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 祛除；把 label 为异常值得祛除（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

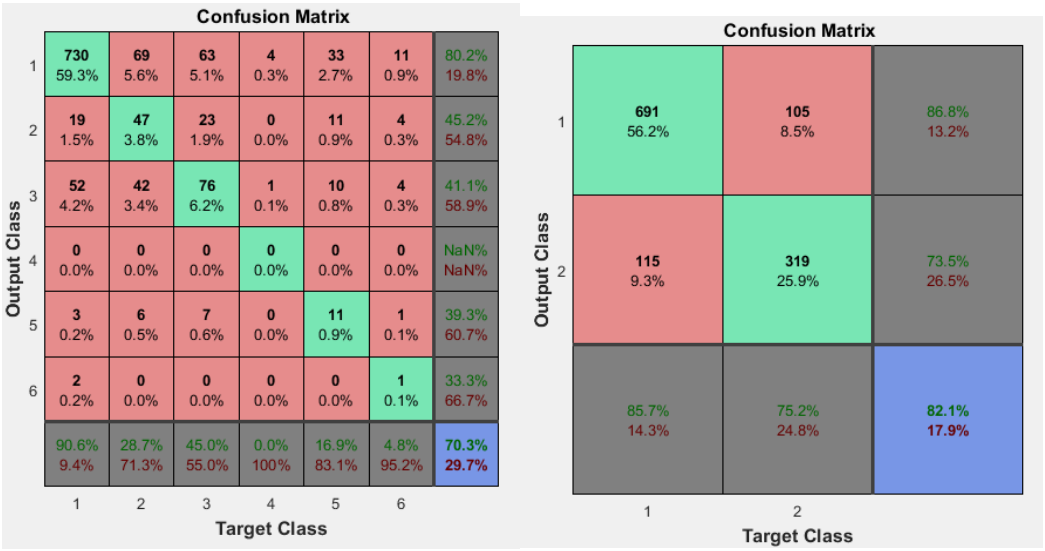
train：1230*70%

validation:1230*15%

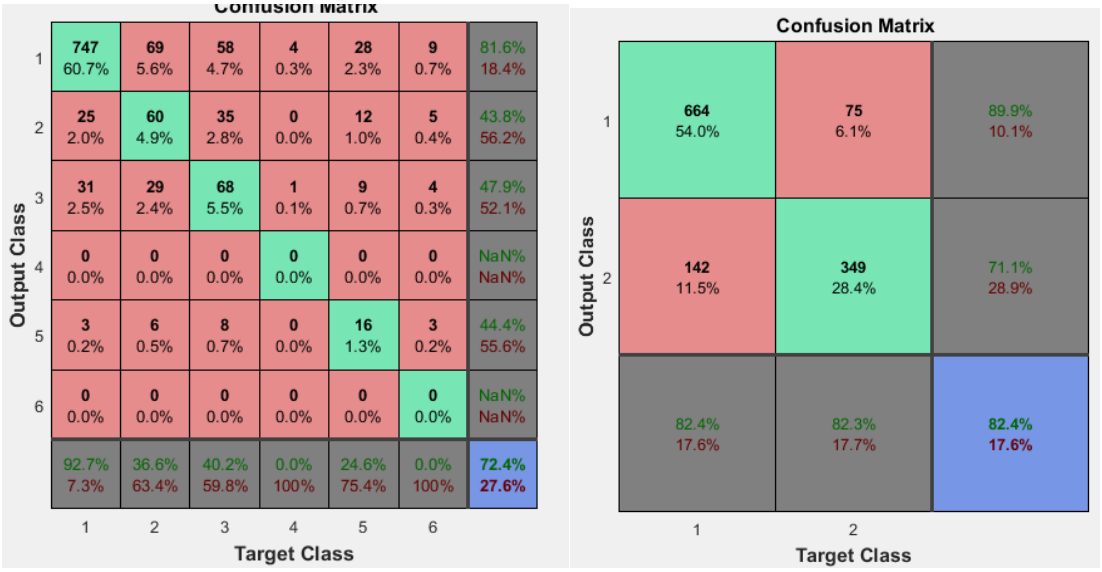
test:1230*15%

五、实验结果

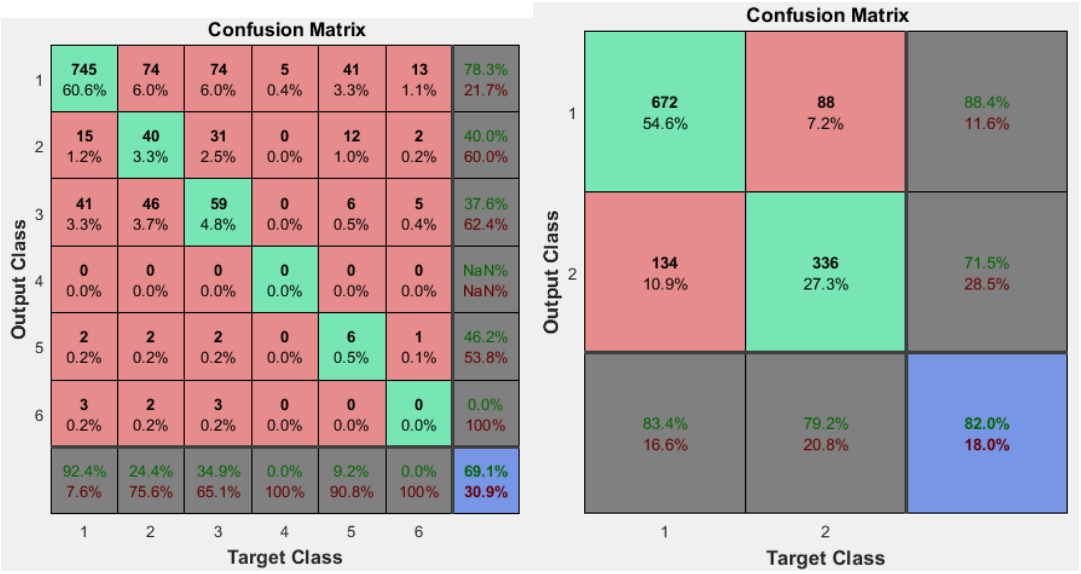
1、参数：hidden layer 90，(左图为多元分类，右图为二元分类)



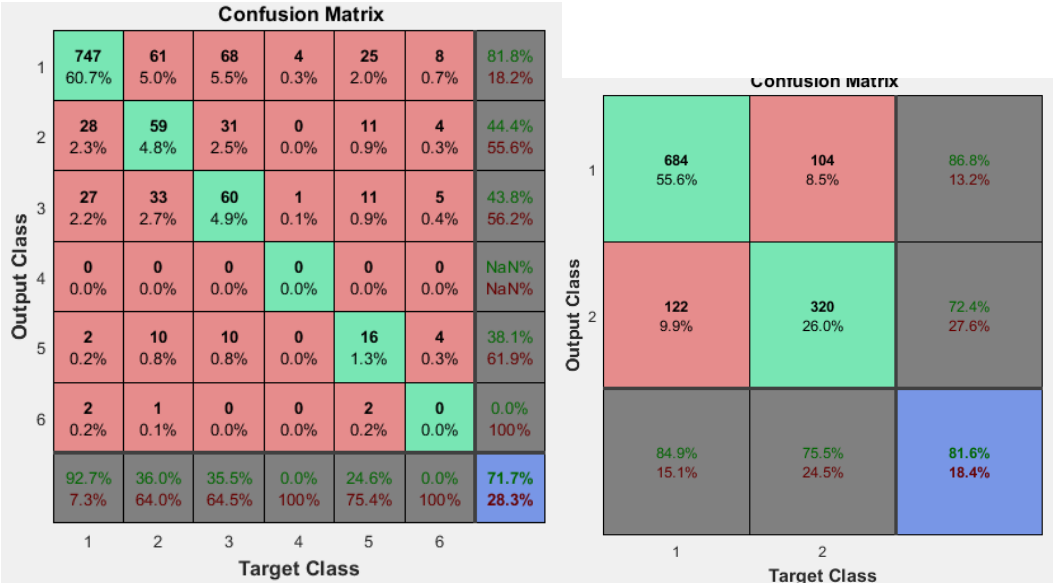
2、参数：hidden layer 9182.　(左图为多元分类，右图为二元分类)



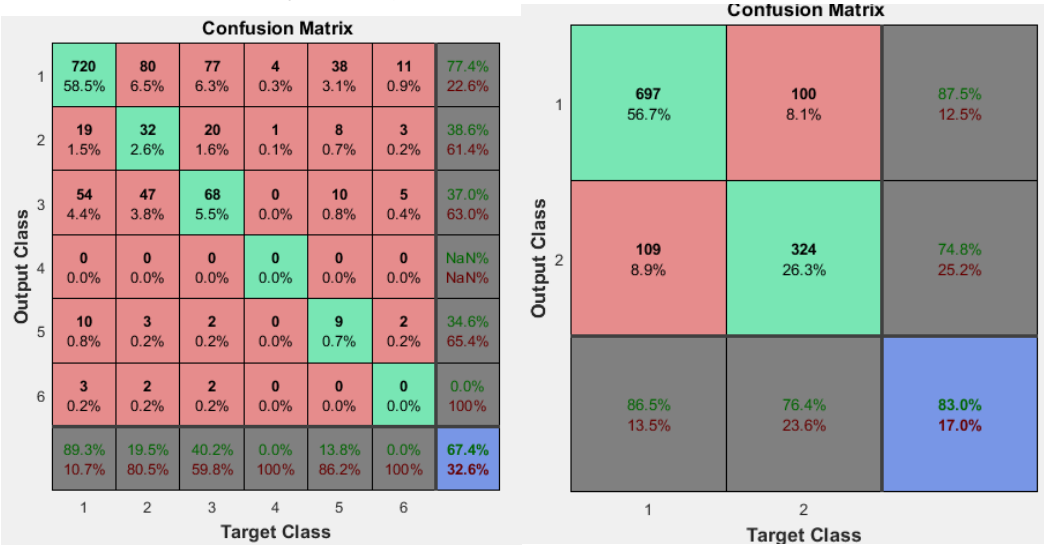
3、参数：hidden layer 92　(左图为多元分类，右图为二元分类)



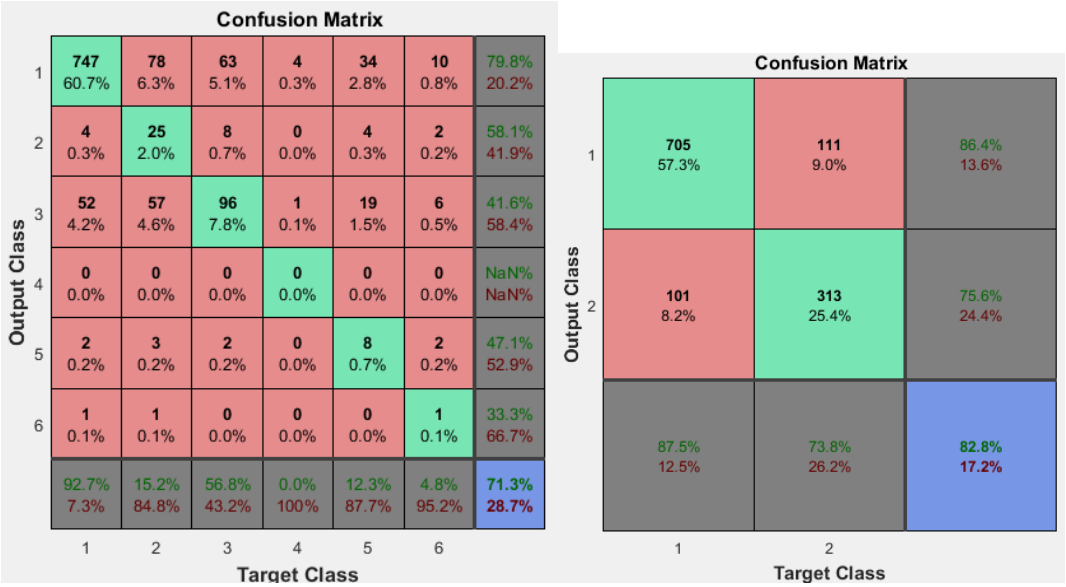
4、参数：hidden layer 93　(左图为多元分类，右图为二元分类)



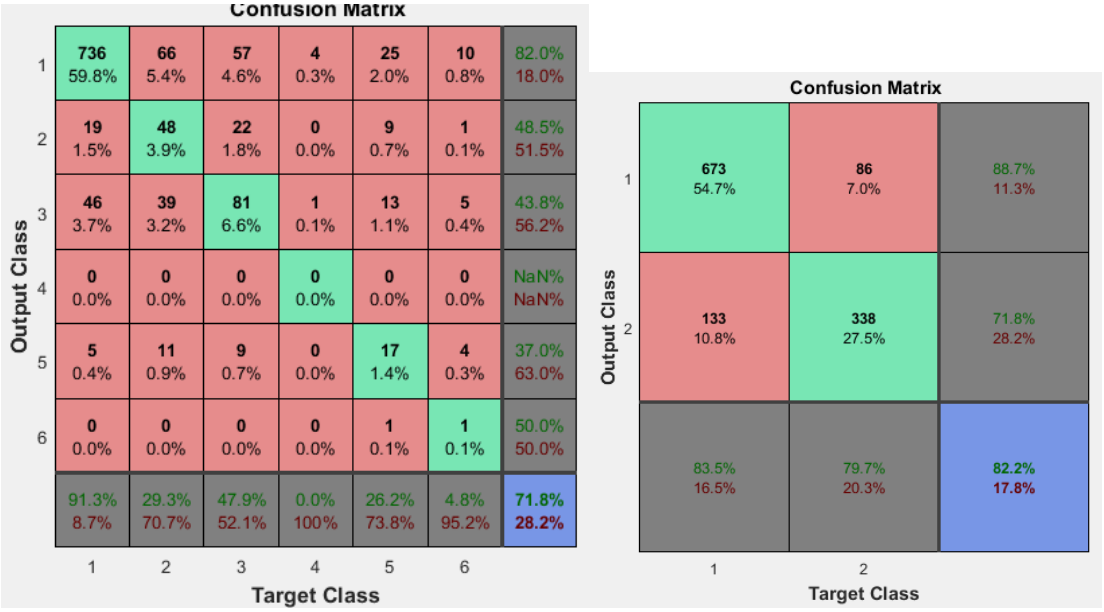
5、参数：hidden layer 94　(左图为多元分类，右图为二元分类)



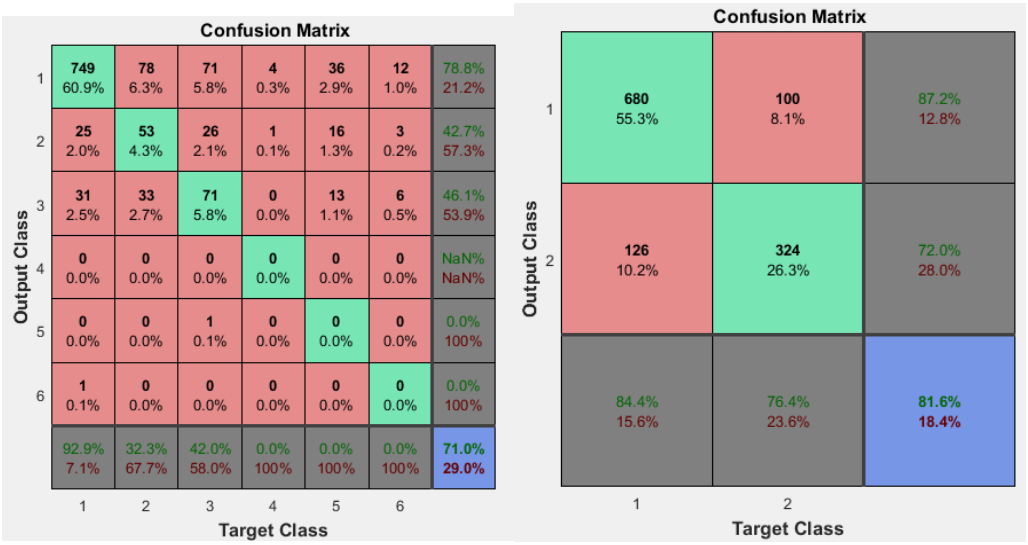
6、参数：hidden layer 95　(左图为多元分类，右图为二元分类)



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