A Hybrid Clustering Technique to Propose the Countries for HELP International

Mahmood A.Mahmood*[1,2], A. A. Abd El-Aziz[2,3] , Karim Gasmi[1,4], and Olfa HRIZI[1,5]

[1]Department of Computer Science, Jouf University, Tubarjal, Kingdom of Saudi Arabia

[2]Department of Information Systems and Technology, Cairo University, Faculty of Graduate Studies for Statistical Research (FGSSR), Egypt

[3]Department of Informtion Systems, Jouf University, Sakaka, Kingdom of Saudi Arabia,

[4]ReDCAD Laboratory, National School of Engineering of Sfax, Sfax University, Tunisia

[5]Laboratory Modelling, Analysis and Control of Systems(MACS), Tunisia

mahmoodissr@cu.edu.eg1, .ahmed@cu.edu.eg 2, kgasmi@ju.edu.sa 3, Oharizi@ju.edu.sa 4

Abstract - HELP International is a charitable nongovernmental organization (NGO) that is committed to fighting poverty and providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities. HELP International has been able to raise around $ 10 million. Therefore, the head of the NGO needs to decide how to use this money strategically and effectively. Hence, the Head requires to make a decision for choosing the countries that are in the direst need of aid. This paper uses a hybrid clustering technique to suggest countries based upon socio-economic and health factors that determine the overall development of the country. The hybrid technique applies K-MEANS clustering and Farthest-First algorithm for clustering the countries. Both techniques are part of unsupervised learning tasks, which group data into multiple clusters. The hybrid technique proposes the countries that are most in need of help to HELP International. Moreover, it helps the head of the NGO in making the decision of choosing the countries that are in the direst need of aid and increase the number of countries without risk according to overall development factors by clustering techniques to HELP International socio-economic and health factors clustering.

Keywords**:** Unsupervised machine learning, K-means, Farthest First, NGO.

1. Introduction

HELP is an altruistic nongovernmental organization (NGO) established by a returning horticultural guide laborer who went through eight years in Africa's emergency and development programs under the U.N., CARE and WUSC. HELP's administrative center is situated at HELP's Center for Ecology Research and Training on the City Farm in Weyburn, Saskatchewan. HELP maintains an overseas office in Nairobi, Kenya [1].

HELP International is a universal philanthropic that is focused on battling destitution and furnishing the individuals of in reverse nations with essential courtesies and alleviation during the hour of debacles and regular disasters [1].

The HELP association is administered by a top managerial staff comprising of individuals from across Saskatchewan just as Washington, including people from the accompanying callings: natural sciences, instruction, universal turn of events, and corporate parts [1].

HELP's Key Crucial to prepare Canadians and universal working together accomplices in driving edge systems in the helping sciences. HELP expands on the characteristic qualities of agrarian culture to make pioneers in helping sciences both at home and globally [1].

HELP International with having had the option to raise around $ 10 million. Presently, the Chief of the NGO needs to conclude how to utilize this cash deliberately and viably. Along these lines, Chief needs to settle on a choice to pick the nations that are in the direct need of help [1]. Henceforth, this paper proposes a hybrid technique to cluster the nations utilizing some financial and wellbeing factors that decide the general advancement of the nation. The proposed hybrid technique performs K-MEANS clustering [2] and Farthest First algorithm [3] for clustering countries. The two techniques are part of unsupervised learning tasks, which group data into multiple clusters. The proposed hybrid technique suggests the nations which the head of the NGO in settling on the choice of picking the nations that are in the direst need of help and improve the precision of bunching methods to HELP Worldwide financial and wellbeing grouping.

The rest of the paper is organized as follows: Section 2 presents a brief introduction about clustering. Section 3 shows the literature review. The proposed hybrid framework is described in Section 4. Finally, in Section 5, the conclusion is summarized.

2. Literature Review

Rao Muzamal Liaqat and al [4] proposed a technique to discover the valuable information from crude data by utilize various unsupervised learning strategies. The proposed technique utilizes the correlation matrix followed by K-mean (fast) to discover the significant pattern as patient state that will assist the practitioner to treat the patient astutely. Adam, Andrzejuk, [5] proposed a conveyance of rural outflows among OECD nations with the assistance of clustering analysis. The author utilized two strategies in the analysis: K-means and HDBSCAN calculations. The two techniques are a piece of unsupervised learning undertakings, which bunch data into numerous clusters. At long last, an evaluation of the acquired arrangements was implemented. Rodrigo, M. Carrillo-Larco, and Manuel Castillo-Cara, [6] used k-means, which is an unsupervised machine learning technique, to create data-driven clusters of nations. The calculation was educated by disease prevalence measurements, metrics of air pollution, financial status, and wellbeing framework inclusion. Ahmed Anwar, and Ussama Yaqub. [7] used K-means clustering to distinguish and comprehend bot movements in twitter conversations. The pervasiveness of Twitter bots has increased noteworthy spotlight as of late because of their abuse in impacting an open conclusions for political additions. R. S. Kamath, and R. K. Kamat, [8] Made an examination utilizing the K- means clustering method. This examination dissected the number of papers, publications, patent applications, and trademarks enlisted concerning the level of Gross Domestic Product spending (GDP) with interest in Innovative work, Research and Development (R&D). Unsupervised learning technique utilized for planning three groups of nations dependent on this dataset. Nations having a place in cluster 0 should concentrate on expanding the number of journal publications. Cluster1 framed by nations must rethink their research and development assets to propel scientists in expanding research profitability.

Aiyshwariya Paulvannan, and al. [9] made an investigation planned to build up a system to even more likely survey natural manageability around the world. The structure introduced meant to evade the deficiencies, for example, weighting, change and positioning, by utilizing an unsupervised data clustering technique known as Self-Organized Maps to discover the inherent patterns present in ecological exhibition data. P.Viveka, F. Kurus, Malai Selvi, [10] proposed a modified k-means algorithm to improve accuracy of the results with low time complexity. Modified k-means algorithm based on two phases are deriving an initial centroid and assigning data points to the closest clusters. The limitations of the modified k-means algorithm are (1) the time complexity increases if the size of data increase, (2) deriving the initial centroid phase takes more time to select the centroids. Weipeng Wang, Shanshan TU, and Xinyi Huang [11] proposed an improved k-means algorithm to solve the problems of distorted center selection and slow iteration convergence in traditional clustering algorithms. This algorithm optimizes the selection process of initial cluster centers by using the proposed abnormal behaviors. Backward of this algorithm has poor results for larger scale data. Wisan Tangwongcharoen, [12] presented a method that solves the movement patterns of shoulder blade problem by k-means and Eular graph. The author used k-means to separating the big data into subgroups to show the density of data in each subgroup. Sungjin Im, and al. [13], presented an approach to solve the k-means clustering with outliers' problem. The approach used greedy algorithms as a preprocessing phase to remove noise data that may be combined with k-means clustering algorithm. The result of this approach is the fastest and more accurate than traditional k-means. The drawback is the complexity time of the preprocessing phase is O(k log n). Mirpouya Mirmozaffari, and al [14], presented a comparison between farthest first and expectation maximization algorithm to get an efficient data envelopment analysis decision-making unit and find the best clustering algorithms for their data set. The farthest first algorithm result is more efficient and good accuracy than other algorithms.

3. Methods

Data mining is the way to knowledge from the enormous measures of data. It can find important information and patterns. Clustering [15] is one of the valuable strategies of data mining for statistical data analysis used in different research fields. According to [16], clustering is an unsupervised machine learning technique, where there are no defined dependent and independent variables. The patterns in the data are used to identify or group similar observations. The objective of any clustering algorithm is to ensure that the distance between data points in a cluster is very low compared to the distance between two clusters. In other words, the members of a group are very similar, and members of different groups are as much as possible dissimilar. In other words, it is a task of organizing datasets of objects into groups of similar objects.

Clustering has wide applications. It is often used as an individual data mining tool to observe the characteristics of each cluster and to focus on a set of clusters for further analysis. Clustering not only can act as an individual tool, but also can serve as a preprocessing step for other algorithms which would then operate on the detected clusters. Fig.1 demonstrates three clusters organized around three different nodes [17].

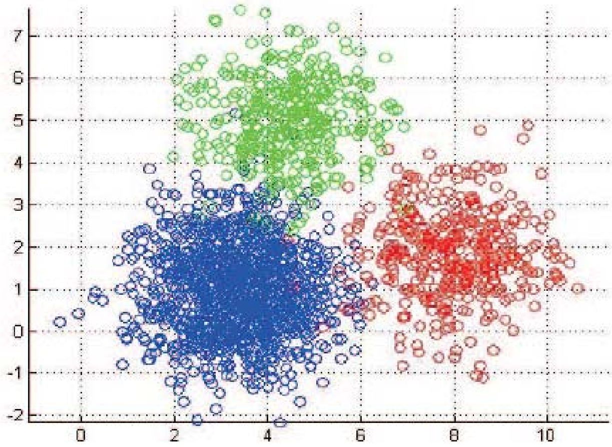

Fig 1: Cluster Analysis [17]

One of such clustering techniques is the popular K-means algorithm. K-mean is a simple algorithm that uses mathematical and statistical methods to solve the clustering problem, yet a widely used approach for clustering. It takes the number of k clusters and statistical vectors as initial random election centroids as an input to deduce the classification model. K-means distribute the m objects into k clusters where each object belongs to the closest cluster as shown in Figure 2. Moreover, the closest cluster computes by mathematical methods called distance measure, such as Euclidean distance; Manhattan distance; or Chebyshev distance [18]. An enhanced k-means algorithm is known as k-means++ [19], in which has a specific mathematical way for choose the initial centroids for k-means algorithm. Firstly, the election the first center randomly and let D(x) denote the shortest distance from a data point to the closest center that already chosen. Moreover, the choose remains k-center by the equation of probability. After that, it proceeds the standard k-means algorithm [19].

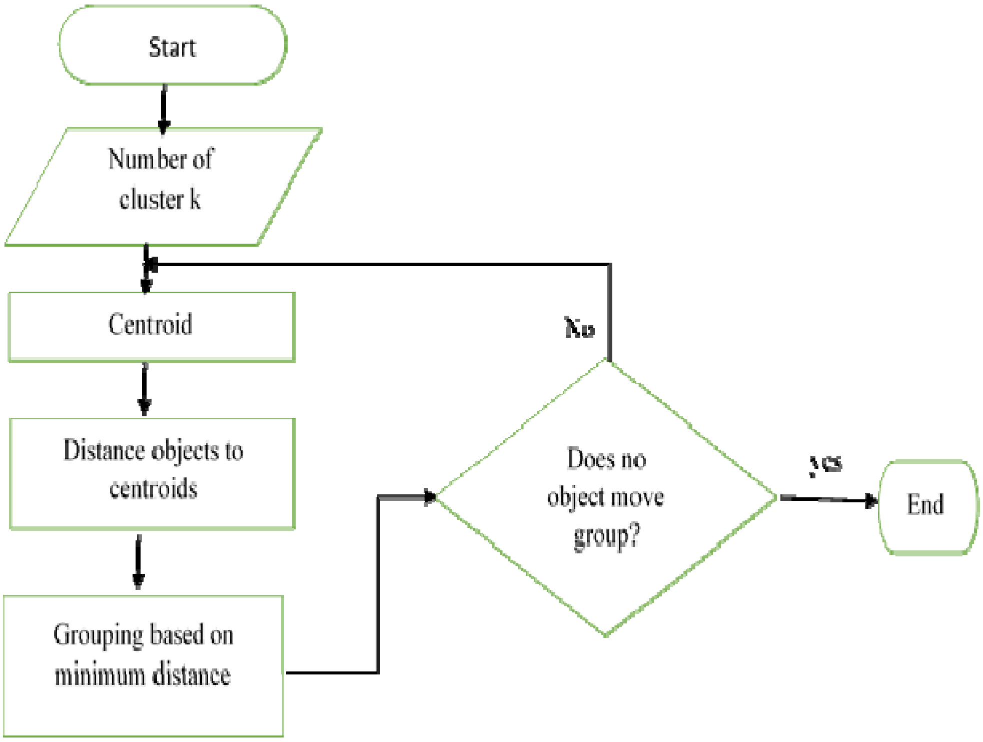
Start
Number of cluster k
Centroid
Distance objects to centroids
Grouping based on minimum distance
Does no object move group?
No
yes
End

Fig 2: K-Means Flow Chart

Hochbaum and Shmoys 1985 proposed another technique called Farthest-First algorithm shown in Figure 3 [3]. Like K-Means, in use the distance mathematical equation to measure the shortest distance, but unlike k-means, it needs only a single pass to cluster a set of data points. Moreover, k-means depend on the geometric centers of clusters, but farthest first all centroids are true data points. Farthest First starts by election a random data position as the original cluster center, during the cluster assignment step, it decides the next center as the data point furthest from the first center. Other centers are selected by the farthest from the set of previously chosen centers. Once a preferred k number of centroids has been selected, the algorithm allocates all other data points to the cluster characterized by the closest centroid and terminates. A hybrid clustering technique starts with the farthest first algorithm to represent the initial centroid and has true data points for k-mean++ clustering algorithm. Then, the hybrid technique proceeds with the k-mean++ clustering algorithm.

```
Input: (D: data set, k: integer)
Begin:
    randomly select first center;
    //select centers
    for (I= 2,…,k) {
        for (each remaining point) { calculate distance
        to the current center set; }
        select the point with maximum distance as new
        center; }
    //assign remaining points
    for (each remaining point) {
        calculate the distance to each cluster center;
        put it to the cluster with minimum distance; }
End
End
```

Fig 3: Farthest First Algorithm

4. Experimental Results and Discussions

The proposed hybrid technique categorizes the countries using clustering techniques based upon socio-economic and health factors that determine the overall development of the country. Hence, it helps the Head of the NGO in making the decision of choosing the countries that are in the direst need of aid and improve the accuracy of clustering techniques to HELP International socio-economic and health clustering. In our experiments, the results of the proposed hybrid technique are compared to the results of the traditional clustering techniques k-means++ and Farthest first separately.

4.1. *Data Acquisition*

HELP International's dataset in Kaggle.com for 167 countries [20]. The description of the dataset is shown in Table 1. We divided the dataset into two subsets, the first subset for health data (child_mort, Health, life_expec, and total_fer) is shown in Table 2, and the other data for socio-economic subset is shown in Table 3. Each subset dataset is clustered into four clusters titled without risk, low risk, medium risk, and high risk. These titles reflect our point of view. Moreover, the original dataset is clustered into clusters with the same titles. The purpose of dividing the datasets is tried to solve the following questions: "What are useful countries without risk for health data?" (i.e., a recognition health problem), "What are useful countries without risk for socio-economic data?" (i.e., a recognition socio-economic problem), and "What are countries without risk for socio-economic and health data?" (i.e., clustering counties for HELP data problem).

Table 1: Dataset Description

Column Name	Description
Country	Name of the country
child_mort	Death of children under 5 years of age per 1000 live births
Exports	Exports of goods and services per capita. Given as %age of the GDP per capita
Health	Total health spending per capita. Given as %age of GDP per capita
Imports	Imports of goods and services per capita. Given as %age of the GDP per capita
Income	Net income per person
Inflation	The measurement of the annual growth rate of the Total GDP
life_expec	The average number of years a newborn child would live if the current mortality patterns are to remain the same
total_fer	The number of children that would be born to each woman if the current age-fertility rates remain the same.
Gdpp	The GDP per capita. Calculated as the Total GDP divided by the total population.

The experiments are done using a laptop device which has Intel core i5 CPU with 2.4GHz with RAM 8GB. We made three experiments titled clustering by health, clustering by socio-economic, and clustering by socio-economic and health. Each experiment is performed by three techniques; k-means++, farthest first and the proposed hybrid approach and the results of the three techniques for each experiment are gathered and compared together.

4.2. *Clustering by Health*

This experiment clusters the health dataset by k-means++, the farthest first and the proposed hybrid technique. K-means++ results are shown in Figure 4, where the black color represents the countries without risk for building projects depending on health, the light gray color represents the country with low risk, the gray color represents the country with medium risk and dark gray color represents the country with high risk.

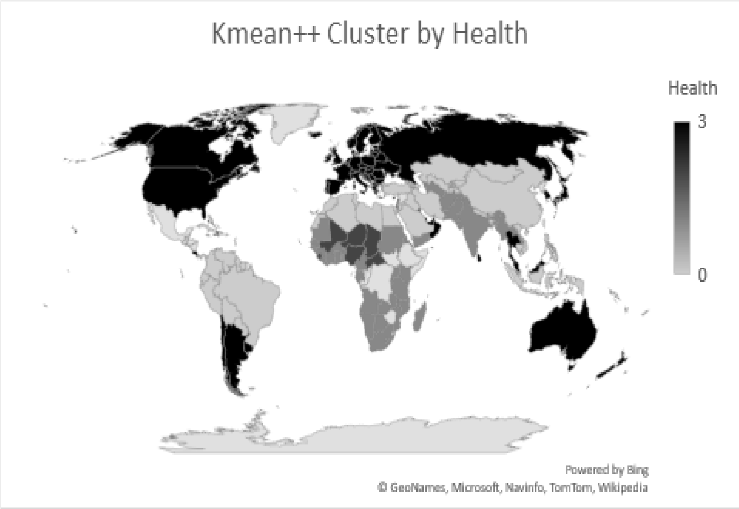
Kmean++ Cluster by Health
Health
3
0
Powered by Bing
© GeoNames, Microsoft, Navinfo, TomTom, Wikipedia

Fig 4: K-means++ Cluster for Health Dataset

Farthest first results are shown in Figure 5, where the black color represents the country without risks for building projects depending on health, the light gray color represents the country with low risk, the dark gray color represents the country with medium risk and gray color represents the country with high risk.

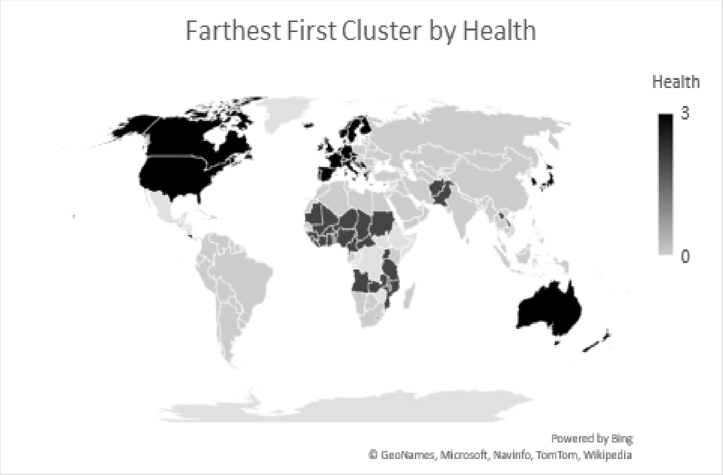
Farthest First Cluster by Health
Health
3
0
Powered by Bing
© GeoNames, Microsoft, Navinfo, TomTom, Wikipedia

Fig 5: Farthest First Cluster for Health Dataset

The proposed hybrid technique results are shown in Figure 6, where the black color represents the country without risks for building projects depending on health, the light gray color represents the country with low risk, the dark gray color represents the country with medium risk and gray color represents the country with high risk.

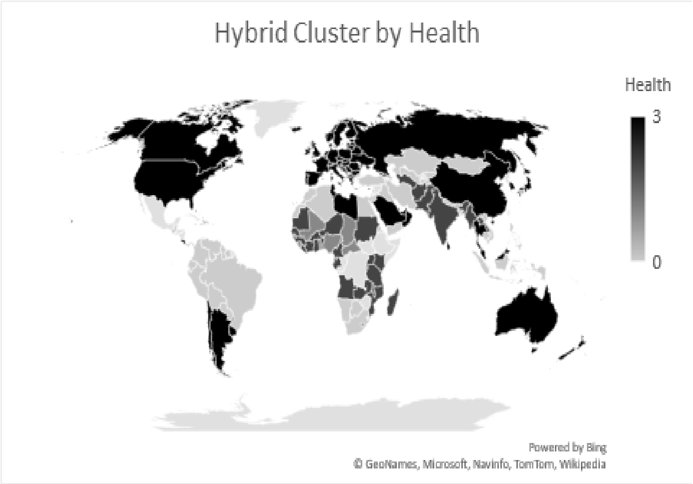

Fig 6: The Proposed Hybrid Cluster for Health Dataset

Table 2 presents the number of countries for each cluster in the health dataset, where the hybrid technique and k-means++ clusters are closest together in countries without risk. As shown in Table 2 the number of countries overall development of health factors increased in hybrid technique

Table 2: Number of Countries for each Cluster in Health Dataset

Techniques/clusters	Without risk	Low risk	Medium risk	High risk
k-means++	66	52	42	7
Farthest First	35	99	32	1
Hybrid	**73**	**50**	**38**	**6**

4.3. *Clustering by Socio-Economic*

This experiment clusters socio-economic dataset by k-means++, the farthest first and the proposed hybrid technique. K-means results are shown in Figure 7, where the black color represents the countries without risk for building projects depending on socio-economic, the gray color represents the country with low risk, the light gray color represents the country with medium risk and dark gray color represents the country with high risk.

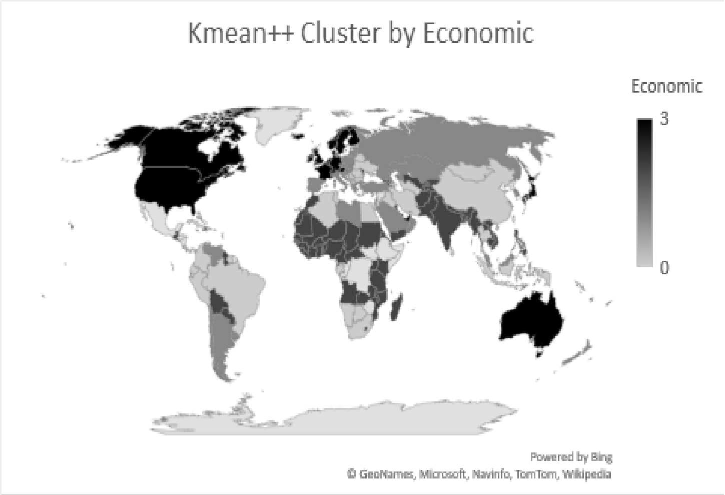
Kmean++ Cluster by Economic
Economic
3
0
Powered by Bing
© GeoNames, Microsoft, Navinfo, TomTom, Wikipedia

Fig 7: K-means++ Cluster for Socio-Economic Dataset

Farthest first result and the proposed hybrid technique results are shown in Figure 8 and Figure 9, respectively. Where the gray color represents the country without risks for building projects depending on socio-economic, the dark gray color represents the country with low risk, the black color represents the country with medium risk and light gray color represents the countries with high risks.

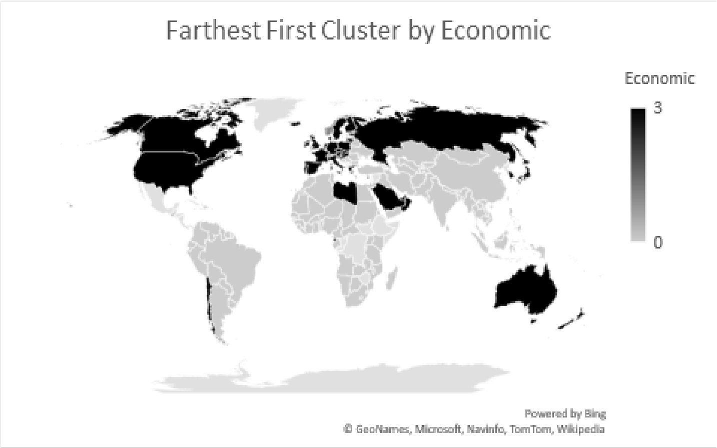
Farthest First Cluster by Economic
Economic
3
0
Powered by Bing
© GeoNames, Microsoft, Navinfo, TomTom, Wikipedia

Fig 8: Farthest First Cluster for Socio-Economic Dataset

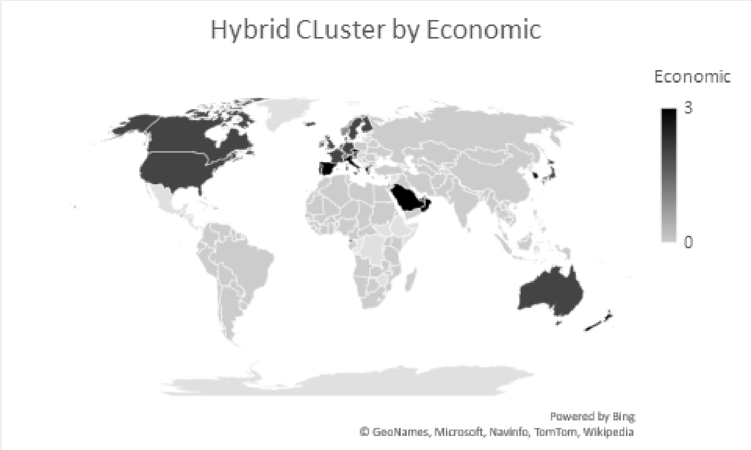
Hybrid CLuster by Economic
Economic
3
0
Powered by Bing
© GeoNames, Microsoft, Navinfo, TomTom, Wikipedia

Fig 9: The Proposed Hybrid Cluster for Socio-Economic Dataset

Table 3 shows the number of countries for each technique for socio-economic data, where the hybrid and farthest first give the same result for countries without risk. As shown in Table 3, the number of countries suggested by K-means++ cluster is more than the others for socio-economic factors.

Table 3: Number of Countries for each Cluster in Socio-Economic Dataset

Techniques/clusters	Without risk	Low risk	Medium risk	High risk
k-means++	23	35	42	67
Farthest First	3	1	44	119
Hybrid	**3**	**1**	**44**	**128**

4.4. *Clustering by Socio-Economic and Health*

This experiment clusters socio-economic and health datasets by k-means++, the farthest first and the proposed hybrid technique. K-means results are shown in Figure 10, where the light gray color represents the country without risks for building projects depending on socio-economic and health, the gray color represents the country with low risk, the black color represents the country with medium risk and dark gray color represents the country with high risk.

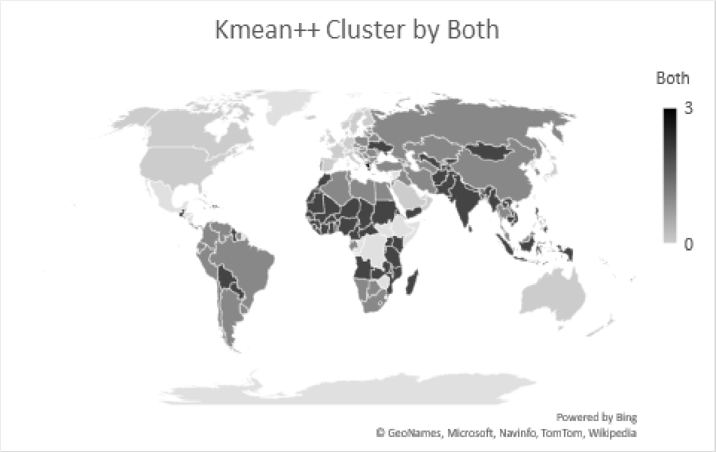

Fig 10: K-means++ Cluster for Socio-Economic and Health Datasets

Farthest first result and the proposed hybrid technique results are shown in Figure 11 and Figure 12, respectively.

Where the gray color represents the country without risks for building projects depending on socio-economic and health, the light gray color represents the country with low risk, the black color represents the country with medium risk and gray color represents the countries with high risks.

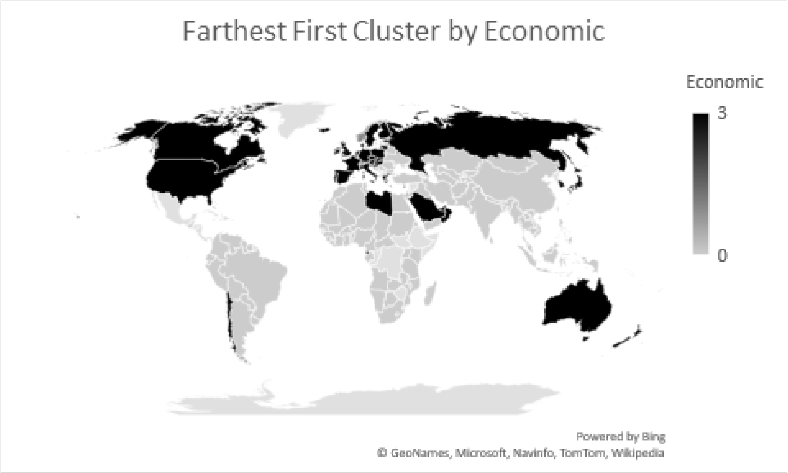
Farthest First Cluster by Economic
Economic
3
0
Powered by Bing
© GeoNames, Microsoft, Navinfo, TomTom, Wikipedia

Fig 11: Farthest First Cluster for Socio-Economic and Health Datasets

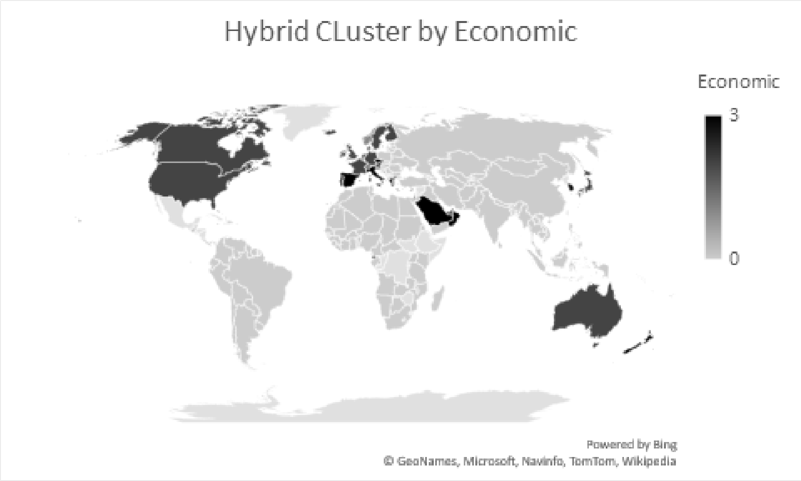
Hybrid CLuster by Economic
Economic
3
0
Powered by Bing
© GeoNames, Microsoft, Navinfo, TomTom, Wikipedia

Fig 11: The Proposed Hybrid Cluster for Socio-Economic and Health Datasets

Table 4 displays the number of countries for each technique in socio-economic and health dataset. As shown in Table 4, the number of countries for overall socio-economic and health factors increased by using the hybrid technique.

Table 4: Number of Countries for each Cluster in Socio-Economic and Health Datasets

Techniques/clusters	Without risk	Low risk	Medium risk	High risk
k-means++	31	61	2	73
Farthest First	7	92	28	40
Hybrid	**32**	**46**	**27**	**62**

5. Conclusion

In this paper, a hybrid technique based on K-mean algorithm and farthest first algorithm is proposed to cluster the countries with HELP International socio-economic and health datasets. For this purpose, a set of data is obtained from Kaggle.com website. K-means++ and farthest first are utilized to categorize the countries based on the collected data. The experiments show that the hybrid performs better than K-means++ and the farthest first for clustering countries separately. The proposed hybrid approach significantly improves the clustering by minimizing the risks related to socio-economic and health data for each country.

6. References

[1] Help international organiztion, https://www.help-international.com/, Accessed 2020,

[2] MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. 1. University of California Press. pp. 281–297

[3] Gonzalez, T. F. (1985), "Clustering to minimize the maximum intercluster distance", Theoretical Computer Science, 38 (2–3): 293–306

[4] Rao Muzamal Liaqat, Bilal Mehboobb, Nazar Abbas Saqibc, and Muazzam A Khand, A “Framework for Clustering Cardiac Patient’s Records Using Unsupervised Learning Techniques“, In Proceeding of the 6th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH 2016).

[5] Adam Andrzejuk, “Classification of Agricultural Emissions Among OECD Countries with Unsupervised Techniques“, Problems of World Agriculture, volume 18 (XXXIII), number 4, 2018: 80–91.

[6] Rodrigo M. Carrillo-Larco and Manuel Castillo-Cara, “Using country-level variables to classify countries according to the number of confirmed COVID-19 cases: An unsupervised machine learning approach“, Wellcome Open Research, 2020.

[7] Ahmed Anwar and Ussama Yaqub, “Bot detection in twitter landscape using unsupervised learning“, In Proceeding of the 21st Annual International Conference on Digital Government Research, June 2020.

[8] R. S. Kamath and R. K. Kamat, “K-MEANS CLUSTERING FOR ANALYZING PRODUCTIVITY IN LIGHT OF R & D SPILLOVER“, International Journal of Information Technology, Modeling and Computing (IJITMC) Vol. 4, No.2, May 2016.

[9] Aiyshwariya Paulvannan Kanmani, Renee Obringer, Benjamin Rachunok and Roshanak Nateghi, “Assessing Global Environmental Sustainability Via an Unsupervised Clustering Framework“, Sustainability 2020, vol. 12(2).

[10] P.Viveka, F. Kurus Malai Selvi, “A Modified K- Means Algorithm for Big Data Clustering”, ADALYA JOURNAL, Volume 9, Issue 2,pp.507-512, 2020.

[11] Weipeng Wang, Shanshan TU and Xinyi Huang,” IKM-NCS: A Novel Clustering Scheme Based on Improved K-Means Algorithm”, Engineering World Journal, volume 1, pp.103-108, 2019.

[12] Wisan Tangwongcharoen, “Comparison of Methods for Analyzing Shoulder Blades”, International Journal of Simulation Systems, Science & Technology ,(IJSSST), Volume 21, Number 3, pp. 3.1-3.8, 2020.

[13] Sungjin Im, Mahshid Montazer Qaem, Benjamin Moseley, Xiaorui Sun, Rudy Zhou,” Fast Noise Removal for k-Means Clustering”, Proceedings of the 23rdInternational Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy.

[14] Mirpouya Mirmozaffari, Azam Boskabadi, Gohar Azeem, Reza Massah, Elahe Boskabadi, Hamidreza Ahady Dolatsara, and Ata Liravian,” A comparison of Machine learning Farthest First and Expectation Maximization Clustering Algorithms Based on the DEA Cross-efficiency Optimization Approach for Banking System”, European Journal of Engineering Research and Science(EJERS), Vol 5, Issue 6, pp. 651-658,2020.

[15] Vivek, S. (2018). “Clustering algorithms for customer segmentation“, https://towardsdatascience.com/clusteringalgorithms- for-customer-segmentation-af637c6830ac. (Access date: 19.09.2018).

[16] K. Mumtaz1 and Dr. K. Duraiswamy “A Novel Density based improved k-means Clustering Algorithm – Dbk-means” International Journal of Computer Science and Engineering ISSN: 0975-3397 213, Vol. 02, No. 02, 2010.

[17] Sharmila and Mukesh Kumar, “An Optimized Farthest First Clustering Algorithm“, In Proceeding of 2013 Nirma University International Conference on Engineering (NUiCONE), 2013.

[18] Heba Ayeldeen, Mahmood A Mahmood, Aboul Ella Hassanien,“ Effective Classification and Categorization for Categorical Sets: Distance Similarity Measures“, Information Systems Design and Intelligent Applications, Advances in Intelligent Systems and Computing, pp. 359-368, springer,2015. DOI 10.1007/978-81-322-2250-7_36.

[19] David Arthur and Sergei Vassilvitskii. 2007. “K-means++: the advantages of careful seeding”. In Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms (SODA ’07). Society for Industrial and Applied Mathematics, USA, pp. 1027–1035, 2007.

[20] [kaggle.com dataset] https://www.kaggle.com/rohan0301/unsupervised-learning-on-country-data, Accessed, july-2020.