ISSN 1816-6075 (Print), 1818-0523 (Online)
Journal of System and Management Sciences
Vol. 13 (2023) No. 5, pp. 169-179
DOI:10.33168/JSMS.2023.0511

Improvement of k-Means Clustering Performance on Disease Clustering using Gaussian Mixture Model

Heru Agus Santoso[1], Su-Cheng Haw[2]

1 Faculty of Computer Science, Dian Nuswantoro University, Semarang, Indonesia
2 Faculty of Computing and Informatics, Multimedia University, Jalan Multimedia, 63100

heru.agus.santoso@dsn.dinus.ac.id

Abstract. k-Means clustering algorithm is an unsupervised learning, provides no opportunity for a data point to be a member of two or more clusters. In fact, a data point can belong to two or more clusters. In our dataset, a set of particular diseases can be member of different cluster locations. Gaussian Mixture Model (GMM) can solve the problem of this k-Means' hard assignment technique. Preprocessing approach on the dataset was also carried out using PCA after the result of Hopkins statistics far from sufficient for clustering purposes. PCA reduces the dimension of dataset, provides the most informative variables that explain the majority of the data. Hopkins test reached 0.958 after performing PCA, indicates the dataset has high tendency to cluster. Improving the performance of k-Means clustering with GMM using Log-likehood, GMM yielded a better result, i.e., 2.217 as compared to k-Means that yielded -606.604. It means GMM outperforms k-means in term of model fitness to the dataset.

Keywords: Clustering, soft-assigment clustering, k-Means, Gaussian Mixture Model.

1. Introduction

All machine learning algorithms generally have three types of learning approaches: supervised, unsupervised, and reinforcement learning (Bishop, 2006). However, current machine learning research is more focused on supervised learning than unsupervised and reinforcement learning. On the other hand, most of the available data for study is unsupervised data(Alvarez et al., 2022). Due to the high amount of unlabeled data produced, this situation is often compared to a birthday cake, where the cake itself is unsupervised learning data. The icing sugar on top of the cake is regarded as supervised learning data, while the cherry on top is considered reinforcement learning data. Therefore, using unsupervised learning algorithms to deal with the increasing amount of data is becoming increasingly crucial for machine learning algorithms. Some of the widely implemented applications using unsupervised learning algorithms include customer segmentation, dimensionality reduction, anomaly detection, image segmentation, and finding communities in social media, among others.

The k-Means algorithm is a widely used unsupervised learning algorithm. This non-probabilistic technique is popular for searching for clustering patterns. It works on unlabelled datasets and is optimized by minimizing the distance of each data point to the cluster centers, also known as centroids (Yuan and Yang, 2019). Additionally, each data point processed with the k-Means algorithm is assigned to a particular cluster. This technique is called "hard assignment," where each data point has only one cluster membership. This "hard assignment" technique works well under certain conditions and has several advantages, including reasonable processing costs, particularly for large volumes of data (Gao et al., 2020). Another advantage of the k-Means algorithm is the ease of interpreting the cluster results because each data point has only one definite membership.

The K-Means algorithm groups patterns visually in the form of a hyper-sphere or circular shape, with centroids as the centers of clusters. The radius of each cluster is calculated by the most distant data point from the centroid. This cluster radius acts as an area determining whether a data point is a member of the group or not. In certain conditions, this cluster pattern allows for overlapping between cluster areas. However, this weakness occurs when the data pattern is complex or non-linear. This condition can reduce the grouping performance in certain applications (Garcia-Dias et al., 2020), such as when a data point can be a member of two or more clusters. In real-world applications, a data point can indeed belong to two or more groups. With the popular K-Means algorithm, this grouping technique provides the potential for inappropriate cluster results. This so-called "hard assignment" of the K-Means algorithm does not provide the probability of a data point to be a member of each possible cluster. In this article, the authors highlight two main weaknesses of the K-Means algorithm, namely the inflexible cluster membership pattern and the exclusion of probabilities in data point membership.

One of the probabilistic clustering techniques to overcome the problem of data point "hard assignment" is the Gaussian Mixture Model (GMM). As a "soft assignment" technique, GMM is used when uncertainty occurs in the process of data point clustering(Li et al., 2017). In reality, data distributions are not always inside a certain definite circle radius. Generally, data in real-world applications is normally or Gaussian distributed. Hence, the idea of combining a number of Gaussian models is expected to approach uncertain data distribution. This technique is capable of being used to represent this uncertainty.

Suppose we have a dataset of endemic diseases in an area, and we extract three Gaussian distributions from the dataset as distinct groups based on the density of the disease. Each of these Gaussian distributions can represent a green, yellow, and red area. At the end of the modeling, we should obtain three different Gaussian distributions on the x-axis. However, the performance of the GMM algorithm can still be optimized (Wang and Jiang, 2021). One such optimization is the Expectation-Maximization (EM) algorithm. This maximum likelihood estimation technique is a density estimation approach that uses a probability distribution on a dataset and looks for optimal parameters. In certain machine learning algorithms, the maximum likelihood is intractable due to the presence of

hidden variables in the dataset known as latent variables. In this case, obtaining maximum likelihood estimation becomes challenging. Therefore, when dealing with datasets that have latent variables, the EM algorithm becomes one of the optimization techniques used to obtain maximum likelihood estimation. Typically, maximum likelihood estimation is achieved through the prediction of the value of the latent variable, followed by repeated optimization of the algorithm until it reaches an optimal solution.

2. Related works

The approach of data clustering to monitor public health in the Italian population based on changes in the health status of the area was carried out using the k-means algorithm. In this study, the k-means clustering used fuzzy membership degrees, which focused on functional data clustering problems. Due to the nature of the k-means algorithm, where the number of clusters has to be stated at the beginning of the process, this can cause poor clustering results. The study considered the possibility that a data point could be owned by more than one cluster at the same time. Therefore, the degree of membership needs to be taken into account in the clustering process. In short, this study proposes a functional fuzzy clustering algorithm to identify the similarity of patterns in functional data in the form of "health composite" indicators in the Italian region between 2010 and 2015. The practical benefit of this research is that it produces methods to monitor the risk of national health imbalances by identifying patterns at the local level. They use the term fuzzy functional classification rather than clustering in this study, even though the algorithm used is unsupervised clustering, namely k-means (Maturo et al., 2020).

Another study in the health sector that used the k-means clustering algorithm was aimed at grouping populations by utilizing data from health insurance (Zahi and Achchab, 2019). The aim of the study was to find patterns in the dataset that could be used to monitor general insurance coverage, especially health insurance. The machine learning algorithm used in this study was able to form clusters of insurance members using a fast partition technique, making it easy to interpret the results. Overall, the study provides a valuable solution for monitoring insurance coverage and improving decision-making in the health sector (Zahi and Achchab, 2019).

A research study was conducted to use clustering techniques for improving food consumption habits, which can support the decision-making process (Baek et al., 2019). Adequate nutrient intake is crucial for maintaining good health, and unhealthy dietary habits can lead to chronic diseases. The goal of this study is to suggest food alternatives that not only satisfy individual preferences but also meet nutritional standards for optimal health. A hybrid clustering and ontology approach was used, utilizing data from three sources, including chronic disease data, nutrition, and a nutrition knowledge base. The euclidean distance metric was employed to calculate distance. The study employed k-means clustering techniques and knowledge base (ontology) to generate food recommendations that satisfy various filters, including the food cluster filter, food similarity filter, preference filter, and feedback (Baek et al., 2019).

The k-means algorithm was used to analyze Covid-19 epidemiological data and the background conditions that influence it. This study aimed to differentiate the spatial variants of COVID-19 by quantitatively analyzing datasets related to socio-demographic and epidemiological data for strategic planning in order to reduce the spread of the disease. The algorithm classified 89 countries into two clusters. The cluster analysis showed that the majority of America, Europe, and Australia were in cluster 2 with a high mortality rate. The k-means algorithm also produced clusters indicating that the higher the percentage of the population infected with COVID-19, the greater GDP was used for health costs. There is a correlation between public health costs and the incidence of COVID-19, and countries must strengthen their public health systems to properly handle COVID-19 (Chandu, 2020)

The disease caused by the new coronavirus has forced people to live in restricted environments, and productive activities have been hindered. Studies on the use of the GMM algorithm for clustering COVID-19 data were conducted in two studies, where models and predictions of the COVID-19 pandemic were constructed. In this study, two models were developed to capture the trend in the number

of cases, as well as to predict the increase in new cases the following day. The data used in this study were COVID-19 cases from India, Italy, and the United States. The study estimated the turnaround date of the trend and predicted the end date of the trend. This study produced promising results as an alternative method to predict and continuously monitor the COVID-19 pandemic (Singhal et al., 2020)

The mixture model has also been used to analyze the distribution of characteristics on the autism spectrum. Today, there are many studies on autism because it is alleged that the number of sufferers continues to increase. Recently, the Finite Mixture Model has been widely applied to the distribution of mixed populations of autism spectrum disorders. However, existing algorithms may not be suitable because of the mixed population conditions between people with autism and non-autism sufferers. One reason is that mixed populations often have an irregular pattern, such as a skewed pattern, so they cannot be resolved with a regular or circular pattern in general (Abu-Akel et al., 2019).

This study utilized a clustering technique to investigate differences in ecosystem health and their underlying factors in China (He et al., 2019). The objective was to gain a better understanding of regional variations in ecosystem health. The dataset consisted of ecosystem health data from various regions in China from 2000 to 2015. K-means clustering was used to analyze the spatial agglomeration data, and the determinants of different ecosystem health levels were examined. The results revealed an increase in ecosystem health from northwestern to southeastern China, with eleven regions exhibiting three distinct types of ecosystems. The study also found that the "moisture index" and "intensity of land use" were significant contributors to the health of the national.

3. Method

The k-means algorithm is an unsupervised machine learning algorithm used for clustering unlabeled data, which means data without information on the target attribute. The algorithm works by first determining the number of clusters to be formed from the population. This number is used to set the value of the k parameter and initialize the centroids. A centroid represents the center of a cluster. Next, each data point is assigned to its closest centroid based on its proximity. This forms the initial groups. The algorithm then iteratively improves the clustering by calculating the mean of all data points within each group and updating the centroid accordingly. The data points are then reassigned to their closest centroid, and the process is repeated until the clusters converge or reach a stopping criterion (“47. In Depth,” n.d.)

Algorithm 1: k-Means clustering

Input:
Dataset D={dp_1,dp_2,...,dp_n}
Number of preferred cluster k, k∈N

Output:
k cluster of D

```
Begin:
    Randomly choose {c1,c2...,ck} as initial centroids
    Repeat
        Compute similarity for dpn to centroids
        Train dpn to be assigned to a cluster
    Until the end of datapoints
    Compute new centroid for each cluster using mean
    If no change of dpn membership to a cluster
    then end
End
```

In the first iteration of the algorithm, centroids are randomly selected from the dataset. For a given data point dp1(a1,b1), the similarity computation with centroid c1(x1,y1) is carried out using a

similarity metric such as cosine similarity or Euclidean distance, as follows ("Euclidean Distance Geometry and Applications | SIAM Review," n.d.):

$$ed\big((a_1, x_1), (b_1, y_1)\big) = \sqrt{(a_1 - x_1)^2 - (a_2 - x_2)^2} \quad (1)$$

3.1. The drawback of k-means clustering

Based on the k-Means algorithm described above, we can analyze its drawbacks. One of them is the random determination of initial centroids. When an inappropriate initial centroid is chosen, it may lead to high computational costs. In addition, the determination of the cluster number also greatly affects the purity of the final cluster results. Furthermore, as mentioned above, k-Means provides no opportunity for a data point to be a member of two or more clusters. However, in real-world applications, a data point can belong to two or more clusters. For example, a student can have both sports and music hobbies and thus can be a member of both the sports and music clusters. Visually, clusters formed by k-Means are circular, whereas in the real world, the shapes of group clusters can be very irregular (Wang and Jiang, 2021)

The objective of the k-means clustering algorithm is to achieve optimal purity of clusters, but this depends on the dataset being processed. There are several optimization techniques to improve the performance of k-Means, including (Yuan and Yang, 2019):

1. Optimization of the centroid initialization technique.
2. Acceleration of the algorithm by reducing unnecessary computation
3. Use of the minibatch technique, which involves shifting the centroid slightly on each iteration
4. Determination of the optimal number of clusters, among others.

With several existing techniques available, k-Means still has a weakness when the distribution of data forms non-spherical shapes. According to research (Bouveyron et al., 2019), the Expectation Maximization (EM) algorithm of Gaussian Mixture Model (GMM) can solve the problem of k-Means' hard assignment technique.

3.2. Optimization of GMM with EM

The Gaussian mixed model (GMM) aims to find the mix of multi-dimensional Gaussian probability distributions that best models the input dataset. In the simplest case, GMM can be used to find clusters in the same way as k-means. GMM is an unsupervised machine learning algorithm that addresses the problem of inflexible circular membership patterns, while also considering the probabilities of datapoint membership. GMM is used to obtain the best Gaussian probability in modeling data clustering. In other words, each data point determines its optimum likelihood. Hence, GMM is a probabilistic clustering technique that takes "soft assignment" into account when assigning data points to appropriate groups (Chatterjee et al., 2022). For example, if we want to determine which distribution a data point comes from a mixture of Gaussian k, we can express it as:

$$P\ (dp_i = k\ |\vartheta) \quad (2)$$

Where:
P: probability of datapoint cluster membership
dpi : i^{th} datapoint
k: k^{th} cluster
ϑ: consist of mean μ, covariance σ and size/weight w

The above expression is used to determine the likelihood of a datapoint coming from Gaussian k. Hence, GMM is a probabilistic clustering approach that assumes the data points come from a mixture of a number of Gaussian distributions with a certain mean, covariance, and weight. If we have a random data point from the dataset, then the GMM distribution can be expressed as a function of Gaussian probability, formulated as follows:

$$f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \quad (3)$$

Where:
x: datapoint
μ :mean
σ:covariance

When dealing with multivariate datapoints, a Gaussian distribution can be considered as a linear combination of variables. This means that the Gaussian probability density function of a multivariate datapoint can be expressed as follows:

$$f(x|\mu,\sum) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} exp\left[-\frac{1}{2}(x-\mu)\Sigma^{-1}(x-\mu)\right] \quad (4)$$

Where:
$\sum$: covariance matrix

Hence, GMM is a linear combination of GM.

4. Result

Our experimental setup involves a dataset of 9702 patient records extracted from electronic health records. It includes five features: ID, age, sex, international code disease (ICD), and sub-district (address). In the data preprocessing phase, we applied techniques such as duplicate removal, outlier detection and removal, and the Hopkins Test (Adolfsson et al., 2019).

Table 1. Example of datapoints

age group	sex	ICD	Location
todler	M	A01.0	Genuk
todler	W	A01.0	Tembalang
todler	W	A01.0	Pedurungan
todler	M	A01.0	Semarang Selatan
todler	W	A01.0	Genuk

count	4009	4009	4009	4009
unique	4	2	938	16
top	TUA	P	A01.0	Tembalang
freq	1206	2323	77	1039

Table 1 provides example of five data points, M and W are male a woman respectively. Name of disease for A01.0 ICD is Typhoid Fever, whereas Location is name of sub-district in Indonesia.

4.1. Hopkins Test

The Hopkins Test is utilized to determine whether a dataset is suitable for clustering by calculating the Hopkins statistic. This statistic provides an indication of the cluster tendency, which refers to how easily the data can be clustered (Adolfsson et al., 2019). If the Hopkins test falls within the range of :

1. {0.01, …, 0.49} : clustering the data may not be appropriate.
2. {0.5,...,0.75}: data needs further investigation as it is moderate tendency.
3. >0.75 it has a high tendency to cluster.

$$H = \frac{\sum_{i=1}^{n} Yi}{\sum_{i=1}^{n} Xi + \sum_{i=1}^{n} Yi} \quad (5)$$

Where:

$\sum_{i=1}^{n} Yi$: sum of distance between each datapoint in dataset.

$\sum_{i=1}^{n} Xi$: sum of distance in randomly generated dataset.

The result of hopkins test for our dataset is 0.59. Therefore, it needs further preprocessing before clustering process. The preprocessing phase was carried out to improve the quality of the dataset before clustering. For example, we eliminated data with an ICD code that was less than 3, but it only resulted in a 0.64 for hopkins test, which is still not sufficient for clustering purposes.

4.2. Principal Component Analysis

To enhance the clustering outcome, we utilize Principal Component Analysis (PCA), which is a machine learning technique for dimensionality reduction. With PCA, the most informative variables that explain the majority of the data's variability are selected. The technique creates new variables, called Principal Components, to reduce the dimensions of the dataset. Here the steps for applying PCA (Feng et al., 2020):

1. Standardize dataset using mean and standard deviation
   std_data = (datapoints - mean) / std_dev
2. Compute covariance
3. Compute eigenvectors and eigenvalues
4. Select set of principal components
5. Transform the data

Fig. 1 shows the scatter plot of data after applying PCA. Number of dimensions reduced into two as the most informative features that explain the majoriy of data. The name of new feature or variable called Principal Components are PC1 and PC2 as depicted in Fig. 1. The Hopkins test yielded a result of 0.958 after applying PCA. As the resulted Hopkins test more than 0.75, this indicates that the dataset has a high tendency to be clustered.

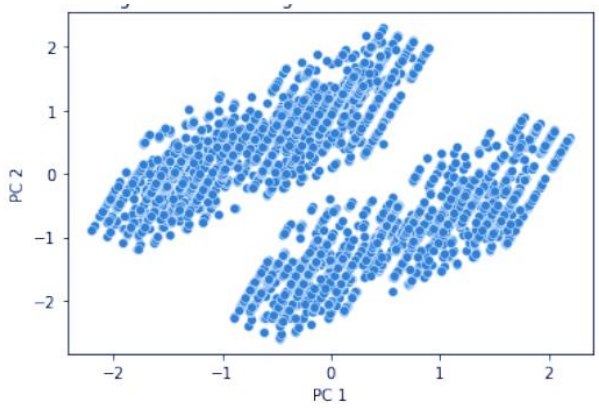

Fig. 1: Scatter plot after applying PCA

4.3. k-Means clustering

Before performing the clustering process with k-means, silhouette analysis (Jajuga et al., 2020) is needed to determine the optimal number of clusters for a given dataset. This analysis is useful for evaluating the quality of clustering.

Silhouette_value = (x - y) / max (x, y) (6)

Where:

x: average distance between data and all other data in its cluster;

y: average distance between data and all data in neighboring cluster.

Next, compute the average of silhouette value for all data in every cluster. Choose k value that gives the highest score. The silhouette value of a data shows the level of similarity of the data with its own cluster as compared to other clusters. The value is at the range of $-1 \leq$ silhouette_value ≤ 1. If the value is close to 1, then it indicates the data fits its own cluster and does not match the other clusters.

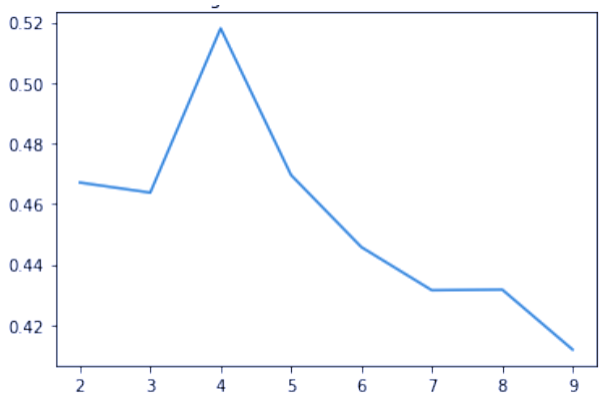

Fig. 2: Silhouette Value

Based on Fig 2, the highest score happens where the number of k = 4. Hence, we run k-Means clustering using number of cluster parameter = 4, and the following parameters in Python: KMeans(algorithm='auto', copy_x=True, init='k-means++', max_iter=50, n_clusters=4, n_init=10, n_jobs=None, precompute_distances='auto', random_state=50, tol=0.0001, verbose=0).

We use Davis-Bouldin Index (DBI) as a metric to evaluate the quality of clustering output. It measures the purity of clustering. The DBI uses lower values as a better clustering output. This indicates that the clusters are well separated and have high intra-cluster similarity and low inter-cluster similarity. In contrast, a higher DBI score indicates that the clusters are poorly separated and have low intra-cluster similarity.The DBI score for k-Means clustering on our dataset yielded a result of 0.79.

4.4. GMM Clustering

Applying GMM for clustering on the dataset, firstly we use both AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) to obtain the number of clusters in the dataset. We plot the result in the Fig 3 below:

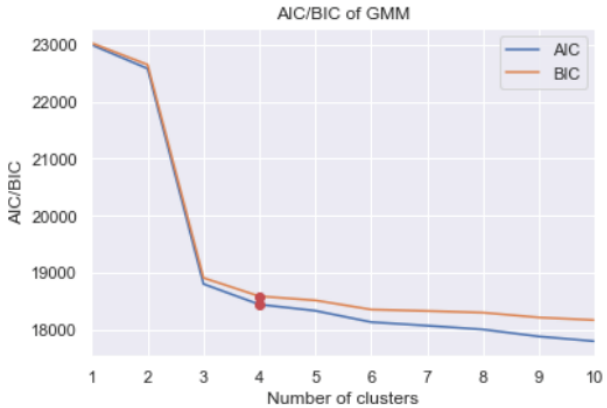

Fig. 3: number of clusters using AIC/BIC

Based on Fig 3, the number of cluster for GMM is 4. DBI for GMM clustering is 1.33. The other testing we perform is Log-likelihood (Chandler and Bate, 2007). It measures the fitness of a model to dataset. The higher value of the Log-likelihood, the better the fitness of model. Fig 4. Shows the comparison of two algorithm, k-Means and GMM based on Log-likelihood performance:

Fig. 4: Log-likelihood performance of GMM and k-Means

Based on Fig. 4, the result of Log-likelihood performance of GMM and k-Means are -2.217 and -606.604 respectively. It means, GMM outperforms k-means in term of model fitness to the dataset as GMM yielded a higher output. To obtain a better the visualization of the result, smotthing technique is used by utilizing sigmoid function.

5. Conclusion

k-Means clustering algorithm is an unsupervised learning, and non-probabilistic technique that works on unlabelled datasets. k-Means provides no opportunity for a data point to be a member of two or more clusters, as a matter of fact, a data point can belong to two or more clusters. Regarding information in our dataset, a set of particular diseases can be member of different cluster locations. The Expectation Maximization (EM) algorithm of Gaussian Mixture Model (GMM) can solve the problem of this k-Means' hard assignment technique.

Preprocessing approach on the dataset was carried out using PCA after the result of Hopkins statistics far from 1, i.e., 0.64. It is still not sufficient for clustering purposes. Hence, to enhance the clustering outcome, PCA is utilized to reduce the dimension of dataset. As a result, the most informative variables that explain the majority of the data's variability are selected. After performing PCA, the resulted Hopkins test reached 0.958. This indicates the dataset has high tendency to cluster.

In term of purity test using DBI, k-Means clustering algorithm yielded a result of 0.79 with the number of k is 4, after performing Silhoustte test to search the appropriate number of cluster. On the other hand, the resulted DBI test after performing GMM is 1.33. It showed that the characteristic of dataset is suitable for k-Means clustering in term of purity testing. However after testing using Log-likehood to measure the performance of clustering, GMM yielded a better result, i.e., 2.217 as compared to k-Means that yielded -606.604.

References

47. In Depth: k-Means Clustering - Python Data Science Handbook, 2nd Edition [Book] [WWW Document], n.d. URL https://www.oreilly.com/library/view/python-data-science/9781098121211/ch47.html (accessed 4.3.23).

Abu-Akel, A., Allison, C., Baron-Cohen, S., Heinke, D., 2019. The distribution of autistic traits across the autism spectrum: evidence for discontinuous dimensional subpopulations underlying the autism continuum. Mol. Autism 10, 24. https://doi.org/10.1186/s13229-019-0275-3

Adolfsson, A., Ackerman, M., Brownstein, N.C., 2019. To cluster, or not to cluster: An analysis of clusterability methods. Pattern Recognit. 88, 13–26. https://doi.org/10.1016/j.patcog.2018.10.026

Alvarez, M., Verdier, J.-C., Nkashama, D.K., Frappier, M., Tardif, P.-M., Kabanza, F., 2022. A Revealing Large-Scale Evaluation of Unsupervised Anomaly Detection Algorithms [WWW Document]. arXiv.org. URL https://arxiv.org/abs/2204.09825v1 (accessed 4.3.23).

Baek, J.-W., Kim, J.-C., Chun, J., Chung, K., 2019. Hybrid clustering based health decision-making for improving dietary habits. Technol. Health Care 27, 459–472. https://doi.org/10.3233/THC-191730

Bishop, C.M., 2006. Pattern recognition and machine learning, Information science and statistics. Springer, New York.

Bouveyron, C., Celeux, G., Murphy, T.B., Raftery, A.E., 2019. Model-Based Clustering and Classification for Data Science: With Applications in R. Cambridge University Press.

Chandler, R.E., Bate, S., 2007. Inference for clustered data using the independence loglikelihood. Biometrika 94, 167–183. https://doi.org/10.1093/biomet/asm015

Chandu, V., 2020. Identification of spatial variations in COVID-19 epidemiological data using K-Means clustering algorithm: a global perspective. medRxiv 2020.06.03.20121194. https://doi.org/10.1101/2020.06.03.20121194

Chatterjee, S., Romero, O., Pequito, S., 2022. Analysis of a generalised expectation–maximisation algorithm for Gaussian mixture models: a control systems perspective. Int. J. Control 95, 2734–2742. https://doi.org/10.1080/00207179.2021.1931964

Euclidean Distance Geometry and Applications | SIAM Review [WWW Document], n.d. URL https://epubs.siam.org/doi/abs/10.1137/120875909 (accessed 4.3.23).

Feng, C., Liu, S., Zhang, H., Guan, R., Li, D., Zhou, F., Liang, Y., Feng, X., 2020. Dimension Reduction and Clustering Models for Single-Cell RNA Sequencing Data: A Comparative Study. Int. J. Mol. Sci. 21, 2181. https://doi.org/10.3390/ijms21062181

Gao, B., Yang, Y., Gouk, H., Hospedales, T.M., 2020. Deep Clusteringwith Concrete K-Means, in: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Presented at the ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4252–4256. https://doi.org/10.1109/ICASSP40776.2020.9053265

Garcia-Dias, R., Vieira, S., Lopez Pinaya, W.H., Mechelli, A., 2020. Chapter 13 - Clustering analysis, in: Mechelli, A., Vieira, S. (Eds.), Machine Learning. Academic Press, pp. 227–247. https://doi.org/10.1016/B978-0-12-815739-8.00013-4

He, J., Pan, Z., Liu, D., Guo, X., 2019. Exploring the regional differences of ecosystem health and its driving factors in China. Sci. Total Environ. 673, 553–564. https://doi.org/10.1016/j.scitotenv.2019.03.465

Jajuga, K., Batóg, J., Walesiak, M., 2020. Classification and Data Analysis: Theory and Applications. Springer Nature.

Li, Q., Xiong, R., Vidal-Calleja, T., 2017. A GMM based uncertainty model for point clouds registration. Robot. Auton. Syst. 91, 349–362. https://doi.org/10.1016/j.robot.2016.11.021

Maturo, F., Ferguson, J., Di Battista, T., Ventre, V., 2020. A fuzzy functional k-means approach for monitoring Italian regions according to health evolution over time. Soft Comput. 24, 13741–13755. https://doi.org/10.1007/s00500-019-04505-2

Singhal, A., Singh, P., Lall, B., Joshi, S.D., 2020. Modeling and prediction of COVID-19 pandemic using Gaussian mixture model. Chaos Solitons Fractals 138, 110023. https://doi.org/10.1016/j.chaos.2020.110023

Wang, J., Jiang, J., 2021. Unsupervised deep clustering via adaptive GMM modeling and optimization. Neurocomputing 433, 199–211. https://doi.org/10.1016/j.neucom.2020.12.082

Yuan, C., Yang, H., 2019. Research on K-Value Selection Method of K-Means Clustering Algorithm. J 2, 226–235. https://doi.org/10.3390/j2020016

Zahi, S., Achchab, B., 2019. Clustering of the population benefiting from health insurance using K-means, in: Proceedings of the 4th International Conference on Smart City Applications, SCA '19. Association for Computing Machinery, New York, NY, USA, pp. 1–6. https://doi.org/10.1145/3368756.3369103