Predictive System for Multivariate Time Series

Jiří Kléma
The Gerstner Laboratory for Intelligent Decision Making and Control
Faculty of Electrical Engineering
Department of Cybernetics
Technická 2, Prague 6, 166 27
email: klema@labe.felk.cvut.cz

Jan Kout, Martin Vejmelka
CertiCon a.s.
Applied Research Department
Václavská 12
Prague 2, 120 00
Czech Republic
email: {kout, vejmelka}@certicon.cz

Abstract

The paper is focused on the analysis and design of multivariate time series prediction systems. It addresses mainly practical issues, the main contribution is the developed and implemented conceptual predictive methodology. It is based on designed data management structures that define basic data flow. Despite the fact that the methodology is inspired by problems common for utility companies that distribute and control the transport of their applicable commodity, it may be considered as a general methodology. Currently, the predictive methodology combines several prediction techniques, such as regression by means of singular value decomposition, support vector machines and neural networks. Data management structures are open to other predictive algorithms as well. The methodology is implemented in the form of a software tool. It is verified on a real-life prediction task—prediction of the daily gas consumption of regional gas utility companies.

1 Introduction

Observing past outcomes of a phenomenon of the interest in order to anticipate the future values is referred to as forecasting or predicting. If a complete and relevant model describing the studied phenomenon is known and if all relevant initial conditions are available then forecasting becomes a trivial task. However, when a model is unknown, incomplete or too complex, a typical alternative way is 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 but not based on *how* and *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 doesn't have to depend on the past data exclusively. There might be also 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 modeling 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 dependence 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 prediction techniques for building predictive models.

The paper deals with the study and analysis of networked resource distribution systems. These systems are usually called *utility companies*. Transported *commodity* might be for example water, electricity, gas, sewage and many others. The paper mainly focuses on the study and analysis of commodity consumptions within utility companies. The main motivation of the study is to develop a methodology to predict the amount/quantity or another parameter of the interest (for example maximal load) of a transported commodity in order to provide an operator of a utility company with relevant information and proper decision support. In general, there are two main goals of time series analysis and prediction:

- Identifying the nature of the phenomenon represented by the sequence of observations in the past and by the sequence of other system variables (supporting attributes) in case of the multivariate time series and
- 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 analyses, 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.

2 State of The Art in Time Series Prediction

Forecasting in time series is a common problem. Different approaches have been investigated in time series prediction over the years [Weigend and Gershenfeld, 1993]. The methods can be generally divided into global and local models.
In global model approach only one model is used to characterize the phenomenon under examination. The local models are based on dividing the data set into smaller sets, each being modeled with a simple model. The global models give generally better results with stationary time series. Stationary series are series that do not change with time.

The first approaches were devoted to linear models for which the theory is known and many algorithms for model building are available. The most used linear regression methods have been the autoregressive (AR) and autoregressive moving average (ARMA) models [Box *et al.*, 1994]. An example of more complex regression method is the multivariate adaptive regression splines [Friedman, 1991]. Using a statistical approach, the integrated autoregressive moving average (ARIMA) methodology has been developed [Box *et al.*, 1994]. The methodology is useful for fitting a class of linear time series models. Statisticians in a number of ways have addressed the restriction of linearity in the Box-Jenkins approach. Robust versions of various ARIMA models have been developed. In addition, a large number of nonlinear time series models is available. The stochastic approach to nonlinear time series that can fit nonlinear models to time series data was developed [Tong, 1990].

However, since almost all measured phenomena are nonlinear, the linear modeling methods are often inappropriate to describe their behavior. This is the reason why many nonlinear methods have become widely utilized. More recently, machine learning techniques (mainly neural networks) have been studied as an alternative to these nonlinear model-driven approaches. The process of constructing the relationships between the inputs and output variables is addressed by certain general-purpose 'learning' algorithms.

Neural networks represent an attractive approach for time series prediction problems [Casdagli and Eubank, 1992], [Drossu and Obradovic, 1996], [Thiesing and Vornberger, 1997] and [Street, 1998]. Neural networks can construct approximations for unknown function by learning from examples. Multi-layered perception (MLP) using the backpropagation algorithm that is described later is the most popular approach. Because of their characteristics, neural networks belong to the data-driven approach, i.e. the analysis depends on the available data, with a little a priori rationalization about relationships between variables and about the models. Some drawbacks to the practical use of neural networks are the possibly long time consumed in the modeling process and the large amount of data required by the present neural networks technology. Speed-up is being achieved due to the impressive progress in increasing the clock rate of present processors. The demands on the number of observations remains however a hard open problem. One cause of both problems is the lack of definite generic methodology that could be used to design a small structure. Most of the present methodologies use networks, with a large number of parameters ("weights"). This means lengthy computations to set their values and a requirement for many observations. Unfortunately, in practice, a model's parameters must be estimated quickly and just a small amount of data is available. Moreover, part of the available data should be kept for the validation and for performance-evaluation procedures. Radial Basis Function (RBF) that also belongs to a group of neural networks is also a popular predictive approach. To sum up, the main problem with neural network is the design of the structure of the network and effectiveness of the learning phase.

Another popular approach is Cased Based Reasoning (CBR) [Klema, 2002]. The main idea of the case based reasoning is in finding similar cases in the past. A new approach called Support Vector Machines (SVM) has become a subject of intensive study [Vapnik, 1995]. There are also first approaches to employ Support Vector Machines for prediction tasks [Muller *et al.*, 1997]. Quite a different approach is based on expert system building from the data [An *et al.*, 1997]. The system is operated by probabilistic decision rules that have been derived from data using Knowledge Discovery in Databases (KDD) approach.

3 Goals of The Paper

This paper is dedicated to functional employment of the machine learning and regression algorithms on the multivariate regression type of prediction problems. The main focus is spent on the analysis of the general approach to similar problem domains as well as on the utilization of real data in the designed and developed predictive methodology. The combination of more methods on the same problem is also discussed.

The utility companies usually collect large amount of data only a part of which is useful for the prediction purposes. Data has to be preprocessed in order to define the appropriate, suitable and relevant input attributes and consequently to reduce the amount of data. The original data might be also not directly useful for the prediction purposes since they need to be transformed from the temporal point of view, e.g., the original time scale does not fit the requirements of the task being solved. The needs for the data preprocessing and processing lead to the development of the open data (pre)processing methodology that can process the input data structures into a problem oriented data structure (it is later called meta-record). Such a problem oriented data structure can reflect particular needs of the problem that is being solved.

There are many predictive algorithms available. The motivation is always to find the best one for the particular needs. Every predictive algorithm requires a different method of data preprocessing. That is why a special algorithm related data structure was designed as the last data processing step. The paper discusses algorithms that are based on statistical model construction as well as algorithms that are based on the machine learning principles. The algorithms are compared not only from the performance perspective but also from the time strenuousness point of view. The initial configurations and the maintenance of the algorithms are also important issues.

One of the main motivations of the predictive methodology is to integrate the data preprocessing and processing (data management) together with several predictive algorithms. Such a system can offer a flexible solution for multivariate regression time series predictions. The final solution is always defined as a combination of individual predictors that are integrated together.

The expected novelty of the presented approach is in:

- Organization of data into specific data management structures permitting transformations in the time scale, data compression, data filtering and custom data processing.
- Design of techniques of flexible transformations of data structures with respect to the prediction tasks at hand.
- Integration of data management structure with several predictive algorithms into the predictive methodology process.
- Application of a weighted combination of predictors, the combination is carried out by a meta-predictor.

4 Data Management

The various tools available differ in their cognitive basis, the expressiveness of the language in which the resulting knowledge is represented and implementation details, including the assumed format of the input data. The gap between the format of data as stored in the data sources and that required by newly developed data mining algorithms must be bridged before these tools can be leveraged to their full potential. Transforming this data into a format appropriate for mining is a key (and often very time consuming) phase of the data mining process called data preprocessing. Practical experience from data mining projects [Pyle, 1999] confirms that more manual and routine effort is usually spent on data management and preprocessing than on modeling itself. The initial phases of the data mining can cost up to 80 % time spent on the project.

There are two main objectives for the preparation of data: organize the data into a standard form that is ready for processing by the prediction algorithms and prepare features that lead to the best predictive performance. It seems to be comparatively straightforward to specify the standard form of data organization. On the other hand, it is much harder to generalize concepts for composing the most predictive features. It is usually supported by domain knowledge.

4.1 Data Structures

The main issue of data processing is data collection and problem definition with respect to collected input data. A multivariate time series is a set of observations on the objects over the time period. These observations are usually assumed to be performed at regular intervals—sampled by time frames. Time frame is the smallest time period that is considered within the task definition. The main feature of time series data is that data is correlated.

The main interest of this chapter is to specify available data structures describing possible tasks to be solved and to predefine a scale of possible desired inputs and outputs for an individual predictor. The overall data flow is shown in Figure 1.

Figure 1 Data flow

Three different data structures are recognized in the system:

- *raw input data* representing system inputs,
- *meta-record* representing the problem definition and
- *records* representing data sets adjusted with respect to needs of individual predictors.

Raw input data file consists of the smallest input units corresponding to the shortest regarded time period (frames), the smallest unit with respect to the data sampling and storage. The purpose of the raw input data is to provide a representation that is identical or is very close to the data that is sampled from the system. Raw input data are represented as a rectangular matrix, columns correspond to individual attributes, rows contain individual cases.

Meta-record represents data structure corresponding to the main transformation step between the raw input data and the data provided to individual predictors. They are introduced by the application of the *transformation functions* to the raw input data. The meta-record defines structure of a feature vector which transformed subset appears on the input of the particular predictor. Meta-record represents a sort of *feature-results pool*. This pool can include all the features as well as partial results available at the given moment. The meta-record can be understood as means for data storage, integration of inputs and outputs. It provides a niche for several processing iterations when the past prediction output is used on the input of the following predictor. It can be used for visualization of data and results as well as for data analysis. Generally, there can be more meta-records derived from the same raw data file.

There are four basic data types distinguished among meta-record data: time related attributes, attributes representing problem inputs (based on raw data), attributes representing problem desired outputs (again based on raw data) and attributes representing outputs of the individual predictors (generated by the predictors). *Time related attributes* define the sequence of a time series. Time identification can consist of more than one attribute (for example day, month, year). There is always an *atomic time attribute* that corresponds to the smallest time unit. Other time attributes are sequentially based on the atomic time attribute (for example a day consists of 24 hours). According to time attributes, missing cases are also recognized. If time attributes are missing or they are not important then the default time series can be defined as a sequence of integer numbers.

Record selects some of the features available in the project meta-record, additional predictor-related transformations can be applied. Normalization is a typical example. There are no time attributes allowed in the records. The output of a predictor is automatically passed to meta-record via the inverse transformation function (for validation, comparison and future utilization by the meta-predictor).

4.2 Transformation Functions

The goal of transformation functions is to guarantee a comfortable and fast transition between the raw input data representation and the meta-record representation as well as between the meta-record representation and the record representation. Transformation functions enable to integrate the data from several consecutive frames into a single record. They are also used to combine several attributes into a new derived attribute. Transformation functions are also able to define time sequence within the meta-record definition.

Time transformation defines namely the basic time period that is represented by a single meta-record. They are used to transform raw input data time attributes into meta-record problem oriented time attributes. The *modified time* definition is introduced via new time attributes MT_1, …, MT_k. They are made as selection or simple transformation of the original time attributes T_1, …, T_i. The transformations can change both time period and scale (granularity).

The second type of the transformation functions implements *transformation of valued data*. The transformation is either applied between the originally sampled variables and the meta-record attributes (denoted as raw input data → meta-record transformation) or between the meta-record and individual predictors (denoted as meta-record → record transformation). The proposed methodology implements a rich set of unary (scaling, power, etc.) binary (addition, percentage difference, etc.) and vector (maximum, sum, etc.) transformations whose description is beyond the scope of this paper. Their detailed description can be found in [Kout, 2003]. The Figure 2 illustrates the idea behind of time and attribute transformations—time scale is changed during the time transformation, vector attribute transformations with different *baselines* are also applied.

4.3 Data Filters

Data filters are designed in order to extract automatically only a specific part of all the available data. Their application is namely expected when transforming meta-record structure into the record structure. Each filter is defined by the parameters *field* (attribute), *condition* (is more than, is between, is outside of, etc.) and *values* (parameter values for the condition). It is possible to combine more filters together via logical connections – conjunction (AND), disjunction (OR) and negation (NOT). For example, the filter *(Temp < 10 AND Season <> 2, 3)* will select all records that have its temperature lower than 10 and season from 2 to 3. Filters are mainly used to prepare local data for a predictor that can create a local model.

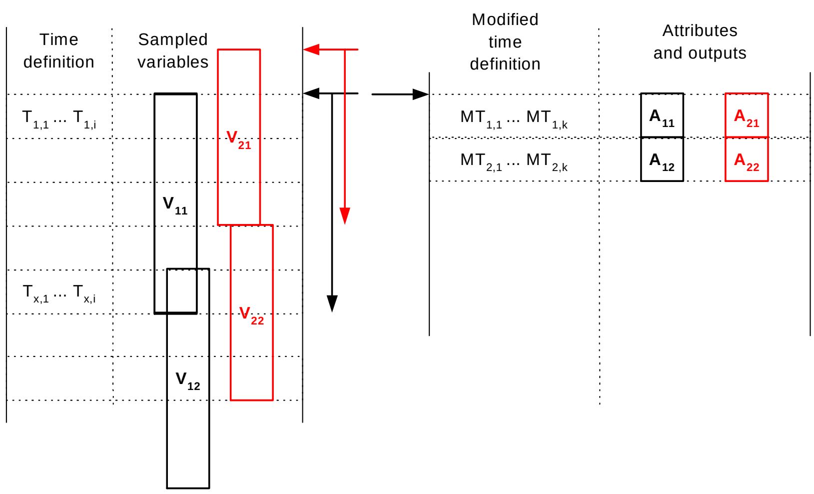

Figure 2 Time and value transformations between raw data (left side) and meta-record (right side)

4.4 Model Ensembles

The instability of a prediction method refers to the sensitivity of the final model to small changes in the training set with regard to the prediction accuracy of the model. Small changes in the learning set may lead to large changes in the prediction performance. An approach to overcome the instability problem is to construct multiple models and combine them to make up the final model. There exist three commonly used methods for constructing ensembles of classifiers 0: boosting, bagging and stacking.

Generally, the regression tasks represent a task group suitable for similar combinations. They can benefit from three basic effects brought by this combination. The first one is compensation of errors having different signs. Having two or more predictive models, the final error of the meta-predictor can be decreased by mutual compensation of the fractional errors of the individual predictive models (e.g., the first module predicts 10% less than the real value is, the second one 6% more – the final error is only 2% provided that final prediction makes average of the fractional predictions). The second effect brings limitation of extreme prediction errors (e.g., the first module predicts 10% less than the real value is, the second one 50% less – the final error is 30% provided that final prediction makes average of the fractional predictions, which means that the extreme error 50% is avoided in the least). The third positive effect asks for a more profound construction of the meta-predictor. The meta-predictor that evaluates comparative advantages of the individual predictors can employ their comparative advantages in the situations that bring the best benefit to the performance of the overall system. This approach relies on the construction of local models of individual predictors.

The overall solution proposed and implemented here relies on a combination of multiple predictors with different types of learning strategies. It means that the heterogeneous learning algorithms represent different types of modeling functions. The meta-predictor is the predictor that integrates all individual results into one result that corresponds to the output of the overall system. First, it can be implemented directly within the meta-record as simple transformation function dealing with the outputs of the individual predictors (averaging, weighted average, combinations of weighting and filtering). Second, an arbitrary predictor can be selected as meta-learner dealing with the outputs of the individual predictors (this scheme represents equivalent of stacking of classifiers).

The overall system currently consists of several relatively independent predictors. Each of sub-units put in use a different predictor – the neural networks (NN), the linear regression (REG), the support vector machines (SVM). The case-based reasoning (CBR) predictor is being implemented now. As the predictive methodology is the main issue of the paper, the individual predictors are not discussed here, details can be found in [Kout, 2003], [Klema, 2002].

5 Predictive Methodology

The data management methodology and predictive algorithms are integrated into the final solution. This solution can be generally applied to those problems whose data contain a time sequence. Original data carrying an original time sequence are used as inputs for the definition of different prediction tasks. Original data is also used to redefine a time sequence into a more descriptive time sequence.

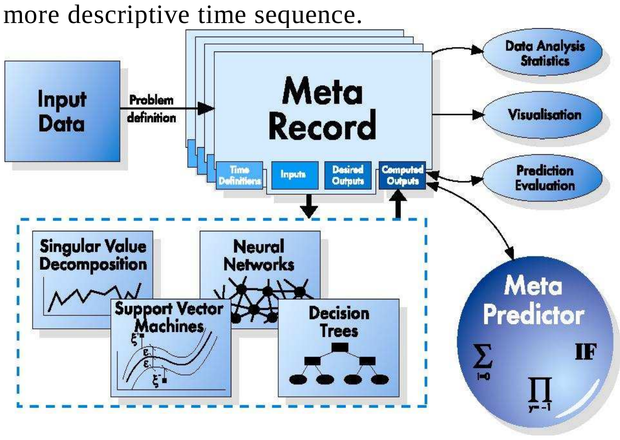

Figure 3 - The final predictive methodology

The predictive methodology is described as a process that is followed when a new particular predictive task is being solved. The description takes care mainly of data processing and the configuration of individual predictors. Each predictor is in some sense unique and that is why special attention should be dedicated to every predictor individually. The predictive process can be summarized into the following steps: definition of inputs, problem definition, configuration of predictors and meta-predictor configuration. The overall methodology is illustrated in Figure 3.

The main advantage of the proposed predictive methodology is in:

- applicability of the same input data structure to various problems,
- redefinition of the time sequence according to predictive needs of a utility company to solve a problem,
- flexible attribute transformations and data filtering,
- reusability of the configured solution structures and predictors for a different prediction task.

6 Conclusion

Precise, relevant and up-to-date prediction is becoming more and more important in many different areas. It can help make appropriate decisions in order to minimize harmful effects of possible problems as well as to maximize profit and customers' satisfaction. The paper is focused on the prediction of multivariate time series. Originally, it was motivated by problems that are common for utility companies. The developed three-level data management framework (raw data, meta-record, records) was tailored to the prediction of various distributed commodities such as water, gas, electricity, sewage and others. Nevertheless, the methodology can be generally employed for any time series regression problem or regression problems in general.

The described data management brings the following benefits: data compression, data filtering, problem definition separated from preprocessing of individual predictors, special built-in transformations or definition of meta-predictor. Data compression is achieved by the transformation of the raw input data file into the meta-record data structure. There are two main reasons for doing that. First, the original data structure can be sampled with higher frequency than is desired. The data management would be complicated to process the prediction on the original data. Moreover, the number of cases of problem-oriented data can be dramatically reduced. This can considerably aid in understanding the data. The second issue of the data compression is that some attributes can be joined together or omitted. The original data size can be radically reduced. Data filtering can help to get rid of suspicious values that are likely to be incorrect. Besides it helps to focus on the most relevant cases in various phases of modeling. These local models can definitely improve system performance as well. Moreover the prediction system can be tuned for the most critical situations individually. Special built-in transformations speed up the preprocessing and improve the data processing scheme (e.g., there are added special built-in features that support calendar-based attributes including their irregularities, arbitrary time aggregates can be derived immediately).

The meta-predictor brings more robustness and precision to the prediction process. Experimental results in several practical domains support the theoretical assumption of the effectiveness of a combination of predictors for regression tasks [Kout and Klema, 1999], [Klema, 2002]. If compared with classification tasks, where there are several disjunctive classes then it is essential that in the case of regression tasks it is possible to employ mathematical operations on numerical output values which can combine results of several predictors. A relatively simple meta-predictor improves the system performance in terms of maximum prediction error and mean absolute percentage error. The system architecture enables simple aggregations of predictors as well as well-known approaches such as stacking.

Acknowledgments

The presented work was supported by the grant of the Czech Ministry of Education, MSM 212300013.

References

[An *et al.*, 1997] An A., Chan C., Shan N., Cercone N., Ziarko W.: *Applying Knowledge Discovery to Predict Water-Supply Consumption*, University of Regina, Saskatchewan, Canada, In IEEE Expert: Intelligent Systems and Their Applications, 1997.

[Box *et al.*, 1994] Box G. E. P., Jenkins G. M., Reinsel G.C.: *Time Series Analysis, Forecasting and Control*, Ed. Englewood Cliffs: Prentice Hall, 3rd edition, 1994.

[Casdagli and Eubank, 1992] Casdagli M., Eubank S.: *Nonlinear Modelling and Forecasting*, Reading MA, Addison-Wesley, 1992.

[Drossu and Obradovic, 1996] Drossu R., Obradovic Z.: *Rapid Design of Neural Networks for Time Series Prediction*, Washington State University, IEEE Computational Science and Engineering, 1996.

[Friedman, 1991] Friedman J.: *Multivariate adaptive regression splines*, Annals of Statistics, 19: 1-142, 1991.

[Klema, 2002] Klema J.: *Prototype Applications of Instance-Based Reasoning*, Doctoral Thesis, Prague, Czech Technical University FEE, Department of Cybernetics, The Gerstner Laboratory, 2002.

[Kout and Klema, 1999] Kout J., Klema J.: *Prediction of Gas Consumption*, In: Systems Integration '99, Prague University of Economy, 1999, pp. 119-128. ISBN 80-7079-059-8, 1999.

[Kout, 2003] Kout, J.: *Predictive Methodology for Utility Companies.* [PhD Thesis]. Prague: CTU FEE, Department of Cybernetics, The Gerstner Lab, 106 p., 2003.

[Muller *et al.*, 1997] Muller K., Smola A., Ratsch G., Scholkopf B., Kohlmorgen J., Vapnik V.: *Predicting Time Series with Support Vector Machines*, ICANN'97, pages 999-1004. Springer, 1997.

[Ng, 1996] Ng K.: *An experimental comparison of stacking, bagging, boosting for combining ensembles of radial bias classifiers*, Term Project 9.520, Spoken Language Systems Group, Laboratory for Computer Science, Massachusetts Institute of Technology, 1996.

[Pyle, 1999] Pyle D.: *Data Preparation for Data Mining*. Morgan Kaufmann, California, 1999.

[Street, 1998] Street N.: *A Neural Network Model for Prognostic Prediction*, Oklahoma State University, In Proceedings of the Fifteenth International Conference on Machine Learning, pages 540-546, San Francisco, 1998.

[Thiesing and Vornberger, 1997] Thiesing F., Vornberger O.: *Forecasting Sale Using Neural Network*, University of Osnabrueck, Computational Intelligence, Bernd Reusch (ed.), Proc. International Conference, 5th Fuzzy Days, Dortmund, Germany, April 28-30, 1997.

[Tong, 1990] Tong H.: *Non-linear Time Series*, Oxford University Press, 1990.

[Vapnik, 1995] Vapnik V.: *The Nature of Statistical Learning Theory*, Springer NY, 1995.

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