Araştırma Makalesi
(Research Article)

Ege Üniv. Ziraat Fak. Derg., 2016, 53 (2):147-151
ISSN 1018 – 8851

Çiğdem TAKMA[1]
Öznur İŞÇİ GÜNERİ[2]
Yavuz AKBAŞ[1]

[1] Department of Animal Science, Faculty of Agriculture, University of Ege, 35100, Izmir, Turkey

[2] Department of Statistics, Faculty of Science, University of Muğla Sıtkı Koçman, 48300, Muğla, Turkey

corresponding author: cigdem.takma@ege.edu.tr.

Clustering of Holstein Friesians Using K-Means Method

Siyah Alacalar'ın K-Ortalamalı Kümeleme Yöntemi İle Sınıflandırılması

Alınış (Received): 06.11.2015 Kabul tarihi (Accepted): 29.02.2016

Key Words:

Holstein, milk yield, cluster, K-means, classification, nonhierarchical

ABSTRACT

By cluster analysis units or variables can be grouped according to similarities or differences in terms of their properties. In this study total of 4496 Holstein Friesian cows were grouped two, three and four clusters according to their herd, age at first calving, lactation length and 305 day milk yields. Nonhierarchical k-means clustering technique is used for this purpose. Related traits were found statically significant for clustering of Holsteins (P<0.01). Holsteins were divided into three clusters and these clusters were found statistically different (P<0.01). The correct classification percentage of cows was 98%. In the third cluster Holsteins which have the highest milk yield, the lowest age of first calving and 305-day lactation period were proposed for breeding programs.

Anahtar Sözcükler:

Siyah alaca, süt verimi, kümeleme, K-ortalamalı kümeleme

ÖZET

Kümeleme analizi ile birim veya değişkenler sahip oldukları özellikler bakımından benzerlik veya farklılıklarına göre gruplandırılabilmektedir. Bu araştırmada 4496 adet Siyah Alaca inek, sürü, ilkine buzağılama yaşı, laktasyon süresi ve 305 günlük süt verimi bakımından aşamalı (hiyerarşik) olmayan k-ortalamalı (k-means) kümeleme yöntemi ile iki, üç ve dört kümeye gruplandırılmıştır. Siyah Alacalar'ın kümelere ayrılmasında inceleme konusu özelliklerin istatistiksel olarak etkili oldukları saptanmıştır (P<0.01). Yine bu özelliklere göre Siyah Alacalar'ın üç farklı kümeye ayrıldığı ve bu üç kümenin istatistiksel olarak farklı olduğu belirlenmiştir (P<0.01). Siyah Alacalar'ın doğru sınıflandırma oranı ise %98 olarak bulunmuştur. Süt veriminin en yüksek, ilkine buzağılama yaşının en düşük ve ideal olan 305 günlük laktasyon süresine sahip üçüncü kümedeki Siyah Alacalar'ın ıslah çalışmalarında kullanılması önerilmiştir.

INTRODUCTION

Determination of individual differences for quantitative and qualitative traits is important in animal breeding. This is also used in the identification and taxonomic classification of species. However the genetic relationships among traits make the determination difficult. Since determination of the differences by molecular genetics techniques are costly, biometrical methods are more advantageous. On the other hand, univariate analysis of multiple traits is inadequate in the identification and classification of populations. So, in the evaluation of morphologic characteristics, it is essential to use multivariate statistical methods (Ruttner et al., 1978; Yakubu and Ugbo, 2010).

In multivariate statistical approach the method that helps to separate units or variables into similar sub-groups is referred to as the clustering method. Clustering method is used widely in the recent years particularly in social, medical and engineering sciences for different purpose. The method employs several approaches while classifying units according to their similarities. These approaches may be divided into two main groups as hierarchical and non-hierarchical methods (Johnson and Wichern, 2005). While in hierarchical clustering starting with each unit as separate cluster and merging the most similar pair of clusters successively to form a new cluster, units are separated into groups that are homogeneous inside and heterogeneous between groups in non-hierarchical clustering (Hair et al., 2006). K-means clustering method is the most widely used one among the non-hierarchical clustering methods (Singh and Singh, 2012).

In animal science the number of studies using clustering methods is fairly low. By using k-means clustering analysis Akıllı and Atıl (2013) classified 100 Holstein Friesians by their lactation order, 305 day milk yield, protein, fat, age at first calving, calving interval, milking day and calving season. In consequence of the analysis, Holstein Friesians are grouped under seven different clusters. Gürcan and Akçapınar (2002) used clustering analysis to examine 236 German Meat Merinos and 238 Karacabey Merinos in terms of ages, live weight, body measurements and fiber diameters. Through hierarchical clustering method, it was determined that the body sizes of the two genotypes are similar. It was determined that Karacabey Merinos have almost the same characteristics as pure German Meat Merinos. Kılıç and Özbeyaz (2010) clustered 100 Karakaya and 100 Bafra (Sakız x Karakaya G1) sheep according to their body dimensions. In their study, Bafra sheep were found to have more heterogeneous body characteristics than Karakaya sheep by using fuzzy clustering. Görgülü (2010) used fuzzy clustering to group 136 Simmental cows according to their lactation order, 305 days milk yield, age at first insemination, age at first calving and dry period and formed four clusters. On the other hand, Küçükönder et al. (2004) researched the characteristics of honeybees that are more effective in determining their breed by using k-means cluster analysis on some morphologic traits of 80 honey bees. At the end of the study it was determined that honeybees can be divided into six clusters in terms of their morphologic traits. Also in another study conducted by Küçükönder et al. (2015), a total of 282 Holstein Friesians were classified by means of fuzzy cluster analysis according to thirteen different traits that are somatic cell count, milk fat (%), milk protein (%), milk lactose (%), casein (%), urea (%), dry matter (%), non-fat dry matter (%), density (g/cm3), acidity (°SH), free fatty acids (mmol/10L), citric acid (%) and freezing point (°C). According to these parameters, Holsteins were classified in two different clusters.

The studies using k-means clustering is not yet common in the field of animal science. There is no other study where k-means clustering is used for dairy cattle. In the present study it was aimed to use k-means clustering method in order to group Holstein Friesians according to their herd, age at first calving, lactation length and 305 day milk yield traits.

MATERIAL and METHODS

Material

Data of this study consist of first lactation records collected within the period from 2001 to 2011 of 4496 Holstein Friesians raised at 214 herds in Isparta province which are the members of the Turkish National Association of Cattle Breeders. For clustering of cows, effects of herd, age at first calving, lactation length and 305 days milk yields were taken into account. Cluster analysis was performed after some restrictions on the data set as age at first calving between 24 and 40 months and lactation length between 220 and 305 days. Thus, cluster analysis was applied to a total of 4496 cows' data from 5355 Holstein Friesian cows.

Method

Cluster analysis is the collection of the techniques that are used to separate units or variables, whose natural grouping is unknown into similar sub groups on the basis of their similarities or dissimilarities. Cluster analysis is similar to discriminant analysis in terms of gathering similar units in the same group and similar to factor analysis in terms of gathering similar variables in the same group. Cluster analysis has also data reduction characteristic (Yim and Ramdeen, 2015). However, it is different from the other mentioned multivariable methods in terms of assumptions. Assumptions such as normality, linearity and homoscedasticity that are valid in many multivariate methods are not regarded in the clustering method. Also cluster analysis does not require any assumption during the determination of cluster number and structure. On the other hand, well sampling of the population and unavailability of multicollinearity among variables are two important points in cluster analysis (McMahon, 2001).

In the present study before clustering Holstein Friesians, the correlations between herd, age at first calving, lactation length and 305 days milk yield were examined. All correlations were found less than 0.70, that's why there is no multicollinearity among variables since if the correlation values among the variables are less than 0.90, it is assumed that there is not multicollinearity practically (Tabachnick and Fidell, 2007).

The k-means method is one of the non-hierarchical clustering methods that assign each unit or variable to a group that has closest each other by observing that similarity among the clusters is low but similarity within cluster is high (Johnson and Wichern, 2005; Çokluk et al., 2010). Similarity within a cluster is measured by the mean value (centroid) of the units within cluster. The steps of k-means algorithm are as follows:

Firstly, decide the number of clusters k. Then select initial centroids at random for k clusters. Assign each unit to the cluster with the nearest centroid. Compute each centroid again as the mean of the units assigned to it. New cluster centers (centroids) are obtained by calculating the mean values of the clusters and by recalculating the units' distances (Equation 1) to the new cluster centers. Repeat previous two steps until no change.

Units' probability to be assigned to clusters is either 0 or 1. When the covariance matrix within clusters becomes minimum, in other words when the distances of the units to the new centroids are the smallest, no further change takes place in clustering and the operation is stopped.

$$W_N = \frac{1}{n}\min \left\| X_i - a_{in} \right\|^2 \text{ and } i=1,\ldots\ldots,N \qquad [1]$$

where W_N is the variance covariance matrix within clusters; X_i are the observation vectors with p variables; a_{in} are the centroids of the clusters and $\left\| X_i - a_{in} \right\|^2$ is a measurement of the similarity between any unit and centroid of the cluster (Chernoff, 1972).

Another point in k-means clustering is the need to specify the *number* of clusters (*k*) in advance. Number of clusters is the most critical aspect in cluster analysis. However, this aspect could not be exactly clarified in the studies done up to now and some approaches do not work in large datasets. On the other hand, cluster analysis can also be combined with other multivariate statistical methods. For instance, cluster analysis can be tested through discriminant analysis in order to evaluate the statistical reliability of the clusters (Vogt and Nagel, 1992).

In the present study, cluster analysis was initially applied with two clusters. The significance of the difference among clusters was tested by discriminant analysis controlling Wilks' Lambda values. In cases where significant differences among clusters were determined, the analysis was repeated by increasing the number of clusters by one. This procedure continued until the difference among the clusters was non-significant and thus the number of clusters was decided. Also the correct classification rate of Holstein Friesians was examined by means of discriminant analysis. Correct classification rate was calculated by proportioning the total estimated number of units to the observed total number of units in clusters. SPSS 20 software was used for k-means cluster and discriminant analyses.

RESULTS

In the present study a total of 4496 Holstein Friesians were classified at first into two, then three and finally into four clusters by using k-means method. The numbers of Holstein Friesians assigned to groups in each clustering are presented in Table 1. From the table it can be seen that the numbers of Holstein Friesians distributed nearly the same into each of the two clusters while the numbers of Holstein Friesians accumulated into the second cluster where cluster numbers were three and four.

Table 1. Numbers of Holsteins within clusters

Number of Cluster	Cluster number	Number of Holstein Friesians
C=2	1	2158
	2	2338
C=3	1	1284
	2	2325
	3	887
C=4	1	694
	2	1722
	3	569
	4	1511

Significance levels of the variables involved in the clustering of Holstein Friesians into two, three and four clusters by variance analysis (ANOVA) were given in Table 2.

The difference between clusters in terms of all variables are statistically significant (P<0.01) for two, three and four clustering (Table 2). In other words, the variables of age at first calving, lactation length and 305 days milk yield are statistically significant variables for clustering Holstein Friesians.

On the other hand, testing the significance of the variables in discriminant analysis is summarized in

Table 3 for each clustering. Wilks' Lambda value of the four-group clustering was not significant (P=0.81). With this finding it was decided that the most suitable number of clusters for the dataset used in the study is three and therefore it does not make sense to create more than three clusters. For three clusters, 51.7% (2325) of the Holstein Friesians were assigned in the second cluster while the remaining 28.6% (1284) and 19.7% (887) were assigned to the first and the third cluster, respectively.

Table 2. The ANOVA results of variables on clustering Holstein Friesians into two, three and four clusters

Number of clusters	Variable	Mean Square of Clusters	df	Mean Square Error	df	F	Prob.
C=2	Age at first calving	194.28	1	14.71	4494	13.21	<0.01
	Lactation length	351049.37	1	513.23	4494	683.99	<0.01
	305 day milk yield	2947954905.94	1	430828.94	4494	6842.52	<0.01
C=3	Age at first calving	142.01	2	14.70	4493	9.66	<0.01
	Lactation length	172274.33	2	514.79	4493	334.65	<0.01
	305 day milk yield	1949682446.15	2	219170.99	4493	8895.71	<0.01
C=4	Age at first calving	111.64	3	14.69	4492	7.60	<0.01
	Lactation length	136888.63	3	500.19	4492	273.67	<0.01
	305 day milk yield	1433649472.13	3	129820.07	4492	11043.36	<0.01

Table 3. Wilks' Lambda test

Number of clusters	Function	Wilks' Lamda	Chi-Square	df	Prob.
C=2	1	0.39	4252.54	4	<0.01
C=3	1	0.20	7316.13	8	<0.01
	2	0.97	121.75	3	<0.01
C=4	1	0.12	9718.39	12	<0.01
	2	0.96	157.65	6	<0.01
	3	1.00	0.43	2	0.81

On the other hand, the crosstable presenting the correct classification rate of the discriminant function for three clusters is presented in Table 4. Nearly all of the Holstein Friesians in the first, second and third clusters in cluster analysis are accumulated again in the first, second and third clusters in discriminant analysis. The correct classification rate was 98% (4404/4496 = 0.98*100) where the sum of diagonal elements was 4404 (1256 + 2319 + 829). In other words, 98% of the Holstein Friesians were assigned correctly to the clusters formed in cluster analysis (Table 4).

Table 4. Cross tabulation for correct classification percentage

Number of clusters	Estimated clusters			
	1	2	3	Total
1	**1256**	28	0	1284
2	5	**2319**	1	2325
3	0	58	**829**	887
Total	1261	2405	830	4496

The mean and standard error of each variable for three clusters are presented in Table 5. Duncan test was also conducted in order to compare the means of the variables between clusters. Mean values of age at first calving, lactation length and 305 days milk yield were found significantly different between clusters (P<0.05). While the age at first calving was the highest in the first cluster, lactation length and 305 days milk yield were the highest in the third cluster.

Table 5. Mean and standard errors of each variable for three clusters

Variables	Number of clusters	Mean	Standard Error
Age at first calving	1	28.96^{a}	0.11
	2	28.59^{b}	0.08
	3	28.22^{c}	0.13
Lactation length	1	261.24^{c}	0.63
	2	279.36^{b}	0.47
	3	282.63^{a}	0.77
305 day milk yield	1	3989.23^{c}	13.00
	2	5325.17^{b}	9.69
	3	6705.08^{a}	15.91

DISCUSSION and CONCLUSION

In the present study Holstein Friesians were grouped by k-means clustering method based on herd, age at first calving, lactation length and 305 days milk yield. As a result Holstein Friesians were clustered into three groups. Before cluster analysis, it is essential to decide on the number of clusters. In our study, the number of clusters used in the classification of Holstein Friesians was determined experimentally. In these experiments two clusters were firstly assumed and the trials continued by adding one more cluster until the most suitable number of clusters was reached. It has been accepted that determining the number of clusters experimentally would be suitable especially in the case where a large number of data are used in classification.

In the present study differences among the clusters were determined in terms of the variables considered. After a successful clustering, the differences in variables among different clusters were expected due to homogeneity within clusters and heterogeneity between clusters. The results from cluster analyses may also vary depending on algorithms used in cluster analysis in addition to the number of clusters and variables used in clustering. In k-means clustering, a unit is either certainly included in a cluster or excluded from that cluster. There are also more flexible clustering methods such as fuzzy clustering where the belongings of each unit to every cluster are calculated separately between 0 and 1 instead of being defined exactly as only 0 or 1 as in k-means clustering. Use of k-means clustering on dairy cattle is rather rare (Akıllı and Atıl, 2013). Some studies (Görgülü, 2010; Küçükönder et al., 2015) on the other hand, are based on the implementation of fuzzy clustering analysis in dairy cattle breeding.

In conclusion, clustering of a large number of animals using their various characteristics is possible by k-means clustering method. Grouping animals on the basis of their characteristics can be used for animal nutrition or breeding etc. Large numbers of Holstein Friesians were grouped under only three clusters considering limited number of variables. Holstein Friesians within the third cluster had the highest milk yield, the lowest age at first calving and approximately 305 days lactation length and better performance in comparison to Holsteins in the other clusters.

REFERENCES

Akıllı A and H. Atıl. 2013. Classification of milk yield characteristics with cluster analysis. 7th International Balkan Animal Conference, BALNIMALCON. Namık Kemal University, Faculty of Agriculture, Department of Animal Science, 3-5 October, Tekirdağ, Turkey.

Chernoff H. 1972. Metric considerations in cluster analysis. Proceedings of the 7th Berkeley. Symposium on Mathematical Statistics and Probability. 1:621-629.

Çokluk Ö., G. Şekercioğlu and S. Büyüköztürk. 2010. Multivariable Statistics for Social Sciences (In Turkish). Pegem Akademi Yayıncılık, Ankara.

Görgülü Ö. 2010. Classification of dairy cattle in terms of some milk yield characteristics using by fuzzy clustering. Journal of Animal and Veterinary Advances 9:1947-1951 DOI 10.3923/javaa.2010.1947.1951.

Gürcan S. and H. Akçapınar. 2002. Alman et ve Karacabey Merinosu koyunlarının canlı ağırlık, vücut ölçüleri ve yapağı inceliği yönünden kümeleme analizi ile incelenmesi. Turkish Journal of Veterinary and Animal Sciences 26:1255-1261.

Hair J, B. Black, B. Babin, R. Anderson and R. Tatham. 2006. Multivariate Data Analysis. 6th ed. Prentice Hall, Upper Saddle River, New Jersey.

Johnson R.A. and D.W. Wichern. 2005. Applied Multivariate Statistical Analysis. 5th ed. Prentice Hall, Upper Saddle River, New Jersey.

Kılıç İ. ve C. Özbeyaz. 2010. Bulanık kümeleme analizinin koyun yetiştiriciliğinde kullanımı ve bir uygulama. Kocatepe Veterinary Journal 3: 31-37.

Küçükönder H, E. Efe, E. Akyol, M. Şahin ve F. Üçkardeş. 2004. Çok değişkenli istatistiksel analizlerin hayvancılıkta kullanımı. 4th National Animal Science Congress, 1-3 September, Isparta, Turkey.

Küçükönder H, T. Ayaşan and H. Hızlı. 2015. Classification of Holstein dairy cattles in terms of parameters some milk component belongs by using the fuzzy cluster analysis. Kafkas Universitesi Veteriner Fakültesi Dergisi, 23: 601-606 DOI 10.9775/kvfd.2015.12987.

McMahon R.G.P. 2001. Deriving an empirical development taxonomy for manufacturing smes using data from Australia's business longitudinal survey. Small Business Economics 17:197–212.

Ruttner F, L. Tassencour and J. Louveaux. 1978. Biometrical statistical analysis of the geographic variability of Apis Mellifera L. Apidologie 9:363-381.

Singh N. and D. Singh. 2012. Performance evaluation of k-means and heirarichal clustering in terms of accuracy and running time. International Journal of Computer Science and Information Technologies 3:4119-4121.

Tabachnick B.G. and L.S. Fidell. 2007. Using Multivariate Statistics. 5th ed. Pearson.

Vogt W. and D. Nagel. 1992. Cluster analysis in diagnosis. Clinical Chemistry 38:182–198.

Yakubu A. and S.B. Ugbo. 2011. An assessment of biodiversity in morphological traits of Muscovy ducks in Nigeria using discriminant analysis. International Conference on Biology, Environment and Chemistry, Singapore, 1:389-391.

Yim O. and K.T. Ramdeen. 2015. Hierarchical cluster analysis: comparison of three linkage measures and application to psychological data. The Quantitative Methods for Psychology 1:8-21.