基于核函数的 PCA 在 QAR 数据分析中的应用

冯兴杰,冯小荣,王艳华
FENG Xing-jie,FENG Xiao-rong,WANG Yan-hua

中国民航大学 计算机科学与技术学院,天津 300300
School of Computer Science & Technology,Civil Aviation University of China,Tianjin 300300,China

FENG Xing-jie,FENG Xiao-rong,WANG Yan-hua.Application of PCA based on kernel function in analysis of QAR data.Computer Engineering and Applications,2009,45(14):207-209.

Abstract: This paper analyzes the drawbacks of general Principal Component Analysis(PCA) firstly,and discusses the Kernel Principal Component Analysis(KPCA) and its drawbacks of high time complexity secondly.Then proposes the kernel function covariance matrix of principal component analysis in the end.Compared to KPCA,the method is fast descending dimension speed. The results show that the proposed method used for QAR data has a good effect of dimension reduction and high rate of correct classification.

Key words: Principal Component Analysis(PCA);kernel function;Kernel Principal Component Analysis(KPCA);covariance matrix

摘 要:分析了传统的主成分分析方法的不足,论述了 KPCA 方法及其时间复杂度高的缺陷。在此基础上,提出基于核函数构造的协方差矩阵的主成分分析,相比 KPCA,该方法具有快的降维速度。实验结果显示:把该方法用于 QAR 数据具有良好的降维效果和高分类正确率。

关键词:主成分分析;核函数;核主成分分析;协方差矩阵
DOI:10.3778/j.issn.1002-8331.2009.14.064 **文章编号**:1002-8331(2009)14-0207-03 **文献标识码**:A **中图分类号**:TP391

1 引言

QAR(Quick Access Recorder)是飞机记载记录系统中的快速存储装置。目前世界上一些发达国家将飞行数据运用于日常监控工作,美国的飞行事故率,从 1950 年的万时率 3.6 降到目前的 0.15。从 1998 年底开始,应民航总局的要求,中国大多数航空公司的飞机上加装了 QAR 系统,QAR 数据成为各个航空公司飞行品质监控、发动机状态检测、诊断飞机系统故障、实现三维动画分析等方面的重要依据[1-2]。但是 QAR 数据维数庞大,并且是一种基于时间序列的多变量的数据。因此在进行数据处理之前,必须进行必要的降维处理。

主成分分析[3-4](Principal Component Analysis,PCA)是一种常用的降维方法, 它通过把数据映射到特征空间的方法,将原始属性线形组合,并根据数据展示在每个特征方向的方差大小排序,再约简掉那些方差很小的特征方向,从而达到降维的目的。传统的 PCA 是一种线性映射, 因此在处理非线性问题时,一般无法取得好的效果。同时它对孤立点或缺失值非常敏感,而孤立点和缺失值会带来残缺或错误的分析结果。为了改进这些缺点专家学者提出了许多新方法,如基于模糊集理论的主成分分析方法[5],由 Hastie 和 Stuetzle 提出的主曲线和主曲面方法[6]。

近年来,随着支持向量机[7](Support Vector Machines,SVM)的研究深入,关于核方法的研究受到重视。由 Scholkopf 提出的基于核函数的主成分分析方法[8-9](KPCA),它通过映射输入变量数据到高维特征空间进行 PCA。KPCA 在计算核矩阵时的时间复杂度和样本的维数以及个数密切相关,为 $O(pN^2)$,其中 p 为样本的维数,N 为样本个数。为了提高时间效率,有关学者提出基于聚类的 KPCA 方法[10],该方法首先通过聚类分析,以聚类中心为新的样本,这样样本数量就大大减少,再施行 KPCA,从而降低时间复杂度。但是该方法也带来了如下缺陷:聚类的个数和新样本都对 KPCA 有一定的影响。针对上述问题,提出一种改进的 PCA,首先通过核函数构造协方差矩阵,然后再进行主成分分析,其时间复杂度为 $O(p^2N)$。实验结果表明这种方法降维效果明显高于 PCA,而不需要预先指定相关参数。在样本数量庞大的情况下,采用适当增加主成分的个数,可以做到和 KPCA 一样的分类精确度,而时间却明显的少于 KPCA。

2 PCA 方法和 KPCA 方法

2.1 PCA 方法

PCA 方法是对描述观测数据的坐标系统的一个正交变换,

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.60672174,No.60776806);中国民航大学基金项目(No.06kym01)。

作者简介:冯兴杰(1969-),男,博士,教授,研究方向:数据库及数据仓库、智能信息处理理论与技术;冯小荣(1980-),男,硕士研究生,研究方向:数据仓库与数据挖掘。

收稿日期:2008-03-17 **修回日期**:2008-06-10

它旨在用原始变量的线性组合获得较少的互不相关的新变量，同时尽可能多地保持原变量的信息。

通过可数变量的几个线性组合来概括大部分原始信息。用 k 个变量的 n 次观测数据代替 $p(k<p)$ 个原变量的几次观测数据，而基本的信息量保持不变。

设有样本容量为 n 的 p 个变量，通过变换将原变量 X_i 转换成主成分（用 F 表示），主成分是原变量的线性组合，且具有正交特征，即将 $X_1,X_2,\cdots,X_p$ 综合成 $k(k<p)$ 个变量 $(F_1,\cdots,F_k)$，可用多项式表示：

$$\begin{cases}F_1=a_{11}X_1+a_{21}X_2+\cdots+a_{p1}X_p\\F_2=a_{12}X_1+a_{22}X_2+\cdots+a_{p2}X_p\\\cdots\\F_k=a_{1k}X_1+a_{2k}X_2+\cdots+a_{pk}X_p\end{cases}$$

这样确定的综合变量 $F_1,F_2,\cdots,F_k$ 分别称作原变量的第1、第 2、…、第 k 个主成分，且 $F_1,F_2,\cdots,F_k$ 在总方差中占的比例依次递减。

2.2 KPCA 方法

PCA 方法忽略了具有较小方差的线性组成部分，保留具有较大方差的项，从而有效减小了数据表示的维数。但它是一种线性变换，只能提取数据中的线性相关特性，对于非线性的问题很难达到预期的降维效果。为此由 Scholkopf 提出的基于核函数的 KPCA 方法。

KPCA 是首先通过一个非线性函数把观测数据映射到一个更高维的特征空间，然后在特征空间中施行 PCA。实际上 KPCA 是通过使用核函数，隐式地定义特征空间，并且特征空间的点积巧妙利用核函数转换到输入空间计算，通过定义核矩阵，再进行主成分分析。

定义 1 核是一个函数 K，对所有的 $x,z\in X$ 满足：$K(x,z)=<\phi(x)\cdot\phi(z)>$，这里 ϕ 是从 X 到特征空间 F 的映射。

定义 2 设 p 维观测数据 X，被非线性的映射到特征空间 F，$K_{ij}=K(\phi(\boldsymbol{x}_i),\phi(\boldsymbol{x}_j))(1\leqslant i,j\leqslant n)$。则称矩阵 $\boldsymbol{K}$ 为核矩阵。其中 $\phi(\boldsymbol{x}_i)$ 为映射后的数据并中心化，$\boldsymbol{x}_i,\boldsymbol{x}_j$ 为数据 X 的第 i,j 个样本，K 为核函数。

有关著作和论文已经给出核函数 K 要满足的 Mercer 条件[7-11]，即通过核函数得到核矩阵 $\boldsymbol{K}$ 必须半正定。

得到核矩阵 $\boldsymbol{K}$ 以后，KPCA 就转化为求解特征值方程 $K\boldsymbol{\alpha}=a\boldsymbol{\alpha}$ 的问题。设 $\lambda_1,\lambda_2,\cdots,\lambda_n$ 为 K 的特征值且依次递减，$\boldsymbol{\alpha}$ 为 λ 所对应的特征向量，为了满足 $\lambda\boldsymbol{\alpha}\cdot\boldsymbol{\alpha}=1$，取 $\|\boldsymbol{\alpha}\|=1/\sqrt{\lambda}$。

则任一观测向量 t 在特征空间 F 主轴的投影为 $\sum_{m=1}^{p}\alpha^m\sum_{i=1}^{n}K(\boldsymbol{x}_i,\boldsymbol{t})$，$K$ 为所选的核函数。

3 基于核函数的协方差矩阵的 PCA 方法（K-PCA）

3.1 K-PCA 方法

通过第 2 章的论述，知道 PCA 是在已知数据样本结构的条件下，基于协方差矩阵进行主成分分析，协方差矩阵的维数等于数据的维数；而 KPCA 是通过非线性的映射后，在不知道数据分布的条件下，基于核函数进行主成分分析，核函数的维数等于数据样本的个数。

KPCA 引用核函数定义核矩阵，计算核矩阵的时间为 $O(n^2)$，当样本容量 n 很大时，KPCA 在时间上就会有困难。实际上，核方法是一种技巧，它可以应用到任何算法中，只要此算法可以用点积表示[12-13]。

在 2.1 节中论述到 PCA 的协方差阵的 $\boldsymbol{S}$，其元素 S_{ij} 恰好为第 i 维数据和第 j 维数据的点积。$S_{ij}=Y_i\cdot Y_j$ $(1\leqslant i,j\leqslant p)Y_i,Y_j$ 是第 i,j 维的数据。

为此，针对 KPCA 在计算核矩阵时的高时间代价，定义基于核函数的协方差矩阵。

定义 3 设 p 维观测数据 $\boldsymbol{X}$

$$\boldsymbol{X}=\begin{pmatrix}x_{11}&\cdots&x_{1p}\\\vdots&&\vdots\\x_{n1}&\cdots&x_{np}\end{pmatrix}=[X_1,\cdots,X_p]$$

称矩阵 $\boldsymbol{S}$ 为基于核函数的协方差矩阵，其中

$S_{ij}=K(\boldsymbol{x}_i,\boldsymbol{x}_j)(1\leqslant i,j\leqslant p)$

$\boldsymbol{x}_i,\boldsymbol{x}_j$ 为数据 X 的第 i,j 维数据并已经中心化，即 $\sum_{i=1}^{n}\boldsymbol{x}_i=0$，$K$ 为所选的核函数。

运用核函数构造协方差矩阵，其大小为 $p\times p$，和数据的记录数没有关系了，不再是 $n\times n$ 的。计算的时间复杂度为 $O(np^2)$，称改进后的方法为 K-PCA。

K-PCA 的主要步骤：

（1）标准化数据样本

$y_{ij}=x_{ij}-\bar{x}_j$，其中 $\bar{x}_j=\frac{1}{n}\sum_{i=1}^{n}x_{ij}$，为每一列指标的平均值。针对变换的结果，得到新矩阵 $\boldsymbol{Y}_{n\times p}=(y_{ij})_{n\times p}$；将矩阵 $\boldsymbol{Y}_{n\times p}=(y_{ij})_{n\times p}$ 中的数据变换成 $y_{ij}^{*}=y_{ij}/s_j$。其中 $s_j=\sqrt{\frac{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}{n-1}}$，为每一列指标的均方差。经过中心化和标准化处理后，得到新的矩阵 $\boldsymbol{Y}_{n\times p}^{*}=(y_{ij}^{*})_{n\times p}$；

（2）通过核函数构造协方差矩阵 $\boldsymbol{K}$；

（3）计算矩阵 $\boldsymbol{K}$ 的特征值 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_p\geqslant0$ 及其对应的单位正交化特征向量 $\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\cdots,\boldsymbol{\alpha}_p$，并且使得 $\lambda_i\boldsymbol{\alpha}_i\cdot\boldsymbol{\alpha}'_i=1$；

（4）建立主成分 F_i 的线性表达式

$$F_i=\sum_{j=1}^{p}\alpha_i^jY_p$$

α_i^j 为向量 $\boldsymbol{\alpha}_i$ 的第 j 个元素。

3.2 常用的核函数

（1）h 次多项式函数

$k(\boldsymbol{x}_i,\boldsymbol{x}_j)=(\boldsymbol{x}_i\cdot\boldsymbol{x}_j+c)^h$，$(h\geqslant2,c\geqslant0)$

（2）Gauss 径向基核

$$k(\boldsymbol{x}_i,\boldsymbol{x}_j)=\exp\left(-\frac{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2}{2\delta^2}\right)$$

（3）双曲正切函数

$k(\boldsymbol{x}_i,\boldsymbol{x}_j)=\tanh(-a(\boldsymbol{x}_i\cdot\boldsymbol{x}_j)-b)$

4 实验数据、方法及结果

4.1 实验数据和方法

实验数据采用某航空公司两组 QAR 数据，第一组数据为同一航线两种不同机型的飞行数据各一次，机型 A 的数据容量为 18 720×120，机型 B 的数据容量为 20 400×120；第二组数据为同一航线和机型，不同飞行人员的飞行数据各一次，飞行员甲（飞行时间 10 000 小时以上）的数据容量为 29 760×120，飞行员乙（飞行时间 5 000 小时以上）的数据容量为 28 080×120。采

用 PCA、KPCA 和 K-PCA 进行降维处理,在计算核矩阵和基于核函数的协方差矩阵时采用次多项式函数,取 c=2,h=2。并用 SVM 方法进行分类,分别采用 2、4、6、8、10 个特征主成分进行分类,把数据等分成 10 份,用其中的 9 份进行训练,1 份进行测试,然后重新选择其他 9 份训练数据,再用剩下的 1 份进行测试,反复进行 10 次,取分类正确率的平均值。

4.2 实验结果

(1)图 1 是分别采用 KPCA 和 K-PCA 计算核矩阵和基于核函数的协方差矩阵所用的时间, 横轴表示选用样本的个数,纵轴表示所用的时间。从图中可以看出,样本个数对 KPCA 的影响非常大,而对 K-PCA 的几乎没有影响,随着样本个数增加,KPCA 的费时越来越多。

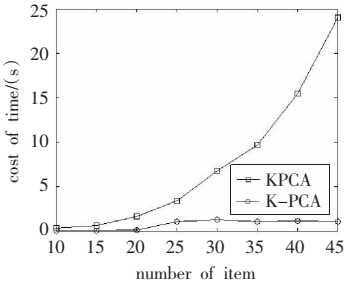

图 1　计算核矩阵的时间比较

(2)图 2 是分别采用 PCA,KPCA 与 K-PCA 对第一组数据进行降维,从图中可以观察到,KPCA 和 K-PCA 比 PCA 在降维效果要好很多。

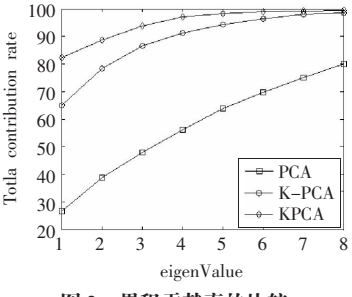

图 2　累积贡献率的比较

(3)图 3 和图 4 是分别对两组数据采用 SVM 方法的分类正确率。改进后的 K-PCA 和 KPCA,比 PCA 的分类正确率高很多。当选择 6 个特征主成分,K-PCA 和 KPCA 的正确率相当,当选择 8 个以上时,两者的正确率几乎一样;而 K-PCA 有较小的时间代价。从图中还可以反映出:图 3 比图 4 的正确率高,说明第一组数据两种机型之间的差距很大,不同机型之间的数据非常不一致,而图 4 说明飞行员具有 5 000 小时以上飞行经验时,对飞机的驾驶影响不是很大。

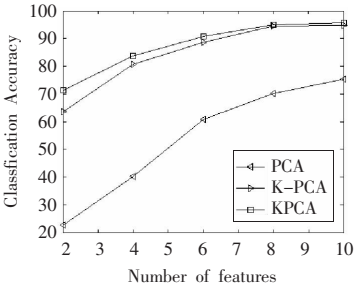

图 3　第一组数据分类正确率

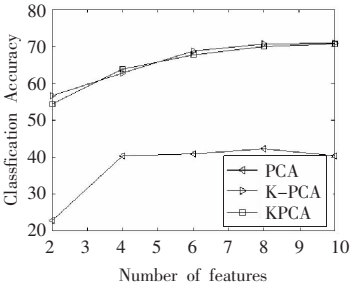

图 4　第二组数据分类正确率

5　结论

通过对协方差矩阵的改进,提出一种基于核函数构造的协方差矩阵的主成分分析方法。从实验结果分析,新方法在降维效果和分类正确率上和 KPCA 相当,有效克服了 KPCA 在计算核矩阵时的高的时间代价。今后将结合 QAR 数据把改进的方法应用在飞行品质、燃油节省和航班延误等方面,并同其他方法,例如 GPCA[14]和 K-LDA[15]方法进行比较。

参考文献:

[1] 黄永芳,黄圣国,孙同江.QAR 数据译码的航班划分[J].交通运输工程学报,2004,4(1):114-117.

[2] 卿立勇,黄圣国,林钰森.基于 QAR 数据的飞机系统故障预测与故障诊断支持系统研究[J].江苏航空,2006(2):11-12.

[3] TJolliffe I.Principal component analysis[M].[S.l.]:Springer,2002.

[4] PoP H F.Principal components analysis based on a fuzzy set approach[EB/OL].(2001).http://citeseer.ist.psu.edu/539714.html.

[5] 林和平,杨晨.模糊主成分分析方法的研究与分析[J].航空计算科学,2006,6:17-20.

[6] Hastie T.Principal curves and surfaces.Laboratory for computational statistics[R].Stanford University Depts of Statistics,1984.

[7] Gunn S R.Support vector machines for classification and regression[R].Faculty of Engineering,Science and Mathematics School of Electronics and Computer Science,1998.

[8] Scholkopf B,Smola A J,Muller K R.Nonlinear component analysis as a kernel eigenvalue problem[J].Neural Computation,1998,10(5):1299-1319.

[9] Muller K R,MiKa S,Ratsch G,et al.An introduction to kernel-based learning algorithms[J].IEEE Trans Pattern Anal Machine Intell,2001,12(2):181-201.

[10] 王和勇,姚正安,李磊.基于聚类的核主成分分析在特征提取中的应用[J].计算机科学,2005:64-66.

[11] 焦李成,刘莉,陈莉,等.智能数据挖掘与知识发现[M].西安:西安电子科技大学出版社,2006.

[12] Park C H,Park H.Nonliner feature extraction based on cancroids and kernel functions[J].Pattern Recognition,2004,37:801-810.

[13] Kim S W,Onmmen B J.On utilizing search methods to select subspace dimensions for kernel-based nonlinear subspaceclassifiers[J].IEEE Trans Patt Anal Mach Intell,2005,27:136-141.

[14] Ye J,Jana Dan R,Li Q.Gpca:An efficient dimension reduction scheme for image compression and retrieval[C]//KDD'04:Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,New York,NY,USA.[S.l.]:ACM Press,2004:354-363.

[15] Yoon H,Yang K,Shahabi C.Feature subset selection and feature ranking for multivariate time series[J].IEEE Trans Knowledge Data Eng-Special Issue on Intelligent Data Preparation,2005,17(9).