Hybrid Intelligent Systems

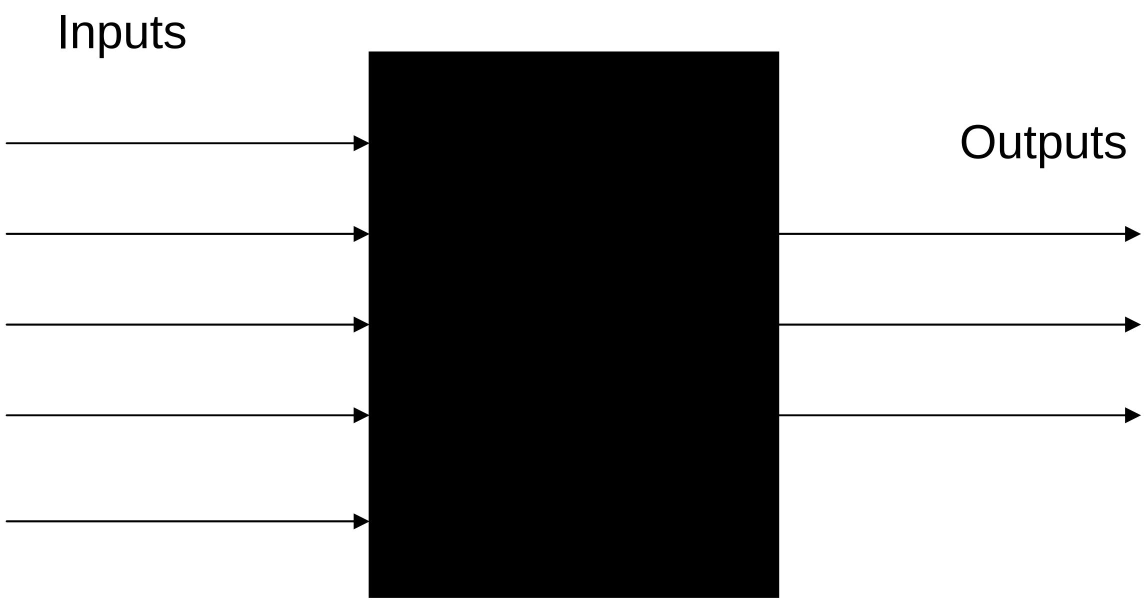
Lecture 10

Rule extraction from neural networks

Outlines

- Black boxes
- Rule extraction
- Neural networks for rule extraction
- Sample problems
- Bibliography

Neural Networks is Black Box



Black-Box Models

- Aims of many data analysis's methods (pattern recognition, neural networks, evolutionary computation and related):
  - building predictive data models
  - adapting internal parameters of the data models to account for the known (training) data samples
  - allowing for predictions to be made on the unknown (test) data samples

Dangers

- Using a large number of numerical parameters to achieve high accuracy
  - overfitting the data
  - many irrelevant attributes may contribute to the final solution

Drawbacks

- Combining predictive models with *a priori* knowledge about the problem is difficult
- No systematic reasoning
- No explanations of recommendations
- No way to control and test the model in the areas of the future
- Unacceptable risk in safety-critical domains (medical, industrial)

Reasoning with Logical Rules

- More acceptable to human users
- Comprehensible, provides explanations
- May be validated by human inspection
- Increases confidence in the system

Machine Learning

- Explicit goal: the formulation of symbolic inductive methods
  – methods that learn from examples
- Discovering rules that could be expressed in natural language
  – rules similar to those a human expert might create

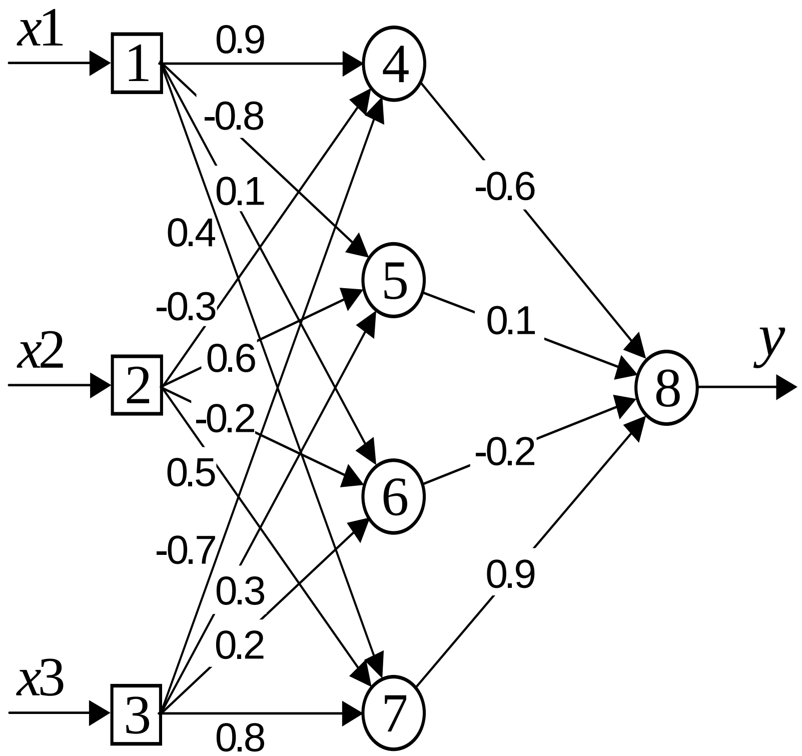
Neural Networks as Black Boxes

- Perform mysterious functions
- Represent data in an incomprehensible way

- Two issues:
    1. understanding what neural networks really do
    2. using neural networks to extract logical rules describing the data.

Sample

From neuron:		1	2	3	4	5	6	7	8
To neuron:	1	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	0	0	0
	3	0	0	0	0	0	0	0	0
	4	0.9	-0.3	-0.7	0	0	0	0	0
	5	-0.8	0.6	0.3	0	0	0	0	0
	6	0.1	-0.2	0.2	0	0	0	0	0
	7	0.4	0.5	0.8	0	0	0	0	0
	8	0	0	0	-0.6	0.1	-0.2	0.9	0



Sample (Cont.)

- If H7 and not H4 then Y
- If X1and not X3 then H4
- If X1 and X2 and X3 the H7

- If H7(90) and not H4(60) then Y
- If X1(90) and not X3(70) then H4
- If X1(40) and X2(50) and X3(80) the H7

Techniques for acquisition of Information from Trained ANN

- Sensitivity analysis
- Neural Network Visualization
- Rule Extraction

Sensitivity Analysis

- Probe ANN with test inputs, and record the outputs
- Determining the impact or effect of an input variable on the output
  - hold the other inputs to some fixed value (e.g. mean or median value), vary only the input while monitoring the change in outputs

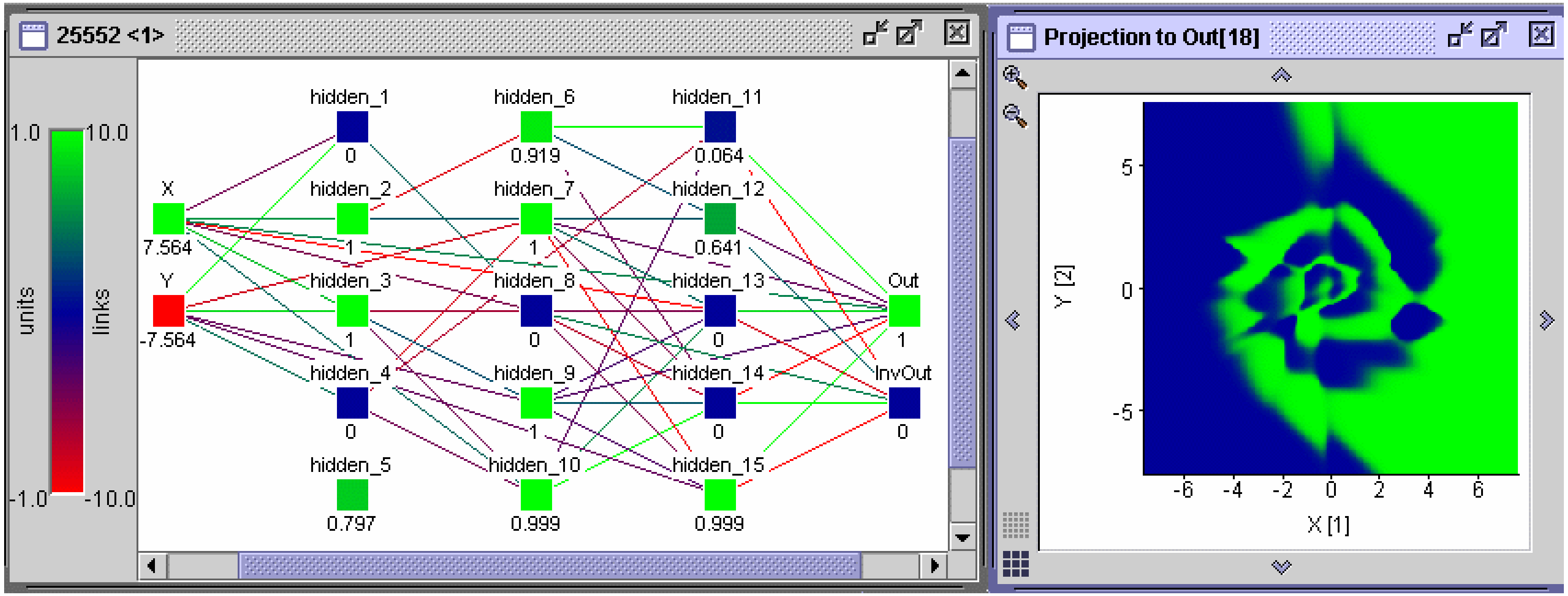
Automated Sensitivity Analysis

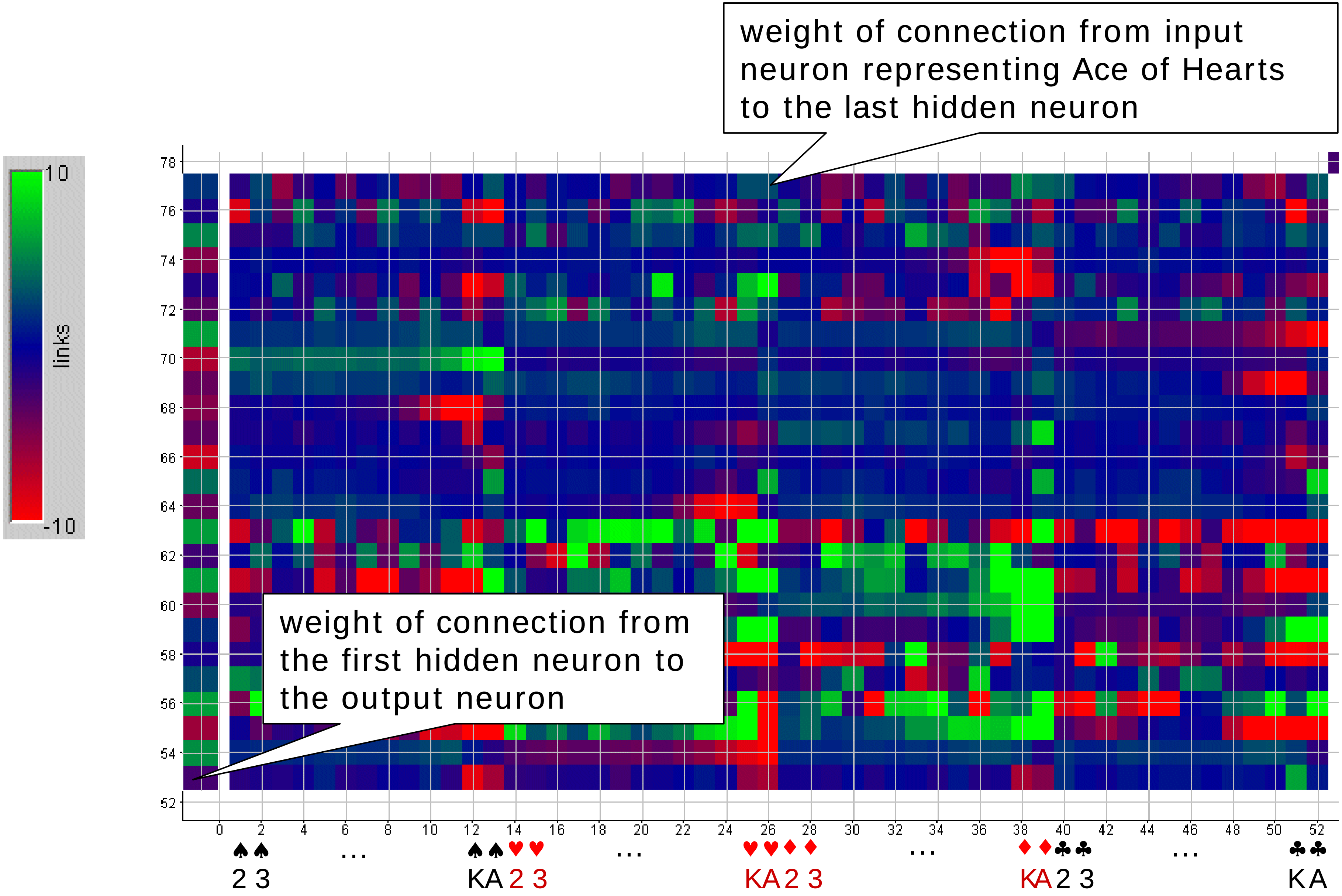
- For backpropagation ANN:
  - keep track of the error terms computed during the back propagation step
  - measure of the degree to which each input contributes to the output error
    - the largest error $\equiv$ the largest impact
  - the relative contribution of each input to the output errors can be computed by acumulating errors over time and normalizing them

Neural Network Visualization

- Using power of human brain to see and recognize patterns in two- and three-dimensional data

Visualization Samples




weight of connection from input neuron representing Ace of Hearts to the last hidden neuron
weight of connection from the first hidden neuron to the output neuron
10
links
-10
♠♠ 2 3
...
♠♠♥♥ K A 2 3
...
♥♥♦♦ K A 2 3
...
♦♦♣♣ K A 2 3
...
♣♣ K A

RULE EXTRACTION

Propositional Logic Rules

- Standard crisp (boolean) propositional rules:

$$\text{IF } x \in X^{(i)} \text{ THEN } Class(x) = C_k$$

- Fuzzy version is a mapping from *X* space to the space of fuzzy class labels
- Crisp logic rules should give precise **yes** or **no** answers

Condition Part of Logic Rule

- Defined by a conjuction of logical predicate functions
- Usually predicate functions are tests on a single attribute
  - if feature **k** has values that belong to a subset (for discrete features) or to an interval or (fuzzy) subsets for attribute **K**

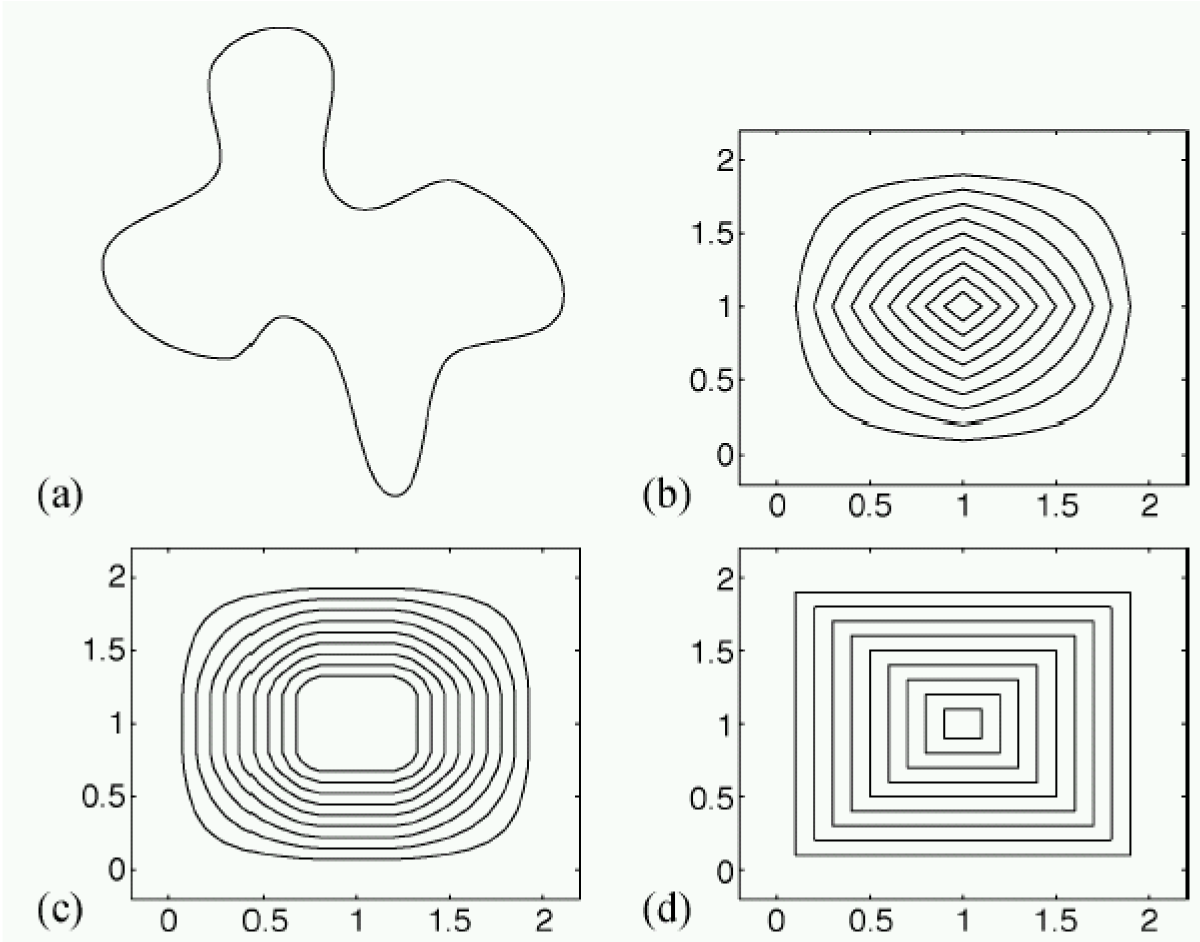
Decision Borders

(a) - general clusters

(b) - fuzzy rules

(c) - rough rules

(d) - crisp logical rules



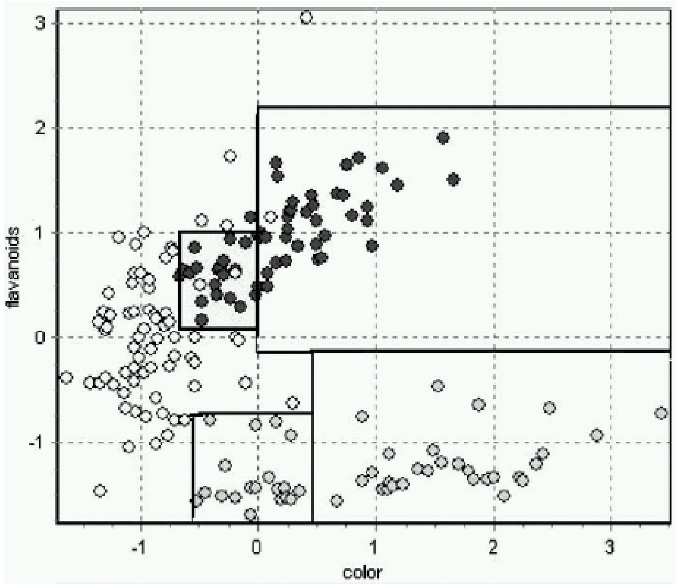
source: Duch et.al, Computational Intelligence Methods..., 2004

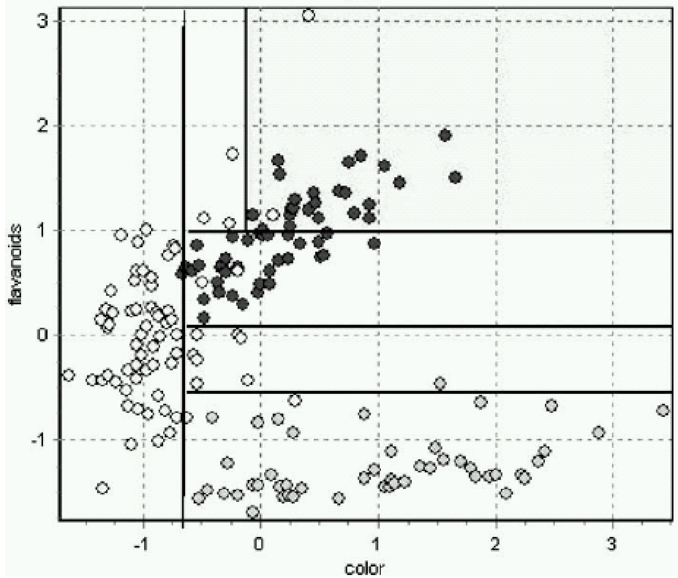
Linguistic Variables

- Attempts to verbalize knowledge require symbolic inputs (called linguistic variables)
- Two types of linguistic variables:
  - context-independent - identical in all regions of the feature space
  - context-dependent - may be different in each rule

Decision Trees

- Fast and easy to use
- Hierarchical rules that they generate have somewhat limited power





source: Duch et.al, Computational Intelligence Methods..., 2004

NEURAL NETWORKS FOR RULE EXTRACTION

Neural Rule Extraction Methods

- Neural networks are regarded commonly as black boxes but can be used to provide simple and accurate sets of logical rules
- Many neural algorithms extract logical rules directly from data have been devised

Categorizing Rule-Extraction Techniques

- Expressive power of extracted rules
- Translucency of the technique
- Specialized network training schemes
- Quality of extracted rules
- Algorithmic complexity
- The treatment of linguistic variables

Expressive Power of Extracted Rules

- Types of extracted rules:
  - crisp logic rules
  - fuzzy logic rules
  - first-order logic form of rules - rules with quantifiers and variables

Translucency

- The relationship between the extracted rules and the internal architecture of the trained ANN
- Categories:
  - decompositional (local methods)
  - pedagogical (global methods)
  - eclectic

Translucency - Decompositional Approach

- To extract rules at the level of each individual hidden and output unit within the trained ANN
  - some form of analysis of the weight vector and associated bias of each unit
  - rules with antecedents and consequents expressed in terms which are local to the unit
  - a process of aggregation is required

Translucency - Pedagogical Approach

- The trained ANN viewed as a black box
- Finding rules that map inputs directly into outputs
- Such techniques typically are used in conjunction with a symbolic learning algorithm
  - use the trained ANN to generate examples for the training algorithm

Specialized network training schemes

- If specialized ANN training regime is required
- It provides some measure of the "portability" of the rule extraction technique across various ANN architectures
- Underlaying ANN can be modified by the rule extraction process

Quality of extracted rules

- Criteria:
  - accuracy - if can correctly classify a set of previously unseen examples
  - fidelity - if extracted rules can mimic the behavior of the ANN
  - consistency - if generated rules will produce the same classification of unseen examples
  - comprehensibility - size of the rules set and number of antecendents per rule must be appropriate

Algorithmic complexity

- Important especially for decompositional approaches to rule extraction
  - usually the basic process of searching for subsets of rules at the level of each (hidden and output) unit in the trained ANN is exponential in the number of inputs to the node

The Treatment of Linguistic Variables

- Types of variables which limit usage of techniques:
  - binary variables
  - discretized inputs
  - continuous variables that are converted to linguistic variables automatically

Techniques Reviews

- Andrews et.al, A survey and critique..., 1995 - 7 techniques described in detail
- Tickle et.al, The truth will come to light ..., 1998 - 3 more techiques added
- Jacobsson, Rule extraction from recurrent ..., 2005, techniques for recurrent neural networks

SAMPLE PROBLEMS

Wisconsin Breast Cancer

- Data details:
  - 699 cases
  - 9 attributes f1-f9 (1-10 integer values)
  - two classes:
    458 benign (65.5%)
    241 malignant (34.5%).
  - for 16 instances one attribute is missing

source: http://www.ics.uci.edu/~mlearn/MLRepository.html

Wisconsin Breast Cancer - results

- Single rule:

  IF f2 = [1,2] then benign else malignant

  – 646 correct (92.42%), 53 errors

- 5 rules for malignant:

  R1: f1<9 & f4<4 & f6<2 & f7<5
  R2: f1<10 & f3<4 & f4<4 & f6<3
  R3: f1<7 & f3<9 & f4<3 & f6=[4,9] & f7<4
  R4: f1=[3,4] & f3<9 & f4<10 & f6<6 & f7<8
  R5: f1<6 & f3<3 & f7<8
  ELSE: benign

  – 692 correct (99%), 7 errors

source: http://www.phys.uni.torun.pl/kmk/projects/rules.html#Wisconsin

The MONKs Problems

- Robots are described by six diferent attributes:
  - x1: head_shape $\in$ round square octagon
  - x2: body_shape $\in$ round square octagon
  - x3: is_smiling $\in$ yes no
  - x4: holding $\in$ sword balloon flag
  - x5: jacket_color $\in$ red yellow green blue
  - x6: has_tie $\in$ yes no

source: ftp://ftp.funet.fi/pub/sci/neural/neuroprose/thrun.comparison.ps.Z

The MONKs Problems cont.

- Binary classification task
- Each problem is given by a logical description of a class
- Only a subset of all 432 possible robots with its classification is given

source: ftp://ftp.funet.fi/pub/sci/neural/neuroprose/thrun.comparison.ps.Z

The MONKs Problems cont.

- M1:
  (head_shape = body_shape) or (jacket_color = red)
  - 124 randomly selected training samples
- M2:
  exactly two of the six attributes have their first value
  - 169 randomly selected training samples
- M3:
  (jacket_color is green and holding a sword) or (jacket_color is not blue and body shape is not octagon)
  - 122 randomly selected training samples with 5% misclassifications (noise in the training set)

M1, M2, M3 – best results

- C-MLP2LN algorithm (100% accuracy):
  - M1: 4 rules + 2 exception, 14 atomic formulae
  - M2: 16 rules and 8 exceptions, 132 atomic formulae
  - M3: 33 atomic formulae

source: http://www.phys.uni.torun.pl/kmk/projects/rules.html#Monk1

BIBLIOGRAPHY

References

- Duch, W., Setiono, R., Zurada, J.M., **Computational Intelligence Methods for Rule-Based Data Understanding**, Proceedings of the IEEE, 2004, vol. 92, Issue 5, pp. 771-805

Surveys

- R. Andrews, J. Diederich, and A. B. Tickle, **A survey and critique of techniques for extracting rules from trained artificial neural networks**, Knowl.-Based Syst., vol. 8, pp. 373–389, 1995
- A. B. Tickle, R. Andrews, M. Golea, and J. Diederich, **The truth will come to light: Directions and challenges in extracting the knowledge embedded within trained artificial neural networks**, IEEE Trans. Neural Networks, vol. 9, pp. 1057–1068, Nov. 1998.

Surveys cont.

- I. Taha, J. Ghosh, **Symbolic interpretation of artifcial neural networks**, Knowledge and Data Engineering vol. 11, pp. 448-463, 1999
- H. Jacobsson, **Rule extraction from recurrent neural networks: A Taxonomy and Review**, 2005 [citeseer]

Problems

- S.B. Thrun *et al.*, **The MONK's problems: a performance comparison of different learning algorithms**, Carnegie Mellon University, CMU-CS-91-197 (December 1991)
- http://www.phys.uni.torun.pl/kmk/projects/rules.html (prof. Włodzisław Duch)