第25卷 第4期
1999年7月

自 动 化 学 报
ACTA AUTOMATICA SINICA

Vol. 25, No. 4
Jul., 1999

短文

基于误差模型的自适应鲁棒主成分分析[1)]

王 松 夏绍玮
(清华大学自动化系 北京 100084)

摘 要 研究了改善主成分分析(PCA)算法鲁棒性的一种实现途径. 通过对误差函数的建模分析,得到一种改进的目标函数. 提出一种新的在线自适应式的鲁棒 PCA 运算规则. 该方法基于单层线性神经网络(NN)结构,但是权值的训练算法是非线性的. 从而在迭代训练中对"劣点"样本加以适当处理来排除对运算精度和收敛性的影响.

关键词 主成分分析(PCA),自适应鲁棒 PCA,劣点,神经网络,极大似然估计.

ADAPTIVE ROBUST PRINCIPAL COMPONENT ANALYSIS BASED ON ERROR MODELING

WANG Song XIA Shaowei
(*Department of Automation, Tsinghua University, Beijing* 100084)

Abstract One way to improve the robustness of principal component analysis (PCA) is studied in the paper. A new adaptive algorithm of robust PCA based on the structure of single layer neural network (NN) is developed with modification of the cost function which can be acquired through modeling of the error function. The new nonlinear robust PCA algorithm can reduce the effects of outliers on the accuracy and convergence of the PCA algorithm through proper processing of them.

Key words Principal component analysis (PCA), adaptive robust PCA, outliers, neural network (NN), maximum likelihood estimate.

1 前言

传统的主成分分析(PCA)是通过对协方差矩阵的特征值分解来实现的[1]. 这是一种批处理式的算法. 近年来,随着人工神经网络理论的引入,建立了多种基于单层线性神经网络的主成分提取算法[2]. 从而克服了传统批处理统计方法需要预知所有学习样本、更新

1)国家自然科学基金资助课题(基金编号 69775001).
收稿日期 1997-01-16 收到修改稿日期 1998-08-25

计算复杂等特点.然而,许多学者已经发现,各种基于简单的单层线性神经网络的算法与传统的批处理统计方法一样,存在严重的鲁棒性问题.本文从消除"劣点"的影响这一角度出发去研究,提出一种鲁棒的主成分分析算法.

2　线性自适应 PCA

设输入 $\boldsymbol{x}$ 为 n 维的零均值的随机向量.传统上的主成分分析是通过对 $\boldsymbol{x}$ 的协方差矩阵进行特征值分解来获得.若利用 $n\times m$ 维($m<n$)变换矩阵 W 实现维数压缩,则 $z=\|\boldsymbol{e}^2\|=\|\boldsymbol{x}-WW^{\mathrm{T}}\boldsymbol{x}\|^2$ 为信号重构误差.定义目标函数

$$J_1(W)=\frac{1}{N}\sum_{i=1}^{N}\|\boldsymbol{x}_i-WW^{\mathrm{T}}\boldsymbol{x}_i\|^2. \tag{1}$$

已经严格证明[2],$J_1(W)$最小所对应的 W 的 m 个列向量就是 $\boldsymbol{x}$ 的前 m 个主方向所张成的 PCA 子空间.用随机梯度下降法进行求解.可得

$$W_{k+1}=W_k+\mu_k(\boldsymbol{x}_k\boldsymbol{e}_k^{\mathrm{T}}W_k+\boldsymbol{e}_k\boldsymbol{x}_k^{\mathrm{T}}W_k), \tag{2}$$

其中　$\boldsymbol{e}_k=\boldsymbol{x}_k-W_kW_k^{\mathrm{T}}\boldsymbol{x}_k$ 为本步重构误差,k 为迭代步数,μ_k 为步长.

经研究发现,式(2)中 $\boldsymbol{x}_k\boldsymbol{e}_k^{\mathrm{T}}W_k$ 的影响较小[3],忽略之,可得

$$W_{k+1}=W_k+\mu_k\boldsymbol{e}_k\boldsymbol{x}_k^{\mathrm{T}}W_k. \tag{3}$$

这就是著名的 Qja 自适应学习规则,这是最早提出的自适应 PCA 算法.

3　基于误差模型的自适应鲁棒 PCA

所谓"劣点"样本通常就是指其重构误差 z 相对平均值过大的样本.其比例通常很小,但往往对 PCA 计算结果产生很大的影响.本文的目的就是削弱其影响,提高 PCA 的鲁棒性.假设训练集中样本的重构误差独立且同分布,就可以针对选择的重构误差模型来求解 W 的极大似然估计.

3.1　负指数分布模型

首先假设 z 服从负指数分布

$$f(z)=\begin{cases}\lambda\mathrm{e}^{-\lambda z}, & z\geqslant 0,\\ 0, & z<0.\end{cases} \tag{4}$$

其中 $\lambda>0$.这样训练样本集合的似然函数为

$$\prod_{i=1}^{N}f(z_i)=\prod_{i=1}^{N}\lambda\mathrm{e}^{-\lambda\|\boldsymbol{x}_i-WW^{\mathrm{T}}\boldsymbol{x}_i\|^2}. \tag{5}$$

由于参数 λ 和 N 是确定的常数,矩阵 W 的极大似然估计应该满足

$$\min_W\{\frac{1}{N}\sum_{i=1}^{N}\|\boldsymbol{x}_i-WW^{\mathrm{T}}\boldsymbol{x}_i\|^2\}. \tag{6}$$

显然,式(6)与式(1)是完全等价的.可见第2节中介绍的线性 PCA 算法实际上就是基于负指数分布重构误差模型的.

图1的曲线1是负指数分布的密度函数.可以看出,随着 z 的增大,密度函数的数值陡降.这意味着负指数误差分布模型认为重构误差很大的样本出现概率极小.这表明负指数

误差分布模型没能充分考虑“劣点”样本的存在和影响.

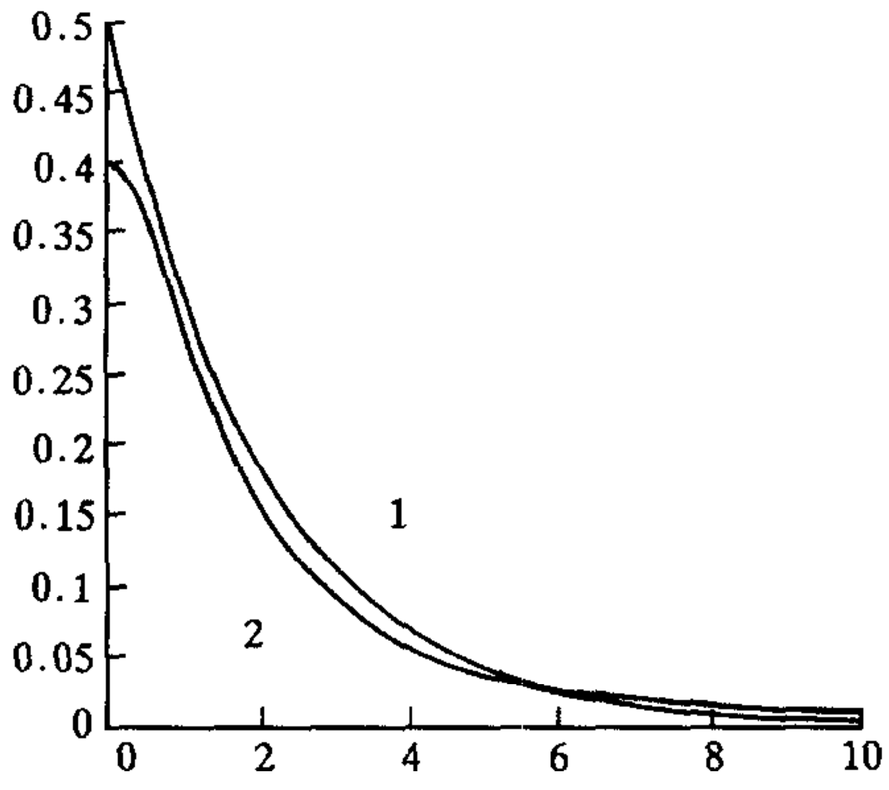

图1 负指数分布和改进柯西分布密度函数

3.2 柯西分布模型

为了提高鲁棒性,这里采用柯西分布作为误差分布.随着误差的增加其密度函数下降速度要慢一些.由于 $z\geqslant 0$,还需要对普通的柯西分布稍作改动,使其变量小于零时密度函数值为零.选择密度函数

$$g(z)=\begin{cases}\dfrac{2}{\pi}\cdot\dfrac{\theta}{\theta^2+z^2}, & z\geqslant 0,\\ 0, & z<0,\end{cases}\tag{7}$$

其中 $\theta>0$.图1中曲线2就是修正的柯西分布密度函数.

由此可以得到似然函数为

$$\prod_{i=1}^{N}g(z_i)=\prod_{i=1}^{N}\frac{2}{\pi}\cdot\frac{\theta}{\theta^2+z_i^2}.\tag{8}$$

由于 θ 为常数,因而,对 W 的极大似然估计可以表示为

$$J_2(W)=\frac{1}{N}\sum_{i=1}^{N}\ln(\theta^2+z_i^2)=\frac{1}{N}\sum_{i=1}^{N}\ln(\theta^2+\|\boldsymbol{x}_i-WW^{\mathrm{T}}\boldsymbol{x}_i\|^4).\tag{9}$$

3.3 自适应鲁棒 PCA 算法

针对目标函数式(9),利用随机梯度下降法,可得到 W 的迭代求解公式

$$W_{k+1}=W_k+\mu_k h(\|\boldsymbol{e}_k\|^2)(\boldsymbol{x}_k\boldsymbol{e}_k^{\mathrm{T}}W_k+\boldsymbol{e}_k\boldsymbol{x}_k^{\mathrm{T}}W_k),\tag{10}$$

其中

$$h(\|\boldsymbol{e}_k\|^2)=\frac{2\cdot(\|\boldsymbol{e}_k\|^2)}{\theta^2+(\|\boldsymbol{e}_k\|^4)}.\tag{11}$$

对应 Oja 规则式(3),忽略式(10)中 $\boldsymbol{x}_k\boldsymbol{e}_k^{\mathrm{T}}W_k$ 的影响可得

$$W_{k+1}=W_k+\mu_k h(\|\boldsymbol{e}_k\|^2)\boldsymbol{e}_k\boldsymbol{x}_k^{\mathrm{T}}W_k).\tag{12}$$

可以看出,修正后的算法与原有的算法相比,引进一个非线性的环节 $h(\|e\|^2)$.

4 仿真结果

在仿真中,随机均匀产生500个主方向为 $(1/\sqrt{2},1/\sqrt{2})^{\mathrm{T}}$ 的零均值二维样本集,再随机插入10个“劣点”.考虑提取其最大主成分,即 $n=2,m=1$.用传统的特征值分解算法的主成分分析结果为 $W=[0.960,0.280]^{\mathrm{T}}$.可见,统计算法鲁棒性很差.

图2是各种自适应算法的计算结果.纵坐标表示迭代中的 W_k 与实际的主方向之间的误差角度.可以看出,线

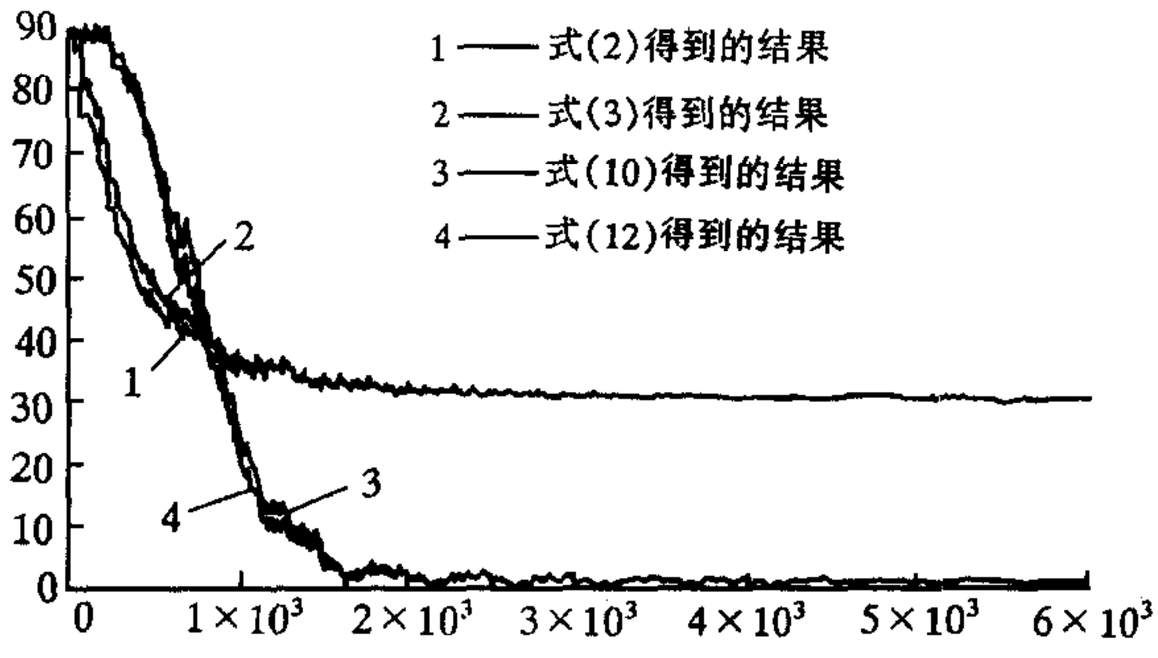

图2 样本集 PCA 仿真结果

性 PCA 的鲁棒性很差. 仅仅2%的"劣点"样本,导致最终求得的主方向与实际值误差达到30°. 而修正的鲁棒 PCA 算法最终能够收敛到正确结果.

5　结论

本文从削弱"劣点"影响的角度发出,采用一种改进的柯西分布作为重构误差的分布模型. 这种模型相对于负指数模型考虑了"劣点"的存在概率. 基于改进的柯西分布误差模型,采用极大似然估计方法. 得到一种改进的目标函数. 然后利用随机梯度下降法得到一种新的自适应的鲁棒 PCA 算法. 实验仿真的结果证明,修正的算法较之简单线性 PCA 算法在鲁棒性上有了明显的改进.

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,1995,73—83

2　Diamantaras K I, Kung S Y. Principal Component Neural Networks: Theory and Applications. New York: John Wiley & Sons, 1996,44—48

3　Xu Lei, Yuille A L. Robust principal component analysis by self-organizing rules based on statistical physics approach. *IEEE Trans. Neural Networks*, 1995,**6**(1):131—143

王　松　男,1998年7月于北京清华大学自动化系获工学博士学位. 主要研究方向是神经网络、模式识别和聚类与分类的理论方法.

夏绍玮　女,北京清华大学自动化系教授,博士生导师. 主要研究方向是系统工程、智能决策和神经网络理论及应用.