NATIONAL RESEARCH
UNIVERSITY

Interval Pattern Concept Lattice as a Classifier Ensemble

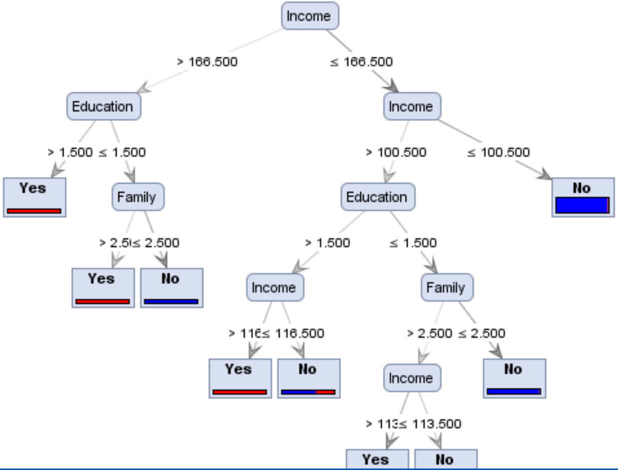
5^{th} Workshop "What can FCA do for Artificial Intelligence"

Yury Kashnitsky and Sergei O. Kuznetsov
{ykashnitsky,skuznetsov}@hse.ru

National Research University
Higher School of Economics (Moscow)

Decision trees for classification

In this work, we address supervised classification, especially decision tree learning.



Problems with decision tree learning

Suboptimality of decision trees

Decision trees:

- Optimize some information criterion (e.g. information gain or Gini impurity)
- Do it greedily
- May not find globally optimal solution

Idea: Why not brute-forcing the search for an optimal hypothesis?

Problems with decision tree learning

Interpretibility & performance

Decision trees:

- \+ Are easily interpreted. Just one (short) rule for each test object
- \- Have poor classification performance

Random forests:

- \+ Often have very good performance
- \- Are much less interpretable. Lots of (long) rules for each test object or just feature importances

Idea: Something in between?

Plan

We are going to:

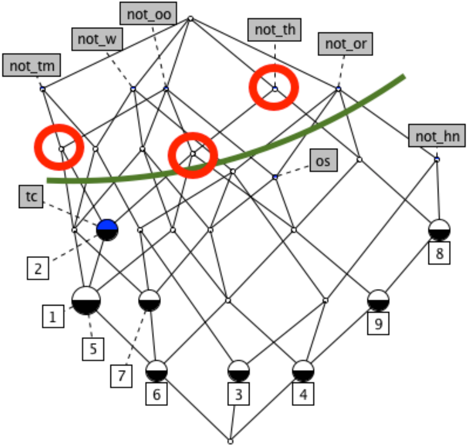
- Find a set of the best hypotheses (in terms of some information criterion)
- Do it for each test instance individually
- Classify each test instance with the set of the best rules
- Handle binary, numeric and complex data

Meanwhile we will:

- Interpret decision trees as the search for a hypothesis in an interval pattern concept lattice.

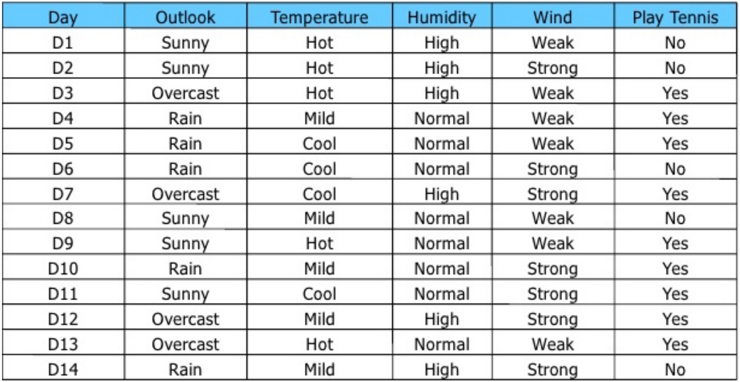
Idea

The main idea: search for classification rules among formal concepts.



Example

A “classical” toy classification problem



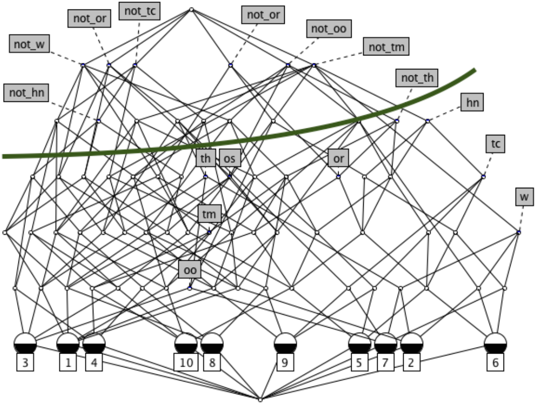
Day	Outlook	Temperature	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	Normal	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	High	Strong	Yes
D8	Sunny	Mild	Normal	Weak	No
D9	Sunny	Hot	Normal	Weak	Yes
D10	Rain	Mild	Normal	Strong	Yes
D11	Sunny	Cool	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Example II

$G\backslash M$	**os**	$\neg os$	**oo**	$\neg oo$	**or**	$\neg or$	**th**	$\neg th$	**tm**	$\neg tm$	**tc**	$\neg tc$	$\neg hn$	**hn**	**w**	$\neg w$	**play**
1	×			×		×	×			×		×	×			×	
2	×			×		×	×			×		×	×		×		
3		×	×			×	×			×		×	×			×	×
4		×		×	×			×	×			×	×			×	×
5		×		×	×			×		×	×			×		×	×
6		×		×	×			×		×	×			×	×		
7		×	×			×		×		×	×			×	×		×
8	×			×		×		×	×			×	×			×	
9	×			×		×		×		×	×			×		×	×
10		×		×	×			×	×			×		×		×	×
11	×			×		×		×	×			×		×	×		?
12		×	×			×		×	×			×	×		×		?
13		×	×			×	×			×		×		×		×	?
14		×		×	×			×	×			×	×		×		?

Example III

A concept lattice with a 0.4 minimal support constraint



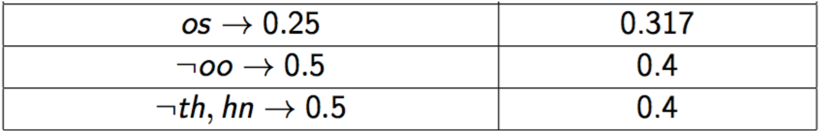
Example IV

10 best classification rules

	Rule	**Gini impurity**
1	$os, \neg tc, \neg hn \rightarrow 0$	0.171
2	$\neg os, \neg w \rightarrow 1$	0.267
3	$\neg oo, \neg tm, w \rightarrow 0$	0.3
4	$os, \neg tc, \neg hn, \neg w \rightarrow 0$	0.3
5	$os, th, \neg hn \rightarrow 0$	0.3
6	$os \rightarrow 0.25$	0.317
7	$\neg oo, \neg tc, \neg hn \rightarrow 0.25$	0.317
8	$\neg or, \neg tc, \neg hn \rightarrow 0.25$	0.317
9	$\neg os \rightarrow 0.83$	0.317
10	$or, \neg th, \neg w \rightarrow 1$	0.343

Example V

3 best rules with Gini impurities to classify the instance *Outlook=sunny, Temperature=mild, Humidity=normal, Windy=true*



$os \rightarrow 0.25$	0.317
$\neg oo \rightarrow 0.5$	0.4
$\neg th, hn \rightarrow 0.5$	0.4

Aggregating this rules yields classifying the instance as negative: $\frac{1}{3}(0.25 + 0.5 + 0.5) \approx 0.4$

How trees deal with continuous attributes

	Age	Default	
1	17	1	19.5
2	22	0	
3	47	0	49.5
4	52	1	
5	67	1	82.5
6	98	0	

New features
$age \leq 19.5$
$age > 19.5$
$age \leq 49.5$
$age > 49.5$
$age \leq 82.5$
$age > 82.5$

Example.
Interval pattern structures and scaling

A simple many-valued context and an interordinal scale

	a
1	4.6
2	4.7
3	4.9
4	5.0
5	5.1

	$a \leq 4.6$	$a \leq 4.7$	$a \leq 4.9$	$a \leq 5.0$	$a \leq 5.1$	$a \geq 4.6$	$a \geq 4.7$	$a \geq 4.9$	$a \geq 5.0$
1	×	×	×	×	×	×			
2		×	×	×	×	×	×		
3			×	×	×	×	×	×	
4				×	×	×	×	×	×
5					×	×	×	×	×

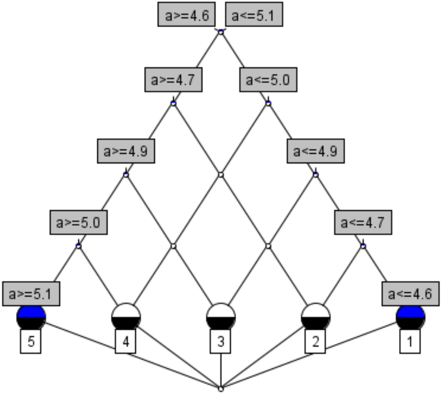
Example.
Interval pattern structures and scaling II

An Interval Pattern Structure here is a triple $IPS = \{G,\ (D,\ \sqcap),\ \delta\}$, where

- $G = \{1, 2, 3, 4, 5\}$ is a set of objects;
- $D = \{[a, b]\}, a, b \in \mathbb{R} \cup \{-\infty, +\infty\}, a \leq b$ – is a set of intervals (descriptions);
- $\sqcap$ – is a semilattice operator for intervals;
- $\delta : G \to D$ – is a mapping that matches an object from G to its description (interval) from D. Here $\delta(1) = [4.6, 4.6], \ldots, \delta(5) = [5.1, 5.1]$.

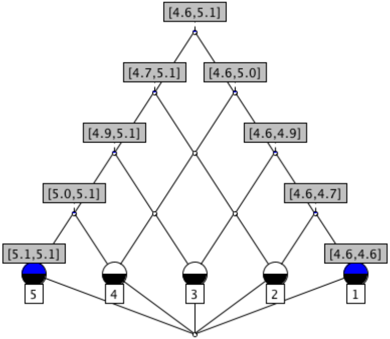
Example.
Interval pattern structures and scaling III

Concept lattice for an interordinaly scaled context.



Example.
Interval pattern structures and scaling IV

A pattern concept lattice corresponding to an interval pattern structure *IPS*.



Isomorphism

In (Kuznetsov, 2009) they show the isomorphism between a concept lattice of an interordinaly scaled context and a concept lattice of the corresponding interval pattern structure.

Pattern structure projection

Definition

A **projection** of a pattern structure $\psi : D \to D$ is a kernel operator that is:

- Monotone: $x \sqsubseteq y \Rightarrow \psi(x) \sqsubseteq \psi(y)$
- Contractive: $\psi(x) \sqsubseteq x$
- Idempotent: $\psi(\psi(x)) = \psi(x)$

Discretizing Pattern Structure Projections

To express decision tree learning in terms of FCA, we introduce discretizing pattern structure projections.

Definition

Let $(G,\ (D,\ \sqcap),\ \delta)$ be an interval pattern structure and let m be the number of attributes. Let $T_i = \{\tau_{i1}, \dots, \tau_{it_i}\}$ $(\tau_{ij} \in \mathbb{R}, i \in [1, m], j \in [1, t_i], t_i \in \mathbb{N})$ be sets of real numbers. Then $\psi(\langle [a_i, b_i] \rangle_{i \in [1,m]}) = \langle [max\{\tau \mid \tau \in T_i \cup \{-\infty, +\infty\}, \tau \leq a_i\}, min\{\tau \mid \tau \in T_i \cup \{-\infty, +\infty\}, \tau \geq b_i\}] \rangle$ is called a **discretizing pattern structure projection** for a pattern structure $(G,\ (D,\ \sqcap),\ \delta)$.

Proposition

A discretizing projection, as introduced here, is a projection according the definition above.

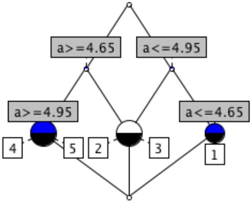
Example

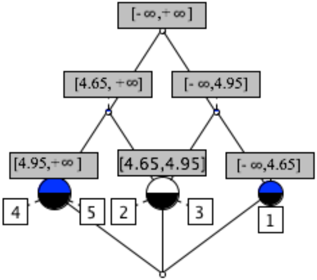
A context formed by discretizing the feature *a* from the previous example with thresholds 4.65 and 4.95.

	$a \leq 4.65$	$a \leq 4.95$	$a \geq 4.65$	$a \geq 4.95$
1	×	×		
2		×	×	
3		×	×	
4			×	×
5			×	×

Example II

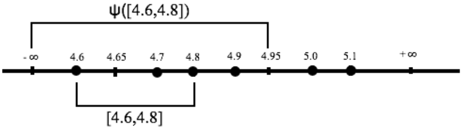
A concept lattice for the above context and an isomorphic pattern concept lattice.





Example III

The intuition behind a dicretizing projection.

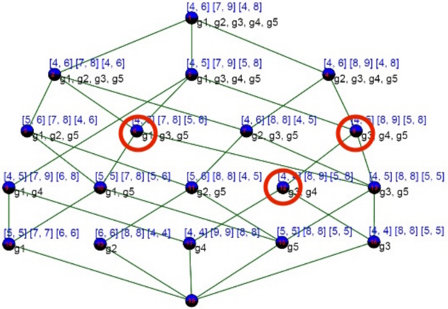


What for?

Thus we interpret the popular decision tree heuristic of discretizing numeric attributes in terms of pattern structures and learning with trees - as the search for a hypothesis in a concept lattice.

Idea

Searching for hypotheses in an interval pattern concept lattice.



Pseudocode

Algorithm 1 Concept Lattice-Based Rule-learner (CoLiBRi)

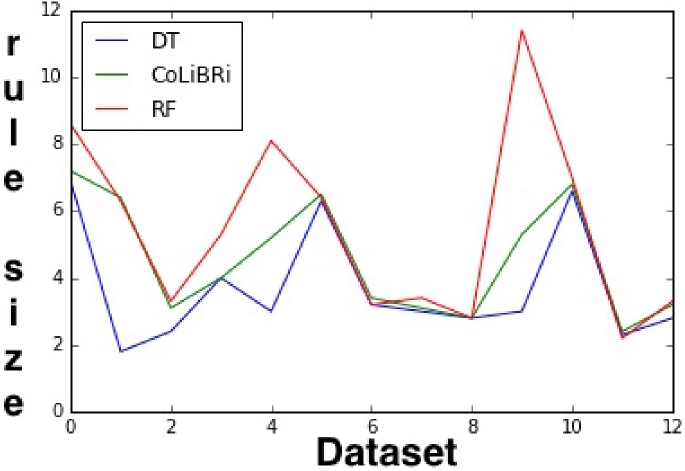
```
Input: $PS_{train} = (G_{train}, ((D, \sqcap), c_{train}), \delta_{train})$
$PS_{test} = (G_{test}, (D, \sqcap), \delta_{test})$
$min\_supp \in \mathbb{R}^+, n_{rules} \in \mathbb{N};$
$CbO_{PS}(PS, min\_supp) : PS \to \mathcal{S};$
$inf : D \times c_{train} \to \mathbb{R};$
$sort(\mathcal{S}, inf) : \mathcal{S} \to \mathcal{S}$

Output: $c_{test}, r_{test}$

  $c_{test} = \emptyset, r_{test} = \emptyset$
  $f_{pos} = \frac{|c'_{train}|}{|G_{train}|}$
  $\mathcal{S} = \{(A, d) : inf(d, c_{train}) \mid A \subseteq G_{train}, d \in D, A^\diamond = d, d^\diamond = A, |A| \geq min\_supp\} =$
  $CbO_{PS}(PS_{train}, min\_supp)$
  $\mathcal{S} = sort(\mathcal{S}, inf)$
  for $g_t \in G_{test}$ do
    $\{d_i\}_{i \in [1, n_{rules}]} = \{d \mid (A, d) \in \mathcal{S}, g_t^\diamond \sqsubseteq d\}$
    $f_i^+ = \frac{|d_i^\diamond \cap c'_{train}|}{|d_i^\diamond|}$
    $r_{test}[i] = \{d_i \to f_i^+\}_{i \in [1, n_{rules}]}$
    $c_{test}[i] = [\sum_{i=1}^{n\_rules} f_i^+ \geq f_{pos} * n\_rules]$
  end for
```

Experiments

dataset	DT acc	RF acc	kNN acc	CoLiBRi acc	DT F1	RF F1	kNN F1	CoLiBRi F1
audiology	0.75	**0.8**	0.63	0.79*	0.71	**0.74**	0.58	**0.74**
balance-scale	0.63	0.66	**0.76**	0.65	0.58	0.63	**0.75**	0.61
breast_cancer	0.7	0.74	0.73	**0.76**	**0.45**	0.42	0.38	0.44*
car	0.75	0.78*	0.71	**0.79**	0.75	**0.76**	0.71	**0.76**
hayses-roth	0.84*	0.83*	0.49	**0.86**	0.84*	0.82	0.49	**0.85**
lymph	0.8	0.83	**0.86**	0.83	0.77	**0.85**	0.84*	0.84*
mol_bio_prom	0.78	**0.83**	**0.83**	0.82*	0.78	**0.84**	0.8	0.83*
nursery	0.64	0.65	**0.72**	0.65	0.62	0.62	**0.7**	0.62
primary_tumor	0.41	**0.46**	0.41	0.45*	0.37	**0.41**	0.37	0.4*
solar_flare	0.7*	0.7*	0.63	**0.72**	0.67	0.69*	0.6	**0.71**
soybean	0.91*	0.91*	**0.92**	0.91*	0.91*	**0.93**	0.92*	0.91*
spect_train	0.61	**0.69**	0.68*	0.7	0.34	0.36*	0.23	**0.38**
tic-tac-toe	0.79	0.79	**0.85**	0.78	0.82	0.86	**0.89**	0.85

Accuracies and F1-scores in classification experiments with the UCI machine learning datasets. “DT acc” and “DT F1” stand for average 5-run 5-fold CV accuracy and F1 score of the CART algorithm, …, “CoLiBRi F1” stands for average 5-run 5-fold CV F1 score of the proposed CoLiBRi algorithm.

Average rule size



Conclusions & Future

Conclusions:

- Throw much computational resources on important tasks!
- Get accurate and interpretable rules!

Anticipations & Discussion:

- Complex structured data as a perspective
- Hope FCA helps in “exhaustive” predictions

Thank you!