COMPUTATIONAL INTELLIGENCE (CI-MAI) - 2016-2017 Project proposal: Modeling Liver Disorders and Yeast data using neurofuzzy and neural network models

Àngela Nebot, Lluís A. Belanche

Abstract

Liver disorders and Yeast data sets are two benchmarks of the UCI machine learning repository. The aim of this project is to develop at least two classification models to solve each problem. One should be a neurofuzy model (an ANFIS) and the other one a neural network model (a MLP). The project should contain an evaluation and discussion section where you assess possible differences between both techniques and discuss their advantages and disadvantages in the context of these modeling problems.

1 Introduction

The two data sets can be found in the following URLs:

`https://archive.ics.uci.edu/ml/datasets/Liver+Disorders`
`https://archive.ics.uci.edu/ml/datasets/Yeast`

The Yeast data set contains 1,484 instances and 9 atributes. The goal is to predict Predicting the cellular localization sites of certain proteins (this is a multiclass classification problem).

The Liver disorders data set contains 345 instances and 7 atributes. The goal is to predict whether some patient has (or has not) a liver problem (this is a binary classification problem).

2 A look at the data

Yeast data set:

The 8 input atributs, all real values, are the following:

1. Sequence Name: Accession number for the SWISS-PROT database
2. mcg: McGeoch's method for signal sequence recognition.
3. gvh: von Heijne's method for signal sequence recognition.
4. alm: Score of the ALOM membrane spanning region prediction program.
5. mit: Score of discriminant analysis of the amino acid content of the N-terminal region (20 residues long) of mitochondrial and non-mitochondrial proteins.
6. erl: Presence of "HDEL"substring (thought to act as a signal for retention in the endoplasmic reticulum lumen). Binary attribute.

7. pox: Peroxisomal targeting signal in the C-terminus.
8. vac: Score of discriminant analysis of the amino acid content of vacuolar and extracellular proteins.

Liver disorders data set:

The 6 input atributs, integer and real values, are the following:

1. mcv: mean corpuscular volume
2. alkphos: alkaline phosphotase
3. sgpt: alamine aminotransferase
4. sgot: aspartate aminotransferase
5. gammagt: gamma-glutamyl transpeptidase
6. drinks: number of half-pint equivalents of alcoholic beverages drunk per day
7. selector: field used to split data into two sets