Clustering for Categorical Data Using Nonlinear Goal Programming

Shimaa Mohee
Assistant Lecturer of Statistics
Faculty of commerce, Al-Azhar University Girl's Branch, Cairo, Egypt.
Email: Dr.ShaimaMohi.team@azhar.edu.eg

Ramadan Hamid
Professor of Statistics
Faculty of Economics and Political Science, Cairo University, Egypt.
Research Professor, Social Research Center, AUC.
Email: ramadanh@aucegypt.edu

Fatma Amin Khalil
Associate Professor of Statistics,
Faculty of commerce, Al-Azhar University Girl's Branch, Cairo, Egypt.
Email: farmakhalil2722.el@azhar.edu.eg

Safia M. Ezzat
Associate Professor of Statistics
Faculty of commerce, Al-Azhar University Girl's Branch, Cairo, Egypt.
Email: safiahamed@azhar.edu.eg

ABSTRACT

Clustering is a popular unsupervised learning method used to group similar data points together. However, handling categorical variables in clustering can be challenging, as most clustering algorithms are designed to work with numerical data. Mathematical programming is a systematic model used for minimizing or maximizing the value of an objective function with respect to a set of constraints. The study suggests a mathematical nonlinear goal programming model for qualitative data clustering. The study clustering qualitative data using the suggested nonlinear goal programming model which has three main advantages: first is the data that is directly used, without the need of being converted to quantitative values, second is the optimal clusters which are automatically obtained by solving the optimization problem and third cluster analysis as an exploratory tool to support the identification of associations within qualitative data, cluster analysis can be helpful in identifying patterns where numerous cases are studied. The study evaluates the performance of the suggested mathematical nonlinear goal programming model using numerical examples by preliminary cases. The results for the suggested mathematical nonlinear goal programming model has to be proved efficient with a general average purity 64.2%.

Keywords:

Clustering, qualitative data, distance measure, nonlinear Goal programming, mathematical programming.

1. Introduction:

Clustering is the process of partitioning a set of observations into subsets. Cluster analysis is a useful tool used to identify patterns and relationships within complex datasets. It involves using algorithms to group data points into groups called clusters. Grouping data points based on their similarities and differences allows researchers to gain insights into the underlying structure of their data, Webb and Copsey (2011), Hair et al. (2013). Cluster analysis is a common technique for statistical data analysis used in many applications such as pattern recognition, image analysis, information retrieval, bioinformatics, computer graphics, biology, security and machine learning. It is typically used when there is no assumption made about the likely relationships within the data, Wang et al. (2021).

Qualitative data describes qualities or characteristics. It is collected using questionnaires, interviews, or observations and frequently appears in narrative form. Qualitative data may be difficult to precisely measure and analyze. Qualitative data consists of words, pictures and symbols that can be examined for patterns or meaning, sometimes with coding. The most common types of qualitative data are nominal data and ordinal data. Qualitative data clustering works with data composed only of qualitative attributes. Bingham (2023).

The goal of mathematical programming is to use mathematical models and particularly optimizing models to assist in taking decisions. The need for mathematical programming allows the formulation of more than one objective on the problem.

The study evaluates the performance of the suggested mathematical nonlinear goal programming model using numerical examples by preliminary cases. The results for the suggested mathematical nonlinear goal programming model has to be proved efficient with a general average purity 64.2%.

This paper is organized as follows. Section 2 describes the review of cluster analysis and qualitative data. Section 3 defines mathematical programming. The suggested nonlinear programming model is presented in section 4. The numerical example is introduced in Section 5 and finally, section 6 provides results and conclusions.

2. Overview of Cluster Analysis and Qualitative Data:

Clustering, the process of dividing a set of entities into homogeneous groups, is a powerful tool for identifying patterns within complex datasets. This technique, which originated in anthropology (Driver and Kroeber, 1932), was introduced to psychology by Zubin (1938) and later applied by Robert Tryon (1939) and Raymond Cattell (1943) for classifying personality traits. Clustering is essential in unsupervised learning and has diverse applications in data analysis. It categorizes data points to maximize within-cluster similarity and minimize inter-cluster similarity. Clustering algorithms generally fall into two categories: hierarchical and optimization clustering (Karthikeyan et al., 2020).

Hierarchical clustering is divided into agglomerative and divisive methods, constructing a hierarchy of clusters to reveal structures within an unlabeled dataset, as noted by Mann and Kaur (2013). Optimization methods differ by allowing entities to relocate and focusing on partitions that optimize a specific criterion, thus not requiring hierarchical structures. Feinstein (1996) and Webb (2003) described these methods as those seeking to maximize (or minimize) predefined measures, aiming for a clustering outcome that best satisfies the clustering criteria.

Distance or similarity measures are fundamental in distance-based clustering algorithms, helping to group similar data points within the same clusters and separate dissimilar or distant points into different ones. These measures use various mathematical models to quantify the proximity or distance between objects. The literature presents a wide range of methods and distance metrics, many of which can be combined in various ways. However, as noted by Shirkhorshidi et al. (2015) and Muniswamaiah et al. (2023), there is no clear guideline or established rule to determine the optimal combination of methods and distance metrics for clustering.

Clustering qualitative data is crucial for analyzing social concepts, values, and behaviors and often involves working with nominal or categorical information, like words, symbols, and images. This type of data is extensively studied for patterns, often through coding, and poses unique challenges, particularly in mixed-type datasets that combine quantitative and qualitative variables.

Caruso et al. (2021) proposed a mixed data clustering methodology to recognize risk groups, combining quantitative and qualitative customer features to enhance bank risk management. Their study highlighted how mixed data clustering can reveal informative patterns that aid in assessing credit risk. Similarly, Seca et al. (2020) introduced a Hierarchical Qualitative Clustering method using Maximum Mean Discrepancy, preserving interpretability in qualitative datasets. Applying this model to Spotify and financial datasets, they illustrated how clustering could group music artists and industries, with applications for investment diversification.

3. Mathematical Programming:

The systematic development of practical computing methods for linear programming began in 1952 at the Rand Corporation in Santa Monica, led by George B. Dantzig. Intensive work on this project continued there until late 1956, achieving significant advancements with first-generation computers. Efforts persisted in Washington and later expanded globally through various individuals and firms. By the late 1960s, advanced mathematical programming systems became widely available software for several computer models.

Mathematical programming seeks to aid decision-making using mathematical models, especially optimization models, allowing for multiple objectives to be addressed within a problem. One of its advantages is the use of sensitivity analysis to observe solution changes based on information shifts. Additionally, mathematical programming does not require assumptions about the distribution of criterion variables.

Mathematical programming is particularly valuable for cluster analysis, as it supports multiple clustering objectives, accommodating diverse criteria for optimal results. Sensitivity analysis enables examination of solution changes with varying data, and assumptions about criterion variable distribution are not necessary. Integer programming, goal programming, and stochastic programming are several models applied in cluster analysis. Arthanari and Dodge (1993).

4. Suggested Nonlinear Goal Programming model:

This section presents the suggested nonlinear goal programming model which is used in cluster analysis of the qualitative data described as follows:

Let $i = 1,2,\dots,n$ and $j = 1,2,\dots,n$ be the set of observations that are to be clustered into m clusters (groups).

For each observation i ∈ N, we have a vector of observations $y_{ik} = \{y_{i1}, y_{i2}, \dots \dots, y_{ip}\}$, where p is the number of variables.

Therefore, we can defin *n*x*n* (0-1) indicator 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 and only if x_{jj} =1, *j*=1,…,n.

These variables need to satisfy the following conditions:

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

$$\sum_{j=1}^{n} x_{ij} = 1 \qquad i = 1,2,\dots,n \qquad (4.1)$$

b. In order to ensure that jth cluster is non-empty only if $x_{jj} = 1$, then this has to be represented as follows:

$$x_{jj} \geq x_{ij} \qquad i = 1,2,\dots,n, \qquad (4.2)$$
$$j = 1,2,\dots,n$$

Note that summing (4.2) with respect to i results in the following set of constraints

$$nx_{jj} - \sum_{i=1}^{n} x_{ij} \geq 0 \qquad j = 1,2,\dots,n$$
(4.3)

c. In order to ensure that the number of non-empty clusters is exactly m, this has to be written as:

$$\sum_{j=1}^{n} x_{jj} = m \qquad (4.4)$$

d. In order to ensure that the minimization of the distance within groups, this has to be written as:

$$\sum_{k=1}^{p} \sum_{i=1}^{n} \sum_{j=1}^{n} (|y_{ik} - y_{jk}| x_{ij}) + d^{-} - d^{+} = 0 \quad (4.5)$$

where $|y_{ik} - y_{jk}|$ define the dissimilarity of two observations *i* and *j* based on the type of variable *p*.

If the p^{th} variable is categorical, then

$$|y_{ik} - y_{jk}| = \begin{cases} 0 \; if \; y_{ik} = y_{jk}, \\ 1 \quad otherwise. \end{cases}$$

If the p^{th} variable is binary, then

$$|y_{ik} - y_{jk}| = \begin{cases} 0 \; if \; y_{ik} = y_{jk} = 1, \\ 1 \quad if \; y_{ik} \neq y_{jk}. \end{cases}$$

From the above discussion, the suggested nonlinear goal programming model for the qualitative data in cluster analysis has to be written as:

Minimize:

$$F = d^{+} \qquad (4.6)$$

Subject to

$$\sum_{j=1}^{n} x_{ij} = 1 \qquad i = 1,2,\dots,n \qquad (4.7)$$

$$nx_{jj} - \sum_{i=1}^{n} x_{ij} = 0 \;, \quad j = 1,2,\dots,n \qquad (4.8)$$

$$\sum_{j=1}^{n} x_{jj} = m \qquad (4.9)$$

$$\sum_{i=1}^{n}\sum_{j=1}^{n} (|y_{ik} - y_{jk}|\, x_{ij}) + d^{-} - d^{+} = 0 \qquad (4.10)$$

each x_{ij} is either 0 or 1

$d^{-}, d^{+} \geq 0$

$m\ is\ number\ of\ cluster$

5. Numerical Example:

This section presents the numerical application for the suggested nonlinear goal programming for qualitative data clustering.

The study uses data whose quality has been proven through previous research in classification. The study applies the suggested nonlinear goal programming to the same data to verify the efficiency of the suggested nonlinear goal programming for clustering the qualitative data. It also discusses the suggested nonlinear goal programming for clustering to assess the accuracy, quality, and usefulness of qualitative research.

The six data sets under consideration have been collected from previous research and the study will apply suggested nonlinear goal programming model as follows:

First, Breast cancer data are used through previous research: Michalski *et al.* (1986), Clark and Niblett (1987), Cestnik *et al.* (1987), and Tan and Eshelman (1988). The previous study evaluated the model on breast cancer detection task, from the UCL machine learning repository where; Michalski *et al.* (1986), the accuracy range is 66%- 72%, Clark and Niblett (1987) the accuracy range is 65%- 72%, Tan and Eshelman (1988) the accuracy range is 68%- 73.5% and Cestnik *et al.* (1987) the accuracy is 78%. The dataset includes 286 observations divided into two clusters. The first cluster contains 201 observations of one class and 85 observations of another class. The instances have to be described by nine variables, three are binary and six are categorical and the class variable. The study chose 30, 100 and 200 as a different sample size and all observations have been applied. See https://archive.ics.uci.edu/dataset/14/ breast+cancer.

Second, nursery data are used through previous research. The past usage is the hierarchical decision model, from which this dataset is derived, was first presented in Olave *et al.* (1989) and within machine learning, the same data was used for the evaluation of HINT (Hierarchy Induction Tool) by Zupan and Bohanec (1997). The dataset includes 12960 observations. The instances are described by eight categorical variables and the class variable. The study chose 30, 100 and 200 observations as different sample sizes. See https://archive.ics.uci.edu/dataset/76/nursery.

Third, Tic-Tac-Toe Endgame data are used through previous research. The database encodes the complete set of possible board configurations at the end of tac- tac- toe games. The past usage of Matheus and Rendell (1989) applied to 100 – instance training and 200 – instance test sets. In a study using various amounts of domain- specific knowledge, its highest average accuracy was 76.6%. Matheus (1990) adding domain knowledge to SBL through feature construction, accuracy reached above 90%. and Aha (1991) used 70% for training, 30% of the instances for testing, evaluated over 10 trials the accuracy range is 82%- 99.1%. The dataset includes 958 observations. The instances are described by nine categorical variables and the class variable. The study chose 30, 100 and 200 observations as different sample sizes. See https://archive.ics.uci.edu/dataset/101/tic+tac+toe+endgame.

Fourth, Lenses data are used through previous research. The past usage of Witten and MacDonald (1988). The dataset includes 24 observations. The instances are described by four attributes, nominal and the class attribute is 3 classes each instance is complete and correct. 9 rules cover the training set. The study applied all observations see https://archive.ics.uci.edu/dataset/58/lenses.

Fifth, Congressional Voting Records data are used through previous research. This data set includes votes for each of the U.S. House of Representative Congressmen on the 16 key votes identified by CQA. The past usage of Schlimmer (1987) the accuracy range is 90%- 95%. The dataset includes 435 observations. The first cluster contains 267 observations of one class and 168 observations of another class. The instances are described by sixteen binary variables and the class variable. The study chose 30, 100 and 231 observations as different sample sizes. See https://archive.ics.uci.edu/dataset/105/congressional+voting+records.

Sixth, Car Evaluation_data are used through previous research. The past usage of Bohanec and Rajkovic (1988) and Zupan *et al.* (1997). The dataset includes 1728 observations. The instances are described by six categorical variables and the class variable. The study chose 30, 100 and 200 observations as different sample sizes. Within machine learning, this dataset was used for the evaluation of HINT (Hierarchy Induction Tool), which was prove to be able to completely reconstruct the original hierarchical model. see https://archive.ics.uci.edu/dataset/19/car+ evaluation.

We are using GAMS software with interaction with the Neos Server for Optimization for solving the nonlinear goal programming problems.

6. Results and Conclusion:

The study applied the previous data with different sample sizes different number of variables with 2 cluster only to evaluate the performance of the suggested nonlinear goal programming for clustering the qualitative data.

Table (1) shows the clustering results for the six data which has been used, number of variables for each case, number of correct observations, the purity clustering, number of clusters, and the types of variables. Breast cancer clustering results for the different four sample sizes are approximately the same and the average purity of clustering range is 51% - 53%. Nursery data clustering results for the different three sample sizes are approximately the same and the average purity of clustering is 71%. Tic-Tac-Toe Endgame data clustering results for the three different sample sizes are approximately the same and the average purity of clustering range is 43% - 60%. The results may be lower than the previous two cases, but we note that the percentage of purity increases with the increase in sample size. Lenses data clustering results for one sample size is satisfied and the purity of clustering is 71%. Congressional Voting Records data clustering results for the three different sample sizes are approximately the same and the average purity of clustering range is 86% -91%. We note that this case gives the highest average purity and compared to the rest of the results, we have found that all variables used in this case were binary. Car Evaluation data clustering results for the different three sample sizes are approximately the same and the average purity of clustering range is 48% - 73%. However, we have noted that the percentage of purity decreases with the increase in sample size, and we have found out that all variables used in this case are ordinal.

Table (1): - The clustering results for the 6 different datasets

Data	N. variable	Sample size	N. Correct	Purity	Average purity	N. Cluster	Type
Breast cancer	9	30	16	53%	52%	C=2	Binary, Nomina, Ordinal
		100	53	53%			
		200	104	52%			
		286	14٧	51%			
Nursery	8	30	21	70%	70%		Nominal
		100	71	71%			
		200	136	68%			
Tic-Tac-Toe	9	30	13	43%	49%		Nominal
		100	22	44%			

		200	**120**	**60%**			
Lenses	**4**	**24**	**١٧**	**71%**	**71%**		**Nominal**
Voting	**16**	**30**	**26**	**86%**	**88%**		**Binary**
		100	**91**	**91%**			
		231	**204**	**88%**			
Car Evaluation	**6**	**30**	**16**	**73%**	**63%**		**Ordinal**
		100	**33**	**69%**			
		200	**58**	**48%**			
Average General Purity					**64.2%**		

The suggested nonlinear goal programming model can be applied as in the previous cases when the sample sizes are small, 24-30, as well as if they are medium, 100, and if they are relatively large, 200-286. Likewise, about the number of variables, the model can be applied when the number of variables ranges between 4 and 16 and the classification is divided into 2 clusters. The six cases whose data were used three types of qualitative data (nominal, ordinal and binary).

The suggested nonlinear goal programming model has been proved an efficient model in the previous cases under study with an average general purity of 64.2%. In spite of a large number of data sets with qualitative attributes, there are few algorithms specifically tailored to cluster qualitative data. Depending on the results of previous preliminary cases, the suggested nonlinear goal programming model could be one of the few models designed to classify qualitative data. The study suggests that to verify and prove the efficiency of the suggested nonlinear goal programming model, this should be done by conducting a simulation method with different sample sizes and different numbers of variables. The study suggests model applying with larger sample sizes in future work.

Reference:

1. Arthanari T.S. and Dodge Y. (1993). *Mathematical Programming in Statistics*. Classic edition, John Wiley and Sons, New York
2. Bingham, A. J. (2023). From data management to actionable findings: A five-phase process of qualitative data analysis. *International journal of qualitative methods*, *22*, 16094069231183620.
3. Caruso, G., Gattone, S. A., Fortuna, F., & Di Battista, T. (2021). Cluster Analysis for mixed data: An application to credit risk evaluation. *Socio-Economic Planning Sciences*, *73*, 100850.
4. Feinstein, A. R. (1996). *Multivariable analysis: an introduction*. Yale University Press.
5. Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2013). *Multivariate data analysis: Pearson new international edition PDF eBook*. Pearson Higher Ed.
6. Karthikeyan, B., George, D. J., Manikandan, G., & Thomas, T. (2020). A comparative study on k-means clustering and agglomerative hierarchical clustering. *International Journal of Emerging Trends in Engineering Research*, *8*(5).
7. Mann, A. K., & Kaur, N. (2013). Review paper on clustering techniques. *Global Journal of Computer Science and Technology*, *13*(5), 43-47.

8. Muniswamaiah, M., Agerwala, T., & Tappert, C. C. (2023). Applications of Binary Similarity and Distance Measures. *arXiv preprint arXiv:2307.00411*.
9. Seca, D., Mendes-Moreira, J., Mendes-Neves, T., & Sousa, R. (2020). Hierarchical Qualitative Clustering: clustering mixed datasets with critical qualitative information. *arXiv preprint arXiv:2006.16701*.
10. Shirkhorshidi, A. S., Aghabozorgi, S., & Wah, T. Y. (2015). A comparison study on similarity and dissimilarity measures in clustering continuous data. *PloS one*, *10*(12), e0144059.
11. Wang, Z., Liu, X., & Li, Q. (2021). A Euclidean Distance Matrix Model for Convex Clustering. *arXiv preprint arXiv:2105.04947*.
12. Webb, A. R. (2003). *Statistical pattern recognition*. John Wiley & Sons.
13. Webb, A. R., & Copsey, K. D. (2011). *Statistical Pattern Recognition*. John Wiley & Sons.