A Deep Learning-Based Feature Selection Process on Leukemia Data

Alessia Auriemma Citarella
Department of Computer Science
University of Salerno
Fisciano (Salerno), Italy
aauriemmacitarella@unisa.it

Alfredo Di Nuccio
Department of Computer Science
University of Salerno
Fisciano (Salerno), Italy
a.dinuccio@studenti.unisa.it

Rita Francese
Department of Computer Science
University of Salerno
Fisciano (Salerno), Italy
francese@unisa.it

Maria Frasca
Department of Computer Science
University of Salerno
Fisciano (Salerno), Italy
mfrasca@unisa.it

Michele Risi
Department of Computer Science
University of Salerno
Fisciano (Salerno), Italy
mrisi@unisa.it

Genoveffa Tortora
Department of Computer Science
University of Salerno
Fisciano (Salerno), Italy
tortora@unisa.it

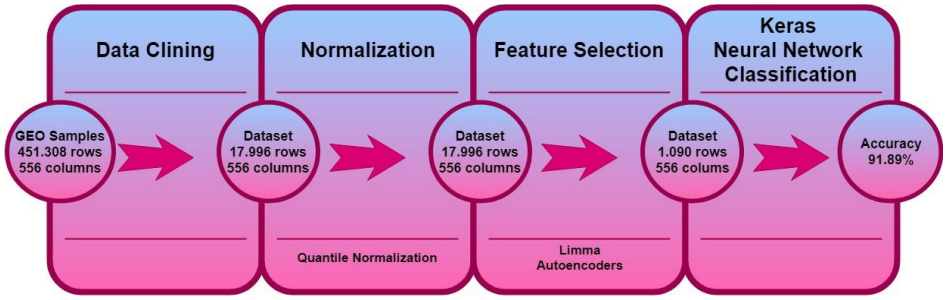

Figure 1: The analysis process

ABSTRACT

This paper aims at defining a feature selection analysis process mainly based on Deep Learning for a particular set of bioinformatics data. In particular, the analysis is carried out on a dataset of 556 patients affected by leukemia extracted from GEO platform public database [1]. The patients belong to two distinct classes: acute lymphoblastic leukemia (ALL) or acute myeloid (AML). Each of them is characterized by a list of identical genes for all the patients. The analysis exploits feature selection techniques aimed at reducing the consistent number of variables (genes).

KEYWORDS

Bioinformatic Data Analysis, Feature Selection, Deep Learning

1 FORMAL MODEL AND APPROACH

The proposed feature selection analysis process aiming at identifying the relevant genes is depicted in Figure 1. The analysis starts with a data clining of the data and removal of the batch effect and proceed with pre-processing techniques aimed at reducing the consistent number of variables (genes), to simplify and speed up the subsequent work of classification of the samples. The reduction in the number of variables (feature selection) is carried out by using specific statistical techniques for the treatment of bioinformatics data extracted from microarray experiments, aimed at classifying genes based on their differential expression between two different biological states (in this case, ALL and AML). Furthermore, we perform an additional feature selection using unsupervised deep learning model, autoencoder, to simplify and speed up the classification. Following the reduction in the number of variables, classification models have been implemented with the use of neural networks.

2 CONCLUSION AND FUTURE WORK

We obtained a classification accuracy of approximately 92%. This result is then compared with the support vector machines (SVM) to provide a broader view of the data classification problem. In addition, the gene enrichment analysis based on the functional annotation of the differentially expressed genes has been conducted. As a result, a differentially expressed pathway [2] between the two pathologies has been detected: the RNA degradation. This pathway is composed of 77 genes, of which 15 are included in the 1,090 genes examined in this work. Future work will be devoted to the analysis of the genes composing this pathway [3].

REFERENCES

[1] GEO Platform public database, https://www.ncbi.nlm.nih.gov/geo/
[2] M. Kanehisa and S. Goto, "Kegg: kyoto encyclopedia of genes and genomes," Nucleic acids research, vol. 28, no. 1, pp. 27–30, 2000.
[3] Aberrant RNA degradation in T-cell leukemia, 2014-2019. [Online]. Available: https://cordis.europa.eu/project/rcn/185655/factsheet/en