RANDOM-BASED INITIALIZATION FOR CLUSTERING MIXED-TYPE DATA WITH THE K-PROTOTYPES ALGORITHM

Rabea Aschenbruck [1], Gero Szepannek[1] and Adalbert F.X. Wilhelm[2]

[1] Stralsund - University of Applied Sciences, (e-mail: rabea.aschenbruck@hochschule-stralsund.de, gero.szepannek@hochschule-stralsund.de)

[2] Constructor University Bremen gGmbH, (e-mail: awilhelm@constructor.university)

ABSTRACT: One of the most popular partitioning cluster algorithms for mixed-type data is the k-prototypes algorithm. Due to its iterative structure, the algorithm may only converge to a local optimum rather than a global one. Therefore, the resulting cluster partition may suffer from the initialization. In general, there are two ways of achieving an improvement of the initialization: One possibility is to determine concrete initial cluster prototypes, and the other strategy is to repeat the algorithm with different randomly chosen initial objects. Different numbers of algorithm repetitions are analyzed and evaluated comparatively. It is shown that an improvement of the cluster algorithm's target criterion can be achieved by an appropriate choice of repetitions, even with manageable time expenditure.

KEYWORDS: k-prototypes, mixed-type data, cluster analysis, initialization.

1 Introduction to the Problem

In the origin initialization, points to be clustered are chosen randomly as initial cluster prototypes. Subsequent iterations lead to a local optimum of the summed squared error minimization problem, but not necessarily to the global minimum for k-prototypes (Huang, 1997). Therefore, the choice of proper starting points is important. In general, there are three different strategies to receive the initial prototypes: The starting points can be determined based on the knowledge of the clustering use case. Otherwise, one can do a mathematical determination or a random-based choice of objects to be clustered. The latter one is probably the most common way in practice, where k objects are randomly selected. These may or may not be good starting points for the iterative algorithm routine. To increase the probability of reaching a global optimum, one can apply the algorithm multiple times on different, randomly

chosen objects. In the following, different numbers of algorithm repetition are compared and evaluated on different data situations with regard to the adjusted Rand index (*short:* ARI; Hubert & Arabie, 1985) and the computation time.

2 Simulation Study on the Random-based Initialization of the *k*-Prototypes Algorithm

Execution of the Simulation Study The aim of this study is to determine an appropriate number of repetitions to obtain a satisfactory cluster partition but at the same time, the number of algorithm repetitions should be as low as possible because of the increasing computation time. In practice, the number of repetitions can be passed to the R function `kproto` (Szepannek, 2018) via the parameter `nstart`. After the algorithm's application on `nstart` randomly chosen prototypes sets, the partition which minimizes the target criterion is used. The simulation study was executed on a Dell PowerEdge R440 server with two Intel Xeon Silver 4216 processors (2 x 16 cores; 2.1 GHz) and 768 GB RAM.

In the simulation study were included 120 different data situations, differing by the variation of the number of observations (500, 1000, 2000), variables (2, 4, 8) and clusters (2, 4, 8), whether the cluster group sizes were equal or unequal, and the ratio of categorical to numerical variables (0.25, 0.5, 0.75). To mitigate the random influence of the generation process 50 data sets were determined for every of the 120 data situations (see Aschenbruck *et al.*, 2022).

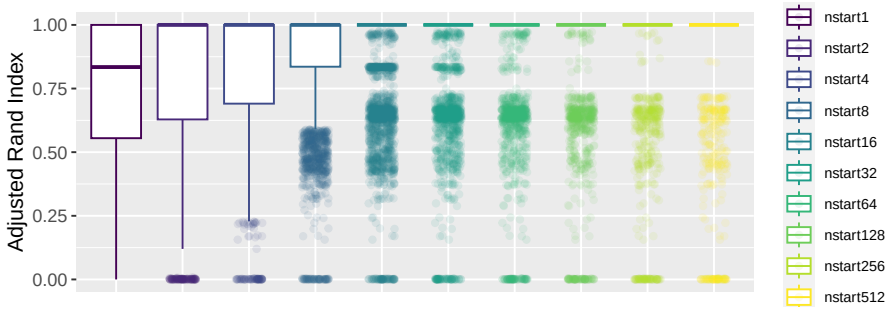

Figure 1. *Boxplots on the adjusted Rand index values of the resulting partitions.*

Comparision of the Different Numbers of Repetitions The higher the number of algorithm repetitions (`nstart`) the higher the rating of the resulting

cluster partitions by the ARI (see Fig. 1). 16 repetitions seem to be a good choice, since for most of the data sets examined, the resulting cluster partition is rated with an adjusted Rand index value of 1, which is the best possible rating and there is virtually not much improvement for more repetitions.

Having a closer look at the rated partitions for the different data situations in Fig. 2, it can be stated that the number of clusters to be determined and whether the clusters are of equal size or not is influencing the need for more repetitions to gain satisfying results. Since an increasing number of algorithm repetitions leads to an increase in computation time, hereafter a determination of the number of repetitions depending on the data situation is proposed.

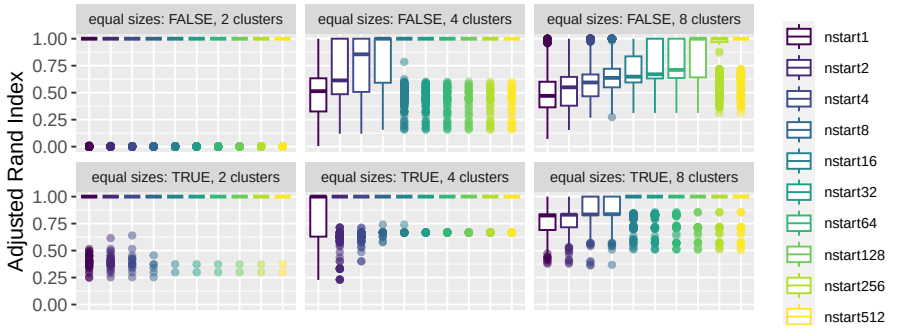

Figure 2. *Boxplots on the ARI of the resulting partitions, shown separately by number of clusters and whether or not the clusters are of the same size.*

Determination of the Number of Repetitions Depending on the Data Situation The data situation based number of repetitions m assures that with a probability of 0.9 at least one of the m sets of initial prototypes contains objects of every cluster group. Considering a geometrically distributed random variable $Z \sim Geo(\pi)$, it follows that the number of repetitions depending on the data situation at hand is

$$m = F_z^{-1}(0.9) \text{ with probability of sucess } \pi^\star = \prod_{i=0}^{k-1} \frac{N - i \cdot \lceil \frac{N}{k} \rceil}{N - i}, \qquad (1)$$

where N is the number of objects to be clustered and k the number of clusters to be determined. Thereby, all clusters are assumed to be of equal size since in practice, the sizes of the clusters to be determined are unknown. Nevertheless, if one suspects a small cluster group it is possible to input k in Eq. (1) as the reciprocal of this size.

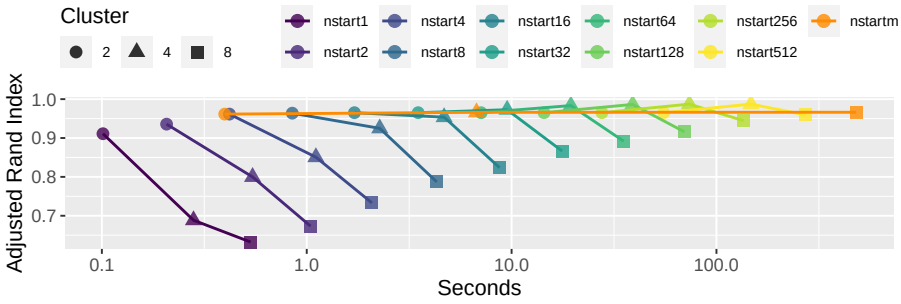

Figure 3. *Relationship between ARI and computation time in seconds (log-scaled), splitted by the initialization approach and the number of clusters in the data.*

Computation Time In Fig. 3 the average computation time for all data situations with the specified number of clusters and the average ARI is given. The influence of the number of repetitions and clusters on the increase in computation time is obvious. The data-based number of algorithm repetitions $m_{k,N}$ ($m_{2,\cdot} = 4$, $m_{4,\cdot} = 24$, $m_{8,500} = 905$, $m_{8,1000} = 931$, $m_{8,2000} = 944$) results in overall good rated partitions while avoiding unnecessary algorithm repetitions.

3 Summary

In this work, a theoretical determination of repetitions was motivated. For a small number of clusters, a few repetitions are sufficient, whereas as that number increases, a strong increase in repetitions is necessary, even at 8 clusters.

References

ASCHENBRUCK, R., SZEPANNEK, G., & WILHELM, A.F.X. 2022. Imputation Strategies for clustering mixed-type data with missing values. *J. Classif.*

HUANG, Z. 1997. Clustering Large Data Sets With Mixed Numeric and Categorical Values. *Pages 21–34 of: Proceedings of the First PAKDD.*

HUBERT, L., & ARABIE, P. 1985. Comparing Partitions. *J. Classif.*, 193–218.

SZEPANNEK, G. 2018. clustMixType: User-Friendly Clustering of Mixed-Type Data in R. *R J.*, **10**(2), 200–208.