New Generation of Machine Learning: Analysis of Single Nucleotide Polymorphism Data with Deep Learning

Erdal Coşgun, Ph.D

Data and Applied Scientist, Microsoft Genomics

ercosgun@microsoft.com

Objective

In recent years, both bioinformatics experts and medical doctors have been working on programming languages such as R, Python, C #. Most of these studies are focused on machine learning, biostatistics and optimization. The biggest problem with these analyzes was that the "computing capacity" of the PCs or on-prem server was low and unsustainable. However, thanks to cloud systems, this problem has also been eliminated. On the other hand, "Machine Learning", which brought a new breath to genome data analysis, has been included in almost every genomic data analysis in the last 10 years. This study aimed to compare Deep Learning, which emerged as "Next Generation Machine Learning", for the analysis of Single Nucleotide Polymorphism (SNP) data with the classical Neural Network and Random Decision Forest methods. In this way, the performance and effect of a new generation method group will be revealed.

Method

The data used in this study has been simulated with PLINK software. Data prepared for different SNP numbers (one, two and three million SNPs) were screened as population-based (equally distributed patient-control). (100,250,500 patient-control) For analysis, *"Microsoft Azure Machine Learning with Microsoft R Server"* and "h2o" package has been used used. Here, all parameters for both groups of methods have been optimized by the "Hyper Search" technique. The results have been compared with accuracy, precision and recall measurements.

Findings

Findings related to the planned scenarios are shown in :Table 1. Results of Deep Learning, Neural Network and Random Decision Forest Methods

#SNP	1 million			2 million			3 million		
#Patient	100	250	500	100	250	500	100	250	500
	Deep Learning								
%									
Accuracy	0.65	0.68	0,58	0.63	0.69	0.64	0.62	0.68	0.62
Precision	0.57	0.61	0,59	0.69	0.59	0.63	0.59	0.61	0.62
Recall	0.60	0.67	0,55	0.68	0.57	0.61	0.64	0.58	0.54
	Neural Network								
%									
Accuracy	0.52	0.57	0,51	0.52	0.55	0.51	0.55	0.55	0.56
Precision	0.50	0.54	0,54	0.56	0.52	0.50	0.56	0.54	0.59
Recall	0.50	0.56	0,56	0.59	0.54	0.53	0.56	0.57	0.51
	Random Decision Forest								
%									
Accuracy	0,61	0.65	0,53	0.78	0.73	0.52	0.67	0.69	0.67
Precision	0,59	0.62	0,56	0.76	0.71	0.55	0.64	0.64	0.64
Recall	0,62	0.63	0,53	0.71	0.73	0.56	0.62	0.63	0.67

Result

Deep Learning, which has a very limited library yet, maintains its performance independently of the number of people with respect to the findings obtained. This is a very important finding. Because the biggest drawback of classical ML methods is the decrease in performance as the number of parameters increase. For example, Random Decision Forest achieved the highest accuracy rate of 78.0% in 2 million SNP scenarios where 100 patients were 100 control subjects . However, no increase in the low SNP number was observed.. The different results here make model optimization difficult. In general, we can state that deep learning is a whole new neural network algorithm. This difference obtained in categorical data type is also important for other biostatistics analyzes. In addition, it will be very helpful for our country's biostatistics experts to follow such alternative and promising methods.

Key words: Deep Learning, Single Nucleotide Polymorphism, Machine Learning

*This study presented at 18th National Biostatistics Conference,2016 as a oral presentation.