Reviewer’s report

Title: PMLB: A large benchmark suite for machine learning evaluation and comparison

Version: 0 Date: 23 Apr 2017

Reviewer: Kate Smith-Miles

Reviewer’s report:

This paper describes the process of collecting, pre-processing and performing some basic analysis on a set of 165 real-world, simulated and toy classification problems (binary and multiclass), mostly obtained from UCI and KEEL (it is never made clear if there are any new datasets that have been created, or it is just pre-processing existing datasets). The stated aims are to facilitate the identification of strengths and weaknesses of ML algorithms, and to identify the diversity of the current datasets. The work is described as "an important first step".

Unfortunately, this is not the first step - there are already a number of papers that have tackled this topic with the same aims, and have developed and analysed dataset repositories for ML. I am quite sure that the authors are aware of these efforts, and it is puzzling that the paper makes no reference to them and claims to be taking the first step. Examples of such relevant papers include:


Most of the literature review in this paper is inadequate, even reference [15] for an argument that new benchmarks are needed, when there are much stronger references available.

The analysis of the 165 datasets is very simple, and the meta-features that have been chosen are really only related to the simple statistical features such as size and class balance. There is 25
years of research on meta-learning that has developed a much richer set of meta-features than these simple few, and it is not really useful to cluster 165 datasets based on such simple features. For example, it is stated that datasets 22, 118 and 164 are hard due to the lack of univariate correlations between features and classes and the high amount of noise, and yet these meta-features are not used to create the clusters. So there is little attempt to gain insights into the connections between dataset complexity and algorithm performance.

Overall the paper reflects some efforts to pre-process some datasets, but fails to deliver on its stated objectives. It leaves the reader feeling disappointed since we expected to see a contribution where a genuinely large collection (way more than 165) of provably diverse datasets have been created to augment the current repositories, with some great new insights into how those new datasets help us understand strengths and weaknesses of ML algorithms. The paper falls very short of this promise, and really offers no new conclusions beyond the existing literature.

**Level of interest**
Please indicate how interesting you found the manuscript:

An article of limited interest

**Quality of written English**
Please indicate the quality of language in the manuscript:

Acceptable

**Declaration of competing interests**
Please complete a declaration of competing interests, considering the following questions:

1. Have you in the past five years received reimbursements, fees, funding, or salary from an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?
2. Do you hold any stocks or shares in an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?
3. Do you hold or are you currently applying for any patents relating to the content of the manuscript?
4. Have you received reimbursements, fees, funding, or salary from an organization that holds or has applied for patents relating to the content of the manuscript?
5. Do you have any other financial competing interests?
6. Do you have any non-financial competing interests in relation to this paper?
If you can answer no to all of the above, write 'I declare that I have no competing interests' below. If your reply is yes to any, please give details below.

I declare that I have no competing interests

I agree to the open peer review policy of the journal. I understand that my name will be included on my report to the authors and, if the manuscript is accepted for publication, my named report including any attachments I upload will be posted on the website along with the authors' responses. I agree for my report to be made available under an Open Access Creative Commons CC-BY license (http://creativecommons.org/licenses/by/4.0/). I understand that any comments which I do not wish to be included in my named report can be included as confidential comments to the editors, which will not be published.

I agree to the open peer review policy of the journal

Do you want to get recognition for reviewing this manuscript? Add a record of this review to Publons to track and showcase your reviewing expertise across the world's journals. Signing up is quick, easy and free!

No