Phenotypes Prediction from Gene Expression Data with Deep Multilayer Perceptron and Unsupervised Pre-training

Blaise Hanczar*, Mathieu Henriette, Toky Ratovomanana, Farida Zehraoui
IBISC, Univ. Evry, University Paris-Saclay, IBGBI, Bd de France, 91000 Evry, France.

* Corresponding author. Tel.: +33 1 64 85 34 61; email: blaise.hanczar@ibisc.univ-evry.fr
Manuscript submitted June 15, 2017; accepted December 22, 2017.
doi: 10.17706/ijbbb.2018.8.2.125-131

Abstract: Machine learning is widely used for phenotype prediction from gene expression data. However, deep learning, that is currently one of the most performant methods, have been very few studied for this problem. In this paper we construct a deep multilayer perceptron using different regularization methods to deal with the problem of small training samples. A large set of unlabeled data is used in an unsupervised pre-training procedure in order to improve the learning of the neural network. The results on several public microarray datasets show that the deep learning improves significantly the performance of the state-of-the-art.

Key words: Deep learning, phenotype prediction, gene expression data.

1. Introduction

The omics technologies, genomic (DNA sequencing), transcriptomic (microarrays), proteomic (protein chips, tissue arrays), allow providing massive molecular-scale information about the patients. The efficient analysis of these types of large scale data to improve the diagnosis and the prognosis, explain the causality of each disease and individualize the treatment for each patient still remains a challenge. This falls within the scope of the personalized medicine [1]. Gene expression chips record RNA transcripts from DNA, allowing studies from clinical samples of complex pathologies. For example, in breast cancer several gene signatures for prognostic have been proposed [2] and few of them has been adopted for clinical diagnosis. Reported predictive gene lists may be unstable and need to be considered with caution [3]. The high dimension of the gene expression data, the insufficient number of samples, the class unbalance and heterogeneity of tumor samples and patient characteristics are the most important challenges in microarray data analysis.

We address these issues by using deep neural networks [4], which is one of the most efficient machine learning algorithms. Deep learning methods have multiple levels of representations corresponding to different levels of abstractions. The high dimension of the gene expression data is taken into account using regularization approaches: the early stopping and the dropout. The last can be considered as an ensemble method with shared parameters. To overcome the problem of the low number of samples, we propose, in addition to the regularization methods, to use an unsupervised model pre-training, which allows to use all the available datasets related to our disease in order to initialize the model parameters. Since a neural network is considered as a black box, we propose to interpret the constructed model using biological

information. This allows to obtain a high reliability on our results.

In the paper, we first address the gene expression analysis problem using deep learning. Then we present our deep learning model with the unsupervised pre-training. We show the efficiency of the algorithm by presenting numerical results on real datasets and give a biological interpretation of some results. We conclude by giving some perspectives of this work.

2. Deep Learning for Gene Expression Analysis

In the last years, deep learning has become one of the most promising methods in machine learning [4]. It is making major advances in solving problems that have resisted the best attempts of the artificial intelligence community for many years. Deep learning methods are representation-learning methods with multiple levels of representation, obtained by composing simple but non-linear modules that each transforms the representation at one level (starting with the raw input) into a representation at a higher, slightly more abstract level. With the composition of enough such transformations, very complex functions can be learned. For prediction tasks, higher layers of representation amplify aspects of the input that are important for discrimination and suppress irrelevant variations. Deep learning is very good at discovering intricate structures in high-dimensional data. Its primary domain of application is image recognition and speech recognition where it has beaten other machine-learning techniques records [5]. Deep learning is promising in many other domains of science and especially in bioinformatics.

At the moment, few works have been published about deep learning for gene expression analysis. Most of these works focus on the analysis of the genes. For example, Chen et al. presents a deep learning model to infer the expression of target genes from the expression of landmark genes [6]. They identify interesting neurons by capturing strong correlations between the landmark genes and target genes. Deep learning is also used for unsupervised tasks [7]. Autoencoders are used to learn a compact representation of the gene expression of the yeast cell cycle. It has been shown that the expression profiles regenerated by autoencoders improve the performance of gene clustering. In another study, unsupervised model like restricted Boltzmann machines and sparse autoencoders are used on yeast transcriptomic data [8]. By integrating knowledge from gene ontology and KEGG in the networks, they show that these models are capable of learning biologically sensible representations of the data and revealing novel insights regarding the machinery regulating gene expression. Very few work has been published about deep learning for phenotype prediction based on gene expression data. The reason comes probably from the difficulty of learning a neural network with small training set. One of the most interesting study uses sparse, stacked autoencoder and PCA to reduce the dimension of the data, then a neural network is constructed for cancer prediction [9].

3. Methods

3.1. Deep Multi-layers Perceptron

The deep learning model that we propose in this paper is based on a multilayer perceptron with an unsupervised pre-training step. In the multilayer network the neurons are organized in layers where each neuron is only connected to all neurons of the previous layer and all neurons of the next layers. The first layer receives the expression profiles, each neuron takes the expression one probe. The last layer gives the probabilities to belong to each class (one neuron for each class). This model is illustrated on the bottom panel of the figure 1. Let's $h_i^{(l)}$ the i-th neuron of the layer l, its activation is defined by:

$$h_i^{(l)} = f\left(\sum_{i=0}^{n_{l-1}} w_{j,i}^{(l)} h_j^{(l-1)}\right)$$

where $w_{j,i}^{(l)}$ is the weight of the connection from the j-th neuron of the layer l-1 to the i-th neuron of the layer l. The weights from the layer l-1 to the layer l are represented by a $n_{l-1} \times n_l$ matrix, where n_i is the number of neurons in the layer l. Note that a bias term is included in the activation computation of the neuron, $h_0^{(l-1)}$ is set to 1, the value of the bias is therefore $w_{0,i}^{(l)}$. f is the activation function, the most common activation function are the sigmoid function $\sigma(x) = {}^{1}/_{1+\exp(-x)}$, the hyperbolic tangent $\tanh(x)$ and the Rectified Linear Unit $ReLU(x) = \max(0, x)$. ReLU is currently the most used activation function in the deep learning community because of the simplicity to compute its derivative. The last layer is a special case, since its gives probabilities, the neuron values have to be in the interval [0,1] and their sum be 1. The softmax activation function is therefore used: $softmax(x_i) = {}^{\exp(x_i)}/_{\sum_i \exp(x_i)}$. The neural networks is fitted in order to find the weights minimizing the cross entropy between the output of the network and the true class vector. The learning algorithm uses a stochastic gradient descent where the training set is divided into several minibatch. The learning rate is automatically adapted to the gradient descent with the adadelta algorithm [10].

We also use the batch normalization that is now a common trick to improve the performance of neural networks [11]. It speeds up the learning of the network and acts as in a legalizer in avoiding the saturation of the neurons. Its principle is to normalize over a batch of training examples, the values received by each neuron i.e. setting the mean to 0 and variance to 1.

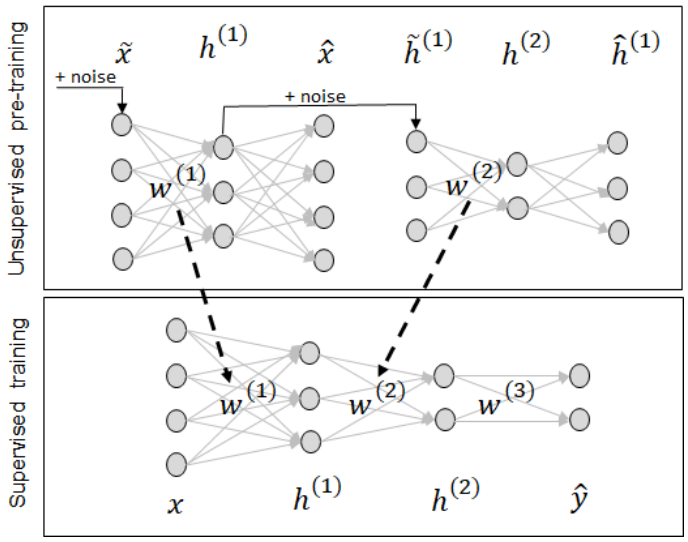

Fig. 1. Deep multilayer perceptron with two hidden layers.

The top panel shows the unsupervised pre-training of the two hidden layers with denoising autoencoder. The bottom panel shows supervised training of the MLP.

3.2. Dealing with the Overfitting

The small number of training examples in the available microarray datasets raises some problems in the learning procedure. The neural networks tend to overfit the training data, i.e. random sampling fluctuations will allow some combination of the input variables to match variable to predict perfectly over the limited samples we have, but the correlations will fall apart for a different set of samples. This overfitting can be greatly reduced by the use of dropout [12]. Dropout consists to switch off a random subset of the inputs and hidden neurons, i.e. set their output to 0. It is similar to train a large ensemble of networks that share

weights and is remarkably effective as a regularizer. Another simple method to reduce the overfitting is to use the early stopping [13]. We monitor the cross-entropy on a validation set and the gradient descent is stopped when validation cross-entropy begin to increase.

3.3. Unsupervised Pre-training

The number of samples for a given phenotype prediction task is generally small, however many other gene expression profiles not related to this phenotype are available. These profiles are grouped together to form a large dataset of samples without labels. This unlabeled dataset cannot be used to predict the phenotype but are useful to construct a hierarchical representation of gene expressions in the neural network. The idea is to find nonlinear combinations of inputs giving useful patterns for gene expression analysis.

The unlabeled dataset is used to initialize the weights of multilayer perceptron before the supervised learning. We pre-train iteratively each hidden layer in learning a denoising autoencoder that reconstruct the output of the previous layer. The principle is illustrated in the top panel of the figure 1. Let be $h^{(l)}$the hidden layer, we construct a network where the input layer is the output of the layer $h^{(l-1)}$with Gaussian noise $N(0,\sigma)$. This input layer is connected to the hidden layer $h^{(l)}$with the weight matrix $w^{(l)}$. The hidden layer is connected to the output layer of the same size than the input layer. The denoising autoencoder tries to reconstruct the input without noise. The gradient descent minimizes the mean square error between the output of the previous layer and the output of the denoising autoencoder. After the learning of the denoising autoencoders, the weights $w^{(l)}$ are reported to the multilayer perceptron. This procedure is repeated for each hidden layer.

4. Results

4.1. Experiments Design

Table 1. Description of the Datasets. The Last Row Describes the Unlabeled Datasets Used for Unsupervised Pre-training

Data ID	Context	#examples	#classes
GEOD 25066_A	Treatment response prediction for breast cancer	388	2
GEOD 25066_B	Prognostic for breast cancer	508	2
GEOD 25066_C	Treatment response prediction for breast cancer	417	3
GEOD 19301_A	Prognostic for Asthma	664	3
GEOD 19301_B	Prognostic for Asthma	651	2
GEOD 68465	Prognostic for lung cancer	442	2
GEOD 5364	Diagnostic of cancer	341	2
TABM 185	Unlabeled data for pre-training	5896	unlabeled

We test deep multilayer perceptron on several tasks of phenotype prediction based on public gene expression datasets. From the ArrayExpress database**,** we get seven gene expression datasets using the microarray HG-U133A. For the pre-training task, we use a large set of unlabeled data containing the expression profiles from many HG-U133A integrated datasets available on ArrayExpress. All datasets are normalized such that the mean and variance of gene expressions are respectively zero and one. The description of these datasets are summarized in the Table 1.

We compare the performance of deep multilayer perceptron with the popular methods from the state-of-the-art i.e. support vector machine, random forest and boosting. These algorithms are applied on the different classification tasks and the classifier accuracy is estimated by 10-fold cross-validation. For the multilayer perceptron, we test different architectures of networks. The best performances are obtained by three hidden layers containing between 500 and 100 ReLU neurons. The dropout rate is 0.2, the batch size

is 16. For the autoencoder, we add a Gaussian noise $N(0,1)$. The performance of the state-of-the-art algorithms may be sensitive to different hyper-parameters. For a fair comparison with the MLP, we test different values of these hyper-parameters. For the support vector machine, we test linear and Gaussian kernel with different values of C and Gaussian variance, for boosting we use adaboost with decision trees with different numbers of iterations and for random forest, we test different numbers and sizes of the trees. All these hyper-parameters are tuned in an intern 10-fold cross-validation procedure. The accuracies obtained by the different methods are reported in the table 2. We see that the deep learning (with or without pre-training) gives the better accuracies for all classification tasks. We can consider that deep learning have the same performance than the state-of-the-art for the datasets GEOD 25066 A and B since the improvement is low (around 1%). For the other datasets the performance of deep learning is significantly better than the state-of-the-art and especially for the datasets GEOD 68465 and 5364 where the improvement is around 10% of accuracy. We also see the impact of unsupervised pre-training. Indeed, in two datasets (GEOD 25066_C and 19301_B), the accuracy without pre-training is slightly higher than that with pre-training. The small difference of accuracy means that the pre-training had not impacted these two datasets. For the five other datasets, the pre-training improves the performance of the mutilayer perceptron.

Table 2. Accuracy Results Obtained by the Different Methods on the Gene Expression Datasets

Data ID	SVM	RF	Boosting	MLP	MLP + pre-training
GEOD 25066_A	79.2	79.9	76.2	80.3	**81.6**
GEOD 25066_B	76.3	76.1	75.1	76.7	**77**
GEOD 25066_C	42.5	46.3	45.1	**49.3**	48.4
GEOD 19301_A	48.8	51.0	52.9	53.2	**56.1**
GEOD 19301_B	50.4	53.3	47.9	**56.8**	55.9
GEOD 68465	58.0	53.9	55.9	65.9	**69.8**
GEOD 5364	68.1	72.0	70.9	73.6	**78.4**

4.2. Biological Interpretation

We propose to analyze and interpret the results of a deep learning network trained with gene expression data using biological information (Gene ontology, KEGG and Disease ontology Lite). We present the results obtained from the deep neural network associated to the breast cancer treatment prediction (GEOD 25066_A dataset). We first select, for each neuron in the hidden and output layers, the lists of genes that contribute to it activation. We have defined a threshold to select the related genes. We focus on the results of the analysis of three interesting neurons from the first hidden layer. Two neurons identify characteristics related to different molecular subclasses of the breast cancer with negative HER2 status and one neuron identify hypermethylated genes in breast cancer. There are 2 molecular subclasses of invasive breast cancer with negative HER2 profile: breast cancers of basal phenotype and so-called breast-like breast cancers.

The results obtained from the first neuron show characteristics evoking basal phenotype cancer. Indeed, the list of genes related to this neuron is enriched in processes related to the epithelium of the mammary gland (GO: 0061180) and to the p53 signaling pathway already identified as implicated in various breast cancers. The most frequent breast cancer is adenocarcinomas, which is developed from epithelial cells in the mammary gland. But still, in more than 80% of the cases, the TP5 gene, which belongs to the list, is mutated for patients with basal - like breast cancer. The second neuron shows features suggestive of breast-like breast cancers. This type of cancer is HER2 negative and is characterized by expression of genes observed in breast tissues and adipose tissues. Having detected an enrichment of the genes involved in the signaling of the ERBB2 pathway, this neuron also identifies enrichment in the adipose tissue development process.

Moreover, in the literature, more than one hundred genes have been reported as methylated hypers in mammary tumors. BRCA1, APC and BCL2 are examples of genes that are silenced by hypermethylation in breast cancer [14], [15]. The third Neuron, having identified these genes as significant, appears to be specialized in the capture of methylated hypersensitive genes in breast cancer.

5. Conclusion

In this paper, we show how to use deep learning for phenotype prediction from gene expression data. The main difficulty comes from the small number of available training examples and the high dimension of the data that leads to overfitting of the model. To reduce the impact of overfitting we use regularization methods like early stopping, dropout and batch normalisation. We also exploit the unlabeled data to pre-train the network in constructing new features adapted to gene expression. The deep learning methods have been tested on several public microarray datasets and we have obtained better performances than the state of the art. We also show that some neurons capture relevant biological information related to the predicted phenotype.

Acknowledgment

We thank for their support Pr Jean-Daniel Zucker, Pr Yann chevaleyre and Dr Edi Prifti from the ICAN institute.

References

[1] Davis, J., Lantz, E., Page, D., Struyf, J., Peissig, P., Vidaillet, H., & Caldwell, M. (2008). Machine learning for personalized medicine: Will this drug give me a heart attack? *ICML Health Workshop*.

[2] Michiels, S., Koscielny, S., & Hill, C. (2005) Prediction of cancer outcome with microarrays: A multiple random validation strategy. *Lancet*, *365(9458)*, 488-492.

[3] Miller, L.D., Smeds, J., George, J., Vega, V. B., Vergara, L., Ploner, A., Pawitan, Y., Hall, P., Klaar, S., Liu, E. T., & Bergh, J. (2005) An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proc Natl Acad Sci USA*, *102(38)*, 13550-5.

[4] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436-444.

[5] Krizhevsky, A., Sutskever, I., Hinton, G. (2012) ImageNet classification with deep convolutional neural networks. *Neural Information Processing Systems*, *25*, 1090–1098.

[6] Chen, Y., Li, Y., Narayan, R., Subramanian, A., & Xie, X. (2016). Gene expression inference with deep learning. *Bioinformatics*, *32(12)*, 1832-1839.

[7] Gupta, A., Wang, H., & Ganapathiraju, M. (2015). Learning structure in gene expression data using deep architectures, with an application to gene clustering. *Proceedings of 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 1328-1335).

[8] Chen, L., Cai, C., Chen, V., & Lu, X. (2016). Learning a hierarchical representation of the yeast transcriptomic machinery using an autoencoder model. *BMC bioinformatics*, *17(1)*, S9.

[9] Fakoor, R., Ladhak, F., Nazi, A., & Huber, M. (2013). Using deep learning to enhance cancer diagnosis and classification. *Proceedings of 30th International Conference on Machine Learning, WHEALTH workshop.*

[10] Zeiler, M. (2012) ADADELTA: An adaptive learning rate method. *CoRR*, abs/1212.5701.

[11] Ioffe, S., & Szegedy, C. (2013) Batch normalization: Accelerating deep network training by reducing internal covariate. *International Conference on Machine Learning ICML*, Vol. 37.

[12] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov R., (2014) Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res., 15(1)*.

[13] Federico, G., Jones, M., & Poggio, T. (1995). Regularization theory and neural networks architectures.

Neural Computation, 7(2), 219–269.

[14] Jovanovic, J., Ronneberg, J. A., Tost, J, & Kristensen, V. (2010). The epigenetics of breast cancer. *Mol Oncol., 4(3)*, 242-54.

[15] Stefansson, O. A., & Esteller, M. (2013). Epigenetic modifications in breast cancer and their role in personalized medicine. *American Journal of Pathology,* 183(4), 1052-63.

Blaise Hanczar is professor at Paris Saclay - Evry Val d'Essonne University, France. He was professor assistant from 2008 to 2015 at the Paris Descartes University. He got his PhD in 2006 at the university Paris 13 on supervised learning for microarray data analysis. His current research interests are about deep learning, reject classification, error estimation and stability. His main domain of application are about medical application, genomics and meta-genomics.

Farida Zehraoui is a professor assistant at Paris Saclay - Evry Val d'Essonne University, France. She is interested in machine learning approaches and their application to biology and personalized/individualized medicine. Her current work focuses on deep supervised and unsupervised neural networks, kernel based methods for heterogeneous data, and combining machine learning with multi-agent systems for medical health care.