Economic Modelling 30 (2013) 514–520

Contents lists available at SciVerse ScienceDirect

Economic Modelling

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

D-GMDH: A novel inductive modelling approach in the forecasting of the industrial economy

Mingzhu Zhang [a], Changzheng He [a,*], Xin Gu [a], Panos Liatsis [b], Bing Zhu [a]

[a] Business School, Sichuan University, Chengdu, 610064, China
[b] School of Engineering and Mathematical Sciences, City University, London EC1V 0HB, UK

ARTICLE INFO

Article history:
Accepted 13 September 2012

JEL classification:
C45
E17
L16

Keywords:
Economic forecasting
Noise
GMDH
Diversity

ABSTRACT

This work proposes a new forecasting model to analyse the economic development of Sichuan province of China. The model, which introduces the concept of diversity, is based on an improvement of the -GMDH algorithm. The new method, called D-GMDH, is compared with two ensemble approaches which are introduced by Dutta (2009), and D-GMDH is better than the two approaches in forecasting accuracy. D-GMDH is also applied to forecast the industrial added value of the Sichuan province. The obtained results are compared with those of the traditional GMDH model, GMDH combination model and the widely used ARMA model. The results show that D-GMDH has good prediction accuracy and is an effective means for economic forecasting when data is contaminated by noise.

1. Introduction

A major yet challenging application of forecasting is the field of economy. Chinese economy demonstrated strong growth in recent years, thus attracting the attention of many researchers (Chow and Li, 2002; Ho and Tsui, 2004; Qin et al., 2007). The economic reform in China is successful in terms of economic growth; however, the economy still undergoes transition. Under this circumstance, the data collection, conversion and entry bring many biases into the data, resulting noisy data. This problem is particularly evident in some regions of China. Thus, according to Chinese economists, Chinese macro-econometric forecasting is characterised by low accuracy. Indeed, it is these two characteristics of the Chinese economy that motivated our research into noise-immune forecasting algorithms.

The Group Method of Data Handling (GMDH) is an inductive modelling method, which was proposed by Ivakhnenko (1968). It has been shown to be robust in forecasting of noise-contaminated data (Müller and Lemke, 2000). The GMDH algorithm and its variants were successfully applied in Chinese economy forecasting with very promising results (He, 2005; He and Ma, 2007). However, these models produced unsatisfactory performances when applied to Sichuan province economic data. Methodical analysis of the GMDH results indicated that it is the generation of initial input models (see Section 3.1) which leads to poor performance.

In this research, we propose a new method to improve the accuracy of the solution and its robustness to noise. The new method, called D-GMDH, introduces diversity which is a key factor in ensemble learning in the GMDH algorithm to generate suitable initial input models. Ensemble learning is a machine learning paradigm, where multiple estimators are trained to solve the same problem (Zhou, 2009). In order to achieve good performance in ensemble learning, the diversity of base estimators needs to be maintained. Experiments have also demonstrated that diversity assists with the task at hand, when the data contains noise (Zhang and Zhang, 2008, Zhang et al., 2008). The new approach D-GMDH is compared with the benchmark ensemble methods, random forests and bagging, which also use diversity (Dutta, 2009). Then, D-GMDH is applied into a real-world dataset. It is compared with the traditional GMDH, GMDH combination model and ARMA in economic dataset of Sichuan province.

The rest of the paper is organised as follows. The next section is the overview of related work, and Section 3 presents the fundamental principles of the GMDH and introduces the concept of diversity. Section 4 describes the main contribution of this research, i.e., the D-GMDH method. Experimental results and comparisons are given in Section 5. The paper provides the conclusion of the work and suggestions for further research in Section 6.

2. Overview of related research

Economy forecasting methods can be roughly divided into three categories: econometric models (Abeysinghe and Rajaguru, 2004; Chow

* Corresponding author. Tel.: +86 28 85418891.
E-mail address: hcz.scu@gmail.com (C. He).

0264-9993/$ – see front matter. Crown Copyright © 2012 Published by Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.econmod.2012.09.021

and Li, 2002; Clements and Hendry, 1995; Ho and Tsui, 2004; Qin et al., 2007), neural network techniques (Lin and Zhu, 2006; Swanson and White, 1997; Yu et al., 2010) and simulation methods (Gu and Tang, 2003; Mai, 2006; Tay and Wallis, 2000; Wang et al., 2010). However, the aforementioned models do not pay attention to the presence of noise in economic data. Giles et al. (2001) demonstrated that the performance of economic forecasting methods, such as econometric methods and artificial neural networks, deteriorates with the level of noise in the data. When data contains noise, the most dangerous thing is over fitting, which implies that models tend to be excessive complex, and have poor generalization (Tan et al., 2005; Wong et al., 2010). Thus, economic forecasting using noisy and volatile datasets is an important topic, which is the focus of the current research.

GMDH method is a means of addressing such problems. The underlying concept of GMDH is based on the principles of self-organisation in biological evolution. It has been widely used for forecasting in various application areas, including energy demand (Howland and Voss, 2003; Srinivasan, 2008), mobile communications (Hwang, 2006), as well as industrial companies, e.g., Boeing, Mobil, Merck(http://gmdh.net). In the area of financial forecasting, He (2005) used GMDH to forecast the GDP of Chengdu (the capital of the Sichuan province). He and Ma (2007) applied GMDH in forecasting macro-economic parameters of the Sichuan province, while Tengger and He (2010) used it to forecast Chinese GDP.

Besides application contributions, there have been numerous works describing approaches in improving the forecasting accuracy of GMDH. For instance, He et al. (2008) proposed an optimal cooperation strategy between external criterion and data division. Analog Complexing algorithm has been introduced by Lemke and Müller (1997) to improve forecasting accuracy. A two-level GMDH algorithm has been shown to provide an effective way of expanding forecasting range (Ivakhnenko and Stepasko, 1985; Tian, 1997). Kim et al. (2009) proposed GMDH-type neural networks to improve forecasting accuracy. The determination of initial input models which is one of the main contributions of the current work is a key question in GMDH modelling and has been considered by various authors (Madala and Ivakhnenko, 1994). The reference functions, such as simple K–G polynomial and harmonic functions (Madala and Ivakhnenko, 1994) use special functions as initial input models, and in this paper we propose an approach to generate the initial input models.

Ensemble learning is a popular subject of machine learning research (Brown et al., 2005), and its generalization ability is usually much stronger than that of base estimators (Zhou, 2009). To achieve high accuracy, maintaining the diversity of base estimators is a key aspect. As Krogh and Vedelsby (1995) first showed, base estimators should be as accurate and diverse as possible. Dutta (2009) proposed metrics for measuring diversity in regression ensemble, namely the correlation coefficient, covariance, chi-square, disagreement measure and entropy. He used the concept of diversity to select the models (base estimators) to be used in the prediction, and compared the forecasting results with those of the larger original ensembles.

Diversity in ensemble learning is helpful when the dataset is contaminated with noise. Melville and Mooney (2004) found that diversity in ensembles results to higher forecasting accuracy even in noisy datasets. Nicolás and Domingo (2010) demonstrated that the approach for ensemble construction based on the use of supervised projections, to achieve both accuracy and diversity of individual classifiers, shows higher levels of robustness in the presence of noise when compared to Adaboost. Xiao et al. (2010) showed that the dynamic classifier ensemble selection strategy (GDES-AD) results to better noise-immunity than other strategies, when considering both accuracy and diversity in ensemble selection. Zhang et al. (2008a, 2008b) found that the modified boosting algorithm based on the use of diversity–accuracy patterns improves the prediction accuracy and robustness to classification noise of the Adaboost. The experiments demonstrated that the proposed Boosting algorithm provides higher prediction accuracy when compared to other methods, with increasing noise levels.

3. Basic theory

GMDH is an inductive modelling method that constructs a hierarchical (multi-layered) network structure to identify complex input–output functional relationships from data. The process of GMDH is analogous to the natural evolution of wheat. To obtain wheat with certain property, a large number of wheat are sown which may have this property. From the harvest of the first generation, wheat which better satisfy the requirements as compared to others is chosen. The seeds of the wheat are sown again. From the second harvest certain seeds are once again selected and sown. After several generations, some wheat will be obtained in which the desired property is more predominant than in others.

Similarly, the process of GMDH is a self-organising process based on sorting-out of gradually complicated models and selection of the best solution by external criterion. GMDH first produces some simple elementary models by reference functions and uses them as initial input models at the start of modelling process. After generating a large number of competition models by these initial input models (inheritance), the algorithm selects certain more optimal intermediate models (selection), so that a large number of new competition models are generated by these intermediate models. Such procedure of inheritance and selection is repeated until an optimal complex model is created. According to the theory of optimal complexity, as the complexity of the model increases, the value of external criterion usually decreases first, then reaches a minimum and later starts to increase again. The GMDH algorithm will stop when the external criterion reaches its minimum and an optimal complex model is obtained (Madala and Ivakhnenko, 1994).

3.1. Initial input models of GMDH

The initial input models v_i of the GMDH algorithm are created by reference functions, typically Kolmogorov–Gabor (K–G) polynomials. Assuming a model with two inputs x_1 and x_2, the corresponding K–G polynomial is given by:

$$y = a_0 + a_1x_1 + a_2x_2 + a_3x_1^2 + a_4x_2^2 + a_5x_1x_2 \tag{1}$$

where the set of initial input models is $\{v_1 = a_0, v_2 = a_1x_1, v_3 = a_2x_2, v_4 = a_3x_1^2, v_5 = a_4x_2^2, v_6 = a_5x_1x_2\}$, and the parameters/weights a_i are real numbers.

Various reference functions can be used in the initial input models depending on the nature of the dataset. For instance, if the sample data appears to show periodicity, harmonic functions are normally chosen (Madala and Ivakhnenko, 1994).

3.2. Existing external criteria in GMDH

For a forecasting model, the dataset W is divided into three disjoint subsets, $W = A \cup B \cup C$, where A, B and C are the training set, the testing set and the validation set, respectively. For each layer, the external criterion values are calculated to select the best model candidates, which become the input models of the next layer.

A central problem of modelling is choosing a "best" model from a class of competing, similar models. The optimal structure of a model is detected by the difference between model output y^M and observed output y that can be measured by:

$$J(\ C) = E\left\{Q\left(y^M, y\ , C\right)\right\} \tag{2}$$

where Q is loss function, E is expected value.

Currently, external criteria have two main categories, i.e., Criteria of accuracy and Criteria of consistency.