HUNGARIAN JOURNAL
OF INDUSTRIAL CHEMISTRY
VESZPRÉM
Vol. 35., pp. 95-99 (2007)

DECISION TREE BASED QUALITATIVE ANALYSIS OF OPERATING REGIMES IN INDUSTRIAL PRODUCTION PROCESSES*

T. VARGA[1], F. SZEIFERT[1], J. RÉTI[2], J. ABONYI[1]

[1]University of Pannonia, Department of Process Engineering, H-8201 Veszprém, P.O.Box 158, HUNGARY
[2]BorsodChem Ltd., H-3700, Kazincbarcika, Bolyai square 6., HUNGARY

The qualitative analysis of complex process systems is an important task at the design of control and process monitoring algorithms. Qualitative models require interpretable description of the operating regimes of the process. This work shows a novel approach to discover and isolate operating regimes of process systems based on process models, time series analysis, and decision tree induction technique. The novelty of this approach is the application of time series segmentation algorithms to detect the homogeneous periods of the operation. Advanced sequence alignment algorithm used in bioinformatics is applied for the calculation of the similarity of the process trends described by qualitative variables. Decision tree induction is applied for the transformation of this hidden knowledge into easily interpretable rule base to represent the operation regions of the process. The whole methodology is applied to detect operating regimes of an industrial fixed bed tube reactor.

Keywords: qualitative analysis, decision tree, operating regime, sequence alignment

Introduction

The improvement of product quality, the need for the reduction of energy and materials waste, and the increased flexibility and complexity of the production systems, process operators require more and more insight into the behaviour of the process. Next to these requirements supporting expert systems should also be able to detect failures, discover the source of each failure, and forecast false operations (e.g. thermal runaway) to prevent from the development of production breakdowns. Data mining of historical process data along advanced process modelling and monitoring algorithms can offer effective solution for this problem.

Quantitative data intensive methods are widely applied because of their statistical nature, but it always claims prior knowledge to analyze the results. Usually prior knowledge is available in the form of qualitative or tendency models of the process. Hence, qualitative analysis of complex process systems is an important task at the design of control and process monitoring algorithms. Qualitative models require the interpretable description not only the historical process data but also the operating regimes of the process.

A common method for decreasing the size of a data set and to get qualitative instead of quantitative information is time series segmentation. Segmentation means finding time intervals where a trajectory of a state variable is homogeneous [1]. Segments can be linear, steady-state or transient, indicative for normal, transient or abnormal operation. Cheung and Stephanopoulos in [2] proposed a second order segmentation method for process trend analysis, the application of episodes with a geometrical representation of triangles. Triangular episodes use the first and second derivatives of a time series on a geometrical basis, hence seven primitive episodes can be achieved as characters. To extract useful feature from time series of the state variables one needs to lower the size and dimension of the data and define a distance measure from a theoretically optimal solution to help operators in their work (i.e. the process trends can be easily compared and evaluated with comparing each sequence of primitive episodes). For sequence comparement, in [3] it was shown as an example that dynamic time warping (DTW) is able to compare DNA sequences if mutation weights (as distances) exist. Going towards this dynamic alignment technique, we applied global pairwise sequence alignment, a well-known technique in bioinformatics developed by [4], to handle not only mutation and substitution but injection and deletion operators in a sequence.

Decision trees are widely used in pattern recognition, machine learning and data mining applications thanks to the interpretable representation of the detected information. This is attractive for a wide range of users who are interested in domain understanding, classification capabilities, or the symbolic rules that may be extracted from the tree and subsequently used in a rule-based decision system. To emphasize how decision trees can be applied to extract useful information from the sequences of process trends, and how they are able to represent the

* An extended version of the lecture presented in 18th ESCAPE Conference, Lyon, France, June 1-4, 2008

operating regimes, an industrial heterocatalytic reactor was analyzed. The results show that the proposed hybrid quantitative - qualitative modelling approach can be effectively used to build a process monitoring and operation support system for industrial reactors.

The paper is organized as follows: in Section 2 the method of qualitative analysis of process trends is briefly introduced, it is followed by the introduction of the developed algorithm for detection of operating regimes. Further sections show an application example and results of the analysis.

A novel qualitative time series analysis algorithm for the detection of operating regimes

Qualitative analysis of process trends

As described in [2], to get from a quantitative to a qualitative representation of a real-valued x(t) function, it has to be reasonable function. It is clear that all the psychical variables in a plant operation are reasonable. It is considered, if we know the value and derivatives of a reasonable function, the state of that function is completely known. The continuous state (CS) over a closed time interval can be defined as a point value, which is a triplet (if x(t) is continuous in t) CS(x, t) ≡ point_value(x, t) = = <x(t), x'(t), x"(t)> Consequently, a continuous trend can be defined as continuous sequence of states. For discrete functions, as an approximation, an underlying continuous function has to be known since the derivatives of single points cannot be performed. These definitions lead to a qualitative description of a state (QS) and trend if x is continuous at t, otherwise it is undefined. QS(x, t) = <[x(t)], [x'(t)], [x"(t)]> where [x(t)], [x'(t)] and [x''(t)] can be {–; 0; +}, depending if they have negative, zero or positive values. Obviously, a qualitative trend of a reasonable variable is given by the continuous sequence of qualitative states. QS(x; t) is called an episode if it is constant for a maximal time interval (the aggregation of time intervals with same QS), and the final definition of a trend of a reasonable function is a sequence of these maximal episodes. An ordered sequence of triangular episodes is the geometric language to describe trends. It is composed of seven primitive notes as {A, B, C, D, E, F, G} illustrated in *Fig. 1*.

Sequence alignment to determine the similarities of the segmented process trends

Sequence alignment is typical expression of bioinformatics, where amino acid or nucleotide sequences have to be compared, how far the evolved new sequences are from the elders, i.e. how old they are, and how many mutation steps were needed to result in the new sequence. The algorithm tries to find the least mutation steps between the elder and offspring sequence applies, that is called minimal evolution. In this paper the most advanced algorithm was used (incorporated in the MATLAB Bioinformatics Toolbox) to determine the minimal sum of transformation weights (which means the similarity of the sequences). For this project therefore we extended the toolbox so it is now not only able to handle amino acid sequences, but the sequences of episodes of time series. For this purpose the similarity of the episodes had to be defined, which becomes the elements of the new transformation matrix.

Visualization and characterization of segments of process trends

Based on these alignment scores (i.e. matching scores), one is able to compare and classify process trends to get a qualitative analysis. The Multidimensional Scaling algorithm (MDS) was applied to visualize the similarity of each process trend to other so the operator can easily check a new trend and in the possession of the necessary a prior knowledge the operator is able to improve the process performance. MDS is a statistical technique for taking the preferences and perceptions of respondents and representing them on a visual grid, called perceptual maps. MDS is a good tool to "rearrange" objects (in our case the process trends) in an efficient manner, so as to arrive at a configuration that best approximates the observed distances (in our case similarities of time series). It actually moves objects around in the space defined by the requested number of dimensions (in our case in three dimension), and checks how well the distances between objects can be reproduced by the new configuration.

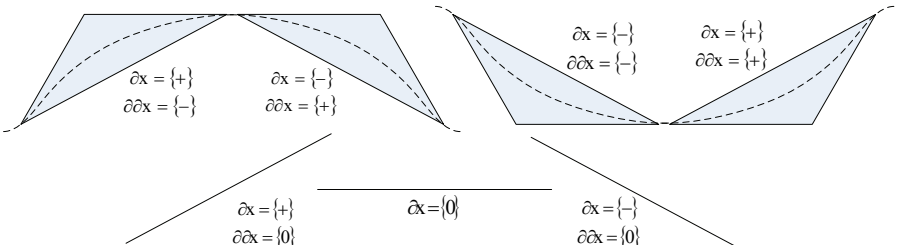


Figure 1: Seven primitive episodes proposed by Cheung and Stephanopoulos

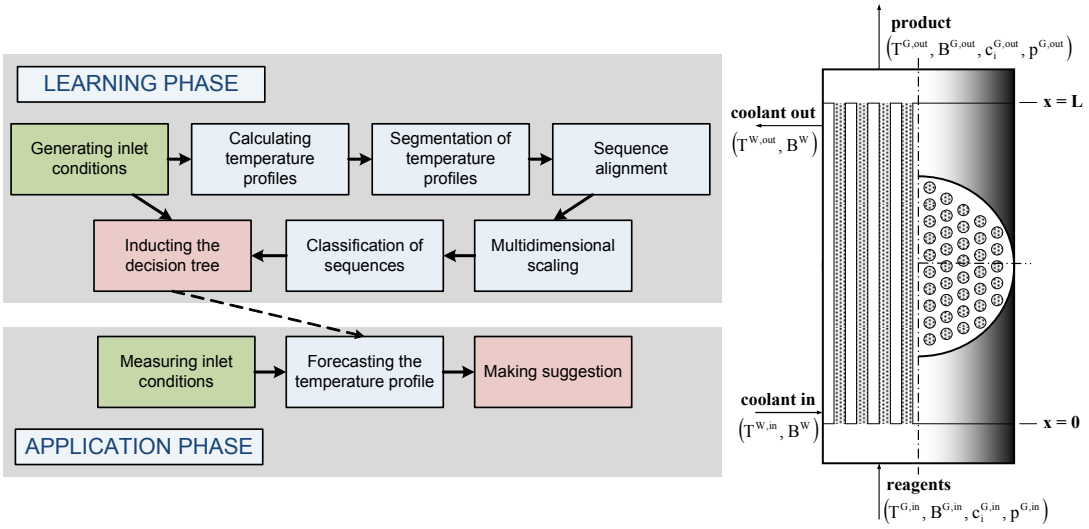


Figure 2: (a) The developed algorithm. (b) Simplified scheme of the studied reactor

Qualitative analysis of operating regimes

The obtained virtual space (shown in *Fig. 2a*) can be easily used to reveal how the process trends are clustered. Since the aim of the proposed methodology is the classification of these process trends and the characterization of the operating regimes of the process variables that affects the shape of these trends, the application of decision trees seems to be a straightforward solution. Binary decision trees consist of two types of nodes: (i) internal nodes having two children, and (ii) terminal nodes without children. Each internal node is associated with a decision function to indicate which node to visit next (e.g. if the temperature is smaller than 235° visit node 25, otherwise visit node 26). Each terminal node represents the output of a given input that leads to this node, i.e. in classification problems each terminal node contains the label of the predicted class (e.g. the 25th terminal node represents reactor runaway). The algorithm has the following basic steps (as shown on Fig. 2a):

- Randomly generating inlet conditions and calculating the temperature profiles;
- Time series segmentation into a sequence of triangular episode primitives;
- Alignment of two episode chains and determining the distance of sequences in a three dimensional virtual space;
- Classifying the time series by a decision tree and based on inlet conditions and the corresponding class of sequence another decision tree is inducted.

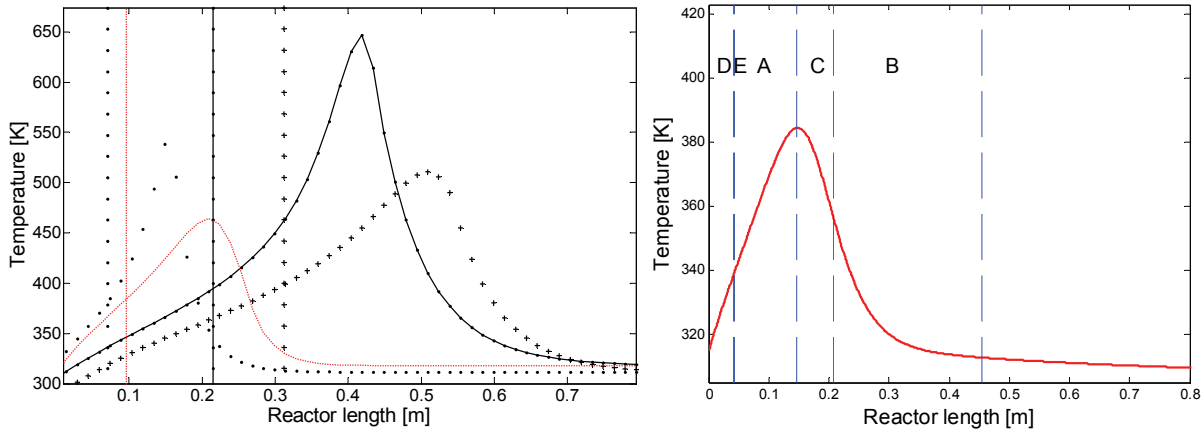
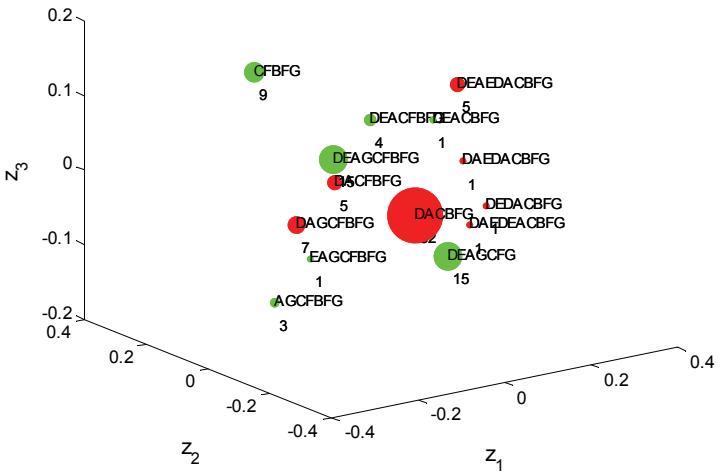


Figure 3: (a) Calculated temperature profiles at some inlet conditions. (b) Example for a segmented process trend. The alphabetic codes of the episodes are also shown.



$T^{W,in}$ <= 289.154 :
| $p^{G,in}$ <= 1.69487 : 4
| $p^{G,in}$ > 1.69487 : 3
$T^{W,in}$ > 289.154 :
| $T^{W,in}$ <= 295.308 :
| | $p^{G,in}$ <= 1.69487 : 2
| | $p^{G,in}$ > 1.69487 : 1
| $T^{W,in}$ > 295.308 :
| | $n^{B,in}$ <= 573.077 : 4
| | $n^{B,in}$ > 573.077 : 1

$T^{W,in}$ – inlet temperature of cooling media [K]
$p^{G,in}$ – inlet pressure of reactor [bar]
$n^{B,in}$ – inlet mass flow rate of component B [mol/s]

Figure 4: (a) Sequences mapped into a three dimensional "virtual" space based on their similarity. (b) The extracted decision tree that represents the operating regimes and able to estimate the class (1-4) of the temperature profiles (shown in *Fig. 5*).

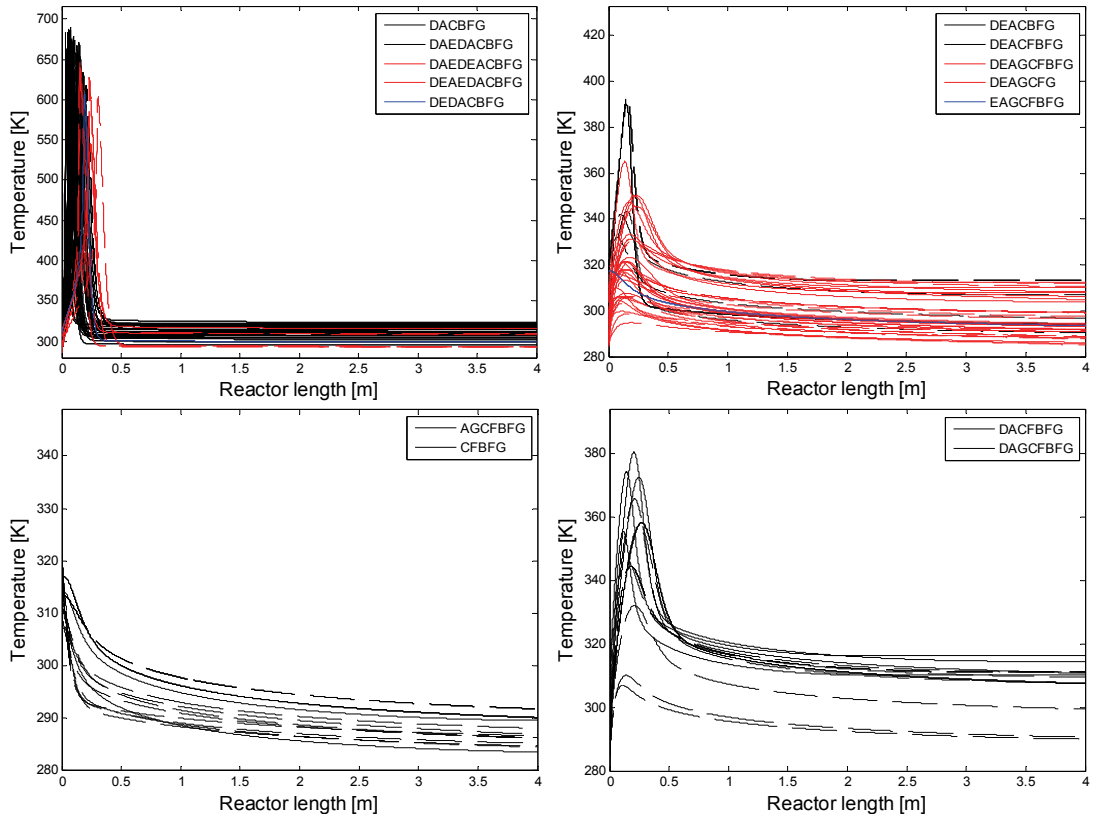


Figure 5: Classified temperature profiles. Four classes of the temperature profiles were detected and the decision tree is able to assign the classes based on the inlet conditions of the reactor.

Application to an industrial fixed bed tube reactor

Process description

To emphasize how decision trees can be applied to extract the relevant information from process trends and how the rules characterize the operating regimes a detailed case study has been worked out based on a sophisticated model of an industrial catalytic fixed bed tube reactor. The studied vertically build up reactor contains a great number of tubes with catalyst (as shown on Fig 2b). Highly exothermic reaction occurs as the reactants rising up the tube pass the fixed bed of catalyst particles and the heat generated by the reaction escapes through the tube walls into the cooling water. Due to the highly exothermic reaction which takes place in the catalyst bed makes the reactor very sensitive for the development of reactor runaway. Reactor runaway means a sudden and considerable change in the process variables. The development of runaway is in very close relationship with the stability of reactor/model. Runaway

has two main important aspects. In one hand runaway forecast has a safety aspect, since it is important for avoiding the damage the constructional material or in the worst case scenario the explosion of reactor; on the other hand it has a technology aspect, since the forecast of the runaway can be used for avoiding the development of hot spots in catalytic bed. The selection of operation conditions is important to avoid the development of reactor runaway and to increase the lifetime of catalyst at same time. The worked out mathematical model has been presented in the previous ESCAPE conference by the authors [5]. The model has been implemented in MATLAB and solved with a low order Runge-Kutta method. The obtained simulator was applied to calculate profiles in case of randomly generated inlet conditions.

Results and discussion

Example for learning samples are plotted on *Fig. 3a* where the vertical lines present where runaway occurs. Such process trends can be easily segmented as it is shown in *Fig. 3b*. It is interesting to note that the algorithm detected that in this case there was no runaway, since it has inserted an E type episode between the D and A episodes, otherwise D-A episodes would mean the change of the sign of the second derivative of the profile that would indicate runaway according to the classical inflection point based runaway detection method. 100 process trends were analyzed. The similarities of the sequences of the episodes generated from these trends were determined by the previously presented sequence alignment. These similarities were used to map the sequences into a three dimensional space to evolve the hidden structure of the trends. A decision tree was inducted to characterize the trends. Four classes were detected. The tree generated based on these new class labels can be seen on *Fig 4b*. On this figure the branches of the tree leading from the root to the leaves should be followed from left to right. In a decision tree the leaves contain the label of the class of the typical temperature profiles. Runaway occurs in case of the first class as shown of Fig 5. Based on this tree the instability regime can be determined ($p^{G,in} > 1.69$ bar and $T^{W,in} > 289$ K). The secondary reduction is directed to the implicit part of the model, only.

Conclusions

This work demonstrated how advanced data mining techniques such as time series segmentation, sequence alignment, and decision tree induction can be used to determine the operating regimes in a heterocatalytic reactor. The results show that the proposed approach is able to distinguish between runaway and non-runaway situations based on a set of linguistic rules extracted from classified process trends obtained by the segmentation of time series generated by the model of the process. The analysis of the extracted rules showed the critical process variables determine the shape of the temperature profiles.

ACKNOWLEDGEMENT

The authors would like to acknowledge the support of Hungarian Research Found (OTKA T049534) and the Cooperative Research Centre (VIKKK) (project III/2). János Abonyi is grateful for the support of the Bolyai Research Fellowship of the Hungarian Academy of Sciences.

REFERENCES

1. KEOGH E., CHU S., HART D., PAZZANI M.: IEEE International Conference on Data Mining, *2001*, 289-296.
2. CHEUNG J. T., STEPHANOPOULOS G.: Computers and Chemical Engineering, *1990*, 14, 495-510.
3. SRINIVASAN R., QIAN M. S.: Chemical Engineering Science, *2006*, 61, 6109-6132.
4. NEEDLEMAN S. B., WUNSCH C. D.: Journal of Molecular Biology, *1970*, 48, 443-453.
5. VARGA T., SZEIFERT F., RÉTI J., ABONYI J.: Computer-Aided Chemical Engineering, *2007*, 24, 751-756.