Comments on "evtree: Evolutionary Learning of Globally Optimal Classification and Regression Trees in R"

3/14/2012

1 Overall Thoughts

The paper did a good job of introducing a new tree-based algorithm overall. A lot of details pertaining to traditional tree models and the concept of evolutionary algorithms were covered, some of which could be clarified further. A clean-up in notation and wording would be greatly beneficial. The code appears easy-to-read and uses parameters consistent with other tree packages.

In the end, the authors gave honest comparisons of their methods to the traditional `rpart` and `ctree` methods, but it is unclear as to when and why `evtree` outperforms its competitors apart from the specific chessboard example given in Section 5.2.

I would recommend the paper for publication, provided that the issues in the remainder of this document are addressed.

2 Spelling/Grammatical Errors

1. (p.3) Above equation (2), "predicition" should be "prediction".
2. (p.7) First paragraph, midway, "spit points" should be "split points"
3. (p.8) Minor split rule mutation paragraph: "less then" should be "less than", and "the split point is change to to..." should be "the split point is changed to".
4. (p.8) Crossover paragraph, last sentence: "nods" should be "nodes".
5. (p.8) Evaluation function first paragraph: should be "The evaluation function represents the requirements to which the population should adapt."
6. (p.14) Middle of first pararaph: "bookclub" should be "book club".

3 Notation

In Section 2, the authors introduce some notation for the tree setup, but the notation seems inconsistent in the paper. The authors write $\theta = \left(v_{n_1}, s_{n_1}, ..., v_{n_{M-1}}, s_{n_{M-1}}\right)$, where $n_r \in \{1, ..., M_{max} - 1\}$ but then also write v_r and s_r where $r = 1, ..., M - 1$. Further, in Section 2.2, the authors talk about s_r and v_r where $r \in \{n_1, ..., n_{M-1}\}$, which is immediately different from the indexing given in equation (3). The authors should use a consistent notation for indexing the v and s.

4 Case Study

The authors have provided very minimal information on the dataset. (p.14) The authors write "Thus, the evolutionary tree uses a different cutoff in `art` for book club members that joined in the last year as opposed to older customers." The text does not explain which variable corresponds to "join date", but based on Figure 3, one can indirectly infer that this variable is `first`. The authors have described te variable `first` as the number of months since the first purchase, but do not say that members join the book club upon making their first purchase.

The table comparing the misclassification and evaluation function values for the different methods identify the methods as evtree, rpart, ctree, rpart2, and ctree2. The paragraph which describes this table instead identifies the methods as ev, rp, ct, rp2, and ct2. It would be better to make sure the two are consistent.

5 Miscellaneous

In addition to the above, I have the following suggestions.

1. (p.7) Section 3.1, "the assignment of one category is flipped to the other terminal node" might be adjusted to "one of the c categories is allocated to the other terminal node, to have the effect of ensuring both terminal nodes are nonempty".

2. (p.7) Split operator paragraphs use r again in 2 different ways: "internal node r" vs. "internal node at position r". As with the remainder of the paper, r is the internal node, not the position.

3. (p.7) Prune paragraph: We might append to the end of the sentence for clarity: "...and no pruning occurs."

4. (p.8) Minor split rule mutation paragraph: The description of the mutation rules is rather complex, with really 4 different cases to consider (nominal/<20 values, nominal/≥ 20 values, continuous/<20 values, continuous/≥ 20 values). It might be good to present this set of rules in a tabular format.

5. (p.8) Crossover paragraph, last sentence: "these are omitted" might be better written as "they are omitted" to be clear that "these" refers to the invalid nodes.

6. (p.8) Evaluation function paragraph: "A suitable evaluation function... minimizes the models' accuracy...and the models' complexity." It appears that "minimizes" should be "optimizes".

7. (p.8) Classification paragraph: "misclassification MC" should be "misclassification rate MC".

8. (p.9) Survivor selection paragraph, last sentence on the page: This sentence is worded awkwardly. Perhaps the authors meant "In the case of a mutation operator, either the solution before modification, θ_i, or after modification, θ_{i+1}, is kept in memory." Without the commas, the sentence is unclear.

9. (p.10) Third paragraph: Ambiguous "they" in "However, as they represent one of the best..., they are both...". Does "they" refer to the parent and offspring trees? Does this mean that despite the (1+1) rule, both the parent and offspring are both selected to the next generation? This seems unclear.

10. (p.17) First paragraph: The authors talked about the demanding memory requirements, for which “400 Mbit” seems very low. Perhaps the authors meant “400 megabytes”?