UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

COS 314: Artificial Intelligence

Assignment 2: Improving ID3 by using a genetic algorithm for feature selection

Due Date: 18 April 2018

On of the most fundamental issues with the decision trees constructed using the ID3 algorithm is their tendency to overfit. The issue of overfitting can be particularly prevalent when the dimensionally of your data patterns becomes very large. Where each dimension correspond to a feature of a data pattern. The issue can be overcome buy only allowing ID3 to see a subset of the features.

The real question then becomes, how do we select this subset? There is clearly no analytic solution to this selection that applies to all problem, as such we have found ourself with an optimization problem to solve. Specifically, we want to find the set of features which allows ID3 to obtain the best generalization ability. Your task for this assignment is to utilize a genetic algorithm (GA) to find the best subset on the given data set.

You have been given three files. Namely, the training set file, the validation set file, and the test set file. Recall that the test set is only to be used to evaluate the performance of your final model, and it should in no way be used to find/build the model itself. The data patterns are each a hundred features long, with each feture having the value of 0 or 1. The class membership of each pattern is indicated with a *True* or *False* at the end of the line

The program and the report must be submitted via the course website. The program must be executable and be able to run without linking to libraries via the IDE (in the case of C++). Please note the programs will not be run in IDEs but as a piece of commercial software.

The report must include:

- A description of the your GA chromosome.
- A description of the GA operators and stopping condition used and why.
- The progression of the optimization process during the run (how your generalization accuracy changed over generations).
- The use of a confusion matrix to illustrate the effectiveness of the final decision tree versus a decision tree that uses all the features.

Total:45