JavaPCA:情报信息压缩抽取分析的一种网络共享软件

齐 艳 红

(中山大学生命科学学院　广州　510275)

摘　要　在情报信息分析工作中,经常会遇到大批量数据要处理。数据量大、信息不确定性高扰乱了信息可呈现的规律。从而需要对数据信息进行压缩,去除虚假信息,从压缩后的数据中抽提出所隐藏的规律。本文研制了用于情报信息压缩抽取分析的一种网络共享软件 JavaPCA,该软件由 6 个 Java 类和一个 HTML 文件组成,有图型用户界面,可在多种 Java 兼容网络浏览器如 Internet Explorer 和 Netscape 上运行。

关键词　情报信息　压缩和抽取　网络计算　共享软件

在情报信息分析工作中,经常会遇到一大批数据要处理和分析。这些数据要么是定量的,要么就是定性的。数据量大,信息也有不确定性,例如,所选取的某些用以表示信息的特征可能有重复或强相关,或次要特征多,扰乱了信息可呈现的规律。在这些情况下,就需要对数据信息进行压缩,去除虚假信息,从压缩后的数据中抽提出所隐藏的规律。迄今为止,已有多种可用于数据信息压缩抽取分析的方法,其中,PCA 是最为有效的算法之一。有关 PCA 的计算程序不少,但多数程序或缺乏友好的图型用户界面,或与特定平台有关,或不能进行网络分布式计算,或不易作为网络嵌入程序,网络信息资源共享的意义较小。信息资源网络化的一个要求是,所有资源尽可能地实现共享。为此,作者研制了用于情报信息压缩抽取分析的一种网络共享软件 JavaPCA。与作者等人研制的其它 Java 软件一样,JavaPCA 具有网络环境中运行、独立于平台、可随时更新等优点,值得进一步检验和应用。

1　算法概要

有多种算法可用于情报信息的压缩抽取分析,PCA 是其中的一类,后者又分若干种算法。本计算软件采用的 PCA 计算方法如下:

假设有 n 个样品,m 个特征,已知 m ×n 取值矩阵 (z_{ij})。构造规范化取值矩阵 (x_{ij}),即取 $x_{ij} = (z_{ij} - z_{bar\,i}) / s_i$,其中,

$$p_i = \sum_{j=1}^{n} z_{ij} / n$$

$$s_i = \left(\sum_{j=1}^{n} (z_{ij} - p_i)^2 / (n - 1) \right)^{1/2}$$

$$i = 1, 2, \cdots, m;\ j = 1, 2, \cdots, n$$

计算矩阵 $R = (r_{ij})$

$$r_{ij} = \sum_{k=1}^{n} (x_{ik} x_{jk}) / (n - 1) \qquad i, j = 1, 2, \cdots, m$$

根据方程 $|pE - R| = 0$,确定本征值 $p_1, p_2, \cdots, p_m$,及对应的本征向量 $u_1, u_2, \cdots, u_m$ 这里 E 为单位阵,且 $\sum_{k=1}^{m} p_k = m$。设 $p_1 > p_2 > \cdots > p_m$,令矩阵 $A = (a_{ij}) = (u_1 \sqrt{(p_1)}\ u_2 \sqrt{(p_2)} \cdots u_m \sqrt{(p_m)})$,则样品 j 的第 i 个成分为:

$$F_i = \sum_{k=1}^{m} (a_{ki} z_{kj}) \qquad i = 1, 2, \cdots, m;\ j = 1, 2, \cdots, n$$

从 p_1 到 p_m,累积达到一定比率,如 95 %时为止的成分为主成分。若主成分 F_1 和 F_2 所占比例较大,就可以用 F_1 和 F_2 分别为 x - 轴和 y - 轴,依各样品的毗邻性,划出若干不同的区,同区内的样品具有相似的特征。

该算法的功能:a. 将数据降维,多个特征被压缩为少数几个主成分,便于在低维下进行分析或做进一步计算;b. 根据每个主成分的特征组成,可分析特征间的相关性,同一主成分中起主要作用的那些特征之间有相似性,分属不同主成分的主要特征之间独立性较大;c. 特征载荷越大,则该特征愈能反映样品之间的差异,即由该特征可抽取的信息量愈大。因此,该算法可从混杂了虚假信息、重复特征或次要特征的数据中抽取出尽可能多的真实信息。

2　软件组成与结构

JavaPCA 用 Java 程序设计工具包 JDK 开发,由 6 个类和一个 HTML 文件组成(http://www.browso.com/sites/reference/mathsoft/pattern-recognition/,浏览器设置方法见 http://www.browso.com/sites/reference/mathsoft/pattern-recognition/demo.htm):

JavaPCA 类。PCA 计算在该类中进行。其 Java Applet 被载入浏览器后,显示输入窗口,内容为输入特征数、样品数、信息量抽取比率(某个大于 0 且小于 1 的值),打开原始数据文件(图 1)。

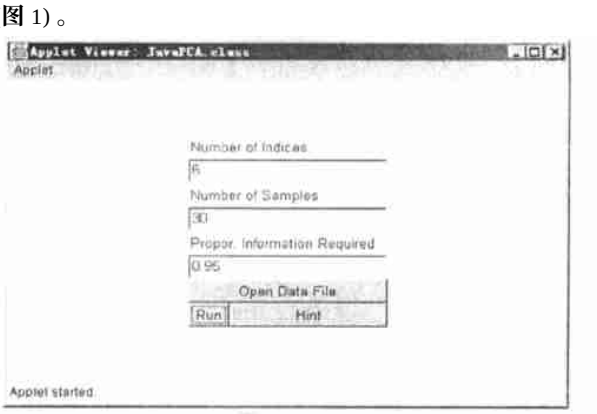

图 1　主窗口(JavaPCA 类)

ResultShow 类(图 2),GraphicsFrame 类,Hint 类,以及 WarningShow 类见文献所述。

PCAGraphics 类。该类用于输出样品在 F1 - F2 平面上的分布图(图 3)。

JavaPCA.html 文件。该文件将 JavaPCA 类载入 Web 浏

览器,并传入窗口大小及组件大小参数。

在原始数据文件中,第一行为各样品的编号,以后每行第一个值为特征编号,该行中其余值为该特征下各样品的取值。取值矩阵文件为普通的 MS - DOS 文本文件(.txt)。

软件运行后输出特征相关性矩阵,各主成分的本征值与本征向量,各主成分的变换方程,各样品的主成分 F_1 和 F_2 值,并输出 F_1 - F_2 平面上的样品分布图

JavaPCA 的特点包括:网络分布式计算,平台无关性,可即时更新,稳定性和安全性好等。

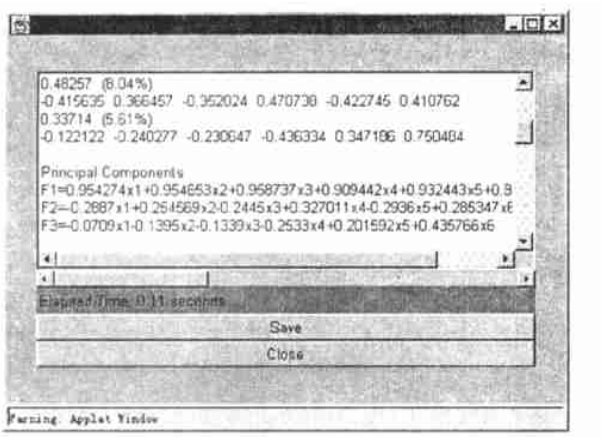

图 2　结果输出窗口(ResultShow 类)

3　计算及分析

取《人大复印报刊资料 G9 分册》收录论文的地区发文统计数据(夏旭,2000;表 5),结果见图 3 所示。其中,样品 1～30 分别为北京,湖北,江苏,上海,广东,天津,四川,河北,湖南,山东,河南,黑龙江,吉林,安徽,辽宁,陕西,浙江,江西,甘肃,广西,福建,内蒙古,山西,宁夏,云南,新疆,西藏,青海,贵州,海南;特征 1～6 分别是第 1 作者数,高校作者数,论文数量,高校论文数量,机构数,高校数。

取定需抽取信息的比率为 95 % (图 1),则软件运行后得到如图 2 和图 4 所示的结果。显然,原来的 6 个特征被压缩为 3 个主成分,即 3 个综合特征。在前 2 个主成分中,各特征的作用大致相同,但第 2 个主成分中高校论文数有较大作用;第 3 个主成分中,第 6 个特征,即高校数起主要作用,高校论文数也有较大作用。因此,高校数是个较为独立的特征,从以下的相关阵也可看出该特征较大的独立性。

1.0	0.8481	0.9952	0.7903	0.959	0.7011
0.8481	0.9999	0.8685	0.9845	0.7889	0.8245
0.9952	0.8685	1.0	0.8268	0.936	0.6914
0.7903	0.9845	0.8268	0.9999	0.7013	0.7588
0.959	0.7889	0.936	0.7013	0.9999	0.7974
0.7011	0.8245	0.6914	0.7588	0.7974	0.9999

由于前 2 个主成分已占到总信息量的 94.12 %,因此,图 4 的样品分布有较高的可信度。根据图 4,北京的收录论文情况与其它省市有质的不同,表现在第一主成分,也表现在第二主成分上;其它地区在第二主成分上无显著差异。湖北也有特殊性,第一主成分值较大;江苏,上海,广东,天津的收录论文情况较为接近;其它地区依样品序号有渐变的特点。由于第一主成分信息量很大(86.08 %),与各特征均为正相关关系,各地区沿 F_1 轴的分布也就反映了收录论文总数及其相关的情况。

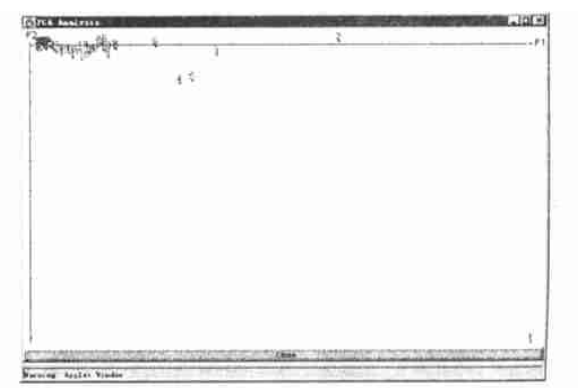

图 4　F_1 - F_7 平面上的样品分布图

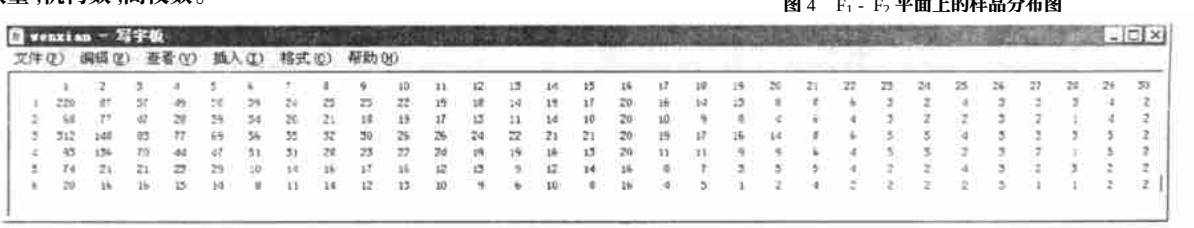

图 3　原始数据及其存放形式

参 考 文 献

1　齐艳红,王德轩,张文军. 1990. 应用计算机分析小麦品种抗赤霉病指标及批量鉴定的初步研究. 植物病理学报,2000;(2)

2　程焕文. 2001. 趋势种种 - 图书馆数字化网络化研究札记. 图书馆论坛,2001;(2)

3　朱淑华. 2001. 论网络环境下图书馆的信息服务工作. 图书馆论坛,2001;(1)

4　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,2002

5　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,2002

6　齐艳红. 图书期刊评定分析的一种网络分布式计算软件. 情报杂志,2003;(1)

7　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,2003

8　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,2003

9　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,2002

10　张文军,齐艳红. Schoenly KG. 泛函连接网络计算软件及其在生物多样性研究中的应用. 生物多样性,2002;(3)

11　沈向若. 网络信息安全隐患及防范对策. 图书馆论坛,2001;(1)

12　袁嘉秀. 普及型用户界面的理想模式. 图书馆论坛,2001;(5)

13　夏　旭.《人大复印报刊资料 G9 分册》收录论文的计量分析. 高校文献信息研究,2000;(1)

(责编:勃王京)