17th European Symposium on Computer Aided Process Engineering – ESCAPE17
V. Plesu and P.S. Agachi (Editors)

Exploring and Improving Clustering based Strategies for Chemical Process Supervision

Rodolfo V. Tona V., Antonio Espuña, Luis Puigjaner

Chemical Engineering Department, Univeritat Politècnica de Catalunya (UPC), E-08028, Barcelona, Spain.email : Luis.puigjaner@upc.edu

Abstract

In last 12 years, Clustering has received much interest for Process Engineering problems. Particularly, the combination of fuzzy clustering with multivariate statistical techniques for Process Supervision Strategies (PSS) has been studied. The above has led to several approaches. However, some clustering associated problems has been ignored. Also, existing PSS have not been compared. In this work, Clustering based PSS (CPSS) are briefly reviewed and a comparison of it is made. This comparison incorporates some novel strategies that adequately treat some identified problems and it is illustrated through several case studies. The results shows the improvements reached with the proposed strategies.

Keywords

Clustering, Fuzzy, Supervision, Multi-operational Processes, Data Mining.

1. Introduction - Clustering for Process Supervision Strategies

Clustering is very popular because offers the chance to obtain information of previously undetected groups from data. The number of proposed clustering techniques is high[1,2]. Their capacities have been explored to support problems of processes supervision. The reported Clustering based Process Supervision Strategies (CPSS) can be grouped as:

- MSTFC strategies: Multivariate Statistical Techniques (MST), like Principal Component Analysis (PCA), are used to obtain Reduced Representations of

Data Process (RRDP). Then, RRDP are analyzed with Fuzzy Clustering (FC) techniques like Fuzzy C-Means (FCM), or Credibilistic FC (CFCM) [3,4,5].

- S_{ACP} strategies: The original data set is divided in groups according to an operational (batch sizes) or a supposed (expected time length faults) criterion. Then, PCA is used to obtain a model of each group. Finally, a PCA-based index (S_{ACP}) is used to measure the similarity among groups [6,7].
- ANN strategies: They are similar to MSTFC because an initial RRDP is obtained (usually with PCA). Then, ANN like Self-Organizing Maps (SOM) or Adaptive Resonance Theory 2 (ART2) are used to analyze the data[8,9].

After a detailed revision of the above approaches, it can be observed that:

- Existing strategies are useful to identify operating regions from historical data. This information is potentially useful to design fault detection/diagnosis systems, to monitor multi-operational processes, to discover causes of past poor performance and so on.
- All clustering techniques are recognized as very sensitive to noise and outliers in Data mining literature [1,2]. The problem of noise has been addressed but nothing has been made with regards to the outliers.
- ANN based clustering highly depend on different parameters. Also, training efforts are frequently high in terms of computing time. The above problems noticeably limit their use within CPSS.
- S_{ACP} strategies are basically useful for cases where data can be divided in groups of equal size.
- MSTFC have been the most explored and applied.
- Comparative studies are needed to establish the real advantages among existing approaches.

In following sections, a comparative study between CPSS is summarised. Only MSTFC are considered for being the most studied in the literature and successfully tested on industrial scenarios.

2. Combining PCA and Fuzzy Clustering for CPSS - MSTFC strategies

2.1. Fuzzy Clustering

In Fuzzy Clustering (FC) it is considered that an object can be a member of different classes at the same time. The classical FC technique is FCM. It is based on minimizing the sum of squared Euclidean distances between data ($\mathbf{X}_k$, k=1,... n) and cluster centers (v_i, i=1,… c).

$$Min \quad J_m(U,V) = \sum_{i=1}^{c} \sum_{k=1}^{n} (\mu_{ik})^m \left\| x_k - v_i \right\|^2 \qquad (1)$$

where $1 \leq m \leq \infty$ is the fuzziness index and c is the number of clusters. The obtained fuzzy c-partition is constrained as follows:

$$\mu_{ik} \in [0,1] \quad \forall i, k, \quad \sum_{i=1}^{c} \mu_{ik} = 1 \quad \forall k, \quad \sum_{k=1}^{n} \mu_{ik} < n, \quad \forall i. \qquad (2)$$

So, FCM identify the center of clusters and calculate membership values (μ) of each data case (k) to each cluster (i). Further FC approaches have been proposed to solve some problems of FCM like handling of clusters with different forms (Gustaffson-Kessel or FCMGK), improving identification by using typicality (τ) instead of membership (Possibilistic C-Means or PCM), and ensuring good identification in front of outliers (CFCM and Fuzzy PCM or FPCM).

2.2. MSTFC strategies

MSTFC strategies for process supervision are discussed in section 1. The basic scheme of them is: 1) Initial data is dimensionality reduced with PCA[4]; 2) *Scores* from PCA are analyzed with an FC technique; 3) Plots, validation index[1] and tables are used to analyze the extracted knowledge.

3. Comparison of MSTFC strategies

In this section, a comparison between MSTFC strategies is presented. Different issues are studied (see section 3.2 and 3.3). MSTFC reported in the literature are considered: PCA combined with FCM (FCM_{PCA}), PCM (PCM_{PCA}) and CFCM ($CFCM_{PCA}$). The FPCM technique, is also used in combination with PCA ($FPCM_{PCA}$). Some additional strategies consisting on versions of the above MSTFC but using an adaptive norm distance as it is proposed by Gustafson and Kessel or GK ($FCMGK_{PCA}$, $CFCMGK_{PCA}$, $FPCM_{PCA}$) are also considered.

3.1. Four case studies

The first two cases (**E1** and **E2**) consist of two dataset with two variables. Case 3 (**E3**) is a CSTR reactor[6] used to produce a single product with different quality degrees. Case 4 (**E4**) is a chemical plant with recycle[10]. It suffers a little change in operating conditions during a long time interval. The clusters number c is know (4 in **E1**; 3 in **E2**; 3 in **E3**; 2 in **E4**).

3.2. Evaluating the partition estimation with different MSTFC

Here, the performance of different MSTFC is evaluated in terms of the quality of estimated clusters. Two validation index are used:

- Cluster Purity (P_k)[6]: For data divided into k clusters, P tries to characterize the purity of each k in terms of how many operating windows or data points of a particular condition are present in that cluster.
- Cluster efficiency (ξ_k)[6]: It is used to characterize the extent to which an operating condition is distributed in different clusters.

Datasets from each **Ei** are processed with MSTFC. P_k and ξ_k are computed for each cluster k and also their average (P_m and ξ_m). The results are shown in tables 1, 2 and 3. Because of similar results with **E1**, table for **E2** is not shown.

Table 1. Validation of clustering results for **E1** case.

			Purity				*Efficiency*			
	P_m	ξ_m	P_1	P_2	P_3	P_4	ξ_1	ξ_2	ξ_3	ξ_4
FCM_{PCA}	99	99	97	97	100	100	100	98	98	100
$FCMGK_{PCA}$	99	99	97	100	100	100	100	98	100	100
$PCMGK_{PCA}$	74	84	63	62	100	72	58	88	100	92
$CFCM_{PCA}$	98	98	98	100	97	98	100	98	96	100
$CFCMGK_{PCA}$	99	99	100	100	97	100	100	98	98	100
$FPCM_{PCA}$	99	99	100	100	97	98	100	98	98	100
$FPCMGK_{PCA}$	99	99	100	100	97	100	100	98	100	100

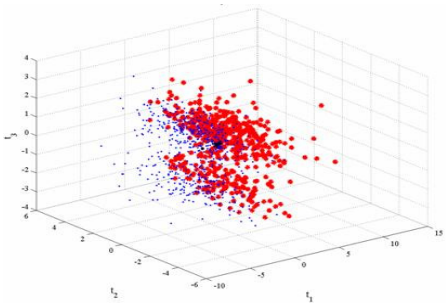

Figure 2a. Partition of **E4**-FPCM.

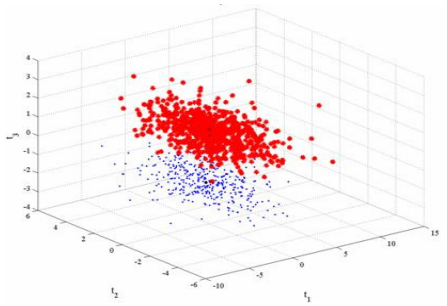

Figure 2b. Partition of **E4**-FPCMGK.

From these tables can be easily concluded that $FCMGK_{PCA}$, $CFCMGK_{PCA}$, $FPCMGK_{PCA}$ leads to better partition estimations. This is because the restriction imposes by the use of Euclidian distances (only clusters of spherical forms can be detected) disappears when an adaptive norm distance approach (GK modification) is adopted. The above is shown in fig. 2a and 2b.

Table 2. Validation of clustering results for **E3** case.

			Purity			*Efficiency*		
	P_m	ξ_m	P_1	P_2	P_3	ξ_1	ξ_2	ξ_3
FCM_{PCA}	87	69	83	78	99	58	51	99
$FCMGK_{PCA}$	99	100	99	98	100	100	100	100
$PCMGK_{PCA}$	*NaN*	97	100	54	*NaN*	91	100	100
$CFCM_{PCA}$	86	69	98	83	78	58	51	99
$CFCMGK_{PCA}$	99	100	98	100	99	100	100	100
$FPCM_{PCA}$	87	69	83	99	78	58	51	99
$FPCMGK_{PCA}$	99	100	99	100	98	100	100	100

It is observed that PCMGK tends to produce unsuitable results in many cases. This observation is important since in previous works[3] the utility of a similar technique for monitoring of a specific problem was reported. Nevertheless, when it is evaluated on different case studies the performance is not always good.

Table 3. Validation of clustering results for **E4** case

			Purity		*Efficiency*	
	P_m	ξ_m	P_1	P_2	ξ_1	ξ_2
FCM$_{PCA}$	67	55	62	72	53	58
FCMGK$_{PCA}$	100	100	100	100	100	100
PCMGK$_{PCA}$	99	99	97	100	98	100
CFCM$_{PCA}$	67	56	72	62	53	58
CFCMGK$_{PCA}$	100	100	100	100	100	100
FPCM$_{PCA}$	67	55	72	62	53	58
FPCMGK$_{PCA}$	100	100	100	100	100	100

3.3. Handling of outliers.

In the above comparison it was assumed that datasets are free of outliers. Now, outliers are considered. Only, the best MSTFC from section 3.3 are used. Furthermore, following extensions of MSTFC strategies are proposed and used:

1. The available process data matrix, **X**, is used to obtain a PCA model.
2. *Scores* from the above model are processed with the choose FC technique.
3. Depending on the choose FC, an *up* measurement is computing (see table 4).
4. Similarly to limits for the *SPE* and T^2 statistics in PCA[6], a $up_{\lim}$ is computed. This limit is based on the empirical distribution of up_i.
5. If $up_i > up_{\lim}$, the corresponding observation "*i*" is rejected as an outliers.

Table 4. up measurements.

OutMI	*MSTFC Strategy*	up_i
OutM1	FCMGK$_{PCA}$	$up_i = \mu_{i,1} \cdot \mu_{i,2} \cdot \ldots \cdot \mu_{i,c}$
OutM2	CFCMGK$_{PCA}$	$up_i = (\psi_i + 0.01)^{-1}$
OutM3	FPCMGK$_{PCA}$	$up_i = \mu_{i,1} \cdot \mu_{i,2} \cdot \ldots \cdot \mu_{i,c}$
OutM4	FCMGK$_{PCA}$	$up_i = d_{i,1} \cdot d_{i,2} \cdot \ldots \cdot d_{i,c}$
OutM5	CFCMGK$_{PCA}$	$up_i = (\sum_{k=1}^{c} d_{i,k} + 0.01)^{-1}$
OutM6	FPCMGK$_{PCA}$	$up_i = d_{i,1} \cdot d_{i,2} \cdot \ldots \cdot d_{i,c}$

3.3.1. Evaluating the performance of the OutMI methods.

The performance of the *OutMI* is set through two proposed index, the Outliers detection Efficiency (*ODEf*) and the Good Data Eliminated (*GDE*).

$$ODEf(\%) = \left(Nodr/Not\right) \cdot 100\% \quad (3)$$

$$GDE(\%) = \left(Nod - Nodr/n\right) \cdot 100\% \quad (4)$$

Where *Nodr* represents the number of outliers detected with an *OutMI*; *Not* represents the real number of outliers presents in the dataset; Nod is the number of good observations incorrectly detected as outliers; *n* is the total number of observations. If all the outliers are detected with an *OutMI*, the corresponding *ODEf* will be highest. *GDE* is compared with *Pot* (real percentage of outliers in data). So:

- If GDE>*Pot*, the *OutMI* has erroneously classifying good data as outliers.
- If GDE=*Pot*, the *OutMI* has only identifying outliers.

From table 5, it is clearly seen that no one of the methods are good for handling the case **E2**. even so, *OutM4* and *OutM6* methods are good for applying clustering together with good handling of outliers data.

Table 5. Performance of Outliers identification methods.

	E1		**E2**		**E3**		**E4**	
	ODEf	GDE	ODEf	GDE	ODEf	GDE	ODEf	GDE
OutM1	33	1.1	0	0	25	2.3	100	1.6
OutM2	66	0,6	100	0	25	0.8	50	0.1
OutM3	66	1.1	100	0	50	2	100	1.7
OutM4	100	0.6	25	3.3	100	0	100	0.1
OutM5	0	1.7	25	6.5	0	0.8	0	1.5
OutM6	100	0.6	25	3.3	100	0	100	0.1

4. Conclusions

In this work, a review of CPSS approaches has been summarized. Still more important, a comparison between different CPSS approaches has been made. The modified approaches included in this comparison allow improving some problems of current CPSS.

Acknowledgements

Financial support received from CEPIMA group (UPC) is fully appreciated.

References

1.Jain, A. K.; M. N. Murty; P. J. Flynn. ACM Computing Surveys, 31(3), 264 (1999).
2.H Han, J.; M. Kamber. Data mining: concepts and techniques. Morgan Kaufmann (2001).
3.Teppola, P., S. Mujunen y P. Minkkinen. Chem. Int. Lab. Syst., 45, 23 (1999).
4.Sebzalli, Y. M. y X. Z. Wang. Eng. App. Artificial Intelligence, 14, 607 (2001).
5.Yoo, C. K., P. A. Vanrolleghem y I. B. Lee. J. Biotechnology, 105, 135 (2003).
6.Singhal, A. y D. E. Seborg. IEEE Cont. Syst. Magazine,(October), 53 (2002).
7.Srinivasan, R., C. Wang, W. K. Ho y K. W. Lim. Ind. Eng. Chem. Res., 43, 2123 (2004).
8.Hwang, D. H. y C. Han. Cont. Eng. Practice., 7, 891. (1999).
9.Li, R. F. y X. Z. Wang. Ind. Eng. Chem. Res., 38, 4345 (1999).
10.Belanger, P. W. y W. L. Luyben. 36(1), pp.706-716. (1997).