Credibility and Interpretability/Explainability in Applications of Machine Learning: An Environment of Granular Computing

Witold Pedrycz
Department of Electrical & Computer Engineering
University of Alberta, Edmonton, Canada
wpedrycz@ualberta.ca

Abstract

Over the recent years, we have been witnessing numerous and far-reaching developments and applications of Machine Learning (ML). With the plethora of applications found in critical areas such as autonomous vehicles, health care, networks, complex decision-making environments. Two interrelated challenges become more apparent, namely credibility and interpretability/explainability. Both of them directly impact the acceptance and usefulness of ML constructs in a real-world environment. The credibility is also of concern to any application, especially the one being associated with a high level of criticality.

The notions of interpretability and explainability are formulated and we show how they are realized through a number of auxiliary models built upon the black models of ML constructs. Model-agnostic explainable models are discussed.

Proceeding with a conceptual and algorithmic pursuits, we advocate that the above problems could be formalized in the settings of Granular Computing. We show that to quantify credibility any numeric result has to be augmented by the associated information granules and the quality of the results is quantified in terms of the characteristics of information granules. Different directions are discussed and revisited including confidence/ prediction intervals, granular embedding of ML models, and granular Gaussian Process models.

When coping with interepretability and explainability of ML, information granules and their processing offer key advantages in a number of ways: (i) by stressing the product instead of product perspective and emphasizing importance of interactivity between the user and the explanation module, (ii) by incorporating suitable levels of abstraction, (iii)by building explanation layers with rule-based computing, (iv) by defining and quantifying stability of interpretation, and (v) by proposing ideas of granular counterfactual explanation.