IDENTIFICATION OF SUBMERSIBLE PUMP TEMPERATURE CHANGES MODEL USING KDD METHODS*

Dominik WACHLA

Silesian University of Technology, Department of Fundamentals of Machinery Design
Konarskiego Street 18A, 41-100 Gliwice, Poland, e-mail: dominik.wachla@polsl.pl

Summary

This paper deals with the problem of the autoregressive model identification using KDD methods. In the considered problem, the autoregressive models are applied to describe dynamics processes of various technical systems. In particular, a method of functional dependencies discovering was presented. The method was designed for exploring data sets gathered by industrial SCADA systems. For the problem of the identification of pump temperature changes model, the method was verified. For this particular reason, a set of data was used which was gathered by submersible pumping station SCADA system. The assumptions, the exemplary results of the conducted research and conclusions were presented, as well.

Keywords: databases, knowledge discovery in databases, system identification, genetic algorithm, support vector machines, attributes selection, SCADA systems.

IDENTYFIKACJA MODELU ZMIAN TEMPERATURY POMPY GŁĘBINOWEJ Z ZASTOSOWANIEM METOD ODKRYWANIA WIEDZY W BAZACH DANYCH

Streszczenie

W artykule poruszono problem identyfikacji modeli autoregresyjnych opisujących dynamikę obserwowanych procesów. W szczególności przedstawiono metodę odkrywania zależności funkcyjnych w zbiorach danych gromadzonych przez przemysłowe systemy SCADA. Opracowaną metodę zweryfikowano dla problemu identyfikacji modelu zmian temperatury pompy głębinowej. W tym celu zastosowano fragment danych zgromadzony przez system rejestracji danych współpracujący pompownią głębinową. Przedstawiono przyjęte założenia, fragmenty uzyskanych wyników oraz wnioski z przeprowadzonych badań.

Słowa kluczowe: bazy danych, odkrywanie wiedzy w bazach danych, identyfikacja systemów, algorytm genetyczny, metoda wektorów wspomagających, selekcja atrybutów, systemy SCADA.

INTRODUCTION

Currently, a lot of technical objects and industrial installations have a Supervisory Control and Data Acquisition (SCADA) systems [6]. They are used for the control and monitoring processes occurring in technical systems. One of the elements of SCADA systems is a database. The database is used for recording a measurement results of the observed system variables. The recorded data can be a source of valuable knowledge about a dynamics of processes occurring in the observed technical systems. A lot of available data and their imperfection require that the methods of Knowledge Discovery in Databases (KDD) should be used in a process of knowledge acquisition.

On the other hand, a typical analysis of process data consists of the identification of the quantitative relation occurring in a set of observed process variables. These relations are described by autoregressive models (**Fig. 1**) [6],[9].
However, the identified autoregressive models don't explain the physical essence of the observed processes. Nonetheless, they can be applied for the control and diagnosis in a case when their accuracy is appropriate.

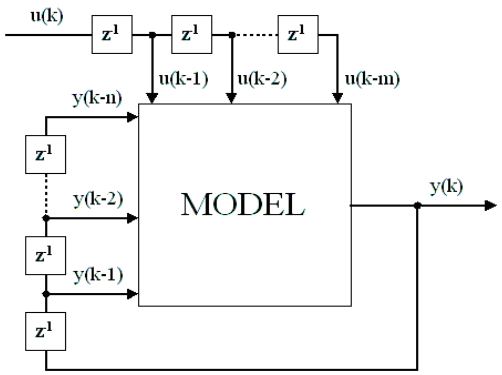


Fig. 1. The general structure of the autoregressive MISO models [10]

The main aim of conducted research was to develop a method of the autoregressive models identification of the observed processes in technical systems. The developed method constitutes

* The research was financed by the Minister of Science and Information Society Technologies (grant No 4 T07B 059 26).

a combination of selected KDD methods and has been designed for the analysis of the data gathered by SCADA systems. The verification of the proposed method was conducted with the application of data gathered by a SCADA system of a submersible pumping station.

1. THE METHOD

In the figure 2 the structure of the proposed method was presented. The method consists of two stages. In the first one, a data transformation into multidimensional space of regressors is done. The data transformation is conducted with the application of the TSI2SI algorithm [10]. The TSI2SI algorithm operation causes change of data description. As a result of TSI2SI algorithm operation, the data in the form of time series are transformed into a set of learning examples (learning vectors):

$$\begin{aligned}&\{a_1(k), a_1(k-1), \ldots, a_1(k-n), \ldots, a_i(k), a_i(k-1),\\ &\ldots, a_i(k-n), \ldots, a_m(k), a_m(k-1), \ldots, a_m(k-n)\}\end{aligned} \quad (1)$$

where: m - number of process variables, n - time horizon. Then, the transformed data can be explored by adequate methods of Data Mining (DM).

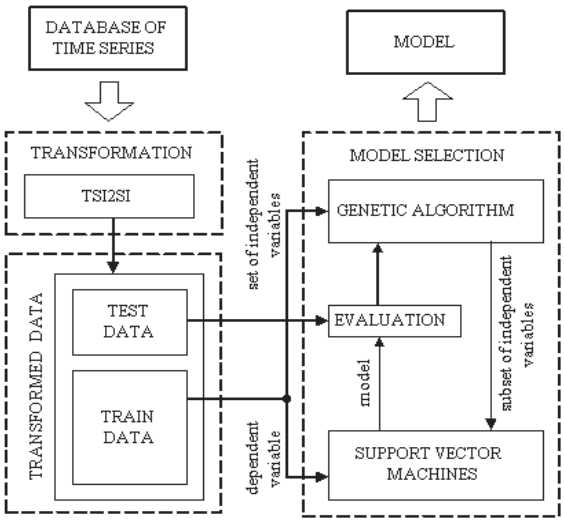


Fig. 2. The structure of the developed method

The application of high values of the parameters m and n in the TSI2SI algorithm induces a big growth in the dimensions of learning vectors (1). For that reason, the second stage of the proposed method is combination of the induction of functional dependencies and the selection of independent variable. This approach is consistent with the wrapper approach [5]. In particular, for the approximation of functional dependencies the Support Vector Machines (SVM) [8] method was chosen. Moreover, the Simple Genetic Algorithm (SGA) [4] for the selection of the independent variables was chosen.

The evaluation function (Fig. 2) is a significant element of the process of the independent variables selection. The general expression for the selection of independent variables using the evaluation function has the following form [6],[9]:

$$\widetilde{A} = \arg\min_{A_i \subseteq A} J(h(A_i)) \quad (2)$$

where: A – set of independent variables, A_i – subset of A, $h(A_i)$ – model $J(\,)$ – evaluation function.

A lot of model evaluation functions based on the theory of information in the domain of system identification was developed. They allow to evaluate identified models when one takes into account the structure complexity and accuracy. One of the most popular and used function is Akaike Information Criterion (AIC) [1],[9]:

$$J_{\mathrm{AIC}} = N \log \mathrm{MSE}\left(h(A_i)\right) + 2\,\mathrm{card}(A_i), \quad (3)$$

where: MSE – the Mean Square Error, card(A_i) – a cardinal number of A_i.

Other approach are based on heuristic rules. In case when the two compared models have a similar accuracy and a different structure complexity, the general rule is to choose a model that have a lesser complexity of structure. Basing on this rule we can construct the heuristic function for models selection in the form:

$$J_{\mathrm{HF}} = N \log \mathrm{acc}\left(h(A_i)\right) - \rho\left(\frac{\mathrm{card}(A) - \mathrm{card}(A_i)}{\mathrm{card}(A)}\right), \quad (4)$$

where: acc( ) - the accuracy function eg. Mean Square Error (MSE), Mean Absolute Percentage Error (MAPE) and etc.

In the equation (4) the parameter ρ has an important role since its value influences on the process of models selection. The ρ parameter allows to determine a degree of the significance of two criteria considered in equation (4), i.e. the model accuracy criterion and model complexity criterion. For the increasing values of ρ the model complexity criterion is more important than the model accuracy criterion and vice versa for the decreasing values of ρ.

2. THE DATABASE

The economic situation in the Polish Mining Industry has caused liquidation of unprofitable coal-mines. As a result of liquidation a lot of pit shafts have remained. Some of the remaining pit shafts are flooded by underground water. This situation has caused a risk of flooding operative coal-mines. To prevent this, submersible pumping stations were installed. Each of the submersible pumping stations consist of four pumps and has SCADA systems for controlling and monitoring water levels and pumps operation. Each of the installed SCADA system has a database. The database is used for recording

measurements of 34 observed process variables. The measurements are conducted with 1 second interval. For the verification of the method a part of data from the database of a submersible pumping station SCADA system was used. The shared data describe a operation of pumping station within a 3,5 months.

From the point of view of the machinery diagnostics, the following attributes are interesting:

- T - the pump temperature,
- C - the pump capacity,
- P - the power of pump motor,
- E - the current intensity taken by pump motor,
- S - the state of pump operation.

Basing on consultations with experts [3] and our own observations, the pump temperature T was established as significant process variable. In connection with this, a review of available data was carried out. As a results of the review, it was stated that data connected with one of the pumps can be used for the identification of pump temperature changes model because these data describe a selected pump in a ,,good" technical state. Then the identified model can be used for the diagnosis of a change of the pump technical state. The basis of the diagnosis is a residual analysis here. Additionally, a correlation analysis among all attributes in the shared database was carried out. The correlation analysis has shown that the pump capacity C, power P, current intensity E and sate of pump operation S are a strongly correlated. In the connection with this a set of considered variables was limited to the temperature T and to the state of pump operation S. Then these variables were used for the identification of a submersible pump temperature changes model.

3. MODEL IDENTIFICATION

For identification of submersible pump temperature changes model the following values of parameters in the proposed method was chosen:

- Transformation *TSI2SI*:
  - n=10,
  - Δt=1,
- Genetic Algorithm [4]:
  - crossover probability $p_c = 0.7$,
  - mutation probability $p_m = 0.01$,
  - size of population: 100,
  - number of iteration: 1000,
  - selection method: roulette,
  - type of succession: trivial.
- Support Vector Machines [7],[8]:
  - SVM algorithm: ν-SVR (ν=0.54,C=1),
  - classes of considered SVM models: (Tab. 1).
- Selection criterion: (Tab. 2).

For calculating a value of acc() function (Tab. 2), the Hold-Out method [2] was chosen. In Hold-out method a set of learning data was divided into two subset. The first subset of learning data is used as a training set and the second subset of learning data is used as a testing set. In the conducted research, sets of the training and testing data had the same number of elements. The complete set of the obtained results of the identification of submersible pump temperature changes model was presented in [10]. In the figure 3, an example of one of discovered models with linear SVM kernel and his statistical evaluations was presented.

Tab. 1. The considered classes of SVM models [10]

Type of model	Kernel	Parameters
MP1	*Linear*	—
MP2	*RBF*	γ=0.001
MP3	*RBF*	γ=0.01
MP4	*RBF*	γ=0.1
MP5	*RBF*	γ=1

Tab. 2. The plan of pump temperature changes model identification [10]

Criterion	*acc*( )	ρ	Type of model
J_{AIC}	MSE	—	*MP1, MP2, MP3, MP4, MP5*
J_{HF}	MAPE	0.1	
J_{HF}	MAPE	0.5	
J_{HF}	MAPE	1.0	

4. SUMMARY

In this paper the method of the data exploration was presented. The method has been designed for analyzing of data gathered by industrial SCADA systems. The essence of the developed method is the data transformation into multidimensional space of regressors. The applied transformation of the process data causes an increase in a number of considered variables. In the connection with this, a variables selection is necessary. For this reason the process of variables selection basing on wrapper approach [5] was applied.

For the identification of the submersible pump temperature changes model the developed method was applied. The part of the data gathered by SCADA system of a submersible pumping station was used as well.

As a results of the conducted research, a lot of submersible pump temperature changes models were obtained. All of the discovered models have a large deviation error in heating and cooling phases of pump operating. Probably, it is caused by using linear and RBF kernels in SVM method only. Nevertheless, the SVM models with linear kernels are more adequate than SVM models with RBF kernel. In the case of nonlinear models (RBF kernel) only the models of MP5 type have a similar degree of reduction of independent variables. The values of

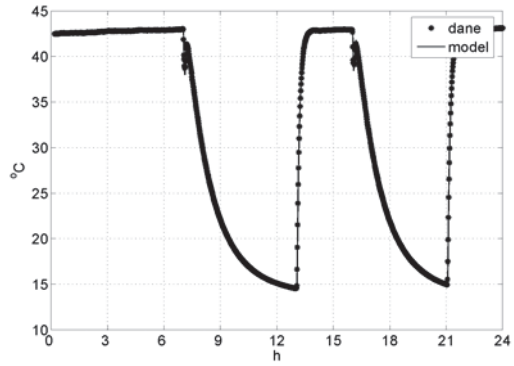
calculated statistics are similar too. However, this accuracy has resulted from an increase in support vectors.

For all of the considered classes of SVM models, the greater values of ρ parameter have caused a decrease in the number of independent variables.

The presented research will be continued to verified the usefulness of different methods of states space searching and other kernels of SVM than linear an RBF. Additionally, we are going to identify a diagnostic model for the recognition of a pump technical state using the knowledge of the submersible pumping station staff.

- SVM model
  kernel function : linear
  kernel parameters : —
  number of SV : 786
  independent variables : *T(k−1), T(k−2), T(k−5), S(k−4)*
  dependent variable : *T(k)*
- Model evaluation
  MSE : 3.1416 E-2
  MAPE : 0.1469
  correlation : 0.9998

a)



b)

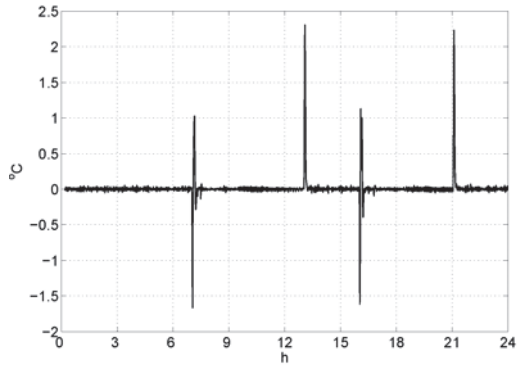


Fig. 3. The exemplary results of the pump temperature changes model identification with the use of SVM linear kernel and $J_{HF}^{\rho=0.5}$ criterion: (a) model (b) residuum [10]

REFERENCES

[1] Akaike H.: A new look at the statistical model selection. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.

[2] Ciupke K.: *Selection and reduction of information in machinery diagnostics*. Vol. 118. Department of Fundamentals of Machinery Design, Silesian University of Technology, Gliwice, 2001 (in Polish).

[3] Gibiec M.: Intelligent health prognostics of machines used in mining industry. *VI National Conference "Diagnostics of Industrial Processes" DPP'05, Rajgród*, ss. 243–245, Warsaw University of Technology, 2005 (in Polish).

[4] Goldberg D.E.: *Genetic Algorithms in Search, Optimization and Machine Learning*. WNT, Warszawa, 1998 (in Polish).

[5] Kohavi R., John G. H.: Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–324, 1997.

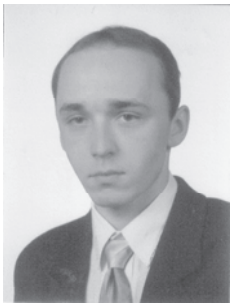
[6] Korbicz J., Kościelny J.M., Kowalczuk Z., Cholewa W. (Eds.): *Diagnostics of Processes. Models, Methods of Artificial Intelligence, Applications*. WNT, Warszawa 2002 (in Polish).

[7] Schölkopf B., Bartlett P. L., Smola A. J., Williamson R.: Shrinking the tube: a new support vector regression algorithm. vol. 11, ss. 330 – 336, Cambridge, MA, 1999. MIT Press.

[8] Schölkopf B., Smola A. J.: *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* MIT Press, Cambridge, MA, USA, 2001.

[9] Söderström T., P. Stoica P.: *System Identification*. PWN, Warszawa, 1997 (in Polish).

[10] Wachla D.: *Identification of Dynamic Diagnostic Models Using Methods of Knowledge Discovery in Databases.* Phd Thesis, Silesian University of Technology, Department of Fundamentals of Machinery Design, Gliwice, 2006 (in Polish).



Dominik WACHLA (Phd Eng.) is an assistant at the Faculty of Mechanical Engineering at Silesian University of Technology at Gliwice. His research is focused of the application of methods of artificial intelligence in the technical diagnostics of machinery and processes.