International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 3 Issue 12 ǁ December. 2015 ǁ PP.26-34

A Mathematical Programming Approach for Selection of Variables in Cluster Analysis

Safia M. Ezzat, Mahmoud M. Rashwan
Department of Statistics, Faculty of Commerce, AL-Azhar University, Girls' Branch-Cairo.
Faculty of Economics and Political Science, Cairo University.

ABSTRACT: 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. Cluster analysis seeks to minimize group variance and maximize between group variance. In this study we formulate a mathematical programming model that chooses the most important variables in cluster analysis. A nonlinear binary model is suggested to select the most important variables in clustering a set of data. The idea of the suggested model 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.

I. 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 [12], [13]. Cluster analysis seeks to minimize group variance and maximize between group variance.

Nevertheless there is a great importance for mathematical 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 mathematical programming does not impose assumptions concerning the distribution of the criterion variables.

In this study we formulate a mathematical programming model that chooses the most important variables in cluster analysis. A nonlinear binary model is suggested to select the most important variables in clustering a set of data. The idea of the suggested model 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.

II. Review of some related clustering methods

In this section we summarize one of the most commonly used clustering methods which is: the k-means method.

k- means method [7], [10]

The algorithm starts by choosing k seeds in some fashion (representing the k clusters) and the rest of the objects are processed sequentially. Each object is compared to all the seed points to see which is the closest and it is put in that cluster. Seed points are updated every time a new object is added to the cluster by assigning cluster mean as the updated seed point.

The process is composed of these three steps:-

- Partition the objects into k initial clusters.
- Proceed through the list of objects; assigning an object to the cluster whose centroid (mean) is nearest. Recalculate the centroid for the cluster receiving the new object and for the cluster losing the object.
- Repeat the previous step until no more reassignments take place.

III. Selection of variables in cluster analysis

Variable selection is defined as the problem of selecting input variables that are most predictive of a given outcome.

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.

There have been many trials for variable selection in cluster analysis ([5], [8], [9])**.**

Brusco and Cradit, **2001** [3] presented a variable-selection heuristic for K-means clustering (VS-KM) based on the adjusted rand index, which is one of the most famous evaluation criteria that are used to compare the performance of clustering methods, which can be applied to large datasets. The heuristic was subjected to

Monte Carlo study designed to test VS-KM. The results indicate that the heuristic is extremely effective at eliminating masking variables.

The method is illustrated as follows: Developing partitions for each individual variable by k-means algorithm. The next step is using the adjusted rand index to compute the degree of agreement between these partitions, and then comparing the results for each variable and the largest value will be chosen. The same process is repeated by using the chosen variable and the other variables, also we use the adjusted rand index to obtain the largest value.

IV. The suggested model

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 $x_{jj} = 1$ $\quad j$=1,…,n.

These variables need to satisfy the following conditions [11]:

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

$$\sum_{j=1}^{n} x_{ij} = 1 \qquad i = 1,\ldots,n. \qquad (4.1)$$

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

$$x_{jj} \geq x_{ij} \qquad i = 1,\ldots,n. \quad j = 1,\ldots,n. \qquad (4.2)$$

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

$$\sum_{j=1}^{n} x_{jj} = m \qquad (4.3)$$

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 (4.2) with respect to *i* results in the following set of constraints

$$n x_{jj} \geq \sum_{i=1}^{n} x_{ij} \qquad j = 1,\ldots,n \qquad (4.4)$$

This can be written as

$$n x_{jj} - \sum_{i=1}^{n} x_{ij} \geq 0 \qquad j = 1,\ldots,n \qquad (4.5)$$

Thus it reduces the number of constraints from n^2 to n as suggested in [1].

To achieve the aim of the study we define a set of variables as follows:

We define p 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:

$$\sum_{s=1}^{p} V_s = r \tag{4.6}$$

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

$$Min \ \sum_{s=1}^{p} (\sum_{i=1}^{n} \sum_{j=1}^{n} (z_{is} - z_{js})^2 x_{ij}) V_s \tag{4.7}$$

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

Since the model aims to select the important variables in cluster analysis (4.7) with respect to the structural constraints (4.1, 4.3, 4.5 and 4.6), the above analysis suggests the objective function F to take the formula given in (4.8).

The mathematical programming model

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

Find the values of x_{ij} ,V_S i,j=1,2,…,n and s=1,2,…,p

which minimize:

$$F = \sum_{s=1}^{p} (\sum_{j=1}^{n} \sum_{i=1}^{n} (z_{is} - z_{js})^2 x_{ij}) V_s \tag{4.8}$$

Subject to

$$\sum_{s=1}^{p} V_s = r \tag{4.9}$$

$$\sum_{j=1}^{n} x_{ij} = 1 \qquad i = 1,2,\ldots,n \tag{4.10}$$

$$nx_{jj} - \sum_{i=1}^{n} x_{ij} \geq 0 \qquad j = 1,2,\ldots,n \tag{4.11}$$

$$\sum_{j=1}^{n} x_{jj} = m \tag{4.12}$$

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

The previous model contains a very large number of variables and constraints. It cannot be solved directly by using the available mathematical programming software. An online site for solving mathematical programming models is used to treat this problem [4].

The solution steps

The mathematical programming model presented above is a binary non linear mathematical 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 standardized values of each variable.

Step 2

We used the GAMS software [2] with interaction with the (Neos Server for Optimization) for solving the binary non linear programming problems [4] with too many variables, constraints or both to solve the final model.

Step 3

Obtain the values of decision variables x_{ij}, V_s and hence state the most important variables and the clustering results.

V. Numerical Example

Ecoli Data Set

According to [6], Nakai considered a set of 220 observations to detect protein Ecoli bacteria using two rule-based expert system, Expert System as 1st cluster and A Knowledge Base as 2nd cluster on five variables (y_1, y_2, y_3, y_4, y_5) as a different method for analysis of the amino acid content of outer membrane and periplasmic proteins. The data set includes 143 in 1st cluster and 77 in the 2nd cluster. In this study, a random sample of 30 observations is chosen [Appendix], 20 observations in 1st cluster and 10 in the 2nd cluster. Since the mentioned four variables have different units of scale, the data is standardized using the famous formula $Z_i = \frac{y_i - \overline{y}}{s_y}$ where $\overline{y}$, s_y are the sample mean and standard deviation values respectively. The resulting standardized variables are z.y_1, z.y_2, z.y_3, z.y_4 and z.y_5. According to the actual clustering, the first cluster contains observations from 1 to 20 while the second includes observations from 21 to 30. The aim of the model is to choose the most important three variables in clustering this set of data. The suggested approach is applied as well as one of the classical methods (VS-KM) [3] on the above data set. The adjusted rand index [14] is used to compare each clustering method with the actual clustering.

Solution Steps

1. Set n=30, p=5: the observed values z_{i1}, z_{i2}, $z_{i3,}$ z_{i4} and z_{i5}, i= 1,2,…,n.
2. Substituting the previous values in the model, we have the problem:
   Find the values of x_{ij}, V_s *i*,*j*=1,2,…,30 and *s*=1, …,5.

which minimize:

$$F = \sum_{s=1}^{5} \left(\sum_{j=1}^{30} \sum_{i=1}^{30} (z_{is} - z_{js})^2 x_{ij} \right) V_s \tag{5.1}$$

Subject to

$$\sum_{s=1}^{5} V_s = 3 \tag{5.2}$$

$$\sum_{j=1}^{30} x_{ij} = 1 \qquad i = 1,2,\ldots,30 \tag{5.3}$$

$$30x_{jj} - \sum_{i=1}^{30} x_{ij} \geq 0 \qquad j = 1,2,\ldots,30 \tag{5.4}$$

$$\sum_{j=1}^{30} x_{jj} = 2 \tag{5.5}$$

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

3. We used the GAMS software [2] with interaction with the (Neos Server for Optimization) for solving the non linear programming problems [4], with too many variables, constraints or both.

4. The following results are obtained:

All the variables are zero except:

$v_3=1$	$v_4=1$	$v_5=1$		
$x_{1,16}=1$	$x_{2,16}=1$	$x_{3,16}=1$	$x_{4,16}=1$	$x_{5,16}=1$
$x_{6,16}=1$	$x_{7,16}=1$	$x_{8,16}=1$	$x_{9,16}=1$	$x_{10,16}=1$
$x_{11,16}=1$	$x_{12,16}=1$	$x_{13,16}=1$	$x_{14,16}=1$	$x_{15,16}=1$
$x_{16,16}=1$	$x_{17,16}=1$	$x_{18,16}=1$	$x_{19,16}=1$	$x_{20,16}=1$
$x_{21,24}=1$	$x_{22,24}=1$	$x_{23,24}=1$	$x_{24,24}=1$	$x_{25,24}=1$
$x_{26,24}=1$	$x_{27,24}=1$	$x_{28,24}=1$	$x_{29,24}=1$	$x_{30,24}=1$

Hence, the three selected variables are v_3, v_4 and v_5, and the two obtained clusters are (1 to 20) and (21 to 30).

These clustering results, together with the results obtained by applying VS-KM method are summarized in the following table (5).

Table (5)

The clustering results of BNLP and VS-KM for the data set of example

		BNLP	**VS-KM**
Variables		y_3, y_4, y_5	y_3, y_4, y_5
Cluster 1		(1,2,3,4,5,6,7,8,9,10,11, 12,13,14,15,16,17,18, 19,20)	(1,2,3,4,5,6,7,8,9,10,11, 12,13,14,15,16,17,18,19,20,21)
Cluster 2		(21,22,23,24,25,26,27,28,29,30)	(22,23,24,25,26,27,28, 29,30)
The center of cluster 1	y_3	0.452	0.451
	y_4	0.315	0.322
	y_5	0.394	0.400
The variance of cluster 1	y_3	0.006	0.005
	y_4	0.009	0.009
	y_5	0.014	0.014
The center of cluster 2	y_3	0.572	0.588
	y_4	0.751	0.782
	y_5	0.779	0.806
The variance of cluster 2	y_3	0.008	0.006
	y_4	0.014	0.005
	y_5	0.010	0.003
% Correct classification		100%	97%
Adjusted Rand index		1	0.865

From these results, it is clear that the VS-KM method and the suggested approach (BNLP) give approximately the same selected variables. The VS-KM method succeeds in classifying about 97% with a value of the adjusted rand index equal to 0.865. The corresponding values for the suggested approach (BNLP) are 100% and 1 respectively and seem to act as better. The advantage of the suggested model is that it takes into consideration all different combinations of variables when choosing the most important variables, this does not happen in the VS-KM method.

VI. Simulation study

The purpose of the simulation study is to assess the performance of the suggested model, and also to compare the proposed model with classical VS-KM method.

In the current simulation study, the following is considered:

1- Sample size is 30 observations.
2- The number of clusters is 3.
3- The number of selected variables is 2 from 3 variables, 3 from 4 and 5 variables and 6 from 10 variables.

Since the model contains a very large number of decision variables ($n^2 + p$) and also the same for number of constants ($2n$+2). It is very difficult to consider the case of large samples in the stimulation study according to the capacity of the available software. So, the study is limited to the case of small sample size (n is taken to be 30) and the number of clusters is limited to (3). The basic factor in the selection model is the number of variables. We consider the cases of p (number of variables) as 3, 4 ,5 and 10.

The overall simulation design could be summarized in the following table:

Combination number	**Sample size**	**Number of clusters**	**Number of variables**	**Number of selected variables**
1	30	3	3	2
2	30	3	4	3
3	30	3	5	3
4	30	3	10	6

For each combination 50 runs are generated. The simulation's results are based on two indices to compare between the suggested model and VSKM method, as follows:

- The correct classification percent
- The adjusted rand index.

These simulated runs have been done through building routines using four packages: Gams, SPSS, MATLAB and Microsoft Excel.

- MATLAB is used to generate simulated data.
- SPSS is used to solve the VSKM method.
- Gams is used to solve the suggested model. It can not be solved directly by using the available mathematical programming software. An online site for solving mathematical programming models is used to treat this problem [4].
- Microsoft excel is used to exchange data and out puts files among the pervious software packages result.

The results of BNLP model together with the results obtained by applying VS-KM method are summarized in following table:

Table (6)

The Correct classification and Adjusted Rand index results of BNLP and VS-KM for the simulation data set

Number of variables	BNLP			VSKM		
	Number of Runs	% Correct classification	Adjusted Rand index	Number of Runs	% Correct classification	Adjusted Rand index
3	39	100	1	38	100	1
	5	96.67	0.898	5	96.67	0.898
	4	93.33	0.792	3	93.33	0.792
	2	90.00	0.705	4	90.00	0.705
4	41	100	1	40	100	1
	6	96.67	0.898	7	96.67	0.898
	3	93.33	0.792	1	93.33	0.792
				2	90.00	0.705
5	40	100	1	40	100	1
	3	96.67	0.898	4	96.67	0.898
	5	93.33	0.792	2	93.33	0.792
	2	90.00	0.705	4	90.00	0.705
10	44	100	1	42	100	1
	5	96.67	0.898	6	96.67	0.898
	1	93.33	0.792	1	93.33	0.792
				1	90.00	0.705

Table (6) summarizes the total result of the simulation date set when using BNLP model in compared with the VSKM method. The BNLP model gives result as 39 of the runs succeed in classifying 100% with a value of the adjust rand index equal to 1 and 11 runs succeed in classifying about 94.24% with a value of the adjust rand index equal to 0.824. While the VSKM method gives result as 38 of the runs succeed in classifying 100% with a value of the adjust rand index1 and 12 runs succeed in classifying about 93.61% with a value of the adjust rand index equal to 0.807 when the number of variables is 3. The other cases 4, 5 and 10 variables are similar as the previous. In most cases, the suggested model acts better than VS-KM.

VII. Conclusion

This paper presents a mathematical programming model that chooses the most important variables in cluster analysis. A nonlinear binary model is suggested to select the most important variables in clustering a set of data. The suggested model seems to be at least equal in performance compared to classical methods.

The suggested approach was compared to one of the classical methods (VS-KM). The comparison was based on published data sets and simulation study. The results show that the suggested approach is promising and at least equivalent to the traditional methods. The advantage of the suggested model is that it takes into consideration all different combinations of variables when choosing the most important variables, this does not happen in the VS-KM method.

VIII. Acknowledgement:

A great appreciation and gratitude to **Prof Nadia Makary** and **Prof Ramadan Hamid** for their moral support, continuous help and valuable discussions.

Appendix

Ecoli Data Set

Obs.	Variables					Standardized Variables				
	Y1	Y2	Y3	Y4	Y5	Zy1	Zy2	Zy3	Zy4	Zy5
1	0.490	0.290	0.560	0.240	0.350	0.671	-1.803	0.696	-0.952	-0.798
2	0.200	0.440	0.460	0.510	0.570	-1.357	0.033	-0.328	0.215	0.223
3	0.420	0.400	0.560	0.180	0.300	0.182	-0.457	0.696	-1.211	-1.029
4	0.250	0.480	0.440	0.170	0.290	-1.007	0.522	-0.532	-1.254	-1.076
5	0.410	0.570	0.390	0.210	0.320	0.112	1.624	-1.044	-1.082	-0.937
6	0.400	0.450	0.380	0.220	0.000	0.042	0.155	-1.147	-1.038	-2.421
7	0.430	0.370	0.530	0.350	0.440	0.252	-0.824	0.389	-0.477	-0.380
8	0.340	0.330	0.380	0.350	0.440	-0.378	-1.314	-1.147	-0.477	-0.380
9	0.400	0.290	0.420	0.350	0.440	0.042	-1.803	-0.737	-0.477	-0.380
10	0.400	0.500	0.450	0.390	0.470	0.042	0.767	-0.430	-0.304	-0.241
11	0.420	0.380	0.540	0.340	0.430	0.182	-0.702	0.491	-0.520	-0.427
12	0.350	0.480	0.560	0.400	0.480	-0.308	0.522	0.696	-0.261	-0.195
13	0.440	0.510	0.470	0.260	0.360	0.322	0.889	-0.225	-0.866	-0.751
14	0.440	0.280	0.430	0.270	0.370	0.322	-1.926	-0.635	-0.822	-0.705
15	0.350	0.370	0.300	0.340	0.430	-0.308	-0.824	-1.966	-0.520	-0.427
16	0.340	0.420	0.410	0.340	0.430	-0.378	-0.212	-0.840	-0.520	-0.427
17	0.340	0.510	0.440	0.370	0.460	-0.378	0.889	-0.532	-0.390	-0.288
18	0.000	0.380	0.420	0.480	0.550	-2.755	-0.702	-0.737	0.085	0.130
19	0.260	0.400	0.360	0.260	0.370	-0.937	-0.457	-1.351	-0.866	-0.705
20	0.160	0.430	0.540	0.270	0.370	-1.637	-0.090	0.491	-0.822	-0.705
21	0.440	0.520	0.430	0.470	0.540	0.322	1.012	-0.635	0.042	0.083
22	0.630	0.470	0.510	0.820	0.840	1.650	0.400	0.184	1.554	1.475
23	0.400	0.500	0.650	0.820	0.840	0.042	0.767	1.618	1.554	1.475
24	0.480	0.450	0.600	0.780	0.800	0.601	0.155	1.106	1.381	1.289
25	0.310	0.500	0.570	0.840	0.850	-0.587	0.767	0.799	1.640	1.521
26	0.330	0.450	0.450	0.880	0.890	-0.448	0.155	-0.430	1.813	1.706
27	0.450	0.400	0.610	0.740	0.770	0.392	-0.457	1.208	1.208	1.150
28	0.710	0.400	0.710	0.700	0.740	2.210	-0.457	2.232	1.035	1.011
29	0.600	0.610	0.540	0.670	0.710	1.441	2.113	0.491	0.906	0.872
30	0.630	0.540	0.650	0.790	0.810	1.650	1.257	1.618	1.424	1.335
M	0.394	0.437	0.492	0.460	0.522	0	0	0	0	0
S.d	0.141	0.080	0.096	0.228	0.212	1.000	1.000	1.000	1.000	1.000

- ***Source.*** **The first fife columns represent a sample of 30 units drawn at random from [6]. The last four columns are calculated by the researchers.**

References

[1.] Arthanari, T.S. and Dodge Y. (1993), *Mathematical Programming in Statistics*, A Wiley Interscience Publication, New York.

[2.] Brooke, A., Kendrick D. and Meeraus A. (1988), *GAMS: A user's gudie*, Redwood City, CA: The Scientific Press.

[3.] Brusco, M.J. and Cradit J.D. (2001), "A Variable Selection Heuristic for K-Means Clustering", *psychometrika*, Vol. 66, No. 2, pp. 249-270.

[4.] Czyzy, J., Mesnier M. and More J. (1998), "The Neos Server", *IEEE Journal on Computational Science and Engineering*, Vol.5, pp.68- 75.

[5.] Fowlkes, E.B., Gnanadesikan, R. and Kettenring, J.R (1988), "Variable Selection in Clustering", *Journal of classification*, Vol.5, pp.205-228.

[6.] Frank, A. and Asuncion, A. (2010), UCI Machine Learning Repository *http://archive.ics.uci.edu/ml*. Irvine, CA: University of

California, School of Information and Computer Science.

[7.] Johnson, R.A. and Wichern D.W. (1998), *Applied Multivariate Statistical Analysis*, Prentice- Hall international, Inc., Fourth Edition.

[8.] Law, M.H., Figueiredo, A.T and Jain, A.K. (2004),"*Simultaneous Feature Selection and Clustering using mixture models*", IEEE Trans. *Pattern Analysis and machine intelligence*,Vol.26, No.9, pp.1154-1166.

[9.] Law, M.H., Jain, A.K., and Figueiredo, A.T. (2003),"*Feature Selection in Mixture – Based clustering*", *Advances in Neural Information Processing Systems*,Vol.15, pp.625-632.

[10.] MacQueen J.B. (1967), "Some Methods for Classification and Analysis of Multivariate Observations", *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, 1:281-297.

[11.] Rashwan, M.M. (2006), *Deterministic and Stochastic Programming for Cluster Analysis*, Unpublished Ph.D. Thesis, Faculty of Economics and Political Science, Cairo University.

[12.] Subhash, S. (1996), *Applied Multivariate Techniques*, John Wiley & Sons, Inc., New York.

[13.] Webb, A.R. (2002), *Statistical Pattern Recognition*, Second Edition, John Wiley & Sons Ltd.

[14.] Yeung, K.Y and Ruzzo W.L. (2001), "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*, pp. 1-6.