The design and validation of an intuitive confidence measure

Jasper van der Waa
TNO
Soesterberg, the Netherlands
jasper.vanderwaa@tno.nl

Jurriaan van Diggelen
TNO
Soesterberg, the Netherlands
jurriaan.vandiggelen@tno.nl

Mark Neerincx
TNO
Soesterberg, the Netherlands
mark.neerincx@tno.nl

ABSTRACT

Explainable AI becomes increasingly important as the use of intelligent systems becomes more widespread in high-risk domains. In these domains it is important that the user knows to which degree the system's decisions can be trusted. To facilitate this, we present the Intuitive Confidence Measure (ICM): A lazy learning meta-model that can predict how likely a given decision is correct. ICM is intended to be easy to understand which we validated in an experiment. We compared ICM with two different methods of computing confidence measures: The numerical output of the model and an actively learned meta-model. The validation was performed using a smart assistant for maritime professionals. Results show that ICM is easier to understand but that each user is unique in its desires for explanations. This user studies with domain experts shows what users need in their explanations and that personalization is crucial.

ACM Classification Keywords

I.2.M Artificial Intelligence: Miscellaneous; I.2.1 Artificial Intelligence: Applications and Expert Systems; I.2.M Artificial Intelligence: Miscellaneous

Author Keywords

Explainability, Machine Learning, lazy learning, instance based, ICM, experiment, user, validation, confidence, measure, certainty

INTRODUCTION

The number of intelligent systems is increasing rapidly due to recent developments in Artificial Intelligence (AI) and Machine Learning (ML). The applications of intelligent systems begin to spread to high-risk domains, for example in medical diagnoses [3], maritime automation [18] and cybersecurity [6]. The need for transparancy and explanations towards end users is becoming a neccessity [8, 4]. This self-explaining capability of intelligent systems allow these to become more effective tools that allow their users to establish an appriopriate level of trust. The field of Explainable Artificial Intelligence (XAI) aims to develop and validate methods for this capacity.

The process of explaining something consists of a minimum of two actors: explainer and the explainee [12]. A large number of studies in XAI focus on the system as the explainer and how it can generate explanations. For example in methods that focus on identifying feature importance [11, 15], those that extract a confidence measure [7], those that search for an informative prototypical feature set [10] or explain action policies in reinforcement learning [9]. Although these are effective approaches to generate explanations, they do not validate their methods with the explainee. A working XAI methods needs to incorporate the user's wishes, context and requirements [5, 13, 1]. As XAI tries to make ML models more transparant, a requirement for XAI methods is to be transparant themselves so the user can understand where the explanation comes from.

The proposed Intuitive Confidence Measure (ICM), is a case-based machine learning model that can predict how likely a given model output is correct in (semi-)supervised learning tasks. ICM is a meta-model that is stacked on top and independent of its underlying ML model. The intuitive idea behind ICM is that it uses past situations and any incorrect or correct outputs in those situations to compute the probability that a given output in some situation is correct. A high confidence is given when the current situation and output is similar to situations in which that output proved to be correct. Since ICM is a case-based or lazy-learning algorithm it allows each outputted confidence to be traced back to items in a data set or memory [2]. For example, the confidence in some output is low because this output is similar to past outputs that proved to be incorrect that were given in very similar situations. This is opposed to a confidence measure that uses active learning where a (possibly large) set of parameters describe learned knowledge that are difficult to explain or understand [8].

Other approaches to estimate a confidence value exist. Several machine learning model types can already provide a probabilistic output such as neural networks with soft-max output layers. However, these confidence estimations can be inaccurate as these models can learn to be very confident in an incorrect output, as a trade-off for general improvement on the overall dataset [14]. Other approaches may not prove to be model agnostic. For example the usage of dropout in neural networks [7].

To test if ICM is indeed easy to understand, we performed an experiment where we compared ICM as a lazy learned meta-model to two different types of certainty or confidence measures: The numerical output of the underlying model itself and an actively learned meta-model approach. We claim that ICM is prefered over these two types because 1) the numerical output of the underlying model is not always available, transparant or accurate [14] and 2) an actively learned meta-model has no clear connection between its outputted confidence and used data [2]. ICM on the other hand is a meta-model and as such independent of the workings of the underlying model except for its outputs and ICM's confidence values are directly related to its training set due to lazy learning.

The experiment was performed within a maritime use case for computer-controlled propulsion, we refer to our earlier paper for a detailed description [18]. Participants had no knowledge about ML and worked in a high-risk maritime domain with extreme responsibilities. In our experiment we simulated the operator's working environment and presented the participant with classification outputs accomponied by a confidence value. Later we interviewed the operators about their experiences and presented them with the three measure types we identified earlier; 1) a numerical model output, 2) an actively learned meta-model and 3), our method, a lazily learned meta-model. We tested the participant's understanding of each of the measures to validate whether ICM, and lazy learned measures in general, are indeed easier to understand and as such prefered over numerical model outputs and actively learned meta-models.

The experiment showed that ICM is indeed easier to understand but each operator had various wishes of when, and even if, a confidence value should be presented and all overestimated their own understanding of complex ML methods. XAI experiments with expert users such as these offer valuable insights in what kind of explanations are required and when.

INTUITIVE CONFIDENCE MEASURE

ICM computes the probability that the given output is correct. It does this by weighing the difference of that output with the ground truths of a set of known past datapoints with the similarity of the current datapoint with those past data points. We visualized this in Figure 1 for a simple example where Euclidean distance can be used as the similarity measure. This figure illustrates the intutive idea that when a situation and output is similar to past situations in which different outputs proved to be correct, confidence will be low. The more similar situations there are with a different and correct output, the lower the confidence. If there are no similar situations, the confidence will be unknown or uniform, depending on the choice of presentation to the user. In the following paragraphs we only explain the vital technical details of ICM, we refer to earlier work for a more technical description and discussion of its advantages and disadvantages [17].

ICM is based on the following three equations, with x as an arbitrary data point, M an arbitrary data set, d the used similarity function, σ as the standard deviation used for the exponential weighting and $M(T = A(x))$ to select all data points in M with the same groundtruth T as the output of model A for x;

$$C(x|\sigma, M) = \frac{\sum_{x_i \in M(T=A(x))} exp\left(-\frac{d(x|x_i)^2}{2\sigma^2}\right)}{\sum_{x_i \in M} exp\left(-\frac{d(x|x_i)^2}{2\sigma^2}\right)} \qquad (1)$$

The memory or dataset M is sequentially sampled according to three aspects from the trainset or during actual usage of the system. This strategy prefers data points with 1) a ground truth least common in the memory, 2) datapoints that are some time apart to mitigate temporal dependencies and 3) datapoints that are relatively dissimilar to the datapoints inside the memory. We refer to the original paper of ICM for a detailed description [17]. This memory is restricted to a fixed size, k, to prevent extreme computational costs. The number of computations increases exponentially with each added data point and to store all data would quickly become unfeasibly for real world cases where the model A and ICM may run for indefinite time.

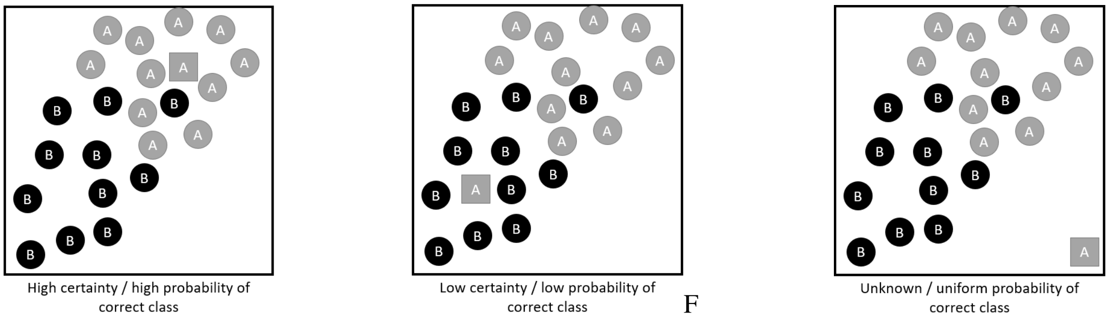
ICM has several properties in common with other lazy learning techniques such as k-Nearest Neighbours (k-NN). In specific ICM is very similar to the weighted k-NN algorithm with an exponential weighting scheme where the normalization garantuees that all weights sum to one. ICM becomes an instance of weighted k-NN for non-linear regression with the model's groundtruth as the dependent variable, the memory M to mitigate computation cost and an arbitrary distance function d.

EXPERIMENT

In a small experiment we compared the understanding of three instances of different types of confidence measures by end-users 1) ICM as a lazily learned meta-models, 2) the approach by Park et al. as actively learned meta-models [16] and 3) the soft-max output as a numerical output of the actual model. The experiment was done based on a recent study with a virtual smart assistant that supports an operator on a ship with situation predictions to aid in his/her monitoring task [18]. We simulated the operator's work environment and the virtual smart assistant and provided realistic scenarios and responses from the assistant including a confidence value for any made predictions. This simulation was used to introduce the participants with the assistant and the numerical confidence values it could provide.

This simulated work environment was followed by an interview during which they received increasingly more information about the three confidence measures. The goal of the interview was to test how well and how quickly the participant understood each of the three confidence measures. The interview went through several stages;

1. First stage
   (a) Brief textual explanations of each measure and opportunity for the participant to rate his/her understanding of the measure.
   (b) Per measurement a moment for the participant to ask questions to allow the supervisor to rate how well the participant understands the examples.


High certainty / high probability of correct class
Low certainty / low probability of correct class
F
Unknown / uniform probability of correct class

(a) (b) (c)

Figure 1: Figure that shows three examples of how ICM works in a 2D binary classification task (class A and B) given a current data point with its output (square) and a set of known data points (circles) with their known ground truths.

   (c) An explanation by the participant for each measure in their own words to rate by a ML expert.

2. Second stage
   (a) Three concrete examples, both visual and textual, for each measure to illustrate its mechanisms where the participant could rate his/her level of understanding for each set of examples.[1]
   (b) Per set of examples a moment to ask questions to the supervisor, to allow the supervisor to rate how well the participant understands the examples.
   (c) An explanation by the participant for each example in their own words to rate by a ML expert.
   (d) The participant's final preference for one of the three confidence measures and an explanation of a given confidence. A ML expert checked whether this explanation overlaps with one of the three measures.

Results

The results of the five participants are shown in table 2. All were experts and potential end-users in the maritime use case. The two users saw no use for a confidence measure believed that predictions should always be correct or otherwise not presented at all. All participants believed that they had some basic to advanced comprehension of each measure and its set of examples, however the experiment supervisor and ML expert disagreed with this for both the 'numerical' and 'active learning meta-model' measures.

Both the supervisor and the ML expert concluded that most participants had some degree of understanding for ICM. Only one participant was not able to comprehend the textual explanation but the understanding of ICM was on average rated higher than that of the 'numerical' and 'active learning meta-model' measures, by both the supervisor and ML expert.

The explanations about the numerical output were lacking because participants had trouble comprehending that a model could learn knowledge and represent it in parameters. They had less difficulty for ICM because its outputs related clearly to past situations. The explanations from the participants regarding the 'meta-model' measure were the most inaccurate. Nearly all participants had the tendency to see this measure as a combination of ICM and a probabilistic output. This was also the reason why three out five participants tended to prefer this measure in the end, even though their own explanations of the confidence values resembled the approach used by ICM.

CONCLUSION

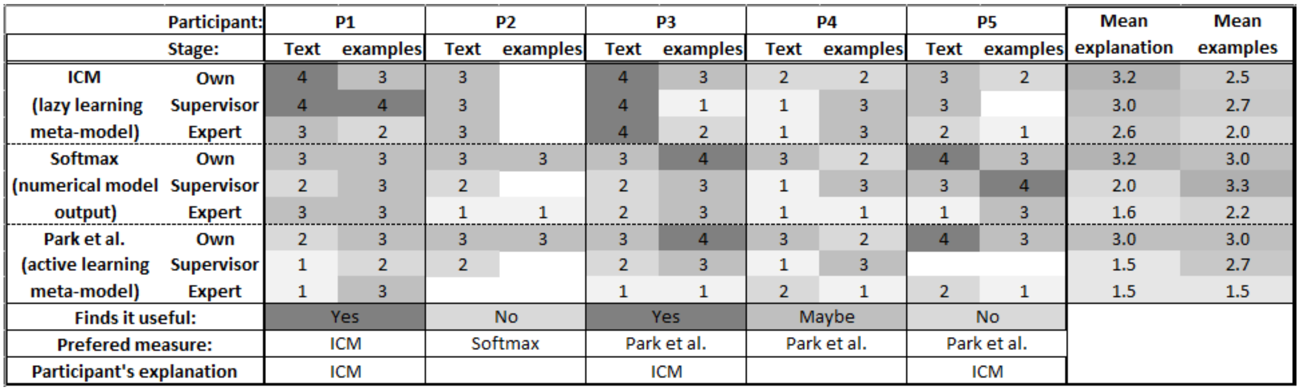
In the introduction we stated that XAI methods should not only be developed but also validated in experiments. We mentioned that XAI methods should be transparant by themselves such that the user can understand where the explanation comes from and why it is given.

The Intuitive Confidence Measure (ICM) was developed as a method to provide a confidence value alongside a machine learning model's output. It uses lazy-learning and intuitive ideas to keep the method relatively simple with clear connections between input and output. We performed a limited usability study with qualitative interviews. These interviews indicated that ICM is relatively simple to explain compared to other confidence measures based on model output (e.g. values from a softmax function) or values from meta-models based on active learning.

The results showed that in a group of similar end-users, there were mixed opinions about the necessity of a confidence measure and how it should be presented. However, most participants thought of ICM as an easy to understand measure and could recall the workings of ICM accurately. Most of the participants were even able to identify advantages and disadvantages of ICM in specific situations, showing a deeper understanding. Future work will focus on a larger study to test the intuitiveness of ICM, technical improvements to ICM to mitigate disadvantages and way on how to generate confidence explanations.

The development of new XAI methods for high-risk domains is important, but their validation in experiments with domain experts is equally important. Like the one presented in this paper, experiments and usability studies with domain experts can help shape the field of XAI.

[1] The textual descriptions and examples can be requested by e-mail.



	Participant:	P1		P2		P3		P4		P5		Mean	Mean
	Stage:	Text	examples	Text	examples	Text	examples	Text	examples	Text	examples	explanation	examples
ICM	Own	4	3	3		4	3	2	2	3	2	3.2	2.5
(lazy learning	Supervisor	4	4	3		4	1	1	3	3		3.0	2.7
meta-model)	Expert	3	2	3		4	2	1	3	2	1	2.6	2.0
Softmax	Own	3	3	3	3	3	4	3	2	4	3	3.2	3.0
(numerical model	Supervisor	2	3	2		2	3	1	3	3	4	2.0	3.3
output)	Expert	3	3	1	1	2	3	1	1	1	3	1.6	2.2
Park et al.	Own	2	3	3	3	3	4	3	2	4	3	3.0	3.0
(active learning	Supervisor	1	2	2		2	3	1	3			1.5	2.7
meta-model)	Expert	1	3			1	1	2	1	2	1	1.5	1.5
Finds it useful:		Yes		No		Yes		Maybe		No			
Prefered measure:		ICM		Softmax		Park et al.		Park et al.		Park et al.			
Participant's explanation		ICM				ICM				ICM			

Figure 2: This table shows the three sets of ratings (min. of 1 and max. of 4): 1) the participant's own belief of understanding (row 'own'), 2) the supervisor's belief and 3) the ML expert's opinion of how well the given explanations from the participant matches the measures and examples. It shows whether the participant found a confidence measure useful, their prefered measure and the best match with their explanation of a confidence value.

ACKNOWLEDGEMENTS

This study was performed as part of the the Early Research Program Adaptive Maritime Automation (ERP AMA) within TNO, an independent research organisation in the Netherlands.

REFERENCES

1. Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *AI Magazine* 35, 4 (2014), 105–120.
2. Christopher G. Atkeson, Andrew W. Moore, and Stefan Schaal. 1997. Locally weighted learning for control. In *Lazy learning*. Springer, 75–113.
3. Arnaud Belard, Timothy Buchman, Jonathan Forsberg, Benjamin K. Potter, Christopher J. Dente, Allan Kirk, and Eric Elster. 2017. Precision diagnosis: a view of the clinical decision support systems (CDSS) landscape through the lens of critical care. *Journal of clinical monitoring and computing* 31, 2 (2017), 261–271.
4. Jordi Bieger, Kristinn R. Thórisson, and B. Steunebrink. 2017. Evaluating understanding. In *IJCAI Workshop on Evaluating General-Purpose AI*.
5. Alan Cooper, Robert Reimann, David Cronin, and Christopher Noessel. 2014. *About face: the essentials of interaction design*. John Wiley & Sons.
6. Sumeet Dua and Xian Du. 2016. *Data mining and machine learning in cybersecurity*. CRC press.
7. Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Int. Conf. on Machine Learning*. 1050–1059.
8. David Gunning. 2017. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency* (2017).
9. Bradley Hayes and Julie A. Shah. 2017. Improving Robot Controller Transparency Through Autonomous Policy Explanation. In *Proc. of the 2017 ACM/IEEE Int. conf. on HRI*. ACM, 303–312.
10. Pang Wei Koh and Percy Liang. 2017. Understanding Black-box Predictions via Influence Functions. *arXiv preprint arXiv:1703.04730* (2017).
11. Scott Lundberg and Su-In Lee. 2016. An unexpected unity among methods for interpreting model predictions. *arXiv:1611.07478 [cs]* (Nov. 2016). arXiv: 1611.07478.
12. Tim Miller. 2017. Explanation in artificial intelligence: Insights from the social sciences. *arXiv preprint arXiv:1706.07269* (2017).
13. Tim Miller, Piers Howe, and Liz Sonenberg. 2017. Explainable AI: Beware of Inmates Running the Asylum. In *IJCAI-17 Workshop on Explainable AI (XAI)*. 36.
14. Anh Nguyen, Jason Yosinski, and Jeff Clune. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*. 427–436.
15. Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. 2017. Feature Visualization. *Distill* 2, 11 (2017), e7.
16. No-Wook Park, Phaedon C. Kyriakidis, and Suk-Young Hong. 2016. Spatial estimation of classification accuracy using indicator kriging with an image-derived ambiguity index. *Remote Sensing* 8, 4 (2016), 320.
17. Jasper van der Waa, Jurriaan van Diggelen, and Mark Neerincx. 2018. ICM: An intuitive model independent and accurate certainty measure for machine learning. In *Proc. of the Int. Conf. on Agents and Artificial Intelligence*.
18. Jurriaan van Diggelen, Hans van den Broek, Jan Maarten Schraagen, and Jasper van der Waa. 2017. An Intelligent Operator Support System for Dynamic Positioning. In *Int. Conf. on Applied Human Factors and Ergonomics*. Springer, 48–59.