ISSN 1479-2621
Plant Genetic Resources: Characterization and Utilization (2011) **9**(4); 523–527
doi:10.1017/S1479262111000827

A comparative performance of clustering procedures for mixture of qualitative and quantitative data – an application to black gram

Rupam Kumar Sarkar[1], A. R. Rao[2]*, S. D. Wahi[3] and K. V. Bhat[4]

[1]*Indian Agricultural Statistics Research Institute, New Delhi 110 012, India,* [2]*Centre for Agricultural Bioinformatics, Indian Agricultural Statistics Research Institute, New Delhi 110 012, India,* [3]*Biometrics Division, Indian Agricultural Statistics Research Institute, New Delhi 110 012, India and* [4]*National Bureau of Plant Genetic Resources, New Delhi 110 012, India*

Received 8 March 2011; Accepted 20 June 2011 – First published online 25 July 2011

Abstract

Knowledge of the genetic diversity of germplasm of breeding material is invaluable in crop improvement programmes. Frequently, qualitative and quantitative data are used separately to assess genetic diversity of crop genotypes. While assessing diversity based on qualitative and quantitative traits separately, there may occur a problem when the degree of correspondence between the clusters formed does not agree with each other. This study compares five different procedures of clustering based on the criterion of weighted average of observed proportion of misclassification in black gram genotypes using qualitative, quantitative traits and mixture data. The INDOMIX- and PRINQUAL-based clustering procedures, i.e. INDOMIX and PRINQUAL methods in conjunction with the *k*-means clustering procedure, show better performance compared with other clustering procedures, followed by clustering based on either quantitative or qualitative data alone. The use of the INDOMIX- and PRINQUAL-based procedures can help breeders in capturing the variation present in both qualitative and quantitative trait data simultaneously and solving the problem of ambiguity over the degree of correspondence between clustering based on either qualitative or quantitative traits alone.

Keywords: cluster analysis; genetic diversity; mixture data; qualitative traits; quantitative traits; RAPD

Introduction

Cluster analysis, in general, is done in an attempt to combine observations into homogeneous groups. The role of classification in the improvement of cultivated plants has long been recognized. For different breeding programmes or for varietal selection, there is a need to identify genetic materials that may contain useful traits. Therefore, it is of great interest to classify or group the accessions according to their trait scores or genetic structure. Different clustering procedures are available to classify accessions based on quantitative (morphological) and qualitative (molecular markers and DNA fingerprints) traits.

The clustering methods used for diversity analysis can be broadly classified as 'hierarchical methods' and 'projection techniques'. Hierarchical clustering methods in general and agglomerative hierarchical methods in particular are more commonly employed in the analysis of genetic diversity in crop species (Mohammadi and Prasanna, 2003). Among the agglomerative hierarchical methods, the unweighted pair-group method using

* Corresponding author. E-mail: arrao@iasri.res.in

arithmetic averages (UPGMA) (Sneath and Sokal, 1973) is the most commonly adapted procedure followed by Ward's minimum variance method (Ward, 1963). The projection techniques, principal components analysis (PCA) and principal coordinate analysis (PCoA), are the methods for displaying (transformed) multivariate data in low-dimensional space (Kolluru *et al.*, 2007). The plot of the two axes of PCA or PCoA in the $X-Y$ plane will allow identification of different clusters formed by the accessions. The third possible approach for this situation, under projection techniques, is the multidimensional scaling. For clustering based on qualitative data, the similarity measures such as simple matching coefficient, Jaccard's coefficient, Dice's coefficient, Russel and Rao's coefficient, Rogers and Tanimoto's coefficient, Sokal and Sneath's coefficient are used to find out the proximity among the accessions (Li, 2006). Whereas for quantitative data, distance measures such as Euclidean, squared Euclidean and Minkowski are used. A recent application of some of the aforementioned methods in the diversity analysis of sorghum and cassava germplasm can be found in Geleta and Labuschange (2005) and Kawuki *et al.* (2011), respectively.

Frequently, the data in germplasm collections contain a mixture of quantitative and qualitative data (Souza and Sorrells, 1991a, b). In practice, the qualitative and quantitative data originate from different sources, and the clustering procedures applied for analysing these data separately are not applicable for the combined (mixture) data. The methods proposed by Gower (1971), Peeters and Martinelli (1989) and Cole-Rodgers *et al.* (1997) for clustering objects based on mixture data require some pre-processing. The quantitative variables are range standardized, while the qualitative ones are method dependent. However, the advantage of the mixture data analysis, with suitable statistical techniques, over solely qualitative or quantitative traits is that it may result in more reliable homogeneous groups. Harch *et al.* (1999) showed that the use of both qualitative and quantitative descriptors can enable precise patterns in the groundnut taxonomy over quantitative descriptors alone. Another use of classification methods based on mixture data can solve the problem of disagreement between groups made solely on the basis of qualitative or quantitative data. de Leeuw and van Rijckevorsel (1980) proposed PCA of mixed variables (PCAMIX) to handle mixture data. Kiers (1989) proposed an alternative to PCAMIX, which is the application of individual difference scaling (Carrol and Chang, 1970) with orthonormality constraints on object coordinate (INDORT) for mixture variables or briefly INDOMIX. Similar to PCAMIX, INDOMIX also yields object coordinates ($= x_{ij}$, the value corresponding to the ith accession of jth principal component; $i = 1, 2,\ldots, m$; $j = 1, 2,\ldots, p$; where p is less than the total number of mixture variables), but does so by optimizing a criterion that differs from that of PCAMIX. Winsberg and Ramsay (1983) proposed the PRINQUAL method, which performs PCA of qualitative, quantitative or mixture data. Expectation maximization (EM) algorithm proposed by Dempster *et al.* (1977) is an unsupervised learning method and can accommodate mixture variables in cluster analysis. Kohonen (1988) proposed self-organizing feature map (SOFM) architecture in artificial neural network (ANN), which is also based on unsupervised learning and can also be deployed for the mixture data. A brief description on the application of PCAMIX, INDOMIX, PRINQUAL, EM and SOFM methods is given in Table S1 (available online only at http://journals.cambridge.org). Application of the first three methods in diversity analysis has not been fully explored. Also, the performance of unsupervised learning methods for diversity analysis is not fully assessed. Use of such efficient methods can enable the breeders to perform single complete analysis on mixture data for clustering accessions. Keeping this in view, an effort is made to study and identify suitable procedures for classification of black gram accessions based on the mixture data.

Materials and methods

In this study, data on 48 accessions of black gram grown during Kharif season in three locations, namely New Delhi, Hyderabad and Amravati, under randomized complete block design with two replications have been obtained from the National Bureau of Plant Genetic Resources, New Delhi. The data consist of information on 11 morphological quantitative characters, namely days to 50% flowering, number of primary branches per plant, number of clusters (on main shoot and branches) per plant, number of pods per cluster, pod length (cm), number of pods per plant, plant height (cm), number of seeds per pod, days to 80% maturity, seed yield per plant, 100 seed weight (g) and 203 random amplification of polymorphic DNA (RAPD) qualitative marker data that are scored 1 and 0 for the presence and absence of RAPD fragments, respectively.

Initially, the black gram accessions were grouped into six ($= k$) known pre-defined clusters based on location (state-wise). The identity of accessions falling in a given pre-defined cluster was observed. Now the accessions were mixed up, and the INDOMIX, PCAMIX and PRINQUAL methods were applied as a basis for clustering accessions. The object coordinates (principal component scores) obtained from each of these methods were then subjected to k-means clustering method, a non-hierarchical technique, for the formulation of homogeneous groups. In this study, k-means clustering method (with $k = 6$)

https://doi.org/10.1017/S1479262111000827 Published online by Cambridge University Press

was used as *a priori* information on number of groups. Furthermore, it is reasonable to suppose that the pre-defined groups represent the true grouping structure of the dataset as the areas from which they derived was geographically diversified and fall under different agroclimatic regions of the country. The constitution of the six clusters was observed for assessing the performance of each of the procedures. Henceforth, the INDOMIX, PCAMIX and PRINQUAL methods together with *k*-means clustering procedure are referred to as INDOMIX-, PCAMIX- and PRINQUAL-based clustering procedures. The EM and ANN procedures were directly applied on the mixture data by specifying the number of clusters as six and constitution of the resultant clusters was observed. All the aforementioned five procedures were assessed for their performance based on the criterion called 'observed proportion of misclassification (p)'. The observed proportion of misclassification, under each clustering procedure, for each of the corresponding pre-defined group was computed as the ratio of the number of accessions that have been wrongly classified into other groups in relationship with the number of accessions originally present in the pre-defined group. On the basis of these observed proportion of misclassification (p_i), an overall proportion of misclassification is calculated for each procedure as the weighted average of observed proportion of misclassification and is given by $\Sigma w_i p_i / \Sigma w_i$, where w_i denotes the weight (ratio of number of accessions in the ith pre-defined group to the total number of accessions) given to the ith pre-defined group such that $\Sigma w_i = 1$. A particular procedure is said to perform well if it has the lowest weighted average of observed proportion of misclassification.

Results

The five different procedures for clustering, described under Materials and methods, have been applied to the mixture data of black gram. Necessary Interactive Matrix Language (IML) code was written, by invoking *proc* IML in SAS 9.1.3 version (SAS, 2005), for the application of INDOMIX, PCAMIX and PRINQUAL methods (Table S2, available online only at http://journals.cambridge.org). The step-wise procedures in the STASTICA 9.0 data mining module have been followed for clustering by the EM and ANN methods. The observed proportion of misclassification of different methods of clustering was computed for all the groups as well as solely based on quantitative data by widely used *k*-means clustering and qualitative data by using the between-group average linkage method as well as the UPGMA method with Jaccard's coefficient of similarity. For qualitative data, Jaccard's coefficient of similarity was used because it was as efficient as that of Nei and Li or modified Roger's measures even in the missing data situations (Kolluru *et al.*, 2007). The weighted average of the observed proportion of misclassification has been computed and presented in Table 1. The results reveal that for cluster I, the observed proportion of misclassification under the INDOMIX-based clustering procedure is zero. Whereas among other procedures, the observed proportion of misclassification is lowest in PRINQUAL followed by PCAMIX-, EM- and ANN-based clustering procedures. Such a trend was strictly not observed in other groups, and this was perhaps due to smaller group size. To overcome the problem of different group size, the weighted average of observed proportion of misclassification has been computed. These proportions of misclassification under INDOMIX- and PRINQUAL-based clustering procedures were found to be least at 0.292 among all the procedures used for mixture, quantitative and qualitative data, whereas the proportion of misclassification for PCAMIX-based procedure has been found to be 0.375, which was lower than the proportion of misclassification by *k*-means method based on quantitative traits alone. However, PCAMIX does not seem to perform better than that of

Table 1. Probability of misclassification by different procedures of clustering

		Mixture data					Quantitative data	Qualitative data	
Cluster	No. of germplasms	PCAMIX[a]	INDOMIX[a]	PRINQUAL[a]	EM	ANN	*k*-means	Average linkage	UPGMA
I	29	0.069	0.000	0.034	0.310	0.207	0.241	0.034	0.137
II	5	0.800	1.000	1.000	0.600	1.000	0.400	1.000	1.000
III	4	0.750	0.000	0.000	0.250	0.000	0.250	0.000	0.000
IV	4	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
V	4	0.750	0.750	0.500	1.000	1.000	1.000	0.750	0.750
VI	2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Weighted average		0.375	0.292	0.292	0.479	0.438	0.417	0.313	0.375

[a] These procedures are applied along with *k*-means clustering procedure.

https://doi.org/10.1017/S1479262111000827 Published online by Cambridge University Press

average linkage method but performs at par with the UPGMA method based on qualitative traits alone. The weighted average of observed proportions of misclassification of EM and ANN procedures was found to be higher compared with other procedures. The possible reason for this could be that the number of observations in the dataset was not adequately large to train the model under these procedures. Moreover, EM and ANN are machine-learning approaches and they need large datasets for better performance.

Discussion

Use of mixture data as opposed to either quantitative or qualitative data alone is that huge amount of information present in the germplasm collection can be used simultaneously for describing the variability in the accessions. In this study, the use of INDOMIX- and PRINQUAL-based clustering procedures has enabled the newly formed groups to be related to the location-based pre-defined groups, which could not have been possible when quantitative or qualitative variables alone are used. The INDOMIX, PRINQUAL and PCAMIX methods quantify the qualitative and quantitative variables to a uniform scale so that the commonly used clustering procedures can be adapted on the object coordinates or principal component scores. The clustering procedures presented in this study are assessed for their performance because of *a priori* information available on the location-based pre-defined groups. Having been identified that INDOMIX- and PRINQUAL-based clustering procedures are performing better over others, the plant breeders can now fruitfully adapt these procedures. Furthermore, these procedures can be useful even in the absence of *a priori* information on '*k*', which can be approximated from the corresponding plots of first two or three principal component scores of INDOMIX or PRINQUAL methods.

The clustering procedures suggested in this study are free from any distributional assumption of the variables in the germplasm data. Although the procedures are demonstrated on black gram data, they can be very well applied to other crops and species. It is worth pointing out at this stage that there exist computer-intensive clustering techniques based on 'partition algorithms', which have not yet been fully explored in studying diversity analysis. Two such robust techniques based on 'partitioning algorithms' are partitioning around medoids and fuzzy analysis. Some work has already been initiated for possible application of these computer-intensive methods in diversity analysis of germplasm by the authors, and hopefully more precise and true representative clusters can be formed from the mixture data. However, at present, it is concluded that the classification of black gram genotypes with mixture data, consisting of quantitative traits and RAPD marker qualitative traits, by the INDOMIX- and PRINQUAL-based procedures may be adapted for obtaining representative homogenous clusters with least observed proportion of misclassification.

Acknowledgements

The authors wish to express their gratitude to referees and editor for important comments and suggestions, which improved the paper substantially.

References

Carrol JD and Chang JJ (1970) Analysis of individual differences in multidimensional scaling via an *N*-way generalization of Eckart–Young decomposition. *Psychometrika* 35: 283–319.

Cole-Rodgers P, Smith DW and Bosland PW (1997) A novel statistical approach to analyze genetic resource evaluations using capsicum as an example. *Crop Science* 37: 1000–1002.

de Leeuw J and van Rijckevorsel JLA (1980) HOMALS and PRINCALS, some generalization of principal components analysis. In: Diday E, Lebart L, Pagès JP and Tomassone R (eds) *Data Analysis and Informatics II*. North Holland/Amsterdam: Elsevier Science Publisher, pp. 231–242.

Dempster AP, Laird NM and Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* 39: 1–38.

Geleta N and Labuschange MT (2005) Qualitative traits variation in sorghum (*Sorghum bicolor* (L.) Moench) germplasm from eastern highlands of Ethiopia. *Biodiversity and Conservation* 14: 3055–3064.

Gower JC (1971) A general coefficient of similarity and some of its properties. *Biometrics* 27: 857–872.

Harch BD, Basford KE, DeLacy IH and Lawrence PK (1999) The analysis of large scale data taken from the world groundnut (*Arachis hypogaea* L.) germplasm collection. II. Two-way data with mixed data types. *Euphytica* 105: 73–82.

Kawuki RS, Ferguson M, Labuschagne MT, Herselman L, Orone J, Ralimanana I, Bidiaka M, Lukombo S, Kanyange MC, Gashaka G, Mkamilo G, Gethi J and Obiero H (2011) Variation in qualitative and quantitative traits of cassava germplasm from selected national breeding programmes in sub-Saharan Africa. *Field Crops Research* 122: 151–156.

Kiers HAL (1989) *Three-way Methods for Analysis of Qualitative and Quantitative Two-way Data*. Leiden: DSWO Press.

Kohonen T (1988) *Self-organizing and Associative Memory*. 3rd edn. New York: Springer-Verlag, Inc.

Kolluru R, Rao AR, Prabhakaran VT, Selvi A and Mohapatra T (2007) Comparative evaluation of clustering techniques for establishing AFLP based genetic relationship among sugarcane cultivars. *Journal of Indian Society of Agricultural Statistics* 61: 51–65.

Li T (2006) A unified view on clustering binary data. *Machine Learning* 62: 199–215.

Mohammadi SA and Prasanna BM (2003) Analysis of genetic diversity in crop plants – salient statistical tools and considerations. *Crop Science* 43: 1235–1248.

Peeters JP and Martinelli JA (1989) Hierarchical cluster analysis as a tool to manage variation in germplasm collections. *Theoretical and Applied Genetics* 78: 42–48.

SAS (2005) *SAS® 9.1.3 Language Reference: Concepts*. 3rd edn. Cary, NC: SAS Institute, Inc.

Sneath PHA and Sokal RR (1973) *Numerical Taxonomy*. San Francisco, CA: Freeman.

Souza E and Sorrells ME (1991a) Relationships among 70 North American oat germplasms. I. Cluster analysis using quantitative characters. *Crop. Science* 31: 599–605.

Souza E and Sorrells ME (1991b) Relationships among 70 North American oat germplasms. I. Cluster analysis using qualitative characters. *Crop Science* 31: 605–612.

Ward JH Jr (1963) Hierarchical grouping to optimize an objective function. *Journal of American Statistical Association* 58: 236–244.

Winsberg S and Ramsay JO (1983) Monotone spline transformations for dimension reduction. *Psychometrika* 48: 575–595.