Unsupervised variable selection for kernel methods in systems biology

Jérôme Mariette [1], Céline Brouard [1], Rémi Flamary [2], and Nathalie Vialaneix [1]

[1] MIAT, Université de Toulouse, INRA, 31326 Castanet-Tolosan, France

[2] OCA Laboratoire Lagrange, Université Côte d'Azur, CNRS, 06000 Nice, France

Introduction

Kernel methods have proven to be useful and successful to analyse large-scale multi-omics datasets [Schölkopf et al., 2004]. However, as stated in [Hofmann et al., 2015, Mariette et al., 2017], these methods usually suffer from a lack of interpretability as the information of thousands descriptors is summarized in a few similarity measures, that can be strongly influenced by a large number of irrelevant descriptors.

To address this issue, feature selection is a widely used strategy: it consist in selecting the most promising features during or prior the analysis. However, most existing methods are proposed in a supervised framework [Tibshirani, 1996, Robnik-Šikonja and Kononenko, 2003, Lin and Tang, 2006]. In the unsupervised framework, the number of proposals is much less important, because there is no objective criterion or value on which to tune the quality of a given feature. Proposals thus aim at preserving at best the similarities between individuals like the **SPEC** approach [Zhao and Liu, 2007] or at recovering a latent cluster structure, like **MCFS** [Cai et al., 2010], **NDFS** [Li et al., 2012] and **UDFS** [Yang et al., 2011].

In this communication, we will present a feature selection algorithm that explicitly takes advantage of the kernel structure in an unsupervised fashion.

Method

In the following, we consider a set of n observations $(\mathbf{x}_i)_{i=1,\dots,n}$, taking values in $\mathbb{R}^p$ ($\mathbf{x}_i = (x_{ij})_{j=1,\dots,p}$) and described by a kernel, K, such that $K : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ is symmetric ($\forall \mathbf{x}, \mathbf{x}' \in \mathbb{R}^p,\ K(\mathbf{x}, \mathbf{x}') = K(\mathbf{x}', \mathbf{x})$) and positive ($\forall\, N \in \mathbb{N}$, $\forall\, (\alpha_i)_{i=1,\dots,N} \subset \mathbb{R}$, $\forall\, (\mathbf{x}_i)_{i=1,\dots,N} \subset \mathbb{R}^p$, $\sum_{i,i'=1}^N \alpha_i \alpha_{i'} K(\mathbf{x}_i, \mathbf{x}_{i'}) \geq 0$). In the sequel, we will denote $K_{ii'} = K(\mathbf{x}_i, \mathbf{x}_{i'})$ and $\mathbf{K}$ the symmetric definite positive $(n \times n)$-matrix with entries $(K_{ii'})_{i,i'=1,\dots,n}$. The feature map associated with K is $\phi : \mathbb{R}^d \to \mathcal{H}$, where $\mathcal{H}$ is the unique Hilbert space that verifies

$$\forall \mathbf{x}, \mathbf{x}' \in \mathbb{R}^p,\ K(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle_{\mathcal{H}}.$$

The variable selection problem can be formulated by introducing a vector of p variables $\mathbf{w} = (w_j)_{j=1,\dots,p}$, taking values in $\{0,1\}^p$ and such that $w_j = 1$ is equivalent to select variable j. A new kernel matrix, $\mathbf{K}^w$, can be defined from K and w by:

$$K^w(\mathbf{x}_i, \mathbf{x}_{i'}) := K(\mathbf{w} \cdot \mathbf{x}_i, \mathbf{w} \cdot \mathbf{x}_{i'}),$$

in which " $\cdot$ " is the elementwise multiplication: $\mathbf{w} \cdot \mathbf{x} := (w_1 x_1, \dots, w_p x_p)^T = \mathrm{Diag}(\mathbf{w})\mathbf{x}$. K^w is the restriction of K to the d variables selected through the definition of $\mathbf{w}$. This gives a natural way to choose $\mathbf{w}$ by searching for values that minimize the distorsion of the original kernel K, as measure by *e.g.* the Frobenius norm:

$$\mathbf{w}^* := \underset{\mathbf{w} \in \{0,1\}^p}{\operatorname{argmin}} \|\mathbf{K}^w - \mathbf{K}\|_F^2 \qquad \text{for } \mathbf{w} \text{ such that } \sum_{j=1}^p w_j \leq s$$

for a given chosen s controlling the sparsity of the solution.

However, when p is large, this problem is hard to solve. To address such problems, [Grandvalet and Canu, 2002] and [Allen, 2013] described approaches using an ℓ_1 penalization that produces a sparse solution. In this paper, we propose to extend them to the unsupervised setting and call the method **UKFS**. More precisely, the problem writes:

$$\mathbf{w}^* := \underset{\mathbf{w} \in (\mathbb{R}^+)^p}{\operatorname{argmin}} \|\mathbf{K}^w - \mathbf{K}\|_F^2 + \lambda \|\mathbf{w}\|_1, \tag{1}$$

in which $\lambda > 0$ is a penalization parameter that controls the trade-off between the minimization of the distorsion and the sparsity of the solution and $\|.\|_1$ is the ℓ_1 norm: $\|\mathbf{z}\|_1 := \sum_{j=1}^p |z_j|$. We propose an efficient gradient based algorithm to solve this problem (not detailed in this abstract).

Results and discussion

To compare our approach against state-of-the-art approaches, *i.e.* **lapl**, **SPEC**, **MCFS**, **NDFS** and **UDFS**, two microarray datasets and a DNA barcoding dataset were analysed on which a ground truth clustering structure is known.

"Carcinom" and "Glioma" datasets respectively contain the expression of 9,182 genes obtained from 174 samples and 4,434 genes from 50 samples. To perform the feature selection on these datasets, **UKFS** was used with the Gaussian kernel $K_{ii'} = e^{-\sigma^* \|\mathbf{x}_i - \mathbf{x}_{i'}\|^2}$ with σ^* chosen so as to minimize the reproduced inertia in the projection on the first two axes of the KPCA with kernel K. "Koren" includes the abundance of 973 operational taxonomic units (OTUs) collected from 43 samples. To address the underlying compositional structure of such dataset, standard pre-processing steps, *i.e.*, total sum scaling normalisation (TSS) and centred log ratio transformation (CLR), were applied before selecting the relevant features

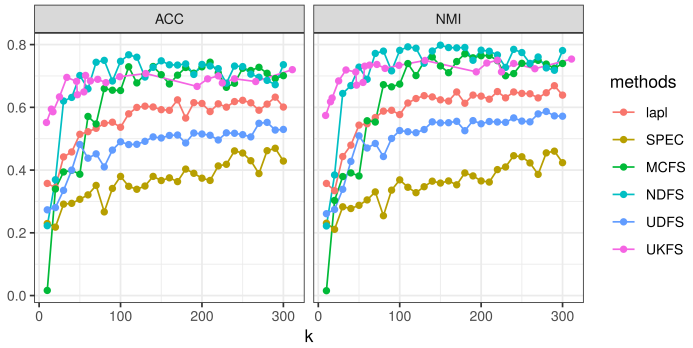

Figure 1: Comparison of the different approaches performances in terms of ACC (left) and NMI (right) computed from kernel k-means results using only k selected features. Presented results are obtained on the "Carcinom" dataset.

with **SPEC**, **MCFS**, **NDFS** and **UDFS**. This pre-processing step is not required by **UKFS** which computes a kernel based on the Bray-Curtis dissimilarity between samples on raw abundances, $d_{BC}(\mathbf{x}_i, \mathbf{x}'_i) = \frac{\sum_{s=1}^{p} |\mathbf{x}_{is} - \mathbf{x}_{i's}|}{\sum_{s=1}^{p} (\mathbf{x}_{is} + \mathbf{x}_{i's})}$, with p the number of OTUs observed.

Methods are evaluated on their ability to recover the dataset underlying classification structure using only a small number of features that they have selected. The true partition is used as ground truth to compute standard clustering performance metrics, *i.e.*, the normalized mutual information (NMI, [Danon et al., 2005]) and the overall accuracy (ACC). Note that our approach is not specifically optimized for this type of problem, contrary to **MCFS**, **NDFS** and **UDFS** which explicitely have a cluster structure assumption and for which we set C, the *a priori* number of clusters of the method, to its true value ($C = 11$ for "Carcinom", $C = 4$ for "Glioma" and $C = 3$ for "Koren").

Results demonstrate a high efficiency of our approach to select features relevant to summarize the structure of the data, in a reasonable computational time and with no a priori on a cluster organization of the data. For the three tested datasets, **UKFS** is in the range of or surpasses results obtained with other methods. More precisely, Figure 1 shows that **UKFS** selects variables allowing to produce clustering with a quality fairly similar to those obtained by two methods designed for such purpose, *i.e.*, **NDFS** and **MCFS**. This observation is confirmed by the results obtained on the other datasets and future work will investigate the biological relevance of selected features.

References

[Allen, 2013] Allen, G. I. (2013). Automatic feature selection via weighted kernels and regularization. *Journal of Computational and Graphical Statistics*, 22(2):284–299.

[Cai et al., 2010] Cai, D., Zhang, C., and He, X. (2010). Unsupervised feature selection for multi-cluster data. In Rao, B., Krishnapuram, B., Tomkins, A., and Yang, Q., editors, *Proceedings of the 16th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD 2010)*, pages 333–342, Washington, DC, USA. ACM, New York, NY, USA.

[Danon et al., 2005] Danon, L., Diaz-Guilera, A., Duch, J., and Arenas, A. (2005). Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 2005:P09008.

[Grandvalet and Canu, 2002] Grandvalet, Y. and Canu, S. (2002). Adaptive scaling for feature selection in SVMs. In Becker, S., Thrun, S., and Obermayer, K., editors, *Proceedings of Advances in Neural Information Processing Systems (NIPS 2002)*, pages 569–576. MIT Press.

[Hofmann et al., 2015] Hofmann, D., Gisbrecht, A., and Hammer, B. (2015). Efficient approximations of robust soft learning vector quantization for non-vectorial data. *Neurocomputing*, 147:96–106.

[Li et al., 2012] Li, Z., Yang, Y., Liu, J., Zhou, X., and Lu, H. (2012). Unsupervised feature selection using nonnegative spectral analysis. In *Proceedings of the 26th Conference on Artificial Intelligence (AAAI 2012)*, pages 1026–1032, Toronto, Ontario, Canada.

[Lin and Tang, 2006] Lin, D. and Tang, X. (2006). Conditional infomax learning: an integrated framework for feature extraction and fusion. In Leonardis, A., Bischolf, H., and Pinz, A., editors, *Proceedings of European Conference on Computer Vision (ECCV 2006)*, volume 3951 of *Lecture Notes in Computer Science*, pages 68–82. Springer, Berlin, Heidelberg.

[Mariette et al., 2017] Mariette, J., Olteanu, M., and Villa-Vialaneix, N. (2017). Efficient interpretable variants of online SOM for large dissimilarity data. *Neurocomputing*, 225:31–48.

[Robnik-Šikonja and Kononenko, 2003] Robnik-Šikonja, M. and Kononenko, I. (2003). Theoretical and empirical analysis of ReliefF and RReliefF. *Machine Learning*, 53(1-2):23–69.

[Schölkopf et al., 2004] Schölkopf, B., Tsuda, K., and Vert, J. (2004). *Kernel Methods in Computational Biology*. MIT Press, London, UK.

[Tibshirani, 1996] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, series B*, 58(1):267–288.

[Yang et al., 2011] Yang, Y., Shen, H. T., Ma, Z., Huang, Z., and Zhou, X. (2011). $\ell_{2,1}$-norm regularized discriminative feature selection for unsupervised learning. In Walsh, T., editor, *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI 2011)*, pages 1589–1594, Barcelona, Spain. AAAI Press.

[Zhao and Liu, 2007] Zhao, Z. and Liu, H. (2007). Spectral feature selection for supervised and unsupervised learning. In Ghahramani, Z., editor, *Proceedings of the 24th International Conference on Machine Learning (ICML 2007)*, pages 1151–1157, Corvalis, OR, USA. ACM, New York, NY, USA.