Kernel Latent regularization for feature selection in tumor classification

Martin Palazzo [1 2 3] **Patricio Yankilevich** [1] **Pierre Beauseroy** [3]

Abstract

The transcriptomics of cancer tumors are characterized with tens of thousands of gene expression features. Feature selection is a useful approach to select the key genes which helps to classify tumors by prognosis. In this work we propose a feature selection method based on Multiple Kernel Learning that results in a reduced subset of genes and a custom kernel that improves the classification performance when used in support vector classification. During the feature selection process this method performs a novel latent regularization that relax the supervised target problem by introducing unsupervised structure obtained from the latent space learned by a non linear dimensionality reduction model. An improvement of the generalization capacity is obtained and assessed by the tumor classification performance.

1. Introduction

Gene expression is considered as an important layer of information to be used as predictor for clinical outcomes of cancer patients patient (Beer et al., 2002) since it can be used to estimate patient prognosis by modeling the problem as a binary classification one between low and high survival. Nevertheless, there are over 20.000 protein coding genes and this high dimensional context increases the complexity of the classification problem in addition to the need to discover gene biomarkers and improve diagnosis. A reduced subset of p genes from an initial feature set of d genes is just necessary to classify between tumor profiles with an acceptable performance where $p << d$. For this reason it is necessary to reduce the initial high dimensional problem while keeping the interpretability of the features involved by using feature selection methods (He & Yu, 2010). In this work we perform gene selection on Breast cancer data from the International Cancer Genome Consortium (ICGC) (Consortium et al., 2010) to classify patients between high and low survival rate with a threshold of 5 years of survival. The data has n = 194 tumor samples characterized by d = 20504 gene features.

Generally, the important features are selected using a pure supervised objective function (Li et al., 2017). In some cases this supervised objective may be too strict and difficult to fulfill in order to obtain a model that could generalize on new unseen data (Reunanen, 2003). Then a major question arises: is it possible to improve the feature selection process by taking advantage of the structure in feature space of training data in the search for better classification and generalization performances?

Our work proposes a feature selection method based on Multiple Kernel Learning (MKL) (Gönen & Alpaydın, 2011) which defines a new kernel by combining multiple kernel functions via a weighting system to optimize an objective function. Additionally the proposed method combines MKL and a nonlinear latent feature extraction model to improve the feature selection problem by a combination of supervised and unsupervised approaches respectively. This combination aims to improve the generalization capacity in classification of the selected features by maximizing the separability between tumor classes while considering simultaneously the latent structure of the training data. The proposed selection method performs what we name a latent regularization using simultaneously the labels of the data and unsupervised latent variables. As consequence, the selected features are the ones that correlates to both the tumor labels and the data latent structure. The proposed method is designed to deal with tumor classification problems where dimensionality $d > 20.000$ and the sample size is lower than $n = 200$ tumor profiles and is named Kernel Latent Regularization Feature Selection (KLR-FS) (Palazzo et al., 2020).

2. Kernel Latent Regularization Feature Selection (KLR-FS)

The KLR-FS has three main steps. First the feature selection strategy using MKL to select a subset of genes by a supervised criteria. Then the latent regularization is introduced. Finally support vector classification is performed on tumor profiles using the selected features.

[1]Biomedicine Research Institute of Buenos Aires [2]Universidad Tecnologica Nacional Buenos Aires [3]Université de Technologie de Troyes. Correspondence to: Martin Palazzo <mpalazzo@ibioba-mpsp-conicet.gov.ar>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 108, 2020.

2.1. Feature selection with Multiple Kernel Learning

Given $\mathcal{X} \subseteq R^d$ a d-dimensional space and a set of n labeled samples such that $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ where $x_i \in \mathcal{X}$ and $y_i \in \{-1, +1\}$ the Kernel Matrix or Gram Matrix K is defined as a $N \times N$ matrix with entries K_{ij}. Every entry of the Kernel or Gram Matrix is defined as $K_{ij} = \langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{H}} = k(x_i, x_j)$ where $\mathcal{H}$ is defined as a Reproducing Kernel Hilbert Space (RKHS). Kernels can be thought as similarity functions. In this work the Radial Basis Function (RBF) Kernel is used and defined as $k(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$. Given two valid kernels K_1 and K_2 over a set of samples N, the alignment A between both kernels is defined as

$$A(K_1, K_2) = \frac{\langle K_1, K_2 \rangle_F}{\sqrt{\langle K_1, K_1 \rangle_F \langle K_2, K_2 \rangle_F}} \quad (1)$$

and measures the similarity between the two kernels using the same sample set N (Cristianini et al., 2002). If tumor labels are used, then K_2 represents an ideal Kernel matrix or target K_{yy} where $K_{yy} = +1$ if $y_i = y_j$ and $K_{yy} = 0$ if $y_i \neq y_j$ and the alignment of a kernel K built on x and the target K_{yy} is known as Kernel Target Alignment (KTA) score. The higher the KTA between a given kernel matrix K and its target K_{yy} the higher the inter-cluster separation between the two classes. Multiple Kernel Learning (MKL) allows the combination of simple kernels to build a more complex one with higher KTA. MKL is defined as the linear combination of multiple kernels to build a final one (Gönen & Alpaydın, 2011) and can be expressed as

$$K_\mu(x, x') = \sum_{i=1}^{n} \mu_i k_i(x, x'), \mu_i \geq 0 \quad (2)$$

where the vector parameter μ corresponds to the weight $\mu_i > 0$ of each kernel k_i and it is directly related to the importance of each kernel in the final solution. This means that the resulting kernel from the combination of the initial ones will present a higher inter-class separability than each individual kernel on its own. The resulting KTA for the kernel K_μ is $A(K_\mu, K_{yy})$.

Given a dataset of n samples characterized by d features, d feature-wise kernels K_i are built. Then using the MKL method a subset of feature-wise kernels P is selected and combined to increase the overall KTA of the resulting kernel K_μ. Only the feature-wise kernels that increases the KTA are included in the final kernel. This approach leads to a sparse solution where the number of selected features p associated to the selected feature-wise kernels is $p << d$. The desired output of the MKL approach is a reduced set of p features associated to the positive weights $\mu_i > 0$ and a kernel K_μ that improve the inter-cluster distance between samples of different tumor classes and thus improve the support vector classification.

2.2. Latent regularization with nonlinear feature extraction

A relaxation of the supervised approach is proposed by mixing the supervised kernel K_{yy} with an unsupervised K_z kernel built on the latent space Z learned by a nonlinear dimensionality reduction algorithm $\phi_z(x)$. Note that the K_{yy} is built from the tumor labels and the K_z from the extracted latent variables. The mixture of both kernels forms a new target kernel K_δ that has supervised and unsupervised information. Kernel-PCA is used to learn $\phi_z(x)$.

Instead of learning a function f from y as the only label $y = \hat{f}(x)$ this work proposes a feature selection model that learns not only from the labels y but also from the latent structure of the training data as $(z, y) = \hat{f}(x)$. Then by a linear combination of the supervised k_{yy} and unsupervised k_z kernels a new hybrid target kernel K_δ is proposed as

$$K_\delta = \delta K_{yy} + (1 - \delta) K_z \quad (3)$$

The target kernel K_δ is used at the MKL step where K_δ contains both the supervised labels and unsupervised latent variables ruled by the parameter $\delta \in [0, 1]$ named *Mixture Coefficient*.

3. Results

We first evaluated the KLR-FS method with $\delta = [0, 0.2, 0.4, 0.6, 0.8, 1]$ by selecting $p = 10$ features. Figure 1 shows how the latent regularization improves the AUC at $\delta = 0.6$ suggesting that learning partially from latent structure improves the classification performance and generalization.

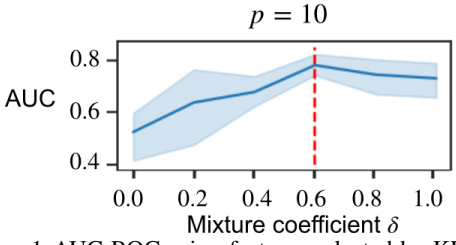

Figure 1. AUC-ROC using features selected by KLR-FS across different values of the δ mixture coefficient.

Then we compared the KLR-FS method ($\delta = 0.6$) against the mRMR (Peng et al., 2005) and HSIC-Lasso (Yamada et al., 2014) feature selection methods with $p = 10$ features by performing support vector classification on the selected features. KLR-FS shows the highest classification performance.

Table 1. Classification AUC for with $p = 10$.

KLR-FS	HSIC-LASSO	MRMR
0.79± 0.1	0.53± 0.11	0.65± 0.08

4. Conclusions

Learning partially from latent space is possible via kernel methods and serves as a regularizer in feature selection tasks. Selected features considering target labels and latent structure improve the classification performance when compared to pure supervised selection tasks.

References

Beer, D. G., Kardia, S. L., Huang, C.-C., Giordano, T. J., Levin, A. M., Misek, D. E., Lin, L., Chen, G., Gharib, T. G., Thomas, D. G., et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature medicine*, 8(8):816–824, 2002.

Consortium, I. C. G. et al. International network of cancer genome projects. *Nature*, 464(7291):993, 2010.

Cristianini, N., Shawe-Taylor, J., Elisseeff, A., and Kandola, J. S. On kernel-target alignment. In *Advances in neural information processing systems*, pp. 367–373, 2002.

Gönen, M. and Alpaydın, E. Multiple kernel learning algorithms. *Journal of machine learning research*, 12(Jul): 2211–2268, 2011.

He, Z. and Yu, W. Stable feature selection for biomarker discovery. *Computational biology and chemistry*, 34(4): 215–225, 2010.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., and Liu, H. Feature selection: A data perspective. *ACM Computing Surveys (CSUR)*, 50(6):1–45, 2017.

Palazzo, M., Yankilevich, P., and Beauseroy, P. Latent regularization for feature selection using kernel methods in tumor classification. *arXiv preprint arXiv:2004.04866*, 2020.

Peng, H., Long, F., and Ding, C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (8):1226–1238, 2005.

Reunanen, J. Overfitting in making comparisons between variable selection methods. *Journal of Machine Learning Research*, 3(Mar):1371–1382, 2003.

Yamada, M., Jitkrittum, W., Sigal, L., Xing, E. P., and Sugiyama, M. High-dimensional feature selection by feature-wise kernelized lasso. *Neural computation*, 26(1): 185–207, 2014.