一种基于 PCA 和系统成团法的聚类软件设计*)

姜 斌[1] 衣振萍[1] 马绍汉[2]

(山东大学威海分校信息工程学院 威海 264209)[1] (山东大学计算机科学与技术学院 济南 250100)[2]

摘 要 提出一种基于主分量分析和系统成团法的快速聚类方法。通过构造主分量空间将分散在一组变量上的高维天体光谱投影到两个主分量上,每一个主分量都是原始变量的线性组合,主分量之间互为正交关系,在剔除冗余信息的同时,得到二维坐标;以此为输入,使用系统成团法进行聚类分析研究,实现高维天体光谱的快速自动分类处理。以上述方法为基础设计天体光谱自动分类软件,实现海量光谱的快速、准确分类。

关键词 高维数据,降维,主分量分析,系统成团法

Design of a PCA and Hierarchical Method Classification Software

JIANG Bin[1] YI Zhen-Ping[1] MA Shao-Han[2]

(School of Information Engineering, Shandong University at Weihai, Weihai Shandong 264209)[1]

(Department of Computer Science, Shandong University, Ji'nan 250100)[2]

Abstract An efficient and quick method based on 2-D pca and hierarchical clustering method is proposed. The coordinates are achieved by projecting the high dimensional celestial objects spectra data to the 2-D space after the principle component space is built, every principle component is the linear combination of the original variables and is irrelevant to each other. The coordinates can be clustered by hierarchical clustering method. An automatic celestial objects spectra classification software is designed based on the method to realize the quick and accurate classification of large numbers of spectra.

Keywords High-dimensional data, Dimensionality reduction, PCA, Hierarchical clustering method

1 引言

高维数据在使用计算机技术实现自动分类处理时,首先要进行降维处理,得到低维数据,再以此为基础进行聚类分析。

以高维的光谱数据为例,国际标准的光谱数据存储格式采用 fits(flexible image transport system)文件[1]形式,其维数可达到 3000 维以上。如果将每一维都做处理,大量光谱数据处理的运算量很大,因此有必要对高维数据在不丢失重要信息的前提下进行降维处理。

降维的原理是通过特征提取和选择,在所有特征中求出最重要的特征,放弃一些次要特征,从而实现对特征空间维数的压缩,达到降维目的。高维数据降维后,就可以进行聚类研究,将有相同特征的对象归为一类。

本文在进行光谱自动识别的研究过程中,研究了恒星等光谱的规律。通过用 PCA 方法构造光谱的主分量,把光谱中的主要特征提取出来,采用主分量为轴,直接把样本点在主分量坐标轴上进行投影,可以得到二维平面上的样本特征点,大大降低了光谱数据的维数。再以此为基础,通过系统成团方法对投影点进行聚类分析,从而实现了对海量光谱数据的快速自动分类。

该方法也可用于统计学等需要进行高维数据处理的领域。

2 主分量分析(PCA)方法

主分量分析法是统计学中分析数据的一种有效方法。其目的是在数据空间中找出一组向量来尽可能地解释数据的方差,用较少数量的特征对样本进行描述来降低特征空间维数,将数据从原来 n 维降低到 m 维($m \ll n$),在降维后保存了数据中的主要信息的同时获得原模式空间的一个最优低维逼近,从而使数据更易于处理。

2.1 PCA 方法的步骤

以恒星光谱分析为例,用主分量分析法进行降维的过程如下[3]:

(1)选取 M 条恒星光谱,记为 p_i($i = 1,\cdots, M$),其中 M 是光谱样本数,构成[$M \times N$]的矩阵, N 是光谱的维数。

(2)对这些光谱数据进行波长统一的预处理,方法采用线性插值。

(3)对每条光谱数据(流量)进行归一化,归一化方程为

$$P_{i,j} = \frac{P_{i,j}}{\sqrt{\sum_{j=1}^{N} P_{ij}^2}}$$

其中 $i \in [1,M]$, $j \in [1,N]$, N 为每条光谱的点数。

(4)构造恒星光谱矩阵 p_{M*N}。该矩阵的每一行代表经过归一化的恒星光谱,共 M 行,每行有 N 个分量,每个分量代表某个波长下的谱线流量。

(5)构造恒星光谱矩阵 P 的相关矩阵 $C_{i*j} = P \times p^T$, p^T

*)本项目受国家重大工程项目 LAMOST 资助。姜 斌 助教,硕士,主要研究方向:模式识别;马绍汉 教授,博士生导师,主要研究方向:算法及其复杂性理论。

为 P 的转置，其中 $i \in [1, M]$，$j \in [1, N]$，C_{i*j} 为 $[M \times M]$ 的方阵。

(6)求相关矩阵 C_{i*j} 的特征值和特征向量，然后将 C_{i*j} 对角化。对角化方程如下：

$$C = R\Lambda R^T \qquad \Lambda = \begin{pmatrix} \lambda_1 & 0 & 0 & L \\ 0 & \lambda_2 & 0 & L \\ 0 & 0 & \lambda_3 & L \\ L & L & L & O \end{pmatrix}$$

其中 R 的每一个列向量 R_i 都是 C 的特征向量。而 Λ 矩阵是一个对角矩阵，对角线上的元素 λ_i ($i \in [1, M]$) 是 C 的特征值，并且按从大到小的顺序排列。

(7)在(6)的基础上构造空间变换矩阵 H，方法如下：选取方差贡献率 μ 大于 95%的对应 C 的特征值的特征矢量，构成特征矩阵 E。方差贡献率 μ 的定义如下：

$$\mu = \frac{\sum_{i=1}^{L} \lambda_i}{\sum_{i=1}^{M} \lambda_i} \qquad L < M$$

在具体实验中仅选取前两个特征值。因为根据经验，方差贡献率大于 95%的特征值总为前两个，而且前两个特征值远大于其余的特征值。这样，特征矩阵 E 是 $[M \times 2]$ 的矩阵。恒星的主分量空间变换矩阵 H 即为特征矩阵 E 的转置和标准化后的 P 的乘积：

$$H = E^T \times P$$

H 为 $[2 * N]$ 矩阵。

(8)在完成上述步骤后，利用空间变换矩阵 H，就可以构造出恒星的主分量空间。矩阵 H 的每一个行向量就是恒星的主分量，而且这些主分量之间相互正交。

通过变换矩阵 H，把每个标准化后的样本投影到二维的主分量空间中，公式为 $P_i \times H^T$，因为 P_i 是 $[1 * N]$，H^T 是 $[N * 2]$，所以得到的是高维光谱数据的二维坐标，投影到二维平面上如图 1。

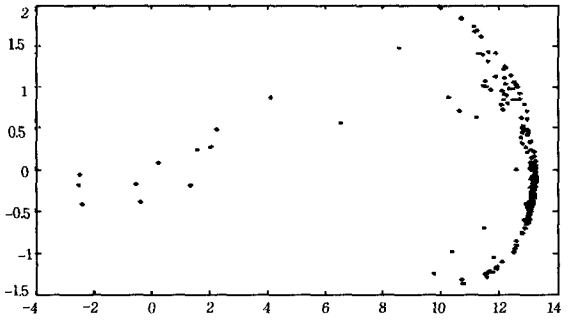

图 1　恒星光谱二维主分量空间投影图

3　系统成团法

系统成团法(Hierarchical Clustering Method)也称系统树法，基本思想是：先将 n 个样品各自看成一类，规定样本之间的距离和类与类之间的距离，选择距离最小的一对并成一个新类，计算新类和其它类的距离，再将距离最近的类合并成一个新类，如此下去，一直到所有样品都归为一类为止。

因样品之间和类与类之间距离有多种定义法。例如，可以定义类与类之间的距离为属于两类的样本间的最近距离，也可以定义为两类的样本之间的最长距离，或者定义为两类重心之间的距离等等，而不同的类间距离就产生了不同的系统成团法。在实验中，采用的是最短距离法(Single linkage 或 Nearest neighbor 法)。现用 d_{ij} 表示样品 (i) 和 (j) 之间的距离，D_{pq} 表示类 G_p 和 G_q 之间的距离，类 G_r 表示类 G_p 和 G_q 合并成的新类，表示为 $G_r = \{G_p, G_q\}$。

类间距离为两类中最近样品之间的距离：

$D_{pq} = \min\{d_{ij}\}$ $i \in G_p, j \in G_q$

计算新类 $G_r = \{G_p, G_q\}$ 与其它类距离的递推公式为

$D_{rk} = \min\{D_{pk}, D_{qk}\}$

以二维坐标作为数据源进行系统树聚类，步骤如下：

(1)计算距离，采用欧几里德距离

$$d_{ij} = \sqrt{\sum_{k=1}^{2} (x_{ik} - x_{jk})^2}$$

算得距离矩阵 $D(0)$——样本间距离的对称 $D(0)$，因为开始时每个样本自成一类，这时 $D_{pq} = d_{pq}$，

D(0)

	G1	G2	G3	G4	G5	G6	G7
G2	2.0						
G3	2.2	2.2					
G4	2.8	2.0	1.0				
G5	6.3	6.0	8.1	8.0			
G6	5.0	4.1	6.3	6.1	2.2		
G7	5.8	5.1	7.3	7.1	1.4	1.0	
G8	8.0	6.7	8.6	7.8	6.7	5.1	5.4

(2)选择 $D(0)$ 中的最小元素 $D_{pq} = 1.0$，它们是 $D_{3,4}$ 和 $D_{6,7}$，故先将 G_3 和 G_4 合并成新类 $G_9 = \{G_3, G_4\}$，再将 G_6 和 G_7 合并，即 $G_{10} = \{G_6, G_7\}$。

(3)计算新类和其它各类的距离，得到

$D_{1,9} = \min\{D_{1,3}, D_{1,4}\} = D_{1,3} = 2.2$

$D_{1,10} = \min\{D_{1,6}, D_{1,7}\} = D_{1,6} = 5.0$

$D_{9,10} = \min\{D_{3,6}, D_{3,7}, D_{4,6}, D_{4,7}\} = D_{4,6} = 6.1$

将 $D(0)$ 中的 p, q 或 $p1, q1, p2, q2$ 行及相应的列(第 3, 4, 6, 7 行和列)删去，加上第 r 行(第 9, 10 行)和第 r 列得到距离矩阵 $D(1)$：

	G1	G2	G5	G8	G9
G2	2.0				
G3	6.3	6.0			
G8	8.5	6.7	6.7		
G9	2.2	2.0	8.0	7.8	
G10	5.0	4.1	1.4	5.1	6.1

(4)重复 2～3 步。在 $D(1)$ 中，$D_{5,10} = 1.4$ 是最小元素，将 G_5 和 G_{10} 并成一类 G_{11}，计算 $D(2)$：

	G1	G2	G8	G9
G2	2.0			
G8	8.5	6.7		
G9	2.2	2.0	7.8	
G11	5.0	4.1	5.1	6.1

在 $D(2)$ 中，$D_{1,2}=D_{2,9}=2.0$ 是最小元素，将 G_1，G_2，G_9 合并成新类 G_{12}，计算 $D(3)$，这时新类与各类的距离

$D_{10,8}=\min\{D_{1,8}, D_{2,8}, D_{9,8}\}=6.7$

$D_{12,11}=\min\{D_{1,11}, D_{2,11}, D_{9,11}\}=4.1$

在 $D(3)$ 中，$D_{12,11}$ 最小，将 G_{11} 和 G_{12} 并成一类 G_{13}，则有

D(3)

	G8	G11
G11	5.1	
G12	6.7	4.1

最后将 G_{13} 和 G_{14} 并成一类 G_{14}。这样，所有的样品到此已并成一类，聚类过程结束。

上述聚类过程可以用聚类图表示，如图 2。图中横坐标表示距离。由图 2 可见 G_2，G_6、G_7 合并成一类，G_1、G_4、G_5 合并成一类，G_3、G_9 合并成一类，G_8 自成一类，全部样本分为四类为宜。

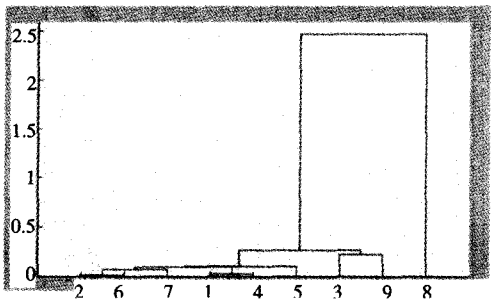

图 2 聚类结果图

在实际分类时，为了达到客观分类的目的，并不需要将聚类过程进行到全部样本并成一类为止，而是给定一个临界值 T，当所有的 $D_{pq}>T$ 时，即认为类与类之间不能再合并了。

4 实验及结果分析

4.1 实验结果

将几类已知分类结果的光谱混在一起使用上述方法进行聚类分析，在相应的阈值下，用已知结果检验分类的正确性。实验数据来自美国的 Sloan Digital Sky Survey (SDSS) 发布的观测数据(DR2)。在其发布的 fits 文件中，用 SPEC_CLN 字段标识其正确分类。为了保证实验有统计意义，每个实验都重复 10 次，并与逐步判别分析方法比较。实验结果如表 1。

表 1 实验结果

NO	光谱数量		正确率	
	Nebula	Quasar	PCA& 系统成团	逐步判别分析
1	500	500	81.70%	64.30%
2	1000	1000	85.70%	73.50%
3	3000	3000	88.54%	77.40%
4	5000	5000	90.60%	82.10%

根据实验结果可以看出，当样本量很大时，用该方法进行聚类的结果正确率较高。

4.2 拟合程度

用相关系数来衡量数据与聚类结构的拟合程度。

相关系数定义为

$$\frac{\sum_{i<j}(Y_{ij}-y)(Z_{ij}-Z)}{\sqrt{\sum_{i<j}(Y_{ij}-y)^2\sum_{i<j}(Z_{ij}-Z)^2}}$$

在上述实验中，相关系数达到 99.667%以上，说明拟合的结果较好。

4.3 时间复杂度分析

算法包括两部分：PCA 降维和系统成团聚类。前者的计算量主要体现在对光谱矩阵等矩阵的计算上，时间复杂度 $O(M*N)$，其中 M 为光谱数量，N 为光谱维数；后者的计算量主要体现在距离矩阵的计算上，时间复杂度为 $O(M^2)$，($M\ll N$)，M 为二维空间投影点，也就是光谱数量。所以算法总的时间复杂度为 $O(M*N+M^2)$。

结束语 本文使用主分量分析法对高维光谱数据进行降维。主分量分析法的一个基本假定是每个方案对应于各个准则的取值服从正态分布。当样本数目较少，或取值的离散化程度较高时，就不能假定准则的取值还服从正态分布。因此只有在大样本的情况下，采用主分量分析法进行降维才有意义。虽然大样本的高维光谱数据使用 PCA 变换有较大的运算量，但由于此方法只需要做一次 PCA 变换，构造出主分量空间后，只需要把待降维的光谱数据投影到此二维空间中即可。实验结果表明，该方法在实际运算时程序本身计算量很小。低维数据为进一步处理数据奠定了基础。使用系统树方法可以迅速发现离群点，同时可以对大样本进行聚类。如何由聚类图来确定合适的分类数和分类结果，目前还没有明确的标准。Bemirmen 提出了以下几点根据聚类图和实际问题的意义来确定适当的分类结果的准则，可供参考：

(1)分类数与实际问题的意义一致。

(2)各类重心之间的距离应很大。

(3)若采用不同的聚类法(如最短距离、最长距离、重心距离)，则在各自的聚类图中应有大致相同的分类结果。

参 考 文 献

1 赵永恒. LAMOST 项目计划书. 北京：国家天文台，2005

2 赵永恒. fits 文件解析. http://www.lamost.org/xoops/modules/intro.html 2006-03-16

3 覃冬梅. 一种基于主分量分析的恒星光谱快速分类法. 光谱学与光谱分析，2003，23(1)：182～186

4 Kurtz M J. Progress in automation techniques for mk classfication. Astrophys，2004(4)：111～117

5 Shai Rong，Salamanca A A. Principal component analysis of synthetic galaxy spectra. Astrophys，2006(24)：41～50

6 Chen OT-C. Motion estimation using a one-dimentional gradient descent search. IEEE Transactions Circuits and System for Video-Technology，2000，10(4)：608～616

7 王文胜. 图像特征抽取的奇异值分解方法. 计算机工程，2006，32(8)：32～36

8 Bailer-Jonea C. Techniques for MK Classfication. Astrophysics and Space Science，2002(24)：21～30

9 董长虹. Matlab 小波分析工具箱原理与应用. 北京：国防工业出版社，2004

10 薛建桥. 神经网络技术与光谱自动分类：[学位论文]. 中国科学院北京天文台，1999