



K-MEANS-BASED ANOMALY DETECTION FOR GAS TURBINE ENGINE PREDICTIVE MAINTENANCE

Uchenna A. Iyke-Anakor

Department of Computer Science
Nnamdi Azikiwe University, Awka

ABSTRACT

Gas turbine engines are important assets in power generation and industrial operations, where unexpected faults can result in equipment damage, unplanned downtime, production losses, and increased maintenance costs. Predictive maintenance provides an opportunity to identify abnormal operating behaviour before it develops into a more serious failure. However, the availability of accurately labelled fault data remains a challenge in many industrial applications. This study presents the application of K-means clustering as an unsupervised machine-learning approach for identifying normal and anomalous operating conditions in gas turbine engine operational data. The study forms part of a broader predictive fault maintenance model developed for a gas turbine engine. Gas turbine operational data were obtained from a publicly available dataset and subjected to preprocessing and normalization before clustering. The operational variables considered include ambient temperature, ambient pressure, ambient humidity, air filter differential pressure, gas turbine exhaust pressure, turbine inlet temperature, turbine after temperature, turbine energy yield, and compressor discharge pressure. Each observation was represented as a sensor measurement vector, and K-means was applied to group observations with similar operating characteristics. Euclidean distance was used to determine the proximity of each observation to the cluster centroids, while the resulting cluster characteristics were used to distinguish normal from anomalous operational patterns. Unlike supervised classification, the approach does not require a pre-existing fault-class label for every observation. The resulting cluster labels provide an operational-state representation that can subsequently support temporal fault prediction in the wider predictive maintenance model. The study demonstrates how unsupervised clustering can provide an effective preliminary anomaly-detection stage for data-driven gas turbine predictive maintenance.

Keywords: Gas Turbine, Predictive Maintenance, Anomaly Detection, K-Means Clustering, Unsupervised Learning, Sensor Data, Operational Condition, Machine Learning.

Introduction

Gas turbine engines are widely used in power-generation and industrial environments because of their ability to produce substantial power continuously. Their performance depends on the interaction of several components and operating parameters, including the compressor, combustion system, turbine, and associated monitoring sensors. Changes in these parameters can indicate variations in the operating condition of the engine and may provide early evidence of abnormal behaviour.



The reliability of a gas turbine is particularly important because an unexpected failure can lead to equipment damage, production interruption, increased maintenance expenditure, and loss of operational availability. Conventional corrective maintenance waits until a fault has occurred, while preventive maintenance commonly schedules maintenance according to predetermined intervals. These approaches may result either in late intervention or unnecessary maintenance. Predictive maintenance attempts to overcome these limitations by using operational data to identify developing abnormal conditions and support maintenance decisions before severe failure occurs.

The increasing availability of sensor data has created opportunities for machine-learning techniques to be applied to predictive maintenance. Gas turbine systems continuously generate measurements relating to environmental conditions, pressure, temperature, compressor operation, turbine operation, and energy production. When these measurements are analyzed collectively, they can reveal patterns associated with different operating conditions.

Anomaly detection is an important component of predictive maintenance because abnormal operating behaviour can provide an early indication that an engine is not functioning within its expected operational range. Narayanan and Bobba (2018), for example, investigated learning-based anomaly detection for industrial robotic-arm applications and demonstrated the use of learned operational behaviour to distinguish anomalous from benign activity. Their work illustrates the broader potential of machine learning for detecting abnormal industrial behaviour.

However, a major challenge in industrial predictive maintenance is the lack of sufficient labelled fault data. In many real-world systems, operational measurements are available in large quantities, while verified records indicating exactly when and why a particular fault occurred may be limited. Consequently, an unsupervised learning method can be useful because it can discover patterns in the available data without requiring every observation to have a predefined fault label.

K-means clustering is one such unsupervised machine-learning technique. It partitions observations into groups according to their similarity and represents each group using a centroid. In the context of gas turbine monitoring, observations with similar sensor characteristics can therefore be grouped together to provide an operational representation of the engine.

This study focuses specifically on the K-means clustering component of a broader predictive fault maintenance model developed for a gas turbine engine. The purpose is to examine how K-means can be used to group gas turbine operating conditions and provide a basis for identifying normal and anomalous observations.

Problem Statement

Gas turbine engines operate under continuously changing environmental and mechanical conditions. Parameters such as ambient temperature, ambient pressure, humidity, compressor discharge pressure,



turbine inlet temperature, and exhaust pressure may vary during operation. Consequently, identifying an abnormal condition based on a single sensor measurement or fixed threshold may not adequately represent the actual operational state of the engine.

A further challenge is the absence of a direct ground-truth fault-class column in the operational dataset used in this study. The available measurements describe the operating characteristics of the gas turbine, but the dataset does not provide a verified fault category for every individual observation.

This creates a limitation for conventional supervised classification because a classifier normally requires labelled examples of the conditions it is expected to recognize. Using an unsupervised approach such as K-means provides an alternative by allowing the operational data to be grouped according to similarity before assigning operational interpretations to the resulting clusters. Therefore, there is a need for an approach that can examine multivariate gas turbine measurements, identify groups of similar operating conditions, and distinguish observations that deviate from the dominant or expected operating patterns. This study addresses this problem through K-means clustering as a preliminary anomaly-detection mechanism.

Aim and Objectives

Aim

The aim of this study is to apply K-means clustering for identifying normal and anomalous operating conditions in gas turbine engine operational data as a component of a predictive fault maintenance model.

Objectives

The objectives are to: obtain gas turbine operational data from a publicly available dataset; preprocess the gas turbine operational measurements; normalize the selected operational variables; represent individual observations as sensor measurement vectors; apply K-means clustering to group similar gas turbine operating conditions; determine the relationship between observations and their corresponding cluster centroids using distance measurements; interpret the resulting clusters in terms of normal and anomalous operating conditions; and provide cluster labels that can serve as an operational-state representation for the broader predictive fault maintenance model.

Related Literature

Predictive Maintenance

Predictive maintenance is a maintenance strategy that uses information about the current or historical condition of equipment to determine whether maintenance intervention may be required. Its purpose is to move maintenance decisions from purely time-based or reactive approaches toward condition-based decisions. For gas turbine engines, predictive maintenance can use sensor measurements to monitor changes in operating behaviour. The ability to identify abnormal patterns before an actual failure occurs can help maintenance personnel investigate the equipment and take appropriate action. The effectiveness



of predictive maintenance therefore depends not only on the ability to predict faults but also on the ability to recognize meaningful changes in equipment operating conditions.

Anomaly Detection in Industrial Systems

Anomaly detection involves identifying observations or patterns that differ from expected behaviour. In industrial systems, an anomaly may indicate unusual operating conditions, sensor abnormalities, component degradation, or a developing fault. Narayanan and Bobba (2018) developed a learning-based anomaly detection framework for industrial robotic arms. Their work classified robot actions as anomalous or benign based on learned operational behaviour and demonstrated the use of anomaly detection in an industrial environment. Although their application was directed toward industrial robotic arms rather than gas turbines, the study supports the general principle that operational behaviour can be learned from machine data and used to detect deviations. The present study applies this general principle to gas turbine operational measurements.

K-means Clustering

K-means is an unsupervised learning algorithm that divides a dataset into a predetermined number of clusters. The objective is to place observations with similar characteristics in the same cluster while separating observations with different characteristics into different clusters. For a dataset containing n observations and k clusters, K-means attempts to minimize the within-cluster sum of squared distances:

$$[J = \sum_{j=1}^k \sum_{x_i \in C_j} |x_i - \mu_j|^2]$$

where:

C_j represents cluster j ;

x_i represents an observation assigned to cluster j ; and

μ_j represents the centroid of cluster j .

The centroid is the mean position of all observations assigned to that cluster.

The distance between an observation and a centroid can be calculated using Euclidean distance:

$$[d(x_i, \mu_j) = \sqrt{\sum_{r=1}^m (x_{ir} - \mu_{jr})^2}]$$

where m represents the number of variables. K-means is appropriate for the present study because it can identify patterns in multivariate operational data without requiring a fault label for every observation.

Methodology

Research Design

This study adopts an unsupervised machine-learning methodology for the detection of anomalous operating conditions in gas turbine engine operational data. The methodology focuses on K-means clustering because the dataset used in the study does not contain a verified fault-class label for each observation. The method therefore first learns groups of similar operating conditions from the available multivariate measurements and subsequently uses the characteristics of these groups to identify operational deviations.



The methodology consists of five principal stages: data acquisition, data preprocessing and normalization, formation of sensor measurement vectors, K-means clustering, and cluster-based anomaly identification. The output of the clustering stage is an operational-state label for each observation. In the broader predictive fault maintenance model, these labels can subsequently be arranged in temporal order and used to support the Simple Recurrent Neural Network (RNN) stage. The present article is limited to the K-means-based anomaly-detection stage.

Dataset Acquisition

The dataset used in this study is the Gas Turbine CO and NOx Emission dataset obtained from Kaggle. The dataset is a copy of the Gas Turbine CO and NOx Emission Data Set originally published through the UCI Machine Learning Repository. The data consist of gas turbine operational measurements collected from a power-generation environment.

Nine operational variables are selected for the clustering analysis: Ambient Temperature (AT), Ambient Pressure (AP), Ambient Humidity (AH), Air Filter Differential Pressure (AFDP), Gas Turbine Exhaust Pressure (GTEP), Turbine Inlet Temperature (TIT), Turbine After Temperature (TAT), Turbine Energy Yield (TEY), and Compressor Discharge Pressure (CDP).

Data Preprocessing and Normalization

The selected operational measurements are first prepared for analysis and then normalized before clustering. Normalization is necessary because the variables have different measurement units and numerical ranges. Since K-means determines cluster membership using distance calculations, allowing variables with larger numerical magnitudes to dominate the distance calculation could distort the resulting clusters.

The normalized data are therefore used as the input to the K-means clustering stage. No fault class is assigned to the observations before clustering. This preserves the unsupervised nature of the method and allows the algorithm to discover groups based on the similarities among the operational measurements.

Formation of Sensor Measurement Vectors

Each observation is represented as a multivariate sensor measurement vector containing the nine selected operational variables:

$$[X_t = [AT_t, AP_t, AH_t, AFDP_t, GTEP_t, TIT_t, TAT_t, TEY_t, CDP_t]^T]$$

where t represents the observation period. The vector provides a combined representation of the gas turbine operating condition at that observation. The use of a multivariate vector is important because an abnormal operating condition may not necessarily be indicated by one variable in isolation. Instead, it may appear as an unusual combination of temperature, pressure, humidity, compressor, turbine, and energy-related measurements. K-means therefore considers the complete measurement vector when determining similarity between observations.



K-means Clustering

K-means clustering is applied to the normalized sensor measurement vectors to group observations with similar operational characteristics. The algorithm requires the number of clusters, K , to be specified. Initial centroids are established, after which the distance between each observation and each centroid is calculated. Euclidean distance is used to measure the similarity between an observation and a centroid:

$$[d(x, \mu) = \sqrt{\sum_{i=1}^m (x_i - \mu_i)^2}]$$

where x_i represents the value of the i^{th} feature of an observation, μ_i represents the corresponding centroid value, and m is the number of selected features.

Each observation is assigned to the centroid with the smallest Euclidean distance. After assignment, the centroid of each cluster is recalculated using the mean of all observations assigned to that cluster:

$$[\mu_k = \frac{1}{|C_k|} \sum_{x_i \in C_k} x_i]$$

where C_k represents cluster k , and $|C_k|$ represents the number of observations assigned to that cluster.

The assignment and centroid-update processes are repeated until the clustering converges or the specified stopping condition is reached. The final centroids represent the learned operational profiles of the identified clusters.

Cluster Distance and Anomaly Identification

After convergence, the distance between each observation and the centroid of its assigned cluster is calculated. This distance provides an indication of how closely the observation corresponds to the representative operational profile of its cluster.

For the anomaly-detection stage, the distance values can be standardized using a z-score:

$$[z = \frac{d_k - \mu}{\sigma}]$$

where d_k is the distance between an observation and its assigned centroid, μ is the mean of the relevant distance values, and σ is their standard deviation.

Observations showing substantial deviation from the established operational pattern can consequently be identified for anomaly analysis. The threshold used for identifying an anomaly must be based on the actual criterion adopted in the study and should not be presented as a verified fault threshold unless it has been validated against actual fault records.

Interpretation of Cluster Conditions

K-means generates clusters according to similarity; it does not inherently assign the meanings "normal" or "fault" to the clusters. Therefore, after clustering, the characteristics of the resulting clusters are examined in relation to the expected operating behaviour of the gas turbine. A cluster representing the dominant or expected operating pattern may be interpreted as a normal operating condition, while a cluster showing substantial deviation from the expected operational pattern may be interpreted as anomalous. The resulting labels are therefore operational-state labels rather than verified mechanical-fault labels.



This distinction is necessary because the dataset used in this study does not contain a ground-truth column identifying the actual fault condition of each observation. Consequently, the K-means stage is regarded as an unsupervised anomaly-detection mechanism rather than a supervised fault-classification model.

Methodology Workflow

The complete methodology can be summarized as follows:

Gas Turbine Dataset → Data Preprocessing → Normalization → Sensor Measurement Vector Formation → K-means Clustering → Centroid Calculation → Euclidean Distance Calculation → Cluster Assignment → Cluster Interpretation → Normal/Anomalous Operational Condition

The resulting cluster labels provide an operational representation of the gas turbine condition.

Discussion

The K-means results demonstrate the usefulness of unsupervised clustering for analyzing gas turbine operational data in situations where explicit fault labels are unavailable. One of the major strengths of the approach is that it considers several operational variables simultaneously. Gas turbine operating conditions are inherently multivariate. A change in turbine behaviour may involve relationships between temperature, pressure, compressor performance, and other parameters rather than an isolated change in a single measurement. By representing each observation as a sensor measurement vector, K-means considers these variables collectively.

The use of normalization before clustering is also important. Because K-means is distance-based, differences in measurement scales can influence the resulting clusters. Normalization helps ensure that the clustering process reflects the combined contribution of the selected operational parameters. Another important aspect of the approach is the use of cluster centroids as representative operational profiles. Instead of treating every observation independently, the clustering process establishes groups that can be interpreted as different operating states. The approach is also appropriate for the data characteristics of the study because the dataset does not provide a direct verified fault-class label for each observation. The UCI documentation identifies clustering as one of the possible tasks for the dataset. The K-means component therefore addresses an important stage of the broader predictive maintenance problem: identifying the operational context in which a measurement occurs.

Contribution to the Predictive Fault Maintenance Model

Although this article focuses exclusively on K-means, the clustering stage was developed as an important component of the user's broader predictive fault maintenance model for a gas turbine engine. The contribution of K-means to the complete model is the generation of operational-state information before temporal prediction. Consequently, K-means is not being presented as the entire predictive fault maintenance system. It is the unsupervised anomaly-detection and operational-state identification component of the larger model.



Knowledge Gap Addressed

Existing research has established the usefulness of machine learning for anomaly detection and predictive maintenance in industrial environments. Narayanan and Bobba (2018), for example, investigated learning-based anomaly detection for industrial robotic-arm applications. Their work demonstrates the potential of learning operational behaviour to identify anomalies, but the application is not specifically concerned with gas turbine operational-condition clustering. Recent gas turbine research has also increasingly explored machine-learning approaches for monitoring, anomaly detection, and predictive maintenance. For example, recent work has investigated intelligent gas-turbine monitoring using approaches including ANFIS and LSTM for real-time anomaly detection and predictive maintenance.

However, there remains a need for relatively straightforward unsupervised approaches that can first organize multivariate gas turbine operational data into meaningful operating-condition groups, particularly when reliable fault labels are unavailable. The present study addresses this gap by applying K-means clustering to gas turbine operational measurements and interpreting the resulting clusters as normal and anomalous operational patterns. The key contribution is therefore the use of multivariate operational-condition grouping as a preliminary anomaly-detection mechanism for gas turbine predictive maintenance.

Advantages of the Proposed K-means Approach

The K-means component provides several advantages.

1 No Requirement for Pre-labelled Fault Classes

Because K-means is unsupervised, the initial clustering process does not require every observation to be assigned a predefined fault class.

2 Multivariate Analysis

The approach considers multiple operational parameters simultaneously rather than relying on a single sensor measurement.

3 Operational-State Representation

The resulting cluster labels provide a compact representation of the operating state of each observation.

4 Identification of Deviating Patterns

Clusters that differ from the dominant operating pattern can provide an indication of potentially anomalous behaviour.

5 Support for Predictive Maintenance

The cluster information can serve as an additional source of operational information for subsequent predictive modelling.

Limitations

The K-means approach also has limitations that should be acknowledged.

First, K-means requires the number of clusters to be specified. The resulting grouping can therefore depend on the selected value of k .



Second, K-means is based on distance and therefore depends on appropriate preprocessing and scaling of the variables.

Third, the algorithm itself does not inherently distinguish between “normal” and “fault.” It only identifies groups of similar observations. The interpretation of a cluster as normal or anomalous must therefore be based on analysis of the cluster characteristics.

Fourth, because the dataset does not contain verified fault labels for every observation, an anomalous cluster should not automatically be interpreted as a confirmed mechanical failure. It represents an operational deviation that requires further examination.

Finally, validation using real-world gas turbine fault records would provide stronger evidence regarding the ability of the clusters to correspond to actual equipment faults.

Implications for Gas Turbine Maintenance

The application of K-means provides a practical foundation for condition monitoring. When new operational measurements become available, they can be normalized using the same preprocessing procedure and compared with the learned cluster centroids. The new measurement can then be assigned to the closest operational cluster. If the observation corresponds to the established normal operating profile, the system can continue monitoring the engine. If the observation corresponds to an anomalous operating profile, the system can generate an indication requiring further investigation. This provides a possible pathway toward real-time condition monitoring: new sensor measurement – Preprocessing – Normalization - Distance to learned centroid - Cluster assignment - Normal/anomalous indication - Maintenance investigation. Such a system can help maintenance personnel identify unusual operating behaviour earlier and potentially reduce dependence on purely reactive maintenance.

Conclusion

This study presented the application of K-means clustering for identifying normal and anomalous operating conditions in gas turbine engine operational data. The approach was developed as the clustering component of a broader predictive fault maintenance model. The study used gas turbine operational variables including ambient temperature, ambient pressure, ambient humidity, air filter differential pressure, gas turbine exhaust pressure, turbine inlet temperature, turbine after temperature, turbine energy yield, and compressor discharge pressure. The data were preprocessed and normalized before being represented as multivariate sensor measurement vectors. K-means clustering was then applied to group observations according to their operational similarity. Euclidean distance was used to determine the relationship between individual observations and cluster centroids, while centroid characteristics provided representative profiles for the identified operational groups.



The resulting clusters were interpreted in terms of normal and anomalous operating conditions. This approach is particularly relevant where explicit fault labels are unavailable because it allows operational patterns to be discovered directly from the sensor data. The main contribution of the study is therefore the development of a clustering-based operational-state identification approach that can provide an anomaly-detection layer for gas turbine predictive maintenance. Rather than claiming that K-means alone predicts mechanical failure, the study establishes K-means as a means of identifying deviations in operating behaviour and generating operational-state information.

Within the broader predictive fault maintenance model, these cluster labels provide useful preliminary information for subsequent temporal modelling using the Simple Recurrent Neural Network. Thus, the K-means stage establishes the operational context upon which subsequent fault prediction can be performed. Future research should validate the identified anomalous clusters against verified gas turbine fault records, investigate methods for selecting the optimum number of clusters, and examine the stability of the identified operational groups under different operating conditions and real-time sensor streams.

REFERENCES

- Kaya, H., Tüfekci, P., & Gürgen, F. S. (2019). Gas Turbine CO and NO_x Emission Data Set. UCI Machine Learning Repository. DOI: 10.24432/C5WC95.
- Narayanan, V., & Bobba, R. B. (2018). Learning based anomaly detection for industrial arm applications. In Proceedings of the 2018 Workshop on Cyber-Physical Systems Security and Privacy (pp. 13–23). Association for Computing Machinery. <https://doi.org/10.1145/3264888.3264894>.
- Gas Turbine CO and NO_x Emission Dataset. (2019). Kaggle. Dataset containing 36,733 hourly gas-turbine observations.