Universidade do Minho
Departamento de Informática

Intelligent Systems
Soft Computing
4º/1º Year, 2º Semester
2019/2020 Edition

Practical Work nº 1

Theme

Artificial Neural Networks in Clinical Examinations

Learning Objectives

With the realization of this practical work, it is intended that groups learn the following procedures:

- Preparation and analysis of datasets;
- Training and validation of learning models, specifically of Artificial Neural Networks (ANN);
- Optimization of learning model parameters;

Problem Statement

This practical work intends to be the starting point for the development of a predictive model using the Python development environment. For this, it will be necessary to develop a solution to the following problem:

Classify whether a mass detected in a mammography exam is benign or malignant, using an Artificial Neural Networks model (ANN).

In this practical work, it is intended to apply the knowledge taught during the curricular unit of Soft Computing for the detection of malignant masses in breast tissue, through the implementation of ANN algorithms. For the problem's resolution, a set of clinical data from real cases (961 instances / exams) are to be applied in order to determine whether there are benign / malignant masses. In other words, each group must implement mechanisms that enable the preparation of the proposed dataset, followed by the training, validation and optimization of the learning models. In this process, groups must consider the model's architecture (number of layers, number of nodes per layer, activation functions, classification threshold, etc.) to maximize the ANN classification performance. For model evaluation, evaluation methodologies and metrics must be proposed and applied, based on its relevance for the problem's resolution. Furthermore, the application of Genetic Algorithms for automatic optimization of the ANN architecture must be applied, as shown in Figure 1: Automatic Optimization Pipeline. Other optimization techniques can also be applied for benchmark analysis.

Based on the problem's statement, a brief summary of the dataset features is shown:

- Number of instances: 961;
- Number of attributes: 6 (1 target, 5 features);
- Features information:

1) BI-RADS Assessment: 1-5 (ordinal);
2) Age: Patient's age accounted in years (integer);
3) Mass shape: round=1, oval=2, lobular=3, irregular=4 (nominal);
4) Mass margin: circumscribed=1, micro-lobed=2, obscured=3, ill-defined=4, spiked=5 (nominal);
5) Mass density: high=1, medium=2, low=3, presents fat=4 (ordinal);
6) Severity: benign=0, malignant =1 (binominal);

- Distribution per class: benign: 516; malignant: 445;

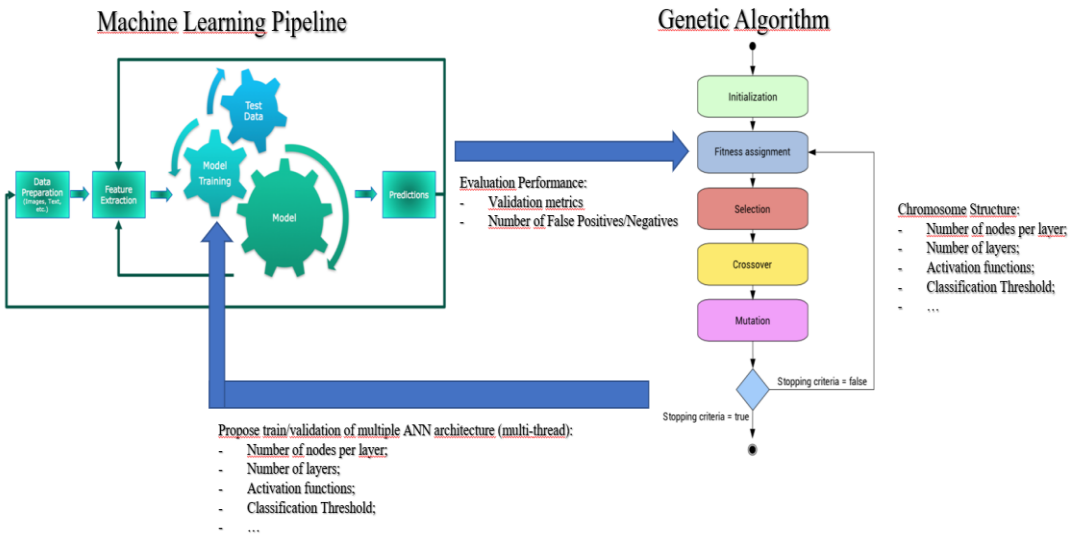

Figure 1: Automatic Optimization Pipeline

The practical work includes the delivery of the developed code and corresponding digital report, describing all the procedures applied and respective justification for its use, based on the demonstration of the results obtained.

Any open-source python library can be applied for the resolution of the exercise. However, all groups are advised to re-use code developed during the course. In addition, consider the application of mechanisms for the mitigation of standard challenges related to machine learning models (i.e., mitigation of overfit/underfit by studying the dataset distribution, application of k-fold cross-validation (k=10), and classification validation based on evaluation metrics and confusion matrix analysis, i.e., analysis of false positives and false negatives classifications). Take also into account the computational resources and time required for each optimization mechanism.

Delivery

The code resulting from this practical work and respective report (in digital format - PDF) must be sent via elearning platform to the respective evaluation item (found in: **Conteúdo – Instrumentos de Avaliação – Practical Work nº 1**), in compressed file (ZIP format), until **14th of April 2020**. The file must be identified in the form "[CN: F1GXX]", in which [GXX] designates the identification number of the group.

The presentation session of the practical work will take place in a format and date to be announced in due course.

Bibliographic References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... & Kudlur, M. (2016). Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)* (pp. 265-283).

Valarmathi, P., & Robinson, S. (2016). An improved neural network for mammogram classification using genetic optimization. *Journal of Medical Imaging and Health Informatics*, *6*(7), 1631-1635.