Course Project

The course project is a group project comprising three separate parts as detailed below. Each group forms a *team* of 4-5 members who will closely work together. The project is to be completed and submitted by **July 30, 2017**. A final presentation follows in the last week of classes, essentially during regular class hours. In addition there will be a short testing session to assess the quality of the technical solution; this will also be arranged in the last week of classes according to availability.

Project Scope. In light of increasing cyber threats, especially advanced persistent threats, and existing vulnerabilities that expose critical infrastructure to a variety of adversarial scenarios, the project explores behaviour-based intrusion detection to enhance cyber situational analysis and facilitate early warning for suspicious anomalies in order to mitigate the impact of attacks by launching countermeasures and also to facilitate cyber forensics.

Methodology. Intelligent monitoring and control of critical infrastructure such as electric power grids, public water utilities and transportation systems produces massive volumes of *time series data* from heterogeneous sensor networks. Whenever comprehensive historic data from past operation of critical infrastructure is available, big data analytics is a sensible approach to advanced anomaly detection and has been studied in the scientific literature. For instance, various types of Hidden Markov Models (HMMs) have been proposed as a formal basis in predictive analytics to represent normal operation in a compact form allowing to effectively differentiate between normal behaviour and anomalies.

Challenges. A number of inescapable 'external factors' can make anomaly detection in time-series data challenging whenever data originates from the operation of a real-world system. Typical examples include: imperfections in the data, such as missing or corrupted values; lack of 'ground truth' in historic data, meaning labels to differentiate normal data points from outliers are unavailable; various types of anomalies to deal with, depending on the application context; striking a good balance between *precision* and *recall*, specifically also reducing the false alarm rate to make anomaly detection practical in a real application context with resource constraints.

Data Source. In this project we use variations of a real data set obtained from monitoring household power consumption for some part of the U.S. electrical power grid. The data is available in compressed form from the course page and comes in several data sets, a training data set and several test data sets (with more test sets coming further down the road). You can use any tool and language. We recommend R or Python to those who are looking for a easy-going and powerful starting point. They have so many analytic packages which make your analysis easier. For example: seqHMM or MHSMM are some popular HMM packages in R; and Python has many HMM packages like np-HMM and so on.

Project Description. The project splits into three different parts, each comprising a number of different tasks as follows:

Part 1: Data Analytics

This part is about understanding the specifics of the data and developing an analytic approach to anomaly detection for different types of anomalies. You may choose to follow the methodical guidelines of the public water utility anomaly detection paper presented in class, or choose a different approach if you are confident this can work as well.

Part 2: Project Report

Each project team is supposed to describe their methodical approach, their experimental analysis, and the key findings from the experiments in the form of a technical report. Details about what is expected from a technical report in terms of overall structure and logical organization will be provided in a separate document. Generally, a technical report is a clearly written, well structured document to communicate technical ideas and insights to an audience with a technical background.

Part 3: Presentation

Each project team will present the outcome of their work in a 15 minutes presentation in class during the last week of classes. This means to properly summarize the essence of your course project in a formal presentation using slides with intuitive textual and graphical illustrations. Only two members of each team will actively present while the other team members only have to answer questions.

The marking of your course project will ultimately take into account all three parts: the breadth and depth as well as the quality of the technical solution, project report, and presentation.

Getting Started. With the information provided here you should be able to get your team organized and your project under way. You will receive additional input verbally and in writing as you proceed through your project. A clear breakdown of tasks and responsibilities among team members certainly helps in developing a clear roadmap allowing the team to work more productively. Please take into account though that the team as a whole is responsible for their project and team members are expected to help each other in managing project tasks.

Finally, we hope that you will find this project an enjoyable and rewarding experience after all.

Thank you in advance for your cooperation!