SPE-180452-MS

Data Mining with Shapelets for Predicting Valve Failures in Gas Compressors

Om P. Patri, Arash S. Tehrani, Viktor K. Prasanna, Rajgopal Kannan, University of Southern California; Anand Panangadan, California State University Fullerton; Nabor Reyna, Chevron Information Technology Company

Abstract

Gas compressor failures are frequently caused by breakdown of valves. Since production is dependent on rotating equipment, it is useful to minimize downtime caused by such valve failures, and try to predict them in advance. This is a challenging problem, which we address using Big Data analysis of the data gathered by a large number of sensors deployed on various parts of the compressor. These sensors take periodic readings (at every few minutes) of various physical properties of the compressors including motor winding temperatures, compressor vibrations, and pressure and temperature for both suction and discharge at various compression stages. We frame this problem as a multivariate time series classification task, and propose a novel machine learning approach to solve it.

Our proposed approach is based on the concept of shapelets, which are discriminative subsequences extracted from time series. This approach does not make assumptions about the nature of the dataset (crucial for real industrial datasets) and has very fast classification times. These shapelets act as a 'signature' capturing the characteristics and differences between sensor data related to normal valve function versus failed valve function. Shapelets are increasingly being used for univariate (single dimension data read by one sensor) time series data mining. But there have been few efforts to solve the problem of multivariate time series classification using shapelets due to the additional challenges emanating from multiple sensors in terms of the size and variety of data. Specifically, the existing approaches make the assumption that the reading of sensors are independent, which is not the case for sensor data in gas compressors as variation or anomaly in a valve affect the reading of adjacent sensors. Since all the sensors record data synchronized in time, the temporal dependencies across them need to be captured.

In this work, we propose a method, which attempts to incorporate these dependencies into the final shapelet-based classification framework. We achieve this using a heuristic of inter-leaving time series data across the sensors. This helps us reduce the multivariate time series data to a univariate format such that existing univariate shapelet extraction methods can be applied directly on the data. We evaluate our approach on real sensor data taken from gas compressors in an oil field in North America. Our results illustrate that time series approaches based on shapelet mining are valuable for fast prediction of failures from sensor data in oil and gas fields. These approaches provide key insights into the functioning of the individual sensors as well as deliver a visual aid to domain experts for further root cause analysis.

1. Introduction

Data mining and machine learning approaches have brought forward a revolution in the process of monitoring, detecting and predicting failures in oilfield equipment [1]. Due to heavy instrumentation in modern oilfield equipment, we have new challenges in dealing with data from multiple sensors (e.g. pressure, temperature and vibrations) to effectively integrate these streams of sensor data and make predictions [5]. Such predictions are useful to detect and prevent equipment failure, reduce downtime, and increase production. The aim of this work is to find signatures in multivariate compressor sensor data, which may aid in the prediction of valve failure, and as a result create a path to prioritize and monitor maintenance schedules for compressors, which are often on remote platforms.

This works builds upon our previous paper SPE-174044-MS [9], where we proposed feature selection approaches for ranking sensor dimensions in sensor data, which is the first step towards performing classification. Here, we propose an advanced approach for classification of multivariate time series sensor data, which is based on using time series 'shapelets'. Time series shapelets [12] focus on extracting discriminative subsequences (or 'signatures') from within the training dataset, that are most relevant for distinguishing between the positive and negative classes.

The advantages of using a shapelet-based approach are: (i) shapelet-based approaches do not make any assumptions about the nature, source or distribution of the input data which is crucial for analysis of complex, real-world oilfield data, (ii) fast classification times because most of the computational complexity is in extracting shapelets from the training data - this training dataset is not needed during classification, (iii) shapelets are visually interpretable and domain experts can use them for deeper investigation and root cause analysis, and (iv) shapelets have been shown to be effective and highly accurate for a variety of time series data mining tasks including classification, clustering, and predictive analytics [6 - 12]. The state of the art supervised univariate shapelet extraction method, which is the one upon which we base our approach, is known as Fast Shapelets [11].

However, most algorithms for shapelet mining only work on univariate time series. There are a few approaches (such as [2, 3, 8]), which propose extensions of the univariate shapelet extraction method for multivariate use cases, but all of them make the assumption that shapelets can be extracted from different sensors independently of each other. This assumption does not hold for real-world oilfield data where one sensor may affect the reading of other sensors (e.g. change in temperature can cause a change in pressure).

In this work, we explore interleaving measurements from multiple sensors as a means of applying univariate shapelet-based classification algorithms to multivariate time-series. Interleaving is treated as a generalization of the approach of concatenating time-series from different sensors. Since the time series we record consists of readings at synchronized time intervals across the sensors, it is intuitive to incorporate the relations between the readings of sensors into the final multivariate shapelet extraction process. In this work, we propose a simple heuristic approach to do this. We demonstrate that this interleaving-based approach provides an effective multivariate time-series classification algorithm, and evaluate it on gas compressor sensor data from an oilfield in North America.

2. Approach

We first illustrate the intuition behind our approach. Let us consider a toy example to see why concatenation is effective in certain tricky scenarios for shapelet-based algorithms, typically for complex, real-world datasets. Consider a case with two sensors where a time-series is to be classified to be in the positive class if a distinct shape (e.g., a peak) appears in Sensor 1's data stream followed by its appearance in Sensor 2. The appearance of these shapes in any other order indicates that the time-series is to be classified as the negative class. Two instances of this type are shown in Figure 1.

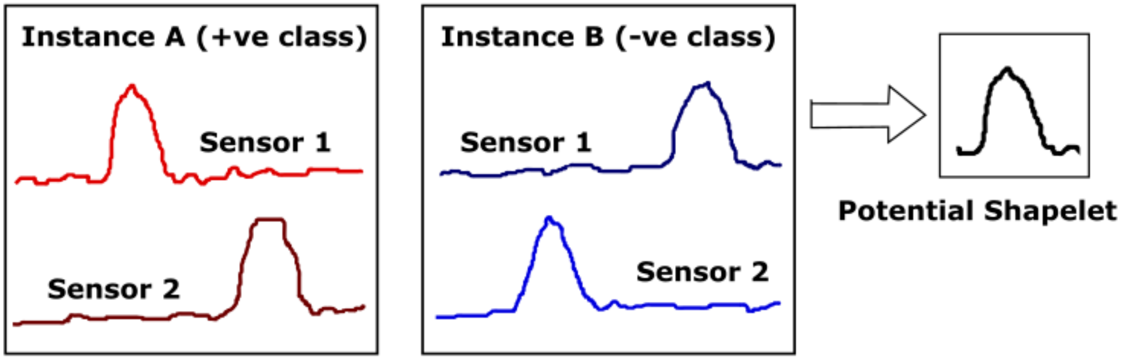

Figure 1: Two multivariate instances each consisting of two time series (two sensors). The peak in the data is a potential shapelet candidate but it is NOT discriminative enough to differentiate between the two instances (which belong to opposite classes), because a similar pattern occurs in all four time series.

In this case, most existing multivariate shapelet-based classification algorithms that extract shapelets from each sensor independently would fail since the mere presence or absence of the shapelets is not sufficiently discriminative. On the other hand, if the two sensor streams are concatenated, a (single) discriminative shapelet composed of the distinct shapes from each sensor can be identified, as shown in Figure 2.

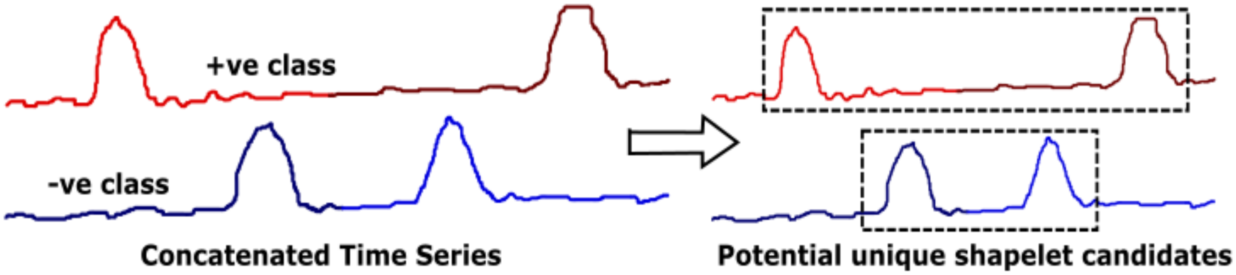

Figure 2: Time series formed by concatenating data from all sensors for each instance in the previous example. Now the potential shapelet candidate from the concatenated time series can discriminate between the two classes.

However, shapelets extracted based on concatenating fixed length segments are not able to capture localized patterns. In addition, the discriminative capacity of such multivariate shapelets depends on the relevance of each sensor stream to the class; the inclusion of time series segments from irrelevant sensors in the concatenated shapelet will increase the error rate in shape pattern matching during classification. We have therefore developed an interleaving and sensor ranking-based approach to make the extracted multi-variate shapelets invariant to the length of segments and number of sensors. Note that even though the concatenated signature or shapelet found may cross boundaries between the two sensors (thus consisting of a pattern which does not actually exist in any one sensor), it does not matter as long as it is a discriminative pattern - since we will apply the same pre-processing to concatenate both the training and testing data.

Our proposed approach is called Inter-leaved Shapelets (ILS), also introduced briefly in [10]. The overall approach is depicted in Figure 3. The idea is to inter-leave time series segments across sensors from multiple dimensions to form the final concatenated one-dimensional time series for each instance. The simplest way to do this is to consider a fixed interval inter-leaving - segment size '*k*' at which we regularly cut the time series and inter-leave the segments, i.e. first *k* elements of the first sensor, then first *k* elements of the second sensor and so on until the first *k* elements of the last sensor, after which we concatenate the *(k+1)*th to *2k*th element of the first sensor again and so on. In the end, we will have the same number of univariate time series as the number of instances.

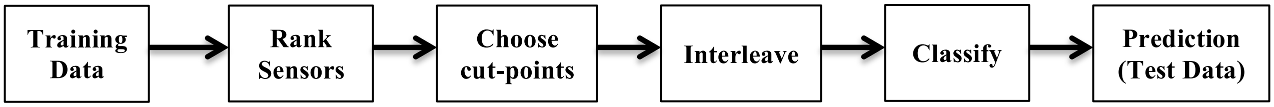

Figure 3: Overview of the proposed Inter-leaved Shapelets (ILS) approach

To improvise upon our naive approach, we rank the sensors before performing the inter-leaving and order the data according to the sensor ranks - from most important to the least. By doing this, we ensure that (i) the shapelet extraction algorithm encounters data from the highly ranked sensors first where a shapelet is more likely to be found, and (i) we can eliminate data from the lower ranked sensors completely leaving them out of the inter-leaving process (they will be at the end of each segment concatenation round). To implement this ranking scheme, we divide the training data into a training and validation set. The ranking of the sensors is based upon using the shapelets extracted from the training set to perform classification on the validation set. We use the Fast Shapelets univariate extraction approach to compute the accuracy on the validation set for each sensor separately, and then the sensor that produces the highest accuracy gets the highest rank.

Note that ILS can be extended by determining the 'cut-points' for inter-leaving automatically depending on detection of change-point events [4] in the data. This can ensure that we do not segment the time series when a change in structure is detected (such as the middle of a data peak) and retain the shape of the change structure. It also reduces the number of cuts if nothing anomalous is detected (in contrast with the current approach of segmenting at fixed intervals).

3. Evaluation

We use a real-world gas compressor sensor dataset from an oilfield in North America, with data from multiple sensors (such as temperature, pressure and vibration), to evaluate the performance of our proposed ILS approach for burnt valve failure prediction.

For our evaluation, we use data from six sensors, which were identified to be important to the functioning of the system by domain experts. We have time series data for a few years collected at the frequency of once every few minutes. We use the pre-processing approaches described in our earlier work in [9] to generate our training and testing time series datasets. We use about one-fourth of the provided training dataset for our validation set. We experiment with a range of segment sizes for fixed interval ILS, denoted by ILS-k where k is the segment size. We also compare our approach to the univariate baseline approach, which ignores the multivariate structure of the data (using Fast Shapelets [11]) as well as Shapelet Forests [8], which ignores the local temporal dependencies across sensors.

We have 211 multivariate training instances (which is equivalent to 211*6 = 1266 univariate time series), and 90 test instances (equivalent to 540 univariate time series). We used 53 of the training instances for the validation set, leaving 158 instances in our final training set. Using the baseline approaches, we achieved an accuracy of 90% on our dataset (486/540 correct predictions). We also observed the same accuracy when using the Shapelet Forests approach.

We experimented with ILS for various segment sizes, and for many of the segment sizes (the 'k' parameter), we were able to achieve the same accuracy of 90%. This proves that our new approach, which works for multivariate time series and can capture more complex shapes in the data, is also able to perform at par with other state-of-the-art approaches. Figure 4 shows a plot of the classification accuracy versus segment size parameter (k) in our evaluation. As we can observe, the accuracy is high and stable in the mid-range of sizes (from 20-35), but is lower for very low segment sizes (below 20) or high segment sizes (above 35). We conclude that given the choice of a good ‘k’ parameter, this approach performs at par with other approaches.

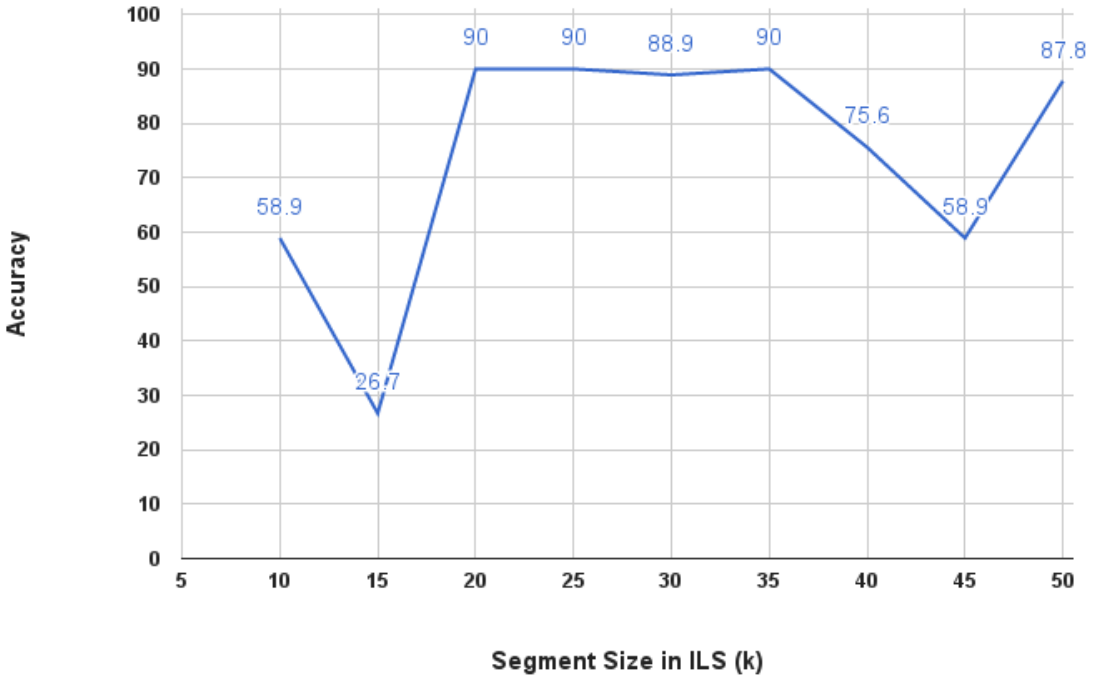

Figure 4: Classification Accuracy on Test Data vs. Segment Size Parameter in ILS (k). The highest accuracy is found for k = 20, 25 or 35.

4. Conclusions

This work presented our attempts towards a multivariate shapelet-based time series classification approach based on an inter-leaving process while performing concatenation of time series. Even with a simple fixed interval cut-point heuristic, we were able to perform at par with the baseline approaches. In the future, we will perform more rigorous evaluation of this approach and work on methods to determine the cut-points automatically through change-point detection, as well as eliminate or filter data from lower ranked sensors as required. Due to the challenging and complex nature of real oilfield sensor data, there are additional data quality issues, which can also be focused on to improve our results.

5. Acknowledgments

This work is supported by Chevron U.S.A. Inc. under the joint project, Center for Interactive Smart Oilfield Technologies (CiSoft), at the University of Southern California.

6. References

[1] A. Abou-Sayed, "Data mining applications in the oil and gas industry," *Journal of Petroleum Technology,* vol. 64, pp. 88-95, 2012.

[2] M. S. Cetin, A. Mueen, and V. D. Calhoun, "Shapelet ensemble for multidimensional time series", in *Proceedings of the 15th SIAM International Conference on Data Mining (SDM)*, 2015.

[3] M. Ghalwash, V. Radosavljevic, and Z. Obradovic, "Extraction of interpretable multivariate patterns for early diagnostics", in *Proceedings of the IEEE 13th International Conference on Data Mining (ICDM)*, pp. 201–210, 2013.

[4] M. Lavielle and G. Teyssiere, "Detection of multiple change-points in multivariate time series", *Lithuanian Mathematical Journal*, vol. 46(3), pp. 287–306, 2006.

[5] O. P. Patri, V. Sorathia, and V. K. Prasanna, "Event-driven information integration for the digital oilfield," SPE 159835-PP presented at the *SPE Annual Technical Conference and Exhibition*, San Antonio, Texas, USA, 2012.

[6] O. P. Patri, A. Panangadan, C. Chelmis, R. G. McKee and V. K. Prasanna, "Predicting Failures from Oilfield Sensor Data using Time Series Shapelets," SPE 170680-MS presented at the *SPE Annual Technical Conference and Exhibition*, Amsterdam, 2014.

[7] O. P. Patri, A. Panangadan, C. Chelmis, and V. K. Prasanna, "Extracting discriminative features for event-based electricity disaggregation," presented at the IEEE Conference on Technologies for Sustainability, Portland, Oregon, USA, 2014.

[8] O. P. Patri, A. Sharma, H. Chen, G. Jiang, A. Panangadan, and V. K. Prasanna, "Extracting Discriminative Shapelets from Heterogeneous Sensor Data", in *Proceedings of the 2014 IEEE International Conference on Big Data (IEEE BigData)*, 2014.

[9] O. P. Patri, N. Reyna, A. Panangadan, and V. K. Prasanna, "Predicting Compressor Valve Failures from Multi-Sensor Data", SPE 174044-MS presented at the *SPE Western Regional Meeting*, Garden Grove, California, USA, 2015.

[10] O. P. Patri, R. Kannan, A. Panangadan, and V. K. Prasanna, "Multivariate Time Series Classification Using Inter-leaved Shapelets", presented at the *Time Series Workshop in Neural Information Processing Systems (NIPS)*, 2015.

[11] T. Rakthanmanon and E. Keogh, "Fast shapelets: A scalable algorithm for discovering time series shapelets," in *Proceedings of the thirteenth SIAM conference on data mining (SDM)*, 2013.

[12] L. Ye and E. Keogh, "Time series shapelets: a new primitive for data mining," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009, pp. 947-956.