Feature extraction Of Genomics Data Using Deep Learning

Raquel Aoki
Phd Thesis
Computing Science
raoki@sfu.ca

Meghna Garg
MSc Thesis
Computing Science
mgarg@sfu.ca

Aniket Mane
MSc Thesis
Mathematics
amane@sfu.ca

Abstract

Due to the great advances in technology, bioinformaticians are now having access to large amounts of genomic and clinical data, which are growing exponentially. In genomics research, one of the biggest challenges we face is to extract features from the datasets. The structures are usually nonlinear and need algorithms to learn high-level representation of features.

1 Introduction

With the advent of deep learning, we have now been granted access to a plethora of optimization techniques that can handle large-scale computations. The field of genomic research presents excellent opportunities to put these techniques to good use. Firstly, due to recent innovations in bioinformatics, we now have an abundant amount of data available to us, allowing more sophisticated analyses. Secondly, due to the complex nonlinear structures of this data, operations over these structures become very difficult. Research studies have been conducted in this area demonstrating that extracting features from genetic data can help obtain better results.

Some researchers believe that working with extracted features is not a good approach if you can work with the original data, working with a subset of your original universe or with a representation in a lower dimension might lose you some information. However, in this work, we point out that working with extracted features can also have advantages. The main advantage is to reduce overfitting and avoid the curse of dimensionality. Other advantages include reduced training time when we use these features in other machine learning algorithms or techniques. So, if the features are well extracted and are in fact a good representation of the original dataset, it might be beneficial to use the new features than the original ones.

In this project we extracted features from a genomic dataset of breast cancer patients. We work with three different and independent methods to extract features, Deep Feature Selection (DFS), Denoising Autoencoder (DA) and Random Forest (RF). In order to evaluate the quality of extracted features, we use them to predict some clinical information such as overall survival rate or type of breast cancer (ER and HER2) that may have affected the patient. We use Logistic Regression to predict the clinical information for the test set, for all three sets of extracted features (one for each clinical information).

We describe the dataset we used, in Section 2. In Section 3 we introduce the methods used to extract features, how they work and their main characteristics. We discuss the experiment results in Section 4. Finally, we analyze these results and obtain our inferences in Section 5.

2 Dataset

We use the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) data as a training set and The Cancer Genome Atlas (TCGA) data as an independent testing set. The METABRIC data set has 1461 patients and the TCGA has 520 patients.

Although the two datasets don't have the same set of genes, we work just with 2520 genes that aree present in both studies. Excluding the genes not present in both datasets is a necessary step that allows us to use one dataset for training and the other as a test set. Each row represents a patient and each column, a different gene. The values inside the datasets represent the gene expression level of that gene in that patient. These values are between 0 and 1.

The clinical information was kept in an independent dataset and we link those datasets by patient id. In this paper, we work just with Overall Survival and the cancer types ER and HER2, that were present on METABRICS and TCGA clinical information.

The datasets (METABRICS and TCGA) were collected from the supplementary material of Tan et al. [2014] available on `http://discovery.dartmouth.edu/~cgreene/da-psb2015/`. The clinical information was collected on the website `http://www.cbioportal.org`.

3 Methods

In literature there are several methods to extract features and the choice about which method should be use depends mainly of characteristics of the dataset. If the relations between the features are simple, linear for example, it is not necessary to use sophisticated or complex methods to extract those features. But if the relations between the features are unknown, complex or non-linear, it is necessary to use more elaborate methods in order to extract features without losing important information present on dataset.

One of the most famous techniques to reduce dimensionality is Principal Component Analysis (PCA), but we chose to not use this approach here for two reasons. Firstly, if we run PCA in training and testing sets separately, we can end up with two different set of features. There is no guarantees or alternative to force the same reduction in both datasets; Secondly, if we combine the two datasets, perform the PCA and split them again, we will have already witnessed the test dataset, which violates the basic caveat of machine learning.

Besides, PCA does not perform well on non-linear datasets and it is well known that the relation between features in genomic datasets is complex and non-linear [Li et al., 2015]. Thus, it is necessary use more complex methods to extract those features. In this project we chose work with Deep Feature Selection(DFS), Denoising Autoencoder(DA) and Random Forest(RF).

It is important to point out that DFS and RF will generate a set of features for each clinical information that we focus on in the evaluation phase. During the training, both these methods work with the clinical information as output/dependent variable. The DA, on the other hand, does not depend on clinical information during the training phase. As a result, we will have just one set of features to work with all clinical information.

We are going to work with 3 clinical information to evaluate the quality of extracted features: overall survival (patient alive or not), ER (positive or negative) and HER2 (positive or negative). Positive ER and HER2 indicate that the cancer is more aggressive and thus they are related with metastasis of cancer. To compare the methods, we fix the dimension of the extracted features to 100 in all three methods.

3.1 Deep Feature Selection

Through piling up hidden layers, deep neural networks are able to model the non-linearity of features. The main advantages of applying deep learning models in this problem is their capacity to process complex systems with non-linear structures.

Based on Multilayer Perceptrons (MLP), which is a deep feed-forward neural network, the Deep Feature Selection (DFS) proposed by Li et al. [2015] adds an extra layer representing the weight of each input feature. This extra layer is a sparse one-to-one layer, as showed in Figure 1.

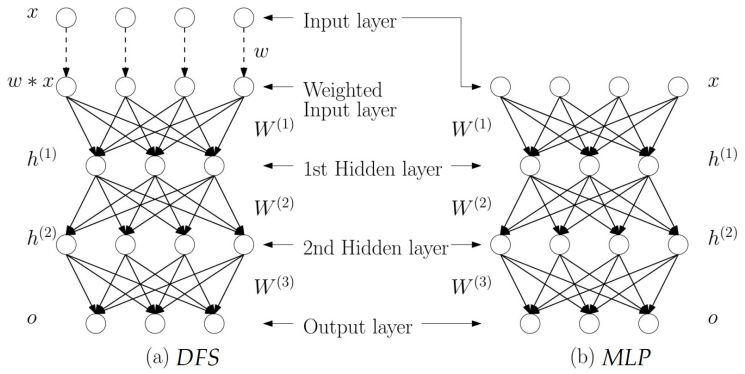

Figure 1: Comparison between a Deep Feature Selection (DFS) network and a Multi Layered Perceptron (MLP) structure

The DFS requires a target output (or a set of target outputs) during the training part. In our case, we use a single target output for every clinical information (overall survival, ER, HER2) that we are interested in. Thus, we train the DFS 3 thrice, once for each clinical information and obtain 3 sets of features. The intuition behind this decision is to be able to isolate genes that are important for every individual clinical trait that we predict. As genes were the features of our original dataset, the features extracted in the DFS are also genes. We choose the top 100 genes with highest weight in the weighted input layer.

Note that since the weight matrix w is to be sparse, we need to use a sparse regularization term on w. This is necessitated by the fact that the features (genes) in our dataset outnumber the samples (patients). We use the same regularizer as used by Li et al. [2015] (elastic net): $\lambda_1\Big(\frac{1-\lambda_2}{2}||w||_2^2 + \lambda_2||w||_1\Big)$. A similar (elastic net) regularizer was used for subsequent, hidden layers with a different choice of λ_1 and λ_2 than for the weighted input layer. Thus, effectively we try to minimize:

$$f(\theta) = l(\theta) + \lambda_1\Big(\frac{1-\lambda_2}{2}||w||_2^2 + \lambda_2||w||_1\Big) + \alpha_1\Big(\frac{1-\alpha_2}{2}\sum_{k=1}^{K+1}||W^k||_2^2 + \alpha_2\sum_{k=1}^{K+1}||W^k||_1\Big)$$

where $\theta = \{w, W^1, b^1, ..., W^{K+1}, b^{K+1}, W^k\}$ is the weight vector for the kth hidden layer and $l(\theta)$ is the log-likelihood function. All bias vectors b^i, $i = 1$ to $K+1$ have been initialized to zero vectors.

Due to the non-convexity of $f(\theta)$ the gradient descent algorithm does not always guarantee the global minimum for higher values of K (higher number of hidden layers). Hence, we perform our experiments with 2 hidden layers, which yields better results than more hidden layers. For every hidden layer the activation layer is sigmoid which performs reasonably better than both ReLU or tanh functions. The first hidden layer contains 128 nodes while the second contains 64 nodes. Finally, the activation function for the output is softmax.

3.2 Denoising Autoencoders

Autoencoder is one of the techniques to reduce dimensionality of datasets. In simple words, an autoencoder aims to learn a representation of the input in a lower dimension using as output/target the input itself. Image and audio analysis are example of fields that use autoencoders to represent the datasets in a lower dimension. Here, we are going to use an autoencoder to represent a large set of genes in a lower dimension.

To force the autoencoder to learn more robust features, we add a noise layer after the input to increase the dataset variation. This autoencoder with the added noise layer is called Denoising Autoencoder [Vincent et al., 2008]. The constructed DA network contains: an input layer, a noise layer, a hidden layer and a reconstructed layer (See Figure 2 for reference). The reconstructed layer and the input

layer have the same dimensions, and the algorithm's goal is minimize the difference between the output and input.

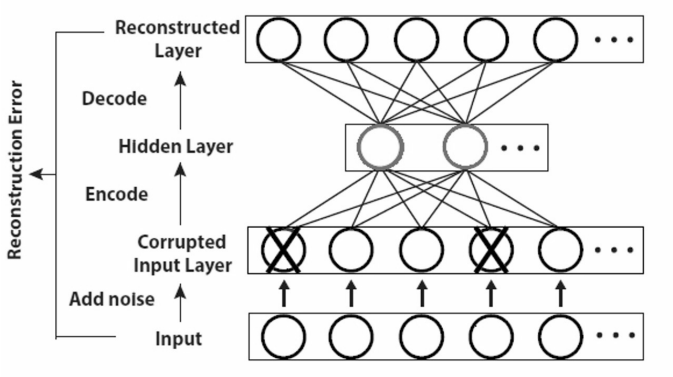

Figure 2: Denoising Autoencoders Illustration

In this method, the features are the result of the product between the genes on the dataset (input) and the hidden layer weights. To obtain this representation, we first train the DA as usual. Then, we obtain the product between the input and the weights extracted from the encoding part, resulting in a representation of the input data in a lower dimension. The size of the hidden layer is 100. The two activation functions were sigmoid and the loss function was the mean squared error.

In subsequent sections, we are going to compare the weight/importance of each gene on those methods. DA demands a different approach to extract the genes importance and we applied the same technique as presented in Tan et al. [2014]. Considering the column c_i in weights matrix (encoding part), the weight matrix is 2520×100. There are 2520 weights (one for each gene) in each column c_i, $\forall i = 1, .., 100$. Each column has a Gaussian distribution with mean $\bar{c_i}$ and standard deviation sd_{c_i}. Then, it assigns the value 1 to gene j if $w_{ij} < \bar{c_i} - 2sd_{c_i}$ or $w_{ij} > \bar{c_i} + 2sd_{c_i}$ and 0 otherwise. We repeat this process for all 100 columns and the final importance of the gene is given by the sum of those 0's and 1's.

We extract the features from DA in two different ways and we are going to test both strategies on testing set. In both cases the feature set have size 100.

1. The features are the product between input and encoding weights;

2. The features are genes with largest importance.

3.3 Random Forest

Random Forest is an ensemble method based on trees. It is a simpler method compared to DFS and DA, but in general presents very good results. Random forest provides a tree based strategies using a majority voting mechanism which gives us a better training and ranking of features. The advantages of this method results from the facts that they do not tend to overfit and generalize well with an internal unbiased estimate. Consequently, they scale well with huge dataset. Figure 3 shows a basic illustration about how a Random Forest works.

We performed the split using 'gini' criterion, where the impurity index at each node was calculated based on the fraction of records that belongs to i^{th} class. A high weight for one class (say i) leads to a stronger decrease of impurity when the number of records for that class increases. We considered the average of weights obtained at each step of 10-folds cross validation, performed on random forest algorithm. Among them, top 100 weighted features have been selected as the most important feature values.

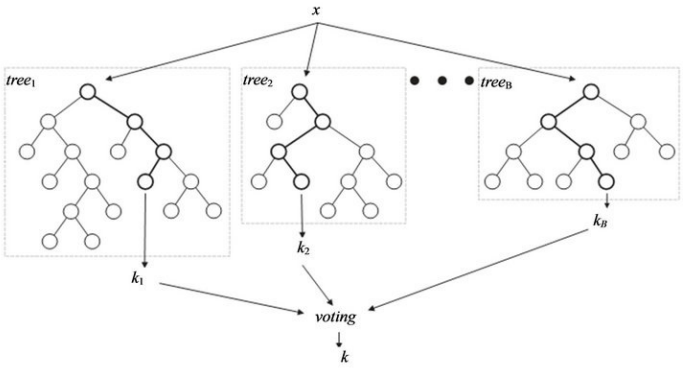

Figure 3: Example of Random Forest for classification

3.4 Approach

After extracting the features from each method described above, we evaluate the quality of each set of features by their predictive capacity. As already cited, the variables that we are going to predict is the patients' clinical information, as described in Table 1.

Table 1: Clinical Information

Abbv.	Description	Type
ER	Estrogen Receptor (ER) status	Binary
HER	Human epidermal growth factor receptor	Binary
OS	Overall Survival	Binary

Due the fact that all three variable be binary, we work with Logistic Regression. We use 100 features extracted from METABRIC to train the Logistic Regression (one model for each variable); we then evaluate the predictive capacity (accuracy) of this model using 100 features from TCGA. The results obtained are showed on the Section 4.

4 Experiment Results

In this section we present our final results. In Subsection 4.1 we show the correlation between the gene weights/importance in the methods and Subsection 4.2 shows the Logistic Regression results.

4.1 Weight Comparison

Figure 4 shows the correlation between the genes weights/importance for each pair of methods and clinical information. From this graphic it is possible to notice that Random Forest and DFS have better correlations. Therefore, they agree about the importance of most genes for predicting ER and HER2.

The comparison between top 100 genes (Table 2) shows that DFS and RF has maximum genes in common considering ER and HER2, thus, perfectly justifying the correlation for weights of 2520 genes between RF-DFS for same types being the highest (Figure 4).

Table 2: Pairwise comparison of three approaches in top 100 important genes

Clinical Information	DFS ∩ DA	DA ∩ RF	DFS ∩ RF
Overall Survival	10	4	8
ER Type	4	5	30
HER2 Type	7	5	23

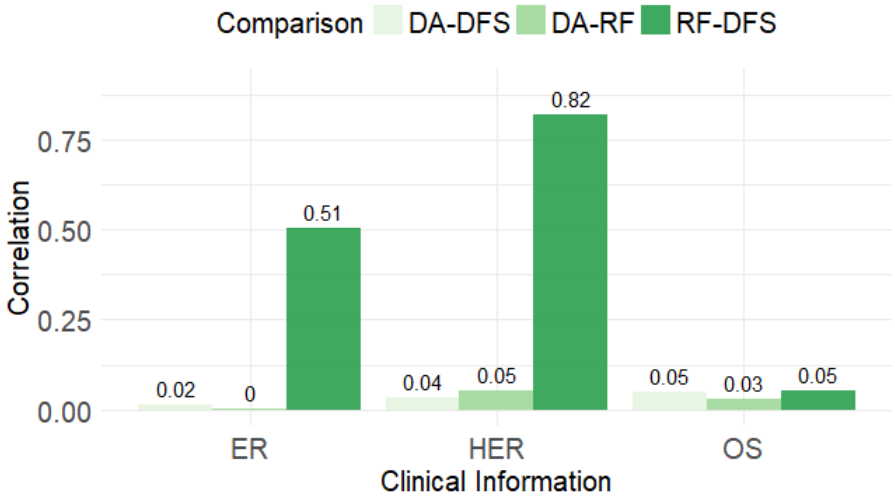

Figure 4: Genes weight/importance correlation

We performed some background research online to get the interpretation of genes obtained from our methodology to confirm our approach and verify if the results observed are significant enough to be highlighted as the most promising genetic characteristic. Below is the description of the genes present in the top 100 that are common to all 3 methods.

- **CLDN11**: Claudin 11 gene had a notable impact in determining the Overall Survival. Erstwhile research (Meng et al. [2016]) indicates higher expression of this gene in patients affected by Infiltrating Ductal Carcinoma (IDC) than in normal cases. Since this is the most common type of breast cancer, CLDN11 is a promising predictor of Overall Survival in breast cancer afflicted patients (Ma et al. [2014]).
- **PGR**: Progesterone is a steroid hormone associated with the female reproductive process, and is critical for normal female development and growth. High expression levels of the Progesterone Receptor gene have been indicative of aggressive cancer growth (Prat et al. [2013]). Malignant cells tend to be ER-positive with high expression levels of PGR thus making PGR a good indicator of this clinical feature.
- **TFF1**: Although typically known to play a role in ovarian cancer, Trefoil factor 1 also displays links with the the development of breast cancer. High expression of TFF1 has been related to proliferation and migration of breast cancer cells (Prest et al. [2002]). All three methods used in our experiments show TFF1 as a significant predictor of the patient being ER-positive.
- **MUCL1**: Mucin-like protein 1 has been identified as a potential marker for the diagnosis of breast cancer. It is related to HER2 and has an active role in cancer cell proliferation (Conley et al. [2016]).

4.2 Predictive Capability

First we evaluate the quality of the features extracted on the training set (METABRIC). Thus, after we fit the logistic regression models, we check their accuracy on testing set. Table 3 shows the accuracy on METABRIC. We noticed that DFS present the best accuracy to predict Overall Survival and HER2; for ER, Random Forest had the best accuracy. Figure 5 shows predictive capacity of those extracted features on the testing set. The features extracted by DFS were the better. For all 3 clinical information, the features from DFS presented the largest accuracy. The results of Random Forest were comparable with DFS results considering the simplicity of the method. The DA features (Hidden Layer representation and Top 100 genes) had good results for ER and HER2,

[1] 'CLDN11'(Top 100 in DA, DFS and RF to predict Overall Survival),'MUCL1'(Top 100 in DA, DFS and RF to predict HER2), and 'PGR','TFF1'(Top 100 in DA, DFS and RF to predict ER)

Table 3: Accuracy in METABRICS (training set)

Clinical Info.	DA - Hidden Layer	DA - Top100	DFS - Top100	RF - Top100	Top 4 genes [1]
Overall Survival	0.678	0.691	0.715	0.689	0.605
ER type	0.977	0.971	0.982	0.983	0.903
HER type	0.949	0.926	0.975	0.930	0.874

but its accuracy for Overall Survival was not very good compared to DFS and Random Forest and it is only slightly better than the results of the Top 4 Genes.

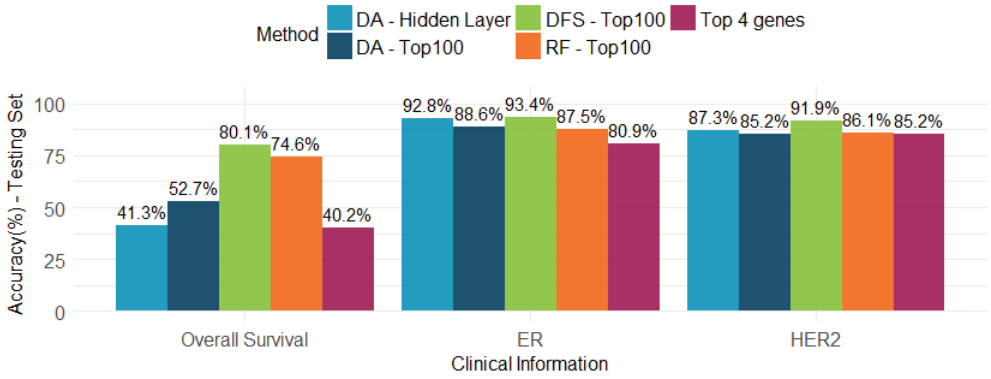

Figure 5: Accuracy in TCGA (testing set)

5 Conclusion

Based on the above evaluation methodology, considering the experiments carried out for DFS , DA and RF algorithms, we can claim that **Deep Feature Selection (DFS)** tends to be the best feature extractor for the given dataset of 2520 genes along with their clinical information used.

The low performance of DAs and Top 4 genes over the same dataset can be explained as the features are extracted from these methods are independent of clinical information and it is just one set of features for all 3 clinical information, while DFS and RF use one set of feature for each clinical information. In the DA, the clinical information is introduced at the time of training the logistic model itself. This also explains the lower correlation values of DA genes weight with the other two algorithm.

Finally, we infer that although all three methods are good for feature extraction, when there is a well defined target variable to predict, DFS and RF perform better. On the other hand, in an open problem, when the features need to generalize more information at same time, the DA method is a better choice.

Appendix

All analysis were developed in python. We used the libraries keras (DA), theano (DFS), sklearn (RF and Logistic Regression), pandas and numpy. The graphics were created at R using ggplot library.

The **contributions** of each group member are:

- Aniket Mane: Deep Feature Selection, interpretation of results, report and presentation;
- Meghna Garg: Random Forest, data collection, gene study and interpretation, report and presentation;
- Raquel Aoki: Denoising Autoencoder, data preparation, logistic regression, graphics, report and presentation

References

S. Conley, E. Bosco, D. Tice, R. Hollingsworth, and Z. Herbst, R.and Xiao. Her2 drives mucin-like 1 to control proliferation in breast cancer cells. volume 35, pages 4225–4234, 2016.

Y. Li, C. Chih-Yu, and W. Wyeth. Deep feature selection: Theory and application to identify enhancers and promoters. pages 205–217, 2015.

F. Ma, X. Ding, Y. Fan, J. Ying, S. Zheng, N. Lu, and B. Xu. A cldn1-negative phenotype predicts poor prognosis in triple-negative breast cancer. PLOS One, 2014.

L. Meng, Y. Xu, C. Xu, and W. Zhang. Biomarker discovery to improve prediction of breast cancer survival: using gene expression profiling, meta-analysis, and tissue validation. In *OncoTargets and Therapy*, 2016.

A. Prat, M. Cheang, M. Martn, J. Parker, E. Carrasco, R. Caballero, S. Tyldesley, K. Gelmon, P. Bernard, T. Nielsen, and C. Perou. Prognostic significance of pgrpositive tumor cells within immunohistochemically defined luminal a breast cancer. volume 31, pages 203–209. Journal of Clinical Oncology, 2013.

S. J. Prest, F. May, and B. Westley. The estrogen-regulated protein, tff1, stimulates migration of human breast cancer cells. volume 16, pages 592–594. FASEB journal : official publication of the Federation of American Societies for Experimental Biology, 2002.

J. Tan, M. Ung, C. Cheng, and C. S. Greene. Unsupervised feature construction and knowledge extraction from genome-wide assays of breast cancer with denoising autoencoders. In *Pacific Symposium on Biocomputing Co-Chairs*, pages 132–143. World Scientific, 2014.

P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103. ACM, 2008.