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, Grandvalet and Canu, 2002] by a large number of irrelevant descriptors.

Kernel methods have proven to be useful and successful to analyse large-scale multi-omics datasets [Schölkopf et al., 2004].

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

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

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

$$K^w(x_i, x_{i'}) := K(w \cdot x_i, w \cdot x_{i'}) ,$$

in which $\cdot$ is the elementwise multiplication: $\mathbf{w} \cdot \mathbf{x} := (w_1 x_1, \ldots, w_p x_p)^T = \text{Diag}(\mathbf{w}) \mathbf{x}$. $K^w$ 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 minimizes the distortion of the original kernel $K$, as measure by e.g. the Frobenius norm:

$$w^* := \arg \min_{w \in \{0,1\}^p} \|K^w - K\|_F^2 \quad \text{for } w \text{ such that } \sum_{j=1}^p w_j \leq s$$

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 $\ell_1$ 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:

$$w^* := \arg \min_{w \in \mathbb{R}^p} \|K^w - K\|_F^2 + \lambda \|w\|_1,$$

in which $\lambda > 0$ is a penalization parameter that controls the trade-off between the minimization of the distortion and the sparsity of the solution and $\| \cdot \|_1$ is the $\ell_1$ norm: $\|x\|_1 := \sum_{j=1}^p |x_j|$. 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 UDFFS, two microarray datasets and a DNA barcoding dataset were analysed on which a ground truth clustering structure is known.

“Carcinom” and “Glomia” 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 $K_{ij} = e^{-\sigma^2 \|x_i - x_j\|^2}$ with $\sigma^2$ 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.
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, \( d_{BC}(x_i, x'_i) = \sum_{l=1}^{p} \frac{|x_{il} - x'_{il}|}{x_{il} + x'_{il}} \), 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 metrics, such as 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 explicitly 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


