ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2.
Available from http://www.i6doc.com/en/.

Sparse k-means for mixed data via group-sparse clustering

Marie Chavent[1], Jerome Lacaille[2], Alex Mourer[1,2,3], and Madalina Olteanu[3]

1- INRIA Bordeaux Sud-Ouest
CQFD team - France

2- Safran Aircraft Engines - Datalab
Villaroche - France

3- SAMM - EA 4543
Université Pantheon Sorbonne - France

Abstract. The present manuscript tackles the issue of variable selection for clustering, in high dimensional data described both by numerical and categorical features. First, we build upon the sparse k-means algorithm with *lasso* penalty, and introduce the group-L_1 penalty – already known in regression – in the unsupervised context. Second, we preprocess mixed data and transform categorical features into groups of dummy variables with appropriate scaling, on which one may then apply the group-sparse clustering procedure. The proposed method performs simultaneously clustering and feature selection, and provides meaningful partitions and meaningful features, numerical and categorical, for describing them.

1 Introduction

Whereas the issue of feature selection through regularization procedures received a great deal of attention in the supervised learning context and resulted in an abundant literature over the last twenty years, it is only much later and relatively recently that it effectively emerged in the unsupervised framework. The first approaches were model-based, these being naturally suited for including *lasso* (L_1) and related penalties, and one may cite [1] for a L_1-penalized EM procedure (the mixture consists of Gaussian distributions with equal variances) or [2] for a detailed review on model-based clustering for high-dimensional data. In the more general framework where no assumption is made on the underlying distribution, a sparse k-means algorithm with L_1 penalty was introduced in [3], and later on extended to feature selection within each cluster and reinforced by consistency results, [4] [5] [6]. Let us also mention that a generalization of the sparse k-means algorithm to overlapping groups of variables was recently introduced in [7] .

That being said, all methods cited above are essentially designed for numerical data, while real data is often made of numerical and categorical features. Some of the authors above touch upon the question of categorical features, by mentioning the possibility of making them numerical using a transformation through dummy variables. However, this processing step is not that immediate, since the Euclidean distance on zero-one vectors is not particularly suited for being mixed with Euclidean distances on numerical variables. Other authors

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2.
Available from http://www.i6doc.com/en/.

implicitly suggest that the proposed algorithms may be written in terms of distances or dissimilarities between input data only, and hence it suffices to use an appropriate distance for categorical features. Nevertheless, the distance-based approaches may rapidly translate into an increased complexity if the size of the data becomes large.

The present manuscript aims at proposing an explicit method for variable selection in a mixed-data framework, using a penalized criterion. Our contribution is two folded: first, we introduce a group-sparse version of the k-means algorithm, which allows one to select among priorly defined groups of numerical variables; second, we use the preprocessing proposed in [8] for mixed data and write each categorical feature as a group of properly scaled numerical features, on which we apply the group-sparse k-means. Eventually, starting from a data set described by both numerical and categorical features, we simultaneously perform k-means clustering and feature selection on both types of features.

2 Method and data preprocessing

We start by introducing the group-sparse k-means algorithm, followed by the particular writing of the problem in the case of mixed data.

For the moment, let us suppose that the data $\mathbf{X}$ consists of n input vectors valued in $\mathbb{R}^p$, and that $X_j \in \mathbb{R}^n$ represents the j-th feature (all features are scaled to zero mean and unit variance). Furthermore, the data may be partitioned into K clusters, with n_k the number and C_k the set of indices of the observations in the k-th cluster. Then, one may define $\overline{X}_{j,k} = \frac{1}{n_k} \sum_{i \in C_k} X_{i,j}$ the average of feature j in cluster k, and $\overline{X}_j = \frac{1}{n} \sum_{i=1}^{n} X_{i,j}$ the average of the jth feature over the entire data set.

Since the k-means algorithm aims at finding the partition $C_1, ..., C_K$ maximizing the between-class variance, let us write this as:

$$\mathcal{B}(\mathbf{X}, C_1, ..., C_K) = \sum_{k=1}^{K} \frac{n_k}{n} \sum_{j=1}^{p} (\overline{X}_{j,k} - \overline{X}_j)^2 = \sum_{j=1}^{p} \mathcal{B}(X_j, C_1, ..., C_K) \ , \qquad (1)$$

where $\mathcal{B}(X_j, C_1, ..., C_K)$ is the between-class variance associated to the j-th feature. In the following, in order to simplify the notations, we shall use $b_j = \mathcal{B}(X_j, C_1, ..., C_K)$, $j = 1, ..., p$.

If one suspects that features do not equally contribute to the clustering, some being more informative than others, she may address this by introducing a weighted version of the between-class variance, penalized by a regularization term:

$$\mathcal{B}(\mathbf{X}, C_1, ..., C_K, \mathbf{w}) = \sum_{j=1}^{p} w_j b_j - \lambda h(\mathbf{w}) = \mathbf{w}^T \mathbf{b} - \lambda h(\mathbf{w}) \ , \qquad (2)$$

where $\mathbf{w}^T = (w_1, ..., w_p)$, $\mathbf{w} \geq 0$, $\|\mathbf{w}\|_2 \leq 1$ is the vector of weights, and $\mathbf{b}^T = (b_1, ..., b_p)$ are the feature-wise between-class variances, for a given partition $C_1, ..., C_K$, and λ is a hyper parameter. In the following, K the number of

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

clusters as well as λ are supposed to be priorly fixed. In the *sparse clustering* algorithm introduced in [3], the regularization term is chosen by analogy with the *lasso* framework, hence $h(\mathbf{w}) = \|\mathbf{w}\|_1$. This leads to a sparse representation of the variables, with the weights w_j directly translating the contribution of the associated features to the clustering.

2.1 Group sparse k-means

We generalize the original *sparse clustering* by introducing a group regularization framework. Suppose that the p features are furthermore divided into L priorly known groups, such that $\mathbf{X} = \left[\mathbf{X}^1|...|\mathbf{X}^L\right]$, with $\mathbf{X}^\ell \in \mathbb{R}^{n\times p_\ell}$, p_ℓ being the size of group ℓ, and $p_1 + ... + p_L = p$. The between-class variance vector $\mathbf{b}$ and the weight vector $\mathbf{w}$ can be also decomposed as $\mathbf{b}^T = (\mathbf{b}_1, ..., \mathbf{b}_L)$ and $\mathbf{w}^T = (\mathbf{w}_1, ..., \mathbf{w}_L)$, where $\mathbf{b}_\ell = (b_{p_0+...+p_{\ell-1}+1}, ..., b_{p_0+...+p_\ell})^T$ and $\mathbf{w}_\ell = (w_{p_0+...+p_{\ell-1}+1}, ..., w_{p_0+...+p_\ell})^T$, with the notation $p_0 = 0$.

For group data, let us define a specific L_1-group penalty, which has been already used in the regression framework [9],

$$h(\mathbf{w}) = \|\mathbf{w}\|_{1,group} = \sum_{\ell=1}^{L} \alpha_\ell \|\mathbf{w}_\ell\|_2 \ , \tag{3}$$

where $(\alpha_\ell)_\ell$ is a vector of weights applied to the groups of variables. In the literature on group-sparse regression, two common choices appear to emerge, either $\alpha_\ell = 1$, $\forall \ell = 1, ..., L$, or $\alpha_\ell = \sqrt{p_\ell}$ $\forall \ell = 1, ..., L$. The latter is penalizing each group by its size, and we shall use it in the following.

With the previous notations, the optimization problem writes as follows:

$$\begin{cases} \max_{\mathbf{w}, C_1, ..., C_K} \mathbf{w}^T\mathbf{b} - \lambda \sum_{\ell=1}^{L} \sqrt{p_\ell}\|\mathbf{w}_\ell\|_2 \ , \\ \|\mathbf{w}\|_2 \leq 1 \ , \ \mathbf{w} \geq 0 \end{cases} \tag{4}$$

For a fixed number of clusters K and for a fixed hyper parameter λ, the problem in Equation 4 is solved using an iterative algorithm. After having set an initial value for the weight vector $\mathbf{w}$ (generally one uses weights with equal norms for the L groups), one is alternating the two following steps until convergence:

1. Keeping $\mathbf{w}$ fixed, find $C_1, ..., C_K$ maximizing $\mathbf{w}^T\mathbf{b}$, which is actually equivalent to performing a usual k-means training on the scaled features $\tilde{X}_j = \sqrt{w}_j X_j$, $j = 1, ..., p$.

2. Keeping $C_1, ..., C_K$ fixed, find $\mathbf{w} \geq 0$ maximizing $\mathcal{B}(\mathbf{X}, C_1, ..., C_K, \mathbf{w})$ and such that $\|\mathbf{w}\|_2 \leq 1$. The solution may be analytically obtained using Lagrange multipliers and the KKT conditions, and may be expressed in terms of group soft-thresholding operators:

$$\mathbf{w}^\star = \begin{cases} \frac{\mathbf{b}}{\|\mathbf{b}\|_2} & \text{if} \quad \lambda = 0, \\ 0 & \text{if} \quad \lambda > 0 \text{ and } \|\tilde{S}_G(\mathbf{b}, \sqrt{\mathbf{p}}\lambda)\|_2 = 0, \\ \frac{\tilde{S}_G(\mathbf{b}, \sqrt{\mathbf{p}}\lambda)}{\|\tilde{S}_G(\mathbf{b}, \sqrt{\mathbf{p}}\lambda)\|_2} & \text{if} \quad \lambda > 0 \text{ and } \|\tilde{S}_G(\mathbf{b}, \sqrt{\mathbf{p}}\lambda)\|_2 \neq 0, \end{cases} \tag{5}$$

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

where

$$\tilde{S}_G(\mathbf{b}, \sqrt{\mathbf{p}}\lambda)^T = \left(S_G(\mathbf{b_1}, \sqrt{p_1}\lambda)^T, ..., S_G(\mathbf{b_L}, \sqrt{p_L}\lambda)^T\right) \in \mathbb{R}^p, \tag{6}$$

$\sqrt{\mathbf{p}}^T = (\sqrt{p_1}, ..., \sqrt{p_L})$, and $S_G(\mathbf{b}_\ell, \sqrt{p_\ell}\lambda) = \frac{\mathbf{b}_\ell}{\|\mathbf{b}_\ell\|_2}\left(\|\mathbf{b}_\ell\|_2 - \sqrt{p_\ell}\lambda\right)_+ \in \mathbb{R}^{p_\ell}$, $\ell = 1, ..., L$, is the group soft-thresholding operator.

The group soft-thresholding operator will take out from the model all groups which have a norm of the corresponding between-class variance $\mathbf{b}_\ell$ smaller than the fixed threshold λ normalized by the size of the group, and shrink by the same amount the norms of the remaining groups of variables.

The hyper parameter λ may be tuned using various criteria for assessing the quality of the model, such as the gap statistic as described in [3], the ratio of explained variance, etc. Usually, one considers a fine grained grid valued between zero and some upper bound chosen as the maximum $\|\mathbf{b}_\ell\|_2$ when training a simple k-means procedure. Choosing the optimal λ and the optimal number of clusters K is not trivial, but, because of the reduced number of pages, we do not discuss here the different criteria in the literature, nor how they fit in the group-sparse framework.

2.2 Sparse clustering for mixed data

Eventually, let us discuss the case where the data is summarized by both numerical and categorical features. If one has n input data $\mathbf{X}$ described by d_1 categorical features and d_2 numerical ones, such that each of the categorical features has p_j possible values, $j = 1, ..., d_1$, she may transform each categorical feature X_j into p_j dummy variables $\tilde{\mathbf{X}}_j = (\tilde{X}_j^1, ..., \tilde{X}_j^{p_j}) \in \{0, 1\}^{n \times p_j}$ and thus define a natural group structure on the transformed data $\mathbf{Y} = [\mathbf{Y}_1|...|\mathbf{Y}_{d_1+d_2}]$, where $\mathbf{Y}_\ell = \tilde{\mathbf{X}}_\ell$ for $\ell = 1, ..., d_1$, $\mathbf{Y}_\ell = X_\ell$ for $\ell = d_1 + 1, ..., d_1 + d_2$, with respective group sizes $\mathbf{p}^T = (p_1, ..., p_{d_1}, 1, ..., 1) \in \mathbb{R}^{d_1+d_2}$.

When training group-sparse k means on the group structure above, this amounts to performing variable selection in a mixed-data context. Before applying the algorithm described in the previous section, $\mathbf{Y}$ must be properly preprocessed. This prior step is described for example in [8]: numerical variables are scaled to zero mean and unit variance, while the dummy variables are centered and normalized by $1/\sqrt{\frac{n}{n_{l,j}}}$, where $n_{l,j}$ is the number of input data taking the j-th value of the lth feature, or equivalently the sum over $\tilde{X}_\ell^j$. The scaling applied to the dummy variables actually leads to using a χ^2 distance on the categorical features, while the numerical features, after scaling, are compared through the usual Euclidean distance.

3 A real-life example

For illustrating the algorithm introduced above, we use the Statlog Heart dataset available in the UCI machine learning data repository [10]. The data consists of 270 inputs and is described by six numerical features (*age*, resting blood pressure

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

- *rest.b.p*, serum cholesterol in mg/dl - *ser.c*, maximum heart rate achieved - *max.h.r.a*, ST depression induced by exercice relative to rest - *oldp*, number of major vessels colored by fluoroscopy - *num.m.v*) and seven categorical ones (*sex*, chest pain type - *chest.p.t*, fasting blood sugar > 120mg/dl - *fast.b.s*, resting electrocardiographic results - *rest.e.r*, exercice induced angina - *exe.i.a*, the slope of the peak exercice ST segment - *slope*, thalassemia -*thal*). One control variable assesses the presence or the absence of a heart disease and gives an a priori partition in two clusters. We used this knowledge to select the number of clusters $K = 2$.

The hyper parameter λ was varied on a grid taking values between 0 and the maximum between-class feature variance, extracted from a k-means algorithm trained without weighting. The regularization paths are illustrated in Figure 1. At the optimum value of the hyper parameter $\lambda^\star$ (selected using a heuristic on the ratio of explained variance), six features only are kept, four numerical, and two categorical. The averages within cluster for the numerical features, and the frequencies within cluster for the categorical ones, are displayed in Table 1 and show that all selected features are significantly different in the clusters. The method thus provides both meaningful clusters, and meaningful features - both numerical and categorical - for describing the clusters. Let us remark here also that the ratio of explained variance, all features included, varies very little when using the partition at the optimal value of λ: only 2% is lost when using the six selected features instead on the initial thirteen.

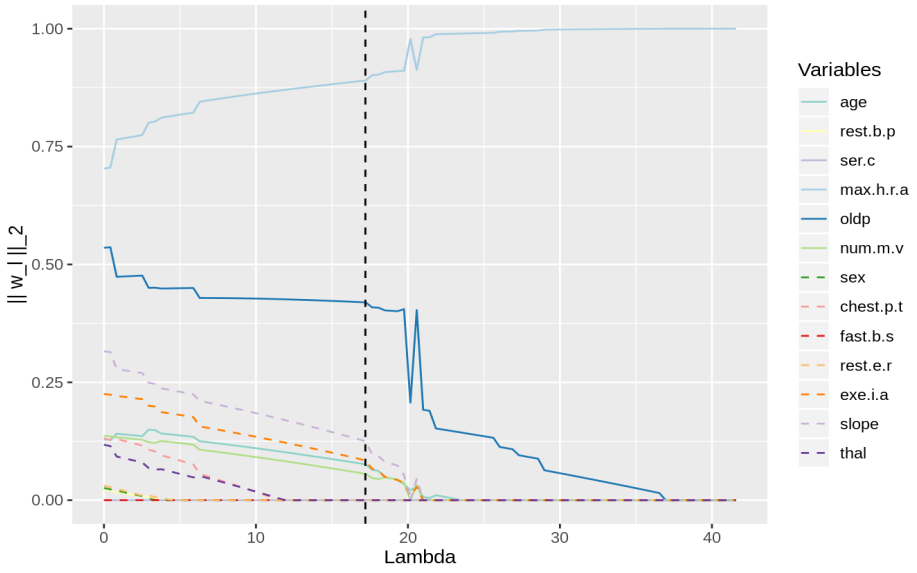

Fig. 1: Feature regularization paths for the heart data. The optimum selected λ is plotted in the vertical dotted black line. The y axis is $\|\mathbf{w}_\ell^\star\|_2$, the norm of the optimal weights associated to each feature.

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

Feature	Cluster 1	Cluster 2	Overall statistics
Maximum heart rate (*max.h.r.a*)	127.1	164.2	149.7
ST depression (*old*)	1.85	0.53	1.05
Slope of the peak (*slope*) - lev. 1	15.1%	69.5%	48.1%
Slope of the peak (*slope*) - lev. 2	73.6%	26.8%	45.2%
Slope of the peak (*slope*) - lev. 3	11.3%	3.7%	6.7%
Ex. induced angina (*exe.i.a*) - lev. 1	41.5%	83.5%	67.0%
Ex. induced angina (*exe.i.a*) - lev. 2	58.5%	16.4%	33.0%
Age (*age*)	58.2	52.0	54.4
Number of major vessels (*num.m.v*)	1.03	0.43	0.67

Table 1: Overall and within clusters average values and average frequencies of the selected features for the optimum hyper parameter $\lambda^\star$. Features are ordered by the decreasing norm of their coefficients, $\|w_\ell^\star\|_2$. All features are significantly different in the two clusters.

4 Conclusion and perspectives

We provide a complete procedure for simultaneously clustering mixed data, and selecting the most relevant features for the clustering. This procedure is illustrated on a real dataset and proves to be efficient. Further work is currently done on finding appropriate criteria for selecting the hyper parameter λ and the number of clusters K. An **R**-package implementing this method is also currently being developed and will be soon made available for the community.

References

[1] W. Pan and X. Shen. Penalized model-based clustering with application to variable selection. *Journal of Machine Learning Research*, 8(May):1145–1164, 2007.

[2] Ch. Bouveyron and C. Brunet-Saumard. Model-based clustering of high-dimensional data: A review. *Computational Statistics & Data Analysis*, 71:52–78, 2014.

[3] D.M. Witten and R. Tibshirani. A framework for feature selection in clustering. *Journal of the American Statistical Association*, 105(490):713–726, 2010. PMID: 20811510.

[4] E. Arias-Castro and X. Pu. A simple approach to sparse clustering. *Computational Statistics & Data Analysis*, 105:217–228, 2017.

[5] S. Chakraborty and S. Das. A strongly consistent sparse k-means clustering with direct l_1 penalization on variable weights. *arXiv preprint arXiv:1903.10039*, 2019.

[6] W. Sun, J. Wang, Y. Fang, et al. Regularized k-means clustering of high-dimensional data and its asymptotic consistency. *Electronic Journal of Statistics*, 6:148–167, 2012.

[7] Z. Huo and G. Tseng. Integrative sparse k-means with overlapping group lasso in genomic applications for disease subtype discovery. *Ann. Appl. Stat.*, 11(2):1011–1039, 06 2017.

[8] M. Chavent, V. Kuentz-Simonet, A. Labenne, and J. Saracco. Multivariate analysis of mixed data: The pcamixdata r package. *arXiv preprint arXiv:1411.4911*, 2014.

[9] Ming Y. and Yi L. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68:49–67, 2006.

[10] A. Frank and A. Asuncion. UCI machine learning repository, statlog (heart) data set, 2010.