Journal of Animal and Veterinary Advances 9 (14): 1947-1951, 2010
ISSN: 1680-5593

Classification of Dairy Cattle in Terms of Some Milk Yield Characteristics Using By Fuzzy Clustering

Ozkan Gorgulu
Ahi Evran University, Mucur Vocational High School,
40000, Kirsehir, Turkey

Abstract: Fuzzy clustering algorithms have been widely studied and applied in a variety of areas. They become the major techniques in cluster analysis. When using conventional clustering techniques, dairy cows can only belong to a group, having a particular performance. But actually, the same cows could be important from different perspectives at the same time to a different degree. Therefore, a fuzzy clustering approach is needed. The objective of the study was to show that whether fuzzy cluster analysis which has been used in different disciplines, may be used in dairy cow breeding studies or not. As a fuzzy cluster method, the fanny algorithm method was applied in this study. In terms of determination of clusters, the parameters were the number of lactations, 305-days milk yield, age at the first insemination, age at the first calving, the length of dry period, and the interval time between calving season. 136 dairy cows divided into four clusters using by fuzzy clustering technique. The four clusters differed significantly ($p<0.05$) from each other. The results show that fuzzy clustering can be used effectively on dairy cows breeding.

Key words: Fuzzy logic, dairy cattle, clustering, lactation, milk yield, Turkey

INTRODUCTION

In 1965 L.A. Zadeh published his famous paper, Fuzzy Set on the journal of Information and Control and the theory of fuzzy set commenced. In the following years, several people visualized this point through successful applications of fuzzy in several areas such as pattern recognition, switching functions, decision making, engineering, medicine, meteorology, manufacturing and more. The developments of this theory in the last several decades have been prosperous.

The main reason contributes to its success is that it is able to handle problems with indefinite nature by providing more transitional information than the crispy set theories for users to support better decisions (Chang and Chang, 2003; Halavati and Shouraki, 2005). Zadeh established the basis of fuzzy logic making a fundamental contribution to the representation of human knowledge, giving the basis for its proper treatment and formalization, taking into accounts the imprecision that governs it and its approximate nature. Within fuzzy logic, two of the principles of classic logic are overcome. The first one is the contradiction principle (i.e., at the same time B and not-B), implying that opposed concepts cannot overlap. In terms of sets, the intersection of the two sets B and not-B is the empty set according to the contradiction principle. The second one is the bivalence law. Any proposition has to be true or false because there are only two truth values: true or false. This implies the middle excluded law: A has to be either B or not-B (in other words, the union of the two sets B and not-B is the universal set). These two laws had kept classic (bi-valued) logic away from some aspects of the real world.

A crisp or classic set is defined so that the elements in a universe are divided in two groups: members (those that certainly belong to the set) and non-members (those that do not belong to it). There is a precise and clear distinction between members and nonmembers of the class or category represented by the set. Yet, it is known that in real world cases, many of the categories that we use do not fit this representation. In natural language we talk about expensive cars, low rents, wealthy regions, solvent customers or important (key) sectors. In these cases, the transition between members and nonmembers appears gradually. To cope with this, a fuzzy set can be mathematically defined to assign a value to each element of the universe that represents its membership to that set. This value is the grade to which an element is compatible with the concept represented by the fuzzy set. In other words: for any element j, its membership to a set i is given a certain value u_{ij}. For a classic or crisp set, u_{ij} has only two values, 0 and 1, meaning, respectively non-membership and membership. For fuzzy sets, u_{ij} can be any value between 0 and 1. As in fuzzy logic, precise reasoning is only a limit case of approximate reasoning so that fuzzy logic is a natural extension of classic logic Diaz *et al.* (2006). Cluster analysis is one of the multivariate techniques which allocation of distinct sets of data is objective and simply based on the similarity or dissimilarity of data sets. Each object can only belong to a determinate set (Chang and Chang, 2003). Fuzzy clustering is a special type of cluster analysis that is Capable of describing ambiguity in the data such as the existence of points that lie between two clusters Lenard *et al.* (2000). In fuzzy cluster analysis it is quite different because there are no sharp boundaries between clusters and an object can simultaneously belong to several clusters. Membership degrees between 0 and 1 are used for each object instead of assigning each element to one of the clusters (Diaz *et al.*, 2006).

Conventional clustering means classifying observations as exclusive subsets (clusters). That is, we can discriminate clearly whether an object belongs to a cluster or not. However, such a partition is insufficient to represent many real situations. Therefore, a fuzzy clustering method is offered to construct clusters with uncertain boundaries, thus this method allows that one object belongs to some overlapping clusters to some degree. In other words, the essence of fuzzy clustering is to consider not only the belonging status to the clusters, but also to consider to what degree do the objects belong to the clusters. There is merit in representing the complex data situations of real data (Sato and Jain, 2006). Suppose that:

$$X = \{x_1, x_2, \ldots\ldots, x_n\} \tag{1}$$

is a given set of n objects and K (K = 1, 2, 3,......,n, K ∈ N) is a number of clusters. Where, N is a set of all natural numbers. Then a fuzzy cluster which is a fuzzy subset in X is defined as follows:

$$\mu_k : X \to [0,1], k = 1,2,.....,k \tag{2}$$

The fuzzy grade for each fuzzy cluster k is denoted as:

$$u_{ik} \equiv \mu_k\, x_i, i = 1,2,.....,n, k = 1,2,.....,k \tag{3}$$

That is u_{ik} shows the degree of belongingness of an object i to a cluster k. In general, u_{ik} satisfies the following conditions:

$$u_{ik} = \in [0.1], \forall_i, k; \sum_{k=1}^{K} u_{ik} = 1, \forall_i \tag{4}$$

In fuzzy clustering, each cluster is a fuzzy set. It has its origins from Ruspini who pointed out several advantages of using fuzzy clustering. That is a membership value equal or close to one would identify core points in a cluster, while lower membership values in a cluster would identify boundary points. Bridge points, in Nagy's terminology may be classified within this framework as undetermined points with a degree of indeterminacy proportional to their similarity to core points. Bridges or strays between sets originated problems of misclassification in classic cluster analysis. Zadeh provides an outline of a conceptual framework for cluster analysis and pattern classification based on the theory of fuzzy sets, since there was still no unified theory at that time. Bezdek subsequently generalized Dunn's research, introducing a fuzzy version of the well known c means clustering algorithm. In Kaufman and Rousseeuw (1990) a different objective function is used. Their dissimilarity or distance measure makes it more robust than fuzzy c means (Rousseeuw, 1995). Hathaway present a generalization of fuzzy c means to allow the use of different distances (Diaz *et al.*, 2006). In this study, the fuzzy clustering algorithm and oclid distance, used to partition the data, as being seen in more detail in (Kauffman and Rousseeuw, 1990). In this study, fuzzy clustering analysis was applied on the data which obtained from Amasya Gokhoyuk State Farm. It was investigated that whether the cows' milk yield characteristics clustered homogeneously by using Fanny algorithm or not. The results of the fuzzy clustering, 136 dairy cows divided into four clusters.

MATERIALS AND METHODS

The collected data were the number of lactations, 305 day milk yield, age at the first insemination, age at the first calving, the length of dry period, the time interval between calving season were obtained from Simmental dairy cows kept in Amasya Gökhöyük State Farm. The data were collected during the period of on 15th January 2007 and 30th December 2008. In the study, S-plus package program was used for the application of fuzzy clustering analysis. Fuzzy cluster analysis can yield useful information on the natural structure of data. It allows for some ambiguity in the data, which often occurs in practice. The fuzzy clustering appears a suitable method if the clusters cannot be separated from each other clearly or if some units in cluster membership are undecided. In fuzzy clustering, each unit is assigned to various clusters and the degree of belonging of a unit to different clusters is quantified by means of membership coefficients which range from 0-1. Sum of the membership coefficients is always equal to 1. So, the unit is assigned to the cluster that has the highest membership coefficient. The membership functions are the functions that characterize the fuzziness in a fuzzy cluster whether or not the elements in cluster are continual or transitory.

The units quietly resembling to each other take place according to the relation of high membership in the same cluster. That's why the fuzzy cluster method calculates the coefficients of the units belonging to the cluster or clusters. Fuzzy cluster has two basic methods. From these, c average cluster method depends on c divisions. The other method depending on fuzzy equality relation is called as a graded cluster method depending on fuzzy equality relation. The similarity structures of cows were found as a basis of Fanny Algorithm that depends on fuzzy equality relation. Fuzzy cluster technique used in this algorithm aims at minimization of objective function below the membership functions in this objective function have these limitations:

$$u_{iv} \geq 0 i = 1,2,.........,n \text{ and } v = 1,2,.....,k \tag{5}$$

$$\sum_{v=1}^{k} u_{iv} = 1 \tag{6}$$

Here, each unit i and each cluster v will be a member of. U_{iv}. U_{iv} is the membership coefficient or the degree of belongingnes of unit i to cluster v. These constraints imply that membership cannot be negative and that each object has a certain total membership distributed over different clusters. By convention, this total membership is normalized to 1. Under these circumstances the objective function is as follows (Mocz, 1995; Tufan and Hamarat, 2003):

$$C = \sum_{v=1}^{k} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} u_{iv}^2 u_{jv}^2 d(i,j)}{2\sum_{j=1}^{n} u_{jv}^2}$$

$$i = 1,2,....,n; j = 1,2,....,n \text{ and } v = 1,2,....,k \tag{7}$$

The Euclidean distance measure (d_{ij}) is used to compute distances between objects and to quantify the degree of similarity for each object. The degree of similarity for each object i and j is computed as:

$$d(i,j)=\sqrt{(x_{i1},-x_{j1})^2+...+(x_{ip}-x_{jp})^2} \tag{8}$$

where the pth measurement of the ith object is given by x_{ip} and d (i, j) is the actual distance between objects i and j (Lenard *et al*., 2000). Fuzzy cluster is evaluated how far it is from certain cluster through Dunn partition coefficient. Dunn's partition coefficient is computed:

$$F_k = \sum_{i=1}^{n}\sum_{j=1}^{n} u_{iv}^2 n \tag{9}$$

Coefficient of Dunn takes values from the minimum 1/k for a completely fuzzy partition i.e., all u_{iv} = 1/k to the maximum of 1 in the case where all u_{iv} = 1 or 0. As this measure is not easily comparable the normalized Dunn coefficient is commonly applied:

$$F_k^{/} = \frac{F_k - (\frac{1}{k})}{1-(\frac{1}{k})} = \frac{kF_k - 1}{k-1} \tag{10}$$

Normalized Dunn coefficient varies from 0-1 independent of number of clusters and is called as Non-fuzziness index. It takes values from 0-complete fuzziness, when membership indices have the same value, to 1-no fuzziness, when each unit is assigned to a certain cluster with the membership coefficient of 1 (Lenard *et al*., 2000; Tufan and Hamarat, 2003). To select the best number of clusters k for the FANNY clustering method we use the silhouette coefficient which not only makes the grouping more robust but also helps to select the suitable number of groups, a question that is always important in cluster analysis.

The silhouette width has values between -1 and +1 and we have $-1 \leq s_i \leq +1$. If s_i = 1, the within dissimilarity is much smaller than the between dissimilarity, therefore object i has been assigned to an appropriate cluster. Lastly the average silhouette coefficient for each cluster and an overall average silhouette coefficient are calculated. The overall average silhouette is average of for all objects in the whole dataset. The largest overall average silhouette indicates the best clustering. Therefore, the number of clusters with maximum overall average silhouette is considered the optimal number of clusters (Tufan and Hamarat, 2003; Diaz *et al*., 2006).

RESULTS AND DISCUSSION

It was investigated that whether the cows milk yield characteristics clustered homogeneously by using Fanny algorithm or not.0 A suitable cluster number was determined by being changed the number of cluster

Table 1: Average and overall average silhouette coefficient for each number of cluster

No. of cluster	Average silhouette coefficient for each cluster s(i) 1	2	3	4	5	6	7	8	9	10	Overall average silhouette coefficients
2	0.6489	0.1924									0.4658
3	0.4746	0.3443	0.6270								0.5015
4	0.3887	0.6142	0.4154	0.6453							0.5285
5	0.1826	0.5349	0.6541	0.4520	0.5188						0.4877
6	0.2073	0.3656	0.6511	0.4340	0.6276	0.5040					0.4676
7	0.1769	0.5505	0.6167	0.3292	0.5046	0.4086	0.3755				0.4311
8	0.0872	0.5888	0.5349	0.3427	0.4329	0.3177	0.2324	0.5417			0.4019
9	0.4907	0.5341	-0.069	0.3427	0.5673	0.3669	0.3336	0.4245	0.1201		0.3573
10	0.4907	0.4863	-0.069	0.3427	0.5411	0.4038	0.2904	0.2840	0.3509	0.2714	0.3483

Table 2: Dunn and normalized dunn coefficients

Parameters	2	3	4	5	6	7	8	9	10
Dunn	0.637	0.5653	0.5524	0.4690	0.4537	0.3735	0.3315	0.3061	0.2816
Normalized Dunn	0.2740	0.3479	0.4033	0.3362	0.3444	0.2691	0.2360	0.2193	0.2018

Table 3: Descriptive statistics of clusters in terms of some milk yield characteristics

Factors	Cluster 1 (n = 38)	Cluster 2 (n = 39)	Cluster 3 (n = 22)	Cluster 4 (n = 37)
No. of Lactation	2.92±0.27^{c}	2.66±0.29^{b}	3.36±0.36^{d}	2.45±0.20^{a}
305 day milk yield	6837.10±143.93^{d}	4684.56±49.73^{b}	3468.31±116.41^{a}	5479.89±36.08^{c}
Age of first insemination,	614.08±14.08^{c}	587.71±8.76^{a}	593.36±8.44^{ab}	602.48±13.50^{b}
Age of first calving	909.51±13.14^{bc}	862.18±20.91^{a}	900.04±13.43^{b}	914.40±17.11^{c}
Dry period	60.13±3.48^{a}	73.33±9.90^{c}	95.63±8.41^{d}	66.54±3.50^{b}

^{a,b,c,d}Means followed by different superscripts differ ($p<0.05$)

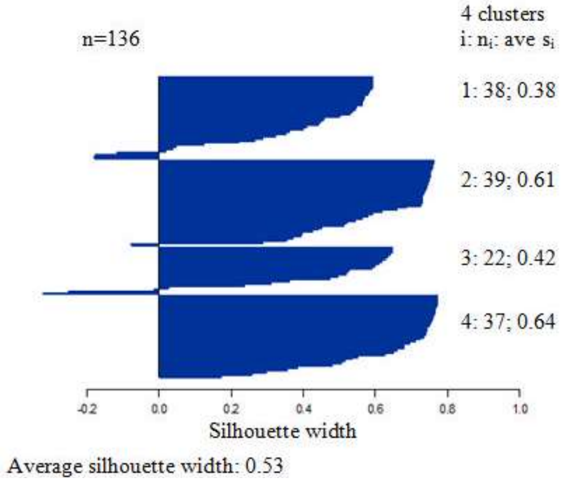

Fig. 1: Average silhouette coefficients for best classification

between 2 and 10. So, the inter-cluster variations of cows started to be observed with the determination of the cluster numbers reflecting the natural classification appearing by coming together of variations. As a rule, the suitable cluster number was determined as 4 in respects of all the separate factors.

We were not interested the number of clusters >10. Because their overall average silhouette coefficients were very low. Fuzzy clustering, in terms of milk yield characteristics; when the clustering cows are considered to their milk yield characteristics, the division coefficient indicating how well the cows are clustered and also the silhouette coefficient, k = 4 was found to be the most useful cluster number. Overall silhouette width values s(i) which signals the quality of clustering were calculated for k = 4.

These values are big, which indicates a good clustering has been formed and s(i) averages are big which indicates a good and certain has been obtained. We can claim that the cows in the 4th cluster are better clustered than those in 1st, 2nd and 3rd cluster. According to the cluster number k = 4 the first cluster obtained cannot be said to be clustered very well. The results are shown in Table 1 and Fig. 1.

We were carried out the fuzzy clustering analysis using by six milk yield characteristics. The best option given by the silhouette plot in the fanny algorithm (Fig. 1) had four clusters with 38, 39, 22 and 37 cows, respectively.

Table 2 shows Dunn and Normalized Dunn coefficient. The normalized value of the coefficient for our clustering with four clusters is 0.4033 which is very much a fuzzy clustering. Descriptive statistics of some milk yield traits for best classification are shown in Table 3. The difference between the clusters was tested with ANOVA.

The four clusters differed ($p<0.05$) for the milk yield characteristics. Duncan Multiple Range test was used to separate the means among the clusters. According to Table 3, the 3rd cluster had the highest average number of lactation. In terms of 305 day milk yield and age of the first insemination, 1st cluster had the highest value. Range of average dry period is between 60.13 days and 95.63 days. According to the values in Table 3 in the cow breeding studies, Silhouette Coefficients different cluster numbers can be used as a helping criterion in genetic improvement (selection) studies.

CONCLUSION

Clustering of cow with respect to their milk yield parameters helps to increase the precision of genetic parameter estimates and also facilitates borderless evaluations. Fuzzy clustering analysis is not an improvement method but it saves time, labor, cost and reliable and objective method due to be advantageous in terms of decision-making at the point of selection studies as a supportive method. Since fuzzy clustering is a specific form of cluster analysis, it can more easily distinguish ambiguity in the dataset that lies between clusters Alam *et al.* (2000). Thus, we have additional information about the structure of the data. The contribution is to the science of biometry was that the use of fuzzy logic in animal selection in a first time in dairy cows to classify them in a fuzzy way. Consequently, this method allows for flexibility when interpreting clusters and provides interesting nuances in data evaluation and commencing on research findings in different way, as we have shown.

ACKNOWLEDGEMENT

I thank Agr. Eng. M. Selim OKYAY and Vet. Cuma MERCAN. This work has been supported in part by general directorate of agricultural enterprises (TIGEM).

REFERENCES

Alam, P., D. Booth, K. Lee and T. Thordarson, 2000. The use of fuzzy clustering and self organizing neural networks for identifying potentially failing banks: An experimental study. Expert Syst. Appl., 18: 185-199.

Chang, Y.C. and B. Chang, 2003. Applying fuzzy cluster method for marine environmental monitoring data analysis. Environ. Inform. Arch., 1: 114-124.

Diaz, B., L. Moniche and A. Morillas, 2006. A fuzzy clustering approach to key sectors of the Spanish economy. Econ. Syst. Res., 18: 299-318.

Halavati, R. and S.B. Shouraki, 2005. Fuzzy learning in zamin artificial world. Fuzzy Sets Syst., 152: 603-615.

Kaufman, L. and P.J. Rousseeuw, 1990. Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley and Sons, New York.

Lenard, M.J., P. Alam and D. Booth, 2000. An analysis of fuzzy clustering and a hybrid model for the auditor's going concern assessment. Decision Sci., 31: 861-884.

Mocz, G., 1995. Fuzzy cluster analysis of simple physicochemical properties of aminoacids for recognizing secondary structure in proteins. Protein Sci., 4: 1178-1187.

Rousseeuw, P.J., 1995. Discussion: Fuzzy clustering at the intersection. Technometrics, 37: 283-285.

Sato, M. and L.C. Jain, 2006. Innovations in Fuzzy Clustering. Springer-Verlag, Netherlands.

Tufan, E. and B. Hamarat, 2003. Clustering of financial ratios of the quoted companies through fuzzy logic method. J. Naval Sci. Eng., 1: 123-140.