European Journal of Operational Research 234 (2014) 127–139

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Stochastics and Statistics

A combination selection algorithm on forecasting

Shuang Cang [a,*], Hongnian Yu [b,1]

[a] School of Tourism, Bournemouth University, Fern Barrow, Poole, Dorset BH12 5BB, UK
[b] School of Design, Engineering & Computing, Bournemouth University, Fern Barrow, Poole, Dorset BH12 5BB, UK

ARTICLE INFO

Article history:
Received 25 September 2012
Accepted 27 August 2013
Available online 8 September 2013

Keywords:
Neural networks
Seasonal autoregressive integrated moving average
Combination forecast
Information theory

ABSTRACT

It is widely accepted in forecasting that a combination model can improve forecasting accuracy. One important challenge is how to select the optimal subset of individual models from all available models without having to try all possible combinations of these models. This paper proposes an optimal subset selection algorithm from all individual models using information theory. The experimental results in tourism demand forecasting demonstrate that the combination of the individual models from the selected optimal subset significantly outperforms the combination of all available individual models. The proposed optimal subset selection algorithm provides a theoretical approach rather than experimental assessments which dominate literature.

1. Introduction

Forecasting has received the considerable research during the past three decades. Three main types of forecasting models (Li, Song, & Witt, 2005; Song & Li, 2008) are *Time series model* (Cao, Ewing, & Thompson, 2012; Cho, 2001; Coshal & Charlesworth, 2011), *Causal econometric model* (Li, Song, & Witt, 2006; Naude & Saayman, 2005; Page, Song, & Wu, 2012; Roget & Gonzalez, 2006) and new emerging *Artificial Intelligence based model*, such as neural network, fuzzy time-series theory, genetic algorithms, and expert systems (Bodyanskiy & Popov, 2006; Cao et al., 2012; Carbonneau, Laframboise, & Vahidov, 2008; Chen & Wang, 2007; Cho, 2003; Hadavandi, Ghanbari, Shahanaghi, & Abbasian-Naghneh, 2011; Law & Au, 1999; Pai & Hong, 2005; Wong, Xia, & Chu, 2010; Wu & Akbarov, 2011). From these studies, researchers often seek to identify the best individual model to generate a forecast. However, combination forecasting has proven to be a highly successful forecasting strategy in many fields, which has been demonstrated by empirical studies.

Forecast combination was pioneered in the sixties by Bates and Granger (1969). Since then it has been demonstrated that forecast combinations are often superior to their constituent forecasts in many fields (Greer, 2005; Hall & Mitchell, 2007; Holden & Peel, 1986; Lessmann, Sung, Johnson, & Ma, 2012; Li, Shi, & Zhou, 2011; Newbold & Granger, 1974; Sánchez, 2008; Timmermann, Elliott, & Granger, 2006; Winkler & Makridakis, 1983; Zheng, Lee, & Shi, 2006). The most widely used and studied combination forecast methods are ensemble methods, such as bagging (Breiman, 1996) and boosting methods. The typical boosting methods are AdaBoost (Freund & Schapire, 1997), LogitBoost (Tibshirani, Friedman, & Hastie, 2000) and MultiBoost (Webb, 2000). These methods which have the learning capability have two steps: *step 1:* construct a set of predication models; *step 2:* predicate a new pattern by taking a weighted vote of their predications. The average or median is used for the continuous outputs, and the majority voting is used for the categorical outputs of the set of predication models from step 1. The most applications are the categorical outputs from the set of predication models. For examples, Wezel van and Pothars (2007) applied ensembles methods (bagging and boosting) to the customer choice modelling problem to improve customer choice predictions. Abellán and Masegosa (2010) proposed the ensemble method using credal decision trees, and showed the good percentage of correct classifications and an improvement in time of processing, especially for large data sets. Finlay (2011) applied bagging and boosting methods to the credit risk assessment to classify consumers as good or bad credit risks, and proposed a new boosting algorithm, 'error trimmed boosting'. Experiments showed that the bagging and boosting methods outperform other multi-classifier systems, and 'error trimmed boosting' outperforms bagging and AdaBoost by a significant margin.

For the continuous outputs from the set of predication models, Li, Wong, and Troutt (2001) proposed an approximate Bayesian algorithm for combining forecasts using several examples. Zou and Yang (2004) developed an algorithm called 'AFTER' to calculate the weights in the combination forecasting with one-step-ahead forecasting, where the weights are updated for each additional

* Corresponding author. Tel.: +44 (0)1202 966758; fax: +44 (0)1202 515707.
E-mail addresses: scang@bournemouth.ac.uk (S. Cang), yuh@bournemouth.ac.uk (H. Yu).
[1] Tel.: + 44 (0)1202 961150.

0377-2217/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.ejor.2013.08.045

observation. The results demonstrated the advantage of the 'AFTER' algorithm. He and Xu (2005) applied the self-organizing algorithm to combine the forecasting models, and demonstrated the superiority by an example of the total retail sales of consumer goods in Chengdu. All individual candidate models are used in the combination for these researches (Li et al., 2001; Zou & Yang, 2004; He & Xu, 2005).

For tourism demand forecasting, the outputs of the individual models are continuous variables. The most common combination forecasting models are linear combination of all available individual forecast models in tourism literature. The researchers (Andrawisa, Atiyaa, & El-Shishiny, 2011; Chan, Witt, Lee, & Song, 2010; Coshal & Charlesworth, 2011; Freitas & Rodrigues, 2006; Lessmann et al., 2012; Menezes de, Bunn, & Taylor, 2000; Shen, Li, & Song, 2011) have demonstrated the efficiency of combination forecasts and the superiority of combination forecasts in contrast to individual forecasts. However all available individual models are used as inputs for the combinations. The question is whether we can optimally select a subset of all individual models instead of all individual models in constructing the combination model. If a subset of individual models as inputs for a combination model can improve performance over using all available individual models as inputs in terms of accuracy and robustness, then this subset of individual models is called as an 'optimal subset'.

One of the important issues is how to select the optimal subset of individual models from all those available individual models without having to try all possible combinations of the individual models. This poses an important challenge as examining all possible combinations of individual models only provides an experimental assessment which does not have a rigorous proof from a theoretical perspective. Furthermore, trying all possible combinations would involve intensive computation and is extremely time-consuming if the total number of individual forecasting models is large. The total number of all possible combinations is $\sum_{m=2}^{M} C_M^m/\Gamma(m+1)$ excluding the individual models for one combination method if there are M individual candidate models available, where $C_M^m = M \times (M-1) \times\ (M-2) \times \cdots \times (M-m+1)$ and $\Gamma(m+1) = m \times (m-1) \times \cdots \times 2 \times 1$. For example, there are 502 possible combinations for one combination method if M equals nine (nine individual models in total).

Combination selection forecasting is rarely studied in the literature. Costantini and Pappalardo (2010) and Kisinbay (2010) employed the encompassing test for combination forecasts algorithms. Costantini and Pappalardo (2010) proposed a hierarchical procedure for the combination, where the procedure was investigated using short-term forecasting models for monthly industrial production in Italy. Kisinbay (2010) demonstrated that the combination forecasts algorithm outperform the benchmark model forecasts using the US macroeconomic dataset, the algorithm developed by Kisinbay (2010) was adopted to analyse US data in the IMF working paper by Baba and Kisinbay (2011).

An optimal subset selection from all individual forecasting models is studied in this paper. The optimal subset may contain one individual model, up to a maximum of all individual models. If the selected subset contains only one single model, this means that the individual model gives the best performance out of all possible combinations of individual models.

An optimal subset selection algorithm using information theory (Mackay David, 2003) is proposed in this paper. The linear combination models proposed by Shen, Li, and Song (2008), Shen et al. (2011) and Wong, Song, Witt, and Wu (2007) are used to examine the optimal subset selection algorithm for this study. The information concepts have never been applied to the selection of individual models as combination models, and all available individual models are used as inputs for the linear combination methods in tourism demand forecasting literature. For this reason, it is useful to explain the developments in information theory that contribute to forecasting.

2. Methodological issues

2.1. Information theory

Traditionally, the best single forecasting model is selected from several individual models in terms of accuracy. In most cases, the best single model may not have extracted all the information that is relevant for the actual output values. The combination models may be able to offer more information to provide a better prediction compared with an individual model. Shannon's information theory (Mackay David, 2003) argues that we can select an optimal subset of all individual models, and this subset contains enough information to forecast the actual outputs. Optimal subset selection using information theory is widely used in other fields such as the pattern recognition and neural networks fields.

Sridhar, Bartlett, and Seagrave (1999) proposed an algorithm using information theory for combining neural network models. This algorithm identifies and combines useful models regardless of the nature of their relationship to the actual output. The algorithm was demonstrated through three examples including the application to a dynamic process modelling problem. The obtained results demonstrated that the algorithm could achieve highly improved performance as compared with a single optimal network or the stacked neural networks based on a linear combination of neural networks.

Many algorithms on feature selection based on mutual information (MI) were developed. The algorithm 'mutual information based feature selection' (MIFS) based on MI between the individual and the class variables was developed by Battiti (1994) for selecting the features in the supervised neural net learning. However this algorithm can only calculate the MI between one single variable with another single variable. Kwak and Choi (2002) analyzed the limitations of the MIFS algorithm (Battiti, 1994) and proposed an 'MI feature selection uniform information distribution' (MIFS-U) algorithm to overcome its limitations. Both MIFS and MIFS-U algorithms can provide better performance compared with the feature selection algorithms such as principal component analysis and neural networks, and have been successfully applied in many experimental design problems. However, both algorithms involve a parameter and it is difficult to determine the range of its value. The fixed parameter is used in the MI based feature selection 'minimal redundancy maximal relevance' (mRMR) algorithm (Peng, Long, & Ding, 2005). The 'normalized mutual information feature selection' (NMIFS) algorithm was proposed in the paper (Estévez, Tesmer, Perez, & Zurada, 2009) based on the normalized MI by the minimum entropy of both features. The average normalized MI is used as a measure of redundancy of the individual feature and the subset of selected features. The experiments demonstrated that the NMIFS algorithm enhances the MIFS, MIFS-U and mRMR algorithms. The parameter is also fixed in the NMIFS algorithm, which is an advantage comparing with the algorithms MIFS and MIFS-U.

In term of speeding, 'fast correlation based filter' (FCBF) is fast due to that a few evaluations of bivariate mutual information are computed. The FCBF is a ranking method combined with the redundancy analysis (Yu & Liu, 2004). Fleuret (2004) proposed the forward selection and 'conditional mutual information maximization criterion' (CMIM) in term of binary feature selection and showed that CMIM is competitive with the FCBF in selecting binary features. Meyer, Schretter, and Bontempi (2008) proposed a 'matrix of average sub-subset information for variable elimination' (MASSIVE) using variable complementarity for microarray