

Model-Based Discriminant Analysis and Two-Step Clustering for Breast Cancer patients

Sundus Al-Aziz and Refah Alotaibi

Mathematical Sciences Department, College of Sciences,
Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia

Abstract: Discriminant analysis (DA) is a classification problem, where two or more groups, clusters, or populations are known a priori and one or more new observations are classified into one of the known populations based on the measured characteristics. The primary objective of this work is to investigate the effect of the covariates of Breast cancer patients to discard covariates that are little related to group distinction and classify cases into groups. Also, Predicting surviving and deceased breast cancer patients, through the factors and determining the most distinct characteristic among patients, or determining discrimination function. Create a model of data using two steps clustering. The data for this project will consist of a study of 5396 patients with advanced breast cancer collected for 9-years (2004 to 2013). Thanks and appreciation to the Research Center at the Specialist Hospital in Riyadh, Saudi Arabia for their cooperation with us in obtaining data. We studied two cases and compared in two steps before the discrimination function was found or the cluster in two steps after finding the discrimination function for breast cancer patients' data.

Key words: Discriminant Analysis · Two-step Clustering · Breast Cancer · Model and classification

INTRODUCTION

The multivariate statistical analysis method and its different ways is based on the description and analysis of multidimensional and multivariable phenomena. If the views shared a set of characteristics and traits among themselves in varying degrees, then the multivariate statistical analysis addresses the study of the data of these views and expresses them through the most influential variables in the phenomenon understudying. It is used to classify individuals in groups on a scale or more, or to distinguish between groups based on the linear assembly of several measures after obtaining a significant Fisher value in the analysis of multivariate test. Discriminatory analysis is also used in the studies that aim to classify individuals into groups based on the number of predictive variables. And it is used in the analysis of multivariate for one dimension as the axial analysis and discriminatory analysis as a tracking measure. Based on the results of the discriminatory analysis, we can predict an individual's membership within a group or a particular set [1].

Several studies have used Discriminatory analysis such as Anandhalli applies the discriminant analysis module in this study to evaluate the influence of faculty member's socioeconomic characteristics on the use of Social networks in KSWU Bijapur. Data were obtained from 105 faculty members selected through stratified sampling method [2], Eloisa Urrechaga investigated the performance of multivariate discriminate analysis to the differential diagnosis of genetic and acquired microcytic anemia [3].

Özge Pasin, Handan Ankaralı,, Osama Abbas compared different clustering algorithms and he has concluded that EM algorithms had better performance from hierarchical clustering methods [4, 15], Kakkar and Parashar compared K-means, hierarchical methods, EM and density based algorithms that used in WEKA in 2014. As a result of their study, they observed that K-means clustering algorithm gave faster results than hierarchical and EM algorithm [5].

Publications that concentrate on clustering analysis such as Rana Walid has attempted to use the concept of clustering for modeling factual data to the ladder of

monthly salaries of teaching staff for in a college in Almousel University [6]. Where she applied the HCM algorithm on these data and the results proved the efficiency of this algorithm in clustering real data and how to represent them in clusters, Darkazanli and alaziz, applied study of the factors affecting age groups of the labor force using two-stage cluster analysis [7].

Other researchers, for example, Shih *et al.* [8] used a two-step method for clustering mixed categorical and numeric data. However, McNichol as used in their study the clustering, classification, discriminant analysis and dimension reduction via generalized hyperbolic mixtures propose a mixture of latent variables model for the model-based clustering, classification and discriminant analysis of data comprising variables with mixed type. This approach is a generalization of latent variable analysis [9].

The aim of the present study was to investigate the prediction factors influencing the survival of breast cancer patients or death of these patients and determine the function of discrimination and classification accuracy of patients with breast cancer (women) in order to find any significant differences among alive and dead breast cancer patients are:

Basic Concepts and Notation

Discriminatory Analysis: It is one of the important (Multivariate Analysis) methods as the variables involved in the model being analyzed in a coherent way, taking into consideration the correlations among these variables. It also seeks to form a statistical model that depicts the correlations among the different variables. Discriminate analysis model is based on access to Discriminate Function, which works on maximizing the differences between the averages of the groups and reducing the similarity of classification errors at the same time, by finding linear combinations of a set of variables [10].

It is well known that the discriminate analysis is one of methods of variables' data analysis. For the purpose of applying this method, two-Step should be held the first is classification step by insulating n from the items of a particular sample or dispersing it into several groups, so that the items within each group are similar as much as possible and are different from the items of the other group. The second is discrimination phase that's upon completion of the classification process for a specific sample or even for a specific community or several groups, by which we know the owner of any new item that did not subject for classification process to a group

according to its characteristics, this means that the classification process precedes the process of discrimination. It turns out that the main objective of this discriminate analysis is the building or the construction of a base derived from the qualities that the views afford, classified into two groups or more for a particular sample so that we can judge by this base on the ownership of the new item (beyond these groups) to one of them. The discriminate function applications are many in the practical life and are used in a lot of Applied Sciences and most of these applications were in the fields of biological, medical and social sciences [1]. In addition, the discriminatory analysis aims to design of linear combinations of the best variables in the subject of the study, check whether there are significant differences between the groups with respect to variables and identify the variables that contribute to the greatest differences among the dependent variable categories. Also, the discriminatory analysis aims to divide conditions among dependent variable categories based on the values of the independent variables and evaluate the classification accuracy (as a percentage) where the discriminant equation:

$$Y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_px_p + \varepsilon$$

where, Y is a latent variable formed by the linear combination of the dependent variable, $x_1, x_2, x_3, \dots, x_p$ are the p independent variables, ε is the error term and $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ are the discriminant coefficients.

The objective discriminant analysis is to test if the classifications of groups in a variable Y depends on at least one of the x_i 's. In terms of hypothesis, it can be written as:

H_0 : Y does not depend on any of the x_i 's.

H_1 : Y depends on at least one of the x_i 's.

OR simply, H_0 : $\beta_i = 0$ for $i=1, 2, \dots, p$ versus H_1 : $\beta_i \neq 0$ for at least one i.

Cut Point: To classify a new item or items, the point that cuts the two groups' values so that if the discriminatory value of the new item is more than the cut point, so this item will be classified for a particular group. However, if the discriminatory value was less the cut point, so this item will be classified for the other group and if the discriminatory value for this item was equal to the cut point, so this item will be classified for either of the two groups randomly [10, 11].

Wilks's Lamda: A ratio of the sum of squares within groups (Also known as the unexplained variation) to the Total sum of squares (The total variation). It represents a part of the overall variation among groups that hasn't been explained by discrimination function; hence, as the value of Wilks Lambda becomes smaller, this indicates the efficiency of discrimination function in the classification of the sample to the groups to which they belong [12].

Clustering: Clustering is one of the scientific developments developed by scientists in the field of knowledge and modern technologies to detect multiple totals. The concept of clustering emerged for the first time by scientist Ronald in 1955. The basic idea of clustering represented in the fragmentation of data into clusters, as the data points for a cluster are more similar to each other compared with those points in other clusters. The clustering can be defined as a programmed and multivariate statistical analysis process, through which the study of similarities among a set of variables of different and multiple models, then comparing these models with each other depending on their content of variables and the arrangement of their relations with each other in clusters form [13, 14].

Cluster Analysis: Cluster analysis is a set of tools for building groups (clusters) from multivariate data objects. The aim is to construct groups with homogeneous properties out of heterogeneous large samples. The groups or clusters should be as homogeneous as possible and the differences among the various groups as large as possible [15]. The Two-Step Cluster Analysis procedure's algorithm can be summarized as follows:

Step 1: The procedure begins with the construction of a Cluster Features (CF) Tree. The tree begins by placing the first case at the root of the tree is a leaf node that contains variable information about that case. Each successive case is then added to an existing node or forms a new node, based upon its similarity to existing nodes and using the distance measure as the similarity criterion. A node that contains multiple cases contains a summary of variable information about those cases. Thus, the CF tree provides a capsule summary of the data file [7, 13].

Step 2: The leaf nodes of the CF tree are then grouped using an agglomerative clustering algorithm. The agglomerative clustering can be used to produce a range of solutions. To determine which number of clusters

is "best", each of these cluster solutions is compared using Schwarz's Bayesian Criterion (BIC) or the Akaike Information Criterion (AIC) as the clustering criterion [7, 13].

The Study Hypotheses:

H_0 : The function does not have the ability to discriminate between Alive and Dead breast cancer patients.

H_1 : The function has the ability to discriminate.

In other words, the null hypothesis states the lack of influence of the independent variables on discriminatory values and it can be expressed mathematically as follows [11, 12]:

$$H_0: \beta_1 = \beta_2 = \beta_p = 0$$

$$H_1: \beta_1 \neq \beta_2 \neq \dots \neq \beta_p \neq 0$$

The Study Methodology and Sample: Dataset: The research used a descriptive analytical approach based on a discriminatory analysis of a sample of breast cancer patients and clustering data. Was extracted data from the years of 2004 to 2013 From the Saudi Cancer Registry (SCR) and the sample amounted to 5396 patients of breast Cancer.

RESULTS AND DISCUSSIONS

The comparison between the two cases by using spss program:

The First Case: Clustering data of breast cancer patients using a clustering method in two-Step, classifying them and finding the discriminatory function in steps:

Two-Step Cluster: Where we clustered data for the variables of (Time, Age, Grade, Extent, Marital Status, Address, Laterality and Case of Death) in two-Step and get to the following results:

As two clusters resulted and silhouette measure was in a high average close to 0.5 shows in Figure (1) and the first cluster ratio was 35.8%, as the second cluster ratio was 64.2%.

The clusters are as follows with the percentages and where two variables were introduced which are (time and age) in the continuous variable field and the (Extent) variables in the classified variables field, Then we entered the other variables in Evaluation Fields field. Comparing the resulting two clusters as follows:

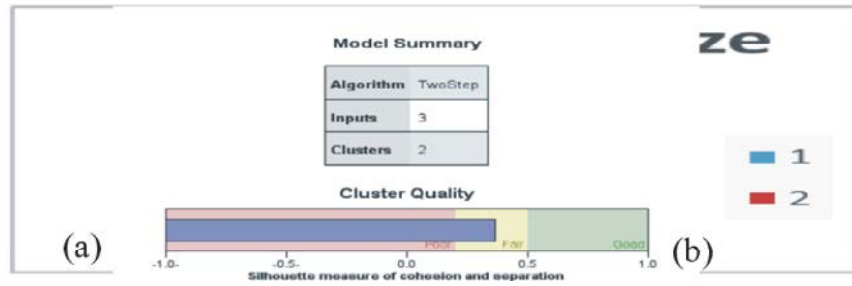


Fig. 1:

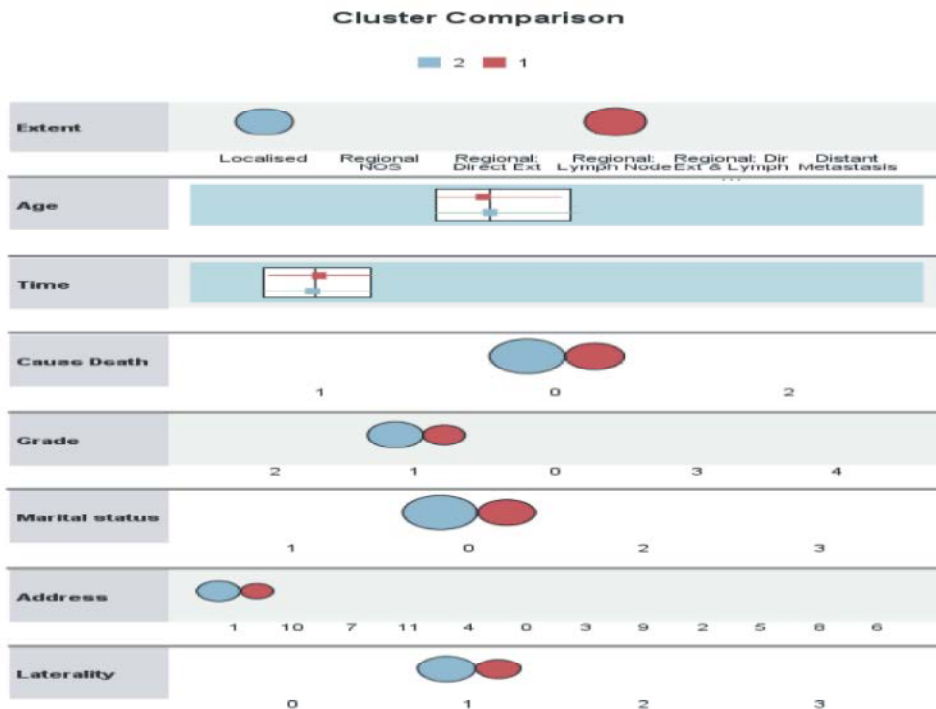


Fig. 2:

It is noted by the comparison that the two clusters for the variable of (Extent) centered in case of Localized and Regional: Lymph Node, as for the variables of (age and time), the two clusters are close, as for the variable of (Cause of Death). Data are clustering in the case of Not Applicable, as for the variable of (Grade) shows in Figure (2).

The following diagrams Figure (3) from (a to r) shows all variables and their percentage included in both clusters can be obtained from the spss program.

After we get a new variable as an output of clustering in two-Step with the dependent variable of (status), discriminant analysis used and we have gotten the following results:

Therefore, Table (1) shows descriptive statistics of the clustering two-step.

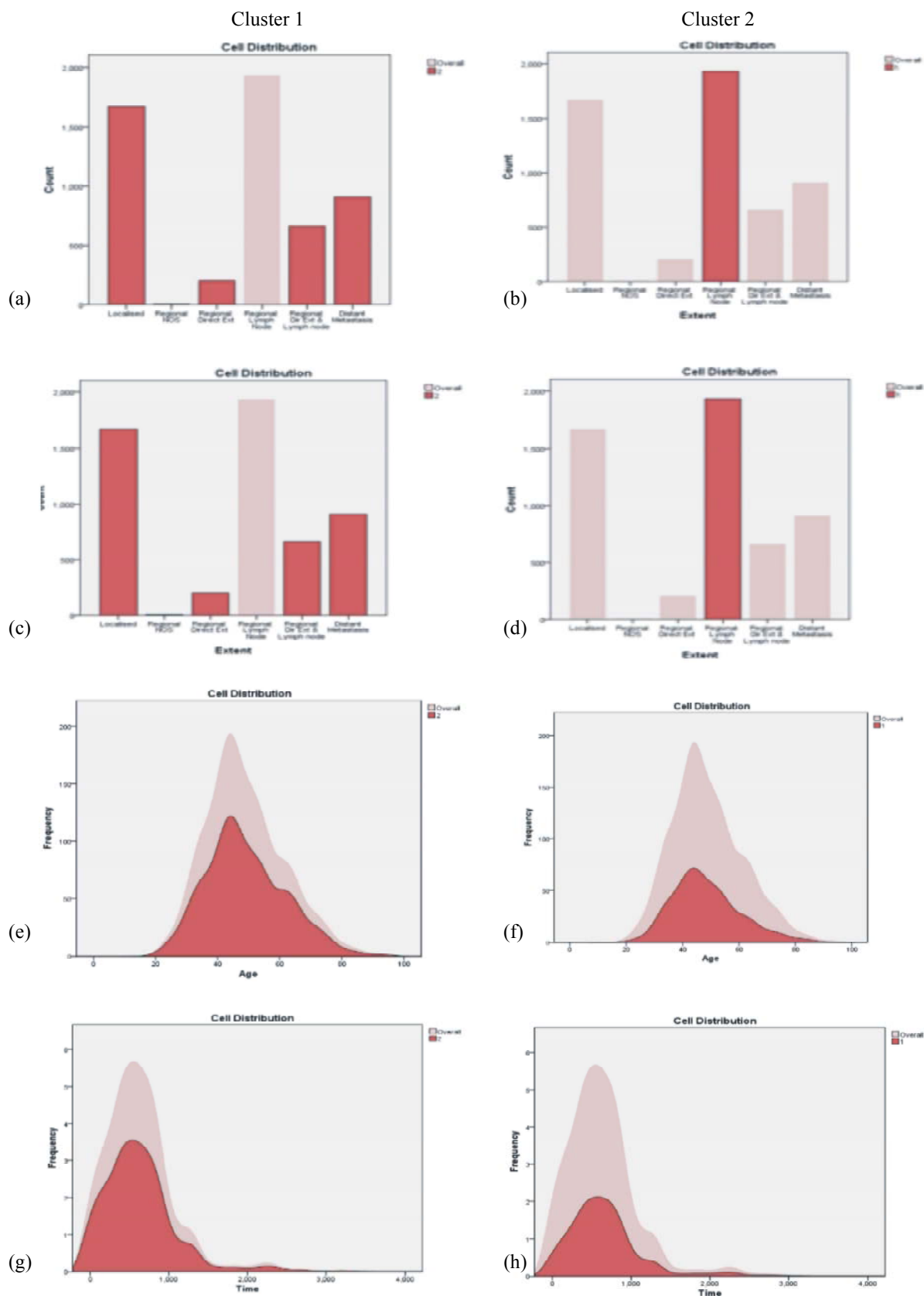
Also from Table (2) Eigenvalue is low but explains the variation of 100% and the value of Correlation is low.

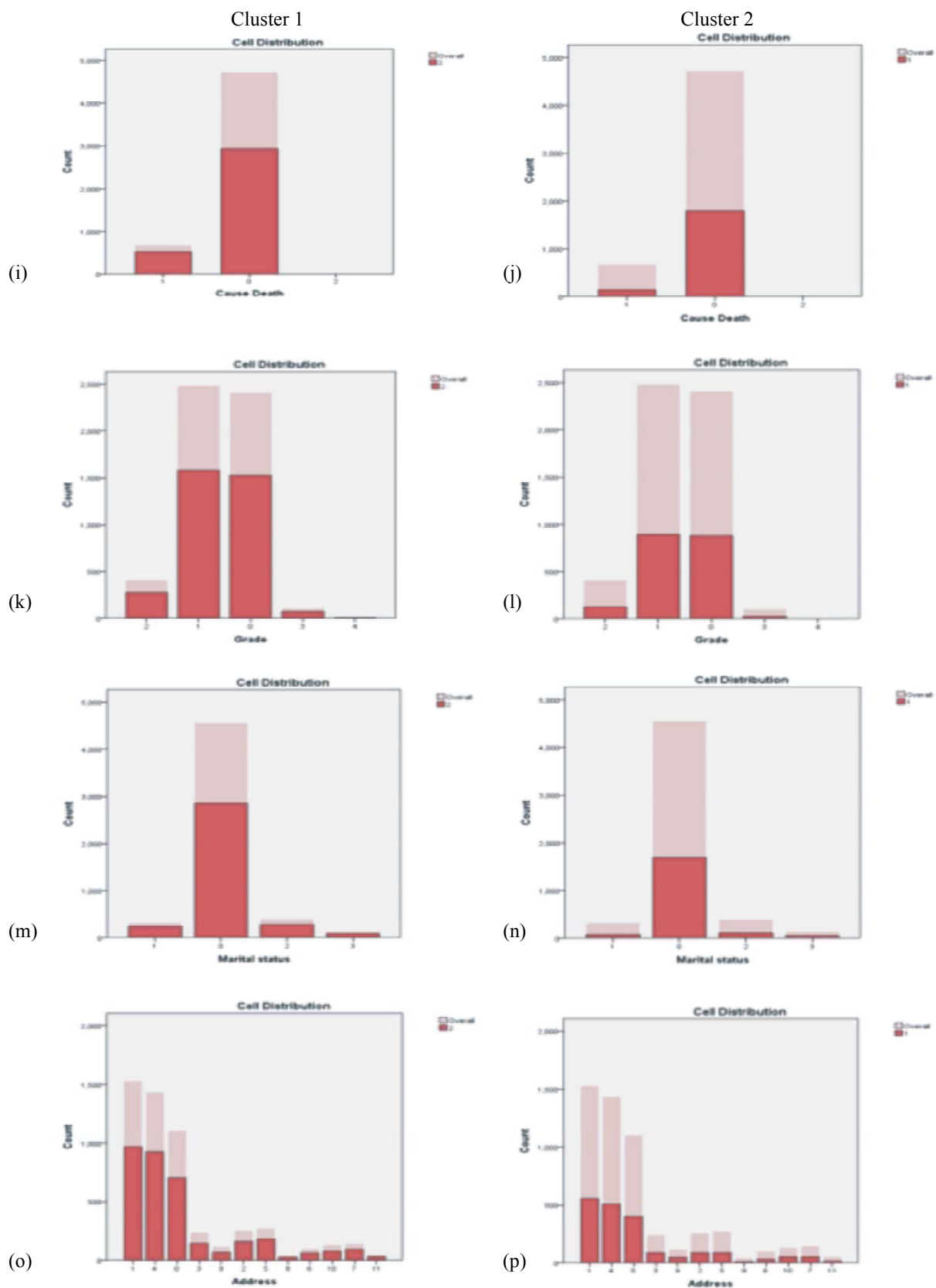
Seen from Table (3) that the value of Wilkes is high and close to the one, but it is significant and rejected the null hypothesis then the dependent variable depended of the x_i 's.

That is, the Canonical Discriminant Function Coefficients, Table (4) used to construct the actual prediction equation that can be used to classify new cases. The Discriminant function of Breast cancer patients after two-step clustering:

$$Y = \beta_0 + \beta_1 x_1$$

$$Y = -3.448 + 2.1x_1$$





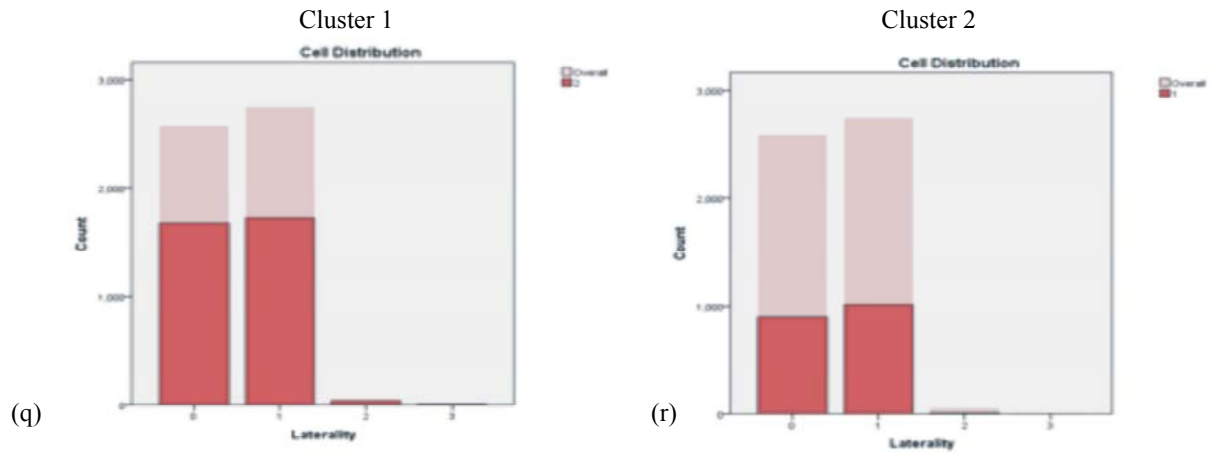


Fig. 3:

Table 1: Group Statistics

Status		Mean	Std. Deviation
Alive	Two-Step Cluster Number	1.62	.485
Dead	Two-Step Cluster Number	1.79	.409
Total	Two-Step Cluster Number	1.64	.480

Table 2: Summary of canonical discriminant functions eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	.014	100.0	100.0	.118

Table 3: Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.986	76.166	1	.000

Table 4: Canonical Discriminant Function Coefficients

	Function
	1
Two-Step Cluster Number(x_1)	2.100
(Constant)	-3.448

Table 5: Functions at Group Centroids

Status	Function
	1
Alive	-.046
Dead	.308

Table 6: Classification Results

		Status	Predicted Group Membership		
			Alive	Dead	Total
Original	Count	Alive	1784	2908	4692
		Dead	149	555	704
	%	Alive	38.0	62.0	100.0
		Dead	21.2	78.8	100.0

Table 7: Group Statistics

Status		Mean	Std. Deviation
Alive	Age	48.66	12.424
	Marital status	2.06	.469
	Address	3.29	2.843
	Grade	2.39	.663
	Extent	4.31	1.798
	Laterality	1.53	.527
	Cause Death	2.00	.000
	Time	610.99	434.871
Dead	Age	48.98	13.604
	Marital status	2.11	.528
	Address	3.11	2.891
	Grade	2.56	.619
	Extent	5.69	1.623
	Laterality	1.55	.569
	Cause Death	1.06	.281
	Time	761.79	577.869
Total	Age	48.71	12.584
	Marital status	2.06	.477
	Address	3.27	2.850
	Grade	2.41	.660
	Extent	4.49	1.836
	Laterality	1.54	.532
	Cause Death	1.88	.333
	Time	630.67	458.834

Table 8: Stepwise Statistics

Exact F				
df1	Statistic	Statistic	Entered	Sig.
1	52612.847	0.093	Cause Death	0
2	26486.403	0.092	Extent	0
3	17670.374	0.092	Age	0

Table 9: Variables in the Analysis

Step		Tolerance	F to Remove	Wilks' Lambda
1	Cause Death	1.000	52612.847	
2	Cause Death	1.000	49233.908	.936
	Extent	1.000	34.379	.093
3	Cause Death	.999	49267.639	.936
	Extent	1.000	34.157	.093
	Age	.999	4.448	.092

We note that the central function values from Table (5), through which we can describe any new item, wither it belongs to ALIVE group, or DEAD group, as its value ranges between (-.046, .308).

The percentage of correct classification of Alive breast cancer patients data from Table (6) was 38% and the percentage of correct classification of Dead breast cancer patients data was 78.8%, while the percentage of the total classification validity amounted to 43.3%.

The Second Case: Classification of breast cancer patients' data and the calculation of the discriminatory function of data.

The variables in Table (8) cause Death, Extent and Age were accepted and the rest of the variables were rejected.

We can notice from Table (9) the variables (Cause Death, Extent and Age) stay in the model.

From Table (10) the big value of the Eigen values (9.831) and the value of the Variation explanation equals 100 and its increase from the first case. The larger eigenvalue and the more of the variance in the dependent variable have been explained by that function.

Table (11) shows Wilkes Lambda Table the decline in its value to .092 close to zero and its significance, smaller values of Wilks' lambda indicate greater discriminatory ability of the function.

Table 10: Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	9.831 ^a	100.0	100.0	.953

Table 11: Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.092	12847.378	3	.000

Table 12: Canonical Discriminant Function Coefficients

variables	Function
	1
Age	-.002
Extent	-.047
Cause Death	9.828
(Constant)	-18.124

Table 13: Functions at Group Centroids

Status	Function
	1
Alive	1.214
Dead	-8.093-

Table 14: Classification Function Coefficients

	Status	
	Alive	Dead
Age	.262	.284
Extent	1.352	1.788
Cause Death	193.482	102.003
(Constant)	-203.462-	-66.786-

Table 15: Classification Results

Original	Count	Status	Predicted Group Membership		Total
			Alive	Dead	
		Alive	4692	0	4692
		Dead	34	670	704
	%	Alive	100.0	.0	100.0
		Dead	4.8	95.2	100.0

The Discriminant function of Breast cancer patients based Table (12) is:

$$Y = -18.124 + 9.828(\text{Cause Death}) - 0.047(\text{Extent}) - 0.002(\text{Age})$$

We can note from Table (13), central function values, through which we can describe any new item, whether it belongs to the ALIVE group, or DEAD group, as its value ranges between (-8.093, 1.214).

The Table (15) shows percentage of correct classification of Alive breast cancer patients data was 100% and the percentage of correct classification of

Dead breast cancer patients data was 95.2%, while the percentage of the original classification validity amounted to 99.4%. Thus, the percentage of correct classification is higher than the first case. Finally, clustering after reaching the predicted discriminatory function variables by using two-Step clustering method:

As two clusters resulted and silhouette measure was very high and close to 1 in Figure (4), thus, we note that there is improvement comparing to the first case and the first cluster ratio was 12.4%, as the second cluster ratio was 87.6%.

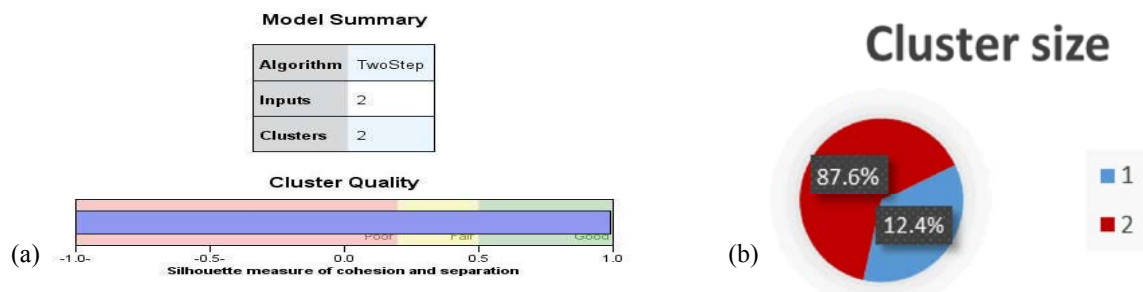


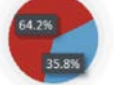
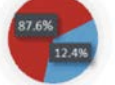
Fig. 4:



Fig. 5:

The comparison between the two cases above are in the following:

<p>The first case: clustering data of breast cancer patients using a clustering method in two-step and then classifying it and find the discriminatory function</p>	<p>The second case: Clustering two-step data of breast cancer patients after find the discriminatory function</p>																																																										
<table border="1"> <thead> <tr><th colspan="4">Classification Results</th></tr> <tr><th rowspan="2"></th><th rowspan="2">Status</th><th colspan="2">Predicted Group Membership</th><th rowspan="2">Total</th></tr> <tr><th>Alive</th><th>Dead</th></tr> </thead> <tbody> <tr><td rowspan="2">Count</td><td>Alive</td><td>1784</td><td>2908</td><td>4692</td></tr> <tr><td>Dead</td><td>149</td><td>555</td><td>704</td></tr> <tr><td rowspan="2">100%</td><td>Alive</td><td>38</td><td>62</td><td>100</td></tr> <tr><td>Dead</td><td>21.2</td><td>78.8</td><td>100</td></tr> </tbody> </table>	Classification Results					Status	Predicted Group Membership		Total	Alive	Dead	Count	Alive	1784	2908	4692	Dead	149	555	704	100%	Alive	38	62	100	Dead	21.2	78.8	100	<table border="1"> <thead> <tr><th colspan="4">Classification Results</th></tr> <tr><th rowspan="2"></th><th rowspan="2">Status</th><th colspan="2">Predicted Group Membership</th><th rowspan="2">Total</th></tr> <tr><th>Alive</th><th>Dead</th></tr> </thead> <tbody> <tr><td rowspan="2">Count</td><td>Alive</td><td>4692</td><td>0</td><td>4692</td></tr> <tr><td>Dead</td><td>34</td><td>670</td><td>704</td></tr> <tr><td rowspan="2">%</td><td>Alive</td><td>100.0</td><td>.0</td><td>100.0</td></tr> <tr><td>Dead</td><td>4.8</td><td>95.2</td><td>100.0</td></tr> </tbody> </table>	Classification Results					Status	Predicted Group Membership		Total	Alive	Dead	Count	Alive	4692	0	4692	Dead	34	670	704	%	Alive	100.0	.0	100.0	Dead	4.8	95.2	100.0
Classification Results																																																											
	Status	Predicted Group Membership		Total																																																							
		Alive	Dead																																																								
Count	Alive	1784	2908	4692																																																							
	Dead	149	555	704																																																							
100%	Alive	38	62	100																																																							
	Dead	21.2	78.8	100																																																							
Classification Results																																																											
	Status	Predicted Group Membership		Total																																																							
		Alive	Dead																																																								
Count	Alive	4692	0	4692																																																							
	Dead	34	670	704																																																							
%	Alive	100.0	.0	100.0																																																							
	Dead	4.8	95.2	100.0																																																							
<p>The percentage of the validity of the total classification by 43.3%. The specificity is 78.8%, and sensitivity is 38% .</p>	<p>The percentage of the validity of the original classification by 99.4%. Thus, the percentage of correct classification is higher than the first case. The specificity is high its 100%, and sensitivity is 95.2.</p>																																																										
<p>Model Summary</p> <table border="1"> <tr><td>Algorithm</td><td>TwoStep</td></tr> <tr><td>Inputs</td><td>2</td></tr> <tr><td>Clusters</td><td>2</td></tr> </table> <p>Cluster Quality</p> <p>Silhouette measure of cohesion and separation</p>	Algorithm	TwoStep	Inputs	2	Clusters	2	<p>Model Summary</p> <table border="1"> <tr><td>Algorithm</td><td>TwoStep</td></tr> <tr><td>Inputs</td><td>2</td></tr> <tr><td>Clusters</td><td>2</td></tr> </table> <p>Cluster Quality</p> <p>Silhouette measure of cohesion and separation</p>	Algorithm	TwoStep	Inputs	2	Clusters	2																																														
Algorithm	TwoStep																																																										
Inputs	2																																																										
Clusters	2																																																										
Algorithm	TwoStep																																																										
Inputs	2																																																										
Clusters	2																																																										
<p>As two clusters resulted and silhouette measure was in a high average close to 0.5, and the first cluster ratio was 35.8%, as the second cluster ratio was 64.2%.</p>	<p>As two clusters resulted and silhouette measure was very high and close to 1, thus, we note that there is improvement comparing to the first case, and the first cluster ratio was 12.4%,as the second cluster ratio was 87.6%.</p>																																																										

Eigenvalues					Eigenvalues				
Function	Eigenvalue	% Of Variance	Cumulative %	Canonical Correlation	Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	.014 ^a	100	100	0.118	1	9.831 _a	100	100	0.953
Wilks' Lambda					Wilks' Lambda				
Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.	Test of Function(s)	Wilks' Lambda	Chi-square	df	
1	0.986	76.166	1	0	1	0.092	12847.38	3	
We note that the value of Wilkes Lambda is too big, close to one					We note that the value of Wilkes Lambda is small, close to zero				
<p style="text-align: center;">Cluster size</p> 					<p style="text-align: center;">Cluster size</p> 				
Size of Smallest Cluster		1933	35.80%		Size of Smallest Cluster		669	12.40%	
Size of Largest Cluster		3453	64.20%		Size of Largest Cluster		4727	87.60%	
Ratio of sizes: Largest Cluster to Smallest Cluster		1.79			Ratio of sizes: Largest Cluster to Smallest Cluster		7.065		
We conclusion is the second case is best.									

CONCLUSION

- Clustering the data of the variables (Time, Age, Grade, Extent, Marital Status, Address, Laterality and Case of Death) for women with breast cancer are using the two-stage cluster method. In addition, we have two clusters, the efficiency of clusters and the second case better than the first where he reached the value of silhouette measure close to one. This indicates an improvement in the cluster process after finding the discriminatory function.
- The value of Wilkes Lambda are improved where the value of 0.092, which is better than the first case where the smaller this value, which indicates the efficiency of the discriminatory function and its ability to distinguish.
- The significant improvement in the value of the canonical correlation, where the value in the second case 0.953 as it reveals the strength of the relationship between the adopted variable and the discriminatory function, which is better than the first case, where the value of 0.118.
- The significant of the test and the ability of the discriminatory function to distinguish and indicates that the dependent variable depends on one of these independent variables.

- Finally, our study proved the power of the distinctive function and its usefulness in the classification of most of the views in the original community where the strength was 99.4% and this means that it was able to categorize 99.4% of the views to the original communities with a total error rate(TER) of 4.8% .

REFERENCES

1. Ramayah1*, T. and Noor Hazlina Ahmad1, 2010. Discriminant analysis. Available online at <http://www.academicjournals.org/AJBM>.
2. Anandhalli, 2018. Discriminant Analysis of the influence of Faculty Members’ Socio-Economic Characteristics on the use of Social networks: a study of KSWU, Bijapur, International Journal of Creative Research Thoughts (IJCRT), 2018.
3. Eloisa Urrechaga, Urko Aguirre and Silvia Izquierdo, 2013. Multivariable Discriminant Analysis for the Differential Diagnosis of Microcytic Anemia. Article ID 457834, pp: 6.
4. Abbas, O.A., 2008. Comparisons between data clustering algorithms. The International Arab Journal of Information Technology, 5: 320-5.
5. Kakkar, P. and A. Parashar, 2014. Comparison of different clustering algorithms using WEKA tool. International Journal of Advanced Research in Technology, Engineering and Science, 1: 20-2.

6. Rana Walid, 2010. Using Clustering for Modeling Monthly Salary Grade. *Iraqi Journal*, 18: 297-320.
7. Dr. Samir Darkazanli and Sundus Alaziz, 2009. An applied study of the factors affecting age groups of the labor force using two-stage cluster analysis. (*University of Aleppo Research Journal* 2009).
8. Ming-Yi Shih, Jar-Wen Jheng and Lien-Fu Lai, 2010. A Two-Step Method for Clustering Mixed Categorical and Numeric Data, *Tamkang Journal of Science and Engineering*, 13(1): 11-19.
9. Paul McNicholas, 2012. Model-based clustering, classification and discriminant analysis of data with mixed type. *Journal of Statistical Planning and Inference*, 142(11): 2976-2984.
10. Alayande S. Ayinla and Bashiru Kehinde Adekunle 2015. An Overview and Application of Discriminant Analysis in Data Analysis, e-ISSN: 2278-5728, p-ISSN: 2319-765X.
11. Garson, G.D., 2008. Discriminant function analysis. <http://www.statsoft.com/textbook/stdiscan.html>.
12. Hardle, W. and L. Simar, 200). *Applied Multivariate Statistical Analysis*. Springer Berlin Heidelberg., pp: 289-303.
13. Dr. Samir Darkazanli and Sundus Alaziz, 2010. Statistical analysis of the data stencil after using the analysis of the basic compounds and the original data node (a comparative study on some diabetics with algorithm applications. (*University of Aleppo Research Journal* 2010).
14. Kevin McGarigal Sam Cushman Susan Stafford, 2000. *Multivariate Statistics for Wildlife and Ecology Research*.
15. Özge Pasin and Handan Ankaralı, 2017. Comparison of EM and Two-Step Cluster Method for Mixed Data. *International Journal of Medical Science and Clinical Inventions*, 4(3): 2768-2773.