Pattern Recognition and Context Prediction of COVID-19 cases in European Countries

Arzu Tosayeva[1,*,†], Ermiyas Birihanu[1] and Tsegaye Misikir Tashu[2]

[1]ELTE Eötvös Loránd University, Budapest, Hungary
[2]Department of Artificial Intelligence, Bernoulli Institute of Mathematics, Computer Science and Artificial Intelligence, University of Groningen, Groningen, The Netherlands

Abstract
The global impact of the COVID-19 pandemic has been significant, which requires data analysis to understand trends and patterns. However, this endeavor is challenging due to the complex transmission dynamics and diverse factors that influence the virus's spread. The data associated with COVID-19 is extensive and constantly evolving, and extracting meaningful insights from it is difficult. Therefore, the objective of this study is to analyze the impact of COVID-19 in various European countries, to identify common patterns, and to make predictions within the relevant context. To accomplish this, we used clustering techniques to reveal patterns in COVID-19 cases among European countries. The implementation involved cluster analysis to estimate labels based on cluster size and density while considering relevant background information. Subsequently, a classification model was applied to the labeled dataset. Using the K-Prototypes algorithm and leveraging the Silhouette score for identification, we determined the optimal number of clusters. These clusters were then combined based on density, and the degree of sparsity was assessed. As a result, two clusters emerged: one labeled as "low chance of infection" and the other as "high chance of infection." Using these results, we implemented a classification algorithm, achieving an accuracy rate of 90%. For this study, we gathered data from five different sources, consolidating them into a single dataset. Our findings demonstrate that combining COVID-19 datasets with diverse features enables trend analysis, while the use of clustering algorithms facilitates successful label identification in unsupervised learning scenarios involving unlabeled data. The density and size of clusters prove valuable in estimating labels, enhancing our overall understanding of the data. Our code is publicly available here.

Keywords
Context prediction, COVID-19, Label estimation, Pattern recognition

1. Introduction

The COVID-19 pandemic has had a profound impact on the global population, causing significant disruptions in healthcare systems, industries, and societies around the world. To control the spread of the virus and alleviate pressure on healthcare systems, numerous countries have implemented strict measures. In Europe, like in other regions, the COVID-19 outbreak emerged in January 2020 and quickly escalated, leading to a surge in cases and fatalities in hospitals [1]. While some European countries are currently experiencing new waves of infections, others are still dealing with a relatively low number of COVID-19 cases. Throughout the epidemic, France, Italy, Spain and the United Kingdom have documented a significant number of cases and fatalities[2]. To combat the virus, the European Union and many member states have implemented various measures such as decontamination, curfews, travel bans, and vaccination campaigns.

As governments implement various containment and social distancing measures, the demand for healthcare systems has increased significantly. This poses a challenging problem in effectively managing infected patients in hospitals. Having an effective modeling method that can identify patterns and predict the spread of the virus within the population would be highly valuable for preparing and formulating health and economic policies for governments, administrators, and decision-makers. This would aid in slowing down or halting the spread of the virus. With the increase in cases of COVID-19 and the availability of more data, several studies have utilized mathematical models [3], [4], [5] to analyze the spread of the virus. In addition, [6], [7] have also used LSTM models to forecast. However, these models often rely on outdated data from the same country, which limits their effectiveness. In a cluster analysis study [8], similarities were observed in the dynamics of the spread of the disease between countries such as Italy, France, and Germany, which implemented similar intervention strategies. In another study [9], supervised machine-learning approaches were used to predict the future of COVID-19. Furthermore, [10] demonstrated the potential of Machine Learning and cloud computing to improve the prediction

ITAT'23: Information Technologies – Applications and Theory, September 22–26, 2023, Tatranské Matliare, Slovakia
*Corresponding author.
†These authors contributed equally.
n0ndni@inf.elte.hu (A. Tosayeva); ermiyasbirihanu@inf.elte.hu (E. Birihanu); t.m.tashu@rug.nl (T. M. Tashu)
0009-0004-0113-9298 (A. Tosayeva); 0000-0001-7081-0365 (E. Birihanu); 0000-0002-4498-2486 (T. M. Tashu)

of epidemic growth.

Most existing works are confined to forecasting within specific countries or regions, overlooking the inclusion of newly reported cases, recovery rates, and mortality numbers that vary across different countries and over time. In this study, we utilized COVID-19 data from diverse sources. We employed clustering techniques for pattern recognition and applied the concept of estimating class labels derived from the work of [11] for estimating class labels. Subsequently, we applied classification methods to develop, train, and assess a supervised model aimed at predicting COVID-19 cases.

The remainder of this study is organized as follows. Section 2 discusses the existing research conducted in the relevant area, along with its limitations. Section 3 outlines in detail the proposed research methodology. Section 4 presents information on the datasets used in the research, along with the preprocessing and analysis steps taken. Section 5 provides a comprehensive overview of the conducted experiments. Finally, Section 6 presents the conclusions drawn from the research and outlines potential future work in the given research area.

2. Related Works

Clustering techniques have been utilized by researchers since the initial spread of COVID-19 cases. These techniques have been instrumental in grouping and identifying distinct patterns to discern differences or similarities among country-specific cases. Cheema et al. [12] proposed the implementation of the K-Means algorithm to establish diverse patterns among countries based on various features such as disease prevalence, health systems, and environmental indicators. The elbow method was employed to determine the optimal number of clusters, considering the sum of squared distances between samples and their closest cluster centers. Their study successfully demonstrated the reliability of Centroid-based Partition clustering in identifying patterns among country-specific cases. However, some limitations of this research include the need to consider both categorical and numerical features, as K-Means may not be the most suitable choice in such cases. Additionally, as the data only covers the year 2021, it may not accurately reflect the current situation for further analysis. Gohari et al. [13] also employed the K-Means algorithm to analyze longitudinal patterns of change in quantitative COVID-19 incidence and mortality rates. They utilized dimension-reduction techniques to identify correlations between features, which enhanced the model's performance. The research successfully identified three distinct patterns through experiments and compared different trajectories. Although this research presents an innovative approach in the field, it is limited to quantitative data and does not consider qualitative data, which can influence the spread of the disease.

In the study conducted by Lai et al. [14], the authors analyzed the incidence and mortality rates of 57 countries in 2020. They used Spearman's rank-order correlation to examine the relationship between cases and deaths. However, this research overlooked trends and patterns, which can be crucial for further analysis. Several other research works have focused on clustering COVID-19 cases for different countries, employing the same K-Means algorithm.

Labeling unlabeled data has been a significant area of research, with various ideas and approaches proposed. According to Fredriksson et al., [15], approximately 80% of engineering tasks in a machine learning (ML) project involve data preparation and labeling. Data preparation and labeling often require extensive effort due to incomplete datasets or the lack of labels for some or all instances. Moreover, even when labels are available, they may not be of good quality, leading to incorrect or partially correct labels for data points. High-quality labels are crucial for successful supervised machine learning, as the model's performance during operations is directly influenced by the quality of the training data.

Different techniques have been suggested by previous researchers to address the labeling challenge. Cui et al. [16] proposed an approach where samples are divided into clusters, and classification models are applied to each cluster based on the dataset's behaviors. The results from each cluster are then combined. To improve classification performance, the swarm algorithm was used for clustering, classification, and ensemble learning. The cluster-based ensemble learning method proved effective with cross-validation practices. However, this research only utilized labeled samples and did not consider unlabeled ones.

In another study by Kusumaningrum et al., [17], Chi-Square was used for labeling with the assistance of K-Means clustering. The homogeneity test was conducted using the Silhouette coefficient, followed by employing the Chi-Square Test for automatic cluster labeling in general.

In the study conducted by Yogesh [9], the aim was to develop an LSTM (long-short-term memory) model for forecasting COVID-19 deaths and cases, specifically for Italy and the United States. The model was subsequently evaluated using data from Germany, France, Brazil, India, and Nepal. On the other hand, Zeroual et al. [18] conducted a comparative study of five deep learning methods (RNN, LSTM, BILSTM, GRU, and VAR) to forecast the number of new cases and recovered cases. However, it should be noted that both researchers focused on a limited set of features during their investigation, potentially limiting the scope and comprehensiveness of their findings.The work by [19] introduced the concept of creating

a transmission dynamics predictor that exploits temporal variations among different countries in relation to the disease's spread. This is significant because certain countries encountered outbreaks before others. However, it's important to note that the data collected by the researchers spanned only a duration of three days.

3. Methodology

3.1. Pattern Recognition

Associating a classification with a label is known as recognition. Pattern recognition, as the science of drawing conclusions based on data, aims to categorize items or events into groups based on shared characteristics. In our work, we are utilizing clustering, which falls under unsupervised learning, to create patterns and uncover commonalities among the data. Clustering is an effective technique for identifying inherent structures or groupings within a dataset without the need for pre-existing labels or categories.

3.2. Label Estimation

Labeling data for COVID-19 manually is a time-consuming task that demands significant human resources. However, our proposed approach tackles this challenge by estimating relevant labels for data points, enabling supervised learning without the need for explicit user-provided labels.

When it comes to defining the exact formula for small clusters, it is important to note that there is no universally accepted formula. In our approach, we consider the number of data objects within a cluster to determine its size. Let us assume that we have k clusters and a cluster set C containing N data points.

Let the number of data points in each cluster be $\frac{N}{k}$ and α be the parameter used to determine whether a cluster is considered small or not. a cluster c is considered small if

$$|c_i| < \alpha \cdot \frac{N}{k} \tag{1}$$

where $|c_i|$ is the number of data points of the cluster. For example, α=0.2 indicates that if a cluster contains less than 20% then the cluster is considered small. As a next step, we are exploring whether the cluster is sparse or dense. For partitioning-based clustering, we used the sum of squares within the cluster ϵ.

In order to find the degree of sparsity, we have β, and we assume that the cluster is sparse when:

$$\epsilon_i < \beta \cdot \text{median}(E) \tag{2}$$

where E is the set of ϵ_i for all i. As a result, if β =2.0 means that ϵ of a cluster is greater than $2.0 \cdot median(E)$ and it is sparse.

In our research, we aimed to assess the prevalence of specific clusters and determine their commonality. To achieve this, we utilized the k-prototypes clustering technique, which is suitable for both categorical and numerical data. We manipulated the size and density of the clusters by varying the values of parameters α and β. By applying the k-prototypes clustering algorithm, we were able to cluster the data effectively. Following the clustering process, we assigned a class label to each cluster based on the features it contained. In COVID-19 cases, infections are categorized as high or low chance of infection. We assign labels to clustered dataset points using two rules: (1) Dense or sparse clusters are designated as "high chance," and (2) while others are categorized as "low chance." This involves classifying clusters into "low" or "high" likelihood of infection groups based on their characteristics. As a result, both the cluster and the data points contained within it are assigned identical labels, reflecting their infection likelihood.

In general, if a cluster exhibits high density or sparsity, it is labeled as having a high chance of infection. On the other hand, if a cluster is not extremely dense or sparse, it can be labeled as having a low chance of infection. This approach allows us to categorize clusters based on their characteristics and assign relevant infection likelihood labels accordingly.

3.3. Context Prediction

To separate the instances in the training data into appropriate classes, SVM is used. SVM utilizes a hyperplane defined as $w^T x_i + b = 0$, where w is the weight vector and b is the bias term. The marginal hyperplanes, $H1$ and $H2$, are given as [20]:

$$H1 : (w^T x_i + b) = 1$$

$$H2 : (w^T x_i + b) = -1$$

Thus, correctly classified points satisfy the inequality:

$$y_i(w^T x_i + b) \geq 1$$

In SVM, the margin refers to the distance between the marginal hyperplanes, also known as the decision boundary. Specifically, for a linear SVM, the margin is equal to $\frac{2}{|w|}$, where w represents the weight vector of the hyperplanes. Support vectors are the data points that lie on either the $H1$ or $H2$ hyperplanes, which define the margin. These data points play a crucial role in determining the position and orientation of the decision boundary.

4. Dataset

4.1. Dataset Description

The datasets used in this paper were collected from the European Centre for Disease Prevention and Control [1]and [2] Our World in Data [2]. In total, five datasets were used for our investigation.

The primary dataset used is titled "Data on the daily number of newly reported COVID-19 cases and deaths by EU/EEA country." This dataset covers the period from February 2020 to October 2022 from 30 European countries. The second dataset contains vaccination information, providing details on the COVID-19 vaccination progress across different countries. The third dataset includes information on the Gross Domestic Product (GDP) of countries impacted by COVID-19. The fourth dataset focuses on travel restrictions implemented by each country. The fifth dataset concentrates on school restrictions in European countries. It considers different levels and time periods of school closures, reopening, and other related restrictions. Finally, we created a comprehensive compilation of all five datasets, resulting in a final dataset with a total of 2,246,240 entries. This comprehensive dataset serves as the basis for our analysis and research findings. It encompasses a diverse range of numerical and categorical units with varying scales.

4.2. Exploratory Data Analysis

Figure 1 presents the changes in COVID-19 cases over time. In the beginning of 2021, the overall number of cases across countries was relatively low, and this trend continued throughout the year. Towards the end of 2021, there was a notable increase in cases, reaching its peak in late 2021 and remaining consistently high throughout 2022. However, there was a significant decrease in cases towards the end of 2022, indicating a decline in COVID-19 cases. Figure 2 shows the overall death figures caused by the spread of COVID-19 from 2021 to 2022. Initially, during the early stages of this period, the number of deaths was notably high. This could be attributed to the limited knowledge and understanding of the pandemic, and the lack of effective solutions and methods to handle it. However, as vaccinations and other measures were implemented, the number of COVID-19 deaths began to decline. Nevertheless, in 2022, there were still considerable rates of deaths reported. Towards the end of 2022, there was a significant decrease in these rates, as several decisive steps were taken in response to the increased severity of the pandemic.

[1]www.ecdc.europa.eu
[2]www.ourworldindata.org

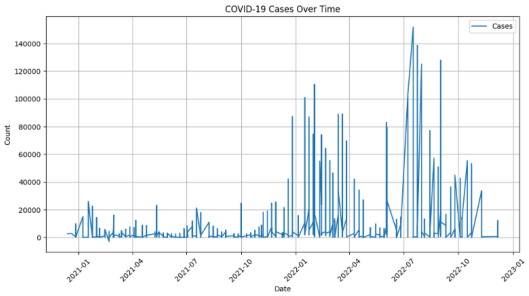

Figure 1: COVID-19 cases over time.

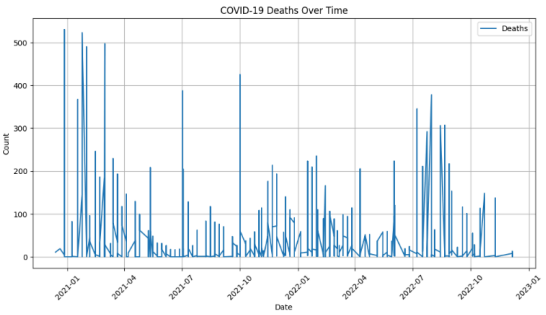

Figure 2: COVID-19 deaths over time.

5. Experimental Setup

5.1. Data Preprocessing

During the data preprocessing phase, we encountered missing values in some features of the dataset. To address these missing values, we applied imputation techniques, which involve estimating the missing values based on the available data. In our case, we found that median and mode imputation were suitable approaches. Therefore, we performed median or mode imputation wherever necessary, substituting the missing values with the median or mode of the corresponding feature. Additionally, we opted to drop insignificant missing values to ensure the integrity of the data.

Lastly, to ensure data integrity, we conducted a check for duplicate columns in the dataset. Any duplicate columns identified were removed from the dataset to avoid redundancy. Additionally, we applied normalization and standardization techniques to scale the data within specific ranges using Z-score standardization. These pre-processing steps help to enhance the quality and comparability of the data for further analysis.

Z-score can be defined as follows:

$$z = \frac{x - \mu}{\sigma} \tag{3}$$

Where z represents standardized version of original value, x is the dataset we want to normalize μ represents mean of dataset or column and σ demonstrates standard deviation of column or dataset.

5.2. Data Analysis

Figure 3 shows the direct impact of vaccination on COVID-19 cases during the entire period of analysis. The results were obtained by calculating the total number of vaccinations, which involved summing the administered doses. We further analyzed the percentage of the population that received at least one dose of the vaccination.

In Figure 4, we observed a correlation between the daily new COVID-19 cases and vaccination rates. When the daily new cases were high, the percentage of vaccinated individuals remained low. Conversely, as the vaccination rates increased, the number of new cases decreased. As time progressed, with a gradual decline in the number of COVID-19 cases, the demand for vaccinations also decreased. These findings demonstrate the significant role vaccination plays in curbing the spread of COVID-19 and reducing the number of cases over time.

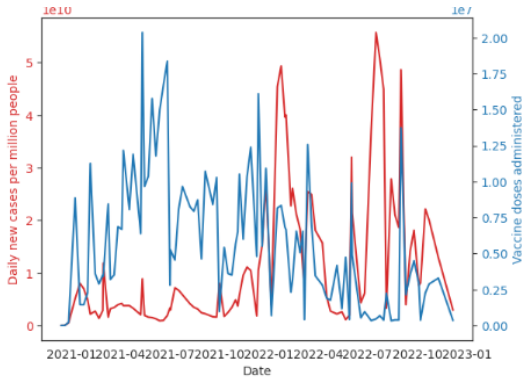

Figure 3: Vaccination Vs COVID-19 cases

In cluster analysis, various methods can be utilized, such as Chi-Square, ExtraTreeClassifier, Forward Feature Selection, and Correlation-based Feature Selection [21]. For our analysis of the clustering method, we employed a correlation graph, specifically utilizing the Pearson Correlation Coefficient. This coefficient is a measure of the strength of the relationship between different features. The Pearson correlation coefficient ranges between -1 and +1. A value close to +1 indicates a strong positive correlation between features. In such cases, if one feature increases, the other feature is also likely to increase, and vice versa. Conversely, a value close to -1 indicates a strong negative correlation. In this scenario, when one feature increases, the other feature tends to decrease, and vice versa. By utilizing the Pearson correlation coefficient and examining the correlation graph, we can gain insights into the relationships between different features and identify patterns within the data. This assists in understanding the underlying structure and dependencies among the variables, thereby aiding the cluster analysis process [21].

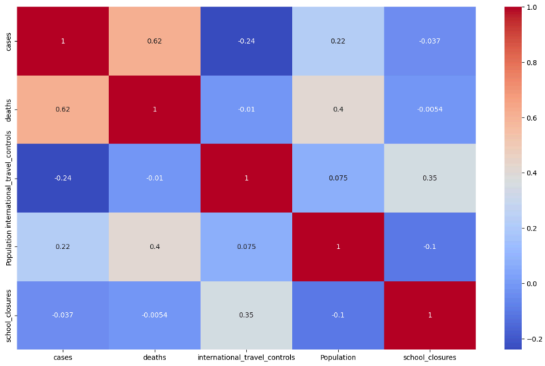

Figure 4: Heatmap for COVID-19 confirmed cases.

In Figure 4, we observe and interpret the correlation between various features. A positive correlation exists between Covid-19 cases and deaths (textbf0.62), as well as between Covid-19 cases and the Population feature (textbf0.22). On the other hand, we find a negative correlation between international school closures and Covid-19 cases (textbf-0.24). This indicates that when international school closures are implemented, there is a reduction in the number of Covid-19 cases. Similarly, a negative correlation exists between travel controls and Covid-19 cases (textbf-0.037), suggesting that as travel controls become more stringent, the number of Covid-19 cases decreases.

These correlations provide valuable insights into the relationships between different factors and Covid-19 cases, deaths, and control measures. Understanding these connections can aid in making informed decisions and formulating effective strategies to manage and mitigate the impact of the pandemic.

5.3. Experimentation

In our experiment, we employed two approaches: K-Prototypes [22] and SVM [23]. For both approaches, we utilized specific hyperparameters, which are listed in Table 1. These hyperparameters were chosen to optimize the performance and accuracy of the models.

Rather than using the default parameters for K-Prototypes and SVM, we chose to determine the best hyperparameters through grid search. The specific parameter values we chose are presented in Table 1.

Table 1
Hyperparameters

Algorithms	Hyperparameters	Values
K-Prototypes	n_clusters	12
	init	Cao
	n_jobs	4
SVM	kernel	linear
	C	0.001

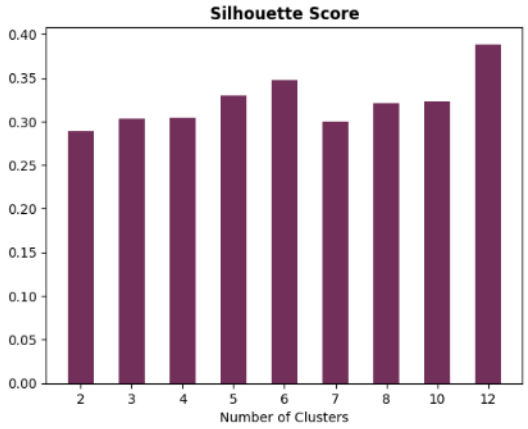

Figure 5: K-Prototypes clustering results.

5.4. Evaluation Metrics

In our clustering evaluation, we utilized the Silhouette Score as a metric to assess the quality and effectiveness of the clustering results. The Silhouette score can be computed as follows:

$$s = \frac{p - q}{\max(p, q)} \tag{4}$$

Where p is the mean distance to the points in the nearest cluster And, q is the mean intra-cluster distance to all the points.

To assess the SVM model's performance, we employed metrics such as accuracy, precision, and recall.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

6. Results

Based on the information provided in Table 2, we observed the distribution of countries across different clusters as follows. Cluster0, Cluster1, Cluster4, Cluster5, Cluster7, and Cluster9 each consist of one country. Cluster10 and Cluster11 contain three countries each. Cluster6 is assigned to two countries and Cluster2 is assigned to four countries. Cluster8 includes nine countries and has the highest mean of confirmed cases. During the Label Estimation phase, all twelve clusters were analyzed based on their sparsity and density. As a result, it was determined that Cluster2 and Cluster8 exhibited factors

Table 2
Clusters produced based on COVID-19 Cases

Cluster	**Cases in country**
Cluster0	Austria
Cluster1	Austria
Cluster2	Austria, Belgium, Bulgaria, Cyprus
Cluster3	Germany, Belgium
Cluster4	Austria
Cluster5	Austria
Cluster6	Spain, Germany
Cluster7	Austria
Cluster8	Finland, Estonia, Greece, Denmark,
Cluster9	Belgium
Cluster10	Spain, Germany, Belgium
Cluster11	Spain, Germany, Greece

associated with a low risk of infection. The specific characteristics of these clusters contributed to a lower risk of COVID-19 infection.

On the other hand, the remaining clusters were combined to represent the high-risk category. These clusters likely had higher case densities and lower overall population vaccinations, indicating a higher risk of COVID-19 spread in those regions. By categorizing the clusters into high and low-risk categories, we gain valuable insights into the different patterns of COVID-19 prevalence and can further explore factors contributing to these variations. As indicated in Table 2, certain countries are present in multiple clusters. This phenomenon can be attributed to the fact that certain regions within a country pose higher risks, while other areas exhibit lower risks. This pattern emerges due to variations in risk levels across different parts of the same country.

For context prediction, we employed the Support Vector Machine (SVM) algorithm. We split the dataset into a testing set comprising 25 percent of the data and a

training set comprising 75 percent. The table below displays the performance metrics of the model. The SVM model achieved an accuracy of 0.84 , indicating that it correctly predicted the context in 85 percent of the cases. The precision, which measures the proportion of correctly predicted positive instances, was 0.90. The recall, representing the proportion of actual positive instances correctly identified, was 0.71. These metrics provide insights into the performance of the SVM model for context prediction.

Table 3
Performance report for SVM

Performance Metrics	Values in percent
Accuracy	84%
Precision	90%
Recall	71%

7. Discussion

Based on the results obtained from our experiments, it is important to acknowledge and discuss the potential challenges and areas for improvement in our work. One significant challenge we encountered was obtaining positive results for the Silhouette Score, which is an important measure to evaluate the quality of cluster grouping. Initially, we faced negative or low scores due to the lack of appropriate scaling in our dataset. Scaling the data correctly is crucial to ensure reliable and meaningful results. Going forward, it is essential to pay attention to data scaling techniques and implement them effectively to enhance the accuracy of our analysis.

Another challenge we faced was the labeling of the dataset. Since the initial dataset did not include labels, we had to undertake the task of labeling the data ourselves. Accurate labeling is essential for cluster analysis, as it provides meaningful interpretation and understanding of the clusters. It required careful analysis of the features of each country to assign appropriate labels. In future research, it would be valuable to explore automated or semi-automated labeling techniques that can expedite the process and enhance the accuracy of labeling. Furthermore, while label estimation techniques were not the focus of our study, it is an aspect that could benefit from improvement. Considering the specific features of the dataset, exploring alternative label estimation approaches can provide more nuanced and detailed analysis objectives. Incorporating advanced label estimation techniques could improve the overall analysis and provide more comprehensive insights into the patterns and dynamics of COVID-19 spread.

Addressing these challenges and limitations in our research will contribute to further improvements in the accuracy and effectiveness of our clustering analysis. By refining data scaling, enhancing labeling techniques, and exploring advanced label estimation approaches, we can advance our understanding of COVID-19 patterns and provide valuable insights for future studies in this field.

8. Conclusion and future work

COVID-19 data has garnered significant attention in various research domains. However, to the best of our knowledge, the integration of clustering techniques for pattern estimation, label exploration, and context prediction has not been previously investigated. We have successfully achieved our objective of grouping countries based on various COVID-19 features, labeling the data using data characteristics, and implementing context prediction. However, there is still much more to be done in this research. One crucial aspect that deserves consideration is the integration of heterogeneity and coherence of the clusters to enhance accuracy. Combining clustering results from multiple algorithms or incorporating ensemble methods can lead to more robust and reliable clustering outcomes.

In our work, we have utilized the K-Prototypes algorithm for handling both numeric and categorical features. However, there are other algorithms, such as the Cluster Ensemble algorithm (CEBMC), that can handle both types of features and may provide additional insights. Exploring and implementing multiple algorithms can lead to a comprehensive understanding of the data and improve the analysis. Furthermore, to enhance the performance of the model, we aim to incorporate multiple clustering algorithms in a combined manner. Ensemble learning techniques, where several clustering algorithms work together, can produce more accurate and stable results. Additionally, we seek to enhance the robustness of label estimation to ensure more accurate results. Exploring and implementing various label estimation techniques, considering the characteristics of the dataset, can lead to more meaningful clustering analysis.

Overall, our research has made significant progress in the domain of COVID-19 data analysis using clustering techniques. However, there are still exciting avenues to explore and improvements to be made. By addressing the limitations and considering the integration of multiple algorithms and ensemble methods, we can advance our understanding of COVID-19 patterns and contribute valuable insights to the research community.

9. Online Resources

Our dataset and code are publicly available here. at https://github.com/Tsegaye-misikir/NCC.

References

[1] S. Pillai, N. Siddika, E. H. Apu, R. Kabir, Covid-19: Situation of european countries so far, Archives of medical research 51 (2020) 723.

[2] S. A. Rizvi, M. Umair, M. A. Cheema, Clustering of countries for covid-19 cases based on disease prevalence, health systems and environmental indicators, Chaos, Solitons & Fractals 151 (2021) 111240.

[3] H. B. Fredj, F. Chérif, Novel corona virus disease infection in tunisia: mathematical model and the impact of the quarantine strategy, Chaos, Solitons & Fractals 138 (2020) 109969.

[4] O. Torrealba-Rodriguez, R. Conde-Gutiérrez, A. Hernández-Javier, Modeling and prediction of covid-19 in mexico applying mathematical and computational models, Chaos, Solitons & Fractals 138 (2020) 109946.

[5] F. Ndaïrou, I. Area, J. J. Nieto, D. F. Torres, Mathematical modeling of covid-19 transmission dynamics with a case study of wuhan, Chaos, Solitons & Fractals 135 (2020) 109846.

[6] A. Tomar, N. Gupta, Prediction for the spread of covid-19 in india and effectiveness of preventive measures, Science of The Total Environment 728 (2020) 138762.

[7] V. K. R. Chimmula, L. Zhang, Time series forecasting of covid-19 transmission in canada using lstm networks, Chaos, Solitons & Fractals 135 (2020) 109864.

[8] S. Ghosal, R. Bhattacharyya, M. Majumder, Impact of complete lockdown on total infection and death rates: A hierarchical cluster analysis, Diabetes & Metabolic Syndrome: Clinical Research & Reviews 14 (2020) 707–711.

[9] F. Rustam, A. A. Reshi, A. Mehmood, S. Ullah, B.-W. On, W. Aslam, G. S. Choi, Covid-19 future forecasting using supervised machine learning models, IEEE access 8 (2020) 101489–101499.

[10] S. Tuli, S. Tuli, R. Tuli, S. S. Gill, Predicting the growth and trend of covid-19 pandemic using machine learning and cloud computing, Internet of Things 11 (2020) 100222.

[11] S. Baek, D. Kwon, S. C. Suh, H. Kim, I. Kim, J. Kim, Clustering-based label estimation for network anomaly detection, Digital Communications and Networks 7 (2021) 37–44.

[12] S. A. Rizvi, M. Umair, M. A. Cheema, Clustering of countries for covid-19 cases based on disease prevalence, health systems and environmental indicators, Chaos, Solitons & Fractals 151 (2021) 111240.

[13] K. Gohari, A. Kazemnejad, A. Sheidaei, S. Hajari, Clustering of countries according to the covid-19 incidence and mortality rates, BMC Public Health 22 (2022) 1–12.

[14] C.-C. Lai, C.-Y. Wang, Y.-H. Wang, S.-C. Hsueh, W.-C. Ko, P.-R. Hsueh, Global epidemiology of coronavirus disease 2019 (covid-19): disease incidence, daily cumulative index, mortality, and their association with country healthcare resources and economic status, International Journal of Antimicrobial Agents 55 (2020) 105946. URL: https://www.sciencedirect.com/science/article/pii/S0924857920300960. doi:https://doi.org/10.1016/j.ijantimicag.2020.105946.

[15] T. Fredriksson, D. Issa Mattos, J. Bosch, H. Olsson, Data Labeling: An Empirical Investigation into Industrial Challenges and Mitigation Strategies, 2020, pp. 202–216. doi:10.1007/978-3-030-64148-1_13.

[16] S. Cui, Y. Wang, Y. Yin, T. Cheng, D. Wang, M. Zhai, A cluster-based intelligence ensemble learning method for classification problems, Information Sciences 560 (2021) 386–409.

[17] R. Kusumaningrum, Farikhin, An automatic labeling of k-means clusters based on chi-square value, Journal of Physics: Conference Series 801 (2017) 012071. URL: https://dx.doi.org/10.1088/1742-6596/801/1/012071. doi:10.1088/1742-6596/801/1/012071.

[18] N. Sharma, N. Gaud, K-modes clustering algorithm for categorical data, International Journal of Computer Applications 127 (2015) 46.

[19] P. Hartono, Similarity maps and pairwise predictions for transmission dynamics of covid-19 with neural networks, Informatics in medicine unlocked 20 (2020) 100386.

[20] E. García-Gonzalo, Z. Fernández-Muñiz, P. J. Garcia Nieto, A. Sánchez, M. Menéndez, Hard-rock stability analysis for span design in entry-type excavations with learning classifiers, Materials 9 (2016) 531. doi:10.3390/ma9070531.

[21] E. Birihanu, J. Mahmud, P. Kiss, A. Kamuzora, W. Skaf, T. Horváth, T. Jursonovics, P. Pogrzeba, I. Lendák, Client error clustering approaches in content delivery networks (cdn), arXiv preprint arXiv:2210.05314 (2022).

[22] M. Shutaywi, N. N. Kachouie, Silhouette analysis for performance evaluation in machine learning with applications to clustering, Entropy 23 (2021). URL: https://www.mdpi.com/1099-4300/23/6/759. doi:10.3390/e23060759.

[23] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, A. Lopez, A comprehensive survey on support vector machine classification: Applications, challenges and trends, Neurocomputing 408 (2020) 189–215. URL: https://www.sciencedirect.com/science/article/pii/S0925231220307153. doi:https://doi.org/10.1016/j.neucom.2019.10.118.