}$$

2.3.2 最长距离法

定义类 G_i 和 G_j 之间的距离为两类最远样品的距离

$$D_{pq}=\max_{X_i\in G_p,X_j\in G_q} d_{ij}$$

最长距离法与最短距离法的并类步骤完全一样，将各样品先自成一类，然后将非对角线上最小元素对应的两类合并。设某一步将类 G_p 与 G_q 合并为 G_r，则任一类 G_k 与 G_r 的距离用最长距离公式为：

① 基金项目：华北电力大学博士学位教师科研基金资助项目（200612005）；河北省高等教育教学教改研究资助项目（2006 – 2009）

$$D_{ij}=\max\{\max_{X_i\in G_k,X_j\in G_p} d_{ij},\max_{X_i\in G_k,X_j\in G_q} d_{ij}\}=\max\{D_{kp},D_{kq}\}$$

再找非对角线最小元素的两类合并，直至所有的样品全归为一类为止。

2.3.3 快速聚类法

快速聚类法又称为动态聚类法，首先将样品根据实际问题的意义粗略地分类，然后再按照某种原则进行修正，直至分类比较合理为止。快速聚类的过程大致可由下图 1 所示【2】：

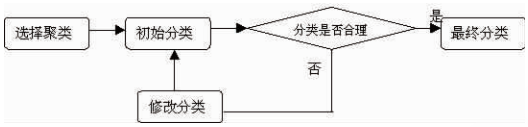

图 1　快速聚类法的过程

3　程序设计

3.1 编程语言的选择

选择 Visual J + +6.0 作为开发工具。

3.2 程序的界面设计

3.2.1 Applet 界面的设计

该界面用 Java Applet 编程实现，使用 this. setLayout(null)进行控件布局非常灵活，可以自由设定各个控件的位置，还可以通过 reshape 自由设定各个控件的大小。程序主界面如下图 2 所示：

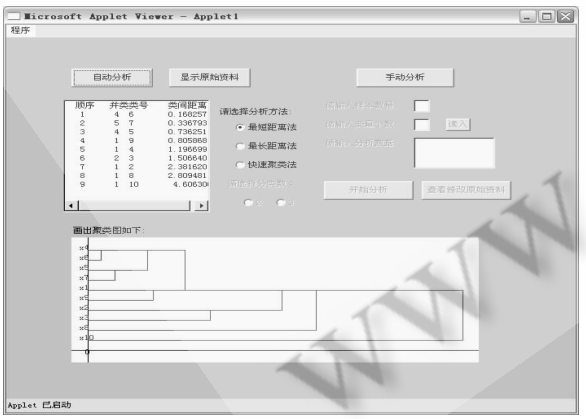

图 2　聚类分析的主界面

3.2.2 网页界面的设计

首先将 Applet 小程序嵌入到网页中，再在网页中写明实验题目、实验原理、方法说明、使用说明、结论及问题等。页面的设计用 Microsoft Office FrontPage 2003 作为开发环境[3]。

3.3 程序中使用的主要控件

Label(标签)；Button(按纽)；TextField(文本框)；TextArea(文本域)；CheckboxGroup(单选按纽组)，在此单选按纽组中定义了三个单选按纽 radio1、radio2、radio3。此外，在此程序中还定义了一块画布 mycanvas，两个弹出式窗口 MyFrame1、MyFrame2。并且画布和弹出式窗口都是通过新建自己的类来实现的[4]。

4　算法设计

4.1 对原始资料的标准化

设

$$X=\begin{pmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{pmatrix}$$

为原始资料矩阵，为消除数据量纲的影响，对原始数据进行标准化，即取

$$x'_{ij}=\frac{x_{ij}-\bar{x}_j}{s_j} \qquad i=1,2,\cdots,n;\ j=1,2,\cdots,m$$

其中，$\bar{x}_j=(\sum_{i=1}^{n}x_{ij})/n$，$s_j=[\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2/(n-1)]^{1/2}$，将标准化后的数据存入数组 x1[ ][ ]中。

4.2 最短(长)距离法聚类过程算法

利用欧氏距离求得距离矩阵 D，找出距离矩阵中非对角线最小元素 x_{pq}，比较距离矩阵中的第 p 行和第 q 列，将相对应的元素的较大值用较小的值代替，然后再删掉矩阵中的第 q 行和第 q 列，将矩阵的行的维数和列的维数分别减 1，继续重复执行，直至矩阵中只剩下一个元素为止，就完成了样本类的合并。此算法的流程图如下图 3 所示：

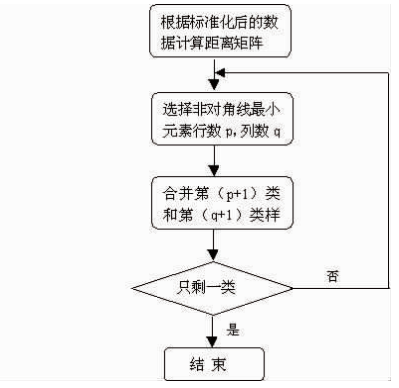

图 3　最短距离法流程图

4.3 快速聚类过程算法

找出距离矩阵 D 中非对角线最大元素，即可设定初始分类的聚点为 x_p 和 x_q，再通过比较各样本与 x_p 和 x_q 之间的距离来确定将样品放入哪一个类，在 n 个样品中，比较 d[i][p]、d[i][q]将样品 i 加入距离较小的类。

根据新分好的类中的样本计算新的聚点：

$$d1' = \frac{1}{count1}\sum_{x_j \in G_1} d[i][p] \text{、} \quad d2' = \frac{1}{count2}\sum_{x_j \in G_1} d[i][p]$$

其中 count1、count2 分别是类 1 和类 2 中已有的样品的个数，重复进行直到划分出一个较合理的类为止【2】。其流程图见图 4

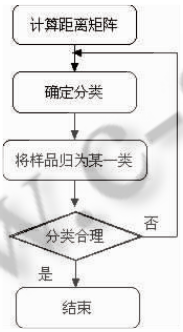

图 4　快速聚类法流程图

5　画图程序算法

5.1 最短、最长距离法的谱系图

在 MyCanvas 中定义数组 lei1[ ]、lei2[ ]通过在 Applet1 中用 MyCanvas. lei1[ ]、MyCanvas. lei2[ ]进行赋值来存放并类过程中每次所合并的样本的序号。用标志变量 flag1 = true 来确定选择的方法是最短距离法。

谱系图的画出是通过画直线来完成的，由于求出的类间距离是根据标准化后的矩阵求出的，所以距离比较小，为了画图的清晰性，将距离扩大 100 倍画出。

5.2 快速聚类法的谱系图

为了直观地显示快速聚类法的结果，首先画出两个方框来存放每一类中的样本。再通过在 Applet1 中赋好值的数组 fenlei1[ ]、fenlei2[ ]来将分好类的样本存放在相应的方框内。

6　程序的输入和输出

6.1 程序的输入

所设计的内容包括自动分析和手动分析两部分内容。在自动分析中，为方便用户学习，系统内已经输入好的一组数据进行分析，这组数据可以通过点击自动分析旁边的显示原始资料按纽显示出来。数据来源为世界各国森林资源的分布数据[1]。

在手动分析中，用户需要输入样本个数 n、变量数 m 及原始数据矩阵 x，系统将自动生成一个 n * m 的矩阵。数据输入格式为：数据 + 空格 + 数据 + 空格 + ……。最后按“读入”按纽就可将数据读入了。

6.2 程序的输出

程序的输出分为文本框输出和画图直观显示。文本框将输出合并的类号和类间距离等，为了更直观的让用户了解聚类分析的过程，还特意加入了谱系图显示。

为了方便用户的使用，进行了网页设计，包括一些必要的文字说明，如实验目的、实验原理、方法步骤、使用说明等，与 Applet 可视化程序一起构成了一个完整的聚类分析实验。

参考文献

1　夏绍玮，杨家本，杨振斌. 系统工程概论. 北京：清华大学出版社，1995. 88 – 97.
2　范金城，梅长林. 数据分析. 北京：高等教育出版社，2006. 176 – 197.
3　王凯，张家才. 网页制作技巧与实例. 北京：冶金工业出版社，1999.
4　杨昭，孙友编. Java 语言程序设计. 北京：中国水利水电出版社，2004.