A Suggested Measure for the Optimum Number of Clusters

Nahed Helmy

Faculty of Commerce

Al – Azhar University, Girls' Branch

Abstract

In this paper a suggested measure for the optimum number of clusters is introduced. There are many existing measures which can be used to determine the optimum number of clusters. For example, c – index measure and the average within cluster distances. A comparison of the previous measures with the suggested one is made. Single – link method using the Euclidean metric distance measure is applied to classify the governorates of Egypt into a number of clusters. The suggested measure to determine the number of clusters is the rule of determining the number of intervals in the case of the quantitative data which is referred to Herbert Sturges (1926). The homogeneity of the resulting clusters is investigated. An application using published data of the most recent population census conduced in A.R.E. in (2006) is introduced.

Key words: *Single – link method, Euclidean metric distance measure, Hartley's test*

1. Introduction

Gorden and Henderson (1977) pointed out that all the stages in both the agglomerative and the divisive methods the number of clusters which is obtained may be regarded as an approximation of the optimum number of clusters.

Glaz and Naus (1983) suggested that an expected variance and approximate distribution of the number of clusters of a given size N (where N represents a sequence

of events which are contained within an interval of length t) can be obtained.

Naoyuki (1997) proposed a method which is used to estimate the number of clusters. When a model that describes the distribution of patterns is defined, the maximum – likelihood estimation can be applied to the parameter estimation and the number of parameters can be optimized by the Akaike information criterion (AIC) or the minimum description length (MDL). Then the number of clusters can be estimated.

Robert, Guenther and Trevor (2001) proposed a method "the gap method" for estimating the number of clusters in a set of data. This method uses the output of any clustering algorithm (e.g., K – means or hierarchical) and comparing the change in within – cluster dispersion with the expected under an appropriate reference null distribution.

Sandrine and Jane (2002) developed a new prediction method to estimate the number of clusters in a data set. They compared the performance of the new and existing methods using simulated data and gene – expression data from four recently published cancer microarray studies. The new method was generally found to be more accurate and robust than the six existing methods considered in their study.

McLachlan and Khan (2004) considered the problem for assessing the number of clusters in a limited number of samples. They proposed to use a normal model – based on approach to the clustering of the samples. It can be used as a test on the smallest number of components in the mixture model compatible with the data.

Seong, Janice and William (2006) suggested a method to predict the number of clusters by applying various agglomerative clustering algorithms. The methods using different indexes are examined and compared based on the concept of agreement (or disagreement) between clusters generated by different clustering algorithms on the set of data.

Mingjin Yan and Keying Ye (2007) proposed the weighted gap and the difference of difference – weighted (DD – weighted) gap methods for estimating the number of clusters in data using the weighted within – clusters sum of errors: a measure of the within – clusters homogeneity.

2. The Main Existing Measures for Determining the Number of Clusters

There are many existing measures that can be used to determine the optimum number of clusters. For example, c – index measure and average within cluster distances. The first measure is called the c – index which is suggested by Bolshakova and Azuaje (2006). This measure is defined as:

$$C = \frac{S - S\min}{S\max - S\min} \qquad (2.1)$$

where

S : is the sum of the distances between all pairs of objects in the clusters.

$S\min$: is the smallest distance.

$S\max$: is the largest distance.

The number of clusters that minimizes c – index is taken as the optimum number of clusters, K.

The second measure is suggested by Thorndike (1953) who plotted average within – cluster distance against number of clusters. The number of clusters that minimizes the value of this measure is an indicator to the optimum number of clusters,K.

3. The Suggested Measure

It is suggested to use the equation which is used in determining the number of intervals in the case of the quantitative data to be a measure for determining the number of clusters. This equation is referred to Herbert Sturges (1926). This measure is given by

$$I = 1 + 3.322 \log(n) \qquad (3.1)$$

where

n: is the number of objects under study

It is noticed that this measure can be used for determining the number of intervals in the case of the quantitative data. Some existing measures which are used to determine the optimum number of clusters need many calculations. Two of these measures are selected to be compared with the suggested one. These measures are the c – index measure and the average within distance measure.

4. The Evaluation Criteria

The resulting clusters are evaluated with respect to homogeneity.

The following methods are used in testing homogeneity in the present study:

4.1 Comparing the variation within the resulting clusters, σ_w^2 with the variation between them, σ_b^2 as described in the following steps:

1. The variation within the resulting clusters, σ_w^2 and the variation between the clusters, σ_b^2 are calculated as follows

$$\sigma_w^2 = \frac{1}{n}\sum_{i=1}^{K} n_i \sigma_i^2 \qquad (4.1)$$

where

n: is the number of objects under study

n_i:n_h: is the number of objects in cluster i,

K: is the number of clusters and

σ_i^2: is the variance of cluster i

$$\sigma_b^2 = \frac{1}{n}\sum_{i=1}^{K} n_i (\bar{X}_i - \bar{X})^2 \qquad (4.2)$$

where

$\bar{X}_i$: is the mean of cluster i and

$\bar{X}$: is the grand mean for all objects

2. The variation within the clusters, σ_w^2 is compared with the variation between the clusters, σ_b^2 . If σ_w^2 is less than σ_b^2 so that the homogeneity within the resulting clusters is increased.

4.2 The Hartley's test

The null and the alternative hypothesis for the Hartley's test are given by

$H_o: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$ *against* H_a: *population variances are not equal*

Hartley's test is developed in the following steps:

1. The variance of the resulting clusters with applying the clustering method are calculated: $S_1^2, S_2^2, \ldots, S_K^2$

The calculated F_{max} is obtained using a quantity given by the equation

$$F_{max} = \frac{S^2_{max}}{S^2_{min}} \qquad (4.3)$$

where

S^2_{max}: is the maximum cluster variance and

S^2_{min}: is the minimum variance of the clusters

2. The value derived in step (1) is compared with $F_{K,n-1}$ using the table of Hartley's test at the level of significance, α where k is the number of clusters and n is the cluster size which has the maximum variance (e.g., is the largest cluster size).
3. If F_{max} is greater than $F_{K,n-1}$; the null hypothesis of equal variances is rejected which means that the homogeneity within the resulting clusters is increased.

When the cluster sizes are equal, n is the size of each cluster but when the cluster sizes are not equal n is the smallest or the largest cluster size.

In the present paper since the cluster sizes are not equal, n is equal to n_{max} (where n_{max} is the largest cluster size which has the maximum variance) (R.Lyman Ott and Michael Longecker).

4.3 The minimization of the trace of the within – cluster scatter or dispersion matrix is described as follows

1. The within – cluster scatter or dispersion matrix is computed as follows

$$W = \sum_{i=1}^{k} W_i = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)(X_{ij} - \bar{X}_i)' \qquad (4.4)$$

where

n_i: is the number of objects in cluster i

X_{ij}: is the vector of object j for cluster i

$\bar{X}_i$: is the mean vector for cluster i and

K: is the number of clusters

2. The trace of this matrix is calculated for the different number of clusters.
3. The stage which has the minimum value of the trace is selected so that the homogeneity of the resulting clusters is increased

5. Application

The application in the present paper using the census data is conducted with assigning the governorates of Egypt to a number of homogeneous non – overlapping clusters according to the degree of urbanization. It is suggested some indicators which affect urbanization. These indicators are degree of industrialization (it is calculated as: the number of objects working in mining and quarrying, manufacturing and electricity divided by the total number of objects working in the different sectors of economic activities), family size (it is calculated as: the total number of persons divided by the total number of households) and educational level (it is calculated as follows: educational status may be divided into the following levels: read and write, qualification less than university degree and university degree and above, these levels of education are given arbitrary weights as: (1) for read and write, (3) for qualification less than university degree and (4) for university degree and above and the educational level is calculated by dividing the number of objects in the three states by the total population). The degree of industrialization is positively related to the urbanization, the family size is negatively related to the urbanization and the

educational level is positively related to the urbanization. It is desired to classify the governorates of Egypt into four different numbers of clusters. The single – link method is used to classify these governorates into different clusters according to the Euclidean metric distance measure.

Three stages are selected. The first stage which has the quarter number of clusters, the second stage which has the half number of clusters and the third one which has the third quarter number of clusters.

By using the suggested measure it is found that the optimum number of clusters is six ones.

The Euclidean metric distance has the following equation

$$d_{ij} = \sqrt{\sum_{L=1}^{h} (X_{iL} - X_{jL})^2}$$

where

$X_{iL} and\ X_{jL}$ are the values of the $L-th$ variable for the $i-th\ and\ j-th$ objects, respectively and h is the number of variables, where the objects are the governorates of Egypt in the census of 2006 under the variables: the educational level, the family size and the degree of industrialization.

6. Results

Table (1) represents the values of both the c – index and the average within distances measures

Table (1)

measure / No. of clusters	C – index	Average within distance
6	50.269	0.513
7	160.394	8.396×10^{12}
14	0.136	5.831×10^{12}
21	262.661	5.977×10^{12}

The C – index measure has the minimum value when fourteen clusters are obtained. So that it is considered the optimum number of clusters in this case. While the average within cluster distance measure has the minimum value when six clusters are obtained. So that it is considered the optimum number of clusters in this case.

Table (2) represents the homogeneity using the Hartley's test, comparing the variation within the resulting clusters with the variation between them and minimization of the trace of the within cluster scatter or dispersion matrix

Table (2)

Test / No. of clusters	Hartley's test	σ_w^2	σ_b^2	Min. of trace of W
6	$F_{max} = 3.848$, $F_{K,n-1} = 3.464$	0.204	0.047	5.457
7	$F_{max} = 3.439$, $F_{K,n-1} = 2.54$	0.167	0.046	4.493
14	$F_{max} = 6$, $F_{K,n-1} = 8.38$	0.017	0.063	0.417
21	$F_{max} = 2.75$, $F_{K,n-1} = 15.5$	0.003	0.009	0.06

The homogeneity within the resulting clusters is tested using the following techniques:

a) **Hartley's test:**

As it is mentioned before, when the null hypothesis of this test is rejected, it means that the homogeneity within the resulting clusters increased.

b) **Comparing the variation within the resulting clusters,**

σ_w^2 with the variation between the clusters, σ_b^2:

If σ_w^2 is less than σ_b^2 then the homogeneity within the resulting clusters increased.

c) **Minimization of the trace of the within – cluster scatter or dispersion matrix, min** $\min tr(W)$

If the trace of the within scatter or dispersion matrix is minimized, then the homogeneity within the resulting clusters increased.

By using the suggested measure it is found that the optimum number of clusters is six:

1. The value of F_{max} is greater than the value of $F_{K,n-1}$, then the homogeneity within the resulting clusters is increased using the Hartley's test.
2. The variation within the resulting clusters is greater than the variation between them, so that the homogeneity within the resulting clusters is decreased.

At the stage which has the quarter number of clusters (i.e., when seven clusters are obtained):

1. The value of F_{max} is greater than the value of $F_{K,n-1}$, then the homogeneity within the resulting clusters is increased using the Hartley's test.
2. The variation within the resulting clusters is greater than the variation between them, so that the homogeneity within the resulting clusters is decreased.

At the stage which has the half number of clusters (i.e., when fourteen clusters are obtained):

1. The value of F_{max} is less than the value of $F_{K,n-1}$, then the homogeneity within the resulting clusters is decreased using the Hartley's test.

2. The variation within the resulting clusters is less than the variation between them, so that the homogeneity within the resulting clusters is increased.

At the stage which has the third quarter number of clusters (i.e., when twenty one clusters are obtained):

1. The value of F_{max} is less than the value of $F_{K,n-1}$, then the homogeneity within the resulting clusters is decreased using the Hartley's test.
2. The variation within the resulting clusters is less than the variation between them, so that the homogeneity within the resulting clusters is increased.
3. The trace of the within cluster scatter or dispersion matrix is minimized, so that the homogeneity within the resulting clusters is increased.

It is concluded that the optimum number of clusters is satisfied by using both the suggested measure and the average within distance measure. In addition the suggested measure is easier in its calculations than the other measures and is more homogeneous.

References

Glaz, J. and Naus, J. (1983)," Multiple Clusters on the Line." Communication in Statistics, Theory and Methods, Vol.12, No. 17, pp. 1961 – 1986.

Gorden, A.D. and Henderson, J.T. (1977), "An Algorithm for Euclidean Sum of squares Classification." Biometrics, Vol. 33, pp. 355 – 362.

Herbert Sturges (1926), "The choice of a class interval." Journal of American Statistician Association, Vol. 21, pp. 65 – 66.

Lei Xu (1997), Bayesian Ying–Yang machine, "Clustering and Number of Clusters", Pattern Recognition Letters, Vol. 18, p. 1167–1178.

McLachlan,G.J. and Khan,N. (2004),"On a Resampling Approach for Tests on the Number of Clusters with Mixture Model – Based Clustering of Tissue Samples", Journal of Multivariate Analysis, Vol. 90, pp. 90 – 105.

Mingjin Yan and Keying Ye. (2007), "Determining the Number of Clusters Using the Weighted Gap Statistic" Biometrics, Vol. 63, pp. 1031 – 1037.

Naoyuki,I. (1997), "Robust Clustering Based on a Maximum Likelihood Method for Estimating a Suitable number of Clusters." Systems and Computers in Japan, Vol.28, No.1, pp. 10 – 23.

R.Lyman Ott and Michael Longnecker (2008), An Introduction to Statistics Methods and Data Analysis, Copyright 2010 Cengage Learning, Inc.

Robert Tibshirani, Guenther Walther, Trevor Hastie (2001), "Estimating the number of clusters in a data set via

the gap statistic", Journal of the Royal Statistical Society - Series B: Statistical Methodology Vol. 63, pp. 411-423

Sandrine Dudoit and Jane Fridlyand (2002), **"**A prediction-based resampling method for estimating the number of clusters in a dataset" *Genome Biology,* No. 3, Vol.7, pp.1–21

Seong S Chae, Janice L DuBien and William D Warde (2006), "A method of predicting the number of clusters using Rand's Statistic" Computational Statistics and Data Analysis, Vol. 50, pp. 3531 – 3546.

Thorndike, R.L. (1953), "Who Belongs in a Family?" Psychometrika,Vol. 18, pp. 267 – 276.