False Positive Ratio and Relevance Sets

The purpose of this note is to show that it is possible to get a very good GINI (ie high) and a very poor FPR (ie high) and that this is related to the ratio of Legal to Fraud in the Relevance Set.

There are various definitions of the FPR but we assume it is defined as:

$$FPR = \frac{FP}{TP}$$

The FPR measures the ratio of incorrect alerts to correct alerts. From an operational point of view we want the FPR to be as low as possible. An acceptable level is considered to be about 30.

Now consider the True Positive Fraction (TPF) and the False Positive Fraction (FPF) taken at a particular threshold:

$TPF = \frac{TP}{N_F}$ and $FPF = \frac{FP}{N_L}$. It is easy to see that $FPR = \frac{FPF}{TPF} \cdot \frac{N_L}{N_F}$ (1)

It is important to note that the ratio $\frac{FPF}{TPF}$ *is independent of both* N_F *and* N_L*.*

To see this consider a classifier that outputs a score *s*.

By recording the score when a transaction is Legal or Fraud it is possible to determine the distributions of $P(s \mid F)$ and $P(s \mid L)$.

For a particular threshold *t* we can then write:

$$TP = N_F \int_0^t P(s \mid F) ds \quad FN = N_F \int_t^1 P(s \mid F) ds$$
$$FP = N_L \int_0^t P(s \mid L) ds \quad TN = N_L \int_t^1 P(s \mid L) ds \qquad (2)$$

Now consider the True Positive Fraction (the Y-axis of the ROC graph) and the False Positive Fraction (the X-axis of the ROC graph).

$$TPF = \frac{TP}{TP + FN} = \frac{\int_0^t P(s \mid F) ds}{\int_0^t P(s \mid F) ds + \int_t^1 P(s \mid F) ds} = \frac{\int_0^t P(s \mid F) ds}{\int_0^1 P(s \mid F) ds} = \int_0^t P(s \mid F) ds \qquad (3)$$

$$FPF = \frac{FP}{FP + TN} = \frac{\int_0^t P(s \mid L) ds}{\int_0^t P(s \mid L) ds + \int_t^1 P(s \mid L) ds} = \frac{\int_0^t P(s \mid L) ds}{\int_0^1 P(s \mid L) ds} = \int_0^t P(s \mid L) ds \qquad (4)$$

So both TPF and FPF can be expressed in terms of the probability distributions and independently of the amount of fraud and legal.

From equations (3) and (4) it is possible to derive an expression for the GINI based on assumptions of the form of the probability distributions. There is no need to make any assumptions about N_F or N_L.

Returning to equation (1), $FPR = \frac{FPF}{TPF} \cdot \frac{N_L}{N_F}$ we can therefore see that, despite a very good GINI, if the ratio of legal to fraud is very large then this will cause the FPR to be very poor.

This observation is important when designing rules and considering their Relevance Sets. If the rule perceives (via its relevance set) a high legal/fraud ratio then although it may have a good GINI it will not behave well.
The GINI is measuring the ability to discriminate but if the relevance set presents very few opportunities to get it right, relative to the number of opportunities to get it wrong, then it will obviously get it wrong more often and have a poor FPR.

Note:

> When looking down a threshold table for a classifier the last entry is for $TPF = FPF = 1$. From equation (1) it is obvious that the FPR column for this entry is the ratio of legal to fraud in the rule's relevance set.

Classifier Optimisation

When adjusting a classifier's parameters using the TPF at an Alert Rate as the performance criteria, is this the best criteria to use ?

Given the $FPR = \frac{FP}{TP} = \frac{FPF}{TPF} \cdot \frac{N_L}{N_F}$ and $AR = \frac{TP + FP}{N_F + N_L}$ then a little algebra gives the following relationships:

$AR = TPF \frac{1 + FPR}{1 + k}$ where $k = \frac{N_L}{N_F}$, and $TPF = AR \frac{1 + k}{1 + FPR}$

So, for a given *AR* and *k*, a minimal *FPR* will give a maximal *TPF*.

When looking at the overall-statistics, as oppose to the performance of individual rules over their relevance-set, then the value of *k* is constant and equal to overall ratio of fraud to legal.

So maximising the TPF is equivalent to minimising the FPR in these circumstances.

Overall GINI

In a rule-based classifier the GINI's of individual rules should be treated with deep suspicion for the reasons outlined above. The same caveats can not be applied to the overall GINI. As there is a known and constant legal/fraud ratio.

