Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2024, 13(4), 909-919
Published Online November 2024 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2024.134093

线性映射模型下的重编码判别分析算法

张恒瑞[1]，董英华[1]，方宇翔[2]，许喆源[1]

[1]南京信息工程大学数学与统计学院，江苏 南京
[2]南京信息工程大学计算机与软件学院，江苏 南京

收稿日期：2024年7月29日；录用日期：2024年11月19日；发布日期：2024年11月28日

摘 要

针对Fisher判别法判别同族群体多中心数据准确率低的问题，提出了线性映射模型下的重编码判别分析算法，将Fisher判别法中的降维思想与重编码方法相结合，采用蒙特卡洛法，通过对伪预测数据的划分。以实映射识别率为目标，确定线性判别函数的待定系数和伪预测数据的划分。实证表明，该算法具有较高的识别率和稳定性。

关键词

判别分析，伪预测，重编码，实映射

Recoding Discriminant Analysis Algorithm for Linear Mapping Models

Hengrui Zhang[1], Yinghua Dong[1], Yuxiang Fang[2], Zheyuan Xu[1]

[1]School of Mathematics and Statistics, Nanjing University of Information Science and Technology, Nanjing Jiangsu
[2]School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing Jiangsu

Received: Jul. 29th, 2024; accepted: Nov. 19th, 2024; published: Nov. 28th, 2024

Abstract

Aiming at the problem of low accuracy of Fisher discriminant method in judging the multicenter data of the same population, a recoding discriminant analysis algorithm under linear mapping model is proposed, which combines the idea of dimension reduction in Fisher discriminant method with recoding method, adopts Monte Carlo method, and divides the pseudo-predicted data. Aiming at the recognition rate of real mapping, the undetermined coefficients of linear discriminant

文章引用: 张恒瑞, 董英华, 方宇翔, 许喆源. 线性映射模型下的重编码判别分析算法[J]. 人工智能与机器人研究, 2024, 13(4): 909-919. DOI: 10.12677/airr.2024.134093

function and the division of false prediction data are determined. The empirical results show that the algorithm has high recognition rate and stability.

Keywords

Discriminant Analysis, False Prognosis, Recoding, Real Map

1. 引言

21 世纪以来，科学技术不断进步。在各种高新技术中，计算机技术的高速发展使数据的收集、存储和传输的能力上了新的一个台阶。生物、航空、金融等各个领域都产生了与数据科学相交叉的子学科。在大数据时代下，为了提高对数据处理的有效性和准确性，不断优化各种统计方法、不断构建各种新式算法成为了非常重要且热门的研究方向。

判别分析是隶属于统计学学科下的一种方法，主要用于分析研究对象的各项特征值。其判别分析的基本原理是按照所需要的判别准则，建立判别函数，用研究对象的大量数据作为测试集确定判别函数的各项待定系数，并计算出相关判别指标，从而确定给定样本属于何类。随着判别分析技术的不断发展，人们根据实际需要已经提出了许多分类方法[1]。比如：基于传统统计学的方法有线性判别法[2]，贝叶斯判别法[3]，K-近邻法[4]；基于机器学习的方法有：BP 神经网络[5]，决策树算法[6]，随机森林算法[7]，元胞自动机[8]，遗传算法[9]，集成算法[10]。不同的聚类算法能很好地解决某些特定的问题，但总体上仍然存在许多等待解决的问题。在 20 世纪 30 年代，数学家 Ronald Fisher 以方差分析的基本思想作为切入点，提出了 Fisher 判别分析法[11]。Fisher 判别法也被称为典则判别，通过构造线性组合的方式，将多维数据映射到低维数据，实现对原数据类别的判别。

一般的 Fisher 判别法作为一种模式识别方法，实现的过程包括数据采集、特征选取、模式选择、训练测试、结果计算和复杂度分析等步骤构成。Fisher 判别法一般从方差分析的角度入手，以组间方差和组内方差的比值的最大值来确定线性映射方向，作为判别函数的系数，构建线性判别函数。但 Fisher’s LDA 算法有以下缺陷：首先，只有在多组群体具有相同的协方差矩阵的时候才可以使用 Fisher 方法；其次，Fisher 方法无法在类别重心重合或同类别多中心的时候将两类区分。

为了克服这些缺陷，国内学者刘小平等人将 Fisher 判别思想与元胞自动机结合，用来模拟土地利用变化[12]；学者郑建峰等人采用快速重编码的方法进行数据处理，在压缩域信息隐藏方案中采用的快速重编码方法[13]，为本文提供了新的思路。

本文提出了一种基于线性映射的重编码(Linear map- ping model and recoding, LMMR)，通过线性函数映射到一维空间，再通过伪预测进行区域划分，计算在不同判别标准下不同判别区域不同属性个体的占比，依次迭代与比较得到判错率最低的目标函数。并且在迭代过程中通过灵活应用蒙特卡洛法同时提高迭代效率和准确性，对算法做出局部优化，具有一定的应用价值和理论意义。

2. Fisher 算法与分析

2.1. Fisher 算法的基本概念

文献[14]指出 Fisher 判别分析通过寻找多变量间的线性法则来构造判别函数，从而描述或解释两组

以及多组群体的区别的一种方式。Fisher 判别分析基于一条假设，即多个群体具有互不相同的均值，但是具有相同的协方差矩阵。Fisher's LDA 分类不同于只是确定分类规则的判别分析，而是在分析的过程中做出预测，将对象分到某一类别之中。Fisher's LDA 分类除了均值不同和协方差矩阵相同的条件外，还要求对两个群体进行判别分析时，假设有 2 个 p 维样本，第 1、2 个样本分别为 $\left(y_{11},y_{12},\cdots,y_{1n_1}\right)$、$\left(y_{21},y_{22},\cdots,y_{2n_2}\right)$，其均值向量分别为 $\overline{y}_1$、$\overline{y}_2$，Fisher 判别法需要计算出对应的 t 统计量，即：

$$t=\frac{a'\left(\overline{y}_1-\overline{y}_2\right)}{\sqrt{\left(\frac{1}{n_1}-\frac{1}{n_2}\right)a'S_{pl}a}} \tag{1}$$

其中 a 为所求线性组合的投影方向，S_{pl} 为两个样本共同的离差矩阵，当 $a=S_{pl}^{-1}\left(\overline{y}_1-\overline{y}_2\right)$ 时，t^2 最大。

在多群体 Fisher 判别分析中，假设存在 g 个群体，设第 k 个群组为 $G_k,k=1,2,\cdots,g$。第 k 个样本 $\left(y_{k1},y_{k2},\cdots,y_{kn_k}\right)$ 的均值为 $\overline{y}_k$。多群体 Fisher 判别分析中，以组间方差和组内方差的比值的最大值来确定投影方向 a，组内方差与组间方差的比值如下：

$$\frac{a'Ba}{a'Wa}=\frac{a'\left[\sum_{k=1}^{g}\left(\overline{y}_k-\overline{y}\right)\left(\overline{y}_k-\overline{y}\right)'\right]a}{a'\left[\sum_{k=1}^{g}\sum_{i=1}^{n_k}\left(y_{k_i}-\overline{y}_k\right)\left(y_{k_i}-\overline{y}_k\right)'\right]a} \tag{2}$$

其中 B 为组间协方差矩阵，W 为组内协方差矩阵之和。

设矩阵 $W^{-1}B$ 有非零特征值 $\lambda_1,\lambda_2,\cdots,\lambda_s$，每个特征值对应的特征向量为 $e_1,e_2,\cdots,e_s$，其中 $s\le\min\left(g-1,p\right)$。

线性组合 $e'_k y$ 为第 k 个判别式。将 s 个判别式作为判别函数，可以将原有高维数据转换到 s 维空间，实现了降维的目的。若要判别一个未知的个体属于哪类群体，只需计算该个体到各类群体中心的马氏距离，马氏距离最短的中心对应的类别，即为该未知个体所属的群体类别。

LDA 判别法对于群体分布没有要求但需要群体有共同的协方差矩阵，比一般的 Fisher 判别法具有更广泛的适用性。作为一种线性判别分析，LDA 判别法从方差分析的思想入手，通过降维实现数据可视化，为其他判别函数的构造提供了启发。

2.2. Fisher 算法的缺陷

本文主要对 LDA 算法的以下两个缺陷进行研究。

1) 异方差性带来的误差问题：Fisher 方法从方差分析的角度切入，计算组内方差与组间方差的比值，从而得到投影方向，只有在多组群体具有相同的协方差矩阵的时候才可以得到使用，具有较大的局限性；其次，由于方差主要通过反映了数据的离散程度来间接反映数据的集中程度，该方法的准确率并不算很高。在很多现实的工程问题中，需要处理的数据群体并不满足 LDA 方法的使用条件。

2) 同类别多中心和多类别同中心的问题：Fisher 方法无法处理同类别多中心问题和多类别同中心。有些数据虽然有明显的划分界限，但是因为某一类的数据中心不止一个，有时各类的中心甚至都落到了对方的区域，产生了多类别同中心的问题，导致无法与其他类之间很好地分开。如图 1 (图中浅蓝色圆点为每个黑色簇的重心，深蓝色圆点为每个红色簇的重心，浅蓝色方点为黑色簇的总重心，粉色方点为红色簇的总重心)，虽然 Fisher 判别法确定了这两类的数据中心，但是这两类的数据中心均落到对方的数据中，并几乎与对方其中一个数据簇的中心重合，而且由于同类别多中心的问题，虽然划分依据使得两类中心的距离最大，但是并没有很好的将两类区分，误差较大。

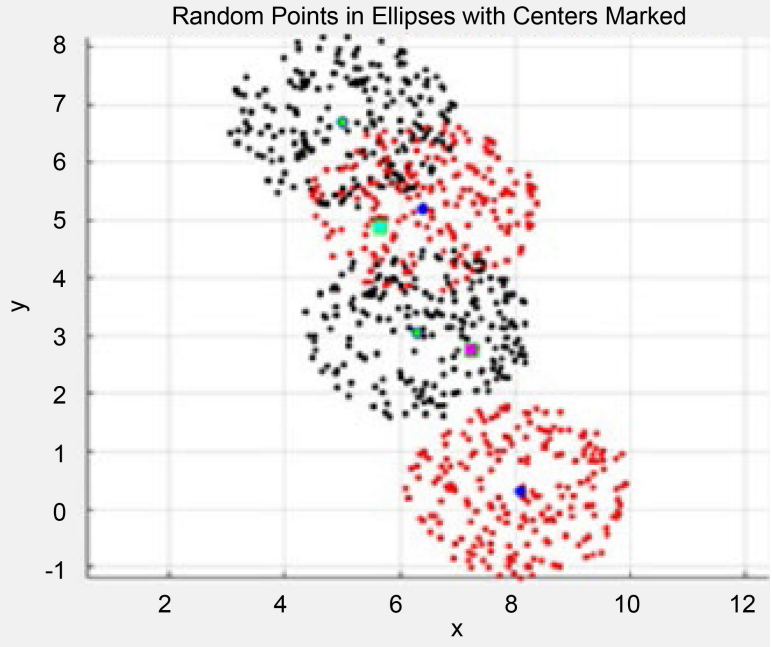


Figure 1. Schematic diagram of multi-centers in the same category
图 1. 多中心同类别示意图

3. LMMR 算法

在计算机科学领域，重映射是指通过一系列规则，将数据从一个结构映射到另一个结构的过程。重映射通常被用来处理数据的转换和格式化。例如：在计算机网络中，通过重映射功能可以把数据从一种协议转换为另一种协议。重映射广泛应用于文件转换、网络通信、图形处理和数据库管理等领域。本文借助重映射思想，将多维数据转化为判别空间，构建了重映射模型。

3.1. 重映射算法

定义 1：伪预测(第一次映射)

设 $x_1, x_2, \cdots, x_n$ 是相互独立的观测，每个观测有 p 维属性，$a = \left(a_1, a_2, \cdots, a_p\right)$ 为系数向量，h 为伪预测(第一次映射)。其函数关系为：

$$\begin{aligned} &h\left(x_j\right) = a' x_j = a_1 x_{j1} + a_2 x_{j2} + \cdots + a_p x_{jp}, j = 1, 2, \cdots, n \\ &a_1 + a_2 + \cdots + a_p = 1, a_i \in [-1, 1], i = 1, 2, \cdots, p \end{aligned} \tag{3}$$

其中 $x_j = \left(x_{j1}, x_{j2}, \cdots, x_{jp}\right), j = 1, 2, \cdots, n$。初始待定系数任意给定，且将该函数表达式的系数做归一化处理，即满足：$a_1 + a_2 + \cdots + a_n = 1$。考虑到系数的重要性及方向性、归一性，限制系数 $a_1, a_2, \cdots, a_n$ 取值范围为[−1, 1]。由于常数项系数对第二次映射无影响，所以不妨忽略常数项系数。这样可以有效将多维线性空间通过线性映射的方法映射至一维线性空间。

定义 2：实映射(第二次映射)

定义实映射 f，表达式如下：

$$f \circ h\left(x_j\right) = \sum_{i=1}^{k} c_i I_{A_i}\left(h\left(x_j\right)\right) \tag{4}$$

其中 A_i 为数轴上划分的第 i 块区间。

通过伪预测和实映射的两次映射，构建重映射模型，以二分类为例，设向量 $c=(c_1,c_2,\cdots,c_k)$，$b=(b_1,b_2,\cdots,b_{k-1})$ 为实数轴上的划分值对应的向量，且 $-\infty=b_0<b_1<b_2<\cdots<b_{k-1}<b_k=+\infty$，$A_i=(b_{i-1},b_i]\cap R$，$h(\cdot)$ 为伪预测的函数，$f(\cdot)$ 为实映射的函数，目标函数 $g(a,b,c)$ 为实映射预测的属性值与实际的属性值差的绝对值之和，使之达到最小的预测方法，称之为 LMMR 模型，具体的最优化模型为：

$$\min g(a,b,c)=\sum_{j=1}^{n}\left|f\circ h(x_j)-y_j\right| \tag{5}$$

其中 $c_i=0$ 或 1，$i=1,2,\cdots,k$，y_j 表示 x_j，$j=1,2,\cdots,n$ 对应的实际属性值。流程图和具体过程见图 2、图 3 所示。

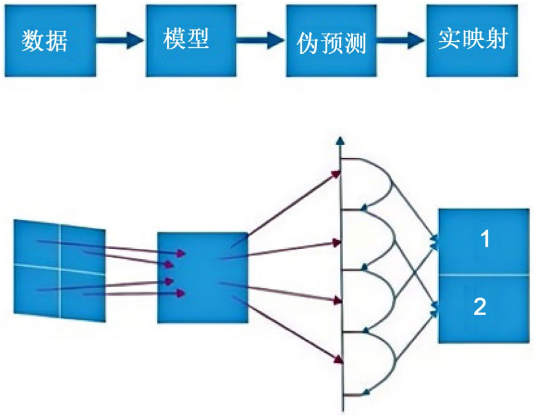
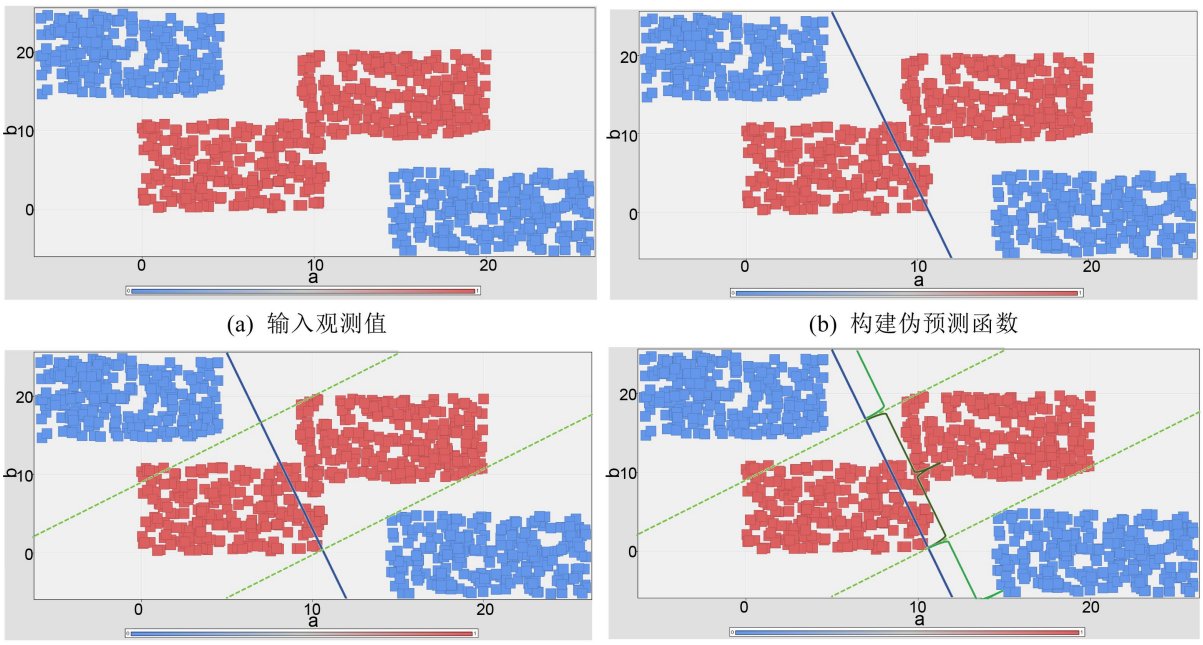


Figure 2. Flowchart of the LMMR algorithm
图 2. LMMR 算法流程图



(a) 输入观测值　(b) 构建伪预测函数

(c) 计算划分值　(d) 实映射

Figure 3. The specific process of the LMMR algorithm
图 3. LMMR 算法的具体过程

3.2. LMMR 算法的具体步骤

基于上述分析，LMMR 的算法的具体步骤如下：

输入数据集 $X=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$ (其中 x_i 为观测向量，y_i 为类别标识)，最大分割数 nd，随机次数 m 输出识别率，分割点最小值，划分点集，分割点最大值，回归参数。

$$\{pun,y_{\min},\{d_1,d_2,\cdots,d_{nd}\},y_{\max},a_1,a_2,\cdots,a_n\}$$

(1) for(i in 1: m){

(2) 用均匀随机数发生器在区间[−1, 1]上生成 $p-1$ 维个的随机数 $a_1,a_2,\cdots,a_{p-1}$，由系数归一性构成系数向量 $a=(a_1,a_2,\cdots,a_p)$。

(3) 将系数 a 和数据集 X 的观测参数值代入伪预测函数 $h(x_j),j=1,2,\cdots,n$。

(4) 计算

$$y_{\min}=\min_{j=1,2,\cdots,n}h(x_j),y_{\max}=\max_{j=1,2,\cdots,n}h(x_j)$$

(5) 用均匀随机数发生器在区间 $[d_0,d_{nd+1}]=[y_{\min},y_{\max}]$ 上按均匀分布随机生成 nd 个分割值：$d_1,d_2,\cdots,d_{nd}$。

(6) do {

(7) for(k in 1: nd){

(8) for(j in 1: nd+1){

(9) 计算在区间 $[d_{j-1},d_j]$ 上各属性个体的个数

(10) 取该区间上个数最多者的属性 c_i 为该区间的属性

(11) 计算在区间上具有属性 c_i 的个体的个数 nn_j

(12) 计算识别率

$$punT=\frac{\sum_{j=1}^{nd+1}nn_j}{length(X)}$$

(13) }

(14) 用梯度下降法在 $[d_{k-1},d_{k+1}]$ 间选取 d_k 使得 $punT$ 识别率最高，记为 pun_k

(15) }

(16)}while (分割值 $d_1,d_2,\cdots,d_{nd}$ 位置发生改变)

(17)计算识别率 $pun=\max_k pun_k$

(18)}

(19)取上述 m 个识别率中最大的识别率 pun 对应的

$$\{pun,y_{\min},\{d_1,d_2,\cdots,d_{nd}\},y_{\max},a_1,a_2,\cdots,a_n\}$$

为最终结果。

4. 实验结果与分析

4.1. 评价指标

为了验证本文算法的有效性，分别对合成数据集和真实数据集进行探索实验，并且用 Fisher 算法和 BP 神经网络算法进行对比。所有算法均采用 R4.1.0 编写程序并且测试，选择 Windows 10 系统作为实验

环境，选择 Intel(R) Core(TM) i5-7200U CPU 处理器，内存为 8.00 GB。

本文分别用蒙特卡洛法和迭代法求伪预测函数系数和划分点值的精确数值解，伪预测函数系数为近似全局最优，划分点值为近似局部最优，其中划分点迭代值的初值为使用蒙特卡洛法得到的近似最优解。

本文采用判别分析算法中最重要的识别率作为模型优劣的评判标准，LMMR 算法通过预测的结果与实际的类别相对比，借助目标函数 $g(a,b,c)$ 得到识别率 R 的计算公式为：

$$R=1-\frac{g(a,b,c)}{N} \tag{6}$$

或者根据 LMMR 具体的算法过程，也可以从分布函数的角度出发，得出识别率的等价计算方法：

定义 4：(分布函数)设 N_i 为类别 i 的个体占数据集的总数，$count_i(y)$ 为在重映射过后，重映射值小于 y 且属于类别 i 的个数，则类别 i 的分布函数的表达式为：

$$F_i(y)=P_i(Y\le y)=\frac{count_i(y)}{N_i} \tag{7}$$

若一共有 $k+1$ 个分割点(包括 $y_{\min},y_{\max}$)，d_j,d_{j+1} 分别为第 j 个和第 $j+1$ 分割点，$N_{j,j+1}$ 为重映射过后，y 介于 d_j,d_{j+1} 中的样本容量，N 为样本总数，则识别率 R 可表达为：

$$R=\frac{\sum_{j=1}^{k}\max\left\{N_{j,j+1}\left(F_i\left(d_j\right)-F_i\left(d_{j-1}\right)\right)\right\}}{2\sum_{i=1}^{n}N_i} \tag{8}$$

4.2. 合成数据集

为了将 LMMR 算法与传统的 Fisher’s LDA、神经网络算法在不同形状簇的数据集上的分类效果对比，本文选取了 3 个具有代表性的二维合成数据集(分别记为 I，II，III)进行对比实验。

Table 1. The synthetic dataset used in the experiment
表 1. 实验中使用的合成数据集

数据集	类别数	α 类重心数	β 类重心数	维度
I	2	1	1	2
II	2	2	2	2
III	2	3	3	2

合成数据集的基本信息如表 1 所示，散点图如图 4 所示。图 4 中不同颜色的区域表示不同的类别，其中，红色点代表 α 类，黑色代表 β 类，蓝色、绿色的圈分别为各区域的重心，黑色的叉、粉色的圈分别为类别 α 和类别 β 的重心。

由图 4 知，传统的 Fisher’s LDA 算法虽然可以识别出每一类的重心，但是对于多重心的问题，每一类的重心可能近似重合，甚至落在了对方类别的区域，由于 Fisher’s LDA 算法能且只能识别每一类的重心，并把各类的重心尽可能地分开，却无法识别每一簇块的中心，导致分割方式与实际情况差别过大，误差偏大。而 LMMR 算法，首先不是通过各簇重心决定分类方式，而是直接以识别率作为目标函数，并且 LMMR 算法允许多次分块。所以是否多重心对它并无影响，可以将互相交错的多簇数据集很好地分类，如图 4(b)、图 4(c)所示。

表 2、表 3 列出了 Fisher’s LDA、LMMR、BP 神经网络 3 种算法在 3 个人工数据集上的分类结果，

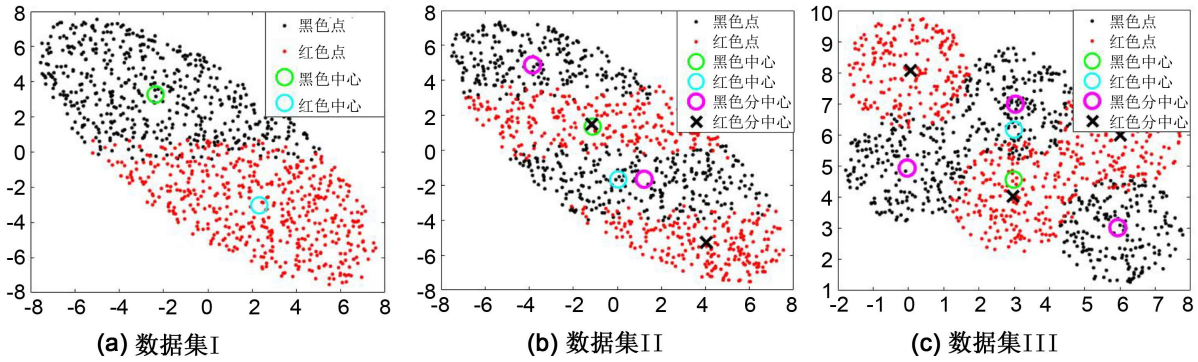


Figure 4. Scatterplot of three 2D synthetic datasets
图 4. 三个二维合成数据集散点图

Table 2. Comparison of recognition rates of three algorithms on synthetic datasets
表 2. 三种算法在合成数据集上识别率的对比

算法	I 识别率	II 识别率	III 识别率
LDA	0.951	0.551	0.6575
LMMR	0.9505	0.9125	0.8063
BP 神经网络	0.955	0.922	0.9275

Table 3. LMMR algorithm in synthetic dataset discriminant result parameters
表 3. LMMR 算法在合成数据集判别结果参数

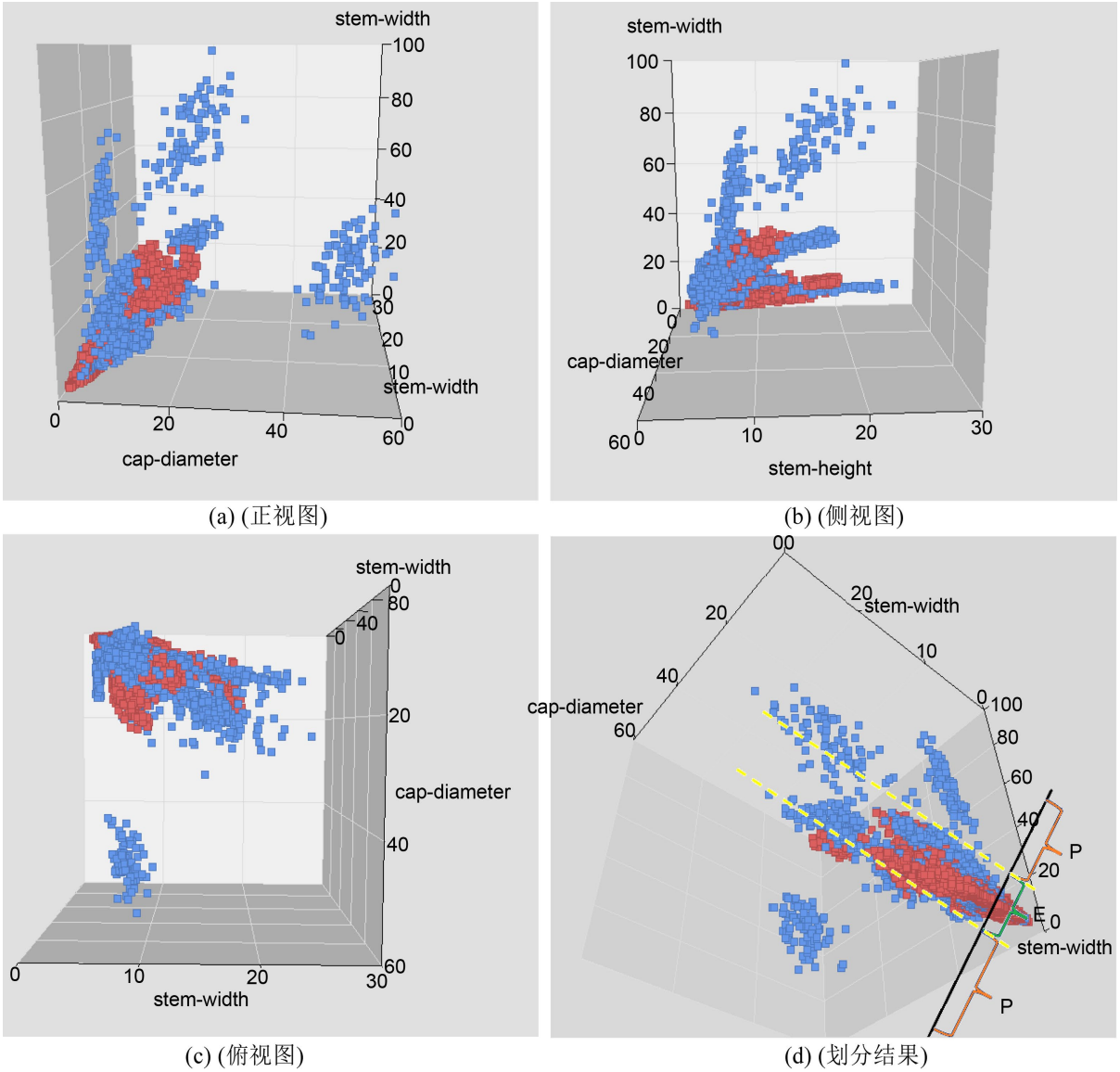
数据集	a_1	a_2	划分值
I	0.064	0.936	{0.185, 6.744}
II	−0.082	1.082	{−3.723, −0.031, −3.031}
III	−0.926	1.926	{1.903, 7.162, 13.663}

可以看出 Fisher's LDA 在多重心的人工数据集上，判别分类结果很差，另外两个算法判别结果都很好，LMMR 算法略逊于 BP 神经网络，但是 LMMR 作为由传统统计学知识构建出的判别分析模型，它的可解释性强，原理清晰，而神经网络仅有判别结果和公式却无法解释模型内在原理。对比各算法的识别率 R 可以发现，LMMR 算法在各个数据集上都有着较好的分类效果。

4.3. 真实数据集

为了验证本文 LMMR 算法在现实数据上的有效性，在 Mushroom data creation [15]真实数据集上与 Fisher' LDA、LMMR 算法以及 BP 神经网络算法进行对比实验，该数据集的基本信息如表 4 所示，其中 e、p 分别表示无毒蘑菇和有毒蘑菇。由于该数据为三维数据，用正视图、侧视图、俯视图对该数据集进行可视化，散点图(正视图、侧视图、俯视图)如图 5 所示，从中可以看出：该数据集是个三中心的数据集(图 5(d)所示)。

表 5 列出了 LDA、LMMR、BP 神经网络这 3 种算法在 Mushroom data creation 数据集中训练集和测试集的分类结果。对于该数据集 3 种算法的分类效果都不是很好，主要是因为该数据集 e 和 p 两类的数据交杂在一起。虽然能大致看出应分成 3 块，但 e 中仍然夹杂了很多 p 类的蘑菇。但通过训练集和测试集的比较，LMMR、BP 神经网络这两种算法相对而言分类效果较好。首先，其训练集识别率相比 LDA



(a) (正视图) (b) (侧视图)

(c) (俯视图) (d) (划分结果)

Figure 5. Scatterplot of the Mushroom data creation dataset
图 5. Mushroom data creation 数据集的散点图

Table 4. Basic information about the Mushroom data creation
表 4. Mushroom data creation 数据集的基本信息

类别	实例数	维数	均值	方差
e	19,028	3	(8.13, 6.8, 15)	(48.7, 10.2, 141)
p	23,771	3	(5.54, 6.3, 9)	(12.7, 8.2, 43.3)

略高。其次，一般而言，测试集的识别率会有所下降，而 LDA 的测试集识别率 *R* 相比训练集竟然下降了近 11%。相比之下，主要是因为 LDA 无法准确地将该数据集分化成 3 类，导致其模型很不稳定；反观 LMMR、BP 神经网络两种算法的测试集识别率仅下降 4%~5%。说明 LMMR 算法不仅拥有接近人工智能算法的识别率，且模型较为稳定，更有传统统计学算法模型易于解释的优点。

Table 5. Comparison of results of three algorithms on Mushroom data creation
表 5. 三种算法在 Mushroom data creation 数据集上结果的对比

算法	LDA	LMMR	BP 神经网络
训练集识别率	0.65	0.68	0.70
测试集识别率	0.54	0.63	0.64
a	(−0.03, −0.13, −0.11)	(−0.46, −0.86, 2.31)	/
划分值	{−0.736}	{20.8, 67.5, 260}	/

4.4. 灵敏度分析

LMMR 算法只有一个参数设置，也就是伪预测函数系数向量 a。为了分析该参数对算法的影响，分别在合成数据集 II 和真实数据集 Mushroom data creation 上进行实验。对于真实数据集，我们将参数 a 的每一个分量，对于合成数据集 II，我们将参数 a 的每一个分量，都做相对于均值 5‰的波动，重复 10 次实验，得出参数 a 的波动对识别率的影响，并画出散点图。

Table 6. Recognition rate of LMMR algorithm datasets after ten perturbations
表 6. 十次扰动后 LMMR 算法数据集的识别率

实验	II	真实数据集
1	0.80791	0.67257
2	0.802083	0.673204
3	0.782916	0.672270
4	0.799583	0.672831
5	0.80125	0.67245
6	0.67264	0.80041
7	0.671779	0.793575
8	0.671546	0.79375
9	0.672527	0.790416
10	0.672527	0.790416
均值	0.6724	0.7956
方差	2.43×10^{-7}	6.85×10^{-5}
相对误差均值	0.9899	0.9868
相对误差方差	5.28×10^{-7}	0.000105

由表 6、图 6 和相对误差可以看出。识别率的波动均在 3%以下，但相对 0.5%的参数扰动，识别率的波动仍有些偏大，说明该模型对伪预测的系数(参数 a)较为灵敏。原因在于，参数 a 是伪预测函数的系数，决定了第一次映射的方向，参数只是发生细微的变化，也可以让第一次映射的角度变化较大，导致对映射值无法进行很好的切割，造成了识别率的快速下降。但识别率的波动也仅有 3%，误差并没有太大，仍在可接受范围之内，说明我们的模型相对稳定。

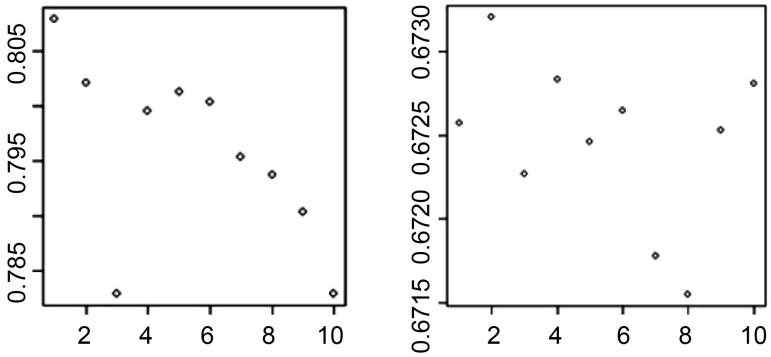


Figure 6. Plot of the fluctuation of the recognition rate of the LMMR algorithm for 10 perturbations of the *a*-value
图 6. *a* 值扰动 10 次 LMMR 算法识别率波动图

5. 结束语

本文针对 Fisher's LDA 算法中处理多中心问题时的误差问题，提出了基于线性映射的重编码的判别分析 LMMR 算法。实验结果表明，与 Fisher's LDA 相比，本文提出的 LMMR 算法具有更好的分类效果。但是在该算法中，对伪预测函数系数很敏感，并且该模型的伪预测也是一维线性映射，而某些多中心的分类问题必须通过高维非线性映射才能解决，否则误差会较大。因此，下一步要研究如何降低模型对伪预测函数系数的敏感性，以及如何基于现在的模型，构造出将伪预测推广到非线性函数，并且进行高维映射的重映射判别分析算法。

参考文献

[1] 刘红岩, 陈剑, 陈国青. 数据挖掘中的数据分类算法综述[J]. 清华大学学报(自然科学版), 2002, 42(6): 727-730.

[2] 李红军, 赵明莉, 母方欣. 基于子空间半监督学习线性判别方法的目标跟踪技术研究[J]. 现代电子技术, 2019, 42(3): 52-55, 50.

[3] 田絮资. 区域地貌中的聚类分析及贝叶斯判别方法[J]. 西北农林科技大学学报(自然科学版), 2003, 31(5): 178-182.

[4] 廖东平, 王书宏, 黎湘. 一种结合 K 近邻法的改进的渐进直推式支持向量机学习算法[J]. 电光与控制, 2010, 17(10): 6-9.

[5] 杨淑娥, 黄礼. 基于 BP 神经网络的上市公司财务预警模型[J]. 系统工程理论与实践, 2005, 25(1): 12-18.

[6] 郭景峰, 米浦波, 刘国华. 决策树算法的并行性研究[J]. 计算机工程, 2002, 28(8): 77-78.

[7] 曹正凤. 随机森林算法优化研究[D]: [博士学位论文]. 北京: 首都经济贸易大学, 2014.

[8] 黎夏, 叶嘉安. 基于神经网络的元胞自动机及模拟复杂土地利用系统[J]. 地理研究, 2005, 24(1): 19-27.

[9] 吉根林. 遗传算法研究综述[J]. 计算机应用与软件, 2004, 21(2): 2911-2916.

[10] 魏平, 徐成贤, 段成德. 全局智能优化集成算法研究[J]. 西安交通大学学报, 2009, 43(12): 60-64.

[11] 刘颖, 穆志纯, 袁立. 基于核函数的 Fisher 判别分析算法在人耳识别中的应用[J]. 微计算机信息, 2006, 22(22): 304-306.

[12] 刘小平, 黎夏. Fisher 判别及自动获取元胞自动机的转换规则[J]. 测绘学报, 2007, 36(1): 112-118.

[13] 郑建峰, 唐建, 郭立. 一种基于快速重编码的 H.264/AVC 压缩域信息隐藏方案[J]. 中国科学技术大学学报, 2013, 43(1): 35-41.

[14] 吴诚鸥, 秦伟良. 近代实用多元统计分析[M]. 北京: 气象出版社, 2007.

[15] Wagner, D., Heider, D. and Hattab, G. (2021) Mushroom Data Creation, Curation, and Simulation to Support Classification Tasks. *Scientific Reports*, **11**, Article No. 8134. https://doi.org/10.1038/s41598-021-87602-3