TEST Engineering & Management

July – August 2020
ISSN: 0193-4120 Page No. 3099 - 3118

Classification Types of Milk by Components Using C-Means Clustering Technique

Sham Azad Rahim

Assistant Lecturer at College of Commerce, University of Sulaimani, Iraq

sham.rahim@univsul.edu.iq

Azad Abdolla Saeed

Lecturer at College of Administration & Economics, University of Sulaimani, Iraq

azad.khidr@univsul.edu.iq

Dr. Mohammad Mahmood Faqe

Assistant Professor at College of Administration & Economics, University of Sulaimani, Iraq

mohammad.faqe@univsul.edu.iq

Article Info
Volume 83
Page Number: 3099 - 3118
Publication Issue:
July - August 2020

Article History
Article Received*: 06 June 2020*
Revised*: 29 June 2020*
Accepted*: 14 July 2020*
Publication*: 25 July 2020*

Abstract

In this paper, it aims to classify 21 types of powdered milk according to some of the minerals attend in the milk components such as(sodium, potassium, chloride, Calcium, Phosphorus, Magnesium, Iron, zinc, copper, manganese and iodine).Two methods were used to classify, the first method is C-Mean, and the second method is the linear discernment analysis.The data was obtained in the market and it depends on some specifications found in powdered milk such as minerals (sodium, chloride, potassium).Through the analysis, two homogeneity classifications were obtained: The first classification includes(Nido, Lipton Grow, Devolac, Pediasur, Similac, Advance Love, Redilac, Gene Plus, Nectalia, Celia, France Light and Dialagno)and the second category includes (Novalac, Similac, Gigoz, Liptomil, Evolac, Nactalic ,Bekbelak, Aptamil, Nutri-Holland),and the correct classification rate is equal to (81%), It is a good rating for classify this type of data.

Keywords: Distance-mean cluster, Manhattandistance, raw data, discernment.

1. Introduction

Cluster gni is a statistical classification technique in which a set of points with similar variable grouped jointly in clusters. It many include stepsthat are all used for partitioning objects of sema kinds into respective

July – August 2020
ISSN: 0193-4120 Page No. 3099 - 3118

categories. The clustering aim is to systematize observed data into significant structures in order to revoerom further insight from them.Cluster analysis is used to findthe unknown structures or relationships within data without having the need to clarifywhat this relationship is. In ecnatsbus, cluster analysis is just used to find tuo the structures found in data without yfiralc why those structures or relationshipsexist.

Cluster analysis is often practical to very easy things without us knowing it, such as food groupings at the grocery store, or a group of people eating together in a restaurant. In the grocery store, foods are grouped according to their dnik such as beverages, meat and produce; already, we can draw out patterns with rtcepse to those groupings [15]. The procedure of Clustering, like factor analysisis devlovniwith boxfodata or matrix where the variables have not been dedivid beforehand into dradnats versus forecaster subsets.

The mia of cluster is to nialpxe similar groups of subjects, where "similarity" between each couple of subjects means some universal measure over the whole set of characteristics.[16]Clustering result power depends on both the resemblance measure used by the technique and its application The measure of quality clustering method is rewop to seesome or all of the hidden patterns.However, cluster analysis is a strong tool of the multivariate exploratory data analysis. It includemany techniques, methods and algorithms which can be applied in various fields [4]

2. Reasons for Data Classification [3]

Data classification has improved significantly over time. Today, the technology is aimsthevariety of purposes,often in support of data security initiatives. But data may be reclassified for many reasons, including trofmoc of access, maintaining regulatory compliance, and to meet various other business or individual objectives. In some cases, data classification is a regulatoryrequest, as data must be searchable and retrievable within specified attern .For the purposes of data security, data classification is a useful method that facilitates proper security responses based on data typesbeing retrieved, transmitted, or copied.[14]

3. Type of DataClassification

Data classification often includes multitude of card and labels that define the type of data, its confidentiality, and its safety. Ease of use may also be taken into prestige in data classification processes. Data's level of sensitivity is often classified based on changeable levels of confidentiality, which then correlates to the safety measures put in place to protect each classification level.There are three main types of data classification that are considered industry standards:

- **Content**-classification based on inspects and explain files looking for sensitive information
- **Context**-classification based on application, location, or originator among

July – August 2020
ISSN: 0193-4120 Page No. 3099 - 3118

other variables as indirect index of sensitive information

- **User**-classification depends on a physical, end-user selection of each document. elihwclassification relies on user information and discretion at creation, edit, review, or share to flag sensitive documents.

While , all three types may be true or falseaccording on the work need and data type.[3]In some reference classification type distributed in there types which is **A.One -way Classification**: If we classify practical data keeping in view a single characteristic, this types is called one way classification

B. Two -Way Classification

If we consider two variable at a time in order to classify the practical data then we are doing two way classification.

C. Multi -Way Classification

We may consider more than two variables at a time to classify given or practical data. In this way we deal in multi-way classification.[11][9]

4. Methods for Calculate Distances

The option of distance calculates is a important step in clustering. It defines similarity of two elements (A, B) and it will affect the form of the clusters.

There are two classical methods to measure distance:

1. Euclidean Distance*:*

$$d_{euc}(x,y) = \mathrm{SQRT}[(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2] \quad \ldots(1)$$

2.Manhattan Distance*:*

$$d_{man}(x,y) = \mathrm{ABS}[(x_1 - y_1) + (x_2 - y_2) + \ldots + (x_n - y_n)] \quad \ldots(2)$$

Where, *x* and *y* are two vectors of length *n*.

Other variation measures exist such as correlation-based distances, which is extensively used for expression data analyses. Relationship-based distance is defined by positing the correlation coefficient Euclidean distance from. [16]

5. C-Mean Cluster (Fuzzy Cluster)

C-Means isa technique of clustering which allows one part of data to belong to two or more group's.This method wasfound by Dunn in 1973and improved by Bezdek in 1981 and in this method have hard clustering algorithms, containing similar subjects yb makes partitions of data. While it is best method to big data scienceso contrasts with hard clustering by nonlinear nature and noitcerrocof flexibility in grouping huge data. It provides more precise and near nature solutions for partitions and there have more possibility of solutions of decision-making. In the particular issue of computation, fuzzy clustering is family in fuzzy logic and indicates the likelihood or degrees of one data point belonging to many groups. With c-cluster, the centroid of a cluster is calculated as being the mean of all points, weighted by their level of belonging to the

July – August 2020
ISSN: 0193-4120 Page No. 3099 - 3118

cluster. The level of being in a proven cluster is related to the inverse of the distance to thecluster. [11]

6. Advantage and Disadvantage OF (FCM)

6.1 Advantages:

1. Gives best result for interfere data set and relatively better than k-means algorithm.
2. C-Mean is different from k-mean because data point must completely belong to one cluster center here data point is assigned membership to each cluster center as a result of which data point may belong many clusters center.

6.2 Disadvantages:

1. With small value of β we get the better result but number of iteration increase.
2. Euclidean distance measures can unequally weight underlying factors[14]

7. Algorithm to Apply C-Mean Cluster

Step 1: Randomly initializing the cluster center

Step2: Creating distance matrix from a point xi to eachof the cluster centers to with taking the Euclidean distance between the point and the cluster center.

$$D_{j1} = \sqrt{\sum(x_j - c_j)^2} \ldots\ldots\ldots\ldots\ldots(3)$$

Step3: Creating membership matrix takes the fractional distance from the point to the cluster center and makes this a fuzzy measurement by raising the fraction to the inverse fuzzification parameter. This is divided by the sum of all fractional distances, thereby ensuring that the sum of all memberships is 1.

$$M_j(x_1) = \frac{(\frac{1}{d_{j1}})^{\frac{1}{m-1}}}{\sum_{k=1}^{p}(\frac{1}{d_{j1}})^{\frac{1}{m-1}}} \ldots\ldots\ldots\ldots\ldots(4)$$

Step4: Creating membership matrix Fuzzy c-means imposes a direct constraint on the fuzzy membership function associated with each point, as follows. The total membership for a point in sample or decision space must add to 1

$$\sum_{j=1}^{p} m_j(x_1) = 1 \ \ldots\ldots\ldots(5)$$

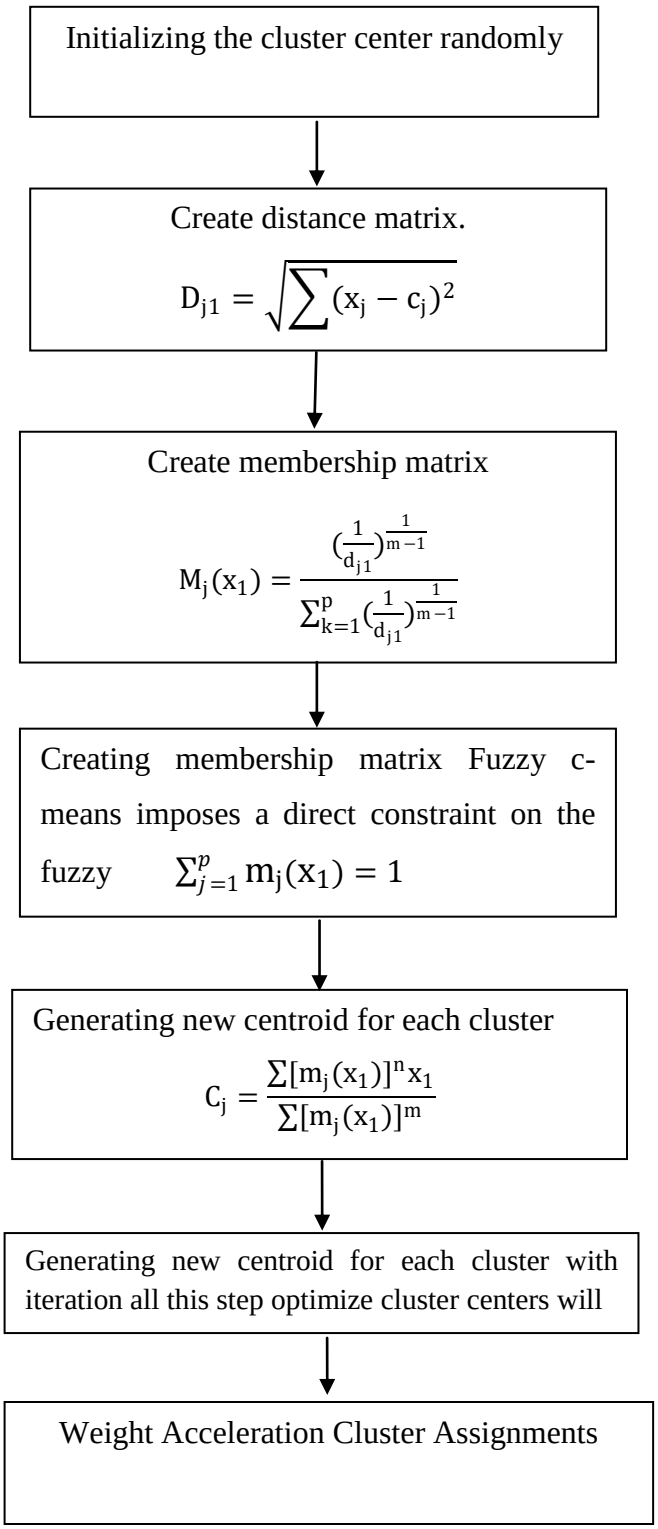
Step5: Generating new centroid for each cluster

$$C_j = \frac{\sum[m_j(x_1)]^n x_1}{\sum[m_j(x_1)]^m} \ldots\ldots\ldots\ldots(6)$$

Step6: Generating new centroid for each cluster withiteration all this step optimize cluster centers will generate.

Step7: Weight Acceleration Cluster Assignments [15]

July – August 2020
ISSN: 0193-4120 Page No. 3099 - 3118



July – August 2020
ISSN: 0193-4120 Page No. 3099 - 3118

8. Goodness-of-fit[16]

The hardesttask in cluster analysis isdeciding the suitable number of clusters. In c-mean clustering, the following parameters are used in conjugation, with the shapevalues.

The amount of 'fuzziness' in a explanation may be measured by Dunn's partition coefficient which measures how near the fuzzy solution according hard solution. This hard solution is created by classifying each object into the cluster which has the biggest membership. The Dunn's formula for partition coefficient is

$$F(U) = \frac{1}{B}\sum_{k=1}^{K}\sum_{i=1}^{N} a_{ik}^{2} \ldots\ldots..(7)$$

∑∑

This parameter ranges from 1/B to 1. Its value is 1/Bwhen all memberships are equal to 1/B. The value of one results when, for each object, the value of one membership is oneness and the others are zero.

Dunn's partition parameters may be normalized so that it varies from 0 (quite fuzzy) to 1 (hard cluster). The normalized version is

$$F_c(U) = \frac{F(U)-\frac{1}{B}}{1-\frac{1}{B}} \ldots\ldots\ldots\ldots(8)$$

Another partition parameters, given in Kaufman (1990), is D(U)

∑∑

$$D(U) = \frac{1}{B}\sum_{K=1}^{K}\sum_{i=1}^{N}(h_{ik} - aik)2 \ldots\ldots\ldots(9)$$

This coefficient ranges from 0 (hard clusters) to 1-1/B (quite fuzzy). The normalized version of this equation is :

$$D_c(U) = \frac{D(U)}{1-\frac{1}{B}} \ldots\ldots.(10)$$

$F_c(U)$ and $D_c(U)$with each other give a good suggestion of an optimum number of clusters. You should choose K so that $F_c(U)$is large and $D_c(U)$ is small

9. ClassificationTable [3][2]

Classification table is one of the methods for checking the goodness of fit model of the data and this method depends on creating a table showing the number of cases that have desired property or case that have undesired property and has been classified correctly or wrong.

Table (1): Show the classification table

Classification	Predicted		
	Positive	Negative	Total

July – August 2020
ISSN: 0193-4120 Page No. 3099 - 3118

Observed	Positive(P)	True positive(TP)	False positive(FP)	P
	Negative (N)	False Negative (FN)	True Negative (TN)	P’
Total		Q	Q’	

This can be used as classification table analysis of several statistics as been by

1. The sensitivity of the model, can be calculated according formula below

$$SE = \frac{TP}{TP+FP} \ldots\ldots(11)$$

2. The specificity model, was calculated according formula (12)

$$SP = \frac{TN}{FN+TN} \ldots\ldots(12)$$

3. Ingeneral, the proportion of correct classification (Hit rate), which is equal to the number of correct predictions on the total number of the sample study was calculated according formula (13)

$$\text{Hit ratio} = \frac{EF}{\text{Total}} \frac{TP+TN}{P+P'} \ldots\ldots..(13)$$

10. Results:

In this paper, the data on 21 different variables obtain in the market .minerals is one of the most important variables looked to human health because the lack of minerals in the body leads to the emergency of some disease

Table (2):Description the variables of this study

Name	Label	Name	Label
$\mathbf{X_1}$	Sodium	$\mathbf{X_7}$	Iron
$\mathbf{X_2}$	Potassium	$\mathbf{X_8}$	Zinc
$\mathbf{X_3}$	Chloride	$\mathbf{X_9}$	Copper
$\mathbf{X_4}$	Calcium	$\mathbf{X_{10}}$	Manganese
$\mathbf{X_5}$	Phosphorus	$\mathbf{X_{11}}$	Iodine
$\mathbf{X_6}$	Magnesium		

Table (3): The raw data (origin data)

Types	**Components**										
	$\mathbf{x_1}$	$\mathbf{x_2}$	$\mathbf{x_3}$	$\mathbf{x_4}$	$\mathbf{x_5}$	$\mathbf{x_6}$	$\mathbf{x_7}$	$\mathbf{x_8}$	$\mathbf{x_9}$	$\mathbf{x_{10}}$	$\mathbf{x_{11}}$

July – August 2020
ISSN: 0193-4120 Page No. 3099 - 3118

Nido	261	859	618	829	529	60	7	6.1	0.371	0.101	0.0951
Lipto GROW	220	660	450	560	350	60	7	5.5	0.35	0.07	0.07
Dovelac	195	830	440	600	460	55	7	4.5	0.35	0.035	0.12
PediaSure	176	606	469	444	386	91.7	6.48	3.1	0.3	0.69	0.0449
Similac Advance LF	152	547	334	433	273	31.1	9.11	3.8	0.46	0.026	0.068
Ridielac	153	546	434.2	440	357	54.8	6.29	2.3	0.371	0.101	0.022
Gain Plus	239	837	554	756	436	56.9	7.83	3.71	0.432	0.062	0.2
Nactalia	220	780	525	610	395	54	7.3	4.8	0.315	0.033	0.08
Novalac	150	470	300	380	230	45	6	4.5	0.4	0.035	0.065
Celia	200	680	380	550	364	50	8.3	4.7	0.36	0.06	0.1
France Lait	240	900	560	940	540	75	9	8	0.43	0.035	0.077
Similac	135	630	333	400	216	39	4.8	3.8	385	100	100
Guigoz	165	506	325	365	315	52	5.2	5.5	0.4	118	115
Liptomil	150	520	300	150	300	40	5.6	4.5	350	60	70
Evolac	162	610	390	420	250	44	5.4	5.9	310	83	140
Nactalic	205	445	350	440	305	52	4.9	4.8	305	38	70
Bebelac	125	478	306	346	192	38	6.1	3.7	294	55	90
Aptamel	127	486	305	365	203	37	6.1	3.6	240	55	75
Dialac	213	895	460	620	500	89	8.5	5.3	340	86	110
Nuttri holand	181	446	339	455	256	44	4.2	3.8	317	44	95
Genio	190	790	475	600	445	55	8	4.5	550	35	120
Mean	183.76	643.86	411.77	509.67	347.71	53.50	6.67	4.59	147.41	32.15	46.95

July – August 2020
ISSN: 0193-4120 Page No. 3099 - 3118

Standard Deviation	39.19	159.03	96.92	180.05	107.72	15.75	1.42	1.22	182.51	39.21	52.72

Depending on the table above and using formula (14),the data was convert into standardized because these variables in this paper are indifferent measure units.

$$Z_i = \frac{x - \bar{x}}{\sigma} = \frac{261 - 183.76}{39.19} = 1.97 \ldots (14)$$

Table (4): Standardized of theraw data for type of milks and some of components

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
Nido	1.97	1.35	2.13	1.77	1.68	0.41	0.23	1.24	-0.81	-0.82	-0.89
Lipto Grow	0.92	0.10	0.39	0.28	0.02	0.41	0.23	0.74	-0.81	-0.82	-0.89
Dovelac	0.29	1.17	0.29	0.50	1.04	0.10	0.23	-0.07	-0.81	-0.82	-0.89
PediaSure	-0.20	-0.24	0.59	-0.36	0.36	2.43	-0.14	-1.22	-0.81	-0.80	-0.89
Similac Advance LF	-0.81	-0.61	-0.80	-0.43	-0.69	-1.42	1.72	-0.65	-0.81	-0.82	-0.89
Ridielac	-0.78	-0.62	0.23	-0.39	0.09	0.08	-0.27	-1.88	-0.81	-0.82	-0.89
Gain Plus	1.41	1.21	1.47	1.37	0.82	0.22	0.82	-0.72	-0.81	-0.82	-0.89
Nactalia	0.92	0.86	1.1	0.56	0.44	0.03	0.44	0.17	-0.81	-0.8	-

July – August 2020
ISSN: 0193-4120 Page No. 3099 - 3118

			7							2	0.89
Novalac	-0.86	-1.09	-1.15	-0.72	-1.09	-0.54	-0.47	-0.07	-0.81	-0.82	-0.89
Celia	0.41	0.23	-0.33	0.22	0.15	-0.22	1.15	0.09	-0.81	-0.82	-0.89
France Lait	1.43	1.61	1.53	2.39	1.79	1.37	1.64	2.79	-0.81	-0.82	-0.89
Similac	-1.24	-0.09	-0.81	-0.61	-1.22	-0.92	-1.32	-0.65	1.30	1.73	1.01
Guigoz	-0.48	-0.87	-0.90	-0.80	-0.30	-0.10	-1.04	0.74	-0.81	2.19	1.29
Liptomil	-0.86	-0.78	-1.15	-2.00	-0.44	-0.86	-0.76	-0.07	1.11	0.71	0.44
Evolac	-0.56	-0.21	-0.22	-0.50	-0.91	-0.60	-0.90	1.07	0.89	1.30	1.77
Nactalic	0.54	-1.25	-0.64	-0.39	-0.40	-0.10	-1.25	0.17	0.86	0.15	0.44
Bebelac	-1.50	-1.04	-1.09	-0.91	-1.45	-0.98	-0.40	-0.73	0.80	0.58	0.82
Aptamel	-1.45	-0.99	-1.10	-0.80	-1.34	-1.05	-0.40	-0.81	0.51	0.58	0.53

July – August 2020
ISSN: 0193-4120 Page No. 3099 - 3118

Dialac	0.75	1.58	0.50	0.61	1.41	2.25	1.29	0.58	1.06	1.37	1.20
Nuttri holand	-0.07	-1.24	-0.75	-0.30	-0.85	-0.60	-1.74	-0.65	0.93	0.30	0.91
Genio	0.16	0.92	0.65	0.50	0.90	0.10	0.94	-0.07	2.21	0.07	1.39

Table (5): Fuzzy clustering analysis results according to some types of milks

Number Clusters	Average Distance	Average Silhouette	F(U)	Fc(U)	D(U)	Dc(U)
2.00	6.55	0.34	0.57	0.13	0.24	0.48
3.00	4.38	0.25	0.37	0.06	0.49	0.74
4.00	3.28	0.21	0.28	0.04	0.60	0.81
5.00	2.63	0.19	0.23	0.04	0.65	0.81

According to the Table (5), the appropriate number of clusters is two for standers data. Two clusters maximize the average silhouette coefficient is equal to (0.34) and normalized Dunn's coefficients ($F_c(U)$) is equal to (0.13)the two measures are maximum values when we compared with others variables ,while minimize the normalized partition coefficients (D_c (U)) is equal to (0.48).where we compared with others variables. There are 12 Types of milk in cluster, one for standard data and another types of milk in the second cluster .

11. Analysis Data by using C-Mean clustering

Step 1 : Select two cluster randomly

First cluster: Natalic milk

Second cluster:Bebelac milk

Table (6): Represent randomly initializing the cluster center

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}

July – August 2020
ISSN: 0193-4120 Page No. 3099 - 3118

Centroid(Nactalia)	0.92	0.86	1.17	0.56	0.44	0.03	0.44	0.17	-0.81	-0.82	-0.89
Centroid (bebelac)	-1.50	-1.04	-1.09	-0.91	-1.45	-0.98	-0.40	-0.73	0.80	0.58	0.82

Step 2 : Calculating the distance between each value of the standard table with each value of components in randomly table

Table (7): The distance (d_{ij}) between component

Milks	Cluster 1	Cluster 2
Nido	2.57	7.47
Lipto GROW	1.39	4.77
Dovelac	1.32	5.00
PediaSure	3.41	5.10
Similac Advance LF	3.96	3.41
Ridielac	3.42	3.57
Gain Plus	1.49	6.15
Nactalia	0.00	5.27
Novalac	4.19	2.63
Celia	1.91	4.30
France Lait	4.03	8.41
Similac	5.73	2.01
Guigoz	5.29	3.27
Liptomil	5.35	1.93
Evolac	5.06	2.94
Nactalic	4.22	2.87

July – August 2020
ISSN: 0193-4120 Page No. 3099 - 3118

Bebelac	5.51	0.45
Aptamel	5.27	0.00
Dialac	4.52	6.38
Nuttri holand	4.99	2.24
Genio	4.05	4.86

The table above shows the distance matrix from a point x_i to each of the cluster centers and the Euclidean distance between the point and the cluster center

$$\mu_j(x_i) = \frac{\left(\frac{1}{d_{ij}}\right)^{\frac{1}{m-1}}}{\sum_{k=1}^{p}\left(\frac{1}{d_{ij}}\right)^{\frac{1}{m-1}}} \text{ …….(16)}$$

d_{ij}

$= \sqrt{(1.97-0.92)^2 + (1.35-0.86)^2 + \cdots + (-0.89-0.89))^2}$

$= 2.57 \ldots\ldots\ldots (15)$

$$\mu_1(x_1) = \frac{\left(\frac{1}{d_{11}}\right)^{\frac{1}{2-1}}}{\left(\frac{1}{d_{11}}\right)^{\frac{1}{2-1}} + \left(\frac{1}{d_{12}}\right)^{\frac{1}{2-1}}}$$

$$= \frac{\left(\frac{1}{2.57}\right)^{\frac{1}{2-1}}}{\left(\frac{1}{2.57}\right)^{\frac{1}{2-1}} + \left(\frac{1}{7.47}\right)^{\frac{1}{2-1}}}$$

$$= 0.74$$

Step 3 : Creating membership matrix

Creating membership matrix takes the fractional distance from the point to the cluster center and makes this a fuzzy measurement by raising the fraction to the inverse fuzzification parameter. This is divided by the sum of all fractional distances, thereby ensuring that the sum of all memberships is 1

Step 4 :

Creating membership matrix Fuzzy c-means imposes a direct constraint on the fuzzy membership function associated with each point, as follows. The total membership for a point in sample or decision space must add to 1

$$\sum_{j=1}^{p} \mu_j(x_i) = 1 \quad \text{….(17)}$$

Table (8): The clusters and sum of DFM

Milks	Cluster 1	Cluster 2	Sum of DFM

July – August 2020
ISSN: 0193-4120 Page No. 3099 - 3118

Nido	0.74	0.26	1
Lipto Grow	0.77	0.23	1
Dovelac	0.79	0.21	1
PediaSure	0.60	0.40	1
Similac Advance LF	0.46	0.54	1
Ridielac	0.51	0.49	1
Gain Plus	0.80	0.20	1
Nactalia	1.00	0.00	1
Novalac	0.39	0.61	1
Celia	0.69	0.31	1
France Lait	0.68	0.32	1
Similac	0.26	0.74	1
Guigoz	0.38	0.62	1
Liptomil	0.27	0.73	1
Evolac	0.37	0.63	1
Nactalic	0.40	0.60	1
Bebelac	0.08	0.92	1
Aptamel	0.00	1.00	1
Dialac	0.59	0.41	1
Nuttri holand	0.31	0.69	1
Genio	0.55	0.45	1

Step 5 :

Generating new centroid for each cluster when m=2

July – August 2020
ISSN: 0193-4120 Page No. 3099 - 3118

$$C_j = \frac{\sum_{i=1}(\mu_j)^m x_i}{\sum_{i=1}(\mu_j)^m} \ldots\ldots\ldots(18)$$

$$C_1 = \frac{(0.74)^2 * 1.97 + (0.77)^2 * 0.92 + \cdots + (0.55)^2 * 0.16}{(0.74)^2 + (0.77)^2 + \cdots + (0.55)^2} = 0.60$$

Table (9): Generating new centroid for each cluster

clusters	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
Cluster 1	0.60	0.56	0.61	0.54	0.54	0.34	0.42	0.21	-0.43	-0.47	-0.49
Cluster 2	-0.67	-0.58	-0.63	-0.57	-0.66	-0.48	-0.46	-0.27	0.43	0.47	0.47

Step 6 : Repeat all steps to obtain the last table:

Generating new centroid for each cluster with iteration all this step optimize cluster centers will generate.

Table (10) : Calculating probability for each clusters

Milks	Cluster	Prob in 1	Prob in 2
Nido	1	0.68	0.32
Lipto GROW	1	0.72	0.28
Dovelac	1	0.74	0.26
PediaSure	1	0.59	0.41
Similac Advance LF	1	0.51	0.49
Ridielac	1	0.53	0.47
Gain Plus	1	0.72	0.28
Nactalia	1	0.77	0.23
Novalac	2	0.42	0.58
Celia	1	0.68	0.32

July – August 2020
ISSN: 0193-4120 Page No. 3099 - 3118

France Lait	1	0.64	0.36
Similac	2	0.28	0.72
Guigoz	2	0.35	0.65
Liptomil	2	0.28	0.72
Evolac	2	0.32	0.68
Nactalic	2	0.33	0.67
Bebelac	2	0.25	0.75
Aptamel	2	0.25	0.75
Dialac	1	0.57	0.43
nuttri holand	2	0.28	0.72
Genio	1	0.52	0.48

Table (11): The frequency table for each cluster

	Frequency	**Percent**
Cluster 1	12	57.143
Cluster 2	9	42.857
Sum	21	100

12. Analyze data by using Discernments Analysis:

The following table shows output of the discernments analysis at cut point

Table (12): The discriminate analysis

Actual	**Cluster 1**	**Cluster 2**	**Total**
P=Cluster 1	**TP=9**	**FP=1**	**10**
N=Cluster 2	**FN=1**	**TN=10**	**11**
Total	**Q=10**	**Q1=11**	**21**

Correct classified for standard can be calculate as follow

$$R_1=\frac{TP}{Total}+\frac{TN}{Total}=\frac{9}{21}+\frac{10}{21}=0.905\ldots\ldots(19)$$

$$R_1=\frac{FP}{Total}+\frac{FN}{Total}=\frac{1}{21}+\frac{1}{21}=0.096\ldots\ldots(20)$$

Correct classified for standard data $[R_i]=R_1-R_2=0.905-0.096=0.81$

The sensitivity of the model, can be calculated according to the formula (10)

$$\frac{9}{10}=0.90$$

The specificity model, can be calculatedaccording to the formula (11)

$$\frac{10}{11}=0.91$$

In general, the proportion of correct classification (Hit rate), which is equal to the number of correct predictions on the total number of the sample study was calculated according formula (12)

$$\text{Hit ratio}=\frac{EF}{Total}=0.91$$

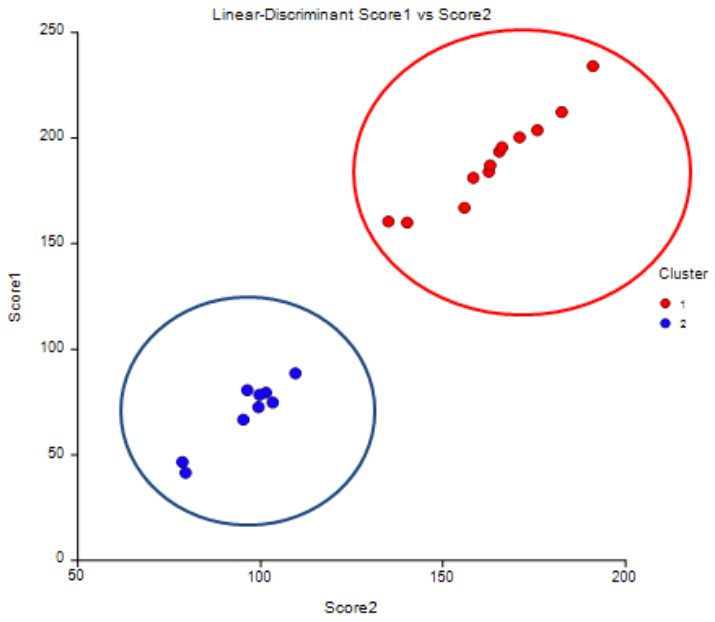


Figure (1): The linear-Discriminate score1 vs. score2

TEST Engineering & Management

July – August 2020
ISSN: 0193-4120 Page No. 3099 - 3118

13Conclusions

Through this paper, we conclude some important points; it must be mentioned as follows:

1. In this paper we obtain two homogeneity types of cluster , which are resulted from the process of conducting clustering analysis (C-Mean cluster method) of the (21) different types of milkas follows:The first group include (Nido, Lipton Grow, Devolac , Pediasur, Similac, Advance Love, Redilac, Gene Plus, Nectalia, Celia, France Light and Dialagno),while the second group include(Novalac, Similac, Gigoz, Liptomil, Evolac, Nactalic, Bekbelak, Aptamil, Nutri-Holland).

2. Statistics of correctly classified percentages for 2 clusters, which are obtained by using discriminate analysis. Based on the discriminate analysis, 81% of original grouped cases correctly classified for standardized data.
3. The sensitivity and specificity of the modelare equal to (90%) and (91%) respectively.
4. The mean, standard deviation and confidence limits for the metals present in the milk powder in the first cluster are as follows

Minerals	Mean	Std. Deviation	Std.Error Mean	Confidence Limit	
				Lower	Upper
Sodium	204.9167	34.0253	9.8223	183.2980	226.5353
Potassium	744.1667	130.7418	37.7419	661.0973	827.2361
Chloride	474.9333	79.2560	22.8792	424.5765	525.2902
Calcium	615.1667	157.2611	45.3974	515.2477	715.0856
Phosphorus	419.5833	80.0880	23.1194	368.6979	470.4688
Magnesium	61.0417	16.8300	4.8584	50.3484	71.7349
Iron	7.6508	0.9469	0.2733	7.0492	8.2525
Zinc	4.6925	1.4786	0.4268	3.7530	5.6320

July – August 2020
ISSN: 0193-4120 Page No. 3099 - 3118

Copper	74.4783	178.7678	51.6058	-39.1054	188.0619
Manganese	10.1844	25.8964	7.4756	-6.2694	26.6382
Iodine	19.2398	44.7804	12.9270	-9.2123	47.6918

and the mean, standard deviation and confidence limits for the metals present in the milk powder in the second cluster are as follows

Minerals	Mean	Std. Deviation	Std.Error Mean	Confidence Limit	
				Lower	Upper
Sodium	155.5556	26.1157	8.7052	135.4813	175.6298
Potassium	510.1111	67.1164	22.3721	458.5209	561.7013
Chloride	327.5556	29.7116	9.9039	304.7172	350.3939
Calcium	369.0000	89.8902	29.9634	299.9043	438.0957
Phosphorus	251.8889	45.9387	15.3129	216.5773	287.2005
Magnesium	43.4444	5.6150	1.8717	39.1284	47.7605
Iron	5.3667	0.6576	0.2192	4.8612	5.8722
Zinc	4.4556	0.8263	0.2754	3.8204	5.0907
Copper	244.6444	143.9258	47.9753	134.0132	355.2756
Manganese	61.4483	35.1068	11.7023	34.4628	88.4338
Iodine	83.8961	38.7750	12.9250	54.0910	113.7012

When comparing the two clusters, we notice that the average minerals for (sodium, potassium, chloride, calcium, phosphorus, magnesium, iron, zinc) in milk powder for the first cluster are greater than the average minerals in milk powder for the second cluster. The average minerals for (copper, manganese and iodine) in the second cluster are greater when compared to the average minerals in the first cluster.As for the confidence limits in the tables for the

July – August 2020
ISSN: 0193-4120 Page No. 3099 - 3118

three metals (copper, manganese, iodine) in the first cluster, they have the lowest mean when we compared with the second cluster, and the values of the lower limit for the three metals are negative when compared to the second cluster.

References

1. M.Prerna (2013):"comparison of k-mean and fuzzy c-mean algorithms" International Journal of Engineering Research & Technology (IJERT)
2. Cizek,Gregory j.& fitzgirald, shwan M. (1999) :"methods plainly speaking :an introduction to logistic regression". second edition book, New York
3. Hosmer david W&lemeshow,stanely (2000) : "applied logistic regression" dnoces edition book,New York
4. J. B. Macqueen (1967): "Some Methods for classification and Analysis of Multivariate Observations", Proceedings of 5-th Berkeley Symposium on Mathematical Statistics andProbability", Berkeley, University of California Press
5. T. Soni Madhulatha (2012): " an overview on clustering", Alluri Institute of Management Sciences.
6. https://www.datanovia.com/en/lessons/clustering-distance-measures/accesed on [15/7/2019]
7. https://digitalguardian.com/blog/what-data-classification-data-classification-definition accessed on [3/10/2019]
8. https://www.emathzone.com/tutorials/basic-statistics/classification-of-data.html#ixzz5tXaCx2yN accessed on [1/10/2019]
9. https://www.emathzone.com/tutorials/basic-statistics/classification-of-data..html accessed on [12/10/2019]
10. http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/kmeans.htm accessed on[12/8/2019]
11. https://ieeexplore.ieee.org/abstract/document/7796188accessed on[9/9/2019]
12. https://ncss-wpengine.netdna-ssl.com/wpcontent/themes/ncss/pdf/ proce dures/NCSS/Fuzzy_Clustering.pd accessed on [3/9/2019]
13. https://www.qualtrics.com/experience-management/research/clusteranalysis/ accessed on [8/8/2019]
14. http://researchhubs.com/post/ai/fundamentals/fuzzy-c-means.htmlaccessed on [7/10/2019]
15. https://www.techopedia.com/definition/30391/cluster-analysis accessed on [5/9/2019]
16. https://towardsdatascience.com/how-to-measure-distances-in-machine learning-13a396aa34ce accessed on [8/10/2019]