ISBN: 978-950-42-0200-4
DOI: https://doi.org/10.33414/ajea.5.779.2020

Reducción de la Dimensionalidad en Perfiles Tumorales con Metodos de Kernel y Redes Neuronales

Dimensionality reduction in tumor profiles with kernel methods and neural networks

Presentación: 06-07/10/2020

Doctorando:
Martin Palazzo
Plataforma de Bioinformatica, Instituto de Investigación en Biomedicina de Buenos Aires—Consejo Nacional de Investigaciones Cientificas y Tecnicas de Argentina (CONICET)—Partner Institute of the Max Planck Society, Argentina. Université de Technologie de Troyes, Francia.
mpalazzo@ibioba-mpsp-conicet.gov.ar

Director/a:
Patricio Yankilevich

Co-director/a:
Pierre Beauseroy

Resumen

Los perfiles tumorales en humanos pueden ser caracterizados por su genómica mediante la expresión de miles de genes. Este tipo de señales pueden ser aprovechadas para detectar estadísticamente patrones que permitan clasificar o agrupar de manera supervisada o no supervisada perfiles tumorales según su fenotipo. En esta presentación y revisión se busca listar dos casos para poder reducir la dimensionalidad genómica y encontrar biomarcadores que permitan realizar las tareas de aprendizaje estadístico mencionadas. Para reducir la dimensionalidad se utilizan funciones de Kernel en combinación con redes Neuronales Artificiales. Los resultados obtenidos en la presente revisión muestran el potencial de estas herramientas para el procesamiento de datos genómicos en pacientes oncológicos.

Palabras clave: Métodos de Kernel, Redes Neuronales Artificiales, Genómica, Cáncer.

Abstract

Human tumors profiles can be characterized by their genomics by expressing thousands of genes. These types of signals can be exploited to statistically detect patterns that make it possible to classify or group different phenotypes in a supervised or unsupervised approach. This presentation seeks to list two application of genomic dimensionality reduction and to find biomarkers that improve the aforementioned statistical learning tasks. In order to reduce dimensionality, Kernel functions are used in combination with Artificial Neural Networks. The results obtained show the potential of these tools for the processing of genomic data in cancer patients.

Keywords: Métodos de Kernel, Redes Neuronales, Genómica tumoral

Introducción

Desde la primera secuenciación completa del genoma humano (Venter, 2001) la cantidad de datos genómicos disponibles en biomedicina se ha incrementado significativamente. Actualmente existen grandes repositorios de datos genómicos de cáncer como The Cancer Genome Atlas (Consortium, 2010) y The International Genome Cancer Consortium (Weinstein, 2013). Estos repositorios permiten acceder al perfil genómico de miles de tumores de pacientes de cáncer y a su información clínica como el subtipo de tumor, el tiempo de sobrevida y el estado del tumor. Adicionalmente los tumores están caracterizados por la expresión de 20.00 genes aproximadamente. Esto permite poder estudiar a los tumores desde su perfil genómico para poder estimar sus condiciones clínicas como el subtipo de tumor mediante tareas de clasificación (Salem, 2017) o agruparlos para poder descubrir subtipos (Young, 2017). El hecho de que existan conjuntos de datos provenientes de una misma población permite la utilización de métodos de aprendizaje automático y estadístico para poder estimar las condiciones clínicas. De todos modos la alta cantidad de genes contemplados cercana a 20.000 implica que el problema este definido en alta dimensión generando que la estimación de atributos clínicos mediante la interpretación de perfiles genómicos sea compleja. Por esta razón uno de los objetivos buscados en este trabajo es poder reducir la dimensionalidad de manera tal de mantener la información latente e intrínseca de los datos originales de expresión genética tumoral. Este objetivo busca reducir la complejidad del problema mediante la selección de un conjunto reducido de biomarcadores (Ang, 2015) y así poder realizar las tareas de aprendizaje automático en la nueva reducción obtenida con menor error. Para poder reducir la dimensionalidad y seleccionar biomarcadores dos familias de métodos son utilizadas y combinadas: métodos de kernel y redes neuronales artificiales.

Desarrollo

Para poder seleccionar genes y reducir la dimensionalidad los métodos propuestos están contemplados de dos modelos. Un primer modelo estará diseñado con el objetivo de aprender una representación de baja dimensionalidad de los datos originales mediante la extracción de características y dimensiones nuevas a partir la combinación no lineal de las variables de entrada. El otro modelo será el encargado de seleccionar el subconjunto de genes que mas se aproxime a la representación aprendida del modelo anterior. Para poder aprender la representación de baja dimensionalidad de los datos de expresión tumorales se utilizan redes neuronales artificiales, particularmente el método de Autoencoder (Way, 2018). Este método consiste en aprender dos funciones, un Encoder y un Decoder. El encoder tendrá la función de reducir la dimensionalidad mediante el aprendizaje de nuevas dimensiones latentes que caracterizarán a los datos de entrada de alta dimensión. El decoder reconstruye los vectores proyectados en las dimensiones latentes al espacio original.

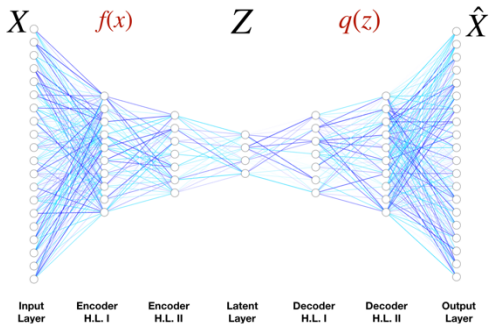

Figura 1: La arquitectura de un autoencoder.

Para poder seleccionar los genes de manera tal de acercarse lo mas posible a la representación de baja dimensión aprendida por el autoencoder los métodos de Kernel son utilizados. Particularmente el método de Multiple Kernel Learning (Aprendizaje de múltiples kernels) (Gönen, 2011). Un kernel es una función de similaridad entre pares de vectores o muestras en un espacio de alta dimensión de Hilbert. Este enfoque permite seleccionar genes tanto para tareas de clasificación como tareas de

agrupamiento en caso que la representación aprendida sea supervisada o no supervisada respectivamente. Mediante la combinación lineal de kernels de la figura 2 se seleccionan un subconjunto de genes que maximiza el alineamiento (Cristianini, 2002) con un kernel objetivo. El kernel objetivo puede ser supervisado o no supervisado. En el caso de un kernel objetivo asociado a la representación obtenida desde el autoencoder será un kernel no supervisado (Palazzo, 2020). En el caso de un kernel objetivo asociado a etiquetas clínicas como subtipo de tumor y en simultaneo combinado linealmente con el kernel objetivo del autoencoder entonces se tendrá una selección que combine un enfoque supervisado y no supervisado en simultaneo (Palazzo M. Y., 2020).

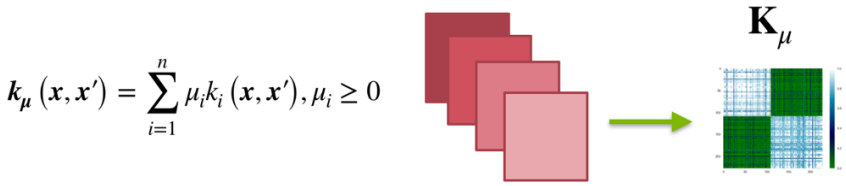

Figura 2: Combinación de kernels por múltiple kernel learning.

El número de genes seleccionados por el método de Multiple Kernel Learning es un hiperparámetro que se define previo a la ejecución del método.

Resultados

En este trabajo se evalúa el sistema propuesto de manera supervisada y no supervisada mediante la clasificación de perfiles tumorales como en agrupamiento de los mismos seleccionando 50 genes. Para evaluar la clasificación se utiliza la métrica Area debajo de la curva ROC en clasificación de subtipos de tumor de cáncer de mamas. Para evaluar el agrupamiento se utiliza la Información mutua de los tumores agrupados en contraste con las etiquetas clínicas de subtipo de tumor de cáncer renal. El proceso se repite 10 veces de manera aleatoria utilizando un 80% de las muestras para entrenamiento y el restante 20% para evaluación independiente para estimar estadísticamente mediante la media el valor de los resultados. En clasificación el método propuesto de Kernel y Autoencoder se compara con el método de mínima redundancia y máxima relevancia (mRMR) (Ding, 2005) para clasificar los perfiles tumorales. En el agrupamiento el método propuesto se compara con el método de k-medias ponderado (WKM) (Tseng, 2007).

Metodo	Genes seleccionados	AUC-ROC
Kernel + Autoencoder	<50	0.77
mRMR	<50	0.6

Tabla 01: media de AUC-ROC de clasificación en el conjunto de muestras de evaluación luego de 10 entrenamientos aleatorios.

Metodo	Genes seleccionados	IM
Kernel + Autoencoder	<50	0.3
WKM	<50	0.19

Tabla 02: media de Información Mutua en agrupamiento luego de 10 entrenamientos aleatorios.

En la tabla 01 y 02 se observa como la combinación de métodos de kernel y redes neuronales no supervisadas mejoran tanto la clasificación como el agrupamiento frente a otros métodos para una reducción y selección de 50 genes.

Conclusiones

Mediante la reducción de dimensionalidad con redes neuronales de perfiles tumorales de alta dimensión es posible generar una representación de referencia lo suficientemente informativa que puede ser utilizada por métodos de kernel para seleccionar genes que expliquen la estructura intrínseca de los datos originales. Los resultados de los dos trabajos revisados (Palazzo M. Y., 2020) (Palazzo M. B., 2020) muestran que la selección de genes mejora tareas de aprendizaje estadístico como la clasificación y el agrupamiento de perfiles genómicos de tumores. Futuros trabajos deberán incluir la combinación de datos multi-omicos como genómicos, transcriptomicos, proteomicos y metabolomicos.

Referencias

[1] Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., ... & Gocayne, J. D. (2001). The sequence of the human genome. science, 291(5507), 1304-1351.
[2] Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., ... & Cancer Genome Atlas Research Network. (2013). The cancer genome atlas pan-cancer analysis project. Nature genetics, 45(10), 1113.
[3] International Cancer Genome Consortium. (2010). International network of cancer genome projects. Nature, 464(7291), 993.
[4] Salem, H., Attiya, G., & El-Fishawy, N. (2017). Classification of human cancer diseases by gene expression profiles. Applied Soft Computing, 50, 124-134. (Salem, 2017)
[5] Young, J. D., Cai, C., & Lu, X. (2017). Unsupervised deep learning reveals prognostically relevant subtypes of glioblastoma. BMC bioinformatics, 18(11), 5-17. (Young, 2017)
[6] Ang, J. C., Mirzal, A., Haron, H., & Hamed, H. N. A. (2015). Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection. IEEE/ACM transactions on computational biology and bioinformatics, 13(5), 971-989. (Ang, 2015)
[7] Gönen, M., & Alpaydın, E. (2011). Multiple kernel learning algorithms. The Journal of Machine Learning Research, 12, 2211-2268.
[8] Way, G. P., & Greene, C. S. (2018) Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. (Way, 2018)
[9] Palazzo, M., Beauseroy, P., & Yankilevich, P. (2020). Unsupervised Feature Selection for Tumor Profiles using Autoencoders and Kernel Methods. arXiv preprint arXiv:2007.06106.
[10] Palazzo, M., Yankilevich, P., & Beauseroy, P. (2020). Latent regularization for feature selection using kernel methods in tumor classification. arXiv preprint arXiv:2004.04866. (Palazzo M. Y., 2020)
[11] Tseng, G. C. (2007). Penalized and weighted K-means for clustering with scattered objects and prior information in high-throughput biological data. Bioinformatics, 23(17), 2247-2255. (Tseng, 2007)
[12] Ding, C., & Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. Journal of bioinformatics and computational biology, 3(02), 185-205.
[13] 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).