

# 手写数字识别

机器学习期末项目汇报

陈 纪程

2020 级 | 数据科学与大数据

## 一. 任务意义与背景

### 1. 数字识别到图像识别

图像识别技术是一门应用广泛,实现方式多样的人工智能技术。在国防科技、农业监控、自动驾驶、医学智能等诸多领域具有广泛和深刻的应用,是人工智能商业化较为成熟的领域。

而手写数字识别,基本上是图像识别的最最基础的部分,其数据规模小,数据特征显著,实现形式多样,且完全具备在纯粹机器学习模块实现的可能。数字识别同时也是文字识别的简化版本在实现的工程量大小和复杂水平上说,手写汉字识别 > 字母识别 > 数字识别,但是三种文字识别的数据基本结构都是相似的,在进行图像处理时,均可以转化为黑白图进行处理,图像数据的维度可以大幅度降低。并且,一定程度上说,三种规模的模式识别都具备等同于标签数量的识别模板,对于手写数字来说,熟悉打印的数字图像,一定程度上就能构成对手写数字图片的理解,并且以拟合的角度来讲,每一个手写的数字图像总是可以抽象为直线和曲线的组合框架,也就是说,可以通过像素点的集中拟合到某个识别模板上去。

图像数据不同于一般意义的数据,其高维的特征使得大部分机器学习算法不能直接地应用于原始数据集上。本次实验力求保全机器学习的基础性、解释性,通过实操探究和学习算法模型的原理性,因而在小样本,小规模的数据集上,对每个图像进行一维展开,直接应用于训练集中,并从对各个标签类的预测准确度,召回率上,可以看到某种特定模型对于不同数字识别的预测能力。

### 2. 数据源

本次实验的竞赛赛事选自 Data Castle 数据竞赛平台,数据链接为:

<https://challenge.datacastle.cn/v3/cmptDetail.html?id=358>,这是一次基础性的赛事,笔者作为团队(团队名 match)的唯一作者参赛,最终模型(集成模型)的预测准确率为 0.9831,在报告的最后会就一些误差图例来看一看不同子模型的假阳率主要出现在哪些数字上。

## 任务意义 与背景

### 1.数字识别到图

### 像识别

### 2.数据源

## 二. 方法调研与问题提出

### 1. 图像识别技术的实现形式

图像识别的难点和重点在于图像数据的特征化，特征化之后的数据，一般是一维的，从而可以应用于广泛的机器学习模型。因此，大部分的算法设计都围绕着图像的特征提取和降维方法来展开。具体可分为以下三种：

#### (1) 模式识别

基于模式识别的特种工程需要以计算机技术为依托，利用数学逻辑运算对图像的形状、字符、格式、曲线等各个特征进行信息评价，最后完成图像识别。模式识别可以依托一些规则有效地降低噪声干扰，也可以经由模板和目标的比对获取分类预测标签。

#### (2) 神经网络图像识别

神经网络在图像识别领域的应用，如 BP 神经网络，CNN 卷积神经网络等，经过这些年来机器学习的发展，某些神经网络已经可以处理一些不经过 transform 或映射降维的图像数据。比如，在卷积神经网络中，图像数据以网格结构进行输入；近期的一些研究表明，GNN 应用于图像识别的经典任务，图片可以被表示为图结构数据的节点。<sup>1</sup>

#### (3) 非线性降维识别技术形式

一般图像识别常用的，是线性降维的形式，线性降维的最突出优势是理解功能，该项技术是对整体的数据集合开展处理，获得的为最优低维度。<sup>2</sup>可以理解为经过线性映射后保留了多数高维特征的算法，主要分为主成分分析（PCA）和线性奇异分析（LDA）两类。线性降维的优点很多，但缺点是算法时间和空间资源消耗大。而相较于此，不必保证“齐次性”和“可加性”的非线性降维，希望在一定程度上低维空间中保持高维结构，更加简单和快捷。

### 2. 算法介绍

本次实验不使用线性降维，但此处为突出图像特征工程的特殊地位，介绍 PCA 和 LDA 两种线性降维方法。需要说明的是，降维算法不仅仅用在模型训练前，在图像数据的处理中，将一些适应于特定特征维度的模型应用于原始维度不同的图片时，同样需要使用数据降维的方法进行预处理后才能进行预测。

#### (1) PCA 主成分分析

## 方法调研 与问题提 出

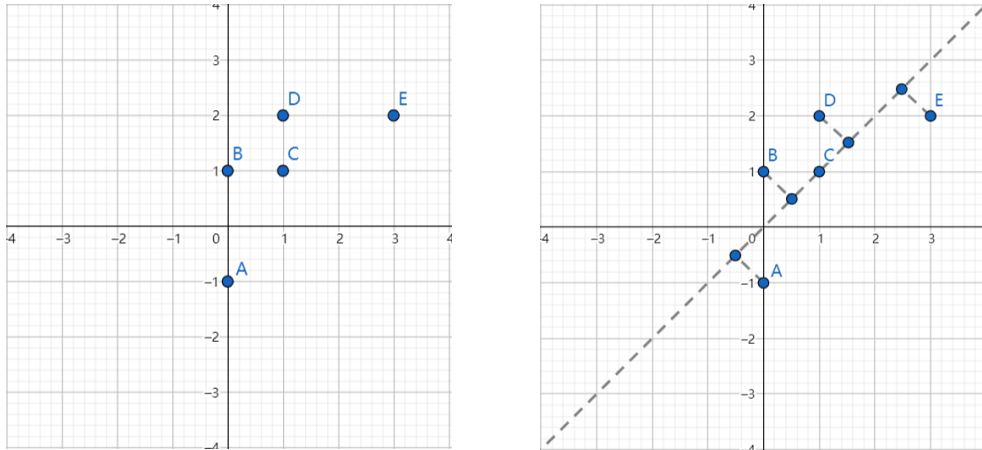
### 1. 图像识别技术 的实现形式

### 2. 算法介绍

### 3. 本次实验所探 究的问题与实 现思路

<sup>1</sup> <https://zhuanlan.zhihu.com/p/539384371>

<sup>2</sup> 藺伟,张驰.人工智能背景下图像识别技术研究[J].无线互联科技,2022,19(14):108-110.



如上图<sup>3</sup>所示，PCA 算法的基本思想是基于最大方差理论，将高维数据投影到一个超平面上，使得各投影点尽量分开，用方差来刻画这一可分性。对于一个高维数据，比如  $n \times n$  的  $X$  阵，要将其映射到  $n \times m$  的  $y$  上，PCA 的核心步骤是基于对去中心化的  $X$  的协方差矩阵所做的特征值分解，而选取的主成分即按前  $m$  大的特征值大小排序的特征向量阵。

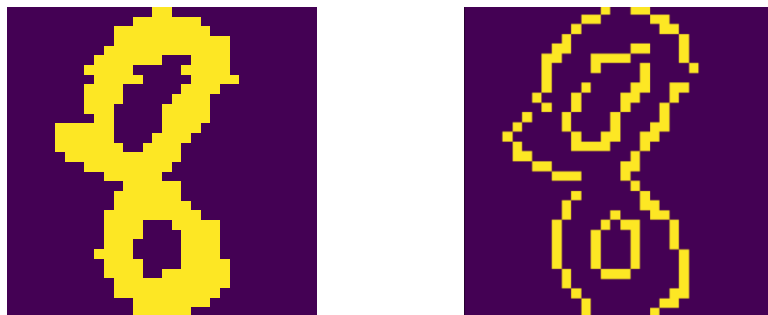
## （2） LDA 线性判别分析

LDA 与 PCA 最大的区别在于 LDA 运用到了标签的信息，对多分类任务相对友好，处理后的数据点具有“同类数据点尽量靠近，方差最小；异类数据中心点尽量分散，方差最大”的特征。因而 LDA 是一类有监督降维算法。

## 3. 本次实验所探究的问题与实现思路

本次实验中，先将二维图像数据简单地拉直成一维向量，这并非线性降维算法，但极大地保留了数据的完整性。

实验希望探索模型对数据特征的学习程度以及泛化能力。问题的源头来自于人类对于数字辨识基本不受数字的填充色影响，比如下图：



两图应该能够被人类清晰地辨认出为同一数字。但是在机器学习模型训练中则很可能由于信息的损失而产生不同性质的判断。因此，本次实验想要探究，一个全部基于图像边缘信息进行训练的模型和一个基于图片全部信息的模型间的预测差异，同时对比模型的不同是否会缩小这种差异。

<sup>3</sup> 图片来源：

[https://mp.weixin.qq.com/s?\\_\\_biz=MzA5ODEzODI2Mg==&mid=2660719181&idx=5&sn=67d0383a9538f9090e1783578be206d9&chksm=8bf5dbb4bc8252a2c5a8250919e1b1ca72574bfaa61330a3d851070bc39987666e41eaeef94b&scene=27](https://mp.weixin.qq.com/s?__biz=MzA5ODEzODI2Mg==&mid=2660719181&idx=5&sn=67d0383a9538f9090e1783578be206d9&chksm=8bf5dbb4bc8252a2c5a8250919e1b1ca72574bfaa61330a3d851070bc39987666e41eaeef94b&scene=27)

因此，实验将在逻辑回归和支持向量机模型中对边缘化前后的数据进行分别训练，对比它们的预测精确度。除此之外，实验特别关注不同模型对于不同数字预测的查准率，并尤其关注对某些数字的预测“不够精确”所带来的后果。

随后的模型中，采取更优的数据集，大都直接在数据全集上进行拟合。因为竞赛所给的测试数据是不含标签的，实验对这些模型的评估，采取了与其他模型预测值的平均差异数的方式，进行比对。

## 三. 实验过程

### 1. 数据转换

竞赛所给原始手写数字数据，均为 2KB 大小， $32 \times 32$  数据量的 txt 文件，为了更好地展示和体现图像数据的特征，因而对用于训练和用于测试的 txt 文件，全部转化为图片文件，储存在 my\_tab 文件夹中。

### 2. 边缘特征提取方法

OpenCV 库对边缘特征的提取方法是基于 Canny 算法的，这一算法依赖于图像梯度的计算，OpenCV 库中，使用 Sobel 核在水平和垂直方向上对平滑的图像进行滤波，以在水平方向和垂直方向上获得一阶导数。在图像边缘图像梯度一定不为 0，但对于彩色图来说，完整图像的內部也有可能出現梯度不为 0 的情况。在 OpenCV 中，利用非局部极大值抑制和磁滞阈值，可以使得算法区分强边缘和过滤弱边缘，所得的图像是仅含有单像素点细线的边缘图像。

值得说明的是，在本次实验中还有一段对于轮廓特征提取尝试，正是基于边缘数据的。在图像处理中，边缘和轮廓是两个概念，轮廓是点集，而边缘却是图像。基于轮廓信息可以计算一些对数字图片较为重要的特征，比如轮廓凸包、特征矩、极点等，还可以基于轮廓特征求轮廓近似，对于采取模式识别的图像识别技术，这一方法无疑是重要的。

#### 不被使用的轮廓特征代码废案

```
ret, thresh = cv2.threshold(edge, 127, 255, 0) #图像像素大于127为255, 否则为0
contours, hierarchy = cv2.findContours(thresh, cv2.RETR_TREE, cv2.CHAIN_APPROX_SIMPLE)
cv2.drawContours(img, contours, -1, (0, 255, 0), 1)
```

特征矩则是重要的图片特征，对于手写数字图像来说，可以基于质心与特征矩对图像做偏移矫正，对于图像歪斜严重影响预测精确度的情况，这一特征工程是重要的。

### 3. 参数调优方法

Sklearn 模块内嵌 Grid\_SearchCV 方法，可以基本实现大多数 Sklearn 内部机器学习模型的参数调优，并集成了交叉验证的方法。在本次实验中具体有逻辑回归，k 近邻，随机森林和神经网络经过了参数调优。注意到，在不同的训练集上，调优的最优参数可能有所不同，但对于同一数据集内的训练集，最优参数大致的分布一定是相近的。

尤其对于神经网络，隐藏层的层数和神经元个数都相对难以确定的时候，应用 Grid\_SearchCV 允许我们去细调，执行下述格式的代码：

## 实验过程

### 1. 数据转换

### 2. 边缘特征提取

### 方法

### 3. 参数调优方法

### 4. 模型的简单评

### 估

```
MLPC_para = {'hidden_layer_sizes':[(i,j) for i in range(160,200,20) for j in range(70,90,10)]}

[119] ✓ 0.5s

+ 代码 + Markdown

mlp_grid = GridSearchCV(MLPClassifier(),MLPC_para,cv = 5,refit=True)
mlp_grid.fit(X_train,y_train)
params = mlp_grid.best_estimator_.get_params()
print(params['hidden_layer_sizes'])

[120] ✓ 35.9s

... (160, 80)
```

可以一步步地实现两层隐藏层下最优神经元个数的确定。其基本思路是，控制 i、j 的上确界和下确界，当最优参数的 i 偏小时，就放松下界，收紧上界；当输出的最优参数 i 偏大时，就放松给定参数范围内 i 的上界，收紧下界，最终达到的最优参数在上确界和下确界之间时，就完成了调参的过程。

4. 模型的简单评估

除了最终模型预测的准确率经竞赛通道的提交取得外，其他模型的评估采取了一系列的方法。

(1) 在训练总集上进一步划分数据，基于有标签的测试子集对边缘化前后的数据进行多维评估

实验在逻辑回归和支持向量机，以及神经网络上实现了这一评估。右图是对在两个数据集上的逻辑回归模型预测精确率（前者为仅在图像边缘数据上进行预测的结果），召回率和 f1 值的综合评估。可以发现，基于图片边缘信息的预测性能各方面都逊于基于原始图像数据的预测，同时，model1 对于“1,8,9”三个数字的预测效果较差，从整体上而言，线性模型对于 1 的预测水平（无论基于哪种数据集），都是比较不好的。

在手写数字的图像集中，有的“1”上面带钩，有的“1”被写得很胖，有的“1”还有一点弯曲，这些对于线性模型来讲，可能是使得预测变得困难的地方。

(2) 仅基于不同数据集的模型预测差异

	precision	recall	f1-score	support
0	0.93	0.90	0.92	79
1	0.75	0.81	0.78	73
2	0.90	0.86	0.88	81
3	0.84	0.86	0.85	78
4	0.82	0.92	0.87	66
5	0.85	0.91	0.88	70
6	0.95	0.97	0.96	76
7	0.93	0.89	0.91	83
8	0.74	0.65	0.69	81
9	0.79	0.75	0.77	87
accuracy			0.85	774
macro avg	0.85	0.85	0.85	774
weighted avg	0.85	0.85	0.85	774

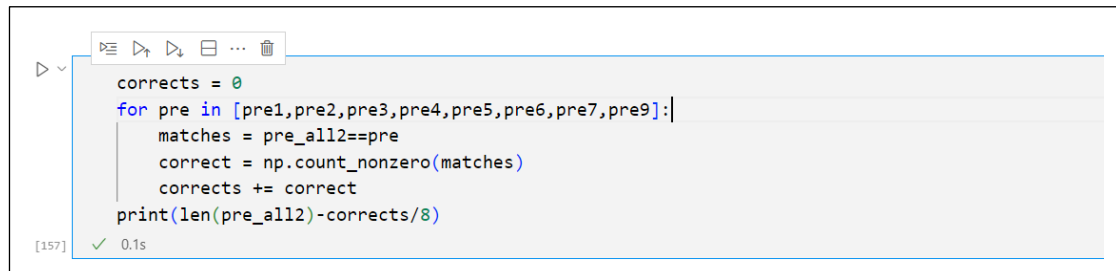
	precision	recall	f1-score	support
0	1.00	1.00	1.00	76
1	0.92	0.91	0.92	80
2	0.92	0.96	0.94	75
3	0.99	1.00	0.99	79
4	0.93	1.00	0.97	69
5	0.97	0.99	0.98	74
6	0.99	1.00	0.99	77
7	1.00	0.99	0.99	81
8	0.96	0.88	0.92	78
9	0.95	0.92	0.93	85
accuracy			0.96	774
macro avg	0.96	0.96	0.96	774
weighted avg	0.96	0.96	0.96	774

由于人对数字的辨识不会随着轮廓内部是否有颜色填充而丧失精度，所以在本次实验中，我们期望某种模型算法的泛化性能越强时，基于两种特征的数据集进行训练的模型，在测试集上体现的差异就越小，在本此实验中，以 difference 来表示这种评估。

在实验中，分别对线性模型、支持向量机以及 stacking 异质集成模型进行了 difference 的计算，三个值分别为 100，74 和 70，这说明相对而言，stacking 模型的泛化性能最好，能够对于仅含有图像边缘信息的数据做出较好的预测，其次是支持向量机，线性模型最差。这一定程度上体现了算法本身对这一分类任务“真相”的获取能力强弱。

### (3) 测试集无标签下的模型评估指标

在 k 近邻，随机森林，梯度优化决策树，神经网络，集成模型训练时，总是基于前几个模型的预测结果，计算 corrects 值，如在集成算法中 corrects 值的算法<sup>4</sup>如下：



```
corrects = 0
for pre in [pre1,pre2,pre3,pre4,pre5,pre6,pre7,pre9]:
    matches = pre_all2==pre
    correct = np.count_nonzero(matches)
    corrects += correct
print(len(pre_all2)-corrects/8)
```

[157] ✓ 0.1s

最终输出的是预测总数减去平均意义下的预测“正确”数，即平均意义下的预测错误数。在已有的评估指标中，平均意义下的预测错误数呈现为：

Stacking 集成 < 神经网络 (mlp2) < 随机森林 < 梯度优化决策树 < k 近邻

这一定程度上说明了复杂算法模型在处理复杂特征数据时的优势。

---

<sup>4</sup> 由于第 8 个模型（神经网络 1，即 mlp1）的性能过差，在此它的预测结果被忽略。



四. 总结与展望

1. 对模型预测有误差数字的反思

下面两图展示了支持向量机和神经网络在两个数据集上的预测表现。

左图是支持向量机的表现，可以发现，支持向量机模型主要对数字“9”的预测准确性较差。

右图是神经网络的表现。基于图像边缘数据训练的神经网络模型效果很差，这可能是由于调参时忽略了隐藏层的层数的优化问题。神经网络算法对于“1”，“4”，“5”，“9”这四个数字的预测性能较差。

			precision	recall	f1-score	support
		0	0.97	0.95	0.96	78
		1	0.92	0.88	0.90	83
		2	0.92	0.89	0.91	81
		3	0.88	0.92	0.90	76
		4	0.85	0.90	0.88	70
		5	0.85	0.96	0.90	67
		6	0.96	0.95	0.96	79
		7	0.95	0.83	0.88	92
		8	0.83	0.82	0.83	73
		9	0.76	0.83	0.79	75
	accuracy				0.89	774
	macro avg		0.89	0.89	0.89	774
	weighted avg		0.89	0.89	0.89	774
			precision	recall	f1-score	support
		0	1.00	1.00	1.00	76
		1	0.97	0.95	0.96	81
		2	0.97	0.99	0.98	77
		3	0.99	1.00	0.99	79
		4	0.97	1.00	0.99	72
		5	0.99	0.99	0.99	75
		6	0.99	1.00	0.99	77
		7	1.00	0.98	0.99	82
		8	0.96	0.95	0.95	73
		9	0.95	0.95	0.95	82
	accuracy				0.98	774
	macro avg		0.98	0.98	0.98	774
	weighted avg		0.98	0.98	0.98	774

			precision	recall	f1-score	support
		0	0.78	0.64	0.70	92
		1	0.47	0.71	0.56	52
		2	0.71	0.61	0.65	90
		3	0.62	0.49	0.55	102
		4	0.47	0.57	0.52	61
		5	0.43	0.47	0.45	68
		6	0.68	0.61	0.64	87
		7	0.57	0.69	0.63	67
		8	0.47	0.46	0.47	74
		9	0.37	0.37	0.37	81
	accuracy				0.56	774
	macro avg		0.56	0.56	0.55	774
	weighted avg		0.57	0.56	0.56	774
			precision	recall	f1-score	support
		0	1.00	0.99	0.99	77
		1	0.91	0.91	0.91	79
		2	0.94	0.96	0.95	76
		3	0.97	0.96	0.97	81
		4	0.91	1.00	0.95	67
		5	0.96	0.96	0.96	75
		6	0.99	0.96	0.97	80
		7	0.99	0.99	0.99	80
		8	0.96	0.93	0.95	74
		9	0.94	0.91	0.92	85
	accuracy				0.96	774
	macro avg		0.96	0.96	0.96	774
	weighted avg		0.96	0.96	0.96	774

从模型预测总的比较中，我们也可以发现一些特例，均表现出模型内部的一些缺陷。



在这个7的例子中，竟然没有一个模型预测正确，经过对训练集数据的观察，发现有大量的“7”中部带有一横杠以体现区分，另一些中间没有一杠但标签值为“7”的，下方的倾斜水平大都较大，而上方一横大都比较水平。从这个角度讲，模型所学习到的一些特征很容易受训练数据的非根本性特征的干扰。



总结与展望

1.对模型预测有误差数字的反思

2.进一步优化的方向

D

成了这一预



训练模型所用的数据全部为  $32 \times 32$  的，但实际生活中的手写数字绝非如此简单，好在对于高精度图片而言，将相近的像素点整理合一化，可以降低图片的像素。对于更低精度的图片，使其具有  $32 \times 32$  的像素就相对困难，可能还要设计专门的修补扩展算法。

本次实验提取了图片边缘特征，却没有更深地挖掘和利用边缘。实际上利用数字的轮廓图，即可以确定成为数字“8”的充分条件，即图像轮廓共有三个闭环曲线。除此之外，利用图像轮廓做一些中心点拟合，即通过单点线的轮廓确定一系列局域中心点，把这些中心点连接起来，即组成图像的单点线骨架。

除此之外，图像的凹凸程度、偏倚和倾斜的角度，都一定程度上可以加入到图像的特征工程中，虽然未必会直接用于模型训练，但是可以集成在图像预处理功能中，达到预处理时对图像去噪声的效果。对于手写数据来说，图像的边缘最容易因为笔误出现一些噪声点，上述特征工程的提出即为实验设计了除噪声的思路。

---

## 参考文献

---

- [1] 孙德刚.BP 神经网络遗传算法的图像识别技术分析与实现[J].信息技术与信息化,2022(05):190-193.
- [2] 李雪芳.基于机器学习的计算机网络图像识别系统[J].信息技术与信息化,2022(08):206-209.
- [3] 蔺伟,张弛.人工智能背景下图像识别技术研究[J].无线互联科技,2022,19(14):108-110.
- [4] 李玉臣.基于 OpenCV 的计算机图像识别技术研究与实现[J].电脑编程技巧与维护,2022(11):147-149+169.DOI:10.16184/j.cnki.comprg.2022.11.033.