Learning Kernels from genetic profiles to discriminate tumor subtypes

Martín Palazzo [12], Pierre Beauseroy [2], Daniel Koile [1], Patricio Yankilevich [1]

[1] Biomedicine Research Institute of Buenos Aires, Max Planck Society Partner Institute, Polo Científico Tecnológico, Buenos Aires, Argentina

{mpalazzo, pyankilevich, dkoile}@ibioba-mpsp-conicet.gov.ar

[2] Institut Charles Delaunay (CNRS), Université de Technologie de Troyes, France
{martin.palazzo, pierre.beauseroy}@utt.fr

Keywords: Kernel Target Alignment, Multiple Kernel Learning, Somatic Mutation, Breast Cancer, Support Vector Classification, Feature Selection.

Several biological data types are getting easier to access. Genomic, Transcriptomic, Proteomic and Metabolomic data are some of the cases among others. Phenotype data also is included, since we can link any biological layer of information mentioned before with a phenotype, like the diagnosis of a disease. Feature selection methods help to determine the main genes and mutations responsible of characterize different Breast Cancer subtype tumors and to perform a better discriminant analysis. Tumor subtype classification has been performed using somatic point mutations [3] and support vector machines classifiers with standard and well-known kernels such us Gaussian. The methods commonly used to select features (genes) are Filter and Wrappers [8]. Our proposal is focused in learning kernels from human genomic data from Breast Cancer patients [4], in order to classify them among different tumor subtypes. Building custom Kernels provides an alternative to improve results. By combining different Kernel functions learned from the training samples we aim to improve the Kernel Target Alignment (KTA) score [1] and thus use the optimized kernel to perform discriminant analysis through support vector classification. Kernel Target Alignment measures the degree of agreement between a reproducing kernel function and a learning task. The higher the KTA, the better the performance of the support vector classification among different classes [2], thus a better discriminant analysis between tumor subtypes for the benefit of a better therapy. Kernel Target Alignment of the kernel 'K' and target label 'y' with respect to the sample 'S' of size 'm' is expressed as:

$$A\left(S,K,yy'\right)=\frac{\left\langle K,yy'\right\rangle_F}{\sqrt{\left\langle K,K\right\rangle_F\left\langle yy',yy'\right\rangle_F}}=\frac{y'Ky}{m\left\|K\right\|_F} \quad (1)$$

The original dataset contains samples of genetic profiles from Breast Cancer patients. Each sample lists all the single base somatic mutations the tumor has. A somatic mutation means a nucleotide variation on the DNA sequence of the tissue where the tumor is located in comparison with the germline sequence of the patient. These mutations are presented in a Variant Calling Format file (. VCF), a list of somatic variants per patient. The dataset consists in 800 breast cancer patients. Patients are categorized between tumor subtypes Triple Negative, ER+ and HER2- [3]. Each patient is linked to a set of mutated genes. Each gene contains the quantity of somatic mutations within it. By this way we count each gene as a feature, and the quantity of mutations within each one as the feature value. Other alternatives are to consider each single somatic mutation as a boolean feature but the sparsity increases considerably. An important characteristic of the data is the high dimensionality and sparsity among features. Originally, our dataset presents 13750 mutated genes including mutated pseudogenes, so our problem presents the curse of dimensionality, since the sample to feature ratio is 0.058. Considering each single point mutation as a feature can decrease considerably the sample to feature ratio and thus intensify the curse of dimensionality. For discriminant analysis and clinical purposes it is required to perform feature selection.

Our work aims to perform the feature selection step on Multiple Kernel Learning [7] by optimizing the Kernel Target Alignment score [2]. It begins by building feature-wise gaussian kernel functions. Then by a constrained linear combination [6] of the feature-wise kernels, we aim to increase the Kernel Target Alignment to obtain a new optimized custom kernel. The linear combination results in a sparse solution where only few kernels survive to improve KTA and consequently a reduced feature subset is obtained. Reducing considerably the original gene set allow to study deeper the selected genes for clinical purposes. The higher the KTA obtained, the better the feature selection, since we want to build custom kernels to use them for classification purposes later [5]. The final kernel after optimizing the KTA is built by a linear combination of 'Ki' kernels, each one associated to a μi coefficient. The μ vector is computed during the optimization process.

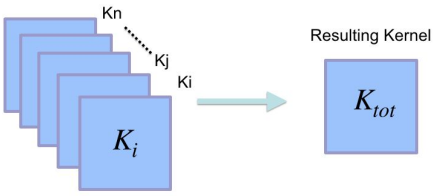

Figure 1. Linear combination of kernels to improve KTA.

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

The objective with the resulting custom kernel and reduced feature subset is to discriminate through support vector classification the different tumor subtypes. Preliminary results in our work reduce the mutated gene set from 13750 to less than 1000 with an improvement of the KTA. We aim to benchmark our proposal with standard kernels like gaussian and with feature selection methods such as Lasso L1 regularization and Recursive Feature Elimination. The proposed method performs two steps in one since it results in a reduced subset of genes determined by the μ vector from one side and in a custom kernel optimized to discriminate the tumor subtypes on the other side.

References

1. Cristianini, N., Shawe-Taylor, J., Elisseeff, A., & Kandola, J. S. (2002). On kernel-target alignment. In Advances in neural information processing systems (pp. 367-373).
2. Ramona, M., Richard, G., & David, B. (2012). Multiclass feature selection with kernel gram-matrix-based criteria. IEEE transactions on neural networks and learning systems, 23(10), 1611-1623.
3. Vural, S., Wang, X., & Guda, C. (2016). Classification of breast cancer patients using somatic mutation profiles and machine learning approaches. BMC systems biology, 10(3), 62.
4. Low, S. K., Zembutsu, H., & Nakamura, Y. (2018). Breast cancer: The translation of big genomic data to cancer precision medicine. Cancer science, 109(3), 497-506.
5. Wang, T., Zhao, D., & Tian, S. (2015). An overview of kernel alignment and its applications. Artificial Intelligence Review, 43(2), 179-192.
6. Pothin, J. B., & Richard, C. (2006, September). A greedy algorithm for optimizing the kernel alignment and the performance of kernel machines. In Signal Processing Conference, 2006 14th European (pp. 1-4). IEEE.
7. Rakotomamonjy, A., Bach, F. R., Canu, S., & Grandvalet, Y. (2008). SimpleMKL. Journal of Machine Learning Research, 9(Nov), 2491-2521.
8. Huang, S., Cai, N., Pacheco, P. P., Narandes, S., & Xu, W. (2018). Applications of support vector machine learning in cancer genomics. Cancer Genomics-Proteomics, 15(1), 41-51.