用ICA提取高维科学数据的特征

聂琨坤　傅　彦
（电子科技大学计算机学院　成都610054）

摘　要　独立分量分析(ICA)是基于数据高阶统计特性的一种线性变换手段。目前，已广泛应用于盲信号分离和图像识别。文章将此技术引入到科学数据挖掘领域，以求解决预处理中高维复杂特征的提取问题。提出了ICA结合主成分分析(PCA)的特征提取步骤，并结合科学数据集量大的特点给出了一种快速收敛算法—FastICA。最后指出ICA特征提取技术可以应用于高维科学数据挖掘，并且较传统的特征提取技术有更高的准确率。

关键词　独立分量分析(ICA)，特征提取，科学数据挖掘

Using ICA to Extract Features from High-Dimensional Scientific Data

NIE Kun-Kun　FU-Yan
(School of Computer Science and Engineering of UESTC, Chengdu 610054)

Abstract　Independent Component Analysis (ICA) is a linear transformation based on high-order statistic properties of the sample data. It has been widely used for image processing and Blind Source Separation (BSS). This paper introduces ICA to the field of scientific data mining and proposes a framework of ICA feature extraction. Furthermore, we give an efficient algorithm—FastICA and explore the application of proposed framework on high-dimensional scientific data. Experiments show that ICA is suitable for feature extraction and performs better with higher accuracy than other traditional ways in the field of scientific data mining.

Keywords　Independent component analysis (ICA), Feature extraction, Scientific data mining

在模式识别或数据挖掘中，常常需要通过事物的特征来识别事物或是提取有用的信息。特征提取是模式识别预处理中的一个特殊步骤，旨在从原有的特征中根据一定的准则提取有意义特征的子集。高维科学数据，数据规模大、特征多且复杂，使得提取特征十分困难。由此，发展有效的特征提取方法具有重要的意义。

独立分量分析(ICA)是一种线性变换手段，经过变换后，输出的分量尽可能统计独立[1]。利用它提取的特征，不但特征数目大大减少给识别带来方便，而且特征各自独立，符合人类认识事物的生理过程，提高了识别的准确性。目前，ICA的研究主要有三个方向：ICA目标函数的研究，如文[2]；ICA学习算法的研究；ICA各种应用技术的研究。ICA应用技术的研究是热点之一，特别是在特征提取方面的应用技术。现在ICA已成功地应用于图像特征提取和盲信号分离[3]，特别是在盲信号处理方面已取得了不菲的成绩。本文试将ICA的应用拓展到科学数据挖掘领域。

1　ICA特征提取步骤

1.1　特征提取过程

特征提取的过程可以公式化地表示为：

$$\Psi:\mathcal{R}^d\to\mathcal{R}^m, m\leqslant d \tag{1}$$

其中Ψ代表变换的过程，特征空间由d维变换到m维($m\leqslant d$)。这是模式识别和数据分析预处理中重要的一步，它减小了分类计算的复杂度，提高了分类器的辨识能力。ICA的特征提取过程可分为三步：提取前的预处理、选取目标函数、通过学习逐步趋于最优点即最大化输出分量的独立性。

1.2　ICA特征提取步骤

假设目标函数表示为J，学习规则表示为C，ICA的特征提取过程就是根据学习规则C，不断地学习，使得J取得最大或最小值的过程。

1.2.1　定义

设数据具有d维的特征，表示为$X=(x_1,x_2,\cdots,x_d)^T$，为了通过线性变换使得输出分量尽可能地独立，我们假设每个特征x_i都是m个独立的子特征$S=(s_1,s_2,\cdots,s_m)^T$的线性组合，且每个s_i都具有零平均

$$X=AS$$

即

$$x_i=a_1s_1+a_2s_2+\cdots+a_ms_m \tag{2}$$

其中$A=[a_1,a_2,\cdots,a_m]\in\mathcal{R}^{d\times m}(a_i\in\mathcal{R}^d)$是一个秩$r(A)=m$的转换矩阵。ICA特征提取问题就是要取得$S$的最近似估计值，我们用$Y$表示$S$的估计值：

$$Y=\hat{S}=WX \tag{3}$$

这里的Y即是提取后的特征，W可认为是A的逆矩阵，整个ICA特征提取的过程就是不断优化W求取Y的过程。

经典的ICA过程，A和W均是方阵，只能将d维的特征消除二阶依赖性后变为d维，不能降到m维，实现特征空间压缩。为此，提出下面的处理过程。

1.2.2　特征提取过程

(1) 预处理。ICA特征提取中通常有两步预处理工作[4,5]。首先，让数据关于零点对称分布，即是让每个x_i均值都为0

$$E\{x_i\}=0 \tag{4}$$

所以，

$$x_i:=x_i-E\{x_i\} \tag{5}$$

然后，进行白化，即是消除数据的二阶相关性。白化后的数据二阶统计独立，并且具有单位方差。假设白化变换矩阵V，白

化后

$$z=Vx, E\{zz^T\}=I \tag{6}$$

这里我们用 PCA 进行白化

$$V=D^{-1/2}E^T \tag{7}$$

其中 E 是 x 的协方差矩阵对应的特征向量组成的矩阵，D 是该矩阵对应的特征值矩阵，即 $E^TDE=E\{xx^T\}$，让 E 除以 $D^{1/2}$的目的是单位化。显然，特征向量间是正交的。

回顾一下(3)式，此时令 $W=BD^{-1/2}E^T$，那么

$$y=Wx=BD^{-1/2}E^Tx \tag{8}$$

由(7)式，V 是由 PCA 得到的，我们将 D 中的特征值由大到小排序，取前 m 个最大的特征值组成新的特征值矩阵 D_m，对应于这 m 个特征值的特征向量组成矩阵 E_m^T，于是

$$V_m=D_m^{-1/2}E_m^T \tag{9}$$

到此，特征空间已由 d 维降到了 m 维，后面的工作便是求取 m 维统计独立的特征了。

(2) 选取目标函数。目标函数(objective function)又称为对比函数(contrast function)，就是判断结果好坏的判据。我们常常通过学习算法，最大化或最小化目标函数，以取得最优解。在选择目标函数时，通常有两种选择：一种是目标函数的评估基于所有样本；另一种是目标函数的评估只基于部分样本或者逐个考察样本。由上面两种选择导出两类目标函数，Multi-unit(多个体)型和 One-unit(单个体)型。前者就是考察整个样本，后者只考察单个样本。两类目标函数中各种类型的目标函数是等同或者类似的，如互信息目标函数和最大似然目标函数同属第一类，它们是等同的[6]。当然不同的目标函数都各有优缺点，这需要在实际应用中斟酌。我们这里选择负熵目标函数的一种近似形式：

$$J_G(y)=c[E\{G(y)\}-G\{G(v)\}]^2 \tag{10}$$

其中 c 是不相关的常数，v 是期望为零、单位方差的高斯变量，$G(\cdot)$是一个非二次的函数，下面给出一个被证明有用的 G 函数：

$$G(u)=-\exp(-u^2/2) \tag{11}$$

式(10)给出的目标函数形式[7]是兼顾负熵和峰度的平衡形式，它具有很多良好的统计特性，特别是健壮性。

(3) 选取算法求取最优解。在选择了判据准则后，我们希望能够选择一个法则去逐步优化准则函数，使其最大化或者最小化。这个实施的法则便是 ICA 的学习算法。通常有三项指标来衡量一个算法：稳定性、收敛速度、占用内存大小。像目标函数一样，关于 ICA 的算法也有很多。同样，可以将它们分为两类：自适应型和批处理型。下面给出基于自然梯度的权值更新规则[7]：

$$\Delta W=[xg(y)-g'(y)]W \tag{12}$$

其中，$y=Wx=BV_mx$ 和 $g(y)=G'(y)$。

经过有限步迭代，求取最优解，此时的输出分量已最大可能地独立了。

2 科学数据挖掘中的 ICA 特征提取

在数值模拟的科学数据挖掘中，数据特征多而且特征彼此关联，因此特征提取是一个艰巨的任务。下面将提到的 ICA 特征提取步骤应用于科学数据挖掘。

数值模拟的科学数据可能是一个物理过程的多次实验记录，也可能是某一物体的多次测量结果。显然，同一事物的记录必然存在关联，彼此关联的高维特征给模式识别和可视化带来了困难。而 ICA 是一种分离技术，经过改进后，它不仅能提取独立的特征而且能压缩特征空间。由此，ICA 特征提取适用于高维科学数据挖掘。如某一个特征 x_i，可分解为：

$$x_i=a_1s+a_2s_2+\cdots+a_ms_m$$

m 个独立的特征 s_i，分解后的特征较原有特征，个数大大减少，而且各自独立，给识别和可视化带来了方便，其准确率也明显提高[8]。

表1 PCA 和 ICA 的正确识别率[8]

	PCA	ICA
fafb	70.71%	78.33%
DupI	35.18%	36.15%
dupII	15.38%	15.81%
fafc	4.64%	6.70%

为了简单起见，可选用(10)式为目标函数，使用 FastICA 算法进行权值更新。之所以叫 FastICA 是因为这种算法比别的算法收敛速度快。现假设只有一个神经元(one-unit)，权值向量为 W，运用 FastICA 去找到权值 W 的更新方向。基本的 FastICA 算法描述如下：

(1) 随机地选择权值向量 W 进行初始化；

(2) 以 $W^+=E\{Xg(W^TX)\}-E\{g'(W^TX)\}W$ 更新权值；

(3) 对更新后的权值 W^+进行规范化 $W=W^+/\|W^+\|$；

(4) 是否收敛?如果否，转向(2)，如果收敛，结束。

其中函数 $g(\cdot)$是(10)中非二次函数 $G(\cdot)$的变体，如 $g(u)=u\exp(-u^2/2)$等。当然，目前 FastICA 算法有很多变体，具体可参见文[9,10]。

结论 我们提出了提取高维科学数据复杂特征的一种方法，并且给出了应用框架，做了简单的应用比较。ICA 确实能应用于科学数据挖掘，在模式识别预处理中提取关键特征。在图像处理和盲信号分离方面，ICA 已表现出优越的分离能力。我们将在现实科学数据集上进一步测试 ICA 的分离性能和研究 ICA 的实现技术。

参 考 文 献

1 Comon P. Independent component analysis, A new concept? [J]. Signal Processing, 1994, 36: 287～314

2 Shun-ichi A. Natural gradient works efficienty in learning[J]. Neural Computation, 1998,10: 251～276

3 Ding Peilu, Zhang Liming. ICA feature extraction-framework and application[J]. Chinese Journal of Electronics,2003, 12(1): 106～110

4 Bell A J, Sejnowski T J. The "independent components" of natural scenes are edge filters[J]. Vision Res., 1997, 37: 3327～3338

5 Hoyer P, Hyv_arinen A. Independent component analysis applied to feature extraction from colour and stereo images[J]. Network: Comput. Neural Systems, 2000, 11 (3): 191～210

6 Cardoso J-F. Infomax and maximum likelihood for source separation [J]. IEEE Letters on Signal Processing, 1997,4: 112～114

7 Hyvärinen A. New approximation of differential entropy for independent component analysis and projection pursuit [J]. Advances in Neural Information Processing System, 1998,10:273～279

8 Baek K,et al. PCA vs. ICA: A comparison on the FERET data set. http:// www.cs.colostate.edu/evalfacerec/papers/cvprip02.pdf,2003

9 Hyvärinen A. Fast and robust fixed-point algorithms for independent component analysis [J]. IEEE Trans. on Neural Networks, 1999,10(3):626～634

10 Hyvärinen A. The fixed-point algorithm and maximum likelihood estimation for independent component analysis [J]. Neural Processing Letters,1999,10(1):1～5