Improving experimental studies about ensembles of classifiers for bankruptcy prediction and credit scoring

这篇论文总体较水，主要说了利用决策树加集成学习来训练数据，实验具体过程几乎没讲，但介绍了一种比较新颖的决策树。

Credal Decision Trees

首先先定义一种计算方式，叫做 S*，在计算一个类别频率的时候，通常做法是将类别数量除以全样本数，但这样的话会导致数值有可能描述不精确，比如 1/5 和 100/500，虽然结果一样但是受到扰动时变化幅度差异很大。为了体现这种差异，在分母加上扰动项 s，通常为 1 或者 2，由此即为 S*（类别名）。

下图为该类别频率区间的计算方式：

$$p(c_j) \in \left[\frac{n_{c_j}}{N+s}, \frac{n_{c_j}+s}{N+s}\right], \quad j = 1, \ldots, k;$$

在构建决策树时，类别选择函数不采用通常的信息增益等公式，而使用叫 imprecise info-gain(IIG) 来划分类别。

$$IIG^{\mathcal{D}}(C, Z) = S^*(K^{\mathcal{D}}(C)) - \sum_i P(Z = z_i) S^*(K^{\mathcal{D}}(C|Z = z_i)),$$

这个算法的特点在于它是基于不确定性最大化的准则，因此人为在分母加上一点扰动。