A supervised methodology to measure the variables contribution to a clustering

Oumaima Alaoui Ismaili[1,2], Vincent Lemaire[2], and Antoine Cornuéjols[1]

[1] AgroParisTech 16, rue Claude Bernard 75005 Paris
[2] Orange Labs, 2 av. Pierre Marzin, 22300 Lannion

Abstract. This article proposes a supervised approach to evaluate the contribution of explanatory variables to a clustering. The main idea is to learn to predict the instance membership to the clusters using each individual variable. All variables are then sorted with respect to their predictive power, which is measured using two evaluation criteria, i.e. accuracy (*ACC*) or Adjusted Rand Index (*ARI*). Once the relevant variables which contribute to the clustering discrimination have been determined, we filter out the redundant ones thanks to a supervised method. The aim of this work is to help end-users to easily understand a clustering of high-dimensional data. Experimental results show that our proposed method is competitive with existing methods from the literature.

1 Introduction

Everyday, huge amounts of data are generated by users via the web, social networks, etc. Clustering algorithms are a tool of choice to explore these high-dimensional data sets. However, their use is often hampered by the lack of understandability of the results. End-users would like to identify the most relevant variables that suffice to explain the observed clusters, but these are not easily detectable once a clustering has been performed. It is therefore crucial to be able to evaluate the contribution of each descriptive variable to the clustering process. Indeed, not all variables are relevant to the clustering: some may be irrelevant, some may be noisy and some may be redundant or (and) correlated.

The purpose of this study is to find a simple way to assist the analysts in their interpretation of a clustering result. The idea is to sort variables according to their contribution to a clustering using a supervised approach. The importance of a variable is evaluated as its power to predict the membership of each object to a cluster. In this paper, we restrict ourselves to an univariate classifiers to obtain an univariate weight for each variable.

The paper is organized as follows: Section 2 describes briefly some related work. Then, Section 3 presents the proposed method to score the contribution of variables to a clustering. This section also presents an alternative method to eliminate redundant variables among the relevant variables. The experimental results are presented Section 4. Finally, the perspectives and the further research are presented as a conclusion in the last section.

2 Related work

Recently, the measure of the importance of the variables has been increasingly studied in the unsupervised learning. The methods proposed in this context can mainly be divided into two categories: *features selection* and *validation indices*.

Features selection methods can be grouped either as *wrapper* or as *filter* approaches. The *wrapper* approach aims to incorporate the feature selection in the clustering process, whereas, the idea of the *filter* approach is first to pre-select the features and then to use the selected features in the clutering process. In the unsupervised context, the *wrapper* methodology was initially proposed by Brodley in [1].

Inspired by the idea given in [1], Zhu et al. presented in [2] a novel method called ULAC. This method is essentially based on the analysis of the correlation among the variables. Moreover, some methods aim at removing the redundancy among variables. Accordingly, they rely on estimations of mutual information or of correlation ([3],[4],[5]). Mitra et al. proposed in [3] a method based on a measure of similarity between variables after elimination of the redundant variables. This measure is defined as the lowest eigenvalue of the correlation matrix. In [4], Vesanto et al. used a visualization tool (SOM-based approach) to detect the correlation between features. The same approach is used by Guerif et al in [5]. The difference between the two approaches is that Guerif et al. integrate a weight criterion in the SOM algorithm to reduce the effect of redundancy.

Other approaches have been presented to evaluate the clustering performance introducing criteria such as validation indices which can be adapted to evaluate the variables importance. Those approaches are divided in two main types: *external* and *internal* [6]. The *external* approaches exploit the supervised information given by the ID-cluster (identification given to each discovered cluster that can be subsequently used as a "label"). Among these approaches, we can cite: Adjusted Rand index [7], F-measure [8] and MMI [9]. The internal approaches use unsupervised criteria like the inertia. Among these methods, we can cite: Davies-Bouldin [10], Silhouette [11], Dunn-index [12], SD [13], XB-index [14], I-index [14] and BIC [15] indices.

3 Contribution

In this section, we propose two supervised approaches which fall within the context of the external validation indices. These approaches allow an interpretation of the clustering output based on relevant variables in case where the clsutering does not suffer from a very bad quality (otherwise there is no sense to interpret the result). In the remainder of this paper, we call this output (or the clustering result) *'the reference clustering'*. The first supervised approach consists in measuring the variables importance with respect to their predictive power regarding the cluster Ids. The second one aims at detecting the redundant variables.

3.1 Variables importance

The objective of this work is to propose a simple way to identify the most relevant features from the output of a clustering. In order to retain all variables, we rank the

variables according to their importance without doing a selection. The main idea is to turn this problem into a supervised classification problem where the cluster membership (ID-cluster) is used as a target class. Then, for each variable, we use a supervised classification algorithm to predict the ID-cluster. We define the importance of variables as their power to predict the ID cluster: a variable is relevant only if it is able to predict correctly the ID cluster obtained from the reference clustering (i.e. clustering using all variables). To measure the importance of each variable, we use two evaluation criteria: Accuracy and Adjusted Rand Index:

- *Accuracy* (ACC) criterion: a variable is considered relevant if the associated accuracy value is high.
- *Adjusted rand index (or ARI)* is a popular cluster validation index proposed by Hubert and Arabie [7]. It can be used to evaluate the performance of the classification as in [16]. In this work, we calculate the ARI between: (i) the reference clustering (ii) the predicted membership (ID-cluster) associated to the variable of which we want to measure the importance. The idea behind this is to compare the reference clustering with each predictive membership associated to each variable. So, a variable is important if the associated predictive ID-cluster is highly similar to the reference clustering, i.e. the ARI value is close to 1.

The algorithm 1 presented below provides a summary of our approach:

Notations:
X: The training database constituted of N examples and d explanatory variables, (X_{ab} is the value of the variable b for the example a)
M: A supervised classifier
CLU: A clustering algorithm
M_{ref}: The reference clustering model
$IdClusters$: A vector of the N memberships
R : Ranking of the d explanatory variables

$XPRE \leftarrow$ `preprocessing` (X)
$M_{ref} \leftarrow$ `train` (CLU, XPRE)
$IdClusters \leftarrow$ Membership($XPRE$, M_{ref})
for *i=1 to d* **do**
  $M_i \leftarrow$ `train` ($XPRE_{.i}$, $IdClusters$)
  $ACC_i \leftarrow$ `computeAccuracy` (M_i)
  $ARI_i \leftarrow$ `computeAdjustedRandIndex` (M_i)
end
$R_{ACC} \leftarrow$ `sortInDescendingOrder` (ACC_i, i=1 to d)
$R_{ARI} \leftarrow$ `sortInDescendingOrder` (ARI_i, i=1 to d)

Algorithm 1: Algorithm for ranking

An interesting measure of importance must allow us to sort variables according to their relevance in a clustering process and the least influent variables should only contain little or irrelevant information to create the clusters. Consequently, the quality of the obtained clustering which is deprived of these variables remains substantially the same or even slightly better (less noise). In contrast, the removal of an important variable deprives the algorithm of important information and leads to a poor clustering result.

To compare our proposed method to other existing methods from the literature, the curve of the ARI values versus the number of variables used will be plotted. This curve is obtained as follows:

For each iteration until one reaches the number of variables:

- Eliminate the less relevant variable with respect to the chosen criterion;
- New partition: run the clustering algorithm without this variable;
- Calculate the ARI value between the reference clustering and the new partition.

The review of the results can be visually made by observing the curve evolution (for example, see Figure 1).

3.2 Redundant variables

Once the variables that are the most informative for the clustering have been identified, it is important to filter out the redundant ones in order to improve the understandability of the result. To solve this problem, we propose a supervised approach.

The concept of redundancy is based on the similarity between partitions obtained using the "predicted ID-Clusters" (using Algorithm 1) for each variable. The assumption is: X_i and X_j are redundant if they produce similar partitions when considering their "predicted ID-Clusters" (using M_i and M_j). A way to measure the similarity between these two partitions is to use the ARI criterion. For example, the ARI criterion will be close to 1 when it calculated between two partitions containing same "predicted ID-Clusters" or between two partitions containing symetric "predicted ID-Clusters". The resulting algorithm is presented below (see Algorithm 2).

Notations:
X: The training database constituted of N examples and d explanatory variables
$M_.$: d supervised classifier models coming from the Algorithm 1
$PredId$: A vector of size N of the predicted ID-Cluster for a given explanatory variable
RE: A matrix of size dxd values

$XPRE \leftarrow$ `preprocessing` (X)
for *i=1 to d* **do**
  $PredId(d) \leftarrow$ `PredictionOfTheMembership` (M_i, $XPRE_{.i}$)
end
for *all pairs of variable (l, m)* **do**
  $RE(l, m) \leftarrow$ `computeAdjustedRandIndex` ($PredId(l)$,$PredId(m)$)
end

Algorithm 2: Algorithm for redundant variables

4 Experimental results

4.1 Protocol

To evaluate the behavior of our approach, we have selected 3 different datasets from the UCI [17]: WINE, PIMA and WAVEFORM datasets. The two first datasets are used to

illustrate the competitiveness of the proposed method to measure the variables importance comparing to two other methods from the literature. Among these methods, we decide to use efficient and often used indexes from the literature: Davies-Bouldin [10] and SD indexes [13]. The last dataset is used to illustrate the behavior of our approach to detect the redundant variables.

We proceed as follows to evaluate the performance of our approach:

- the pre-processing used is standardization[3];
- to obtain the reference clustering, the K-means algorithm [18] has been used where:
  - K is equal to the number of target class for each used datasets (as in [19]);
  - the method used to initialize the centroids is K-means++ algorithm [20];
  - the number of replicates is 25 [4].
- a decision tree (CART) [21] has been used to predict the ID-cluster[5].

4.2 Variables contribution

The first experimentation to test our approach is made using the WINE dataset which is constituted of $N = 178$ sample points described with $d = 13$ variables and associated with three different classes. The ARI obtained between the reference clustering (using K-means algorithm, where K=3) and the target class is equal to 0.91. Figure 1 presents the evolution of the ARI curve for the three approaches (SD, DB and the supervised approach using ARI or ACC to measure the contribution of variables in the clustering results) versus the number of variables. The table 1 (left part) presents the list of the ranked variables (from the most important to the least important) for the three approaches.

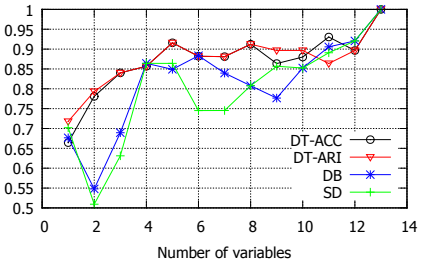

Fig. 1: Evolution of the ARI criterion for the 4 methods (K=3)

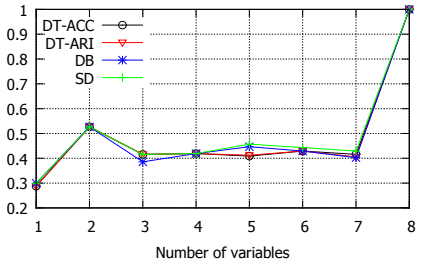

Fig. 2: Evolution of the ARI criterion for the four methods (K=2)

[3] All the experimentation have been realized using R (`http://www.r-project.org/`) and are easily reproducible.

[4] The initialization process and the nature of the K-means algorithm does not guarantee to reach a global minimum. Therefore the algorithm has to be run several times.

[5] To evaluate the importance of the variables for the clustering, we need to choose a classifier which does not modify the representation used to elaborate the reference clustering; i.e the data after the pre-processing step.

Table 1: Ranking of the variables

Index	Wine													Pima							
DB	V7	V6	V10	V1	V12	V13	V9	V8	V4	V5	V3	V2	V11	V8	V2	V1	V3	V6	V5	V4	V7
SD	V7	V6	V10	V1	V12	V9	V8	V13	V5	V4	V3	V2	V11	V8	V2	V1	V3	V6	V7	V4	V5
ARI-Tree	V7	V13	V12	V1	V10	V6	V11	V2	V9	V4	V8	V5	V3	V8	V2	V1	V3	V5	V6	V4	V7
ACC-Tree	V7	V13	V12	V1	V10	V6	V11	V2	V9	V4	V5	V8	V3	V8	V2	V1	V3	V5	V6	V4	V7

The PIMA data dataset contains $N = 768$ sample points described with $d = 8$ variables which are associated with two different classes. The ARI obtained between the reference clustering (using K-means algorithm, where $K = 2$) and the target class is equal to 0.11. Table 1 (right part) and Figure 2 present respectively the list of the ranked variables (from the most important to the least important) and the evolution of ARI curve for the three approaches (DB, SD and the proposed approach).

The results obtained on PIMA and WINE show that the proposed method is competitive with regards to DB and SD approaches on these two datasets.

4.3 Redundant variables

To test the ability of our approach to detect the redundant variables, we use the WAVEFORM dataset. This dataset consists of $n = 5000$ sample points described with 40 variables and associated with three different classes: only the first 21 variables are real attributes for this database and most of these are relevant to a classification problem whereas the last 19 variables are noisy standard centered Gaussian variables (for more details see [21], page 43 - 49). Figure 3 shows that the proposed method identifies the irrelevant set of variables $W = V1, V21 - V37, V39, V40$. The remaining variables are all relevant variables for the clustering.

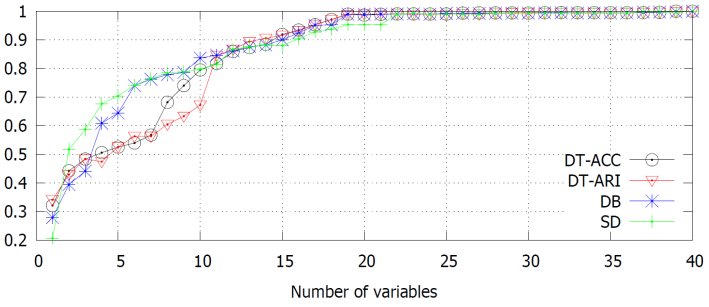

Fig. 3: Evolution of ARI criterion for the three methods (K=3)

To identify the redundant variables, we use the method described in Section 3.2. Table 2 presents the ARI values calculated between two pairs of relevant variables (the 20 variables identified by the proposed method using the ACC criterion). In this table, if we consider only the values above 0.96 to focus the attention on the high values of redundancy. The set of redundant variables is then : $R = V38, V2, V20, V19, V3$. Finally the set of relevant variables is $V = V4 - V18$. These obtained results are similar to those obtained using RD-MCM selection features method (see [19]). The ARI value obtained between the predicted ID-cluster using all variables (41 variables) and the predicted ID-cluster using the relevant variable (18 variables) is equal to 0.935.

Table 2: ARI values between pairs of relevant variables

	V7	V15	V8	V14	V16	V6	V13	V12	V17	V9	V5	V10	V4	V18	V11	V3	V19	V20	V2	V38
V7	1,00	0,51	0,42	0,42	0,39	0,41	0,41	0,38	0,37	0,36	0,35	0,34	0,34	0,34	0,32	0,32	0,32	0,32	0,32	0,32
V15		1,00	0,66	0,62	0,59	0,59	0,58	0,54	0,52	0,50	0,49	0,46	0,46	0,46	0,44	0,44	0,44	0,44	0,44	0,44
V8			1,00	0,76	0,72	0,70	0,67	0,62	0,59	0,56	0,54	0,51	0,51	0,51	0,49	0,48	0,48	0,48	0,48	0,48
V14				1,00	0,81	0,78	0,75	0,67	0,62	0,59	0,57	0,54	0,53	0,53	0,51	0,51	0,50	0,51	0,50	0,50
V16					1,00	0,84	0,77	0,71	0,66	0,61	0,60	0,56	0,56	0,56	0,53	0,53	0,52	0,53	0,52	0,52
V6						1,00	0,86	0,75	0,68	0,63	0,62	0,58	0,57	0,58	0,55	0,54	0,54	0,54	0,54	0,54
V13							1,00	0,79	0,72	0,66	0,65	0,60	0,60	0,60	0,58	0,57	0,57	0,57	0,57	0,57
V12								1,00	0,88	0,81	0,78	0,74	0,73	0,73	0,69	0,69	0,68	0,68	0,68	0,68
V17									1,00	0,89	0,84	0,81	0,79	0,79	0,75	0,74	0,74	0,74	0,74	0,74
V9										1,00	0,91	0,86	0,85	0,84	0,80	0,79	0,79	0,79	0,79	0,79
V5											1,00	0,89	0,87	0,86	0,83	0,82	0,81	0,82	0,82	0,82
V10												1,00	0,94	0,91	0,88	0,87	0,86	0,86	0,86	0,86
V4													1,00	0,94	0,89	0,88	0,87	0,87	0,87	0,87
V18														1,00	0,92	0,89	0,88	0,88	0,89	0,88
V11															1,00	0,95	0,93	0,92	0,92	0,92
V3																1,00	**0,96**	0,95	0,93	0,94
V19																	1,00	**0,97**	0,95	**0,96**
V20																		1,00	**0,97**	**0,97**
V2																			1,00	**1,00**
V38																				1,00

5 Conclusion

This paper has presented a supervised method to measure the importance of the variables used in a clustering. This method turned the problem into a supervised classification problem to sort variables according to their importance at the end of the clustering convergence. The experimental results corroborated the competitiveness of the method comparing to other methods from the literature. It has been incorporated successfully in the process of marketing service in the french Orange company. Future works will be done to incorporate the method in the convergence of the clustering algorithm and to measure the variables importance as a multivariate supervised classification problem.

References

1. Dy, J.G., Brodley, C.E.: Feature selection for unsupervised learning. J. Mach. Learn. Res. **5** (2004) 845–889

2. Liu, P., Zhu, J., Liu, L., Li, Y., Zhang, X.: Application of feature selection for unsupervised learning in prosecutors' office. In: Fuzzy Systems and Knowledge Discovery (FSKD. (2005) 35–38
3. Mitra, P., Murthy, C.A., Pal, S.K.: Unsupervised feature selection using feature similarity. IEEE Trans. Pattern Anal. Mach. Intell. **24**(3) (March 2002) 301–312
4. Vesanto, J., Ahola, J.: Hunting for correlations in data using the self-organizing map. In: ICSC Academic Press, Proceeding of the International ICSC Congress on Computational IntelligenceMethods and Applications (CIMA 99 (1999) 279285
5. Guérif, S., Bennani, Y., Janvier, E.: μ-SOM : Weighting features during clustering. In: Proceeding of the 5th workshop on self-organizing maps (WSOM05). (2005) 397–404
6. Halkidi, M., Batistakis, Y., Vazirgiannis, M.: On clustering validation techniques. J. Intell. Inf. Syst. **17**(2-3) (2001) 107–145
7. Hubert, L., Arabie, P.: Comparing partitions. Journal of classification **2**(1) (1985) 193–218
8. Larsen, B., Aone, C.: Fast and effective text mining using linear-time document clustering. In: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (ICDM), New York, NY, USA, ACM (1999) 16–22
9. Alok, A.K., 0001, S.S., Ekbal, A.: A min-max distance based external cluster validity index: MMI. In: HIS, IEEE (2012) 354–359
10. Davies, D.L., Bouldin, D.W.: A cluster separation measure. Pattern Analysis and Machine Intelligence, IEEE Transactions on **PAMI-1**(2) (1979) 224–227
11. Rousseeuw, P.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. (1) (1987) 53–65
12. Dunn, J.C.: A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. Journal of Cybernetics **3**(3) (1973) 32–57
13. Halkidi, M., Vazirgiannis, M., Batistakis, Y.: Quality scheme assessment in the clustering process. In: Principles of Data Mining and Knowledge Discovery. Volume 1910 of LNCS. Springer Berlin Heidelberg (2000) 265–276
14. Xie, X.L., Beni, G.: A validity measure for fuzzy clustering. IEEE Trans. Pattern Anal. Mach. Intell. **13**(8) (August 1991) 841–847
15. A.Raftery: A note on Bayes factors for log-linear contingency table models with vague prior information. Journal of the Royal Statistical Society (1986) 249–250
16. Jorge M., S., Mark, E.: On the use of the adjusted rand index as a metric for evaluating supervised classification. In: Proceedings of the International Conference on Artificial Neural Networks (ICANN). (2009) 1027–1035
17. Blake, C.L., Merz, C.J.: Uci repository of machine learning databases (1998)
18. MacQueen, J.B.: Some methods for classification and analysis of multivariate observations. In Cam, L.M.L., Neyman, J., eds.: Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability. Volume 1., University of California Press (1967) 281–297
19. Celeux, G., Martin-Magniette, M.L., Maugis, C., Raftery, A.E.: Comparing model selection and regularization approaches to variable selection in model-based clustering. Journal de la Société Française de Statistique (2014)
20. Arthur, D., Vassilvitskii, S.: K-means++: The advantages of careful seeding. In: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms. SODA '07 (2007) 1027–1035
21. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and regression trees. Wadsworth International Group (1984)