論 文 の 内 容 の 要 旨

論文題目 A STUDY ON DATA VISUALIZATION AND STATISTICAL PROCESS MONITORING FOR NONLINEAR SYSTEMS（非線形システムを対象にしたデータの可視化および統計的プロセス監視に関する研究）

氏 名 デ・ソウザ・エスコバール・マテウス　Matheus de Souza Escobar

1 Introduction

Machine learning can be used to highlight important characteristics of a system by analyzing the relationship of variables and samples. One particular element is explored in this thesis: fault detection. A combination of human knowledge, experience and statistical analysis is used, leading to information that can be used for future assessment. One might argue, however, on how reliability of the information available, since they can lead to misinformed decisions and false conclusions regarding the nature of the dataset and quality of obtained results.

Initially, the distinction between supervised and unsupervised approaches is explored. To challenge this notion, an unsupervised monitoring methodology is proposed, combining Generative Topographic Mapping (GTM)[1] and Graph Theory[2]. GTM highlights system features, reducing variable dimensionality. Graph Theory visualizes this information, through a network, clustering similar samples. Simulation data sets, Tennessee Eastman Process[3] and a real industrial case study are presented for validation of this approach, comparing it against supervised and unsupervised Principal Component Analysis (PCA)[4], Dynamic Principal Component Analysis (DPCA)[5], Kernel Principal Component Analysis (kPCA)[6] and GTM. Complementary, the study also focuses on applicability domain (AD) and fault detection, analyzing how the definition of proper training and test data sets affect modeling. Distinct data splitting scenarios are created, evaluating predictive modeling and anomaly detection capabilities[7]. A flour dataset is used for assessment, allied to Genetic Algorithm Partial Least Squares (GAPLS)[8] and Genetic Algorithm-based Wavelength Selection (GAWLS)[9], which are used for modeling. Their non-deterministic nature is essential for AD evaluation, analyzing variation and average of prediction errors.

2 GTM & Graph Theory Combined Approach

2.1 Literature Review

2.1.1 Generative Topographic Mapping

GTM is a widely used technique applied for visualization of high-dimensional data. It consists of a probabilistic non-linear approach, where a low-dimensional latent variable **z** is represented in a 2D space, so to approximate original data **x** as a high-dimensional manifold on the original data space. Once the map is trained, it is possible to determine for each sample a PD profile in the latent grid. PD profiles are unique for each sample, which allows the similarity assessment of all samples on the same basis.

2.1.2 Graph Theory (GT)

Graphs are symbolic representations of networks that model pairwise relations between objects[2]. When it comes to process monitoring, exploring the potential of samples as the main elements of a network is proposed [10], leading to insights on fault detection and data visualization. Graphs share two elements: nodes and edges. The former represents observations (samples) and the latter, connections between those observations. Given the similarity assessment aforementioned, an adjacency matrix (AM) can formalize this web of links, by representing all connections.

2.2 Proposed Strategy

The main methodology explored in this section involves two elements: extraction of essential information and effective data clustering. This is achieved by combining GTM and Graph Theory[10]. Primarily, GTM highlights relevant information. Every sample in the latent space has a unique PD profile, which is used for similarity assessment. Each assessment between samples fills one element of the AM. With the AM built, LCF can cluster data into similar groups. For unsupervised fault identification, it is assumed that faults are less frequent than normal states. Normal data, on the other hand, represents a stable majority of the samples available. Normal samples, therefore, have a higher number of connected nodes combined with higher connection density. It is also important to notice that anomalous data might be detected as not one cluster, but several, representing different fault characteristics or different states over time.

2.3 Results and Discussion

Different case studies were analyzed. TEP, for example, is a realistic virtual industrial process, with 53 variables. In order to evaluate fault detection capabilities, 21 preprogrammed faults are available. Nine scenarios were chosen for analysis, tackling different anomalies present in chemical plants. The proposed methodology was compared against unsupervised PCA, DPCA, kPCA and GTM, outperforming them all. The next step relies on comparing GTM+GT against their supervised counterparts. Table 1 shows false negative (FN) anomaly detection values, where close to zero values indicate that few anomalies were wrongly detected as normal data. DPCA had the best performance overall, where PCA, kPCA and GTM did not discriminate F5 well. GTM+GT, overall, performed better than the other techniques presented, even as an unsupervised methodology.

Table 1: FN in percentage for unsupervised, supervised and proposed approaches.

	F1	F2	F5	F6	F7	F8	F12	F13	F17
Uns. PCA	71.2	22.9	85.2	16.0	81.0	64.0	67.7	75.6	58.3
Uns. DPCA	50.8	23.1	85.4	16.2	80.6	74.4	67.3	79.2	58.1
Uns. KPCA	40.3	18.7	60.4	10.7	81.3	55.7	59.3	62.4	57.6
Uns. GTM	100	100	100	100	100	100	100	100	100
GTM+GT	0.6	0	0	0	0	3.4	6.4	3.0	5.6
Sup. PCA	0.42	2.50	49.2	0	0	2.5	5	3.12	11.25
Sup. DPCA	0	2.08	0	0	0	1.87	1.87	2.08	12.92
Sup. KPCA	0.80	1.81	41.6	1.20	0	2.81	4.22	3.01	12.05
Sup. GTM	0	2.08	51.7	0	0	2.71	3.75	3.12	10

The industrial case study explores an Exhaust Gas Denitration Process at Mitsui Chemicals, Inc. 37 variables were used for assessing normal and anomalous states. Two distinct case studies were considered, each with 2000 samples, consisting of normal and anomalous data. For GTM+GT, Figure 1 shows the networks obtained.

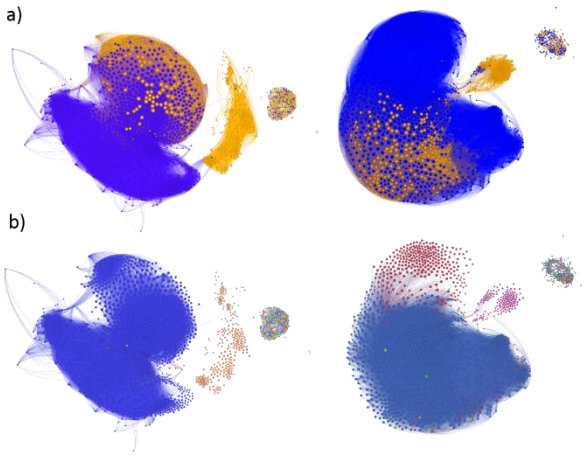

Figure 1. GTM+GT results for cases A and B, showing reference labels and clustering results.

Initially, one can notice that discrimination is fair for case A, but is has poor performance for case B, mainly due to clustering performance issues. Moreover, it can be noticed how some outliers are actually similar to the normal cluster, to the point of correlating with lots of samples. This is an indication that the labels defined by the operators do not match necessarily the actual states in the plant, which could be assessed, though, via GTM+GT. As for the other approaches, most of the detection results for PCA, DPCA, kPCA and GTM performed poorly when it comes to discriminating normal and anomalous data, based on operator's reference labels, except for supervised kPCA and GTM. T^2 has good discrimination for kPCA, while Remapping Error benefits GTM.

3 Data Splitting & Fault Detection

3.1 Literature Review

3.1.1 Applicability Domain

AD represents to what extent a model can be used to predict other data. Analyzing AD can lead to multivariate statistical monitoring, where model accuracy can be monitored through its AD over time[11]. From this premise, the concept of prediction trustworthiness is important. In the realm of distance-based methodologies, less robust, yet simpler, approaches can be seen. T^2 and Q statistics[12], for example, are used for assessing AD. All PCA based methods use this approach for monitoring. Another alternative relies on exploring the standard deviation of prediction errors, where the same sample is predicted several times using different models whose methodological source is similar. Genetic algorithms are used for the assessment. The use of standard deviation allows the evaluation of prediction anomalies. Low average prediction errors should lead to small standard deviations and high prediction errors, to big ones. When high prediction errors have low standard deviation, however, values are being predicted poorly consistently, indicating measurement error in Y-variables.

3.2 Proposed Strategy

For the work presented in this thesis, GAPLS and GAWLS are used for the generation of several predictions for each sample. Standard deviation of output variable prediction errors against the average prediction error plots can, thus, be easily obtained for all samples, as expressed in Equations 1 to 3,

$$\bar{y}_{pred_i} = \sum_{k=1}^{L} y_{pred_{ik}} / P \tag{1}$$

$$\sigma_i = \sqrt{[1/(P-1)] * \sum_{k=1}^{P} \left(y_{pred_{ik}} - \bar{y}_{pred_i}\right)^2} \tag{2}$$

$$\Delta y_i = \left|\bar{y}_{pred_i} - y_{test_i}\right| \tag{3}$$

where P is the number of available predictions, y_{pred_i} is the i^{th} predicted output value, y_{test} is the i^{th} output reference value and σ_i is the standard deviation of the i^{th} sample prediction errors. σ_i shows how consistent the prediction is for recurrent models being tested against the test data set.

3.3 Results and Discussion

The initial data set analyzed consists of 34 samples from different brands and types of flour, where the protein content was measured in triplicate using two distinct techniques: Fluorescence and NIR spectroscopy. The analysis portrayed here has an intimate connection with a basic pre-processing philosophy that is ignored by many researchers: knowing how to define your training and test data for soft sensor modeling. Different data split set-ups were devised, following a training:test ratio 0.7:0.3, so to evaluate the impact of different data sets on prediction. Due to its non-deterministic approach, 30 GAWLS and GAPLS predictive models were generated for each scenario, so to assess the reliability of each data splitting, where GAWLS is applied only to NIR. Figure 2 shows the standard deviation of prediction errors plot against the average prediction error for all scenarios.

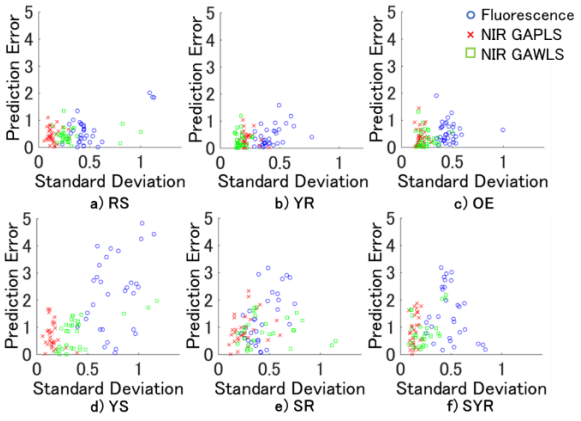

Figure 2. Standard deviation and prediction errors plot for Fluorescence GAPLS models and NIR GAPLS and GAWLS models.

Two different aspects of data affect the soft sensor performance: data splitting and outlier contamination. For all ES scenarios, training data set is contaminated with two anomalous experiments. Due to the proximity of both training and test data set, however, the prediction is not really affected by it. Looking exclusively at prediction accuracy, thus, would be deceptive, since the anomaly would never be detected. For SS scenarios, separating entire flour samples decreased the overall accuracy of the model, but allowed a better visualization of the true nature of the system. Furthermore, for SYR it showed that one entire triplicate was faulty. This affects SYR prediction error, since the model is trying to predict samples that do not belong to the AD of the model created Similarly, for all ES scenarios, one experiment of the same sample (the one in the test data) was detect as anomalous. This confirms, thus, the potential of this methodology for anomaly detection..

4 Conclusion

Regardless of the application, relying on the data being used for assessment is desirable. To what extent can one trust on the already available information is challenged here by showing different aspects of modeling. In order to support that, a combined GTM and Graph Theory was proposed initially, so to allow a fresh perspective on process data. TEP and the real data case studies aimed to justify the use of an unsupervised approach. GTM extracts relevant information for similarity assessment of distinct samples, while minimizing the impact of unnecessary features. Graph Theory takes this information and encapsulates similar data through the core relationship between samples. The results presented here for both case studies reveal the potential of such unsupervised methodology.

Complementarily, one should be careful on how training and test data sets are created, considering both prediction reliability and fault detection. The way data is selected impacts the final accuracy and predictive capability of models developed. Biased data sets affect model's performance and constrain their application. GAPLS and GAWLS non-deterministic structure led to several models generating different predictions for each sample. By evaluating the average prediction error for each sample against the variation of such errors, the AD of all scenarios could be assessed and anomalies on Y-values could be detected.

Acknowledgements

The author acknowledges the support of the Core Research for Evolutionary Science and Technology (CREST) project 'Development of a knowledge-generating platform driven by big data in drug discovery through production processes' of the Japan Science and Technology Agency (JST)

References

[1] C. M. Bishop, M. Svensn, and C. K. I. Williams. Gtm: The generative topographic mapping. 10(1):215–234, 1998. [2] [2] F. Harary. *Graph Theory*. Perseus Books, 1994.
[3] J. J. Downs and E. F. Vogel. A plant-wide industrial process control problem. *Computers & Chemical Engineering*, 17(3):245–255, 1993. [4] I.T. Jolliffe. *Principal Component Analysis*. Springer, 2002.
[5] W. Ku, R. H. Storer, and C. Georgakis. Disturbance detection and isolation by dynamic principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 30(1):179–196, 1995.
[6] J. Lee, C. Yoo, S. W. Choi, P. A. Vanrolleghem, and I. Lee. Nonlinear process monitoring using kernel principal component analysis. *Chemical Engineering Science*, 59(1):223–234, 2004.
[7] M. S. Escobar, H. Kaneko and K. Funatsu. Flour concentration prediction using gapls and gawls focused on data sampling issues and applicability domain. *Chemometrics and Intelligent Laboratory Systems*, 137(0):33–46, 2014.
[8] K. Funatsu, K. Hasegawa and Y. Miyashita. Ga strategy for variable selection in qsar studies: Application of ga-based region selection to a 3d-qsar study of acetylcholinesterase inhibitors. *Journal of Chemical Information and Modeling*, 37(2):5, 1997.
[9] K. Funatsu, M. Arakawa and Y. Yamashita. Genetic algorithm-based wavelength selection method for spectral calibration. *Journal of Chemometrics*, 25(1):10, 2010. - John Wiley & Sons, Ltd.
[10] M. S. Escobar, H. Kaneko and K. Funatsu. Combined generative topographic mapping and graph theory unsupervised approach for nonlinear fault identification. *AIChE Journal*, pages n/a–n/a, 2015.
[11] H. Kaneko, M. Arakawa and K. Funatsu. Applicability domains and accuracy of prediction of soft sensor models. *AIChE Journal*, 57(6):1506–1513, 2011.
[12] Q. Chen, U. Kruger, M. Meronk and A. Y. T. Leung. Synthesis of t2 and q statistics for process monitoring. *Control Engineering Practice*, 12(6):745–755, 2004.