Indian J. Genet., 74(2): 189-196 (2014)
DOI: 10.5958/0975-6906.2014.00155.2

Nonlinear principal component based fuzzy clustering: A case study of lentil genotypes

Girish K. Jha*, Chiranjit Mazumder, Jyoti Kumari[1] and Gajab Singh

Indian Agricultural Research Institute, New Delhi 110 012; [1]National Bureau of Plant Genetic Resources, New Delhi 110 012

(Received: October 2013; Revised: February 2014; Accepted: March 2014)

Abstract

Cluster analysis is frequently used by the plant breeders in grouping germplasm collections into a few homogeneous groups in order to identify accessions with specific property of potential relevance for their plant improvement programs. The set of descriptors for the germplasm accessions consists of both numerical and categorical descriptors. In such situations, the standard principal component analysis will not be appropriate for feature extraction of data using all descriptors because it deals with only numeric variables. In this paper, nonlinear principal component analysis was used to analyse the descriptors of lentil accessions which can handle mixture of measurement types. The first two nonlinear principal components were used as input to fuzzy *c*-means algorithm in grouping 518 lentil genotypes into four clusters based on their agronomic and morphological traits.The study demonstrated that the proposed nonlinear principal component based fuzzy clustering has a promising potential in agriculture as a tool for evaluation and efficient grouping of germplasm collections.

Key words: FCM algorithm, Mixture of data types, Nonlinear principal component analysis, Lentil and Validity measures

Introduction

Cluster analysis, an unsupervised pattern recognition technique, includes methods and algorithms for grouping objects according to measured or perceived intrinsic characteristics or similarity [1]. There are two main approaches to clustering. One approach is crisp (or hard) clustering and the other one is fuzzy clustering [2]. A characteristic of crisp clustering is that the boundary between clusters is fully defined. In contrast, the boundaries between clusters generated by a fuzzy clustering algorithm are vague. This means that each pattern or object data of a fuzzy partition belongs to different classes with membership degrees between 0 and 1 indicating their partial membership. In many real world applications, fuzzy clustering is more natural than hard clustering. Fuzzy *c*-means (FCM) algorithm proposed by Dunn [3] and extended by Bezdek [4] is one of the most well-known methodologies in clustering analysis. Literature suggests that the performance of FCM algorithm is superior to K-means algorithm, Self Organization Map (a neural network based algorithm) in the presence of correlated variables, overlapping clusters and outliers [5].

Traditionally, principal components analysis (PCA) is used for dimensionality reduction or feature extraction for successful implementation of cluster analysis. However, PCA assumes linear relationships between variables and its interpretation is only sensible if all of the variables are assumed to be scaled at the numeric level. In the agricultural sciences, these assumptions are frequently not justified, and therefore, PCA may not always be the most appropriate method of analysis. For instance, the set of descriptors for the accessions of most germplasm collections consists of both numerical and categorical descriptors. To circumvent these limitations, an alternative, referred as nonlinear principal components analysis has been developed [6]. This alternative method has the same objectives as traditional principal components analysis, but is suitable for variables of mixed measurement levels, which may not be linearly related to each other. At this juncture, it is worth noting that this novel and useful technique has been rarely used in the field of plant breeding. We came across a single such study

*Corresponding author's e-mail: girish.stat@gmail.com
Published by the Indian Society of Genetics & Plant Breeding, F2, First Floor, NASC Complex, PB#11312, IARI, New Delhi 110 012
Online management by indianjournals.com

by Kroonenberg *et al*. [7] during review of literature. They used nonlinear principal component analysis to analyze the categorical and numerical descriptors of Australian groundnut accessions.

This paper focuses on grouping lentil (*Lens culinaris)* germplasm collections based on its phenotypic variability. Germplasm collections contain large numbers of accessions on which several quantitative and qualitative characteristics are measured. Collections of germplasm in gene banks are often so large that their size interferes with achieving the main goals for which the collections have been established, namely, the conservation and utilization of the genetic diversity of a crop species and its relatives [8]. To solve these problems, Frankel [9] and Brown [10] proposed the establishment of a core collection which involves the selection of a subset 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.

The FCM algorithm for classification has been used very successfully in many agricultural applications including clustering of chickpea genotypes [11-12]. However, no effort has been made to combine nonlinear PCA with FCM to take advantage of the unique strength of both these techniques for clustering. In this paper, an effort has been made to use first two nonlinear principal components, based on categorical and numerical descriptors, as input to FCM algorithm in grouping 518 lentil genotypes and devise a new unexplored approach for the classification of germplasm. The proposed technique will provide not only an insight into the relationships between the descriptors, but also enable to estimate variability in the germplasm and genetic relationship between accessions.

Methods and material

Nonlinear principal component analysis

Nonlinear principal component analysis is an extension of standard principal component analysis that can handle variables measured on nominal, ordinal and ratio scales of measurement. In nonlinear PCA, linear and nonlinear transformation of variables and linear combination of transformed variables are obtained using the method of alternating least squares such that the average squared correlation of the transformed variables and the linear combination is maximized. The correlation of the categorical variables with the component is achieved by assigning numeric values to the categories through a process called optimal quantification. Optimal quantification replaces the category labels with category quantification in such a way that the newly quantified variable will have as high a correlation with the first component as possible, given the other variables. In case of ordinal variable, any monotonic transformation by scoring the ordered categories so that the order is preserved may be used. With numeric variables, linear transformation is used so that equidistant observed values remain equidistant after transformation. In addition, nonoptimal transformations like logarithm and other power transformation may be used. Literature suggests that it is advantageous to group numerical variables to 6-10 categories of equal interval except for the end points [7]. It is further suggested that for balanced analyses, most categories should not have frequency less than 5. The detailed account on the history, statistical theory and the variation of the technique is described by Gifi [13]. In case, more than one component is extracted then two possibilities occurs with respect to quantification that is the same quantification for the categories of a variable for all components (called single quantification) or a separate quantification for each component (called multiple quantification). All analysis related to nonlinear principal component analysis was performed using CATPCA of SPSS and PROC PRINQUAL of SAS software.

Fuzzy c-means algorithm

The fuzzy *c*-means (FCM) algorithm generalizes the hard *K*-means algorithm to allow a pattern or datum to partially belong to multiple clusters [4]. Therefore, it produces a constrained soft partition for a given dataset. FCM aims to determine cluster centers v_i (i = 1, 2, … , c) and the fuzzy partition matrix U by minimizing the objective function J defined as follows

$$J(U,\nu_1,\nu_2,\cdots,\nu_c;X)=\sum_{i=1}^{c}\sum_{j=1}^{n}\{u_{c_i}(x_j)\}^m d_{ij}^2$$

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 > 1 is the fuzzy exponent that determine the degree of fuzziness of the final partition, d_{ij}^2 is the squared

distance between the vector of observations of object *j* to the vector of *c* cluster centers [2]. Fuzzy partitioning is carried out through an iterative optimization of the objective function *J*, with the update of the membership $u_{c_i}(x_j)$ and the cluster center v_i by,

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

$$v_i = \frac{\sum_{j=1}^{n}((u_{c_i}(x_j))^m x_j}{\sum_{j=1}^{n}((u_{c_i}(x_j))^m}$$

The FCM clustering algorithm was implemented using the fuzzy logic toolbox of the MATLAB software.

Cluster validity functions for the FCM

Cluster validity functions are often used to determine the appropriate number of clusters for a given dataset (the choice of parameter *c* for FCM). A number of cluster validation functions for the FCM have been proposed in the literature [14-16] 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 [17]. 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).

Data description

The experimental data for this study comprised of 518 lentil (*Lens culinaris*) accessions of which 206 entries were exotic collections and 312 were indigenous collections including 59 breeding lines. These accessions were grown in augmented block design with five checks during *rabi* season, 2006-07 at Indian Institutes of Pulses Research, Kanpur. Accessions were evaluated for 19 descriptors varying in measurement type following the Biodiversity International and national Distinctness, Uniformity and Stability (DUS) descriptors guidelines. Data on qualitative traits, days to 50% flowering (days from

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

sowing to appearance of 50% flowering) and days to maturity (days from sowing to 90% maturity) were recorded as single value on plot basis. On the other hand, quantitative traits like plant height, pods/plant, yield/plant, biological yield/ plant etc. were recorded from 5 plants which had been randomly chosen in each plot and arithmetic mean of these traits were used for the analysis. The 19 descriptors have been grouped into three data types namely, 4 binary, 5 ordinal and 10 numerical variables (Table 2).

Results and discussion

In this study, the descriptors for lentil germplasm included variables of different levels of measurement. Feature extraction or dimension reduction is one of the important steps in fuzzy clustering techniques. Linear or non-linear combinations of the original variables are called features and the process of generating them is called feature extraction. In this study, nonlinear PCA has been used as a method of feature extraction. As indicated earlier, binary variables were transformed by optimally scoring the categories and ordinal variables were transformed monotonically by scoring the ordered categories so that the order is preserved. For instance, the absence of stem pigmentation was assigned a value of –2.48 and the presence of stem pigmentation was assigned a value 4.03 through optimal scoring in order to have maximum correlation with component. Besides, all numerical variables were grouped into 6-10 categories in such a way that except for end points, the new categories were of equal interval and no category contained less than 5 accessions. In case of ordinal descriptors, categories were collapsed with their neighbouring categories in case they contained less than 5 accessions. Categories were collapsed to prevent rare categories unduly influencing the analysis.

Table 2: Descriptors observed from the 518 lentil germplasm accessions

Code	Name	Description
		Binary Descriptors
SP	Stem pigmentation	absent=1, present=2
TL	Tendril length	rudimentary=1, prominent=2
PP	Pod pigmentation	absent=1, present=2
CC	Cotyledon colour	yellow=1, red=2
		Ordinal Descriptors
SV	Seedling vigour	poor=1, medium=2, good=3, very good=4
LS	Leaf size	small=1, medium=2, large=3
LP	Leaf pubescence	low=1, medium=2, high=3
LC	Leaf colour	light green=1, ash green=2, green=3, dark green=4
PGH	Plant growth habit	erect=1, horizontal=2
		Numeric Descriptors
DF	Days to 50% flowering	1=58-62, 2= 63-67, 3= 68-72, 4= 73-77, 5= 78-82, 6= 83-87, 7= 88-92, 8= 93
PH	Plant height	1=17-20.3, 2= 20.4-23.7, 3= 23.8- 27.1, 4= 27.2- 30.5, 5= 30.6 - 33.9, 6= 34- 37.3, 7=37.4-40.7, 8=40.8
DM	Days to 90% maturity	1= 114-117, 2= 118-121, 3= 122-125, 4=126-129, 5=130-133 6=134
SW	100 Seed weight	1<= 1.8, 2=1.8- 2.1, 3= 2.2- 2.5, 4= 2.6- 2.9, 5= 3.0- 3.3, 6=3.4
BYP	Biological yield/plant	1= 7, 2= 7.1-10.1, 3= 10.2- 13.2, 4= 13.3- 16.3, 5= 16.4-19.4, 6= 19.5
PB	Primary branch/plant	1= 2- 2.7, 2= 2.8- 3.5, 3= 3.6- 4.3, 4= 4.4- 5.1, 5= 5.2- 5.9, 6 =6.0-6.7, 7=6.8
SB	Secondary branch/ plant	1= 4- 5.5, 2= 5.6- 7.1, 3= 7.2- 8.7, 4= 8.8- 10.3, 5= 10.4- 11.9, 6= 12.0- 13.5, 7= 13.6- 15.1, 8=15.2
PPP	Pods/plant	1= 35.0, 2= 35.1- 66.6, 3= 66.7- 98.2, 4= 98.3- 129.8, 5= 129.9- 161.4, 6= 161.5- 193.0, 7= 193.1 - 224.6, 8= 224.7- 256.2, 9 = 256.3
YPP	Yield/plant	1= 1.7, 2= 1.8- 3.4, 3= 3.5- 5.1, 4= 5.2- 6.8, 5= 6.9- 8.5, 6= 8.6
PHLP	Plant height at lowest pod	1= 4.7, 2= 4.8- 6.6, 3= 6.7- 8.5, 4=8.6-10.4, 5=10.5 - 12.3, 6= 12.4 - 14.2, 7= 14.3-16.1, 8= 16.2

The correlations between optimally quantified descriptors and components as well as the squared multiple correlations for the regression of the descriptors on the components for 518 accessions are presented in Table 3. The squared multiple correlations for the regression of the descriptors on the components (often called communalities) indicate how well the components succeed in accounting for the variability of the quantified descriptors. The proportional variance accounted for by the component is the average of the squared multiple correlations with the component. The overall proportion of variance accounted for by the two nonlinear components was 0.32 which is the average of the squared multiple correlations (Table 3). The relatively low percentage of variance accounted for can be partly attributed to the presence of descriptors with a limited number of categories and should not be taken as an indication of a lack of structure.

Table 3: Correlations between optimally quantified variables and components (loadings) for 518 lentil accessions

Descriptor	Component*		Variance accounted for (%)
	1	2	
Stem pigmentation	-.170	-.068	.034
Tendril length	-.165	-.330	.136
Pod pigmentation	.075	-.056	.009
Cotyledon colour	-.226	-.113	.064
Seedling vigour	-.080	.088	.014
Leaf size	.229	.084	.060
Leaf pubescence	-.252	-.480	.294
Leaf colour	.328	.484	.342
Plant growth habit	-.099	-.092	.018
Days to 50% flowering	.429	**.656**	.614
Plant height	.463	.305	.307
Days to 90% maturity	.293	**.501**	.337
100 seed weight	-.253	-.414	.235
Biological yield/plant	**.649**	**-.543**	.715
Primary branch/plant	**.600**	-.382	.505
Secondary branch/plant	**.807**	-.208	.695
Pods/plant	**.772**	-.282	.675
Yield/plant	**.696**	-.433	.672
Plant Height at lowest pod	.488	.421	.415
Variance accounted for (%)	19.215	13.109	32.324

*****Values larger than 0.50 are set in bold

Table 3 clearly revealed that the descriptors like biological yield/plant, yield/plant, primary branches/plant, secondary branches/plant, days to 50% flowering, days to maturity are important in distinguishing between the accessions while stem pigmentation, pod pigmentation, plant growth habit, cotyledon colour and seedling vigour are not. It can be noted from the table that none of the binary and ordinal quantified descriptors had correlations value more than 0.5 with components. However, it is worth reporting that quantified ordinal descriptors namely leaf pubescence and leaf colour had correlations value 0.48 with components.

In order to get a better understanding of the relationships among variables, biplot of component loadings has been presented in Figure 1, which, clearly showed that descriptors like biological yield/plant, yield/plant, primary branches/plant, secondary branches/plant and pods/plant are highly correlated to each other as their arrows all point in the same direction. Similarly descriptors like plant height, plant height at lowest pod, days to 50% flowering, days to maturity and leaf colour are also correlated with each other. Figure 1 also indicated that days to 50% flowering and 100-seed weight is negatively correlated. Furthermore, the lengths of the arrows of the descriptors generally indicate the importance of the descriptors for distinction between the accessions. Descriptors like seedling vigour, leaf size, plant growth habit etc. with their small arrow are not important while yield/plant, days to 50% flowering and seeds weight are important. The traits

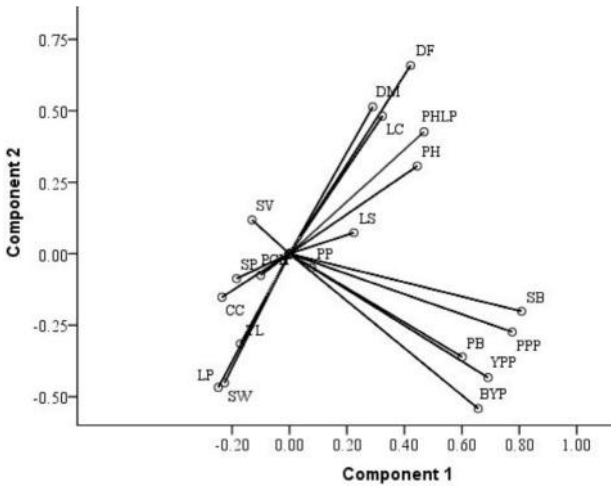

Fig. 1. Plot of the component loadings for 19 descriptors along the 1st and 2nd nonlinear principal component vectors based on 518 lentil accessions

with large arrow are quantitative traits which has got major role in depiction of genetic diversity. Similar kind of results was obtained by Li *et al.* [18] in estimation of genetic diversity while developing core collection in rice.

The biplot of accession scores is depicted in Fig. 2. In Fig. 2, the majority of the accessions (about 470) roughly form an ellipse with the major axis running from north-east to south-west direction, with increased saturation indicating large number of overlapping accessions. Most of the indigenous genotypes/land races (microsperma types) overlapped together due to their narrow genetic base. Introgression of useful genes from macrosperma germplasm of the Mediterranean region, which is the center of origin and diversity, has been suggested to widen the genetic base of lentil in South Asia [19]. In this study, genetic stocks developed by crossing with exotic material were found as outlying accessions in Fig. 2 along with exotic lines.

Further, first two dimensions of nonlinear component analysis were extracted for using as input

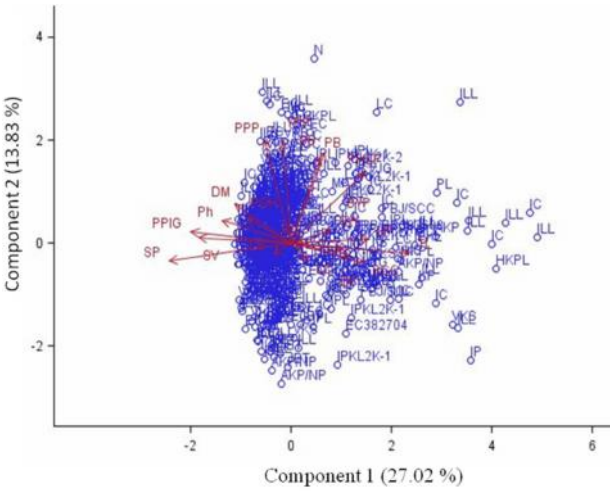

Fig. 2. Plot of accession scores along the 1st and 2nd nonlinear principal component vectors for 518 lentil accessions

for fuzzy clustering. 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 (ε), m = 2 was considered by taking clue from the literature [5]. A value of 0.001 was used for the stopping criterion (ε). Euclidean distance was chosen as the distance measure, which gives an equal weight to both the components. Moreover, the fuzzy partition matrix (U) need to be initialized.

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. In this study, we searched the optimal number of clusters using validity measures. The upper bound for the number of clusters (c_{max}) 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 4. It is worth mentioning that no validation measure is reliable by itself, that is why all the four measures were used in the analysis and the optimum number of clusters was obtained with the comparison of all the results. It may be noted that the optimal value of the four measures is not attained simultaneously for our dataset. Therefore, a fused result for the four indices has been used to determine the best number of clusters as four. After deciding the number of clusters, the FCM algorithm was implemented for grouping the lentil accessions.

The 518 feature vectors used in this study were grouped into four clusters based on the FCM method. In Fig. 3, the first two nonlinear principal component scores were plotted with cluster centers and membership contours overlaid to aid visualization of group differences. Fig. 3 clearly showed that there are

Table 4: 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
V_{PC}	0.696	0.614	0.545	0.488	0.469	0.456	0.446	0.437	0.426	0.418	0.410
V_{PE}	0.466	0.682	0.863	1.023	1.113	1.179	1.243	1.296	1.354	1.402	1.449
V_{SC}	0.010	0.006	0.006	0.006	0.005	0.004	0.005	0.004	0.004	0.004	0.004
V_{XB}	8.041	12.041	8.612	8.850	4.277	4.482	4.860	3.652	3.420	2.912	4.177

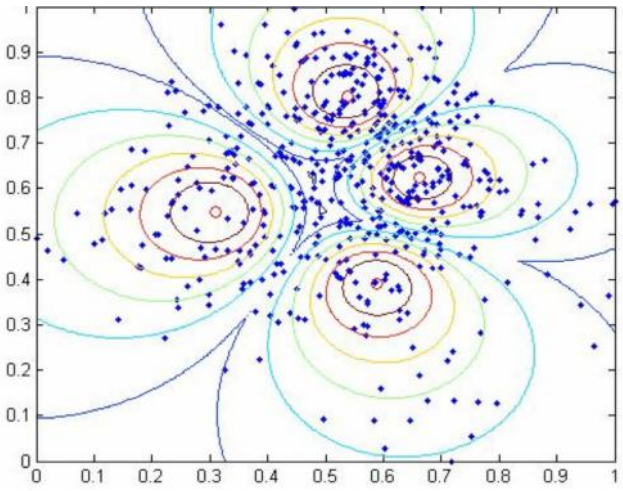

Fig. 3. Clustering by Fuzzy c-means algorithm of lentil genotype data

four clusters based on their centers' Euclidian distance from each other. It is evident from the figure 3 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 defuzzifying the results, we see that cluster I had 111 genotypes, cluster II had 145, cluster III and Cluster IV had 100 and 162 genotypes respectively. Mean value along with standard deviation of 10 numeric descriptors for four clusters is shown in Table 5. From table 5, we observed that genotypes in cluster IV were late in maturity with the medium level of pod /plant and low level of yield while genotypes in Cluster I had the maximum yield capacity and also moderate in maturity but high pods/plant. Genotypes in cluster III were late in maturity and medium yielding with very high pods/plant level. In cluster II, genotypes were low yielding and pods/plant was also very low but they are early maturing with large seed size. Clustering results of the study showed that if breeders are looking for larger sized lentil seeds which are mostly preferred in export market, then accessions from cluster II will be suitable for breeding purposes. Cluster IV contains most of the exotic accessions which are unadapted to Indian condition hence late in maturity with low yield/plant.

Conclusions

Summarizing the phenotypic variability in germplasm data using multivariate statistical technique is a common practice by the plant breeders in order to identify accessions of potential relevance for their crop improvement programs. The set of descriptors for the accessions of germplasm collections consists of both numerical and categorical descriptors. This poses problems for a combined analysis of all descriptors because few statistical techniques deal with mixtures of measurement types. In this paper, nonlinear principal component analysis was used to analyze the descriptors of 518 lentil accessions which can handle descriptors of different levels of measurement. Further, first two nonlinear principal components were used as variables for fuzzy clustering method for grouping lentil germplasm collections. The appropriate number of clusters was obtained with the help of validity measures. The results (accessions plot) of the study showed that most of the indigenous genotypes/land

Table 5. Mean and standard deviation of 10 numeric characters for four clusters

Characters	Cluster I	Cluster II	Cluster III	Cluster IV
No. of genotypes	111	145	100	162
Days to 50% flowering	70.44±9.43	68.92±9.47	83.00±6.00	86.59 ± 5.92
Plant height	30.10±3.93	27.41±4.80	32.41±3.91	32.21±4.63
Days to maturity	123.26±4.80	123.51±5.39	127.45±3.11	128.23±3.54
100-seed weight	2.73±0.47	2.75±0.45	2.32±0.46	2.13±0.43
Biological yield/plant	16.75±3.46	10.83±2.76	15.49±2.59	11.37±2.67
Primary branch/plant	4.49±1.11	3.06±0.77	4.26±1.05	3.40±0.65
Secondary branch/plant	11.57±1.94	7.82±1.67	11.74±1.91	9.81±1.83
Pods/plant	146.71±46.70	72.39±29.85	147.78±38.00	102.05±32.96
Yield/plant	5.18±1.70	2.54±1.25	4.60±1.18	2.95±0.87
Plant height at lowest pod	9.80±2.41	9.19±1.99	11.61±1.81	11.61±2.05

races overlapped together due to their narrow genetic base. Most of the outlying accessions belong to exotic origin or breeding lines derived from crosses with exotic lines. It can be concluded from this study that the nonlinear principal component based fuzzy clustering techniques were successfully applied to classify lentil genotypes on the basis of agro-morphological traits. The clustering pattern will also provide the knowledge of genetic relationship among the accessions and enable breeders to select diverse accessions for crossing programme and genetic enhancement to improve the yield potential of crop through breeding for superior recombinants. Further it will also be very useful for gene bank curators for core set development and enhancing utilization of genetic resources and their management in case of large germplasm collection. Hence, fuzzy clustering has a promising potential in agriculture as a tool to evaluate, understand, predict, and manage crop production.

References

1. **Jain A. K.** 2010. Data clustering: 50 years beyond K-means. Pattern Recognition Letters, **31**: 651-666.
2. **Yen J. and Langari R.** 2006. Fuzzy logic: Intelligence, control and information. Pearson Education, Inc., New Delhi.
3. **Dunn J. C.** 1974. Some recent investigations of a new fuzzy partition algorithm and its application to pattern classification problems. J. Cybernetics, **4**: 1-15.
4. **Bezdek J. C.** 1981. Pattern recognition with fuzzy objective function algorithms. Plenum Press, New York.
5. **Mingoti S. A. and Lima J.** 2006. Comparing SOM neural network with Fuzzy c-means, K-means and traditional hierarchical clustering algorithms. European J. Oper. Res., **174**: 1742-1759.
6. **Kruskal J. B. and Shepard R. N.** 1974. A nonmetric variety of linear factor analysis. Psychometrika, **39**: 123-157.
7. **Kroonenberg P. M., Harch B. D., Basford K. E. and Cruickshank A.** 1997. Combined analysis of categorical and numeric descriptors of Australian groundnut accessions using nonlinear principal component analysis. J. Agric. Biol. Env. Stat., **2**: 294-312.
8. **Franco J., Crossa J., Warburton M. L. and Taba S.** 2006. Sampling strategy for conserving maize diversity when forming core subsets using genetic markers. Crop Sci., **46**: 854-864.
9. **Frankel O. H.** 1984. Genetic perspectives of germplasm conservation. In Genetic Manipulation: Impact on Man and Society. Edited by Arber WK, Llimensee K, Peacock WJ, Starlinger P. Cambridge: Cambridge University Press: 161-170.
10. **Brown A. H. D.** 1989. The case for core collection. Genome, **31**: 818-824.
11. **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. Comp. Electr. Agric., **71**: 107-127.
12. **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.
13. **Gifi A.** 1990. Nonlinear multivariate analysis. Wiley, United Kingdom.
14. **Bezdek J. C.** 1974. Cluster validity with fuzzy sets. Cybernatics, **3**: 58-73.
15. **Gunderson R.** 1978. Application of fuzzy ISODATA algorithms to start-tracker pointing system. In Proc. of 7th Triannual World IFAC Cong., Helsinki, 1319-1323.
16. **Xie X. and Beni G.** 1991. A Validity Measure for Fuzzy Clustering, IEEE Trans. Pattern Analysis and Machine Intelligence, **13**: 841-847.
17. **Pal N. R. and Bezdek J. C.** 1995. On cluster validity for the fuzzy c-means model. IEEE Transactions on Fuzzy systems, **3**: 370-379.
18. **Li X., Lu Y., Li J., Xu H. and Shahid M. Q.** 2011. Strategies on sample size determination and qualitative and quantitative traits integration to construct core collection of Rice (*Oryza sativa*). Rice Sci., **18**: 46-55.
19. **Erskine W., Chandra S., Chaudhary M., Malik I. A., Sarker A., Sharma B., Tufail M. and Tyagi M. C.** 1998. A bottleneck in lentil: widening its genetic base in South Asia. Euphytica, **101**: 207-211.