PCA和相融性度量在聚类算法中的应用

姜 斌[1]，潘景昌[1]，郭 强[2]，衣振萍[1]

(1. 山东大学威海分校信息工程学院 山东 威海 264209； 2. 上海大学计算机工程与科学学院 上海 闸北区 200436)

【摘要】提出一种基于主分量分析和相融性度量的快速聚类方法。通过构造主分量空间将高维数据投影到两个主成分上进行特征提取，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，实现高维数据降维，得到二维坐标，以此作为聚类分析的输入；提出相融性度量的定义，用相融性度量描述一个样本与训练集相融合的程度，设计一种基于相融性度量的分类器。以该方法为基础设计的光谱自动分类系统可实现快速、准确地分类。

关 键 词 相融性度量； 降维； 高维数据； 主分量分析
中图分类号 TP391 文献标识码 A

Application of PCA and Coherence Measure in Clustering Algorithm

JIANG Bin[1]，PAN Jing-chang[1]，GUO Qiang[2]，YI Zhen-ping[1]

(1.School of Information Engineering, Shandong University at Weihai Weihai Shandong 264209;
2. School of Computer Science, Shanghai University Zhabei Shanghai 200436)

Abstract An efficient and quick method based on 2-D Principal Component Analysis (PCA) and coherence measure is introduced. The coordinates are achieved by projecting the high dimensional data to the 2-D space after the principle component space is built and feature extraction is finished at one time. Every principle component is the linear combination of the original variables and is irrelevant to each other. A novel coherence measure is introduced and designed for effectively measuring the coherence of a new specimen of unknown type with the training samples. The spectrum can be classified quickly and exactly by the classifier.

Key words coherence measure; dimensionality reduction; high-dimensional data; principal component analysis

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[1-2]形式，其维数可达到3 000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，提出了一种基于主分量分析(Principal Component Analysis，PCA)和相融性度量的快速聚类算法。

1 主分量分析方法

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，将数据从原来的n维降低到m维($m<<n$)。降维后保存数据中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。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=1,2,$\cdots$,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) 对每条光谱数据(流量)进行归一化，归一化方程为：$P_{ij}=P_{ij}\Big/\sqrt{\sum_{j=1}^{N}P_{ij}^2}$。其中$i\in[1, M]$，$j\in[1, N]$；

收稿日期：2007 – 09 – 14
基金项目：国家重大工程LAMOST项目
作者简介：姜斌(1977 –)，男，硕士，主要从事模式识别及其应用方面的研究.

N为每条光谱的点数。

(3) 构造恒星光谱矩阵$\boldsymbol{P}_{M\times N}$。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强度。

(4) 构造恒星光谱矩阵$\boldsymbol{P}$的相关矩阵$\boldsymbol{C}_{i\times j}=\boldsymbol{P}\times\boldsymbol{P}^{\mathrm{T}}$，$\boldsymbol{P}^{\mathrm{T}}$为$\boldsymbol{P}$的转置。其中，$i\in[1, M]$，$j\in[1, N]$，$\boldsymbol{C}_{i\times j}$为$[M\times M]$的方阵。

(5) 求相关矩阵$\boldsymbol{C}_{i\times j}$的特征值和特征向量，然后将$\boldsymbol{C}_{i\times j}$对角化，对角化方程如下：

$$\boldsymbol{C}=\boldsymbol{R}\boldsymbol{\Lambda}\boldsymbol{R}^{\mathrm{T}},\qquad \boldsymbol{\Lambda}=\begin{pmatrix}\lambda_1 & 0 & 0 & \cdots\\ 0 & \lambda_2 & 0 & \cdots\\ 0 & 0 & \lambda_3 & \cdots\\ \cdots & \cdots & \cdots & \ddots\end{pmatrix}$$

式中　$\boldsymbol{R}$的每一个列向量$\boldsymbol{R}_i$都是$\boldsymbol{C}$的特征向量；$\boldsymbol{\Lambda}$矩阵是一个对角矩阵，对角线上的元素$\lambda_i(i\in[1, M])$是$\boldsymbol{C}$的特征值，并且按从大到小的顺序排列。

(6) 在过程(5)的基础上构造空间变换矩阵$\boldsymbol{H}$，方法如下：选取方差贡献率μ大于98%的对应$\boldsymbol{C}$的特征值的特征矢量，构成特征矩阵$\boldsymbol{E}$。方差贡献率为：

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

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98%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。因此特征矩阵$\boldsymbol{E}$是$[M\times 2]$的矩阵。恒星的主分量空间变换矩阵$\boldsymbol{H}$即为特征矩阵$\boldsymbol{E}$的转置与标准化后的$\boldsymbol{P}$的乘积，$\boldsymbol{H}=\boldsymbol{E}^{\mathrm{T}}\times\boldsymbol{P}$，$\boldsymbol{H}$为$[2\times N]$矩阵。

(7) 在完成上述步骤后，利用空间变换矩阵$\boldsymbol{H}$，就可以构造出恒星的主分量空间。矩阵$\boldsymbol{H}$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。

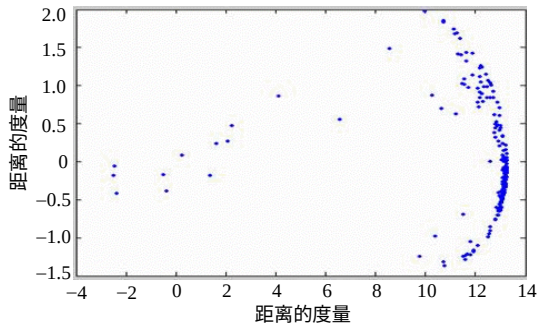

图1　恒星光谱二维主分量空间投影图

通过变换矩阵$\boldsymbol{H}$，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$\boldsymbol{P}_i\times\boldsymbol{H}^{\mathrm{T}}$。因为$\boldsymbol{P}_i$是$[1\times N]$矩阵，$\boldsymbol{H}^{\mathrm{T}}$是$[N\times 2]$矩阵，所以得到的是高维光谱数据的二维坐标，用Matlab投影到二维平面上，如图1所示。

2　相融性度量

2.1　基本思想

根据样本与训练样本集的融合程度而不是距离进行聚类，如果一个样本的局部离散度与附近训练样本的局部离散程度比较接近，就将它与训练样本集聚为一类。

2.2　相融性度量的定义

2.2.1　样本的局部离散度$V_L(X)$

假设K是一个大于1的整数，样本X的K个近邻记为$X_1^*,X_2^*,\cdots,X_k^*$，属于第l类的有K_l个样本，记为$X_{l,1}^*,X_{l,2}^*,\cdots,X_{l,k}^*(l=1,2,\cdots,c)$，样本$X$关于第$l$类的局部离散度$V_L(X)$[4-5]定义为：

$$v_l(x)=\frac{1}{k_l}\sum_{i=1}^{k_l}(x_i^l-x)^{\mathrm{T}}(x_i^l-x)\tag{1}$$

或

$$v_l(x)=\frac{1}{k_l+1}[(x-m_l(x))^{\mathrm{T}}(x-m_l(x))+\sum_{i=1}^{k_l}(x_{l,i}^*-m_l(x))^{\mathrm{T}}(x_{l,i}^*-m_l(x))]\tag{2}$$

式中　$m_l(x)=\frac{1}{k_l+1}\left(x+\sum_{i=1}^{k_l}x_{l,i}^*\right)$。$V_L(X)$称为样本$X$的第$L$类近邻离散度。

2.2.2　样本的相融性$\mathrm{co}_l(x)$

假定有关于c个类别的分类问题，第i类有N_i个训练样本$\{x_j^l, j\le N_l, l\le c\}$，训练样本总数$N=\sum_{l=1}^{c}N_l$。样本$X$的第$l$类相融性度量定义为[6]：

$$\mathrm{co}_l(x)=\frac{\left(\sum_{i=1}^{k_l}v_l(x_{l,i}^*)\Big/k_l\right)}{v_l(x)}\tag{3}$$

或

$$\mathrm{co}_l(x)=\frac{\max\limits_{1\le i\le k_l}\{v_l(x_{l,i}^*)\}}{v_l(x)}\tag{4}$$

$$\mathrm{co}_l(x)=\frac{\min\limits_{1\le i\le k_l}\{v_l(x_{l,i}^*)\}}{v_l(x)}\tag{5}$$

2.3　分类规则

t是相融性度量参数的阈值，用来衡量待分类样本与训练样本集是否有效地融为一体。样本X的K个近邻记为$X_1^*,X_2^*,\cdots,X_k^*$，其中，属于第t类的K_l个样本记为$X_{1l},X_{2l},\cdots,X_{kl}l$。而且假定近邻个数满足$K_l\ge K_0$

的类为$l_1,l_2,\cdots,l_r$，令：

$$\mathrm{co}_{l_s}(x)=\max_{i=1,2,\cdots,r}\mathrm{co}_{l_i}(x)$$

分类规则如下[7-8]：

(1) 如果$\mathrm{co}_{l_s}(x)\geq t$，则决策样本$X$来自第$l_s$类。

(2) 否则，决策样本X来自未知类。

从上面的分类方法以及对$\mathrm{co}_l(X)$的定义知道，每次分类时都需要计算训练样本X关于第$l(x)$类的近邻离散度$V_{l(x)}^{(x)}$。实际项目中，分类器训练时可以提前把它们都计算出来并保存，每次分类时只要直接使用即可，其中$l(x)$表示训练样本X所属的类别。

2.4 分类算法

算法训练分类器的设计如下：

(1) 选定参数K和K_0，它们是满足下列条件的整数：$K\geq 3$，$K_0\geq 2$，$K\geq K_0$。

(2) 对每个训练样本X，计算它关于第$l(X)$类的近邻离散度$V_{l(x)}(x)$。

(3) 确定相融性阈值参数t，它满足$t\geq 0$，而且一般取$t<1$。

(4) 根据式(3)计算$\mathrm{co}_l(x)$，并按照上面的分类规则进行分类决策。

从分类规则可以看出，相融性度量$\mathrm{co}_l(x)$的三种定义方式反映了不同的倾向性：式(4)更倾向于将未知类别样本决策为来自已知类别，式(5)相反。式(3)稳定性好，式(4)、(5)容易受个别离群训练样本影响。

3 实验结果及分析

将几类已知分类结果的光谱作为实验数据使用上述方法进行聚类分析，PCA使用的阈值是能使样本分为N类的阈值，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(DR5)，在其发布的fits文件中，用SPEC_CLN字段标识其正确分类，由此可以决定N值。实验中，分类器的训练过程包括以下步骤：

(1) 确定分类器参数K、K_0，实验中K=7，K_0=3。

(2) 对每个训练样本X，计算第$L(X)$类近邻离散度$V_{L(X)}(X)$，并存储下来，以备搜索K个近邻时调用。其中$l(x)$定义如下：

$$l(x)=\begin{cases}1 & \text{如果}x\text{来自BL类}\\ 2 & \text{如果}x\text{来自NL类}\\ 3 & \text{如果}x\text{来自LINER类}\end{cases}$$

实验中，用式(3)计算第$l(x)$类近邻离散度。为了保证实验有统计意义，每个实验都重复10次，并与K近邻分类方法比较(K=N)，实验结果如表1、2所示。

表**1** 基于相融性度量方法的实验结果

t	识别率/(%)	误识率/(%)	拒绝率/(%)
0.7	91.22	3.23	5.55
0.6	92.17	2.56	5.27
0.5	93.01	2.34	4.65
0.4	93.75	2.07	4.18
0.3	95.52	1.79	2.69
0.2	96.40	1.01	2.59
0.1	97.20	0.97	1.83

表**2** 基于***K***近邻分类器的实验结果

识别率/(%)	误识率/(%)	拒绝率/(%)
93.50	6.50	0

从实验结果可知，K近邻分类器的误识率是6.50%；而基于相融性度量方法的分类器，当相融性阈值参数t=0.1时识别率可以达到0.97以上。所以，该方法有较好的识别率。

4 结 论

本文首先使用PCA方法对高维光谱数据进行降维。PCA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[9]。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于正态分布。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，大量的计算工作可以交给MATLAB处理[10-11]，为进一步处理数据奠定了基础。相融性度量根据样本与训练样本集的融合程度进行决策，判断它来自已知类别还是未知类别。对在K近邻法中不论样本距离训练样本集多远，最终都决策为某一具体的已知类型做了改进。对于不同类别交叠区域中的样本，相融性度量更好地描述了它属于不同类的可能性大小，比K近邻分类器单纯依靠近邻中包含每类样本的数目进行分类更为合理。以此方法设计的恒星光谱自动分类软件具有较好的效率和准确率。对方法中如何确定相融性阈值参数t是解决问题的关键，也是今后需要进一步研究的问题。

参 考 文 献

[1] 赵永恒. LAMOST项目计划书[R]. 北京: 国家天文台, 2005.

[2] 赵永恒. fits文件解析[EB/OL]. www.lamost.org, 2007-02-19.
[3] 覃冬梅. 一种基于主分量分析的恒星光谱快速分类法[J]. 光谱学与光谱分析, 2003, 23(1): 182-186.
[4] 李乡儒. 几个学习算法及其在星系光谱分类中的应用[D]. 北京: 中国科学院北京天文台，2007.
[5] 边肇祺, 张学工. 模式识别[M]. 北京: 清华大学出版社, 2000.
[6] 苏金明. Matlab工具箱应用[M]. 北京: 电子工业出版社, 2004.
[7] 薛建桥. 神经网络技术与光谱自动分类[D]. 北京: 中国科学院北京天文台, 1999.
[8] KURTZ M J. Progress in automation techniques for mk classfication[J]. Astrophys, 2004, (4): 111-117.
[9] CHEN OT-C. Motion estimation using a one-dimentional gradient descent search[J]. IEEE Transactions Circuits and System for VideoTechnology, 2000, 10(4): 608-616.
[10] BAILER-JONEA CAL. Techniques for mk classfication[J]. Astrophysics and Space Science, 2002, (24): 21-30.
[11] 王文胜. 图像特征抽取的奇异值分解方法[J]. 计算机工程, 2006, 32(8): 32-36.

编　辑　漆　蓉

(上接第1288页)

4　结　论

本文介绍了孤立点检测的传统算法，并在此基础上，提出了平均密度的概念，平均密度接近物理学上的关于密度的定义，使人们对孤立点的认识更自然；在平均密度概念的基础上，给出了基于平均密度的孤立点检测方法，该方法对孤立点的检测更加自动化，通常情况下，它不依赖于用户输入参数。

和基于密度的或基于距离的大多数孤立点检测算法一样，该方法的时间复杂度是$O(n^2)$，在数据规模较大时，需考虑抽样来确定平均密度S_i和平均距离D，再对各数据对象进行孤立点检测。本文在传统孤立点定义的基础上，拓展了新的视点，在算法自动化上作了一定的探索。

参　考　文　献

[1] HAN J, KAMBER M. Data mining: concepts and techniques[M]. [S.l.]: Morgan Kaufmann Publishers, Inc. 2001.
[2] HAWKINS D M. Identification of outliers[M]. London: Chapman and Hall, 1980.
[3] BARNETT V, LEWIS T. Outliers in statistical data[M]. New York: John Wiley & Sons, 1994.
[4] KNORR E M, NG RT. Algorithms for mining distance-based outliers in large datasets[C]//In: Proceedings of the 24th VLDB Conference. New York: [s.n.], 1998: 392-403.
[5] RAMASWAMY S, RASTOGI R, SHIM K. Efficient algorithms for mining outliers from large data sets[C]//In: Proceedings of the ACM SIGMOD Conference. [S.l.]: [s.n.], 2000: 473-438.
[6] 孙焕良, 鲍玉斌, 于　戈, 等. 一种基于划分的孤立点检测算法[J]. 软件学报, 2006, 17(5): 1009-1015.
[7] BREUNIG M M, KRIEGEL H, NG R T, et al. LOF: Identifying density-based local outliers[C]//In: Proc of the 2000 ACM SIGMOD Int'1 Conf on Management of Data. Dallas: ACM Press, 2000: 93-104.
[8] PAPADIMITIROU S, KITAGAWA H, GIBBONS P B, et al. LOCI: Fast outlier detection using the local correlation integral[C]//In: Proc of the 19th Int'1 Conf on Data Engineering. [S. l.]: IEEE Computer Society Press, 2003.
[9] 蒋盛益, 李庆华, 王　卉, 等. 一种增强的局部异常挖掘方法[J]. 计算机研究与发展, 2005, 42(2): 210-216.
[10] HARDIN J, ROCKE D M. Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator[J]. Computational Statistics and Data Analysis, 2004, 44: 625-638.
[11] 邵峰晶, 孙仁诚, 郭振波. 基于孤立点发现的彩色图像人脸边缘提取算法[J]. 计算机科学, 2006, 33(9): 201-203.

编　辑　张　俊