International Journal of Forecasting 34 (2018) 507–528

Contents lists available at ScienceDirect

International Journal of Forecasting

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

Determining analogies based on the integration of multiple information sources

Emiao Lu *, Julia Handl, Dong-ling Xu

Alliance Manchester Business School, Manchester, UK

ARTICLE INFO

Keywords:
Analogy
Bayesian pooling
Kalman filter
Model selection
Multicriteria clustering

ABSTRACT

Forecasting approaches that exploit analogies require the grouping of analogous time series as the first modeling step; however, there has been limited research regarding the suitability of different segmentation approaches. We argue that an appropriate analytical segmentation stage should integrate and trade off different available information sources. In particular, it should consider the actual time series patterns, in addition to the variables that characterize the drivers behind the patterns observed. The simultaneous consideration of both information sources, without prior assumptions regarding the relative importance of each, leads to a multicriteria formulation of the segmentation stage. Here, we demonstrate the impact of such an adjustment to segmentation on the final forecasting accuracy of the cross-sectional multi-state Kalman filter. In particular, we study the relative merits of single and multicriteria segmentation stages for a simulated data set with a range of noise levels. We find that a multicriteria approach consistently achieves a more reliable recovery of the original clusters, and this feeds forward to an improved forecasting accuracy across short forecasting horizons. We then use a US data set on income tax liabilities to verify that this result generalizes to a real-world setting.

1. Introduction

Forecasting approaches such as the cross-sectional multi-state Kalman filter algorithm (C-MSKF: Duncan, Gorr, & Szczypula, 1993) exploit information from analogies or analogous time series in order to increase the accuracy of point forecasts for a target time series. The identification of suitable analogies is crucial to these approaches; nevertheless, surprisingly little research has investigated appropriate analytical modeling approaches for judging similarities between time series (Lee, Goodwin, Fildes, Nikolopoulos, & Lawrence, 2007) and supporting the principled selection of analogies (Armstrong, 2001).

The identification of analogous time series typically involves the use of segmentation approaches for partitioning a set of time series into a set of homogeneous clusters (e.g., Duncan, Gorr, & Szczypula, 2001). Segmentation approaches can be applied widely in areas such as economics, finance, operational research, and public budgeting. Segmentation is typically used to identify meaningful sub-groups (e.g., customers, businesses and countries), and can be useful for identifying, understanding and targeting these groups. The sub-groups identified during segmentation may feed forward into further analysis, including the development of cluster-specific forecasting strategies. Segmentation is often modeled as a single-criterion problem, both in the traditional marketing literature and in practice, but it is inherently a multicriteria problem, as clusters are typically desired to be homogeneous with respect to a set of both explanatory and response variables (Liu, Ram, Lusch, & Brusco, 2010; Myers, 1996; Smith, 1956). Similarly, in the context of forecasting, we may view the segmentation as involving multiple information sources, as one must consider both past realizations of a given time series (response variables) and the associated causal factors (explanatory variables) that describe the underlying causal relationships

* Corresponding author.
E-mail address: emiao.lu@gmail.com (E. Lu).

https://doi.org/10.1016/j.ijforecast.2018.02.002
0169-2070/

for the co-movement of the analogous time series (Duncan et al., 2001). For example, a set of products may be considered a group due to sharing the same sphere of influence, similar consumer preferences, promotion levels, or local trends. Ignoring one of these sources of information during the segmentation stage may lead to clusters that are not differentiated sufficiently in terms of either time series patterns or causal factors, and thus lead to sub-optimal results in further analysis. In order to obtain meaningful groups of analogies for forecasting, we need to ensure that we are identifying clusters that are interpretable at a domain level (represented by similarities in the values of a set of shared causal factors), but also simultaneously show similarities in their time-based patterns.

Here, we experiment with a simple prediction process that outlines this idea and contrasts the performances of single-criterion and multicriteria segmentation approaches in the context of forecasting analogous time series, for which both time-based patterns and potential causal factors are known. We demonstrate that the segmentation approach using both information sources is preferable, in the sense that it can generate, and usually identify, segmentations that boost the performance of pooling in terms of the forecasting accuracy.

The remainder of the paper is structured as follows. Section 2 surveys related work in the literature, including pooling approaches and popular segmentation approaches. Section 3 proposes our three-stage prediction process. Section 4 presents experiments that investigate the impacts of different segmentation approaches on the performances of different pooling approaches. In particular, we use simulated data to investigate the sensitivity of the approaches to changes in the relative reliability of the two information sources. Section 5 then summarizes the results on a data set of personal income tax liability data. Finally, Section 6 concludes.

2. Previous research

Analogies have been employed widely in the forecasting field in an attempt to improve the forecasting accuracy (Armstrong, 2006; Green & Armstrong, 2007; Piecyk & McKinnon, 2010). According to Duncan et al. (2001), analogies can be defined as time series that exhibit similarities in time-based patterns due to shared underlying causal factors. They typically co-vary, and thus, are correlated positively over time.

Analogies have been utilized most commonly in the context of judgmental approaches to forecasting (i.e., forecasting by analogy and related work, as per Nikolopoulos, Litsa, Petropoulos, Bougioukos, & Khammash, 2015; Savio & Nikolopoulos, 2013). These methods use analogies for the purpose of adjusting statistical forecasts (Webby & O'Connor, 1996), since this may reduce the biases due to optimism or wishful thinking (Armstrong, 2001; Petropoulos, Makridakis, Assimakopoulos, & Nikolopoulos, 2014). There has also been work done on the development of statistical methods that can exploit the information available from analogies. A well-established model is the Bass model (Bass, 1969; Nikolopoulos, Buxton, Khammash, & Stern, 2016), which has been used to forecast the sales of products that are yet to be launched, through the use of information from similar products (Goodwin, Dyussekeneva, & Meeran, 2013). An alternative way of exploiting analogies is to use Bayesian pooling approaches, such as the cross-sectional multi-state Kalman filter (C-MSKF: Duncan et al., 1993, 2001), which requires a relatively small number of parameters. This method borrows strength from groups of analogous time series in order to increase the accuracy of point forecasts.

Time series forecasting with respect to the demand for products or services often needs to be robust in situations that are characterized by structural change (i.e. changes to the trend of the time series), due, for example, to external influences such as the action of a competitor. Methods such as exponential smoothing (Brown, 2004) and the multi-state Kalman filter (MSKF: Harrison & Stevens, 1971), which revise model parameter estimates over time, have been developed for dealing with such situations. Such methods must compromise between two different needs, namely a responsiveness to change and the accuracy of the forecasts. By utilizing additional information from analogies, the C-MSKF method combines the MSKF's ability to yield accurate forecasts with a quick responsiveness to change. This approach has proven effective in a number of challenging applications, such as the forecasting of the churn in telecommunications networks (Greis & Gilstein, 1991), infant mortality rates (Duncan, Gorr, & Szczypula, 1995) and tax revenue (Duncan et al., 1993). The C-MSKF can draw strength from the availability of multiple data points for the same time period across different analogous series, which lends it robustness with respect to outliers. In general, C-MSKF has been said to be competitive with conventional time series forecasting methods, such as the damped exponential smoothing (Damped) methods, exponential smoothing (ETS), MSKF, the naïve drift method (Drift), the random walk (RW) or the Theta model in situations that satisfy the following three conditions (Duncan, Gorr, & Szczypula, 1994; Duncan et al., 2001): (i) the number of points that are suitable for extrapolation is small (due to either size or a structural change); (ii) analogies are present across several time series; and (iii) at least three observations are available after a structural change due to the impact of an external influence. Finally, a key assumption behind C-MSKF is that time series that are classed as analogous (i.e., that exhibit co-movement during the investigation's estimation period) do not diverge frequently in the forecasting periods. This requirement underlines the importance of determining analogies accurately as the first step of the analysis.

The homogeneity of the underlying set of analogous time series is fundamental to the effectiveness of pooling approaches (Stimson, 1985). Previous research (Duncan et al., 2001) has demonstrated that pooling across a homogeneous set of time series gives a superior forecasting accuracy relative to pooling across a heterogeneous set. In this context, three general approaches have typically been considered for identifying analogies, namely correlational co-movement, i.e., the grouping of time series based on the correlations between the time series patterns observed; the grouping of time series using model-based approaches (Frühwirth-Schnatter & Kaufmann, 2008); and