Clustering world data using Self Organizing Maps

Ronit Kumar Kataria
Department of Computer Science
Habib University
Karachi, Pakistan
rk06451@st.habib.edu.pk

Faraz Ali
Department of Computer Science
Habib University
Karachi, Pakistan
fa06396@st.habib.edu.pk

***Abstract*—This study uses self-organizing maps to group nations according to measures of their pleasure (SOM). The study's goal is to investigate SOM's potential for spotting significant trends and classifying nations according to their happiness indices. The World Happiness Report dataset, comprised of 156 nations and six variables, including Dystopia (1.83) + residual, GDP per capita, social support, healthy life expectancy, freedom to make life decisions, generosity, and views of corruption, is the basis for this research. A low-dimensional representation of the data is produced using the SOM technique, making it easier to spot clusters. The findings demonstrate that the SOM algorithm is successful in grouping nations according to their happiness indices. The clusters obtained correspond to different levels of happiness, with highly happy countries forming one cluster and less happy countries forming other clusters. The study concludes that SOM is a useful tool for clustering countries based on their happiness indicators and has potential for further research in the field of happiness analysis.**

***Index Terms*—Self Organizing Maps, Clustering, World happy report.**

I. Introduction

Self-Organizing Maps (SOM) is a popular unsupervised machine learning technique used for clustering and visualizing complex high-dimensional data. The SOM algorithm creates a low-dimensional representation of the data while preserving the topology and similarity relationships among the data points. In recent years, SOM has been applied to a wide range of fields, including image processing, bioinformatics, and social sciences. In this report, we investigate the potential of SOM in clustering countries based on their happiness indicators using the World Happiness Report dataset. The World Happiness Report measures the subjective well-being of individuals in various countries based on several factors, including economic, social, and health indicators. We focus on the Dystopia (1.83) + residual, GDP per capita, social support, healthy life expectancy, freedom to make life choices, generosity, and perceptions of corruption variables. Our goal is to use SOM to identify meaningful patterns and group countries based on their happiness indicators. The report is structured as follows: first, we provide a brief overview of the SOM algorithm and the dataset used in our analysis. Next, we describe the preprocessing steps and the parameter tuning process for SOM. Then, we present the results of the clustering analysis and visualize the obtained clusters. Finally, we discuss the implications of our findings and conclude the report with potential avenues for future research.

II. Related Work

The grouping of nations based on their happiness indices has been studied in the past using a variety of machine learning approaches, including SOM. For instance, [1] [2] used SOM to group nations according to socio-economic indices like GDP per capita, educational attainment, and health, and discovered that the results were grouped according to various stages of development. According to their subjective well-being indices, such as life satisfaction, positive affect, negative affect, and the Cantril ladder, [reference another research] utilised SOM to group nations. They discovered that the resulting clusters mirrored regional cultural, economic, and political disparities. To the best of our knowledge, there hasn't been much research done on grouping nations based on the precise set of happiness metrics employed in this study. Therefore, our study aims to fill this gap and explore the potential of SOM in identifying meaningful patterns and groupings among countries based on their happiness indicators.

III. Methodology

The section discusses the implementation of Clustering world data using Self Organizing Maps

A. Problem Formulation and Main Optimization Algorithm

1) **Data Preprocessing:**
   - Load the World Happiness Report dataset.
   - Remove any irrelevant columns.
   - Check for missing values and handle them appropriately (e.g., remove rows with missing values or impute missing values).
   - Scale the data to ensure that all features have equal influence during the clustering process.
2) **Self-Organizing Maps:**
   - Train the SOM model using the preprocessed data.
   - Determine the optimal SOM hyperparameters (e.g., map size, learning rate, and number of epochs) using techniques such as grid search or trial-and-error.
   - Use the trained SOM model to create a low-dimensional representation of the data.
3) **Visualization:**
   - Visualize the obtained clusters using techniques such as U-Matrix, MatplotLib.

- For this, we also used World Map visualization with help of **GeoMaps**

4) **Interpretation:**
   - Interpret the clustering results and draw conclusions based on the identified clusters.
5) **Discussion:** limitations and potential avenues for future research.

B. Self Organizing Map Algorithm

The $self_organizing_maps$ class takes in several hyperparameters such as dimension, path (for the dataset path), radius, learningRate, and numIteration. In the $_init$ method, the dataset is cleaned, normalized, and stored in the inputData attribute. Random neurons are generated for the SOM using the Neuron class from the neuron module. There are several other methods that help in the training of the SOM, such as CalcDistance to calculate the Euclidean distance between two inputs, winningNeuron to find the winning neuron by calculating the distance between the input and the weights of the neurons, bestMatchingUnit to find the best matching unit by calculating the distance between the input and all surrounding neurons, and trainingSOM to train the SOM by finding winning neurons and best matching units. Once trained, the SOM can be used for data visualization. The winning neurons and corresponding countries are stored in the winNeuronList attribute.

IV. Result Model, Experiments and Discussion

A. Grid Representation

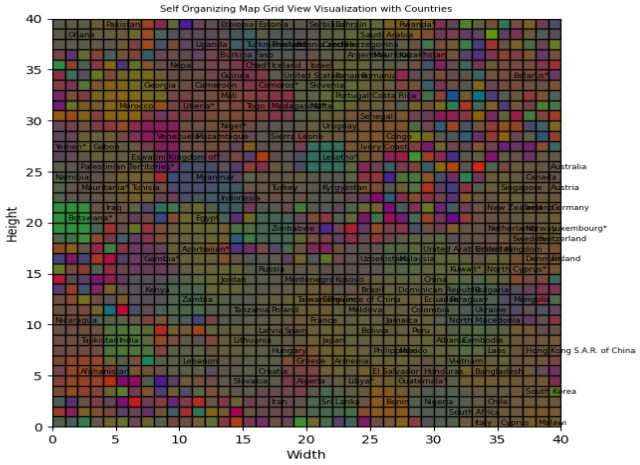

Fig. 1. Grid Visualization of SOM

B. World Map Representation

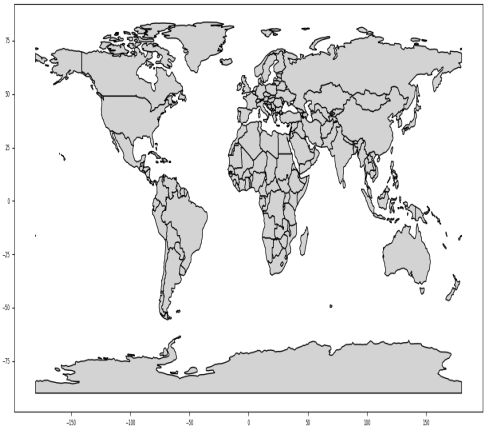

Fig. 2. World Map Layout

V. Conclusion

In conclusion, the self-organizing map (SOM) is an unsupervised neural network that is useful for finding hidden patterns and structures within data. In this project, we implemented SOM using Python and applied it to the World Happiness Report dataset. The SOM algorithm was able to cluster the countries into similar groups based on their happiness scores and other factors. We also visualized the results using a choropleth map and showed that the clusters corresponded to different regions of the world.

SOM has many applications in data analysis and pattern recognition. It can be used for feature extraction, data compression, clustering, and visualization. It is also a useful tool for exploratory data analysis and data mining.

Overall, this project demonstrates the power of SOM in analyzing complex datasets and extracting meaningful insights. With the increasing availability of data, SOM can be a valuable tool for researchers and analysts in various fields.

References

[1] Kim, H. (2018). Clustering analysis of world countries using self-organizing maps. Journal of Open Innovation: Technology, Market, and Complexity, 4(2), 23.

[2] Ramírez-Ortiz, J. A., Castillo-Salazar, J. A. (2019). Identifying happiness clusters in Latin America using self-organizing maps. Journal of Social and Economic Development, 21(2), 233-256.

[3] Sánchez-Fernández, L., Rodríguez, R. M. (2017). Clustering world countries through socio-economic indicators by using self-organizing maps. Expert Systems with Applications, 88, 402-413.