Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 69(3)2015 307-314

Principal Component based Fuzzy *c*-means Algorithm for Clustering Lentil Germplasm

Chiranjit Mazumder[1], Girish K. Jha[2], Rajender Parsad[1], Anshu Bharadwaj[1] and Jyoti Kumari[3]
[1]*ICAR-Indian Agricultural Statistics Research Institute, New Delhi*
[2]*ICAR-Indian Agricultural Research Institute, New Delhi*
[3]*ICAR-National Bureau of Plant Genetic Resources, New Delhi*

Received 07 July 2014; Revised 23 September 2015; Accepted 31 December 2015

SUMMARY

Cluster analysis is used extensively to organize data into groups based on similarities among the individual data items, leading to a crisp or fuzzy partition of sample space. Fuzzy *c*-means (FCM) is a clustering algorithm which all owsone data point to be long to two or more clusters. In this paper, principal component based fuzzy *c*-means clustering is applied for classifying 518 lentil genotypes based on their numeric agronomic and morphological traits. The appropriate number of clusters is obtained with the help of validity measures. Results of the study revealed that the genetic divergence is not highly related to geographical origins as exotic and indigenous lentil genotypes are distributed in all the four clusters.

Keywords: FCM algorithm , Fuzzy clustering, Lentil and Validity measures.

1. INTRODUCTION

Collections of plant genetic resources in gene banks are currently facing problems caused by the large size of collections, and the resultant costs of the irmaintenance (Franco *et al*., 2006). The large size of the germplasm collections of ten has hindered their evaluation and utilisation for specific breeding purposes. To solve these problems, Frankel (1984) and Brown (1989) proposed the establishment of a core collection which involves these lection of a sub set from the whole germplasm by certain methods of classification in order to capture the maximum genetic diversity of the whole collection while minimizing accessions and redundancy. The success of the development of core collection mainly depends on the optimal classification strategies of whole collection. Clustering is frequently used by the plant breeders for grouping germplasm collections into a few homogeneous groups.

Cluster analysis, an unsupervised pattern recognition technique, includes methods and algorithms for grouping objects according to measured or perceived intrinsic characteristics or similarity (Jain, 2010). There are two major categories of cluster analysis, hierarchical and non-hierarchical. Hierarchical methods cluster sequentially, by starting with the most similar pair of objects and forming higher clusters step by step. Thenon-hierarchical cluster analysis methods evaluate over all distributions of object pairs, so an initial number of clusters are assumed. A characteristic of all these classical clustering techniques is that the boundary between clusters is fully defined, that means, each object or pattern belongs to exactly one cluster. All these classical clustering techniques

Corresponding author: Girish K Jha
E-mail address: girish.stat@gmail.com

perform better and give good approximation for well characterised datasets with compact and well separated clusters. In practice, however it is observed that the boundaries between clusters are not always clearly defined. In many real-world clustering problems, some objects partially belong to multiple clusters, rather than to a single cluster exclusively. Such overlapping clusters motivated the development of fuzzy clustering algorithm. In fuzzy clustering, an object or entity is allowed to belong to many clusters with different degrees of membership (Yen and Langari, 2006). A type of fuzzy clustering is known as fuzzy *c*-means (FCM) algorithm (Bezdek, 1981).

Literature suggests that the performance of FCM algorithm is superior to K-means algorithm, Self Organization Map (a neural net work based algorithm) in the presence of correlated variables, overlapping clusters and outlier (Mingoti and Lima, 2006). The FCM algorithm and its various extensions have been used very successfully in many agricultural applications including clustering of chickpea genotypes (Khazaei *et al*., 2008; Huang *et al*., 2010). For genotype clustering, researchers are confronted with a large number of variables and some of them may be correlated. Researchers wish to reduce this large number of variables to a small numbers of uncorrelated features with as little loss of information as possible. Traditionally, principal component analysis (PCA) is considered to be an appropriate method of feature extraction for mapping the data in to a space of lower dimensionality.

In this paper, principal component based fuzzy *c*-means clustering is applied for classifying 518 lentil genotypes based on their numeric agronomic and morphological traits. This paper focuses on classifying lentil (*Lens culinaris)* germplasm collections maintained at national level and illustrates a new unexplored approach for the efficient classification of germplasm that can be used for the development of core collection. We begin with materials and methods in Section 2. Results and discussion are provided in Section 3.

2. MATERIALS AND METHODS

The experimental data for this study comprises 518 lentil accessions, of which 206 entries are exotic collections and 312 are indigenous collections including 59 breeding lines. These accessions were grown in augmented block design with five checks during *rabi* season, 2006–07 at ICAR-Indian Institute of Pulses Research, Kanpur. Accessions were evaluated for 19 descriptors, including plant characteristics and seed characteristics following the biodiversity and national Distinctness, Uniformity and Stability (DUS) descriptors guidelines. Out of these 19 descriptors, 4 were binary in nature, 5 were ordinal types and 10 were numeric in nature.

In this study, only numerical descriptors have been used for clustering as qualitative descriptors were found non-informative in distinguishing between the accessions (Jha *et al*., 2014). Hence, the data matrix X consisted of 518 rows and 10 columns. Each observation representing lentil genotype consist of 10 measured variables, grouped into 10 dimensional row vector $x_k=\{x_{k1}, x_{k2},\ldots,x_{k10}\}^T$, $x_k \in R^{10}$. A set of 518 observations is denoted by $\mathbf{X}= \{x_k/k=1,2,\ldots,518\}$and is represented as 518×10matrix:

$$\mathbf{X} = \begin{bmatrix} X_1X_1 & \cdots & X_1X_{10} \\ \vdots & \ddots & \vdots \\ X_{518}X_1 & \cdots & X_{518}X_10 \end{bmatrix}$$

In pattern recognition terminology, the rows of **X** are called patterns or objects and the columns are called as the features or attributes. In our data, rows of **X** represent the accessions and the columns are quantitative traits.

The fuzzy *c*-means (FCM) algorithm generalizes the hard K-means algorithm to allow a pattern or datum to partially belong to multiple clusters. FCM aims to determine cluster centers $v_i(i=1,2,\ldots,c)$ and the fuzzy partition matrix U by minimizing the objective function J defined as follows:

$$J\,(U, v_1, v_2, \ldots, v_c; X) = \sum_{i=1}^{c}\sum_{j=1}^{n}\{u_{c_i}(x_j)\}^m\, d_{ij}^2$$

subject to the condition $\sum_{i=1}^{c} u_{c_i}(x_j) = 1, j = 1, 2, \ldots, n,$ where n is the sample number, $u_{c_i}(x_j)$ is the degree of membership of object x_j to the cluster i, m is the fuzzy exponent that determine the degree of fuzziness of the final partition, or in other words the degree of overlap between groups, d_{ij}^2 is the squared distance between the vector of observations of object *j* to the vector of *c* cluster centers. Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of the membership $u_{c_i}(x_j)$ and the cluster center v_i by,

$$u_{c_i}(x_j) = \frac{1}{\sum_{i,k=1}^{c} \left(\frac{d_{ij}}{d_{kj}}\right)^{\frac{2}{(m-1)}}} (m \neq 1) \text{and} v_i = \frac{\sum_{j=1}^{n} ((u_{c_i}(x_j))^m x_j}{\sum_{j=1}^{n} ((u_{c_i}(x_j))^m}$$

Like hard K-means algorithm, the desired number of cluster *c* has to be predefined and *c* initial seeds of the cluster are required to perform the FCM (Bezdek, 1981). These seeds are modified in each stage of the algorithm and for each object a degree of membership to each *c* clusters are estimated. A metric isused to compare every object to the cluster seed but the comparison is made using a weighted average that takes into account the degree of membership of the objects to each cluster. In the end of the algorithm, a list of the estimated degree of membership of the objects to each of the *c* clusters is printed. The objects are assigned to the cluster for which the degree of membership is higher. Contrary to the K-means method FCM is more flexible because its shows those objects that have some interface with more than one cluster in the partition. The FCM clustering algorithm was implemented using the Fuzzy Logic Tool box of the MATLAB software.

In order to determine the best number of clusters for the given dataset, cluster validity measures have been used. A lot of cluster validity criteria have been proposed in the literature (Bezdek, 1974; Gunderson, 1978; Xie and Beni, 1991) which can be grouped into two important types. One is based on the fuzzy partition of sample set and the other is on the geometric structure of sample set. Table 1 lists four cluster validity functions which have been used in our study to determine the appropriate number of clusters. The partition coefficient (V_{PC}) and partition entropy (V_{PE}) are membership based validity measures. Some empirical studies have shown that maximizing (V_{PC}) (or minimizing (V_{PE}) often leads to a good interpretation of the sample data (Pal and Bezdek 1995). The major drawbacks of these two measures are that they do not consider geometrical properties such as the degree of separation between clusters and their monotonic tendency with *c*. These limitations motivated the development of the second type of validity measures such as Gunderson's separation coefficient (V_{SC}) and Xie-Beni function (V_{XB}) which uses both the dataset and the prototypes (cluster centers).

Table 1. Four Validity Functions for the Fuzzy *c*-means

Validity Index	Functional Description	Optimal Partition
Partition coefficient	$V_{PC} = \frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n}\{u_{c_i}(x_j)\}^2$	Max
Partition Entropy	$V_{PE} = -\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n}\{u_{c_i}(x_j)\log u_{c_i}(x_j)\}$	Min
Separation coefficient	$V_{SC} = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n}\{u_{c_i}(x_j)\}^m \|x_j - v_i\|^2}{n_{\min(i,k)}\|v_k - v_i\|^2}$	Min
Xie and Beni's function	$V_{XB} = \left(\frac{\sum_{i=1}^{c}\left(\sum_{j=1}^{n}\{u_{c_i}(x_j)\}^2 \|x_j - v_i\|^2\right)}{n\left(min_{i\neq j}\|x_j - v_i\|^2\right)}\right)$	Min

3. RESULTS AND DISCUSSION

The characteristic features, that is, mean, coefficient of variation (CV), maximum and minimum value of sample data used for analysis is presented in Table 2. High coefficient of variation was observed in characters like pods / plant (40.96%) and yield / plant (43.29%). This indicated that there is a large amount of diversity among lentil genotype accessions with regard to these traits. For other traits except days to maturity, coefficient of variation was recorded in the range of 14–28%. Low values of coefficient of variation limit the scope of selection for those traits. A very low value of CV was observed in case of days to maturity because high temperature during late reproductive phase forces simultaneous maturity for most of the genotypes. As indicated earlier, principal component analysis has been used for feature extraction. The first step in PCA consists of testing whether the variables show a sufficient level of correlation. To this end, the correlation matrix for sample data has been analyzed and is given in Table 3. Significant correlation at the 0.05 and 0.01 levels have been made bold in the Table. Table clearly revealed that most of the numerical traits are highly linearly related. Yield per plant had significant and high positive association with pods per plant (r=0.807), biological yield per plant (r=0.681), secondary branches (r=0.511), primary branches (r=0.379), while it is negatively correlated with days to flowering and days to maturity. Similar kind of findings was reported by other researchers also (Hegazy *et al*. 2012; Ramgiry *et al*., 1989). Seed weight was negatively correlated with most of the characters except biological yield per plant and seed yield per plant which was low positively significant. This is contradictory to some earlier findings that might be due to low coefficient of variance in seed weight of studied germplasm.

Table 2. Descriptive Statistics for 10 Numerical Characters in 518 Lentil Accessions

Character	Code	Unit	Mean	Minimum	Maximum	CV(%)
Days to 50% flowering	DF	d	78.69	58.00	106.00	13.67
Plant height	PH	cm	30.79	17.00	47.60	15.57
Days to maturity	DM	d	126.03	114.00	140.00	3.73
100-seed weight	SW	g	2.43	1.20	4.10	21.77
Biological yield plant^{-1}	BYP	g/plant	13.37	4.20	28.00	28.03
Primary branch plant^{-1}	PB	no.	3.76	2.00	9.00	28.22
Secondary branch plant^{-1}	SB	no.	10.22	4.00	18.00	23.42
Pods plant^{-1}	PPP	no.	116.16	3.70	309.30	40.96
Yield plant^{-1}	YPP	g/plant	3.72	0.20	10.70	43.29
Plant height at lowest pod	PHLP	cm	10.74	1.00	19.00	21.46

Table 3. Correlation between 10 Numerical Descriptors in 518 Lentil Genotypes

Character	DF	PH	DM	SW	BYP	PB	SB	PPP	YPP	PHLP
DF	1	.208	.479	-.519	-.037	.080	.272	.143	-.018	.343
PH	.208	1	.164	-.100	.163	.059	.189	.224	.188	.327
DM	.479	.164	1	-.212	-.032	.016	.155	.123	-.011	.270
SW	-.519	-.100	-.212	1	.003	-.014	-.188	.151	.119	-.221
BYP	-.037	.163	-.032	.003	1	.503	.608	.680	.681	.057
PB	.080	.059	.016	-.014	.503	1	.742	.398	.379	.057
SB	.272	.189	.155	-.188	.608	.742	1	.567	.511	.338
PPP	.143	.224	.123	-.151	.680	.398	.567	1	.807	.212
YPP	-.018	.188	-.011	.119	.681	.379	.511	.807	1	.153
PHLP	.343	.327	.270	-.221	.057	.057	.338	.212	.153	1

Note: Significant correlations are bold.

1.1 Feature Extraction

Feature extraction is one of the important steps in clustering techniques. Linear or non-linear combinations of the original variables are called features and the process of generating the miscalled feature extraction. In this study, principal component analysis has been used as a method of feature extraction to obtain the new variables for the fuzzy clustering. PCA reduces a large number of variables to a much smaller number of uncorrelated linear combination of original variables, called principal components or features. The PCA also avoids the effect of colinearity among descriptors, which is present in the sample data as evident from Table 3.

Quantitative traits of lentil accessions data are of heterogeneous type. Some of them are numbers (primary branch per plant, pods per plant, etc…), others are weights (100-seed weight, biological yield per plant, etc…), or lengths (plant height, plant height at lowest pod), or duration (days to 50% flowering, days to maturity, etc.). In addition quantitative traits such as plant height, seeds per plant and 100-seed weight are of different order of magnitude. So normalization of each variable is required before applying principal components analysis. Values of each variable was normalized between [0, 1] by dividing with its maximum value. The normalized data was then used to calculate the correlation coefficient matrix. In general, principal component with eigen value greater than one are retained. This criterion is based on the fact that a component having an eigen value higher than unity contains more information than one original variable, so that the principle of dimensionality reduction is ensured. In this case, eigen value for first three principal components were more than one and able to explain 67.35% variability present in the dataset. Hence, first three principal components were used as the inputs for the fuzzy clustering algorithm to get the appropriate number of clusters. The principal component loadings represent the correlation of the variables with an axis and loadings of the retained principal components were presented in Table 4. The first component accounting for 35.76% of the total variance showed high positive loading (> 0.80) on secondary branch and pods per plant where as moderate loading (> 0.60) on primary branch, biological yield and yield per plant. The second component accounting for 21.03% of the total variance showed moderate positive loading (> 0.60) on days to maturity and days to 50% flowering, moderate negative loading on seed weight, whereas the third component capturing 10.56% of the variance, was positively correlated with plant height.

Table 4. Principal Component Loadings of Variables from Correlation Matrix for the Dataset

Variable	Component		
	1	2	3
Days to 50% flowering	0.30	0.78	-0.20
Plant height	0.35	0.30	0.66
Days to maturity	0.22	0.63	0.01
100-seed weight	-0.21	-0.64	0.36
Biological yield plant^{-1}	0.79	-0.37	-0.01
Primary branch plant^{-1}	0.67	-0.21	-0.43
Secondary branch plant^{-1}	0.86	0.02	-0.26
Pods plant^{-1}	0.85	-0.14	0.12
Yield plant^{-1}	0.78	-0.36	0.24
Plant height at lowest pod	0.39	0.51	0.36
Variance accounted for(%)	35.76	21.03	10.56

2.1 Parameters of the FCM Algorithm

For the implementation of FCM algorithm, the following parameters must be specified by the experimenter: the number of clusters *c*, the fuzziness exponent, *m*, the termination tolerance, and the norm-inducing matrix, ***A*.** Moreover, the fuzzy partition matrix, *U*, need to be initialized. The fuzziness exponent *m* is an important parameter because it influences the fuzziness of the resulting partition. For *m* equal to one, the partition becomes hard and as *m* approaches to infinity, the partition becomes completely fuzzy. In this study $m = 2$ was considered by taking clue from the literature (Wag *et al*., 2004). The FCM algorithm stops iterating when the norm of the difference between $\boldsymbol{U}$ in two successive iterations is smaller than the termination parameter ϵ. Here, $\epsilon = 0.001$was used for analysis. The shape of the clusters is determined by the choice

of the matrix **A** in the distance measure. The common choice of identity matrix for **A** provides the standard Euclidean norm which induces hyper spherical clusters.

The number of clusters *c* is the most important parameter in the sense that the remaining parameters have less influence on the resulting partition. However, *c* is rarely known *apriori* while clustering real data due to the lack of knowledge about the underlying structures in the data. Hence, one needs to make assumption about the number of underlying clusters. Two main approaches, one is based on validity measures and other is based on iterative merging or insertion of clusters, are adopted for determining the appropriate number of clusters in the data. In this study, we searched the optimal number of clusters using validity measures. The upper bound for the number of clusters, c, was estimated 12 with the help of lentil accessions data and validity measures. Then all the four validity measures were run with fuzzy *c*- means algorithm for each $c \in \{2, 3,\ldots, c_{max}\}$ Values of the validity measures depending on the number of cluster are presented in Table 5.

It is worth mentioning that no single validation index is reliable by itself, that is why four indices were used in the analysis and the optimum number of clusters was obtained with the comparison of their results. We considered that partition with less clusters are better, when the differences between the values of a validation index are minor. On the basis of PC and PE, the number of clusters can be 5 while SC hardly decrease at $c = 4$ (Table 5). Considering that SC is more useful, when comparing different clustering methods with the same *c*, we chose the optimal number of clusters to 4, which was confirmed by the Xie and Beni's index too. It may be noted that the optimal value of the four indexes is not attained simultaneously for our dataset. Therefore, a fused result for the four indexes has been used to determine the best number of clusters. After deciding the number of clusters, the FCM algorithm was implemented for grouping the lentil accessions.

Table 5. Values of Validity Measures using Fuzzy *C*-means Algorithm for different Cluster Numbers

c	2	3	4	5	6	7	8	9	10	11	12
PC	0.7468	0.6262	0.5521	0.5066	0.4845	0.4743	0.4685	0.4582	0.4552	0.4498	0.4428
PE	0.4021	0.6565	0.8462	0.9880	1.0764	1.1427	1.1944	1.2547	1.2863	1.3262	1.3619
SC	0.0066	0.0058	0.0053	0.0052	0.0050	0.0049	0.0044	0.0044	0.0036	0.0040	0.0033
XB	12.9651	11.9134	10.2186	10.0489	9.8191	8.6679	6.1124	5.4716	3.6932	4.1779	3.9550

At this juncture, it is worth mentioning that the result of applying FCM algorithm to a given dataset depends not only on the choice of parameters *m* and *c*, but also on the choice of initial prototypes. Accordingly, we used multiple retries, with different initial prototypes in order to avoid local minima and find the global minimum for the objective function. The 518 feature vectors used in this study we regrouped into four clusters based on the FCM methods. In figure 1, the first two principal components cores were plotted with cluster centers and membership contours overlaid to aid visualization of group differences. Fig. 1 clearly showed that there are four clusters based on their centers' Euclidian distance from each other. It is evident from the figure 1 that all four clusters are clearly separated in our analysis. Thus, we are able to define groups of genotypes that are significantly different from each other for character of interest. After defuzzyfying the results, cluster I consisted of 61 genotypes, cluster II of 111, cluster III of 165 and cluster IV of 181 genotypes. The mean value along with the standard deviation for each cluster is shown in Table 6.

Table 6 revealed that genotypes in cluster I were late in maturity with the highest pod/ plant and the highest yield/ plant, while genotypes grouped in cluster II were early in maturity and had the lowest pod / plant, the lowest yielding but the highest seed-weight (bold seeded). Genotypes in cluster III showed late in maturity, the lowest

seed weight and had medium yielding. The information on clusters with particular genotypes and interesting traits will assist in looking extensively for more genotypes with similar traits (Upadhyaya *et al*., 2001). Genetic relationship among the germplasm revealed by clustering can guide crop researchers to choose diverse parent for crossing programme to get superior recombinants. It was observed from the results of the study that genetic divergence is not highly related to geographical origins as exotic and indigenous lentil genotypes were present in all the four clusters. These groups can be used for the development of core collection, small number of germplasm accessions with maximal diversity, which is an important research area for crop improvement.

Table 6. Mean and Standard Deviation of 10 Characters for four Clusters

Characters	Cluster I	Cluster II	Cluster III	Cluster IV
No. of Genotypes	**61**	**111**	**165**	**181**
Days to 50% flowering	79.16±10.40	74.83±12.66	80.65±9.19	78.96±10.36
Plant height	31.38±3.93	29.96±5.09	31.52±4.25	31.63±4.80
Days to maturity	126.36±4.75	124.77±5.57	126.82±3.97	125.98±4.59
100-seed weight	2.42±0.55	2.65±0.56	2.32±0.48	2.41±0.51
Biological yield plant^{-1}	17.98±3.74	9.97±2.96	14.92±2.68	12.51±2.56
Primary branch plant^{-1}	4.58±0.99	3.12±0.89	4.03±1.09	3.62±0.90
Secondary branch plant^{-1}	12.57±2.21	8.18±2.23	11.08±1.90	9.85±1.83
Pods plant^{-1}	202.15±32.40	54.89±16.72	141.98±13.60	101.06±12.85
Yield plant^{-1}	6.10±1.51	2.05±1.14	4.46±1.08	3.27±0.69
Plant height at lowest pod	11.28±2.36	9.74±2.09	10.96±2.31	10.98±2.24

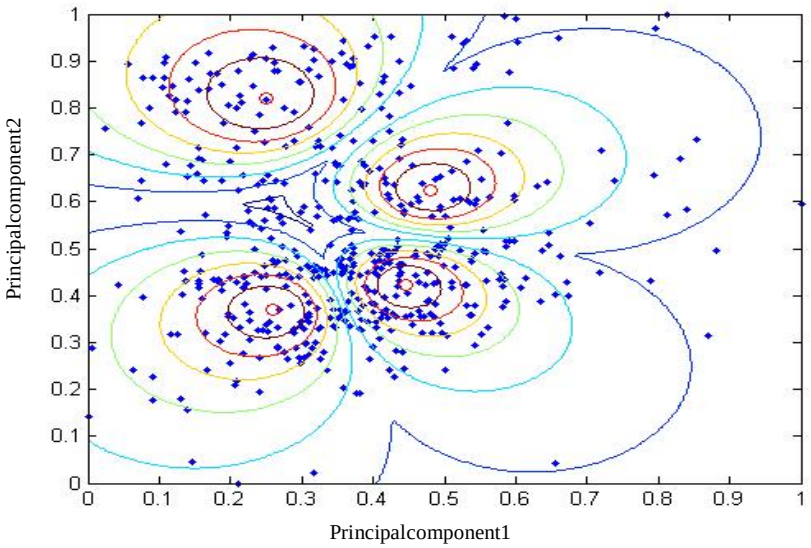

Fig. 1. Clustering by PCA based Fuzzy *c*- Means Algorithm of Lentil Genotype Data

4. CONCLUSIONS

Fuzzy set theory provides useful concepts and methods to deal with vagueness and imprecision, which is a characteristic feature of agricultural data. In this study, 518 lentil accessions have been grouped into four clusters on the basis of 10 numerical agro-morphological descriptors using principal component based fuzzy *c*-means algorithm. The principal component analysis was used for dimensionality reduction and also avoided the ill-effects of collinear descriptors. The appropriate number of clusters was obtained with the help of validity measures. Results of the study revealed that genetic divergence is not highly related to geographical origins as exotic and indigenous lentil genotypes were distributed in all the four clusters. However the clustering pattern will provide the knowledge of genetic

relationship among the accessions and help select diverse accessions for crossing programme and genetic enhancement to improve the yield potential of crop through breeding for superior recombinants. It can be concluded from this study that the principal component based fuzzy clustering techniques were successfully applied to classify lentil genotypes in terms of agronomic traits. Since the proposed technique is not data specific as the descriptors for germplasm accession of any crop consists of quantitative descriptors, these techniques can be utilized for other agricultural crops. The success of the development of most representative core collection is mainly depends on the reliable grouping of whole collection. Hence, fuzzy clustering has a promising potential in agriculture as a tool to evaluate, understand, predict, and manage crop production.

REFERENCES

Bezdek, J.C. (1974). Cluster validity with fuzzy sets. *Cybernatics* **3**, 58–73.

Bezdek, J.C.(1981). Pattern recognition with fuzzy objective *Function Algorithms*. Plenum Press, NewYork.

Brown, A.H.D. (1989). The case for core collection. *Genome*, **31**, 818–824.

Franco, J., J. Crossa, M.L. Warburton and S. Taba. (2006). Sampling strategy for conserving maize diversity when forming core sub sets using genetic markers. *Crop Science* **46**, 854–864.

Frankel, O.H. (1984). Genetic perspectives of germplasm conservation. In Genetic Manipulation: Impact on Man and Society. Edited by Arber W.K., Llimensee K, Peacock W.J., Starlinger P. Cambridge: Cambridge University Press, 161–170.

Gunderson, R. (1978). Application of fuzzy ISODATA algorithms to start-tracker pointing system. In *Proc. of 7th Triannual World IFAC Congr.*, Helsinki, 1319–1323.

Hegazy, S.R.E., Selim, T. and E.A.A. El-Emam. (2012). Correlation and path coefficient analyses of yield and some yield components in lentil. *Egypt Journal of Plant Breeding* **16(3)**, 147–159.

Huang, Y., Lan, Y., Thomson, S.J., Fang, A, Hoffmann, W.C. and Lacey R.E. (2010). Development of soft computing and applications in agricultural and biological engineering. *Computers and Electronics in Agriculture* **71**, 107–127.

Jain, A.K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters* **31**, 651–666.

Jha, G.K., Mazumder, C., Kumari, J. and Singh, G. (2014). Non linear principal component based fuzzy clustering: A case study of lentil genotypes, *Indian Journal of Genetics and Plant Breeding*, **74(2)**, 189–196.

Khazaei, J., Naghavi, M.R., Jahansouz, M.R., and Salimi-Khorshidi, G. (2008). Yield estimation and clustering of chickpea genotypes using soft computing techniques. *Agron. J.*, **100**, 1077–1087.

Mingoti, S.A. and Lima, J. (2006). Comparing SOM neural network with Fuzzy *c*-means, K-means and traditional hierarchical clustering algorithms. *Europ. J. Operat. Res.* **174,** 1742–1759.

Pal, N.R. and Bezdek, J.C.(1995).On cluster validity for the fuzzy *c*-means model. *IEEE Trans. Fuzzy Sys.* **3**, 370–379.

Ramgiry, S.R., Paliwal, S., Tomar, S.K. (1989). Variability and correlation of yield and other quantitative characters in Lentil. *Lens Newsletter* **16(1)**,19-21.

Upadhyaya, H.D., Brume, P.J. and Singh, S. (2001). Development of chickpea core subset using geographical distribution and qualitative traits. *Crop Sci*, **41,** 206–210.

Wang, X., Wang Y. and Wang L. (2004). Improving fuzzy *c*-means clustering based on feature-weight learning. *Pattern Recog. Lett.*, **25**, 1123–1132.

Xie, X.L. and Beni, G.A. (1991). Validity measures for the fuzzy clustering. *IEEE Trans. Pattern Anal. Machine Intel.* **3(8**), 841–846.

Yen, J. and Langari, R. (2006). Fuzzy logic: Intelligence, control and information. Pearson Education.