Real-time Anomaly Detection on Large-scale Data

Yuqian Song  
Barclays PLC.  
1 Churchill Place  
London, E14 5HP  
yuqian.song@barclays.com

Eirini Spyropoulou  
Barclays PLC.  
1 Churchill Place  
London, E14 5HP  
eirini.spyropoulou@barclaycard.com

Bassel Ojjeh  
LigaDATA  
Palo Alto, CA  
bassel@ligadata.com

ABSTRACT
Anomaly detection has become an emerging topic of great importance across a wide range of business units in financial institutions. Examples are trading surveillance, cyber security, fraud detection and human resources. The major challenge when building anomaly detection solutions in the context of is that they need to be real-time, in order to support real time decisioning, while being able to support training over huge amounts of data. This paper aims to introduce a real-time anomaly detection solution built on open source big data technologies along with two business use cases on large-scale data: one is for trading surveillance to identify unauthorized trading behaviour and the other is related to cyber security, identifying anomalies in network traffic or human behaviour. The solution also demonstrates the design of a generic framework applicable for different use cases in the future.

Keywords
- real-time
- anomaly detection
- open source
- big data
- machine learning

1. INTRODUCTION
This paper describes an anomaly detection solution powered by an open source real-time decision engine [1] and combining the insights extracted via hybrid machine learning approach with domain expertise. This solution has been applied to two use cases and implemented using open source big data technologies.

2. Unauthorized Trading Use Case
The unauthorized trading use case aims to identify the abnormal behaviours of individual traders based on their historical data and peer comparison. In this use case, the trading data is highly enriched and aggregated with a wide range of reference data in order to gain more comprehensive view of the behaviour of individual traders. Models are created based on trader/book level data to evaluate potential risk in real time.

3. Cyber Security Use Case
Here we present a system that consumes various internal and data sources and builds models which are able to identify anomalous network traffic and human behaviour. Semi-supervised learning models are built for this purpose using historical data. Human curated rules are used to label the data.

4. Real-time Anomaly Detection Framework
A general framework is designed for the real-time anomaly detection solution (see Figure 1).

5. Future Work
The development and implementation of this solution are still in progress and it is expect to be evaluated and benchmarked with large-scale of financial data in different use cases.

6. REFERENCES