Title: Symbolic Knowledge Extraction from Trained Neural Networks Governed by Lukasiewicz Logics

Author(s): Leandro, Carlos; **Pita, Hélder**; **Monteiro, Luis**

Source: Computational Intelligence **Book Series:** Studies in Computational Intelligence

Volume: 343 **Pages:** 45-58 **Published:** 2011

Document Type: Proceedings Paper

Language: English

Abstract: This work describes a methodology to extract symbolic rules from trained neural networks. In our approach, patterns on the network are codified using formulas on a Lukasiewicz logic. For this we take advantage of the fact that every connective in this multi-valued logic can be evaluated by a neuron in an artificial network having, by activation function the identity truncated to zero and one. This fact simplifies symbolic rule extraction and allows the easy injection of formulas into a network architecture. We trained this type of neural network using a back-propagation algorithm based on Levenderg-Marquardt algorithm, where in each learning iteration, we restricted the knowledge dissemination in the network structure. This makes the descriptive power of produced neural networks similar to the descriptive power of Lukasiewicz logic language, minimizing the information loss on the translation between connectionist and symbolic structures. To avoid redundance on the generated network, the method simplifies them in a pruning phase, using the "Optimal Brain Surgeon" algorithm. We tested this method on the task of finding the formula used on the generation of a given truth table. For real data tests, we selected the Mushrooms data set, available on the UCI Machine Learning Repository.

KeyWords Plus: Algorithm

Reprint Address: Leandro, C (reprint author), Inst Politecn Lisboa, Inst Super Engn Lisboa, Area Cient Matemat, Lisbon, Portugal.

Addresses:
1. Inst Politecn Lisboa, Inst Super Engn Lisboa, Area Cient Matemat, Lisbon, Portugal

Publisher: Springer-Verlag Berlin
Publisher Address: Heidelberger Platz 3, D-14197 Berlin, Germany

ISSN: 1860-949X

ISBN: 978-3-642-20205-6

Citation: LEANDRO, Carlos; PITA, Hélder; MONTEIRO, Luis - Symbolic Knowledge Extraction from Trained Neural Networks Governed by Lukasiewicz Logics. Computational Intelligence. ISSN 1860-949X. ISBN 978-3-642-20205-6. Vol. 343 (2011) p. 45-58.