基于随机森林的手写数字识别

范仕欣 周 丽

四川大学锦城学院计算机与软件学院 四川 成都 611731

【摘 要】随机森林是一种高效灵活的机器学习算法。近年来手写数字识别的研究引起广泛关注，基于随机森林的特点，本文将基于随机森林算法，进行手写数字识别实验，并将其与决策树算法下的手写数字识别进行比较，以确保手写体的数字识别。结论是，在手写符号的数字识别中，随机森林更为有效和准确。

【关键词】手写数字识别随机；森林决策树

1 引言

手写数字识别技术(handwritten numeral recognition)是光学字符识别技术（optical character recognition)的一个分支[3]，其主要的研究内容是：利用计算机等电子设备自动识别手写数字体。本文将通过采取随机森林算法实现手写数字识别，同时将随机森林与决策树算法所实现的手写数字识别结果进行比较，从而展示随机森林在手写数字识别中的优越性。

2 MNIST 数据集的介绍

本文所采用的数据集是 MNIST，该数据集来源于美国国家标准与技术研究所，其中有 60000 个字符作为训练集，10000 个字符作为测试集。在作为测试的 10000 张图片里，每一张图片都显示着 0-9 里的任意一个数字，且每张图片都有一个与之相对应的标签，其标签集包含了 0、1 2 3 4 5 6 7、8、9 共 10 个分类类别，其中每一张图片都是 28*28 像素的灰度图像。[2]

实验所采用的训练数据集如下所示，一次性将所要训练的数据集传入，以下将给出部分数据为例：

图 1：MNIST 手写数据集

3 随机森林算法

中国有句古话叫做“三个臭皮匠顶个诸葛亮”，对于传统的决策树算法是通过算法自身所具有的知识对新的数据进行分类处理，其所包含的每一个决策树都作为一个分类器，有多少棵决策树就有多少个分类结果。而随机森林可以说是采用集成学习的改进决策树，随机森林算法通过“Bagging”算法将多个 CART 树集成根据所有决策树投票获得最终结果。[1] 希望以此能够使得最终的分类效果能够超过单一结果选取的效果。

4 随机森林在手写数字识别中的训练流程

在此实验的过程当中，随机森林对于手写数字识别的具体流程大致如下：

4.1 数据准备与模型搭建

首先在读入数据集之后保存 100 张手写字符图片，为了减少计算量笔者对所传入的数据再一次进行了归一化的处理。

其次，引入随机森林的分类器，进而用其构建一个随机森林的模型用于手写数字识别的训练。在模型的构建过程中，包括几个重要的参数，n_estimators 代表决策树的个数对于随机森林算法决策树的个数越多越好本文取值为 100，虽然可能随着决策树数目的增多性能会受到一定的影响但是至少要 100 个左右、criterion：计算分类结果好坏的标准选值为‘gini’，以便来选择最合适的节点、max_depth 设置树的最大深度选值为 10、max_features：用于选择最合适的属性时所划分的特征不能超过此值、verbose：表明任务是否还在进行，此处设置为 1 。

4.2 训练过程

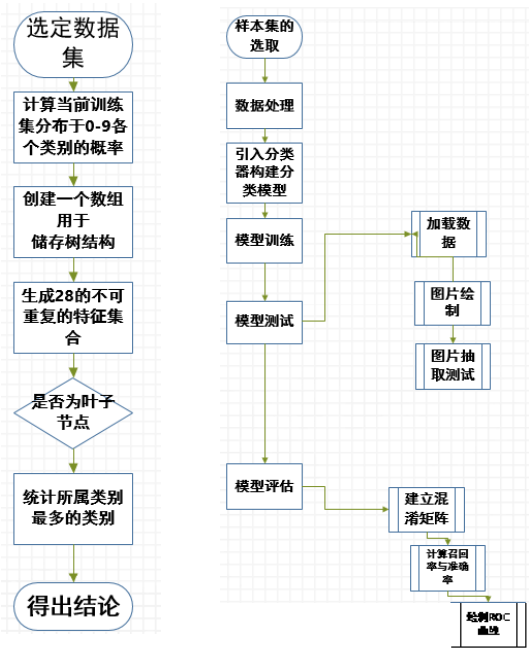

图 2：实验流程图 3：训练过程

在训练过程中涉及到一个有关于叶子结点的判断问题，当所遍历到的节点不是叶子节点时那么我们不能对其进行统计，那么将选取样本集特征集合中的一个特征,利用所有样本集中该特征值的大小对样本进行排序，将其全部放入右分类中并遍历左分类计算 infoGain 的值，继而遍历整个特征值集合然后创建左右节点，遍历其左右节点直到遍历到其叶子节点，然后进行所属类别的统计。

5 模型测试

具体的模型测试流程可以参照上图 2 中的流程，在此实验中笔者在测试的过程中是每一次测试对四张图片进行测试，在测试时将数据放入已经训练好的模型当中，为了避免预测结果被遮挡的情况所以每一次只用随机抽取函数在测试集中随机抽取两张图片，只进行两个字符图片的绘制，笔者选取某一组的图片测试的具体结果进行举例如下所示：

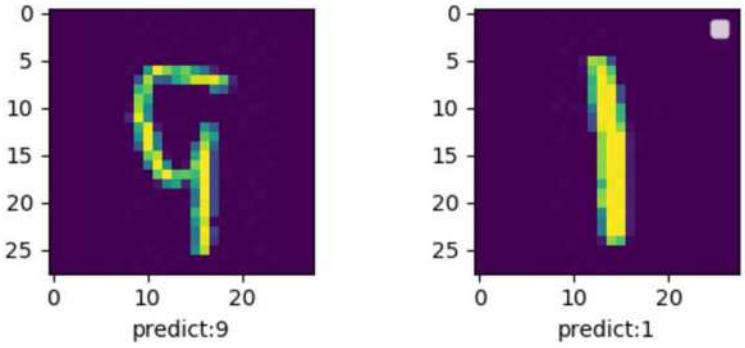

图 4：测试结果

可以看出在此时的测试结果里，应用随机森林的手写数字识别的图像已经相对比较清楚了，即就是在此时手写数字识别已经能够取得一个较好的识别效果了。

6 结果分析与模型的评估比较

为了证明随机森林在手写数字识别中有较好的效果，笔者将通过采用混淆矩阵的建立，召回率以及准确率的计算以及 ROC 曲线的绘制，来对随机森林算法模型与决策树算法模型进行对比评估，证明随机森林应用于手写数字识别当中相比于决策树算法效果更好。

6.1 混淆矩阵的建立

混淆矩阵对一个分类问题首先将其划分为正类与负类两个类别根据真实类别与预测类别构建混淆矩阵，对于本文所采用的 MNIST 数据集而言，其包括了 0-9 十个类别的数字识别，利用随机森林与决策树算法构建的混淆举证如下：

表 1：随机森林混淆矩阵

	0 类	1 类	2 类	3 类	4 类	5 类	6 类	7 类	8 类	9 类
0 类	115	0	0	0	0	0	0	0	0	0
1 类	0	123	0	0	0	0	0	0	0	0
2 类	47	0	76	0	0	0	0	0	0	0
3 类	0	0	0	116	2	0	0	0	0	0
4 类	0	0	1	8	84	0	0	0	0	0
5 类	0	0	0	0	1	117	1	0	0	0
6 类	0	0	0	0	0	0	103	1	0	0
7 类	0	0	0	0	0	0	1	105	0	0
8 类	0	0	0	0	0	0	0	0	67	56
9 类	0	0	0	0	0	0	0	0	0	115

表 2：决策树的混淆矩阵：

	0 类	1 类	2 类	3 类	4 类	5 类	6 类	7 类	8 类	9 类
0 类	106	1	0	0	0	0	0	0	0	0
1 类	48	75	0	0	0	0	0	0	0	0
2 类	0	12	79	1	0	0	0	0	0	0
3 类	0	4	0	106	6	0	0	0	0	0
4 类	0	0	1	0	90	2	0	0	0	0
5 类	0	0	0	0	0	115	4	0	0	0
6 类	0	0	0	0	0	0	100	4	0	0
7 类	0	0	0	0	0	0	5	99	0	0
8 类	0	0	0	0	0	0	0	0	60	63
9 类	0	0	0	0	0	0	0	0	0	110

6.2 召回率与准确率的计算

召回率是测试样本中正确分类为正类的样本数占实际为正类样本数的比例。而准确率所代表的是测试样本中正确分类的样本数占总测试的样本数的比例。根据混淆矩阵所得的结果采用如下公式：

$$Recall = \frac{TP}{TP + FN} \quad \text{（公式 1）}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad \text{（公式 2）}$$

根据两个算法所得的混淆矩阵将其结果带入公式中进行计算，所得结果如下所示：

表 3：召回率与准确率

	召回率	准确率
随机森林	0.948638362	94.92%
决策树	0.874982536	87.65%

6.3 ROC 曲线的绘制

在 ROC 曲线的绘制之前为了方便将传入的数据所带的标签进行二值化的处理。在随机森林与决策树模型中首先设置 x，y 上下限避免和坐标轴的边缘重合，设置中文需要的导入库设置标签名，同时涉及到三个重要参数的计算 TPR，FPR 以及 thersholds 通过采用公式 TPR = TP / (TP + FN) 以及 FPR = FP / (FP + TN) 对于样本数据进行计算得到其值同时计算最佳阈值，所得的 ROC 曲线如下：

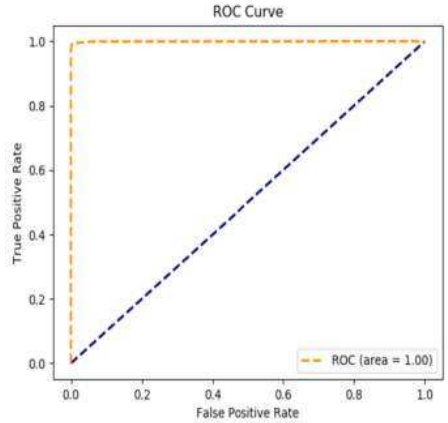

图 5：随机森林 ROC 曲线

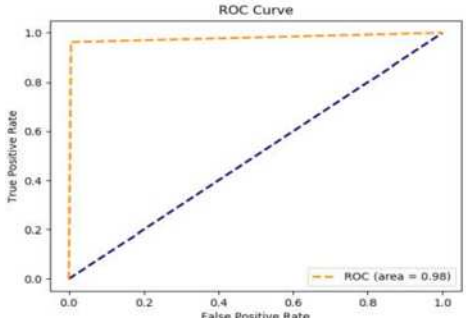

图 6：决策树 ROC 曲线

由上结果可以看出，无论是从混淆矩阵、召回率与准确率的计算还是 ROC 曲线的绘制都是随机森林算法应用于手写数字识别的效果更好。在混淆矩阵中随机森林的准确识别率高于决策树，而在召回率与准确率的计算中很明显是随机森林模型的值更加接近于 1，根据以上 ROC 绘制的结果而言随机森林关于 MNIST 数据集的 ROC 曲线相比于决策树的 ROC 曲线更加接近于左上角，说明这个模型的性能更好。同时还有未呈现出来两者的训练时间，在 MNIST 数据集上决策树算法的训练时间相比于随机森林的训练时间要多出了将近 17 秒的时间，那么在一点上随机森林也是远远优于决策树算法的。

7 结语

通过上述实验的过程，文章中提出的基于随机森林对于手写数字识别分类的实现方法。其主要思路是通过在 MNIST 数据集上利用随机森林算法的基本原理，对于此数据集的识别有一个更加准确与高效的训练，对数据集进行训练，对测试集进行测试，同时通过多方面的指标进行模型的评估与对比，通过实验证明了随机森林不仅可以实现手写数字识别的分类同时也获得了较好的效果。

【参考文献】

[1] 范芳菲．基于 KNN 对手写数字的识别 [J]．电子制作，2020(24)：53-54+74.

[2] 李想．基于 PCR 与 SVM 的手写数字识别效果对比研究 [J]．科技创新与生产力，2021(03)：65-69.

[3] 陈鸿宇．基于 KNN 算法手写数字识别技术的研究与实现 [J]．信息通信，2020(12)：28-32.

[4] 薛强． 基于信任随机森林的不确定手写数字识别研究 [D]．成都理工大学，2019.

[5] 赵力衡．基于决策树的手写数字识别的应用研究 [J]．软件，2018,39(03)：90-94.

[6] 张勇，马亚州，侯益明．基于 KNN 算法的手写数字识别研究 [J]．无线互联科技，2020,17(14)：111-112+115.

[7] 贯宸，杨云峰．基于四种算法下的手写数字识别准确率对比 [J]．新型工业化，2020,10(07)：1-3.

[8] 王爱丽，薛冬，吴海滨，王敏慧．基于条件生成对抗网络的手写数字识别 [J]．液晶与显示，2020,35(12)：1284-1290.

[9] 杨刚，贺冬葛，戴丽珍．基于 CNN 和粒子群优化 SVM 的手写数字识别研究 [J]．华东交通大学学报，2020,37(04)：41-47.

[10] 王泽原，赵丽，胡俊．大数据环境下利用随机森林算法和决策树的贫困生认定方法 [J]．湘潭大学自然科学学报，2018,40(06)：115-120.