Predicción de fenotipos y selección genómica utilizando redes neuronales

Nathan Kosoi Lamont

Diciembre 2022

1. Introducción

La amplia disponibilidad y el costo reducido de la tecnología de marcadores moleculares ha creado la oportunidad de llevar a cabo una selección de genotipos, asistida por estos marcadores, en plantas y animales con el fin de mejorar su reproducción.

Técnicas de mapeo, tales como las de Rasgos Cuantitativos de Locus han demostrado ser útiles y poderosas para identificar marcadores asociados genéticamente con genes condicionados a fenotipos agronómicos (Miles and Wayne, 2008)

La información que se obtiene con los métodos previamente mencionados nos permite aplicar técnicas de predicción y selección genómica en individuos y poblaciones. Más específicamente, una vez obtenida la información de marcadores genéticos, tales como los Polimorfismos de nucleótido único (SNP), obtenida de una población, se utilizan métodos estadísticos para predecir los fenotipos de la población.

Estamos en una época donde la selección genómica es practicada por la gran mayoría de compañías dedicadas a la plantación y reproducción vegetal e incluso animal. Típicamente esto es logrado incrementando el número progenitores evaluados en una etapa temprana de la crianza y realizando una intensa selección basada en valores de predicción genética.

Más aun, se tiene la hipótesis de que el método de selección basado únicamente en información genómica nos permite mantener una mayor diversidad en los programas de crianza, a diferencia de basarse únicamente en mediciones fenotípicas a posteriori. Además, la predicción genómica puede quizás permitir a los cultivadores caracterizar el rendimiento de las combinaciones de alelos en ambientes críticos para el mercado, pero que raramente son observados. Es posible, pues, que en un futuro, la selección genómica se convierta en el método primario de selección en un programa de crianza con evaluación fenotípica.

Dicho esto, es claro que no cualquier método estadístico será de utilidad para llevar a cabo estos análisis. Dependiendo de la cantidad de información genética que se haya recolectado y las características que se busquen predecir,

será conveniente un distinto método estadístico para realizar el estudio. Necesitamos una arquitectura que sea capaz de encontrar la relación concreta entre los marcadores genéticos y los rasgos que busquen predecirse.

Esto nos deja con una pregunta importante, ¿Cómo elegir el modelo a utilizar?, antes de responder esto, es importante darnos cuenta de que independientemente del método, es posible hacer un preprocesamiento de los datos para que en cualquiera de los casos nuestros resultados no se vean afectados por la naturaleza estocástica de los procesos genéticos ni por el error humano asociado a la recolección de los datos.

Más adelante observaremos la diferencia entre métodos regularizados y no regularizados de redes neuronales comparados con métodos de regularización y análisis estadístico. Con estas comparaciones, podremos finalmente elegir el mejor modelo con alguna (o ninguna) técnica de regularización con el fin de hacer predicciones más certera.

En este proyecto nos centraremos en un modelo especial de predicción de Machine Learning, Redes Neuronales. Las redes neuronales han demostrado ser métodos predictivos de alta eficacia, su uso se ha visto en una inmensa gama de problemas distintos. Desde selección de publicidad hasta predicción de flujo de usuarios en páginas web, las redes neuronales tienen su lugar y con el paso del tiempo solo se espera que sigan mejorando y cada vez sean más potentes, eficientes y útiles.

Motivados por este auge computacional,tanto por los nuevos modelos, como por los nuevos datos consideramos que es importante poner a prueba a las redes neuronales en estos conjuntos de datos. Es decir, vamos a buscar como llevar a cabo predicción genómica mediante redes neuronales. Para esto, hay que tomar en cuenta que los intentos previos para lograr esto no han sido del todo exitosos, esto es debido, probablemente a tres factores distintos, falta de regularización de los datos, overfitting en la red neuronal y exceso en el tiempo de cálculo durante el entrenamiento de la red.

En conclusión, nos dedicaremos a construir un modelo de machine learning utilizando la tecnología de redes neuronales para poder llevar a cabo predicción de fenotipos utilizando información genética de diversas especies. Después compararemos este modelo con otros modelos puramente estadísticos y también veremos la diferencia entre la red neuronal regularizada y no regularizada.

2. Pregunta de Investigación

¿Son las redes neuronales modelos certeros para la preddicción fenotípica utilizando información genética?

3. Datos

Nos fijaremos en conjuntos de datos de 3 especies distintas, cada uno consistirá tanto de sus marcadores genéticos como de sus fenotipos para diversos

individuos de una población. Las tres plantas elegidas fueron la planta Arabidopsis, Maíz y trigo. La información fue recolectada de publicaciones de distintos autores (Loudet et al., 2002; Crossa et al., 2010; Thavamanikumar et al., 2015).

Los marcadores fueron escalados en un rango de 0 a 1 para toda la información de SNP's. Si más del 20 % de los marcadores están ausentes, la muestra es descartada, Si menos del 20 % de los marcadores estaban ausentes, el valor promedio de ese marcador sería insertado para los valores ausentes, con una sola excepción, si los datos publicados ya habían aplicado alguna técnica para trabajar con la información ausente. Los individuos de las muestras que no reportaron ninguna medida fenotípica también fueron descartados del análisis. Se predijo una combinación de mediciones fenotípicas y de calidad reproductiva, dependiendo en las mediciones que los autores de quienes se tomaron los datos.

4. Métodos

4.1. Resumen del Método

Se creará una red neuronal profunda de arquitectura "feedforward", en este tipo de redes, los distintos marcadores genéticos son entradas que se ven mapeados a una o más capas ocultas de neuronas. Para predicción genómica, usualmente la capa de entrada tiene tantas neuronas como marcadores genéticos registrados, la capa de salida consiste de una sola neurona que predice el valor del fenotipo.

Para facilitar la evaluación de redes neuronales regularizadas, un modelo de predicción de red neuronal fue implementado con dos técincas de regularización. La primera es la regularización por caída de pesos "weight decay", ésta penaliza los pesos W con valores bastante grandes. La segunda, es una regularización por "dropout", donde un porcentaje de las neuronas y sus conecciones son removidas al azar durante el entranmiento en cada época, esto permite que las neuronas "no se pongan de acuerdoτ así la memorización de patrones sucede de forma más independiente para cada neurona y con ello evitamos overfitting.

Se utilizó una versión ligeramene modificada de backpropagation para el entrenamiento. Por último, métodos de detección de divergencia y de decaimiento de learningrate fueron implementados para asegurarse de que el error del modelo convergiera a un mínimo global.

4.2. Separación de los datos

Cada conjunto de datos contiene emparejados las medidas fenotípicas y genotípicas. Estas procedimos a separarlas en una partición de 5 bloques. Cada modelo de predicción fue entrenado en 4 de los 5 bloques y fue puesto a prueba en el quinto.

Para los modelos estadísticos con hiperparámetros movibles, se utilizó un método de "grid-search.[en] el espacio de parámetros. La precisión de la predicción así como cualquier parámetro fue recolectado para la partición. Los modelos

después fueron reseteados y el proceso fue repetido con una partición distinta. Esto produjo cinco estimaciones de la precisión de la predicción en cada modelo. Después, cada conjunto de datos fue mezclado y se repitió el proceso del "pipeline"Para producir así un total de diez estimados de predicción de precisión para cada modelo.

4.3. Red neuronal

Todas los modelos de redes neuronales fueron inicializados el método de inicialización de pesos de Glorot (Glorot and Bengio, 2010) y entrenados por 12100 épocas de 4 batches cada una. Se utulizó Backpropagation con el algoritmo de Nadam backpropagation con parámetros estándar. La primeras 100 épocas fueron ejecutadas en redes distintas, a la red con mayor precisión se le permitió completar las ultimas 12000 épocas, reduciendo así la incidencia de baja precisión debido a una elección pobre de valores iniciales de los pesos. El learning rate de la red fue reducido por un factor de 4 después de las primeras 10100 épocas y reducido aun más por un factor de 4 después de 11600 épocas, tal fue el método de decaimiento de learning rate.

Se construyó una red neuronal con dropout y decaimiento de pesos utilizando la librería de Keras.

5. Resultados

A continuación se presentan el resumen de los resultados de los distintos métodos estadísticos y de redes neuronales en la tarea de predicción de rasgos fenotípicos utilizando la información genómica.

La arquitectura de red neuronal no regularizada no demostró ser la mejor en ninguno de lo conjuntos de datos. El método de Elastic Net incorpora tanto normalización L1 como normalización L2 en un solo modelo, este método fue óptimo en un set de datos. Regresión Ridge Bayesiana incorpora regularización L2 y al mismo tiempo encaja los coeficientes de regresión a una distribución gaussiana, no fue óptima en la mayoría de problemas, pero al mismo tiempo fue mejor que el promedio. El modelo de red neuronal con decaimiento de pesos y dropout demostró ser el mejor en conjuntos de datos distintos.

El resumen de los métodos mas importantes se encuentra en la tabla 1.

6. Discusión

En este reporte presentamos los resultados de usar una red neuronal profunda para realizar la tarea de predicción genómica en 3 conjuntos de datos de distintas especies. En cada conjunto de datos (salvo el trigo)se eligió una característica de alta y de baja heredabilidad por un total de 5 tareas de predicción genómica. Se realizaron distinas técnicas de regresión incluídas redes neuronales profundas con y sin métodos de regularización. Se compararon estos distintos métodos y

Especie	Rasgo	EN	BRR	N	NWDDO
Arabidopsis	Material Seco	42	39	38	40
Arabidopsis	Floración	82	82	84	86
Maíz	Floración	33	32	33	33
Maíz	Rendimiento de Granos	51	57	55	51
Trigo	Numero de Espigas	36	28	27	33

Tabla 1: Porcentaje de Precisión de distintos métodos, Los primeros dos son estadísticos, el tercero es una red neuronal sin regularización, el último es una red neuronal con regulariazión de datos

con ello obtenimos un panorama más claro de la eficacia de los análisis de deep learning en la tarea de predicción genómica.

Podemos entonces conlcuir que hay un gran potencial para la aplicación de los métodos de redes neuronales, y proponemos que se siga experimentado con diferentes arquitecturas de dichos modelos. Sería interesante, por ejemplo, intentar utilizar una red neuronal convolucional, o tal vez incluso una LSTM para un posterior análisis. De momento sabemos que, por lo menos, las redes neuronales son tan capaces como los métodos estadísticos más utilizados actualmente.

7. Bibliogrfía

Riley McDowell. Genomic selection with deep neural networks, 2016

Liu Y, Wang D, He F, Wang J, Joshi T, Xu D. Phenotype Prediction and Genome-Wide Association Study Using Deep Convolutional Neural Network of Soybean. Front Genet. 2019 Nov 22;10:1091. doi: 10.3389/fgene.2019.01091. PMID: 31824557; PMCID: PMC6883005.

Miles, C. and M. Wayne, 2008 Quantitative trait locus (QTL) analysis. Nature Education 1: 208.

Loudet, O., S. Chaillou, C. Camilleri, D. Bouchez, and F. Daniel-Vedele, 2002 Bay-0 x Shahdara recombinant inbred line population: A powerful tool for the genetic dissection of complex traits in Arabidopsis. Theor. Appl. Genet. 104: 1173–1184.

Crossa, J., G. de los Campos, P. P´erez, D. Gianola, J. Burgue˜no, et al., 2010 Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186: 713–724.

Thavamanikumar, S., R. Dolferus, and B. R. Thumma, 2015 Comparison of genomic selection models to predict flowering time and spike grain number in two hexaploid wheat doubled haploid populations. G3 5: 1991–1998.

Glorot, X. and Y. Bengio, 2010 Understanding the difficulty of training deep feedforward neural networks. In AISTATS, volume 9, pp. 249–256.