Short Communications

Indian Journal of Agricultural Sciences **82** (2): 161–3, February 2012

Classification of maize genotypes by artificial neural network-based method: self organizing feature map*

SUKANTA DASH[1], S D WAHI[2] and A R RAO[3]

Indian Agricultural Statistics Research Institute, New Delhi 110 025

Received: 21 September 2010; Revised accepted: 25 August 2011

Key words: ANN, City-block distance, Cluster analysis, Euclidean distance, Maize

The summarization of large quantities of multivariate data is being increasingly practiced in various branches of agricultural science. The number of multivariate statistical techniques, namely, cluster analysis, principal component analysis, factor analysis is being widely used for classification purposes. One of the basic problems faced by the plant breeders is to classify large number of genotypes / lines into fewer manageable homogeneous groups / clusters. Cluster analysis is a classificatory technique for grouping objects into unknown number of homogeneous groups. There are large number of clustering methods and dissimilarity measures available in literature for making homogeneous groups. One of the main problems faced by the breeder is to choose a suitable method of clustering and dissimilarity measure among the different methods and dissimilarity measures available as there is hardly any information available in literature on the performance of these clustering methods and dissimilarity measures. Moreover, recently application of artificial neural networks (ANNs) for classification purposes has been increasingly recognized (Zupan 1994, Tudu *et al.* 2007). Kohonen's network is a 'self-organizing' system, which automatically adapts itself in such a way that the similar input objects are associated with the topological close neurons in the ANN. So it is of great interest to know the performance of ANN vs classical clustering methods. Keeping this in view, the present paper deals with comparing the performance of six different clustering methods, including ANN and two commonly used dissimilarity measures for clustering genotypes into homogeneous groups.

The secondary data on maize crop collected from Annual Progress Report 2005–06 of All India Coordinated Maize Improvement Project, Directorate of Maize Research, IARI campus, New Delhi has been used in the present investigation. Seventyseven maize genotypes were grown in seven different locations, i e Bajaura, Kangra, Ludhiana, Karnal, Varanasi, Jashipur, Ambikapur. The data was collected on ten morphological characters such as grain yields (kg/ha) at 15% moisture (X_1), days to 50% pollen shed (X_2), days to 50% silking (X_3), days to 50% dry husk (X_4), moisture percentage at harvest (X_5), plant aspect in gm/plant (X_6), ear aspect gm/plant (X_7), Husk cover kg/plant (X_8), Plant height in cm (X_9), Ear height in cm (X_{10}).

The commonly used classical methods of clustering fall into two general categories: Hierarchical and Non-hierarchical (Johnson and Wichern 2006). The two distance measures, four hierarchical and one non-hierarchical clustering methods considered in the present study are described below:

The squared Euclidean distance between two p-dimensional observations $\boldsymbol{x} = [x_1, x_2, \ldots, x_p]'$ and $y = [y_1, y_2, \ldots, y_p]'$ is

$$d(x,y) = \sum_i (x_i - y_i)^2$$

City-block distance: This distance is simply the average difference across dimensions. The city-block distance is computed as:

$$d(x, y) = \sum_i |x_i - y_i|$$

It is a method of calculating distances between two clusters computed as the distance between the two closest elements in the two clusters.

Mathematically, the linkage function – the distance D(X,Y) between clusters X and Y – is described by the expression:

$$D(X,Y) = \min_{x \in X,\, y \in Y} d(x, y)$$

where X and Y are any two sets of elements considered as clusters, and D(X,Y) denotes the distance between the two

*Short notes

Based on complete M.Sc. thesis of the first author, submitted to PG School, IARI, New Delhi in 2008

[1]PhD Student, Sukhatme Hostel, IASRI; (e mail: sukanta.iasri@gmail.com); [2]Principal Scientist, Biometrics Division, IASRI; (e mail sdwahi@iasri.res.in); [3]Senior Scientist, Centre for Agricultural Bioinformatics, IASRI; (e mail: arrao@iasri.res.in)

elements x and y.

It is a method of calculating distances between clusters as the maximum distance between a pair of objects, one in one cluster, and one in the other.

Mathematically, the complete linkage function—the distance D(X,Y) between clusters X and Y—is described by the following expression:

$$D(X,Y) = \min_{x\in X,\ y\in Y} d(x, y)$$

where d(x,y) is the distance between elements $x\in X$ and $y\in Y$; X and Y are any two sets of elements considered as clusters.

Average linkage (Alink) treats the distance between two clusters as the average distance between all pairs of elements, where one member of pair belongs to each cluster.

Mathematically, the average linkage function — the distance D(X,Y) between clusters X and Y — is described by the following expression :

$$D(X,Y) = \frac{1}{|X|\cdot|Y|} \sum_{x\in X} \sum_{y\in Y} d(x, y)$$

where d(x,y) is the distance between elements $x\in X$ and $y\in Y$; X and Y are any two sets of elements considered as clusters and |X| and |Y| are the number of elements in X and Y respectively.

Ward's method minimizes the total within-cluster variance. At the initial step, all clusters are singletons (clusters containing a single point). At each step the pair of clusters with minimum cluster distance is merged. To implement this method, at each step a pair of clusters that leads to minimum increase in total within-cluster variance after merging is found. This increase is a weighted squared distance between cluster centres.

Non-hierarchial clustering method is a method of cluster analysis which aims to partition n observations into k clusters in which each element belongs to the cluster with the nearest mean.

Mathematically, for a given a set of elements (x_1, x_2, …, $\mathbf{x}_n$), where each element is a P-dimensional real vector, k-means clustering aims to partition the n observations into k sets ($k \le n$)

$\mathbf{S} = \{S_1, S_2, \ldots, S_k\}$ so as to minimize the within-cluster sum of squares (WCSS):

$$\arg_s \min \sum_{i=1}^{k} \sum_{\mathbf{x}_j\in S_i} \left\|\mathbf{x}_j - \boldsymbol{\mu}_i\right\|^2$$

where $\pm_i$ is the mean of points in S_i.

The sixth method of clustering used here is an ANN based classification method given by Kohonen (1988) - self organizing feature map (SOFM), which is capable to solve the unsupervised problems rather than the supervised problems. In unsupervised problems (like clustering) it is not necessary to know in advance the membership of the objects of different clusters. During the training in the Kohonen's ANN, the P-dimensional neurons are 'self-organized' themselves in the two-dimensional plane in such a way that the objects from the P-dimensional measurement space are mapped into the plane of neurons with respect to some internal property correlated to the P-dimensional measurement space of objects.

SAS 9.1 software with proc cluster has been used for analysis by different clustering methods except ANN, for which MATLAB 7.0 software has been used.

Three homogeneous clusters based on multivariate data on 77 maize genotypes were obtained based on the results of different clustering procedures on consensus basis. The mean vectors and dispersion matrices were used as population parameters for simulation purposes. Four different hierarchical methods of clustering, given in material and methods, using squared Euclidean and City-block dissimilarity measures, K-means non-hierarchical method and ANN are compared on the basis of different simulated multivariate normal populations with given mean vectors and dispersion matrices. The performance of these methods is compared on the basis of average probability of mis-classification obtained from 20 different simulated samples. Further, the consistency of these methods is judged on the basis of standard errors of these probabilities. The samples of multivariate data for three clusters of small (=30), moderate (=60) and large (=150) sample size are simulated. The average probabilities of mis-classification along with standard errors for different methods of clustering are given in Table 1. The results of Table 1 show that the percentage of mis-classification is least (5.66) for ANN method, followed by Ward's method with squared Euclidean distance (9.33) and K-means (9.67) in case of small sample size. Whereas on comparing the hierarchical clustering methods (with City-block distance), K-means and ANN methods, it was found that ANN is best followed by K-means method. The performance of ANN method is found to be the best for moderate as well as large samples, followed by K-means method, Average linkage and Ward's method. Among the different methods of clustering, irrespective of dissimilarity measures used, the ANN is the most consistent method with least standard error, followed by K-means method for all the sample sizes. Hence, it is concluded that the performance of ANN method is the best among the six methods of clustering irrespective of the sample size. In case of small sample size the Ward's method with Squared Euclidean distance is found to be the second best procedure. Further, in case of medium and large samples the K-means method is found to be the second best clustering procedure.

SUMMARY

Seventy seven maize (*Zea mays* L.) genotypes collected from Annual progress report 2005-06 of All India Coordinated Maize Improvement Project, Directorate of Maize Research are classified by 6 different clustering methods including

Table 1 The average percentage probability of misclassification (P) and its standard error (SE) for different methods of clustering based on Squared Euclidean distance and City-block dissimilarity measures

Sample size	Distance		Clustering Method					
			Slink	Clink	Alink	Ward's	K-means*	ANN**
Small	Euclidean	P	16.501	11.500	11.001	9.333	9.667	5.666
		S.E.	2.196	1.718	2.196	1.338	1.277	0.598
	City-block	P	16.000	12.333	11.167	10.334		
		S.E.	1.737	1.512	1.330	1.182		
Medium	Euclidean	P	29.167	11.085	7.918	9.168	6.984	5.417
		S.E.	3.613	1.264	0.879	0.909	0.663	0.399
	City-block	P	35.417	11.476	8.499	9.698		
		S.E.	3.030	1.28	0.813	0.938		
Large	Euclidean	P	47.416	8.500	7.801	9.201	6.899	4.534
		S.E.	2.354	0.893	0.686	0.677	0.567	0.330
	City-block	P	47.566	8.699	8.301	9.600		
		S.E.	2.229	0.875	0.637	0.683		

* - K-means method is based on nearest centroid (mean)

** - ANN method is based on Kohonen's network

ANN and compared based on probability of misclassification. The percentage probability of misclassification for small, moderate and large sample sizes based on ANN method was 5.666, 5.417 and 4.534 respectively. The second best method for small sample size was Ward's method with 9.333 as percentage probability of misclassification. Whereas for moderate and large sample sizes K-means method was the second best method with 6.984 and 6.899 as percentage probability of misclassification. Hence, it can be concluded that the performance of ANN method is the best among the six methods of clustering irrespective of the sample size and dissimilarity measures used.

REFERENCES

Johnson R A and Wichern D W. 2006. *Applied Multivariate Statistical Analysis*. 5th edn., London, Inc. Pearson Prentice Hall.

Kohonen T. 1988. An introduction to neural computing. *Neural Networks*, **1**: 3–16.

Tudu B, Bhattacharya N, Jana A, Ghosh D and Bandyopadhyay R. 2007. Self-organizing map based classification of smell stages of black tea Fermentation process using flectronic nose. *3rd Indian International Conference on Artificial Intelligence*, 1626–35.

Zupan J. 1994. Introduction to Artificial Neural Network (ANN) Methods. *Acta Chemical Slovenica* **41** (3): 327–52.