# 方法简介

随机森林：我们从字面意思上理解，该学习模型具有随机性且由多棵决策树构成。随机性主要体现在随机抽样（行随机）和随机选择特征（列随机）两方面，可以很好地防止过拟合现象发生；多棵决策树可以防止模型泛化能力低现象发生，这些现象在我们平常做研究的时候是很常见的。

随机森林流程：

假设数据集D = {（xi，yi）} i = 1,...N. 特征空间X特征的个数为H，迭代M轮，第j轮随机抽样形成的样本集Dj，第H轮训练的回归分类器（基函数）为Gj；

for j to M do:

1、从D中随机抽样得到抽样样本集Dj；

2、根据Dj训练CART得到回归分类器Gj；

3、如果是回归问题，F（x） += Gj / M；

endfor

4、得到模型F（x）

根据Dj训练CART得到回归分类器Gj：

1、随机选择hj个特征（hj <<H）,这里这个特征也是放回的，也就意味着树之间可能有共同的特征；

2、根据随机选的hj个特征和CART生成树准则训练回归分类树，截至条件是结点上的样本都分成了一类或者分裂时两个特征连续出现；

随机选择hj个特征，相对于使用全部特征而言，树的方差减小有效防止过拟合现象， 但偏差会增大，所以hj选择必须得当，可以用搜索法和交叉验证得到hj。

原文链接：https://blog.csdn.net/fengziyihang/article/details/79638851

# 应用思路/步骤

选取语音信号的五个特征，将提取出的特征向量导入excel表格

导入特征信息，调用matlab中的TreeBagger函数，作为随机森林分类器

前20个信息作为训练集，最后一个作为测试集

将分类结果与原本标签作比较，得出语音信号识别的准确率

运行100次求均值，得出平均准确率

# 运行代码

train\_data = xlsread('Character.xls','sheet1','A1:E200') ;

train\_label = xlsread('Character.xls','sheet1','F1:F200') ;

test\_data = xlsread('Character.xls','sheet1','A201:E210') ;

test\_label = xlsread('Character.xls','sheet1','F201:F210') ;

% 随机森林分类器（Random Forest）

nTree = 5

result=zeros(100,1);

for j=1:100

B = TreeBagger(nTree,train\_data,train\_label);

predict\_label = predict(B,test\_data);

for i=1:10

if ischar(predict\_label{i,1})

predict\_label{i,1}=str2num(predict\_label{i,1});

end

end

m=0;

for i=1:10

if predict\_label{i,1}==test\_label(i,1)

m=m+1;

end

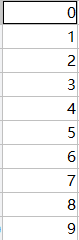
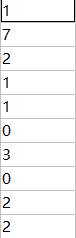
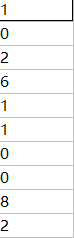
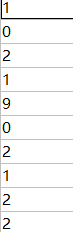
end

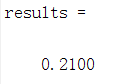
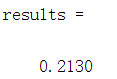
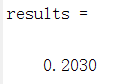
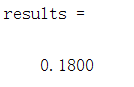
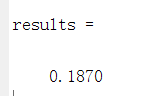
result(j)=m/10;

end

results=mean(result)

# 结果（截图+数据）评价分析

**原始标签 预测标签（选取三组）**



**平均准确率（运行5次）**

>> randomforest

nTree =

5

results =

0.1870

优点：

1、由随机森林基本原理，我们可以发现基模型之间的独立性很强，所以该算法可以并行处理， 也就是基于随机森林和大数据平台可以很好地应对大数据时代的到来。

2、随机抽样和随机抽取特征可以很好地防止过拟合现象；

3、随机抽取特征也可以克服特征维度过高问题；

4、模型结构相对简单；

缺点：

1、对于数据噪音过大的样本集，容易产生过拟合现象；

2、对于有不同取值的属性的数据，取值划分较多的属性会对随机森林产生更大的影响，所以随机森林在这种数据上产出的属性权值是不可信的。

原文链接：https://blog.csdn.net/fengziyihang/article/details/79638851