

A k-means clustering approach based in mathematical programming: a bovine animal application

Alejandra Inga Quezada

Faculdade de Ciências Aplicadas- UNICAMP
Rua Pedro Zaccaria, 1300, Limeira, SP, 13484-350, Brasil
a211717@dac.unicamp.br

Washington Alves de Oliveira

Faculdade de Ciências Aplicadas- UNICAMP
Rua Pedro Zaccaria, 1300, Limeira, SP, 13484-350, Brasil
waoliv@unicamp.br

RESUMO

As técnicas de aprendizado de máquina buscam padrões estruturais nos dados para captar e fornecer informações importantes. Essas técnicas poderiam ser abordadas usando programação matemática (otimização) para fornecer novas possibilidades numéricas de resolução. Neste trabalho, usamos um modelo de clustering para o problema k -means, baseado em programação matemática inteira. O teste computacional foi realizado em dados simulados para um banco de dados de animais bovinos contendo características físicas e de raça de cada animal, mas com padrões desconhecidos. O interesse é identificar padrões entre os diferentes animais, agrupando-os a partir das soluções de um modelo de otimização linear inteiro. Os resultados computacionais são estudados a partir de um conjunto de procedimentos estatísticos descritivos para validar esta pesquisa.

PALAVRAS CHAVE. Clusterização, Programação Matemática, Aprendizado de máquina.

Programação Matemática, Pesquisa Operacional na Agricultura e Meio Ambiente

ABSTRACT

Looking for patterns in data to try to separate them into different groups by using, for example, similarity or dissimilarity rule is a common problem solved through clustering. This could be approached as a mathematical programming model that minimizes the distance between the data inside a cluster thus each will be as more homogeneous as possible. In this work, we use a clustering model for the k -means problem-based in integer mathematical programming. The computational test was conducted on simulate data for a bovine animal database containing physical and breed characteristics of each animal but with unknown patterns. The interest is to identify patterns among the different animals by grouping them from the solutions of an integer linear optimization model. The computational results are studied from a set of descriptive statistical procedures to validate this research.

KEYWORDS. Clustering, mathematical programming, decision making

Mathematical Programming, Optimization Research in Farming and Environment

1. Introduction

As technology advances, it becomes easier to capture and save different human activities such as buying a product, renting a movie, or going to the cinema. The information volume converts to data that are stored in large databases, usually used by companies to understand their customers preferences. Thus new products/goods and services are adjusted and offered for costumers according to these preferences.

Machine learning (ML) arise as a set of techniques with the purpose of understanding and extracting relevant information from data by seeking patterns in it, so future predictions can be made. Also, optimization (OP) is well known as a large area of study which techniques, seek to reach a specific objective due to a set of limitations or constraints.

Even though, both ML and OP belong to different areas, according to [Song et al., 2019] they are based fundamentally on artificial intelligence for which they can interact with each other and with themselves to overcome their own limitations by working together. In that way, ML can be expressed as an optimization model that uses experiences or example data whose patterns are taken as an input to learn from and optimize.

In this paper, we approach the k-means clustering problem based in mathematical programming proposed by [Gnagi e Baumann, 2017]. The aim for this model is to find the best number of groups such that data points inside of a specific group (similarity) are more similar than data points outside of this group (dissimilarity). We use simulated data from a farming company which works with bovine animals. The goal here is to seek the best number of groups or clusters of bovine animals based in three characteristics: ECC, frame and weight.

The remainder of the paper is organized as follows. In the next section, we present a short review of previous literature about clustering and its k -means algorithm. In section 3, we present the methodology and describe the model presented by [Gnagi e Baumann, 2017]. In section 4, we formally describe the problem and the characteristics of the bovine animal data. Finally, in section 5 we evaluate the results of the formulation and in section 6 we draw some conclusions and discuss promising future research directions.

2. Literature Review

According to Bishop [2006], clustering techniques belong to unsupervised learning, which operates with input data whose label or answer are unknown. The aim here is to find some patterns in the input data, see how often they occur and learn what generally happens or not. Some of the most popular unsupervised algorithms are K -means clustering, hierarchical clustering, and principal component analysis (PCA); the first two pursue to group similar examples according to which variables the data is conformed, and the last algorithm, allows to make a dimension reduction by combine variables and reduce them to the most representative ones.

Several applications have been made: Kakushadze e Yu [2017] use a k-means clustering algorithm to group types of cancers, Qureshi e Ahamad [2018] propose a cluster method using k-means clustering based image segmentation with neutrosophic logic, Liu e Ge [2018] applied a hierarchical clustering based in weight random forests to classify complex industrial process, and Wei et al. [2019] study several hierarchical divisive clustering for categorical data.

For clustering, the goal is to partition an unlabeled dataset into homogeneous groups or clusters considering just the information obtained solely from the data. It has applications in data mining [Berkhin, 2006], pattern classification [Sah et al., 2018], image segmentation [Dhanachandra e Chanu, 2017], among others. As mentioned in the previous section, this paper uses the k -means clustering, in which a data set is partitioned in k distinct non-overlapping clusters, which are specified in advance, where each observation is assigned to exactly one of the clusters. Moreover,

the clustering seeks to find the best k clusters while minimizing the distance of each point to its cluster centroid.

The k -means clustering can be define by given a dataset $D = \{x_i\}_{i=1}^n$ and a k number of clusters, so its clustering is denoted as $C = \{C_1, \dots, C_k\}$. It seeks to minimize the total sum of the variations between observations in each cluster after the data get partitioned in k clusters (1); in this case, we propose the sum squared errors (SSE) as a measure as it is the most common used one.

$$\underset{C_1, \dots, C_k}{\text{Minimize}} \quad \sum_{i=1}^k \sum_{x_j \in C_i} \|x_j - \mu_i\|^2. \quad (1)$$

3. Methodology

In section 2, the algorithm for k -means finds all the k^n ways to partition n observations into k clusters but for an n and k bigger it is almost impossible. To made it easier, an alternative approach is proposed which finds a local optimum instead of the globally optimum clustering [Gambella et al., 2019]. Consider the binary variable y_{ij} :

$$y_{ji} = \begin{cases} 0, & \text{if data point } j \text{ belong to cluster } i \\ 1, & \text{otherwise} \end{cases} \quad (2)$$

The k -means clustering is formulated as a mixed interger nonlinear program:

$$\underset{}{\text{Minimize}} \quad \sum_{j=1}^n \sum_{i=1}^k y_{ji} \|x_j - \mu_i\|^2 \quad (3)$$

$$\text{subject to} \quad \sum_{i=1}^k y_{ji} = 1, \quad \forall j \in n. \quad (4)$$

$$\sum_{j=1}^n y_{ji} \leq 1, \quad \forall i \in k. \quad (5)$$

Where the objective function (3) minimize the sum of the distance between all data points j and the center of their cluster i , considering the decision variable y_{ij} . The constraint (4) ensures that each point j is assigned only to one cluster i ; and the constraint (5) ensures that each cluster i have at least one point.

This formulation helps us to understand how machine learning interacts with optimization techniques, that according to Song et al. [2019], they interact into ways: machine learning for optimization and optimization for machine learning. To the propose of this research, we use the second one, as optimization is used to improve machine learning design errors, by decrease human participation.

The model proposed by Gnagi e Baumann [2017] uses the structure in (3)–(5) to develop a clusterization model based in mathematical programming to solve the k -means clustering problem. It uses object assignation with its respective cluster center through minimization of distances between objects.

Considering a set of observations with size n and d attributes with a number of clusters k

that is established by the modeler.

$$\text{Minimize} \quad z = \sum_{i=1}^n \sum_{j=1}^n \delta_{ij} z_{ij} \quad (6)$$

$$\text{subject to} \quad \sum_{i=1}^n z_{ij} = 1, \quad j = 1, \dots, n, \quad (7)$$

$$\sum_{j=1}^n z_{ij} \leq n y_i, \quad i = 1, \dots, n, \quad (8)$$

$$\sum_{i=1}^n y_i = k, \quad (9)$$

$$y_i, z_{ij} \in \{0, 1\}, \quad i = 1, \dots, n, \quad j = 1, \dots, n, \quad (10)$$

where δ_{ij} is the distance between object i and j , and z_{ij} is the decision variable on which an object i is chosen as cluster center of object j . Further, y_i represents which object i is design as cluster center.

Here, the objective function (6) seeks to minimize the distance between each object i and its assigned cluster center j while it considers three set of constraints: (7) guarantees that each object is only assigned to one cluster center; (8) assures that at most n objects are assigned to a center cluster of object i ; and (9) ensures that just k objects are selected as cluster centers.

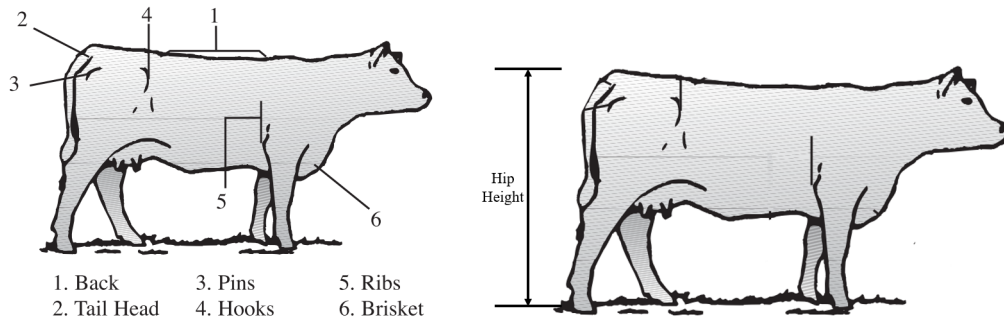
However, since this model is object-oriented, the number of variables and constraints increases as the size of the data does, so it also spend more running time to obtain an optimal solution. To overcome this issue, an scaling approach of the previous model is proposed by Gnagi e Baumann [2017] which seeks a set of representatives that represent correctly the original objects to decrease the number of variables and constrains.

4. The bovine animal cluster problem

To test the previous model, we work with a simulate database of bovine animals for a farming company whose objective is to obtain groups of their animals with the most homogeneous characteristics per cluster. This database has 102 cases and four characteristics: breed type, ECC, frame, and weight. The first has two types of breed cattle called: CA and NEL; the second is a body condition score (BCS or ECC as its initials in Portuguese) that describes the nutritional value through visual validation and is a scale from 1 to 9; the third is a numerical skeletal description called “Frame”, which has a range from 1 to 9; and the fourth measures the total weight of each beef in Kg.

The ECC suggests the relative fatness of the body composition of the cow where 1 is associated when a cow has a thin body and 9, is when the animal has extreme fatness. This score is an excellent indicator of the nutritional status in beef cows [Eversole et al., 2005] which will lead to establishing a new nutrition plan for each cow or group of cows to achieved optimal values. The ECC score is based in six key areas for evaluation: backbone, tail head, pins, hooks, ribs, and brisket (Figure 1(a)).

Likewise, Frame score describes the skeletal size of the cattle [McDonald, 1982] based on the hip height (Figure 1(b)) were 1 is the smallest and 9 represent the largest cow size. As our problem is to group the cattle by similarity, this score will indicate which one has an optimal weight to be a slaughter. Prediction of the animal maturity also can be done by seeking an optimal weight that matches the frame value considering that this score usually does not change in time.



(a) Visual areas used to determine BCS in beef cows [Eversole et al., 2005] (b) Proper position for correctly hip height measure.

Figure 1: Specific points in a cattle to determinate the ECC and Frame scores.

Previous statistical analysis was made, in order to look for a presumed linear relationship between variables that could impact the algorithm development. Figure 2 shown a pairwise plot for both breed categories and it indicates that neither feature has any linear relationship with the others so it is possible that no correlation exists too.

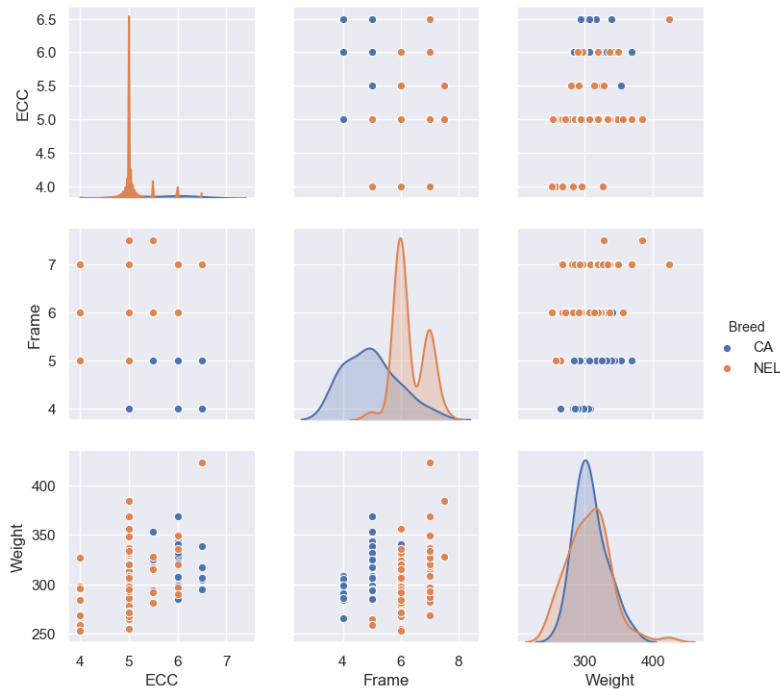


Figure 2: Pairwise plot of bovine animal dataset

The Algorithm 1 shows the procedure for the model in (6)–(10). It receives in input the set D of n bovine animals, the number of k clusters, and the maximum time for iteration as stopping criteria. Its output is the set C of cluster centers with their corresponding points for each bovine animal that belongs to them. In our work, before beginning the clustering process, standardization of D must be made so each characteristic contributes equally and we decide to use the euclidean

distance for each point of D to each cluster center.

Algorithm 1: Clustering algorithm based in mathematical programming

input : $D = \{x_1, \dots, x_n\}$ (set of data points to be clustered)
 k (number of clusters)
 $MaxTime$ (maximum time for iteration)
output: $C = \{c_1, \dots, c_k\}$ (set of cluster centers)
Label of points assign to one cluster c_k

Standardize D with mean = 0 and standard deviation = 1
Compute euclidean distance between each point of D called δ_{ij}
while $MaxTime$ **do**
 for $i = 1, \dots, n$ **do**
 for $j = 1, \dots, n$ **do**
 calculate $\sum_{i=1}^n \sum_{j=1}^n \delta_{ij} z_{ij}$
 if $\sum_{i=1}^n z_{ij} = 1$ **and** $\sum_{j=1}^n z_{ij} \leq ny_i$ **and** $\sum_{i=1}^n y_i = k$ **then**
 Find optimal minimum
 return vector $C = \{c_1, \dots, c_k\}$

5. Results

In this section, we present the results for a bovine animal database using the model in (6)–(10). All the routines were implemented in Julia language with CPLEX solver 12.9, and the computational tests were executed on a PC with an Intel Core i7-3770, 7.9 GB of RAM and Linux operating system.

The objective was to obtain the optimum number of k cluster that has the most homogeneous characteristics between them. For this, a cluster analysis is made by an internal measure based on sum square error (SSE).

$$SSE = \sum_{i=1}^k \sum_{x \in C_i} dist(c_i, x)^2 \quad (11)$$

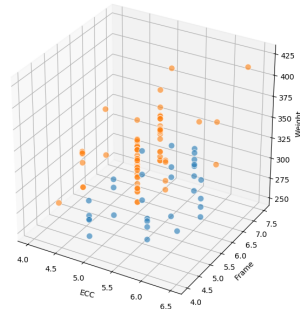
where x represent a data point in cluster C_i and c_i represent its cluster center. The SSE seeks to measure the goodness of a clustering structure by using distances (in this case, the Euclidean distance) between each observation and its cluster center.

Each feature gives independent information, thus the Algorithm 1 can be used correctly changing the number of clusters k as its SSE is calculated. The graphic representation for the first three tests is shown in Figure 3.

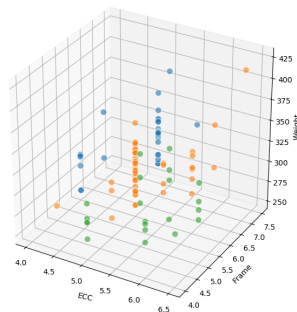
The principal objective of this application is to improve the sales of a farming company as with the correct number of groups or clusters, it would be easier to identify which group owns the best characteristics so it could be sold first at a better price. And which group has the worst characteristics, so maybe it can be improved before selling them.

To evaluate the optimum number of clusters k , a plot is generate between different values of k and their respective SSE. For this case, it is represent in Figure 4.

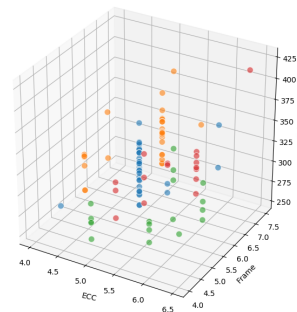
As can be seen, the lower SSE, the more similar data points are between each o inside a cluster, so the best number of clusters is where the graphic had a knee point. In this case, is when the number of clusters k takes a value of four.



(a) Two cluster case



(b) Three cluster case



(c) Four cluster case

Figure 3: 3D plot with two, three, and four clusters.

The average values of each feature for the four clusters are shown in Table 1. According to Eversole et al. [2005], the optimal ECC value is when it ranges between 5 and 7 because the cattle cow has a good overall appearance. When ECC is lower than 5, the cow loses more muscle tone because less fat is available to supply energy to sustain vital bodily functions [Gadberry, 2012]. From the results, the objects in cluster C_2 , have an average of 4.74 ECC, which means that there will be some cows in this cluster, which need a change in their nutrition plan so they can reach the ideal amount of fat at the moment of slaughter.

Cluster	ECC	Frame	Weight
C1	5.051	6.026	300.103
C2	4.741	6.828	311.483
C3	5.861	4.444	307.222
C4	5.688	5.688	321.563

Table 1: Average values of ECC, Frame and weight in each cluster

McDonald [1982] mentions in his article the range of weight a cattle should have depending on the Frame value to be considered as a potential market sell (Table 2). However, unlike ECC score, the higher the Frame score gets, does not mean that the cow is better; it is related to how long the cow will take to reach maturity. The experiment results in Table 1, clusters C_1 , and C_2 have an average of 6.026 and 6.828 Frame values respectively; nevertheless, their weight values do not

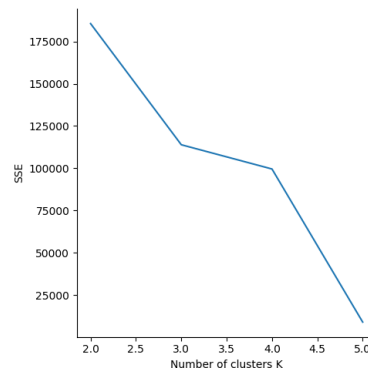


Figure 4: Total SSE by number of cluster

reach the 350 Kg in order to be considered ideal to slaughter. So it is appropriate to conclude that the cows belonging to those clusters, would take more time to gain weight and reach their corresponding maturity. Although, this can be accelerated by modify the nutrition plan or change the propose of the cattle, for example, using it for reproduction.

Frame	Weight (Kg)
1-2	150 - 180
3-5	200 - 350
6-9	350 - 450

Table 2: Optimal cattle weight according to its Frame score

6. Conclusions

In this work, we studied the k -means clustering problem with an optimization application perspective. Even though this technique has a machine learning background, the recent implementation of mathematical programming (or optimization techniques) for machine learning starts to grow such that one technique will improve the other.

The results of a clustering model based on mathematical programming for the bovine database allows seeking the best distribution in each cluster by minimizing the sum square error (SSE). Thus, each group will be as homogeneous as possible in order to improve the sales for the ones with the best characteristics, in this case based in their ECC, Frame and weight values and to look for new strategies to help the cows with worst development as for example, change the nutrition plan or use the cattle for other propose different than the slaughter.

As a future research, we intend to enlarge the computational tests with a bigger database, for which a large-scale method must apply a dimension reduction like Principal Component Analysis (PCA). Also, actual and new rates for clustering analysis will be used and then, compare with the ones obtained in the first numerical results.

References

Berkhin, P. (2006). A survey of clustering data mining techniques. In *Grouping multidimensional data*, p. 25–71. Springer.

- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag.
- Dhanachandra, N. e Chanu, Y. J. (2017). A survey on image segmentation methods using clustering techniques. *European Journal of Engineering Research and Science*, 2(1):15–20.
- Eversole, D. E., Browne, M. F., Hall, J. B., e Dietz, R. E. (2005). Body condition scoring beef cows.
- Gadberry, S. (2012). Feeding beef cows based on body condition scores. In *MP (University of Arkansas (System). Cooperative Extension Service)*. University of Arkansas Libraries, Fayetteville.
- Gambella, C., Ghaddar, B., e Naoum-Sawaya, J. (2019). Optimization models for machine learning: a survey. *arXiv preprint arXiv:1901.05331*.
- Gnagi, M. e Baumann, P. (2017). Large-scale clustering using mathematical programming. p. 789–793.
- Kakushadze, Z. e Yu, W. (2017). K-means and cluster models for cancer signatures. *Biomolecular detection and quantification*, 13:7–31.
- Liu, Y. e Ge, Z. (2018). Weighted random forests for fault classification in industrial processes with hierarchical clustering model selection. *Journal of Process Control*, 64:62–70.
- McDonald, A. (1982). Frame scoring of beef cattle. *Agnote-Melbourne. Dept. of Agriculture*.
- Qureshi, M. N. e Ahamad, M. V. (2018). An improved method for image segmentation using k-means clustering with neutrosophic logic. *Procedia computer science*, 132:534–540.
- Sah, S., Gaur, A., e Singh, M. P. (2018). Evaluating pattern classification techniques of neural network using k-means clustering algorithm. In *Next-Generation Networks*, p. 563–588. Springer.
- Song, H., Triguero, I., e Özcan, E. (2019). A review on the self and dual interactions between machine learning and optimization. *Progress in Artificial Intelligence*, 8(2):143–165.
- Wei, W., Liang, J., Guo, X., Song, P., e Sun, Y. (2019). Hierarchical division clustering framework for categorical data. *Neurocomputing*, 341:118–134.