

摘要：随着信息化社会的到来，函数型数据变得越来越重要。一般而言，一元函数型数据用于研究一个变量的统计规律，多元函数型数据用于研究多个变量之间的相互依赖关系以及统计规律，所以多元函数型数据对研究对象的分析更加深入。目前，针对多元函数型数据的研究方法较少，特别是基于多元函数型数据的识别问题更是鲜有研究。因此，为了有效抓取多元函数型数据的特征，文中给出了多元函数主成分的数值算法，并将其应用于英文手写体数据集、中文手写体数据集和运动数据集的识别任务中。实验表明，与各种基函数展开法进行比较，多元函数主成分分析这一特征抽取方法的识别效果更好，为多元函数型数据的识别学习奠定了基础。

主要贡献

- 1) 给出多元函数主成分的数值计算方法，为多元函数型数据的特征抽取提供了思路。
- 2) 转化英文手写体数据、中文手写体数据和运动数据为多元函数型数据，然后运用多元函数主成分进行特征抽取，找到关键区分特征。
- 3) 结合随机森林算法分类预测，实验表明多元函数主成分这一特征抽取方法优于其他的多元函数型数据的特征抽取方法。

具体算法

算法步骤

Step1：

对样本数据集进行B样条基展开处理，得到函数曲线。

Step2：

在函数曲线上进行采样，得到新数据集。采样点尽可能密集，提高函数主成分的精度。

Step3：

对新数据集进行标准化，消除不同变量之间量纲的影响。

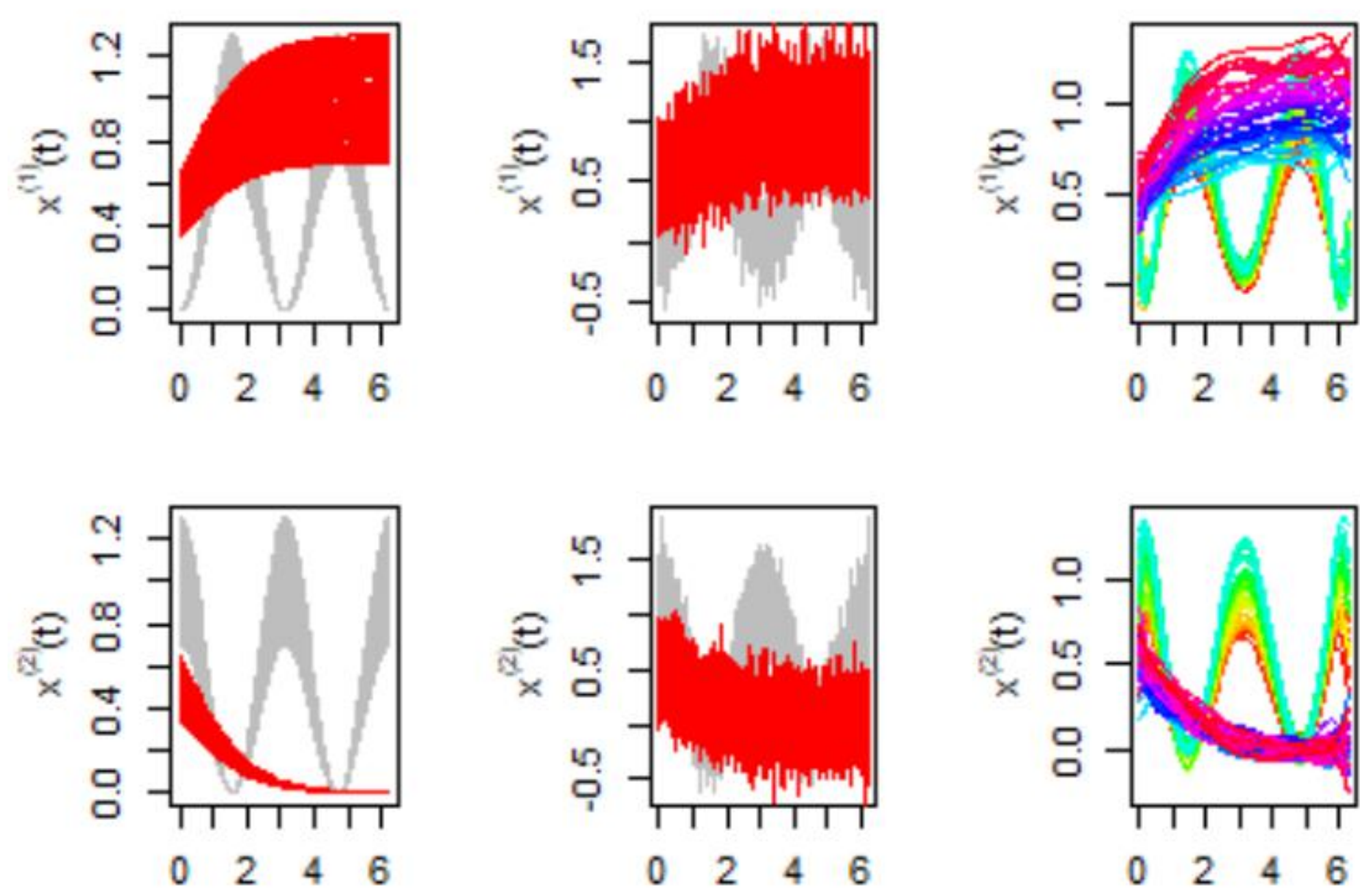
Step4：

完成数据标准化后，采用多元函数主成分分析的数值计算方法，得到多元函数主成分的特征值、特征向量以及各主成分得分。

Step5：

将各主成分特征向量平滑处理得到多元主成分函数。

模拟实验



(a) Generated curve (b) Original curve (c) Fitting curve
图1 数据的生成和拟合(两行分别描述第一、二维函数型数据)

模拟实验数据集分为两类，第一类数据集是由 $y = \sin^2 x$ 和 $y = \cos^2 x$ 构成；第二类数据是由 $y = \frac{e^x}{1+e^x}$ 和 $y = \frac{1}{1+e^x}$ 构成。经处理两个维度的生成数据集、原数据集和拟合后的函数型数据集见图1。

模拟数据集运用多元函数主成分分析数值计算方法，得到主成分函数和主成分贡献率表。选择适合的主成分，计算其得分，见图2。据此将数据集进行分类，采用随机森林算法，模拟100次随机抽样实验，得到分类预测结果，其分类准确率达到1。模拟实验表明多元函数主成分分析在该分类问题中具有很好的识别效果。

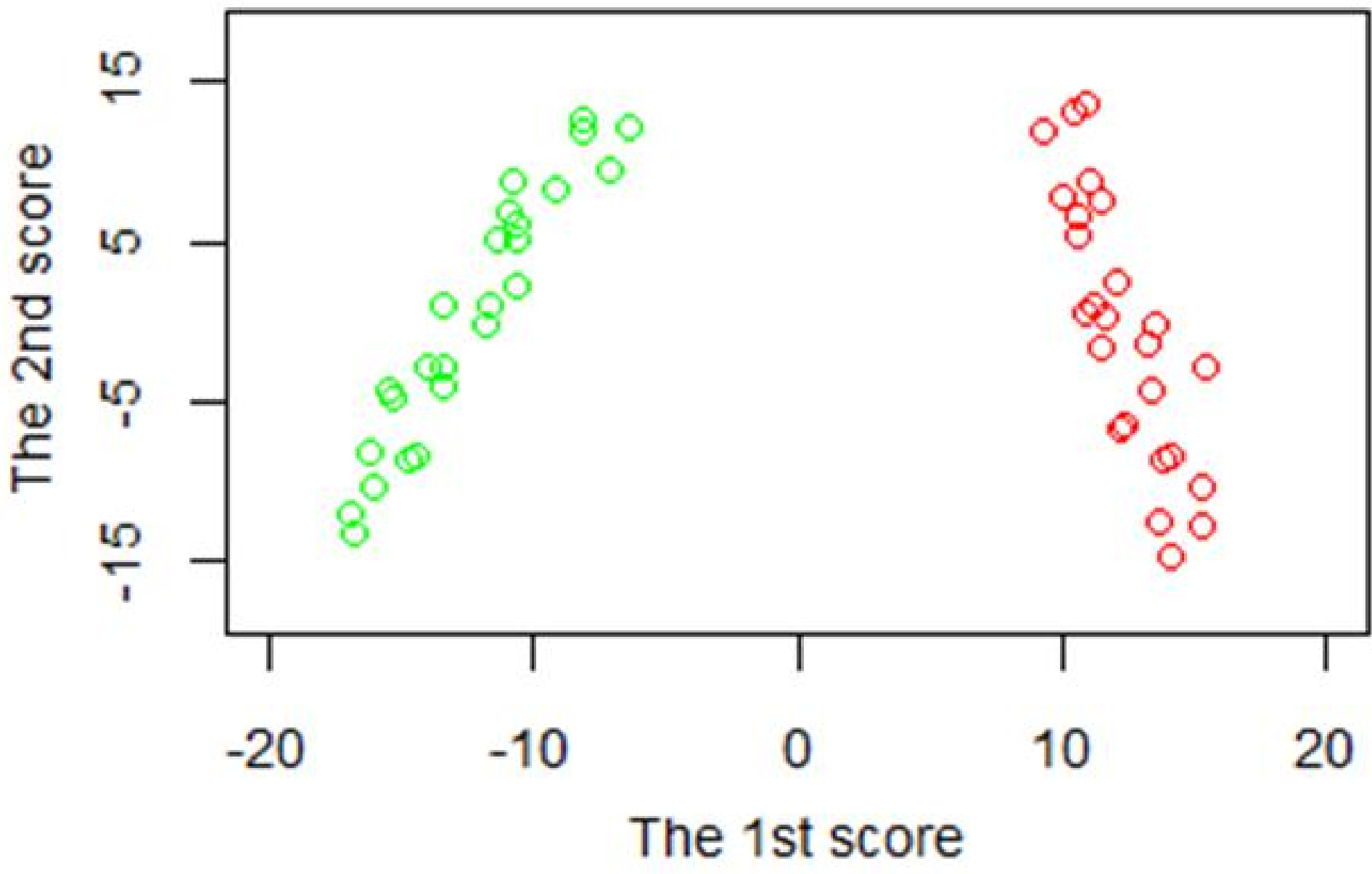


图2 各观测对象主成分得分

真实案例

- 案例1：英文手写数据集，该数据集是一个草书笔迹样本，由20个重复的手写体“fda”组成，每个手写样本有1401对坐标。
- 案例2：Statistical Science in Chinese，该数据集是重复手写简体中文“统计学”50次笔尖的三维坐标值，每次手写过程包含601个采样点。
- 案例3：UCF iPhone 运动数据集，该数据集是不同有氧运动状态下iphone手机测量记录的数据。观测对象数为114个，运动状态分为bike，climbing，gymbike三个类别。数据集的维度为4，每条记录包含500个采样点。

以上三个数据集采用不同特征提取方法进行分类预测。结果如表1所示（表1仅展示运动数据集分类预测结果），经实验效果比对发现，多元函数主成分方法的分类效果明显优于其他方法。

表1 运动数据集的分类正确率

Feature extraction methods	Mean	Std
B-spline	0.9228	0.0616
Fourier	0.9283	0.0554
Polygonal	0.9328	0.0608
Exponential	0.9072	0.0647
Monomial	0.8622	0.0837
Power	0.8589	0.0787
MFPCA	0.9544	0.0532

总结

本文面向多元函数型数据的识别问题展开研究。首先将观测数据转换为多元函数型数据，然后对多元函数型数据进行主成分表示，最后将主成分特征输入到随机森林这一分类器中，得到识别结果。实验结果表明，多元函数主成分的分类效果明显优于其他基展开方法的分类效果，为多元函数型数据在分类问题中的特征抽取提供了新思路。未来可以将多元函数主成分分析推广到视频、音频等数据类型的研究中。

投稿人 姓名：孟银凤、李庆方
单位：山西大学