Variable Selection for Clustering and Classification of Data with Missing Values

Brynn O'Connell and Dr. Brian Franczak, Mathematics and Statistics, MacEwan University

Introduction

- The objective of this project is to evaluate the performance of the variable selection for clustering and classification (VSCC) approach on two different incomplete data sets.
- Specifically, we randomly remove 0%, 25%, 50%, and 75% of the values from the iris and seeds data sets.
- For each incomplete data set, we consider three imputation methods before applying VSCC.

Methods

Imputation Methods

- We consider three imputation procedures:

1. Mean Imputation (MI)
2. Predictive Mean Matching (PMM)
3. Random Sample from Observed Values (RS)

Variable Selection for Clustering and Classification (VSCC)

- We will employ the VSCC technique on the data sets which works by finding a subset of variables that simultaneously minimize the within-group variance and maximize the between-group variance.
- The VSCC clustering algorithm works as follows:

1. performs initial clustering using *mclust*
2. use the estimated weights of the individual observations belonging to each group to initialize VSCC
3. perform *mclust* on the (up to) five variable subsets given by VSCC
4. select the best model using the BIC
5. select the best variable subset according to the total model uncertainty.

Model Evaluation

- The Bayesian information criterion will be used to assess model fit.
- The adjusted Rand index (ARI) will be used to assess classification performance.

Data Set Description

Iris

- The iris data set (Becker et al., 1988) is comprised of 150 observations, 4 quantitative variables, and 1 categorical.
- The quantitative variables give the sepal length, sepal width, petal length, and petal width in centimeters.
- These measurements were obtained from 50 flowers of each of the three iris species: setosa, versicolor, and virginica.

Seeds

- The seeds data set is comprised of 210 observations from three different varieties of wheat grain: Kama, Rosa, and Canadian (Giordani et al., 2020).
- There are seven measurements of the geometrical properties of the wheat kernels:

1. Length of the kernel
2. Width of the kernel
3. Asymmetry coefficient
4. Length of kernel groove
5. Area
6. Perimeter
7. Compactness

Results

Iris

- In Table 1, the highest BIC values are given for the 25% case for MI and RS; however, the highest BIC value is given for the 75% case for PMM.
- In Figure 1, across the different cases, the highest ARI for MI, PPM, and RS is the 0% case.

Seeds

- In Table 2, for PMM and RS the highest BIC values are given for the 25% case, whereas for MI, the highest BIC is given for the 75% case.
- In Figure 2, the highest ARI for the MI and RS is given for the 0% case, and for PMM it is from the 25% case.

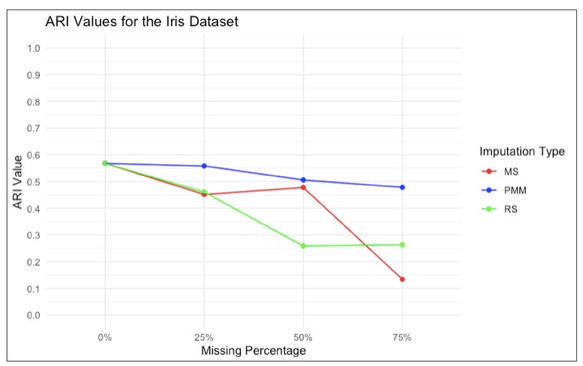

Figure 1. ARI values for the iris dataset across the three types of imputation

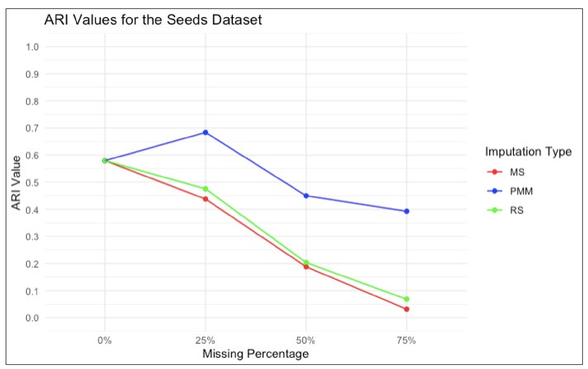

Figure 2. ARI values for the seeds dataset across the three types of imputation

Table 1. BIC and ARI values for the besting fitting models.

	MI				PMM			RS		
	0%	25%	50%	75%	25%	50%	75%	25%	50%	75%
# Vars	4	1	4	4	4	4	4	1	3	4
Relation	Full	1	2	2	Full	Full	2	1	1	Full
BIC	-790.6956	-290.19	-1315.722	-1352.074	-833.8029	-842.3195	-632.9451	-245.3039	-1006.417	-1480.267
Model	VVV	V	EEE	EVV	VVV	VVV	VVV	V	VVI	VVI
# Groups	2	2	7	5	2	2	2	2	6	6
ARI	0.568	0.425	0.478	0.134	0.558	0.506	0.478	0.461	0.259	0.263

Table 2. BIC and ARI values for the besting fitting models

	MI				PMM			RS		
	0%	25%	50%	75%	25%	50%	75%	25%	50%	75%
# Vars	7	6	7	1	7	4	7	7	7	6
Relation	3	2	Full	1	Full	2	3	3	3	4
BIC	179.9839	-1148.616	-2128.415	-564.4613	-735.6096	-1776.559	-1293.076	-1267.491	-2517.963	-2868.933
Model	EEV	VVV	VVE	E	VVV	VEE	VEE	VVV	VVV	VVE
# Groups	4	5	7	3	3	2	5	4	4	3
ARI	0.58	0.438	0.188	0.032	0.684	0.45	0.393	0.475	0.204	0.068

Conclusions

Iris

- The highest ARI across all data sets cases was observed from the data sets submitted to PMM. As expected, the highest ARI was found for the original data set with no missing values.

Seeds

- We observed the highest ARI on data sets submitted to PMM for imputation. Interestingly, the highest ARI value across all considered trials was produced for the data set with 25% of its values missing after PMM.

Limitations

- We did not consider multiple imputation within the PMM and RS frameworks.
- Further, we did not consider multiple permutations of the missing values.

Future Work

- Address the limitations.
- Consider other imputation procedures, clustering procedures, and data sets.

References

Giordani, P., Ferraro, M.B., Martella, F.: An Introduction to Clustering with R. Springer, Singapore (2020)

Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988) The New S Language. Wadsworth & Brooks/Cole. (has iris3 as iris.)