Fuzzy clustering in the problem of technological state assessment for phosporose production

Maxim Dli[a], Andrey Puchkov[a], Tat'yana Kakatunova[a]

[a] *Branch of the National Research University Moscow Power Engineering Institute in Smolensk, Energetichesky proezd 1, Smolensk, 214013, RussiaUniversity 1, Address, City, Index, Country*

Abstract

The paper proposes a method for assessing the state of a technological process on the basis of fuzzy clustering of its parameters, and then calculation of the Kullback–Leibler divergence for clustering results taken at discrete time points. A description of the software that implements the method in question as well as the results of a simulation experiment are presented.

Keywords 1

Fuzzy clustering, Kullback–Leibler divergence, technological process state assessment

1. Introduction

Modern complex technological processes (TP) are characterized by a large number of multiply connected parameters which need monitoring and controlling in accordance with the specified algorithms to reach the objective indicators for production. A representative in this group of processes is the production of yellow phosphorus from apatite-nepheline ores wastes accumulating in large volumes in the dumps of mining and processing plants and representing a serious environmental hazard to nearby territories. Solving the problem of waste recovery and conducting related research to create information support (infoware) for the control system of this process is an urgent research problem [11].

The presence of many different technological zones and equipment leads to the fact that mathematical models for individual units and technological areas are developed by separate research teams, they have high complexity, a different set of limitations, initial conditions, a different methodological apparatus, and consequently inconsistency in the accuracy of the solutions obtained. Besides, quality assessments of individual models for technological objects are carried out according to criteria that are not agreed upon, which leads to significant difficulties in creating complex models of the entire technological system for solving optimal control problems.

Optimal control of TP begins with an assessment of its condition, which is difficult to perform in accordance with the above mentioned aspects of models building for individual engineering units and zones. The state assessment should be based on a unified approach to modeling of all TP elements, taking into account accuracy equalization of individual fragments in the complex model. Under the conditions considered, it is advisable to use methods that provide a solution to the problem of assessing the state without penetrating into the physics of the operation of a particular technological unit, but they should allow providing the required accuracy.

This opportunity is provided by machine learning methods, the spectrum of which is quite wide and continues to improve: training with a teacher (artificial neural networks, the support vector method), without a teacher (various types of clustering), with reinforcement (genetic algorithms). Each of these approaches has its own advantages and limitations, therefore, currently combined methods are widely used, implemented in boosting technologies – a consistent improvement of the obtaining solution [4]. However, in each specific application, it is necessary to create an architecture for a technological information processing system and adapt well-known methods to a specific application, which makes

Russian Advances in Fuzzy Systems and Soft Computing: selected contributions to the 8-th International Conference on Fuzzy Systems, Soft Computing and Intelligent Technologies (FSSCIT-2020), June 29 – July 1, 2020, Smolensk, Russia

EMAIL: midli@mail.ru(A. 1); putchkov63@mail.ru(A. 2); tatjank@yandex.ru(A. 3)

the problem of developing methods for assessing the state of a given TP for phosphorus production urgent.

In the proposed method for processing technological information coming from the control equipment of CPES for the production of phosphorus from apatite-nepheline ore wastes, at the beginning, clustering of the incoming information at discrete time points and accumulation of its results over a certain interval are performed. Then the differences between the clustering results are calculated, on the basis of which it becomes possible to make conclusions about the current state of TP and forecast its development.

2. Materials and methods

Clustering methods are widely used in various subject areas where the problem is to identify regularities in the data coming from instrumentation installed on a real object, simulation results, and other sources.

Clustering of technological information and analysis of its results is an integral part of the algorithmic support of TP intelligent monitoring, the improvement of which goes not only along the way of hardware upgrading, but also in the direction of complicating data processing procedures. This allows control systems to be adjusted more accurately by both the separate units and the integral TP and it allows reducing the probability of emergency operation for the equipment operation [2,13].

Clustering problems in principle do not have a unique solution, therefore, there are many clustering methods, among these methods graph methods, taxonomy (hierarchical clustering), statistical methods, Kohonen networks and fuzzy clustering can be distinguished. The choice of a method depends on the research objectives for which clustering is carried out. There is a large number of modifications for the procedure [7,8], which allows analyzing data structures, however, this is not a priority when processing technological information on phosphorus production in which there is no data hierarchy. It is significant that the other methods, like hierarchical clustering, do not contain parameters in their results that would characterize the confidence that a point of the attribute space belongs to each found cluster.

Similar information can be provided by a group of fuzzy clustering methods such as Fuzzy C-Means, Kernel Fuzzy C-Means, Gustafson – Kessel and their modification [6] including the use of ensemble clustering [10]. However, the results analysis for clustering in each case requires additional efforts to interpret them for the subject area under consideration and is carried out either using expert methods or using any other approaches.

The proposed method for analyzing the results of clustering is based on the assumption that when observing TP for a certain time, some points of the attribute space can "move" from one cluster to the other, and the dynamics of this movement can serve as a characteristic of the TP state. Clustering application allows reducing the size of the processing data to carry out the further analysis.

A simplified structural flowchart for a chemical power engineering system (CPES) of phosphorus production (circled by a dashed line) and the proposed step-by-step processing of technological information is presented in Figure 1. The CPES includes technological units (control objects): CO1 – a granulator that prepares raw pellets from waste of apatite-nepheline ores, CO2 – a multi-chamber indurating machine, where roasted pellets are produced at the exit, they enter the ore-thermal furnace indicated as CO3. All technological information is accumulated in the data base (DB) from which it goes to the block of clustering C and KLD block, calculating Kullback–Leibler divergence used to track the changes of the clustering results at various time intervals.

The vector of input/output data reflecting the composition of data flows between the control objects have the following composition: $X_1 = (D, u^1, V_1^s)$, $X_2 = (r_1, \varepsilon, u^o , V_2^s)$, $X_3 = (r_2, \sigma_k, \eta , V_3^s)$, where D – dispersion of apatite-nepheline ores waste; u^1 – mass fraction of moister in the fusion mixture; V_i^s – parameters vector, characterizing heat and physical, lithographic and granulometric properties of the corresponding material flows for the converted raw material, $i = 1, 2, 3$; r_1 and r_2 – radiuses for the raw pellet at the output of CO1 and roasted pellet at the input from CO2, respectively; u^o, ε – moister content and porosity of the pellet; σ_k – ultimate strength of the pellet; η – the degree of response in the decarbonization reactions. The output parameter for CPES is γ_p – purity of the obtained phosphorus.

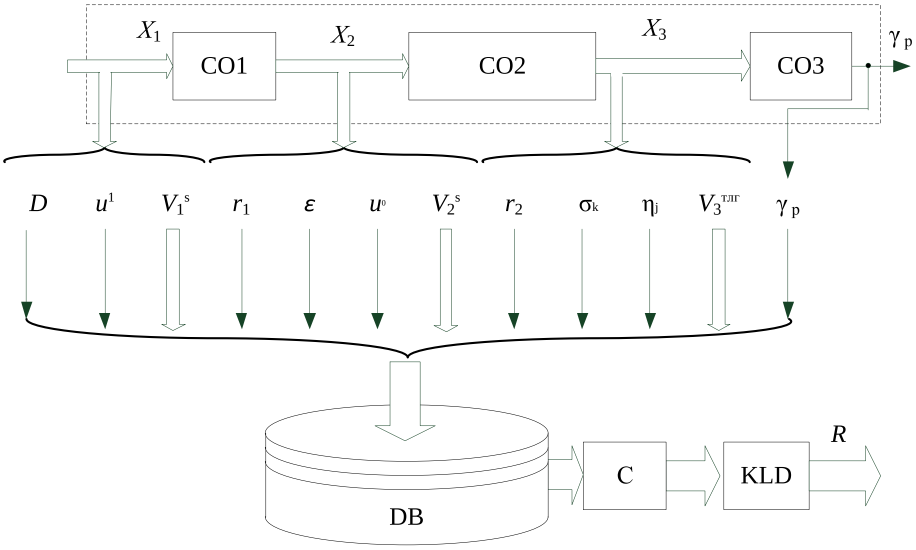

Figure 1: The structure for information processing of TP state

The statement of the problem for assessing the TP state is as follows. Given: the space of objects $X = (X_1, X_2, X_3, \gamma_p) = R^n$, where n – total number of technological parameters; m – numbers of clusters, which can be various, but it is fixed at each new start of the analysis procedure. There is a learning sample of objects $\{X^i\}$, i=1, …, l, where l – the size of learning sample. Based on this information, it is required to conduct an assessment of the CPES state at time point t_f.

The CPES can be characterized by the range number where the deviation γ_p (%) from the nominal value $\gamma_{p,nom}$ (another indicator or functional can be used) is fell into. The set of states forms the multiplicity $R=\{R_1, R_2, \ldots, R_r\}$, where r – number of states. Each number from R can have a linguistic meaning, for example, "nominal condition" (NC), "small deviation from NC", "mean deviation from NC" and etc.

If TP is carried out optimally, then it is possible to assume that the values for its parameters ζ_j from X are distributed standardly with the meansquare deviation $\sigma_\zeta(\zeta_j)$ determined from the formula: $|\zeta_{j\,\min} - \zeta_{j\,\mathrm{nom}}|/\sigma_\zeta(\zeta_j) = U_{q/2}$, where $\zeta_{j\,\min}$ – lower boundary for the field of tolerance zone parameter ζ_j, $U_{q/2}$ – quantile of $q/2$, q – permissible deviation from the nominal value. Let $(\zeta_j)_i$ and $(\zeta_j)_{i+1}$ are the values for the output parameter ζ_j of TP at i-th and (i+1)-th points of time, respectively. Then, under the assumptions made, the sequence $(\zeta_j)_i$ will be a random process with discrete time. Any deviation from the standard distribution may be a sign of violations in the preparation of raw materials and violation in the TP conditions.

The presented problem of TP states assessment is proposed to be divided into two stages: at the first stage, a fuzzy clustering of TP parameters is performed, at the second stage, the Kullback–Leibler divergence (KLD) is calculated, which allows estimating the distance of two probability distributions, obtained for different TP time intervals, from each other [12].

The theoretical aspects of fuzzy clustering are well developed, and for its implementation there are ready-made software solutions, for example, such as MatLAB, scientific libraries for Python, R and others. In these software tools, it is possible to change the methods and parameters of the clustering procedure itself, which makes it possible to carry out engineering research of the proposed algorithmic structures.

The widely known method of fuzzy clustering of C-means (Fuzzy C-means) is a generalization of the k-means algorithm, and fuzziness appears in it when describing clusters as fuzzy sets with a core in the center of the cluster [3]. Each point is included in all clusters with varying grades of membership, the sum of which is equal to one. At the output of this method, matrix U is formed containing the values of the membership functions of each point $\{X^i\}$ to the identified clusters in (1)

$$U = \begin{pmatrix} \mu_1^1 & \mu_2^1 & \cdots & \mu_m^1 \\ \mu_1^2 & \mu_2^2 & \cdots & \mu_m^2 \\ . & . & . & . \\ \mu_1^l & \mu_2^l & \cdots & \mu_m^l \end{pmatrix}, \tag{1}$$

where μ_i^j – the membership function of data X^j to cluster number i , m – the number of clusters.

However, Fuzzy *C*-means has a significant limitation, it breaks into clusters incorrectly in the case when they have different dispersion along different axes in the feature space. In order to eliminate this drawback, clustering GMM (Gaussian mixture models) is used in this method [1].

Data for clustering are parameter value packages PX_k, k=0, 2, …, K, where K – number of packages. To form set PX_k time interval T is given, in which at intervals $\Delta t << T$ the sensors of the control and measuring equipment are polled, and $T=(K-1)\Delta t$. The question of choosing T and Δt is solved basing on the characteristics of the TP, its response time, in particular, as well as the requirements for analyzing the state of the TP.

Line l in k–th package PX_k contains normalized values of TP parameters to the range [–1, 1] at the time point $l\Delta t$ *in* (2)

$$PX_k = \begin{pmatrix} x_1^1 & x_2^1 & x_3^1 & \gamma_p^1 \\ x_1^2 & x_2^2 & x_3^2 & \gamma_p^2 \\ . & . & . & . \\ x_1^l & x_2^l & x_3^l & \gamma_p^l \end{pmatrix}, \quad (2)$$

For each PX_k clustering is carried out and as a result a sequence of matrix (1) appears: U_1, U_2, ..., U_{K-1}. Mismatching between U_1 and U_k is calculated on the base of KLD in (3)

$$KL_k = \sum_{j=1}^{m}\sum_{i=1}^{l}(U_{i,j})_k \log\frac{(U_{i,j})_k}{(U_{i,j})_1}, \quad (3)$$

The sequence of scalar KL_k values obtained on the basis of (3) reflects the behavior of the process at time (variability compared with the results of initial clustering) and makes it possible to track and predict the reproducibility and stability of TP, since *KL* is a generalized characteristic obtained for PX_k data packages.

It is noted that the arrangement of TP parameters into matrix (2) can be different, for example, to fill in parameters not in one column, but in several columns at once, the main thing is that this order should not change from one PX_k to the other.

The choice of the time interval *T* should be also carried out taking into account time scales and TP characteristics. In CPES of phosphorus producing the beginning and the end of the *T* interval is advisable to take at the beginning and at the end of the next technological cycle, respectively.

3. Results and discussions

The proposed method for assessing the state of TP was tested on a software model developed in the MatLAB 2019a environment. Full-scale experiments on CEPS were not carried out, since the algorithmic part of the control system is still under development. The simulation of parameters values for TP was software-based, in terms of mathematical models of CEPS [9]. The simulation results were grouped into a file in CSV format (a text format for representing tabular data), which acted as DB in the structure shown in fig. 1. In order to test the described method for assessing the TP state, a trend was introduced into the initial data, leading to a change in the phosphorus content at the output of the ore-thermal furnace. To calculate GMM model the function *GMModel*=fitgmdist(*PX,m*) was used introduced into MatLAB, beginning with R2014a version. Clusters were built on the basis of the obtained model by the function [*...,U*]=cluster(*GMModel,* ***PX***), which returns matrix *U* containing posteriori probabilities that replace membership functions in (1).

Figure 2 shows the results of simulation experiment: at the top there are “heat maps” for PX_k, visualizing the changes of TP parameters. In the middle part of fig. 2 there is a graph of KLD indicating technological conditions, and at the bottom there is a graph of the change γ_p - output of CPES.

The conducted experiments showed that the use of KLD in the presented version to identify tendencies in the state of TP is an informative characteristic, since it allows taking into account the complex relationship of the process parameters, their mutual «location» and to identify negative scenarios of TP development at earlier stages. Thus, Figure 2 reflects the situation when the output value of CPES is still in the zone of the nominal condition while KLD has big deviation from it.

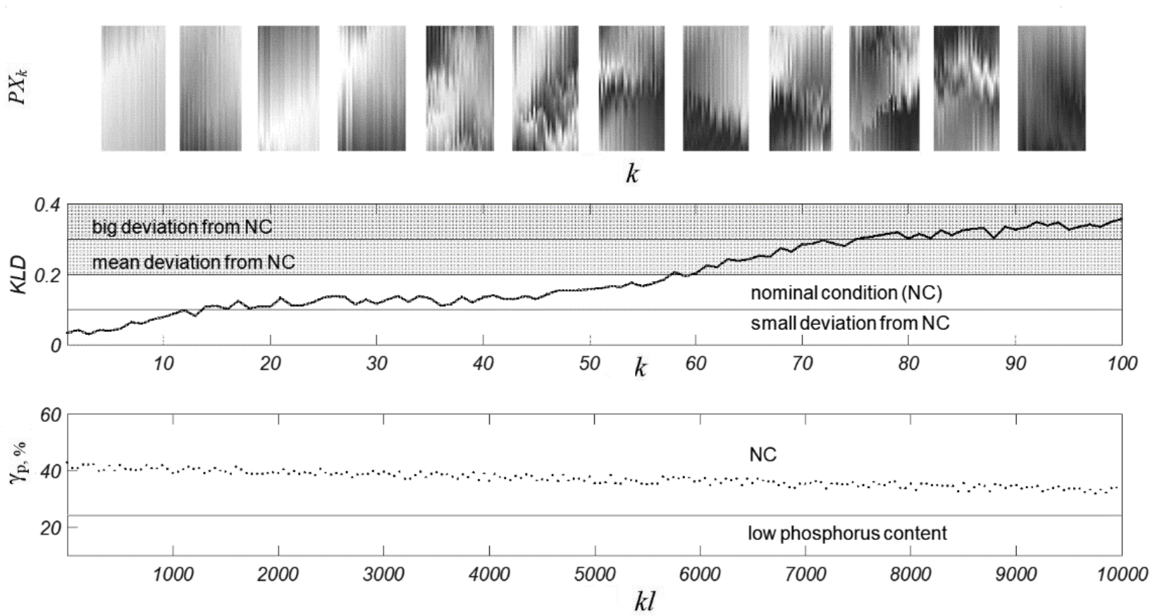

Figure 2: Visualized data packages PX_k, graph for KLD and γ_p

It should be noted, that a change in the initial number of clusters *m* (in experiments from 3 to 6) did not affect the KLD trend, but it only influences the dispersion of its values. Therefore, the proposed method can be considered slightly sensitive to the parameter *m*.

4. Conclusion

The method for assessing the state of the technological process presented in the work uses a two-stage procedure: fuzzy clustering of data and calculation of the Kullback–Leibler distance between the clustering results. Such an approach allows creating a scalar process that aggregates the variability of all technological parameters, which provides an opportunity to build simpler procedures for assessing the state of the technological process and its forecast. The developed method is quite universal and can find its application in the algorithmic support of various control systems.

5. Acknowledgements

The reported study was funded by RFBR, project No.18-29-24094.

6. References

[1] S. Adams, P. A. Beling, A survey of feature selection methods for Gaussian mixture models and hidden Markov models. Artif Intell Rev, 52 (2019) 1739-1779. doi:10.1007/s10462-017-9581-3.

[2] M. Chalouli, N. Berrached, M. Denai, Intelligent Health Monitoring of Machine Bearings Based on Feature Extraction. J Fail. Anal. and Preven, 17 (2017) 1053-1066. doi:10.1007/s11668-017-0343-y.

[3] L. M. Goyal, M. Mittal, J. K. Sethi, Fuzzy model generation using Subtractive and Fuzzy C-Means clustering. CSIT, 4 (2016) 129-133. doi:10.1007/s40012-016-0090-3.

[4] Y. Guo, X. Wang, P. Xiao et al., An ensemble learning framework for convolutional neural network based on multiple classifiers. Soft Comput, 24 (2020) 3727-3735. doi:10.1007/s00500-019-04141-w.

[5] V. I. Rutsependic, E. S. Yakovleva, O. V. Permyakova, Technological processes state assessment, Processing of solid and layered materials., 1 (2010) URL: https://cyberleninka.ru/ article/n/otsenka-sostoyaniya-tehnologicheskih-protsessov.

[6] Y. Li, G. Yang, H. He et al., A study of large-scale data clustering based on fuzzy clustering. Soft Comput, 20 (2016) 3231-3242. doi:10.1007/s00500-015-1698-1.

[7] Z. Li, S. Wang, W. Yin, Determining optimal granularity level of modular product with hierarchical clustering and modularity assessment. J Braz. Soc. Mech. Sci. Eng., 41 (2019) 342. doi:10.1007/s40430-019-1848-y.

[8] X. Liu, Y. Liu, Q. Xie et al., A potential-based clustering method with hierarchical optimization. World Wide Web, 21 (2018) 1617-1635. doi:10.1007/s11280-017-0509-2.

[9] V. P. Meshalkin, V. I. Bobkov, M. I. Dli, S. V. Khodchenko, Computer simulation of a chemical-energy-technological process for roasting a moving phosphorite pellets multilayer mass, Academy of Science reports. 477, 5 (2017) 559-562.

[10] M. Mojarad, S. Nejatian, H. Parvin, et al., A fuzzy clustering ensemble based on cluster clustering and iterative Fusion of base clusters. Appl Intell, 49 (2019) 2567-2581. doi:10.1007/s10489-018-01397-x.

[11] A. Yu. Puchkov, M. A. Vasilkova, System analysis and assessment of the potential use for apatite-nepheline ores waste from mining and processing enterprises. Mordern high technologies, 7 (2019) 85-89.

[12] S. V. Vimala, K. A. Vivekanandan, Kullback–Leibler divergence-based fuzzy C-means clustering for enhancing the potential of an movie recommendation system. SN Appl. Sci. 1 (2019) 698. doi:10.1007/s42452-019-0708-9.

[13] M. Wang, Z. Zhang, K. Li et al., Research on key technologies of fault diagnosis and early warning for high-end equipment based on intelligent manufacturing and Internet of Things. Int J Adv Manuf Technol, 107 (2020) 1039-1048. doi:10.1007/s00170-019-04289-7.