Rule Extraction from Artificial Neural Networks

Martin Svatoš

December 1, 2016
White-boxes and black-boxes

white-box

- interpretable
- understandable by human (if the model is small)
- decision tree
- rule set

black-box

- good results
- any guarantee on errors?
- random forest
- complex ensemble
- ANNs
Motivation for creating white-boxes out of black-boxes

sometimes there is a customer’s requirement to explain decision of the model

- god and bad debts payers (remember the crisis in 2008)
- self-driving cars (since 1989)

*dark side of deep networks – high number of parameters*

- AlexNet (60M)
- GoogLeNet (5M)
- ResNet (152 layers)

reasons for rule/knowledge extraction

- interpretability vs. explanation
- compression of model
- discovery of latent concepts learned inside a black-box
Motivation for creating white-boxes out of black-boxes

what has been done
- rule extraction from ANNs - NP-hard [3]
- rule extraction from SVM [1]
- seeing the forest through the trees [7]

applications of rule extraction
- control systems
- air pollution levels
- quality of cotton yarn
- fraud detection
- recognizing various hand gesture
- predicting derivative use for financial risk hedging
- theory refinement, neural-symbolic learning cycle
Rule Extraction (RE) from ANNs

properties of ideal RE algorithm

- independent of network’s structure, activation functions, weights’ learning algorithm

properties of a model found by an RE algorithm

- high fidelity (how well the model mimics ANNs decisions)
- small model

basic approaches

- pedagogical - considers only ANNs’ inputs and outputs
- decompositional - considers ANNs’ activation functions, . . .
- eclectic - mix of previous

besides rule sets, also decision trees are mined from ANNs
the first approach for RE from ANNs in 1988

- SUBSET, MofN, CGA, RX, Re-RX, KT, VIA, RuleNet, RULEX, RULENEG, BRAINNE, DEDEC, Glare, NeuroRule, OSRE, HYPINV, CRED, FERNN, BIO-RE, TACO-miner
- Trepan, ExTree
- FRENGA, IGART-FIS, FNES, FuNe I, fuzzy-MLP
Basic approaches

naive pedagogical approach

- try all combinations of inputs nodes, group by output class
- does not say anything about latent concepts
- need for pruning the network before the process (RxREN)

SUBSET, MofN, KT

- decompositional approaches
- for each hidden and output node: find every combination of incoming edges that activate that node (e.g. sum of incoming edges must be greater than bias)
- substitute these rules instead of nodes, transform it to a rule set

TREPAN

- pedagogical approach
- rule extraction as learning
- oracle based method
- produces \( M\text{-of-}N \) decision tree using beam search
- good news: there is still a working implementation

last five years

- NN-LFIT
- MNIST dataset
- first-order extension of TREPAN for CILP++ [2]
Experiments with TREPAN

Train fidelity:
- DNC
- CasCor
- KBANN
- TopGen
- REGENT

Test fidelity:
- mushroom
- monks2
- monks3
- monks1
- ionosphere
- labor
- iris
- wine
- promoters
- votes
- breastCancer1
- breastCancer2
- breastCancer3
- lenses
- diabetes
- glass
- horse
- segmentation
- soybean
- splice


Son N Tran and A d’Avila Garcez. “Knowledge extraction from deep belief networks for images”. In: IJCAI-2013 Workshop on Neural-Symbolic Learning and Reasoning. 2013.
