s c c h
software competence center
hagenberg

FUZZY.LOGIC.LAB.LINZ

Advances in Knowledge-Based Technologies

Proceedings of the Master and PhD Seminar

Winter term 2019/20, part 1

Softwarepark Hagenberg
SCCH, Room 0/2
13 December 2019

Software Competence Center Hagenberg
Softwarepark 21
A-4232 Hagenberg
Tel. +43 7236 3343 800
Fax +43 7236 3343 888
www.scch.at

Fuzzy Logic Laboratorium Linz
Softwarepark 21
A-4232 Hagenberg
Tel. +43 7236 3343 431
Fax +43 7236 3343 434
www.flll.jku.at

Program

Session 1 — Chair: Susanne Saminger-Platz

09:00 Katrin Treitinger:
Transfer Learning with Fuzzy Systems

09:30 Florian Sobieczky:
Accuracy vs. fidelity of explainable AI in the presence of an interpretable base model

Session 2 — Chair: Bernhard Moser

10:00 Werner Zellinger:
Mathematics of Deep Learning: Insights from the Oberwolfach Seminar

Transfer Learning with Fuzzy Systems

Katrin Treitinger
LCM - Linz Center of Mechatronics
Linz, Austria
katrin.treitinger@lcm.at

December 13, 2019

Abstract

Transfer learning is the remarkable human ability to apply knowledge that has already been learned before to a different topic, like making rational decisions, learning a new ability or instrument or recognizing patterns and so on. This is done easily by humans every day and many scientists are trying to imitate this kind of learning with computers. They use statistical, stochastic, functional models or process operations to simulate human thinking or to describe the reality or any other process. These mathematical models base on the two-valued Boolean logic and are an idealization of the real world, because they cannot cope with imprecise linguistic terms, vague concepts or fuzzy information. Being able to allow a partial fulfillment of an attribute and to define sets over the membership degrees of objects, and not over the objects themselves, leads to the concept of fuzzy sets and fuzzy logic. This made it possible to capture objects with unclear boundaries, linguistic terms and expressions and so on. In order to perform logical operations in this environment, the fuzzy conjunction, fuzzy disjunction, fuzzy negation and fuzzy implication operators are defined with the help of triangular norms and triangular conorms. Therefore the intersection, the union and the complement of fuzzy sets can be determined.

Creating mathematical models, which make use of fuzzy sets, fuzzy logic and of the corresponding mathematical framework led to fuzzy systems which consist of a collection of so-called "IF ... THEN ..."-rules, for example the linguistic model, the Takagi-Sugeno fuzzy model, the Takagi-Sugeno-Kang fuzzy model and many more. While the linguistic models have fuzzy sets as inputs and as outputs, which need to be defuzzified to get a crisp output, the Takagi-Sugeno fuzzy models and the Takagi-Sugeno-Kang fuzzy models have a "built-in-defuzzification" as the output is calculated by affine functions and respectively by polynomial or any non-linear functions.

In the further course Takagi-Sugeno fuzzy models are used for transfer learning. The initiated fuzzy model needs to be learned which is realized by clustering the data samples in the source task. Afterward the parameters of the affine function are calculated by the "Weighted Least Squares" WLS method for every rule individually (local learning) or by the "Least Squares" LS method for one solution over all rules (global learning). This determined Takagi-Sugeno model serves as a starting point for a joint optimization (over Source and Target Task) for transfer learning. Three variants for a joint optimization are presented to establish transfer learning in TS fuzzy systems, all of which are relying on the concept of feature space representation learning through distribution matching of rule activation levels:
Variant 1 will match the distributions only in the consequent space, local for each rule separately, while the WLS Part will minimize the error on the source task.
Variant 2 will match the distribution in the consequent space and in the antecedent space, local for each rule, while the WLS part will minimize

the error on the source task.
Variant 3 will match the distribution in the consequent space and in the antecedent space, global over all K rules, while the LS part will minimize the error on the source task.
The local variants are expected to run more stable and quickly, because they have to deal with smaller optimization problems, but they can be "blind" for the global performance and perhaps won't find the global optimum. Matching only in the consequent space is expected to produce more precise results, because of a linear optimization problem, but only if the rules are close and only the functional tendency is different. The global variant will probably need more processing power because it has to deal with high-dimensional distributions.

Accuracy vs. fidelity of explainable AI in the presence of an interpretable base model

Florian Sobieczky - Software Competence Center Hagenberg

After reviewing AI scenarios involving reference to the physical sciences, we propose a model agnostic scheme to achieve interpretability for an AI learning model which delivers its explanations in the framework of an underlying interpretable model.

AI is used here to predict the base model's error and to improve/correct the prediction accuracy. In order to also provide interpretability of the correction, the base model is trained before and after the correction is applied to the labels of the training data. The difference of the two base models parameters allows indicating which features are most prominently used in the correction. We call this approach BAPC="Before and After correction Parameter Comparison" [1]. It can most effectively be applied as a local surrogate. The restriction of AI acting only in a "correcting" fashion on top of an existing base model (i.e. under the allows for rigorous estimates of the neighbourhood of instances on which explanation fidelity is guaranteed.

Related work:

[1] F.Sobieczky: Interpretierbarkeit vs. Genauigkeit von KI in Produktionsprozessen, Predictive Analytics 2019, Slides Login:pran19 PW:datatechnology

[2] Carvalho, Pereira, Cardoso: Machine Learning Interpretability: A Survey on Methods and Metrics, Ele c- tronics 2019, 8, 832

[3] E. P. Denadai: Model Interpretability of Dee p Neural Networks (DNN), Towards Data Science https://towardsdatascience.com/interpretability-of-deep-learning-models-9f52e54d72ab