Symbolic Time Series Analysis for Predictive Maintenance

Omri Y Ram

Débora Corrêa & Rachel Cardell-Oliver
Computer Science and Software Engineering
The University of Western Australia

Charlie Musca & Kevin Winchester
RCT

Abstract

Decreasing costs and increasing availability of sensors and computing power allowed wide adoption of the Industrial Internet of Things enabled predictive maintenance. Challenges are still posed by data collected in an industrial setting which is often unlabeled. Despite the accrued attention, there are still not many works detailing unsupervised approaches to recognize degraded performance. This study seeks to explore the use of a D-Markov chain constructed on symbolized time series data to identify time series segments exhibiting degraded performance in industrial machinery. To date, the research has established that the methodology is capable of identifying time series segments corresponding to degraded performance on a benchmark dataset. Further assessment by domain experts is still necessary to verify that the method works on real world data.

1. Introduction

Decreasing costs and increasing availability of sensors, computing power and other Internet of Things (IoT) related technologies have enabled wide adoption of the Industrial Internet of Things (IIoT) (Coleman et al., 2017, pp.12-14). Prior to the advent of IIoT, three primary strategies existed for machine maintenance: reactive maintenance, planned maintenance, and proactive maintenance (Susto, Beghi, & De Luca, 2012). IIoT enabled predictive maintenance strategy (PdM) uses analytics and machine learning algorithms to diagnose problems, optimize resource allocation and improve maintenance scheduling (Coleman et al., 2017, pp.5-9). Concomitantly, PdM has garnered increased interest in the industrial sector.

RCT Technology delivers a range of machine solutions tailored for the mining and industrial sectors. They also provide regular maintenance services to their clients, who can benefit from advancements in predictive maintenance by optimizing resource allocation and maintenance scheduling. RCT's Guidance Control Unit (GCU) is a product offered to their clientele. When incorporated into loaders utilized at mine sites, the GCU facilitates navigation along predefined routes. The automation logs from the GCU document both the projected and actual articulation and speed of the vehicle. These logs, obtained during technician assessments, assist in diagnosing problems with the vehicle or the GCU. This suggests that the data captures the vehicle's operational state, implying potential for the development of a PdM framework.

A systematic literature review by Carvalho et al. (2019) outlines an increasing trend in publications over the period 2009 – 2018. Despite growing interest, challenges related to data quality remain a significant obstacle. Within the review, Carvalho outlines the most popular algorithms implemented within predictive maintenance as supervised, requiring labeled data. RCT data collected from the GCU does not contain labels indicating which segments of the time series correspond to degraded performance, thus unsupervised methods must be used to develop a model capable of identifying the segments within the time series data which indicate a problem with the machine. This research project proposes the use of a symbolic time series analysis approach to extract rules exhibited by the system's anticipated nominal behavior. Subsequent data can then be juxtaposed against these established rules to discern deviations which may coincide with degraded performance.

Symbolic time series analysis transforms time series data into a discrete representation to uncover the coarse-grained rules or structures inherent to the system. Symbolization is achieved by partitioning the possible values in the original segmented time series and assigning a label to each partition. The set of labels or symbols constitutes the alphabet (Daw, Finney & Tracy, 2003). Various methods exist for clustering time series segments, providing meaningful partitioning of the data (Liao, 2005). One such method is the Gaussian Mixture Model (GMM) which partitions data into a predetermined number of components (analogous to clusters). Each component represents a mixture of Gaussian distributions responsible for producing the observations assigned to it (Géron, 2022). The component labels associated with each segment can be used to construct a symbolic sequence. Information theoretic, markovian and other methods may then characterize the dynamics of the underlying system (Das et al., 1998).

This project employs the use of a D-Markov chain, a fixed-order fixed-state Markov chain (Ray, 2004). In a D^{th} order Markovian process, the probability distribution of subsequent states is entirely dependent on the previous D states. For example, in a first order process the next state depends on the current state and in a second order process the next state depends on the current state and the previous state. In this project's context, the 'alphabet' refers to the potential states that the system can exhibit, and the symbolic representation of the time series is the evolution of the states of the system over time. A D-Markov chain transition matrix can be constructed by calculating the observed probability of transitioning from any one state or sequence of states to the next (Rajagopalan et al., 2007). A D-Markov chain, constructed from a system with nominal behavior, can identify non-nominal state transitions as those with a 0% probability in a nominal system. In this way, a predictive maintenance framework can be employed for recognizing time series segments exhibiting poor performance.

The project's objective is to devise a predictive maintenance framework that assists RCT technicians in assessing machine performance. To this end, data from GCU logs are symbolized and a D-Markov chain is constructed from a system exhibiting optimal behavior and a system known to exhibit suboptimal behavior. In order to identify segments demonstrating degraded performance of the machine, the two resulting transition matrices were compared to identify transitions which have 0% probability of occurring in an optimal system. A time series segment in the RCT data, previously known to exhibit degraded performance, was identified using this methodology. Additionally, the approach has been validated and performance metrics derived using a labeled benchmark dataset.

2. Process

The procedure comprises three stages: the time series data undergoes preprocessing and segmentation, after which the segments are split into historical and operational sets. The historical set has been previously examined by a domain expert and is known to exhibit optimal behavior, while the unexamined operational set might contain instances of degraded performance. Both sets are partitioned to construct the symbolic time series representation. Last, a D-Markov chain is developed to identify non-nominal behavior in the operational set.

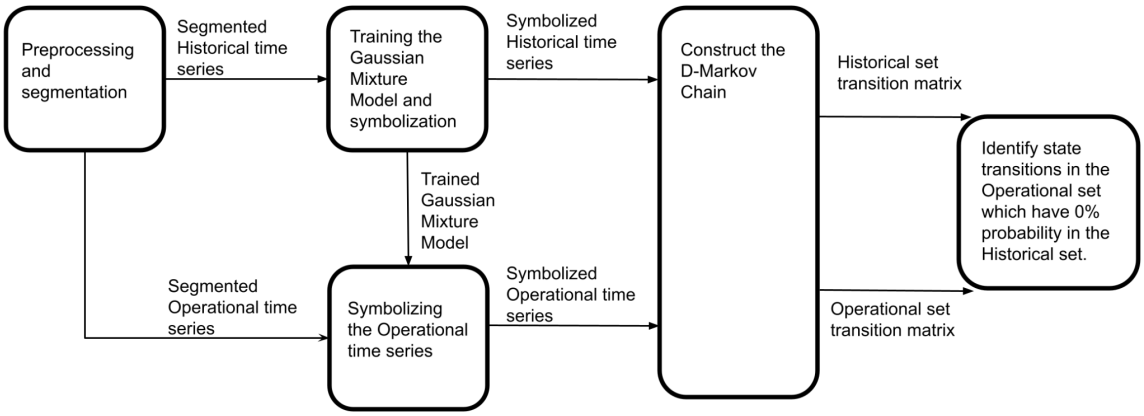

Figure 1 Diagram illustrating the process developed for the predictive maintenance framework.

2.1 Data Preparation

The data is normalized to have 0 mean and unit variance before being segmented into equally sized fragments. Some domain knowledge is necessary for determining the optimal size of the time series segments. Longer segments retain less information from the original time series, however segments which are too short contain more noise. Non-overlapping segments are used instead of a sliding window as it was demonstrated that time series subsequences form meaningless clusters (Keogh & Lin, 2005). Two distinct data representations can be fed into clustering algorithms: 1) preprocessed original data, and 2) feature vectors containing segment-specific statistics such as minimum, maximum, mean, and variance. The segments are then split into a historical set of data exhibiting optimal behavior and an operational set which may exhibit degraded performance.

2.2 Deriving a Symbolic Representation for Time Series Data

The GMM is first fit on the historical subset of the data. The fit of a model to a dataset can be assessed using the Bayesian Information Criterion (BIC). The BIC is beneficial because a better score indicates that the model's parameters explain the observed data well. Additionally, the BIC penalizes models with an excessive number of components. GMMs with 2 to 20 components were evaluated on both the original data and feature vectors according to their BIC scores. The selected model is used to symbolize the operational and historical sets by assigning the component label to each segment.

2.3 Symbolic Time Series Analysis

The historical set transition matrix is constructed by calculating the observed probability of transitioning from each state to the next in the symbolized time series. As the historical set contains data reflecting the system's optimal behavior, transitions with 0% probability of occurring can be used to identify segments which should not exist in an optimal system. The symbolized operational set is then used to construct a D-Markov chain for comparison with the historical set transition matrix. Any state transitions occurring in the operational set transition matrix with 0% probability in the historical set transition matrix will be classified as positive cases of degraded performance.

3. Results and Discussion

The methodology was verified on two datasets, the RCT data collected from the GCU automation logs and an additional open source anomaly benchmark dataset sourced online. The Skoltech Anomaly Benchmark (SKAB) dataset (Katser & Kozitsin, 2020) was curated to evaluate unsupervised anomaly detection models. It captures experiments on a water pump, including both perturbations and standard cycling conditions. This dataset contains measurements detailing the motor's state and the water it circulates. Experimenters manually annotated anomalous points within the time series, making it suitable for gauging the performance of the developed framework. As such, the framework's performance on the SKAB data can be assessed similarly to a supervised binary classifier. To assess the model's performance on the RCT data, a technician is consulted to examine the segments classified as positive.

3.1 SKAB dataset results

The SKAB dataset is segmented into 5-second intervals with 5 observations each. It was not possible to consult domain experts on which variables should serve as the best indicators of the water pump's performance. However, current measurements (amperage) to the motor serve as an indicator of the motor's workload and the resulting water's flow rate are expected to characterize the system well.

Markov Order	Second Order	First Order	Second Order
Representation	Feature vector	Preprocessed original data	Preprocessed original data
Accuracy	67.5%	65%	68%
Precision	90.9%	100%	58.75%
Recall	2.9%	1.4%	27.8%

Table 1 Performance of identifying anomalies in the SKAB dataset. The feature vector is derived by obtaining the minimum, maximum, mean and variance of each feature in the time series segment.

The first-order Markov chain created with feature vectors didn't have any state transitions with 0% probability. As a result, three models were evaluated for comparison. Given that a third of the segments contained anomalies, the accuracy—defined as the proportion of correctly classified segments—was only marginally better than a naive classifier with 65% accuracy. However, the precision—defined as the proportion of correct positive classifications—of all

three models exceeded that of a random classifier. No single data representation offered a clear advantage. Across all three models, a trade-off was observed between precision and recall, where recall is the proportion of positive cases correctly identified.

3.2 RCT Dataset Results

The RCT data is sampled at a high frequency (averaging 40 ms) and irregularly, is adjusted by downsampling - taking the median value every 250 ms. This process also aids in outlier removal. For segmentation, the data is divided into segments of 2.5 seconds, each containing 10 observations. Consultations with RCT technicians have established that the speed and articulation angle of the loader's pivot point are important for determining the performance state. As such, a first and second order markov chain were constructed.

Historical transition matrix

Current state \ Next state	1	2	3	4
1	60%	5.4%	2.3%	11%
2	12%	70%	12%	6.5%
3	41%	19%	29%	11%
4	25%	0.0%	17%	58%

Operational transition matrix

Current state \ Next state	1	2	3	4
1	64%	2.9%	22%	12%
2	3.3%	80%	9.1%	7%
3	39%	16%	30%	15%
4	23%	1.3%	12%	64%

Figure 2 First order markov chain transition matrices derived from the RCT data on the historical and operational time series. Each tuple exhibits the probability of transitioning from the current state to the next.

The first order Markov chain constructed identified two time series segments in the historical set (transitions from state 2 to state 4) corresponding to a machine exhibiting odd behavior. However, the machine was not in an operational environment in both segments which does often display odd behavior.

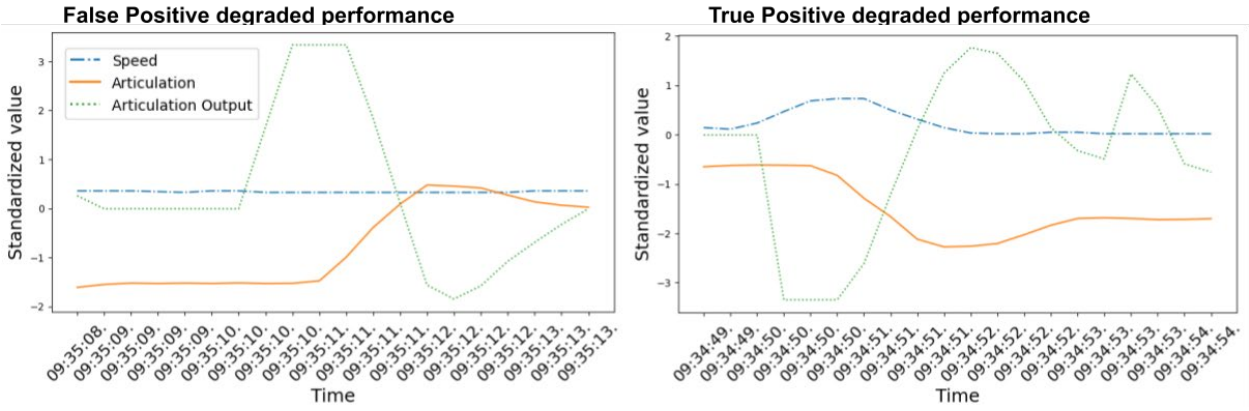

Figure 3 The figure displays segments identified by a second-order Markov chain. The segment labeled 'True Positive' exhibits a snaking pattern in the articulation output. This snaking pattern is associated with overshooting, a condition where the machine turns excessively and needs to reverse direction. The articulation output serves as the control signal directing the machine to articulate either to the left or to the right and the articulation variable is the actual articulation angle.

The second order Markov chain classified 34 segments as positive out of a total of 664. One of two instances of degraded performance previously identified by a technician was successfully classified as such. Additionally, a segment classified as positive exhibiting a similar problem in the machine was uncovered (illustrated in Figure 3). The false positive case seemed to correspond to the vehicle mistakenly beginning a turn before continuing straight resulting in an odd and sudden change in articulation. As such, the robustness of the methodology against classifying odd maneuvers as degraded performance requires further work.

4. Conclusions and Future Work

A proof of concept has been established for the potential of developing an unsupervised predictive maintenance framework based on symbolic time series analysis. While it was demonstrated on the benchmark dataset that the framework does not have high accuracy, the method does exhibit higher precision than would be expected from a random classifier. In the context of predictive maintenance, precision might be more valuable for identifying if any poor performance indicators are at all present with high confidence rather than identifying all such instances. Future work encompasses testing the performance of the methodology on different representations of the data, validating the findings with the RCT technicians and testing other partitioning methods for deriving the time series symbolization.

5. Acknowledgements

I extend my gratitude to client mentors Charlie Musca and Kevin Winchester for their invaluable mentorship throughout this project. My appreciation also goes to RCT technicians Brodie McLennan and Andres Koiv, as well as RCT Data Scientist Yulia Novskaya, for their insights and constructive feedback. I commend Jeremy Leggoe and Kimberlie Hancock for orchestrating the CEED program and ensuring a rewarding experience for all participants. Lastly, heartfelt thanks to my academic supervisors Débora Corrêa and Rachel Cardell-Oliver for their support and counsel.

6. References

Coleman, C., Damodaran, S., Chandramouli, M., & Deuel, E. (2017). Making maintenance smarter: Predictive maintenance and the digital supply network. Deloitte Insights.

Das, G., Lin, K. I., Mannila, H., Renganathan, G., & Smyth, P. (1998, August). Rule Discovery from Time Series. In KDD (Vol. 98, No. 1, pp. 16-22).

Daw, C. S., Finney, C. E. A., & Tracy, E. R. (2003). A review of symbolic analysis of experimental data. Review of Scientific instruments, 74(2), 915-930.

Géron, A. (2022). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. " O'Reilly Media, Inc.".pp.260-274.

Katser, I. D., & Kozitsin, V. O. (2020). Skoltech anomaly benchmark (skab). Kaggle.

Keogh, E., & Lin, J. (2005). Clustering of time-series subsequences is meaningless: implications for previous and future research. Knowledge and information systems, 8, 154-177.

Liao, T. W. (2005). Clustering of time series data—a survey. Pattern recognition, 38(11), 1857-1874.

Rajagopalan, V., Ray, A., Samsi, R., & Mayer, J. (2007). Pattern identification in dynamical systems via symbolic time series analysis. *Pattern Recognition*, *40*(11), 2897-2907.

Ray, A. (2004). Symbolic dynamic analysis of complex systems for anomaly detection. *Signal processing*, *84*(7), 1115-1130.

Susto, G. A., Beghi, A., & De Luca, C. (2012). A predictive maintenance system for epitaxy processes based on filtering and prediction techniques. IEEE Transactions on Semiconductor Manufacturing, 25(4), 638-649.