Principles for Development of Predictive Stability Models of Social and Economic Systems on the basis of DTW

*Aleksey Kislyakov**, and *Natalya Tikhonuyk*

Russian Academy of National Economy and Public Administration, Vladimir branch, Gorky Str., 59a, 600017 Vladimir, Russia

Abstract. This paper presents the concept for the development of predictive models of social and economic system evolution providing the necessity of combining solution search optimization algorithms and methods of regressive and clustering analysis for the adequate description of system attribute space. The rationale for the selection of metrics on the basis of a dynamic time-warping algorithm which allows to carry out clustering of the system attribute space. The example of solution of description task for COVID-19 pandemic development attribute for a particular country or region is considered. The developed concept formulates main provisions and indicators that can be used in order to increase the algorithm efficiency for the development of predictive complicated system models.

1 Introduction

The modern research methodology of social and economic systems represents the interpenetration process of principle, models and methods of related sciences [1] and is based on the concept of data driven managerial decision-making and accompanying by extensive studies of market status for stable development and detection of possibilities of effective warping in the changeable environment. In the context of the market economy digitalization, tasks for development of predictive models of social and economic system, as goods and services markets based on the consumer behavioural inquiry acquire critical importance and special actuality.

Any complicated system tendency to sustainability and self-organization [2], enables the summarization of approaches to the description and prediction of evolutionary processes in complicated systems [3]. The self-organization as the fundamental property of complicated self-developing system is expressed in the structural complexity of physical and biological systems, their evolutionary changes [4].

The absence of unifying concept and approaches to the description of evolutionary processes and the development of predictive models of complicated social and economic systems using the data mining algorithm and machine-learning tools results in the development of insufficiently correct, adequate and interpreted mathematic models.

* Corresponding authors: ankislyakov@mail.ru

The objective of this paper is to study the concept for development of predictive models of social and economic system evolution considering the symmetry breaking principle on the basis of time series cluster analysis.

The solution of scientific challenge is important not only from the point of basic science, but for the solution of challenge spectrum related to the management in various sectors of economy and separate organization in terms of infrastructure and organization changes in order to increase the resilience of social and economic systems to external factors' changes.

2 Methods

Key issues arising in the description of social and economic processes are the high degree of information redundancy and as a consequence the low degree of obtained results' interpretability. The development of complicated system predictive models requires from mathematic models and software tools increasingly more versatile and flexible approaches allowing to provide the model adaptation or learning in the fluid external conditions.

Usually, the system status characteristics are reflected as time series of separate variable (attribute) changes, in the result of which the prediction task reduces to the factor and regression analysis [5]. As for the task for sequential optimization of objective path, the model degenerates into the model of multiple time series for which several vectors of dynamically developing attributes are traced [6,7].

However, it is important not to just summarize which factors and to which extent impact the process on the basis of correlation analysis that summarizes the degree of relation between two attributes, but to perform the optimum comparison on the basis of economic attribute space dynamics analysis that enables to consider how factors acted relative to each other in different time intervals. The most interesting is the detection of factor similarity that allows to group them up, i.e. provide the transition to the cluster analysis [6,8].

Currently, there are the number of works [1,4], connecting evolutionary processes of complicated systems with the symmetry breaking event in the context of the interaction structure random variability between elements (subsystems) in physical, biological, social and economic systems expressed as time series. The complicated system symmetry breaking is of primary importance in studying the early prevention of financial crisis and economic risk accounting. In this case the key issue is the assessment of studied process noisiness by random event and detection of sustainable attributes, relations between elements of the complicated system.

The symmetry breaking is intended to explain the formation of behavioral patterns or structural changes in complicated systems. Provided that the primary system evolutionary resource is the mechanism for broken symmetry recovery and the system tendency to self-organization.

The effective approach to the development of predictive models of social and economic system evolution requires the combining not only genetic algorithms and novelty search algorithms, but the combining of cluster and regression analysis [9,10] for the adequate description of system attribute space.

In order to teach the predictive model and optimization algorithm, the selection of adequate metrics is required, based on which regression and clustering tasks will be solved.

In the literature [11] can be found the wide variety of distance metrics used in both cluster and regression analysis and learning tasks without the teacher. Among the standard measures are the Euclidean distance, the Manhattan distance, the Minkowskian distance, the Chebyshev distance and the Pearson correlation coefficient.

However, the comparison of time series requires metrics enabling the qualitative comparison the similarity of time variations of several attributes in different scale. In this

case, the particular importance is taken on the task of parallel comparison of such parameters on the basis of DTW-distance measure use. The name of this measure is derived from the dynamic time-warping algorithm [12].

The DTW algorithm allows to transform and scale the timeline in order to achieve the optimal comparison and/or optimal alignment of two time series to the intent that the distance between them would be minimum.

Since the number of possible warping paths increases exponentially with the length of compared sequences, then in order to find the optimal way during the finite time, it is required to introduce the series of restrictions.

- 1) The warping path shall be monotonic. This restriction is set using so-called step patterns, templates determining the direction of authorized transitions between table cells at each step.
- 2) The second restriction relates to the warping window width, within which it is allowed to make the warping path. There are multiple types of such window, but the most commonly the Sakoe-Chiba window is used [15,16] that covers a certain symmetric domain along the local loss matrix diagonal.

Fig. 1. An example of Sakoe-Chiba window

The Figure 2 shows one way to visualize the matrix, the found warping path and compared time series. Such transformation is called BandDTW [15] and used for acceleration of DTW algorithm operation by adding restrictions forcing the warping path to fall within the domain around the incidence matrix: if the optimal path crosses the line, the DTW distance will not be optimal.

The basic idea of the mentioned approach is in the dynamic use of the similarity and/or the correlation between time series. The more similarity between series is, the less space is require to calculate the DTW between them.

3 Results and Discussion

Let us consider this algorithm work on the example of actual in 2020 task of COVID-19 pandemic development predictive model for a particular country or region. As the initial data were used the shared data set in the system for data analysis competition organization and is available at: <https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset>.

The data set includes the number of affected cases, deaths and recovery by countries/regions within 5 months (from January 22 until June 28, 2020). In order to demonstrate the mentioned approach COVID-19 affected cases from 102 countries in which the number of affected cases exceeded 100 000 people as of end of June, 2020 were

considered. All time series are considered in a logarithmic scale for affected cases axis for the more qualitative clustering. The maximum Sakoe-Chiba window width in all cases amounted to 7 counts.

Figure 3 shows results of comparison and visualization of optimal warping path of time series reflecting in logarithmic scale the number of people affected by COVID-19 by countries: the standard – Russia, the query – India (a), China (b) and USA (c), and Figure 3 shows results of optimal comparison of time series.

Fig. 2. Examples of comparison and visualization of optimal warping path of time series reflecting in logarithmic scale the number of people affected by COVID-19 by countries: the standard – Russia, the query – India (a), China (b) and USA (c)

Where: Индексный номер запроса - Standard index number; Индексный номер эталона - Query index number.

Figure 4 shows that in case of the most similar development of the situation, e.g. as for pair Russia-India, the more symmetric warping path is observed rather than pair Russia-USA, where the similarity of series is minimum and the optimal warping path is asymmetric. For pair Russia-China the development identity is observed only for the last 100 days. Gray dotted lines show points only on the warping path. Thus, the optimal warping path is represented by the response function of either parameter to the other parameter change with certain time intervals that allow to carry out clustering with the bigger number of attributes.

Fig. 3. Results of optimal comparison of time series of COVID-19 affected cases in Russia (black solid line) and India (a), China (b) and USA (c) (red dashed lines)

Where: Индексный номер запроса - Standard index number; Индексный номер эталона - Query index number; Запросный ряд - Query series.

Fig. 4. Results of hierarchic agglomerative clustering by countries

It shall be noted that the number of clustering groups is unknown in advance, so it is not reasonable to use the clustering on the basis of the k-means method. The clustering using algorithms based on the decision tree requires the test learning sample. However, in the mentioned example the sample can dynamically change, for this reason hierarchic clustering methods are used as the most appropriate method in this paper (Fionn Murtagh, Pedro Contreras, 2011; Shreya Tripathi, Aditya Bhardwaj, Poovammal E., 2018), not requiring the learning sample and enabling time series division into groups.

Fig 5. An example of time series clustering for 102 countries

Where: Количество заболевших, чел (логарифмический масштаб) - Affected cases, people (logarithmic scale); Отсчет времени по дням - Time count by days.

The clustering results represented by dendrogram have already allowed to evaluate the degree of time series similarity and visually detect anomalies in time series dynamics (for example, for China). Figure 6 shows the clustering results as dendrogram for 102 countries. Dotted lines show time series representing cluster centroids. As an example, was considered the division into 9 clusters.

The cluster No. 3 consists of only one time series – China, which affected cases’ dynamics differs from other countries worldwide. In such case Russia will fall within cluster No. 2 with Brazil, Canada, India, Turkey, etc.

The obtained results allow to assess the development of the most common situations and time series of which countries are similar to each other over time. Thus, the described concept of time series prediction on the basis of attribute dynamics study and allow to evaluate common trends and the process dynamics.

4 Conslusions

The presented indices allow to form the unified concept of complicated system study on the basis of time series cluster analysis considering the system uncertainties consisting of the following conclusions:

1. The concept of predictive model development on the basis of symmetry breaking allows to detect the attribute stability and relations between elements of the complicated system.
2. The use of DTW-metrics allows to assess the similarity and/or the correlation between time series in different time intervals.
3. The cluster analysis allows to describe the complicated system attribute space considering the most prominent interactions.

4. The symmetry breaking of optimal warping paths shows directions of the system structural changes and its response to environmental changes and the division of clusters allows to describe system state attributes by groups in general.

The shown clustering example is executed for the pre-set number of clusters, the choice of which was not rigorously substantiated that complicates the further interpretation of results. Despite of the conducted study many applied and theoretical issues of symmetry breaking remain in abeyance setting the stage for further studies in this sphere.

Acknowledgements

The study was conducted under the sponsorship of the grant of the Russian Foundation for Basic Research No. 18.07.00170.

References

1. V.G. Rau, K.A. Gorshkov, S.V. Polyakov, T.F. Rau, A.N. Kislyakov, I.A. Togunov, N.E. Tikhonyuk. The multi-authored monograph: *The Study of Broken Symmetry Group Theory in Natural, Biological, Social and Economic Systems*, 67 (2020)
2. M.A. Deryabina, *Economic Issues*, **7**, 73 (2019)
3. Ya. Omelyanenko, *Evolutionary Neural Networks in Python*, 123 (2020).
4. G. Nikolis, I. Prigozhyn, *Perception of Sophistic. The Introduction*, 50 (2012)
5. Frank Nielsen, *Introduction to HPC with MPI for Data Science*, 282 (2016)
6. E.A. Pechenyy, N.K. Nuriev, S.D. Starygin, *Mathematical Methods in Engineering and Technologies: Collection of International Scientific Conference Papers*: in 12, 3, 19 (2019)
7. Mishra Brojo, Hazra Deepannita, Tarannum Kahkashan, Kumar Manas, *Business Intelligence using Data Mining techniques and Business Analytics*, 544 (2016)
8. Alboukadel Kassambara, *Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning (Multivariate Analysis)*, **1**, 1st Edition, 188 (2017)
9. A.N. Kislyakov, *Bulletin of NSUEM*, **1**, 64 (2020)
10. A.N. Kislyakov, *Proceedings of Saint-Petersburg State University of Economics*, **3(123)**, 110 (2020)
11. V.K. Shytikov, S.E. Mastitsky, *Classification, Regression, Data Mining Algorithms using* (2017)
12. S.E. Mastitsky, *Time Series Analysis using* (2020)
13. T. Giorgino, *Journal of Statistical Software*, **31(7)**, 1 (2009)
14. Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, *An Introduction to Statistical Learning with Applications in R*, 436 (2013)