FILIÈRE RECHERCHE

Literature Review

Elisabeth Lahalle

Ricardo Chiquetto do Lago
Maxsuel Fernandes de Almeida
Samara Ndiaye
Ana Cecília Reghini
Lucas Tramonte

September 2024

Monitoring anomalies in gamma radiation is essential for early detection of potential hazards, ensuring prompt action to safeguard public health and the environment. Therefore, it is critical to enhance detection systems to balance the need for sensitivity with minimizing false alarms. One example of its use was during the 2010 Winter Olympics. Sharma et al. (2012) [1] explore the use of machine learning algorithms for health-related radiological monitoring instead of gamma-ray spectrometers, that were used by the Canadian government and generated a significant number of false positives. The authors use the Mahalanobis distance as the anomaly detector, but they conclude that further work should include an appropriate threshold, and that a limiting factor in Machine Learning is the scarcity of data of some classes.

Geiger-Müller tube (GMT) probes are one technique used by the European Radiological Exchange Platform (EURDEP) to centralize data from radiation monitoring networks for environmental gamma dose rate (GDR) monitoring. These inexpensive sensors have a limited spectrum information, which makes anomaly detection more difficult. Count rates are usually adjusted for background noise and temperature sensitivity in GDR measurements.

Environmental factors, particularly precipitation, pose challenges in radiation monitoring. Rain can transport radionuclides, causing spikes in dose rates, with variations in precipitation and other factors complicating accurate anomaly identification. Breitkreutz et al. (2023) [2] investigate solutions to these issues. The study's algorithm improves GDR anomaly detection through a series of steps:

- Data Augmentation: Combines GDR and weather data, flagging invalid entries.
- Precipitation Peak Removal: Uses a modified Livesay model to eliminate precipitation-induced GDR peaks.
- GDR Prediction: Employs an LSTM neural network to predict baseline GDR based on cleaned data.
- Anomaly Detection: Compares actual GDR with predicted baselines using Extreme Value Theory (EVT) and hierarchical clustering.

Although precipitation removal is effective, accuracy is decreased when local precipitation measurements are replaced by cloud reflectivity data. In terms of forecasting GDR baselines, Long Short-Term Memory (LSTM) networks do well; aggregated data without autoregression or snow masking yields the best results. However, the DSPOT algorithm frequently performs better than SPOT, especially for short-term abnormalities, because of sporadic LSTM inaccuracies. Rather than abrupt shifts, the SST method works better at detecting slow-evolving anomalies.

Poirier et al. (2022) [3] propose two new methods for detecting anomalies in gamma time series: a Matrix Profile-based method and a Crossmatch-based approach. Whereas the Crossmatch approach concentrates on finding similarities in time-series data, the Matrix Profile method lowers computing complexity. Both are compared to a standard moving standard deviation method, and while the Crossmatch approach suffers with noise but still works in other cases, the Matrix Profile method has short calculation time and performs well in noisy conditions. When motifs mimic background noise, the conventional technique is less successful even though it is fast and steady. These approaches demonstrate promise for real-time environmental monitoring when tested on simulated radiological data, and the article suggests that integrating them could improve signal monitoring systems.

One eminent application of unsupervised anomaly detection is in spacecraft systems, due to the particularities of space missions, such as its remote locations. Gao et al. (2012) [4] present an unsupervised method based on normal behavior clustering that employs large amounts of unlabelled telemetry data to construct a normal behavior model by identifying and eliminating abnormal data clusters. Subsequently, the model monitors real-time telemetry data, detecting any variations as possible anomalies. This approach is useful since it identifies unknown anomalies, as it does not rely on prior knowledge. The authors also discuss the limitations of traditional techniques that are too simplistic or heavily based on expert knowledge. In contexts where manual labeling and comprehensive knowledge modeling are difficult, the unsupervised technique offers a strong foundation for anomaly identification by clustering normal behavior and utilizing a nearest-neighbor algorithm to find outliers.

Hundman et al. (2018) [5] developed a new method for detecting anomalies in spacecraft systems, using the ability of LSTM to model long time dependency and a groundbreaking method for setting dynamically the threshold above which a sequence in considered as anomalous. The authors give also a new way to mitigate the high number of false positives obtained with traditional methods, and explore the limits of traditional methods have been highlighted mostly in the case of collective and contextual anomalies.

According to Ergen et al. (2017) [6], one noteworthy strategy of unsupervised learning is the application of Long Short-Term Memory (LSTM) neural networks. This research presents algorithms based on LSTM for anomaly detection that are able to deal with variable-length data. The process consists in running sequences through an LSTM network to produce representations of fixed length, which are then used as inputs to Support Vector Data Description (SVDD) or One-Class Support Vector Machines (OC-SVM), and this joint use is achieved by gradient-based training techniques. This method allows the system to learn the best boundaries for anomaly detection and also the temporal dependencies present in the data. Comparing this strategy against conventional procedures, there have been noticeable performance gains, especially when applied to time-series data.

The use of LSTM networks for anomaly detection is further explored by Malhotra et al. (2015) [7], who apply the technique to four real-world datasets: power demand, multi-sensor engine data, ECG signals, and Space Shuttle Marotta. Based on historical data, the model predicts future time steps, and anomalies are detected when the prediction error goes over a threshold, modeled as a multivariate Gaussian distribution. The results show that LSTM networks perform better than conventional recurrent neural networks (RNNs), particularly in situations where there are long-term dependencies. The authors find that stacked LSTM networks can accurately simulate the behavior of regular time series and identify abnormalities, which makes them appropriate for a range of anomaly detection applications.

Another way to address the anomaly detection task is to trait such problem as a change-point problem. Lavielle (2005) [8] explores this problem by proposing a statistical framework that leverages penalized contrasts functions of the type $J(\boldsymbol{\tau}, \boldsymbol{y}) + \beta \cdot pen(\boldsymbol{\tau})$ to estimate the number of change points and their locations in data sequences. In this framework, $J(\boldsymbol{\tau}, y)$ is the contrast function that measures the fit of the model $\boldsymbol{\tau}$, $pen(\boldsymbol{\tau})$ is a penalty term that depends on the dimension of the model and β is a penalization parameter that controls the balance between the fit and the penalty. The key contribution of this work is the introduction of an adaptive procedure to select the penalization parameter.The goal of this adaptive process is to determine the ideal penalization parameter value that yields a reliable and accurate estimate of the number of change points. The penalized contrast approach is highly suitable for anomaly detection in data, as anomalies often manifest as abrupt changes in the statistical properties of a sequence. By automatically determining the optimal number and location of change points, this method provides a powerful tool for identifying unexpected events in time series data, making it applicable to a wide range of anomaly detection tasks.

Kawahara and Sugiyama (2012) [9] address the change-point problem using the direct density-ratio estimation technique. The authors propose a non-parametric method that directly estimates the ratio of probability densities between reference and test intervals. This method is computationally efficient, suitable for online detection, and avoids the difficulties associated with density estimation. The approach is validated using both artificial and real-world datasets, demonstrating its flexibility and accuracy in detecting changes in various time-series scenarios.

Sinaga and Yang (2020) [10] propose a new algorithm, the U-k-means, an unsupervised clustering method that addresses the limitations of traditional k-means by eliminating the need for initialization and parameter selection. This method allows users to focus on analyzing the data without the burden of predefining the number of clusters, making U-k-means a robust choice for clustering tasks in diverse fields, with the proposition of a new objective function::

$$J_{U-k-means}(z, A, \alpha) = \sum_{i=1}^{n}\sum_{k=1}^{c} z_{ik}\|x_i - a_k\|^2 - \beta n \sum_{k=1}^{c} \alpha_k \ln \alpha_k - \gamma \sum_{i=1}^{n}\sum_{k=1}^{c} z_{ik} \ln \alpha_k$$

where $\|\mathbf{x}_i - \mathbf{a}_k\|$ is the Euclidean distance between the data point $\mathbf{x}_i$ and the cluster center $\mathbf{a}_k$ and z_{ik} is a binary variable indicating if the data point $\mathbf{x}_i$ belongs to the k-th cluster.

References

[1] Shiven Sharma, Colin Bellinger, Nathalie Japkowicz, Rodney Berg, and Kurt Ungar. Anomaly detection in gamma ray spectra: A machine learning perspective. In *2012 IEEE Symposium on Computational Intelligence for Security and Defence Applications*, pages 1–8, 2012.

[2] Harald Breitkreutz, Josef Mayr, Martin Bleher, Stefan Seifert, and Ulrich Stöhlker. Identification and quantification of anomalies in environmental gamma dose rate time series using artificial intelligence. *Journal of Environmental Radioactivity*, 259-260:107082, 2023.

[3] Lisa Poirier-Herbeck, Elisabeth Lahalle, Nicolas Saurel, and Sylvie Marcos. Unknown-length motif discovery methods in environmental monitoring time series. In *2022 International Conference on Electrical, Computer and Energy Technologies (ICECET)*, pages 1–5, Prague, Czech Republic, 2022.

[4] Yu Gao, Tianshe Yang, Minqiang Xu, and Nan Xing. An unsupervised anomaly detection approach for spacecraft based on normal behavior clustering. In *2012 Fifth International Conference on Intelligent Computation Technology and Automation*, pages 478–481. IEEE, 2012.

[5] Kyle Hundman, Valentino Constantinou, Christopher Laporte, Ian Colwell, and Tom Soderstrom. Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 387–395, New York, NY, USA, 2018. Association for Computing Machinery.

[6] Tolga Ergen, Ali H. Mirza, and Suleyman S. Kozat. Unsupervised and semi-supervised anomaly detection with lstm neural networks. *arXiv preprint arXiv:1710.09207*, 2017.

[7] Pankaj Malhotra, Lovekesh Vig, Gautam M. Shroff, and Puneet Agarwal. Long short term memory networks for anomaly detection in time series. In *The European Symposium on Artificial Neural Networks*, 2015.

[8] Marc Lavielle. Using penalized contrasts for the change-point problem. *Signal Processing*, 85(8):1501–1510, 2005.

[9] Yoshinobu Kawahara and Masashi Sugiyama. Sequential change-point detection based on direct density-ratio estimation. *Statistical Analysis and Data Mining*, 5(2):114–127, 2012.

[10] Kristina P. Sinaga and Miin-Shen Yang. Unsupervised k-means clustering algorithm. *IEEE Access*, 8:80716–80727, 2020.