

K-MEAN CLUSTERING OF HOLSTEIN FRIESIAN DAIRY CATTLE USING GENOMIC BREEDING VALUES

Buğra HOŞGÖNÜL^{1*}, Hasan ÖNDER¹


¹Ondokuz Mayıs University, Faculty of Agriculture, Department of Animal Science, 55139, Samsun, Türkiye


Abstract: Clustering refers to algorithms to uncover such clusters in unlabeled data. Data points belonging to the same cluster exhibit similar features, whereas data points from different clusters are dissimilar to each other. The identification of such clusters leads to segmentation of data points into a number of distinct groups. In this study it was aimed to classify the 492 Holstein Friesian dairy cattle with determining the optimum number of clusters using the genomic breeding values (GBVs) calculated with 13250 SNPs using GBLUP for milk yield (kg), milk fat (%), milk protein (%), milk lactose (%), and milk dry matter (%). Results showed that the optimum number cluster was determined as two for the genomic breeding values. Determining the most appropriate number of clusters, it provides great convenience in the selection of breeding animals after determining the animals that can provide optimum efficiency in the herd or the animals that need to be eliminated from the existing herd. As a result, it can be said that the k-means method can be used successfully in clustering animals for genomic breeding values, but for this, at first, the optimum number of clusters must be determined.

Keywords: K-mean clustering, Breeding value, Genomic selection, Dairy cattle

*Corresponding author: Ondokuz Mayıs University, Faculty of Agriculture, Department of Animal Science, 55139, Samsun, Türkiye

E mail: honder@omu.edu.tr (H. ÖNDER)

Buğra HOŞGÖNÜL  <https://orcid.org/0009-0002-9548-3457>

Hasan ÖNDER  <https://orcid.org/0000-0002-8404-8700>

Received: December 15, 2024

Accepted: January 08, 2024

Published: January 15, 2025

Cite as: Hoşgönül B, Önder H. 2025. K-mean clustering of Holstein Friesian dairy cattle using genomic breeding values. BSEng Sci, 8(1): 263-267.

1. Introduction

Clustering, as a generic tool for finding groups or clusters in multivariate data, has found wide application in biology, agriculture, psychology and economics (Kodinariya and Makwana, 2013). Cluster analysis encompasses different methods and algorithms for grouping objects of similar kinds into respective categories (Frades and Matthiesen, 2010). Clustering analysis method is one of the main analytical methods in data mining, the method of clustering algorithm will influence the clustering results directly (Na et al., 2010). Clustering is the separation of data with similar characteristics into groups. The general purpose of cluster analysis is to ensure homogeneity within the cluster and heterogeneity between the clusters. In other words, it is desired that the variance within the cluster is low and the variance between the clusters is high (Çolak et al., 2015). It is used to divide units or variables into homogeneous groups by using some measures calculated based on similarities or differences between variables. This allows similar individuals to be collected in the same cluster. Grouping ungrouped data according to their similarities helps the researcher to obtain appropriate usable summarizing information (Kodinariya and Makwana, 2013).

Clustering is the unsupervised, semi supervised, and supervised classification of patterns into groups (Frades and Matthiesen, 2010). Unsupervised learning is done by

grouping (clustering) only the elements that have similar properties in the data, without labeling the data set as cause-effect, input-output (Çolak et al., 2015).

One of the main difficulties for cluster analysis is that, the correct number of clusters of different types of datasets is seldom known in practice. However, most of clustering algorithms are designed only to investigate the inherited grouping or partition of data objects according to a known number of clusters. Thus, identifying the number of clusters is an important task for any clustering problem in practice albeit it must be faced with many operational challenges. A tractable way for cluster analysis is to ask the end user to input the number of clusters in advance, which needs the expert domain knowledge over the underlying datasets. On the other hand, many statistical criteria or clustering validity indices have been investigated in the sense of automatically selecting an appropriate number of clusters (Kodinariya and Makwana, 2013).

Today, there are hundreds of clustering methods and they are classified in various ways. According to a widely used classification, clustering methods can be examined in three groups as hierarchical methods, k-mean (partitioning) methods and mixed methods that combine them in various ways (Çolak et al., 2015).

K-mean clustering analysis has been perhaps one of the most widely used segmentation methods for more than 50 years. It has been among the most widely used



methods in almost every field such as economics, customer management, marketing, bioinformatics and engineering research, as well as in informatics applications such as object classification, image segmentation, data mining, and machine learning (Cebeci et al., 2015).

In animal breeding, it is important to separate animals into groups using breeding values determined for more than one character. K-mean clustering is a one of the important tool for this purpose. In this study, we aimed to classify the 492 Holstein Friesian dairy cattle with determining the optimum number of clusters using the genomic breeding values (GBVs) of milk yield (kg), milk fat (%), milk protein (%), milk lactose (%), and milk dry matter (%).

2. Materials and Methods

2.1. Materials

In this study, we used the genomic breeding values (GBVs for 13250 SNPs) of milk yield (kg), milk fat (%), milk protein (%), milk lactose (%), and milk dry matter (%) estimated using GBLUP for 492 Holstein Friesian dairy cattle from a previous published study (Önder et al., 2023).

2.2. K-mean Cluster Analysis and Distance Measures

2.2.1. K-mean cluster analysis

K-mean clustering algorithms are algorithms that divide/partition datasets into k subsets (or clusters). Therefore, one of the most studied issues is the selection of k, which must be known at the very beginning of the analysis, before any algorithm is run. This parameter indicates the number of clusters into which the data set will be clustered, in other words, it indicates the number of clusters present in the data set. A successful or correct clustering depends on the optimal choice of k. Because, regardless of k, partitioning algorithms will produce a valid or invalid clustering result. However, since the aim is to obtain a valid clustering result, finding and using the actual number of clusters or the number closest to it is necessary to ensure accurate results. In other words, the correct selection of k is the main determining factor for a successful cluster analysis.

According to the working mechanism of the k-means algorithm, firstly k objects are randomly selected to represent the center point or mean of each cluster. The remaining objects are included in the clusters to which they are most similar by taking into account their distances from the mean values of the clusters. Then, the mean value of each cluster is calculated and new cluster centers are determined and the distances of the objects to the center are examined again (Na et al., 2010; Kodinariya and Makwana, 2013).

The process of k-means algorithm can be defined as the input and the output that the input is number of desired clusters, k, and a database $D=\{d_1,d_2,\dots,d_n\}$ containing n data objects and the output is a set of k clusters (Na et al., 2010; Cebeci et al., 2015).

The algorithm basically consists of four stages:

The first step is randomly selecting k data objects from dataset D as initial cluster centers (Figure 1).

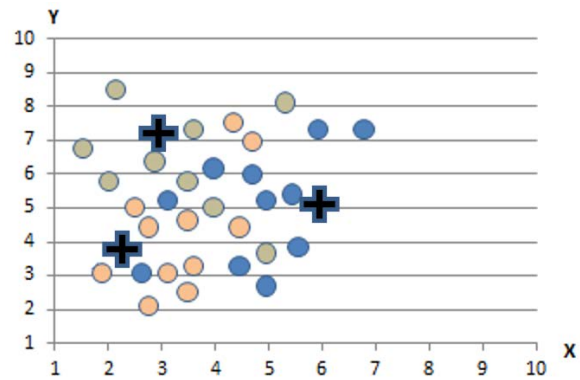


Figure 1. Randomly selected k cluster centers.

The second step is calculating the distance between each data object $d_i(1 \leq i \leq n)$ and all k cluster centers $c_j(1 \leq j \leq k)$ and assign data object d_i to the nearest cluster.

The step three is determining new centers according to the clustering (or shifting old centers to the new center) (Figure 2).

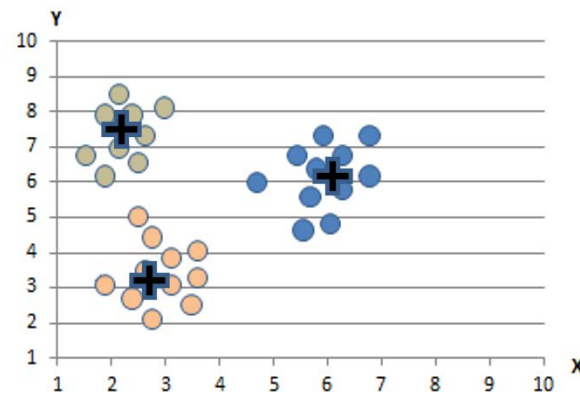


Figure 2. The new centers according to the clustering.

The fourth step is repeating steps two and three until a stable state is reached (Na et al., 2010; Çolak et al., 2015).

2.2.2. Distance measures

K-means is one of the clustering algorithms frequently used in the literature because it is simple and fast. The K-means algorithm divides the data set into k clusters and represents each cluster with a centroid (cluster center). The algorithm assigns the data to the closest centroid by using the squared distances between the data and the centroids.

Most clustering methods are based on the calculation of distances between observation values. Therefore, there is a need for relations that calculate the distance between two points. Euclidean distance can be calculated using Pearson or Manhattan distance formulas for distances between units in a data matrix containing continuous variables (Cebeci et al., 2015).

In the clustering phase, first the distance matrix is obtained. Distance measures can be used directly in clustering units or variables, or they can be used to calculate similarities and differences between units or variables (Figure 3).

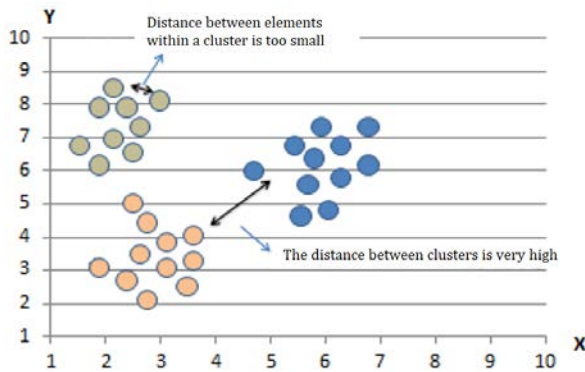


Figure 3. Data points and distances between clusters.

Euclidean distance

Euclidean distance (Equation 1) is a measure that determines the distances between the *i*th and *j*th observations in an *n*x*p* dimensional data matrix directly in the unit of measurement (Kurnaz and Önder, 2021).

$$d_{(i,j)} = \sqrt{\sum_{k=1}^p (X_{ik} - X_{jk})^2} \quad (1)$$

	A	B	C	D	E	F
1 NO		Milk	Fat	Pro	Lac	DM
2 X1		-761620627737,3250	93418419396,3995	1221379862090,4100	-888807597638,8460	-337628362351,5640
3 X2		-1280510865552,5700	346261253333,2680	1117933713690,1700	-1730445064381,5400	-564502156321,4430
4 X3		-711112554546,64200	313714512418,69700	649898437272,50700	-1293411366237,08000	-168220977207,51900
5 X4		-725803386369,30000	1017073959427,91000	1195837945823,68000	-1090929447420,98000	99357406828,63050
6 X5		-480576929324,79100	74826404622,23070	989828817931,66300	-1280047199605,54000	-193788801592,98200
7 X6		-864713241328,37900	392120314592,03600	1055314867556,73000	-538395630898,36200	-37459224667,76590
8 X7		-623171839016,41400	738436075054,16200	944613005602,30100	-858624793267,75500	50669825524,17770
9 XR		-802025203330,03400	583420057140,17800	461863278397,39200	-901929274436,06200	244452780346,11300

Figure 4. Sample data.

With the breeding values we obtained, the “readr” and “factoextra” libraries were used for K-mean clustering analysis in R software and the code used is given below.

```
library(readr)
library(factoextra)

mydata <- read.delim(file.choose()) #call the data file (txt).

str(mydata)

baru <- mydata[,-1] # deletion of the first column (animals) from the data.

rownames(baru) <- make.names(mydata$No, unique = TRUE) #determining the animals as clustered objects

databaru <- baru[sample(nrow(mydata)),c(2:5)]

fviz_nbclust(databaru, kmeans, method = "silhouette") # determining the optimum number of cluster
```

here *i*=1,2,...,*n*; *j*=1,2,...,*n* and *k*=1,2,...,*p*. *n* is the number of units and *p* is the number of variables.

Pearson distance

Pearson distance (Equation 2) is the Euclidean distance proportional to the variance of the variable. Pearson distance is also called standardized Euclidean distance (Immink et al., 2018).

$$d_{(i,j)} = \sqrt{\sum_{k=1}^p (X_{ik} - X_{jk})^2 / S_k^2} \quad (2)$$

Manhattan distance

Manhattan distance is a distance type calculated by taking the sum of the absolute distances between units. The Manhattan distance matrix (*D_M*) elements are calculated as follows (Kurnaz and Önder, 2021).

$$d_{M(i,j)} = \sum_{k=1}^p (|X_{ik} - X_{jk}|) \quad (3)$$

2.2.3. k-mean cluster analysis

For the genomic breeding values (GBVs for 13250 SNPs) of milk yield (kg), milk fat (%), milk protein (%), milk lactose (%), and milk dry matter (%) estimated using GBLUP for 492 Holstein Friesian dairy cattle were given as a small sample in Figure 4.

```
kmeans.awal <- kmeans(databaru,2) # the number (2) is applied number of cluster
```

```
kmeans.awal
```

```
fviz_cluster(kmeans.awal, databaru) # graphing the clusters
```

3. Results

The optimum number of clusters was determined using the breeding values we obtained and is shown in Figure 5.

According to the results obtained, the optimum number of clusters was determined to be two. In this study, results for three and four clusters were also given in order to show the effects of using different cluster numbers than the optimum number of clusters.

The graphed clusters were given in Figure 6, Figure 7, and Figure 8 for two, three and four clusters, respectively.



Figure 6. K-mean clusters results for two clusters.

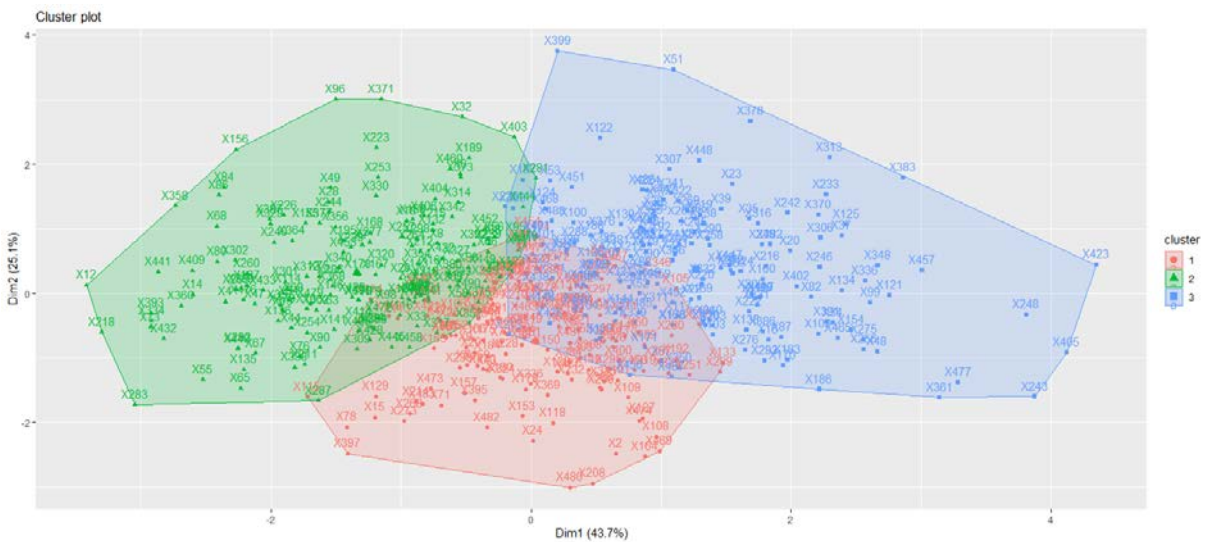


Figure 7. K-mean clusters results for three clusters.



Figure 8. K-mean clusters results for four clusters.

4. Discussion and Conclusion

Results showed that the optimum number cluster was determined as two for the genomic breeding values (GBVs for 13250 SNPs) of milk yield (kg), milk fat (%),

milk protein (%), milk lactose (%), and milk dry matter (%) estimated using GBLUP for 492 Holstein Friesian dairy cattle (Figure 5). When the Figure 5 was interpreted, it is understood that the two clusters are not

separated by definition and that some animals are located in the intersection of these two clusters. When we consider the number of clusters of three and four, it was easily understood that that some clusters are located within others and huge intersection of the clusters observed.

According to these results the population can be divided two groups such as high and low combined breeding values. If more than two clusters wanted to use the results is getting unclear. Determining the most appropriate number of clusters, it provides great convenience in the selection of breeding animals after determining the animals that can provide optimum efficiency in the herd or the animals that need to be eliminated from the existing herd. János et al. (2021) indicated in their study, they showed that cluster analysis had a positive effect on the herds they grouped in the breeding, feeding and breeding bull selection of Limousin breed cattle and that it would be beneficial for breeders. Doğan (2002) stated that it would be appropriate to use Cluster Analysis as a method in animal breeding, especially when making selection.

As a result, it can be said that the k-means method can be used successfully in clustering animals for genomic breeding values, but for this, the optimum number of clusters must first be determined.

Author Contributions

The percentages of the authors' contributions are presented below. The authors reviewed and approved the final version of the manuscript.

	B.H.	H.Ö.
C	70	30
D	70	30
S	30	70
DCP	70	30
DAI	70	30
L	70	30
W	70	30
CR	70	30
SR	70	30

C=Concept, D= design, S= supervision, DCP= data collection and/or processing, DAI= data analysis and/or interpretation, L= literature search, W= writing, CR= critical review, SR= submission and revision.

Conflict of Interest

The authors declared that there is no conflict of interest.

Ethical Consideration

Ethics committee approval was not required for this study due to there is no experimental study on research material.

References

- Cebeci Z, Yıldız F, Kayaalp GT. 2015. Choosing an optimal k in k-means clustering. 2. Ulusal Yönetim Bilişim Sistemleri Kongresi, October 8-10, Erzurum, Türkiye, pp: 231-242.
- Çolak B, Durdağ Z, Erdoğan P. 2015. Automatic clustering with k-means. *El-Cezeri J Sci Eng*, 3(2): 315-323.
- Doğan İ. 2002. Selection by Cluster Analysis. *Turk J Vet Anim Sci*, 26: 47-53.
- Frades I, Matthiesen R. 2010. Overview on techniques in cluster analysis. In: Matthiesen R (eds) *Bioinformatics Methods in Clinical Research. Methods in Molecular Biology*, vol 593. Humana Press. https://doi.org/10.1007/978-1-60327-194-3_5
- Janos T, Natasa F, Marton S. 2021. Determining the type of Limousin candidate bulls by cluster analysis. *Nat Resour Sust Devel*, 11(1): 113-120.
- Immink KAS, Cai K, Weber JH. 2018. Dynamic threshold detection based on Pearson distance detection. *IEEE Transact Commun*, 66(7): 2958-2965.
- Kodinariya TM, Makwana PR. 2013. Review on determining number of cluster in k-means clustering. *Int J Adv Res Comput Sci Manag Stud*, 1(6): 90-95.
- Kurnaz B, Önder H. 2021. Distance based regression models. II. *International Applied Statistics Conference*, June 29 – July 2, Tokat, Türkiye, pp: 120-126.
- Na S, Xumin L, Yong G. 2010. Research on k-means clustering algorithm: An improved k-means clustering algorithm. *Third International Symposium on Intelligent Information Technology and Security Informatics*, April 22, Jian, China, pp: 63-67.
- Önder H, Sitskowska B, Kurnaz B, Piwczynski D, Kolenda M, Sen U, Tırınk C, Çanga Boğa D. 2023. Multi-trait single-step genomic prediction for milk yield and milk components for Polish Holstein population. *Animals*, 13: 3070. <https://doi.org/10.3390/ani13193070>