Scalable and Generic Framework for Automated Time-Series Anomaly Detection

Dr. Chetan J. Shelke[1], Suraj Dhole[2], Miss. Ankita S. Nathe[3]
Department of Computer Science & Engineering
P. R. Pote (Patil) College of Engineering and Management, Amravati

Abstract - Early detection of anomalies plays a key role in maintaining consistency of person's data. Current state of the art anomaly detection approaches suffer from scalability, use-case restrictions, difficulty of use and a large number of false positives. The Extensible Generic Anomaly Detection System enables the accurate and scalable detection of time-series anomalies. EGADS separates forecasting, anomaly detection and alerting into three separate components which allows the person to add in their own models into any of the components. EGADS is the first comprehensive system for anomaly detection that is flexible, accurate, scalable and extendible by comparing automated generic and scalable time series approach against other anomaly detection systems on real and synthetic data with varying time-series characteristics.

Keywords - Anomaly detection, Extensible, Generic, Outlier, Time-series

I. INTRODUCTION

To protect our data from malicious attack numbers of applications are available still lot's of software getting disturb by malicious attack. Anomaly detection is about finding the normal practice patterns from the audit data, while misuse detection is on the subject of encoding and matching the intrusion patterns via the audit data. Anomaly detection refers to the important problem of finding non-conforming patterns or behaviors in live traffic data. These non-conforming patterns are often known as anomalies, outliers, discordant observations, exceptions, aberrations, surprises, peculiarities or contaminants in different application domains.

Efficient identification of outlier is useful in many applications like credit card fault, medical diagnoses and other [1].Detecting outliers or anomalies in data have been studied in the statistics community as early as the 19th century [Edgeworth 1887]. Over time, a variety of anomaly detection techniques have been developed in several research communities. Many of these techniques have been specifically developed for certain application domains, while others are more generic [4]. Recent approaches suffer from lots of false positive.

Use-case, anomaly detection model may provide low false positive rate [4] but still it's not able to detect anomaly at high rate. Whereas EGADS (Extensible Generic Anomaly Detection System) enables the accurate and scalable detection of time-series anomalies. EGADS separates forecasting, anomaly detection and alerting into three separate components which allows the person to add her own models into any of the components. It uses a set of default models that are tuned to reduce the number of false positives, which by itself suffices for the average user [4].

The anomalies of interest may vary in magnitude, severity or other parameters which are unknown a priori and depend on the use-case. For this reason the alerting component of EGADS uses machine learning to select the most relevant anomalies for the consumer.

II. IMPLEMENTATION

The EGADS framework consists of three main components: the time-series modeling module (TMM), the anomaly detection module (ADM) and the alerting module (AM). There are a number of anomaly detection systems [1,3,4] but they are use-case specific and are not extendible [1, 3]They do not service the needs of many projects at Yahoo which require different types of anomaly detection with the additional requirements that the models be easily configurable, scalable, and dependence-free for simple deployment. Whereas EGADS is build in such a manner so that it can easily integrate with Yahoo Monitoring System (YMS). YMS processes millions of data points every second. So it's critical to provide accuracy, scalability and automated anomaly detection. EGADS detect multiple types of anomalies. It is composed of three stages forecasting, anomaly detection and filtering. In this system features are divided into two parts first one is time series feature and second one is model features.

System Integration - EGADS is integrated with YMS. A key constraint of YMS is scale; the platform needs to evaluate millions of data points per second. As a result, many of the integration architecture decisions are focused on optimizing real-time processing. The integration with YMS is shown in Figure 1. Several support components are required to drive action based on detected anomalies. First of all, all anomaly detection models are generated in batch and applied in real time.

The batch flow is comprised of three steps:

1. Telemetry (i.e. the monitored time-series) data are stored in bulk on a Hadoop cluster.
2. A batch model generator runs against these data and builds models for targeted time-series.
3. The models are stored in a model database.

The online flow then utilizes the stored models:

1. Data flows into a Storm stream-processing topology.

2. One of the bolts (modules) in the topology calls the EGADS ADM to evaluate incoming data points based on models stored in the model database.
3. If an anomaly is present, this is fed to a secondary rule flow, consisting of combinatorial rules and other use case specific logic.
4. Based on the rules, if the anomaly is an alert event, the event is generated, stored in a status database, and forwarded to an alert routing system.
5. The alert routing system applies routing configuration rules to send the alert to the appropriate support staff.

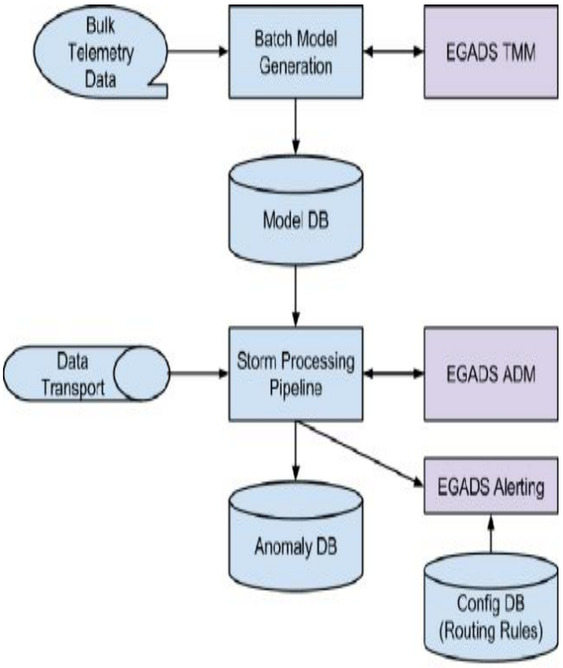

Figure 1: EGADS-YMS Architecture

III. TECHNIQUES UES FOR ANOMALY DETECTION

Currently EGADS is capable of detecting three classes of anomalies outlier, change point and anomalous time-series.

A. Outlier Detection - EGADS uses two methods for detecting outlier plug-in-method and decomposition method.

i). Plug-in-method: Detecting outlier is one of the important function in many monitoring application. This is the first method for outlier detection. This method detect outlier by clearly monitoring normal behavior with little difference consider as outlier, to measure normal behavior EGADS uses wide range of time series model and forecasting model.

Time Series is defined as a set of observations taken at a particular period of time. For example, having a set of login details at regular interval of time of each user can be categorized as a time series. On the other hand, when the data is collected at once or irregularly, it is not taken as a time series data.

Plug-in-method consists of two components TMM and ADM in EGADS. TMM can make prediction on bases of some rule based system which gives expert knowledge about how anomalies time series data behave in particular time span.

ADM compute some notation of deviation which are refer to as deviation matrix. The simplest measure of deviation is the prediction error. If the error falls outside some fixed thresholds, an alert is issued.

ii). Decomposition-based methods: It’s a second method for outlier detection based on idea of time-series decomposition. Decomposition of time series is done into three parts trend, seasonality and noise on bases of time domain and frequency domain.[7, 8]

B. Change Point Detection - EGADS provide another approach for change point detection by using relative or model based methods.

C. Detecting Anomalous Time-series - In EGADS clustering the time-series by using various time-series features including trend and seasonality, spectral entropy etc. into set of cluster. On that set intra or inter-cluster anomalous time series is perform by measuring deviation within or among the cluster.

IV. ALTERING

The final state in anomaly detection is to detect accurate and timely alert. For that purpose EGADS uses two process threshold selection and filtering.

A. Threshold Selection - The threshold selection is use to select proper threshold on deviation matrix. EGADS uses two approaches for threshold selection parametric and non-parametric. The parametric approach is use when deviation matrix is properly distributed here assumption is made that data is normally distributed and by using ‘three sigma rule’ one can select the appropriate threshold. The second approach specifies that the deviation matrix is not properly distributed. The basic idea is to find low density regions of the deviation metric distribution. One approach is to use an algorithm such as Local Outlier Factor (LOF) [3] which is based on a concept of a local density, where locality is given by nearest neighbors, whose distance is used to estimate the density. By comparing the local density of an object to the local densities of its neighbors, one can identify regions of similar density, and points that have a substantially lower density than their neighbors. These are considered to be outliers.

B. Filtering - Filtering performs the last stage post-processing on the anomalies which are then delivered to the consumer. While the candidate anomalies, which are the input to the filtering stage, are statistically significant, not all of them will be relevant for a particular use-case [2].

V. CONCLUSION

Thus this paper introduce the system called EGADS(Extensible Generic Anomaly Detection System), which enable accurate ,scalable, flexible and extendible detection of time-series anomalies by separating forecasting, anomaly detection and altering into three separate components. It is mostly unsupervised. Basically it is divided into two main features time-series feature and model feature. Both the data and the framework are being open-source.

VI. REFERENCES

[1]. V. J. Hodge and J. Austin, "A Survey of Outlier Detection Methodologies," *Artificial Intelligence Review*, vol. 22, no. 2, pp. 85–126, 2004.

[2]. N. Laptev, S. Amizadeh, and I. Flint, "Generic and Scalable Framework for Automated Time-series Anomaly Detection," *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 15*, 2015.

[3]. "Introduction," *Fourier Analysis of Time Series Wiley Series in Probability and Statistics*, pp. 1–8, 2004.

[4]. B. Ng, "Survey of Anomaly Detection Methods," Dec. 2006.

[5]. Y. Kawahara, T. Yairi, and K. Machida, "Change-Point Detection in Time-Series Data Based on Subspace Identification," *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, 2007.

[6]. V. Moskvina and A. Zhigljavsky, "An Algorithm Based on Singular Spectrum Analysis for Change-Point Detection," *Communications in Statistics - Simulation and Computation*, vol. 32, no. 2, pp. 319–352, Jun. 2003.

[7]. M. M. M. Fuad, "A Haar Wavelet-based Multi-resolution Representation Method of Time Series Data," *Proceedings of the International Conference on Agents and Artificial Intelligence*, 2015.

[8]. S. Aminikhanghahi and D. J. Cook, "A survey of methods for time series change point detection," *Knowledge and Information Systems*, vol. 51, no. 2, pp. 339–367, Aug. 2016.