REVIEW OF MASTER THESIS

Name of the student: Jakub Malý
Thesis title: Automatic event recognition for Higgs boson detection
Name of the reviewer: Boris Flach
Institution: Czech Technical University in Prague, Faculty of Electrical Engineering

1. Results of the work and thesis structure

The master thesis presented by Jakub Malý aims at testing and analysing different machine learning approaches for classifying particle collision events obtained in the ATLAS detector of the CERN large hadron collider. In particular, it focuses on detecting events that may produce Higgs bosons. The prime motivation of the thesis is to analyse standard machine learning approaches and their suitability for predicting these rare events with high precision.

After a short introduction, the author gives an overview of the experimental setup and some relevant background of elementary particle physics. The next two chapters describe the structure of the available data and recall the basics of statistical pattern recognition. Chapter 6 describes the considered machine learning approaches, applies them on the data and validates their results. The spectrum of analysed approaches ranges from simple methods like k-neighbors classifiers over AdaBoost to more complex approaches like random forests and neural networks. The simulated data used for training and validation of the learning approaches were obtained from CERN. The last two chapters describe changes and adaptations that where necessary for applying the methods on real data obtained by the author from CERN in the last period of his work.

The thesis concludes with a summary of the authors findings and a personal comment on the impact of the pandemic restrictions on his work.

2. Comments

The thesis is written in a linguistically competent way. Its overall structure is appropriate. It reveals, however, weaknesses on a finer level that make it difficult to read. The main reason is that author has scattered concept details across several chapters. For instance, I would have expected to find all important facts about the data and their structure as well as all details about the chosen pre-processing in Chapter 4. (Data). Instead, relevant pieces are scattered in chapters 4-6 with partially wrong reference links.

The choice of the loss function/validation criterion is an important, application dependent design option for predictors and machine learning approaches. The author has chosen to follow CERN recommendations and proposes to use a significance score. Unfortunately, the thesis part explaining this loss is not well written and details remain unclear as a consequence. Again, I would have expected to find all related facts in Chapter 6. (Classification) and not scattered over several chapters.

Almost all machine learning approaches considered by the author are discriminative methods, i.e. methods that either learn a predictor by empirical risk minimisation or learn predictive posterior class probabilities in some model class. It remains unclear for me, why these approaches require class weights in case of unbalanced training sets, as long as the prior class probabilities do not change.

The list of references is appropriate. However, some references have missing bibliographical details.

3. DEFENSE QUESTIONS

Q1: Give a concise explanation of the significance score loss proposed in your thesis.

Q2: A predictor has been trained for classifying patterns by predicting the posterior class probabilities $p(y \mid x)$. Let us assume the 0/1 loss. Consider the situation that the class frequencies in the training set differ from the true prior class probabilities (at inference time). Explain how to use the predictor without re-training it.

Q3: Give possible options for losses that can be used in situations where we want detect rare events of some class on the background of possibly several other classes.

4. CONCLUSIONS

The thesis reflects a substantial amount of work performed by the author. Despite the weaknesses mentioned above, it fulfills the criteria of a graduation thesis. Moreover, I can imagine that working with data from a large research entity (as CERN), may require extra efforts due to possibly rigid regulations and procedures. Therefore I recommend to accept the thesis for the defense and grade it with 'C' (good).

Dresden, 15.06.2020 Dr.rer.nat.habil. Boris Flach