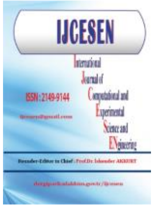
International Journal of Computational and Experimental Science and ENgineering
(IJCESEN)
Vol. 5-No.1 (2019) pp. 23-26
http://dergipark.gov.tr/ijcesen



ISSN: 2149-9144

Research Article

Clustering of European Countries in terms of Healthcare Indicators

Billur ECER[1], Ahmet AKTAS[2*]

[1] Ankara Yıldırım Beyazıt University Industrial Engineering Department, Ankara, Turkey
[2] Gazi University Industrial Engineering Department, Ankara, Turkey

* **Corresponding Author :** aaktas@gazi.edu.tr
ORCID: 0000-0002-4394-121X

Article Info:

DOI: 10.22399/ijcesen.416611
Received : 18 April 2018
Accepted : 20 January 2019

Keywords

Data mining
Clustering
Healthcare indicators
EU countries

Abstract:

Health is always considered as one of the most important issues related to human being. Due to this importance, governments should primarily provide the best healthcare services to their citizens. Some indicators can show the quality of healthcare services in the country. However, one country can have a higher value of one indicator and can have a lower value of another. Thus, countries can be categorized in terms of quality of healthcare services. Clustering is a useful tool for comparing countries and defining the similar countries in terms of healthcare services. In this study, 28 European Union (EU) countries were evaluated on 14 health factors and the number of clusters was determined by the generally accepted rule of thumb. To cluster countries, k-means clustering method is run in WEKA software for two cluster numbers and four different initial solution approaches. The resulting clusters were evaluated according to the Spearman rank correlation coefficient using the order of the GDP per capita values of the countries in each cluster. It seems using four clusters with Canopy initial solution approach is the most appropriate way of clustering.

1. Introduction

There are national and international health organizations working on prevention and treatment for diseases at national and international level. Apart from these, countries are also obliged to provide their citizens health services in the best possible way. Health care is important among the basic needs. It is possible to compare the level of development with the current situation in terms of health services.

In Table 1, a summary of literature for data mining studies about health subject is presented. For this reason, in this study, European Union member countries are clustered by the k-means method considering healthcare indicators. The rest of the paper organized as follows. The application of clustering with k-means method is given in the second section. The paper concluded with presenting clustering comparisons of countries and suggestions for future works in the third part.

2. Clustering of EU Countries via K – Means Algorithm

K – means clustering algorithm is developed by MacQueen in 1967 [1]. This method allows each unit to belong to only one cluster. With the K-means method, each element is assigned to the cluster that has the closest cluster center to itself. The average of values related to all elements in the cluster is the cluster center. K-means method assumes that the center point represents the cluster. The aim is to determine k clusters of elements with the least squared center distance for each cluster. In the K-means method, k-number of clusters is determined, and then clustering is performed considering distance criteria for clusters depending on the number of clusters. The disadvantage of the method is that it is difficult to determine the number of k sets. To overcome this difficulty, rule of thumb for determining cluster number k is

Table 1. *Literature summary for data mining studies in healthcare.*

Author	Year	Subject	Method
Ersöz [2]	2008	Evaluation of countries	Multi-dimensional scaling
Lorcu et al. [3]	2012	Evaluation of countries	Multi-dimensional scaling, hierarchical clustering
Alptekin [4]	2014	Evaluation of countries	Fuzzy clustering
Lewandowski et al. [5]	2014	Psychiatric diseases	K-means, Ward method
Moser et al. [6]	2014	Heart diseases	Hierarchical clustering, Ward method
Tsumotoa vd. [7]	2015	Nursing	Dual clustering
Olson et al. [8]	2016	Old patients	Hierarchical mass clustering, failure discovery rate

approximately calculated by the following formula [9] where k is the number of clusters, and n is the sample size:

$$k \approx \sqrt{\frac{n}{2}} \qquad (1)$$

In the study, clustering of the European Union member countries in terms of health factors is done with Weka software by the k-means method. A data set consisting of the mean values of the last 20 years of the values of 14 variables listed below related to health factors of 28 EU countries is used [10]. The data set used is derived from the World Bank World Development Indicators database. There is no lost value. The following variables are considered:

- Health expenditure per capita (US $)
- Hospital beds (per 1000 people)
- Immunization, Diphtheria Whooping cough Tetanus (% of children aged 12-23 months)
- Immunization, Measles (% of children aged 12-23 months)
- Life expectancy at birth (year)
- Risk of maternal mortality (%)
- Number of deaths under the age of five
- Number of surgical procedures (per 100,000 population)
- Prevalence of anemia among children (% of children under 5)
- Prevalence of anemia in non-pregnant women (% of women aged 15-49)
- Prevalence of anemia in pregnant women (%)
- Prevalence of anemia in women of reproductive age (% of women aged 15-49)

Table 2. *Obtained clusters for trials.*

	3 Clusters				4 Clusters			
Country	Random	K-Means++	Canopy	Farthest first	Random	K-Means++	Canopy	Farthest first
Germany	1	1	1	1	1	1	4	1
Austria	1 *	2 *	1 *	3 *	1 *	2 *	1 *	3 *
Belgium	1	2	1	1 *	1	2	1	1 *
Bulgaria	3 *	3 *	2 *	2 *	3 *	3 *	2 *	4 *
Czech Republic	3	3	2	2	3	3	2	4
Denmark	1	2	1	1	1	2	1	1
Estonia	3	3	2	2	4 *	3	2	4
Finland	1	1 *	1	1	1	1 *	1	1
France	1	1	1	1	1	1	4 *	1
Croatia	2	3	2	2	2	3	2	4
Netherlands	1	2	1	1	1	2	1	1
England	1	2	1	1	1	2	1	1
Ireland	1	2	1	3	1	2	1	3
Spain	1	2	1	1	1	2	1	1
Sweden	1	2	1	1	1	2	1	1
Italy	1	2	1	1	1	2	1	1
Cyprus	2 *	3	3	1	2 *	3	3 *	1
Latvia	3	3	2	2	4	3	2	2 *
Lithuania	3	3	2	2	4	3	2	4
Luxemburg	1	1	1	1	1	1	1	1
Hungary	3	3	2	2	3	4 *	2	4
Malta	1	2	1	3	1	2	1	3
Poland	3	3	2	2	3	4	2	4
Portugal	1	2	1	1	1	2	1	1
Romania	3	3	2	2	4	4	2	2
Slovakia	3	3	2	2	3	3	2	4
Slovenia	2	3	3	2	2	3	3	4
Greece	1	2	1	1	1	2	1	1

- The surviving woman until the age of 65 (% in the community)
- Male surviving until age 65 (% in community)

To create clusters, Simple K-Means method is used. The Weka program uses four different initial clustering approaches to start clustering in the K-means method. These approaches are defined as Random, K-Means ++, Canopy and Farthest First. In order to determine the number of clusters K, the rule of thumb is applied equal to 28. In this case, k is approximately equal to 3.741. Using the 4 different approaches mentioned above for the K-means methods, the results for 3 and 4 cluster formation situations are evaluated. Obtained clusters in that trials are given in Table 2. The optimal clustering should be determined by considering the effect of the different initial methods used on clustering results. The groups obtained by 4 different methods and 2 cluster numbers are evaluated by using the Spearman Rank Correlation Coefficient given in Equation (2).

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \qquad (2)$$

The evaluation of the clusters is done by using the rank of GDP per capita in the corresponding countries. Average GDP values and country ranks are presented in Table 3. Taking these sequence numbers into account, the Spearman rank correlation coefficient obtained for each clustering experiment is given in Table 4.

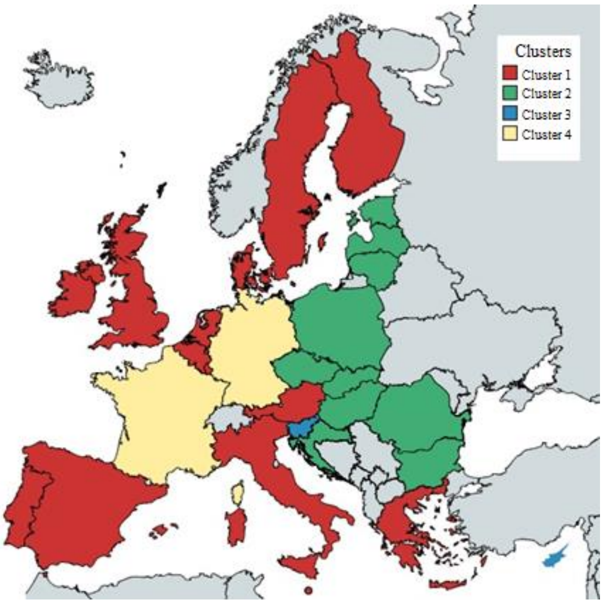


Figure 1. *Representation of clusters on the European map.*

Table 3. *GDP values and ranks of countries.*

Country	Average GDP	Rank
Germany	35470.5	10
Austria	37934.8	7
Belgium	36000.8	9
Bulgaria	3825.43	28
Czech Republic	12921.9	19
Denmark	46984.3	2
Estonia	9447.16	23
Finland	37107.2	8
France	34133.3	11
Croatia	9625.45	22
Netherlands	41110.9	5
England	38068.2	6
Ireland	46127	3
Spain	25092.3	13
Sweden	41478.5	4
Italy	30745	12
Cyprus	23183	14
Latvia	6964.52	26
Lithuania	7610.89	25
Luxemburg	76462.3	1
Hungary	10288.6	21
Malta	14969.4	18
Poland	8336.02	24
Portugal	18355.9	16
Romania	4707.28	27
Slovakia	11974.2	20
Slovenia	17371.6	17
Greece	20431.43	15

Since we have made an evaluation based on the sum of the intraclass distances, we can say that the value of the Spearman Correlation Coefficient obtained is reasonable for the larger value test. In the clustering of EU Member States in terms of health factors, 4 clusters obtained with the Canopy trial can be used. These clusters are presented in Figure 1.

3. Conclusion

In the study, European Union member states are clustered in terms of health services. The application is performed using the Weka software and the K-means method.

Table 4. *Spearman Correlation Coefficients for trials.*

Trial	Number of Clusters	Initial Clustering Approach	Spearman Correlation Coefficient
1	3	Random	0.7893
2	3	K-Means++	0.7126
3	3	Canopy	0.7969
4	3	Farthest first	0.7682
5	4	Random	0.7920
6	4	K-Means++	0.7184
7	4	Canopy	**0.8035**
8	4	Farthest first	0.7693

Spearman rank correlation coefficient is calculated to evaluate clusters obtained by using different initial clustering approaches and cluster numbers. 4 clusters with the Canopy approach seems the best.

For future studies, different clustering approaches can be tried in the same data set, results can be obtained with different variables. The clusters can be evaluated according to the Pearson correlation coefficient using numerical variables, not rank variants.

References

[1] J. MacQueen, Proc. Fifth Berkeley Symposium on Mathematical Statistics and Probability, 21 June – 18 July, 1965 and 27 December, 1965 - 7 January, 1966 Berkeley-USA.

[2] F. Ersöz Analysis of health levels and expenditures of Turkey and OECD countries , Journal of Statisticians: Statistics & Actuarial Sciences, 2 (2008) 95–104.

[3] F. Lorcu, B. A. Bolat, A. Atakisi, Examining Turkey and Member States of European Union in Terms of Health Perspectives of Millennium Development Goals, Quality & Quantity, 46 (2012) 959-978. DOI: 10.1007/s11135-011-9648-1

[4] N. Alptekin, Comparison of Turkey and European Union Countries' Health Indicators by Using Fuzzy Clustering Analysis, International Journal of Business and Social Research, 4 (2014) 68-74. DOI: 10.18533/ijbsr.v4i10.607

[5] K. E. Lewandowski, S. H. Sperry, B. M. Cohen, D. Öngür, Cognitive variability in psychotic disorders: a cross-diagnostic cluster analysis, Psychological Medicine, 44 (2014) 3239-3248. DOI: 10.1017/S0033291714000774

[6] D. K. Moser, K. S. Lee, J. R. Wu, G. Mudd-Martin, T. Jaarsma, T. Y. Huang, X.Z. Fan, A. Strömberg, T. A. Lennie, B. Riegel, Identification of symptom clusters among patients with heart failure: An international observational study, International Journal of Nursing Studies 51 (2014) 1366-1372. DOI: 10.1016/j.ijnurstu.2014.02.004

[7] C. H. Olson, S. Dey, V. Kumar, K. A. Monsend, B. L. Westra, Clustering of elderly patient subgroups to identify medication-related readmission risks, International Journal of Medical Informatics 85 (2016) 43-52. DOI: 10.1016/j.ijmedinf.2015.10.004

[8] S. Tsumoto, S. Hirano, H. Iwata, Mining Schedule of Nursing Care based on Dual-Clustering, Procedia Computer Science, 55 (2015) 1203-1212. DOI: 10.1016/j.procs.2015.07.125

[9] T. M. Kodinariya, P. R. Makwana, Review on determining number of Cluster in K-Means Clustering, International Journal of Advance Research in Computer Science and Management Studies, 6 (2013) 90-95.

[10] World Bank (26.03.2016). World Development Indicators. Link: http://data.worldbank.org/data-catalog/world-development-indicators Access data: 26.03.2016.