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

一、 实验目的

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）

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：python 2.7; Matlab 2015a;libsvm

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)

2、用 matlab 把清理后的数据归一化后转换为 libsvm 的输入格式，并保存为 breast_cancer_data

3、对数据进行缩放: （Window command line）

```
svm-scale -l -1 -u 1 breast_cancer_data > breast_cancer_data.scale
```

4、对 scale 后的数据进行参数寻优，主要是找 c 与 g 的值，训练出 c=8.0；g=0.5（Window command line）

```
python grid.py breast_cancer_data.scale
```

5、交叉验证：（把数据集分成 10 份进行交叉验证）用高斯核模型。作用：对上面得到的参数进行交叉验证，然后的到一个 accuracy.不能训练出 model,只能用于 accuracy 比较（Window command line）,accuracy 为交叉验证时产生的各种模型的 accuracy 的平均值

数据集：共 1230 个数据，43 个特征

svm-train -c 8.0 -g 0.5 -v 10 breast_cancer_data.scale

6、自定义交叉验证得到最好的 model，用于 prediction

五、实验结果

实验一结果：(多元分类)

1、参数选择：-c 8.0 -g 0.5

```
.*.*
optimization finished, #iter = 46
nu = 0.179272
obj = -17.378133, rho = -0.704455
nSV = 17, nBSV = 1
Total nSV = 679
Cross Validation Accuracy = 71.3008%
```

经过交叉验证后，分类结果为：平均准确率：71.1273%

建立模型

- 随机选出 1000 个样本作为训练集，剩下的作为测试集
- 用 1000 个样本训练出模型，然后用 230 个样本进行测试

```
F:\riverchuan\DataMining\Code\libsvm\windows>rem predict the accuracy on the tes
t data_set. and the output file is the estimate using the model

F:\riverchuan\DataMining\Code\libsvm\windows>svm-predict -b 0 breast_cancer_data
_scale_test breast_cancer_data_scale_train.model breast_cancer_data_scale_test_p
redict
Accuracy = 75.6522% (174/230) (classification)

F:\riverchuan\DataMining\Code\libsvm\windows>_
```

准确率：75.6522%，这个可以生成 breast_cancer_data_predict 数据进行分析

2、参数选择：-c 2048.0 -g 0.00195

```
[local] 13 3 64.1525 (best c=2048.0, g=0.001953125, rate=71.2084)
[local] 13 -9 69.8297 (best c=2048.0, g=0.001953125, rate=71.2084)
[local] 13 -3 66.5045 (best c=2048.0, g=0.001953125, rate=71.2084)
2048.0 0.001953125 71.2084
```

```
.*.*
optimization finished, #iter = 12
nu = 0.113464
obj = -580.895656, rho = -0.667420
nSV = 4, nBSV = 0
Total nSV = 584
Cross Validation Accuracy = 71.2084%

F:\riverchuan\DataMining\Code\libsvm\windows>
Microsoft Pinyin 半 :
```

经过交叉验证后，分类结果为：平均准确率：71.2084%

3、参数选择：-c 524288.0 -g 7.62939453125e-6

```
[local] 27.0 -13.0 67.2344 (best c=524288.0, g=7.62939453125e-06, rate=71.2895)
[local] 27.0 -3.0 63.1792 (best c=524288.0, g=7.62939453125e-06, rate=71.2895)
[local] 27.0 -23.0 68.9376 (best c=524288.0, g=7.62939453125e-06, rate=71.2895)
[local] 27.0 3.0 64.1525 (best c=524288.0, g=7.62939453125e-06, rate=71.2895)
524288.0 7.62939453125e-06 71.2895

F:\riverchuan\DataMining\Code\libsvm\windows>
```

```
.*.*
optimization finished, #iter = 13
nu = 0.113223
obj = -148416.432590, rho = -0.666945
nSV = 4, nBSV = 0
Total nSV = 566
Cross Validation Accuracy = 71.2895%

F:\riverchuan\DataMining\Code\libsvm\windows>
```

经过交叉验证后，分类结果为：平均准确率：71.2%

实验二结果：（二元分类）

参数寻优

```
[local] 13 1 72.7642 (best c=32768.0, g=0.00048828125, rate=82.8455)
[local] 13 -11 82.3577 (best c=32768.0, g=0.00048828125, rate=82.8455)
[local] 13 -5 82.2764 (best c=32768.0, g=0.00048828125, rate=82.8455)
[local] 13 -15 80.8943 (best c=32768.0, g=0.00048828125, rate=82.8455)
[local] 13 3 70.5691 (best c=32768.0, g=0.00048828125, rate=82.8455)
[local] 13 -9 81.7073 (best c=32768.0, g=0.00048828125, rate=82.8455)
[local] 13 -3 80.5691 (best c=32768.0, g=0.00048828125, rate=82.8455)
32768.0 0.00048828125 82.8455
```

参数选择：-c 32768.0 -g 0.00048828125

```
.....................*.....................*.........*
optimization finished, #iter = 48873
nu = 0.378982
obj = -13202737.987338, rho = 181.480004
nSV = 454, nBSV = 391
Total nSV = 454
Cross Validation Accuracy = 82.5203%
```

经过交叉验证后，分类结果为：平均准确率：82.5203%

建立模型

- 随机选出 1000 个样本作为训练集，剩下的作为测试集
- 用 1000 个样本训练出模型，然后用 230 个样本进行测试

```
F:\riverchuan\DataMining\Code\libsvm\windows>rem predict the accuracy on the tes
t data_set. and the output file is the estimate using the model

F:\riverchuan\DataMining\Code\libsvm\windows>svm-predict -b 0 breast_cancer_data
_binary_scale_test breast_cancer_data_binary_scale_train.model breast_cancer_dat
a_binary_scale_test_predict
Accuracy = 84.3478% (194/230) (classification)
```

六、交流及讨论

1、可分析特征之间的关联性，通过设置特征之间的权重，来提高准确度。

2、数据集数量有限，是否可以用更多可用数据集来训练出更好的模型来提高准确度；

3、数据集更大是否会影响训练出来的模型?

4、-c 2084 与 -g 是两个可调变量，从上面实验结果可知调这两个参数没有质的飞跃了。

5、看 libsvm/tools/readme, libsvm/readme/, libsvm/matlab/readme

6、实验二的二元分类结果比多元分类结果好一些，但依然没有 90%以上。

附件是：

breast_cancer_data.xlsx 清理后的数据

breast_cancer_data 用 myExercise 转换为 libsvm 格式的文件

-c 8.0 -g 0.5\build_breast_cancer_data_model.bat 产生 model 的文件

-c 8.0 -g 0.5\breast_cancer_data_scale_train 训练 model 样本

-c 8.0 -g 0.5\breast_cancer_data_scale_test 测试样本

-c 8.0 -g 0.5\breast_cancer_data_scale_test_predict 预测值

-c 8.0 -g 0.5\cross_validation_breast_cancer_data.bat 交叉验证产生平均 accuracy 文件

-c 8.0 -g 0.5\breast_cancer_data_scale.png 参数寻优图片