Canonical linear GP for classification

CSCI6506 Sandbox 1

Assignment due date: 20th January, 2015

Task definition

- Your goal is to develop an implementation for canonical GP for multi-class classification using a linear representation.
- It is recommended that you first read the paper by [1].[1]
- Your instruction set should consist of the FOUR two argument arithmetic operators: $R[x] = R[x] < op > R[y]$ where $< op > \in \{+, -, /, \times\}$
  - Note that you will need to consider how to protect operators or trap exceptions. See for example the discussion by [3].
- You should identify appropriate variation operators and support the following two types of selection operator:

  1. steady state tournament;
  2. proportional section.

  The deployment of the selection operator is parameterized such that an entire run is performed with a single case of the operator. One of your goals will be to discover which selection operator is more effective.
- In the case of the steady state tournament this has the form:

  1. Choose FOUR individuals with uniform probability from the population (these are the members of the tournament).
  2. Evaluate the fitness of each individual participating in the tournament.
  3. Apply the variation operators to the best TWO individuals from the tournament.
  4. Replace the worst TWO individuals from the tournament with the children from step 3 (thus updating the population).

[1] A more detailed account is available in [2], however, this is only recommended as additional information.

Table 1: Source data for sandbox 1

Name	Data Source
Iris	`http://archive.ics.uci.edu/ml/datasets/Iris`
Tic-tac-toe endgame	`http://archive.ics.uci.edu/ml/datasets/Tic-Tac-Toe+Endgame`

- You will also need to specify
  - the form assumed for your fitness function.
  - stop criterion.
  - the framework used to initialize your population of candidate programs.
- You will use two tasks for benchmarking your canonical GP classifier, specified as per Table 1. In each case 20% of the data should be reserved for test (i.e., not used for training your model) with exemplars selected to reflect the class distribution of the training partition.

  For example, if exemplars in a 4 class problem appear at a frequency of 50%, 25%, 10%, 15% across the dataset, then you want to partition your data into independent training and test partitions such that this distribution is maintained. Note that an exemplar can only appear in one of the partitions, *not* both!

Reporting

- Write a short 2 page summary answering the following questions:
  - the approach you adopted for designing the fitness function;
  - identify the benchmarking practice you assumed to identify which selection operator you preferred;
  - identify your best classifier, post training, on the *training partition* and then report the percentage of each class this classifier identifies under the *test partition*. Why is such a policy assumed (as opposed to identifying the best model from the test partition)?
  - be sure to identify all parameter settings that you assumed in your experiments and make a case for the variation operators you decided to support.
- Your report should be emailed to `mheywood@cs.dal.ca` before midnight on the sandbox due date.

References

[1] M. Brameier and W. Banzhaf. A comparison of linear genetic programming and neural networks in medical data mining. *IEEE Transactions on Evolutionary Computation*, 5(1):17–26, 2001.

[2] M. Brameier and W. Banzhaf. *Linear genetic programming*. Springer, 2007. `http://link.springer.com/book/10.1007%2F978-0-387-31030-5`.

[3] J. Ni, R. H. Drieberg, and P. I. Rocket. The use of an analytic quotient operator in genetic programming. *IEEE Transactions on Evolutionary Computation*, 17(1):146–151, 2013.