Predictive system for multivariate time series

Jan Kout, Tomas Vlcek, Jiri Klema

Prediction is one of the tasks being solved by people everyday. When encountering a complex and possibly dynamic domain containing a large number of possible dependencies, the predictive task can become too difficult to be solved by usual common sense reasoning. This article introduces Open Prediction System (OPS) – a system that helps to develop predictive models automatically. It represents a general predictive system applicable in a wide range of problem domains. The special attention is paid to the tasks of prediction in multivariate time series motivated by problems common for utility companies that distribute and control the transport of their applicable commodity. General issues of forecasting are discussed together with OPS predictive methodology implemented. Examples of case studies of such system are also included.

1. Time series forecasting

Observing past outcomes of a phenomenon of the interest in order to anticipate the future values is referred to as forecasting or predicting. When a complete and relevant model describing the studied phenomenon is known and all relevant initial conditions are available, then forecasting becomes a trivial task. However, when a model is unknown, incomplete or too complex, then it is typical to build a model that takes into account past values of this phenomenon. In other words, the model is based on "*what the system does*" and not based on "*how or why the system does it*". The past values of the phenomenon over the time form so called a time series of a phenomenon. The prediction of future values of a phenomenon does not have to depend on the past data exclusively. There might also be additional and more relevant information available in the present state of the environment. For example, prediction of water consumption does not depend on past values of water consumption only, but also on other relevant information available, such as outdoor temperature, measurement time, season, and others. Such a multidimensional time series dependency is usually referred to as multivariate time series.

In addition, due to the evolution of rapid processing systems in recent decades the research has been focused on the development of intelligent systems that can design predictive (or generally data) models of phenomena automatically. The problem of empirical modelling is becoming very important in many diverse engineering applications. The performance of a constructed model depends on the quantity as well as on the quality of the observations used during the model construction process. However, in most cases, data is finite and sampled in a non-uniform way. Moreover, due to the highly dimensional nature of many problems, the data is only sparsely distributed in the input space. The learning problem is then considered as a problem of finding a desired dependency using the limited number of observations available. All the mentioned reasons resulted in attention being given to the use of machine learning techniques and statistical predictive techniques for building predictive models.

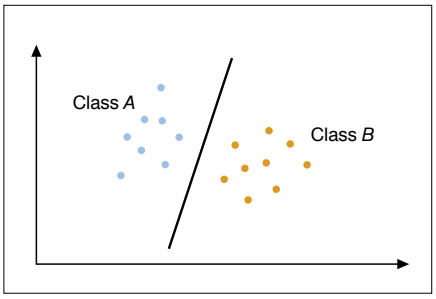

Fig. 1. Classification task – prediction of states

In general, there are two main goals of time series analysis and prediction:

- Determining - identifying the nature of the phenomenon represented by a sequence of observations from the past and, in case of the multivariate time series, by a sequence of other system variables (supporting attributes).
- Forecasting – predicting future values of the phenomenon.

Both of these goals require that the pattern of observed time series data is identified and more or less formally described. Once the pattern is established, it can be interpreted and integrated with other data. Regardless of the depth of the understanding and the validity of the created model of the phenomenon, it is possible to extrapolate the identified pattern to predict future values. As in most other analysis, in time series analysis it is assumed that the data consists of a systematic pattern (usually a set of identifiable components) and random noise (error) that usually makes the pattern difficult to identify.

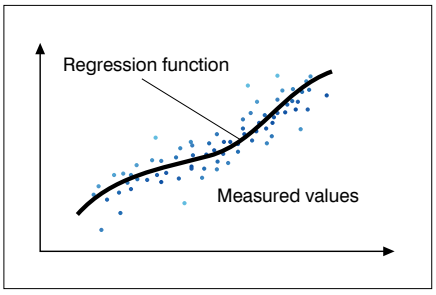

Fig. 2. Regression task – prediction of numerical values

2. State of the art

Forecasting in time series is a common problem. Different approaches have been investigated over the years [1]. In principle, global and local models can be distinguished. In the global approach, only one model is used to characterize the phenomenon under examination. The local approach is based on dividing the data set into smaller sets, and each set is being modelled with a simple model. The global models give generally better results with stationary time series that do not change with time.

The models themselves can be linear or non-linear. The linear models for which the theory is known and many algorithms for model building are available can offer simplicity and comprehensibility. However, since almost all measured phenomena are non-linear, many non-linear methods have become widely utilized. More recently, machine learning techniques have been studied as an alternative to these non-linear model-driven approaches. The process of constructing the relationships between the input and output variables is addressed by certain general-purpose learning algorithms, such as neural networks, support vector machines, case-based reasoning, and others.

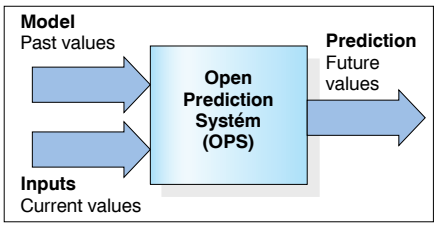

Fig. 3. Predictive modelling in principle

3. Predictive methodology

The designed methodology [2] provides a general predictive methodology allowing predict numerical values (real numbers) as well as discrete values (states, classes) (*fig. 1*, *fig. 2*). The predicted output is based on current data and on a built model. The model is

built by applying the predictive algorithm to the collected historical data. As reliable and bigger collected dataset is available, a more precise prediction is achieved. It is necessary to define potentially relevant data as inputs and the corresponding outputs before the learning phase. Such algorithms where historic inputs and output are given are categorised as a supervised-learning. When providing the current input data, the built model will provide an output.

The participation of a problem domain expert during the development of a representative model is usually an option, since the expert naturally understands which data is potentially important and which is irrelevant to the problem being solved.

To sum up, the development of the prediction system can be divided into three main phases. The first, the problem definition and data acquisition part defines the prediction problem and objectives. This phase also analyzes and transforms the available data. The main goal of the second phase – feasibility study – is to create and validate prediction models using the collected data. The simulation of on-line regime quarantines the stability and reliability of the solution. The last phase integrates the prediction system with an existing information system or a stand alone application is developed.

4. Open Prediction System

The Open Prediction System (OPS) [3] is a software tool employing the technology that finds solutions for prediction and data mining tasks (*fig. 3*). The OPS is used to determine relevant features from collected databases. The features are used to build models in order to detect potential harmful situations, predict future values and trends, as well as to support strategic planning and resource allocation. The prediction engine offers flexible data structures and combines several machine learning algorithms. The building and verification of predictive models are completed effectively and with a minimal effort.

The OPS is useful for feasibility studies to decide about the applicability and suitability of collected data with respect to the defined problem. The models may be customised and configured quickly and effortlessly. The OPS provides many presentation capabilities such as reports or graphs to simplify the evaluation of algorithms. Once the performance of the designed solution meets the requirements of the customer, the customised predictive module(s) may be integrated into the target information or control system of the customer.

The final prediction module typically employs several models based on various machine learning techniques to achieve a higher stability and reliability of the prediction. The final module is based on the configuration resulting from OPS based study/evaluation phase.

The prediction modules are the customer oriented part of OPS. They work completely autonomously within an information system they are designed for. The prediction modules are integrated into an information system that provides prediction modules with all inputs. The modules regularly update the prediction according to a defined interface protocol automatically without any other assistance. Generally, a custom prediction module can be designed as a stand alone application.

There are the following main features of OPS:

- Flexible data pre-processing including data transformation, filtration, and meta-data definition.
- Support of time series data pre-processing, typical for utility companies.
- Availability of wide range of model building algorithms.
- Combination of diverse solution support stability and better performance.
- Fast development cycle.

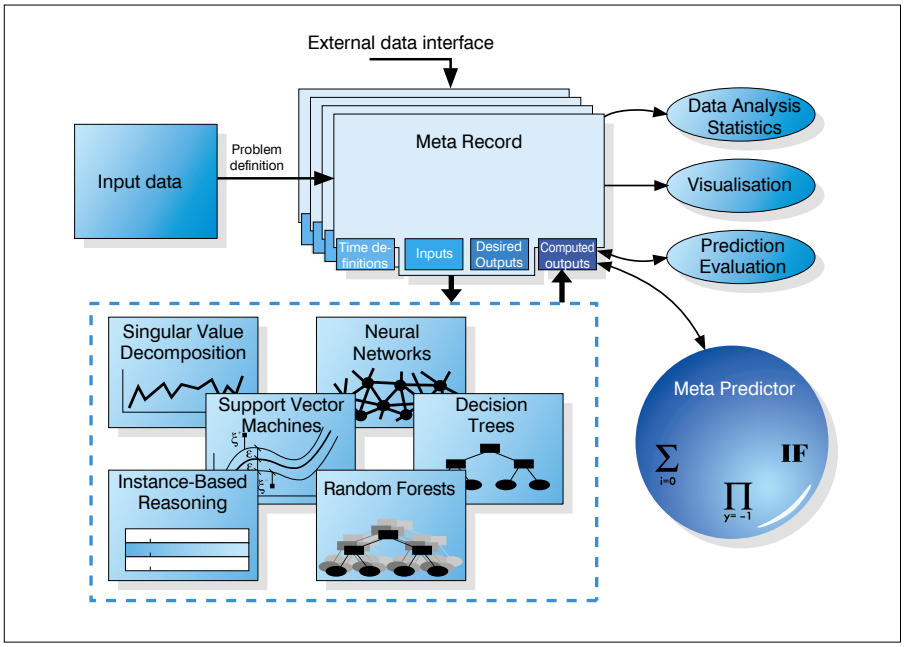

Fig. 4. OPS predictive methodology

5. Case studies

5.1 Utility companies

A typical practical problem of multivariate time series prediction can be connected with networked resource distribution systems. These systems are usually called utility companies. Transported commodity might be for example water, electricity, gas, sewage, and many others. Unsubstantiated decisions on capacity allocation usually lead to substantial problems relating to reputation and to profit. That is why reliable answers to questions such as: "*How much? Where? When?*" provide a company with a considerable competitive advantage. The reputation of a company hinges on delivering high quality goods to the right place at the right time. Utility companies are usually bound by two contracts: one to the global supplier and the other to end-customers. It is important to negotiate cost-effective contracts with both sites. Predictions show where conditions should be pushed and where compromises can be accepted. The OPS has a built-in data processing support that especially suits typical needs of utility companies. The original data set can be processed in order to meet the demands of the prediction task being solved.

The gas distribution companies have enormous demand for gas consumption prediction within a day. If they can foresee (predict) larger than the defined consumption then they have the possibility to react so that they do not exceed the daily consumption limit. When the limit is exceeded, then a local distributor is charged an expensive penalty. A number of local distributors or their customers have the possibilities to switch to an alternative energy source to overcome short periods of peak consumption. The prediction system provides an operator with predictions of gas consumption for the day every hour. The prediction becomes more precise every hour as new data is collected. The predicted future consumption value is passed to the information system and gives the main guideline in timely peak consumption identification.

Heat production and long-distance distribution via heating and distribution of a fluid media (water) greatly depends on various effects. The energy consumption and losses are effected by the input media itself, past and immediate environment conditions, network topology, and types of customers. These relations are far from trivial as they intensely interact and deal with numerous time constants.

Short-time heat consumption prediction can help to optimise the temperature of the distributed media while its long-term counterpart can serve to optimize its storage.

5.2 Processing of biological signals

The implantable pacemakers are accepted as an alternative and effective therapy for preventing sudden cardiac death due to heart arrhythmia. The early detection of heart failures can improve the functionality of the pacemaker. The atrial fibrillation is the most common

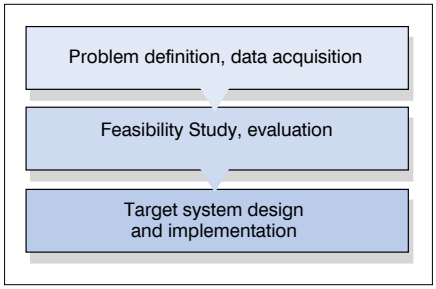

Fig. 5. Prediction system development

supraventricular arrhythmia associated with a considerable risk of morbidity and mortality. The main objective of the study was to discriminate (classify) several types of atrial arrhythmias using the atrial and ventricular intracardiac signals. The OPS was used to analyze and pre-process the raw measured data, to develop a set of definite features, and to construct classifiers.

5.3 Industrial diagnostics

Condition monitoring of induction motor driven pumps involves detecting of commonly encountered faults associated with the pumps before these faults become very serious (resulting e.g., in the shutdown of a process plant). The project was dedicated to timely detection of such faulty conditions. The main goal was to develop a robust decision tool transforming values of the on-line measured signal features (current consumption, frequency, etc.) into a distinct state classification (e.g., normal operation, cavitation, and blockage).

Classification of events in the heating system represents a task focused on the intelligent analysis of specific pump utilization and the optimal pump control. One of the main objectives of the project was to detect events in the heating system and subsequently adjust the pump operating point in order to minimize its power consumption while maximizing flat-dwellers' comfort.

6. Conclusion

All-encompassing advancements in the fields of computers, measurement devices, and telecommunications enable companies and institutions to collect and store their extensive operational data effectively and inexpensively. These data are a potential source of valuable knowledge. Data warehouses and their on-line analytical processing can answer simple and well-posed questions of operative control. Nevertheless, the field of informatics provides a great variety of additional methods to investigate and (more or less) formally express complex associations and patterns. In parallel, problem-oriented predictive models can also be inferred and applied. Such predictive models can be used to improve effectiveness, flexibility and timeliness of control, and production or distribution processes. The OPS represents a general tool that helps in rapid prototyping of problem-oriented predictive systems from the particular operational or other data.

Acknowledgement: This work was partially supported by the Ministry of Education of the Czech Republic under the Project Centre of Applied Cybernetics No.1M6840770004.

References:

[1] WEIGEND, A. – GERSHENFELD, N. (eds): *Time Series Prediction – Forecasting the Future and Understanding the Past*. Addison-Wesley, Reading, Massachusetts, 1993.

[2] KLEMA, J. – KOUT, J. – VEJMELKA, M.: *Predictive System for Multivariate Time Series*. In Cybernetics and Systems 2004. Vienna: Austrian Society for Cybernetics Studies, vol. 1,2, pp. 723—728, 2004

[3] *Open Prediction System – Principles and Methodology,* available on http://www.certicon.cz

Jan Kout, Tomas Vlcek, CertiCon, a. s., Applied research, CAK (kout@certicon.cz)
Jiri Klema, Gerstner laboratory for intelligent decision making and control, department of cybernetics, FEE CTU (klema@labe.felk.cvut.cz)

Jan Kout (1972), received his PhD degree in Artificial Intelligence and Biocybernetics in 2003. He has been working as an applied research manager of CertiCon Corp. He has co-operated on various international projects focused on application of AI techniques in the industry (Vitatron Medical, Grudfos A/S, Rockwell Automation). He also participates on EU Research and Technological Development programs.

Jiri Klema (1971), is an assistant professor at the Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague. He obtained his PhD degree in Artificial Intelligence and Biocybernetics in 2002. His research interests are machine learning and data mining, in particular he is concerned with mining of sequential and temporal data. He participated in multiple application projects solved in co-operation with Rockwell Automation, Grundfos A/S, Abbott Laboratories, Health Data Research, and IKEM Praha.

Tomas Vlcek (1964), graduated in Technical Cybernetics (1988) and received PhD in Artificial Intelligence and Biocybernetics in 1993. Since 1989 he has been working as an assistant professor at Czech Technical University in Prague and since 1999 as a managing director of CertiCon Corp. He has been responsible for a number of international projects in the field of artificial intelligence and decision support systems, both industrial and within EU Research and Technological Development programs.