

An Enhanced K-means Firefly for Health Care Cluster Analysis of Philippines' COVID-19 Datasets

Jerico N. Contreras^{a,1}, Vince Andrei D. Isip^{b,1},
Raymund Dioses^{c,1}, Dan Michael Cortez^{d,1}

^a jncontreras2018@plm.edu.ph

^b vadisip2018@plm.edu.ph

^c rmdioses@plm.edu.ph

^d dmacortez@plm.edu.ph

¹ Computer Science Department, College of Engineering and Technology
Pamantasan ng Lungsod ng Maynila (University of the City of Manila)
Intramuros, Manila 1002, Philippines

Abstract

The development of digital health technologies during the COVID-19 outbreak sets the foundation for different research initiatives, including cluster analysis to better present and interpret the COVID-19 dataset. One of the cluster analysis algorithms often used is the K-means that various researchers applied in different applications. The application includes, but is not limited to, analyzing tourism attractions and restaurants, developing a machine learning model to track the virus's progression, and categorizing a country's cities or provinces based on its COVID-19 records. This paper intends to extend the cluster analysis of the prior research by enhancing a novel clustering algorithm known as K-means Firefly, which has improved the conventional K-means. The proponents achieved their objectives and were able to solve the limitation of the algorithm by improving the initialization of the algorithm's data and clusters, specifically by employing Principal Component Analysis for dimension reduction of data, the Calinski-Harabasz Index for automatically determining the number of clusters, and K-means++ for obtaining the initial clusters. The enhanced algorithm clustered the cities in the Philippines based on their COVID-19 datasets containing 32 healthcare-related features. As a result, the algorithm can handle real-world datasets with multiple features by reducing its dimensionality. It can automatically determine the optimal number of clusters and the initial location of the centroids. In terms of internal validity metrics, the enhanced algorithm also performed better than the previous implementation, with a percentage difference of 90.16% for the Silhouette Coefficient, 72.99% for the Davies-Bouldin Index, and 68.98% for the Calinski-Harabasz Index. The proposed algorithm may use for various applications, such as data dashboards and real-time tracking websites, due to its extended dynamic features in handling datasets and producing substantial clusters.

Keywords: K-means firefly; cluster analysis; health care; algorithm enhancement; Philippines, COVID-19

1. Introduction

The global proliferation of COVID-19 paved the path for digital health technologies, which have become the essential tools for addressing pandemic-related challenges. One of the innovations brought about by these technologies is the emergence of data dashboards and trackers with the help of machine learning, artificial intelligence (AI), and big data, which allow for the reporting and monitoring of the pandemic's impacts via statistical and graphical visualizations. According to Whitelaw et al. (2020), data dashboards and migration mapping enable monitoring of disease activity in real-time, adopted by countries like China, Singapore, Sweden, Taiwan, and the USA. Machine learning and AI predictive models were constructed to estimate the regional transmission patterns of the virus and guide the border inspections and surveillance. However, they needed to be trained using the COVID-19 datasets.

The development of these digital health technologies on a local and global scale lays the groundwork for various research initiatives involving cluster analysis to enhance the presentation and analysis of the COVID-19 dataset. The K-means clustering algorithm became the popular approach that has been used for cluster analysis. Several researchers clustered provinces, cities, or countries based on their COVID-19 data, like mortality rates, cases, incidences, facilities, environmental indicators, etc., using k-means clustering and presented significant clusters.

The previous research has utilized the conventional K-means cluster, which various recent studies have improved. This study attempts to improve K-means firefly, a novel clustering technique created for healthcare data, by expanding its capabilities for real-world datasets to be used dynamically. The improved algorithm will cluster cities in the Philippines based on their health-care features using the COVID-19 datasets.

2. Related Works

2.1. Research efforts during the COVID-19 pandemic

During the COVID-19 pandemic, cluster analysis emerged as a general approach in research. Countries were categorized by Rizvi, Umair, and Cheema (2021) into one of four groups according to their correlations with the 18 feature variables derived from illness prevalence, health system indicators, and environmental indicators. Both Abdullah et al. (2020) and Virgantari and Faridhan (2020) used the same methodology to study the use of K-means to cluster provinces in Indonesia using their COVID-19 examples were effective in classifying the underlying groups in the dataset. Several researchers produced a cluster analysis of the datasets by using each country's COVID-19 incidence and death rates as their primary variables (Gohari et al., 2022; Hutagalung, 2021). In addition, specific research has been applied in the tourist and business sectors for various goals. Ocampo et al. (2021) conducted research in which they grouped IVIF datasets using K-means and evaluated each tourist area or restaurant based on their perceived exposure to COVID-19.

2.2. The combination of PCA, K-means, and K-means++

Li (2018) experimented with classifying the efficacy of the three conventional initialization approaches, e.g., Random, K-means++, and PCA-based K-means, in enhancing the performance of K-means. The study concluded that the initialization strategies locate almost identical cluster centroids. However, the PCA-based K-means technique improves running time substantially quicker than the other options. Similar approaches are used by Xiangyuan, Siyuan, and Hao (2020) to enhance the K-means algorithm initialization by pre-processing the data with Principal Component Analysis (PCA) and obtaining a superior procedure for picking the first

centroid with the original k-means++ algorithm. In addition, Zubair et al. (2020) introduced an effective clustering approach that uses PCA and the Percentile method to select the ideal starting centroids of the K-means clustering algorithm. They utilized it to categorize countries depending on the quality of their healthcare systems during COVID-19.

3. Existing Algorithm

3.1. Background

In their paper entitled Cluster Analysis of Health Care Data Using Hybrid Nature-Inspired Algorithms, Ahmed P and Agrawal (2020) propose the K-means firefly method for cluster analysis. It combines the most beneficial aspects of the K-means and firefly algorithms. It omits their negative steps to create a unique and efficient clustering algorithm with a 54 percent improvement in performance over the classic K-means method.

Based on their approaches, they implemented the firefly algorithm by using the inter-cluster and intra-cluster element distances and the density-distance relationship to calculate the movement of the fireflies. In addition, they developed the glowing coefficient (GL) equation, which determines the point's position and whether it will move or remain in the same cluster throughout the iterations. The GL equation is defined in Equation 1.

$$GL = \frac{Density(c,j)}{D(i,C_i)} \quad (1)$$

where,

$Density(c,j)$ = the number of elements in the cluster j , $i \in$ all elements, and $j \in$ all clusters

$D(i, C_i)$ = the distance between particle and cluster center j

3.2. Pseudocode

```

Generate K random cluster centroids
repeat
  for all instance i in S
    shortest ← 0
    membership ← null
    for all Centroid c do
      distance ← Distance(c)
      if dist < shortest then
        shortest ← dist
        membership ← c
      end if
    end for
  end for
  Recalculate Centroids(c)

  for all instance i in S
    shortest ← 0
    membership ← null
    for all Centroid c do
      GL ← Total Members(c) / Distance(C)
      if GL < shortest then
        shortest ← dist
        membership ← c
      end if
    end for
  end for
  Recalculate Centroids(c)
until convergence
end procedure

```

Fig. 1. K-means Firefly Pseudocode

3.3. Limitations of the Algorithm

Since K-means firefly is a combination of the standard K-means and firefly clustering algorithms, it inherits the initialization issue of K-means. Reddy and Vinzamuri (2018) stated that the selection of initial centroids and the determination of the number of clusters, which comprise the first step of the algorithm, have a significant impact on the performance of K-means. Random selection of starting centroids is the approach in K-means firefly in which the initial centroids are dependent on the dataset and the number of data points to the cluster. When these factors, such as outliers and noises, are not favorable, they can never produce desirable output, even after repeated runs (Kumar et al., 2018). Furthermore, the number of clusters (K) in K-means firefly has to be determined and known beforehand, restricting its ability to cluster the real-world datasets automatically.

The proponents of K-means firefly aim to apply their algorithm to real-time applications and health care datasets, as it was only tested on typical machine learning datasets, such as Iris and Diabetes. However, they posed a challenge over the data characteristics that the algorithm will be processed. According to Berisha et al. (2021), digital health data has an emerging problem of continuously having more features in the long run, leading to a phenomenon known as the 'curse of dimensionality'. With the growing variety of clinical datasets, data redundancy may prevent learning the relevant relationships within the data due to extraneous, non-contributing features. Machine learning models become complex when there are a lot of dimensions in the data, and they tend to lose their intelligibility, accuracy, and clarity over their predictions (Patra et al., 2021).

4. Proposed Methodology

4.1. Pseudocode of Enhanced K-means Firefly

```

Data Pre-processing using Principal Component Analysis (PCA)

max_CHS ← 0
max_K ← 2

for K in range 2 to 15
  Generate K cluster centroids using k-means++
  repeat
    for all instance i in S
      shortest ← 0
      membership ← null
      for all Centroid c do
        distance ← Distance(c)
        if dist < shortest then
          shortest ← dist
          membership ← c
        end if
      end for
    end for
    Recalculate Centroids(c)

    for all instance i in S
      shortest ← 0
      membership ← null
      for all Centroid c do
        GL ← Total Members(c) / Distance(C)
        if GL < shortest then
          shortest ← dist
          membership ← c
        end if
      end for
    end for
    Recalculate Centroids(c)
  until convergence

  curr_CHS ← Calinski-Harabasz Score(cluster)

  if curr_CHS > max_CHS
    max_CHS ← curr_CHS
    max_K ← K
  end if
end for
end procedure

```

Fig. 2. Enhanced K-means Firefly Pseudocode

4.2. Enhancements to the Algorithm

4.2.1. Data Dimensionality Reduction using Principal Component Analysis

Principal Component Analysis will be used to solve the problem of data dimensionality. PCA can detect existing multi-collinearity between features/variables and is highly recommended for denoising and data reduction on data with large feature dimensions (Loukas, 2021). The algorithms perform better when the dimensionality, or amount of data features, is reduced because it can eliminate unimportant features and minimize noise. A further advantage is that dimensionality reduction might result in a more intelligible model. With a decrease in dimensionality, the data mining technique requires less time and computing capacity (Kumar et al., 2018).

4.2.2. Automatic Determination of K with Calinski-Harabasz Index

The proponents adopted the method of Singh and Bhatia (2011) to automatically calculate the required number of clusters (K) across the datasets by performing the algorithm for a set of K values. The algorithm will be executed through an iterative approach to acquire a 15-cluster solution as adoption to the evaluation of the Calinski-Harabasz algorithm in producing K clusters (Orlov, n.d.; Baruah, 2020). Based on the 15-cluster solution, the ideal K value for the algorithm will be determined by selecting the K value that created clusters with a higher score based on the Calinski-Harabasz validation index.

4.2.3. Initialization of Starting Centroids using K-means++

The proponents will use the K-means++ to generate the initial cluster centroids. It produces better clusters and allows the algorithm to converge more quickly, even though it is more expensive computationally (Sharma, 2019; Thakur, 2020). The technique employs a straightforward probability-based approach, with the first centroid chosen randomly. The next centroid selected is the one that is farthest away from the present centroid. This decision is based on a weighted probability score. The selection process is repeated until it gets K centroids, after which the algorithm is performed on these centroids.

5. Results and Discussion

5.1. Handling of the COVID-19 Datasets

The enhanced K-means firefly algorithm is used to cluster the cities and provinces in the Philippines, specifically the region of National Capital Region (NCR), CALABARZON (Region IV-A), and MIMAROPA (Region IV-B), based on their weekly COVID-19 data with 32 numerical healthcare-related features. The Principal Component Analysis (PCA) has reduced the data into 15 features, eliminating the extraneous features and producing substantial clusters.

```
[8.3457e+04 1.1700e+02 8.2006e+04 1.3340e+03 8.4000e+01 2.9400e+02
0.0000e+00 1.4200e-01 3.5000e+01 6.0000e+00 1.7140e-01 2.9000e+01
8.2860e-01 3.2200e+02 1.3600e+02 4.2240e-01 1.8600e+02 5.7760e-01
6.8000e+01 4.0000e+00 5.8800e-02 6.4000e+01 9.4120e-01 1.3000e+01
3.9800e-01 3.5700e+02 1.4200e+02 2.1500e+02 1.2000e+01 1.0000e+00
0.0000e+00 0.0000e+00]
```

Fig. 3. Sample row from the Philippines' COVID-19 dataset

```
[ 4.19907119 -1.27380955 -1.19180534 1.65167993 0.14466359 -0.72588077
-0.10825027 -0.4947769 -0.76073112 0.37876609 0.78204481 -0.41004766
-0.39745011 -0.79578182 0.51379981]
```

Fig. 4. Sample row after applying the Principal Component Analysis

The enhanced algorithm can be applied to real-world datasets without manually providing the number of clusters. Further explanation for this improvement will be found in Section 5.2.

5.2. Initialization of Centroids and Number of Clusters (K)

Using the K-means++ algorithm, the proponents slightly enhance the number of iterations needed before the convergence by 12.50%. The resulting clusters were desirable based on the comparison table shown in Table 1 due to the proper initialization of the centroids.

```
K-means firefly iterations count: 8
Enhanced K-means firefly iterations count: 7
```

Fig.5. Number of iterations before convergence for K-means Firefly and Enhanced K-means Firefly

The number of clusters was also determined automatically by the algorithm using an iterative approach with the help of the Calinski-Harabasz Index (CHI). It looks up to the 15-cluster solution that has been created in the iterations and selects the K value with a higher score for the CHI evaluation measure. This improvement allows the algorithm to be used internally in the applications without providing input for the number of clusters and to handle the real-world datasets dynamically.

k= 2		CHI =	27.02806848040209
k= 3		CHI =	112.05907647965424
k= 4		CHI =	82.9249753439035
k= 5		CHI =	82.54943946397833
k= 6		CHI =	202.62569971488244
k= 7		CHI =	166.25757557917822
k= 8		CHI =	181.08407605324757
k= 9		CHI =	114.02281927940098
k= 10		CHI =	130.90006356939114
k= 11		CHI =	116.75137177286601
k= 12		CHI =	110.39944371195207
k= 13		CHI =	102.49321340230235
k= 14		CHI =	92.99080496656062
k= 15		CHI =	84.8339526057967

Fig.6. The 15-cluster solution with the corresponding Calinski-Harabasz Index

5.3. Comparative Analysis

The proponents evaluated the performance of the conventional K-means firefly algorithm and the enhanced variation through the following internal validation indices: Silhouette Coefficient, Davies-Bouldin Index, and Calinski-Harabasz Index. The purpose of these metrics is to measure the uniqueness of the points inside the clusters compared to the points in other clusters by evaluating how directly associated the objects in a cluster are to one another and how different or well-separated a cluster is from other clusters (Xiong & Li, 2018).

Table 1. Comparison of K-means firefly and Enhanced K-means Firefly using internal validation metrics

Metrics	K-means Firefly	Enhanced K-means Firefly	Improvement (%)
Silhouette Coefficient	0.05729857583886415	0.5825827170554937	90.16%
Davies-Bouldin Index	2.4731157562171155	0.6679160989231555	72.99%
Calinski-Harabasz Index	62.84741593782496	202.62569971488244	68.98%

The Silhouette Coefficient (SC) and Calinski-Harabasz Index (CHI) show a better clustering outcome by maximizing their values, which is the opposite interpretation of the Davies-Bouldin Index (DBI). Table 1 shows that the enhanced K-means algorithm performed better in all validation indices than the K-means firefly, with a significant performance based on its percentage change. The results indicate that the improved algorithm produced a much better cluster.

6. Conclusion

The proponents of this study improved the K-means firefly algorithm by extending its capability to be used in real-world healthcare datasets while also improving the quality of the cluster analysis. This was done to address the algorithm's limitations discussed in its previous development. The enhanced K-means firefly is much more suitable for utilizing it in a real-time application, such as data dashboards and tracking websites, and data processing-related tasks.

The proponents suggest assessing the proposed algorithm with various internal or external validation indices for future consideration. In addition, the technique may be enhanced by studying its computational complexity and performance benchmarks once it has been implemented in software applications. Additional data pre-processing techniques and swarm intelligence should also be employed to enhance the algorithm further.

Acknowledgments

The researchers would like to acknowledge their families, friends, and colleagues for their endless support for the researchers' academic endeavors. The researchers would also like to thank their adviser, Raymund Dioses, and the faculty of Pamantasan ng Lungsod ng Maynila, College of Engineering and Technology - Computer Science Department, for their knowledge and guidance. Lastly, the researchers would also like to appreciate the support of Ricaflor De Leon for her contributions that progress this paper.

References

- Abdullah, D., Susilo, S., Ahmar, A. S., Rusli, R., & Hidayat, R. (2022). *The application of K-means clustering for province clustering in Indonesia of the risk of the COVID-19 pandemic based on COVID-19 data*. *Quality & quantity*, 56(3), 1283–1291. <https://doi.org/10.1007/s11135-021-01176-w>
- Ahmed P, K., Agrawal, R. (2020). Cluster Analysis of Health Care Data Using Hybrid Nature-Inspired Algorithms. In S. De, S. Dey, Siddharthal, B. (Ed.). *Recent Advances in Hybrid Metaheuristics for Data Clustering* (pp. 101-111). John Wiley & Sons Ltd.
- Baruah, I. D. (2020, October 25). Cheat sheet for implementing 7 methods for selecting the optimal number of clusters in Python. Towards Data Science. <https://towardsdatascience.com/cheat-sheet-to-implementing-7-methods-for-selecting-optimal-number-of-clusters-in-python-898241e1d6ad>
- Berisha, V., Krantsevich, C., Hahn, P. R., Hahn, S., Dasarathy, G., Turaga, P., & Liss, J. (2021). *Digital medicine and the curse of dimensionality*. *Npj Digital Medicine*, 4(1), 153. <https://doi.org/10.1038/s41746-021-00521-5>
- Gohari, K., Kazemnejad, A., Sheidaei, A., & Hajari, S. (2022). *Clustering of countries according to the COVID-19 incidence and mortality rates*. *BMC Public Health*. <https://doi.org/https://doi.org/10.1186/s12889-022-13086-z>
- Hutagalung, J., Ginantra, N. L. W. S. R., Bhawika, G. W., Parwita, W. G. S., Wanto, A., & Panjaitan, P. D. (2021). *COVID-19 Cases and Deaths in Southeast Asia Clustering using K-Means Algorithm*. IOP Publishing. <https://doi.org/10.1088/1742-6596/1783/1/012027>
- Kumar, V., Tan, P.-N., Steinbach, M., & Karpatne, A. (2018). *Introduction to Data Mining*. Pearson.
- Li, B. (2018). An experiment of k-means initialization strategies on handwritten digits dataset. *Intelligent Information Management*, 10(02), 43–48. <https://doi.org/10.4236/iim.2018.102003>
- Ocampo, L., Aro, J.L., Evangelista, S.S., Maturan, F., Selerio, E. Jr., Atibing, N.M., Yamagishi, K. (2021). *On K-Means Clustering with IVIF Datasets for Post-COVID-19 Recovery Efforts*. *Mathematics*; 9(20):2639. <https://doi.org/10.3390/math9202639>
- Orlov, K. (n.d.). Kirill's SPSS Macros Page | Raynald's SPSS Tools. Kirill's SPSS Macros Page | Raynald's SPSS Tools; www.spsstools.net. Retrieved May 15, 2022, from <https://www.spsstools.net/en/macros/KO-spsmacros/>
- Patra S. S., Harshvardhan, G. M. Gourisaria, M. K. Mohanty, J. R., Choudhury, S. (2020). Emerging Healthcare Problems in High-Dimensional Data and Dimension Reduction. In S. Roy, L. M. Goyal, M. Mittal (Ed.). *Advanced Prognostic Predictive Modelling in Healthcare Data Analytics* (pp. 25-49). Springer Nature Singapore Pte Ltd.
- Reddy, C. K., Vinzamuri, B. (2018). A Survey of Partitional and Hierarchical Clustering Algorithms. In V. Kumar (Ed.). *Data Clustering: Algorithms and Applications* (pp. 87-110). Chapman & Hall/CRC.
- Rizvi, S. A., Umair, M., & Cheema, M. A. (2021). *Clustering of Countries for COVID-19 Cases based on Disease Prevalence, Health Systems and Environmental Indicators*. Cold Spring Harbor Laboratory Press. <https://doi.org/10.1101/2021.02.15.21251762>
- Sharma, P. (2019, August 16). *The Most Comprehensive Guide to K-Means Clustering You'll Ever Need*. Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/>
- Singh, R. V., & Bhatia, M. P. S. (2011). Data clustering with modified K-means algorithm. 2011 *International Conference on Recent Trends in Information Technology (ICRTIT)*, 717–721. <https://doi.org/10.1109/ICRTIT.2011.597237>
- Thakur, N. K. (2020, April 19). *Comparison of Initialization strategies for k-Means*. Medium; medium.com. <https://medium.com/analytics-vidhya/comparison-of-initialization-strategies-for-k-means-d5ddd8b0350e>
- Virgantari, F., Faridhan, Y. E. (2020). *K-Means Clustering of COVID-19 Cases in Indonesia's Provinces* (Vol. 5, Issue 2). *International Journal of Natural and Engineering Sciences*.
- Whitelaw, S., Mamas, M. A., Topol, E., & Van Spall, H. (2020). *Applications of digital technology in COVID-19 pandemic planning and response* (Vol. 2, pp. 435–440). doi:10.1016/ S2589-7500(20)30142-4.
- Xiangyuan, H., Siyuan, L., & Hao, W. (2020). A survey on k-means initialization methods. <https://www.comp.nus.edu.sg/~amab/randalg20/HLW.pdf>
- Xiong, H., Li, Z. (2014). Clustering Validation Measures. In V. Kumar (Ed.). *Data Clustering: Algorithms and Applications* (pp. 571-602). Chapman & Hall/CRC.
- Zubair, Md., Asif Iqbal, Md., Shil, A., Haque, E., Moshui Hoque, M., & Sarker, I. H. (2021). An efficient k-means clustering algorithm for analysing covid-19. In A. Abraham, T. Hanne, O. Castillo, N. Gandhi, T. Nogueira Rios, & T.-P. Hong (Eds.), *Hybrid Intelligent Systems* (Vol. 1375, pp. 422–432). Springer International Publishing. https://doi.org/10.1007/978-3-030-73050-5_43