Expert Systems With Applications 42 (2015) 9105–9116

Contents lists available at ScienceDirect

Expert Systems With Applications

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

An improved grey relational analysis approach for panel data clustering

Xuemei Li [a,b,c], Keith W. Hipel [b,d,*], Yaoguo Dang [c]

[a] School of Economics, Ocean University of China, Qingdao, 266100, PR China
[b] Department of Systems Design Engineering, University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada
[c] College of Economics and Management, Nanjing University of Aeronautics and Astronautics, Nanjing, 211106, PR China
[d] Centre for International Governance Innovation, Waterloo, Ontario, N2L 6C2, Canada

ARTICLE INFO

Keywords:
Clustering
Panel data
Grey relational analysis
Chinese panel data

ABSTRACT

An enhanced grey clustering analysis method based on accumulation sequences using grey relational analysis (AGRA) is put forward for specifying hierarchies of clusters in panel data. The clustering method can handle panel data containing N samples, each of which has m time series of indicators for which the observations for a given time series can be measured at different times than other series and contain different numbers of data points compared to other series. The overall clustering approach, which is called the Mean-AGRA clustering method, contains three main parts: a sequence of transformations of each separate time series; appropriate pairwise comparisons of the grey relational degree of an AGRA model for each pair of samples, across all samples as well as appropriate combinations thereafter, for three specific types of grey relational degrees; clustering all samples according to their AGRA degrees. To demonstrate how this new clustering method can be utilized in practice, it is applied to panel data consisting of 12 natural environmental indicators and 8 societal time series ($m = 20$) for 30 provinces ($N = 30$) in mainland China. The findings clarify how, for example, the provinces in China can be meaningfully categorized according to topography into two main groups consisting of plateaus and plains. The new method can handle different lengths of time series within a sample and across samples, which is useful when values occur at different times when comparing any two series. Moreover, the new clustering method avoids the problem of combining two samples having a limited degree of similarity, which exists in the traditional method. Consequently, the AGRA model and Mean-AGRA clustering method have expanded the scope of application of grey relational and clustering analysis.

1. Introduction

Grey relational clustering analysis clusters objects according to their relational degrees and groups similarly clustered objects under one category. When dealing with two-dimensional data, the existing grey relational analysis (GRA) models form the basis for grey relational clustering analysis (Du, Shi, Liu, & Fang, 2010; Guo & Zhang, 2010; Li, Dang, & Wang, 2015a; Liu, Yang, Cao, & Xie, 2013; Sun & Dang, 2008). Different kinds of GRA models can be selected depending on practical issues. GRA has been employed in various fields, where it has produced promising results, as it is not limited to data distribution and quantity (Yin, 2013). GRA works effectively in combination with clustering (Abdulshahed, Longstaff, Fletcher, & Myers, 2015; Guo & Zhang, 2010; Wu, Lin, Peng, & Huang, 2012), decision-making (Jin, Mi, Xu, Wang, & Wei, 2012; Li, Chen, & Xiang, 2015; Wang & Dang, 2009; Wei, 2011; Zhang & Liu, 2011), evaluation (Jung & Kwon, 2010; Rajesh & Ravi, 2015; Wong & Hu, 2013), optimizing methods (Ke, Liu, Chen, & Fang, 2010; Pophali, Chelani, & Dhodapkar, 2011), fuzzy theory (Tang, 2015; Yang & Huang, 2012), forecasting method (Hus, 2011; Jiang & He, 2012; Wang, Wang, & Jiang, 2015), case adaption (Zhang, Li, & Wang, 2015) and extreme learning machine (Yu, Hui, & Choi, 2012). Additionally, the GRA model has been widely applied in the study of societal economics and engineering practice problems, such as in determining regional economy development gaps (Wang, 2010), identifying the factors that influence service brand equity (Wei, Zhang, Liang, & Yao, 2011), tuning PID control parameters for micro-piezo-stage (Lin, Chiang, & Lin, 2011), the identification of factors influencing a company's financial performance (Kung & Wen, 2007), the evaluation of software (Huang, Chiu, & Chen, 2008), assessing the relationship between China's economic development and energy consumption (Yuan, Liu, Fang, & Xie, 2010), examining the utilization (Ma, Wang, Wu, & Tu, 2014) and the sustainability (Liu, Baniyounes, Rasul, Amanullah, & Khan, 2013) of renewable energy, determining the representative elementary volume (REV) of rock mass (Tan, Chen, Que, & Zhang, 2012), establishing the law of maximum flood peak in the upstream of the Yellow River (Wang, Chen, & Yan, 2013), investigating the distributed

* Corresponding author at: Department of Systems Design Engineering, University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada. Tel.: +1 519 888 4567 32830.

E-mail addresses: xuemeiliwinter@gmail.com (X. Li), kwhipel@uwaterloo.ca (K.W. Hipel), iamdangyg@163.com (Y. Dang).

http://dx.doi.org/10.1016/j.eswa.2015.07.066
0957-4174/

multiple flight path (Yi, Zhang, Cai, & He, 2013), 2D object recognition (Sun, Horng, Liu, & Tien, 2009) and failure mode and effects analysis (Liu, You, Fan, & Lin, 2014).

1.1. Related work

There are two issues that need to be handled during grey relational clustering analysis of panel data. One issue is developing a GRA model for panel data for which there has been some research. Zhang and Liu (2010) discussed similarity of geometric features in three-dimensional space and provided new ideas about grey relational clustering analysis of panel data. Wu and Liu (2013) defined the degree of convexity using a Hessian matrix, characterized the similarity between samples according to convexity and proposed grey relational clustering analysis of panel data based on grey relational analysis of convexity. Qian, Wang, and Dang (2013) provided a grey matrix relational analysis model to determine clusters for samples. Liu, Dang, Qian, and Zhou (2014) proposed a grey grid relational model according to the geometrical characteristics of panel data. Li, Dang, and Wang (2015b) made a grey relational model for panel data to cluster indicaters.

The second issue is designing a clustering method based on grey relational degree. The existing grey relational clustering analysis methods for panel data use the clustering technique called the critical value clustering method put forward by Liu, Dang, and Fang (2004). Similar to grey relational clustering analysis, most of the expansions of other traditional clustering analysis propose new distance or similarity measures for panel data (Assmann & Boysen, 2011; Juárez & Steel, 2010; Li, Dai, & He, 2013; Zhu & Chen, 2007). Some clustering methods first reduce dimension or transform panel data before calculating clusters (Xu & Fang, 2010).

1.2. Research motivation

Research on grey relational analysis and grey relational clustering analysis is required in order to make these approaches more comprehensive. Some problems and related improvements are as follows: (1) The results of calculations for some grey relational degree procedures may change if the order of the samples is altered (Liu et al., 2014; Wu & Liu, 2013; Zhang & Liu, 2010). This is because these methods only consider comparisons between adjacent samples. To overcome this, all pairwise comparisons of samples must be considered in the calculations. (2) The existing relational, evaluating and decision models for panel data are limited to time series data having equally-spaced observations at exactly the same time and over the same time period. If the number of observations in the time series varies, some researchers only use these observations for the time series that correspond to the shortest series. When a gap exists in a time series, some researchers estimate the missing value by using the average of the two neighboring observations or some other type of interpolation technique. Hence, any trend that may exist in a time series across all observations may not be properly taken into account. (3) Existing relational models may not preserve certain mathematical characteristics such as what are referred to as the "index property" and the "inner property" in the original observations of the given time series. (4) The critical value clustering method put forward by Liu et al. (2004) can be used as a kind of fast clustering method. However, sometimes faulty results may be obtained. (5) When clustering samples according to distance, incorrect clustering may occur in some circumstances. Therefore, in this research similarities in trends are used for clustering purposes.

Liu et al. (2013) state that grey relational theories and practical achievements mainly concentrate on cases in which the research objects and their behavioral characteristics are described by real number sequences. Additionally, the study of high-dimensional models is just starting. In fact, a large number of practical and scientific problems remain to be solved through analytical methods based on panel data, matrix data, matrix sequence data and high-dimensional data. Panel data contain observations of multiple phenomena obtained over multiple time periods for the same firms or individuals, so individuals and indicators can both be clustered if panel data is obtained. Panel data clustering is an indispensable component in decision making and expert analysis.

1.3. Solution in this paper

In order to solve the aforementioned problems, the AGRA model based approach for panel data clustering of samples is proposed in this paper. As indicated in the right column of Fig. 1, the overall idea of this model is to firstly transform the original time series to obtain an accumulated sequence, which is then simulated. Then, the generation rate sequences for all samples, which can be obtained using the former simulation functions, are used to calculate grey relational degrees. The AGRA model is composed of three kinds of grey relational degrees. As shown in the bottom rectangle in the right column of Fig. 1, the Mean-AGRA model is used for carrying out the clustering calculations.

As depicted on the left in Fig. 1, the paper is structured as follows. Data transformations are studied in Section 2. In Section 3, the AGRA Model for Panel Data is established. Next, in Section 4, the clustering method for panel data based on the AGRA model is put forward and the steps followed in the Mean-AGRA clustering method for panel data are provided. An application using the model is furnished in Section 5, while in Section 6 appropriate conclusions are drawn.

2. Data transformation for panel data

Some basic concepts pertaining to data transformations for panel data are defined in this section. These definitions are needed to establish the AGRA model in the subsequen sections.

Definition 1. (*panel data*) (Zhang & Liu, 2010). The value of indicator $i\,(i = 1, 2, \ldots, m)$ for sample $s\,(s = 1, 2, \ldots, N)$ at time $t(t = 1, 2, \ldots, t_i^s)$ is $x_s^{(0)}(i, t)$. Then, the matrix containing the m indicators time series for sample s $(s = 1, 2, \cdots, N)$ is

$$X_s^{(0)}(i, t) = \begin{bmatrix} x_s^{(0)}(1,1) & x_s^{(0)}(1,2) & \ldots & x_s^{(0)}(1,t_1^s) \\ x_s^{(0)}(2,1) & x_s^{(0)}(2,2) & \cdots & x_s^{(0)}(2,t_2^s) \\ \vdots & \vdots & \ddots & \vdots \\ x_s^{(0)}(m,1) & x_s^{(0)}(m,2) & \cdots & x_s^{(0)}(m,t_m^s) \end{bmatrix} (s = 1, 2, \ldots, N)$$

Panel data $X = (X_1^{(0)}(i,t), \ldots, X_s^{(0)}(i,t), \ldots, X_N^{(0)}(i,t))$ $(i = 1, 2, \ldots, m, t = 1, 2, \ldots, t_i^s)$ constitutes the matrix representation of all samples.

Research on GRA models indicates that the existing relational models may not preserve certain mathematical characteristics like the "index property" and the "inner property" in the original observations of the given time series. From the viewpoint of grey systems, although a particular system may appear to be complex, the system always possesses an operational capability or function and follows inherent laws. In fact, there is a method called grey generation which can explore the inner property of the original system as reflected in observations regarding the system. Accumulation generation is an effective kind of grey generation, which whitens grey processes and reveals the development tendency or trend in the accumulation process. Original data may seemingly not adhere to regular laws but emerge as a result of integrated characteristics of accumulation generation (Liu et al., 2004). Because economic, ecological and agricultural systems can be treated as generalized energy systems,