DOI 10.18699/BGRSSB-2018-069

Search for a hidden structure in genomic data based on a compressing autoencoder

A. Zarubin[1, 2]*, A. Markov[1], V. Stepanov[1], V. Kharkov[1]
[1] *Tomsk National Research Medical Center RAS, Research Institute of Medical Genetics, Tomsk, Russia*
[2] *Siberian State Medical University, Tomsk, Russia*
* *e-mail: aleksei.zarubin@medgenetics.ru*

Key words: compressing autoencoder, genomic data, neural networks

Motivation and Aim: Recently, the problem of genomic data processing is becoming more urgent. The main difficulty lies in the "curse of dimension" – because the number of polymorphisms determined is usually several orders of magnitude greater than the number of genotyped samples. Usually the principal component analysis (PCA) is used to solve this problem. PCA also allows analyzing the hidden data structure by finding new variables. But it has some drawbacks, especially in analyzing complex interactions. Therefore, the application of a compressing autoencoder for genomic data as an approach that can approximate nonlinear interactions might be promising.
Methods and Algorithms: In order to evaluate how well the autoencoder manages to find the hidden structure in the data, we used a dataset on genotypes of 894 people from 28 populations from Russia and neighboring countries [1]. After filtering out the missing values, the remaining 113749 polymorphic variants were used as a training sample. Artificial neural networks of different architectures were modeled in the R software environment using the Keras library [2].
Results: In the selection and evaluation of hyperparameters and the architecture of the neural network, the linear activation function for the output layer of the encoder and exponential linear unit for all the fully-connected and convolutional layers were most successful. The most effective algorithm of optimization was the Adam algorithm. As the final test model, we selected a 7 layered fully-connected perceptron with a total of 117646423 parameters and two linear outputs from the encoder. The training was gradual in 20 iterations with batch size 20. Seven populations are separated into mono groups quite well and quickly, but due to some non-linearity of axes, it is necessary to reduce the speed of training, or gradually exclude from training the samples, which have already clearly separated into a separate cluster and repeat the learning process in a smaller sample. The remaining populations are more difficult to differentiate, although they form a number of clusters. An increase in the number of output neurons from the encoder to 3 makes it possible to isolate up to 11 populations. In turn, using PCA, only up to 6 populations can be clearly identified.
Conclusion: Compressing autoencoder shows a higher efficiency comparing to PCA for searching the hidden structure in the genomic data and lowering the dimension. A reliable differentiation of populations requires the determination of many more hyperparameters the most effective use of linear and piecewise linear activation functions, Adam as an optimization algorithm and a reduced learning rate.

References

1. Triska P. et al. (2017) Between Lake Baikal and the Baltic Sea: genomic history of the Gateway to Europe. BMC Genetics. 18(1):110.
2. Chollet F.(2015) Keras. https://github.com/fchollet/keras