Neural Network 对 breast_cancer_data 数据进行分类判断病理结果

一、 实验目的

用 Neural Network 对 breast_cancer_data 数据进行分类来判断病理结果(0 良性, 1 原位癌, 2 恶性, 3 恶性 1 级, 4 恶性 2 级, 5 恶性 3 级)

二、实验内容和要求

1、用 libSVM 对 breast_cancer_data 数据进行分类来判断病理结果(0 良性,1 原位癌,2 恶性,3 恶性 1 级,4 恶性 2 级,5 恶性 3 级)

数据集: 1230 个数据样本, 42 个特征

良性样本: 806 个

原位癌样本: 164个

恶性: 169个

恶性1级:5个

恶性 2 级: 65 个

恶性 3 级: 21 个

2、用 libsvm 对 breast_cancer_data_binary 数据进行二元分类 (0 没有患病, 1 有患病)

数据集: 1230 个数据样本, 42 个特征

没有患病: 806

有患病: 424

三、实验主要仪器设备和材料

实验环境

硬件环境: 个人台式机 Microsoft Windows 8.1

软件环境: Matlab 2015a

四、实验步骤:

1、用 Excel 清理数据,把列为空值得删除;把特征为为空值的 item 补零;把没有 lable 的 item 袪除;把 label 为异常值得袪除(label>5)(没有考虑特征的关联性),最后保存为 breast_cancer_data.xlsx.(思考,进一步用程序实现自动化清理数据.

3、用 matlab 自带的 Neural Network 训练,验证模型并测试。(没有用 cross-validation)数据集:共1230个数据,42个特征

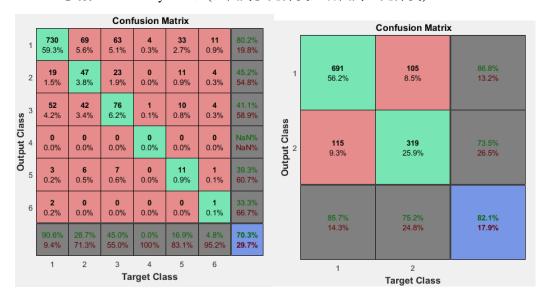
train: 1230*70%

validation:1230*15%

test:1230*15%

五、实验结果

1、参数: hidden layer 90, (左图为多元分类,右图为二元分类)



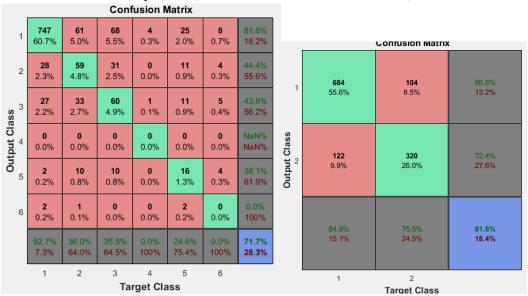
2、参数: hidden layer 9182. (左图为多元分类,右图为二元分类)



3、参数: hidden layer 92 (左图为多元分类,右图为二元分类)

Confusion Matrix									Confusion Matrix			
1	745 60.6%	74 6.0%	74 6.0%	5 0.4%	41 3.3%	13 1.1%	78.3% 21.7%		672	88	88.4%	
2	15 1.2%	40 3.3%	31 2.5%	0 0.0%	12 1.0%	2 0.2%	40.0% 60.0%	Output Class	54.6%	7.2%	11.6%	
Output Class	41 3.3%	46 3.7%	59 4.8%	0 0.0%	6 0.5%	5 0.4%	37.6% 62.4%		134 10.9%	336 27.3%	71.5% 28.5%	
	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	NaN% NaN%					
no 5	2 0.2%	2 0.2%	2 0.2%	0 0.0%	6 0.5%	1 0.1%	46.2% 53.8%		83.4% 16.6%	79.2% 20.8%	82.0% 18.0%	
6	3 0.2%	2 0.2%	3 0.2%	0 0.0%	0 0.0%	0 0.0%	0.0% 100%					
	92.4% 7.6%	24.4% 75.6%	34.9% 65.1%	0.0% 100%	9.2% 90.8%	0.0% 100%	69.1% 30.9%					
	1 2 3 4 5 6 Target Class								1	2 Target Class		

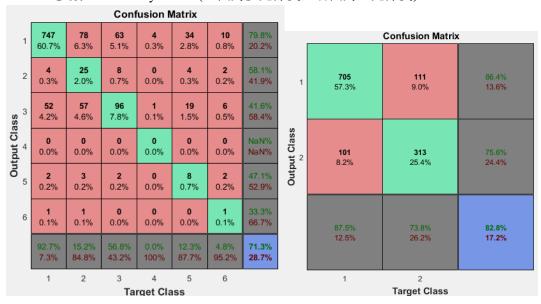
4、参数: hidden layer 93 (左图为多元分类,右图为二元分类)



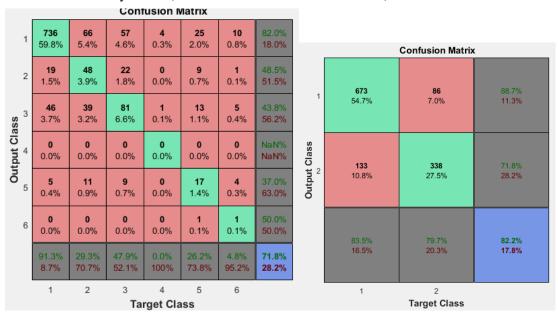
5、参数: hidden layer 94 (左图为多元分类,右图为二元分类)

										Confusion Matrix	
Confusion Matrix											
1	720 58.5%	80 6.5%	77 6.3%	4 0.3%	38 3.1%	11 0.9%	77.4% 22.6%	Output Class	697 56.7%	100 8.1%	87.5% 12.5%
2	19 1.5%	32 2.6%	20 1.6%	1 0.1%	8 0.7%	3 0.2%	38.6% 61.4%				
sse 3	54 4.4%	47 3.8%	68 5.5%	0 0.0%	10 0.8%	5 0.4%	37.0% 63.0%		109 8.9%	324 26.3%	74.8% 25.2%
Output Class	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	NaN% NaN%				
no 5	10 0.8%	3 0.2%	2 0.2%	0 0.0%	9 0.7%	2 0.2%	34.6% 65.4%		86.5% 13.5%	76.4% 23.6%	83.0% 17.0%
6	3 0.2%	2 0.2%	2 0.2%	0 0.0%	0 0.0%	0 0.0%	0.0% 100%				
	89.3% 10.7%	19.5% 80.5%	40.2% 59.8%	0.0% 100%	13.8% 86.2%	0.0% 100%	67.4% 32.6%				
	1 2 3 4 5 6 Target Class								1	2 Target Class	

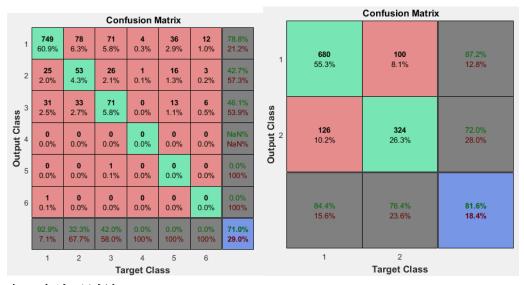
6、参数: hidden layer 95 (左图为多元分类,右图为二元分类)



7、参数: hidden layer 96 (左图为多元分类,右图为二元分类)



8、参数: hidden layer 97 (左图为多元分类,右图为二元分类)



六、交流及讨论

- 1、从实验一结果来看.用神经网络做多元分类平均准确率 70.62%
- 2、从实验二结果看,用神经网络做二元分类平均准确率 82.215%
- 2、之前训练效果比较差点的原因是,原始数据 mapping 到[01,]区间,训练效果差,现在是把原始数据 mapping 到[-1,1]区间。
- 3、你之前说改 W 和 b,可以在自定义,但一般是先 BP 算法,默认 W,B 都是零矩阵,我觉得没有必要改了,matlab 自带的例子也是默认参数,定义 10 层网络准确率都能达到 97.2% 还有 W,与 B 更新的学习率 alpha,没有提供更改的接口,改不了。
 - 4、总之,以上方法及参数训练出来的神经网络应该是较优的了。

附件

my_breast_cancer_project 文件夹是训练神经网络的 matlab 代码。

Reference:

9. http://ufldl.stanford.edu/wiki/index.php/Backpropagation_Algorithm