:已标记的数据集

：未标记的数据集

和独立同分布

:随机森林中所有的N个分类器

:随机森林中的第i个分类器 (i = 1,2,...N)

：随机森林中除了第i个分类器外的所有其他分类器，也称作的*concomitant ensemble， 伴随集成*

*:* 未标记数据新标上标签后组成的新的已标记数据集

：预先设置的置信度阈值

：中的第i个样本(i=1,2,...M,M为中总的样本个数)

:噪声样本比例

：训练数据集的大小

：最坏情况下的错误率

：utility function（效用函数）

CoForest

在每一轮迭代中，都会评估中的每一个样本，如果中的分类器得到标签的投票的总数比例超过了，那么就把这个样本标记上，然后把这个样本和标签复制到**中，然后在这次迭代中用中的已标记的数据去重新训练第i个分类器(注意：样本虽然已经打上标签并且复制到*，但是*并没有从中移除，所以还可能被其他())所选到，或者在下一轮迭代中又被选到)

当所有的样本被评估的置信度都高于，这些样本都会加到*中，这样*的规模就会非常大，甚至在极端情况下*等于*的大小，但是learned hypothesis（也就是学习到假设，就是训练出来的分类器）没有完全捕获到数据隐藏的分布情况，特别是在刚开始的前几次迭代，这个时候使用如此大量的自动标记的数据可能就会影响learned hypothesis的性能。这个时候就可以通过给每一个没有标记的样本一个权重，这个权重不是固定的，它是由伴随集成所预测的置信度来衡量的，这样做的好处是：一方面对没有标记的样本的给以权重减小了使用那些大量没有标记的而占主导统治的样本所带来的潜在的负面影响。另一方面，赋予未标记样本权重使得算法对参数就会不是很敏感，即使很小，这种赋予权重的策略也会限制那些低预测置信度样本的影响。

不仅如此，使用分类器的集成方法(ensemble)，不仅仅是提供了一个简单的方法去避免使用复杂的置信度评估方法，并且对于未标记的样本打上标签的准确度上也比单一的分类器好很多。虽然集成分类器比单一的分类器好很多，但是对于未标记的样本的错误分类还是不可避免的，所以不时的还是会收到噪声样本(未标记的样本打上了错误的标签复制到**)，这样就会使得的再一次迭代训练的分类器改良受到影响，解决的办法就是扩增已标记的数据集，在一定条件下将充分大量的新标记的样本加入到已标记的数据集中。

假设训练数据的大小为，噪声样本比例为，learned hypothesis在最坏情况下的错误率，如果三者的关系满足

 （1）

其中c为常量，如果满足(1)式关系，那么learned hypothesis 就会最小化噪声训练样本序列的不同从而已概率为1收敛到实际真正的hypothesis 

重写改善一下(1)式，就能得到效用函数

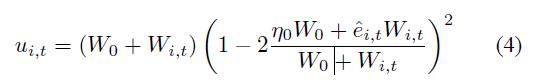
 （2）

在第t次迭代中，(*i* = 1*...N* ) 通过原始的已标记的数据集(样本数量为)和新标记上的已标记的数据集(样本数量为)，其中是由伴随集成决定和打上标签的，为在数据集上的错误率，中被错误打上标签的所有样本的权重为，

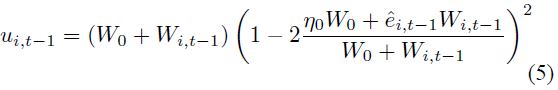
其中=，是对中的样本的预测置信度。为了表示上的统一，重写为由权重组成的，=，在扩增训练集中，噪声样本包含中的噪声样本和中那些被错误打上标签的噪声样本，因此在中的噪声样本比例可以这么计算：

 （3）

用(3)和扩增训练集的大小(+)分别代入（2）中的和，这样在第t次迭代中的效用函数就可以表示为：



类似的，在第t-1次迭代中的效用函数就可以表示为：



由（2）我们知道，效用函数和最坏情况的错误率的平方成反比，因此为了减小每个最坏情况的错误率，那么的效用函数值就必须在每一次迭代中增大，如，闲杂我们假设中几乎没有噪声样本存在并且每个分类器都满足弱分类器的要求，例如<0.5,那么我们比较（4）和（5）的右边就知道当<和>成立的时候就有，所以可以总结为：



根据（6）知道，<和>应该同时满足，但是即使这个条件都满足了，<也可能不成立，因为可能远远大于，为了让（6）在这种情况下成立，就必须进行下采样使得小于。