FAULT DETECTION ON BIG DATA: A NOVEL ALGORITHM FOR CLUSTERING BIG DATA TO DETECT AND DIAGNOSE FAULTS

Avery J. Smith, Kody M. Powell*
University of Utah
Salt Lake City, UT 84108

Abstract Overview

With computer technology improving exponentially, data will grow incomprehensibly in size, complexity, and noise. However, latent within the data, valuable signals are hidden that, if discovered and analyzed, can offer abundant benefits, such as fault detection. Traditionally, principal component analysis has been used to perform fault detection in large, multivariate systems. However, these methods often struggle to find the true origin, as they are susceptible to contribution smearing. In this work, a chemical plant system was analyzed and a novel cluster and detect method for fault detection utilizing machine-learning clustering algorithms was created in aim to improve fault detection time and diagnosis. Plant data containing complex variables were simulated, clustered into groups, and analyzed through principal component analysis as individual groups. This approach often resulted in quicker identification and more accurate diagnosis than the traditional principal component analysis method.

Keywords

Process Monitoring and Diagnostics, Fault Detection, Fault Diagnosis

Introduction

Faults in chemical plant systems can be immensely detrimental as they increase plant downtime, decrease product yield, cause environmental problems, and raise exposure to plant safety hazards. Occasionally, faults are obvious and can be easily discovered. Other faults remain hidden and are even difficult to notice they have occurred. Data-driven fault detection techniques provide insight to how the plant is operating, even without any engineering-based process knowledge or first-principle models.

The industrial Internet of Things, expanded computer power, and next generation wireless sensors have caused manufacturing data such as flow, concentration, temperature, pressure, and dozens of other measurements to be taken multiple times per second. In larger settings, this data can be overwhelming. One tactic to approaching large manufacturing data sets is to use clustering algorithms to partition the data tags into smaller, more digestible chunks, then to apply fault detection algorithms on those individual groupings. This allows clear vision when something has gone wrong in the process, and what tags are directly contributing.

Fault detection through PCA (and other multivariate means) are well studied. Braatz, Chiang, and Russell have provided a detailed overview of several multivariate fault detection methods (Braatz et al., 2000). Methods where the plant is divided into separate groups and analyzed individually is referred to as decentralized multiblock modeling. Qin used multiblock analysis on an industrial

* To whom all correspondence should be addressed

polyester firm where the blocks were determined through process knowledge (Qin et al., 2001). Yan recently developed a way to perform the blocking in a purely data-driven method through mutual information values to perform fault detection efficiently (Jiang & Yan, 2014).

In this work, sensor data from a simulated chemical plant are clustered using a novel clustering algorithm to group the sensors into blocks. Traditional principal component analysis (PCA) is then applied on the smaller groups with Hotelling's T^2 being monitored. The proposed method proved at least equally proficient with respect to detection time and considerably improved fault isolation and diagnosis capabilities compared to the traditional PCA approaches.

Clustering

A custom-clustering algorithm was created based on correlation values between variables. After preprocessing the data through mean-centering, and normalizing by standard deviation, the absolute value of the Pearson correlation of the sensor was taken. A matrix was established comparing each sensor's correlation to the others. The rows of the matrix were analyzed and all sensor relations below a defined correlation relation threshold were deemed insignificant. If duplicate arrays existed, they were absorbed into one, single array. Furthermore, a sub-space threshold was set to determine the minimum relation between groups to be considered associated. The arrays were respectively compared to the others and any array that shared a fraction of sensors at or above sub-space threshold, were considered a match and cumulatively combined into a larger array, which was repeated until convergence. Lastly, a minimum size threshold was defined to limit lower end of the array size.

In this data set, the algorithm split the 376 sensors into nine specific groups within the data, with an additional 10th as a cumulative pot of previous groups that did not reach the minimum array size threshold. These different groups were created through purely data driven decisions, however, they resemble the individual units on the chemical plant simulation as if process knowledge was used. For instance, group one only includes tags from the second continuously stirred tank reactor (CSTR2) while group seven contains tags only associated with first plug flow reactor (PFR1) exclusively.

The different clustering groups can also be evaluated by examining a heat map of the sensor's Pearson's correlations in the order of the corresponding groups with a one value representing exact correlation, and zero as no correlation. To fully understand how the proposed clustering algorithm has given shape to the data, it is important to first review the correlation heat map with the sensors in random order (Figure 1). The random order heat map is highly disorganized and difficult to read. Without any process knowledge or a clustering algorithm, it is challenging to tell how the sensors relate.

The heat map after clustering allows the interconnected relationships between different parts of the plant to be seen (Figure 2). The first nine groups can be seen as the yellow boxes going along the diagonal on the heat map. In the bottom left corner, you will see a series of four, small yellow near identical squares, outlined in a red box. These represent the four independent CSTR's at the beginning of the process. The CSTR's (group one through four) are surrounded by very dark blue showing that each respective reactor is completely independent from the others. The sixth yellow square is comprised of sensors from part of the distillation column and is loosely correlated with all units that come before it. Square seven and square eight are the PFR's. They contain a large number of variables, indicated by their family square size. The variables within the respective PFR's are highly correlated, but are completely

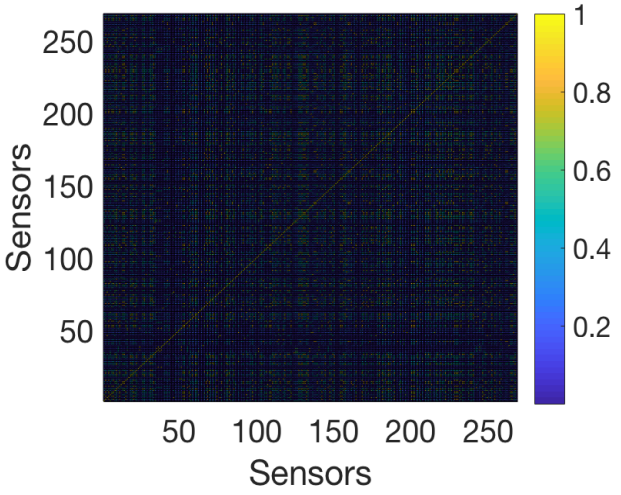

Figure 2. Sensor correlation heat map in random order

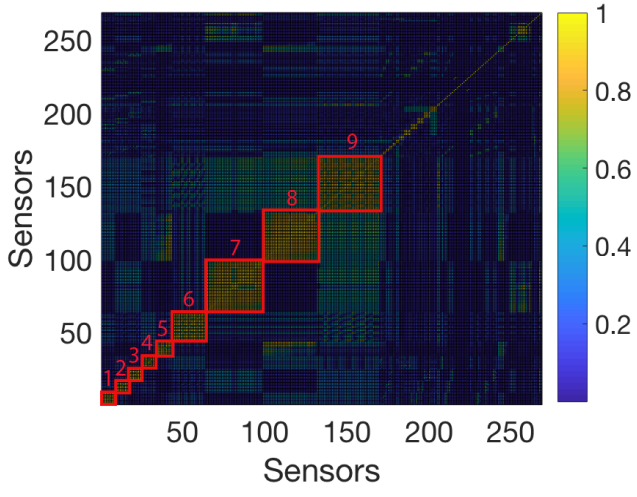

Figure 1. Sensor correlation heat map after organized by clustering method

independent of the other PFR by the two equi-sized dark squares that lie next to the PFR families. The final ninth box demonstrates the second distillation column group. The rest of the graph comes from group ten (the cumulative, small correlation groups) as well as the sensors that did not correlate above the correlation threshold with any other sensors.

Fault Detection Through PCA

A mean T^2 value as well as its standard deviation were taken across a span of normal operating behavior for the entire plant for the global method and the individual groups for the proposed clustering method. An abnormality threshold was defined to indicate when the process might have a fault. Abnormality was defined by a limit of standard deviations above the mean to signal anomaly. In this study, the abnormality threshold was defined as three standard deviations above the mean value, corresponding to a 99% confidence level using the F-distribution. Faults were considered detected when the T^2 value crossed the abnormality threshold and stayed above it for a consecutive, *M,* minutes. Once the fault has been detected, the cause of the fault can then be evaluated. Through the global method, contribution scores were calculated and the top five sensors in magnitude were reported as the primary culprits. For the novel, cluster and detect method, the T^2 values were monitored for the individual 10 groups. If a fault was detected in one of the groups through the same parameters mentioned above, that individual group was delegated responsibility for causing that fault. Contribution scores were also calculated within this group and the top five were reported. To further evaluate the difference between the global and clustering method, T^2 ratios were calculated to show relative magnitude of responses in anomaly severity compared to normal operating conditions. Finally, for this data set, no false alarms occurred. From Table 1, it can be noted that for the majority of the simulated faults, the cluster and detect method was able to either find the fault more quickly (time), with a greater surety (larger T^2 ratio), or in a more accurate location (cluster source) that the traditional global PCA method.

Conclusions

PCA is widely used to transform original process tags to reduced dimensions of principal components to perform fault detection with the use of Hotelling's T^2statistic. However, a new method is proposed that applies a clustering algorithm to group the process sensors into several groups before the PCA is performed to simplify the complexity and give operators specific locations that isolate the root causes for the fault. The algorithm was able to create ten individual clusters with high correlation within the group and extremely low with the other groups. With this clustering pre-fault detection, not only did this method show potential to find faults at a quicker rate, but also increase accuracy and speed in finding the source of the faults in a way that is bereft of the dismantling effects of contribution smearing.

References

Braatz, R. D., Chiang, L. H., & Russell, E. L. (2000). *Data-driven Methods for Fault Detection and Diagnosis in Chemical Processes* (1st ed.). Springer-Verlag London. https://doi.org/10.1007/978-1-4471-0409-4

Jiang, Q., & Yan, X. (2014). Plant-wide process monitoring based on mutual information–multiblock principal component analysis. *ISA Transactions*, *53*(5), 1516–1527. https://doi.org/10.1016/J.ISATRA.2014.05.031

Qin, S. J., Valle-Cervantes, S., & Piovoso, M. J. (2001). On unifying multiblock analysis with application to decentralizedprocess monitoring. *Journal of Chemometrics*, (15), 715–742.

Table 1. Results of cluster and detect method and traditional global PCA method for simulated faults

Fault Information				Global PCA Results			Cluster and Detect Results			
Fault #	***Fault Name***	***Associated Equipment***	***Original Source***	***Global Time***	***Global Contribution Source***	***Global T^2 Ratio***	***Cluster Time***	***Cluster Source***	***Cluster Contribution Source***	***Cluster T^2 Ratio***
1	Concentration Step	PFR 1	RX401BCC4PV	17	RX401BQPV RX401BCC5PV RX401AT1PV RX401CCD3PV RX401BT1PV	9.73	17	RX401	RX401BCC4PV RX401ACC4PV RX401ACC3PV RX401ACD5PV RX401ACC5PV	67.05
2	Heater Failure	CSTR 3	RX103QPV	17	RX101CBFPV HX103BTBPV FT202FCPV HX401TCPV RX102CAPV	72.61	45	RX103	RX103TRSP RX103CCSP HX103ATA1PV RX103TRPV RX103QPV	2,143.96
3	Reflux Valve Stuck	DC 1	DC601FRPV	15	RX104CAPV TK501TPV RX104CBPV RX103TRPV TK303LPV	465.01	15	DC601 (First Family)	DC601XC1PV DC601XC3PV DC601FRPV DC601FVPV DC601QPV	527.31
4	Gradual Fouling	CSTR 1	UA (HX101A)	1192	HX103ATAPV FT204CBPV FT203FCPV HX102ATAPV FT204CCPV	1.72	688	HX101	HX101BTB1PV HX101ATA1PV HX101BTCPV RX101CCSP RX101TRSP	16.12