Cluster analysis and Discriminant analysis on Sina -Weibo Influencers

Our aim is to define different roles of social media influencers quantitatively by the help of multivariate analysis techniques.

Statistical methods such as cluster analysis and discriminant analysis are used to classify the Sina-Weibo data of influencer according to their popularity and activity.
The results are as follows: 1) by hierarchical cluster method and k-means clustering method, the roles can be divided into four categories: "superstar", "influencer", "artist" and "actors"; 2) the correct rate of Fisher's linear discriminant function was 94.5%.

Background

With the development of social networks, the entertainment industry, as the focus of public attention, also relies on this kind of new media to gradually become transparent, interactive and popular. In the era of "trending search" and "nationwide Weibo", not only actors and artists, entertainment bloggers, but also the official agents of studios have attracted wide attention in Sina-Weibo. Whether it is a nobody or a superstar, everyone is likely to become a trending center.

Therefore, the evaluation of a star's Sina-Weibo influencing ability depends not only on its popularity in the corresponding fields, but also on the how attractive their blogs are. Thus, analysis of a star's influencing ability helps us better position the star's "money absorbing ability" in the era of "fan economy".

Introduction to sample data

Sina-Weibo is the website with the most complete collection of stars and the widest user group. We select 1747 pieces of data, including information about actors, artists, athletes, influencer, etc.

Dependent variable:
Microblog account: text format, the name of the star microblog account.

Features:

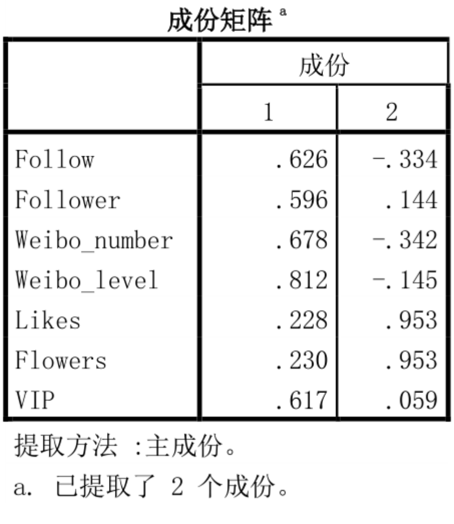
- Following: continuous variable (unit: number), value range [04833], is the number of other people followed by a star. It can be used to reflect the microblog characteristics of the star.
- Number of Blog posted: continuous variable (unit: number), value range [2,193299], as of the date of data collection. The number of tweets that are still kept in time can be used to reflect the microblog activity of the star.
- Account rank: continuous variable (unit: rank LV), value range [2,46], according to the number of active days of users. It is confirmed that users with higher level can enjoy more privileges, which can be used to reflect the activity of the star's microblog.

- Followers: continuous variable (unit: person), value range [2239,90928261], which can be used to reflect stars how popular it is.
- Membership level: continuous variable, value range [1,6], member level of star registered microblog, need to recharge actively, can enjoy more privileges, to a certain extent, can reflect the influence of the star.
- Adoration value: continuous variable (unit: score), value range [22,5626047]. Fans need to sign in, or send flowers and gifts to the star to increase the adoration value. The value can reflect the popularity and influence of the star.
- Number of flowers received: continuous variable (unit: bundle), value range [11,2763629. Flowers and gifts need to be purchased, so this value can reflect the popularity and influence of the star.

Data manipulation

Cleaning
1) Delete duplicate data: mark observations with the same posting name and delete duplicate information;
2) Dealing with missing values: We subjectively classify the data with missing values. For the stars that have many followers, the incomplete data is filled in manually. For the users with little influence, we delete their missing data.

Feature engineering
Firstly, we use **principal component analysis (PCA)** to do factor analysis on seven variables. The results are as follows:

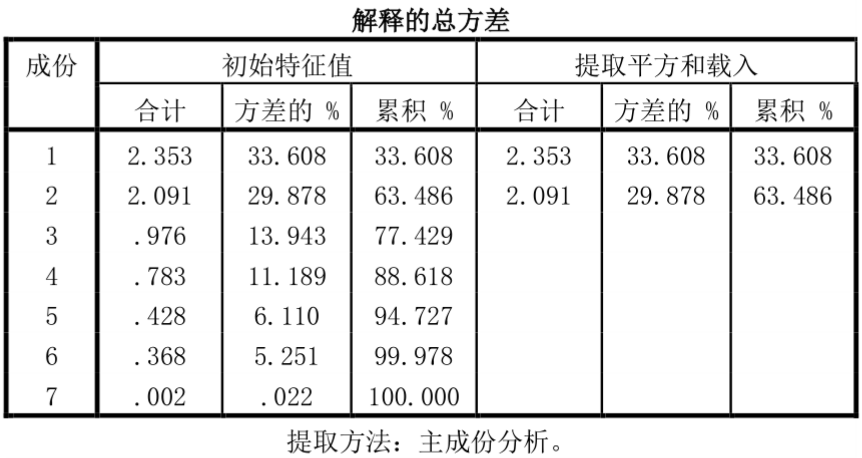


成份矩阵[a]

	成份	
	1	2
Follow	.626	-.334
Follower	.596	.144
Weibo_number	.678	-.342
Weibo_level	.812	-.145
Likes	.228	.953
Flowers	.230	.953
VIP	.617	.059

提取方法 :主成份。

a. 已提取了 2 个成份。



解释的总方差

成份	初始特征值			提取平方和载入		
	合计	方差的 %	累积 %	合计	方差的 %	累积 %
1	2.353	33.608	33.608	2.353	33.608	33.608
2	2.091	29.878	63.486	2.091	29.878	63.486
3	.976	13.943	77.429			
4	.783	11.189	88.618			
5	.428	6.110	94.727			
6	.368	5.251	99.978			
7	.002	.022	100.000			

提取方法：主成份分析。

The component matrix suggests that we extract two principal components, but their cumulative total variance of interpretation is only 63.486%. This can not reflect the important information of the original data, that is, the ability of each principal component to concentrate the original variable information is not significant different, so we do not choose the method of PCA for dimension reduction.

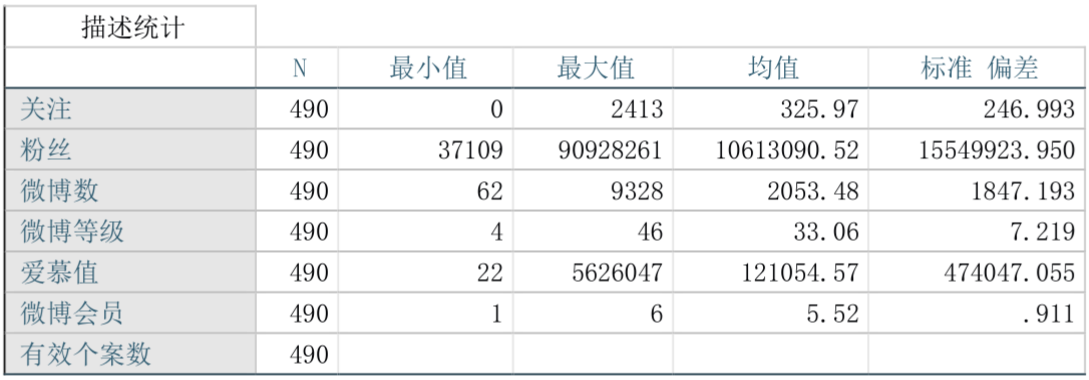
In addition, in order to verify the above conclusions and eliminate overlapping information, we conducted correlation analysis on these variables. The result shows that the correlation of most variables are less than 0.3, but there is a strong correlation between the "adoration value" and the "number of flowers received", with the Pearson correlation coefficient is as high as 0.998.

Since fans sending flowers to the stars is one of the ways to increase the star's "adoration value"), "number of flowers received" and the "adoration value" reflect the same information. Therefor, we delete the variable of "the number of flowers received".

Cluster analysis

Hierarchical cluster method

In order to determine the details of within-group distance measurement and among-group distance measurement, we need to know more about the the sample data.

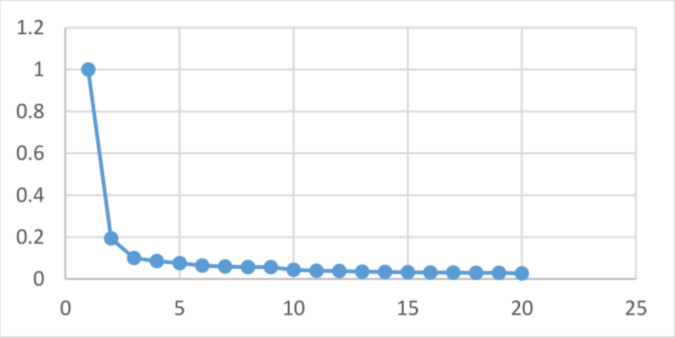


描述统计

	N	最小值	最大值	均值	标准 偏差
关注	490	0	2413	325.97	246.993
粉丝	490	37109	90928261	10613090.52	15549923.950
微博数	490	62	9328	2053.48	1847.193
微博等级	490	4	46	33.06	7.219
爱慕值	490	22	5626047	121054.57	474047.055
微博会员	490	1	6	5.52	.911
有效个案数	490				

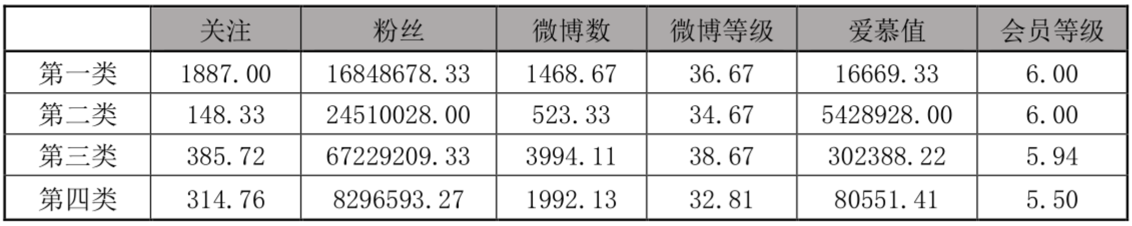
Descriptive statistics show that each variable has great difference in value. In order to avoid the impact of magnitude difference on clustering results, we first use standardize variables, and model the cleaned data using inter-group-link clustering, with the square Euclidean distance as the measurement.

Now we see how the aggregation coefficient changing with the number of classifications.



When the number of categories is less than or equal to 3, the aggregation coefficient decreases rapidly with the increase of the number of categories; when the number of categories is greater than or equal to 4, the aggregation coefficient tends to zero gently. Therefore, the number classification should be 3 or 4.

Four groups of mean values were extracted for comparison:



	关注	粉丝	微博数	微博等级	爱慕值	会员等级
第一类	1887. 00	16848678. 33	1468. 67	36. 67	16669. 33	6. 00
第二类	148. 33	24510028. 00	523. 33	34. 67	5428928. 00	6. 00
第三类	385. 72	67229209. 33	3994. 11	38. 67	302388. 22	5. 94
第四类	314. 76	8296593. 27	1992. 13	32. 81	80551. 41	5. 50

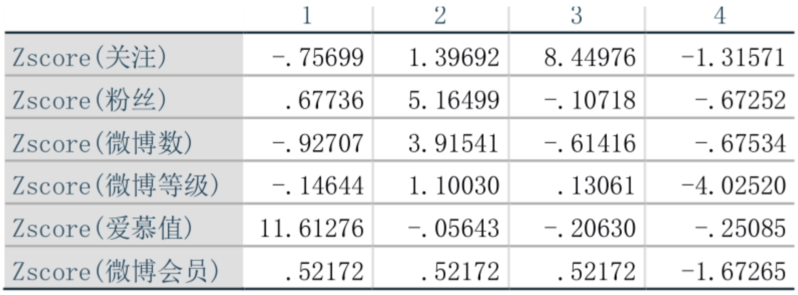
There is no significant difference between the four groups in the level of microblog and membership, and there are significant characteristics in other indicators.

k-means clustering

Define k

In the k-means clustering method, the determination of k value of the number of clusters depends on practical experience and the reference to the clustering results of the hierarchical clustering method. Thus we pick k = 4.

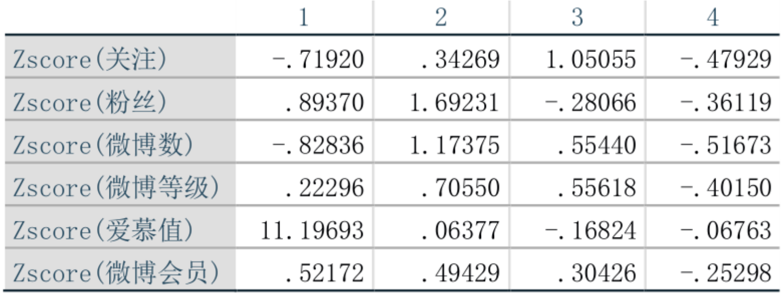
k-means++: The Advantages of Careful Seeding



	1	2	3	4
Zscore(关注)	-. 75699	1. 39692	8. 44976	-1. 31571
Zscore(粉丝)	. 67736	5. 16499	-. 10718	-. 67252
Zscore(微博数)	-. 92707	3. 91541	-. 61416	-. 67534
Zscore(微博等级)	-. 14644	1. 10030	. 13061	-4. 02520
Zscore(爱慕值)	11. 61276	-. 05643	-. 20630	-. 25085
Zscore(微博会员)	. 52172	. 52172	. 52172	-1. 67265

From the table above, it can be seen that the distance between the four centroids (seed points) is very large, and thus it is almost impossible to be in the same class, so the clustering results are relatively reliable.

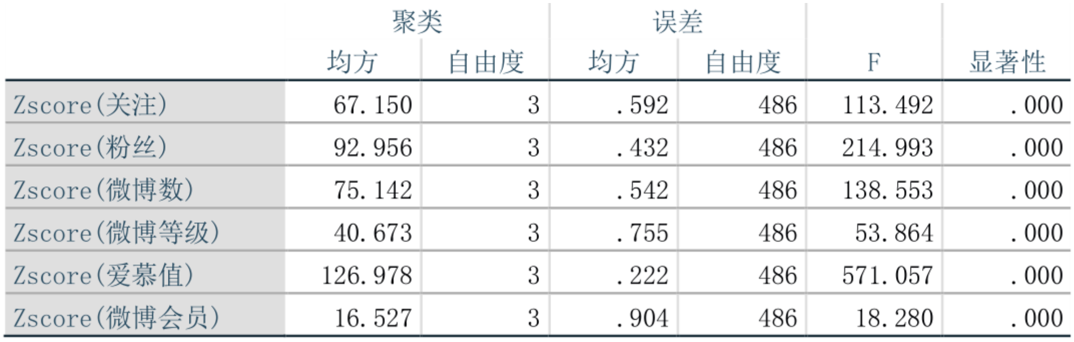
Final centroids



	1	2	3	4
Zscore(关注)	-.71920	.34269	1.05055	-.47929
Zscore(粉丝)	.89370	1.69231	-.28066	-.36119
Zscore(微博数)	-.82836	1.17375	.55440	-.51673
Zscore(微博等级)	.22296	.70550	.55618	-.40150
Zscore(爱慕值)	11.19693	.06377	-.16824	-.06763
Zscore(微博会员)	.52172	.49429	.30426	-.25298

From the above "final cluster center" table, we can see that the four types of microblog user samples have their own characteristics in activity and popularity, which are consistent with our intuitive expectation.

Analysis of Variance



	聚类		误差			
	均方	自由度	均方	自由度	F	显著性
Zscore(关注)	67.150	3	.592	486	113.492	.000
Zscore(粉丝)	92.956	3	.432	486	214.993	.000
Zscore(微博数)	75.142	3	.542	486	138.553	.000
Zscore(微博等级)	40.673	3	.755	486	53.864	.000
Zscore(爱慕值)	126.978	3	.222	486	571.057	.000
Zscore(微博会员)	16.527	3	.904	486	18.280	.000

From the ANOVA table that the six variables selected have significant contributions to the classification, which indicates that previous variable selection process is scientific and the results are reliable.

In particular, compared with the hierarchical clustering results, the results obtained by k-means are quite different in the specific number of each class, but the characteristics and classification ideas of each class are almost the same. The conclusions of the two are similar and complementary to each other. The reliability of the two clustering results is supported by each other.

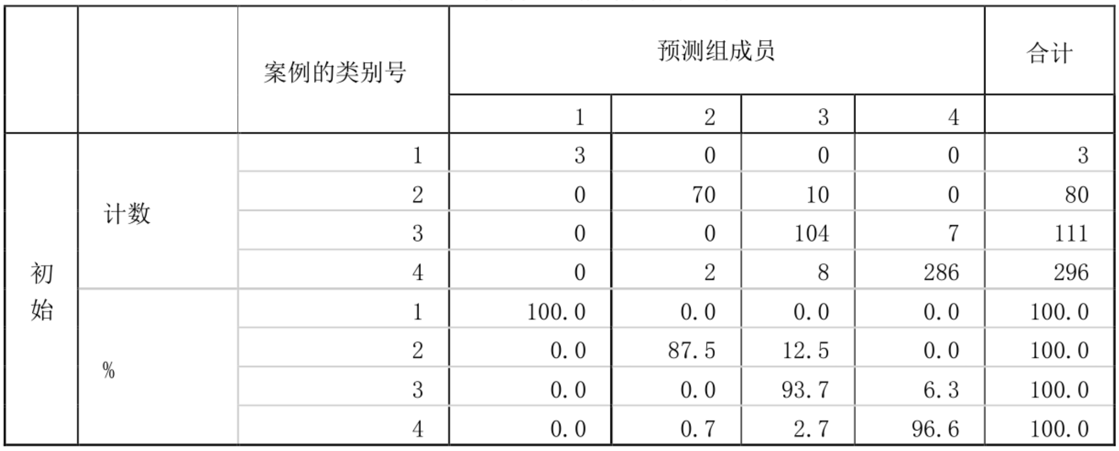
Discriminant analysis

After clustering Sina-Weibo influencer, we collect several trending influencer data, and try to identify the categories of these stars by discriminant analysis. (In order to eliminate the influence of dimension, we have standardized all the data here.)

Decide prior probability

In order to verify the feasibility of discriminant analysis and determine the prior probability, a discriminant analysis is carried out with the original data.

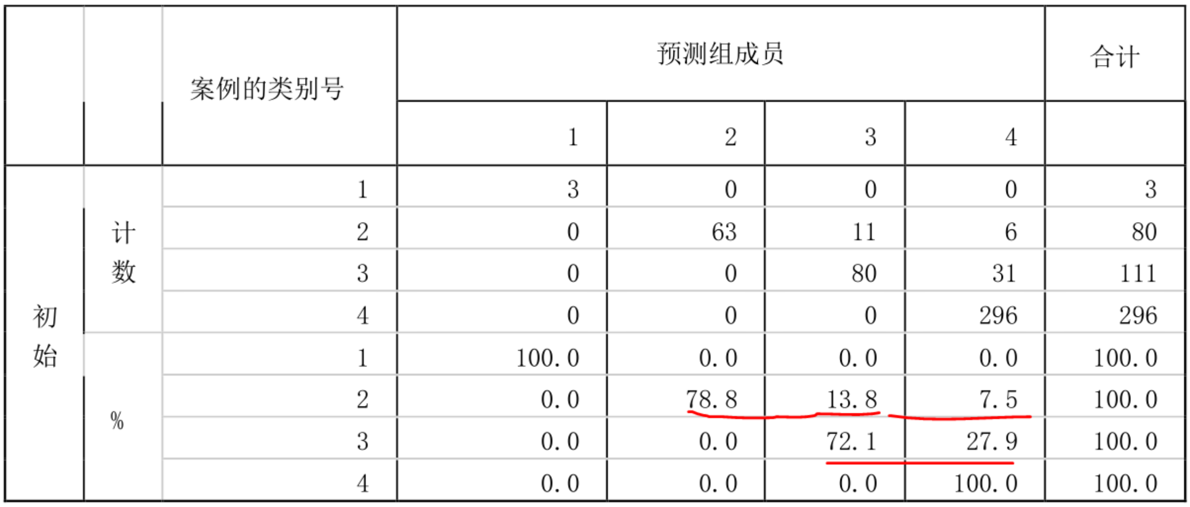
- **Using equal prior probability**



		案例的类别号	预测组成员				合计
			1	2	3	4	
初始	计数	1	3	0	0	0	3
		2	0	70	10	0	80
		3	0	0	104	7	111
		4	0	2	8	286	296
	%	1	100.0	0.0	0.0	0.0	100.0
		2	0.0	87.5	12.5	0.0	100.0
		3	0.0	0.0	93.7	6.3	100.0
		4	0.0	0.7	2.7	96.6	100.0

The results show that 94.5% of the grouped cases are classified correctly, so we think that the method is feasible from the perspective of accuracy.

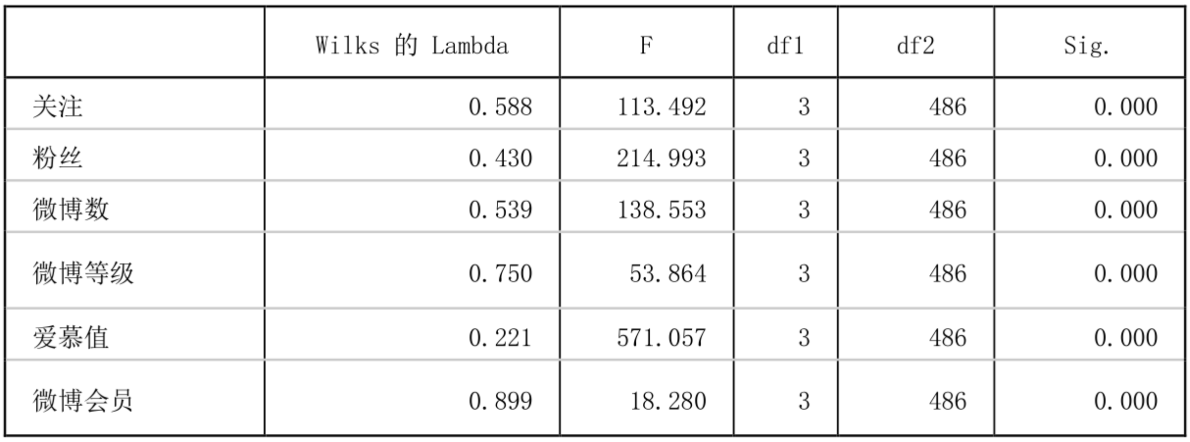
- **Using weighted prior probability**



		案例的类别号	预测组成员				合计
			1	2	3	4	
初始	计数	1	3	0	0	0	3
		2	0	63	11	6	80
		3	0	0	80	31	111
		4	0	0	0	296	296
	%	1	100.0	0.0	0.0	0.0	100.0
		2	0.0	78.8	13.8	7.5	100.0
		3	0.0	0.0	72.1	27.9	100.0
		4	0.0	0.0	0.0	100.0	100.0

Only 90.2% of the grouped cases are correctly classified by this method, so we choose equal priori probability for discrimination.

Test for the equality of within-group means



	Wilks 的 Lambda	F	df1	df2	Sig.
关注	0.588	113.492	3	486	0.000
粉丝	0.430	214.993	3	486	0.000
微博数	0.539	138.553	3	486	0.000
微博等级	0.750	53.864	3	486	0.000
爱慕值	0.221	571.057	3	486	0.000
微博会员	0.899	18.280	3	486	0.000

Since the centroids of all features in each group are significantly different, the discrimination analysis is meaningful.

Fisher Canonical discriminant functions

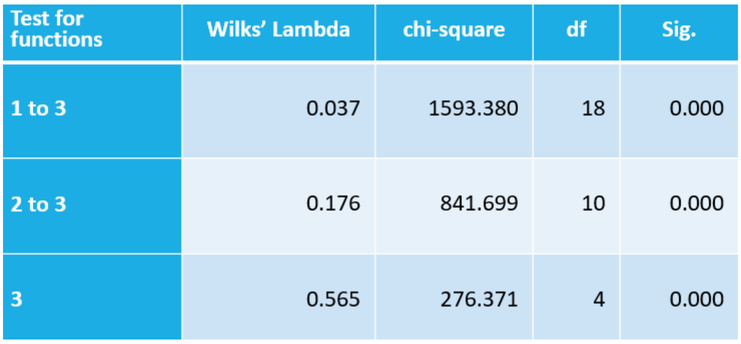
$$y_1=0.014x_1-0.32x_2+0.091x_3+0.062x_4+3.922x_5-0.138x_6+0.182$$

$$y_2=0.282x_1+1.06x_2+0.879x_3-0.027x_4+0.121x_5+0.244x_6-0.002$$

$$y_3=1.047x_1-0.972x_2+0.244x_3+0.043x_4+0.288x_5+0.242x_6+0.005$$

where x_1 is following, x_2 is fans, x_3 is number of posts, x_4 is account rank, x_5 is adoration value, and x_6 is membership level.

Description of discriminant functions



Test for functions	Wilks' Lambda	chi-square	df	Sig.
1 to 3	0.037	1593.380	18	0.000
2 to 3	0.176	841.699	10	0.000
3	0.565	276.371	4	0.000