3rd.International Conference on
Theoretical and Applied Computer Science and Engineering
(ICTACSE, 2018)
Ankara, Turkey, November 23-24,2018

Clustering of Countries using Computational Models

Batuhan Albayrak, Ayşenur Gilik, Arif Selçuk Öğrenci
Faculty of Engineering and Natural Sciences
Kadir Has University
İstanbul, Turkey

***Abstract*—The member countries of the Organization for Economic Co-operation and Development are clustered based on 11 social and economic indicators for the years 2000 and 2015. Two methods are utilized: k-means and self-organizing map. The outcomes of the clustering effort for different number of clusters are compared between the methods and years so that transition of countries among clusters are obtained. The results indicate that both methods give similar clustering outputs hence it can be concluded that clustering based on those indicators is an alternative method of classification for the countries.**

Keywords—country clustering; k-means; Self-Organizing Map; OECD data.

I. INTRODUCTION

There are several international organizations that maintain a classification system of countries based on their development level. The most widely cited organizations are the UNDP (United Nations Development Program), the World Bank, the IMF (International Monetary Fund), the OECD (Organization for Economic Co-operation and Development), and the WTO (World Trade Organization) [1]. All of them base their evaluations on similar development taxonomies where a dichotomous classification (developed vs developing) or a trichotomous system (developed, middle, and developing) is obtained as the result. The classification is essentially based on a ranking. The thresholds for class boundaries are based on ad hoc assumptions and judgement. There are also attempts to carry out the classification of countries using a transparent methodology where the decision for threshold values is based on the data [1].

Other popular issues related to the classification of countries are the change of positions within time and the relative position of countries with respect to others. Hence, clustering of countries into a number of groups based on their characteristics, offers a potential to shed light on understanding the classification issue. The crucial factor in this process is that clustering will only be based on the data which are considered to be representative for countries. That is, the clustering process will be carried out deploying computational models that can take a subset of more than 500 indicators available for countries in various datasets. There are research articles in the literature that deal with the classification and clustering of countries for various purposes. For example, the work in [2] deals with the clustering and then classification of 54 countries according to their progress in becoming a knowledge economy. Three ordinal classifiers and a support vector regression system are compared for the clustering job where eleven indicators from the World Bank are used. For a smaller set of 8 European countries, fuzzy c-means, spectral clustering, and expectation maximization are compared to analyze daily electrical loads [3].

In another work, SOM (Self Organizing Map) has been utilized to cluster 25 transition economies based on 14 socio-economic indicators selected from the World Development Indicators (WDI) of the World Bank [4]. Different machine learning approaches such as multilevel regression trees and boosting are used to develop methods for the analysis of PISA test scores of 9 countries [5]. Recently, hierarchical cluster analysis is performed for 30 OECD countries using 14 variables obtained in public databases in order to group the countries by their innovation output in healthcare [6].

Considering the existing literature, this work focuses on clustering of all OECD countries (36 in total) based on 11 publicly available socio-economic indicators where two research questions are addressed:

1. Is there a difference in clustering output for two different computational methods, namely k-means clustering and SOM?
2. How do the clusters change in time from 2000 to 2015?

The outcomes of the clustering analysis will be a major contribution to comment on the classification of countries and the test the validity of other classification approaches. The next section will describe the data and the methodology employed. Then, results of clustering for the two years are given and the paper is concluded by a comparison of methods and by a detailed analysis of clustering outcomes.

II. DATA AND METHODOLOGY

A. Data

OECD has 36 member countries which are dispersed in a wide spectrum with respect to geographical location, social and economic status. The fundamental objective of OECD is "to promote policies that will improve the economic and social well-being of people around the world." Towards that mission, OECD maintains a publicly available database where a total of 274 indicators in historical windows are stored for OECD countries [7]. As a starting point, a reasonable subset of those indicators have been selected for analysis. Two criteria are employed in the decision for inclusion: The indicator should have valid data for most, if not for all, of the countries for both the calendar years 2000 and 2015. Moreover, the indicators should reflect the general social and economic status of the respective countries. The list of the indicators along with their descriptive statistics are given in Table I.