Doctoral Thesis Assessment: Computational Intelligence Methods in Metalearning, by Jakub Šmíd

Joaquin Vanschoren

SYNOPSIS

This thesis addresses the problem of calculating attribute-level characteristics in datasets, using them to define distance metrics between those datasets and leveraging these distances for metalearning, and more specifically, for algorithm and hyperparameter recommendation. The main contribution lies in proposing 'attribute assignment' algorithms that align the attributes of two datasets, so that both datasets can be expressed as a point in the space spanned by these attributes, and a distance between them can be measured. This is founded theoretically by proving that the resulting distance is a metric, and evaluated empirically by using this distance to recommend classification algorithms on a large set of classification problems from OpenML. Other work explores the use of genetic programming to generate new distance functions, even though these currently do not yet outperform the previous algorithms.

NOVELTY OF SCIENTIFIC RESULTS

The problem of handling attribute-level characteristics (meta-features) is a well-known and important issue in meta-learning. This work proposes a completely new and well-founded approach to solve this. Prior work tried to aggregate attribute-level meta-features into global meta-features, e.g. histograms of attribute properties [64]. This work is, as far as I know, the only work that actually defines a new distance function for the attribute-level properties. This, in turn, can be combined with previously defined distance functions. The candidate is also to be commended for proving that this distance is actually a metric and evaluating it on a very large set of datasets and meta-features from OpenML. The work on using genetic programming to define new metric is also certainly novel. While it does not yet yield state-of-the-art results, it is a possibly fruitful avenue for future research.

IMPORTANCE FOR THE AREA

Including attribute-level information is an important open problem in meta-learning, since a lot of information is lost when aggregating this information for use in current (mainly propositional) algorithm selection techniques. Hence, the proposed attribute-based distance can be used to improve metalearning and algorithm selection systems. While further research is needed, for instance the inclusion of landmarkers, the results presented in this thesis look promising. Also the thorough evaluation of the technique on OpenML meta-data (a dump of which is provided by the authors) sets a good example for future studies, and allows the community to easily compare with and build on the new technique. I would also recommend to include the attribute level metafeatures as an extension to OpenML, so that others can easily reuse them.

APPLICATIONS TO OTHER AREAS

Given the ever growing importance of machine learning in all aspects of society, and the lack of trained experts, being able to accurately recommend machine learning algorithms is a very important problem. Indeed, it will allow more people to use machine learning more effectively in science, industry and society, using better algorithms or simply saving time to arrive at those algorithms. In addition, having a trustworthy distance metric between datasets is also very useful: it can help scientists, for instance, to find similar datasets created by other scientists and reuse the analysis techniques used there.

PRESENTATION AND FORM

Overall, the thesis is easy to read. There are consistent grammatical mistakes in the use of articles, singular/plural, and sometimes conjugations, but overall the meaning of the text is clear. The text is also quite well-structured, with only few parts that seem out of place (e.g. the discussion of active testing on p. 24 would fit much better under section 2.9, and section 2.9 could be structured better). Some figures are missing proper labelling, e.g. Figure 6.5 is very hard to interpret, but overall everything is well illustrated. The definitions and proofs are also presented carefully, with a few exceptions (e.g. Definition 2 is quite imprecise).

ABILITY FOR CREATIVE SCIENTIFIC WORK
The thesis clearly presents creative scientific work. As far as I know, it is very original, executed independently, and the results seem promising.

OPEN QUESTIONS

* You use a novel distance measure to be able to include attribute-level information, and thus sidestep the issue of propositional meta-learners such as kNN. What about using your attribute-level descriptions together with a non-propositional meta-learner, such as a graph-based learner (e.g. using a graph kernel)?
* You present multi-objective evaluation function such as A3R, but only evaluate your method using rank correlations. Have you tried evaluating your method with A3R as well?
* You never compare with state-of-the-art recommendation techniques such as AutoWEKA or active testing, but only use the global ranking as a baseline. Why is this? Would an AutoWEKA run be much slower that the attribute alignment? It is not clear how fast this alignment actually is. Would it make sense to include your new distance measure in techniques such as AutoWEKA?
* You exclude landmarkers from your study, but it is not clear to me why. I understand that you want to compare your measure independently as well, but it would be very good to compare your method with a standard meta-learning system using landmarkers, since they carry so much intrinsic information. You want to prove that your distance carries ADDITIONAL information not captured by the landmarkers. Also, you write that you exclude them because of concerns that they include information on the test set. I don't understand this, because on the meta-level, you would simply compute the landmarkers on the new test datasets, without knowing what the best algorithm is (it never accessed that meta-information).
* There are other existing techniques to align sequences, such as Dynamic Time Warping or using dynamic programming, which could be used to align attributes, too. Did you consider these? I know that these as meant for sequences that are ordered, but you also impose an order by sorting attributes prior to alignment in your first method. Hence, would it not make sense to use existing sequence alignment techniques here, without requiring dummy attributes?
* Kalousis [64] recommended the use of histograms to capture more information about the attribute distributions. Did you ever compare against that approach?

Joaquin Vanschoren