大数据分析实验报告——分类

苏致成 201250104

使用方法

J48

J48是C4.5在weka中的称谓，这是用在分类问题的经典算法，目标是监督学习。具体描述为：通过学习，找到一个从属性值到类别的映射关系，并且这个映射能够用于对新的类别未知的实体进行分类。其相比于ID3算法的改进如下：

1. 通过信息增益率而不是信息增益选择分裂属性。
2. 能够将连续性的属性进行离散化处理。
3. 构造决策树之后进行剪枝操作。
4. 能够处理具有缺失属性值的训练数据。

算法流程

计算类别信息熵：

$$Info(D) = -\sum_{i=1}^{m} p_i log_2(p_i)$$

计算每个属性的信息熵：

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} * Info(D_j)$$

计算信息增益：

$$Gain(A) = Info(D) - Info_A(D)$$

计算增益率：

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} * log_2 \frac{|D_j|}{|D|}$$

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo(A)}$$

根据如上信息增益率进行排序可以选出合适的分裂属性。分裂之后，若无其他分裂点则将其定义为子节点。重复上述过程即可得到决策树。

NaiveBayes

朴素贝叶斯算法的优点在于在数据量较小的时候比较有效，可处理多类别问题。

算法

用贝叶斯公式将多特征分类问题表达如下：

$$P(y|x_1,\dots x_n)=\frac{P(y)P(x_1,\dots,x_n|y)}{P(x_1,\dots,x_n)}$$

如果存在数据缺失问题，尤其是特征 x 越多该问题越突出，统计这些特征出现的概率则越困难，因此，朴素贝叶斯算法作出如下假设，即特征之间相互独立，互不影响，则可以简化为以下式子来求解某个特征的似然度。

$$P(x_i|y,x_1,\dots,x_{i-1},x_{i+1},\dots,x_n)=P(x_i|y)$$

数据集处理思路

J48和NaiveBayes数据集处理思路相同，因此一同描述。

1. 数据集中存在缺失值，因此先进行数据处理：若存在缺失值，则直接删除该条目。因为此处存在着众多的分类数据，无法像连续数据一样进行均值化处理。
2. 调用 $setClassIndex()$ 将最后一列属性不纳入分类的因素考虑。
3. 将数据集随机分为80%的测试集和20%的训练集。
4. 将预测结果与实际结果对比，计算分类准确率。

实验结果

结果显示，$J48$ 的分类准确率比 $NaiveBayes$ 的分类准确率更高。

J48

NaiveBayes


NaiveBayes_
"C:\Program Files\Java\jdk-15.0.2\bin\java.exe" ..
原数据集大小:48842
10月 10, 2022 1:59:10 下午 com.github.fommil.netlib
警告: Failed to load implementation from: com.githu
10月 10, 2022 1:59:10 下午 com.github.fommil.jni.Jn
信息: successfully loaded C:\Users\Dudu\AppData\Loc
剔除缺失值后数据集大小:45222
训练集大小:36178
测试集大小:9044
NaiveBayes分类方法，分类准确率： 0.8290579389650597