MISSISSIPPI STATE UNIVERSITY

HIGH PERFORMANCE DATA ANALYTICS

DASI

Anomaly/Event Detection in High-Velocity Streaming Data

CHALLENGE

With an ever-increasing necessity of smarter threat detection systems in cyber security, a data-driven approach towards pre-cognition of events/anomalies can provide time-critical reactivity to counter complex and real-time threats. In such high throughput systems, traditional batch processing approaches are insufficient to provide real-time insights into the data. Such high-velocity data streams require Big-Data based parallelization strategies for detecting trends. Architectures modeled after such an approach, can leverage the predictive capabilities to build smarter threat/intrusion detection systems with a proactive approach towards security of systems.

CURRENT PRACTICE

Traditional models of trend detection are applicable to a wide variety of security systems such as intrusion detection, security event management, and threat detection systems. However, most of them are based on parametric statistical models in which the single feature of bursty traffic is compared to a pre-determined threshold. Traditionally Poisson processes are used to establish a baseline Poisson rate for a sliding window, and any deviation from the baseline is categorized as an event. Another significant portion of approaches rely on signature-based event-detection, where a previously identified event pattern is matched with a new pattern to report occurrence of an event.

The weakness in both the methods is the inability to identify events before they become a trend. With the onset of high throughput systems generating large volumes of high velocity data, such traditional approaches cannot be used for reacting (enforcing necessary security policies or counter measures) to cyber attacks in real-time.

Developing highly parallelizable architectures for ananomaly/event prediction in high-velocity streaming

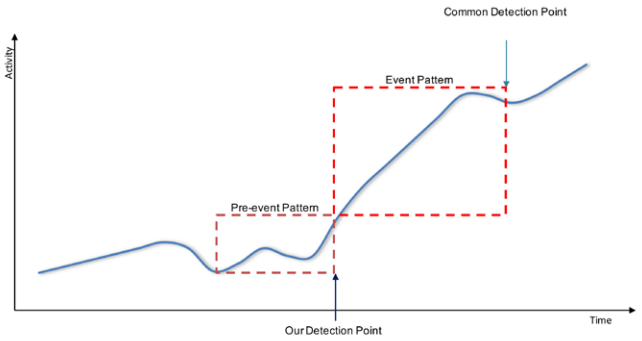

Figure 1: Anomaly/Event detection in current state-of-art vs the proposed research approach

TECHNICAL APPROACH

The research aims at developing real-time multi-dimensional event/anomaly prediction systems based on pre-event data-trained machine learning models. More specifically, the proposed research plans to create a highly scalable supervised machine-learning models based on the training of pre-event patterns leading up to an event. The training set for the models will be extracted from the test datasets (described later) and stored with a binary classification of: 1) non-event patterns, and 2) event patterns. An occurrence of a new pattern from the streaming data can then be compared with the previously classified patterns, and distance vectors from

each of the older patterns will be computed. These distance vectors (Euclidean) will serve as the input to the machine-learning model for codification/ classification of the class to which the observed pre-event pattern belongs to. Using the classification, the model will then be able to predict the upcoming trend (event/non-event, type/pattern of event) of the new observation.

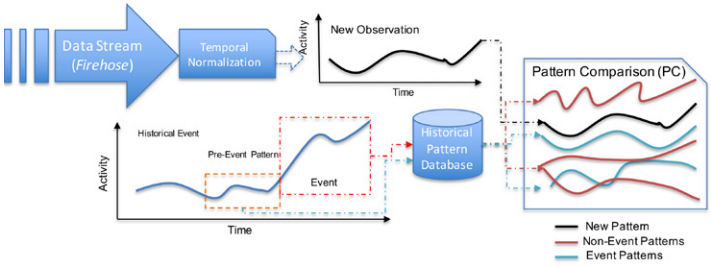

Figure 2: Data-driven approach towards anomaly/ event prediction

IMPACT

The research envisions the multi-dimensional non-parametric models can be applied to gain critical insights in the trends of other large-scale streaming data sources. For example, anomalies in sensor systems, web-servers, honeypots, packet routers, etc. can be tracked to determine increases in potential hacking activity. Online-learning models can be utilized to create models that can learn in real-time instead of relying on sets of pre-trained features. Applicability of the prediction model to develop smarter intrusion/threat detection systems can also be evaluated. Other application areas include implementation of game theory algorithms on top of the prediction models to develop self-adjusting intelligent decision engines (for example, traffic congestion, load-balancing severs, etc.). Using a scalable/distributed machine learning approach will also help researchers understand and establish paradigms for the applicability of disparate machine learning algorithms in large-scale streaming data.

Towards the overarching goal of the project, the research challenges that need to be addressed are:

- Identification of features/dimensions in traffic for extraction, normalization and prediction
- Identification of machine-learning models for high-dimensional data
- Parallelization and scaling strategies using High Performance computing at Mississippi State University (HPC²) test bed
- Evaluation of online learning schemes for dynamic models
- Describing a generic pipeline for universal applicability

Contacts

Somya D.Mohanty
Assistant Research Professor
662.325.3356
mohanty@dasi.msstate.edu

Funding provided by a grant from Pacific Northwest National Laboratory.