Thesis Review

Cooperative Game Theory for Machine Learning Tasks

Master Thesis by Bc. Jan Pecka, 2020

It has been a pleasure to read the thesis by Bc. Jan Pecka. The addressed field is close to my area of expertise which includes feature selection and recently the local explainability of neural models. Both fields are related to each other, as well as to coalitional game theory, as studied in detail in the thesis.

The thesis addresses the problem of explaining verdicts of black-box predictive models using techniques based on and derived from the concept of Shapley values. Shapley values themselves are considered the standard tool for explanation in recent years. However, I have not seen yet as thorough formal study of the Shapley values concept and related concepts, as provided in this thesis. The first chapter provides formal introduction into the theory of coalitional games and covers all foundations needed to build up the case for further study of related explanation techniques. The next chapter provides exceptionally well covered axiomatic apparatus leading to Shapley, Banzhaf and related types of values. Computational aspects are covered well too, including the concept of Adaptive Shapley value sampling. Banzhaf values are normally not considered in literature as suitable for explanation. Chapter 3 provides experimental comparison of Shapley and Banzhaf and provides a case for further Banzhaf study in explanation context. The last chapter illustrates the covered concepts on a real world examply study on brain activity mapping data. The choice of problem makes sense and provides interesting insight into the applicability of explanation techniques.

The text of the thesis is throughout excellently written, with unusually strong command of formally correct mathematical language. At the same time the text – despite its complex subject – is easy to read. In this sense the thesis is exceptional, I have not seen too many Master theses written that well. Singular mis-formulations can be found but do not harm the overall impression. Another property of the text worth praise is the completeness of the presented apparatus and related discussions. Clearly the author strived to cover all possible formulations, interpretations and consequences of each presented concept. This in itself would be enough for the thesis to succeed, as it provides a unifying view of the subject that is original and helpful. I particularly enjoyed some interesting points being made, e.g.,: the explanation of the relation of Shapley to local approximation through linear regression, the description of adaptive sampling based estimation of Shapley values, or even the excellent technical discussion of hyper-parameter estimation through Bayesian processes for the purpose of the final experiment.

There were only two moments where I felt the thesis lacking. The outcome of Chapter 3 feels a bit half-way. It justifies further study of Banzhaf by showing its very similar performance to Shapley. It would be nice to get a bit more insight into what is the core of applicability difference between Shapley and Banzhaf. But I agree with the author that this is a complex subject beyond the scope of this thesis. The second questionable point is a bit more notable. The construction of training data set that is at the core of Chapter 4 brain activity analysis, appears slightly flawed. Concatenating measurements from 84 patients into one time series inevitably must introduce at least 83 points in which the time series information can not be fully correct (potentially spreading the error to 83*k time windows if size of window is k), unless some additional assumption is adopted about the pool of patients. This potential problem is completely ignored in the following. I admit that it most likely does not invalidate the results later achieved (to be more precise, the effect of the injected error most likely remains very small). It is just a bit surprising that this applicational concern is

overlooked by the author when everything else in the thesis is studied with unusual attention to detail.

To summarize, the exceptional qualities of the thesis significantly over-weigh its flaws, hence I fully recommend the thesis to be accepted as Master thesis. Having hasitated a bit between suggesting mark A or B (due to the treatment of data in Chapter 4), I finally lean towards **A** to praise the exceptional formal qualities of the core body of the text.

In Prague, 16 July 2020
RNDr. Petr Somol, Ph.D.