Real-Time Anomaly Detection System for Highway Traffic

Milos Kotlar; km175003p@student.etf.rs,[1] Zaharije Radivojevic,[1] Milos Cvetanovic,[1] Marija Punt,[1] and Veljko Milutinovic[2]

[1] Computer Engineering, School of Electrical Engineering, Belgrade 11000, Serbia.
[2] Department of Computer Science, Indiana University, Bloomington 47405, USA.

Abstract

This paper proposes a real-time anomaly detection system for highway traffic data, which is able to detect well-known rule-based anomalies, defined by the domain expert, as well as anomalies or novelties that deviate from the data distribution. The proposed system gathers data from highway toll stations, and detects anomaly points and anomalous patterns using unsupervised and semi-supervised machine learning algorithms.

Introduction

Anomaly detection is important part of any security system, where anomalies diverge from an overall pattern of activity. In highway traffic data, criminal activities such as smuggling, trafficking, kidnapping, terroristic attacks, and many others could be considered as anomalous activities and should be detected in its early stages. This paper proposes real-time anomaly detection system that is able to detect early stage of suspicious activities using highway traffic data.

The proposed system is able to detect both univariate and multivariate anomalies, as well as anomalous patterns. Univariate anomalies present extreme values of a single feature, where values deviate from the rest of distribution. Multivariate anomalies present deviation of a data where multiple features combined together deviates from the rest of distribution. Depending on the environment, anomalies could be of three kinds: point, contextual, and collective. Point anomalies are single data points that lay far from the rest of the distribution. Contextual anomalies are data points that are anomalous in a specific context. Collective anomalies are data points that are not anomalies individually, but related together they are anomalies.

In the following sections, this paper discusses state-of-the-art anomaly detection algorithms existing in the open literature and describes details of the proposed system.

Existing Solutions

In the open literature, there are papers related to anomaly detection algorithms used in various applications in medicine, geoscience, intrusion detection systems, road traffic, and others [1,2,3]. For example, Wong et al. [1] proposed a rule-based anomaly detection algorithm for early disease outbreak detection. Most of these applications utilize unsupervised and semi-supervised machine learning algorithms, as well as static rules. We would like to tackle the problem of criminal activities based on highway traffic data.

Problem Statement

We defined the following anomalous patterns that could be extracted from the highway traffic data:

1. Smuggling/Trafficking - Trucks or buses heading out of country using alternative routes
2. Kidnapping/Robbery - Luxury cars speeding with no tendency to leave the country
3. Car Hijacking - Luxury cars speeding with tendency to leave the country
4. Protests – Tolls with capacity higher than usual
5. Terroristic Attack – A group of suspicious vehicles heading to the same direction
6. Criminal Organizations - A group of suspicious vehicles heading in different directions
7. Missing Vehicles - A vehicle that does not have an exit record on a particular toll
8. Suspicious Activity – A suspicious vehicle that drives the same route too often
9. Novel Activity for a Single Vehicle
10. Novel Activity for a Group of Vehicles

Proposed Solution

Real-time anomaly detection system proposed in this paper is based on python programming language and consists of two parts: simulator and detector. Simulator generates data based on simulation setup and simulates highway traffic flow for a certain period of time.

Data Model

Figure [1] presents UML data model diagram. Vehicle contains plate number, a unique identifier of a vehicle, model of vehicle, type (car, truck, or bus), and whether a vehicle is luxury or not. Toll contains name, a unique identifier, type (checkpoint or network), position (in, out, mid) that indicates where the toll is located in country, average capacity, distance and balance for other connected tolls. Pass presents an event which contains speed and timestamp beside vehicle and toll details. Vehicle also contains information about next toll and distance to the toll, which is used by simulator to simulate the highway traffic flow.

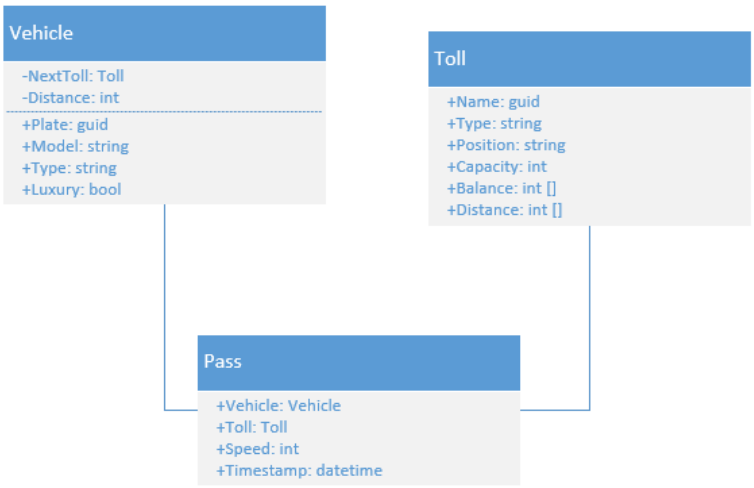

Figure 1: UML Data Model Diagram

Tolls network is an undirected graph where each path has distance and probability for taking this direction. An example of highway network where each node presents toll is shown in Figure 2.

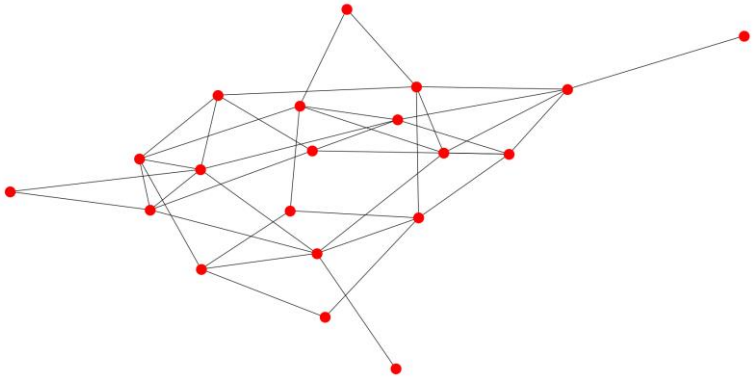

Figure 2: An example of highway network graph

Figure 3 presents BPMN diagram of proposed system. Simulator and detector runs simultaneously, where simulator generates new data. Detector fetches the generated data, runs checks, and raise alarm if anomaly is detected.

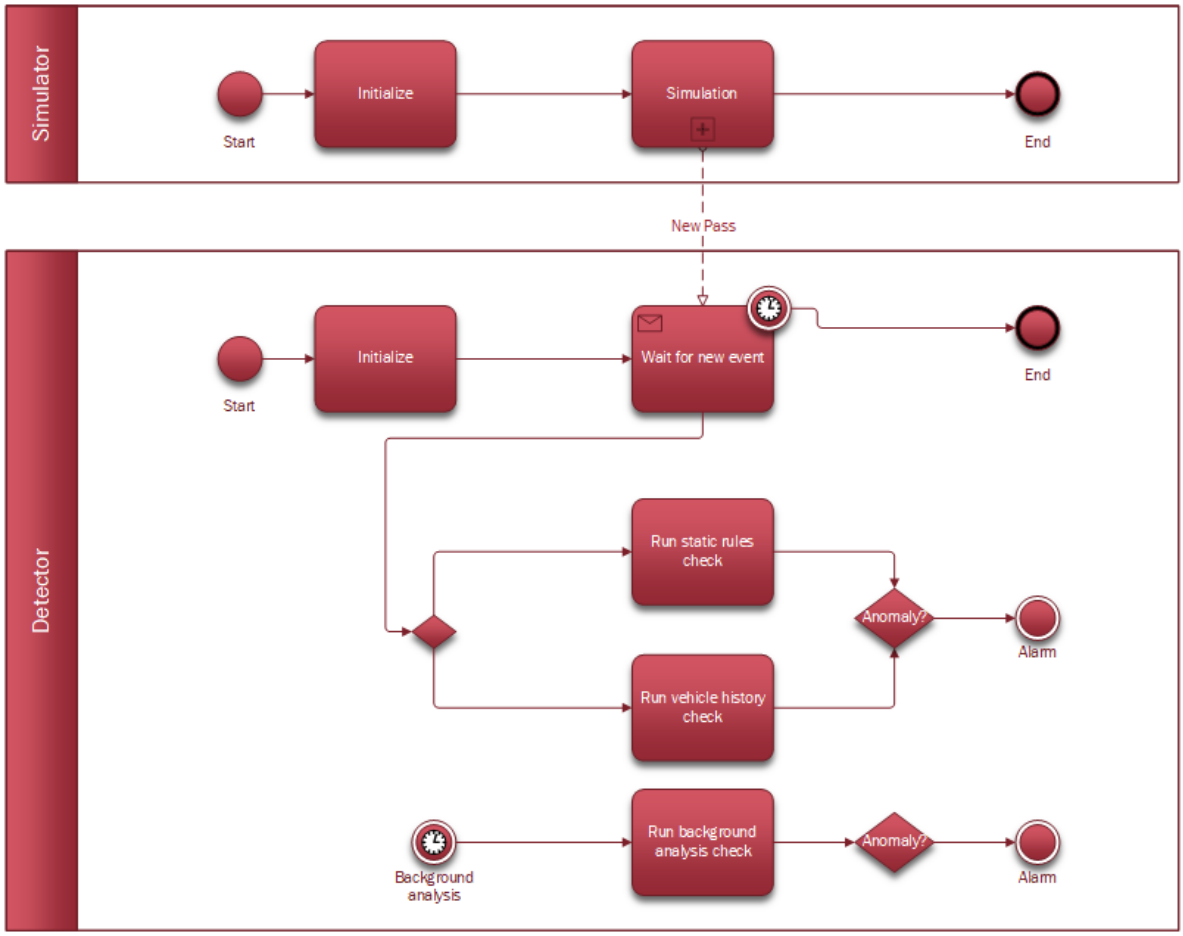

Figure 3: BPMN diagram of the proposed real-time anomaly detection system

Simulator

Simulator contains environmental details and properties of simulation, such as duration in hours, toll count, connectivity, distance range, vehicle count, and other properties about vehicles. When a vehicle passes a toll, new event is captured and saved into shared queue.

Detector

Detector reads shared queue that contains new events generated by the simulator and perform two types of analysis: static rules check and vehicle history check. Simultaneously, every *X* units of time, background analysis check is performed. Static rules are defined by the domain expert and does not involve machine learning algorithms. Static rules for the proposed anomaly detection system are presented in Table 1.

Table 1: Static rules defined by the domain expert.

Static Rule	Detection Factor		Detected Activity
Trucks or buses heading out of country using alternative routes	Toll Position Vehicle Type Capacity	In/Out Truck/Bus Low	Smuggling/Trafficking
Luxury cars speeding with no tendency to leave the country	Vehicle Speed Vehicle Type Toll Position Vehicle Luxury	High Car Mid Yes	Kidnapping/Robbery
Luxury cars speeding with tendency to leave the country	Vehicle Speed Vehicle Type Toll Position Vehicle Luxury	High Car Out Yes	Car Hijacking

Tensor Decomposition

Vehicle history check takes all records and using unsupervised machine learning algorithm named tensor decomposition detect anomalies. If anomaly is detected, alarm is raised. Tensor is intuitive structure for representing multidimensional data with temporal factor. *PARAFAC* is an algorithm that decomposes a tensor into 1-rank tensor with their components, as shown in Figure 4.

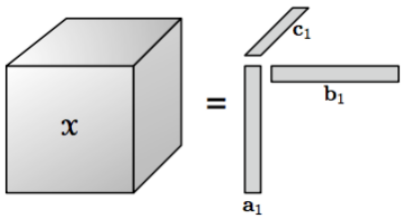

Figure 4: Tensor decomposition (PARAFAC algorithm)

Using tensor decomposition, the system can detect anomalies without prior knowledge what is normal activity pattern. Figure 5 presents tensor decomposition for vehicles and their movements. In order to form tensor *X(t, T1, T2)* we discretize timestamp in bins of one hour.

One entry of the tensor presents whether a vehicle passed *T1* coming from *T2* in hour *t*. Using such an approach we can determine whether a group of suspicious vehicles was at the same place, and prevent potential terroristic attack or identify criminal organization.

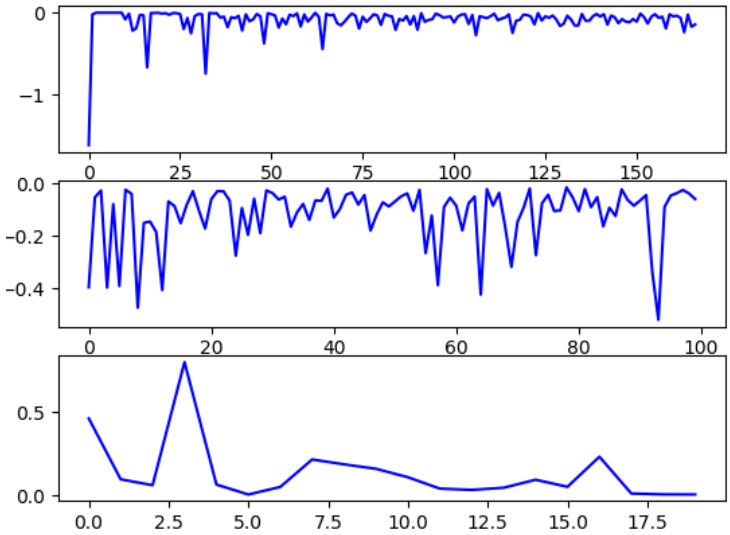

Figure 5: Tensor decomposition of tensor X(t,T1,T2)

Multivariate Gaussian Distribution

Background analysis check utilizes several unsupervised or semi-supervised algorithms. If data could be described by gaussian distribution, then we can apply multivariate gaussian distribution algorithm to detect anomalies. Model is fitted to data distribution and mean and variance is estimated for each feature. After that, using multivariate gaussian distribution we can calculate probability of each example, how similar it is to data distribution. Using such an approach we can detect if traffic volume is increased on a particular toll and thus detect anomalous patterns, like protests. Threshold could be calculated dynamically using cross-validation and F1-score, but training data is required.

K-Means

Clustering is another technique that is used for anomaly detection for collective anomalies. Initially, data is clustered into several clusters where data points with usual combination of features are together. Data points that are far from the cluster could be considered as anomalous data points. In the proposed system, k-means algorithm is used to cluster data into several clusters, and detect anomalies based on Euclidean distance, as shown in Figure 6.

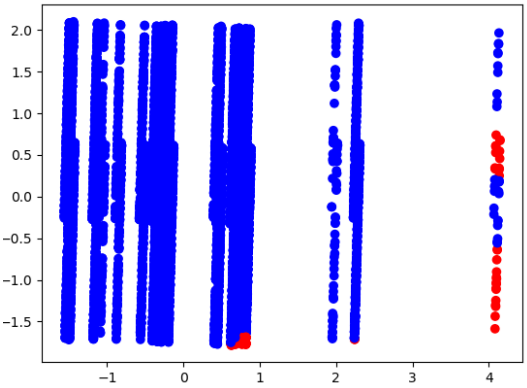

Figure 6: Detected anomalies using k-means algorithm

OneSVM and Isolation Forest

Low data density based approach, such as classification, could be used for anomaly detection. OneSVM is classifier that can predict only one class, normal class without anomalies, and thus if a data point is anomaly, it doesn't fit well into existing classification. Isolation forest, modified random forest algorithm, could be also used for anomaly detection, by measuring path length from root to the predicted class. If the path is too short, it means that data point is far from data distribution, and thus could be considered as an anomaly. The proposed system uses both algorithms for anomaly detection. In the open literature there are many other techniques that relies on low data density approach, which could be used for anomaly detection such as kNN, LOF, DBSCAN, and others.

All of these techniques should use windowing methods in order to increase overall performance. In order to support supervised learning, training data is needed, which is unavailable in this domain. The proposed system could support data labelling, which means after a certain period of time, supervised techniques could be applied. Figure 7 illustrates how different types of learning methods could be used for anomaly detection.

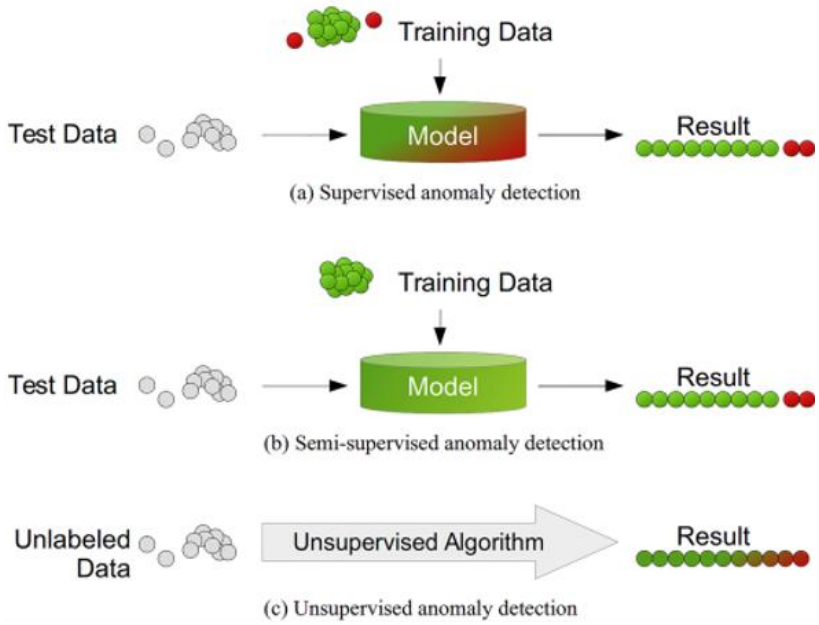

Figure 7: Machine learning for anomaly detection

[Figure taken from Goldstein M, Uchila S (2016) A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data]

Results

TBD

Conclusion

TBD

References

[1] W. K. Wong and A. Moore and G. Cooper and M. Wagner, "Rule-Based Anomaly Pattern Detection for Detecting Disease Outbreaks," *American Association for Artificial Intelligence*, pp. 217–223, 2002.

[2] V. Vercruyssen and W. Meert and G. Verbruggen et al., "Semi-supervised Anomaly Detection with an Application to Water Analytics," *International Conference on Data Mining*, 2018.

[3] D. Bruns-Smith and M. M. Baskaran and J. Ezick and T. Henretty and R. Lethin, "Cyber Security Through Multidimensional Data Decompositions," Cybersecurity Symposium, 2016.