Spoofing Detection Documentation

Sadrach Pierre

January 7, 2019

1 Data Collection

Collect order book data with timestamps, order IDs, status (NEW, CANCELLED, MATCHED), price, side, and quantity. For the time being data was artificially generated from random normal distributions and used to train the gaussian model.

2 Data Cleaning

For the next Data Scientist Hire Once the required Level 2 order book data is obtained follow these steps in order to clean the data. First calculate the difference in price between cancellation price and best ask/bid price. As prescribed by Leangarun et al. define the cancellation price as $P^{cancel}(t)$ and the matched price as $P^{matched}(t)$ of buy/sell orders. Current bid/ask price is represented by $P(t)$. $V^{cancel}(t)$ and $V^{matched}(t)$ is denoted as the cancellation volume and matched volume of buy/sell order, respectively. The delta price of each time frame is calculated using the following: This is a measure of the price difference between the best bid/ask and the cancelled price.

$$P_{TF} = \sum_{t=1}^{n} \frac{|P^{cancel}(t) - P(t)|V^{cancel}(t)}{V^{cancel}(t)} \tag{1}$$

where $t = 1, 2, 3, , \ldots, n$ is the length of the sliding window. The expected value of the matched price is

$$\mathrm{E}[P^{matched}] = \sum_{t=1}^{n} \frac{P^{matched}(t)}{n}. \tag{2}$$

Finally ΔP is defined as,

$$\Delta P = P_{TF}/\mathrm{E}[P^{matched}] \tag{3}$$

Discard samples with large values of ΔP. Values with large ΔP correspond to cancelled orders that are far from the best bid/ask.

In the end you'll have one-minute snapshots of order book data with small ΔP's which means the cancelled and match orders are close to the best bid/ask.

2.1 Feature Engineering

The key features for detecting spoofing will be one-minute snapshots taken of $V^{cancel}(t)$ and $V^{matched}(t)$ taken from the cleaned data.

2.2 Labeling Data

The data should be labelled such that feature input containing matched and cancelled volumes far from the average matched and cancelled volumes are cases of spoofing.

2.3 Gaussian Model

The corresponding expectation values will be used in the gaussian model:

$$\mathrm{E}[V^{matched}] = \sum_{t=1}^{n} \frac{V^{matched}(t)}{n}. \tag{4}$$

$$\mathrm{E}[V^{cancelled}] = \sum_{t=1}^{n} \frac{V^{cancelled}(t)}{n}. \tag{5}$$

The probability density function of a 2-dimensional Gaussian distribution is given by,

$$p(x;\mu,\Sigma) = \frac{1}{(2\pi)|\Sigma|^{1/2}}\exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)) \tag{6}$$

where μ is a vector $[\mathrm{E}[V^{matched}], \mathrm{E}[V^{cancelled}]]$ and Σ is a covariance matrix for $V^{matched}$ and $V^{cancelled}$.