FRAUD DETECTION THEORY DOC

MICHAEL MCLEOD

1. The Problem

We want to be able to estimate the probability of a given transaction being a fraud. Each transaction is tested against a set of rules, any number of which it may trigger. The goal is to estimate the fraud probability given the rules which have been triggered, and some aggregated historic data from past transactions. This historic data contains the total number of transactions, the total number of fraudulent transactions, and for each rule in the rule set the number of times the rule was triggered by fraudulent transactions and the number of times the rule was triggered by genuine transactions.

2. A Simple Bayesian Approach

Let f be the proposition that the current transaction is a fraud,

$$f = \begin{cases} 1, & \text{transaction is a fraud} \\ 0, & \text{transaction is genuine.} \end{cases} \tag{1}$$

We want to calculate

$$P(f|S,H), \tag{2}$$

where S is the set of rules triggered and H is the historic data. Let

$$R = \{r_0, ..., r_n\} \tag{3}$$

be the set of rules available in the system; then

$$S \in 2^R. \tag{4}$$

Using Bayes' theorem we then want to calculate the posterior probability of a fraud

$$P(f|S,H) = \frac{P(S|f,H)P(f|H)}{P(S|H)}. \tag{5}$$

We shall tackle each term in turn.

2.1. **The Prior.**

The prior term is

$$P(f|H), \tag{6}$$

the probability of the transaction being a fraud given the historic data, but in lieu of rule information (or, rather, integrated over it). Here we will make an assumption that there exists some overall fraud rate F with which frauds occur, such at $F \in [0, 1]$ represents the fraction of fraudulent transactions. Then

$$P(f|F) = F. \tag{7}$$

$$P(f|H) = \int_0^1 P(f|F)P(F|H)\, dF = \int_0^1 FP(F|H)\, dF \tag{8}$$

Since F is unknown to us we shall assign it a flat prior, $P(F) = 1$, and therefore

$$P(F|H) = \frac{P(H|F)P(F)}{P(H)} = \frac{P(H|F)}{\int_0^1 P(H|F)dF} \tag{9}$$

Assuming that transactions are all independent then

$$P(H|F) = \binom{n_f + n_g}{n_f} F^{n_f}(1 - F)^{n_g}\, dF, \tag{10}$$

where n_f is the total number of fraudulent transactions, and n_g is the total number of genuine transactions. The total number of transaction overall is therefore $n_f + n_g$.

Then we have for the denominator

$$\int_0^1 P(H|F)P(F) = \binom{n_f + n_g}{n_f} \int_0^1 F^{n_f}(1-F)^{n_g}\, dF = \binom{n_f + n_g}{n_f} B(n_f + 1, n_g + 1), \tag{11}$$

where $B(x, y)$ is the Beta function. Expanding the Beta function in terms of the Γ-function, and substituting the Γ-function for factorial we have:

$$\int_0^1 P(H|F)P(F) = \frac{(n_f + n_g)!}{n_f!\, n_g!} \frac{n_f!\, n_g!}{(n_f + n_g + 1)!} = \frac{1}{n_f + n_g + 1} \tag{12}$$

Substituting equation 12 and equation 9 into equation 8 leaves us with

$$P(f|H) = (n_f + n_g + 1)\binom{n_f + n_g}{n_f} \int_0^1 F^{n_f+1}(1 - F)^{n_g}\, dF. \tag{13}$$

This is integrated similarly to yield

$$P(f|H) = (n_f + n_g + 1)\frac{(n_f + n_g)!}{n_f!\, n_g!} \frac{(n_f + 1)!\, n_g!}{(n_f + n_g + 2)!} = \frac{(n_f + 1)}{(n_f + n_g + 2)}. \tag{14}$$

The probability of a genuine transaction $\bar{f}$ is calculated in the same way to yield

$$P(\bar{f}|H) = \frac{(n_g + 1)}{(n_f + n_g + 2)}, \tag{15}$$

which satisfies the equation $P(f|H) + P(\bar{f}|H) = 1$ as necessary.

2.2. **The Likelihood.**

This term is

$$P(S|f, H), \tag{16}$$

the probability of the given rule set S (and only S) being triggered by the transaction, given that the transaction is fraudulent and the historic data.

Assuming that each rule is triggered independently, the

$$P(S|f, H) = \prod_{r \in S} P(r|f, H) \times \prod_{r \in R-S} \left(1 - P(r|f, H)\right), \tag{17}$$

i.e. the product of the probability of being triggered for each triggered rule (in S) and the probability of not being triggered for each rule not triggered (in R but not in S).

Again we do not know the probability of a fraudulent transaction triggering a given rule, so we take the same approach as we did above and assume that such a rate exists, and then marginalise over it.

$$P(r|f, H) = \frac{n_{r,f} + 1}{n_f}, \tag{18}$$

where $n_{r,f}$ is the number times rule r was triggered by a fraudulent transaction.

2.3. **The Evidence.**

This term is

$$P(S|H) = P(S|f, H)P(f|H) + P(S|\bar{f}, H)P(\bar{f}|H) \tag{19}$$

The term $P(S|\bar{f}, H)$ is calculated in the same was as the likelihood, but for the probability of rules being triggered by a genuine transaction,

$$P(S|\bar{f}, H) = \prod_{r \in S} P(r|\bar{f}, H) \times \prod_{r \in R-S} \left(1 - P(r|\bar{f}, H)\right), \tag{20}$$

$$P(r|\bar{f}, H) = \frac{n_{g,r} + 1}{n_g + 2}. \tag{21}$$

The posterior may be calculated from a combination of the terms above.