用Bayes分类器做0-9数字手写体识别

陈天阳 2018-05-24

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 参数  设置 | 图像尺寸 | | 28x28 | —— | —— | —— | —— | —— | —— | —— | —— | —— | —— | —— | |
| 训练集 | | 10x1000 | 10x5000 | —— | —— | —— | —— | —— | —— | —— | 60000 | 60000 | 60000 | |
| 帧数 | | 7x7 | —— | —— | —— | —— | —— | 14x14 | 28x28 | 4x4 | —— | 14x14 | 28x28 | |
| 帧尺寸 | | 4x4 | —— | —— | —— | —— | —— | 2x2 | 1x1 | 7x7 | —— | 2x2 | 1x1 | |
| 分类法 | | 二分 | —— | —— | —— | —— | —— | —— | —— | —— | 多分(0-16N) | 多分(0-4N) | —— | |
| 图像格式 | | gray 0-255 | —— | —— | —— | —— | —— | —— | —— | —— | BW | BW | BW | |
| 阈值 | | 4/16(1020) | —— | 1/16(255) | 6/16(1530) | 8/16(2040) | 12/16(3060) | 1/4(255) | 1/1(255) | 12/49(3060) | 不设置 | 不设置 | 不设置 | |
| 实验  次数 | 数字 | 测试集 | Lab1 | Lab2 | Lab3 | Lab4 | Lab5 | Lab6 | Lab7 | Lab8 | Lab9 | Lab10 | Lab11 | Lab12 | |
| 正确率  (%) | 0 | 980 | 86.94 | 86.12 | 83.27 | 85.10 | 85.61 | 78.16 | 89.29 | 90.20 | 48.27 | 88.98 | 90.20 | 90.51 | 90.51 |
| 1 | 1135 | 94.98 | 94.45 | 95.15 | 94.98 | 94.89 | 94.45 | 96.30 | 94.54 | 81.76 | 96.56 | 95.86 | 95.86 | 95.59 |
| 2 | 1032 | 75.10 | 75.97 | 69.09 | 69.19 | 69.09 | 59.59 | 82.46 | 83.33 | 42.34 | 82.56 | 83.14 | 82.66 | 82.56 |
| 3 | 1010 | 72.18 | 71.49 | 59.80 | 75.25 | 78.22 | 68.51 | 81.49 | 82.87 | 52.18 | 81.68 | 82.97 | 84.55 | 83.56 |
| 4 | 982 | 68.94 | 69.55 | 57.84 | 68.53 | 65.27 | 50.41 | 78.00 | 79.12 | 36.56 | 79.33 | 82.28 | 90.22 | 80.96 |
| 5 | 892 | 50.00 | 52.80 | 47.65 | 53.92 | 53.25 | 48.21 | 66.03 | 70.63 | 10.76 | 64.69 | 70.85 | 65.47 | 70.29 |
| 6 | 958 | 82.25 | 82.46 | 84.55 | 76.93 | 77.97 | 70.25 | 87.16 | 89.14 | 73.17 | 90.81 | 90.71 | 87.79 | 88.83 |
| 7 | 1028 | 72.37 | 73.93 | 65.27 | 77.04 | 78.89 | 72.57 | 82.20 | 83.66 | 68.19 | 85.31 | 85.51 | 85.21 | 84.73 |
| 8 | 974 | 64.89 | 65.30 | 60.99 | 63.76 | 60.16 | 51.33 | 74.02 | 76.08 | 48.87 | 73.31 | 78.03 | 62.94 | 77.82 |
| 9 | 1009 | 74.93 | 71.75 | 65.91 | 73.93 | 73.14 | 51.83 | 81.57 | 84.74 | 69.38 | 82.06 | 83.55 | 22.89 | 83.55 |
| Total | 10000 | 74.77 | 74.86 | 69.46 | 74.36 | 74.17 | 65.12 | 82.21 | 83.71 | 53.98 | 82.90 | 84.59 | 77.13 | 84.13 |

**注：所有“——”均以Lab1为样本，代表该处参数取值与实验一对应相同。**

**数据来源：MNIST**

Lab1：正确率太低。

Lab2：无明显改进，有数字甚至不升反降，而计算量加大。进一步实验发现，综合考虑计算量和正确率时，1000应是一个比较好的数。

Lab3,4,5,6：无明显改进，绝大部分数字的识别率随阈值的增大先增大，后减小，区别在于最大值点处。大部分的最大值点位于阈值小于或等于6处，因此整体识别率也先增大后减小，在阈值等于4处恰恰效果最好。这在直观上也是可以理解的。

Lab7：帧减小为原先的1/4，效果明显改进。考察单元细化以致效果改善，这在直观上也是可以预期的。

Lab8：帧再次减小为原先的1/4，变为逐像素计算，无明显改进。

Lab8+：阈值设为0.1，所有结果和实验八相同，可见虽然图像是gray，但是基本都是0或255，可以视作灰度值。

Lab9：处理过于粗犷，可以预料的差。

Lab10：考察更加细致，在于第BW图像帧内多分，分类梯度为1个像素点，这样也就无所谓阈值。与对比Lab1作对比可以发现在帧尺寸相同的情况下，多分的效果明显好于二分，这也是可以直观理解的。

Lab11：进一步减小帧尺寸，相比于同样帧尺寸的二分法，只提升了2.3个百分点。这也是方便理解的，因为帧越小，内部的分类梯度就越小，提升空间就越小。

Lab12：帧尺寸减小到1个像素点，做到这一步，与Lab8在本质上无差别，在本方法下的性能提升空间为零，且消耗巨大的时间和运算资源。正确率反而下降的原因可能在于在处理概率连乘的时候为避免0概率而令分母分子同时加1这一操作导致的误差；还有一点更为关键，此时需要进行28x28，即784个概率连乘，导致了大量极大数甚至Inf，而\*sqrt(10)的操作已无法较好地加以平衡。其中第二点原因应当更为关键。需要指出的是，做到这一步，耗时极长(Intel Core i7:57.4s for 10000 images)。

在进行了一些改进后(补偿常数设置为sqrt(2))，避免了概率值出现Inf，总体正确率回升到84%附近。