Training cost-sensitive neural networks with methods addressing the class imbalanced problem

对于代价敏感的机器学习算法，决策树很好实现，但在神经网络中不大好处理，因此需要借助其他方法来达到将类别不平衡以及代价敏感因素加入考虑的效果。文中提到了如下四种方法：

1. 过采样 oversampling
   通过将部分样本采样多次，来改变样本总体的类别分布趋势，使得没个样本的代价可以从其样本的个数来判断。
   先由如下公式算出基准类别的个数，

$$\lambda = \arg\min_j \frac{\frac{Cost[j]}{\min_c Cost[c]} N_{\arg\min_c Cost[c]}}{N_j}.$$

   再由下式算出每个类别所需要采样的样本个数

$$N_k^* = \left\lfloor \frac{Cost[k]}{Cost[\lambda]} N_\lambda \right\rfloor.$$

   如果需要采样的样本个数大于该类别样本总数，那么就从该类别样本中重复采样以使得满足所需数量。这样一来，训练的样本数往往会大于原样本个数，导致训练时间加长。
2. 欠采样 undersampling
   欠采样是过采样的反例，同样是改变类别分布趋势，但欠采样是通过降低每一个类别的样本数量而改变分布，基准类别个数由如下公式计算得到：

$$\lambda = \arg\max_j \frac{\frac{Cost[j]}{\max_c Cost[c]} N_{\arg\max_c Cost[c]}}{N_j}.$$

   可以看出，是过采样的逆。所需样本个数往往比该类别样本个数要少，这样可以加快训练速度。
3. 阈值变换 threshold-moving
   通过改变神经网络输出节点的阈值，来使得代价高的样例更难被误分类。
4. Hard-ensemble & soft-ensemble
   使用集成投票的方式来决定类别，两种方式区别在于 hard 是采用二值变量进行投票，soft 采用实数型变量进行投票。

关于这四种处理方法的性能，作如下讨论：

在二分类问题中，类别不平衡会加大代价敏感学习的难度，过采样和欠采样效果不错，但不如阈值变换和集成方法，阈值变换效果最好，soft-ensemble 其次。

在多分类问题中，采样法经常失效甚至造成负面影响，尤其是在极度不平衡的样本上，阈值变化有时也会失效，但有时也会取得效果，soft-ensemble 和阈值变换效果类似。