Journal of Multidisciplinary Engineering Science and Technology (JMEST)
ISSN: 2458-9403
Vol. 3 Issue 3, March - 2016

Variables Selection In Cluster Analysis Using Goal Programming

Ramadan Hamid Mohamed
Professor of Statistics,
Faculty of Economics and Political Science,
Cairo University

Elham Abdul-Razik Ismail
Professor of Statistics,
Faculty of Commerce,
Al-Azhar University (Girls' Branch)

Mahmoud Mostafa Rashwan
Lecturer of Statistics,
Faculty of Economics and Political Science,
Cairo University

Safia Mahmoud Ezzat
Assistant Lecturer of Statistics,
Faculty of Commerce,
Al-Azhar University (Girls' Branch)

***Abstract*—Data clustering is a common technique for statistical data analysis; it is defined as a less of statistical techniques for classifying a set of observations into completely different groups. Variable selection becomes increasingly important in modern data analysis. In this study suggested a nonlinear goal programming approach that select the most important variables in clustering a set of data. The study evaluate the performance of the nonlinear goal programming suggested approach for selection variables in cluster analysis used numerical example. The suggested nonlinear goal programming approach selects the most important variable in cluster analysis simultaneously and the results are satisfied.**

Keywords—Cluster analysis; Variable selection; Goal programming; Nonlinear programming.

1- Introduction

Data clustering is a common technique for statistical data analysis; it is defined as a class of statistical techniques for classifying a set of observations into completely different groups Webb [1], Yeung and Ruzzo [2]. Cluster analysis seeks to minimize group variance and maximize between group variance.

Nevertheless there is a great importance for goal programming in treating cluster analysis problem because it enables to formulate more than one objective for clustering, and hence takes in consideration different criteria for achieving the optimal clustering. Moreover goal programming does not impose assumptions concerning the distribution of the criterion variables.

Recently variable selection becomes more important for a lot of research in several areas of application, since datasets with tens or hundreds of variables are available and may be unequally useful; some may be just noise, thus not contributing to the process.

Variable selection plays an important role in classification. Before beginning designing a classification method, when many variables are involved, only those variables that are really required should be selected; that is, the first step is to eliminate the less significant variables from the analysis. There can be many reasons for selecting only a subset of the variables instead of the whole set of candidate variables: (1) It is cheaper to measure only a reduced set of variables, (2) Prediction accuracy may be improved through exclusion of redundant and irrelevant variables, (3) The predictor to be built is usually simpler and potentially faster when fewer input variables are used and (4) Knowing which variables are relevant can give insight into the nature of the prediction problem and allows a better understanding of the final classification model. Research in variable selection

started in the early 1960s. Over the past four decades, extensive research into feature selection has been conducted. Much of the work is related to medicine and biology. The selection of the best subset of variables for building the predictor is not a trivial question, because the number of subsets to be considered grows exponentially with the number of candidate variables.

There have been many trials for variable selection in cluster analysis. Liu el at. [3] proposed a new variable selection method using the information criterion. The primary characteristic of the proposed method is that it works like hierarchical clustering where each variable is considered as a cluster and the between-cluster and within-cluster distances are measured by mutual information and the coefficient of relevancy respectively. Consequently, the final aggregated cluster is the selection result, which has the minimal intra-distance and the maximal inter-distance with the class cluster.

Shen el at. [4] proposed a regression method for simultaneous supervised clustering and variable selection over a given undirected graph, where homogeneous groups or clusters are estimated as well as informative predictors, with each predictor corresponding to one node in the graph and a connecting path indicating a priori possible grouping among the corresponding predictors. The method seeks a parsimonious model with high predictive power through identifying and collapsing homogeneous groups of regression coefficients.

Boutsidis and Ismail [5] presented the deterministic variable selection algorithm for K-means clustering with relative error guarantees. Their result improves upon this in two ways. First, their algorithms are deterministic; second, by using their deterministic algorithms in combination with this randomized algorithm. They can select features and obtain a competitive theoretical guarantee.

Benati and Garcia [6] focused on binary data where proposed a combinatorial model for clustering that selects simultaneously the best set of variables, the best set of median and the optimal data partition when the criterion used is the minimization of the total distance inside the cluster between the median of the cluster and the units that belong to the cluster. Instead of developing new solution tools for this nonlinear model, this approach is to study two different mixed integer linearization and to determine which one is the most efficient. The first is a direct linearization formulation of the initial quadratic model and the second is based on the so-called radius formulation of the p median problem.

The idea of the suggested approach depends on clustering data by minimizing the distance between observations within groups. Indicator variables are used to select the most important variables in the cluster analysis. This study suggested a nonlinear goal programming model that select the most important variables in cluster analysis.

The organization of the study is as follows: In Section 2 the study described cluster analysis by linear goal programming. In Section 3 the suggested nonlinear goal programming approach which used in this study. the study applied data analysis to evaluate the performance of the suggested approach under consideration in Section 4. Finally, concluding remarks are provided in Section 5.

2- Cluster analysis by linear goal programming

In Rashwan [7] presented a linear goal programming approach to obtain the optimal number of clusters. The idea of the approach is to minimize the total within groups distance and maximize the between groups distance.

Cluster analysis by linear goal programming approach described as follows:

Let i = 1,2,…,n be the set of observations that are to be clustered into m clusters (groups).

For each observation $i \in N$, we have a vector of observations $y_i = \{y_{i1}, y_{i2}, \ldots\ldots, y_{ip}\} \in R^p$, where p is the number of variables.

If the data is standardized using the formula $Z_k = \frac{(y_k - \mu_k)}{\sigma_k}$, Then we have the corresponding vector of observations $z_i = \{z_{i1}, z_{i2}, \ldots\ldots, z_{ip}\} \in R^p$.

Since we aim to construct *m* clusters, we start by defining *n* clusters fictitiously, *n-m* of which will be empty.

Therefore we define *nxn* (0-1) variables x_{ij} such that

$$x_{ij} = \begin{cases} 1 & \text{if the } i^{\text{th}} \text{ element belongs to the } j^{\text{th}} \text{ cluster} \\ 0 & \text{otherwise} \end{cases}$$

Where cluster j is non empty if x_{jj} =1 *j*=1,…,n.

These variables need to satisfy the following conditions Subhash [8]:

1- In order to insure that each element belongs only to one non empty cluster, then the following constraint is needed:

$$\sum_{j=1}^{n} x_{ij} = 1 \qquad i = 1,2,\ldots,n \qquad (1)$$

2- In order to insure that j^{th} cluster is non empty only if x_{jj} =1, then this can be represented as follows:

$$x_{jj} \geq x_{ij} \qquad \begin{matrix} i = 1,2,\ldots,n \\ j = 1,2,\ldots,n \end{matrix} \qquad (2)$$

3- In order to insure that the number of non empty clusters is exactly m, then this can be written as:

$$\sum_{j=1}^{n} x_{jj} = m \qquad (3)$$

For example if x_{77}=1, then cluster 7 is non empty and if it includes element number 1 and 3 in addition to element number 7, then we have (1,3,7) as cluster 7 and x_{17}= x_{37}=x_{77}=1, while x_{71}=x_{73}=x_{13}=x_{31}=0 and also x_{11}=x_{33}=0.

Note that summing (2) with respect to *i* results in the following set of constraints

$$nx_{jj} \geq \sum_{i=1}^{n} x_{ij} \qquad j = 1,2,\ldots,n \qquad (4)$$

Thus it reduces the number of constraints from n^2 to *n* as suggested in Arthanari and Dodge [9].

Rashwan [7] considered the following model given by

Minimize

$$F = (\sum_{i=1}^{n} d_{1i}^{-} + \sum_{i=1}^{n} d_{1i}^{+} + \sum_{j=1}^{n} d_{2j}^{-} + d_3^{-} + d_4^{+}, d_5^{+}) \qquad (5)$$

Subject to

$$\sum_{j=1}^{n} x_{ij} + d_{1i}^{-} - d_{1i}^{+} = 1 \qquad i = 1,2,\ldots,n \qquad (6)$$

$$nx_{jj} - \sum_{i=1}^{n} x_{ij} + d_{2j}^{-} - d_{2j}^{+} = 0 \qquad j = 1,2,\ldots,n \qquad (7)$$

$$\sum_{j=1}^{n} x_{jj} + d_3^{-} - d_3^{+} = 2 \qquad (8)$$

$$\sum_{k=1}^{p} \sum_{i=1}^{n} \sum_{j=1}^{n} (z_{ik} - z_{jk})^2 x_{ij}) + d_4^{-} - d_4^{+} = 0 \qquad (9)$$

$$\sum_{k=1}^{p} \sum_{\substack{r' \in I \; r \in I \\ (r'>r) \\ I=\{i: x_{ij}=1\}}} \left(\frac{\sum_{i=1}^{n} z_{irk} x_{ir}}{\sum_{i=1}^{n} x_{ir}} - \frac{\sum_{i=1}^{n} z_{ir'k} x_{ir'}}{\sum_{i=1}^{n} x_{ir'}} \right)^2 + d_5^{-} - d_5^{+} = 0 \qquad (10)$$

each x_{ij} is either 0 or 1

$$d_4^{-}, d_4^{+}, d_5^{-}, d_5^{+} \geq 0$$

$d_{1i}^{-}, d_{1i}^{+}, d_{2j}^{-}, d_{2j}^{+}, d_3^{-}, d_3^{+}$ are nonnegative integers.

3- Goal programming approach for selection variables in cluster analysis

The following approach is an extension of work presented by Rashwan [7]. The suggested nonlinear goal programming approach select the most important variable in cluster analysis simultaneously. The following section describe the suggested approach:

p the number of variables V_s such that:

$$v_s = \begin{cases} 1 & \text{if the } s^{\text{th}} \text{ variable is important} \\ 0 & \text{otherwise} \end{cases}$$

For s= 1,……,*p*

These variables need to satisfy the following condition:

In order to insure that the selected number of important variables is exactly *r*, then:

$$\sum_{s=1}^{p} V_s = r \qquad (11)$$

To obtain the most important variables, the suggested version to achieve this aim is the minimization of the total sum of square deviations within groups by minimizing the weighted total sum of squares of distance between all observations within each cluster. The suggested weights are the indicator variables V_s.

This objective may be written as

$$Min \sum_{s=1}^{p}(\sum_{i=1}^{n}\sum_{j=1}^{n}(z_{is}-z_{js})^2 x_{ij})V_s \qquad (12)$$

where z_{is} is the standardized i^{th} observations of the s^{th} variable.

The corresponding fourth goal, in which d^+ and d^- represent the non-negative deviational variables and for which d^+ needs to be minimized, can be written as:

$$\sum_{s=1}^{p}(\sum_{j=1}^{n}\sum_{i=1}^{n}(z_{is}-z_{js})^2 x_{ij})V_s + d^- - d^+ = 0 \qquad (13)$$

Since the model aims to select the important variables in cluster analysis (13) with respect to the structural constraints (1,3 ,5 and 11), the above analysis suggests the achievement function *F* to take the formula given in (14).

From the above discussion, the goal nonlinear programming model for the selection of variables in cluster problem takes the form:

Find the values x_{ij}, V_s, d^+ of *i,j*=1,2,…,*n* and *s*=1,2,…,*p* .

Which Minimize:

$$f = d^+ \qquad (14)$$

Subject to

$$\sum_{s=1}^{p} V_s = r \qquad (15)$$

$$\sum_{s=1}^{p}(\sum_{j=1}^{n}\sum_{i=1}^{n}(z_{is}-z_{js})^2 x_{ij})V_s + d^- - d^+ = 0 \qquad (16)$$

$$\sum_{j=1}^{n} x_{ij} = 1 \qquad i = 1,2,...,n \qquad (17)$$

$$nx_{jj} - \sum_{i=1}^{n} x_{ij} \geq 0 \qquad j = 1,2,...,n \qquad (18)$$

$$\sum_{j=1}^{n} x_{jj} = m \qquad (19)$$

each x_{ij}, V_s is either 0 or 1

4- Data analysis and results

This section discusses three different real data to evaluate the performance for the suggested approach. Fisher Iris data set, Leaf data set and Education data set which used to evaluate the performance of the nonlinear goal programming suggested approach for selection variables in cluster analysis. The steps which applied to evaluate the performance of the nonlinear goal programming suggested approach (NLGP) for selection variables in cluster analysis as follows:

1- The data used different number of variables with between small and fairly large.

2- The data used in the real case and the standardization case also when the actual cluster known and not.

3- Solved the suggested model in two case clustering with all variable and select the most important variable.

4- The results are based on the most three commonly validity measures to evaluate the suggested model, as follows:

- The correct classification percent.
- The adjusted rand index Yeung and Ruzzo [2] is external criterion, which evaluates the results of a clustering method using a pre-specified

structure imposed on a data set. The largest value of this index the better performance of the clustering method.

• Davies-Bouldin Index Webb [1] is internal criterion, which evaluates the clustering results in terms of quantities obtained from the data set itself. It is being more close to zero indicates a better clustering.

The correct classification percent and the adjusted rand index used if the actual clustering is known while the Davies-Bouldin Index used if the actual clustering is not known.

The solution steps

The goal nonlinear programming model presented above is a nonlinear goal programming model. The following steps are suggested as a technique to solve this problem:

Step 1

Specify the number of observations (*n*) and number of variables (*p*), then enter the real or the standardized values of each variable. The variables are standardized using this formula

$$Z_k = \frac{Y_k - \bar{Y}_k}{S_k}$$

where $\bar{Y}_k$ and S_k are the mean and standard deviation values respectively.

Step 2

The LINGO software used for solving the nonlinear goal programming problems with too many variables, constraints or both to solve the final model.

Step 3

Obtain the values of decision variables δ_j and hence state the most important variables and the clustering results.

Fisher Iris data set:

Fisher data [10] applied to evaluate the performance of the goal programming suggested approach for selection variables in cluster analysis. It considered a set of 150 objects to illustrate linear discriminate analysis. The data set describes three species (clusters) of Iris flowers: Setosa, Versicolor and Virginica on four variables on each plant (lengths (l) and widths (w) of sepals(s) and petals (p)). i.e. we have four variables s.l, s.w, p.l and p.w. The data set includes 50 plants in each cluster. In this study, a random sample of 30 observations is chosen. Ten observations are drawn from each of the three clusters. According to the actual clustering: the first cluster contains objects (1,2, 3,4,5,6,7,8,9,10), the second contains(11,12,13,14,15,16,17,18, 19,20) and finally the third cluster contains (21, 22,23,24,25,26,27,28,29,30).

The results are summarized in the following table (1).

Table (1)

The clustering results of NLGP for fisher data set

	Real data		Standardization data	
	All variable	Selection three variable	All variable	Selection three variable
Variables	V_1, V_2, V_3, V_4	V_2, V_3, V_4	V_1, V_2, V_3, V_4	V_1, V_3, V_4
Cluster 1	1,2,3,4,5, 6,7,8,9, 10	1,2,3,4,5, 6,7,8,9, 10	1,2,3,4,5, 6,7,8,9, 10	1,2,3,4,5, 6,7,8,9, 10
Cluster 2	11,12,13, 14,15,16, 17,18,19, 20, 30	11,12,13, 14,15,16, 17,18,19, 20	11,12,13, 14,15,16, 17,18,19, 20, 30	11,12,13, 14,15,16, 17,18,19, 20,27,30
Cluster 3	21,22,23, 24,25,26, 27,28, 29	21,22,23, 24,25,26, 27,28, 29,30	21,22,23, 24,25,26, 27,28, 29	21,22,23, 24,25,26, 28, 29
% Correct classification	96.67%	100%	96.67%	93.33%
Adjusted Rand index	0.898	1	0.898	0.808

Table (1) shows the suggested model (NLGP) give approximately the same results in two case clustering with all variable and select three variable. In this data selection variables is not necessary because the number of variables is small but it used to clarify cluster analysis.

Leaf data set:

According to Silva et al. [11] presented database comprises 40 different plant species (clusters) considered a set of 340 observations on 13 variables (Eccentricity, Aspect Ratio, Elongation, Solidity, Stochastic Convexity, Isoperimetric Factor, Maximal Indentation Depth, Lobedness, Average Intensity, Average Contrast, Smoothness, Third moment, Uniformity, Entropy). This data applied to evaluate the performance of the goal programming suggested approach for selection variables in cluster analysis. In this study, a random sample of 52 observations is chosen, 12 observations from the 1st cluster, 10 from the 2nd cluster, 10 from the 3th cluster, 8 the 4th cluster and 12 the 5th cluster. The listed results are summarized in the following table (2).

Table (2)

The clustering results of NLGP for leaf data set

	Real data		Standardization data	
	All variable	Selection nine variable	All variable	Selection nine variable
Variables	V_1,V_2 ,V_3 ,V_4,V_5,V_6, V_7,V_8,V_9, V_{10} , V_{11} ,V_{12},V_{13}	V_3,V_4,V_5, V_6,V_7 ,V_9 ,V_{10},V_{11},V_{12}	V_1,V_2,V_3, V_4,V_5,V_6, V_7 ,V_8,V_9, V_{10} ,V_{11} ,V_{12},V_{13}	V_1, V_2 ,V_3 ,V_4,V_5,V_6 ,V_7,V_8, V_{12}
Cluster 1	2,3,4,5,6,7 ,8,9,11,12, 24,26,30, 37, 38	3,5,7,8,9, 12,16,19, 22,38	2,3,4,5,6, 7,8,9,11, 12,24,37, 38	2,3,4,5,6,7 ,8,9,11,12, 16,19,22, 35, 38
Cluster 2	10,13,14,1 5,16,17,18 ,19,20,21	10,13,14,1 5,17,18,20 , 21	10,13,14,1 5,16,17,18 ,19,20,21	10,13,14,1 5,17,18,20 , 21
Cluster 3	43,44,45,4 6,47, 48	23,25,27,2 8,29,30,31 ,32,35,36	23,25,27,2 8,29,31,32	23,25,27,2 8,29,30,31 ,32
Cluster 4	22,23,25, 27,28,29, 31,32,33, 34,35,36, 39,40	1,2,4,6,11, 24,33,34, 37,39,40	1,22,26,30 ,33,34,35, 36,39,40	1,24,26,33 ,34,36,37, 39,40
Cluster 5	41,42,49, 50,51,52	41,42,43, 44,45,46, 47,48,49, 50,51,52	41,42,43, 44,45,46, 47,48,49, 50,51,52	41,42,43, 44,45,46, 47,48,49, 50,51,52
% Correct classification	59.62%	75%	84.62%	82.69%
Adjusted Rand index	0.288	0.484	0.588	0.566

In table (2) when leaf data set was used with fairly large variables. The results for the suggested model (NLGP) when the model selected 9 variables is satisfied, than the suggested model when it is used all variables.

Education data set:

Education data applied to evaluate the performance of the goal programming suggested approach for selection variables in cluster analysis. The annual statistical report of Ministry of Education 2007/2008 [12] includes many variables about education. The suggested model used to group the Egyptian governorates in six clusters together with selecting ten variables from the set of 18 variables related to basic education. The values of eighteen variables defined as follows:

1) The failure percentage in the primary certificate ***(v_1)***.

2) The failure percentage in the preparatory certificate ***(v_2)***.

(The failure percentage is the number of failing pupils divided by the total number of pupils who already attended final exams for each certificate).

3) The percentage of teachers having intermediate qualifications in the primary stage ***(v_3)***.

4) The percentage of teachers having intermediate qualifications in the preparatory stage ***(v_4)***.

(Teachers having intermediate qualifications are those who have a 3 or 5 years diploma certificate .i.e. not having a university level qualification certificate).

5) The percentage of non educational teachers in the primary stage ***(v_5)***.

6) The percentage of non educational teachers in the preparatory stage ***(v_6)***.

(Teachers who are not graduates of the faculty of education are called non educational).

7) The average class density in the primary stage

(v_7).

8) The average class density in the preparatory stage ***(v_8)***.

(The average class density is the total number of pupils divided by the number of classes for a given stage).

9) The percentage of non full day schools in the primary stage ***(v_9)***.

10) The percentage of non full day schools in the preparatory stage ***(v_{10})***.

11) The percentage of multi-shift schools in the primary stage ***(v_{11})***.

12) The percentage of multi-shift schools in the preparatory stage ***(v_{12})***.

(The multi-shift schools are those schools having two or three periods per day).

13) The percentage of defective school buildings ***(v_{13})***.

(Defective school buildings are those that have too old buildings or those that are about to collapse in both stages).

14) The illiteracy rate in the governorate ***(v_{14})***.

(The illiteracy rate is defined by the number of illiterate persons in the age group (10+) as a ratio of population in the same age group).

15) The females ratio of pupils in the primary stage ***(v_{15})***.

16) The females ratio of pupils in the preparatory stage ***(v_{16})***.

17) The percentage of special education schools in the primary stage ***(v_{17})***.

18) The percentage of special education schools in the preparatory stage ***(v_{18})***.

(Special education schools are those specified for disabled children).

The data set includes 27 objects (areas) which are the 27 governorates. The results are summarized in the following table (3).

Table (3)

The clustering results of NLGP for edu data set

	Real data		Standardization data	
	All variable	Selection ten variable	All variable	Selection ten variable
Variables	$V_1,V_2,V_3,V_4,V_5,V_6,V_7,V_8,V_9,V_{10},V_{11},V_{12},V_{13}V_{14},V_{15},V_{16},V_{17},V_{18}$	$V_1,V_2,V_7,V_8,V_{11},V_{12},V_{15},V_{16},V_{17},V_{18}$	$V_1,V_2,V_3,V_4,V_5,V_6,V_7,V_8,V_9,V_{10},V_{11},V_{12},V_{13}V_{14},V_{15},V_{16},V_{17},V_{18}$	$V_1,V_7,V_8,V_9,V_{10},V_{11},V_{12},V_{13},V_{14},V_{16}$
Cluster 1	1,2,14	1,9,11,12, 13,22, 25,26	1,2,14,25, 27	1,9,11,12, 13,22, 25,26
Cluster 2	3,4,5,10, 11,16,17, 18,19,20, 21,26	2,7,14,15	3,4,17,18	2,6,7,8,14, 15
Cluster 3	7,8,9,12, 13,24	3,4,17,18, 19	5,10,16,19 , 20,21	3,4,17,18, 19
Cluster 4	6,15,22	5,6,8,10, 16,20,21	6,7,8,15, 22	5,10,16,20 , 21
Cluster 5	23	23	23	23
Cluster 6	25,27	24,27	9,11,12,13 ,24,26	24,27
D. B. index	2.382	0.932	0.551	0.287

For education data set the study used the Davies-Bouldin index as a performance measurement. Table (3), shows the results for the suggested model (NLGP) when the model selected 9 variables is satisfied, than the suggested model when it is used all variables.

Conclusion

From the previous results when the study used three different types of real data the suggested have the following advantages:

- The suggested nonlinear goal programming approach selects the most important variable in cluster analysis simultaneously.
- The suggested nonlinear goal programming approach used when three different types of real data are used and the results is satisfied.
- The suggested nonlinear goal programming approach can be used with standardization and non standardization data and the result is satisfied.

References

1- Webb, A.R., Statistical Pattern Recognition, Second Edition, John Wiley & Sons Ltd, 2002.

2- Yeung, K.Y and Ruzzo W.L., "Details of the Adjusted Rand Index and Clustering algorithms

supplement to the paper "An empirical study on Principal Component Analysis for Clustering gene expression data", Bioinformatics, 2001, pp. 1-6.

3- Liu, H. Wu X. and Zhang S., "Feature Selection using Hierarchical Feature Clustering", the 20th ACM international conference on information and knowledge management, CIKM '11, ISBN 1450307175, 2011, pp. 979 – 984.

4- Shen, X., Huang H. and Pan W., "Simultaneous supervised clustering and feature selection over a graph", Biometrika, Vol. 99, Issue 4, 2012, pp. 899 – 914.

5- Boutsidis, C. and Magdon-Ismail M., "Deterministic Feature Selection for k-Means Clustering", IEEE Transactions on Information Theory, Vol. 59, Issue 9, 2013, pp. 6099 – 6110.

6- Benati, S. and Garcia, S., "A mixed integer linear model for clustering with variable selection", Computers & operations research, Vol. 43, ISSN 0305-0548, 2014, pp. 280 – 285.

7- Rashwan, M. M., Deterministic and Stochastic Programming for Cluster Analysis, Unpublished Ph.D. Thesis, Faculty of Economics and Political Science, Cairo University, 2006.

8- Subhash, S., Applied Multivariate Techniques, John Wiley & Sons, Inc., New York, 1996.

9- Arthanari, T.S. and Dodge Y., Mathematical Programming in Statistics, A Wiley Interscience Publication, New York,1993.

10- Fisher, R.A, "The use of Multiple Measurements in Taxonomic Problems", Annals of Eugenics, Vol.7, 1936, pp.179-188.

11- Pedro F. B. Silva, Andre R.S. Marcal, Rubim M. Almeida da Silva, "Evaluation of Features for Leaf Discrimination", Springer Lecture Notes in Computer Science, Vol.7950, 2013,197-204.

12- Ministry of Education, statbook, http://services.moe.gov.eg/books/A_0708/main_book0708.html,5/3/2010.