ATLANTIS PRESS

Advances in Intelligent Systems Research, volume 166

7th Scientific Conference on Information Technologies for Intelligent Decision Making Support (ITIDS 2019)

Data Filtration and Clustering for Purposes of Petroleum Quality Indicators Computation Using Situational Models

Alexander Verevkin
Industrial Process Automation Department
Ufa State Petroleum Technological University
Ufa, Russian Federation
apverevkin@mail.ru

Timur Murtazin
Industrial Process Automation Department
Ufa State Petroleum Technological University
Ufa, Russian Federation
tm.murtazin@mail.ru

Sergei Denisov
Industrial Process Automation Department
Ufa State Petroleum Technological University
Ufa, Russian Federation
s89173422202@yandex.ru

Konstantin Ustyuzhanin
Industrial Process Automation Department
Ufa State Petroleum Technological University
Ufa, Russian Federation
ustyuzhanin.ky@gmail.ru

***Abstract*—Generally, advanced process control systems (APC-systems) base on using technological process models tolerating output quality indicators (OQI) and technical and economic indexes (TEI) forecasting "on the fly". There are many techniques of OQI and TEI evaluation in existence, except one used to work with static information (for example, results of passive experiments) for parametric identification. In addition, control systems store operating parameter's data in its database in a shape of time sequences without any validation or testing for homogeneity. Inhomogeneity of the data drops model quality to the state, when data makes it impossible to develop a situational model without pre-processing and cluster separation, according to which one creates the situational model. This article considers filtration and clustering techniques of APC historical data, including information about technological mode and used OQIs. Described filtering and clustering solutions based on parity models and technological measures cross-correlation techniques; their implication presented on the example of multioutput fractionating tower.**

Keywords—advanced control, homogeneous data situational modeling, clustering, data analysis

I. INTRODUCTION

Problems of advanced control occur from OQI-based in-process control, and TEI-based optimization of operating practices. APC uses models allowing them to estimate and predict feature values of OQI and TEI.

There are known tested methods of developing situational models for a majority of an oil refinery and petrochemical industry processes [1-4].

To find model parameters one use results of static experiments from a real plan [5-7].Pre-processing of the data includes a division of data into sectors (data clusters) which have similar or consequent object behavior; to define static profiles of the unit (modelling object) deriving from time trend analysis and internal mechanisms are the beginning of defining a situational model of OQIs and TEIs [1].The problems that arise are common for a time series forecasting cases in different areas [7-9], but at the same time there are differences for oil refinery facilities.

The specificity of the data obtained from the APCS is their heterogeneity caused by the multi-connectivity and non-stationarity of the object, the presence of transient regimes, the nonlinearity of the links, the high influence of the uncertainty factor due to the presence of uncontrolled disturbances (for example, the composition of raw materials), as well as the uniformity of the measurement intervals of individual process parameters, OQI and TEI, the presence of distorted data and omissions.

Data of operating parameters in modern distributed control systems are archived in the database in the form of a time sequence with periodicity from 1s to several tens of minutes, and, as a rule, without special means of checking for reliability.

In homogeneity of data dramatically reduces the quality of models [5], therefore such data cannot be applied for the development of models without pre-processing and clustering, which inherent characteristics of homogeneity.

There are software solutions known for APC development, for instance: PACE (Platform for Advanced Control and Estimation-Yokogawa Electric Corporation), Profit®Suite (Honeywell), which has tools for data pre-processing. However, the provided functionality of this software mainly delivers an ability to visualize time series, evaluate process statistics and exclude "manually" abnormal, according to the developer, data segments. In addition, the modeling data preparation procedure has no automation, and the basic operations are performed manually. A deeper analysis of the data which provides by these packages involves correction by interpolated values, by the rate of change of the parameter using a Kalman filter. These tools are rarely used due to the peculiarities of application in practical tasks. Problem of automated segmentation (clustering) on the basis of uniformity (homogeneity) has no solution at all. In the same time, data clustering allows to

carry out a situational approach, good for modeling nonlinear objects [1, 11].

Thus, the automation procedure of data pre-processing improves efficiency of APC development. It should be noted that the task of data filtering in the development of calculation models of OQI and TEI is not new; many approaches have been proposed in this field [5-8]. In this paper presents approaches to data preparation and clustering used mainly in the development of models of multioutput fractionating towers, primary oil refining plants, but the ideas can be use in other processes. The considered methods are brought to realization in the form of a software.

II. CONCEPTUAL APPROACHES OF TECHNOLOGICAL DATA SERIES PROCESSING

The volume of training samples, as a rule, should be big enough and can reach tens of thousands of measurements for each parameter .Under these conditions, the need for automated processing of statistical information including filtering incoming APCS data ("unloading" of data), bringing them to a form convenient for standardized automated processing, splitting it into clusters in which it can be considered homogeneous is obvious. Ultimately, one can obtain situational models with sufficient accuracy for each cluster [1, 8].

Regardless of the object and nature of the time series, each time series is usually analyzed for the following components [5]:

1) A series or a long-term tendency in a development of a series: for instance, a change of process characteristics due to catalyst deactivation;

2) Periodical components: some effects in the series dynamics, which are repeated after certain periods; the appearance of such components is typical for processes with recycling [8];

3) Intervention: sharp changes in the nature of the process behavior under the influence of any reasons. Such reasons for oil refining processes are changes in the load, composition of raw materials;

4) Random residue or non-systematic random effect.

Division of the dynamics of the time series into the described components determines the choice of mathematical methods used to identify the corresponding components.

So, to identify an analyze the trend one is using the apparatus of regression analysis and moving averages.

To analyze periodic component one uses smoothing, auto-regression, and spectral analysis.

A special class of models is designed to identify and predict the consequences of interventions. The practical meaning of such models is that in the early stages indirect indicators determines a significant change in the technological mode (for example, when immeasurable changes in the composition of a plant's raw input), which allows the use of predictive control methods.

One use cross-analysis to draw out liaisons between different time series; a list of these relationships in the future will be used in predicting OQIs.

If the problem of time series analysis for periodical, systematic and random revenues has a feasible solution in traditional techniques [5], then the problem of intervention detection requires taking into account the features of the modeling object.

III. DATA CLUSTERING TECHNIQUE

The basis of the methods of intervention detection for technological objects of oil refining lies within presumption that conditionally static modes have an interrelation between technological parameters that cannot sustain radical changes. With this in mind, it is proposed to use two main approaches that will allow data clustering:

1) Analysis of residuals using the reference model (RM);

2) Using pair-wise correlation coefficients between parameters of the sample collection.

The idea of the first method is based on the analysis of residuals between the experimental data and the data obtained from the RM of the interrelations between parameters.

It should be noted that this method is also used in the problems of detection of measurement errors [10].As a result of processing of statistics, in general cases a multiple regression can be used as RM. As a rule, one use only inputs of the reference model which a priori have a relationship with the calculated output parameter of the model, that is, the structure of RM is determined phenomenologically. In a particular case, the dual regression model is used while an independent parameter of the model is selected out of the analysis of the pair correlation coefficients. There can be several such models. For an instance, in a reformer any of the temperatures in the catalyst bed can be taken as an output parameter, and the remaining temperatures and raw material consumption — as input.

In the absence of heuristic knowledge about the object, correlation analysis can be carried out in advance, the results of which for each of the output parameters are selected their subsets of the input parameters.

The first step is to build one or more models Mi, i=1, 2,… based on the initial statistics. Next, according to the i-th model, one forms a set p* off iltered data that satisfy the condition:

$$|p - p^{calc}| < \varepsilon^{max}, \quad (1)$$

Where p is a set of technological variables of the unit; p^{calc} is a set of related variables from the model M(p); ε^{max} is a set of permissible error so f calculation of parameters which value sare calculated by models on training samples. One chooses values of heuristically as a result of iterative execution of the first step. A set of filtered (or clustered) data is formed as an intersection of the sets obtained from separate models for each element p when the number of models is huge.

A stable relationship of several technological variables characterizing the technological state of the unit is typical to for a fairly wide class of technological objects. In this case, RM can be an operator transforming such manifold into a formal attribute or a filtration rule [12]:

$$\Psi\{p\} \xrightarrow{J} \eta, \qquad (2)$$

where $\Psi\{*\}$ is an operator that transforms a set of technological parameters pinto a manifold of attributes; J is the criterion helping to select a common attribute or a rule η from the manifold.

For a n instance, a multioutput fractionating tower may have an RM in a form of a boundary approximation of the temperature profile by the height of the column which provides a formal criterion of attachment of a particular technological mode to a cluster. Fig. 1 shows a division (central solid line) of a set of all the temperature profiles into two clusters: the one above the line and the one below the line. In this case, operator $\Psi\{*\}$ is a temperature profile defined from the following temperatures: overhead temperature (number 1on the horizontal axis), fore cut temperature (number 2),temperature of the second running (number 3), temperature of the third running (number 4) and the bottom fraction temperature (number 5).Since the temperature gradient in the height of the column varies little for stationary modes of multioutput fractionating towers, the criterion J for the formulation of the clusters is the condition under which all the temperatures of the current profile lie in the same area selected by the curve of the boundary temperature profile. Fig. 1 shows the curve of the temperature profile of a certain mode (dashed line), which does not meet this criterion, so the values of the parameters of this mode will be filtered.

The idea of the second technique — filtration by the pair-wise correlation coefficients of the sample parameters is as follows. In the first step, one divides time trends into intervals with a sufficiently large number of samples. For an instance, if the parameter counts are archived every 10 minutes, it will be 144 values per day, and for a week it will be about a thousand, that is, the usual amount of data on the interval is several hundred or even thousands of observations. The criterion for choosing the value of the partition intervals is the low variability of the technological mode on the interval .For most oil refining processes, these intervals typically range from a few days to weeks. Then, onecalculates pair correlation coefficients $r_{\text{pi,pk}}$ between parameters $p_i$andp_k, i≠k for each interval*t*.

On the next step, one calculates mean correlation coefficient M[*t*,$r_{\text{pi,pk}}$] (it also may be an estimation of expectation or a median) for each parameter separately.

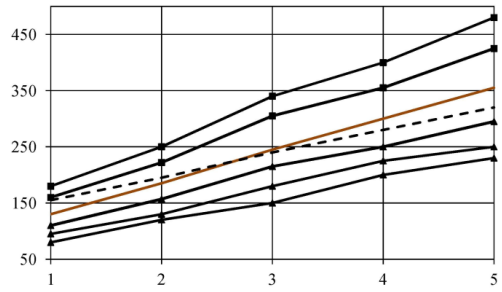

Fig. 1. Temperature profiles of the vacuum tower for different technological modes

A set of clustered data p^1 is a set of intervals which satisfy following:

$$r_{pi,pk}(t)\big| > \big[\mathrm{M}\big(\big|r_{pi,pk}\big|\big) + \Delta\big], \qquad (2)$$

whereΔ is apredetermined deviation of the values of the pair-wise correlation coefficient from its mathematical expectation (which also can be mean or median) which for the problems under consideration is 3...5% of the mathematical expectation $\mathrm{M}(|\,\mathrm{r}_{\text{pi,pk}}\,|)$. One assumes that in a stationary mode a pair-wise correlation coefficient for two random parameters cannot sustain critical changes within a random period.

Thus, according to the characteristic relationship of the parameter values, one obtains some subsets of homogeneous values p_s^*, where *s*is an index of the set where the condition (1) and (or) condition (2) is positive; the subset of these values forms a cluster of technological parameters which can be used in the future to obtain situational models when each subset ***s*** has a separate situational model [1,12].

IV. SOFTWARE PROFILE FOR DATA PROCESSING CASES

The following python modules have been made for automation of the described techniques (see Table I).

The module 1 called “Formation of standard data tables” creates “*.xlsx” files separately for technological measures and laboratory analysis data.

The module 2 performs data checking for exceeding a specified limit values.

The module 3 called “Clustering data by the temperature profile of the tower” creates a temperature chart of the tower profiles. In this case, the temperature values for different modes are recalculated to the total base pressure, which can be taken, for example, as the pressure of the overhead.

Module 5 called “Tower temperature profiles analysis” does further processing. For the module to work one need to provide the boundaries of the temperature profile corresponding to the conditionally static (basic) technological mode which automatically creates a dataset of this mode, data in which is considered homogeneous. It also excludes modes for which the temperature profile crosses more than one cluster.

TABLE I. SOFTWARE MODULES FOR APCS DATA PROCESSING

Block index	Block purpose	Name
0	Project initialization	init
1	Formation of standard data tables	import
2	Limit-wise filtration	specbound
3	Clustering data by the temperature profile of the tower	specclusters-temps
4	Filtering by the value of the residuals of the reference model	specfiltrate-model
5	Tower temperature profiles analysis	profiles
6	Filtration by the cross-correlation tables	specfiltrate-corrs

Module 4 called “Filtering by the value of the residuals of the reference model” creates RMs in a shape of regression for selected set of variables and set the residual value by which the data will be filtered out; homogeneous data will be sampled automatically. This module visualizes forecasts

for the model, errors of the forecasts (residuals), combined plots for comparison of the calculated values of the model with sample data, and the plots of the model residuals. It should be emphasized that the removed values can also be analyzed in order to form a separate group of homogeneous data with its own RM corresponding to another basic technological mode.

Module 6 called "Filtration by the cross-correlation tables" works similar to module 4. Here one uses pair-wise correlation coefficients between technological parameters instead of a reference model. A user should define the length of the period within the specified limits which will be used for correlation coefficient evaluation. In the end, module visualizes changes of the pair-wise correlation in a comparison with their median values for the selected intervals. Module 6 creates a dataset with pair-wise correlation lies within the permissible deviation from the mathematical expectation of the correlation coefficients of the series.

V. Efficiency evaluation for proposed clustering techniques

Let us illustrate the results of data clustering using reference model and cross-correlation tables with multioutput fractionating tower. The initial sample contains the tower operation data for one year with the periodicity of recording the values of 10 min., a total number of values is 52 560. For ease of analysis Fig. 2 also contains the average value of the correlation coefficient for 10 days (solid line).

A reference model in a shape of a function T2=M(T1) has been made using the same data. A dashed line on Fig. 2 represents a change of the data density (in relative units) for 10 days, each one of which satisfy the condition (1) by the limit value of the discrepancy between the initial and calculated RM-data. Assuming, thediscrepancyisN_f/N, where N_f is a number of observations that satisfy the condition (1), and N is a total number of observations in the selected period. One can see that the quality of statistical information begins to decline significantly from about the twentieth numbers of counts. Correlation between the plot values from Fig. 2 that satisfy passing condition is 0.52 in Cheldoke scale which corresponds to a moderate liaison.

The average correlation coefficient ratio and average density data (as %) that passed through filtering and forming a single cluster makes 0,92/83,48% for data from a group before index 20, and 0,78/33,61% for data from index 20 and further.

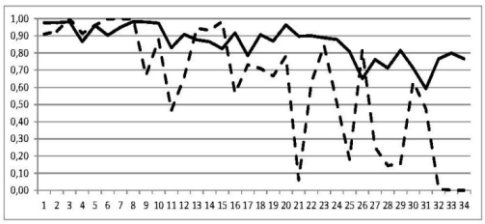

Fig. 2. Changes of the average value of the pair-wise correlation for neighboring temperatures of side cuts of a multioutput fractionating tower (solid line) and density of the sample data satisfying the condition of filtering with a reference model (dashed line)

The average correlation coefficient ratio and average density data (as %) that passed through filtering and forming a single cluster makes 0,92/83,48% for data from a group before index 20, and 0,78/33,61% for data from index 20 and further.

VI. Conclusions

On the basis of the considered example, the following conclusions can be drawn.

1. The considered filtering and clustering techniques allow to provide the formation of samples, which can be used to obtain models for calculating OQIs and TEIs of acceptable quality.
2. Good correlation of the results of the two considered clustering methods confirms their effectiveness, and each of them can be used either separately or together.

References

[1] Verevkin A.P., Kiryushin O.V. Avtomatizaciya texnologicheskix processov i proizvodstv v neftepererabotke i nefteximii. – Ufa: Izd-vo UGNTU. 2005. – 171 p.

[2] Dozorcev V.M., Iczkovich E`.L., Kneller D.V. Usovershenstvovannoe upravlenie texnologicheskimi processami (ARS): 10 let v Rossii // Avtomatizaciya v promy`shlennosti.2013. №1. pp. 12-19.

[3] Fortuna, L., Graziani, S., RizzoA. and Xibilia M.G. Soft sensors for monitoring and control of industrial processes. – London: Springer-Verlag. 2007. – 271 p.

[4] Terrence Blevins, Willy K. Wojsznis, Mark Nixon. Advanced Control Foundation: Tools, Techniques and Applications. ISA. 2012.– 556 p.

[5] Orlova I.V., Polovnikov V.A. E`konomiko-matematicheskie metody` i modeli: komp`yuternoe modelirovanie. – M.: Vuzovskij uchebnik, 2007. — 365 p.

[6] Kadlec P., Gabrys B. and StrandtS. Data-driven soft sensors in the process industry. Computers and Chemical Engineering. 2009. Vol. 33. pp. 795–814.

[7] Huynh N., Mahmassani H. S., Tavana H. Adaptive speed estimation using transfer function models for real-time dynamic traffic assignment operation // Transportation Research Record. (1783). 2002. P. 55-65.

[8] Verevkin A.P, Kalashnik D.V., Xusniyarov M.X. Modelirovanie operativnogo opredeleniya indeksa rasplava dlya upravleniya processom proizvodstva polie`tilena //Bashkirskij ximicheskij zhurnal.Ufa: UGNTU. 2013. T. 20. № 1.

[9] Verevkin A.P., Matveev D.S., Xusniyarov M.X., Chikurov A.V. Postroenie matematicheskoj modeli trubchatoj pechi piroliza dlya celej optimizacii rezhimov i diagnostiki progarov zmeevika // Neftegazovoedelo.2010. T.8. №1.– pp. 70-73.

[10] Ocenivanie sostoyaniya e`lektroe`nergeticheskoj sistemy`: algoritmy` i primery` resheniya linearizovanny`x zadach / Gurina L.A., Zorkal`cev V.I., Kolosok I.N., Korkina E.S., Mokry`j I.V. – Irkutsk: ISE`M SO RAN, 2016. – 37p.

[11] Verevkin A.P., Murtazin T.M. Adaptaciya modelej dlya operativnogo upravleniya texnologicheskimi processami po texniko-e`konomicheskim pokazatelyam // Territoriya neftegaz. 2016. №11. – pp. 14-19.

[12] Axmetov S.A., Ishmiyarov M.X., Verevkin A.P. i dr. Texnologiya, e`konomika i avtomatizaciya processov pererabotki nefti i gaza: uch. posobie. Pod red. Axmetova S.A. – M.: Ximiya. 2005. – 736 p.