PhD. Assignment: Detecting Collective Anomalies in Weather Dataset

Abdul Hakmeh
abdul.hakmeh@gmail.com

***Abstract*—Anomaly detection is a technique used to identify the important and the critical events in time series data. In this work, we deal with real multivariate time series data to identify intervals which contains extreme outlier and categorize them based on the intensity level of each outlier. First, a pointwise detector is used to identify a group of outliers in the data. Then, a sliding window with predefined parameters is applied to the proposed intervals containing the intended outliers. The suggested intervals are then fed into a Kullback-Leibler divergence algorithm to quantify the degree of divergence between the distributions of a given interval and the rest of the data. After that, a list of top intervals is created to order anomalies based on their score. The results show that the algorithm was able to detect numerous outlier intervals, however, due to the lack of labeled data we could not evaluate the algorithm performance quantitatively.**

1. INTRODUCTION

Anomaly detection generally refers to the problem of finding the patterns in the data that do not contribute to expected behavior;these patterns are often named as anomalous observations or outliers. Anomaly detection is used in various applications, such as credit card fraud detection and cybersecurity intrusions. Identifying anomalies in the data is a critical task as these anomalies can be exploited to extract meaningful and important information. For example, an abnormal traffic pattern in a computer network could mean that a hacked computer sends sensitive data to an unauthorized destination and this makes the analysis of such events interesting [2].

Anomaly detection is not straightforward task due to numerous challenges such as the lack of labeled data and the similarity between anomalous data and noise. Therefore, in anomaly analysis, it is important to distinguish between anomaly detection and noise removal. Noise removal aims to remove unwanted objects before any data analysis is performed. However, methods proposed for noise removal are often used for anomaly detection and vice versa [2]. Most of the existing anomaly detection methods tend to solve a particular formulation of the problem. These formulations are usually determined by several factors, such as the availability of labeled data, the nature of the anomalies, and the application domain in which the anomaly must be detected. An additional level of complexity is added when the data contains rare events in large multivariate or in Spatio-temporal data. To overcome these challenges,

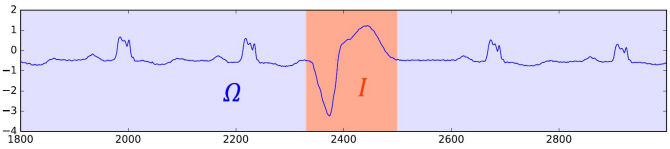

Figure 1. Schematic explanation of collective anomalies term, where a collection of data instances is anomalous I compared to the rest of the time series Ω [1]

numerous methods and techniques are presented to detect anomalies in the data. These methods can be classified into classification, nearest neighbor, clustering, and statistical techniques, among others [2].

In this work, we target the problem of detecting anomalies in a real multivariate time series data that contains three marine variables collected over 6 months. For that, we detect several short multivariate intervals with a maximum length of 7 days that represent anomalous events (Collective Anomalies). Then, we rank the detected intervals according to the intensity of the anomaly, i.e., how much these intervals differ from other normal intervals in the data. Since making a full scan of the data to identify the outliers is a demanding task, point-wise outlier detection is applied. Then, a sliding window technique is used to create intervals among the outliers points. A Kullback-Leibler divergence (KL) algorithm is used to quantify the degree of divergence between two distributions based on a closed-form expression. In the end, the extracted intervals are ordered based on their KL result.

The remainder of the paper is organized as follows: Section 2 describes our proposed method in detail. Section 3 gives an experimental evaluation with a corresponding analysis of the results. Finally, section 4 summarizes and concludes the report.

This report is organized as follows. In the next section, the basic knowledge for this work is presented by recalling the methods used. Then, an experimental evaluation is given with a corresponding analysis of the results. The last section contains a summary of our work with a conclusion.

2. METHODOLOGY

This section briefly explains the methods and techniques used in this work. Firstly, the Kullback-Leibler divergence algorithm employed to evaluate the intervals generated by a sliding window-based technique is described. Then, the local outlier factor, which is used to identify the pointwise outliers in the data, is explained.

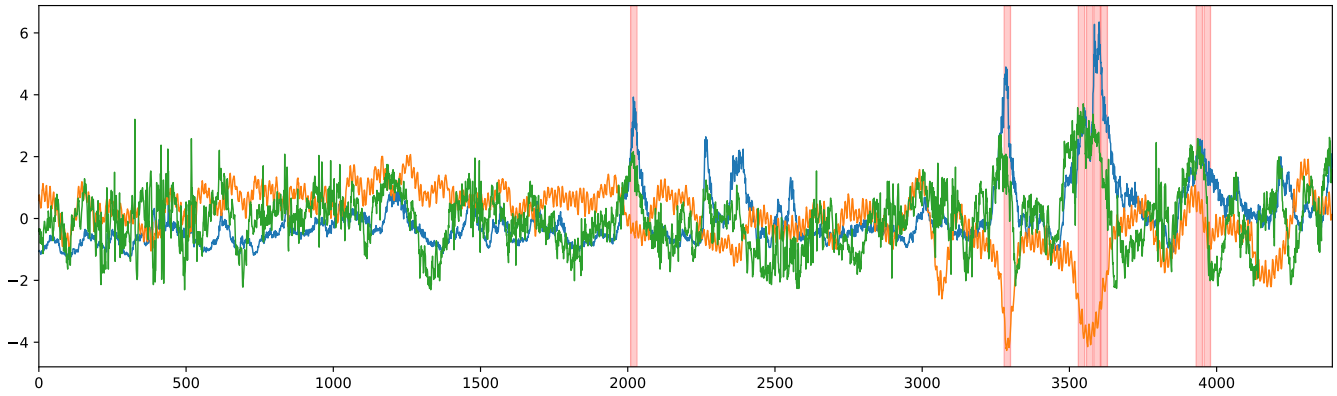

Figure 2. An example illustrates the phenomenon of multiple detections of minimum size when using the original KL divergence (note the thin lines between each detection)

2.1. Kullback-Leibler

Given a series $(x_t)_{t=1}^{n}$, where $x_t \in \mathbb{R}^d$ of n time steps. we look for intervals $I = \{t \in \mathbb{N} \mid a \leqslant t \leqslant b\} = [\,ab)$ in which its data distribution is most different from the distribution of the data in the rest of the time series. This seies of data known as collective anomalies (Feg. 1). For this task an approximated Kullback-Leibler divergence (KL) is used in a closed-form solution to quantify the degree of divergence between two distributions p_I and p_Ω, where $\Omega = [\,1, n] \setminus I$ and p is the probability density functions $p_I = \mathcal{N}(\mu_I, \sigma_I)$ and $p_\Omega = \mathcal{N}(\mu_\Omega, \sigma_\Omega)$, where μ, σ related to mean and covariance matrices respectively. The standard KL (relative entropy) takes two arrays of probabilities corresponding to two different distributions. There is a special case of KL when the two distributions being compared follow a Gaussian (bell-shaped) distributed. In such a situation, all we need to compute $KL_{(I,\Omega)}$ is the means of the two distributions and their standard deviations. The Authors of [3] have introduced a variational approximation of KL to quantify the degree of divergence between two Gaussians according to Eq. 1, where Tr related to sum along diagonals of an array and d to the number of features. High similarity of I, Ω distributions can be observed when $KL_{(I,\Omega)}$ is near zero, and vice versa. In this way we plan to order the intervals by creating a list of top anomalies, so for example the top three intervals are the intervals that have the height scores over all intervals. The number of top anomalies remains a predefined variable.

$$KL_{(I,\Omega)} = \tfrac{1}{2}[\log \tfrac{|\sigma_\Omega|}{|\sigma_I|} + \; Tr\left[\sigma_\Omega^{-1}\sigma_I\right] - d \\ +(\mu_I - \mu_\Omega)^T \sigma_\Omega^{-1}(\mu_I - \mu_\Omega)] \tag{1}$$

A full scan of the data looking for extreme outlier intervals is uninteresting and irrelevant because the anomalies are rare. To minimize the number of intervals that must be fed into the KL algorithm, we used a simple technique of selecting intervals of interest based on a set of anomaly points identified using a local outlier method. Continuous intervals are generated which contain the detections from the point-based method. The intervals are created using a sliding window with a step size of one hour and an arbitrary window length in a range between the minimum and maximum interval length specified by the user. Then, the extracted interesting intervals are entered into KL for detailed evaluation. This technique is based on the assumption that many samples within an anomalous interval will also have a high pointwise score.

In the initial experiments, we found that KL generally leads to detections of the predefined minimum size, such that larger anomalies are split among numerous successive detections (Feg. 2). This is due to the different number of samples used to estimate the interval distribution. Assuming that the data are taken from a Gaussian curve, KL becomes a random variable, but its mean depends on the length of the interval. To solve this problem, we adopted the modification proposed by [1].

$$KL_{U-KL(I,\Omega)} = 2.\,|I|\,.KL_{(I,\Omega)} \tag{2}$$

The unbiased KL divergence is based on the use of a statistical test to test the hypothesis that a given set of data points were sampled from a Gaussian with known parameters [1]. The resulting relationship is demonstrated in Eq. 2.

2.2. Local Outlier Factor

The local outlier factor is an unsupervised anomaly detection technique for finding outliers in a multidimensional time series. The method introduces a regional outlier for each object in a given data set, indicating the degree of outliers. It quantifies how far away an object is from its neighbour, when the outlier factor is local, a limited neighborhood of each object is considered. The approach is loosely related to density-based clustering, but it does not require an explicit or implicit notion of clusters.

LOF first computes the $K4$-distance between a point and its K_th nearest neighbor (k-neighbor), given by $N_k(A)$,

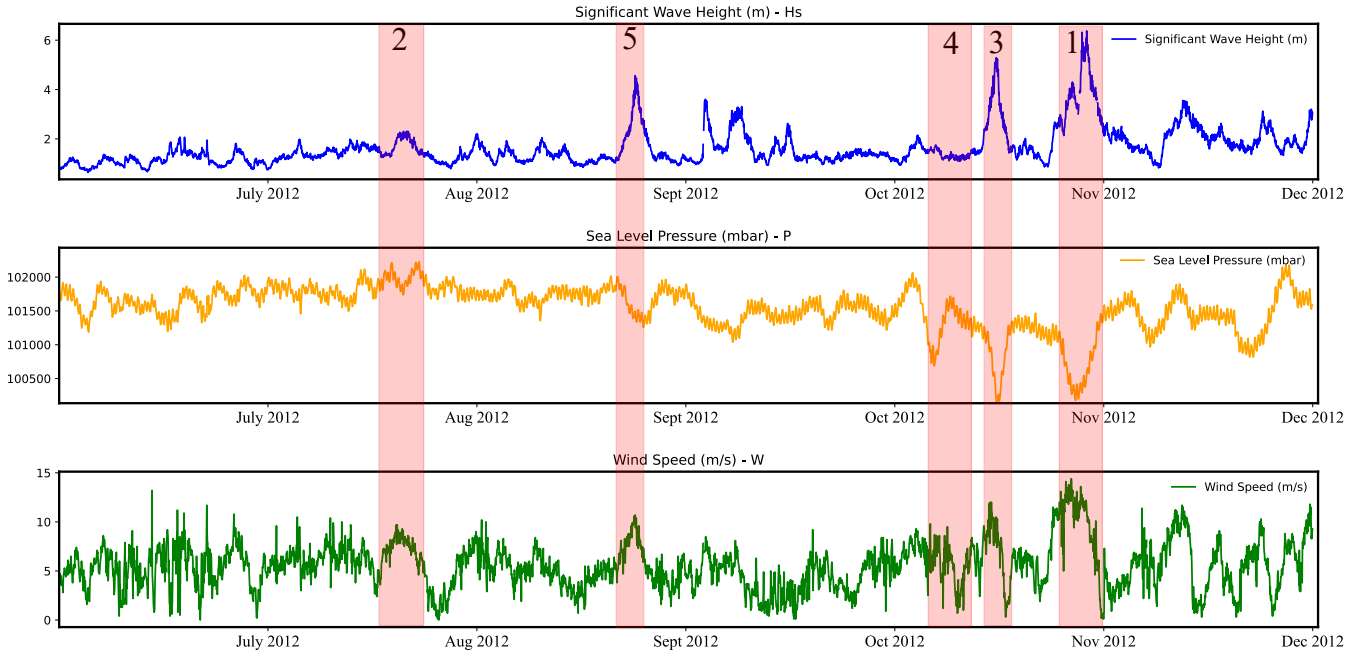

Figure 3. Results of weather dataset. The red area represents the detected intervals. The numbers illustrate the order of the intervals based on their score (1 has the highest intensity)

including a group of points lies on the circle of radius K-distance. Then the reachability density RD is computed, which defines a maximum of K-distance of X_j and the distance between X_i and X_j. Using RD, we compute the local reachability density LRD. This indicates how far a point is from the nearest cluster. Low values of LRD mean that the nearest cluster is far from the point under consideration. After calculating the LRD of each point, we compare it with the average LRD of its K neighbors. LOF is the ratio between the average LRD of A's K neighbors and the LRD of A itself. Thus, if the point is not an outlier, the ratio of the average LRD of the neighbors is approximately equal to the LRD of the point. This means that the density of the point and its neighbors are approximately equal. And if the LRD of a point is smaller than the average LRD of its neighbors, the point is indicated as an outlier [4].

3. EXPERIMENTAL EVALUATION

In this section, we evaluate the KL algorithm explained earlier using real time series data. Due to the lack of ground-truth data, quantitative evaluation using performance metrics is not possible. Therefore, we attempt to match the most evaluated intervals with known anomalies in weather events that occurred near the location where the data were recorded.

3.1. Weather Dataset

To evaluate the methodology, we conduct experiments with real-world multivariate time series data recorded from ocean observing buoys provided by the National Data Buoy Center (NDBC). The dataset covers six months of hourly data, beginning in June 2012 and ending in November of the same year. This period corresponds to the Atlantic hurricane season, which was particularly active this year with 19 tropical cyclones. 10 of them became hurricanes (winds over 64 km/h) [1]. This information can help to interpret and evaluate the results by matching the extracted interval with the already known time window of hurricanes since the dataset does not provide ground-truth data. Due to the limited time, we skip the matching.

The variables provided are measurements of significant wave height Hs, wind speed W, and sea level pressure P. The data were collected at a site near the Bahamas in the Atlantic Ocean [1] (23.866° N, 68.481° W).

As with any machine learning method, the quality of the result depends heavily on the data fed into it. Therefore, the data should be preprocessed to get it into a clean and complete form ready for the following learning task. The provided data contains a small percentage of missing values ($< 0.5\,\%$); therefore, we exclude the rows from the data where the missing values are present. Due to the massive differences between the values of the variables, the data are normalized by subtracting the mean and dividing by the standard deviation. The normalization is performed for each variable separately. The experiments are limited by the size of the intervals to be searched between one and a maximum of six days.

1. https://www.ndbc.noaa.gov/station$_page.php?station$ =41046

3.2. Results

Figure 3 illustrates the results of applying the KL algorithm based on generated intervals containing pointwise outliers. The figure shows three time series curves representing the features of the data set. The red color highlights the area where the outlier intervals are detected. The numbers at the top indicate the order of the intervals based on their intensity. Number 1 has the highest intensity. The results show that the algorithm is able to detect numerous outlier intervals, most of which could be identified by eye. Others, like interval number 4 are hard to identify directly. However, we could not confirm the results due to the lack of labeled data.

4. CONCLUSION

This work is carried out as a PhD. application assignment with a duration of 10 days. The task is to identify and sort possible outlier intervals from a multivariate time series weather dataset. We use a K-nearest neighbour based method to search for anomalous data points in the data. We then create intervals between the sampled points using the sliding window technique. An approximation to Kullback-Leibler divergence (KL) is used to quantify the degree of divergence between two Gaussian distributions (the generated interval distribution and the distribution of the remaining data). At the end, a list of the top intervals is generated to rank the anomalies based on their scores. The results show that the algorithm is able to detect numerous outlier intervals. However, due to the lack of labeled data, we could not numerically evaluate the performance of the algorithm.

References

[1] Rodner, Erik, et al. "Maximally divergent intervals for anomaly detection." arXiv preprint arXiv:1610.06761 (2016).

[2] Chandola, Varun, Arindam Banerjee, and Vipin Kumar. "Anomaly detection: A survey." ACM computing surveys (CSUR) 41.3 (2009): 1-58.

[3] Hershey, John R. and Peder A. Olsen. "Approximating the Kullback Leibler Divergence Between Gaussian Mixture Models." 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP: IV-317-IV-320.

[4] Breunig, Markus M., et al. "LOF: identifying density-based local outliers." Proceedings of the 2000 ACM SIGMOD international conference on Management of data.