Predictive K-means with local models Supplementary material

Vincent Lemaire[1], Oumaima Alaoui Ismaili[1],
Antoine Cornuéjols[2], Dominique Gay[3]

[1] Orange Labs, Lannion, France
(vincent.lemaire,oumaima.alaouiismaili)@orange.com
[2] UMR MIA-Paris, AgroParisTech, Université Paris-Saclay, 75005, Paris, France
antoine.cornuejols@agroparistech.fr
[3] LIM-EA2525, Université de La Réunion
dominique.gay@univ-reunion.fr

Abstract. We give here a supplementary material concerning the paper: "Predictive K-means with local models", V. Lemaire et al. Workshop "Learning Data Representation for Clustering" (LDRC) held at Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD) 2020.

1 Second axis: Comparison with related works to predictive clustering

In the literature, the terms "predictive clustering" or "supervised clustering" cover different kinds of methods. Briefly, we may distinguish between (*i*) supervised clustering algorithm, (*ii*) methods that try to fit simultaneously the clustering and linear regression as local model (clusterwise regression), and (*iii*) methods that incorporate a partial knowledge on the labels aka. semi-supervised clustering.

In the *first type of approaches*, the algorithms seek to maximize the purity of the clusters in terms of the labels of the instances that belong to them under the constraint that the clusters have high intra similarity and high extra dissimilarity. For example AL-Harbi et al. ([2]) suggest replacing the standard Euclidean distance by a ponderated Euclidean distance to differentiate the significances of the features in relation to each cluster. A weight is assigned to each feature and it estimated through the maximization of the percentage of the well classified instances. In ([1]), Aguilar et al. present a new predictive clustering algorithm based on the nearest neighbor method. The algorithm starts by initializing m clusters with m the number of instances. Then, clusters with identical neighbors are merged (i.e, the nearest instances which have the same target class). In ([7]) four algorithms of predictive clustering are proposed by Rick et al.

In the *second type of approaches*, clusterwise methods aim at partitioning m instances into K clusters, where each group is characterized by it specific regression coefficients in a linear regression model. The idea is to assume that

there is an underlying group structure of the instances and that each cluster can be revealed by the fit of a specific regression model. Formally, the clusterwise approach aims to group instances with similar regression effects into K clusters such that the overall sum of squared residuals within clusters is minimal. Several algorithms have been proposed to achieve this goal and to estimate the regression coefficients, see for instance ([6, 9]).

The *third type of approaches* is well illustrated by the MPCK-MEANS algorithm ([3]). It uses examples for which the labels are known. MPCK-MEANS is a semi-supervised clustering algorithm derived from k-means that incorporates both metric learning and the use of pairwise constraints in a principled manner.

To gain additional insights about the performances of the algorithm PKM_{SNB}, we compare it here with four well-known algorithms of "predictive clustering": SRIDHCR ([7]), a supervised clustering editing process proposed by Al-Harbi ([2]) (K-means with metric learning), MPCK-Means (k-means with contraints) ([3]) and some preliminary results compared to the recent algorithm COBRA ([10], [11]). For fair comparisons, we retained the values of the parameters of these algorithms which give the best results.

Accuracy comparisons with SRIDHCR ([7])			
Datasets	**K**	**SRIDHCR**	**PKM$_{\mathbf{SNB}}$**
Glass	6	0.63	0.94
Heart	2	0.74	0.82
Iris	3	0.97	0.93
Accuracy comparisons with Al-Harbi's ([2])			
Datasets	**K**	**Al-Harbi**	**PKM$_{\mathbf{SNB}}$**
Auto	2	0.92	0.85
Breast	2	0.97	0,97
Pima	2	0.74	0.74
Pairwise F-measure comparisons with MPCK-Means ([3])			
Datasets	**K**	**MPCK Means**	**PKM$_{\mathbf{SNB}}$**
Iris	3	0.96	0.95
Digit 389	3	0.96	0.98
Wine	3	0.96	0.98
Letter IJL	3	0.82	0.84
Ionosphere	2	0.76	0.86
ARI comparisons with COBRA ([10])			
Datasets	**K**	**COBRA**[4]	**PKM$_{\mathbf{SNB}}$**
Segmentation	7	0.85	0.66
Glass	6	0.40	0.82
Breast	2	0.80	0.88

Table 1. comparaison of our algorithm with 4 algorithms using different criteria.

Table 1 reports the predictive performance of our method PKM_{SNB} along with the performances of the four algorithms mentioned above. We used the same experimental conditions than those employed by the authors of these papers.

We also used the same datasets and the performance criterion they used in their papers[5]. The results show that our algorithm, PMK_{SNB}, is very competitive[6].

2 Methodology to use the algorithm: interpretable power

The above discussion centered on the predictive performances of the algorithms and the ease of use of $\texttt{PKM}_{\texttt{SNB}}$. This section deals with the ability of the algorithm to offer interpretable clusters. For instance, the Vehicle database is interesting. It contains 846 instances described by 18 explanatory features and one target descriptor with 4 classes (*bus*, *opel*, *saad* and *van*). In this illustrative study, the entire data set is used in the training phase with $K = J = 4$. The meaning of the features can be found in the UCI repository [8]. For lack of space, only the 6 most informative and important features (`A5`, `A6`, `A15`, `A16`, `A10`, `A12`) in the building of the initial clustering are presented here.

The interpretation of the $\texttt{PKM}_{\texttt{SNB}}$ model is based on a two-level analysis (see figure 1).

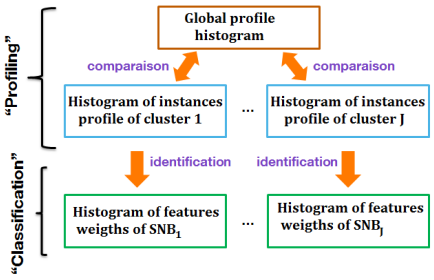

Fig. 1. A two-level analysis.

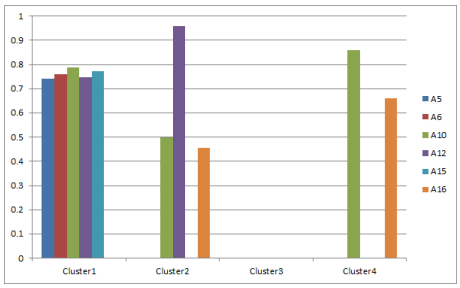

Fig. 2. Histogram : variable weights.

The first analysis consists in analyzing the profile of each cluster using histograms. Figure 3 thus presents both the average profile of the overall population (each bar representing the percentage of instances having a value of the corresponding interval). The number of intervals and the thresholds are obtained during the preprocessing step which computes the redescription of the data.

This visualization is easy to understand and allows the method to be widely usable. For instance, it can be seen that the feature `A12` is highly discriminating for 'cluster 4' because 100% of the instances belonging to this cluster have `A12` <

[5] That means that we did not reconduct their experiments but we just draw the values in their paper

[6] Note : Here again we assumed that the number of clusters is equal to the number of classes. This is a limiting assumption. We think that our method might works even better if we lose that assumption.

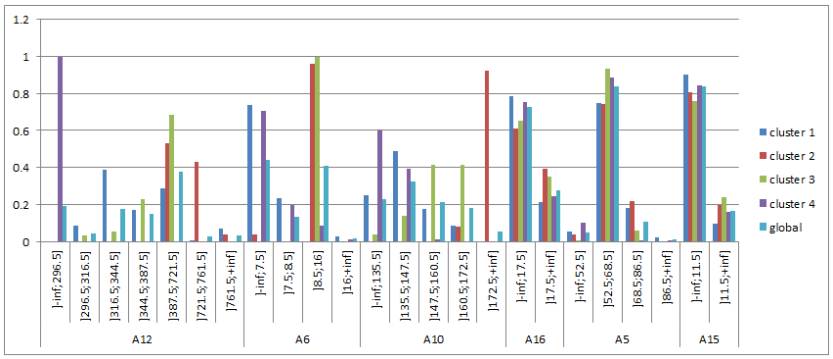

Fig. 3. Histogram: mean profiles of the persons globally and for each cluster.

296.5. The attributes in the histogram presented in Figure 3 are sorted[7] allowing the user to focus on the important attributes (when the number of attribute is large). For instance, the variable `A12` is more discriminative to find "good" clusters than `A6` and others. `A15` is the last one, the less discriminative as is apparent in Figure 3.

The second analysis allows one to have an idea about the features importance in each cluster and thus to know the reason of the local classification. In this case, a histogram is provided for each cluster. It gives the weight of the features of the local SNBs (see figure 2 which gives the variable weights (A5, A6, A10, A12, A15, A16) for the local models in each cluster. These weights correspond to the terms W_i in the equation of the SNB. For instance, from this figure, we can see that: *i*) the cluster 3 has only 0 as a value of weight for all used features. That means that the majority vote has been retained. *ii*) the features in the cluster 1 have very close weights and *iii*) the feature A12 is the most important in the cluster 2 etc. This second level of analysis contains elements which are not available in the majority vote approach $\texttt{PKM}_{\texttt{MV}}$.

One of the benefits of incorporating local model into predictive K-Means algorithm is to improve its descriptive power (results interpretation). Indeed, the local models allow one to know, in the training phase, the features, which contribute the most both to the global clustering construction, and to the construction of each cluster. Therefore, it is easy to know the reasons behind the prediction of each new instance.

References

1. Aguilar-Ruiz, J.S., Ruiz, R., Santos, J.C.R., Giráldez, R.: SNN: A supervised clustering algorithm. In: International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems. pp. 207–216 (2001)

[7] The supervised preprocessing methods ([4,5]) that we use provides a criterion which expresses the ability of each attribute to predict the class when estimating the conditional distribution.

2. Al-Harbi, S.H., Rayward-Smith, V.J.: Adapting k-means for supervised clustering. Applied Intelligence **24**(3), 219–226 (2006)
3. Bilenko, M., Basu, S., Mooney, R.J.: Integrating constraints and metric learning in semi-supervised clustering. In: Proceedings of the Twenty-first International Conference on Machine Learning (ICML) (2004)
4. Boullé, M.: A Bayes optimal approach for partitioning the values of categorical attributes. Journal of Machine Learning Research **6**, 1431–1452 (2005)
5. Boullé, M.: MODL: a Bayes optimal discretization method for continuous attributes. Machine Learning **65**(1), 131–165 (2006)
6. DeSarbo, W.S., Cron, W.L.: A maximum likelihood methodology for clusterwise linear regression. Journal of Classification **5**(2) (1988)
7. Eick, C.F., Zeidat, N., Zhao, Z.: Supervised clustering - algorithms and benefits. In: International Conference on Tools with Artificial Intelligence. pp. 774–776 (2004)
8. Lichman, M.: UCI machine learning repository (2013), http://archive.ics.uci.edu/ml
9. Saporta, G.: Clusterwise methods, past and present. In: 61 World Statistics Congress. Marrakech, Maroc (July 2017)
10. Van Craenendonck, T., Dumancic, S., Blockeel, H.: COBRA: A fast and simple method for active clustering with pairwise constraints. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI. pp. 2871–2877 (2017)
11. Van Craenendonck, T., Dumancic, S., Van Wolputte, E., Blockeel, H.: COBRAS: fast, iterative, active clustering with pairwise constraints. In: Proceedings of Intelligent Data Analysis (2018)