ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2023, Vol. 84, No. 7, pp. 891–899.

AUTOMATION IN INDUSTRY

Soft Sensors Based on Digital Models

A. A. Chereshko

Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia
e-mail: chereshko@phystech.edu

Received May 16, 2022
Revised May 20, 2022
Accepted June 17, 2022

Abstract—The article proposes a method for creating soft sensors using identification models obtained by associative search algorithm. The method consists in constructing an approximating hypersurface of the space of input vectors and their corresponding one-dimensional outputs at each time instant. Case studies are presented and the advantages of the author's method over traditional approaches are evaluated are revealed.

Keywords: soft sensor, model predictive control (MPC), identification, machine learning, clustering methods, associative search

DOI: 10.25728/arcRAS.2023.93.50.001

1. INTRODUCTION

Product quality is the most important parameter in the process optimization. However, quality metrics are difficult to measure in real time because product properties can change over time due to many factors. At modern manufacturing enterprises, various methods and means of physical and chemical physical and chemical analysis of product samples are widely used. Nevertheless, the use of embedded analysis systems is not always justified due to their inertia and high cost. Thats why soft sensors have gained popularity as a low-cost alternative to complex analytical systems. Soft sensors make predictions of product quality in MPC [1, 2]. They allow to control product properties in real time with an acceptable accuracy at relatively modest deployment and maintenance costs. Soft sensors can control physical and chemical properties, that cannot be controlled by conventional analyzers. They can also be used to monitor product quality, where the use of inline analyzers is economically feasible or technically impossible [1]. Soft sensors based on linear regression with automatic fit of free term according to laboratory control data have performed well in practice, but for some nonlinear objects they give inadequate prediction. The article offers a new approach to predicting quality metrics for a wide class of processes described by non-linear models. Models based on associative search algorithms use the knowledge base of processes to build, at each time step, the best model by the least squares method (LSM). A case studies was conducted based on ore grinding process data, the advantages of the proposed method in comparison with the traditional ones are shown.

2. CLASSICAL METHODS FOR SOFT SENSORS CONSTRUCTING

Model Predictive Control (MPC) allows direct control of the products quality, that is evaluated by soft sensors. Usually, soft sensors in MPC systems are realized in the form of simple regression models:

$$Y = \sum_{1}^{N} b_i x_i + b_0, \tag{1}$$

where Y is product quality, x_i are the input variables, b_i are the coefficients for each input variable, b_0 is the free term of linear regression.

Model (1) is developed on the basis of historical process data. Factory data often contain many outliers, which can lead to poor modeling. It is easy to show that a least squares method (LSM) model is highly distorted by even a few noticeable outliers. There is a so-called weighted least squares (WLS) method, an improved method of constructing a linear regression, when individual outliers in the data do not distort the constructed model so much.

For soft sensors of the form (1) we often use the algorithm for constructing a free linear regression term on the basis of the laboratory control data. For this purpose, the coefficient b_0 is recalculated as follows:

$$b_{0new} = b_0 + k(Y_{lab} - Y_{model}), \tag{2}$$

where Y_{lab} is the laboratory quality score, Y_{model} is the quality score calculated by a model, b_{0new} is the new value of free term b_0, k is the weight coefficient of accounting for incoming laboratory control data. Soft sensors based on linear regression with automatic fit of free term b_0 proved to be a good solution to various practical problems. For nonlinear processes, such models may not be satisfactory. In this case, it seems reasonable to use an approach to predicting quality metrics based on data mining of the object functioning and formation of an inductive knowledge base for it.

3. SOFT SENSORS BASED ON INTELLIGENT ANALYSIS OF TECHNOLOGICAL DATA

Models based on associative search algorithms use the inductive knowledge base of process to build, at each time step, the best model by the LSM. The concept of inductive knowledge — the regularities extracted from the data of the object functioning was introduced by V.N. Vapnik [3]. Such models are formed at each time step by the system identifier on the basis of the analysis of the information on the process dynamics accumulated by current time instant. This information makes possible to replenish inductive knowledge base and additionally train the system.

The model, generated by the soft sensors at a certain time instant, replenishes at each time step the appropriate "library of models" in the knowledge base. This digital model is fully characterized by a set of values of the following attributes: inputs, output, coefficients. Further, these models can be used in the traditional MPC scheme.

The models can be derived using well-known identification methods, such as LSM. Prediction with the associative search method [4] for solving control problems has high accuracy of identification model for a wide class of nonlinear and nonstationary objects [5, 6]. In addition, pre-training (clustering) in real time provides the algorithm with a speed gain, which may be important for a certain class of process control problems.

The identification model is fully described by the sets of inputs and corresponding outputs of the system, that are stored in the archive, and in this aspect, it can be considered as a digital model. The combination of statistical data sets (feature values) gives a "digital portrait" of the process dynamics.

4. DESCRIPTION OF THE ASSOCIATIVE SEARCH ALGORITHM

The process of inductive knowledge is reduced to the restoration (associative search) of knowledge by its fragment [7, 8]. In this case knowledge can be interpreted as an associative connection between images. As an image, we will use the vectors of inputs, that is, input variables.

The criterion of closeness of images can be formulated in different ways. In the most general case, it can be represented as a logical function, that is, a predicate. In the particular case, when the sets of features are vectors in n-dimensional space, the proximity criterion may be a distance in this space.

Associative search can be performed as a process of either restoring an image according to partially specified features (or restoring a fragment of knowledge under conditions of incomplete information. This process is simulated in various models of associative memory), or searching for other images related associatively to the given one, but representing other time instants.

Various schemes of associative search are known [9]. Thus, in frame-based systems, the search task is implemented in the form of matching frames. In semantic networks the search is performed by matching fragments of the network and the graph-query.

The approach based on the method of verbal analysis of decisions proved effective for solving discrete multi-criteria selection problems [10]. This approach involves decomposition of the description of objects by many criteria into their partial descriptions of smaller dimensions, which are offered to the decision maker for comparison (under the assumption of pairwise equal evaluations by the criteria that are not included in such descriptions).

There is a well-known model that describes the process of associative thinking as a sequential recall based on the application of associations — pairs of images characterized by their own set of features. This model appears to be an intermediate stage between neural network models and logical models used in classical artificial intelligence systems.

The statement of the Y signal modeling problem is a result of operator's influence (in general case, nonlinear) on vector signal $x_1, x_2, \ldots, x_n$ in discrete time [11]. At any chosen time instant a new linear model in the neighborhood of the operating point is created (instead of approximation of real signal in time).

The linear dynamic model has the following form:

$$Y = \sum_{i=1}^{m} a_i y_{N-i} + \sum_{j=1}^{n} \sum_{s=1}^{S} b_{js} x_{N-j,s},\ \forall j \in 1, \ldots, N, \tag{3}$$

where y_N is the output prediction at the moment N, x_N is the input vector, m is the memory depth for output, n is the memory depth for input, S is the length of the vector of inputs.

This model is not a classical regression model: not the entire dynamic chronological "tail" is selected, but only certain inputs, in accordance with a given criterion. The coefficients at other inputs are assumed to be zero.

To construct a virtual model corresponding to a certain point in time, the input vectors, close to the current one according to the criterion, are selected from the archive. Then, based on the classical (nonrecurrent) LSM, the value of the output at the next point in time is calculated.

Described algorithm does not create a single approximating model of the real process for all moments of time, but builds a new model for each fixed moment t (Fig. 1). In this case each point of the global nonlinear regression surface is formed as a result of using "local" linear models.

In contrast to classical regression models, for each fixed point in time, input vectors are selected from the archive that are close to the current input vector according to a certain criterion rather than in chronological order by steps backwards.

The selection criteria are called associative search criteria or associative pulses. Thus, in (3) n represents the number of vectors from the archive (from time moment 1 to moment N) selected by the associative search criterion. At each time interval $\{N-1, N\}$ a certain set of n vectors is chosen, $1 \leqslant n \leqslant N$.

The criterion for selecting input vectors from the archive to build a virtual model at a given time by the current state of the object can be as follows. Let the distance in R^p between the points

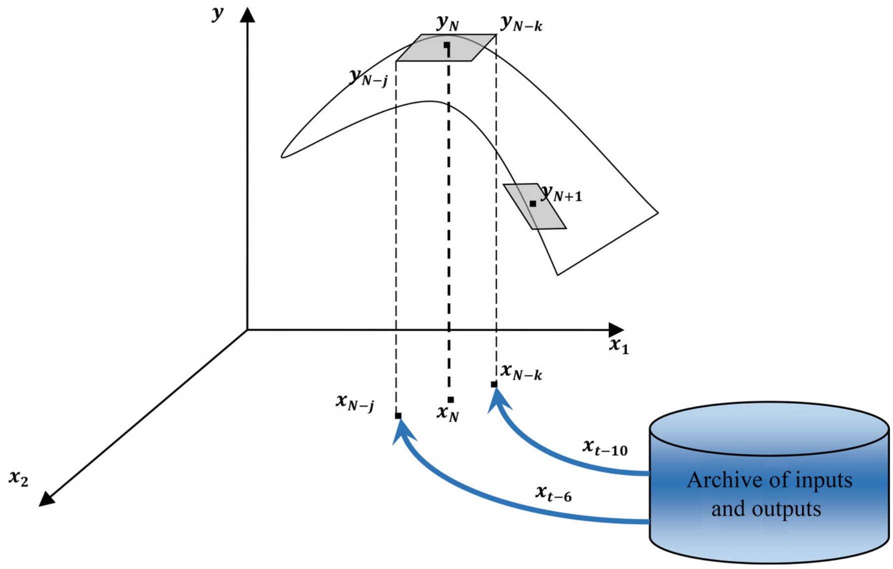

Fig. 1. The approximating hypersurface of the space of input vectors and one-dimensional output.

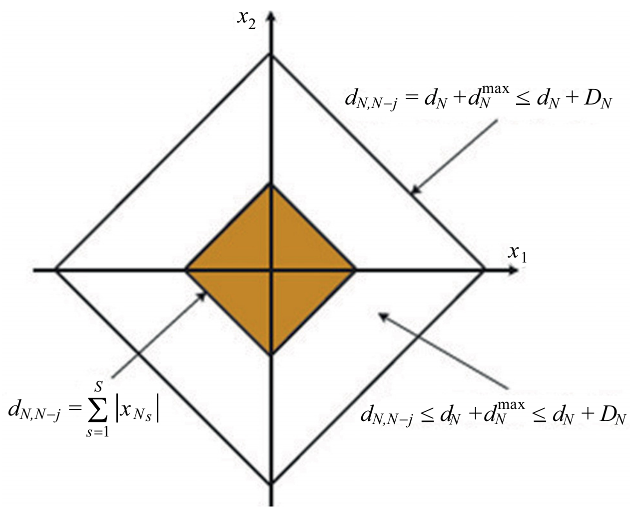

Fig. 2. The area of input vectors acceptably close to the input vector.

of the input space of dimension P be like:

$$d_{t,t-j} = \sum_{p=1}^{P} |x_{tp} - x_{t-j,p}|\,, \quad j = 1, \ldots, s, \tag{4}$$

where $s < t$, x_{tp} are the components of input vector on current time moment t. Suppose that for the current input vector x_t:

$$\sum_{p=1}^{P} |x_{tp}| = d_t. \tag{5}$$

To construct the approximating hypersurface x_t, let us choose such vectors x_{t-j}, $j = 1, \ldots, s$, from the archive of historical data, that for a given D_t the following condition is satisfied (Fig. 2):

$$d_{t,t-j} \leqslant d_t + \sum_{p=1}^{P} |x_{t-j,p}| \leqslant d_t + D_t, \quad j = 1, \ldots, s. \tag{6}$$

The preliminary value of D_t is determined on the basis of the knowledge of the process. If the chosen domain does not contain enough inputs for the application of LSM, i.e., the corresponding system of linear algebraic equations has no solution, then the given point criterion can be relaxed by increasing the threshold value D_t.

In order to increase the speed of the identification algorithm (both at the stage of training and at the subsequent operation of the object), one of the methods of intelligent data analysis, that is, clustering (dynamic classification, automatic grouping of data, "learning without a teacher"), is used. There are a lot of clustering methods, namely hierarchical algorithms, k-means algorithm, minimum covering tree algorithm, nearest neighbor method and others. All of them determine to which region, into which this space is divided, point in a multidimensional space belongs to.

As a result, at each time instant each investigated point in the multidimensional space can be assigned to some group and gain special cluster label. In the associative search problem for selection of input vectors close to the current one, the cluster label is defined according to the criterion of associative selection of input vectors from the archive. To build soft sensors, vectors are selected within the corresponding cluster.

For a dynamic linear model, the following algorithm is used to determine the unknown coefficients. The model is formed as:

$$y_N = \sum_{i=1}^{Q} a_i \hat{x}_i, \tag{7}$$

where $\hat{x}_i = (\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_r)$, $r = m + nS$, $\hat{x}$ is extended input vector for which:

$$\{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_m\} = \{y_{N-1}, y_{N-2}, \ldots, y_{N-m}\}; \tag{8}$$

$$\{\hat{x}_{m+1}, \hat{x}_{m+2}, \ldots, \hat{x}_{m+nS}\} = \{y_{N-1,1}, y_{N-1,2}, \ldots, y_{N-1,S}, \ldots, y_{N-n,S}\}; \tag{9}$$

α is an extended vector of input coefficients:

$$\{\alpha_1, \alpha_2, \ldots, \alpha_m\} = \{\alpha_1, \alpha_2, \ldots, \alpha_m\}; \tag{10}$$

$$\{\alpha_{m+1}, \alpha_{m+2}, \ldots, \alpha_{m+nS}\} = \left\{b_{1,1}, b_{1,2}, \ldots, b_{1,S}, \ldots, b_{n,S}\right\} \tag{11}$$

To build the model (7), the input vectors, which are close to the current one according to chosen criterion, are selected from the process data archive. After vector selection, a matrix of extended input vectors is compiled:

$$\hat{X} = \begin{pmatrix} \hat{x}_1^1 & \cdots & \hat{x}_r^1 \\ \vdots & \ddots & \vdots \\ \hat{x}_1^P & \cdots & \hat{x}_r^P \end{pmatrix}; \quad P \gg r. \tag{12}$$

To find the coefficients α_i you need to solve a system of linear equations:

$$\hat{X}\alpha = \hat{y}, \tag{13}$$

where $\hat{y}$ is system output at next time step for the selected extended vectors of process inputs.

When solving the system of linear Eqs. (13), assuming that $\operatorname{rank} \hat{X} = r$, LSM can be applied to find the estimate $\hat{\alpha}$:

$$\left(\widehat{y} - \hat{X}\hat{\alpha}\right)^T \left(\widehat{y} - \hat{X}\hat{\alpha}\right) = \min_{\alpha} \left(\widehat{y} - \hat{X}\alpha\right)^T \left(\widehat{y} - \hat{X}\alpha\right). \tag{14}$$

Assuming that $\hat{X}$ it is a matrix of full rank:

$$\hat{\alpha} = (\hat{X}^T \hat{X})^{-1} \hat{X}^T \hat{y}. \tag{15}$$

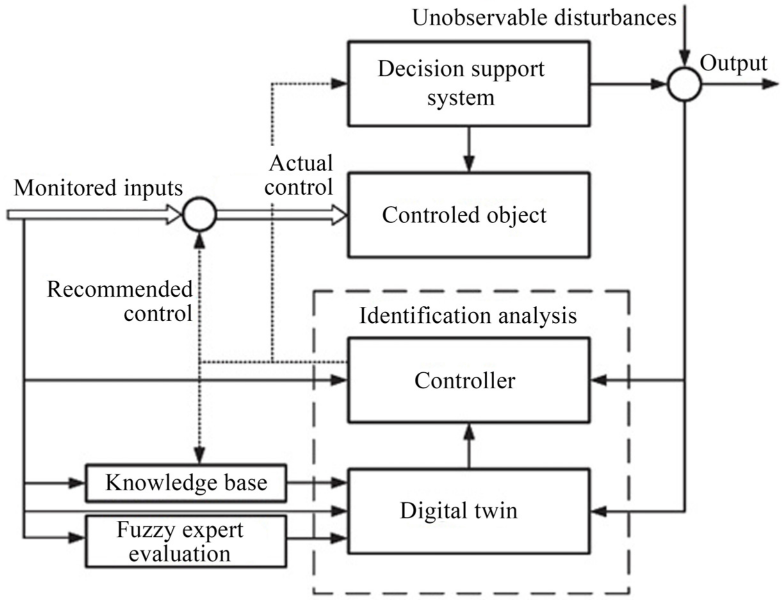

Fig. 3. Regulation scheme with an identifier in the feedback circuit.

$\hat{\alpha}$ is an LSM-estimate and, according to the Gauss–Markov theorem, has minimal dispersion in the class of unbiased linear estimates of the parameter α.

For dynamic models, we have the case of poor matrix conditioning due to statistical dependence of the components of the extended vector of inputs, so the estimation (15) may be inadequate. In this case we propose to use the following Moore–Penrose procedure. It is proposed to search estimation of $\hat{\alpha}_0$ such that:

$$\hat{\alpha}_0^T \hat{\alpha}_0 = \min \hat{\alpha}^T \hat{\alpha}. \tag{16}$$

According to the Moore–Penrose theorem, the estimation $\hat{\alpha}$ minimizes the left part of (14) if and only if it is represented in the form:

$$\hat{\alpha} = \hat{X}^+ \hat{y} + (1 - \hat{X}^+ \hat{X})p, \tag{17}$$

where p is some r-dimensional vector.

Asymptotically the normal estimation $\hat{\alpha}_0$ (16) is in the form:

$$\hat{\alpha}_0 = \hat{X}^+ \hat{y}. \tag{18}$$

The same procedure can be applied to form digital identification models in a closed loop control of the process.

The associative search algorithm makes it possible to obtain models, which for each time step are described by synchronized sets of values of inputs, outputs, control actions, coefficients, i.e., digital models, which are formed by the identifier in the feedback circuit of the MPC. In this case, the identifier in the feedback circuit of the control system is a digital twin (Fig. 3), because it forms a digital predictive model based on the current and statistical data of the process.

For non-stationary processes, as we know from [11], the associative search method also offers a constructive solution to the identification problem, namely, wavelet transform. This approach has demonstrated efficiency both for nonstationary input signal and in the case of impossibility of modeling internal dynamics of the control object. To apply the associative search algorithm for the purpose of predicting the dynamics of nonstationary processes, it is necessary to select from the technological archive the vectors that are close to the current one according to the criterion formed for the coefficients of the multiple-scale wavelet decomposition.

5. CONSTRUCTION OF INTELLIGENT SOFT SENSOR BASED ON ASSOCIATIVE MODELS

We propose the following method of developing a software and algorithmic complex for the formation of real-time models of associative search — associative soft sensor.

It is necessary to take into account the specifics of laboratory work in production for process, when building the model. Often the laboratory analysis of products is performed not on one-time, but on averaged samples. Modeling should provide not only a sufficiently accurate description of the process class, but also adequately reflect the specifics of the production situation and the characteristics of a particular process. It is important to use all the available a priori information as much as possible. In particular, it is necessary to take into account all the constraints, determined by both technological regulations and expert opinions. The decision maker (operator or technologist) acts as an expert analyzing the situation.

Thus, the following main elements are highlighted in the development of soft sensors for process.

Description of the process: its features, allowing to formalize it with the help of certain mathematical models; accounting for certain data of the technical documentation, in particular, the technological regulations (or similar documents), which allows you to determine all the necessary restrictions:

— Base of technological regulations contains (in formalized form) the description of the following items: equipment, technological standards, rules of operations in various situations, detailed order of process, mode rules, mode parameters;

— Database of libraries for formalized representation of processes and their mathematical models.

The scheme of information flows of the investigated process, formed by the user of the system using the interactive interface, which allows to formalize the description of the simulated process in the form of differential or finite-difference equations.

Process inductive knowledge base must contain an archive of "production experience" of a particular process: from time-synchronized inputs and their corresponding outputs to the archives of configured models and archives of formalized situations ("coded" features and characteristics of the current state). The main elements of the system for formation and storage of knowledge, interpreted as patterns characterizing the process, are the following:

— Database of the functioning process;

— Data of technological equipment — actual values at specific time instants of technological parameters: consumption, pressures, temperatures, etc.; possible deviations from standard situations (set of patterns), additional limitations;

— Base of constructed point process models (archive of constructed models): sets of values of inputs and controls, as well as their corresponding outputs (finished products, by-products, waste) according to monitoring data;

— Assessments of results and recommendations for management (including formalized ones), that is, evaluations obtained by associative identification algorithms, as well as formalized values of experts' assessments (e.g., by means of fuzzy models).

The identifier in the feedback circuit of the automatic / automated process control system (digital twin) generates a digital model at each moment of time. Various elements of the formalized description of both the process itself and its current state come to the input of the identifier.

6. BUILDING AN INTELLIGENT SOFT SENSOR FOR A CONCENTRATOR OF MINING PRODUCTION

An enrichment plant is a mining facility for the primary processing of solid minerals in order to obtain technically valuable products suitable for industrial use. By means of various technologies (flotation, magnetic separation and others), a concentrate is obtained from the mined ore, in which the content of the useful component is much higher than in the original raw material. Concentrators process enrich ores of ferrous and nonferrous metals, nonmetallic minerals, and coal.

Below we compare results of laboratory control of product quality (iron concentration) with the quality metrics generated by soft sensors, developed on the basis of multiple linear regression and regression with associative search. The performance quality criteria of soft sensors are the mean absolute error (MAE) of the model and Pearson's correlation coefficient.

Values of laboratory analysis and soft sensors predictions of different types on iron concentration are presented on Figs. 4 and 5. Table shows the corresponding performance quality criteria of two types of soft sensors.

The formula for the linear regression model of the iron concentration:

$$F = 76.34 + 0.00545D, \tag{19}$$

where F — iron content in the sectional concentrate, D — drain density in hydrocyclones.

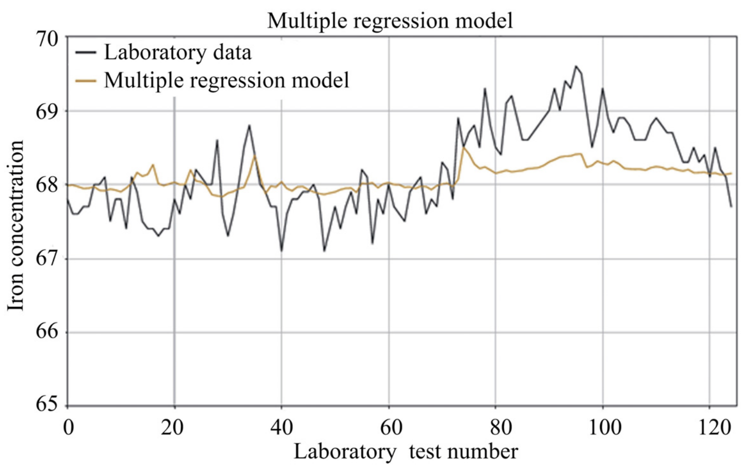

Fig. 4. Comparison of soft sensors values (multiple linear regression) with laboratory control data.

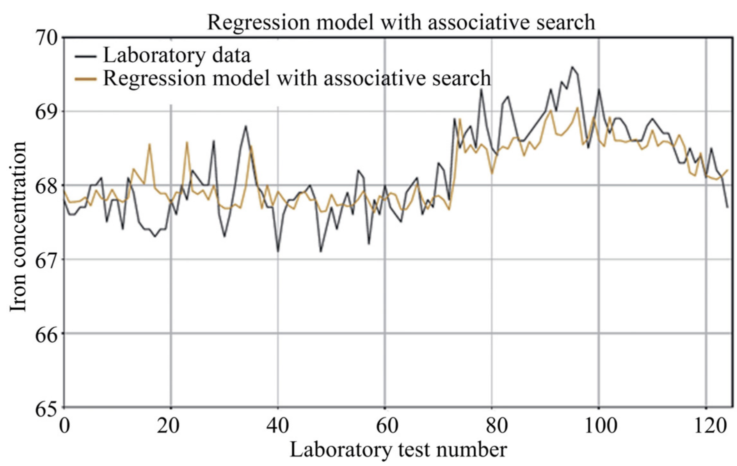

Fig. 5. Comparison of soft sensors values (regression with associative search) with laboratory control data.

Summary table of evaluation parameters

Model quality parameter	Linear regression	Regression with associative search
Mean absolute error (MAE)	0.416	0.282
Mean square error (MSE)	0.285	0.129
Pearson's correlation coefficient	0.479	0.803

Table shows that MAE and MSE for soft sensor based on associative model are smaller, and Pearson correlation coefficient is larger in comparison with soft sensor based on regression model. Thus, the regression model with associative search gives a more accurate prediction than the linear regression.

7. CONCLUSION

In this work the method of building soft sensors using intelligent identification algorithm, which forms real-time models of process by means of intelligent data analysis and machine learning is proposed. The new algorithm of soft sensors based on regression with associative search in inductive knowledge base is presented. Numerical simulation of the proposed methods for the process of a concentrator of mining production was conducted. The results demonstrate the higher accuracy and efficiency of regression model with associative search over classical regression models.

FUNDING

The reported study was funded: by the Russian Science Foundation, project no. 19-19-00673, by RFBR, according to the research project by RFBR and NSFC, project no. 21-57-53005.

REFERENCES

1. Bakhtadze, N.N., Virtual Analyzers: Identification Approach, *Autom. Remote Control*, 2004, vol. 65, no. 11, pp. 1691–1709.
2. Lototsky, V., Chadeev, V., Maksimov, E., and Bakhtadze, N., Prospects of Application of Virtual Analyzers in Production Control Systems, *Automation in Industry*, 2004, no. 5, pp. 23–29.
3. Vapnik, V., *Vosstanovlenie zavisimostei po empiricheskim dannym* (Reconstructing Dependencies from Empirical Data), Moskow: Nauka, 1979.
4. Bakhtadze, N., Kulba, V., Lototsky, V., and Maximov, E., Identification-based Approach to Soft Sensors Design, *Proceedings of IFAC Workshop of Intelligent Manufacturing Systems*, 2007, vol. 40, no. 3, pp. 86–92.
5. Bakhtadze, N., Sakrutina, E., and Pyatetsky, V., Predicting Oil Product Properties with Intelligent Soft Sensors, *IFACPapersOnLine*, 2017, vol. 50, no. 1, pp. 14632–14637.
6. Bakhtadze, N., Sakrutina, E., Pavlov, B., Lototsky, V., and Zaikin, O., Knowledge-based Prediction in Process Control Systems under Limited Measurement Data, *Procedia Computer Science J.*, 2017, vol. 112, pp. 1225–1237.
7. Bakhtadze, N., Chereshko, A., Elpashev, D., Suleykin, A., and Purtov, A., Predictive Associative Models of Processes and Situations, *IFACPapersOnLine*, 2022, vol. 55, no. 2, pp. 19–24.
8. Chereshko, A. and Titkina, M., Application of Associative Search Algorithms in Control Systems with a Predictive Model, *Avtomatizatsiya v Promyshlennosti*, 2022, no. 6, pp. 58–62.
9. Patel, V. and Ramoni, M., Cognitive Models of Directional Inference in Expert Medical Reasoning, in *Expertise in Context. Human and Machine*, MIT Press, 1997, pp. 67–99.
10. Razumkov, M., Verbal Analysis Methods: Research and Comparison, *Fundamental'nye Issledovaniya*, 2016, no. 10-3, pp. 642–646.
11. Bakhtadze, N., Lototsky, V., Vlasov, S., abd Sakrutina, E., Associative Search and Wavelet Analysis Techniques in System Identification, *Proceedings of the 16th IFAC Symposium on System Identification*, 2012, vol. 45, no. 16, pp. 1227–1232.