第37卷 第14期 计 算 机 工 程 2011年7月
Vol.37 No.14 Computer Engineering July 2011

·工程应用技术与实现· 文章编号：1000—3428(2011)14—0239—03 文献标识码：A 中图分类号：TP301.6

# 基于相似度的核主元分析方法及其应用研究

张传标，倪建军，苗红霞，韩光洁

(河海大学计算机与信息学院，江苏 常州 213022)

**摘 要：**常规核主元分析(KPCA)方法在对大样本数据分析建模时，存在运算复杂度高、建模时间长以及所需存储空间大等缺点。为此，提出一种基于相似度函数的快速核主元分析(SF-KPCA)方法。建立大样本数据间的相似度函数矩阵，分析数据样本间的相似程度，剔除冗余数据，再利用优化数据样本建立核主元分析模型，对数据样本进行分析。将SF-KPCA方法应用于高压断路器故障诊断中，实验结果证明了该方法的快速性和有效性。

**关键词：**大样本数据；相似度函数；快速核主元分析；高压断路器；故障诊断

# Study on Kernel Principal Component Analysis Method Based on Similarity and Its Application

**ZHANG Chuan-biao, NI Jian-jun, MIAO Hong-xia, HAN Guang-jie**

(College of Computer and Information, Hohai University, Changzhou 213022, China)

**【Abstract】** In order to overcome the shortcomings of conventional Kernel Principal Component Analysis(KPCA) method in modeling and analyzing of large sample data(e.g., high computational complexity, long time modeling and large storage space etc.), a fast KPCA method based on Similarity Function(SF-KPCA) is proposed. The similarity function matrix of a large data samples is established to analyze the similarity between data samples, and the redundant data is eliminated. The KPCA model using the optimized data samples is built. The data samples are analyzed. The method is applied to the fault diagnosis of high voltage circuit breaker. Simulation results show the proposed method's rapidity and effectivity.

**【Key words】** large sample data; similarity function; fast Kernel Principal Component Analysis(KPCA); high voltage circuit breaker; fault diagnosis

**DOI:** 10.3969/j.issn.1000-3428.2011.14.081

## 1 概述

主元分析(Principal Component Analysis, PCA)方法是一种多元统计分析方法[1]，可以将高维、含噪声以及高度相关的数据投影到包含大量原始数据信息的低维子空间内再进行分析，这种降维特性使其得到了广泛的应用[2-5]。基于PCA方法只能解决线性问题，并不适合分析非线性问题。为了提高PCA方法的适用性，文献[6]提出了核主元分析(Kernel PCA, KPCA)方法，通过“核技巧”将输入空间映射到高维特征空间，从而将原输入空间的非线性问题转换为特征空间的线性问题。KPCA方法展现出解决非线性问题方面的优势，使其也得到了广泛的应用，文献[7]将KPCA方法应用于非线性故障诊断中，利用KPCA进行数据特征提取，再进行数据分析与诊断，该理论的应用领域还有模式识别[8]、图像处理[9]、信号处理[10]等。常规核主元分析方法虽然解决了多变量和非线性问题，却不适合用于大样本数据分析，因为在大样本核矩阵运算中，需要大量运算，并占据大量存储空间，为此，文献[11]利用数据挖掘理论，通过利用特征选择技术来减少特征数量，从而得到最优特征子空间，再对数据进行核主元分析，提取数据的非线性特征；文献[12]提出一种基于特征子空间的KPCA方法(FS-KPCA)，通过在特征空间上构建具有较小维数的正交基来简化核矩阵，从而降低运算和分析复杂度。在文献[12]中，每个初始数据样本均参与核矩阵运算，并需要不断求解核矩阵的行列式值，在一定程度上增加了运算复杂度。该文献并没有分析重要参数合理选取原则，即在保证获得满意故障诊断准确度前提下，如何合理优化数据样本。

本文提出一种基于相似度函数的快速核主元分析(SF-KPCA)方法，利用相似度函数的方法，分析2组数据样本相似程度，合理约减相似度高的数据样本，优化建模数据样本，再进行KPCA建模与数据分析。

## 2 SF-KPCA原理

核主元分析方法是一种非线性主元分析方法，通过非线性映射将输入空间映射到特征空间，在特征空间上进行主元分析，具体相关知识参阅文献[6-7]。当对大样本数据进行核主元分析时存在运算量大、实时性差等问题，因此，必须有效减少输入样本的冗余信息，优化建模数据样本。

本文利用相似度函数的方法优化输入样本，依次计算任何2组输入数据的相似度，得到相似度矩阵 $\boldsymbol{R}$，然后对矩阵 $\boldsymbol{R}$ 进行分析。对于相似度高的2组数据，可认为近似相同，由此来约减其中一组数据，减少数据冗余，本文选取相似度函数为：

$$R_{ij}=\exp\left(-\frac{1}{\delta}\left\|x_i-x_j\right\|^2\right)$$

其中，$x_i\in\Re^m, i=1,2,\mathrm{L}\ ,n$ 为数据样本；$\|\cdot\|$ 表示矩阵的2范数；

**基金项目：**河海大学常州校区创新基金资助项目(XZX/09B002-02)；常州市输配电及节电技术重点实验室开放课题基金资助项目(CS0904)

**作者简介：**张传标(1983－)，男，硕士研究生，主研方向：智能算法，故障诊断；倪建军，副教授、博士；苗红霞，讲师、博士研究生；韩光洁，副教授、博士

**收稿日期：**2011-01-27 **E-mail：**zhangcbhhuc@163.com

$m$ 为变量个数；$n$ 为样本个数；$\delta$ 为相似度函数归一化参数；$R_{ij}$ 表示第 $i$ 组数据样本与第 $j$ 组数据样本的相似程度。

数据样本优化的原则如下：

$$\begin{cases} R_{ij} < \varepsilon & \text{保留第}i,\ j\text{组数据} \\ R_{ij} \geqslant \varepsilon & \text{保留第}i\text{组数据，剔除第}j\text{组数据} \end{cases}$$

样本约简运算结束后，得到优化数据样本集，将该数据样本集作为核主元分析建模的输入样本，建立核主元分析模型。具体算法流程如图 1 所示。

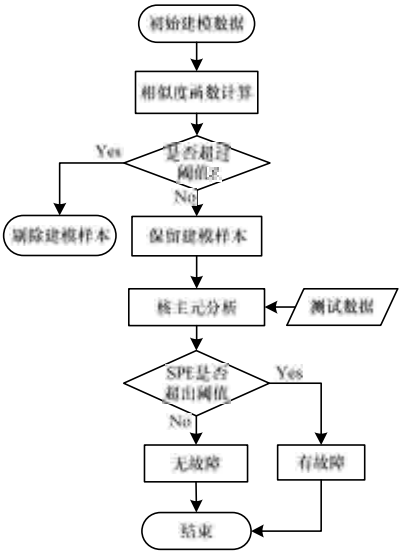


**图 1 本文算法流程**

## 3 仿真与分析

为了验证 SF-KPCA 方法的有效性，选取 LW6B-252 高压断路器作为研究对象，通过采集并分析电气回路数据信息，实现断路器故障诊断。通过对其故障以及故障原因研究与分析，在正常工况下选取大量初始数据样本，利用相似度函数计算数据样本间相似度函数矩阵，选取阈值 $\varepsilon$ 得到合理的优化建模数据样本，利用优化后的数据样本建立 KPCA 模型，提取数据特征，然后使用测试数据样本对所建模型进行测试，分析模型的故障诊断准确度。

### 3.1 仿真实验

本文根据断路器正常工况下的运行参数，选取初始数据样本 1 000 组，每组数据包含 6 个运行参数。如果使用该数据样本直接建立 KPCA 模型，计算后的核矩阵维数为 1 000×1 000，并且计算复杂度为 $O(1\ 000^3)$，严重影响了特征提取和数据分析的时效性，因此，必须对该数据样本进行合理的剔除冗余处理，优化数据样本，既节省存储空间也提高运算效率。

根据相似度函数约减数据样本的步骤对该 1 000 组数据进行样本冗余约减处理，以优化后数据样本作为核主元分析的输入样本建立核主元分析模型，选取 500 个测试数据样本进行模型的测试，并利用平方预测误差(SPE)来分析有无故障存在。

### 3.2 参数选取

针对正常工况下的信号数据输出情况，并尽量减少分析与诊断误差，选取相似度函数归一化参数 $\delta$ 计算公式为：

$$\delta = \sum_{i=1}^{6} \max D_i - \sum_{i=1}^{6} \min D_i$$

其中，$D_i$ 为第 $i$ 个参数的数据值，根据上式计算得到 $\delta$ 的合适数值为 60。

核函数选取高斯函数为：

$$K(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$$

其中，本文选取 $\sigma = 10$。

下面详细分析不同阈值 $\varepsilon$ 对诊断性能的影响。

(1)阈值 $\varepsilon$ 对样本约减效率的影响

通过实验，针对阈值 $\varepsilon$ 对数据样本约减的效率，对应剩余数据样本量如图 2 所示。

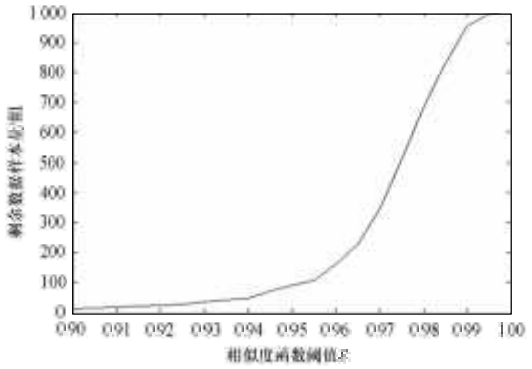


**图 2 阈值与剩余样本量关系曲线**

由图 2 可看出，随着阈值的不断增加，约减的数据样本越来越少，剩余样本量越来越多。阈值在区间[0.97,0.99]取值时，曲线斜率最大，说明约减的数据样本效率最高，也即在此范围内存在的冗余信息最多，因此，合适的阈值应该在此区间选取。

(2)阈值 $\varepsilon$ 对后续故障诊断准确性的影响

通过大量实验，分析不同阈值 $\varepsilon$ 对基于 KPCA 模型下的故障诊断有效性的影响，选取 SPE 阈值为 0.006，即超过该值判定有故障发生，否则判定无故障发生。

对选取的 500 组测试数据样本进行模型测试，实验测试结果如表 1 所示。

**表 1 不同阈值下的误诊断率**

| 阈值$\varepsilon$ | 误诊断数/组 | 误差率/(%) |
|---|---|---|
| 0.96 | 21 | 4.2 |
| 0.97 | 11 | 2.2 |
| 0.98 | 3 | 0.6 |
| 0.99 | 2 | 0.4 |

从表 1 可以看出，阈值 $\varepsilon$ 变化会影响诊断的准确性，当 $\varepsilon$ 阈值低于一定的值，如 $\varepsilon = 0.96$ 时，数据样本约减过多，剩余数据样本无法反映整体的数据特征，因此利用该样本建立 KPCA 模型进行故障诊断，误诊断偏高，无法高效、准确地检测出故障并进行分析。

不同阈值 $\varepsilon$ 和故障误诊断数的拟合曲线如图 3 所示。

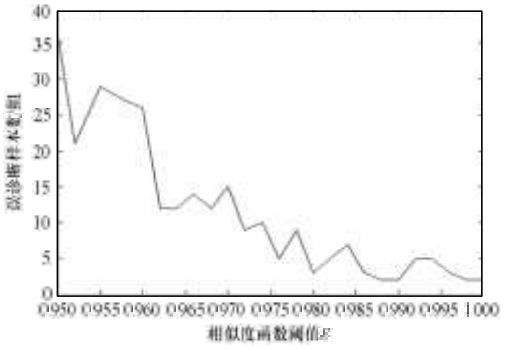


**图 3 不同阈值下 KPCA 模型误诊断拟合图**

在图 3 中，曲线反映随着阈值 $\varepsilon$ 的不断增大，故障误诊断数目总体上是下降趋势，曲线并不是圆滑的下降曲线是因为初始数据样本是在正常数据范围内随机生成的，因此，在误诊断数目上会有细微的波动，但总体趋势是下降的。

建模数据样本由于阈值的变化而变化，由图 2 中可以看出，阈值越大，约减的样本越少，冗余信息越大。由于建模数据样本数的不同，使得建模和分析时间也不同，具体阈值与总耗时(样本约减、KPCA 建模和故障诊断时间之和)的关系曲线如图 4 所示。

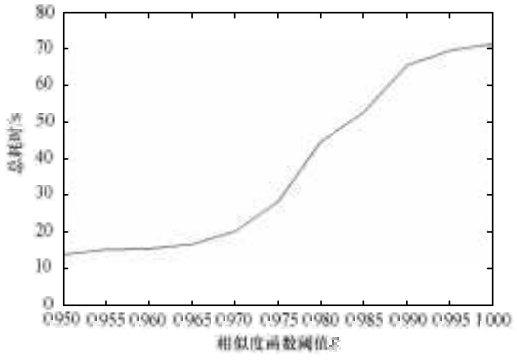


**图 4　阈值与总耗时关系曲线**

在图 4 中，阈值 $\varepsilon$ 的值在[0.97,0.99]之间时，曲线的斜率最大，反映建模时间变化最明显，这也可从图 2 中推理出：数据样本越少，建模时间越短。

由于数据样本的不断减少，使得在进行核主元建模时核矩阵运算量不断减少，从而不断缩短建模时间，但模型的误诊断率也会随之不断增加。综合考虑建模时间和诊断准确度，并结合表 1、图 2、图 4，选取相似度函数阈值 $\varepsilon$ 的取值范围为[0.975,0.985]，在该区域内选值，建模时间相对来说比较短，诊断正确率比较高，满足实际的应用要求。

如果直接利用初始 1 000 组数据样本进行建模分析，则总耗时为 61 s，而如果使用优化数据样本的方法，在[0.975, 0.985]的范围内选择阈值 $\varepsilon$，总耗时为[24 s, 50 s]。根据文献[12]提出的 FS-KPCA 方法，本文使用相同样本数据仿真与分析，并与 SF-KPCA 方法进行比较分析，得到剩余样本与耗时的关系如图 5 所示。

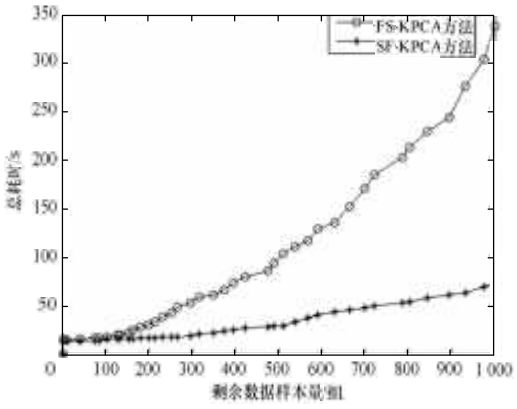


**图 5　2 种方法分析耗时比较曲线**

由图 5 中 2 种方法比较曲线可以看出，SF-KPCA 方法在总耗时上明显低于 FS-KPCA 方法，因为 FS-KPCA 方法在进行降秩处理时需要进行核矩阵行列式运算，剩余数据样本越大，运算复杂度越高，总消耗时间越长，分析实时性越差，而 SF-KPCA 方法仅需要计算相似度函数矩阵，计算复杂度低，实时性较好。而且随着样本的不断增加，总耗时的优势越明显。

由分析看出，SF-KPCA 方法在数据分析与诊断耗时上明显少于未优化样本和 FS-KPCA 方法，验证了本文方法的快速性和有效性。

## 4　结束语

本文针对 KPCA 模型不适合大样本数据分析的缺点，提出了基于相似度函数的 KPCA 方法，利用相似度函数对数据样本进行合理约减，减少冗余数据信息，并详细分析了参数阈值 $\varepsilon$ 选取对数据分析准确性和时效性的影响。最后将该方法引入到高压断路器故障诊断中，仿真结果显示，在大样本数据下，采取 SF-KPCA 方法，通过分析并确定合适的阈值选取范围，可以在保证数据分析准确性的前提下，大大缩短数据分析时间，并将本方法与 FS-KPCA 方法进行比较分析，总耗时明显优于 FS-KPCA 方法。

## 参考文献

[1] Wise B M, Gallagher N B. The Process Chemometrics Approach to Process Monitoring and Fault Detection[J]. Journal of Process Control, 1996, 6(6): 329-348.

[2] 王志征, 余岳峰, 姚国平. 基于主成分分析法和自适应神经模糊推理系统的电力负荷预测[J]. 电力自动化设备, 2003, 23(9): 39-41.

[3] Kumar D, Kumar S, Rai C S. Feature Selection for Face Recognition: A Memetic Algorithmic Approach[J]. Journal of Zhejiang University: Science A, 2009, 10(8): 1140-1152.

[4] 李冬辉, 王乐英, 李　晟. 基于 PCA 的空调系统传感器故障诊断[J]. 电工技术学报, 2008, 23(6): 130-136.

[5] 汤红忠, 肖业伟, 黄辉先, 等. 基于 PCA 矢量形态学的彩色图像分割方法[J]. 计算机工程, 2009, 35(12): 201-203.

[6] Scholkopf B, Smola A J, MJuller K. Nonlinear Component Analysis As a Kernel Eigenvalue Problem[J]. Neural Computation, 1998, 10(5): 1299-1399.

[7] Choi S W, Lee C, Lee J M. Fault Detection and Identification of Nonlinear Processes Based on Kernel PCA[J]. Chemometrics and Intelligent Laboratory Systems, 2005, 75(1): 55-67.

[8] Kim K I, Jung K, Kim H J. Face Recognition Using Kernel Principal Component Analysis[J]. IEEE Signal Processing Letters, 2002, 9(2): 40-42.

[9] Santhanam A, Rahman M M. Kernel PCA in Detecting Moving Vehicle from Its Viewpoint[C]//Proc. of International Conference on Computing: Theory and Applications. Kolkata, India: [s. n.], 2007: 665-670.

[10] Teixeira A R, Tomé A M, Lang E W. Greedy KPCA in Biomedical Signal Processing[C]//Proc. of the 17th International Conference on Artificial Neural Networks. Porto, Portugal: [s. n.], 2007: 486-495.

[11] Cho H W. A Data Mining-based Subset Selection for Enhanced Discrimination Using Iterative Elimination of Redundancy[J]. Expert Systems with Applications, 2009, 36(2): 1355-1361.

[12] 付克昌, 吴铁军. 基于特征子空间的 KPCA 及其在故障检测与诊断中的应用[J]. 化工学报, 2006, 57(11): 2665-2669.

编辑　任吉慧