Zofia M. ŁABĘDA-GRUDZIAK
INSTITUTE OF AUTOMATIC CONTROL AND ROBOTICS, WARSAW UNIVERSITY OF TECHNOLOGY
ul. św. Andrzeja Boboli 8, 02-525 Warszawa

Identification of dynamic system additive models by KDD methods

M.A. Zofia M. ŁABĘDA-GRUDZIAK

A Faculty of Mathematics and Information Sciences at Warsaw University of Technology Graduate, and working towards a Ph.D. at the Institute of Automatic Control and Robotics at the Faculty of Mechatronics at Warsaw University of Technology. Her research interests are in the general areas of fault diagnosis, industrial process modelling based on data mining technique including both theory and applications, computational process modelling and simulation, and applied mathematics.

e-mail: Z.Labeda@mchtr.pw.edu.pl

Abstract

The goal of this paper is to present a new way of knowledge discovery data (KDD) application to construct a statistical model that describes dynamic systems. This includes presentation of data mining as an iterative and adaptive process, from communication of the research problem through data collection, data preprocessing, model building, model evaluation, and finally, model deployment. The types of models discussed in this paper are in form of additive models and can be used for prediction of process outputs, for calibration, or for diagnostics purposes. The backfitting algorithm with nonparametric smoothing techniques was used for estimation of the additive model. The example of application of the methods, conclusions and remarks are presented as well. The research was carried out based on archival process data recorded in the Lublin Sugar Factory S.A.

Keywords: identification, additive model, databases, knowledge discovery data, dynamic systems.

Identyfikacja addytywnych modeli obiektów dynamicznych metodami odkryć wiedzy w bazach danych

Streszczenie

Celem niniejszej pracy jest zaprezentowanie nowego podejścia do identyfikacji modeli obiektów dynamicznych metodami odkryć wiedzy w bazach danych. W szczególności przedstawiono eksplorację danych jako proces iteracyjny i adaptacyjny, od zrozumienia uwarunkowań badawczych, przez zebranie danych, przygotowanie danych, modelowanie, ewaluację modelu do jego wdrożenia. W badaniach wykorzystano addytywny model regresji, który może posłużyć do przewidywania wartości wyjściowych procesu, kalibracji, a także w celach diagnostycznych. Do wyznaczenia parametrów modeli addytywnych zastosowano algorytm dopasowania wstecznego i nieparametryczne techniki estymacji. Badania przeprowadzono na podstawie archiwalnych danych pomiarowych zarejestrowanych w Cukrowni LUBLIN S.A.

Słowa kluczowe: identyfikacja, model addytywny, bazy danych, odkrywanie wiedzy z danych, obiekty dynamiczne.

1. Introduction

Contemporary industrial installations often accomplish very complex production processes. Nowadays each industrial system can access a large quantity of data and information about itself and its environment. This data has the potential to predict the evolution of interesting variables or trends and allows us to create models based on measured data and expert's knowledge about the object.

With the growth of computer technology, software and hardware continually offer more power at lower cost, allowing easier access and transfer, storing data acquisition, speed of their processing. The integration of automated or semi-automated data acquisition systems, e.g. Supervisory Control and Data Acquisition (SCADA), into production systems resulted in enormous data volumes. The science of extracting useful information from large data sets or databases is known as data mining, and is the most well-known branch of knowledge discovery, also known as Knowledge Discovery in Databases (KDD) [1,2,3]. The key task of this discipline for diagnostic purposes is the analysis of observational data sets, and how to find the relevant information quicker and more accurately, which aids the decision making about the recognition of changes of the state of the process during its operation.

In order to meet reliability requirements of safety-critical processes, properly maintained industrial process control increased efficiency, safety, profitability, and help the ecology. Moreover, the process economy requires that the number of breaks, switch-offs and the service costs are as low as possible. For this reason, the knowledge-based applications facilitate the proper and optimal decisions on the emergency and corrective actions and on repairs.

For many types of data analysis problems there are no more than a couple of general approaches to be considered on the route to the problem solution. Within the different approaches for a specific problem type, there are usually at most a few competing tools that can be used to obtain an appropriate solution. The bottom line for most types of data analysis problems is that selection of the best method to solve the problem is largely determined by the goal of the analysis and the nature of the data. The classical way of identification of the industrial dynamic systems was usually based on the analytical models or models based on fuzzy logic, artificial neural networks and fuzzy neural networks [4, 5]. For many systems, the model study based on differential and algebraic physical equations was either very difficult or almost impossible, and model parameters identification yields further difficulties. Moreover, an increasing number of input signals rapidly increases the computational costs and number of rules, in neural network and fuzzy logic modelling, respectively.

In this paper, an alternative technique which overcomes the limitation of the multivariate modelling is presented. This important class of flexible models arises in form of additive models [6]. In the paper the author explores a new way of additive models and knowledge discovery data application, in the context of dynamic systems identification. Detailed information on how to collect data, prepare data, construct appropriate models, evaluate the models delivered in the modelling phase, and make use of the models is covered in the following sections. The final section of the paper contains a case study that illustrates the general information presented in the first five sections using archival process data recorded in the Lublin Sugar Factory S.A.

2. Knowledge discovery process

Knowledge discovery is defined as "the non-trivial extraction of implicit, unknown, and potentially useful information from data" [2]. The knowledge discovery process is a series of activities from defining objectives to evaluating results. Here are its six phases: research understanding phase, data understanding phase, data preparation phase, modeling phase, evaluation and deployment phase. The next phase in the sequence often depends on the outcomes associated with the preceding phase.

2.1. Research understanding phase

Definition of the objectives involves defining the aims of the analysis. First of all we must formulate the project objectives and requirements from a research perspective. This includes a discussion of what process modeling is, the goals of process

modeling, such as answering a scientific or engineering question, and translates these goals into the formulation of a knowledge discovery data problem definition. A clear statement of the problem and objectives that need to be achieved are the prerequisites for setting up the analysis correctly. Finally, researchers must prepare a preliminary strategy for achieving these objectives.

2.2. Data understanding phase

Once the objectives of analysis have been identified, it is necessary to select data for the analysis. First of all, it is necessary to identify the data sources. In many cases there are a lot of archival SCADA data available to use, that are cheaper and more reliable. It is often useful to set up an analysis on a subset or sample of available data. Working with samples allows us to check the model validity against the rest of the data, giving an important diagnostic tool. Second, we must try to relate data with each other and then find hidden trends in data.

2.3. Data preparation phase

When the database is available, it is often necessary to undergo preprocessing in the form of data cleaning and data transformation. Much of the row of databases may contain missing values, outliers, and incorrect data.

Missing data is a problem in data analysis methods. A common method of handling missing values is to delete records or fields with missing values. The disadvantage of this approach is reduction of the sample size of data and that the records with missing values may be different than those without missing values (e.g., missing values that are non-random), so we have a non-representative sample after removing the cases with missing values. Alternatively, data analysts have turned to methods that would replace the missing value with some constant, or with the field mean, and or with a value generated at random from the variable distribution observed [1, 3]. If any essential information is missing, it will then be necessary to review the phase that highlights the source.

Outlier is an observation that is numerically distant from the rest of the data. Outlier points can therefore indicate faulty data, erroneous procedures, or areas where a certain theory might not be valid. However, in large samples, a small number of outliers is to be expected and not due to any anomalous condition. One graphical method for identifying outliers is to examine a histogram of the variable. In another way we can use Z-score standardization [1]. Usually, an outlier can be identified because it is much faster than 3 standard deviations from the mean and therefore has a Z-score standardization that is either less than -3 or greater than 3.

Variables tend to have ranges that vary greatly from each other. Therefore, the analyst should normalize their numerical variables, by e.g. min-max normalization or Z-score standardization.

Data cleaning is not just about removing bad data, but about finding hidden correlations in the data, identifying sources of data that are the most accurate, and determining which columns are the most appropriate for use in analysis. Thus, it is necessary to select the input signal and output signals for identifying a dynamic system, i.e. the physical measurable quantities influencing the process and resulting from the process operation, respectively.

2.4. Modeling phase

Model building, however, is different from most other areas of statistics with regard to method selection. There are more general approaches and more competing techniques available for model building than for most other types of problems. There is often more than one statistical tool that can be effectively applied to a given modeling application. The large menu of methods applicable to modeling problems means that there is both more opportunity for effective and efficient solutions. Thus, the analyst must select and apply appropriate techniques, and then calibrate model settings to optimize results. Indirect information on the effectiveness of a process modeling analysis can be obtained by checking the validity of the underlying assumptions. Confirming that the underlying assumptions are valid helps ensure that the methods of analysis were appropriate.

The most basic assumption inherent to all statistical methods for process modeling is that the process to be described is actually a statistical process. Thus, in order to successfully modeling a process using statistical methods, it must include random variation. Another assumption is that the mean of the random errors at each combination of explanatory variable values is zero and the random errors have a constant standard deviation. With most process modeling methods, inferences about the process are based on the idea that the random errors are drawn from a normal distribution. One reason this is done is because the normal distribution often describes the actual distribution of the random errors in real-world processes reasonably well. The overall goal of all statistical procedures is to enable valid conclusions to be drawn from noisy data. In the most difficult processes the signals must be filtered. Noise reduction is the process of removing noise from a signal and forms a problem in many engineering systems. A method of cancelling noisy data is the finite impulse response (FIR) filter.

To produce a final decision it is necessary to choose the best model of data analysis form the statistical methods available. We must first choose an appropriate functional form of the model according to the aim of the analysis. Given the form or structure of a model, we choose appropriate values for its parameters by estimation – that is, by minimizing or maximizing an appropriate score function measuring the fit of the model to the data. In theory, there are as many different ways of estimating parameters as there are objective functions to be minimized or maximized. The two major methods of parameter estimation for process models are maximum likelihood and least squares or weighted least squares.

2.5. Evaluation and deployment phase

Before we deploy a model into an industrial environment, we want to test how well the model performs. Also, when we build a model, we typically create multiple models with different configurations and test all models to see which yields the best results for our problem and our data. After that, we make use of the models created to predictions, which we can then use to make diagnostic decisions.

In this paper, an alternative technique which overcomes the limitation of multivariate nonparametric modeling is presented. This important class of flexible models arises in form of additive regression models [6, 7].

3. Additive model

Considering the structure with $p>1$ input signals $X_1, X_2, ..., X_p$, and one output signal Y, the additive model is defined by

$$Y = \alpha + \sum_{j=1}^{p} \varphi_j(X_j) + \varepsilon, \tag{1}$$

where error ε is a sequence of independent and identically distributed random variables (*iid*) with the mean $E(\varepsilon)=0$ and the finite variance $Var(\varepsilon)=\sigma^2$, where σ is standard deviation. The φ_j s are unknown, arbitrary univariate functions one for each predictor X_j. Functions φ_j s can be, for example, roots, logarithms or trigonometric functions. Since each variable is represented separately, model (1) summarizes the contribution of

each predictor with a single coefficient and provides a simple method for predicting new observations.

For a pair $\{(x_{ij}, y_i)_{i=1}^{n}\}_{j=1}^{p}$ of a random sample, where y_i represent measurements of the variable Y and x_{ij} are the n observed values of the variable X_j, the additive model can be estimated by minimization of the residual sum of squares

$$\arg\min_{\{\alpha,\varphi_j\}} \sum_{i=1}^{n}(y_i - \alpha - \sum_{j=1}^{p}\varphi_j(x_{ij}))^2 , \qquad (2)$$

which is the discrepancy between the data and our estimation model. The estimators of function φ_j are found by use of nonparametric estimators, such as natural cubic splines or local polynomials. These estimators, called smoothers, have a single smoothing parameter. For choosing the smoothing parameter, automatic selection was used by the generalized cross-validation or a graphical method helping us to choose the appropriate value [6, 7].

We want the functions to be fitted simultaneously so we need the unconventional estimation methods of the additive model. The most general method is the iterative backfitting algorithm [6, 7, 8].

4. Backfitting algorithm

Denote by $\hat{\Phi}_j = (\hat{\varphi}_j(x_{1j}),..., \hat{\varphi}_j(x_{nj}))^T$ and $\mathbf{y} = (y_1,..., y_n)^T$, the vector of predictor and measured response at n design vectors $\mathbf{x}_j = (x_{1j},..., x_{nj})^T$, appropriately. Determine the natural cubic spline or local polynomial smoother $\mathbf{S}_j$, which denotes a smooth of the response $\mathbf{y}$ against the predictor $\mathbf{x}_j$.

Estimation of the additive model runs in the following steps of the backfitting algorithm:

backfitting algorithm

i. Initialize $\hat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\hat{\Phi}_j^{(0)} \equiv 0$ for $j = 1,..., p$.

ii. For $j = 1,..., p$ calculate

$$\hat{\Phi}_j^{(m)} = \mathbf{S}_j(\mathbf{y} - \hat{\alpha} - \sum_{k \neq j}\hat{\Phi}_k^{(m-1)}) .$$

iii. Continue step (ii) until individual functions don't change:

$$\| \hat{\Phi}_j^{(m)} - \hat{\Phi}_k^{(m-1)} \| < \rho,$$

where ρ is fixed, small number and $\|\cdot\|$ is fixed norm in functional space. Thus, at each step, one component is estimated keeping the other components fixed, the algorithm proceeding component by component and iterating until convergence. The initial functions might be any sensible function, for example the linear regression of $\mathbf{y}$ on the predictors.

The convergence to uniqness and independent of the starting value solution was proved [6, 8].

5. Example

The example of applications of the knowledge discovery methods has been carried out based on the actuator chosen for benchmark in the Lublin Sugar Factory S.A. All research has been performed with use of the R-project designed to advanced statistical calculations [9].

5.1. Define the problem

The main technological task of a sugar evaporation station is to thicken the beet juice being just after the filtering and cleaning processes. The evaporation station consists of seven evaporators grouped in five sections. The first five evaporators work with natural juice circulation and the last two have another construction and work with juice circulation forced by pumps. The juice condensation process is performed using steam and vapour. Steam is produced by water-steam-boiler and is mainly delivered to the first section of the evaporation station. Whereas, vapour as recyclable medium, is produced in each evaporator and heat accumulator.

The planned aim is identification of a model of the control valve with a servomotor and a positioner, by using the KDD methods. This actuator (final control elements) connected with the evaporation station is situated on the inflow of thin juice into evaporation station. Properly maintained control valves increase efficiency, safety, profitability. Moreover, the correct monitoring of the valve facilitates the proper and optimal decisions on the emergency and corrective actions and on repairs.

5.2. Exploratory data analysis

This example has used the 20 process data between October 29 and November 18, 2001 (with sampling time 1s), acquired from SCADA system suited for grant DAMADICS entitled: "Development and Application of Methods for Actuator Diagnosis in Industrial Control Systems". Each data file contains data acquired from one day. Data are structured in a form of a matrix (86400 rows x 33 columns) of real numbers. To get the list of process variables, please refer to [10]. Based on the expert's knowledge, the variables used in the modeling process were chosen form database and given in Table 1.

Tab. 1. Variables used in the modeling
Tab. 1. Zmienne użyte w modelowaniu

Variable symbol	Variable description	Range	Unit
F	Juce flow (1st evaporator inlet)	0-500	m³/h
CV	Control value (controller output)	0-100	%
*P*1	Juce pressure (valve inlet)	0-1000	kPa
*P*2	Juce pressure (valve outlet)	0-1000	kPa
*T*1	Juce temperature (valve outlet)	50-150	°C

Before knowledge discovery data algorithms can be used, a target data set must be assembled. This set must be large enough to contain the patterns already present in the data. After that, the records must be divided into two sets, the "training set" and the "test set". The training set is used to "train" the data mining algorithms, while the test set is used to verify the accuracy of any patterns found. It reduces the risk that the statistical method might adapt to irregularities and loses its ability to generalize and forecast. In the example, the training set consists of data from the different periods of the first 14 days of installation work (29.10.2001 - 12.11.2001), approximately 17000 samples. The rest of the analyzed data were used for testing.

To be useful for knowledge discovery purposes, the databases need to undergo a preprocessing, in the form of data cleaning, data transformation and identifying outliers. Some of the field values were missing for certain records (30.11.2001) and some were the most extreme observations. All fields that have missing data or outliers in at least one of the selected variables were omitted in the analysis. Such omission does not make the sample less representative, because the database is very large. Moreover, the monitored variables are usually subjected to random noise. A practical experience shows that the result has a relatively large variance due to noise and deviations between the process and its

model. Therefore, the signals must be filtered. In order to do it, the finite impulse response of 4th order FIR filter was applied [11]. The higher order filters did not produce a noticeable improvement in quality of modeling.

In this paper the actuator was determined by the following additive model:

$$\begin{aligned} F_t = \alpha &+ \varphi_1(CV_{t-1}) + \varphi_2(CV_{t-2}) + \varphi_3(CV_{t-3}) + \varphi_4(CV_{t-4}) + \\ &\varphi_5(P1_{t-1}) + \varphi_6(P1_{t-2}) + \varphi_7(P1_{t-3}) + \varphi_8(P1_{t-4}) + \\ &\varphi_9(P2_{t-1}) + \varphi_{10}(P2_{t-2}) + \varphi_{11}(P2_{t-3}) + \varphi_{12}(P2_{t-4}) + \\ &\varphi_{13}(T1_{t-1}) + \varphi_{14}(T1_{t-2}) + \varphi_{15}(T1_{t-3}) + \varphi_{16}(T1_{t-4}) + \varepsilon_t, \end{aligned} \tag{3}$$

where ε_t, for $t = 5,...,n$, are *iid* random errors.

5.3. The modeling results

The additive model (3) was fitted by the backfitting algorithm and the natural cubic spline with degrees of freedom *df*=4, used as a smoothing parameter. Based on the training sample, the author obtained the estimated flow values (predicted F), and the real flow values from the process (F) whose graph is shown in Figure 1.

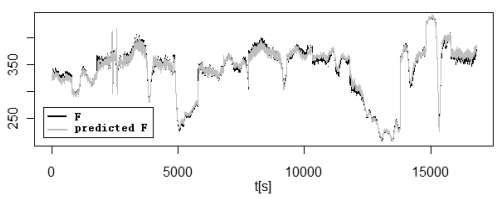

Fig. 1. Measured and modeled signals for the training sample
Rys. 1. Przebieg sygnałów pomierzonego i modelowanego dla próby uczącej

In order to examine the received model, test samples consisting of the data from the normal process behaviour were applied. Figure 2 shows graphs of the signals measured and modeled, as well as residuals (standardized(F-predicted F)) for exemplary fragments of the test samples.

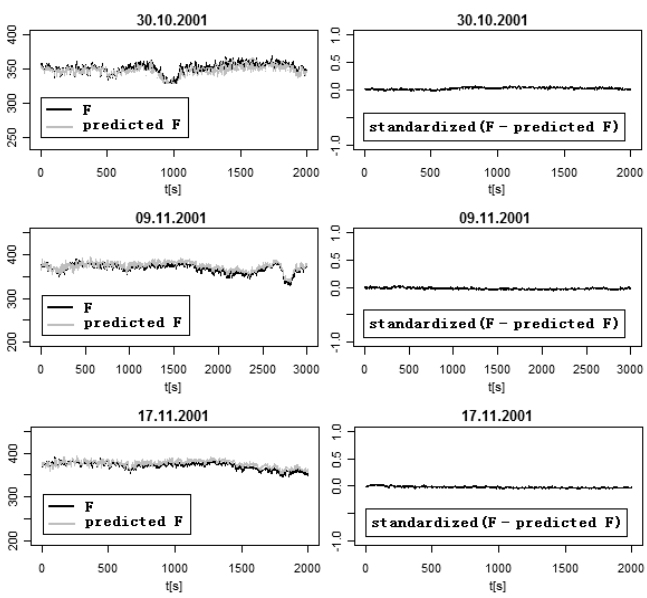

Fig. 2. Measured and modeled signals as well as residuals for test samples
Rys. 2. Przebiegi sygnałów, pomierzonego i modelowanego oraz residuów dla prób testowych

On their basis we can clearly see that the deviations from the normal process behaviour are oscillating around zero. Thus, the model (3) does fit well.

Model checking is a procedure which leads to evaluation of the model delivered in the modeling phase for quality and effectiveness. In practice we have to determine measures of variability which will describe how the measurements data are spread out in relation to the prediction data. Using the mean squared error criterion and the mean absolute deviation error [4] the errors were not higher than 4.52 and 5.1%, respectively. The variance for all fits is slightly higher than 6.4.

6. Conclusions

The knowledge discovery database is an interdisciplinary field, because it encompasses a wide variety of topics in computer science and statistics. However, the precise boundaries of the KDD process are difficult to state, so the KDD is easy to do badly. The analyses carried out on unpreprocessed data can lead to erroneous conclusions, while the inappropriate analysis can lead to big failures.

In this paper, an effective method of the additive models identification by KDD methods has been presented. This is a new way in the industrial process identification for which the difficulty associated with the problem of dimensionality is substantially reduced. Moreover, additive nonparametric regression allows researchers to evaluate data without the necessity to postulate for the relationship between the output and input signals, and to combine flexible nonparametric modeling of multidimensional inputs with a statistical precision that is typical of one-dimensional explanatory variable. The received results are satisfactory because the additive model produced the fits that closely match the true relationship between the real and predicted flow values. Therefore, the knowledge discovery data is a useful tool for identification of dynamic models of the industrial process in the analyzed structures.

This work was supported in part by the Grant from the Ministry of Science and Higher Education, no. N N514 238337, "Zastosowanie addytywnego modelu regresji do generacji residuów dla potrzeb detekcji uszkodzeń".

7. References

[1] Larose D.T.: Discovering Knowledge in Data: An Introduction to DATA MINING. Wiley, 2005.
[2] Frawley W.J., Piatetsky-Shapiro G. and Matheus C.: Knowledge Discovery In Databases. AAAI Press/MIT Press, Cambridge, 1991.
[3] Hand D., Mannila H. and Smyth P.: Principles of Data Mining. MIT Press, 2001.
[4] Kościelny J.M.: Diagnostyka procesów przemysłowych. EXIT, Warszawa, 2001.
[5] Korbicz J., Kościelny J.M., Kowalczuk Z., Cholewa W. (red): Diagnostyka procesów. Modele, metody sztucznej inteligencji, zastosowania. WNT, Warszawa, 2002.
[6] Hastie T., Tibshirani R.: Generalized additive models. Chapman and Hall, 1990.
[7] Łabęda Z.M.: Wykorzystanie addytywnego modelu regresji w eksploracyjnej analizie danych. VI Sympozjum Modelowanie i Symulacja Komputerowa w Technice, Łódź, 2008.
[8] Łabęda Z.M.: The backfitting and marginal integration estimators for additive models. IV Konferencja Naukowo-Techniczna Doktorantów i Młodych Naukowców, Warszawa, 2009
[9] Good P.I.: Introduction to statistics through reasampling methods and R/S-PLUS. Wiley, 2005.
[10] Bartyś M., Syfert M.: Lublin Sugar Factory data description, http://diag.mchtr.pw.edu.pl/pub/diamadics/Lublin/damadics-lublin-data-description-v02March2002.zip, 2002.

otrzymano / received: 01.09.2010
przyjęto do druku / accepted: 02.02.2012 *artykuł recenzowany*