Operations Research and Fuzziology 运筹与模糊学, 2019, 9(2), 156-164
Published Online May 2019 in Hans. http://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2019.92018

Hans汉斯

Linear Discriminant Analysis for Functional Data

Xintong Wang

College of Mathematics and Systems Science, Xinjiang University, Urumqi Xinjiang
Email: 2477224765@qq.com

Received: May 6th, 2019; accepted: May 20th, 2019; published: May 27th, 2019

Abstract

In this paper, functional linear discriminant analysis method is proposed for the classification problem of input as functional data. By introducing the functional norm to measure the distance within-class and between-class, an optimization model of functional linear discriminant analysis is constructed. Furthermore, by using the basis function method to transform the infinite dimensional function space into a finite dimensional optimization model, then this model is easy to solve. Since the data is functional, the first derivative or the second derivative of the function can be found. The classification result can be further improved by using the data after the derivative. Finally, the numerical experiments show the feasibility and effectiveness of the functional linear discriminant analysis method.

Keywords

Functional Data, Linear Discriminant Analysis, Classification

函数型线性判别分析

王馨彤

新疆大学数学与系统科学学院，新疆 乌鲁木齐
Email: 2477224765@qq.com

收稿日期：2019年5月6日；录用日期：2019年5月20日；发布日期：2019年5月27日

摘　要

本文针对输入为函数型数据的分类问题提出了一个函数型线性判别分析方法。通过引入函数范数来度量类内距离和类间距离，从而构造了函数型线性判别分析的优化模型。进一步，通过利用基函数方法将无

文章引用：王馨彤. 函数型线性判别分析[J]. 运筹与模糊学, 2019, 9(2): 156-164.
DOI: 10.12677/orf.2019.92018

穷维函数空间优化模型转化为有限维优化模型，从而使模型易于求解。由于数据被函数化后，可对函数求一阶导数或二阶导数。利用求导数后的数据可进一步提高分类效果。最后，数值实验部分展示了函数型线性判别分析方法的可行性和有效性。

关键词

函数型数据，线性判别分析，分类问题

1. 引言

机器学习[1]是一门多领域的学科，主要研究计算机如何模拟和学习人类的各种活动，并在储存信息的过程中不断达到自我完善。现如今，机器学习已经广泛应用在实际生活中。分类问题是机器学习领域中最常见的一类问题，也是机器学习领域的研究热点之一，如模式识别[2]、文本分类[3]、手写字体识别[4]、人脸图像识别[5]等领域。然而，通常的分类问题的输入是向量的形式呈现的，传统的分类学习机也只是解决输入为向量的分类问题。事实上有些数据是随着时间变化的，应当以动态的角度来看待数据，寻找出数据中隐含的某种函数特性，如果我们利用函数型数据分析技术找出数据隐含的连续函数，也可进一步对函数求一阶或者二阶导数，能挖掘数据中隐含的更多信息。本论文主要针对输入为函数型数据的分类问题提出了一个新的函数型线性判别分析算法。

线性判别分析[6] (linear discriminant analysis, LDA)是 Fisher 在 1936 年首次提出，在人脸识别[7]、食品安全[8]、医学[9]等领域有广泛的应用。在模式识别领域中，贝叶斯分类效果最优，但由于其概率密度函数难以估计，使得线性分类器(如 Fisher 线性判别分析)被广泛的应用在实际生活中。Li (2011) [10]等首先将数据利用线性判别分析方法把训练样本投影到子空间中提取数据的相应特征，再利用支持向量机进行分类。实验结果表明，该方法不仅起到降维的作用，同时提高了分类准确率，大大缩短了计算的时间。Li [11]提出了 2－维线性判别分析(2-dimensional linear discriminant analysis, 2D-LDA)的方法，该方法从图像矩阵中提取出重要的特征，进而计算类间散度矩阵和类内散度矩阵。而线性判别分析和 2－维线性判别分析存在稀疏性的问题，在特征提取中存在遗漏重要数据信息的情况。但以上的分析方法都无法解决以函数型数据为输入的分类问题。从而需要对函数型数据的内在结构和特性进行了解和分析。

函数型数据的研究是 Ramsay [12]在 1982 年首次提出，阐述了函数型数据不再是传统的静态数据，而被视为动态数据，当原始数据信息丢失或者缺损时，可利用数据的函数特性进行填补。1991 年，Ramsay [13]提出了一些适用于具有时间序列变化数据的分析方法，如加拿大气象站的日降水量分布情况和温度变化的联系，用函数型主成分分析方法解决了这一实际问题。随后，Ramsay 在《Functional Data Analysis》[14]中对函数型数据进行了详细的概括和总结，其中包括函数型数据的概念，处理函数型数据的方法，并将主成分分析、典型相关分析、判别分析及线性模型等经典方法引入到函数型数据分析中。

针对函数数据的分类问题，本文提出函数型线性判别分析，该方法把函数型数据表示成光滑曲线或者连续函数，从而我们可以考虑到函数的特性(如连续、求导)。函数型线性判别分析的主要思想是通过给定的训练集，寻找一条投影函数，使同类样本的投影点距离尽可能的接近，同时又使异类样本的投影点尽可能的远离。基函数法是常见的将原始数据转化为函数的平滑技术之一，即将函数型数据在一组基下

进行展开。因此寻找这条投影函数，就变成寻找基函数的系数向量。函数型线性判别分析的优点之一是可以求出导数曲线或者微分曲线，并通过求导或者微分我们可从数据中挖掘出隐藏的重要信息，从而得到更好的分类结果。

本文的组织结构如下：第二章：给出了一些准备工作，对线性判别分析进行了简单的回顾。第三章：详细的介绍了本文提出的方法。第四章：把函数型判别分析与线性判别分析方法进行比较，用数值实验说明本文方法的可行性和优势。最后第五章是本文的结论与展望。

2. 相关工作

线性判别分析方法的主要思想是将样本投影到一条直线上，使其同类样本的投影点更聚集、不同类样本的投影点尽可能的分离，从而最终确定投影方向 w，其主要过程如下：

给定训练集 $D=\left\{\left(x_j,y_j\right)\right\}_{j=1}^{m}$，$x_j\in R^n, y_j\in\{1,2\}$，令 $X_i,\mu_i,\sum_i$ 分别表示第 $i\in\{1,2\}$ 类样本的集合、均值向量、协方差矩阵。则第 i 类均值为 $\mu_i=\frac{1}{n_i}\sum_{x\in X_i}x$，其中 n_i 表示第 i 类样本的个数。

类内散度矩阵表示该类样本到中心的距离即：

$$S_w=\sum_{x\in X_1}\left(x-\mu_1\right)\left(x-\mu_1\right)^T+\sum_{x\in X_2}\left(x-\mu_2\right)\left(x-\mu_2\right)^T \tag{1}$$

类间散度矩阵表示各类中心到总体中心的距离即：

$$S_b=\left(\mu_1-\mu_2\right)\left(\mu_1-\mu_2\right)^T \tag{2}$$

为使每类在 w 方向投影后的样本距离尽可能的大，则优化模型为：

$$\max J=\frac{w^TS_bw}{w^TS_ww} \tag{3}$$

其中式(3)是类间散度矩阵 S_b 相对于类内散度矩阵 S_w 的广义 Rayleigh 熵。由于式(3)的分子和分母都为二次项，因此与长度无关，只与方向有关。令 $w^TS_ww=1$ (不失一般性)，则式(3)等价于最小化 $-w^TS_bw$，利用拉格朗日乘子法，最终式(3)等价于 $S_bw=\lambda S_ww$，即为一般特征值问题。

3. 函数型线性判别分析

函数型数据两分类问题描述如下：

给定训练集 $D=\left\{x_j(t),y_j\right\}_{j=1}^{N},y_j\in\{1,2\}$，我们的目的是找一条投影曲线 $w(t)$，使同类样本的投影点尽可能接近，使异类样本的投影点尽可能的远离。找到投影曲线后，再对新样本进行分类，将其投影到同样的曲线上，最终根据投影点的位置来确定新样本的类别。则优化模型如下：

$$\max J=\frac{\left\|\int w(t)\mu_1(t)-\int w(t)\mu_2(t)\right\|_2^2}{Var\int w(t)x_1(t)+Var\int w(t)x_2(t)} \tag{4}$$

其中 $\mu_i(t)=N^{-1}\sum_{i=1}^{N}x_i(t)$，$Var\int w(t)x_i(t)=N^{-1}\sum_{i=1}^{N}w(t)x_i(t)$ 分别表示第 $i\in\{1,2\}$ 类样本的均值函数，协方差矩阵。将数据投影到函数 $w(t)$ 上，两类样本的中心在函数上的投影分别为 $\int w(t)\mu_1(t)$ 和 $\int w(t)\mu_2(t)$；则让类中心之间的距离尽可能的大，即 $\left\|\int w(t)\mu_1(t)-\int w(t)\mu_2(t)\right\|_2^2$ 尽可能的大。将所有样本点都投影到这条曲线上,则同类样本的协方差矩阵分别为 $Var\int w(t)x_1(t)$ 和 $Var\int w(t)x_2(t)$。我们需使同类样本的距离尽

可能的接近，即让两类样本的协方差的和尽可能的小。

通过上式(4)求得投影函数为 $w^*(t)$，则对新来的样本 $x_j^*(t)$ 投影到函数 $w^*(t)$ 下的得分定义为 f_j，即 $f_j=\int w^*(t)x_j^*(t)\mathrm{d}t$。最终计算得分 f_j 与均值函数 $\mu_i(t)$ 的距离，根据所在投影点的位置确定新样本的类别；也就是说，新的样本与哪一类中心点越接近，就被分到这一类，即 $\min\left\|f_j-\mu_i(t)\right\|^2$。

由于函数型数据是无穷维的，则需对函数型数据进行降维处理。常见的降维方式为函数型数据在一组基下进行展开。具体如下：由 K 个已知的基函数的线性组合来拟合已知的曲线样本 $x_j(t)$，公式如下

$$\hat{x}_j(t)=\sum_{k=1}^{K}c_{jk}\phi_k(t) \tag{5}$$

其矩阵形式表示如下

$$\hat{x}_j(t)=C_j^T\Phi=\Phi^TC_j \tag{6}$$

其中 $\Phi=\left(\phi_1(t),\cdots,\phi_K(t)\right)^T$ 表示 K 维基函数，其系数向量为 $C_j=\left(c_{j1},\cdots,c_{jK}\right)^T$。当 K 取的越大，近似的精度越高。本文的目的是寻找函数型线性判别分析下的投影函数 $w^*(t)$，则将 $w^*(t)$ 同样用基函数线性组合的形式展开为 $w(t)=\sum_{k=1}^{K}d_kt_k(t)=d^T\Phi$，因此寻找投影函数就变成了寻找基函数所对应的系数向量 $d=\left(d_1,\cdots,d_K\right)^T$。

因此通过基底函数法，把函数型数据用基函数的线性组合展开，所得第 1 类的协方差矩阵表示如下：

$$\begin{aligned}\mathrm{var}\int w(t)x_{j1}(t)&=N^{-1}\int w(t)x_{j1}(t)\int w(t)x_{j1}(t)\\&=N^{-1}\int \mathrm{d}^T\Phi\Phi^T\bar{C}\int\left(\mathrm{d}^T\Phi\Phi^T\bar{C}\right)^T\\&=N^{-1}\mathrm{d}^T\int\Phi\Phi^T\bar{C}\bar{C}^T\int\Phi\Phi^T\mathrm{d}\\&=\mathrm{d}^TJV_0J\mathrm{d}\end{aligned} \tag{7}$$

同样地，$i=2$ 类的协方差矩阵可以表示为：

$$\begin{aligned}\mathrm{var}\int w(t)x_{j2}(t)&=N^{-1}\int wx_{j2}\int wx_{j2}\\&=N^{-1}\int \mathrm{d}^T\Phi\Phi^T\hat{C}\int\left(\mathrm{d}^T\Phi\Phi^T\hat{C}\right)^T\\&=N^{-1}\mathrm{d}^T\int\Phi\Phi^T\hat{C}\hat{C}^T\int\Phi\Phi^T\mathrm{d}\\&=\mathrm{d}^TJV_1J\mathrm{d}\end{aligned} \tag{8}$$

其中矩阵 J 为 K 阶对称阵其元素表示为 $J_{ij}=\int\Phi_i\Phi_j^T$。$V_0,V$ 分别是关于系数矩阵 $\bar{C},\hat{C}_1$ 的协方差矩阵。同样地，把 $\left\|\int w(t)\mu_1(t)-\int w(t)\mu_2(t)\right\|_2^2$ 也用基底函数法进行转换公式表示如下：

$$\begin{aligned}\left\|\int w(t)\mu_1(t)-\int w(t)\mu_2(t)\right\|_2^2&=\int w(t)\left(\mu_1(t)-\mu_2(t)\right)\mathrm{d}t\int w(t)\left(\mu_1(t)-\mu_2(t)\right)\mathrm{d}t\\&=\int \mathrm{d}^T\Phi\Phi^Tm\int\left(\mathrm{d}^T\Phi\Phi^Tm\right)^T\\&=d^T\int\Phi\Phi^Tmm^T\int\Phi\Phi^T\mathrm{d}\\&=d^TJVJ\mathrm{d}\end{aligned} \tag{9}$$

其中 $\mu_i(t)$ 是第 i 类样本的均值函数，表示为 $\mu_i(t)=m_i^T\phi\left(i\in\{1,2\}\right)$，$m=\left(m_1,\cdots,m_K\right)$ 为 $\mu_0(t)-\mu_1(t)$ 的系数向量，则优化模型为

$$\max J(d)=\frac{d^T JVJd}{d^T J(V_0+V_1)Jd} \tag{10}$$

由于式(10)的分子和分母都是关于 d 的二次项，因此式(10)的解与 d 的长度无关，只与其方向有关。不失一般性，令 $d^T J(V_0+V_1)Jd=1$，则式(10)等价于：

$$\begin{aligned}&\min_d && -d^T JVJd\\ &\text{s.t} && d^T J(V_0+V_1)Jd=1\end{aligned} \tag{11}$$

定义拉格朗日函数：

$$L(d,\lambda)=-d^T JVJd+\lambda\left[d^T J(V_0+V_1)Jd-1\right] \tag{12}$$

其中 $\lambda=(\lambda_1,\cdots,\lambda_K)$ 为拉格朗日乘子，并利用 KKT 条件可得：

$$\frac{\partial L}{\partial d}=-2JVJd+2\lambda J(V_0+V_1)Jd=0 \tag{13}$$

由拉格朗日乘子法，上式等价于

$$JVJd=\lambda J(V_0+V_1)Jd \tag{14}$$

令 $L=J(V_0+V_1)J$，即求 $L^{-1}JVJd=\lambda d$ 的一般特征值问题。

综上所述，函数型线性判别分析的算法为：

算法：函数型线性判别分析算法

输入：训练集 $D=\{x_j(t),y_j\}_{j=1}^N, y_j\in\{1,2\}$

输出： d

1) 选择基函数，求解样本函数的系数 $\bar{C},\hat{C}$。

2) 计算协方差矩阵 V_0,V_1,V 和基矩阵 J。

3) 计算(14)式得出 λ。

4) 代入 $f_j=\int wx=\int w(t)x_j(t)\mathrm{d}t$ 并计算 $\min\left\|f_j-\mu_i(t)\right\|^2$。

4. 数值实验

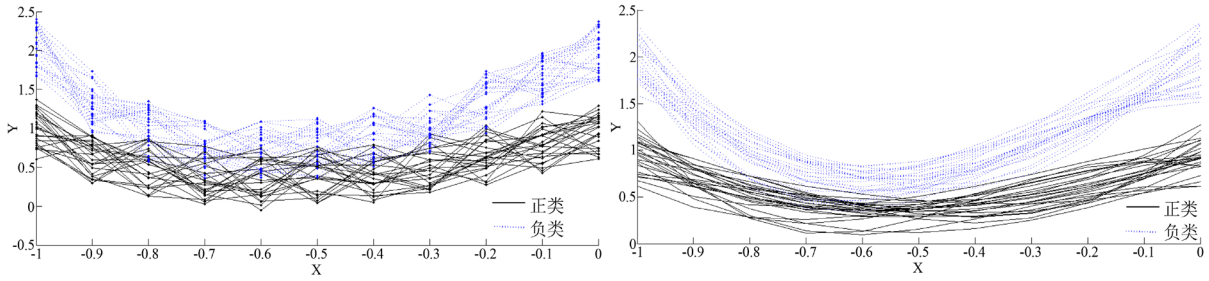
4.1. 人工数据的数值实验

我们用人工数据测试本文的方法，本节数值实验分为两部分，第一部分为线性可分的，第二部分为线性不可分的；首先介绍本节数值实验的第一部分，我们构造 2 个多项式曲线，多项式的最高次数为 4 次幂，其区间为$[-1,0]$并且随机加了噪声，为了使模型几何意义更加明确，我们选取一组多项式系数构成多项式曲线。

图 1(a)表示由离散点构成的 50 条不同类曲线，每类各有 25 条曲线，黑色线表示正类曲线，蓝色表示负类曲线。并对曲线随机加上噪声。从图 1(a)可看出我们很难用肉眼分清哪条曲线属于哪一类。图 1(b)表示通过函数型数据拟合后的曲线，我们可以很显然的看出经过基函数法后的曲线不仅变得光滑而且能清晰地区分两类，其中图中蓝色虚线表示负类，黑色实线表示正类。

图 2 表示随机一次五折交叉验证过后的分类情况，其中“△”表示正类的训练集，“〇”表示负类的训练集。我们可从图 2 看出同类样本点的距离都很接近，并且不同类样本点的类间距离很远，能够清晰地区分

两类。



(a) (b)

Figure 1. (a) Curves of 50 different types of artificial data sets, (b) Curves fitted by basis function method
图 1. (a) 50 条不同类的人工数据集构成的曲线，(b) 用基函数法拟合过后的曲线

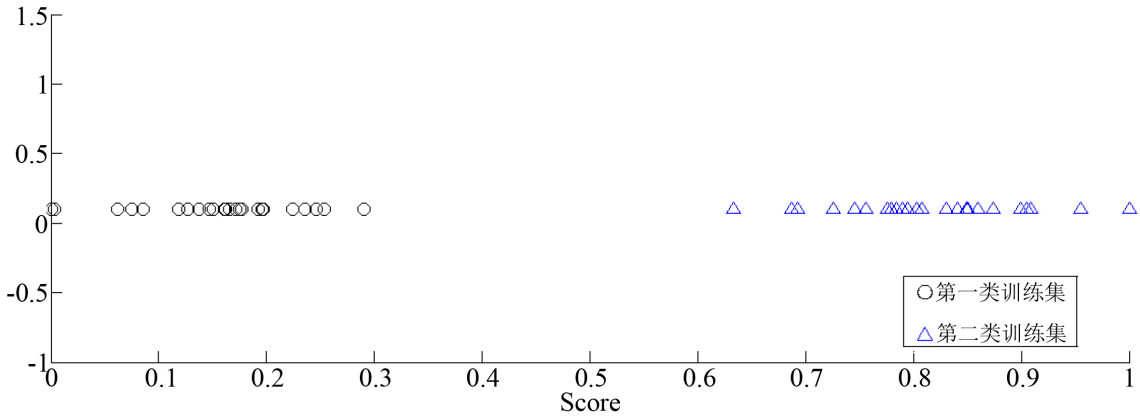
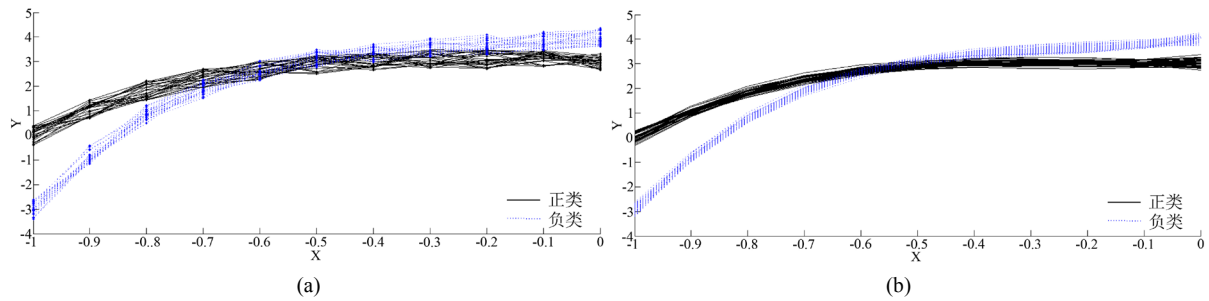


Figure 2. Classification of 50 curves using functional linear discriminant analysis
图 2. 50 条曲线运用函数型线性判别分析方法后的分类情况

第二部分数值实验是线性不可分的曲线，如图 3(a)，我们从肉眼角度观察曲线有重叠的部分，很难区分各个曲线属于哪一类，图 2(b)经过基函数法过后曲线变得较为光滑，能够大致看清分成两类的曲线趋势。表 1 为经过函数型线性判别分析方法过后的两次人工数据的准确率，线性可分的准确率接近于100%，说明在本文方法中，很好的把类内的样本点聚在一起，同时，不同类的样本点的类间距离较远；线性不可分的准确率为 91.5%，由此可见在曲线重叠的部分导致类间距离变小，从而影响了线性不可分曲线的准确率。而把函数型数据离散化后，用原始的线性判别分析方法进行分类，无论是线性可分的数据还是线性不可分的数据，结果都没有函数型线性判别分析方法的准确率高。



(a) (b)

Figure 3. (a) Curves of 50 different types of artificial data sets, (b) Curve fitted by basis function method
图 3. (a) 50 条不同类的人工数据集构成的线性不可分曲线，(b) 用基函数法拟合过后的曲线

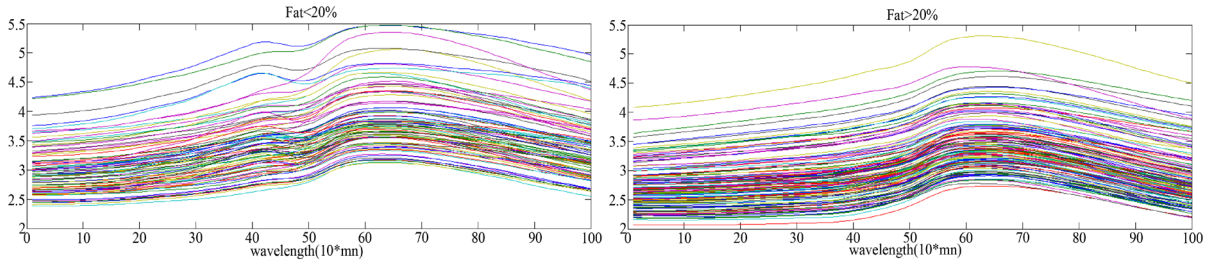
Table 1. Data set details
表 1. 数据集的详细信息

Dataset	LDA	FLDA
	准确率	准确率
人工数据(线性可分)	0.9713	0.998
人工数据(线性不可分)	0.9032	0.915

4.2. Spectrometric 数据集

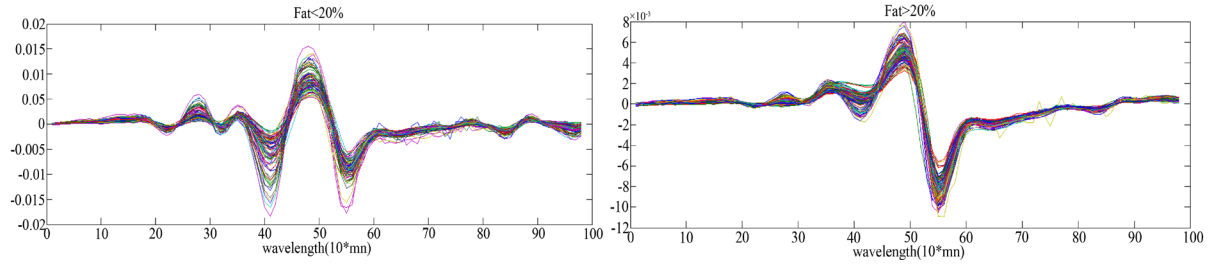
本节介绍的 *Spectrometric* 数据集来自食品工业，每个观测值是肉类样本(经过精细切碎后)的近红外吸收光谱，一共有 215 个光谱样本。观测值中包括了在波长为 850~1050 nm 范围内的 100 个吸收光谱。此数据集的分类问题在于从脂肪含量低(低于 20%)的肉类样品分出高脂肉类(高于 20%)。

图 4 表示一阶求导后的图像，图 5(a)表示脂数低于 20%的曲线，图 5(b)表示脂数高于 20%的曲线，因为图 5(a)、图 5(b)中的曲线波动基本一致，我们无法从一阶求导后的曲线中肉眼看出类别。而图(5)表示二阶求导后的图像，从图中可看出经过二阶求导后的函数曲线有明显不同的波动。因此我们可从求导后的曲线得出隐性的函数条件，从而运用函数型线性判别分析的方法进行分类。



(a) (b)

Figure 4. First-order derivation of spectrometric data
图 4. 对 spectrometric 数据进行一阶求导



(a) (b)

Figure 5. Second-order derivation of spectrometric data
图 5. 对 spectrometric 数据进行二阶求导

图 6 取 20%的 spectrometric 数据一阶求导后的分类情况，发现虽然同类样本距离较为接近,但异类样本的距离没有区分开，因此分类不是很明显。图 7 取 20%的 spectrometric 数据二阶求导后的分类情况，明显地，我们可以看出求完二阶导后的异类样本区分开了，但仍然会有 3 个异常点分错类别。表 2 是 5 次五折交叉验证后的准确率，其平均准确率为 98.30%。

Table 2. Data set details
表 2. 数据集的详细信息

Dataset	准确率				
spectrometric	1.0000	0.9787	0.9574	1.0000	0.9787

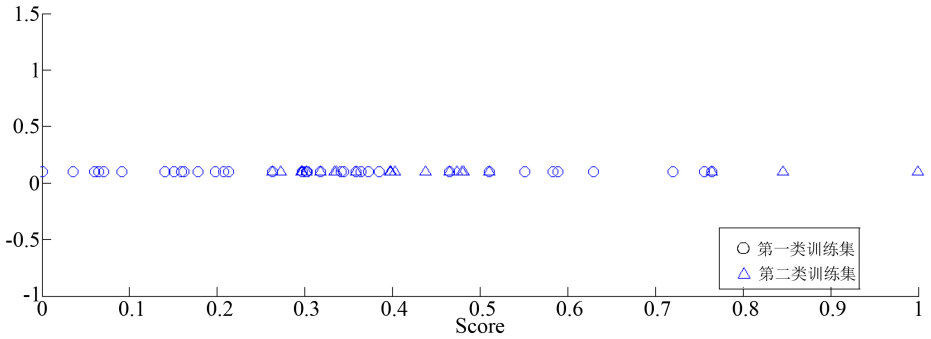


Figure 6. Take 20% of the spectrometric data after the first-order derivation classification
图 6. 取 20%的 spectrometric 数据一阶求导后的分类情况

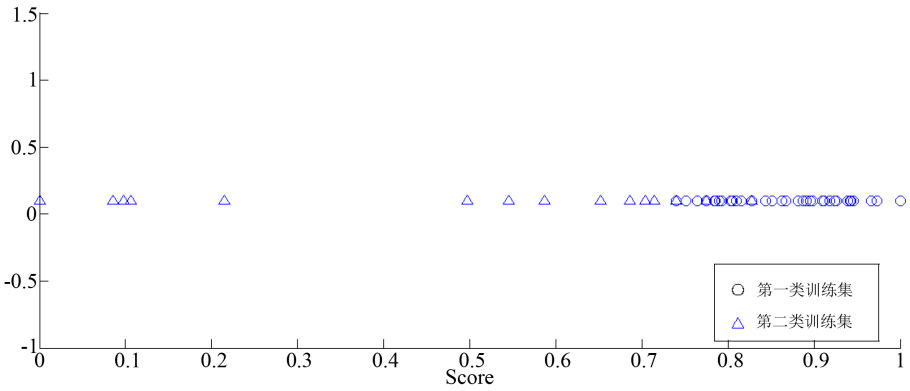


Figure 7. Take 20% of the spectrometric data after the second-order derivation classification
图 7. 取 20%的 spectrometric 数据二阶求导后的分类情况

5. 结论

本文是在传统的线性判别分析的基础上，针对函数型数据，提出了函数型线性判别分析算法来解决函数型数据的两分类问题。该方法证明函数型数据可以挖掘其函数特征，例如连续，求导。实验结果表明，spectrometric 数据集经过二阶求导过后分类效果才明显提高。该方法不仅能对函数型数据分类，还可以对离散数据进行分类，通过拟合估计出的函数转换成函数型数据，同理对其数据进行分类。

参考文献

[1] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016.

[2] 张学工. 关于统计学习理论与支持向量机[J]. 自动化学报, 2000, 26(1): 36-46.

[3] 平源. 基于支持向量机的聚类及文本分类研究[D]: [博士学位论文]. 北京: 北京邮电大学, 2012.

[4] 田盛丰, 黄厚宽. 基于支持向量机的手写体相似字识别[J]. 中文信息学报, 2000, 14(3): 37-41.

[5] 邹建法, 王国胤, 龚勋. 基于增强 Gabor 特征和直接分步线性判别分析的人脸识别[J]. 模式识别与人工智能, 2010, 23(4): 477-482.

[6] Fisher, R.A. (1936) The Use of Multiple Measurements in Taxonomic Problems. *Annals of Human Genetics*, **7**, 179-188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x

[7] Belhumeur, P.N., Hespanha, J.P. and Kriegman, D.J. (1997) Eigenfaces versus Fisherfaces: Recognition Using Class Specific Linear Projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **19**, 711-720. https://doi.org/10.1109/34.598228

[8] Rezzi, S., Giani, I., Heberger, K., *et al*. (2007) Classification of Gilthead Sea Bream (*Sparus aurata*) from 1H NMR Lipid Profiling Combined with Principal Component and Linear Discriminant Analysis. *Journal of Agricultural and Food Chemistry*, **55**, 9963-9968. https://doi.org/10.1021/jf070736g

[9] 张新新, 李雨, 等. 主成分-线性判别分析在中药药性识别中的应用[J]. 山东大学学报, 2012, 50(1): 143-146.

[10] 栗科峰, 卢金燕, 等. 基于子图分割与多类支持向量机的人脸识别方法[J]. 科技通报, 2018(8): 1001-7119.

[11] Li, M. and Yuan, B. (2005) 2D-LDA: A Statistical Linear Discriminant Analysis for Image Matrix. *Pattern Recognition Letters*, **26**, 527-532. https://doi.org/10.1016/j.patrec.2004.09.007

[12] Ramsay, J.O. (1982) When the Data Are Functions. *Psychometrika*, **47**, 379-396. https://doi.org/10.1007/BF02293704

[13] Ramsay, J.O. and Dalzell, C.J. (1991) Some Tools for Functional Data Analysis. *Journal of the Royal Statistical Society*, *Series B*, **53**, 539-572. https://doi.org/10.1111/j.2517-6161.1991.tb01844.x

[14] Ramsay, J.O. and Silverman, B.W. (1997) Functional Data Analysis. Springer-Verlag, New York. https://doi.org/10.1007/978-1-4757-7107-7