Neurocomputing 73 (2010) 2006–2016

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Meta-learning for time series forecasting and forecast combination

Christiane Lemke *, Bogdan Gabrys

Smart Technology Research Centre, School of Design, Engineering and Computing, Bournemouth University, Poole House, Talbot Campus, Poole BH12 5BB, UK

ARTICLE INFO

Available online 7 March 2010

Keywords:
Forecasting
Forecast combination
Time series
Time series features
Meta-learning
Diversity

ABSTRACT

In research of time series forecasting, a lot of uncertainty is still related to the task of selecting an appropriate forecasting method for a problem. It is not only the individual algorithms that are available in great quantities; combination approaches have been equally popular in the last decades. Alone the question of whether to choose the most promising individual method or a combination is not straightforward to answer. Usually, expert knowledge is needed to make an informed decision, however, in many cases this is not feasible due to lack of resources like time, money and manpower. This work identifies an extensive feature set describing both the time series and the pool of individual forecasting methods. The applicability of different meta-learning approaches are investigated, first to gain knowledge on which model works best in which situation, later to improve forecasting performance. Results show the superiority of a ranking-based combination of methods over simple model selection approaches.

1. Introduction

Time series forecasting has been a very active area of research since the 1950s, with research on the combination of time series forecasts starting a few years later. During this time, many empirical studies on forecasting performance have been conducted to assess performance of the continuously growing numbers of available algorithms, for example in [42,33]. These studies, however, fail to provide consistent results as to which actual method performs best, which is not surprising considering the variety in investigated time series forecasting problems. Robert J. Hyndman described the future challenges for time series prediction [36] in the following words: "Now it is time to identify why some methods work well and others do not".

But what is it that determines the success or failure of a forecasting model? The well-known no-free-lunch theorem, for example described in [49], states that there are no algorithms that generally perform better or worse than random when looking at all possible data sets. This implies that no assumptions on the performance of an algorithm can be made if nothing is known about the problem that it is applied to. Of course, there will be specific problems for which one algorithm performs better than another in practice. In accordance to this, this work investigates approaches to relax the assumption that nothing is known about a problem by automatically extracting domain knowledge from a data, linking it to well-performing methods and drawing conclusions for a similar set of time series.

Traditionally, experts visually inspect time series characteristics and fit models according to their judgement. This work investigates an automatic approach, since a thorough time series analysis by humans is often not feasible in practical applications that process a large number of time series in very limited time.

A classic and straightforward classification for time series has been given by Pegels [37]. Time series can thus have patterns that show different seasonal effects and trends, both of which can be additive, multiplicative or non-existent. Gardner [19] extended this classification by including damped trends. Time series analysis in order to find an appropriate ARIMA model has been discussed since the seminal paper of Box and Jenkins [6]. Guidelines are summarised in [34] and rely heavily on visually examining autocorrelation and partial autocorrelation values of a series.

The idea of using characteristics of univariate time series to select an appropriate forecasting model has been pursued since the 1990s. The first systems were rule based and built on a mix of judgemental and quantitative methods. Collopy and Armstrong use time series features to generate 99 rules [12] for weighting four different models; features were obtained judgementally, by both visually inspecting the time series and using domain knowledge. Adya et al. later modify this system and reduced the necessary human input [1,2], yet did not abandon expert intervention completely. Vokurka et al. [47] extract features automatically to weight between three individual models and a combination in a rule-base that was built automatically, but required manual review of the outputs. Completely automatic systems have been proposed in [4], where a generated rule base selects between six forecasting

* Corresponding author. Tel.: +44 1202 595298; fax: +44 1202 595314.
E-mail addresses: clemke@bournemouth.ac.uk (C. Lemke), bgabrys@bournemouth.ac.uk (B. Gabrys).

0925-2312/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2009.09.020

methods. Discriminant analysis to select between three forecasting methods using 26 features is used in [41].

The phrase "meta-learning" in the context of time series was first used in [39] and represents a new term for describing the process of automatically acquiring knowledge for time series forecasting model selection that was adopted from the general machine learning community. Two case studies are presented in [39]: In the first one, a C4.5 decision tree is used to link six features to the performance of two forecasting methods; in the second one, the NOEMON approach [25] is used for ranking three methods. The most recent and comprehensive treatment of the subject can be found in [48], where time series are clustered according to their data characteristics and rules generated judgementally as well as using a decision tree. The approach is then extended to determine weights for a combination of individual models based on data characteristics. Table 1 summarises some facts about the related work presented here for better overview of approaches and methods used. The calculation of features and meta-learning method listed are implemented automatically if not otherwise stated.

Some time series features presented in this work are similar to the ones used in literature, but new and different features are introduced extending previous work published in [28,30]. In particular, features concerning the diversity of the pool of algorithms are included, which is facilitated by adding a number of popular forecast combination algorithms to the feature pool. In addition to the original question of which model to select, this work also tries to find evidence for features being useful for guiding the choice of whether to pick an individual model or a combination. In an initial exploratory experiment, decision trees are generated to find evidence for the existence of a link between time series characteristics and the performance of models. Leaving aside the requirement for interpretable rules and recommendations, four meta-learning techniques are compared in another empirical experiment, assessing potential performance improvements.

The paper is structured as follows: Section 2 will present the underlying empirical experiments and results. Section 3 begins treating the model selection problem as a classification task and describes an extensive number of time series characteristics which are necessary to link performances of algorithms to the nature of the time series. The different experiments using meta-learning techniques are evaluated in Section 4.

2. Performance of forecasting and forecast combination methods

This part of this work presents empirical experiments that provide the basis for further meta-learning analysis. Individual predictors are diverse, but are kept relatively simple and, more importantly, automatic. These methods perform often just as well as more complex methods [33] are more efficient in terms of computational requirements and also more likely to be employed in practical applications, especially when no expert advice is available.

2.1. Data sets

Two data sets both consisting of 111 time series have been used in this study; they were obtained from the NN3 [13] and NN5 [14] neural network forecasting competitions. NN3 data include monthly empirical business time series with 52–126 observations, while the NN5 series are daily time series from cash machine withdrawals with 735 observations each. The competition task was to predict the next 18 or 56 observations, respectively. While NN3 data did not need specific preprocessing, NN5 data included some missing values, which were substituted by taking the mean of the value of the corresponding weekday of the previous and the following week. The last 18 or 56 values of each series were not used for training the models to enable out-of-sample error evaluation.

2.2. Forecasting methods

Available forecasting algorithms can be roughly divided into a few groups. Simple approaches are often surprisingly robust and popular, for example those based on exponential smoothing [20,33]. Statisticians and econometricians tend to rely on complex ARIMA models and their derivatives [6]. The machine learning community mainly looks at neural networks, either using multi-layer-perceptrons with time-lagged time series observations as inputs as, for example, in [50,16], or recurrent networks with a memory, see, for example [26]. As not all of the algorithms provide native multi-step-ahead forecasting, some of them are implemented using two approaches: An iterative approach, where the last prediction is fed back to the model to obtain the next forecast, or a direct approach, where n different predictors are trained for each of the 1 to n steps ahead problem. The selection of models used in this work is presented in the next paragraphs.

2.2.1. Simple forecasting models

Many algorithms for forecasting time series are considered simple, yet they are usually very popular and can be surprisingly powerful. In the latest extensive M3 competitions [33], an exponential smoothing approach was considered a good match for the most successful complex method while providing a better trade-off between prediction accuracy and computational complexity. Simple methods used for this experimental study

Table 1
Time series model selection—overview of literature.

Year	Authors	Features	Meta-learning method	Time series	Model pool
1992	Collopy and Armstrong [12]	18 (judgemental)	Rule base (judgemental)	126 (M1)	Exp. smoothing (Holt and Brown), random walk, linear regression
1996	Vokurka et al. [47]	5	Rule base (partly automatic)	126 (M1)	Exp. smoothing (single and Gardner), structural and a combination of the three
1997	Arinze et al. [4]	6	Rule base	67	Exp. Smoothing (Holt and Winter), adaptive filtering, three "hybrids" of the previous
1997	Shah [41]	26	Discriminant analysis	203 (M1)	Exp. smoothing (single and Holt-Winter), structural
2000	Adya et al. [2]	26 (mainly automatic)	Rule base (judgemental)	3003 (M3)	Exp. smoothing (Holt and Brown), random walk, linear regression
2004	Prudencio and Ludermir [39]	6/7	Decision tree/NOEMON	99/645 (M3)	Exp. smoothing, neural network/random walk, Holt's smoothing, auto-regressive
2009	Wang et al. [48]	9	Decision tree	315	Random walk, smoothing, ARIMA, neural network