Mining Implications from Lattices of Closed Trees

José L. Balcázar, Albert Bifet and Antoni Lozano

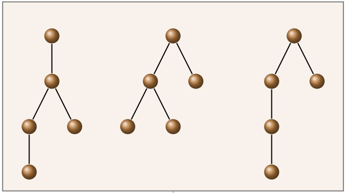
Universitat Politècnica de Catalunya

Extraction et Gestion des Connaissances EGC'2008
2008 Sophia Antipolis, France

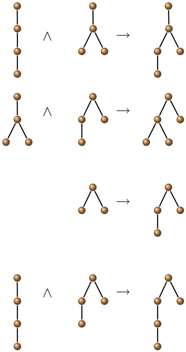
Introduction

Problem

Given a dataset $\mathscr{D}$ of rooted, unlabelled and unordered trees, find a “basis”: a set of rules that are sufficient to infer all the rules that hold in the dataset $\mathscr{D}$.



$\mathscr{D}$

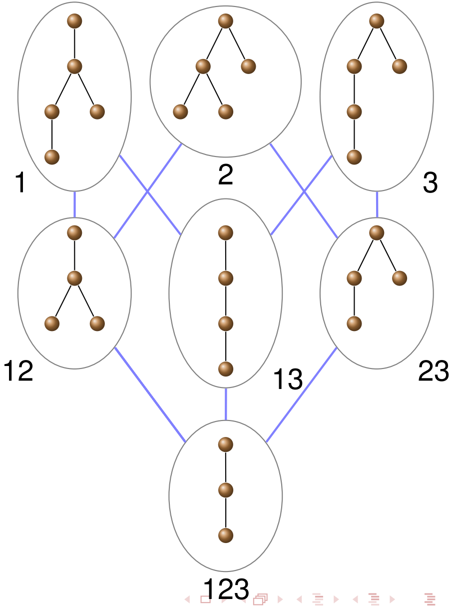


Introduction

Set of Rules:

$$A \rightarrow \Gamma_{\mathscr{D}}(A).$$

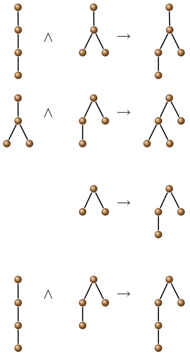
- antecedents are obtained through a computation akin to a hypergraph transversal
- consequents follow from an application of the closure operators

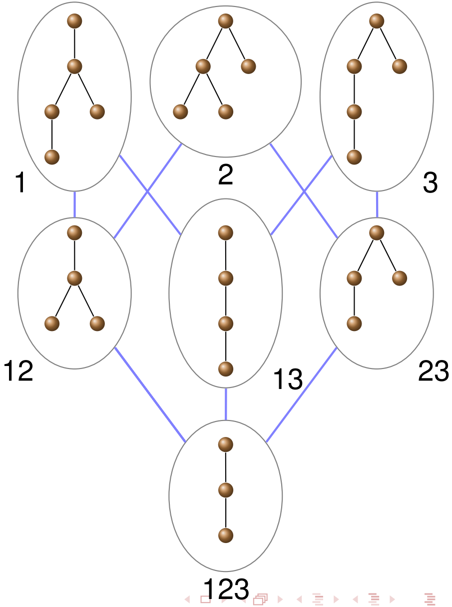


Introduction

Set of Rules:

$$A \longrightarrow \Gamma_{\mathscr{D}}(A).$$





Trees

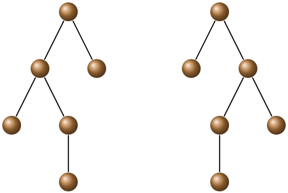
Our trees are:

- Rooted
- Unlabeled
- Unordered

Our subtrees are:

- Induced
- Top-down

Two different ordered trees
but the same unordered tree



Deterministic association rules

- Logical implications are the traditional mean of representing knowledge in formal AI systems. In the field of data mining they are known as **association rules**.

M	a	b	c	d
m_1	1	1	0	1
m_2	0	1	1	1
m_3	0	1	0	1

$$\begin{aligned} a &\rightarrow b, d \\ d &\rightarrow b \\ a, b &\rightarrow d \end{aligned}$$

- Deterministic association rules are implications with 100% confidence.
- An advantatge of deterministic association rules is that they can be studied in purely logical terms with **propositional Horn logic**.

Propositional Horn Logic

M	a	b	c	d
m_1	1	1	0	1
m_2	0	1	1	1
m_3	0	1	0	1

$$\begin{array}{rl} a \rightarrow b, d & (\bar{a} \vee b) \wedge (\bar{a} \vee d) \\ d \rightarrow b & \bar{d} \vee b \\ a, b \rightarrow d & \bar{a} \vee \bar{b} \vee d \end{array}$$

- Assume a finite number of variables.
  - $V = \{a, b, c, d\}$
- A clause is **Horn** iff it contains at most one positive literal.
  - $\bar{a} \vee \bar{b} \vee d$ $\quad$ $a, b \rightarrow d$
- A **model** is a complete truth assignment from variables to $\{0, 1\}$.
  - $m(a) = 0, m(b) = 1, m(c) = 1, \ldots$
- Given a set of models M, the Horn theory of M corresponds to the conjunction of all Horn clauses satisfied by all models from M.

Propositional Horn Logic

Theorem

Given a set of models M, there is exactly one minimal Horn theory containing it. Semantically, it contains all the models that are intersections of models of M. This is sometimes called the **empirical Horn approximation.**

We propose

- Closure operator
- translation of tree set of rules to a specific propositional theory

Closure Operator

- $\mathscr{D}$: the finite input dataset of trees
- $\mathscr{T}$: the (infinite) set of all trees

Definition

We define the following the Galois connection pair:

- For finite $A \subseteq \mathscr{D}$
  - $\sigma(A)$ is the set of subtrees of the A trees in $\mathscr{T}$

$$\sigma(A) = \{t \in \mathscr{T} \mid \forall t' \in A (t \preceq t')\}$$

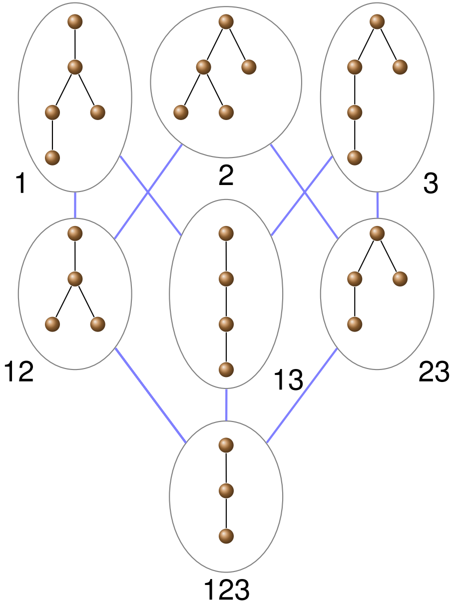
- For finite $B \subset \mathscr{T}$
  - $\tau_{\mathscr{D}}(B)$ is the set of supertrees of the B trees in $\mathscr{D}$

$$\tau_{\mathscr{D}}(B) = \{t' \in \mathscr{D} \mid \forall t \in B (t \preceq t')\}$$

Closure Operator

The composition $\Gamma_{\mathscr{D}} = \sigma \circ \tau_{\mathscr{D}}$ is a closure operator.

Galois Lattice of closed set of trees



Model transformation

Intuition

- One propositional variable v_t is assigned to each possible subtree t.
- A set of trees A corresponds in a natural way to a model m_A.
- Let m_A be a model: we impose on m_A the constraints that if $m_A(v_t) = 1$ for a variable v_t, then $m_A(v_{t'}) = 1$ for all those variables $v_{t'}$ such that $v_{t'}$ represents a subtree of the tree represented by v_t.

$$\mathscr{R}_0 = \{ v_{t'} \to v_t \mid t' \preceq t, t \in \mathscr{U}, t' \in \mathscr{U} \}$$

Model transformation

Theorem

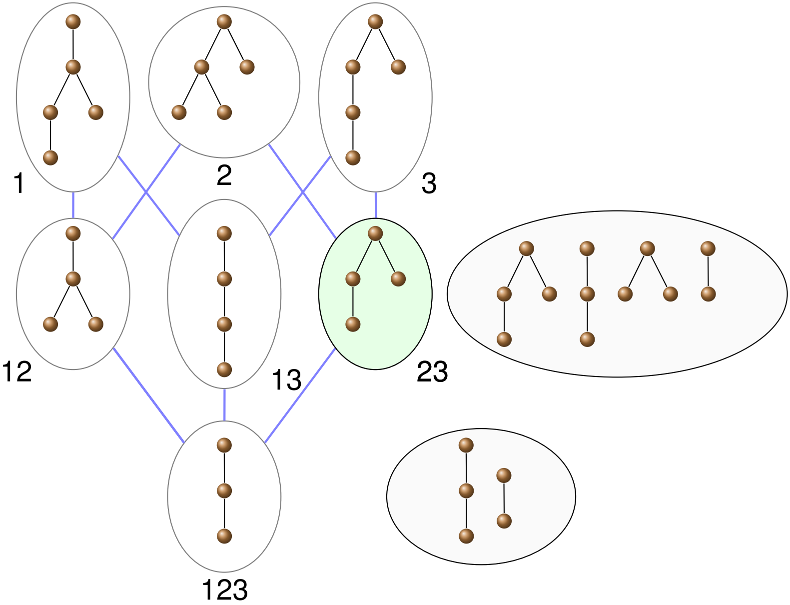
The following propositional formulas are logically equivalent:

- the conjunction of all the Horn formulas that are satisfied by all the models m_t for $t \in \mathscr{D}$
- the conjunction of $\mathscr{R}_0$ and all the propositional translations of the formulas in $\mathscr{R}'_{\mathscr{D}}$

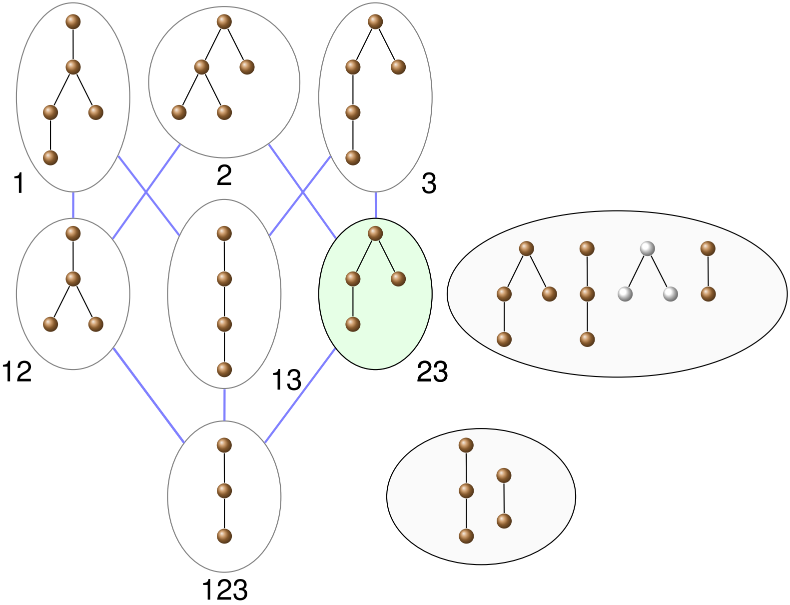
$$\mathscr{R}'_{\mathscr{D}} = \bigcup_{\mathscr{C}} \{A \rightarrow t \mid \Gamma_{\mathscr{D}}(A) = \mathscr{C}, t \in \mathscr{C}\}$$

- the conjunction of $\mathscr{R}_0$ and all the propositional translations of the formulas in a subset of $\mathscr{R}'_{\mathscr{D}}$ obtained transversing the hypergraph of differences between the nodes of the lattice.

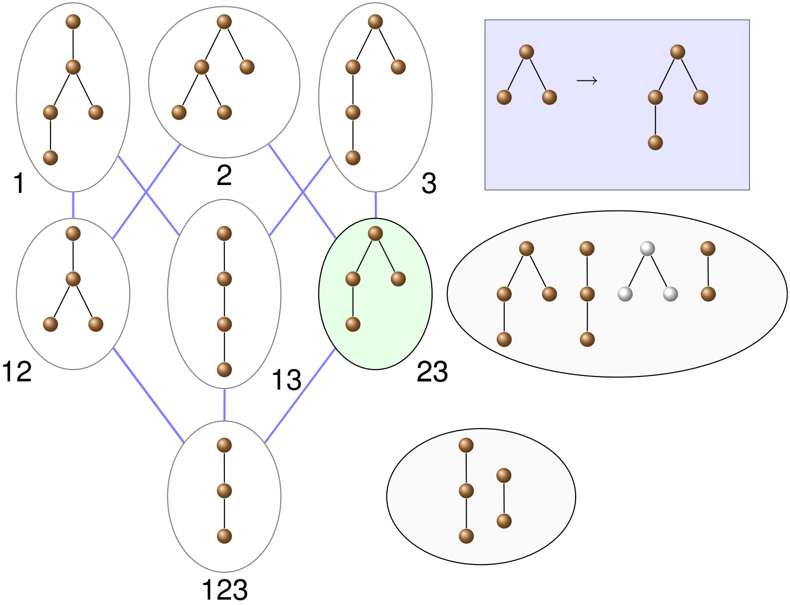
Association Rule Computation Example



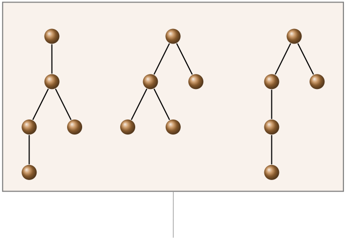
Association Rule Computation Example



Association Rule Computation Example

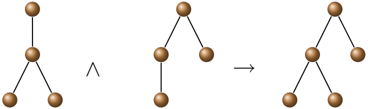


Implicit rules



$\mathscr{D}$

Implicit Rule

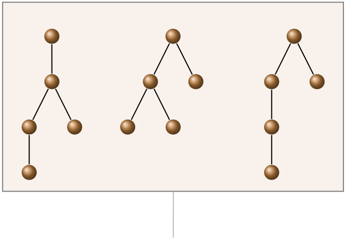


Given three trees t_1, t_2, t_3, we say that $t_1 \wedge t_2 \rightarrow t_3$, is an *implicit Horn rule* (abbreviately, an *implicit rule*) if for every tree t it holds

$$t_1 \preceq t \;\wedge\; t_2 \preceq t \;\leftrightarrow\; t_3 \preceq t.$$

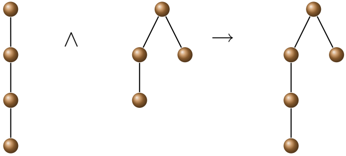
t_1 and t_2 have implicit rules if $t_1 \wedge t_2 \rightarrow t$ is an implicit rule for some t.

Implicit rules



$\mathscr{D}$

Implicit Rule

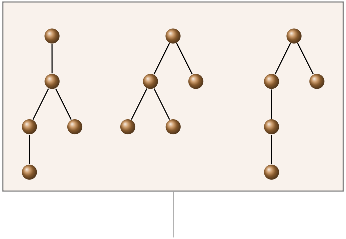


Given three trees t_1, t_2, t_3, we say that $t_1 \wedge t_2 \rightarrow t_3$, is an *implicit Horn rule* (abbreviately, an *implicit rule*) if for every tree t it holds

$$t_1 \preceq t \ \wedge \ t_2 \preceq t \ \leftrightarrow \ t_3 \preceq t.$$

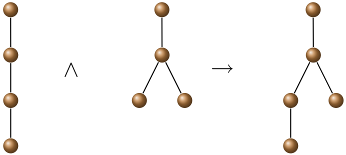
t_1 and t_2 have implicit rules if $t_1 \wedge t_2 \rightarrow t$ is an implicit rule for some t.

Implicit rules



$\mathscr{D}$

NOT Implicit Rule

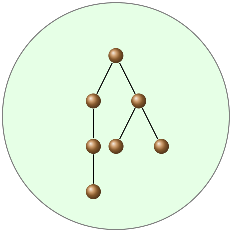


Given three trees t_1, t_2, t_3, we say that $t_1 \wedge t_2 \rightarrow t_3$, is an *implicit Horn rule* (abbreviately, an *implicit rule*) if for every tree t it holds

$$t_1 \preceq t \;\wedge\; t_2 \preceq t \;\leftrightarrow\; t_3 \preceq t.$$

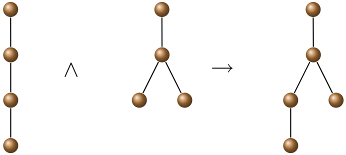
t_1 and t_2 have implicit rules if $t_1 \wedge t_2 \rightarrow t$ is an implicit rule for some t.

Implicit rules

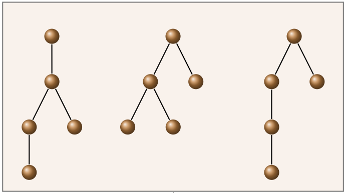


This supertree of the antecedents is **NOT** a supertree of the consequents.

NOT Implicit Rule

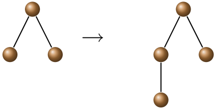


Implicit rules



$\mathscr{D}$

NOT Implicit Rule



Given three trees t_1, t_2, t_3, we say that $t_1 \wedge t_2 \rightarrow t_3$, is an *implicit Horn rule* (abbreviately, an *implicit rule*) if for every tree t it holds

$$t_1 \preceq t \ \wedge \ t_2 \preceq t \ \leftrightarrow \ t_3 \preceq t.$$

t_1 and t_2 have implicit rules if $t_1 \wedge t_2 \rightarrow t$ is an implicit rule for some t.

Implicit Rules

Theorem

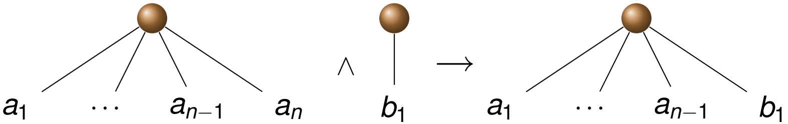
All trees a, b such that $a \preceq b$ have implicit rules.

Theorem

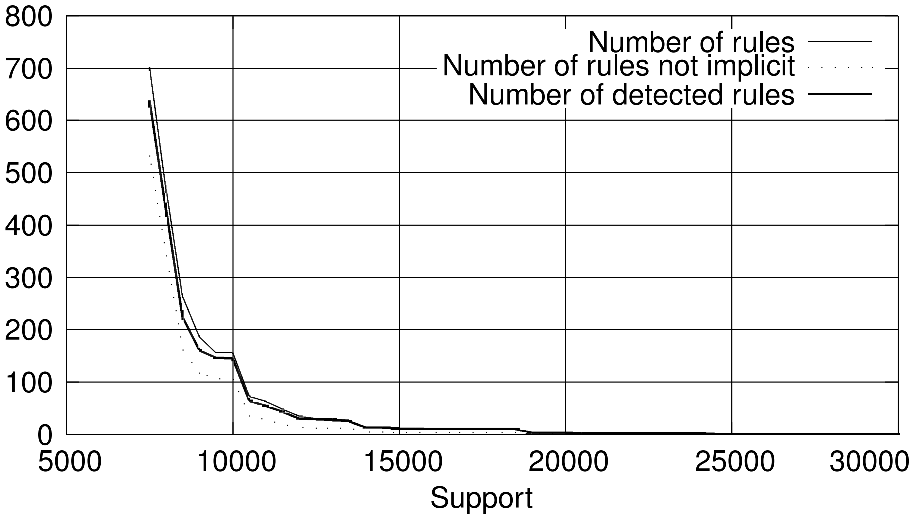
Suppose that b has only one component. Then they have implicit rules if and only if a has a maximum component which is a subtree of the component of b.

- for all $i < n$

$$a_i \preceq a_n \preceq b_1$$



Experimental Validation: CSLOGS



Summary

Conclusions

- A way of extracting high-confidence association rules from datasets consisting of unlabeled trees
  - antecedents are obtained through a computation akin to a hypergraph transversal
  - consequents follow from an application of the closure operators
- Detection of some cases of **implicit rules**: rules that always hold, independently of the dataset