Designing a Machine-learning strategy to predict types of cancers

Hugo MARTIN[1],

[1]Aix-Marseille Université, M2 DLAD, Marseille, F-13288, France;

Abstract: In this study, we proposed to study the training performance of a neural network in the aim to predict types of cancers for patients affected by this pathology. Although we could reduce the overfitting effect of the trained neural network model by changing some parameters, it is still possible to improve this model by varying other parameters we don't change during this study and using other machine learning models to compare efficiencies of different models on the dataset learning.

Keywords:

Machine-learning, UMAP, PCA, Neural Network

INTRODUCTION

Machine-learning techniques are more and more used in the biology field. The use of these algorithms allowed to improve diagnostics and data exploration. All these techniques look for minimizing a loss function in the aim to have the best accuracy on new data never encountered. We proposed here to study the learning performances of the use of a neural network which is a supervised learning technique on the Cancer Genome Atlas (TCGA) dataset to predict types of cancers for more than 800 patients. (1)

MATERIALS AND METHODS

Data Exploration

In order to investigate the data organisation, we compared two techniques of Dimension reduction which are Principal Component Analysis (PCA) PCA and Uniform Manifold Approximation and Projection (UMAP) to highlight characteristics of the dataset. (2, 3) Dimension reduction consists of projetting data on a two-dimensional space. Data were scaled with a standardization function provided by scikit-learn (4) before using the Dimension Reduction techniques in the aim to prevent from no-scaling effects. We used the scikit-learn and umap-learn Python packages in the aim to build these dimension reductions (3, 4). UMAP parameters were left by default to compute the Dimension reduction.

Neural Network Training

We built several neural networks which we have compared accuracy and loss performances by plotting accuracy curves and loss curves. Neural networks were trained by using keras and tensorflow Python packages (5). In all, we realised seven configurations of neural networks by changing batch size, features size, Lasso regularization term and Ridge regulation term.

Given that the dataset contains an important number of features (20531 genes) for more than 800 patients, we selected the 2264 genes having the more important variances because the analysed dataset is a RNA-Seq dataset and genes with the higher variances are those which have the higher probabilities to cause cancer. Then, we reduced this number to 397 genes for the last network configuration.

	1st Config	2nd Config	3rd Config	4th Config	5th Config	6th Config	7th Config
Batch Size	32	32	32	32	32	32	6
Features size	2264	2264	2264	2264	2264	2264	397
Lasso Regularization Term	0	0.01	0.1	0	0.01	0.1	0
Ridge Regularization Term	0	0	0	0	0	0	0.001
Number of Layers	4	4	4	3	3	3	4
Number of Neurons	1624	1624	1624	560	560	560	17
Dropout Layer	0	0	0	0	0	0	1(0.01)

Figure 1: Array of Neural Network Configurations

RESULTS

Data Exploration

As we can see on the figure (Figure 2), the UMAP Dimension Reduction Technique is the best technique here to visualise our data. It can be explained by the fact PCA doesn't allow to visualise all the data from the dataset and this one doesn't allow to visualize no-linear relations between data contrary to the UMAP. Compared to the PCA, the UMAP owns several parameters (number of neighbours, minimal distance between data, number of components and metric to use to compute distance in the new space) which is possible to modify, according to the studied data. This is also the main reason UMAP is not used for analysis such GWAS, the diversity of parameters for this analysis is so important that UMAP can't be used as covariable (6, 7). Then, we can see on the UMAP that patients having the same types of cancer are clustered between them on the Dimension Reduction.

Moreover, we can see some exceptions exist which one can also be visualised on the PCA. But in general, patients having the same type of cancer tend to be clustered between them certifying that the main relation linking patients is their types of cancers which is very important for the learning process by the neural network.

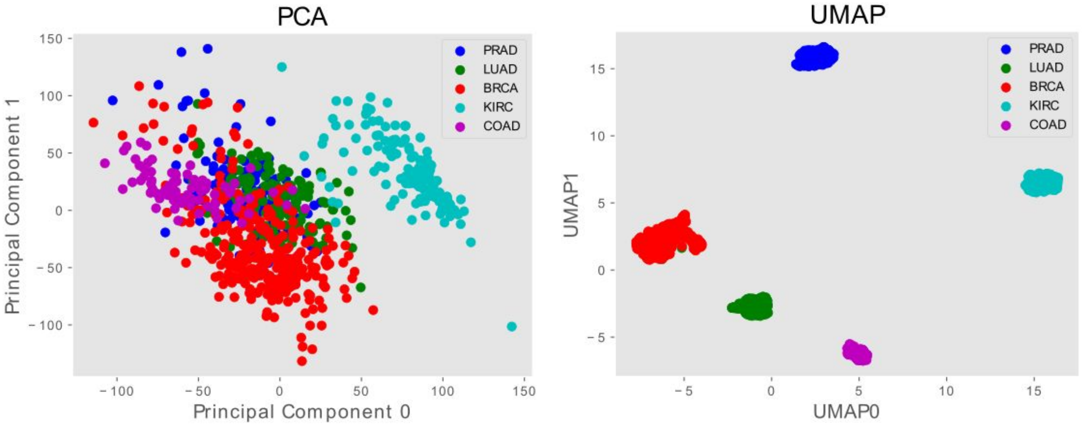

Figure 2: Comparison of UMAP and PCA Dimension Reduction techniques

Neural Network Training

It could be noticed by studying loss and accuracy curves that the fourth configuration of the neuronal network gives an overfitting effect (Supplementary Data). Compared to the obtained results with this configuration, we could observe that adding a Lasso Regularization Term (to prevent overfitting) on the activity (layer output) allowed to produce loss curves similar to those obtained for the 1st, 2nd and 3rd neural networks (Supplementary data). We can see on these loss curves that test loss curves remain to 0 and the training loss curves decrease suddenly. By watching the accuracy curves on these configurations, we could see that these neural network configurations give an overfitting learning. To correct this overfitting effect, we reduced the number of features as well as the batch size (the number of features to train per pass on the training dataset) and the number of neurons (17 neurons). Moreover, we added a dropout layer with a dropout rate of 0.01, the dropout layer allowing to select random neurons which are ignored during the training. A Ridge regulation term was applied on the weights of the network. After the construction of this 7th network, we could notice that the overfitting effect could be reduced but the neural network always tends to overfit on data.

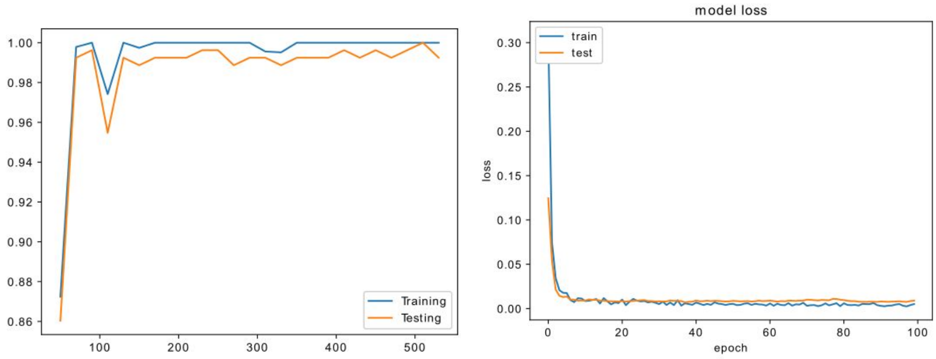

Figure 3: Accuracy curves and loss curves for the 7th Neural Network Configuration
For the Accuracy curves, the x-axis represents the size of the training dataset and the y-axis represents the accuracy

DISCUSSION

Neural Networks are models which are known to be efficient to make predictions, this is why we chose to study this model to predict types of cancers for patients from the Cancer Genome Atlas dataset. Although the overfitting effect could be reduced, the network can be further improved. During this study, we chose to build neural networks with the Stochastic Gradient Descent optimizer to optimize the loss function. Moreover, we could in a further study apply other optimizers in the aim to verify these one allow to prevent overfitting. In addition, we only used binary cross entropy loss function to train neural networks. The binary cross entropy loss function is mainly efficient for predictions of binary variables, it could be interesting to test others loss functions which could be more efficient to predict values of multi-classified variables (predict several types of cancers). Then, we could compare several machine learning models to choose the model maximizing the learning efficiency on the Cancer Genome Atlas dataset to select the best model to learn this dataset.

SUPPLEMENTARY DATA

A Readme is available in the working space.

REFERENCES

1. Weinstein,J.N., Collisson,E.A., Mills,G.B., Shaw,K.M., Ozenberger,B.A., Ellrott,K., Shmulevich,I., Sander,C. and Stuart,J.M. (2013) The Cancer Genome Atlas Pan-Cancer Analysis Project. *Nat. Genet.*, **45**, 1113–1120.
2. Lever,J., Krzywinski,M. and Altman,N. (2017) Principal component analysis. *Nat. Methods*, **14**, 641–642.
3. McInnes,L., Healy,J. and Melville,J. (2020) UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv180203426 Cs Stat*.
4. Pedregosa,F., Varoquaux,G., Gramfort,A., Michel,V., Thirion,B., Grisel,O., Blondel,M., Müller,A., Nothman,J., Louppe,G., *et al.* (2018) Scikit-learn: Machine Learning in Python. *ArXiv12010490 Cs*.
5. keras-team/keras (2021) Keras.
6. Diaz-Papkovich,A., Anderson-Trocmé,L., Ben-Eghan,C. and Gravel,S. (2019) UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. *PLoS Genet.*, **15**.
7. A review of UMAP in population genetics | Journal of Human Genetics.