MCIT
Modeling, control & information technologies

On One Approach to Predictive Modeling Based on Monitoring Data

https://doi.org/10.31713/MCIT.2020.21

Dmytro V. Stefanyshyn
Department of natural resources
Institute of Telecommunications and Global Information Space of the NAS of Ukraine, ITGIP of the NASU
Kyiv, Ukraine
d.v.stefanyshyn@gmail.com

Abstract — **The paper proposes one approach to predict the behavior of complex dynamic systems based on monitoring data presented as time series. The general idea of the approach is a decomposition of a prediction task of the behavior of a complex dynamic system possessing substantial structural and parametric uncertainty. The prediction task is performed in two stages. The first stage realizes the retrospective situational modeling for individual time periods resulting in a set of situational models in the form of simple regressions. The second stage comprises inductive modeling based on the previous situational modeling results. According to the approach, the prediction by means of such combined situational-inductive modeling is aimed at establishing relevant situational models of regression type being adequate in the future within certain limited time periods. The approach allows using simultaneously both the principle of optimization in modeling and the principle of adaptation to situational changes occurring in dynamic systems.**

Keywords—mototoring data; predictive modeling; situational and inductive models; time series

I. Introduction

Until recently, a shortage of monitoring data was often considered to be one of the main shortcomings and limitations in controlling the condition of complex dynamic systems. The implementation of automated monitoring systems (AMSs) has totally changed the situation. Nowadays, such systems can provide the acquisition and reliable storage of large arrays of different input data (so-called big data). As a consequence, another problem has arisen regarding the proper handling of these big data [1, 2]. On the one hand, AMSs can support the necessary quantitative volumes and quantitative-qualitative parameters for collection, storage, and retention of data. On the other hand, as the number of data increases, so does the complexity of their interpretation and predictive modeling based on collected data because of heterogeneity and non-stationarity of data. As a result, traditional interpretation and predictive models do not work successfully. Big data require an increase in model dimension with taking into account new factors, parameters, non-linear effects, etc., and, as a consequence, this leads often to disruption of stability in solutions and building of inadequate prognostic models.

First of all, this applies to traditional regression modeling based on monitoring data. Practice shows, the construction of more sophisticated adequate regression models can become a challenging problem [3–5]. Better interpretation models may have a tendency to overpredict (or underpredict) low values and underpredict (or overpredict) high ones [4]. Both simple and complex regressions may be getting lost in their accuracy and attractiveness in extrapolation prediction even in simple cases [3]. As a result, better interpretation models can easily overemphasize patterns that are not reproducible and demonstrate the essential instability of extrapolation in the prediction zone, and a researcher may be unaware of the fatal prediction faults until the next set of samples appears [4, 5].

II. Advantages and Challenges of Regression Modeling

Regression modeling is widely used for interpretation and prediction of the behavior of complex technogenic, ecological, and economic systems and relationships being under monitoring [1–5]. This approach enables us to simplify significantly tasks of modeling and predicting based on empirical data giving the possibility of making decisions expeditiously avoiding the development of much more complex system models, both deterministic and stochastic.

Formerly, when data collection was done manually, they were usually considered to be limited and insufficient to develop more sophisticated regression models. Automated monitoring has boosted the possibilities of collecting needed empirical data but, a large amount of different statistical data has created both new opportunities and new challenges in regression modeling. It has turned out the more sophisticated structure of a regression model is, the more difficult it is to ensure its accuracy from the point of view of prediction. Possibly, the main predictive problem while regression modeling is that regression models are traditionally built as models that should suit the best way to all collected data following the principle of optimization. However, the increase in the number of observation data may complicate the execution of important limit restrictions of regression modeling, especially if there is more than one predictor and a researcher has to consider the several explanatory variables and the relationships among them.

Moreover, increasing the dimension of the regression model by introducing the additional explanatory variables cannot usually solve the problem of heteroscedasticity (heterogeneity of variance). An essential additional problem may be the presence of multicollinearity of the explanatory variables, when the coefficients of regression models become unstable to small changes in the data, which violates the stability of solutions. At the best, more sophisticated regression models turn out to be more successful for data interpretation, yet not for prediction.

Nevertheless, the regression models are quite convenient models to solve practical prediction tasks based on monitoring data [4, 5]. They can easily be formalized and adapted to different experimental data. As well as, if it is necessary, regression models admit various modifications depending on the peculiarities and complexities of prediction problems [4, 5]. All this supports their popularity in predicting practice.

III. Fundamentals of Situational and Inductive modeling

A. Situational modeling

The main idea of situational modeling is that the evolution of a dynamic system is modeled in the context of its "movement" through a series of situations resulting from various actions. A complete description of the infinite set of all possible situations of the functioning of the system is replaced by a certain finite set of generalized model situations that reproduce to a certain degree its possible states [6–8]. These model situations (by R. Reiter [7]) do not determine literally appropriate states of the system; they are presumed to show only the history of certain real events as completed sequences of actions in certain periods of time.

Since real situations cannot be described totally, and it is possible to consider only some of their aspects, the non-monotonic output rule is used to describe the evolution of the dynamic system. It is assumed (by J. McCarthy [6]) that on the basis of past facts, by which past model situations are described, and on using some general rules or assumptions, according to which actions and events within those situations take place, it is possible to predict some similar situations that will appear in the future.

Nowadays, situational modeling has become quite popular in economics, medicine, military affairs, forensics, politics, and other similar spheres, as well as especially in artificial intelligence, where the development of a logical approach to modeling the behavior of complex dynamic systems and processes gave impetus to the creation of special situational calculus theory [8].

In many applied studies situational models may be presented by the simplest single-factor regression models where every regression is adapted to an individual model situation connected with a limited time period. In essence, these simplified regression models built from monitoring data are peculiar situational models reflecting different situations in the past. Accordingly, the result of situational modeling is a set of relevant regressions, where each of them is considered to be adequate within only a limited time period. The uncontrollable factors capable affecting the structure and parameters of the regression models are considered as peculiar predictive backgrounds [3, 9, 10] reflecting unknown or undetermined conditions in which the system exists in a certain period of time. Every predictive background may also include unknown predictors. As the constancy of a predictive background relates to a limited period of time, so it identifies the only specific situation and the only specific situational model.

B. Inductive modeling

Inductive modeling is known as an original scientific approach to modeling based on experimental data that was proposed by the Ukrainian scientist O.G. Ivakhnenko. In particular, this approach found their theoretical and practical reflection in the widely famous method to be named the Group Method of Data Handling (GMDH) [11, 12].

It is now used to solve different problems the pattern recognition, the structural-parametric identification of mathematical models of complex systems, the simulation and forecasting of complex processes, and systems based due to experimental data. According to this approach [11, 12], on the basis of available empirical data, a hypothesis about a possible class of models is put forward, the procedure of automatic generation of a set of alternative models belonging to this class is formed (the set may consist of thousands and tens of thousands of models), and the criterion of choosing the best model is established. Since most routine work is transferred to a computer, it is assumed that the impact of human mistakes on the final result of modeling may be minimized.

Nowadays, the GMDH method is also considered as one of the most appropriate and advanced information technologies to obtain knowledge from experimental data, or as one of the most effective methods of intellectual (or intelligent) data analysis [13]. However, the main theoretical and practical result of this approach to modeling based on monitoring data is that the complexity of optimal predictive model depends on the level of uncertainty in the data: the higher this level is (e.g. due to noise or their abundance), the simpler must be the optimal model (with less number of predictors) [14]. Otherwise, the quite successful model for data interpolating can get lost its accuracy and attractiveness in predicting.

According to our approach to predictive modeling, the definition "inductive model" is related to a model obtained from a set of situational models. In other words, the inductive models are models of "levels", which determine the behavior of a dependent variable taking account of results of situational modeling for some fixed values of predictors [3, 9, 10, 14].

Depending on the results of situational modeling and prediction tasks inductive models may have various structures (compositions). They can be presented of a set of trends [9] if the time factor is essential or taken into account, or in the form of a set of regression models [9, 14], if the time factor is not taken into account or it is unessential. More general inductive models may consist of trends and random "balances" after the extraction of these trends, trends, and regression models for random "balances" [14], and so on. Inductive models may be modified in an appropriate way if new data and tendencies appear [9, 14]. As well as, if necessary, time or transportation lags between model variables can be taken into account [3, 9, 14]. In addition, some techniques of adaptive modeling [3–5, 13] can be applied too.

IV. Combined Situational-Inductive Predictive Modeling

The practice of mathematical modeling based on experimental data shows that different models may fit the system being modeled. This modeling is known to relate to solving inverse, not well-posed, or ill-posed problems [15]. It means a lot of quasi valid solutions may exist corresponding to the same inputs (observational data). And, whatever sophisticated a system model based on monitoring data is it will not be fully adequate to reality. In fact, any model built due to empirical data is only partially determined. What is more, not necessarily the sophistication of a mathematical model is the key to its adequacy.

According to L. Ljung [16], such problems should be solved sequentially. At first, the task of the system operational particularities identification is considered. Next, the task of the system's structure identification is posed. Lastly, the task of parameter identification is considered. It is emphasized, to solve the problem of the identification of a system model, it is not always necessary to follow accurate mathematical methods. However, it's important to draw on the whole information including the implicit information that data sets contain.

A. The general algorithm to solve the prediction problem

The general algorithm to solve the prediction problem based on monitoring data by means of combined situational-inductive modeling includes four main stages, namely:

- Preliminary modeling (pre-modeling);
- Fragmentation of monitoring data time series and retrospective situational modeling;
- Inductive modeling based on time series presenting retrospective situational modeling results;
- Prediction: determination of perspective situational models due to inductive models of "levels".

B. Preliminary modeling

Preliminary modeling (or pre-modeling) is an important integral part of modeling based on monitoring data. The pre-modeling applies to the choice of model variables ensuring the adequacy of accepted models, both dependent (resultant, endogenous) variables and explanatory (exogenous) variables (predictors). The result of pre-modeling can be a significant simplification of the model structure and removing some predictors. First of all, it is about the removal of intercorrelated predictors. Removing some of them might lead to a more parsimonious model without compromising the performance of the regression model [4]. In some cases, the regression model simplification in an above-mentioned manner may also be regarded as a practical way to improve it [3–5, 9, 10, 14].

C. Fragmentation of time series and retrospective situational modeling

Big time series can be broken into separate time sections, each of which contains the needed quantity of sample data to provide appropriate constraints and assumptions of regression modeling. The shorter time series of a dependent variable compared to the total duration of the monitoring, the more monotonous and homogeneous it may be, in particular, because of fewer influential factors [3, 9, 14].

Admittedly, there are several essential advantages of removing some predictors in regression modeling due to time series fragmentation [4]. Firstly, as fewer predictors are, so is less the need for model complexity and computational time. Secondly, simplified regression models can largely be improved in their performance; their stability will significantly increase without the problematic explanatory variables. It should also be mentioned that the main idea of regression modeling is that regression occurs when a dependent, endogenous variable depends not only on some independent, exogenous, explanatory variables (predictors) but also on some uncontrolled unknown factors. Yet, all simplifications should be justified in terms of predicting.

It can be assumed that the behavior of the endogenous variable of the regression model over an as short as possible but still sufficient interval of observations to support restricts and assumptions of modeling can depend on the minimum number of predictors. So, ideally, that dependence on only one influential predictor can be established. Undoubtedly, the longer the time interval is the more likelihood of the appearance of other factors that need to be taken into account, in particular by introducing additional explanatory variables, which cannot be further neglected.

Consequently, to perform retrospective situational modeling, sample time series describing the behavior of dependent and independent model variables within separate time periods are prepared. These separate samples (or clusters) have to meet the established criteria of situational modeling adequacy. In particular, it should be taken into account the behavior of key predictors in individual situations, namely:

- Non-stationary oscillations with monotonically increasing trends. Time periods characterized by the relatively slow or relatively rapid monotonous growths of their trends may also be allocated;
- Non-stationary oscillations with monotonically decreasing trends. Time periods characterized by relatively slow or relatively rapid monotonous declines of their trends may also be allocated;
- Random stationary (quasi-stationary) variations.

D. Inductive modeling and prediction

Inductive models built on the results of retrospective situational modeling create the basis for predicting perspective situational models. Usually, inductive models are models of "levels", which determine the behavior of dependent variables for some fixed values of predictors according to situations to possibly appear in the future. According to the adaptation principle, inductive models can extend over both the entire monitoring period and cover some specific time periods. In particular, taking into account the seasonal factor or peculiarities of the behavior of key predictors [9] by appropriate fragmentation of time series allows improving the performance of inductive models to predict relevant perspective situational models. Inductive models may be also modified in an appropriate way depending on the appearance of new data and tendencies.

Actually, the task of situational predictive modeling based on monitoring data is the well-known extrapolation task, which can be defined as the identification of the most probable situational model that will meet some expected situation in the future depending on situations that appeared in the past. Inductive models create an appropriate basis for predicting perspective situational models. Predicted situational models may be unambiguous or the result of their predicting will be a certain set of perspective situational models corresponding to various expected situations in the future.

Situational models of future periods may be established unambiguously if corresponding inductive models are represented by trends with high determinations. When applying composite-type inductive models, the previous situation can uniquely determine the next situational model if there is a transport lag between adjacent situations. It is also assumed that the smaller the duration of a model situation in time is, the more likely the constancy of the relevant predictive background can be expected and conditions of predicting can be supported more accurately. As well as, the results of our test studies [3, 9, 14] show that the more completeness of the monitoring data during short time intervals is the easier to provide the monotony of observations series, the homogeneity of the data samples, and the independence of the endogenous variable from the less significant factors.

It should be also noted the next. If inductive models are built on the basis of situational models of past periods that cover data of similar clusters of monitoring data (for example, taking into account the seasonal factor, etc.), the accuracy of the situational prediction on the basis of inductive models may increase significantly. Situational regression models can also be single-factor models, which determine the dependence of an endogenous variable from one exogenous variable. The most suitable independent variable for such situational models may be an operating variable. However, the main thing is that each of the situational models is better adapted to its particular situation (to its forecast background).

V. Conclusions

Proposed is an approach to predict the behavior of complex dynamic systems based on monitoring data in the framework of combined situational-inductive modeling. The approach realizes the idea of decomposition of complex prediction tasks of the behavior of complex dynamic systems possessing substantial structural and parametric uncertainty.

According to the approach, the predictive modeling based on monitoring data consists of establishing relevant situational models of regression type being adequate in the future within certain limited time periods. The approach allows using simultaneously both the principle of optimization in modeling and the principle of adaptation to situational changes occurring in dynamic systems.

Finally, the proposed approach to predictive modeling agrees with three basic monitoring principles formulated by R.A. Collacott [17]: 1) Consistency and regularity (continuity) of measurements for parameters and characteristics selected for the control; 2) Detection of changes in the behavior of these parameters and characteristics over time; 3) Prediction of future situations taking into account those changes.

References

[1] M. Berthold, Ch. Borgelt, F. Höppner, and F. Klawonn, “Guide to intelligent data analysis: how to intelligently make sense of real data”, Springer-Verlag, London, 2010.

[2] M.A. Wani, and S. Jabin, “Big Data: Issues, Challenges, and Techniques in Business Intelligence”, In: V. Aggarwal, V. Bhatnagar, and D. Mishra (eds), “Big Data Analytics, Advances in Intelligent Systems and Computing”, 654, Springer, Singapore, 2018.

[3] D.V. Stefanyshyn, “A method of forecasting of indexes of dynamic system that evolves slowly, based on time series analysis”, Proc. of 4th ICIM, pp. 221–224, 2013.

[4] M. Kuhn, K. Johnson, “Applied Predictive Modeling”, Spr. Sc.+Bus. Media, NY, 2013.

[5] S. Geisser, “Predictive Inference: An Introduction”, NY, Chapman & Hall, 2016.

[6] J. McCarthy, “Situations, actions, and causal laws, Memo 2”, Stanford Univ-ty Artificial Intelligence Project,1963.

[7] R. Reiter, “Knowledge in Action: Logical Foundations for Specifying and Implementing Dynamical Systems”, MIT Press, 2001.

[8] S.J. Russell, and P. Norvig, “Artificial Intelligence: A Modern Approach”, 3rd ed. Pearson Educ., Inc., Upper Saddle River, New Jersey, 2010.

[9] D.V. Stefanyshyn, “Improving diagnostic models for forecasting the behavior of dams equipped with automated monitoring systems”, MME, 9, pp. 50–61, 2017.

[10] D.V. Stefanyshyn, V.M. Korbutiak, Y.D. Stefanyshyna-Gavryliuk, “Situational predictive modelling of the flood hazard in the Dniester river valley near the town of Halych”, Env. safety and nat. resources, 29, pp. 16–27, 2019.

[11] A.G. Ivakhnenko, “Polinomial theory of complex systems”, IEEE Trans. on Syst. Man and Cyb., 4, pp. 364–378, 1971.

[12] H.R. Madala, and A.G. Ivakhnenko, “Inductive Learning Algorithms for Complex System Modeling”, CRC Press, 1994.

[13] M. Berthold, Ch. Borgelt, F. Höppner, and F. Klawonn, “Guide to intelligent data analysis: how to intelligently make sense of real data”, Springer-Verlag, London, 2010.

[14] D.V. Stefanyshyn, “An approach to predicting based on monitoring data by means of combined situational-inductive modeling (the main idea and expected results)”, MME, 17, pp. 69–81, 2019.

[15] A.N. Tikhonov, and F.P. Vasil’ev, “Methods of solution of ill-posed extremal problems”, Banach Centr. Publs., 3, PWN, Warszawa, pp. 297–342, 1978.

[16] L. Ljung, “System identification: Theory for the user”, Prentice-Hall, Inc., 1987.

[17] R.A. Collacott, “Structural Integrity Monitoring”, Chapman and Hall: London, New York, 1985.