Applied Mathematics & Information Sciences Letters
An International Journal

http://dx.doi.org/10.18576/amisl/060102

Clustering Olive Oil Aspect of Similarity in the Compositions of Fatty Acid

Mohammad Ali Afshar Kazemi, Neda Kiani and Sima Hajian*

Department of Industrial Management, Central Tehran Branch, Islamic Azad University, Tehran, Iran

Received: 29 Apr. 2017, Revised: 13 Aug. 2017, Accepted: 19 Aug. 2017
Published online: 1 Jan. 2018

Abstract: In the international trades, the composition of fatty acid is the factor determining the olive oil price. Using the composition of fatty acid, one can investigate the quality of olive oil and also the probability of its impurity through determining the percentage of the fatty acid. As Italian olive oil is of high quality, this study tens to cluster the nine districts producing olive oil in Italy based on the similarities of the composition of fatty acid and study the level of fatty acid in every district. This paper tends to first identify and recognize the data based on the CRISP-DM cycle in data mining and cluster the nine districts based on the similarities of the composition of fatty acid through removing the unnecessary records in the stage of preparing a model based on the K-means method. Also, it is intended to determine the level of influence of the fatty acid in every cluster on the quality of olive oil in the district. According to the results obtained, the olive oil produced in the eastern districts of Liguria and Umbria is of higher quality.

Keywords: saturated acid, clustering, CRISP-DM cycle

1 Introduction

Being nutritious and being used for health, medical and industrial purposes, olive has always received significant attention. The reason for this is not only because it is natural, but also because of the unsaturated fatty acid, especially oleic acid in it [1].

The different kinds of oil and fat differ from each other with the kind and the amount of the saturated fat in them. Approximately 98 percent of olive oil is made of triglyceride which is a combination of saturated and unsaturated fatty acid and is reported to be of different proportions. This is because of the influence of different agricultural and environmental conditions. In general, of the several kinds of vegetable oil, olive oil is more desirable considering the composition of its fatty acid [2]. In the international trades, the composition of fatty acid is the factor determining the olive oil price. In the European market, the Tunisian olive oil is the cheapest because of the low level of oleic acid. According to the researches done on the structure of the fat, there are two kinds of olive oil containing low oleic acid and linoleic acid and high palmitic acid which the second kind obviously has a lower price [3].

Olive tree is one of the oldest trees planted through the history and in fact, it was planted before the written history [4]. Olive oil is one of its most important products and 93 of the olive produced worldwide are sent to the factories to make olive oil [5]. The composition of olive oil includes Triacylglycerol that makes about 97 of the natural oils and the rest are other trivial compositions. Next, there are free fatty acids with 0.5 to 1 percent of Non-Glyceride substances. These compositions of low amount are important for the sustainability and the taste of the olive oil. A quantitative analysis is the factor determining the originality of different kinds of olive oil [6].

98 to 99 percent of the composition of olive oil is made of a combination of saturated and unsaturated triglycerides. The composition of olive oil in the differential figures and different planting conditions varies significantly, but during the growth period the composition of olive oil is the same in all the kinds [7]. The comparison of the fatty acids in olive oil with the ones in the other vegetable oils shows that the total of the saturated and unsaturated fatty acids in olive oil is similar to the other ones, but what makes it distinct is the composition of its unsaturated fatty acids, meaning most of the composition of olive oil is made of oleic acid (olive

* Corresponding author e-mail: Neda_kiani81@yahoo.com

Table 1: The Data

class	Region	sample
1	North Apulia	25
2	Calabria	56
3	South Apulia	206
4	Sicily	36
5	Inner Sardinia	66
6	Coastal Sardinia	33
7	Eastern Liguria	56
8	West Liguria	56
9	Umbria	50

oil has an average of 55 to 83 percent of oleic acid). The composition of fatty acid varies in different kinds of oil based on the area they are produced. There are factors influencing on the composition of fatty acid in the olive oil like latitude, altitude, climate, variety and the level of ripeness [8]. Using the composition of fatty acid and determining the percentage of fatty acids, one can decide the quality and the probable impurity of the olive oil [4].

Researches have shown that the amount and the composition of fatty acids affect the quality of the olive oil. This study tends to investigate the similarities in the composition of fatty acids in the olive oil obtained from nine regions in Italy.

2 .Description of data and method

2.1 .The set of data on olive oil

The researcher, the case study is the clustering of 572 kinds of olive oil based on eight kinds of fatty acid in nine regions in Italy. The analysis of the data is presented in the paper by Forina, Arrnanino, Lanteri and Tiscornia [9]. This problem has been chosen, because it is a multi-class problem with nine classes that are not ordinal. The term 'non-ordinal' means that the class numbers are not hierarchical and do not reflect a natural ordering. This data consists of the percentage composition (100) of eight fatty acids (Palmitic, Palmitoleic, Stearic, Oleic, Linoleic, Linolenic, Arachidic and Ecosenic) found in the lipid fraction of 572 Italian olive oils. (An analysis of this data is given in Forina, Arrnanino, Lanteri and Tiscornia) [10]. There are nine collection areas, Are shown in Figure 1, including four from southern Italy (North and South Apulia, Calabria, Sicily), two from Sardinia (Inland and Coastal) and three from northern Italy (Umbria, East and West Liguria). The data arise from a study conducted to determine the authenticity of an olive oil, are shown in table 1.

The main question is that which regions are similar to each other and can be put in the same cluster based on the composition of fatty acid in the oil?

Fig. 1: The nine collection areas

2.2 .Data mining

The researcher, the analysis of the data is based on CRISP-DM cycle. The common steps of this model in data mining are shown in Figure 2 [4].

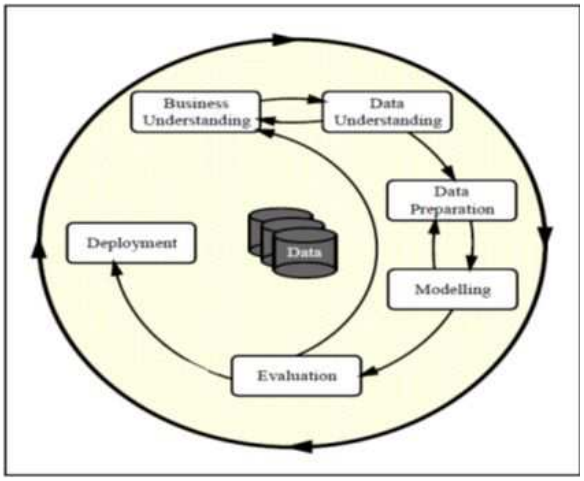

Fig. 2: CRISP-DM Cycle in data mining

What is important is that there should be a proper understanding of the CRISP-DM cycle steps, different modeling techniques should be selected and used and the parameters should be evaluated according to the optimum values. There are several modeling techniques in every data mining issue and some data require a special format. Due to the close relationship between the data preparation and modeling and mostly an initial understanding of the data on an issue and acquiring an initial idea for the upcoming limitations, this cycle contributes a lot in

choosing the kind of the model and doing the modeling [11,19].

After the preparation and identifying the model based on the main goal which is clustering the data based on the similarities of the composition of fatty acid in nine regions in Italy, the K-Means method was used. The last phase is CRISP-DM cycle that provides the strategy for developing the model. Identifying the clusters is not the end of the data mining project [12]. The main task is to understand that how the different regions based on the compositions of the clustering material and also the important finding should be used in for to refrain from high expenses. Therefore, every strategy will include a supervision or advice in order to improve the performance of the production process. The detail will provide a way to improve the operation based on using a similar composition in the nine regions in the analysis and conclusion part section [13].

K-Means clustering

K-Means clustering is one of the traditional learning algorithms which is studied well and solves the fundamental problems of clustering. This algorithm tries to discover the possible categories in the information through organizing the objects in groups with similar members. Therefore a cluster refers to a set of objects that are similar to each other and dissimilar to the ones in other groups. K-Means clustering can be considered the most important learning approach [11].

This method, although simple, is a basic one of many of the other clustering methods (like the fuzzy clustering). This is an exclusive and flat method [12]. There have been many forms for this algorithm, but all of them have a repetitive trend that, for a fixed number of the clusters, try to estimate the following:

Obtaining points as the contours of the clusters. These points are actually the main points belonging to every cluster.

Assigning every sample of data to a cluster whose center is the closest to the data.

In a simple kind of this method, first there are randomly selected some points as the number of the desired cluster. Then the data are assigned to one of the clusters based on closeness (similarity). Therefore, there are obtained new clusters. Through repeating this, one can calculate new centers in every repetition through getting the mean of the data and again assign them to new clusters. This continues until there are no changes in the data. The following function is the target function [14]:

$$J = \sum_{j=1}^{K} \sum_{i=1}^{N} \left\| x_i^{(j)} - c_j \right\|^2 \qquad (1)$$

In which the standard $\|\cdot\|$ Is the distance between? c_j And the center of the jet cluster.

Clustering using K-Means algorithm

Let $X = \{x_1, x_2, x_3, \ldots, x_n\}$ be the set of data points and $V = \{v_1, v_2, \ldots, v_c\}$ be the set of centers.

1. Randomly select $'c'$ cluster centers.
2. Calculate the distance between each data point and cluster centers.
3. Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers.
4. Recalculate the new cluster center using:

$$v_i = \frac{1}{c_i} \sum_{j=1}^{c_i} x_i \qquad (2)$$

   Where, $'C_i'$ represents the number of data points in i^{th} cluster.
5. Recalculate the distance between each data point and new obtained cluster centers.
6. If no data point was reassigned then stop, otherwise repeat from step 3 [15]

Although K-Means algorithm is common, but the solutions it provides depends on the initial values of the cluster centers [16].

The k-means clustering method has the following potential advantages [2]: (1) dealing with different types of attributes; (2) discovering clusters with arbitrary shape; (3) minimal requirements for domain knowledge to determine input parameters; (4) dealing with noise and outliers; and (5) minimizing the dissimilarity between data. However, for a successful application of the k-means clustering, we have to overcome its shortcomings: (1) the way of initializing means is not specified. One popular way to start is to randomly choose k of the samples; (2) the results produced depend on the initial values for the means, and it frequently happens that suboptimal partitions are found. The standard solution is to try a number of different starting points; (3) it can happen that the set of samples closest to some cluster centers is empty, so that they cannot be updated. This is an annoyance that must be handled in an implementation; (4) the results depend on the metric used to measure the dissimilarity between a given sample and a certain cluster center. A popular solution is to normalize each variable by its standard deviation, though this is not always desirable; and (5) the results depend on the value of k [17].

3 .Implementation of data

In this study, the implementation of the data mining process is done in two steps of preparation and modeling which are to be discussed: Preparation process:

This step includes the initial investigation of the data, correcting the current mistakes and increasing the quality of the data set. In this paper, the analysis of the data is

presented in the paper by Forina, Arrnanino, Lanteri and Tiscornia [18]. The data is purged and at this step one starts to investigate the lost values. The qualitative reporting of Data Audit is used for the investigation of lost values. Are shown in Figure 3.

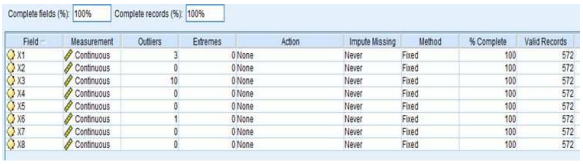

Fig. 3: Data Audit

As seen above, there are no lost values in the data set. However, one can see that some fields are full of straggled values. To investigate the straggled values in a multivariate way, the Anomaly Detection clustering technique was used to determine the different records of the original data body, Are shown in Figure 4.

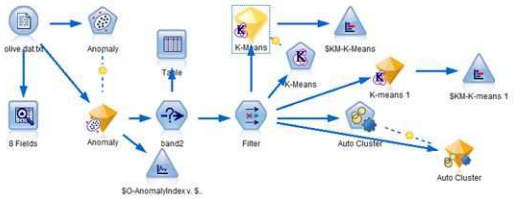

Fig. 4: Body Model

In this method first the data are clustered. Then through determining the norm in every cluster and the distance of every record from the cluster norm, the straggled values are analyzed. Figure 5 shows the distance of every record from the cluster norm individually:

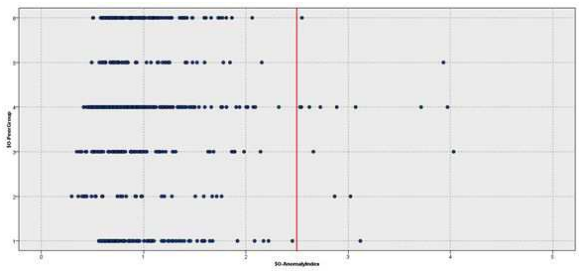

Fig. 5: The distance Records

As shown in the figure 5, considering the distribution of the Anomaly Index values, the 2.48 threshold value is determined for the identification of the straggled records. Based on this, 15 observations are considered as the straggled records of the data set and removed.

Modeling step

In this step, in order to answer the question of how to understand which olive oils from the different regions in Italy have similar characteristics based on the composition of the fatty acids in them? Here, the clustering model is used. Eight kinds of fatty acid (palmitic, palmitoleic, Stearic, Oleic, linoleic, linolenic, Arachidic and Ecosenic) which are referred to with the variable $X_1 - X_8$, are the input fields of the clustering model. To do the clustering, the K-Means algorithm is used, Are shown in Figure 6. The number of algorithms, 5 to 9, is determined according to the data volume and seems to be appropriate.

Fig. 6: K-Means Algorithms

This process shows that 5 clusters having the silhouette of 0.55 by the K-Means algorithm have the highest separation quality in the clusters. The level of importance in every input field is shown in Figure 7:

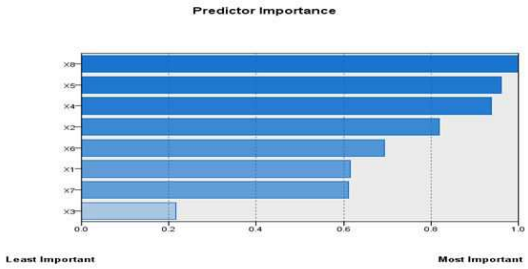

Fig. 7: The level based on predictor importance

As it can be seen the fatty acids X_8, X_5 and X_4 (Ecosenic, Linoleic, Oleic) are influential in making homogeneous clusters. In figure 8 there is shown the distribution of the clusters based on the numbers of the records related to every cluster.

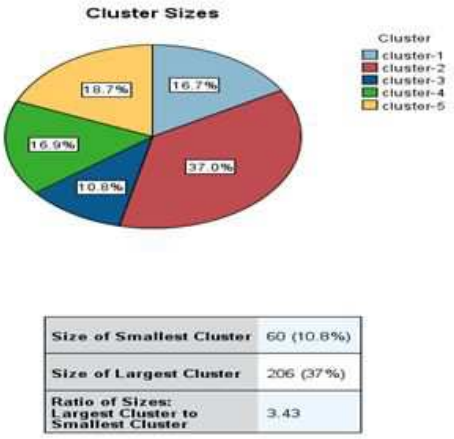

Fig. 8: the distribution of the clusters

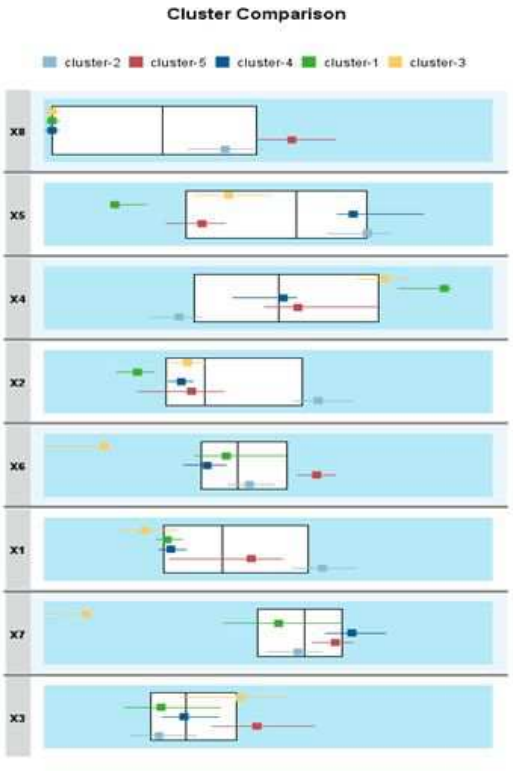

Fig. 9: cluster Comparison

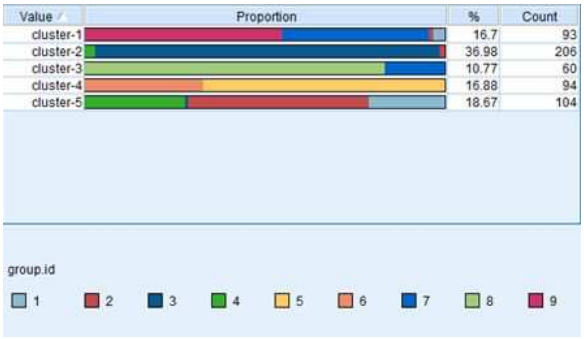

Fig. 10: the distribution of the regions

The biggest cluster contains 206 samples of olive oil equal to 37 percent of the total samples under study and it also is 3.43 times bigger than the smallest cluster with 60 samples equal to 10.8 percent of the whole. In figure 9 one can see the distribution of every fatty acid in each of the clusters.

In the above figure, there is shown the box plot of every fatty acid. The individual distribution of fatty acids in the clusters is shown by small box plots. The first cluster includes the samples low in Palmitic, Palmitoleic, Linoleic and Ecosenic acids (X_8, X_5, X_2 and X_1) And high in oleic acid (X_4). The second cluster includes the samples high in Linoleic, Palmitoleic, and Palmitic (X_5, X_2 and X_1) And low in oleic acid (X_4). The third cluster includes the samples low in Ecosenic, linolenic, palmitic and Arachidic (X_8, X_6, X_1 and X_7) And high in Oleic and Stearic acid (X_4 and X_3). The samples in the fourth cluster are low in Ecosenic, Linolenic, Palmitoleic and Lalmitic (X_8, X_6, X_2 and X_1) And high in Arachidic and Linoleic (X_7 and X_5). The samples in the fifth cluster are high in Ecosenic, Linolenic, Arachidic, and Stearic (X_8, X_6, X_7 and X_3) And low in linoleic acid (X_5). In this graph, Are shown in figure 10, the distribution of the regions in the clusters is shown by colors.

The purity of the colors in clusters indicates the high quality of the clustering. It means that the regions are separated from each other based on the fatty acids and the regions with similar fatty acids are put in the same cluster so that most of the clusters in one region are in the same cluster. The first cluster contains the samples from the regions 7 and 9 (the East Liguria and Umbria). The second cluster has the oils from region 3 (South Apulia). Most of its samples in the third cluster are from the region 8 (West Liguria). The fourth cluster has the samples from the regions 5 and 6 (Inner Sardinia and Coastal Sardinia). Finally, the fifth cluster contains the olive oils from the regions 4, 2 and 1 (Sicily, Calabria, and North Apulia Calabria).

4 .Conclusion

To identify the similarities in the nine regions in Italy for the quality of the olive oil based on the level of the fatty acids, clustering method and data mining were used as guidelines. The K-Means algorithm was performed on eight fields of the fatty acid level and the following results were obtained: the quality of the olive oil from the East Liguria and Umbria is the same in that they are low in Palmitic, Palmitoleic, Linoleic and Ecosenic and high in Oleic acid. The olive oil from Inner Sardinia and Coastal Sardinia have the same quality in that they are low in Palmitic, Palmitoleic, Linolenic and Ecosenic and high in Linoleic and Arachidic. The olive oil from Sicily, Calabria, and North Apulia, Calabria are high in Ecosenic,L, Arachidic and Stearic and low in Linoleic. Based on the researches done on the structure of the fat, there are two classes of olive oil. One of them is high in Oleic and low in Linoleic and Palmitic, and the other is low in Oleic and high in Linoleic and Palmitic. The second one has a lower price. Based on what was mentioned, the olive oils from the East Liguria and Umbria regions are of higher quality.

References

[1] Stati Ben, M., D. Gerasopoulis, I. Metzidakis, 1994. The effect of harvest maturity, temperature, modified atmosphere and salt on the olive quality of stored. Koroneiki. Sostanze Grass, LXXI: 235-241.

[2] Cristina. R, David A. Pelta, 2014,"Application of data mining methods for classification and prediction of olive oil blends with other vegetable oils", Anal Bioanal Chem 406:25912601.

[3] Rana, M. S. And Ahmed, A. A, 1981, Characteristics and composition of Libyan olive oil. JAOCS 58 (5) 630-631.

[4] Piravi-vanak, Z., Ghavami, M., Ezzatpanah, H., Arab, J., Safafar, H. And Ghasemi, J. B, 2009, Evaluation of Authenticity of Iranian olive oil by fatty acid and triacylglycerol Profiles, Journal of American Chemistry Society, 86: 827-833.

[5] Surinder K. And T.R. Sharma, 1991, Fatty Acid. Composition of Himachal. Olive Oil. Journal of Food Science Technology, India 28 (3) 171-173.

[6] Tous, J., Romero, A, 1994, Cultivar and location effects on the olive oil quality in catatonia (Spain). Acta Hort. 356:323-327.

[7] Fedeli.E. 1977, Lipids of olive prog. Chem. Fast and other lipids, 15:57.

[8] Firestone, D., KL. Carran, and Reina RJ.1998. Update on control of olive oil adulteration and misbranding in the United States, J. Am. Oil Chem, Soc.65:782-788.

[9] Forina, M., Armanino, C., Lanteri, S. & Tiscornia, E. (1983), Classification of olive oils from their fatty acid composition, in H. Martens & H. Russwurm Jr., Eds, Food Research and Data Analysis, Applied Science Publishers, London, pp. 189214.

[10] Forina, M., Lanteri, S. Armanino, C., Casolino, C., Casale, M., Oliveri, P, 2008, V-PARVUS. An Extendible Package of programs for explorative data analysis, classification and regression analysis. Dip. Chimica e Tecnologie Farmaceutiche ed Alimentari, Universit di Genova.

[11] Berry, M.J.A.; Linoff, G, 1997, Data Mining Techniques. For Marketing, Sales and Customer Support. Wiley Computer Publishing.

[12] E. Alpaydin, 2004, "Introduction to Machine Learning", The MIT Press.

[13] Jongsawas, Chongwatpol, 2015, "Prognostic analysis of defects in manufacturing", Industrial Management & Data Systems, Vol. 115 Iss 1 pp. 64-87.

[14] S.Z. Selim, M.A. Ismail, 1984, K-means type algorithms: a generalized convergence theorem and characterization of local optimality, IEEE Trans. Pattern Anal. Mach. Inteli 6 : 8187.

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

[16] J. Sander, 2003, "Principle of Knowledge Discovery in Data: Clustering I", Department of Computing Science University of Alberta, Tutorial Slides.

[17] Azzalini A., Torelli N, 2007, clustering via nonparametric density estimation. Statistics and Computing, 17, 71-80

[18] O. Zion, Principles of knowledge discovery in databases, Chapter 8: Data Clustering, Lecturing Slides for CMPUT 690, University of Alberta, 1999.

[19] Rudiger, W., Jochen, H, 2000, CRISP-DM: Towards a Standard Process Model for Data Mining, 6:11

Mohammad Ali Afshar Kazemi has received his Ph.D in Operations Research from Science and Research Branch, Islamic Azad University in Iran, in 2003. He has published more conference papers and journal papers. He is a faculty member of Islamic Azad University, Central Tehran Branch and has an associate degree.His main research interests are: dynamical systems, optimization theory, intelligent network, neural network, Data Mining, simulation, Queue theory.

Neda Kiani is Ph.D. student in Industrial Management Department from Central Tehran Branch, Islamic Azad University in Iran.Her main research interests are: dynamical systems, optimization theory, intelligent network, neural network, Data Mining.

Sima Hajian is Ph.D. student in Industrial Management Department from Islamic Azad University Tehran Central Branch in Iran.Her main research interests are: optimization theory, Data Mining.