K-Means algorithm: An Unsupervised Clustering Approach using various Similarity/Dissimilarity measures

Surendra Singh Patel[1,2][0000-0001-5775-0513], Navjot Kumar[1,2], Aswathy J[1,2], Sai Krishna Vaddadi[1,2], SA Akbar[1,2], PC Panchariya[1,2]

[1] Academy of Scientific and Innovative Research (AcSIR), Ghaziabad, UP, India
[2] CSIR-Central Electronics Engineering Research Institute, Pilani, Rajasthan, India
patelsurendra333@gmail.com

Abstract. Clustering is an unsupervised method of classifying data objects into similar groups based on some features or properties usually known as similarity or dissimilarity measures. K-Means is one of the most popular method of clustering falls under the category of hard clustering. In this clustering method, any data object can belong to a single cluster. On the other hand, in soft clustering methods (e.g. fuzzy c-means clustering), the data object can be clustered in more than one cluster with some degree which is specified by the membership value with limitation imposed as the summation of these membership values should be equal to one. Although K-Means clustering technique is fairly old approach but still enjoy immense popularity in terms of being used in data grouping applications and machine learning. In this paper K-Means approach with five different distance measures e.g. Euclidean, Squared Euclidean, Half Squared Euclidean, Cosine and City Block distance has been explored and a comparative study is made based on the performance of these similarity criterions on real time Edible oil dataset acquired using MIR spectroscopy. Furthermore, it is also tried to investigate which similarity measure performs well for a particular set of data carrying unique pattern. The K-Means algorithm with various similarity-dissimilarity measures have been formulated and implemented in MATLAB R2015b environment provided by Mathworks.

Keywords: Supervised/Unsupervised learning, Clustering, Classification, K-Means clustering, Similarity and Dissimilarity measures.

1. Introduction

Clustering is an unsupervised method of classification, which divides the data samples into similar groups, possess similar characteristics. Clustering can be categorized into two subgroups i.e. hard clustering and soft clustering. The hard clustering means an individual data point can belong to a single group while the soft clustering allows the data points in more than one cluster with some degree of membership. Sensible grouping of data objects is to be considered an important part of data analysis in order to

realize the underlying pattern of unlabeled data. Data clustering is the means of understanding the characteristics or features of data objects by assigning them into similar groups based on these features. The unsupervised methods of data mining is found a bit complex as compared to supervised methods. Then one might be inquisitive why anyone is intended to address such unpromising problem. Some of the reasons could be identified as collecting the samples and labeling may surprisingly be expensive approach and in many data mining applications properties of the data might vary with time and subsequently the pattern may altered specifically in food applications where the food quality may vary due to season change. Hence these changes in the characteristics of pattern can be tracked in unsupervised learning methods and eventually improved performance can be observed [1]. Measuring the distance or similarity between two data objects is considered as an essential step in data mining applications [2]. Without being affected with the fact that complexity in cluster analysis is greater as compared to classification, there is an enormous scope of clustering techniques in any type of multivariate data analysis has been reported [3]. With the evolution of high-end storage and computational capability instruments and modern mathematical approaches, it has become possible to handle unprecedented amount of data. The goal of data mining is to extract the dominant features of the data so that underlying structure of the data can be identified and natural cluster can be formed [4]. The methods like segmentation, classification and clustering seems to be similar but they are really not. Since classification is a supervised method of grouping the data by assigning them into already defined groups while the clustering methods split the data objects which are not previously defined and segmentation on the other hand is the method of splitting the data objects based on similar characteristics [5]. Preprocessing considered as one of the major step for extracting the meaningful information from huge amount of data from spectroscopic analytical instruments. Navjot et al. used k-means algorithm for comparing the performance of various preprocessing techniques which are often used in multivariate data analysis [6]. The selection of initial cluster center is a sensitive step while using the k-means algorithms since it eventually affects the assignment of data points in the desired groupings. Four initialization methods have been studied to understand the effect of initial cluster center assignment [7]. In addition to it, some optimization based methods are also proposed to initialize the cluster center in [8] and refining initial starting conditions for k-means algorithm to avoid convergence to numerous local minima [9]. Khan et.al. proposed an algorithm to initialize the cluster center in order to make the convergence to a better local minima [10]. There are several methods proposed in the literatures to improve the efficacy of traditional k-means algorithm e.g. fuzzy entropy to improve the performance of traditional k-means [11], hybrid genetic algorithm [12][13]. Image segmentation is an important step in digital image processing. Clustering can also help in identifying the best segmentation technique to get the better insights of an image [14].

2. K-Means Clustering Algorithm

K-Means is one of the hard clustering method of classification. It splits the whole data samples into similar groups based on their similarity measure. Euclidean distance based similarity measure is the most commonly used method in these techniques. The basic algorithm have following steps:

Step 1: Chose the initial cluster centroids C_1, C_2.....C_k (where k is the number of clusters) from the given data set X_1, X_2, X_3.....X_n either randomly or using any analytical method.

Step 2: For each point X_i, find nearest centroid C_j

$$\underset{j}{\operatorname{argmin}}\, D(x_i, c_j)$$

Where D: Distance measure between data point x_i and cluster center c_j.

- Assign the point X_i to the cluster j.

Step 3: Compute new cluster center as:

$$C_j^* = \frac{1}{n_j} \sum_{x_i \to C_j} x_i$$

Step 4: If $C_j^* = C_j$ then the algorithm converges otherwise repeat from step 2.

Step 5: If algorithm does not converge in step 4 then it indicates that the algorithm executed maximum number of iterations.

The purpose of calculating distance between any two points is basically to check whether these two point are nearer to each other then we can infer that these two points are similar and vice-versa. In the following section, the definition and mathematical expression of five different types of similarity/dissimilarity measures are presented:

2.1. Euclidean Distance

The Euclidean distance is also known as L2 norm. The Euclidean distance between two points p1 and p2 having 'n' dimensions can be mathematically expressed as:

$$D_{Eucl.} = \sqrt{\sum_{i=1}^{n}(p1_i - p2_i)^2}$$

2.2. City Block or Manhattan Distance

The city block distance is also termed as L1 norm and this is the special case of Minkowski (Lp) distance with p = 1. The distance can be calculated between two points p1 and p2 as:

$$D_{city} = \sum_{i}^{n}|(p1_i - p2_i)|$$

2.3. Squared Euclidean Distance

In K-Means clustering, the results may not differ with squared euclidean distance and euclidean distance but since in this distance metric the square root need not to be calculated, the process of clustering comes out to be faster. The distance metric can be calculated with the following mathematical expression:

$$D_{Sq\_Eucl.} = \sum_{i=1}^{n}(p1_i - p2_i)^2$$

2.4. Half Squared Euclidean Distance

In half squared euclidean distance, also one has not to calculate the square root instead squared euclidean distance has to be made half. So, mathematically it can be expressed as:

$$D_{HSq\_Eucl.} = \frac{1}{2}\sum_{i=1}^{n}(p1_i - p2_i)^2$$

2.5. Cosine Similarity and Cosine Distance

Cosine similarity basically calculate the cosine angle between two points instead of distance calculation which can mathematically be expressed as:

$$\text{Cosine similarity} = \frac{\sum_i^n p1_i * p2_i}{\sqrt{\sum_i^n {p1_i}^2}\sqrt{\sum_i^n {p2_i}^2}}$$

And eventually,

Cosine distance = 1-Cosine similarity

3. Results and Discussion

In this study, seven edible oils (Mustard, Ricebran, Sunflower, Soya bean, Groundnut, Cottonseed and Olive oil) data is being used to compute the efficacy of various similarity/dissimilarity measures in k-means algorithm. The edible oils data was acquired using PYREOS make Attenuated Total Reflection (ATR with ZnSe crystal surface) based MIR spectrophotometer in the spectral range of 1818-909 cm^{-1}. A raw MIR spectrum of seven different oils is depicted in figure 1.

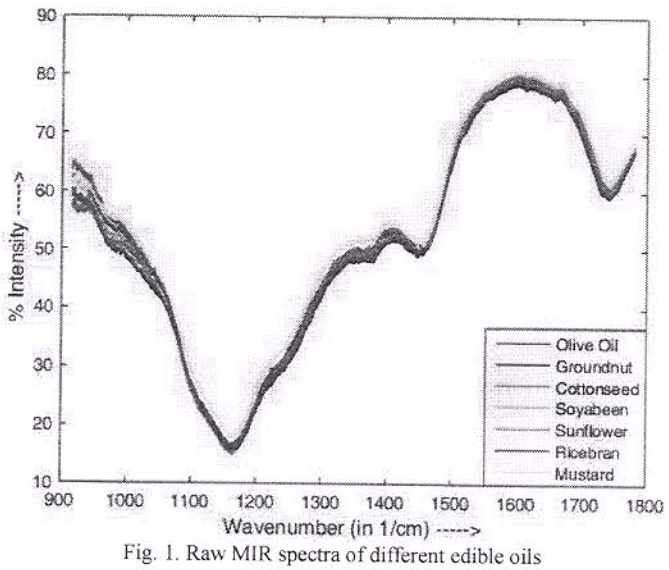

Fig. 1. Raw MIR spectra of different edible oils

The MIR spectra of oil data contains 128 variables, which is quite a large number in the view of computational complexity. Hence, principal component analysis, one of the most commonly used variable reduction technique [15, 16] utilized to improve the computational cost when dealing with multivariate data. In addition to this PCA is commonly used to visualize the high dimensional data for further data exploration [17]. The first seven principal components explaining approximately 99% of the information was used in the calculation of k-means in order to identify the efficient clusters. The PCA score plot for seven different edible oils is shown in fig. 2. PCA method can also be considered as a preliminary step in exploratory data analysis while performing clustering techniques. So in this work, PCA is performed on raw data to identify whether data clustering was possible on the data or not. The PCA score plot clearly indicates the grouping of different edible oils.

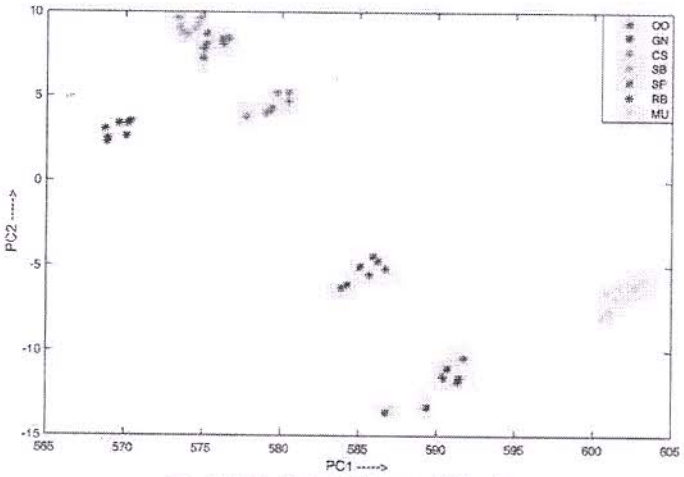

Fig. 2. PCA Plot of different edible oils

Confusion Matrix: Confusion matrix is considered as a performance measure for any clustering or classification technique. A perfect grouping of similar data points can be understood with the help of confusion matrix having only diagonal elements and all other elements as zero. The confusion matrix indicating the predicted class and actual class for seven different oils are presented in the tables (1-5).

TABLE 1. Confusion matrix for Euclidean Distance

Predicted / True	RB	GN	CS	SB	SF	RB	MU
RB	7	0	0	0	0	0	0
GN	0	7	0	0	0	0	0
CS	0	7	0	0	0	0	0
SB	0	0	4	3	0	0	0
SF	0	0	0	0	7	0	0
RB	0	0	0	0	0	7	0
MU	0	0	0	0	0	0	7

TABLE 2. Confusion matrix for city block distance

Predicted / True	RB	GN	CS	SB	SF	RB	MU
RB	7	0	0	0	0	0	0
GN	0	4	0	0	0	0	0
CS	0	7	4	0	0	0	0
SB	0	0	4	5	0	0	0
SF	0	0	0	0	5	0	0
RB	0	0	0	0	0	7	0
MU	0	0	0	0	0	0	2

TABLE 3. Confusion Matrix for Squared Euclidean Distance

Predicted / True	RB	GN	CS	SB	SF	RB	MU
RB	7	0	0	0	0	0	0
GN	2	5	0	0	0	0	0
CS	0	2	5	0	0	0	0
SB	0	0	2	5	0	0	0
SF	0	0	0	0	7	0	0
RB	0	0	0	0	0	7	0
MU	0	0	0	0	0	0	7

TABLE 4. Confusion Matrix for Half Squared Euclidean Distance

Predicted / True	RB	GN	CS	SB	SF	RB	MU
RB	7	0	0	0	0	0	0
GN	0	3	4	0	0	0	0
CS	0	0	0	7	0	0	0
SB	0	0	0	7	0	0	0
SF	0	0	0	0	7	0	0
RB	0	0	0	0	0	7	0
MU	0	0	0	0	0	0	7

TABLE 5. Confusion Matrix for Cosine Distance

Predicted / True	RB	GN	CS	SB	SF	RB	MU
RB	7	0	0	0	0	0	0
GN	0	7	0	0	0	0	0
CS	0	0	5	2	0	0	0
SB	0	0	0	5	2	0	0
SF	0	0	0	0	5	2	0
RB	0	0	0	0	0	7	0
MU	0	0	0	0	0	5	2

4. Conclusion

Various distance measures are reported in literatures and can be found in many standard statistical analysis software but in this paper, commonly used five-distance measures are implemented and tested on seven different category edible oil data. It can be observed from the confusion matrices (Table 1-5) and classification accuracy of individual distance measures presented in table 6, that for this particular data, half-squared Euclidean distance measure produced the best classification accuracy in terms of grouping the similar category oils in one cluster as desired.

TABLE 6. Classification Accuracy

Distance Measures	Classification Accuracy
Euclidean	77.55
City Block	69.39
Squared Euclidean	87.76
Half Squared Euclidean	77.55
Cosine Distance	77.55

It is also observed that while running the algorithm for several times, the inconsistent cluster centers and subsequently uncertain clusters are being formed which concluded that the K-means algorithm is not deterministic. This non-deterministic behavior of K-Means algorithm resulted due to random initialization of cluster center from one run to another. In addition, the K-means clustering is affected with the outliers. This effect of outliers can be minimized using median calculation in place of mean calculation of the data points while calculating the distance. The algorithm run was performed several times for each distance measurement to obtain the best-optimised cluster centre and classification accuracy. The average of values obtained in every run is presented in the form of the confusion matrix along with the classification accuracy.

Acknowledgements. We are thankful to the Director, CSIR-CEERI, Pilani for his continuous encouragement to carry out this work. We would also like to express gratitude towards all members of Signal Analytics Group (SAG) and CEERI Jaipur center for their cooperation and help.

References

1. Richard O. Duda, Peter E. Hart DGS (2001) Pattern classification. 738
2. Boriah S Similarity Measures for Categorical Data : A Comparative Evaluation
3. Jain AK (2010) Data clustering: 50 years beyond K-means. Pattern Recognition Letters 31:651–666. https://doi.org/10.1016/j.patrec.2009.09.011
4. Brunton SL, Kutz JN (2017) Data Driven Science & Engineering - Machine Learning, Dynamical Systems, and Control. 572
5. SAKTHIVEL E. (2015) Application of Fuzzy Logic in Data Mining
6. Kumar N, Panchariya PC, Patel SS, et al (2018) Application of Various Pre-Processing Techniques on Infrared (IR) Spectroscopy Data for Classification of Different Ghee Samples. In: 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA). pp 1–6
7. Lozano JA, Larra P (1999) An empirical comparison of four initialization methods for the K -Means algorithm. Pattern Recognition Letters 20:1027–1040
8. Babu GP, Murty MN (1994) Simulated annealing for selecting initial seeds in the k-means algorithm. 85–94
9. Bradley PS, Bradley PS (1998) Refining Initial Points for K-Means Clustering One Microsoft Way Refining Initial Points for K-Means Clustering. 91–99

10. Khan SS, Ahmad A (2004) Cluster center initialization algorithm for K -means clustering. 25:1293–1302. https://doi.org/10.1016/j.patrec.2004.04.007
11. Sharma J, Panchariya PC, Purohit GN (2013) Clustering Algorithm based on K-means and Fuzzy Entropy for E-nose Applications. 3–5
12. Krishna K, Murty MN (1999) Genetic K-Means Algorithm. IEEE Transaction on Systems, Man and Cybernetics - Part B: Cybernetics 29:433–439
13. Babu GP, Murty MN (1993) A near-optimal initial seed value selection in K-means algorithm using a genetic algorithm. 14:763–769
14. Manoharan S (2020) Performance Analysis of Clustering Based Image Segmentation Techniques. Journal of Innovative Image Processing 02:14–24
15. Kumar N, Panchariya PC, Kiranmayee AH, et al (2017) Rapid classification of different types of ghee using mid infrared spectroscopy. Proceedings - TIMA 2017: 9th International Conference on Trends in Industrial Measurement and Automation. https://doi.org/10.1109/TIMA.2017.8064802
16. Preview A (2010) Principal Components Analysis. 374–377
17. Gewers F, Ferreira GR, Arruda HF De, Silva FN (2018) Principal Component Analysis : A Natural Approach to Data Exploration