Predictive Modeling of Large-Scale Sequential Curves Based on Clustering

Wen Long[1] and Huiwen Wang[2]

[1] Research Center on Fictitious Economy & Data Science, Chinese Academy of Sciences, Beijing 100080, China
longwen@gucas.ac.cn
[2] School of Economics & Management, Beijing University of Aeronautics & Astronautics, Beijing 100083, China
wanghw@vip.sina.com

Abstract. Traditional approach to predict large-scale sequential curves is to build model separately according to every curve, which causes heavy and complicated modeling workload inevitably. A new method is proposed in this paper to solve this problem. By reducing model types of curves, clustering curves and modeling by clusters, the new method simplifies modeling work to a large extent and reserves original information as possible in the meantime. This paper specifies the theory and algorithm, and applies it to predict GDP curves of multi-region, which confirms practicability and validity of the presented approach.

Keywords: curves clustering, predictive modeling, large-scale curves, SOM.

1 Introduction

A series of sequential data can draw a curve, which depicts dynamic information clearly. Nowadays the amount of data which is stored in databases increases very fast; and the situation of large-scale sequential curves often happens in the analysis work, such as GDP curves of multi-regions, sales volume curves of products, observed clinical data of large sample. How to analyze and predict a large set of curves is worthy to be further studied.

Traditional solution to the problem is building model separately according to every curve, that is, one curve needs one model. The method can obtain the predictive results accurately, however modeling work will become rather heavy and complicated inevitably while the analysis objects are large scale.

Some illuminating work seems to have opened a new prospect. G. Hebrail [1, 2] has performed clustering to 2,665 electric load curves in order to distinguish diverse electric consumption patterns, and got pleasant results. Other related work [3, 4] also proved that clustering is an effective approach to describe and analyze huge data sets or sequential curves.

However, this paper is focused on predictive modeling on a large set of curves, which is different from their work. Therefore, it's expected that the shape of curve can display some visible regularity so as to be convenient for choosing appropriate predictive model. Yet in fact, original data usually exhibit abundant curve configuration, which makes it difficult to choose predictive model.

M. Bubak et al. (Eds.): ICCS 2008, Part II, LNCS 5102, pp. 486 – 493, 2008.

Consequently, this paper proposes a new approach to solve the problem of predictive modeling of large-scale curves. Basic idea of the new approach is to perform some transformation to original data for the sake of eliminating scale and reducing types of models, and clustering to the preprocessed curves based on some certain clustering technique, then build predictive models by clusters, finally calculate the predictive value of original data through inverse algorithm.

This paper is organized as follows. In Sect. 2, the method of curves clustering modeling will be introduced, which will be specified by three parts, including reduction of models' types, curves clustering and predictive modeling by cluster. Sect. 3 will state the method of Self-Organizing Map briefly, which is used in the curves clustering. In Sect. 4, applying this new approach, a case will be studied concerning predicting GDP of 133 countries and regions. At last Sect. 5 summarizes the results.

2 Predictive Modeling of Large-Scale Curves Based on Clustering Method

2.1 Reducing Types of Models

One problem of existing approach to predictive modeling of large-scale curves is that modeling work is too much to apply in the applications. Since types of models decide number of modeling, the models' types must be decreased.

In the problem of predictive modeling of curves, the model's type depends upon two factors, scale of original data and shape of original curves. Therefore, the reduction of models' types must be based on the unification of scale of original data and simplification of shape of original curve.

In the social and economic cases, usually the data sequences are positive, so this paper only discusses the situation while $x_{it} > 0$ ($i = 1,\cdots,n; t = 0,1,\cdots,T$).

Given original data sequences described by a set of vectors $\boldsymbol{X}_i = \{\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \cdots, \boldsymbol{x}_{iT}\}$, $(i = 0,1,\cdots,n)$. Then the development speed with link relative method of individual i at the time t is defined as

$$a_{it} = \frac{x_{it}}{x_{i(t-1)}}, \quad (t = 1,\cdots,T) \ . \tag{1}$$

Accumulate development speeds with link relative method, obtain

$$\begin{aligned} b_{i1} &= a_{i1}, \\ b_{it} &= b_{i(t-1)} + a_{it}, \quad (t = 2,\cdots,T) \ . \end{aligned} \tag{2}$$

Here, accumulated development speed curve $\boldsymbol{b}_i = \left(b_{i1}, b_{i2}, \cdots, b_{iT}\right)$ is an increasing one whose scale has been eliminated and tendency grows steadily.

After calculating the velocities of development with link relative method $\boldsymbol{a}_i = \{a_{i1}, \cdots, a_{iT}\}$ and accumulated the velocities of development $\boldsymbol{b}_i = \{b_{i1}, \cdots, b_{iT}\}$, the diversity of original curves due to scale has been obviated, and accumulated curves $\boldsymbol{b}_i$

demonstrates good configuration with increasing steadily. As a result, the types of curves have been largely decreased, and the accumulated curves $\boldsymbol{b}_i$ will be more convenient to clustering.

2.2 Curves Clustering

Clustering can be loosely defined as the process of organizing objects into groups whose members are similar in some way. The standard way of analyzing a large set of curves is to perform a clustering of the curves, so that experts look at a small number of classes (i.e. clusters) of similar curves instead of at the whole set of curves [2].

The direct objective of curves clustering in this paper is to build models by cluster. Therefore, what are used to perform clustering are not original curves $\boldsymbol{x}_{it}(t=0,1,\cdots,T)$, but accumulated velocities of development curves $\boldsymbol{b}_{it}(t=1,\cdots,T)$ that benefit modeling. If curves belonging to the same cluster display similar curve features and evident cluster effect, the result of clustering is regarded as pleasant.

In this paper, the method of Self-Organizing Map will be applied to perform clustering, that will be introduced at Sect. 3.

2.3 Modeling by Cluster

2.3.1 Basic Idea of Modeling by Cluster

A hypothesis must be given before modeling by cluster that those curves which belong to one cluster have same or similar dynamic trends, that is, they grow at the same or similar pattern.

Perform curves clustering to accumulated velocities of development curve $\boldsymbol{b}_{it}(t=1,\cdots,T)$. However, if the clustering result is not pleasant and the curves belonging to a certain cluster are dispersed each other, for the sake of lowering the loss of original information, a second clustering should be performed to them, till the pleasant result of clustering bas been obtained.

As for the clusters with good clustering result, the work of modeling by cluster can be performed. Provided that n curves are sorted into l classes ($l \leqslant n$), and cluster k includes n_k curves, define

$$y_t^k = \overline{b}_t^k = \frac{1}{n_k}\sum_{i=1}^{n_k} b_{it}^k, \quad (t=1,\cdots,T) \ . \tag{3}$$

Build predictive model $y^k = f_k(t)$ according to $\boldsymbol{y}^k = \{y_1^k, y_2^k, \cdots, y_T^k\}$, $(k=1,\cdots,l)$. Based on the hypothesis of modeling by cluster stated above, this model can explain all the curves contained in cluster *k*.

Consequently, the amount of modeling has been reduced from the amount of individuals n to that of classes l, which has simplified the modeling work to a large extent, also reserved original information as possible.

2.3.2 Algorithm of Returning to Original Data

If $\hat{y}_{T+1}^k$ denotes the predictive value of $\boldsymbol{y}^k$ at future time T+1, $\hat{y}_{T+1}^k$ can be calculated by the predictive model of cluster j, that is $y^j = f_j(t)$.

Calculate the development speed of cluster j at time T+1 according to $\hat{y}_{T+1}^{j}$, obtain

$$a_{T+1}^{j} = \hat{y}_{T+1}^{j} - y_{T}^{j}, \quad (j = 1, \cdots, l) \ . \tag{4}$$

Then, the predictive value of individual i at the time $T+1$ can be obtained as follow

$$\hat{x}_{i(T+1)}^{j} = x_{iT}^{j} \times a_{T+1}^{j}, \quad (j = 1, \cdots, l, \ \ i = 1, \cdots, k) \ . \tag{5}$$

3 Neural Network of Self-organizing Map

In the method introduced in Sect. 2, neural network of Self-Organizing Map is applied to perform curves clustering, which can exhibit visualized results [5].

Self-Organizing Map（SOM）is a method of artificial neural network, proposed by Prof. Teuvo Kohonen of Finland in 1981. This network can simulate self-organizing map function of nervous system of brain. It is a competitive learning network that it can perform unsupervised self-organizing learning and learn from complex, multi-dimensional data and transform them into visually decipherable clusters. The main function of SOM networks is to map the input data from an n-dimensional space to a lower dimensional (usually one or two-dimensional) plot while maintaining the original topological relations [6].

The SOM network typically has two layers of nodes, input layer and competitive layer.

The nerve cells in the input layer are one-dimensional, and those in the competitive layer are two-dimensional. All the nerve cells both in the two layers connect each other. As the training process proceeds, the nodes adjust their weight values according to the topological relations in the input data. The node with the minimum distance is the winner and adjusts its weights to be closer to the value of the input pattern [7, 8].

Justification for weight vector can be explained by this formula as follows:

$$W_i(t+1) = W_i(t) + h_{c(x),i}(X(t) - W_i(t)) \ . \tag{6}$$

Here t denotes iterative number of input vector; $W_i(t)$ is weight vector and $X(t)$ is observed vector X at t^{th} iteration; $h_{c(x),i}$ is neighborhood function and $c(x)$ represents the winner.

Define neighborhood function $h_{c(x),i}$ as

$$h_{c(x),i} = \alpha(t) \exp(-\left\| r_i - r_c \right\|^2 / 2\delta^2(t)) \ . \tag{7}$$

$\alpha(t)$ denotes learning rate varying in [0,1] and descending with increasing of iterations. r_i, r_c are position vectors, corresponding to W_i, W_c, $r_i \in R^2$, $r_c \in R^2$. $\delta(t)$, descending with increasing of iterations, corresponds to the width of neighborhood function $h_{c(x),i}$ denoting $N_c(t)$.

The algorithm of SOM can be briefly described as follows.

1. Initialization of weight values by random and give an initial radius of neighborhood.
2. Input a new vector $X_j \in R^p, j = 1, 2, \cdots, n$.
3. Calculated the distances between X_j and all the output nodes.
4. Find the node *c* whose weight vector is closest to the current input vector X_j.
5. Train node *c* and all nodes in some neighborhood of *c*, and then modify the weight vector using formula (6).
6. Return to step 2 for $X_j^t, t = 1, \cdots, T$.

In the course of development of SOM, originally initial weight vector is chosen by random, that indicates SOM can self-organize toughly even though initial situation is out-of-order. But in practice, if initial weight vector can be selected by PCA rules, the constringency of algorithm will be accelerated.

4 Case: Predicting GDP Curves of 133 Countries and Regions

Based on the new approach proposed in this paper, a case is studied to predict the future GDP of 133 countries and regions according to historical GDP data from 1990 to 2002 [9].

Calculate the accumulated GDP velocities of development curves of 133 countries (or region). Then perform clustering to the accumulated curves using the method of Kohonen. Considering two factors of decrease of modeling and reservation of original information, this paper classified the 133 accumulated curves into 10 clusters (Fig. 1).

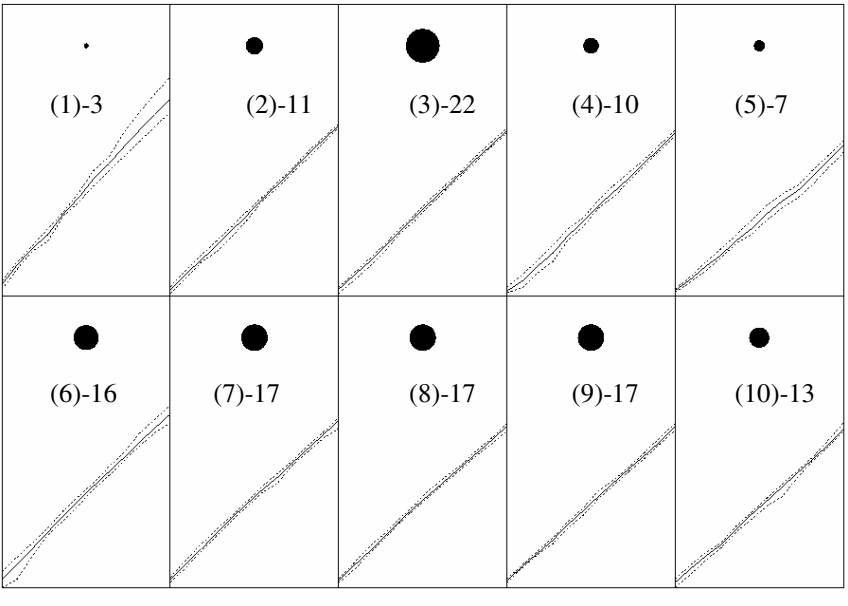

Fig. 1. Clustering result of 133 accumulated GDP development speeds curves

An example of cluster 3 is illustrated, which contains the most curves among these clusters. Cluster 3 includes 22 countries and regions, and there exists a great disparity in the size of GDP among them. Seen from Fig. 2, due to the GDPs of U.S.A. and Britain counted by $1,000 billions, being much greater than others, especially influenced by U.S.A., the curves of other 20 countries and regions pile up near the abscissa axis. In fact, even in the 20 countries and regions, there is a wide gap of GDP. For instance, GDPs of India, Norway and Saudi Arabia are above $100 billions; and those

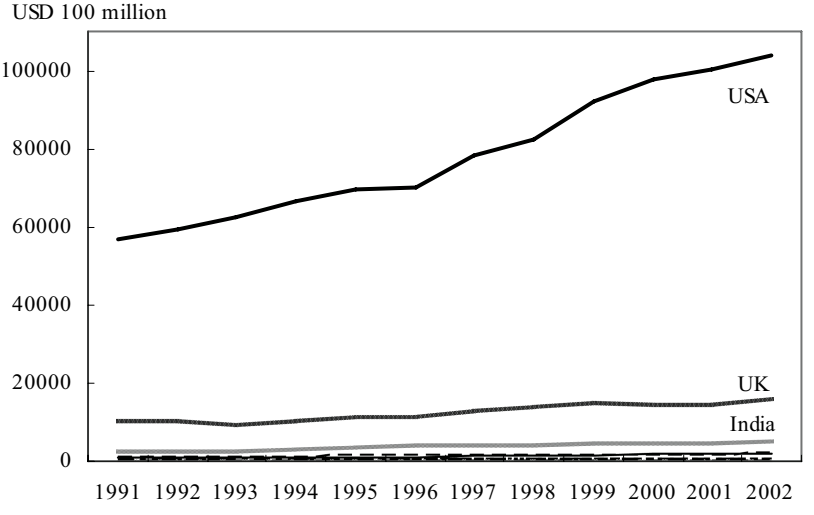

Fig. 2. 22 observed GDP curves of cluster 3

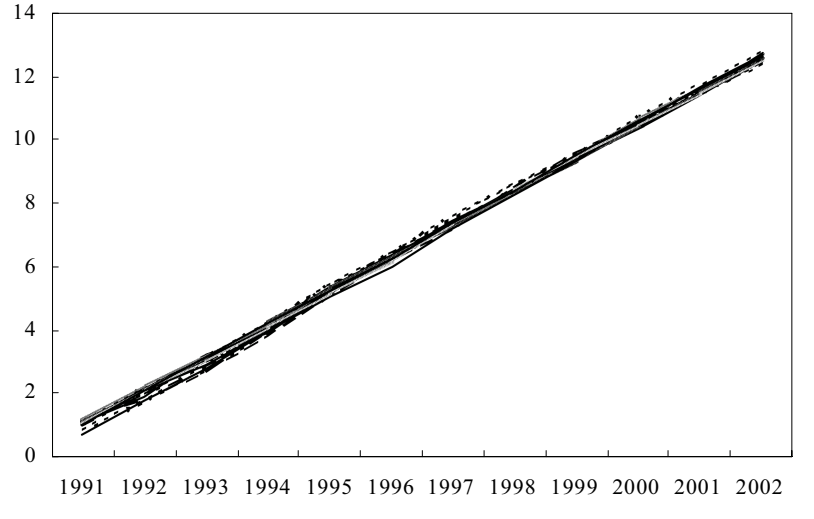

Fig. 3. 22 accumulated development speed curves of cluster 3

of Bengal, Hungary and Nigeria are counted by $10 billions; yet Haiti, Nepal and Malta produce GDP counted by billions. In addition, the configuration of curves doesn't show evident regularity.

However, we find that the curves display visible regularity after calculating accumulated development speeds of GDP (Fig. 3). 22 curves congregate closely, not only avoiding the influence of scale, but also presenting evident increasing trends. Obviously it becomes much easier to modeling in this instance.

Build models separately to 10 clusters of accumulated curves, and fitting values of every cluster can be obtained. Then the predictive GDP of each country (or region) can be worked out by the means mentioned at Sect. 2.3.

In order to test precision of the result calculated using the new method, this paper will measure the error by comparing the fitting values of GDP with observed ones. Table 1 has given the relative errors of GDP fitting values in 2001 and 2002.

Table 1. Relative error of GDP fitting values of 133 countries and regions in 2001 and 2002

Range of relative error		≤10%	10~15%	15~20%	>20%
proportion of countries (or regions) included	2001	80.5%	13.5%	3.0%	3.0%
	2002	73.7%	15.0%	7.5%	3.8%

Seen from Table 1, the accuracy of GDP fitting values in 2001 and 2002 seems satisfied and the relative error of most countries (or regions) limits in the range of 10%, such as U.S.A. (Fig. 4), whose observed data and fitting ones are very close. That further confirms the new approach has obtained pleasant results, which decreases the amount of modeling from 133 to 10 under the circumstances of keeping original information as possible.

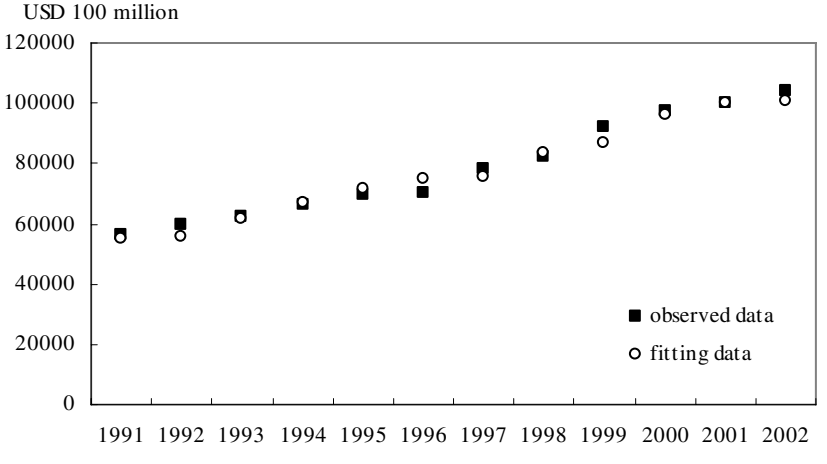

Fig. 4. Comparison between observed GDP data and fitting ones of U.S.A

5 Conclusions

This paper introduces a new approach to predict a large set of timing curves. The predictive precision of this method relates to two factors, one is amount of classes during clustering; the other is choice of model during predictive modeling.

In general, the more classes the curves are divided into while performing clustering to accumulated curves, the more information will be reserved while modeling. If the amount of clusters equals to that of individuals, the new approach will become to build model according to every curve, which is just the traditional method. Consequently there exists a contradiction between preserving original information and diminishing workload of modeling. The analyst must make a comprehensive decision to how many clusters should be classified into based on the specific data and clustering results.

The models of modeling by cluster also influence the final predictive precision. More the error of model for accumulated velocities of development shows, more the error of predictive data which passes from the model magnifies. So choosing model while perform modeling by cluster must be deliberate.

To sum up, the approach proposed in this paper offers an efficient and effective solution to predictive modeling of a large set of timing curves, which is also applicable to the related problems.

Acknowledgments. This research was supported by National Science Fund of China (NSFC). The authors wish to thank Prof. Georges Hébrail of ENST and Prof. Ruoen Ren of BUAA for helpful comments.

References

1. Chantelou, D., Hébrail, G., Muller, C.: Visualizing 2,665 electric power load curves on a single A4 sheet of paper. In: International Conference on Intelligent Systems Applications to Power Systems (ISAP 1996), Orlando, USA (1996)
2. Debrégeas, A., Hébrail, G.: Interactive interpretation of Kohonen maps applied to curves. In: International Conference on Knowledge Discovery and Data Mining (KDD 1998), New York (1998)
3. Guo, H., Renaut, R., Chen, K., Reiman, E.: Clustering huge data sets for parametric PET imaging. BioSystems 71, 81–92 (2003)
4. Jank, W.: Ascent EM for fast and global solutions to finite mixtures: An application to curve-clustering of online auctions. Computational Statistics & Data Analysis 51, 747–761 (2006)
5. Mingoti, S.A., Lima, J.O.: Comparing SOM neural network with Fuzzy c-means, K-means and traditional hierarchical clustering algorithms. European Journal of Operational Research 174, 1742–1759 (2006)
6. Kiang, M.Y.: Extending the Kohonen Self-Organizing Map Networks for Clustering Analysis. Computational Statistics and Data Analysis 38, 161–180 (2001)
7. Kohonen, T.: Self Organized Formation of Topologically Correct Feature Maps. Biological Cybernetics 43(1), 59–69 (1982)
8. Kohonen, T., Oja, E., Simula, O., Visa, A., et al.: Engineering application of the self-organizing map. Proceeding of the IEEE 84, 1358–1384 (1996)
9. International statistical yearbook. China Statistics Press, Beijing (1997-2004)