

## 基于最大间隔超平面的增强特征提取算法

侯 勇<sup>1,2\*</sup>, 郑雪峰<sup>1</sup>

(1. 北京科技大学 计算机与通信工程学院, 北京 100083 2. 山东经贸职业学院 科学与人文学院, 山东 潍坊 261011)

(\* 通信作者电子邮箱 aspnetcs@163.com)

**摘 要:**核主成分分析(KPCA)与多层感知器(MLP)是流行的特征提取算法,但这些算法存在效率低下与易陷于局部最优解等问题。针对KPCA与MLP算法存在的问题,提出了一个新颖的特征提取算法——基于最大间隔超平面的增强的特征提取算法(EFE)。该算法独立于输入样本的概率分布,通过采用间隔最大化且两两正交的最大分割超平面,将输入样本映射到超平面的法线所张成的子空间中,实现输入样本的特征提取。在对现实世界数据集wine与AR的特征提取的实验表明,基于最大间隔超平面的增强特征提取算法在执行效率、识别准确率方面均超出了KPCA与MLP的执行效率与识别准确率。

**关键词:**特征提取;降维;核主成分分析;多层感知器;最大间隔超平面;内在维数

**中图分类号:** TP339 **文献标志码:** A

### Margin maximizing hyperplanes based enhanced feature extraction algorithm

HOU Yong<sup>1,2\*</sup>, ZHENG Xuefeng<sup>1</sup>

(1. School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083 China;

2. College of Humanities and Science, Shandong Vocational College of Economics and Business, Weifang Shandong 261011, China)

**Abstract:** Kernel Principal Component Analysis (KPCA) and Multi-Layer Perceptron (MLP) neural network are popular feature extraction algorithms. However, these algorithms are inefficient and easy to fall into local optimal solution. The paper proposed a new feature extraction algorithm — margin maximizing hyperplanes based Enhanced Feature Extraction algorithm (EFE), which can overcome the problem of KPCA and MLP algorithm. The proposed EFE algorithm, which maps the input samples to the subspace spanned by the normals of hyperplanes through adopting the pairwise orthogonal margin maximizing hyperplanes, is independent of the probability distribution of the input samples. The results of these feature extraction experiments on real world data set — wine and AR show that EFE algorithm is beyond KPCA and MLP in terms of the efficiency of the implementation and accuracy of recognition.

**Key words:** feature extraction; dimensionality reduction; Kernel Principal Component Analysis (KPCA); Multi-Layer Perceptron (MLP); Margin maximizing hyperplanes; intrinsic dimension

## 0 引言

现实世界中,诸如声音信号、数字图像、fMRI扫描<sup>[1]</sup>等通常都具有高维的数据信息,在处理此类高维数据之前,通常要对其进行降维,即特征提取。特征提取<sup>[2]</sup>就是将高维数据转换成有意义的低维数据。理想的情况下,映射后数据的低维表示应该有一个维数称为数据的内在维数。数据的内在维数是用来考虑所观察数据性质的最小维数。由于特征提取减轻了维数灾难和其他高维空间的意外性质,所以特征提取在许多领域里有重要的应用。因此降维有利于对高维数据进行分类、压缩与可视化操作。流行的特征提取算法有核主成分分析(Kernel Principal Component Analysis, KPCA)与多层感知器(Multi-Layer Perceptron, MLP)等。然而这些线性技术不能充分地处理复杂的非线性数据。

在缺乏一个比较系统的特征提取理论的推动下,对KPCA与MLP特征提取算法进行了比较研究。提出了一种新颖的特征提取算法——增强的特征提取方法(Enhanced Feature Extraction algorithm, EFE)。实验结果表明该算法的效率与识别准确率均能够超越KPCA与MLP特征提取算法。

## 1 特征提取

特征提取问题的定义如下:假设有一个数据集,用 $n \times D$

的矩阵 $X$ 表示,即该数据集由 $D$ 维 $n$ 个数据向量组成。进一步假设,这个数据集有其内在维数 $d$ (其中 $d \ll D$ )。从数学的角度上讲,此处的内在维数意味着数据集 $X$ 中的点位于 $d$ 维流形上或附近处,该 $d$ 维流形嵌入在 $D$ 维空间中。特征提取技术将 $D$ 维数据集 $X$ 转化成一个新的 $d$ 维数据集 $Y$ ,同时尽可能多地保留该数据集的几何性质。在一般情况下,数据流形的几何性质与数据集 $X$ 的内在维数 $d$ 都是未知的。因此,特征提取(降维)是一个病态问题,只能通过假设数据具有某些特性(例如其内在维数),才能对其进行有效的特征提取。

### 1.1 KPCA 特征提取算法<sup>[3]</sup>

KPCA是用核函数将传统的线性主成分分析(Principal Component Analysis, PCA)扩展到高维空间中。近年来,人们用“核技巧”对线性特征提取技术进行改进,并成功地提出了核岭回归<sup>[4]</sup>,支持向量机。

由于KPCA是基于核的非线性映射方法,KPCA的映射性能依赖于核函数的选择,常用的核函数包括:线性核(线性核使KPCA等同于线性PCA)、多项式核与高斯核<sup>[5]</sup>。

KPCA已经成功地应用于许多领域,如人脸识别<sup>[6]</sup>、语音识别与异常检测<sup>[7]</sup>。KPCA的一个重要弱点是:核矩阵的大小同数据集中实例数的平方成正比。

### 1.2 MLP 特征提取算法<sup>[8]</sup>

MLP特征提取算法是隐含层有奇数个神经元的前向神

收稿日期:2012-10-19;修回日期:2012-12-10。

作者简介:侯勇(1978-),男,山东蓬莱人,讲师,博士研究生,主要研究方向:数据挖掘、网络安全、机器学习;郑雪峰(1951-),男,福建福州人,教授,主要研究方向:网络安全。

神经网络。其中中间的隐含层有  $d$  个节点,输入与输出层有  $D$  个节点。该网络的目标函数是最小化输入值与输出值的均方差。为了使自动编码器能够将数据从高维空间完美地映射到低维空间,通常在网络中使用 sigmoid 激活函数。

MLP 的各层之间通常有大量的连接。因此,训练网络的误差后向传播算法<sup>[9]</sup>的收敛速度很慢,并且极有可能陷入局部极小值<sup>[10]</sup>。

MLP 已经成功应用于许多领域,如丢失数据的归集,HIV 病毒<sup>[11]</sup>的分析等。

## 2 增强的特征提取方法

针对 KPCA 与 MLP 特征提取算法存在效率低下与易陷于局部最优解等问题,本章给出所提出的特征提取算法——增强的特征提取方法。该算法不依赖于输入样本的概率分布<sup>[12]</sup>,通过采用最大分割超平面<sup>[13]</sup>,将输入样本映射到由一组间隔最大化且两两正交的超平面的法线所张成的子空间中,实现输入样本的特征提取,再用提取出的特征训练出具有差异性的基分类器。

增强的特征提取算法的细节如下所示:

设定  $X, y$  是训练数据,  $X = (x_1, \dots, x_n)$  是输入样本 ( $x_j \in \mathbf{R}^d, j = 1, 2, \dots, n$ ),  $y \in \{-1, +1\}^n$  是相应的类标签。假设  $(x_i, y_i) (i = 1, \dots, n)$  是相互独立,同分布随机变量。通过求解下列优化问题(1)得到最大间隔超平面:

$$\begin{aligned} \min_{w, b, \varepsilon_i} & \|w\|^2 + b^2 + C \sum_{i=1}^m \varepsilon_i^2 \\ \text{s. t. } & y_i(w^T \varphi(x_i) + b) \geq 1 - \varepsilon_i, i = 1, \dots, m \\ & u_q^T w = 0; q = 1, \dots, s \\ & u_q = w_q / \|w_q\| \end{aligned} \quad (1)$$

其中:  $w^T \varphi(x_i) + b$  是要求解的超平面;  $\varphi$  是映射函数;  $k, \varepsilon = [\varepsilon_1, \dots, \varepsilon_m]^T \in \mathbf{R}_+^m$  是误差松弛变量;  $C$  是正则化参数,用来权衡误分类代价。

通过引入拉格朗日乘子  $\alpha = [\alpha_1, \dots, \alpha_m]^T \in \mathbf{R}_+^m$  与  $\gamma = [\gamma_1, \dots, \gamma_s]^T \in \mathbf{R}^s$ , 并应用拉格朗日乘子方法,可得到问题(1)的对偶式。

$$\max_{\alpha, \gamma} (2\alpha^T I - \alpha^T \hat{K} \alpha - 2\alpha^T Y \Phi^T U \gamma - \gamma^T U^T U \gamma); \alpha \geq 0 \quad (2)$$

其中:

$$U_{dxs} = [u_1, \dots, u_s]$$

$d$  是核引导特征空间的维数。

$$Y_{m \times m} = \text{diag}(y_1, \dots, y_m) \quad (4)$$

$$\hat{K}_{m \times m} = Y(K + II^T + I/C)Y \quad (5)$$

其中:  $K_{m \times m} = \Phi^T \Phi, \Phi = [\varphi(x_1), \dots, \varphi(x_m)]$ ,  $\hat{K}$  是核矩阵,核函数是  $k(z_i, z_j) = y_i y_j k(x_i, x_j) + y_i y_j + \delta_{ij} y_i y_j / C, z_i = (x_i, y_i), z_j = (x_j, y_j)$ , 且  $\delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$ , 对应于  $k$  的特征提取为:

$$\hat{\varphi}(z_i) = [y_i \varphi(x_i)^T, y_i, y_i / \sqrt{C} e_i^T]^T \quad (6)$$

其中:  $e_i \in \mathbf{R}^m$  的第  $i$  个元素是 1,其余的元素为 0。很明显,式(2)是一个二次规划问题。

基于 KKT (Karush-Kuhn-Tucker) 条件,原始问题的最优解可由最优值  $\alpha$  与  $\gamma$  表示,即

$$w = \sum_{i=1}^m \alpha_i y_i \varphi(x_i) + \sum_{q=1}^s \gamma_q u_q \quad (7)$$

$$b = \sum_{i=1}^m \alpha_i y_i$$

$$\xi_i = \alpha_i / C$$

第  $s$  次迭代所获得的最优解  $\alpha_s$  是  $\alpha_s = [\alpha_{s1}, \dots, \alpha_{sm}]^T$

$$w_1 = \sum_{i=1}^m \alpha_{i1} y_i \varphi(x_i) = \Phi Y \alpha_1 \quad (8)$$

其中:  $\Phi = [\varphi(x_1), \dots, \varphi(x_m)]$ , 核函数  $K = \Phi^T \Phi$ , 因此式(8)是  $\varphi(x_1), \dots, \varphi(x_m)$  的线性组合。基于式(7)与(8),则可得:

$$w_2 = \Phi Y \alpha_2 + \gamma_1 u_1 = \Phi Y (\alpha_2 + \gamma_1 \alpha_1 / \|\Phi Y \alpha_1\|)$$

则  $w_2$  仍是  $\varphi(x_i)$  的线性组合,通过归纳可得出,算法的每次迭代,所求得的权值  $w_i$  都是  $\varphi(x_1), \dots, \varphi(x_m)$  的线性组合。

基于上面的分析,下面给出增强的特征提取方法的过程:

给定问题(1)的参数  $(X, y, C)$ , 依据式(7)与(8),得到最优解  $\alpha, \gamma$  与  $w_1$ , 则第一个提取的特征  $f_1(x) = w_1^T x$ 。转换样本数据,即将样本数据投影到同  $w_1$  正交的空间中。为简单起见,假设  $w_1$  已被规范化,即  $\|w_1\|_2 = 1$ , 则映射后的数据为

$$\begin{aligned} x'_i &= x_i - (w_1^T x_i) w_1; i = 1, \dots, n \\ x' &= (x'_1, \dots, x'_n) \end{aligned} \quad (9)$$

用  $X'$  表示矩阵  $(x'_1, \dots, x'_n)$ , 应用式(9)求解出的参数为  $(X', y, C)$  的(1)的解为  $(w_2, \gamma_2)$ , 则第二个提取出的特征为  $f_2(x) = w_2^T x'$ , 其中  $x' = x - (w_1^T x) w_1$ 。

根据所需要提取的特征维数,重复上述特征提取过程。

## 3 仿真实验

第一个实验,展示了增强的特征提取算法的性能。本次实验采用 wine 数据集。应用 KPCA、MLP 与 EFE 对该数据集进行特征提取。对该数据集进行 2 维特征提取,结果如图 1 所示。用 KPCA 与 MLP 对 wine 数据集进行特征提取时,该数据集被映射到对应于两个最大特征值的特征向量上。由于增强的特征提取算法对应于二分类操作,因此,对多分类操作, EFE 采用一对多<sup>[14]</sup>的方式对不同的类进行分组,即当对一个未知样本进行分类时,将该类划分为具有最大分类函数值的那类。本次实验增强的特征提取算法首先将第一类与第二类划分到一个组,然后将第二类与第三类划分成一组。

本次实验,将所有算法都将 wine 数据集都映射到 2 维空间,结果如图 1 与表 1 所示。

图 1(a)是将 wine 数据集映射到前两维(酒精维与苹果酸维)所张成的子空间的散点图,其余三个子图分别是经 KPCA、MLP 与 EFE 算法转换后的散点图<sup>[15]</sup>。

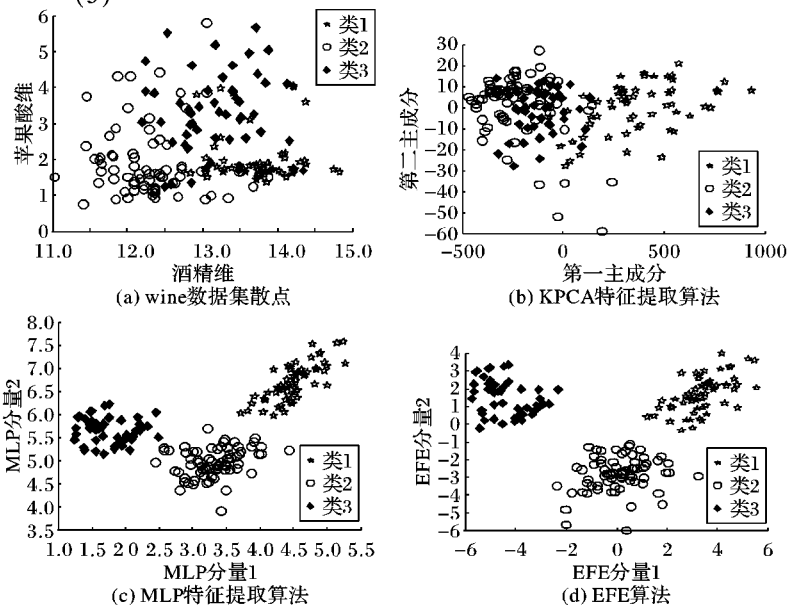


图 1 wine 映射到二维子空间的散点图

表1 算法时间与距离

算法	耗用时间	类1-2 距离	类2-3 距离	类1-3 距离
KPCA	0.667 67	0.000 04	NaN	0.000 04
MLP	0.019 79	0.440 78	0.233 99	1.263 82
EFE	0.000 03	1.655 18	2.842 26	3.860 89

从图1与表1可以看出,在映射空间中,MLP与EFE算法的特征提取,都实现了不同类的线性分组,基于EFE算法所对应的不同类的划分间隔最大,所用的时间最短,因此,同MLP、KPCA相比,EFE的性能最好的。

MLP的特征提取算法对wine数据集的三类与二类的划分间隔最大,特征提取达到了最优,而对二类与一类,三类与一类的划分的距离很小,即特征提取陷入了局部最优解。

从一定程度上讲,不同类的分离间隔可用正则化参数 $C$ 控制,本次实验,EFE算法的 $C$ 的取值是6,高斯核函数是 $k(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x} - \mathbf{y}\|^2 / \beta)$ ,  $\beta = \frac{1}{m^2} \sum_{i,j=1}^m \|\mathbf{x}_i - \mathbf{x}_j\|^2$ ,  $m$ 是训练样本的数量。

接下来,在AR人脸数据集上对MLP、KPCA与EFE算法的性能进行了比较,使用 $K$ 近邻分类器<sup>[16]</sup>。

AR是一个用于人脸识别的数据集,它包含了126个人(70个男人与56个女人)的不同表情图,共1638幅图像。每幅图像的大小是 $768 \times 576$ ,为了便于处理,通过采样将图像大小降低到 $700 \times 500$ ,并将该数据集的方差与均值分别标准化成0与1。

在基于MLP的特征提取算法中,MLP采用3层结构,分别是输入层\隐含层与输出层,激活函数是sigmoid函数。 $K$ 近邻算法的 $k$ 设置成1。KPCA算法采用高斯核,宽度参数 $\sigma = 100000$ 。EFE算法的正则化参数 $C$ 取值是0.618,并采用固定的高斯核函数,即 $k(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x} - \mathbf{y}\|^2 / \beta)$ ,  $\beta = \frac{1}{m^2} \sum_{i,j=1}^m \|\mathbf{x}_i - \mathbf{x}_j\|^2$ ,  $m$ 是训练样本的数量。

实验结果如表2所示。

表2 各个算法的性能

提取的特征数	识别的准确率			耗用时间/s		
	KPCA	MLP	EFE	KPCA	MLP	EFE
13	0.8544	0.85117	0.91747	11	12	1
14	0.8629	0.86810	0.95462	14	13	3
15	0.9315	0.92358	0.97578	16	17	4
16	0.9494	0.95838	0.98706	22	23	6
17	0.9560	0.96026	0.99224	28	45	8
18	0.9489	0.94610	0.99600	36	59	10

随着所提取特征数的增加,所耗用的CPU时间也随之增加,识别准确率呈上升趋势,但是当提取的特征数增加到一定程度时,识别准确率不再上升,甚至出现下降的趋势,这是因为提取出的特征数已经超过了该数据集的内在维数。

最后可以看出,EFE的效率与识别准确率始终能够超过MLP与KPCA的效率与识别准确率,并且不会陷入局部最优解。

由于在AR数据集的图像中包含相当大面积的太阳眼镜/围巾,这使得KPCA与MLP特征提取算法效率与识别准确率低下。另一个现象是KPCA与MLP特征提取算法在性能上相差无几。

#### 4 结语

在介绍KPCA与MLP特征提取算法时,给出了这些算法

的缺陷,并提出了一种新颖的特征提取算法——增强的特征提取算法。在实验中,对这些算法的性能进行了比较。基于得到的结果,可以得出如下结论:增强的特征提取算法在效率与识别准确率方面均超出了KPCA与MLP特征提取算法执行效率与识别准确率,并且没有陷入局部最优解。在未来,研究重点是将增强的特征提取算法应用到具有可被优化的目标函数求解上。

#### 参考文献:

- [1] SUN F, MORRIS D, LEE W, *et al.* Feature-space-based fMRI analysis using the optimal linear transformation[J]. *IEEE Transactions on Information Technology in Biomedicine*, 2010, 14(5): 1279–1290.
- [2] ZHANG J, ZHU L P, ZHU L X, *et al.* On a dimension reduction regression with covariate adjustment[J]. *Journal of Multivariate Analysis*, 2012, 104(1): 39–55.
- [3] OGAWA T, HASEYAMA M. Missing intensity interpolation using a kernel KPCA-based POCS algorithm and its applications[J]. *IEEE Transactions on Image Processing*, 2011, 20(2): 417–432.
- [4] KIM N, JEONG Y S, JEONG M K, *et al.* Kernel ridge regression with lagged-dependent variable: applications to prediction of internal bond strength in a medium density fiberboard process[J]. *IEEE transactions on systems, man and cybernetics, Part C: Applications and reviews*, 2012, 42(6): 1011–1020.
- [5] MA X, ZABARAS N. Kernel principal component analysis for stochastic input model generation[J]. *Journal of Computational Physics*, 2011, 230(19): 7311–7331.
- [6] WEN Y, HE L, SHI P, *et al.* Face recognition using difference vector plus KPCA[J]. *Digital Signal Processing*, 2012, 22(1): 140–146.
- [7] GU Y, LIU Y, ZHANG Y. A selective KPCA algorithm based on high-order statistics for anomaly detection in hyperspectral imagery[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2008, 5(1): 43–47.
- [8] PARK J, DIEHL F, GALES M J F, *et al.* The efficient incorporation of MLP features into automatic speech recognition systems[J]. *Computer Speech and Language*, 2011, 25(3): 519–534.
- [9] OH S-H. Error back-propagation algorithm for classification of imbalanced data[J]. *Neurocomputing*, 2011, 74(6): 1058–1061.
- [10] MOALLEM P, AYOUGH S A. Removing potential flat spots on error surface of MultiLayer Perceptron (MLP) neural networks[J]. *International Journal of Computer Mathematics*, 2011, 88(1/2/3): 21–36.
- [11] KIM G, KIM Y, LIM H-S, *et al.* An MLP-based feature subset selection for HIV-1 protease cleavage site analysis[J]. *Artificial Intelligence in Medicine*, 2010, 88(2/3): 83–89.
- [12] HU G X, WNAG K. The estimation of probability distribution of SDE by only one sample trajectory[J]. *Computers and Mathematics with Applications*, 2011, 62(4): 1798–1806.
- [13] PLASTRIA F, CARRIZOSA E. Minmax - distance approximation and separation problems: geometrical properties[J]. *Mathematical Programming*, 2012, 132(1/2): 153–177.
- [14] KHAN N M, KSANTINI R, AHMAD I S, *et al.* A novel SVM + NDA model for classification with an application to face recognition[J]. *Pattern Recognition*, 2012, 45(1): 66–79.
- [15] VIAU C, MCGUFFIN M J, CHIRICOTA Y, *et al.* The FlowViz-Menu and parallel scatterplot matrix: hybrid multidimensional visualizations for network exploration[J]. *IEEE Transactions on Visualization and Computer Graphics*, 2010, 16(6): 1100–1108.
- [16] JIANG J Y, TSAI S C, LEE S J, *et al.* FSKNN: multi-label text categorization based on fuzzy similarity and  $k$  nearest neighbors[J]. *Expert Systems with Application*, 2012, 39(3): 2813–2821.