Soft Sensor Change Point Detection and Root Cause Analysis

Liang Cao[1†], R. Bhushan Gopaluni[1], Lim C. Siang[2], Yankai Cao[1], Jin Li[2]

[1]Department of Chemical and Biological Engineering, University of British Columbia, BC, V6T 1Z3, Canada
(E-mail: bhushan.gopaluni@ubc.ca)

[2]Department of Process Control Engineering, Burnaby Refinery, BC, V5C 1L7, Canada

Abstract: Soft sensor has been playing an indispensable role in the process monitoring of key process variables. How to know if deployed soft sensor models are still performing well is a challenging but crucial topic for the industry. If there exists change points in soft sensor predictions, it indicates abrupt and significant changes in the process conditions. The presence of change points may require us to rebuild the model to ensure that it does not drift. Root cause analysis plays an important role in process monitoring when a change point occurs. Fast and accurate change point attribution is essential for timely recovery of model performance. This work proposes a straightforward way to detect the change points and find the root causes of changes. Off-line change point detection is used to detect changes by formulating change point detection as a discrete optimization problem. Then, we work on understanding which feature or combination of features that are shifting soft sensor predictions. Shapley additive explanations (SHAP) is adopted to explain the predictions of soft sensor model. It connects optimal contribution distribution with local explanations using the classic Shapley values. Finally, the effectiveness of proposed algorithms is validated on a real industrial data.

Keywords: Change point detection, Root causal analysis, Soft senor, Shapley additive explanations

1. INTRODUCTION

The real industrial process often has different temporary working conditions, where the change points occur irregularly. Therefore, it is challenging for engineers to quickly and accurately locate change points, and find the root cause of changes [1]. Change point detection can be regarded as dividing the time series into multiple segments with different piece-wise stationary distributions [2]. Due to the lack of well-labeled samples and complex process conditions, machine learning is more promising on rapid localization and discovery of change points than relying on human experts. Change point detection has received extensive attention from researchers and has driven the development of fields such as finance [3], biology [4], and signal processing [5], etc.

How do we know there exists change point in the deployed soft sensor model? One obvious and simple way is to calculate the residuals between the soft sensor predictions and the lab results (after aligning their timestamps, whether it's the squared error, absolute error or some other metric). Residual detection [7] directly depends on the accuracy of the model, but the inaccuracy of the model is inevitable, because it is very likely to encounter process conditions that were not present during the model training. This makes it difficult for us to distinguish the change points from the residuals. Even though the residual is correctly calculated, this does not provide much information about detecting change points. For example, how bad should the residuals be before we consider it as a change point? It is not easy to set a threshold for residuals to ensure excellent performance under various working conditions.

To identify change points with very little assumptions on the underlying distribution of data, we can formulate change point detection problem as a discrete optimization problem [2, 6]. The algorithm is described by three elements: a cost function, a search method and a constraint on the number of changes. The detailed discussion is given in section 2 .

After detecting change points, it is important to automatically locate the causes of those change points. Soft sensors have great potential for providing accurate predictions while poor performance in explaining their predictions [8]. In some cases, we do not care why a decision was made, it is enough to know that the predictive performance on a test dataset was good. However, in industrial process monitoring, knowing the 'why' is very important. With the stringent requirements for product quality and cost, the complexity and automation degree of the industrial process are continuously growing. As the scale of plants grows, it is urgent to enhance the safety, reliability, and robustness of the process. Any change points may lead to disastrous consequences. Therefore, it is not enough to detect the change point, the model must also explain the reason why change point occurs. The current root cause analysis is mainly done with engineers' expertise, which is not feasible in complex processes.

Multivariate statistical process monitoring (MSPM) models, such as principal component analysis (PCA) [9], partial least squares (PLS) [10], independent component analysis (ICA) [11], are efficient in complex processes fault diagnosis and have received widespread attention from scholars. MSPM generally assumes that the normal process data X obeys a statistical distribution $p(X)$ in a steady state, and uses normal operating condition data to estimate $p(X)$. The corresponding threshold is then determined based on hypothesis testing at a given significance level α. Within MSPM framework, the contribution of each feature to the statistics is calculated to identify the root causes. MSPM is based on the assumption of

† is the presenter of this paper.

steady state of the process, however, industrial processes usually are non-stationary, such as catalyst activity reduction, migration of process conditions, etc.; if the steady state model is used to explain the process, the credibility of the results will inevitably be seriously affected.

Considering the characteristics of real industrial processes, a new method is needed to interpret the output of soft sensors with less assumptions about the model and data. The Shapley value [12, 13] is a method to fairly distribute contribution of each player in cooperative game theory, which can be used to explain the prediction of any machine learning model. In soft sensor modeling, a player is a individual input or a group of inputs of soft sensor, a game is the prediction of soft sensor. It provides not only global explanations but also local explanations. If the Shapley value attribution is represented as a linear additive feature model, then it will be Shapley additive explanations (SHAP) model [14]. As [14] mentioned, the Shapley value is the only machine learning explanation method that guarantees a fair distribution among the features.

This work aims to introduce Shapley value into the analysis of soft sensor change points. The remaining part of this article is organized as follows. In Section 2, detailed explanations of change point detection is given. Section 3 reviews the Shapley value and the SHAP method. Section 4 presents a case study on the real process data to verify the effectiveness of the proposed method. Section 5 closes the paper with a summary.

2. CHANGE POINT DETECTION

Define signal $y = \{y_1, \cdots, y_T\}$ and assume y to be piece-wise stationary, meaning that some characteristics of the process change abruptly at some unknown instants $t_{1:K} = \{t_1, t_2, \cdots t_K\}$, where $t_1 < t_2 < \cdots < t_K$. Change point detection is to estimate those instants when y is observed. Define $V(\tau, y)$ as the total cost when choosing possible segmentation τ and it can be given as follows:

$$\begin{aligned} V(\tau, y) := \sum_{k=0}^{K} c\left(y_{t_k . t_{k+1}}\right) = c\left(\{y_\tau\}_1^{t_1}\right) + c\left(\{y_\tau\}_{t_1+1}^{t_2}\right) \\ + \cdots + c\left(\{y_\tau\}_{t_k+1}^{t_{k+1}}\right) + \cdots + c\left(\{y_\tau\}_{t_{K-1}+1}^{t_K}\right) \end{aligned} \tag{1}$$

where $\{t_1, t_2, \cdots, t_k, \cdots\}$ represents the change point indexes, $c\left(y_{t_k . t_{k+1}}\right)$ is a cost function for segment $y_{t_k . t_{k+1}} = \{y_{t_k}, y_{t_k+1}, y_{t_k+2}, \cdots, y_{t_{k+1}}\}$, K is the number of change points.

The segment cost $c\left(y_{t_k . t_{k+1}}\right)$ is expected to be low if the segment is homogeneous (without change points within the segment) and high if the segment is heterogeneous (with change points within the segment). Many cost functions have been proposed, such as l_1 cost, l_2 cost, Poisson cost, kernel based cost, etc. In this work, segment cost function $c\left(y_{t_k}\right)$ is defined as l_2 cost:

$$c\left(y_{t_k . t_{k+1}}\right) := \sum_{t=t_k}^{t_{k+1}} \left\| y_t - \bar{y}_{t_k . t_{k+1}} \right\|_2^2 \tag{2}$$

where $\bar{y}_{t_k . t_{k+1}}$ is the mean of segment $y_{t_k . t_{k+1}}$.

For soft sensors, the number of change points K is unknown, a regularizer $pen(\bullet)$ on the number of segment is required to reduce overfitting. The choice of penalty is strongly associated to the magnitude of the detected change. A small penalty will result in detection of many change points, even noise-induced fluctuations. A large penalty, on the contrary, will lead to only a few important change points being detected, or even none. There are several choices for penalty, such as, linear penalty, Bayesian Information Criterion(BIC), Akaike Information Criterion (AIC), etc. In this work, liner penalty $pen(\tau)$ is chosen and it is defined as follows:

$$pen(\tau) := \beta \left|\tau\right| \tag{3}$$

where β represents the regularization coefficient. The smaller β, the weaker the penalty. To search the change points, pruned exact linear time (Pelt) is chosen as search method with linear penalty. It is fast with $O(t)$ computational efficiency, which is ideal for large-scale industrial data. Fig.1 gives an example of change point detection. Finally, for any $y_t \in \tau$, the change point detection problem can be formulated as a discrete optimization problem:

$$\min_{\tau} V(\tau, y) + pen(\tau) := \sum_{k=0}^{K} \sum_{t=t_k}^{t_{k+1}} \left\| y_t - \bar{y}_{t_k . t_{k+1}} \right\|_2^2 + \beta \left|\tau\right|$$

Fig. 1 An example of change point detection

3. SHAPLEY ADDITIVE EXPLANATIONS

The goal of our work is to distribute the contribution of each feature to the prediction. Let $w_x(S') = \frac{|S'|!(M - |S'| - 1)!}{M!}$, then the Shapley value of feature i is defined as follows:

$$\phi_i(f, x) = \sum_{S' \subseteq M \setminus i} w_x(S') \left[f\left(S' \cup \{i\}\right) - f(S') \right] \tag{4}$$

where f is the black box model, x is the input datapoint, $\phi_i(\bullet)$ is the Shapley value of feature i under model f, M is the number of input features, S' is a subset of the features. For $w_x(S')$, the denominator $M!$ represents all possible feature combinations; the numerator $|S'|!(M - |S'| - 1)!$ means the appearance times of

$S^{'} \cup \{i\}$ in all $M!$ combinations; $f\left(S^{'} \cup \{i\}\right) - f(S')$ indicates the expected marginal contribution of feature i in one combination. The Shapley value of a feature i is the weighted average contributions under all feature combinations.

If the Shapley value attribution is represented as a linear additive feature model, then it will be Shapley additive explanations model, which is given as follows:

$$f(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j \tag{5}$$

where $f(\bullet)$ is the explanation model, ϕ_0 is the base prediction without knowing any input information, M is the number of input features, and ϕ_j is the the distributed contribution for feature j, $z' \in \{0,1\}$ is the subset features vector, 1 indicates that the corresponding feature is present while 0 is absent. From this definition, we can conclude that if all features are present, $f(z') = \phi_0 + \sum_{j=1}^{M} \phi_j$ and if all features are absent, $f(z') = \phi_0$. To prove Eq.5, assume a linear model $\hat{f}(x)$ with p features:

$$\hat{f}(x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p \tag{6}$$

where β_j is the coefficient of feature j. For one feature j, its contribution on the model prediction $\hat{f}(x)$ can be calculated as:

$$\phi_j(\hat{f}) = \beta_j x_j - E(\beta_j X_j) = \beta_j x_j - \beta_j E(X_j) \tag{7}$$

where $E(\beta_j X_j)$ is the average estimated effect of feature j. If all of p features are present, the total contribution of p features on data point x can be given as follows:

$$\begin{aligned} &\sum_{j=1}^{p} \phi_j(\hat{f}) = \sum_{j=1}^{p} (\beta_j x_j - E(\beta_j X_j)) \\ &= (\beta_0 + \sum_{j=1}^{p} \beta_j x_j) - (\beta_0 + \sum_{j=1}^{p} E(\beta_j X_j)) \\ &= \hat{f}(x) - E(\hat{f}(X)) = \hat{f}(x) - \phi_0 \end{aligned} \tag{8}$$

Fig.2 shows an example of SHAP analysis, the black box model has 4 inputs and the model output is 1.34. Based on the SHAP analysis, feature 1 has the largest contribution while feature 3 has the smallest contribution.

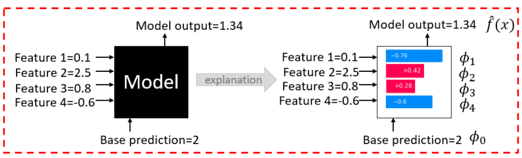


Fig. 2 An example of SHAP analysis

4. SHAP ANALYSIS FOR SOFT SENSOR CHANGE POINTS

In this work, we propose an efficient method to detect the soft sensor change points and find the root causes of changes. The framework of proposed method is given in the Fig.3. By converting change point detection problem to discrete optimization problem, the position of change points in soft sensor predictions is detected. Then, Shapley additive explanations (SHAP) is adopted to explain and locate the root causes of the changes with the Shapley value.

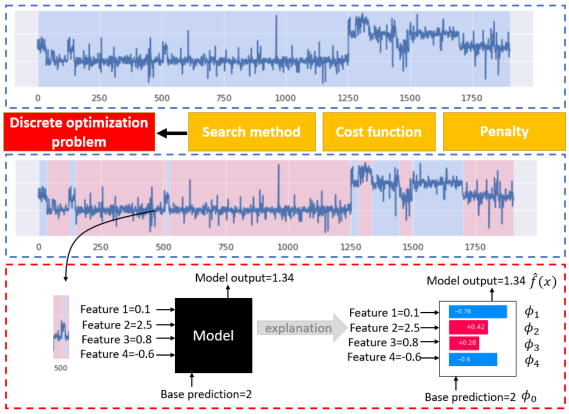


Fig. 3 The framework of proposed method

5. CASE STUDY

In this section, a real soft sensor data from Parkland refinery in Burnaby, British Columbia,Canada, is used for case study. Considering the data security issues, the feature names are anonymized and the data is preprocessed. The commercial process data is filtered by removing data that is beyond a certain threshold and the outliers. 3-σ limits are used to set the upper and lower threshold limits. Then, data standardization is adapted to rescale the range so that standardized data X has zero-mean and unit-variance. Fig.4 shows the data after preprocessing. The output of the soft sensor is the key performance indicator (KPI), and the inputs are feature 1-9. Fig.5 gives the correlation between all of the process variables.

5.1. Change point detection

For KPI, it is easy to find that there exists some change points, which represents abrupt changes in soft sensor predictions. For change point detection of the KPI, the penalty β is set as 10, the cost function is chosen as l_2 and change point is searched by Pelt method. Fig.6 gives the change point detection result of KPI, 31 change points are detected.

With the presence of change points, it is essential to find the root cause of the change. If the change is caused by some faults or variations in the inputs, rebuilding the soft sensor is required to ensure a stable and reliable prediction.

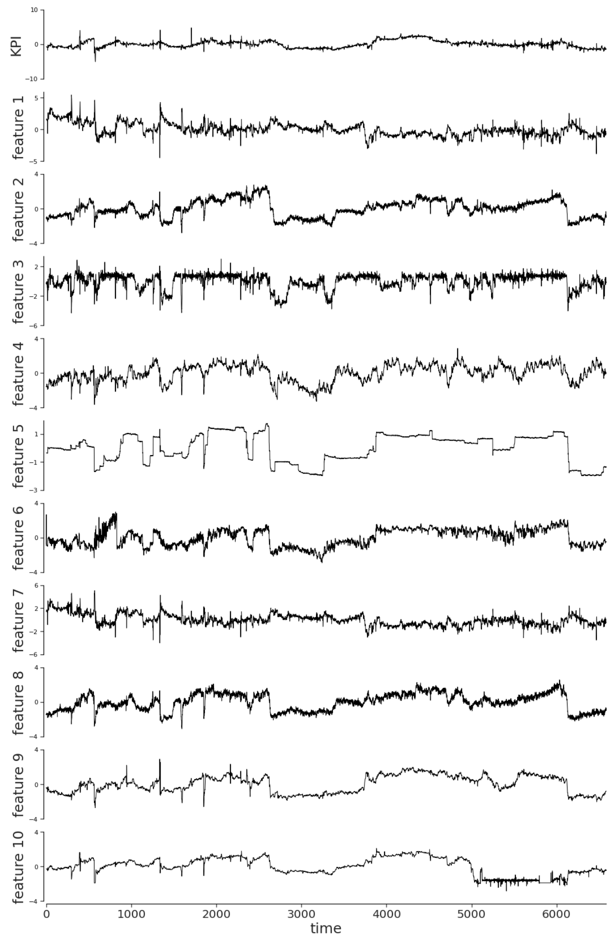
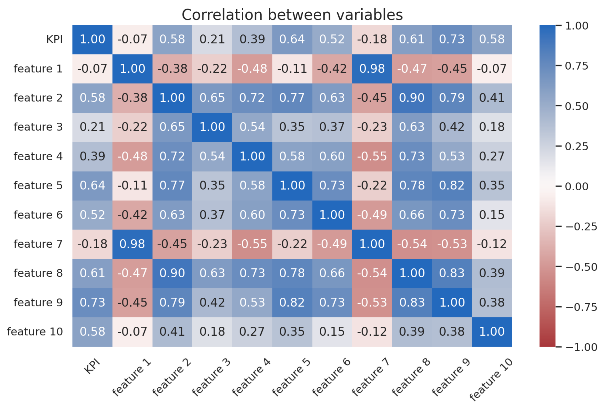


Fig. 4 Graphical representation of real process data



	KPI	feature 1	feature 2	feature 3	feature 4	feature 5	feature 6	feature 7	feature 8	feature 9	feature 10
KPI	1.00	-0.07	0.58	0.21	0.39	0.64	0.52	-0.18	0.61	0.73	0.58
feature 1	-0.07	1.00	-0.38	-0.22	-0.48	-0.11	-0.42	0.98	-0.47	-0.45	-0.07
feature 2	0.58	-0.38	1.00	0.65	0.72	0.77	0.63	-0.45	0.90	0.79	0.41
feature 3	0.21	-0.22	0.65	1.00	0.54	0.35	0.37	-0.23	0.63	0.42	0.18
feature 4	0.39	-0.48	0.72	0.54	1.00	0.58	0.60	-0.55	0.73	0.53	0.27
feature 5	0.64	-0.11	0.77	0.35	0.58	1.00	0.73	-0.22	0.78	0.82	0.35
feature 6	0.52	-0.42	0.63	0.37	0.60	0.73	1.00	-0.49	0.66	0.73	0.15
feature 7	-0.18	0.98	-0.45	-0.23	-0.55	-0.22	-0.49	1.00	-0.54	-0.53	-0.12
feature 8	0.61	-0.47	0.90	0.63	0.73	0.78	0.66	-0.54	1.00	0.83	0.39
feature 9	0.73	-0.45	0.79	0.42	0.53	0.82	0.73	-0.53	0.83	1.00	0.38
feature 10	0.58	-0.07	0.41	0.18	0.27	0.35	0.15	-0.12	0.39	0.38	1.00

Fig. 5 Correlation between process variables

5.2. SHAP analysis

SHAP is a powerful method to fairly distribute the contribution of each feature to the change. There exists a Python package developed by Lundberg et al [14] that can easily visualize the output of any machine learning model. Bayesian Ridge regression is utilized to build soft sensor and linear explainer is chosen to explain the output of Bayesian Ridge model.

Fig.7 displays the impact of inputs on the soft sensors prediction. Each instance is represented by a single dot on the feature row with the SHAP value on x the axis.

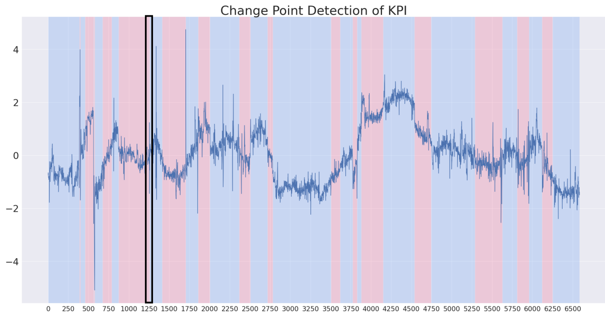


Fig. 6 Change point detection result of KPI

The color bar represents the raw values of the feature, and the dots ”stacked” along each feature row represent the density. Fig.8 presents the average of absolute SHAP values over all samples. The y axis is the feature importance order and x axis is the Shapley value of each feature. The red means the feature possesses positive impact while the blue is negative impact. Here the base prediction $\phi_0(E(\hat{f}(X)))$ is 0.087. To explain the global SHAP feature importance, take feature 9 and feature 5 as an example: for all data, the first most important feature 9 will contribute 1.01 to the soft sensor output compared to the base prediction while the second most important feature 5 will contribute -0.82 to the soft sensor output.

Fig. 7 Global SHAP feature importance

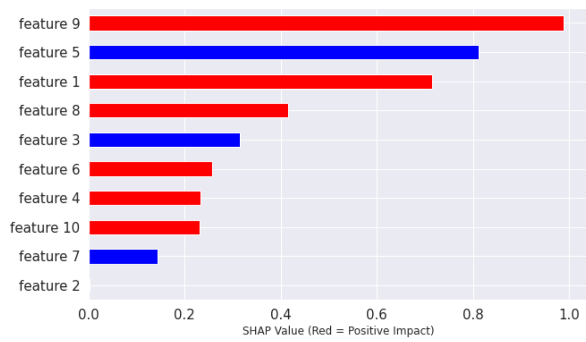


Fig. 8 The average of absolute SHAP values

Since the global SHAP feature importance plot contains no more information beyond the importance, and the features that are most globally important are not necessarily the ones causes change. To find the root cause of change, a more informative plot with local explanations of inputs is needed.

Fig.9 is the Shapley value summary plot of each feature at each instance. $f(x)$ is the soft sensor prediction, the blue means the feature at that instance has a negative contribution for the prediction while the red means the positive contribution. Take the first instance (red zone) as an example, feature 9 has positive contribution on the soft sensor prediction while feature 1 has negative contribution. The darker the color, the greater the contribution. For example, feature 1, 9, etc have greater contribution than feature 2, 5, etc.

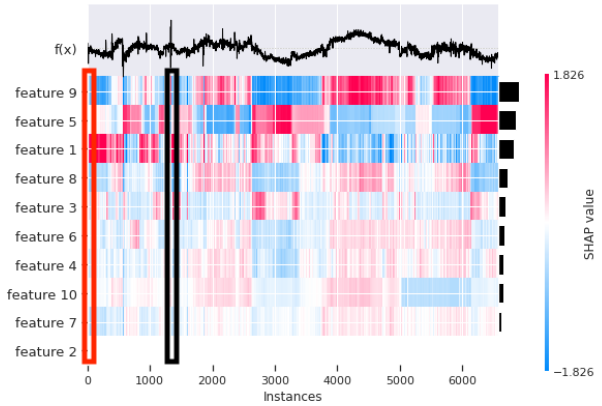


Fig. 9 Local SHAP values for all data points

After detecting the location of change points, understand what tags that are shifting inferential predictions is important. Take one change point that detected in 1260 (black zone) as an example, figure 10 shows a set of Shapley value of different input features. The Shapley value of feature 5 is -0.76, which means the feature 5 has the largest negative contribution. Interestingly, feature 9 has negative contribution (-0.6) at instance 1260 while the global contribution (1.01) is positive. When adding up all Shapley values (-0.25), the base prediction 0.087 ($E(\hat{f}(X))$) will become the final model output -0.163 ($f(x)$).

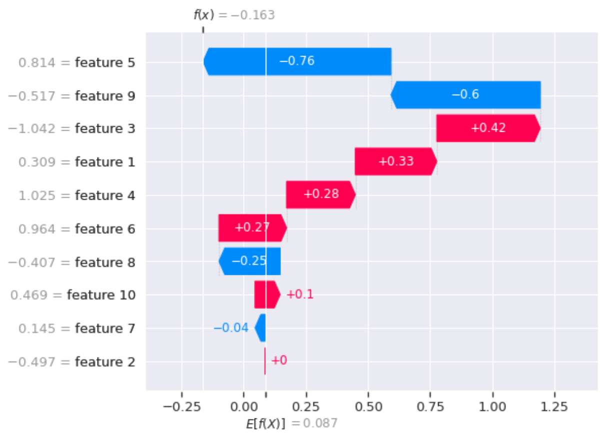


Fig. 10 Local SHAP values for instance 1260

From Fig.10, we are able to conclude that feature 5 is more likely to be the root cause of change point 1260. The reason we label feature 5 as root cause is that it has the largest contribution. The black zone of Figs. 11-14 indicates the values of features 1,3,5,9 at data point 1260. We can easily find that there is a large change in the value of feature 5. At the same time, the values of features 1 ,3, 9 have relatively obvious drifts, which means that they also have a large contribution to the change.

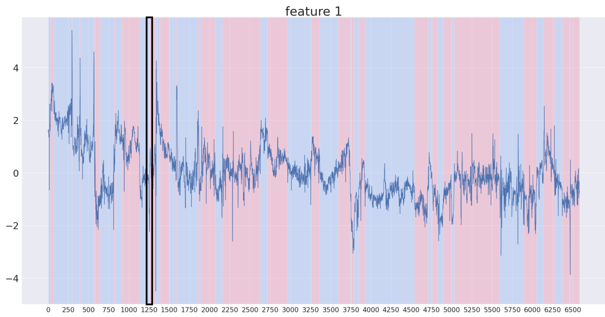


Fig. 11 Change point detection result of feature 1

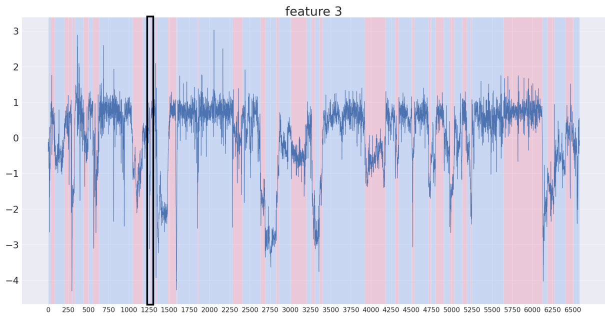


Fig. 12 Change point detection result of feature 3

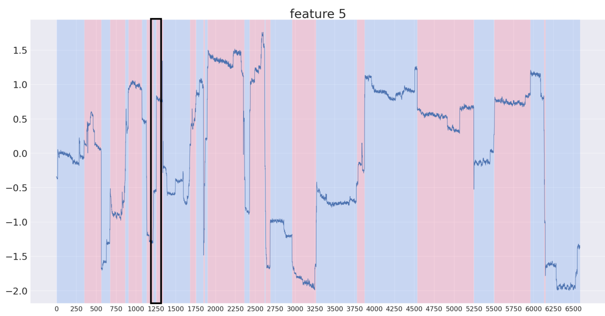


Fig. 13 Change point detection result of feature 5

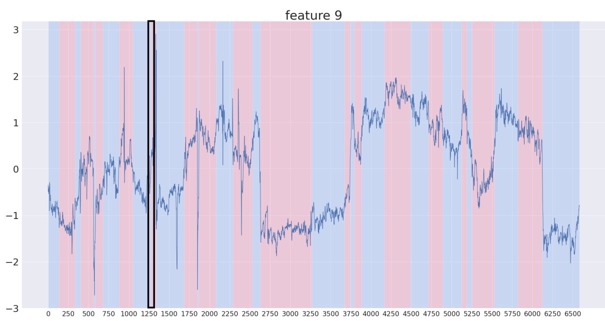


Fig. 14 Change point detection result of feature 9

5.3. Discussion

The overall target of this work is to assist engineers to better evaluate the soft sensor performance and locate the root cause. Change point detection analysis and SHAP analysis are able to guide the human expert to evaluate and improve the performance of soft sensors. However, various practical problems still exist.

First, it is important to point out the SHAP interpretation method does not exactly mean causality in most cases, the Shapley value of each feature is different from the true causal effect. SHAP could fairly distribute the contributions of each feature and discover informative relationships between the input features and the prediction, but the interpretation would be misleading if the distribution of training data and test data are dramatically different [15, 16]. In our future work, we would like to introduce causal models and causal inference into SHAP analysis.

Second, simultaneous occurrences of multiple root causes is a tough problem [17]. As we can see in case study, for KPI change point 1260, input features 1,3,5 and 9 have changed significantly at the same time. It is difficult to tell exactly whether just one feature or a combination of features caused the change. Therefore, conducting temporal analysis to infer whether a change point includes multiple root causes is essential.

6. CONCLUSION

In this work, a novel method to detect the soft sensor change points and find the root causes of changes is proposed. Change points of soft sensor are detected by solving a discrete optimization problem. As an important contribution, this work uses SHAP to fairly distribution the contribution of each feature. Without requiring strict assumptions on the model and data distribution, the proposed algorithm is able to perform well on any model. The real-world commercial refinery case study validates the effectiveness of the proposed method.

ACKNOWLEDGEMENTS

Liang Cao and Bhushan Gopaluni would like to thank MITACS for financial support. We thank Dr.Jianping Su and Mr. Yixiu Wang for technical support and guidance. We gratefully thank colleagues at Parkland for providing data and computing resources.

REFERENCES

[1] Y. Luo, B. Gopaluni, Y. Xu, L. Cao, Q.X. Zhu, "A novel approach to alarm causality analysis using active dynamic transfer entropy", *Industrial Engineering Chemistry Research*, Vol. 59, No. 18, pp. 8661–8673, 2020.

[2] C. Truong, L. Oudre, N. Vayatis, "Selective review of offline change point detection methods", *Signal Processing*, Vol. 167, 2020.

[3] K. Frick, A. Munk, H. Sieling, "Multiscale change point inference", *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, Vol. 76, pp. 495–580, 2014.

[4] T. Hocking, G. Schleiermacher, I. Janoueix-Lerosey, V. Boeva, J. Cappo, O. Delattre, F. Bach, J.P. Vert., "Learning smoothing models of copy number profiles using breakpoint annotations.", *BMC Bioinformatics*, Vol. 14, pp. 164, 2013.

[5] V. Jandhyala, S. Fotopoulos, I. Macneill, P. Liu, "Inference for single and multiple change-points in time series", *Journal of Time Series Analysis*, Vol. 34, No. 4, pp. 423–446, 2013.

[6] R. Maidstone, T. Hocking, G. Rigaill, P. Fearnhead, "On optimal multiple changepoint algorithms for large data", *Statistics and Computing*, Vol. 27, No. 2, pp. 519–533, 2017.

[7] R. Isermann, "Model base fault detection and diagnosis methods", *Proceedings of 1995 American Control Conference*, Vol.3, pp. 1605–1609, 1995.

[8] L. Cao, F. Yu, F. Yang, Y.K. Cao, B. Gopaluni, "Data-driven dynamic inferential sensors based on causality analysis", *Control Engineering Practice*, Vol.104, 2020.

[9] Z.J.Lou, Y.Q.Wang, S.Lu, P.Sun, " Process Monitoring Using a Novel Robust PCA Scheme", *Industrial Engineering Chemistry Research*, Vol.60, pp. 387–398, 2021.

[10] C. Shang, X.Q. Gao, F.Yang, W.X.Lyu, D.X.Huang " A comparative study on improved DPLS soft sensor models applied to a crude distillation unit", *IFAC-PapersOnLine*, Vol.48, pp. 234–239, 2015.

[11] J.M.Lee, S.J.Qin, I.B. Lee, " Fault detection and diagnosis based on modified independent component analysis", *AIChE Journal*, Vol.52, pp. 3501–3514, 2006.

[12] S. Lloyd, " Notes on the n-Person Game – II: The Value of an n-Person Game", *Santa Monica, Calif.: RAND Corporation*, 1951.

[13] M. Sundararajan, A. Najmi, " The Many Shapley Values for Model Explanation", *Proceedings of the 37th International Conference on Machine Learning*, pp. 9269–9278, 2020.

[14] S.M. Lundberg, S.I. Lee, " A unified approach to interpreting model predictions", *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 4768–4777, 2017.

[15] D. Janzing, L. Minorics, P. Bloebaum, " Feature relevance quantification in explainable AI: A causal problem", *Proceedings of the 23th International Conference on Artificial Intelligence and Statistics*, pp. 2907–2916, 2020

[16] B.W Huang, K. Zhang, J.J. Zhang, J. D. Ramsey, R. S.Romero, C. Glymour, B. Schölkopf, " Causal Discovery from Heterogeneous/Nonstationary Data", *Journal of Machine Learning Research*, Vol.21, pp.1–53, 2020.

[17] X. Zhang, L. Xiong, N. Sun, M. Wang, H. Tang, Y. Zhao, " Accurate Inference of Unseen Combinations of Multiple Rootcauses with Classifier Ensemble", *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 9306–9310, 2022