Multi-country analysis of the COVID-19 pandemic typology using machine learning and neural network algorithms

Vladimir Malugin
Belarusian State University,
Minsk, Belarus,
malugin@bsu.by

, Akim Sergeev
Belarusian State University,
Minsk, Belarus,
akvise@gmail.com

Aleksandr Solomevich
Belarusian State University,
Minsk, Belarus,
alexandr.solomevich@gmail.c

***Abstract* – The paper presents the results of a multi-country analysis of the intensity typology of the COVID-19 pandemic in 30 countries of the European region based on publicly available and regularly updated panel data for the entire period 2020-2022 of high pandemic activity. In the generated space of classification features, using cluster analysis algorithms, all countries are divided into three classes, which differ in the intensity of the epidemic process. Based on the obtained country ratings, an integral statistical indicator of the COVID-19 pandemic is constructed. A set of discriminant analysis machine learning and neural network algorithms are used to estimate current as well predict the expected class of the epidemic state based on the newly acquiring data.**

Keywords — COVID-19 typology, multi-country analysis, country ratings, integral pandemic indicator, machine learning, neural network.

I. Introduction

The problem of analyzing the COVID-19 pandemic in various aspects is given considerable attention in the world scientific literature [1]. An important direction in the ongoing research is the development of methods for statistical analysis of the COVID-19 pandemic based on the data available in the mode of regular updating. Both simulation [2] and statistical models [3] are used to analysis and short-term predict the epidemic process at the level of individual countries. Considerable attention is paid to the tasks of analyzing the COVID-19 pandemic in a multi-country aspect [1, 4].

Previously, in [5] the following main problems were solved: development of statistical methods for classifying countries by the intensity of the COVID-19 pandemic based on the available unclassified data, starting from the first wave; construction of statistical indicators characterizing the intensity of the pandemic at the country and multi-country levels. The purpose of this study is to assess the current class of the epidemic state on the base of real-time updating data. Various machine learning and neural network algorithms estimated on the training sample are used and compared for classification accuracy.

II. Multi-country analysis of the COVID-19 typology using unclassified sample

A. Problem statement

It is assumed that the available panel data include the values of N indicators of epidemic process obtained for some sample of countries of volume n at time moments

$$t(t=1,\dots,T):$$
$$x_{i,t}=(x_{i1,t},\dots,x_{iN,t})'\in\Re^N\,(i=1,\dots,n,t=1,\dots,T).$$

In the context of the COVID-19 analysis, panel data have a heterogeneous cluster structure. It is supposed that the most important factor of heterogeneity is the difference between countries by a latent feature, which characterizes the intensity of the COVID-19 epidemic process. According to this property, countries can be assigned to one of the L classes. This property is expressed by a discrete random variable $d_{it}\in\{1,\dots,L\}$, indicating the class number for country *i* at time *t*. Class numbers $\{d_{it}\}$ are interpreted as country ratings of the intensity of the epidemic process.

The problem of statistical classification: to divide the sample $\{x_{i,t}\}$, heterogeneous in terms of latent feature, into L homogeneous subsamples (classes) that differ in the space of classification features by the degree of intensity of the epidemiological process. The solution to this problem is the classification matrix

$$D=\{d_{i,t}\}(i=1,\dots,n,t=1,\dots,T).$$

B. Initial data

We use the daily data for 30 countries of the European region (Armenia, Austria, Azerbaijan, Belarus, Bulgaria, Croatia, Czech, Denmark, Estonia, Finland, France, Georgia, Germany, Great Britain, Greece, Hungary, Ireland, Italy, Kazakhstan, Latvia, Lithuania, Moldova, Netherlands, Poland, Romania, Russia, Slovenia, Spain, Switzerland, Turkey) from March 1, 2020 to April 18, 2022 [6]. The list of indicators used includes: total number of infections (Total – T(*t*)), number of active cases of infection (Active – I(*t*)), number of recovered (Recovered – R(*t*)), number of deaths (Deceased – D(*t*)).

C. Classification features.

The following classification features are used:

- the ratio of the number of closed cases to the total number of infected (Closed to Total);
- the ratio of the number of closed cases to the number of active cases (Closed to Active);
- daily growth rate of the total number of infections or the ratio of the current value of cases to the previous one (Total Infections Daily Rate);

- mortality rate – the proportion of deaths from the total number of officially registered cases of COVID-19 (Death Rate).

D. *Used approach and algorithms.*

Since there is no training sample and, accordingly, the number of classes L is not known, it is necessary to use panel data classification algorithms in the self-learning mode. To solve this problem, it is proposed the approach to the analysis of panel data with a cluster structure by means of machine learning algorithms [7].

This approach includes the following steps:

- preliminary statistical analysis of sample and outlier detection;
- censoring and scale transformation of features to interval (0,1) in such a way that values close to zero correspond to a more favorable course of the epidemic process and vice versa;
- cluster analysis of initial non-classified sample and COVID indicators calculation;
- discriminant analysis and forecasting base on SVM and neural network algorithms.
- analysis of pandemic statistics at country and multi-country levels.

E. *Constructed COVID Indicators and their economic analysis.*

Based on the estimated classification matrix $D=\{d_{i,t}\}(i=1,...,n,t=1,...,T)$ the following indicators of the COVID-19 pandemic are used:

$d_{it}\in\{1,...,L\}$ – *daily country rating (DCR)*, which characterizes the degree of intensity of the epidemic for country i at time t: rating values 1 and L correspond to the lowest and highest degree of intensity of the epidemic process;

ACR_i – *Average Country Rating for the entire time interval*:

$$ACR_i=\frac{1}{T}\sum_{t=1}^{T}d_{it}\in(1,L),\ i=1,...,n;$$

IMI_t – *Integral Multicounty Indicator of COVID-19* at time $t=1,...,T$:

$$IMI_t=\frac{1}{n}\sum_{i=1}^{n}d_{it}\in(1,L),\ t=1,...,T.$$

Fig. 1 show the results of ranking all countries according to the ACR rating for the entire observation period.

Table 1 shows the values for GDP Annual Growth Rate (at the end of 2020) [7] for countries classified in groups with “low level” (1<ACR≤1,5). “average level” (1,5<ACR≤2.5) and “high level” (2,5<ACR≤3) of the epidemic intensity. It is obvious that the intensity of the epidemic process in every country is largely due to the ongoing anti-COVID policy. Based on the classification results (Figure 1) and Table 1, certain conclusions can be drawn about the effectiveness of anti-COVID measures applied in different countries [8].

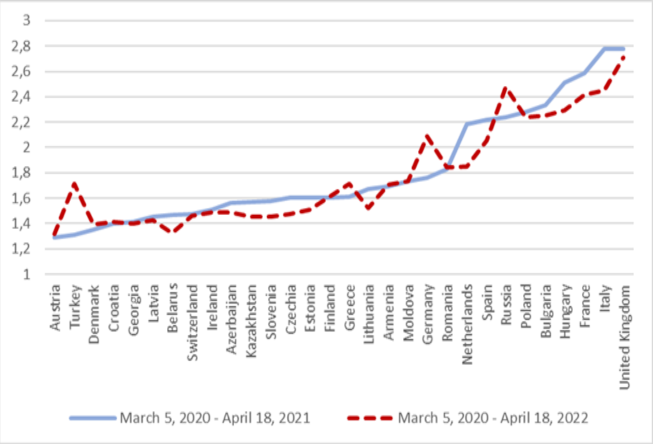

Fig.1. The results of ranking all countries according to the ACR rating for the entire observation period

TABLE I. THE EPIDEMIC INTENSITY AND GDP GROWTH RATES

Epidemic intensity based on ACR	**GDP Annual Growth Rate**
Low	-3,610
Average	-4,063
High	-7,716

TABLE II. ASSESSING THE QUALITY OF THE ALGORITHMS

Algorithm	**Accuracy**	**Balanced Accuracy**	**F1 Score**	**Time Taken**
SVM	0.71	0.70	0.71	19.40
BaggingClassifier	0.71	0.69	0.70	0.89
ExtraTreesClassifier	0.71	0.69	0.70	2.22
KNeighboursClassifier	0.68	0.66	0.68	0.26
DecisionTreeClassifier	0.67	0.65	0.67	0.16
LGBMClassifier	0.67	0.64	0.67	0.87

III. ANALYSIS OF PANDEMIC TYPOLOGY USING MACHINE LEARNING ALGORITHMS

To assess the class of epidemic intensity (daily country ratings) $d_{it}\in\{1,...,L\}$ $(I=1,\ldots,n,\ t=T+1,\ T+2,\ \ldots)$ on the base of new real-time updating data we use more than 10 machine learning and neural networks algorithms, realized in the Python package Scikit-Learn [9]. The main characteristics of the accuracy and performance for the best 6 algorithms are presented in Table 2.

As can be seen from table 2, the Support Vector Machine algorithm (SVM) occupies a leading position among other classification algorithms. With the SVM algorithm, it is possible to compare algorithms based on decision trees such as ExtraTreeClassifier and simple DecisionTreeClassifier using the idea of weighted voting, which has approximately the same metrics as the support vector method.

Besides classical machine learning algorithms, neural networks were applied. The best results were achieved using the with Dropout and BatchNorm1d layers. The classification accuracy of this algorithm for all classes reached 0.65. Other metrics of this algorithm for individual classes are presented in Table 3.

TABLE III. CLASSIFICATION RESULTS USING A NEURAL NETWORK ALGORITHM

Class	Precision	Recall	F1 Score
1	0.67	0.73	0.70
2	0.64	0.61	0.62
3	0.63	0.56	0.59

As we can see from the Table 3 the smaller recognition capabilities takes plays for the third class and better performs are achieved for the first and second classes. This is explained by the fact that the classification accuracy of the neural network algorithm significantly depends on the size of the training sample in general and for individual classes in particular. The number of countries with a rating of 3, which belong to class 3, is significantly less than the number of countries in classes 1 and 2.

IV. CONCLUSION

The results of the study of various statistical classification algorithms in the training mode allow us to conclude that the SVM algorithm has the best capabilities, which makes it easy to adapt it to new data. It also uses a fairly small amount of memory, making it efficient when working with large data sets. Based on the classification results obtained on the base of neural networks we can say that the model works almost 2 times better than “random guessing”, that is, random class selection, assuming that the probability of choosing each of the classes is 0.33.

Along with the tasks of multi-country pandemic analysis based on panel data classification algorithms for classified and unclassified samples, the problems of modeling and forecasting the epidemic process in the Republic of Belarus in 2020-2022 are considered. In [3, 10-12] were presented econometric models that are recommended to be used for modeling and forecasting the number of active infections within individual wave (vector error correction model - VECM COVID-19 RB) [3, 10] and for the entire period of the epidemic process (Markov switching linear regression model with autocorrelation errors – MS-LR-AR COVID-19 RB) [10–12].

A necessary condition for building these models is the stability and controllability of the epidemic process.

REFERENCES

[1] Cao Longbing and Liu Qing, “COVID-19 Modeling: A Review,” SSRN papers, 2021, https://ssrn.com/abstract=3899127.

[2] W. Kermack and A. McKendrick, “Contributions to the mathematical theory of epidemics,” Bulletin of Mathematical Biology, 53 (1–2), 1991, pp. 33–55.

[3] Yu. Kharin, V. Valoshka V, O. Dernakova, V. Malugin and A. Kharin, “Statistical forecasting of the dynamics of epidemiological indicators for COVID-19 incidence in the Republic of Belarus,” Journal of the Belarusian State University. Mathematics and Informatics, №3, 2020, pp. 36–50 (in Russian).

[4] S. Jang, L. Hussain-Alkhateeb and R. Rivera Ramirez, “Factors shaping the COVID-19 epidemic curve: a multi-country analysis,” BMC Infect Dis 21, 2020, https://doi.org/10.1186/s12879-021-06714-3

[5] V. Malugin, “Typology analysis, forecasting and assessment of the impact of the COVID-19 pandemic on economic growth rates,” V.I. Malugin, Economics. Modeling. Forecasting., 2023, Vol. 17, pp. 220-231 (in Russian).

[6] Worldometers.info [Electronic resource] – Mode of access: https://www.worldometers.info/coronavirus. – Date of access: 27.02.2022.

[7] V. Malugin, N. Hryn and A. Novopoltsev, “Statistical analysis and econometric modelling of the creditworthiness of non-financial companies,” Int. J. Computational Economics and Econometrics, 2014, Vol. 4(1/2), pp. 130-147.

[8] The World Bank Group [Electronic resource] – Mode of access: https://data.worldbank.org/ – Date of access: 02.03.2022.

[9] Scikit-Learn package in Python - https://scikit-learn.org/stable/supervised_learning.html

[10] V. Malugin, A. Kornievich and V. Potapovich, “Statistical analysis and econometric modeling of the COVID-19 pandemic,” Proc. of the International Congress on «Computer Science: Information Systems and Technologies» (CSIST’2022): Minsk, October, 6-10, 2022, p. 137–141.

[11] V. Malugin and A. Novopoltsev, “Statistical Estimation and Classification Algorithms for Regime-Switching VAR Model with Exogenous Variables,” Austrian Journal of Statistics, Vol. 46, 2017, pp. 47-56.

[12] V. Malugin, “Discriminant analysis of multivariate autocorrelated regression observations under conditions of parametric heterogeneity of models,” Informatics, 2008, №3(19), pp. 17–28 (in Russian).