I see what you did there: Understanding when to trust a ML model with NOVA

Tobias Baur, Alexander Heimerl, Florian Lingenfelser, Elisabeth André
Human Centered Multimedia Lab
Augsburg University Germany
Augsburg, Germany
{baur,heimerl,lingenfelser,andre}@hcm-lab.de

***Abstract*—In this demo paper we present NOVA, a machine learning and explanation interface that focuses on the automated analysis of social interactions. NOVA combines Cooperative Machine Learning (CML) and explainable AI (XAI) methods to reduce manual labelling efforts while simultaneously generating an intuitive understanding of the learning process of a classification system. Therefore, NOVA features a semi-automated labelling process in which users are provided with immediate visual feedback on the predictions, which gives insights into the strengths and weaknesses of the underlying classification system. Following an interactive and exploratory workflow, the performance of the model can be improved by manual revision of the predictions.**

***Index Terms*—annotation tools, cooperative machine learning, explainable AI**

I. Introduction

In various research disciplines the annotation of social behaviours is a common task. This process includes manually identifying relevant behaviour patterns in audio-visual material and assigning descriptive labels. Generally speaking, segments in the signals are labelled using sets of discrete classes or continuous scores, e. g., a certain type of gesture, a social situation, or the emotional state of a person. To automatically detect social signals from raw sensory input it is common practice to apply machine learning (ML) techniques. However, the performance of a ML-System is largely dependent on the amount and quality of the annotated training data. Especially in the field of social signal processing, annotating enough data can be a lengthy and cumbersome task.

An obvious solution to this problem is exploitation of computational power to accomplish some of the annotation work automatically. To ensure the quality of the predicted annotations this still requires human supervision to identify and correct errors. To keep the human effort as low as possible, it is useful to understand why a model makes wrong assumptions. Therefore, it is not only important to provide tools that ease the use of semi-automated labelling, but also to increase the transparency of the decision process. By visualising the predictions with model-agnostic explainable AI methods, even non ML experts get an idea about the strengths and weaknesses of the underlying classification model and can immediately decide which parts of a prediction are worth keeping. Ideally, the system even guides the user's attention towards parts where manual revision is necessary. Once an annotation has been revised, the model can be retrained to improve its performance for the next cycle. This procedure can be repeated until a desired performance is reached.

This work has received funding from the BMBF under FKZ 01IS17074, FMLA, and from the DFG under project number 392401413, DEEP.

II. Cooperative Machine Learning

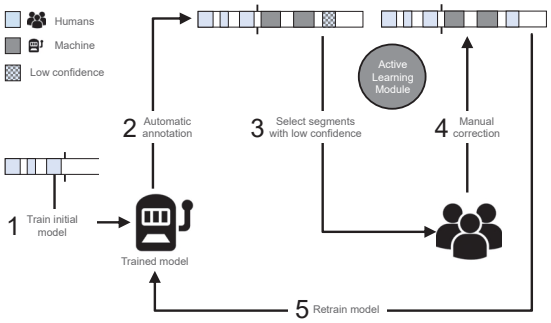

Fig. 1. The scheme depicts the general idea behind Cooperative Machine Learning (CML): (1) An initial model is trained on partially labelled data. (2) The initial model is used to automatically predict unseen data. (3) Labels with a low confidence are selected and (4) manually revised. (5) The initial model is retrained with the revised data.

In this work, we subsume learning approaches that efficiently combine human intelligence with the machine's ability of rapid computation under the term *Cooperative Machine Learning* (CML). In Figure 1, we illustrate our approach to CML: an initial model is trained (1) and used to predict unseen data (2). An active learning module then decides which parts of the prediction are subject to manual revision by human annotators (3+4). Afterwards, the initial model is retrained using the revised data (5). Now the procedure is repeated until all data is annotated. By actively incorporating the user into the loop it becomes possible to interactively guide and improve the automatic predictions while simultaneously obtaining an intuition for the functionality of the classifier.

However, the approach not only bears the potential to considerably cut down manual efforts, but also to come up with a better understanding of the capabilities of the classification system. For instance, the system may quickly learn to label

some simple behaviours, which already facilitates the work load for human annotators at an early stage. Then, over time, it could learn to cope with more complex social signals as well, until at some point it is able to finish the task in a completely automatic manner. To evaluate the efficiency of the integrated CML strategy, we performed a simulation study on an audio-related labelling task. Following this approach we were able to reduce the initial annotation labour by 37.23% [1]. We opt to make the described strategy an integral part of the NOVA tool which we will describe in more detail in the next section.

III. NOVA Tool

The NOVA user interface has been designed with a special focus on the annotation of continuous recordings involving multiple modalities and subjects. An instance of a session is shown in Figure 2.

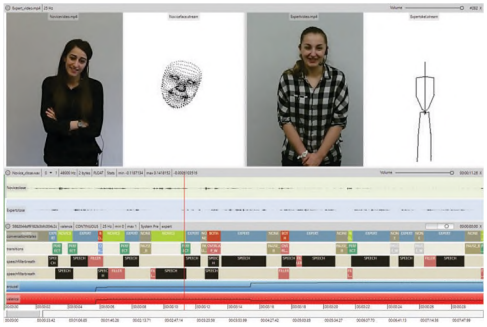

Fig. 2. NOVA allows to visualise various media and signal types and supports different annotation schemes.

On the top, several media tracks are visualised and ready for playback. Note that the number of tracks that can be displayed at the same time is not limited and various types of signals (video, audio, facial features, skeleton, depth images, etc.) are supported. In the lower part, we see multiple annotation tracks of different types (discrete, continuous and transcriptions) describing the visualised content.

To support a collaborative annotation process, NOVA maintains a database back-end, which allows users to load and save annotations from and to a MongoDB database. This gives annotators the possibility to immediately commit changes and follow the annotation progress of others. NOVA provides instruments to create and populate a database from scratch. New annotators, schemes and additional sessions can be added. NOVA provides several functions to process the annotations created by multiple human or machine annotators. For instance, statistical measures such as Cronbach's α or Cohen's κ can be applied to identify inter-rater agreement.

A typical ML pipeline starts by preprocessing data to input data for the learning algorithm, a step known as *feature extraction*. An XML template structure is used to define extraction chains for individual feature extraction components. Finally, a classifier, which may also be added using XML templates, can be trained. To automatically finish an annotation, the user either selects a previously trained model or temporarily builds one using the labels on the current tier.

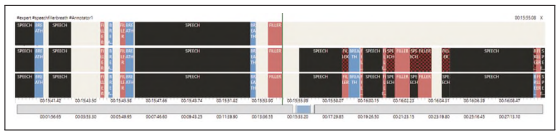

Fig. 3. The upper tier shows a partly finished annotation. ML is now used to predict the remaining part of the tier (middle), where segments with a low confidence are highlighted with a red pattern. The lower tier shows the final annotation after manual revision.

An example before and after the completion is shown in Figure 3. Note that labels with a low confidence are highlighted with a pattern. This way, the annotator can immediately identify the quality of the prediction. In case that the user wishes to gain additional insight on the classifier's prediction, NOVA provides the possibility to create visual explanations by incorporating the two explanation frameworks LIME [2] (see Figure 4) and iNNvestigate [3].

Fig. 4. Explanations for the top four classes in a facial emotion recognition task, generated in NOVA with the usage of LIME.

IV. Conclusion

Summing up, the described methodology offers transparency on multiple levels. By observing the output of the classifier, the user can assess its performance and trace how it changes with new input. In addition, visualising the relevant parts for the model's decision hints why a prediction was successful in one place but failed in another. For instance, the user may find out that predictions were wrong due to irrelevant features. This way, users also learn in which situations they can trust the model. We subsume this approach under the term *eXplainable Cooperative Machine Learning* (xCML). NOVA is open-source and available on Github: https://github.com/hcmlab/nova.

References

[1] J. Wagner, T. Baur, Y. Zhang, M. F. Valstar, B. Schuller, and E. André, "Applying cooperative machine learning to speed up the annotation of social signals in large multi-modal corpora," *arXiv preprint arXiv:1802.02565*, 2018.

[2] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should I trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, 2016, pp. 1135–1144.

[3] M. Alber, S. Lapuschkin, P. Seegerer, M. Hägele, K. T. Schütt, G. Montavon, W. Samek, K. Müller, S. Dähne, and P. Kindermans, "innvestigate neural networks!" *CoRR*, vol. abs/1808.04260, 2018.