Rule Extraction from Artificial Neural Netwroks

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

RE methods

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-of-N* decision tree using beam search
- good news: there is still a working implementation

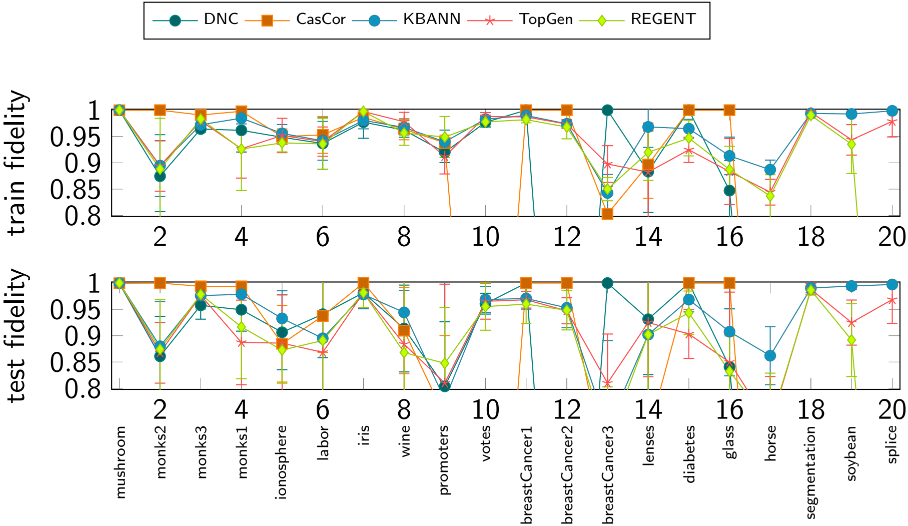
Current Approach for Deep Networks

first RE from deep network in 2000 [4] – 2 hidden layers

last five years

- NN-LFIT
- MNIST dataset
- DeepRED [8] based on CRED [5]
- RE from Deep Belief Networks [6] – RBM for images
- first-order extension of TREPAN for CILP++ [2]

Experiments with TREPAN



Q & A

Bibliography I

Joachim Diederich. *Rule extraction from support vector machines*. Vol. 80. Springer Science & Business Media, 2008.

Manoel Vitor Macedo França, Artur S d'Avila Garcez, and Gerson Zaverucha. "Relational Knowledge Extraction from Neural Networks". In: (2015).

M Golea. "On the complexity of rule extraction from neural networks and network querying". In: *R ules and N et w orks* (1996), p. 5.

DaeEun Kim and Jaeho Lee. "Handling continuous-valued attributes in decision tree with neural network modeling". In: *European Conference on Machine Learning*. Springer. 2000, pp. 211–219.

Makoto Sato and Hiroshi Tsukimoto. "Rule extraction from neural networks via decision tree induction". In: *Neural Networks, 2001. Proceedings. IJCNN'01. International Joint Conference on*. Vol. 3. IEEE. 2001, pp. 1870–1875.

Bibliography II

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.

Anneleen Van Assche and Hendrik Blockeel. "Seeing the forest through the trees: Learning a comprehensible model from an ensemble". In: *European Conference on Machine Learning*. Springer. 2007, pp. 418–429.

Jan Ruben Zilke, Eneldo Loza Mencıa, and Frederik Janssen. "DeepRED–Rule Extraction from Deep Neural Networks". In: *International Conference on Discovery Science*. Springer. 2016, pp. 457–473.