文章编号：1005-8451（2016）06-0055-04

主成分分析与奇异值分解技术在铁路数据预处理中的应用

徐贵红，郭剑峰，杨涛存，东春昭
（中国铁道科学研究院　铁路大数据研究与应用创新中心，北京　100081）

摘　要：数据预处理是在数据建模之前对采集到的原始数据进行的一些前期处理工作，能够滤除原始数据存在的噪声干扰、降低数据维度进而提取数据的时域特征。铁路运输行业在生产过程中累积的大量数据往往包含着噪声干扰，并且经常是海量高维的，无法直接用于数据建模、分析和挖掘。主成分分析与奇异值分解作为线性代数中一种重要的矩阵分解技术，已经成为近年来常用的数据时域预处理方法，本文主要论述主成分分析与奇异值分解技术在铁路数据预处理中的应用。

关键词：主成分分析；奇异值分解；数据预处理
中图分类号：U29-39　　**文献标识码：**A

Principal component analysis and singular value decomposition technologies in railway data preprocessing

XU Guihong, GUO Jianfeng, YANG Taocun, DONG Chunzhao
(Research and Application Innovation Center for Big Data Technology in Railway, China Academy of Railway Sciences, Beijing　100081, China)

Abstract: Data preprocessing is a preliminary work before data modeling. It can filter the noise interference of original data and can reduce data dimension to extract features of the data in time domain. The large amount of data accumulated in the production process of railway transport industry often contains noise interference. Besides, it is often massive and high dimensional, which cannot be directly used for data modeling, analysis and mining. Principal component analysis and singular value decomposition are important matrix decomposition technologies in linear algebra. They have been commonly used in data time-domain preprocessing in recent years. This paper mainly discussed the use of principal component analysis and singular value decomposition technologies in railway data preprocessing of the application.

Key words: principal component analysis; singular value decomposition; data preprocessing

铁路运输行业在生产过程中积累了大量的数据，包括检测车采集的各种动态检测数据；工务、电务、机务、车辆、供电专业的静态检查数据；安监问题数据；计划统计数据等[1~2]。由于在获得这些数据时使用了大量的传感器等检测设备，使得采集到的原始数据极易受到噪声干扰。此外，来自各种传感器数据的彼此之间还存在着复杂的相关关系和共线性关系，所以原始数据的质量较低。若直接使用这些低质量的原始数据进行建模分析将会得出错误的挖掘结果。因此，在开展大数据建模、计算、分析和挖掘工作之前，需要对原始数据进行预处理。广义的数据预处理概念包括数据清洗、数据集成、数据变换和数据归约等方法[3]。针对上述通过传感器采集获得的结构化的铁路数据，本文的数据预处理主要指数据去噪、数据降维、时域特征提取等内容。主成分分析与奇异值分解作为线性代数中一种重要的矩阵分解技术，已成为近年来常用的时域数据预处理方法[4]，非常适用于解决数据去噪、数据降维等预处理过程中的问题。

收稿日期：2016-06-15
基金项目：中国铁路总公司科技研究开发计划课题（2015X003-F）；中国铁道科学研究院院基金重大项目（1551DZ8004）。
作者简介：徐贵红，副研究员；郭剑峰，助理研究员。

1　主成分分析与奇异值分解

主成分分析与奇异值分解技术作为基础的数学分析方法，在各领域中的应用已非常广泛，本节主

要阐述主成分分析与奇异值分解的基本原理。

1.1 主成分分析

主成分分析（PCA，Principal Component Analysis，），也称作主元分析，其基本思想是采用坐标系旋转的方法，找出几个少数的综合变量代替原来的多个变量，使这些少数的综合变量彼此之间互不相关而且尽可能的保留原始变量中的信息。通常，数学上的处理方法是把原始变量做线性组合。原始变量中的信息可以用方差度量，其方差值越大，保留的信息就越多。记第 1 个线性组合得到的综合变量为 Y_1，其具有最大方差 var(Y_1)，称其为第 1 主成分；若第 1 主成分不足以代表原来的 p 个变量的信息，就再选取第 2 个线性组合 Y_2，称其为第 2 主成分，依此类推 [5]。

主成分分析法的数学表示如下：实际数据 $X=(X_1, X_2,\cdots, X_p)$ 中有 p 个随机变量，其协方差矩阵 $\sum_X$ 如下：

$$\sum_X=(\sigma_{ij})_p=E[(X-E(X))(X-E(X))^T] \tag{1}$$

设协方差矩阵 $\sum_X$ 的特征值为 $\lambda_1 \geqslant \lambda_2,\cdots, \lambda_{p-1} \geqslant \lambda_p \geqslant 0$，对应单位正交特征向量：

$$U=[u_1, u_2,\cdots, u_p]=\begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1p} \\ u_{21} & u_{22} & \cdots & u_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ u_{p1} & u_{p2} & \cdots & u_{pp} \end{bmatrix} \tag{2}$$

式中：$u_k=(u_{1k}, u_{2k},\cdots,u_{pk})^T$, $k=1,2,\cdots, p$

原始数据 X 对应的第 k 个主成分为：

$$Y_k=u_kX=u_{1k}X_1+u_{2k}X_2+\cdots+u_{pk}X_p \tag{3}$$

主成分有如下两个性质：

$$\begin{cases} \mathrm{var}(Y_k)=u_k^T\Sigma_X u_k=\lambda_k \\ cov(Y_k, Y_j)=u_k^T\Sigma_X u_j=0 \end{cases} \tag{4}$$

即：得到的新指标 $Y_1,Y_2,\cdots,Y_p$ 能够充分反映原来的指标集合中的信息而且相互独立，去除了原变量中的相关性和多重共线性干扰。

1.2 奇异值分解

主成分分析的分解方法是奇异值分解（SVD，Singular Value Decomposition）的一种变形和弱化 [6]，对于 m•n 阶矩阵 X，其中的元素全部属于实数域或复数域。则存在一个分解，使得：

$$X = USV^H \tag{5}$$

式中，U 是 m•m 阶酉矩阵；S 是半正定 m•n 阶对角矩阵；V 的共轭转置 V^H 是 n•n 阶酉矩阵。该分解称为 X 的奇异值分解。矩阵 S 对角线上的元素 S_{ii} 称为 X 的奇异值，奇异值由大至小排列。

在矩阵 X 的奇异值分解中，酉矩阵 U 的列组成一组对 X 正交输入的基向量，这些向量是由 XX^H 构成矩阵的特征向量；酉矩阵 V 的列组成一套对 X 正交输出的基向量，这些向量是 X^HX 构成矩阵的特征向量；S 对角线上的元素是奇异值，其作用是在输入与输出之间进行标量的膨胀控制，同时也是 X^HX 及 XX^H 的特征值，并与 U 和 V 的列向量相对应。

2 PCA与SVD在铁路数据预处理中的应用

针对检测车采集的铁路动态检测数据；工务、电务、机务、车辆、供电专业的静态检查数据等结构化的铁路数据，本节主要阐述主成分分析与奇异值分解在数据降维、特征提取、数据去噪等领域中的应用方法。

2.1 数据降维

铁路运输行业在生产过程中形成的数据包含了工务、电务、机务、车辆、供电等多专业的动、静态检测数据，这些不同专业的数据往往是庞大的、高维的，并且在高维数据内部和不同数据之间都存在着较强的相关性和多重共线性等复杂关系。例如：轨道不平顺检测数据中的轨向不平顺和车辆动态响应检测数据中的横向加速度之间存在着较强的相关关系；车辆动态响应检测数据中的轴箱垂向加速度和车辆动力学检测数据中的轮轨垂向力之间存在着较强的相关关系和共线性，原始数据中的这种关系如图 1 所示。

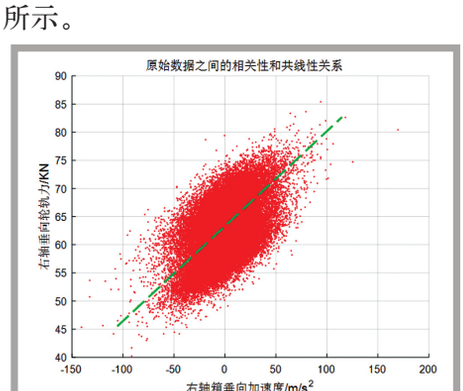

图1 原始数据之间存在的相关性和共线性关系

原始数据中的相关关系和多重共线性会影响建立数据模型的准确性，将这些数据组合在一起还有可能带来维数灾难，所以不能或无法直接使用这些原始数据建立数据模型。为此，可以使用主成分分析与奇异值分解技术将这些多源、高维、海量的铁路各专业原始数据加以融合，获得这些数据的主成分。用融合之后形成的几个少数的低维综合变量来代替原始数据高维变量中的复杂信息，这些少数的几个低维综合变量覆盖了原始数据中的绝大部分信息，因此，也可以视为原始数据在时域中的一种特征。由于主成分分析与奇异值分解得到的特征向量都是互不相关的，所以这种预处理方法不仅降低了数据的维度，同时还去除了原始数据中的相关性和多重共线性干扰，可为数据建模、分析、挖掘提供可用的特征数据。

2.2　数据去噪

在对铁路数据进行预处理时，除了要降低数据维度，消除数据中的相关性和共线性干扰，铁路运输行业在生产过程中形成的数据主要利用各种传感器、惯性测量和数字信号处理设备采集，由于外界阳光、雨水、异物等干扰、振动传感器的频响以及数据传输误差等原因，铁路数据中常存在着冲击噪声和白噪声。去除冲击噪声有诸多处理方法，如中值滤波等[7]，但滤波会损失掉数据中的部分有用信息，如数据的均值趋势项等主要成分。此外，白噪声的频谱是布满整个频域的，一般不容易直接滤除，需要分析信号的频域特征后，通过设定带通滤波器予以滤除。使用传统的滤波方法滤除冲击噪声会减少数据中的有用信息并降低数据的质量。根据冲击噪声和白噪声具有无固定指向性的随机特点，可以通过主成分分析或奇异值分解舍去最后几个特征值或奇异值小的随机成分予以去除。基于主成分分析和奇异值分解的铁路数据预处理去噪方法的原理可以从其几何意义中得出。

探究主成分分析与奇异值分解的几何意义，有助于更加直观的理解该技术在数据去噪中发挥的重要作用。以任意的二阶矩阵为例，通过 PCA 或 SVD 都可以将一个相互垂直的网格坐标系变换到另外一个相互垂直的网格坐标系，这种坐标系的旋转变换并没有改变原向量的主要成分，如图 2 所示。

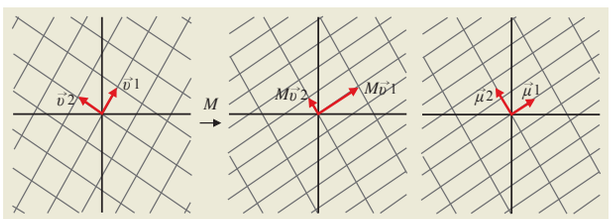

图2　PCA和SVD的几何意义

因此，以二维空间为例[8]，在使用 PCA 和 SVD 进行数据去噪时，n 个满足正态分布的样本数据在二维空间中的分布大致可以表示为一个椭圆，其中每个样本有 2 个变量。由于在奇异值分解和主成分分析中奇异值和特征值越小的主成分包含的信息量越小，而且冲击噪声和白噪声也是随机发生的，即没有固定的指向性，因此可以通过舍去最后几个较小的主成分来达到去除噪声的目的[9]，如图 3 所示。

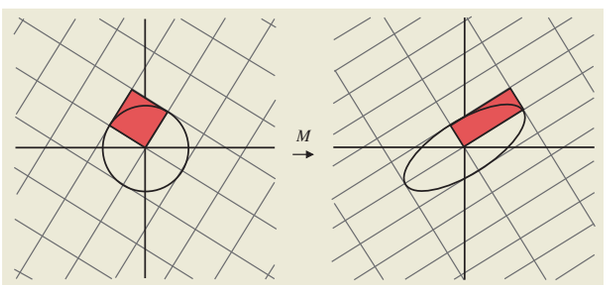

图3　PCA和SVD去噪的图形解释

利用基于主成分分析和奇异值分解的数据预处理噪声去除方法对实测的铁路轨道轨向不平顺数据进行去噪分析，试验的结果如图 4 所示。

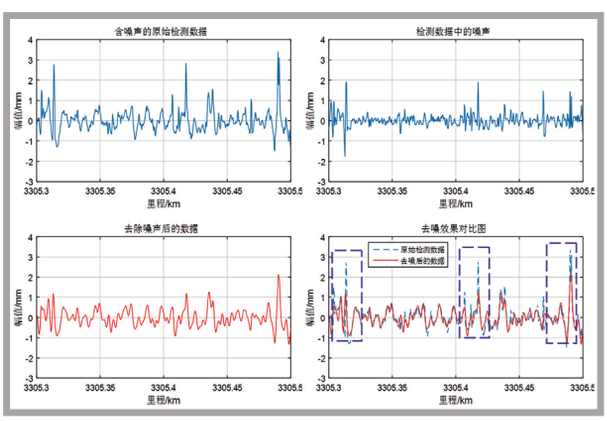

图4　铁路数据去噪试验结果

从图 4 的测试结果中可以看出，基于主成分分析和奇异值分解的噪声去除方法不仅能去除

（下转 P62）

故的频率越高。从图 4 中可以看出，结果比较符合事故的空间分布情况，能够比较好地满足实际应急维修需要，且在线网上均衡性较好。可知，本文研究所构建的模型具有实际可行性。

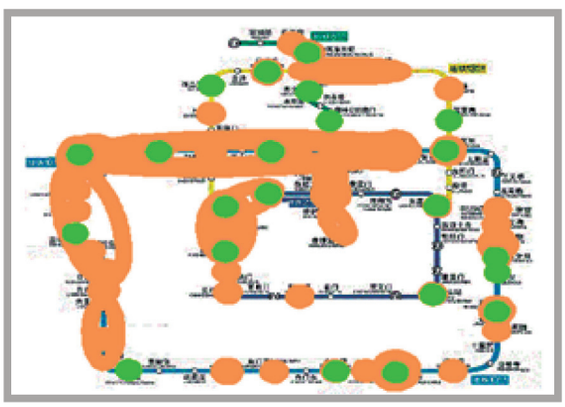

图4　选址结果与事故情况对比图

6　结束语

本文提出的分配策略研究方法同时考虑了历史事故发生情况与线网现状，可以得出较为合理的分配方案，结果能够同时满足实际维修和均衡性的要求，具有较好的参考价值。但仅对应急维修人员进行了路网上的分配研究，而在对城市轨道交通的运营设备故障引起的事故进行应急维修时，除需要相应的维修人员外，还需要相应的维修物资。今后还需要在建立与应急救援物资的联合布局方面进行进一步的研究。

参考文献：

[1] Rashmi S.Shetty. An eventenvironment[D]. Tampa Bay:single game solution for resource allocation in a multi-crisis University of South Florida. 2004.

[2] 孙晓临．城市轨道交通网络应急救援站设置与资源配备优化研究 [D]. 北京：北京交通大学，2012.

[3] 曾敏刚，崔增收，李　双．一种多受灾点的灾害应急救援资源分配模型 [J]. 工业工程，2010, 13（1）：85-89.

[4] 孙彩红．基于网络化的地铁应急救援资源配置 [D]. 北京：北京交通大学，2010.

[5] 徐之恒．地铁突发事件应急救援资源的配置研究 [D]. 北京：北京交通大学，2012.

责任编辑　徐侃春

（上接 P57）

K3305+319、K3305+421、K3305+493 处的冲击噪声干扰，而且还去除了该区段数据中的白噪声干扰。这些噪声是由于外界的阳光反射、雨水冰雪、异物入侵干扰、振动传感器的频响干扰以及数据传输误差、电磁干扰等原因造成的。去除这些噪声干扰后保留了原始数据主成分中的有用信息，可为后续的数据建模、分析、计算和挖掘等工作提供有效的实验数据。

3　结束语

主成分分析与奇异值分解可以将多源、高维、海量的铁路各专业原始数据融合，获得这些数据的主成分，降低数据的维度，去除原始数据中的相关性和多重共线性干扰，为数据建模、分析、挖掘等工作提供可用的特征数据，是一种有效的铁路数据预处理方法。

该技术可用于铁路数据预处理中的去噪环节，不仅能够去除数据中的冲击噪声干扰，还可以去除数据中含有的白噪声干扰，因此，具有较好的推广前景及应用价值，可用于多种铁路数据的预处理过程中。

参考文献：

[1] 王卫东，徐贵红，刘金朝，等．铁路基础设施大数据的应用与发展 [J]. 中国铁路，2015（5）：1-6.

[2] 徐贵红，陶　凯，刘金朝，等．铁路工务安全生产管理分析系统 [J]. 铁路技术创新，2015（2）：27-30.

[3] García S, Luengo J, Herrera F. Data preprocessing in data mining [M]. Switzerland: Springer, 2015.

[4] Wall M E, Rechtsteiner A, Rocha L M. Singular value decomposition and principal component analysis[M]. A practical approach to microarray data analysis. Springer US, 2003: 91-109.

[5] Jolliffe I. Principal component analysis[M]. John Wiley & Sons, Ltd, 2002.

[6] Van Loan C F. Generalizing the singular value decomposition[J]. SIAM Journal on Numerical Analysis, 1976, 13(1): 76-83.

[7] 刘金朝，王卫东，孙善超．铁路轨道几何数据冲击噪声小波有序中值滤波方法 [J]. 振动与冲击，2014, 33（10）：29-33.

[8] 周宪英，高成文，曹建华．主成分分析法及其在数据降噪中的应用 [J]. 兵工自动化，2014, 33（9）：55-58.

[9] 钱征文，程　礼，李应红．利用奇异值分解的信号降噪方法 [J]. 振动、测试与诊断，2011, 31（4）：459-463.

责任编辑　陈　蓉