UNSUPERVISED ANOMALY DETECTION

Suliman Alnutefy and Ali Alsuwayh

School of Technology and Innovation, Marymount University –
Ballston Center ,United States.

ABSTRACT

This research focuses on Unsupervised Anomaly Detection using the "ambient_temperature_system_failure.csv" dataset from Numenta Anomaly Benchmark (NAB). The dataset contains time-series temperature readings from an industrial machine's sensor. The aim is to detect anomalies indicating system failures or aberrant behavior without labeled data. Various algorithms, such as K-means, Gaussian/Elliptic Envelopes, Markov Chain, Isolation Forest, One-Class SVM, and RNNs, are applied to analyze the temperature data. These algorithms are chosen for their ability to identify significant deviations in unlabeled datasets. The study explores how these techniques enhance anomaly understanding in time series data, relevant in manufacturing, healthcare, and finance. This research's novelty lies in employing unsupervised learning techniques on a real-world dataset and understanding their adaptability in anomaly detection. The results are expected to contribute valuable insights to the field, showcasing the practicality and effectiveness of these algorithms across various scenarios.

KEYWORDS

Unsupervised Anomaly Detection, Time Series Data, Numenta Anomaly Benchmark, Industrial Machine Sensor Data, Algorithm Analysis, Machine Learning

1. INTRODUCTION

Anomaly detection in time-series data is a critical task in many industries, with applications ranging from predictive maintenance in manufacturing to fraud detection in finance. The unique nature of time-series data, where information is collected over a period, presents specific challenges and opportunities for anomaly detection. Traditional data analysis methods might fail to capture the dynamic characteristics of time-series data, making specialized approaches necessary. This study focuses on the ambient_temperature_system_failure.csv dataset, a time-series dataset from the Numenta Anomaly Benchmark (NAB), to explore the effectiveness of various unsupervised learning algorithms in detecting anomalies in industrial machine temperature data. The rationale behind this focus is to understand how deviations in temperature readings, which might indicate potential failures or malfunctions, can be identified effectively using advanced machine learning techniques.

Unsupervised learning, a branch of machine learning where algorithms learn from data without being explicitly programmed, is particularly well-suited for anomaly detection in unlabeled datasets. In this study, several unsupervised algorithms, including Cluster-based anomaly detection (K-means), Gaussian/Elliptic Enveloppe, Markov Chain, Isolation Forest, One-Class SVM, and Recurrent Neural Networks (RNNs), are applied to the NAB dataset. These algorithms are selected for their ability to identify outliers or unusual patterns in data without the need for

David C. Wyld et al. (Eds): IOTBS, NLTM, AIMLA, DBDM -2024
pp. 145-154, 2024. CS & IT - CSCP 2024 DOI: 10.5121/csit.2024.140210

predefined labels. The choice of algorithms reflects a comprehensive approach, encompassing both traditional statistical methods and modern deep learning techniques. This variety allows for a thorough examination of each algorithm's strengths and limitations in the context of anomaly detection in time-series data.

The practical implications of this research are vast. In industrial settings, accurately detecting anomalies in machine sensor data can prevent costly downtimes and catastrophic failures. In the financial sector, anomaly detection can signal fraudulent activities or market irregularities. Beyond these direct applications, the study contributes to the broader field of machine learning by providing insights into the adaptability and efficiency of various unsupervised learning algorithms in different contexts. The research also aims to bridge the gap between theoretical algorithmic understanding and practical application, demonstrating how complex data sets like those in the NAB can be effectively analyzed using a range of machine learning techniques. This exploration is expected to pave the way for more sophisticated and accurate anomaly detection systems, applicable across various domains where time-series data plays a crucial role.

2. RELATED WORK

The field of anomaly detection has witnessed significant advancements, particularly in the realm of electronic health records (EHR). Philipp Röchner and Franz Rothlauf's study is a notable contribution [1]. This research delves into the application of unsupervised learning techniques for identifying implausible records in cancer registries, an area fraught with challenges due to the sensitive and complex nature of medical data. The study's unique approach to handling EHR data underscores the evolving landscape of anomaly detection, where the focus is shifting towards more specialized domains like healthcare, demanding tailored algorithmic strategies.

Goldstein's study provides a comprehensive overview of the current state and future directions of unsupervised anomaly detection [2]. This work is crucial in understanding the broad spectrum of applications and methodologies within this field. Goldstein’s review encapsulates various studies and findings, offering a bird's-eye view of the advancements and challenges in unsupervised anomaly detection. It sheds light on the diverse contexts in which these methods are applied, from industrial systems to complex network data, and discusses the evolving nature of algorithms in tackling real-world anomaly detection problems.

The research by Belay et al. marks a significant stride in the domain of IoT and time series data [3]. This study critically examines the intersection of IoT technology and unsupervised anomaly detection, highlighting the challenges and potential solutions in handling the vast, complex data generated by IoT devices. Their work emphasizes the importance of multivariate analysis in time series data, a crucial aspect for many modern applications, ranging from environmental monitoring to smart city management. The paper serves as an essential reference for understanding the intricacies of IoT data analysis and the role of unsupervised learning in this rapidly expanding field.

Röchner and Rothlauf's research is pivotal in addressing the specific challenges posed by electronic health records (EHR) [1]. Their methodology not only focuses on the detection of anomalies but also underscores the importance of handling sensitive health data responsibly. This research is particularly relevant in the context of patient data privacy and the need for accurate detection mechanisms in healthcare. It highlights the growing need for specialized anomaly detection techniques that are adaptable to the nuances of different data types, especially in critical sectors like healthcare.

Goldstein's editorial contribution provides a panoramic view of the unsupervised anomaly detection landscape [2]. This work is instrumental in identifying the gaps and future research directions in the field. By collating various studies, Goldstein's review serves as a roadmap for researchers and practitioners, indicating the areas that require more in-depth exploration and where the potential for innovation is ripe. It also brings to light the interdisciplinary nature of anomaly detection, demonstrating its relevance across a spectrum of fields.

Belay et al.'s paper on IoT-based multivariate time series analysis is a forward-looking piece that paves the way for future research in this area [3]. Their analysis of existing solutions and performance in the context of IoT highlights the ever-increasing importance of this technology in daily life and the consequent need for robust anomaly detection mechanisms. This study serves as a guide for developing advanced algorithms capable of handling the complexity and scale of IoT-generated data, an area that is set to become even more critical in the near future.

3. DATASET DESCRIPTION

The project at hand utilizes the "ambient_temperature_system_failure.csv" dataset from the Numenta Anomaly Benchmark (NAB) collection, specifically from the realKnownCause folder. This dataset is instrumental for exploring unsupervised anomaly detection methods in a real-world context. The data, sourced from Kaggle, presents a compelling case study for the application of anomaly detection algorithms to time-series data from temperature sensors.

3.1. Dataset Characteristics

The dataset comprises 7267 entries, each with two primary columns: timestamp and value. The timestamp column records the time at which each temperature reading was taken, and the value column provides the corresponding temperature measurement. Initially, the temperature values are recorded in Fahrenheit, which are later converted into Celsius for standardization and ease of analysis. The data range covers a comprehensive timeline, providing a robust basis for detecting anomalies in temperature readings.

3.2. Preliminary Data Analysis

An initial examination of the dataset includes standard data exploration techniques such as checking the data types and the range of values. The analysis begins with converting the timestamp data from an object type to a datetime format, enabling more effective time-series analysis. The average temperature is calculated to gain a basic understanding of the dataset's central tendency. This preliminary analysis forms the foundation for more advanced feature engineering and anomaly detection techniques.

4. METHODOLOGY

The methodology employed in this project involves a series of data preprocessing steps, feature engineering, and the application of various unsupervised machine learning models for anomaly detection.

Feature engineering plays a critical role in enhancing the effectiveness of anomaly detection algorithms. In this project, several new features are derived from the original dataset, including the hour of the day (hours), a binary indicator of daylight (daylight), the day of the week (DayOfTheWeek), and a binary indicator of weekdays (WeekDay). These engineered features

provide additional context to the model, allowing it to discern patterns based on time of day, day of the week, and light conditions.

The core of the project revolves around the application of various unsupervised anomaly detection techniques:

Clustering (K-Means): The dataset is standardized and reduced to its principal components for efficient clustering. K-Means clustering is applied to group data points, and the distances from these clusters are used to identify anomalies.

Markov Chain Models: These models are utilized to analyze the transitions between different states in the data, providing insights into potential anomalies based on these transitions.

Distance and Threshold-Based Anomaly Detection: After clustering, the distance of each point from its nearest cluster centroid is calculated. Points that lie beyond a certain threshold distance are flagged as anomalies.

These methodologies are complemented by various visualization techniques, including plotting temperature over time and histogram representations of data under different conditions, to provide a comprehensive understanding of the data and the anomalies detected.

The combination of these sophisticated techniques in Python, leveraging libraries such as Pandas, NumPy, and Matplotlib, illustrates the power of modern data science tools in extracting meaningful insights from complex datasets.

5. ANALYSIS

The analysis of the dataset involves three key sections of Python code. Each section contributes to understanding the dataset's anomalies through different approaches and techniques.

5.1. Temperature Data Clustering and Anomaly Detection

In the first part of the analysis, the focus is on clustering the standardized temperature data and detecting anomalies based on their distance from cluster centroids.

```
# Standardizing and clustering temperature data

temperature_data = df[['temperature', 'hours', 'light', 'WeekdayIndicator', 'WeekendIndicator']]

scaler = preprocessing.StandardScaler()

scaled_data = scaler.fit_transform(temperature_data)

temperature_data = pd.DataFrame(scaled_data)

# Applying PCA for dimensionality reduction

pca_reducer = PCA(n_components=2)

reduced_data = pca_reducer.fit_transform(temperature_data)
```

```
# Clustering with K-Means and determining the optimal number of clusters
num_clusters = range(1, 20)
kmeans_models = [KMeans(n_clusters=i).fit(reduced_data) for i in num_clusters]
loss_scores = [kmeans_models[i].score(reduced_data) for i in range(len(kmeans_models))]
plt.figure()
plt.plot(num_clusters, loss_scores)
plt.show()

# Assigning clusters to the data and calculating distance to nearest centroids
df['cluster_group'] = kmeans_models[14].predict(reduced_data)
 df['feature1'] = reduced_data[0]
df['feature2'] = reduced_data[1]

# Function to calculate the distance of each point to its nearest centroid
def calculateCentroidDistance(data, kmeans_model):
centroid_distance = pd.Series()
for i in range(len(data)):
point = np.array(data.iloc[i])
centroid = kmeans_model.cluster_centers_[kmeans_model.labels_[i]-1]
centroid_distance.at[i] = np.linalg.norm(point-centroid)
return centroid_distance

distance = calculateCentroidDistance(reduced_data, kmeans_models[14])
```

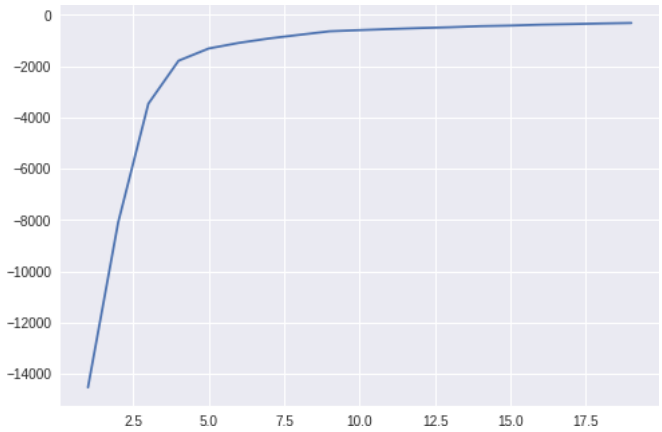

Figure 1. Outliers based on their distance from cluster centroids - elbow method

The clustering approach helps in grouping similar temperature readings, allowing us to identify outliers based on their distance from cluster centroids. The application of PCA assists in reducing the dimensions of the data, thereby simplifying the clustering process. By determining the optimal number of clusters using the elbow method, we can ensure that the data is not over or under-clustered, which is crucial for accurate anomaly detection.

5.2. Histogram Analysis of Temperature Data

The second part of the analysis involves creating histograms for different categories of data to visualize the distribution of temperature readings.

```
# Categorizing data for histogram analysis

category_values = df['WeekendIndicator']*2 + df['light']

categories = df['temperature'].groupby(category_values)


# Plotting histograms for each category

plt.figure()

for category, values in categories:

heights, bins = np.histogram(values)

plt.bar(bins[:-1], heights, width=(bins[1]-bins[0])/6, label=f'Category {category}')

plt.legend()

plt.show()
```

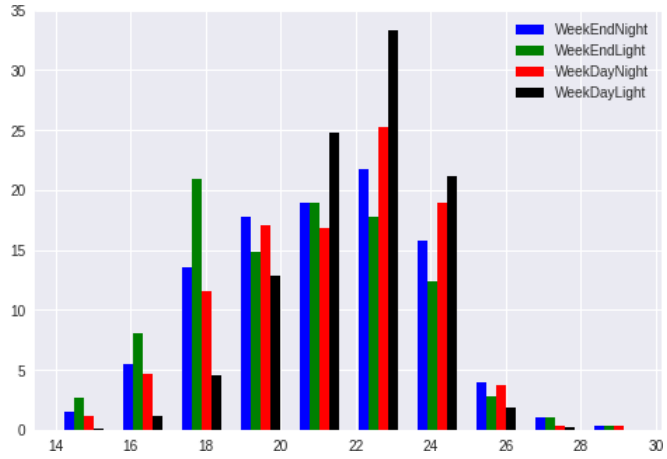

Figure 2. temperature data distribution across different times of the day and week

This section provides a visual representation of the temperature data distribution across different times of the day and week. The histograms for each category (like weekdays, weekends, day, and night) reveal patterns and anomalies in temperature readings, offering insights into potential irregularities or unusual temperature behaviors.

5.3. Anomaly Detection Over Time

In the third part, anomalies are plotted over time to visualize their occurrence in the dataset.# Detecting and plotting anomalies over time

```
outlier_fraction = 0.01

number_of_outliers = int(outlier_fraction * len(distance))

 threshold_value = distance.nlargest(number_of_outliers).min()


df['temp_anomaly'] = (distance >= threshold_value).astype(int)


plt.figure()

anomalies = df[df['temp_anomaly'] == 1]
```

```
plt.plot(df['time_epoch'], df['temperature'], color='blue')

 plt.scatter(anomalies['time_epoch'], anomalies['temperature'], color='red')

plt.show()
```

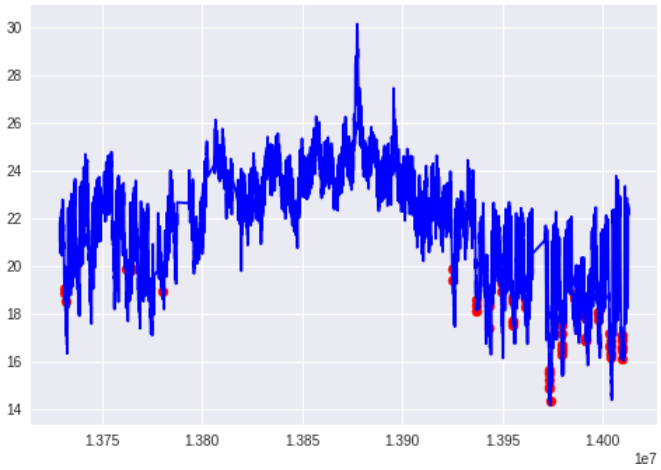

Figure 3. Anomaly Detection Over Time

This section effectively visualizes the detected anomalies over time, providing a clear picture of when these irregularities occur. By setting an appropriate threshold for distance, anomalies are identified based on their deviation from normal patterns. The temporal aspect of this analysis is crucial in understanding the context and potential causes of the anomalies, such as equipment malfunctions or environmental factors.

6. DISCUSSION – EXPLANATION

The analytical approach undertaken in this study has yielded a multifaceted understanding of the temperature dataset, with each visual output contributing unique insights into the system's behavior. The elbow graph serves as the preliminary step in our cluster analysis, suggesting an optimal cluster count. The stabilization of within-cluster variance beyond five clusters indicates a diminishing return on model complexity, suggesting a potential balance at this cluster number. However, a deeper dive into the graph reveals a secondary elbow at approximately 15 clusters. This could imply a more complex underlying structure within the data, possibly reflecting different operational modes of the monitored system, changes in environmental conditions, or the transition between different states of the system.

The histogram analysis further delineates the data into discrete temporal categories, revealing a pronounced variance in temperature readings between different times of the day and week. The WeekEndNight category, for instance, displays a wider distribution of temperatures, indicating greater variability or a potential calibration issue with sensors during these periods. Conversely, the WeekDayLight category exhibits a tighter distribution, implying a more stable and predictable operational environment. These patterns suggest that the system experiences different states of thermal behavior, which are influenced by both the natural diurnal/nocturnal cycle and human-induced operational patterns.

The time-series graph, marked with anomalies, shows sporadic occurrences of extreme temperature values that could be indicative of system malfunctions or external disturbances. A

cluster of anomalies within a short time frame, particularly noticeable in certain periods, raises questions about possible systemic issues or external events. For example, a sequence of anomalies coinciding with known maintenance schedules or external temperature spikes due to weather events could be inferred from the data.

In discussing the findings, it is crucial to note the limitations of the approach. The unsupervised nature of the analysis, while powerful in detecting patterns without prior labeling, does not provide insight into the causality behind the anomalies. The marked deviations could result from both benign and critical events; distinguishing between these requires a contextual understanding of the system's operational environment.

7. CONCLUSION

In conclusion, the application of unsupervised anomaly detection techniques to the dataset has successfully identified patterns and anomalies within the temperature readings. The insights derived from clustering, histogram analysis, and time-series visualization have potential implications for predictive maintenance and operational optimization in industrial settings. By identifying periods prone to anomalies, organizations can allocate resources more efficiently, perform targeted maintenance, and potentially prevent system failures.

However, the study is not without its limitations. The elbow method, while useful, provides a somewhat subjective means of determining the number of clusters. The true complexity of the data may require a more nuanced approach or the incorporation of domain knowledge to decide on the appropriate number of clusters. Furthermore, the histogram analysis, though illustrative, does not account for the multivariate nature of the data, which could conceal complex interactions between different variables influencing temperature readings.

Finally, while the marked anomalies provide a clear indication of deviations from the norm, the lack of labeled data means that the true nature of these anomalies cannot be ascertained without additional information. This underscores a broader limitation of unsupervised learning; while it can highlight potential areas of interest, it cannot always provide definitive conclusions without further investigation or corroborating data. Future work could involve integrating supervised learning techniques with labeled data to validate and refine the findings of this study, thereby enhancing the robustness and applicability of the anomaly detection models used.

ACKNOWLEDGEMENTS

I would like to express my profound gratitude to all those who have contributed to the success of this research on Unsupervised Anomaly Detection. Firstly, I extend my sincere thanks to my academic advisor and mentor, whose invaluable guidance, persistent support, and insightful critiques have been pivotal in shaping this research. Their expertise and encouragement were crucial in navigating the complex aspects of this study.

Special appreciation goes to the members of the research team who contributed their time, effort, and expertise. Their dedication and collaborative spirit were indispensable in conducting the research and analyzing the data using various unsupervised learning algorithms. I am grateful for their unwavering

Finally, I must express my deepest gratitude to my family and friends for their unwavering support, patience, and belief in my abilities. Their moral support and encouragement were my pillars of strength during challenging times.

This research is not just a product of my efforts but a testament to the collaborative and supportive spirit of all those mentioned above. Thank you.

REFERENCES

[1] Philipp Röchner and Franz Rothlauf, "Unsupervised anomaly detection of implausible electronic health records: a real-world evaluation in cancer registries," BMC Medical Research Methodology, vol. 23,no. 1, May 2023, doi: https://doi.org/10.1186/s12874-023-01946-0.

[2] M. Goldstein, "Special Issue on Unsupervised Anomaly Detection," Applied sciences, vol. 13, no.10, pp. 5916–5916, May 2023, doi: https://doi.org/10.3390/app13105916.

[3] M. A. Belay, S. S. Blakseth, A. Rasheed, and P. Salvo Rossi, "Unsupervised Anomaly Detection for IoT-Based Multivariate Time Series: Existing Solutions, Performance Analysis and Future Directions,"Sensors, vol. 23, no. 5, p. 2844, Mar. 2023, doi: https://doi.org/10.3390/s23052844.

AUTHORS

Suliman Alnutefy, Candidate in doctoral of science in cybersecurity, School of Technology andInnovation, Marymount University – Ballston Center ,United States.

Project manager in King Abdullah Scholarships program, Since 2010. Advisor at Accreditation & Authentication Department, Saudi Arabia Cultural Missio – Fairfax, Virginia, USA Since 2018.

Ali Alsuwayh, Candidate in doctoral of science in cybersecurity, School of Technology and Innovation,Marymount University – Ballston Center ,United States.

Head of Ticket Department , Ministry of higher Education , Riyadh KSA, Since 2006. Director of theTuition Fees Department, Student Allowances Department, and Director of the Medical Insurance Department Since 2019. Academic Advisor Since 2022