Chapter 17

Time series prediction

Amaury Lendasse, Francesco Corona, Federico Montesino-Pouzols, Patrick Bas, Antti Sorjamaa, Mark van Heeswijk, Laura Kainulainen, Eric Severin, Elia Liitiäinen, Tuomas Kärnä, Yu Qi, Yoan Miche, Emil Eirola, Dušan Sovilj, Olli Simula

17.1 Introduction

Amaury Lendasse

What is Time series prediction? Time series prediction (TSP) is a challenge in many fields. In finance, experts forecast stock exchange courses or stock market indices; data processing specialists forecast the flow of information on their networks; producers of electricity forecast the load of the following day. The common point to their problems is the following: how can one analyze and use the past to predict the future? Many techniques exist: linear methods such as ARX, ARMA, etc., and nonlinear ones such as artificial neural networks. In general, these methods try to build a model of the process. The model is then used on the last values of the series to predict the future values. The common difficulty to all the methods is the determination of sufficient and necessary information for an accurate prediction.

A new challenge in the field of time series prediction is the Long-Term Prediction: several steps ahead have to be predicted. Long-Term Prediction has to face growing uncertainties arising from various sources, for instance, accumulation of errors and the lack of information.

Our contributions in TSP research. The TSP group is a new research group. It has been created in 2004. A notable achievement has been the organization of the first European Symposium on Time Series Prediction (ESTSP'08) on September 2008 in Porvoo [1]. For this symposium, a time series competition has been organized and a benchmark has been created.

In the reporting period 2006 - 2007, TSP research has been established as a new project in the laboratory. Nevertheless, TSP research has already been extended to a new direction: Chemoinformatics.

This Chapter starts by introducing some theoretical advances undertaken during the reporting period, including the presentation of the ESTSP'08 competition. Also the problem of input selection for TSP is reported. The applications range includes Chemoinformatics.

In 2010, as we believe that i'The Times They Are A-Changin", the TSP group will evolve and will become the "Environmental and Industrial Machine Learning Group".

17.2 European Symposium on Time Series Prediction

Amaury Lendasse, Olli Simula and Timo Honkela

Introduction

Time series forecasting is a challenge in many fields. In finance, one forecasts stock exchange courses or stock market indices; data processing specialists forecast the flow of information on their networks; producers of electricity forecast the load of the following day. The common point to their problems is the following: how can one analyze and use the past to predict the future? Many techniques exist including linear methods such as ARX or ARMA, and nonlinear ones such as the ones used in the area of machine learning. In general, these methods try to build a model of the process that is to be predicted. The model is then used on the last values of the series to predict future ones. The common difficulty to all methods is the determination of sufficient and necessary information for a good prediction. If the information is insufficient, the forecasting will be poor. On the contrary, if information is useless or redundant, modeling will be difficult or even skewed. In parallel with this determination, a suitable prediction model has to be selected. In order to compare different prediction methods several competitions have been organized, for example, the Santa Fe Competition, the CATS Benchmark Competition and the ESTSP'07 Competition.

After the competitions, their results have been published and the time series have become widely used benchmarks. The goal of these competitions is the prediction of the subsequent values of a given time series (3 to 100 values to predict). Unfortunately, the long-term prediction of time series is a very difficult task. Furthermore, after the publication of results, the real values that had to be predicted are also published. Thereafter, it becomes more difficult to trust in new results that are published: knowing the results of a challenge may lead, even unconsciously, to bias the selection of model; some speak about "data snooping". It becomes therefore more difficult to assess newly developed methods, and new competitions have to be organized.

This text is based on papers presented at the joined ESTSP'08 (European Symposium on Time Series Prediction) [1] and AKRR'08 (Adaptive Knowledge Representation and Reasoning) conferences. This shared event took place in Porvoo, Finland, from 17th to 19th of September, 2008. The goal of joining these conferences was to create an interdisciplinary forum for researchers who may widen their scope of attention beyond the usual scope of research. The crossfertilization took place, for instance, by offering the attendees shared keynote talks. Prof. Marie Cottrell (Paris University 1) gave a talk on data analysis using Self-Organizing Maps. Prof. José Príncipe (University of Florida) described information theoretic learning and kernel methods. Dr. Harri Valpola (Helsinki University of Technology) explained how to extract abstract concepts from raw data using statistical machine learning methods. One specific shared theme of interest was anticipation, i.e., how an agent makes decisions based on predictions, expectations, or beliefs about the future. Anticipation is an important concept when complex natural cognitive systems are considered.

ESTSP'08 Competition

The goal of the ESTSP'08 competition was to predict the future of three very different Time Series[1]. Firstly, the length and the sampling period of the time series are very different. Secondly, the origin of each time series varies. The data and the origins, i.e., environment, electric load, and internet traffic, are described below in more detail. In order to provide the participants an equal opportunity for success, the origins of the three time series were kept secret until the end of the competition.

Data sets

Chemical descriptors of environmental condition

This series is part of a multidimensional time series of monthly averages of different chemical descriptors of a certain area of the Baltic Sea. The series is made of 354 samples and spans for 29.5 years. This competition data set is shown in Figure 17.1. For this time series, the goal was to predict the next 18 values of the third time series, using the two other one as exogenous variables.

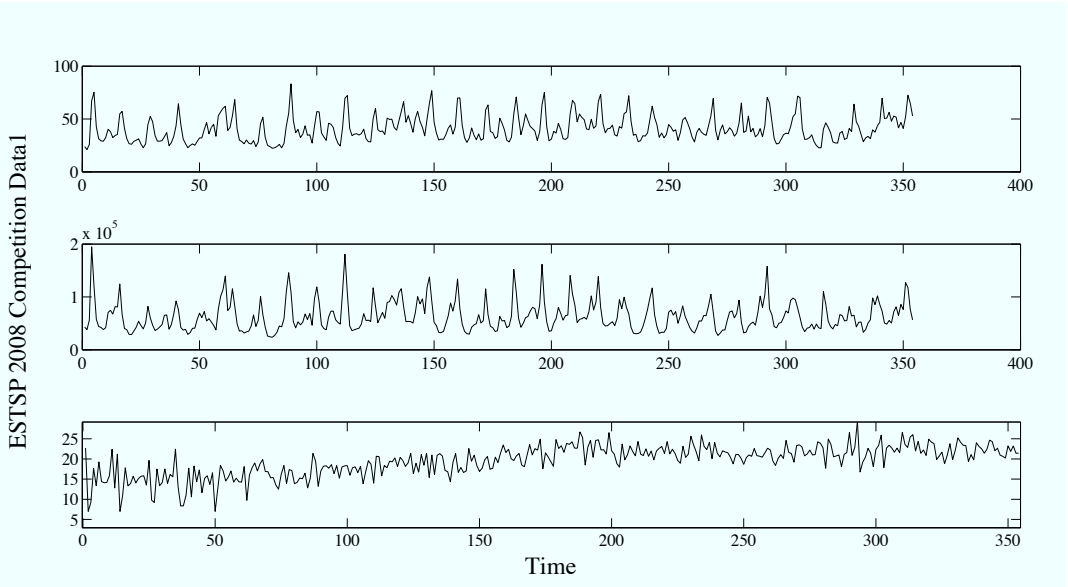

Figure 17.1: ESTSP 2008 competition data 1.

Traffic in a data network

The second dataset from the ESTSP 2008 competition is a univariate time series consisting of 1300 samples that describe the daily average amount of traffic in a data network. The competition data set 2 is shown in Figure 17.2. For this time series, the goal was to predict the next 100 values of the time series.

Electric load

The third dataset was a univariate time series consisting of 31614 samples that describe the daily average amount of eletric load. The competition data set 2 is shown in Figure 17.3. For this time series, the goal is the prediction of the next 200 values of the time series.

[1]The data sets can be downloaded from
http://www.cis.hut.fi/projects/tsp/index.php?page=timeseries.

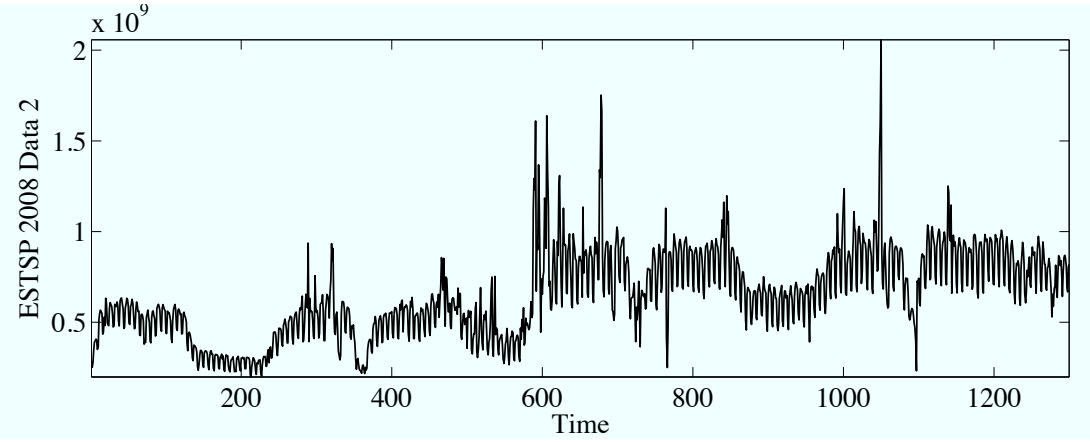

Figure 17.2: ESTSP 2008 competition data 2.

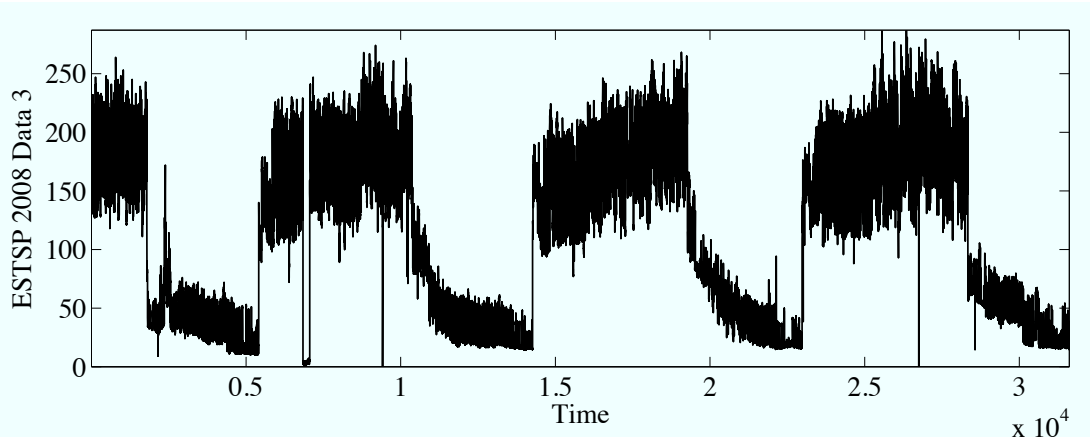

Figure 17.3: ESTSP 2008 competition data 3.

Results

Twenty sets of predictions have been submitted to the competition. The results in Table 17.1 present the Normalized Test Mean Squared Error for the 3 predictions respectively. We present only the results of the participants that agreed to have their results published. The winners of the competition were Rubio, Herrera, Pomares, Rojas and Guillen.

Summary of the Best Papers

The papers can be classified in 3 distinct categories:

1. The authors that participated to the competition.

	Kourentzes	Bontempi	Olteanu	Wyffels	Espinoza	Adeodato	Rubio	Montesino
Data 1	0.07	0.12	0.19	0.157	0.112	0.151	0.079	0.16
Data 2	0.212	0.431	0.359	0.529	0.266	0.49	0.208	0.4
Data 3	0.25	1.802	1.655	1.582	0.464	1.611	0.036	1.344
Total	0.178	0.785	0.735	0.756	0.281	0.751	0.107	0.635

Table 17.1: Competition Results: Test NMSE for Each Data Set.

2. The papers that presents new methods for the analysis and/or prediction of Time Series but did not participate in the competition.

3. The papers that participated in the ESTSP08-AKRR'08 Special Session on Prediction for Finance organized by Prof. Eric Séverin.

Competition Papers

Crone and Kourentzes propose a data driven, fully automated methodology to specify multilayer perceptrons for time series prediction using a combination of iterative (neural network) filters and wrappers. Their approach is capable of identifying unknown time series frequencies, multiple overlying seasonality, and additional relevant features without human expert intervention. The approach has shown promising performance in forecasting by ranking second in the ESTSP competition.

Pouzols and Barriga deal with an automatic methodology for clustering-based fuzzy inference models. A number of clustering methods are compared and an extension of Improved Clustering for Function Approximation is proposed. The appraoch yields compact models and its accuracy and speed compare favorably against MLP, LS-SVM and ELM models for a diverse set of time series benchmarks.

Ben Taieb, Sorjamaa and Bontempi present a new multiple-output approaches for Multi-Step-Ahead Time Series Forecasting and compares it to state-of-the-art approaches. The extensive validation made with the series of the NN3 competition shows that the multiple-output paradigm is very promising and able to outperform conventional techniques.

Reservoir Computing has been shown to perform well in chaotic time series prediction. **Wyffels and Schrauwen** extend these results by a comparison of multiple Reservoir Computing strategies for time series prediction (including research on regularization, influence of reservoir size and decomposition) in the domain of noisy, seasonal time series prediction for industrial purposes. They compare their approach to standard approaches such as ARIMA modeling and NAR modeling using LS-SVMs.

Rubio, Herrera, Pomares, Rojas and Guillen present a kernelized version of the weighted k-nearest neighbours method (KWKNN) for regression problems and address the creation of specific-to-problem kernels for time series data. This unified framework for kernel and k-nearest neighbours methods allows for a comparison of KWKNN with LSSVM using time series prediction examples with interesting results. Additionally, a parallel implementation of KWKNN, developed in order to speed up the method and make it practical for large datasets, is proposed and applied to a large scale problem.

General Papers

Sovilj, Sorjamaa, Yu, Miche and Séverin present a methodology for long-term time series prediction that can also be applied to standard regression tasks. The methodology consists of two main steps: (1) input variable scaling or projection with Delta Test, optimized with Genetic Algorithm, and (2) prediction on the projected data using two models, Optimally-Pruned Extreme Learning Machine and Optimally-Pruned k-Nearest Neighbors. The methodology is tested on two time series prediction tasks and one financial regression problem.

Nybo provides an applied perspective from the petroleum industry. Normally conservative, this industry nonetheless shows an increasing interest in machine learning and data mining. The paper gives a taste of the new opportunities in this industry and goes

on to show how a successful choice of machine learning algorithms becomes governed by the industry's work processes and the user's behavioural mode.

Souza and Barreto provide a comprehensive performance evaluation of the use of vector quantization (VQ) algorithms to building local models for inverse system identification. Statistical hypothesis testing is carried out through the Kolmogorov-Smirnov test in order to study the influence of the VQ algorithms on the performances of the local models. Tests on four benchmarking input-output time series reveal that the resulting local models achieve performances superior to standard global MLP-based model.

Lemke and Gabrys describe how the performance of the time series forecasting algorithms differ depending on the data set used. However, for a limited data set of similar time series, it can be possible to determine one particular method or combination of methods that performs best. Following this idea, the article presents an empirical study extracting characteristics of time series in order to generate domain knowledge. This knowledge is then used to dynamically select or combine different forecasting algorithms.

Mateo, Sovilj and Gadea present a method that uses genetic algorithms to select an optimum set of input variables that minimizes the Delta Test on a dataset. The nearest neighbor computation has been speeded up by using an approximate method. The scaling and projection of variables has been addressed to improve the interpretability.

Guillen, Herrera, Rubio, Pomares, Lendasse and Rojas present a totally new approach for the problem of filtering the outliers, reducing the noise and defining a good subset of samples. The approach is based in the concept of Mutual Information with the advantage of just having one parameter to be tuned. The simple idea is efficient and easy to implement, providing satisfactory results within a wide range of problems.

Korpela, Mäkinen, Nöjd, Hollmén and Sulkava present a Markov-switching autoregressive model. Its performance is compared with other statistical and machine learning methods in a new kind of real-world change detection problem with environmental time-series.

Financial Prediction Papers

du Jardin presents two main results. It is shown that a neural-network-based model for predicting bankruptcy performs better when designed with appropriate variable selection techniques than when designed with methods commonly used in the financial literature. Furthermore, it has been found that there is a relationship between the structure of a prediction model and its ability to reduce Type I errors.

Séverin deals with the advantages of the self-organizing map algorithm in the field of corporate finance. Not only the SOM method is able to improve the classical method for bankruptcy prediction but it also questions the scoring models.

17.3 Tools for long-term prediction of time series

Amaury Lendasse, Yu Qi, Yoan Miche, Emil Eirola, Dusan Sovilj, Olli Simula and Antti Sorjamaa

The time series prediction problem is the prediction of future values based on the previous values and the current value of the time series (see Equation 17.1).

$$\hat{y}_{t+1} = f_1(y_t, y_{t-1}, ..., y_{t-M+1}). \tag{17.1}$$

The previous values and the current value of the time series are used as inputs for the prediction model. One-step ahead prediction is needed in general and is referred as Short-Term Prediction. But when multi-step ahead predictions are needed, it is called Long-Term Prediction problem.

Unlike the Short-Term time series prediction, the Long-Term Prediction is typically faced with growing uncertainties arising from various sources. For instance, the accumulation of errors and the lack of information make the prediction more difficult. In Long-Term Prediction, performing multiple steps ahead prediction, there are several alternatives to build models. Two variants of prediction strategies are studied and compared:the Direct (see Equation 17.2) and the Recursive Prediction Strategies (see Equation 17.1).

$$\hat{y}_{t+k} = f_k(y_t, y_{t-1}, ..., y_{t-M+1}). \tag{17.2}$$

In order to perform Long-Term Prediction, several tools have been studied and developed:

- Nonparametric noise estimation
- Imputation of Missing Data
- OP-ELM and Ensembles of ELM

17.4 Nonparametric noise estimation

Elia Liitiäinen, Francesco Corona, Emil Eirola, Olli Simula and Amaury Lendasse

The residual variance estimation problem (or Nonparametric noise Estimation) is well-known in machine learning and statistics under various contexts. Residual variance estimation can be viewed as the problem of estimating the variance of the part of the output that cannot be modeled with the given set of input variables. This type of information is valuable and gives elegant methods to do model selection. While there exist numerous applications of residual variance estimators to supervised learning, time series analysis and machine learning, it seems that a rigorous and general framework for analysis is still missing. For example, in some publications the theoretical model assumes additive noise and independent identically distributed (iid) variables. The principal objective of our work is to define such a general framework for residual variance estimation by extending its formulation to the non-iid case. The model is chosen to be realistic from the point of view of supervised learning. Secondly, we view two well-known residual variance estimators, the Delta test and the Gamma test in the general setting and we discuss their convergence properties.

Contributions:

Minimizing the Delta test for variable selection in regression problems

There exists a wide variety of models that are able to approximate any function such as Radial Basis Function Neural Networks , Multilayer Perceptrons, Fuzzy Systems , Gaussian Process , Support Vector Machines (SVM) and Least Square SVM, etc. however, they all suffer from the Curse of Dimensionality. As the number of dimensions d grows, the number of input values required to sample the solution space increases exponentially, this means that the models will not be able to set their parameters correctly if there are not enough input vectors in the training set. Many real life problems present this drawback since they have a considerable amount of variables to be selected in comparison to the few number of observations. Thus, efficient and effective algorithms to reduce the dimensionality of the data sets are required. Another aspect that is improved by selecting a subset of variables is the interpretability of the designed systems.

The literature presents a wide number of methodologies for feature or variable selection although they have been focused on classification problems. Therefore, specific algorithms for regression must be designed. Recently, it has been demonstrated in how the Delta Test (DT) is a quite powerful tool to determine the quality of a subset of variables. The latest work related to feature selection using the DT consisted in the employment of a local search technique such as Forward-Backward. However, there are other alternatives that allow to perform a global optimization of the variable selection like Genetic Algorithms (GA) and Tabu Search (TS). One of the main drawbacks of using global optimization techniques is their computational cost. Nevertheless, the latest advances in computer architecture provide powerful clusters without requiring a large budget, so an adequate parallelization of these techniques might ameliorate this problem. This is quite important in real life applications where the response time of the algorithm must be acceptable from the perspective of a human operator. Our reserach proposes several new approaches to perform variable selection using the DT as criterion to decide if a subset of variables is adequate or not. The new approaches are based in local search methodologies, global optimization techniques and the hybridization of both [2].

Residual variance estimation in machine learning

The problem of residual variance estimation consists of estimating the best possible generalization error obtainable by any model based on a finite sample of data [3, 4, 5]. Even though it is a natural generalization of linear correlation, residual variance estimation in its general form has attracted relatively little attention in machine learning. In our research, we examine four different residual variance estimators and analyzed their properties both theoretically and experimentally to understand better their applicability in machine learning problems. The theoretical treatment differs from previous work by being based on a general formulation of the problem covering also heteroscedastic noise in contrary to previous work, which concentrates on homoscedastic and additive noise. Secondly, we demonstrate practical applications in input and model structure selection. The experimental results show that using residual variance estimators in these tasks gives good results often with a reduced computational complexity, while the nearest neighbor estimators are simple and easy to implement.

17.5 Imputation of missing data in climatology and finance

Antti Sorjamaa, Olli Simula and Amaury Lendasse

Meteorology and climate research are two rapidly growing fields with an increasing need for accurate and large measurement datasets. The African continent represents a clear example of the current challenges in these fields. The drought and humidity imbalance create extreme conditions for both the people on the continent and the very necessary research. Lake Tanganyika is located in the African Rift in the center of the African continent. It is an important source of proteins for the people around it and the fish industry provides not only the food for the people, but also gives thousands of workers a job.
The importance to the people and the extraordinary size and shape of the lake make it really valuable for the climate research, but the size brings also difficulties. The size and the shape of the lake make it hard to adequately measure the bio-geo-hysical parameters, such as surface temperature. But due to the current political and economical situation in Africa, the satellite is the only valid option. The data measured by satellite includes a vast number of missing values, due to clouds, technical difficulties and even heavy smoke from forest fires. The missing values make a posteriori modeling a difficult problem and the filling procedure a mandatory preprocessing step before climate modeling.
A great number of methods have been already developed for solving the problem by filling the missing values, for example, Kriging and several other Optimal Interpolation methods, such as Objective Analysis. One of the emerging approaches for filling the missing values is the Empirical Orthogonal Functions (EOF) methodology. The EOF is a deterministic methodology, enabling a linear projection to a high-dimensional space. Moreover, the EOF models allow continuous interpolation of missing values even when a high percentage of the data is missing. In our research, an improvement to the standard EOF method is presented, called EOF Pruning. It enhances the accuracy of the EOF methodology and even speeds up the calculation process [6].
Academics as well as practitioners often face the problem of missing data in financial time series. Non-quotation date, too recent inception date, intention not to report a bad performance or mistake of data provider are some of the reasons why missing values occur recurrently in financial databases. Moreover, in order to achieve good performance, most financial models need complete and cylindrical samples. Thus, most of the time, imputation methods have to be applied before running the model. A number of methods have been developed to solve the problem and fill the missing values, both commercial and academical. The methods in both sectors can be classified into two distinct categories : deterministic methods and stochastic methods. Self-Organizing Maps (SOM) aim to ideally group homogeneous individuals, highlighting a neighborhood structure between classes in a chosen lattice. The SOM algorithm is based on unsupervised learning principle where the training is entirely stochastic, data-driven. No information about the input data is required. Recent approaches propose to take advantage of the homogeneity of the underlying classes for data completion purposes. Furthermore, the SOM algorithm allows projection of high-dimensional data to a low-dimensional grid. Through this projection and focusing on its property of topology preservation, SOM allows nonlinear interpolation for missing values.
Empirical Orthogonal Functions (EOF) are deterministic models, enabling linear projection to high-dimensional space. They have also been used to develop models for finding missing data. Moreover, EOF models allow continuous interpolation of missing values, but are sensitive to the initialization. Our research proposes a new method, which combines the advantages of both the SOM and the EOF. The nonlinearity property of the SOM

is used as a de-noising tool and then continuity property of the EOF method is used to recover missing data efficiently [7].

17.6 OP-ELM and ensembles of ELM

Yoan Miche, Antti Sorjamaa, Mark van Heeswijk, Tiina Lindh-Knuutila, Timo Honkela, Erkki Oja, Olli Simula and Amaury Lendasse

,

The amount of information is increasing rapidly in many fields of science. It creates new challenges for storing the massive amounts of data as well as to the methods, which are used in the data mining process. In many cases, when the amount of data grows, the computational complexity of the used methodology also increases.
Feed-forward neural networks are often found to be rather slow to build, especially on important datasets related to the data mining problems of the industry. For this reason, the nonlinear models tend not to be used as widely as they could, even considering their overall good performances. The slow building of the networks comes from a few simple reasons; many parameters have to be tuned, by slow algorithms, and the training phase has to be repeated many times to make sure the model is proper and to be able to perform model structure selection (number of hidden neurons in the network, regularization parameters tuning. . .).
Guang-Bin Huang et al. propose an original algorithm for the determination of the weights of the hidden neurons called Extreme Learning Machine (ELM). This algorithm decreases the computational time required for training and model structure selection of the network by hundreds. Furthermore, the algorithm is rather simplistic, which makes the implementation easy.
In our research, a methodology called Optimally-Pruned ELM (OP-ELM), based on the original ELM, is proposed. The OP-ELM methodology is compared using several experiments and two well-known methods, the Least-Squares Support Vector Machine (LS-SVM) and the Multilayer Perceptron (MLP). Finally, a toolbox for performing the OP-ELM has been developed [8].

Ensembles of ELM have also been used for Time Series Prediction. A large number of application areas of time series prediction involve nonstationary phenomena. Therefore, contrary to the stationary case, one cannot assume that one can use what has been learned from past data and one has to keep learning and adapting the model as new samples arrive. Possible ways of doing this include: 1) retraining the model repeatedly on a finite window of past values and 2) using a combination of different models, each of which is specialized on part of the state space.
Besides the need to deal with nonstationarity, another motivation for such an approach is that one can drop stationarity requirements on the time series. This is very useful, since often we cannot assume anything about whether or not a time series is stationary.
Ensemble methods have been applied in various forms (and under various names) to time series prediction, regression and classification. A non-exhaustive list of literature that discusses the combination of different models into a single model includes bagging, boosting, committees, mixture of experts, multi-agent systems for prediction, classifier ensembles, among others.
In order to construct the ensemble model, a number of Extreme Learning Machines (ELMs) of varying complexity are generated, each of which is individually trained on the data. After training, these individual models are combined in an ensemble model. The output of the ensemble model is a weighted linear combination of the outputs of the individual models. During the test phase, the ensemble model adapts this linear combination over time with the goal of minimizing the prediction error: whenever a particular model has

bad prediction performance (relative to the other models) its weight in the ensemble is decreased, and vice versa. In our first experiments, we tested the performance of this adaptive ensemble model in repeated one-step ahead prediction on a time series that is known to be stationary (the Santa Fe A Laser series). The main goal of this experiment is to test the robustness of the model and to investigate the different parameters influencing the performance of the model. In the second experiments, the model is applied to another time series (Quebec Births) which is nonstationary and more noisy than the Santa Fe time series [9].

17.7 Chemoinformatics

Francesco Corona, Elia Liitiäinen, Tuomas Kärnä, Olli Simula and Amaury Lendasse

Many analytical problems related to spectrometry require predicting a quantitative variable through a set of measured spectral data. For example, one can try to predict a chemical component concentration in a product through its measured infrared spectrum. In recent years, the importance of such problems in various fields including the pharmaceutical, food and textile industries have grown dramatically. The chemical analysis by spectrophotometry rests on the fast acquisition of a great number of spectral data (several hundred, even several thousands).

In spectrometric problems, one is often faced with databases having more variables (spectra components) than samples; and almost all models use at least as many parameters as the number of input variables. These two problems, colinearity and risk of overfitting, already exist in linear models. However, their effect may be even more dramatic when nonlinear models are used (there are usually more parameters than in linear models, and the risk of overfitting is higher). In such high-dimensional problems, it is thus necessary to use a smaller set of variables than the initial one.

We have proposed methods to select spectral variables by using concepts from information theory [10, 11, 12, 13, 14]:

- the measure of mutual information
- the measure of topological relevance on the Self-Organizing Map
- the Functional Data Analysis (FDA)
- Nonparametric Noise Estimation

One particular application has been studied in the field of Oil Production.

In this industrial application, there has been applied process data. The aim has been to get new empirical modeling tools, which are based on information technology. The outcome has been emphasized on tools, which are suitable in fast data mining from large data sets. The test cases have included:

- Analysis of instrumental data, on-line monitoring data and quality data
- Non-linear processes
- Identification of delays between stages in industrial processes
- Robust variable selection methods

17.8 Steganography and steganalysis

Yoan Miche, Amaury Lendasse, Patrick Bas and Olli Simula

Steganography has been known and used for a very long time, as a way to exchange information in an unnoticeable manner between parties, by embedding it in another, apparently innocuous, document. For example, during the 80's, Margaret Thatcher decided to have each word processor of the government's administration members changed with an unique word spacing for each, giving a sort of invisible signature to documents. This was done to prevent the continuation of sensitive government information leaks.

Nowadays steganographic techniques are mostly used on digital contents. The online newspaper, Wired News, reported in one of its articles on steganography that several steganographic contents have been found on web-sites with very large image database such as eBay.

Most of the time research about steganography is not as much to hide information, but more to detect that there is hidden information. This reverse part of the steganography is called steganalysis and is specifically aimed at making the difference between genuine documents, and steganographied – called stego – ones. Consequently, steganalysis can be seen as a classification problem where the goal is to build a classifier able to distinguish these two sorts of documents.

During the steganographic process, a message is embedded in an image so that it is as undetectable as possible. Basically, it uses several heuristics in order to guarantee that the statistics of the stego content (the modified image) are as close as possible to the statistics of the original one. Afterwards, steganalysis techniques classically use features extracted from the analyzed image and an appropriately trained classifier to decide whether the image is genuine or not.

In our work, a widely used and known set of 193 image features has been used. Theses features consider statistics of JPEG compressed images such as histograms of DCT coefficients for different frequencies, histograms of DCT coefficients for different values, global histograms, blockiness measures and co-occurrence measures. The main purpose of this high number of features is to obtain a model able to detect about any steganographic process.

The usual process in steganalysis is then to train a classifier according to the extracted features. Consequently a set of 193 features for each image of the database is obtained, giving an especially high dimensionality space for classifiers to work on. Earlier research about these high dimensionality spaces has shown that a lot of issues come out when the number of features is as high as this one.

The main idea behind the carried out work [15, 16] is to give insights on proper handling and use of such high dimensionality datasets; indeed, these are very common in the steganography/steganalysis field and users tend not to respect basic principles (for example having a sufficient number of samples regarding the dimensionality of the problem).

17.9 Bankruptcy prediction

Yu Qi, Laura Kainulainen, Eric Severin, Olli Simula and Amaury Lendasse

Bankruptcies are not only financial but also individual crises which affect many lives. Although unpredictable things may happen, bankruptcies can be predicted to some extent.

This is important for both the banks and the investors that analyze the companies, and for the companies themselves. The aim of our research is to see, whether new machine learning models combined with variable selection perform better than traditional models: Linear Discriminant Analysis, Least Squares Support Vector Machines and Gaussian Processes. They form a good basis for comparison, since LDA is a widely spread technique in the financial tradition of bankruptcy prediction, LSSVM is an example of Support Vector Machine classifiers and Gaussian Processes is a relatively new Machine Learning method.

Since all the possible combinations of the variables cannot be evaluated due to time constraints, forward selection may offer a fast and accurate solution for finding suitable variables.

Our main results can be found in [17, 18, 19].

References

[1] A. Lendasse. European Symposium on Time Series Prediction, ESTSP'08, Amaury Lendasse editor, ISBN 978-951-22-9544-9.

[2] A. Guillén, D. Sovilj, F. Mateo, I. Rojas and A. Lendasse, Minimizing the Delta Test for Variable Selection in Regression Problems International Journal of High Performance Systems Architecture, Vol. 4, pp. 269-281, 2008.

[3] E. Liitiäinen, M. Verleysen, F. Corona and A. Lendasse, Residual variance estimation in machine learning, Neurocomputing, October 2009, pp. 3692-3703.

[4] E. Liitiäinen, F. Corona, A. Lendasse, On non-parametric residual variance estimation Neural Processing Letters, December 2008,pp. 155-167.

[5] E. Liitiäinen, A. Lendasse andF. Corona, Bounds on the mean power-weighted nearest neighbour distance, Proceedings of the Royal Society A, September 2008, pp. 2293-2301.

[6] A. Sorjamaa, A. Lendasse, Y. Cornet and E. Deleersnijder, An improved methodology for filling missing values in spatiotemporal climate data set, Computational Geosciences, January 2010, pp. 55-64.

[7] A. Sorjamaa, P. Merlin, B. Maillet and A. Lendasse, A Non-Linear Approach for Completing Missing Values in Temporal Databases, European Journal of Economic and Social Systems, November 2009, pp. 99-117.

[8] Y. Miche, A. Sorjamaa and A. Lendasse, OP-ELM: Theory, Experiments and a Toolbox, LNCS - Artificial Neural Networks - ICANN 2008 - Part I, September 2008, pp. 145-154.

[9] M. van Heeswijk, Y. Miche, Tiina Lindh-Knuutila, Peter A.J. Hilbers, Timo Honkela, E. Oja and A. Lendasse, Adaptive Ensemble Models of Extreme Learning Machines for Time Series Prediction, ICANN'09, Lecture Notes in Computer Science, pp. 305-314.

[10] F. Corona and E. Liitiäinen and A. Lendasse and L. Sassu and S. Melis and R. Baratti, A SOM-based approach to estimating product properties from spectroscopic measurements, Neurocomputing, 2008, pp. 71-79.

[11] T. Kärnä, F. Corona and A. Lendasse, Gaussian basis functions for chemometrics, Journal of Chemometrics, 2008, pp. 701–707.

[12] F. Corona, S.P. Reinikainen, K. Aaljoki, A. Perkkio, E. Liitiäinen, R. Baratti, A. Lendasse and O. Simula, Wavelength selection using the measure of topological relevance on the Self-Organizing Map, Journal of Chemometrics, 2008, pp. 610-620.

[13] F. Corona , M. Mulas, R. Baratti and J. A. Romagnoli, On the topological analysis of industrial process data using the SOM, Elsevier, Computer Aided Chemical Engineering: Proceedings of PSE 2009 International Symposium on Process Systems Engineering, Salvador Bahia (Brazil), August 16-20 2009, pp. 1173-1178.

[14] F. Corona, E. Liitiäinen, A. Lendasse, R. Baratti and L. Sassu, Delaunay tessellation and topological regression: An application to estimating product properties. , Elsevier, Computer Aided Chemical Engineering: Proceedings of PSE 2009 International

Symposium on Process Systems Engineering, Salvador Bahia (Brazil), August 16-20 2009, pp. 1179-1184.

[15] Y. Miche, P. Bas, A. Lendasse, C. Jutten and O. Simula, A Feature Selection Methodology for Steganalysis, Traitement du Signal, May 2009, pp. 13-30.

[16] Y. Miche and P. Bas and A. Lendasse and C. Jutten and O. Simula, Reliable Steganalysis Using a Minimum Set of Samples and Features, EURASIP Journal on Information Security, March 2009, 13 pages.

[17] Q. Yu, A. Sorjamaa, Y. Miche and E. Séverin, A methodology for time series prediction in Finance, ESTSP'08, September 17-19, 2008, pp. 285-293.

[18] Q. Yu, A. Sorjamaa, Y. Miche, A. Lendasse, A. Guillén, E. Séverin and F. Mateo, Optimal Pruned K-Nearest Neighbors: OP-KNN - Application to Financial Modeling , Eighth International Conference on Hybrid Intelligent Systems, September 2008, pp. 764-769.

[19] Q. Yu, A. Sorjamaa, Y. Miche, E. Séverin and A. Lendasse, OP-KNN for Financial regression problems, Mashs 08, Computational Methods for Modelling and learning in Social and Human Sciences, Creteil France, June 5-6, 2008.