Feature Extraction in Genomic Datasets using Deep Learning

Raquel Aoki, Aniket Mane and Meghna Garg

Simon Fraser University

INTRODUCTION

Evaluate the selected features from 3 standardized methods listed below and measure the relevance with clinical data:

- **Deep Feature Selection (DFS)**
- **Denoising autoencoders (DA)**
- **Random Forest (RF)**

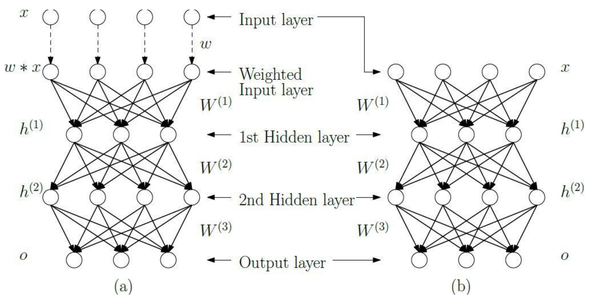
Why work with features extraction? It is way to reduce overfitting, have shorter training times in other algorithms and it also helps to avoid the curse of dimensionality.

Our contribution: Comparison of the features extracted from the above 3 algorithms in Genomic Datasets.

Why genomic dataset? Due to their complex nonlinear structures, the task of extracting features from thousands of genes based on their non-coding DNA sequence is found to be challenging in bioinformatics.

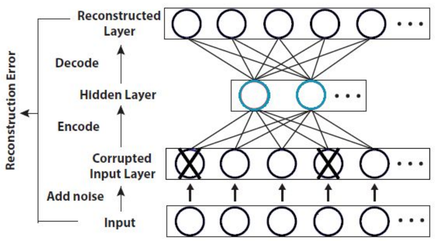
DEEP FEATURE SELECTION

- DFS adds an extra layer representing the weight of each input feature;
- Takes advantages of deep structures to model nonlinearity.



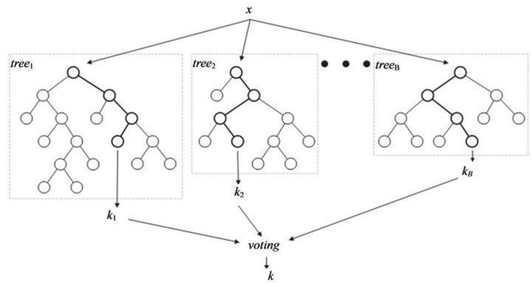
DENOISING AUTOENCODERS

- Autoenconders are used to learn a representation of data in lower dimensions;
- DAs learns a compact and efficient representations from input data by incorporating noise during training;
- A procedure which generates robust features.

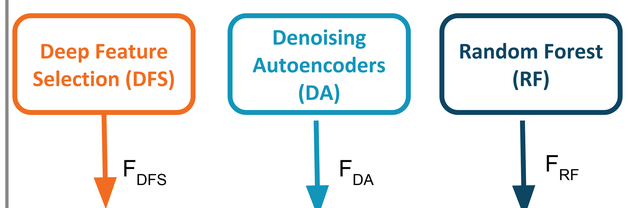


RANDOM FOREST

- Ensemble learning method for classification;
- The tree-based strategies uses a majority voting mechanism which give us a better training and a ranking of the features;
- Does not overfit and generalizes well with an internal unbiased estimate.



PROBLEM OVERVIEW



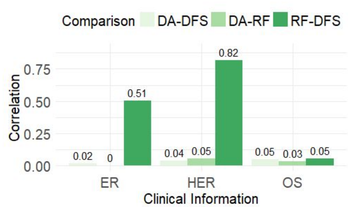
F_{DFS}	100 Genes with highest weight in DFS
F_{DA}	Representation on the DA hidden layer size 100
F_{RF}	100 Genes with highest feature importance in RF

Breast Cancer Datasets:

- Metabrics: 2520 genes and 1461 patients
- TCGA: 2520 genes and 520 patients

FEATURE ANALYSIS

- The correlations of the 2520 genes weights/importance between RF-DFS for type cancer ER and HER2 are the highest.



- The comparison between the top 100 genes shows that DFS and RF has 30 and 23 genes in common considering types cancer ER and HER2 respectively

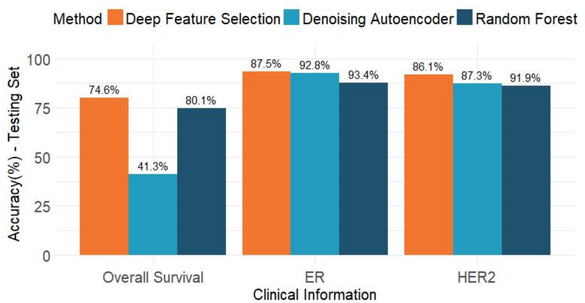
Clinical Inf.	DFS∩ DA	DA∩RF	DFS∩RF
OS	10	4	8
ER	4	5	30
HER2	7	5	23

EXPERIMENT RESULTS

- METABRICS was used as training set and TCGA as testing set.
- The features extracted were tested by their predictive capacity of Overall Survival, ER and HER2 (clinical features that predict aggravated breast cancer growth).
- We use Logistic Regression to link the features and the clinical information. In METABRICS (training set) the DFS had the best accuracy.

Clinical Inf.	DFS	DA	RF
OS	**71.5%**	67.8%	68.9%
ER	98.2%	97.7%	**98.3%**
HER2	**97.5%**	94.5%	93.0%

- Considering the results, ER had the highest accuracy among clinical information in all 3 methods.
- In TCGA(testing set) the DFS give the best set of features in all 3 clinical informations considered.



Conclusion: Based on the above evaluation among DFS,DA and RF, we claim DFS to be the best feature extractor for the given dataset of 2520 genes along with their clinical info.

REFERENCES

Unsupervised Feature Construction and Knowledge Extraction from genome-wide assays of Breast Cancer with Denoising Autoencoders. *Tan et al, Pacific Symposium Biocomputing 2015.*

Deep Feature Selection: Theory and Application to Identify Enhancers and Promoters. *Li et al, Journal of Computational Biology, 2016.*