第 38 卷 第 2 期 计 算 机 工 程 2012 年 1 月
Vol.38 No.2 Computer Engineering January 2012

·人工智能及识别技术· 文章编号：1000—3428(2012)02—0175—03 文献标识码：A 中图分类号：TN911.72

基于核熵成分分析的数据降维

黄丽瑾，施 俊，钟 瑾

(上海大学通信与信息工程学院，上海 200072)

摘 要：针对高维数据的维灾问题，采用核熵成分分析方法降维数据，并与主成分分析及核主成分分析方法进行对比。降维后的数据利用支持向量机算法进行分类，以验证算法有效性。实验结果表明，KECA 在较低的维数时仍然能获得较好的分类精度，可以减少后续的处理复杂度和运行时间，适用于机器学习、模式识别等领域。

关键词：降维；核熵成分分析；核主成分分析；支持向量机

Data Dimension Reduction Based on Kernel Entropy Component Analysis

HUANG Li-jin, SHI Jun, ZHONG Jin

(School of Communication and Information Engineering, Shanghai University, Shanghai 200072, China)

【Abstract】Aiming at the curse of dimensionality, the kernel entropy component analysis(KECA) is used to reduce the dimension of data, which is compared with Principal Component Analysis(PCA) and Kernel PCA(KPCA). The low dimensional data after dimension reduction are classified by Support Vector Machine(SVM) algorithm to compare the accuracy. Experimental results indicate that high classification accuracy can be obtained at low dimension number with KECA, which reduces the processing complexity and running time. It suggests that KECA-based dimension reduction algorithm has the feasibility to be applied in the fields of machine learning, pattern recognition, etc.

【Key words】dimension reduction; Kernel Entropy Component Analysis(KECA); Kernel Principal Component Analysis(KPCA); Support Vector Machine(SVM)

DOI: 10.3969/j.issn.1000-3428.2012.02.057

1 概述

数据降维是解决维灾问题的有效手段[1-3]。降维技术分为线性和非线性 2 大类。常见的线性降维方法包括主成分分析(Principal Component Analysis, PCA)、多维缩放、因子分析、投影追踪、线性判别分析、局部保留投影、独立分量分析等。常见的非线性降维方法包括自组织映射网络、核主成分分析(Kernel Principal Component Analysis, KPCA)、Isomap、局部线性嵌入、流形学习等[2-3]。这些方法在不同的应用领域都取得了很好的效果[1-3]。PCA 是应用广泛的经典降维方法之一[4]，它借助于一个正交变换，将其分量相关的原随机向量转化成分量不相关的新随机向量。PCA 的缺点之一是存储空间大，计算复杂度高，且线性映射在一定程度上影响其效果。而 KPCA 是 PCA 的非线性推广[5-6]，可处理大量非线性问题，且更为简洁高效。核熵成分分析(Kernel Entropy Component Analysis, KECA)是一种新的数据变换方法[7-8]。它是在 KPCA 的基础上引入熵的概念，在特征空间进行熵成分分析以实现数据变换，具有很好的非线性处理能力。目前，KECA 已初步应用于数据聚类和图像去噪，并取得了很好的效果，但还未有将 KECA 应用于降维处理的较为全面的研究。

本文基于 KECA 理论，结合支持向量机(Support Vector Machine, SVM)分类算法，研究分析 KECA 应用于降维的可行性与降维性能。

2 核熵成分分析原理与降维算法

2.1 KECA 基本原理

KECA 的原理介绍如下[8]：

给定 N 维样本 x， $p(x)$是概率密度函数，则 Renyi 熵为：

$$H(p)=-\lg\int p^2(x)\mathrm{d}x \tag{1}$$

令 $V(p)=\int p^2(x)$，采用 Parzen 窗 $\hat{p}(x)=\frac{1}{N}\sum_{x_t\in D}K_\sigma(x,x_t)$，以均值对 $V(p)$ 进行估计，可得到下式：

$$\hat{V}(p)=\frac{1}{N}\sum_{x_t\in D}\hat{p}(x_t)=\frac{1}{N}\sum_{x_t\in D}\frac{1}{N}\sum_{x_{t'}\in D}K_\sigma(x,x_{t'})=\frac{1}{N}\mathbf{1}^{\mathrm{T}}\boldsymbol{K}\mathbf{1} \tag{2}$$

其中，$\mathbf{1}$ 为 $(N\times1)$ 的向量；$\boldsymbol{K}$ 为 $(N\times N)$ 的核矩阵。

假设 $k(k<N)$维数据通过 $\boldsymbol{\Phi}$ 映射到子空间 U_k，当且仅当子空间与 Renyi 熵建立联系时，并根据熵的大小将特征值和特征向量进行重新排序，产生 KECA 的映射 $\boldsymbol{\Phi}_{eca}$：

$$\boldsymbol{\Phi}_{eca}=P_{U_k}\boldsymbol{\Phi}=\boldsymbol{D}_k^{\frac{1}{2}}\boldsymbol{E}_k^{\mathrm{T}} \tag{3}$$

其中，$\boldsymbol{D}$ 是特征值 $\lambda_1,\lambda_2,\cdots,\lambda_N$ 的对角矩阵；$\boldsymbol{E}=[\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_N]$。

转换成求解最小值问题，即：

$$\boldsymbol{\Phi}_{eca}=\boldsymbol{D}_k^{\frac{1}{2}}\boldsymbol{E}_k^{\mathrm{T}}:\min\hat{V}(p)-\hat{V}_k(p) \tag{4}$$

结合式(2)，可将式(4)改写为：

$$\min\frac{1}{N^2}\mathbf{1}^{\mathrm{T}}(\boldsymbol{K}-\boldsymbol{K}_{eca})\mathbf{1} \tag{5}$$

基金项目：国家自然科学基金资助项目(60701021)；上海市教育委员会科研创新基金资助项目(09YZ15)；上海市教委重点学科建设基金资助项目(J50104)；上海大学研究生创新基金资助项目(SHUCX112137)

作者简介：黄丽瑾(1986－)，女，硕士研究生，主研方向：信号处理；施 俊，副教授；钟 瑾，硕士研究生

收稿日期：2011-07-18 **E-mail：**junshi@staff.shu.edu.cn

其中，$\boldsymbol{K}_{eca}=\boldsymbol{\Phi}_{eca}^{\mathrm{T}}\boldsymbol{\Phi}_{eca}=\boldsymbol{E}_k\boldsymbol{D}_k\boldsymbol{E}_k^{\mathrm{T}}$。

由式(5)可知，KECA 通过在特征空间中进行类似 KPCA 的相关操作，转换为最小值的优化问题以实现数据变换。两者最大的不同在于 KECA 在求解过程中需要计算特征值与特征向量的 Renyi 熵。KECA 不需要重新建立核矩阵的中心，且它突破了仅依赖于前 k 个特征值限制的局限性，具有降低计算机复杂程度、简单有效的优点。

2.2 支持向量机原理

为检验 KECA 的降维性能，对于降维后的数据，采用 SVM 算法进行分类，以评估分类精度。SVM 是在统计学习理论和结构风险最小化原理基础上建立起来的一种机器学习方法[9-10]。它较好地解决了小样本、非线性、高维数、局部极小点等实际问题，具有很强的泛化能力。因此，本文采用 SVM 作为分类器，其具体算法见文献[9]。

2.3 降维算法

如图 1 所示，基于 KECA 的降维算法具体流程如下：

(1)输入 N 维向量数据集 $\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N]$；

(2)求数据集 X 的欧式距离，通过 Parzen 窗函数建立核矩阵 $\boldsymbol{K}_{eca}$；

(3)通过非线性映射在特征空间计算特征值和特征向量；

(4)计算特征值与特征向量的 Renyi 熵，并将特征值和特征向量根据熵值的大小进行重新排序，产生 KECA 映射 $\boldsymbol{\Phi}_{eca}$；

(5)根据降维需要，选择前 $k(k<N)$维映射，建立核矩阵与映射之间的关系，从而实现高维到低维的数据变换；

(6)在求出新的降维数据后，输入分类器进行降维性能测试。

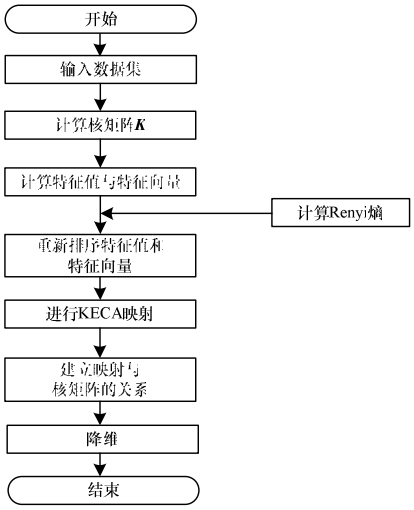

图 1 KECA 降维流程

3 实验测试

为测试 KECA 降维算法的性能，采用 UCI 机器学习数据库中的数据进行测试[11]。如表 1 所示，选取数据为原始特征维数较多的威斯康辛肿瘤数据库(WDBC Dataset)、电离层数据库(Ionosphere Dataset)和肺癌数据库(Lung Cancer Dataset)。

表 1 数据库信息描述

数据库	样本数	原始特征维数
WDBC	569	32
Ionosphere	351	35
Lung Cancer	32	57

为验证 KECA 的降维效果，本文选择了同为谱降维算法的 PCA 和 KPCA 进行对比实验。对每一个数据库的数据，分别利用这 3 种降维算法从原始特征维数或原始样本数开始逐一递减降维，并将每次降维后的数据输入 SVM 进行分类。SVM 算法采用简单易用且有效的 LIBSVM 软件包[12]。同时，利用交叉验证方法进行寻优。

4 实验结果与分析

图 2 所示为 3 种降维算法对 WDBC 数据降维后分类的结果。由图可见，从第 2 维到 30 维特征，KECA 降维以后都表现出了稳定的分类结果，即使是当维数低于 5 维(大于 1 维)时，其分类精度也在 93%左右，体现了很好的稳定性和准确性。PCA 的降维性能与 KECA 差不多，但它是在维数超过 10 维以后才在部分维数时体现出较好的分类精度，而且稳定性不如 KECA，在不同维数时分类精度的波动较大。KPCA 不仅分类精度低，在不同维数时的分类精度波动也很大。

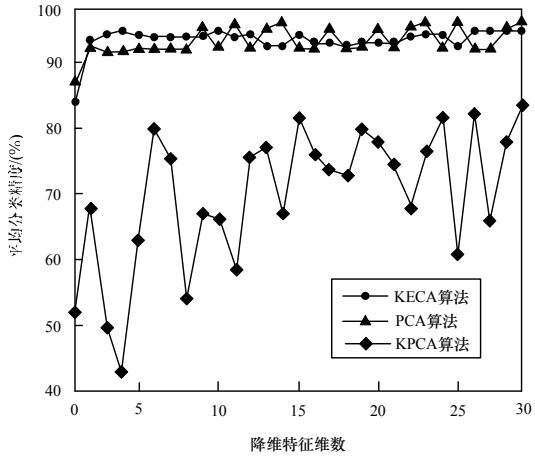

图 2 WDBC 数据库的降维算法性能比较

图 3 为 Ionosphere 数据降维后的分类结果。当维数降至 2 维~7 维之间时，KPCA 的降维性能明显优于 KECA 和 PCA；当维数在 8 维~15 维之间时，KECA 的降维性能与 KPCA 相当，均优于 PCA 降维，而且在 15 维左右时，KECA 降维算法率先达到最优性能；当特征维数大于 15 维以后，KPCA 降维后的分类精度要高于其他 2 种方法，但波动较大。

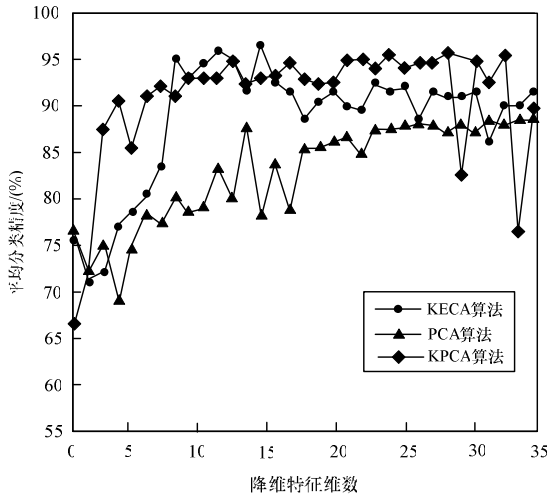

图 3 Ionosphere 数据库的降维算法性能比较

图 4 为肺癌数据降维后的分类结果。肺癌数据库具有明显的维灾现象。由图 4 可见，当特征维数降至第 9 维时，KECA 与 PCA 同时达到最优性能。由于肺癌数据需要进行 3 种分类，因此 3 种降维算法的波动性都较大。但总体来说，KECA 相对更为稳定，也能取得较好的分类精度。

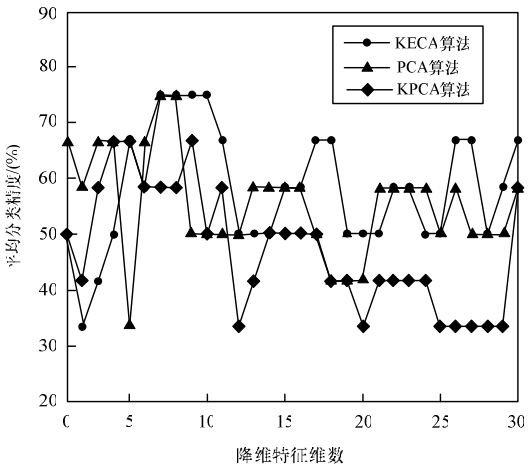

图 4 肺癌数据库的降维算法性能比较

表 2 定量比较了 3 种算法降维后的结果。对于 WDBC 数据，KECA 的最优分类精度只比 PCA 低了一个百分点，但此时对应的维数却是远低于 PCA 和 KPCA 的维数，只有 PCA 的 41.7%，KPCA 的 17.2%。对于 Ionosphere 数据，KECA 不仅获得了最高分类精度，而且其最优分类时对应的维数也远低于 PCA 和 KPCA 的维数，只有 PCA 的 23.3%，KPCA 的 74.9%。对于肺癌数据，KECA 和 PCA 的最优分类精度和对应的特征维数一样，分类精度远高于 KPCA 降维后的分类精度，而维数只比 KPCA 多了 3 维。由定量的数据比较发现，KECA 表现出优越的降维性能，它一般能在较小的特征维数时反而获得最优的分类精度。

表 2 3 种降维算法的比较

数据库	KECA 最优维数	PCA 最优维数	KPCA 最优维数	原始分类精度/(%)	KECA 精度/(%)	PCA 精度/(%)	KPCA 精度/(%)
WDBC	5	12	29	91.96	94.64	95.98	83.48
Ionosphere	17	30	23	91.00	96.50	88.50	95.50
Lung Cancer	7	7	4	58.33	75.00	75.00	66.67

从表 2 还可以发现，通过降维处理后再进行分类，分类精度一般要优于直接对原始高维数据进行分类。这是因为高维数据在包含有效特征的同时，也有着很多冗余的无效信息，这些冗余特征并不能帮助提高分类的精度，甚至反而会降低分类精度[13]。而高维特征经过降维后，获得更加紧凑和经济的数据表示方式，所包含的特征能更好地表达数据集的特点，从而有利于分类精度的提高。更为重要的是，降维有效减少了特征维数的数量，明显降低了计算复杂程度，从而可以减少后续分类算法的运行时间，提高运行的效率。

从理论上讲，真实数据是高度非线性的流形，非线性降维技术的降维能力应优于线性降维技术。但实际上，有研究表明：对于实际的高维数据，大部分的非线性降维技术的降维效果并不比线性 PCA 降维好[3]。在本文实验中，KPCA 降维效果也不是都比 PCA 好。但是，作为非线性的降维算法，KECA 却获得了媲美甚至优于 PCA 的降维效果。这是由于 KECA 中的特征向量能保证熵的减少最小，而在其他方法中，最大的特征值却不能保证熵的减少达到最少。KECA 根据熵值最大程度地判断并保留主要特征信息，从而更好地保留了输入高维数据的原始特征，在较低维数时即能呈现较好的降维结果，具有较强的非线性处理能力。

在实际问题中，经常出现会小样本问题。目前最简单的解决方法是通过降维将样本维数减至小于或等于样本个数为止。研究表明，当高斯核函数的宽度参数 σ 过小时，KPCA 降维性能受小样本问题影响较大[3]。本文实验中的肺癌数据就是一个小样本问题，KPCA 的降维性能与文献[3]研究结果相符。但同样基于核方法的 KECA 却体现了解决此类问题的优越性，显著提高了分类精度。

5 结束语

本文基于 KECA 实现了数据降维，通过与传统的谱降维方法进行比较发现，KECA 降维算法在不同的维数下具有相对稳定的分类精度，且能在较小的维数时，仍然保持较高的分类精度。当然，任何降维算法都有其局限性，关键在于满足分类精度和执行时间的前提下，选择正确的降维算法。而 KECA 降维算法的出现，为高维数据降维提供了一种新的选择。

在下一步的研究中，将扩大测试数据的范围，从而更为全面地评估 KECA 降维算法，更好地应用于机器学习、模式识别和生物信息学等领域。

参考文献

[1] Burges C J C. Dimension Reduction: A Guided Tour[J]. Foundations and Trends in Machine Learning, 2010, 2(4): 275- 365.

[2] Tsai F S. Comparative Study of Dimensionality Reduction Techniques for Data Visualization[J]. Journal of Artificial intelligence, 2010, 3(3): 119-134.

[3] Maaten L J, Postma E O, Herik H J. Dimensionality Reduction: A Comparative Review[R]. Tilburg University, Technical Report: TiCC-TR 2009-005, 2009.

[4] Xiao Lin, Sun Jun, Boyd S. A Duality View of Spectral Methods for Dimensionality Reduction[C]//Proc. of the 23rd International Conference on Machine Learning. New York, USA: ACM Press, 2006: 1041- 1048.

[5] Kruger U, Zhang Junping, Xie Lei. Developments and Applications of Nonlinear Principal Component Analysis — — A Review[C]//Proc. of Lecture Notes in Computational Science and Engineering. [S. l.]: Springer, 2008: 1-43.

[6] 郭 飞, 王 成. 基于 LMP 和 KPCA 的人脸识别[J]. 计算机工程, 2010, 36(24): 183-186.

[7] Jenssen R, Storaas O. Kernel Entropy Component Analysis Pre-images for Pattern Denoising[C]//Proc. of the 16th Scandinavian Conference on Image Analysis. Berlin, Germany: Springer-Verlag, 2009: 626-635.

[8] Jenssen R. Kernel Entropy Component Analysis[J]. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2010, 32(5): 847-860.

[9] Cortes C, Vapink V. Support Vector Network[J]. Machine Learning, 1995, 20(3): 273-297.

[10] Burges C J C. A Tutorial on Support Vector Machines for Pattern Recognition[J]. Data Mining and Knowledge Discovery, 1998, 2(2): 1-43.

[11] University of California Irvine. UCI Machine Learning Repository[EB/OL]. (2010-09-13). http://archive.ics.uci.edu/ml.

[12] Chang Chih-Chung, Lin Chih-Jen. LIBSVM——A Library for Support Vector Machines[EB/OL]. (2010-03-01). http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[13] Kotsiantis S B, Zaharakis I D, Pintelas P E. Machine Learning: A Review of Classification and Combining Techniques[J]. Artificial Intelligence Review, 2006, 26(3): 159-190.

编辑 顾逸斐