Automated Classification of Bioprocess Based on Optimum Compromise Whitening and Clustering

doi: 10.15255/CABEQ.2014.2090

J. Kukal,* J. Mareš, J. Náhlík, P. Hrnčiřík, and M. Klimt
Faculty of Chemical Engineering, University of Chemical Technology in Prague, Technicka 5, 166 28 Prague 6, Czech Republic

Original scientific paper
Received: August 6, 2014
Accepted: December 2, 2015

The proposed methodology of technological state classification is based on data smoothing, dimensionality reduction, compromise whitening, and optimum clustering. The novelty of our approach is in the stabile state hypothesis which improves initialization of c-mean algorithm and enables interleaved cross-validation strategy. We also employ the Akaike information criterion to obtain the optimum number of technological states that minimize it, but using as many as possible clusters and components. The general approach is applied to state classification of *Pseudomonas putida* fed-batch cultivation on octanoic acid.

Key words:
fed-batch cultivation, technological state classification, dimensionality reduction, clustering, stabile state hypothesis

Introduction

Biotech unit operations are often characterized by a large number of inputs (operating parameters) and outputs (performance parameters) along with complex correlations amongst them. A typical biotech process starts with the vial of the cell bank, ends with the final product, and has anywhere from 15 to 30 such unit operations in series. The aforementioned parameters can impact process performance and product quality, as well as interact amongst each other. Chemometrics presents one effective approach to gather process understanding from such complex data sets. The increasing use of chemometrics is fuelled by the gradual acceptance of quality by design and process analytical technology among the regulators and the biotech industry, which require enhanced process and product understanding.

From the standpoint of industrial operation, it is important to note that the type of metabolism used by the cultivated microorganism for the processing of substrates has a decisive impact on process performance measured by indicators, like productivity and yield. Therefore, the design of bioprocess control strategies is to be focused not only on the issue of cell environment control, but should ideally also aim at the control of the cell physiology itself. This issue has been addressed by the introduction of a control concept referred to as technological state control by Konstantinov and Yoshida[1]. In contrast to conventional control strategies operating in closed loop in respect to the cell environment, the physiological state control scheme creates a closed loop in respect to the cell state. Consequently, the environment is not a goal but a tool for manipulating cell physiology.

From all the subtasks of a general technological state control strategy, the task of on-line recognition of the physiological state of the cultivated microorganism is of key importance. The classification schemes involved in this task are usually based to some extent on expert knowledge representation and are frequently implemented using various artificial intelligence techniques[1,2].

Review of current state

The chemical industry was early in recognizing and adopting chemometrics as a quick and economical method of extracting real-time information from data and, thus, leading to improved process monitoring and control. Visible spectroscopy, near-infrared (NIR) spectroscopy, mid-infrared (MIR) spectroscopy, nuclear magnetic resonance (NMR) spectroscopy, and Fourier transform infrared (FTIR) spectroscopy are some of the commonly used process analyzers that have been used in the chemical industry. Principal component analysis (PCA), partial least squares (PLS) regression, principal component regression (PCR), canonical variable analysis (CVA), and modified soft independent modeling of the class analogy (SIMCA) are some of the statistical tools that have been used to facilitate analysis and modeling of the abundant data that are provided by the aforementioned process analyzers.

*Corresponding author: jaromir.kukal@vscht.cz

Chemometric tools have been used in the cell culture operations in the last decade. PCA has been used for detection and diagnosis of abnormal process conditions in an industrial fed-batch cell culture process. The model was successfully able to detect abnormal process conditions, which resulted from three known fault types, namely irregular thermal heating, elevated dissolved oxygen values, and large variation in agitation[3]. PLS calibration models of NIR spectra have been utilized for the measurement of glucose, lactate, glutamine, and ammonia in undiluted serum-based cell culture media[4]. Robust, analyte-specific models were generated, and the low values of standard errors of prediction for each analyte demonstrate that the models can be used to (off-line) determine the important nutrient and byproduct content in a serum-based cell culture medium. A novel PLS approach called evolving PLS has been compared with the traditional PLS using data from an industrial fed-batch mammalian cell culture process for prediction of intermediate and final quality variable values[5]. Use of in situ 2D fluorometry in combination with chemometrics has been evaluated for monitoring the concentration of viable cells and the concentration of recombinant proteins in mammalian cell culture[6]. PCA was used to filter the large volumes of redundant spectral data, while PLS correlated the reduced data with the target state variables. Both viable cells density and glycoprotein concentration were accurately estimated, which strongly suggests that the combination of 2D fluorometry with suitable chemometric techniques is a consistent technique for the monitoring of a cell culture medium. Modeling and monitoring of batch processes using neural networks, where the principal component analysis is used for the problem dimensionality reduction is described in Kulkarini *et al*[7]. Data mining and fuzzy modeling described in Ganzle *et al*[8] uses principal component analysis parameters reduction and correlation. Principal component analysis is used for real time monitoring during real experiments, where the enzyme peniciline G acylase was produced, more in Nucci, Cruz and Giordano[9].

Materials and methods

Microorganism and cultivation conditions

The *Pseudomonas putida* KT2442 strain was kindly provided by Dr. M. A. Prieto from CSIB-CSIC. The inocula for fed-batch cultivations were prepared at 30 °C in shaking flasks in a rotary incubator (incubation duration: 1618 h). Composition of the incubation medium per litre: 4.7 g $(NH_4)_2SO_4$, 0.8 g $MgSO_4 \cdot 7H_2O$, 12 g $Na_2HPO_4 \cdot 7H_2O$, 2.7 g KH_2PO_4, 3 g nutrient broth. Productive medium for the fed-batch phase contained per litre: 4.7 g $(NH_4)_2SO_4$, 0.8 g $MgSO_4 \cdot 7H_2O$, 9 g $Na_2HPO_4 \cdot 7H_2O$, 2.03 g KH_2PO_4, 1 g octanoic acid and 10 mL trace element solution (composition per litre: 10 g $FeSO_4 \cdot 7H_2O$, 3 g $CaCl_2$, 2.2 g $ZnSO_4 \cdot H_2O$, 0.5 g $MnSO_4 \cdot 4H_2O$, 0.3 g H_3BO_3, 0.2 g $CoCl_2 \cdot 6H_2O$, 0.15 g $Na_2MoO_4 \cdot 2H_2O$, 0.02 g $NiCl_2 \cdot 6H_2O$, 1 g $CuSO_4 \cdot 5H_2O$).

Experimental setup

The fed-batch cultivations (*Pseudomonas putida* KT2442) were carried out under the following conditions: temperature 30 °C, pH = 7, stirrer speed 900 min^{-1}, air flow 9.5 L min^{-1}. Base (14 % NH_4OH) and acid (17 % H_3PO_4) solutions were added to the cultivation medium to control pH. Following the initial batch phase, octanoic acid was continually supplied with a feeding rate set by the operator. Feeding strategies varied by individual cultivation runs, generally there was a phase of an exponential feeding followed by underfeeding and starvation, respectively.

All cultivations were carried out in a 7-litre laboratory bioreactor (newMBR, Switzerland) at the Bioprocess Control Laboratory at the Department of Computing and Control Engineering of the University of Chemical Technology in Prague (UCT Prague). The bioreactor was equipped with an IMCS 2000 analogue control unit (temperature, pH, stirrer speed, antifoam level, and airflow control), a programmable logic controller (Modicon Compact PC-E984-265, Schneider Electric, France), and the proprietary Biogenes II control system (based on Factory Suite 2000 software package, Wonderware, USA). The dissolved oxygen tension was measured by an oxygen probe (Mettler Toledo); the oxygen and carbon dioxide concentrations in the off-gas were measured by SERVOMEX 1100 and 1440 analysers, respectively. For the substrate supply to the bioreactor, a DP200 peristaltic pump (New Brunswick) was used. Control variables feeding rate, acid, base and antifoam addition were also recorded.

Classification of technological states

The main aim of this paper was automated classification of biotechnological system states. However, the proposed methodology of classification is more general and consists of several steps.

Primary data preprocessing

Let $T_s > 0$ be sampling period, $N \in \mathbf{N}$ be number of observable variables, $\xi_i \in \mathbf{R}^N$ be i^{th} sample of observable variables at time $t = T_s i$, and $\{\boldsymbol{\xi}_i\}_{i=0}^{m-1}$ be time series of complete primary data consisting

of m samples. The only one aim of preprocessing is data smoothing. Let $r \in \mathbf{N}$ be order of symmetric linear smoothing, which is based on formula

$$\boldsymbol{x}_k = \frac{1}{2r+1} \sum_{j=-r}^{+r} \boldsymbol{\xi}_{k+j+r-1}$$

of simple but optimal robust linear smoother[10]. The final result of data preprocessing is the pattern vector $\boldsymbol{x}_k \in \mathbf{R}^N$ which is shifted via r steps. Therefore, the series of patterns is $\{\boldsymbol{x}_k\}_{k=1}^{M}$ where $M = m - 2r$. The time delay of smoothing is $T_s r$ which must be approximately equal to the large time constant of given technological process.

Dimensionality reduction and Data Whitening

Dimensionality reduction is based on the Principal Component Analysis (PCA)[11]. Firstly, we calculate the mean value vector as

$$\boldsymbol{x}_0 = \frac{1}{M} \sum_{k=1}^{M} \boldsymbol{x}_k$$

and covariance matrix as

$$\boldsymbol{C} = \frac{1}{M-1} \sum_{k=1}^{M} (\boldsymbol{x}_k - \boldsymbol{x}_0)(\boldsymbol{x}_k - \boldsymbol{x}_0)^{\mathrm{T}}.$$

The EigenValue Decomposition (EVD) is based on solving of the equation $(\boldsymbol{C} - \lambda I)\boldsymbol{v} = \boldsymbol{0}$ with constraint $\|\boldsymbol{v}\| = 1$, where $\lambda \geq 0$ is the eigenvalue and $\boldsymbol{v} \in \mathbf{R}^N$ is corresponding eigenvector. The solutions of EVD can be ordered as $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_N \geq 0$ and corresponding eigenvectors are $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_N$. The traditional PCA of order $D \in \mathbf{N}, D \leq N$ is based on the formula for the output vector

$$\boldsymbol{p}_k = \boldsymbol{W}^{\mathrm{T}}(\boldsymbol{x}_k - \boldsymbol{x}_0) \in \mathbf{R}^D$$

where

$$\boldsymbol{W} = (\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_D) \in \mathbf{R}^{N \times D}.$$

Data Whitening (DWH)[12] is improved PCA technique which guarantees unit covariance matrix of the resulting vector $\boldsymbol{w}_k = \boldsymbol{L}^{-1/2}\boldsymbol{W}^{\mathrm{T}}(\boldsymbol{x}_k - \boldsymbol{x}_0) \in \mathbf{R}^D$ where

$$\boldsymbol{L} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_D) \in \mathbf{R}^{D \times D}$$

under the supposition of

$$\lambda_D > 0.$$

Being inspired by PCA and DWH, we develop novel technique of Compromise Whitening (CWH) which is based on the formula $\boldsymbol{y}_k = \boldsymbol{L}^{-\alpha/2}\boldsymbol{W}^{\mathrm{T}}(\boldsymbol{x}_k - \boldsymbol{x}_0) \in \mathbf{R}^D$ where $\alpha \in [0,1]$ also enables to realize PCA and DWH as extreme cases. Meanwhile, the PCA does not perform data standardization, the DWH does, but it also increases the noise level of higher order components. Therefore, we suppose that CWH as a compromise between these two disadvantages, can be a more efficient tool than PCA and DWH of the same dimension D in particular cases, related to the cluster analysis as the final step of data processing.

Cluster analysis of time series

Whitened time series $\{\boldsymbol{y}_k\}_{k=1}^{M}$ can be split into H disjoint sub-series which correspond with the technological states for given $H \geq 2$. The resulting sub-series are compact data segments in the ideal case when the technological process stays in the first state for several steps and then sequentially moves to the second and the other states. But the reality of the technological changes is more complex in general. Let $q_k \in \{1,2,\ldots,H\}$ be class (technological state) membership of k^{th} pattern $\boldsymbol{y}_k$ and $\{\boldsymbol{q}_k\}_{k=1}^{M}$ be adequate time series of class memberships. We use a modification of K-means clustering algorithm[13] consisting of three operations: cluster initialization, update of cluster centers, and revision of cluster composition. The traditional K-means randomly initialize the clusters which can cause convergence problems. Our modification is based on the equilateral initialization according to the idea of idealized state sequence. We postulate the initial state as

$$q_k = \left\lceil \frac{kH}{M} \right\rceil$$

which approximately guarantees equal cardinality of clusters. The cluster centers are updated via traditional formula

$$\boldsymbol{t}_j = \frac{\sum_{k:q_k=j} \boldsymbol{y}_k}{\sum_{k:q_k=j} 1}$$

for $j = 1,\ldots,H$. Novel pattern memberships are then recalculated as

$$q_k^{\mathrm{new}} \in \arg\min_{1 \leq j \leq H} \left\| \boldsymbol{y}_k - \boldsymbol{t}_j \right\|$$

for $k = 1,\ldots,M$. When $\{\boldsymbol{q}_k^{\mathrm{new}}\}_{k=1}^{M}$ differs from $\{\boldsymbol{q}_k\}_{k=1}^{M}$, we perform the next update of the cluster centers and adequate membership revision until pattern membership stabilization, as usual.

Let SSQ be residual sum of squares of the final clustering. Classical Akaike Information Criterion[14], divided into M, has the form

$$AIC = \frac{SSQ}{M} + 2H$$

which is useful for the determination of the optimum cluster number, i.e. the number of technological states in our case. The cluster number which minimizes *AIC* is declared to be optimal. The choice of *AIC* is motivated by its cross-validation properties[14]. The other criteria of clustering quality are not recommended due to their poor relation to the cross-validation process.

Optimum parameter setting

The proposed classification method of technological states has the optional parameters as the pattern vector length N, the order of smoothing r, the number of whitened components D, CWH index α, and the number of states H. But the parameters N and r are strongly connected with the biotechnological process, its monitoring, and time constants. Therefore, they are not subjects of optimization with three aims:

– optimal clustering with minimal value of *AIC* as declared above,

– maximal number of clusters H for detailed analysis of the technological states

– maximal number of CWH components D for saving of the data dimensionality.

The optimization task seems to be a multi-criteria one, but its solution would be sensitive to the choice of *AIC*, H, and D weights in the compromise objective function. We prefer the three-stage hierarchical optimization process driven by three optimization aims. To avoid multiplicities, we define the left minimum (left maximum) as the lower possible value of the scalar variable, which guarantees the minimal (maximal) value of the given function.

The optimum values of H, D, and α can be obtained for the given data set and fixed N, r as follows:

Optimum H*(D,α) is obtained by the minimization of AIC(H) as the left minimum for constant D, α.

Optimum D*(α) is obtained by the maximization of H*(D) as the left maximum for constant α.

Optimum α* is obtained by the maximization of D*(α) as the left maximum.

Therefore, α*, D*(α*), and H*(D*(α*),α*) are the obtained optimal values, which form the main result of our novel approach to the automated state classification of biotechnological process from measured time series.

Interleaved Cross-validation Strategy

We split the original pattern series $\{\boldsymbol{x}_k\}_{k=1}^{M}$ into training series $\{\boldsymbol{x}_{2k-1}\}_{k=1}^{Q}$ of odd patterns denoted as TS, and verification series $\{\boldsymbol{x}_{2k}\}_{k=1}^{Q-1}$ of even patterns denoted as VS, where $Q = \lfloor (M+1)/2 \rfloor$. We then perform CWH and cluster analysis only on the TS yielding odd pattern membership series $\{q_k^{\text{odd}}\}_{k=1}^{Q}$ together with CWH and clustering parameters, which are directly used for processing of even patterns from VS. The resulting even pattern membership series is $\{q_k^{\text{even}}\}_{k=1}^{Q-1}$.

The novel Interleaved Cross-validation Strategy (ICVS) is based on the hypothesis of stabile technological states, which rarely changes to another state in the meaning of small change probability. When $q_k^{\text{odd}} = q_{k+1}^{\text{odd}}$, the neighbor states of TS are the same, and according to the hypothesis of stabile technological states, the adequate interleaved state of VS must be also the same, i.e. $q_k^{\text{even}} = q_k^{\text{odd}} = q_{k+1}^{\text{odd}}$. Let S be the number of stabile cases when $q_k^{\text{odd}} = q_{k+1}^{\text{odd}}$, and E be the number of misclassified cases when $q_k^{\text{odd}} = q_{k+1}^{\text{odd}} \neq q_k^{\text{even}}$. Therefore, the classification error can be defined as $ERR = E/T$ as a good post-optimization indicator of the parameter setting quality in the case of a single technological experiment.

Final Cross-validation Strategy

The final cross-validation methodology is based on general assumptions[15]. Having data from two independent technological experiments, we can easily perform traditional cross-validation, which is based on a separate parameter setting from both experiments. We define inner classification as ICVS using the data from the given experiment and optimum values H*, D*, α*. The outer classification is defined as ICVS using the data from the other experiment, which vectors $\boldsymbol{x}_0$, $\boldsymbol{t}_1$, ..., $\boldsymbol{t}_H$ and matrices $\boldsymbol{L}$, $\boldsymbol{W}$ are applied to the data from the given experiment to obtain a potentially different state classification.

The final cross-validation strategy consists of:

(i) Optimization of H, D, and α in case of the first experiment.

(ii) Evaluation of CWH and cluster analysis vectors and matrices in the case of the first experiment.

(iii) Inner classification of technological states according to (ii) in the case of the first experiment.

(iv) Evaluation of CWH and cluster analysis vectors and matrices in the case of the second experiment.

(v) Inner classification of technological states according to (iv) in the case of the second experiment.

(vi) Outer classification of technological states according to (ii) in the case of the second experiment.

(vii) Comparison of inner (v) and outer (vi) classification in $D \times D$ contingency table.

(viii) Evaluation of the classification error from the contingency table.

Results

The general methodology was directly applied to the real data from two experiments with the given biotechnological process. Sampling period was 60 sec. The order of smoothing r was set to 10, which corresponds with the time constant of the given biotechnological process. The first data set consists of 2096 points. The influence of CWH parameter α on parameter D is demonstrated in Table 1. The dimensionality D of CWH varies between 1 (poor) and 4 (rich) with maximum for $\alpha^* = 0.54$, which is recommended as the optimum value for the best dimensionality $D^* = 4$, and consequently $H^* = 5$ as the optimum number of technological states. The traditional cases of PCA ($\alpha = 0$) and DWH ($\alpha = 1$) obtain poor dimensionality $D = 1$. Therefore, the novel CWH technique obtains better results than the referential methods on the given task. Time dependency of these four components in the first experiment is depicted in Fig. 1 as traditional PCA. Adequate classification into the five technological states is demonstrated in Fig. 2. As seen in Table 1, the classification error *ERR* obtained via ICVS varies between 2.96 % and 19.76 %, but this criterion was not a subject of minimization. The optimum CWH has interleaved cross-validation error *ERR* = 19.08 % in the first experiment.

Table 1 – *Optimum compromise whitening for the first experiment*

α	D	H	*AIC*	*err* [%]
0.00	1	8	107.4890	14.14
0.10	1	8	52.5981	14.14
0.20	1	8	30.6401	14.14
0.30	2	8	24.0317	14.25
0.40	2	6	17.8663	19.76
0.42	2	6	17.0290	19.24
0.44	3	6	16.9660	14.56
0.46	3	6	16.3167	14.40
0.48	3	5	15.5427	15.86
0.50	3	5	14.8464	17.42
0.52	3	5	14.2488	16.69
0.54	4	5	13.9929	19.08
0.56	3	4	12.9685	19.40
0.58	3	4	12.3289	18.62
0.60	1	3	8.4479	5.20
0.70	3	4	9.9722	21.48
0.80	2	3	6.9752	11.44
0.90	1	2	4.3509	2.96
1.00	1	2	4.1404	2.96

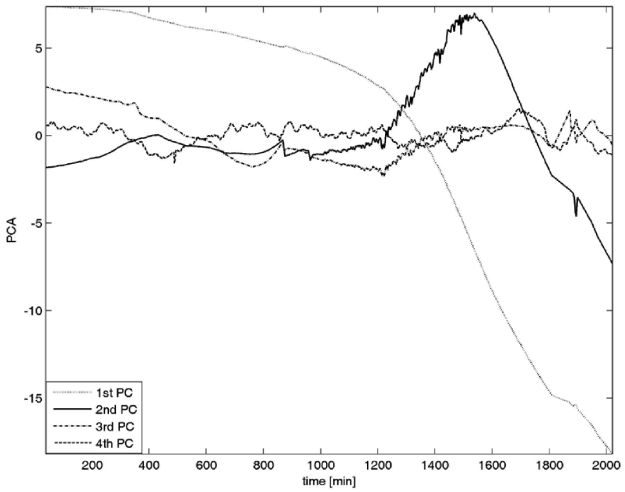

Fig. 1 – *Optimum components for the first experiment*

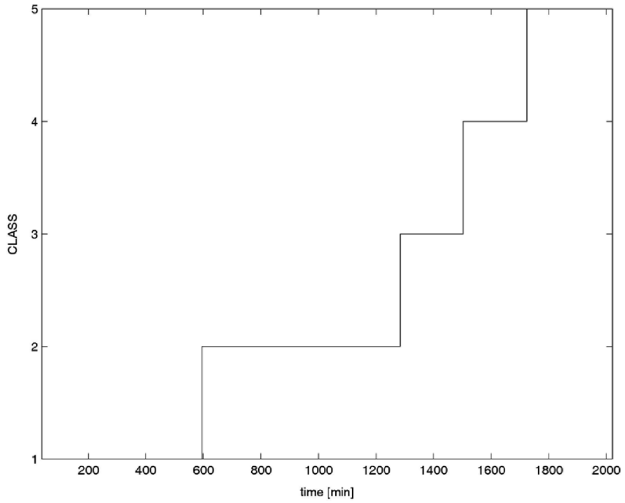

Fig. 2 – *Optimum clustering for the first experiment*

Therefore, the technological process was classified into five states. The first state describes the beginning of the experiment, i.e., the batch phase and the beginning of fed-batch phase. The second and third states describe the production phase of the experiment where the optimal feeding presents 90 % of the span, of which underfeeding is represented by 10 %. The beginning of the fourth state comes with the concentration of dissolved oxygen at 17 % and with the evident decreasing in the dissolved oxygen difference. Finally, the fifth state is

represented by the dissolved oxygen and the difference of dissolved oxygen equal to zero.

In the case of the second experiment, the data set consists of 1923 points. The optimum classification parameters $\alpha^* = 0.54$, $D^* = 4$, and $H^* = 5$ from the first experiment were used for the inner and outer classification of data points in the second experiment. Meanwhile, the inner classification formed new CWH components and cluster centers from the second data set, the outer classification used the old CWH components and cluster centers from the first experiment to data from the second one. The results of the final cross-validation are collected in Table 2 in the form of a contingency table. There are only 416 missclassified data points, meaning a cross-validation error of 21.63 % using independent data sets. This quality measure is very similar to interleaved cross-validation error of 19.08 %. Therefore, the novel methodology of optimum CWH with cluster analysis is able to generalize the relationships from one experiment to be applicable to another one.

Table 2 – *Final cross-validation on the second experiment*

		Outer states			
Inner states	437	6	0	0	0
	0	550	0	0	0
	0	130	191	0	0
	0	25	185	136	0
	0	0	0	70	193

Conclusions

The general methodology based on compromise data whitening, cluster analysis, interleaved cross-validation, and cross-validation on the second data set, was applied to two data sets from the biotechnological process of *Pseudomonas putida* fed-batch cultivation on octanoic acid. The optimum compromise data whitening outperformed both PCA and traditional data whitening in the given cases. The resulting fourth component system localized five technological states with an interleaved cross-validation error of 19.08 % and final cross-validation error of 21.63 % on the second data set. The resulting states have biotechnological interpretation. The proposed methodology of biotechnological state analysis can be used in similar cases.

List of abbreviations

CVA – Canonical Variable Analysis
CWH – Compromise Whitening
DWH – Data Whitening
EVD – EigenValue Decomposition
FTIR – Fourier Transform Infrared spectroscopy
ICVS – Interleaved Cross-validation Strategy
MIR – Mid-infrared spectroscopy
NIR – Near-infrared spectroscopy
NMR – Nuclear Magnetic Resonance spectroscopy
PC – Principal Component
PCA – Principal Component Analysis
PCR – Principal Component Regression
PLS – Partial Least Squares regression
SIMCA – Soft Independent Modeling of the Class Analogy
TS – Training Series
VS – Verification Series

List of symbols

AIC – Akaike Information Criterion
D – dimension of PCA, DWH, and CWH
E – number of misclassified cases
ERR – classification error
H – number of technological states
i – index of primary data samples
j – index of filtering
k – index of pattern
M – number of valid patterns
m – number of samples
N – number of observable variables
Q – number of training patterns
r – order of smoothing
S – number of stabile cases
SSQ – residual sum of squares
t – time
T_s – sampling period
α – CWH parameter
λ – eigenvalue
$\boldsymbol{C}$ – covariance matrix
$\boldsymbol{I}$ – identity matrix
$\boldsymbol{p}$ – PCA vector
$\boldsymbol{q}$ – class membership vector
$\boldsymbol{q}^{new}$ – improved class membership vector
$\boldsymbol{t}$ – cluster center vector
$\boldsymbol{v}$ – eigenvector
$\boldsymbol{W}$ – PCA matrix
$\boldsymbol{w}$ – DWH vector
$\boldsymbol{x}_0$ – mean value vector
$\boldsymbol{x}$ – pattern vector
$\boldsymbol{y}$ – CWH vector
$\boldsymbol{\xi}$ – sample vector of observable variable
^{T} – transposition
* – optimal value
$\mathbf{N}$ – set of natural numbers
$\mathbf{R}$ – set of real numbers

References

1. *Konstantinov, K. B., Yoshida, T.*, Physiological state control of fermentation processes, Biotechnol. Bioeng. **33** (1989) 1145. doi: http://dx.doi.org/10.1002/bit.260330910
2. *Konstantinov, K. B.*, Monitoring and control of the physiological state of cell cultures, Biotechnol. Bioeng. **52** (1996) 271. doi: http://dx.doi.org/10.1002/bit.260520203
3. *Gunther, J. C., Conner, J. S., Seborg, D. E.*, Fault detection and diagnosis in an industrial fed-batch cell culture process, Biotechnol. Prog. **23** (2007) 851. doi: http://dx.doi.org/10.1002/bp070063m
4. *Rhiel, M., Cohen, M. B., Murhammer, D. W., Arnold, M. A.*, Non destructive near-infrared spectroscopic measurement of multiple analytes in undiluted samples of serum-based cell culture media, Biotechnol. Bioeng. **77** (2002) 73. doi: http://dx.doi.org/10.1002/bit.10093
5. *Gunther, J. C., Conner, J. S., Seborg, D. E.*, Process monitoring and quality variable prediction utilizing PLS in industrial fed-batch cell culture, J. Process Control **19** (2009) 914. doi: http://dx.doi.org/10.1016/j.jprocont.2008.11.007
6. *Teixeira, A. P., Portugal, C. A. M., Carinhas, N., Dias, J. M. L., Crespo, J. P., Alves, P. M., Carrondo, M. J. T., Oliveira, R.*, In situ 2D fluorometry and chemometric monitoring of mammalian cell cultures, Biotechnol. Bioeng. **102** (2009) 1098. doi: http://dx.doi.org/10.1002/bit.22125
7. *Kulkarini, S. G., Chaudhary, A. K., Nandi, S., Tambe, S. S., Kulkarini, B. D.*, Modeling and monitoring of batch processes using principal component analysis assisted generalized regression neural network, Biochem. Eng. J. **18** (2004) 193. doi: http://dx.doi.org/10.1016/j.bej.2003.08.009
8. *Ganzle, M. G., Kilimann, K. V., Hartmann, C., Vogel, R., Delgado, A.*, Data mining and fuzzy modeling of high pressure inactivation pathways of Lactococcus Lactis, Innovative Food Sci. Emerging Technol. **8** (2007) 461. doi: http://dx.doi.org/10.1016/j.ifset.2007.04.003
9. *Nucci, E. R., Cruz, A. J. G., Giordano, R. C.,* Monitoring bioreactors using principal component analysis: production of penicillin G acylase as a case study, Bioprocess Biosyst. Eng. **33** (2010) 557. doi: http://dx.doi.org/10.1007/s00449-009-0377-y
10. *Maronna, R. A., Martin, R. D., Yohai, V. J.*, Robust Statistics: Theory and Methods, John Wiley & Sons, New York, 2006, pp 107-109. doi: http://dx.doi.org/10.1002/0470010940
11. *Jolliffe, I.*, Principal component analysis, Springer, Heildberg, 2002, pp 37-45.
12. *Eldar, Y., Oppenheim, A. V.*, MMSE whitening and subspace Whitening, IEEE Trans. Inform. Theory **7** (2003) 1746. doi: http://dx.doi.org/10.1109/tit.2003.813507
13. *Hartigan, J., Wang, M. A.*, K-means clustering algorithm, Appl. Stat. **28** (1979) 100. doi: http://dx.doi.org/10.2307/2346830
14. *Fang, Y.*, Asymptotic equivalence between cross-validations and Akaike information criteria in mixed-effects models, J. Data Sci. **9** (2011) 15.
15. *Geisser, S.*, Predictive Inference, Chapman and Hall, New York, 1993, 178-196. doi: http://dx.doi.org/10.1007/978-1-4899-4467-2