# RF vs Extrem RF

Extremtreeclassify:简直神奇，第一速度相当于RF的十分之一，而精确度相差无几！！

在原理上，etc比rf更进一步随机，

As in random forests, a random subset of candidate features is used, but instead of looking for the most discriminative thresholds, thresholds are drawn at random for each candidate feature and the best of these randomly-generated thresholds is picked as the splitting rule. This usually allows to reduce the variance of the model a bit more, at the expense of a slightly greater increase in bias:【http://scikit-learn.org/stable/modules/ensemble.html#forest】

【P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. In *Machine Learning*, volume 63, pages 3–42, 2006.】

# proximity矩阵

randomForest中的亮点，获取相似性矩阵。可以用于聚类（knn，Hierarchical等）。

原理：是如果两个观测实例落在同一棵树的同一个叶子节点的次数越多，则这两个观测实例的相似度越高。这个矩阵进行了归一化，就是除以ntree，你种了多少颗树。很简单的思想。

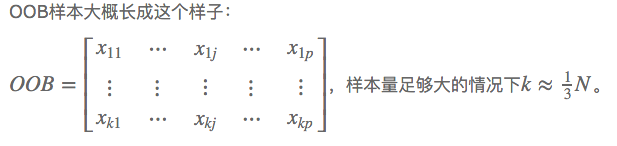
# 特征选择：

Decrease GINI： 【http://www.cnblogs.com/hrlnw/p/3850459.html】。

对于分类问题（将某个样本划分到某一类），也就是离散变量问题，CART使用Gini值作为评判标准。定义为Gini=1-∑(P(i)\*P(i)),P(i)为当前节点上数据集中第i类样本的比例。例如：分为2类，当前节点上有100个样本，属于第一类的样本有70个，属于第二类的样本有30个，则Gini=1-0.7×07-0.3×03=0.42，可以看出，类别分布越平均，Gini值越大，类分布越不均匀，Gini值越小。在寻找最佳的分类特征和阈值时，评判标准为：argmax（Gini-GiniLeft-GiniRight），即寻找最佳的特征f和阈值th，使得当前节点的Gini值减去左子节点的Gini和右子节点的Gini值最大。

对于回归问题，相对更加简单，直接使用argmax(Var-VarLeft-VarRight)作为评判标准，即当前节点训练集的方差Var减去减去左子节点的方差VarLeft和右子节点的方差VarRight值最大。

Decrease Accuracy：【http://www.loyhome.com/%E2%89%AA%E7%BB%9F%E8%AE%A1%E5%AD%A6%E4%B9%A0%E7%B2%BE%E8%A6%81the-elements-of-statistical-learning%E2%89%AB%E8%AF%BE%E5%A0%82%E7%AC%94%E8%AE%B0%EF%BC%88%E5%8D%81%E5%85%AD%EF%BC%89/】。



K个样本，p个特征。然后随机改变OOB样本的第j列：保持其他列不变，对第j列进行随机的上下置换，得到误差2。至此，我们可以用误差1-误差2来刻画变量j的重要性。当然这里loss function可以自己定。这里的大致思想就是，如果一个变量j足够重要，那么改变它会极大的增加测试误差；反之，如果改变它测试误差没有增大，则说明该变量不是那么的重要。

# 样本均衡

balanced RF，采是用分层重抽样的方法，即对minority class采取up-sampling,对于majority class 进行down-smapling。

weighted RF，