Indian Journal of Agricultural Sciences **82** (12): 1055–8, December 2012

Performance of clustering procedures for grouping germplasms based on mixture data with missing observations

RUPAM KUMAR SARKAR[1], A R RAO[2], S D WAHI[3] and K V BHAT[4]

Indian Agricultural Statistics Research Institute, New Delhi 110 012

Received: 7 March 2011; Revised accepted: 4 September 2012

ABSTRACT

Occurrence of missing observations in mixture of qualitative and quantitative trait data is a common feature in breeding experiments. However, it becomes difficult to cluster the germplasms in presence of missing data. In the present study, five different clustering methods, six different ways of imputing missing data and three levels of missing observations have been considered in order to compare the performance of clustering procedures meant for mixture data. It was found that all the clustering methods are robust against imputation up to 5% missing observations. The INDOMIX and PRINQUAL methods in conjunction with k-means clustering with imputation of missing observations by (i) mean substitution in quantitative traits and frequency substitution in qualitative traits and (ii) multiple imputation in quantitative traits and 0 imputation in qualitative traits found to perform better than EM, ANN and PCAMIX methods for classification of germplasms. This study has been conducted during 2009–10 at Indian Agricultural Statistics Research Institute and for illustration purpose data has been obtained from National Bureau of Plant Genetic Resources.

Key words: Cluster analysis, Imputation, Missing data, Mixture data, Qualitative traits, Quantitative traits, Random Amplified Polymorphic DNA (RAPD)

In breeding trials, the qualitative and quantitative data which comes from different sources are often combined to form more accurate homogeneous groups of large number of germplasms. Occurrence of missing data is a common phenomenon in breeding trials and is primarily due to the result of non-response of germplasms for quantitative (morphological) and qualitative (molecular marker) traits. It may be also due to constraints like, non-germination, non-initiation of flowers, pest and disease attack. In case of RAPD based molecular marker qualitative data, where the profiles are visually scored as 0 and 1 against a possible absence or presence of band, it may be possible that some profiles are not so much distinguishable. Moreover, many a time RAPD markers fail to produce any kind of expression for some genotypes. So, in such situations one may face the problem of handling missing observations. In literature, limited procedures are available on cluster analysis in situations involving missing observations. Till 1970's, missing data has been viewed as a hurdle and were normally deleted thereby resulting in loss of information. There are different ways of averaging these missing data, like EM algorithms (Dempster *et al.* 1977) and multiple imputation techniques (Little and Rubin 1987). Most of the commonly used Statistical software packages like SPSS and SAS provide analysis under missing value situations. Troyanskaya *et al.* (2001) showed that K-Nearest Neighbors impute (KNNimpute) method provides a more robust and sensitive method for missing value estimation than Singular Value Decomposition impute (SVDimpute) method and both SVDimpute and KNNimpute performs better than the commonly used row average method. Bo T H *et al.* (2004) concluded that Least Squares impute (LS impute) produce estimates that are consistently more accurate than those obtained using KNNimpute and are also at least as accurate as Expectation Maximization (EM) impute algorithm. Kolluru (2007) proposed fuzzy clustering method to classify sugarcane genotypes with AFLP marker data containing missing observation.

Clustering of germplasms based on mixture data in presence of missing observations is challenging task and needs to be tackled through sound statistical techniques. Keeping this in view, the present investigation aims to identify suitable methods of imputing missing observations in mixture data and to assess the performance of different clustering procedures against different levels of imputed missing observations.

[1]Ph D Student (e mail: rupamiasri@gmail.com), [2]Senior Scientist (e mail: arrao@iasri.res.in), [3]Principal Scientist (e mail: sdwahi@iasri.res.in), IASRI, New Delhi; [4]Principal Scientist (e mail: kvbhat2001@yahoo.com), NBPGR, New Delhi

MATERIALS AND METHODS

The data on 48 genotypes of blackgram with 11 morphological characters, namely, days to 50% flowering, number of primary branches/plant, number of clusters (on main shoot and branches)/plant, number of pods/cluster, pod length (cm), number of pods/plant, plant height (cm), number of seeds/pod, days to 80% maturity, seed yield/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, obtained from National Bureau of Plant Genetic Resources, New Delhi was used in this investigation. These genotypes have been raised during 2002 *kharif* season at three locations, namely, New Delhi, Hyderabad and Amravati in a Randomized Complete Block Design with two replications. The above data has been used to illustrate/study the robustness of different clustering procedures at three levels of missing observations (1%, 5% and 10%), by deleting observations randomly. The above procedure was repeated 50 times to generate 50 samples each for three levels of missing observations.

Imputation of missing observations

As the data set contains mixture of qualitative and quantitative data, different combination of methods of imputing missing observations is adopted as shown in Table 1. Depending upon the nature of missing observations, i e quantitative or qualitative traits, the observations were imputed by different procedures. In case of missing quantitative data, the observations were imputed either by mean substitution or by multiple imputation procedure. Whereas, in case of missing qualitative data three different procedures of imputation, viz. substitution by 0, substitution by 1 and substitution by 1 or 0 depending upon the maximum frequency of 1 or 0 for a given marker.

Mean imputation method simply replaces the missing observation by the mean value for that particular quantitative variable. In multiple imputation, each missing observation was replaced with a set of probable values that represent the uncertainty about the right value to impute (by invoking MI procedure of SAS). This process results in valid statistical inferences that properly reflect the uncertainty due to missing values.

Table 1 Combination of different imputation procedures

Combination	Qualitative	Quantitative
C1	0	Mean imputation
C2	0	Multiple imputation
C3	1	Mean imputation
C4	1	Multiple imputation
C5	1, if freq(1) > freq(0) or 0, otherwise	Mean imputation
C6	1, if freq(1) > freq(0) or 0, otherwise	Multiple imputation

Five different clustering methods, viz. INdividual Difference scaling with Orthonormality constraints on object coordinate for MIXture variables or briefly INDOMIX (Kiers 1989), Principal Components Analysis for mixture data or PCAMIX (de Leeuw and van Rijckevorsel 1980), PRINcipal components analysis for QUALitative data or PRINQUAL (Winsberg and Ramsay 1983), Expectation-Maximization or EM (Dempster *et al.* 1977) and Artificial Neural Network or ANN (Kohonen 1988) are considered for identification of homogeneous groups in the blackgram germplasms with mixture of quantitative and qualitative data. The performance of the clustering procedures for mixture data are assessed under different levels of missing observations against different methods of imputation techniques. Analyses of performance by different methods, viz. INDOMIX with k-means clustering, PCAMIX with k-means clustering, PRINQUAL with k-means clustering, EM, ANN by different combinations of imputation procedures (C1 to C6) under different levels of missing observations are assessed on the basis of the criterion given below:

Criterion for assessing the performance of different clustering techniques

For this purpose, initially the germplasms are grouped into six known pre-defined clusters based on location (State-wise). Then the probability of misclassification (p) for each pre-defined group is computed as the ratio of the number of germplasms that have been wrongly grouped into other groups in relation to the number of germplasms in the pre-defined group by applying five different methods in conjunction with K-means clustering procedure. On the basis of these probabilities of misclassification an over all probability of misclassification is worked out for each method as weighted average probability of misclassification and is given by $(\sum w_i p_i) / (\sum w_i)$, where w_i denote the weight (ratio of number of germplasms in the i^{th} pre-defined group to the total number of germplasms) given to the i^{th} pre-defined group such that $\sum w_i = 1$. A particular method is said to perform well if it has lowest weighted average probability of misclassification.

RESULTS AND DISCUSSION

The samples with randomly created missing values, to the extent of 1%, 5% and 10% in the original data set, were imputed by the method of mean substitution and multiple substitution for quantitative traits and imputation by zero, one and frequency substitution for qualitative traits. The mixture data thus generated through all possible combinations of data imputation techniques (C1 – C6) were analyzed by using different clustering procedures.

The five different methods for clustering have been applied on the mixture data of blackgram. Necessary

Interactive Matrix Language (IML) code in SAS 9.1.3 (SAS 2005) for application of INDOMIX, PCAMIX and PRINQUAL methods have been developed and are given in Supplementary data. The step-wise procedures in STASTICA 9.0 data mining module have been followed for clustering by EM and ANN methods. In case of INDOMIX, PCAMIX and PRINQUAL the principal component scores have been obtained initially. These scores were then subjected to k-means clustering procedure for further clustering purpose. All the germplasms have been grouped into six clusters by each of the clustering methods. Six clusters have been considered because of prior information available from the state wise known distribution of germplasms.

The genotypes falling under different state-wise groups were compared with the newly formed six groups under each method for the purpose of computing the probability of misclassification. The weighted average probabilities of misclassification are also computed for all the methods. The weighted average probabilities of misclassification averaged over 50 generated samples are worked out for different combination of data imputation techniques against different percentage of missing observation and are presented in Table 2.

It has been observed from the results that the weighted probability of misclassification at 1% of missing observations under different imputation techniques are equal to that obtained from 0% missing data irrespective of clustering procedures adopted. As the percentage of missing observation increase from 1% to 10%, the weighted probability of misclassification increases. This is true under all combinations of imputation methods. Further it is observed that the weighted probability of misclassification increases at a higher rate when the number of missing observations increases from 5% to 10%; however the rate of misclassification is comparatively low as the missing observation increases from 1% to 5%. The probability of misclassification is quite high in case of EM and ANN methods and do not show significant change with increase in percentage of missing observations from 0 to 10. Moreover, it is also observed that the treatment of missing observations by (i) mean substitution in quantitative traits and frequency substitution in qualitative traits (C3) and (ii) multiple imputation in quantitative traits and 0 imputation in qualitative traits (C4) perform better than all other methods of imputation for a given percentage of missing observations. It is observed that among all the clustering procedures, INDOMIX and PRINQUAL in conjunction with k-means clustering show the weighted probability of misclassification averaged over 50 generated samples is around 0.3, which is much lower than other methods, particularly, EM and ANN. The possible reason for high weighted average probability of misclassification under different levels of imputed missing observation by EM and ANN could be that the number of observations in the data set is not sufficient enough to train the models.

Table 2 Weighted average probability of misclassification of different methods for different levels of missing observations and imputation procedures

Per cent missing observations	Method					
	C1	C2	C3	C4	C5	C6
			PCAMIX*			
0	0.375	0.375	0.375	0.375	0.375	0.375
1	0.375	0.375	0.417	0.344	0.375	0.375
5	0.417	0.406	0.448	0.396	0.417	0.438
10	0.458	0.417	0.500	0.448	0.458	0.469
			INDOMIX*			
0	0.292	0.292	0.292	0.292	0.292	0.292
1	0.313	0.292	0.323	0.313	0.292	0.292
5	0.396	0.338	0.347	0.365	0.417	0.458
10	0.500	0.448	0.448	0.458	0.469	0.469
			PRINQUAL*			
0	0.292	0.292	0.292	0.292	0.292	0.292
1	0.292	0.292	0.292	0.292	0.281	0.292
5	0.396	0.344	0.323	0.337	0.344	0.365
10	0.431	0.438	0.438	0.427	0.437	0.406
			EM			
0	0.479	0.479	0.479	0.479	0.479	0.479
1	0.479	0.458	0.479	0.448	0.500	0.469
5	0.531	0.490	0.500	0.458	0.563	0.521
10	0.583	0.542	0.531	0.479	0.573	0.542
			ANN			
0	0.438	0.438	0.438	0.438	0.438	0.438
1	0.458	0.417	0.430	0.417	0.438	0.430
5	0.468	0.479	0.457	0.479	0.458	0.479
10	0.498	0.498	0.479	0.500	0.521	0.500

*These methods are applied along with k-means clustering procedure

It can be concluded from the results that the INDOMIX and PRINQUAL methods in conjunction with k-means clustering with imputation of missing observations by (i) mean substitution in quantitative traits and frequency substitution in qualitative traits and (ii) multiple imputation in quantitative traits and 0 imputation in qualitative traits found to perform better than the other methods of clustering germplasms. Also, the above suggested methods are free from distributional assumptions of the variables in the germplasm data. Even though these procedures are demonstrated on blackgram data they can be very well applied to other crops.

REFERENCES

Bo T H, Dysvik B and Jonassen I. 2004. LSimpute: Accurate estimation of missing values in micoarray data with least squares method. *Nucleic Acids Research* **32**(3): 34.

de Leeuw J and van Rijckevorsel J L A. 1980. HOMALS and PRINCALS, some generalization of principal components

analysis. (*in*) *Data Analysis and Informatics II*, pp 231–42, Diday E, Lebart L, Page's J P and Tomassone R (Eds). Elsevier Science Publisher, North Holland / Amsterdam.

Dempster A P, Laird N M and Rubin D B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* **39** (1): 1–38.

Kiers H. 1989. *Three - Way Methods for the Analysis of Qualitative and Quantitative Two – Way Data*, p 185. DSWO Press, University of Leiden, Netherlands.

Kohonen T. 1988. *Self-organizing and Associative Memory*, edn 3, 312. Springer-Verlag Inc., New York, USA.

Kolluru R, Rao A R, Prabhakaran V T, 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**(1): 51–65.

Little R J A and Rubin D B. 1987. *Statistical Analysis with Missing Data*, edn 2, p 385. John Wiley and Sons, New York, USA.

SAS. 2005. *SAS 9.1.3 Language Reference: Concepts,* 3rd edn. SAS Institute Inc., USA.

Statistica. 2009. *Statistica 9.0: Statistica Data Miner.* StatSoft Inc., OK, USA

Troyanskaya O, Cantor M, Sherlock G, Brown P, Has T, Tibshirani R, Botstein D and Altman R B. 2001. Missing value estimation methods for DNA microarrays. *Bioinformatics* **17**(6): 520–5.

Winsberg S and Ramsay J O. 1983. Monotone spline.