**基于WEKA的数据挖掘算法学习（一）**

1. **Weka Explorer基本知识了解**

**1.1 Explorer标签页**

1. Preprocess，选择和修改要处理的数据。

2. Classify，训练和测试关于分类或回归的学习方案。

3. Cluster，从数据中学习聚类。

4. Associate，从数据中学习关联规则。

5. Select attributes，选择数据中最相关的属性。

6. Visualize，查看数据的交互式二维图像。

**1.2 状态栏**

状态（Status）栏在窗口的最底部，显示一些正在做的信息。

**1.3 Log 按钮**

一个可拖动的文本区域，在WEKA 中执行某种操作时， 该日志就会记录发生的操作及时间。

**1.4 WEKA 状态图标**

状态栏的右边是 WEKA 状态图标。X表示并发进程发生的数量。

1. **数据的准备与格式的转换**

1.下载完备数据集：waveform.data.z。

2.在waveform.data首行添加定义，并存为.csv文件。

3.在weka explorer中save保存为weka中可以使用的.arrf。

4.在文本编辑器中设置class为{0,1,2}:



这里属性的申明格式为：@attribute <attribute-name> <datatype>

1. **测试过程与结果**

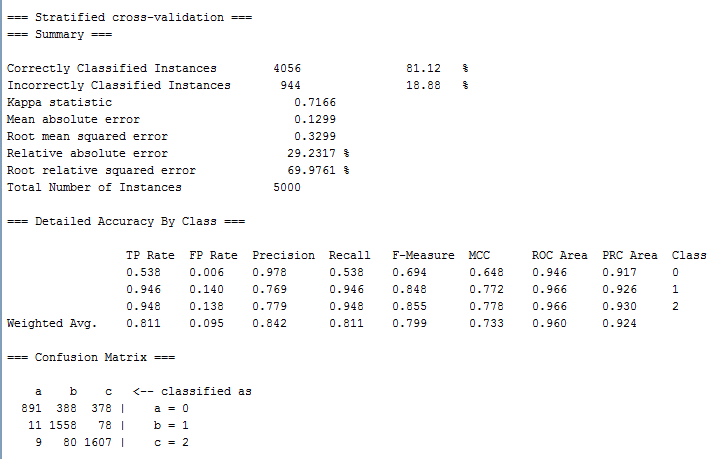
**3.1朴素贝叶斯(NaiveBayes)**

**3.1.1 思想**

采用了属性条件独立性假设，即假设每个属性独立地对分类结果发生影响。基于这个思想，可以得到：

简化了所有属性联合概率的求解，但在实际训练样本中往往属性并不独立，对结果会产生影响。

**3.1.2朴素贝叶斯实验结果截图**

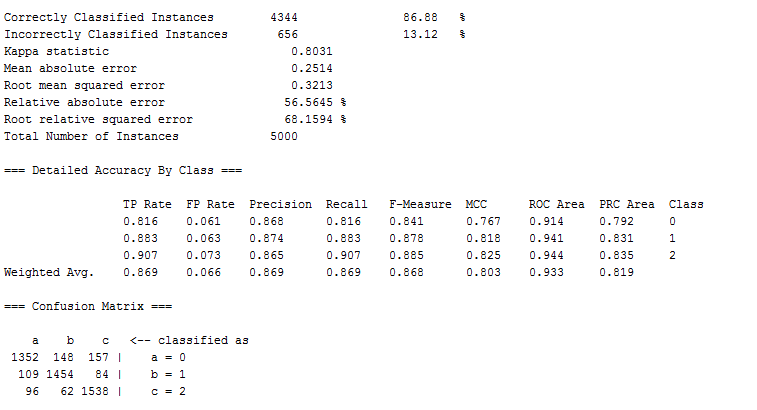


**3.2 SVM(SMO支持向量机分类)**

**3.2.1思想**

训练集在样本空间中找到一个划分超平面，将不同类别的样本分开。超平面由线性方程决定，寻找其法向量与位移。

**3.2.2 SMO实验结果截图**

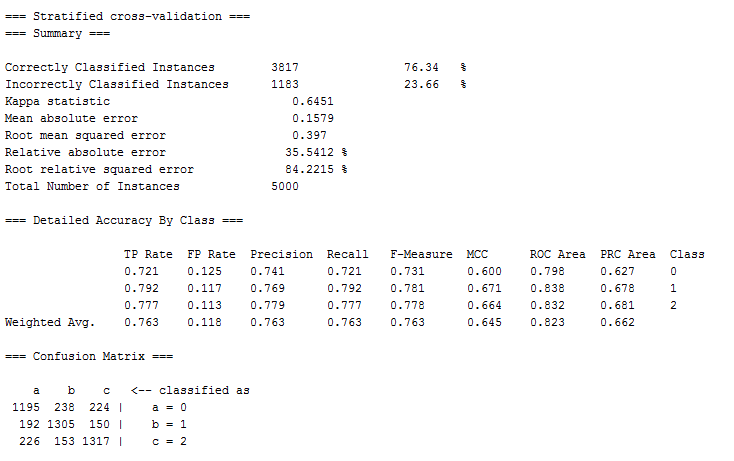


**3.3 KNN (IBk:k最近邻分类)**

**3.3.1思想**

如果一个样本在特征空间中的k个最相邻的样本中的大多数属于某一个类别，则该样本也属于这个类别，并具有这个类别上样本的特性。即，KNN是基于训练数据集对特征空间的一个划分，训练集、距离向量、k值、分类决策规则确定后，结果唯一确定。

**3.3.2 KNN实验结果截图**



此时是k值为1的情况，当调大k值，发现正确率会稍微增加，且查准率和查全率均有提高，原因是当k值小时，k近邻模型更加复杂，相反，k值大时，k近邻模型更简单。所以在实际中选择了k值小的情况。

**3.4 算法结果分析**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 算法 | Recall值 | F值 | AUC值 | Precision值 | Time |
| NaiveBayes | 0.811 | 0.799 | 0.960 | 0.842 | 0.05s |
| SVM | 0.869 | 0.868 | 0.933 | 0.869 | 0.91s |
| KNN | 0.763 | 0.763 | 0.823 | 0.763 | 0.01s |

从对比的结果来看，SVM的查准率和查全率均高于NaiveBayes、KNN，但运行效率却比二者低。

通过网上查阅资料得知，SVM是最好的现成的分类器，并且能够得到较低的错误率，SVM可以对训练集之外的数据点做很好的分类决策。

而朴素贝叶斯通过计算概率来进行分类，可以用来处理多分类问题，对小规模的数据表现很好，适合多分类任务，适合增量式训练，算法也比较简单，由于朴素贝叶斯的“朴素”特点，所以会带来一些准确率上的损失，需要计算先验概率，分类决策存在错误率。

而k近邻则对异常值不敏感，数值很大的时候不用这个，计算量太大。但是单个样本又不能太少 否则容易发生误分。

**四、本周实验心得体会**

本周对weka的界面做了简单的熟悉，对Classify进行了简单的使用。跑了三个经典的分类算法，但算法的理论还尚未清楚，只是跑了跑数据集，对实验的结果的分析也不好。