

# Deduction Schemes for Association Rules

José L. Balcázar

2008: Technical University of Catalonia – 2009: University of Cantabria

Discovery Science – Budapest 2008

# Association Mining

## Key Notion in Data Mining

One very popular instance of **pattern elicitation** from data.

- ▶ Chosen notion of **pattern** is a sort of **implication**.
- ▶ Implications are the motivating concept of the study of Logic along 2350 years.

# Association Mining

## Key Notion in Data Mining

One very popular instance of **pattern elicitation** from data.

- ▶ Chosen notion of **pattern** is a sort of **implication**.
- ▶ Implications are the motivating concept of the study of Logic along 2350 years.
- ▶ Implications seem to help to suggest hopefully useful further actions, to be developed by some end user.
- ▶ In fact, their very syntax embodies a suggestion of a cause-effect relationship.

# Association Mining

## Key Notion in Data Mining

One very popular instance of **pattern elicitation** from data.

- ▶ Chosen notion of **pattern** is a sort of **implication**.
- ▶ Implications are the motivating concept of the study of Logic along 2350 years.
- ▶ Implications seem to help to suggest hopefully useful further actions, to be developed by some end user.
- ▶ In fact, their very syntax embodies a suggestion of a cause-effect relationship.
- ▶ **But:**

“5.1361 Der Glaube an den Kausalnexus ist der Aberglaube”  
(LW, T L-Ph)

# Association Mining

## Key Notion in Data Mining

One very popular instance of **pattern elicitation** from data.

- ▶ Chosen notion of **pattern** is a sort of **implication**.
- ▶ Implications are the motivating concept of the study of Logic along 2350 years.
- ▶ Implications seem to help to suggest hopefully useful further actions, to be developed by some end user.
- ▶ In fact, their very syntax embodies a suggestion of a cause-effect relationship.
- ▶ **But:**
  - “5.1361 Der Glaube an den Kausalnexus ist der Aberglaube”  
(LW, T L-Ph)
- ▶ Although Science is full of advances arising from the progress in understanding causal nexus.

# Implications, I

## A Variant of Definite Horn Formulas

### Alternative Notational Approaches:

Items and “transactions” (or rows) of a relational database table, propositional variables and models, attributes and objects in Formal Concept Analysis...

- ▶ Propositional variables, items: Boolean-valued.
- ▶ Models, binary strings, transactions: provide a Boolean value for each variable or item.
- ▶ Definite **Horn** Clause: one single positive disjunct, like  $\neg a \vee \neg b \vee c$ ; equivalent form as implication, like  $a \wedge b \Rightarrow c$ .
- ▶ Definite Horn Formula: conjunction of Definite Horn Clauses.
- ▶ Implication, paraclause: conjunction of Definite Horn Clauses with the same antecedent:  $(\neg a \vee \neg b \vee c) \wedge (\neg a \vee \neg b \vee d)$ , that is,  $a \wedge b \Rightarrow c \wedge d$ ; we write  $ab \Rightarrow cd$ .

# Implications, II

A.k.a. Deterministic Association Rules

## Main Property

A dataset can be axiomatized by a Horn Formula if and only if it is closed under bitwise conjunction.

# Implications, II

A.k.a. Deterministic Association Rules

## Main Property

A dataset can be axiomatized by a Horn Formula if and only if it is closed under bitwise conjunction.

Examples from a “ML abstracts” dataset:

carlo  $\implies$  monte

monte  $\implies$  carlo

descent  $\implies$  gradient

hilbert  $\implies$  space

margin support  $\implies$  vector

# Implications, II

A.k.a. Deterministic Association Rules

## Main Property

A dataset can be axiomatized by a Horn Formula if and only if it is closed under bitwise conjunction.

## Examples from a “ML abstracts” dataset:

carlo  $\implies$  monte

monte  $\implies$  carlo

descent  $\implies$  gradient

hilbert  $\implies$  space

margin support  $\implies$  vector

## Proper Implications or Softer Constraints?

It is unclear that proper implications are **really** very useful on large, real-life datasets.

# Implications, III

## Some Important Advantages of Implications

But, in fact, implications are rather well understood:

- ▶ clear, crisp notion of redundancy, namely **entailment**;
- ▶ algorithms to construct canonical axiomatizations (also called **bases**) of provably minimum size  
(for us today, “size” is number of rules);
- ▶ deductive **calculus**: the three **Armstrong schemes**,
  - ▶ **Reflexivity**: if  $Y \subseteq X$  then  $X \implies Y$ ,
  - ▶ **Augmentation**: if  $X \implies Z$  and  $Y \implies W$  then  $XY \implies ZW$ ,
  - ▶ **Transitivity**: if  $X \implies Y$  and  $Y \implies Z$  then  $X \implies Z$ .

Armstrong schemes are sound and complete:

An implication  $X \implies Y$  can be obtained from a set  $\mathcal{R}$  of implications by means of the Armstrong schemes **if and only if every dataset** in which the rules of  $\mathcal{R}$  hold satisfies as well  $X \implies Y$ .

# From Implications to Associations

## Standard Confidence-and-Support Setting

### Association Rules:

We aim at extending these advantages of implications into the usual notion of association rules based on **confidence** and **support** thresholds:

- ▶  $s(X \rightarrow Y)$  (support of the rule): number (or ratio) of transactions including the union  $XY$  of  $X$  and  $Y$ ;
- ▶  $c(X \rightarrow Y)$  (confidence of the rule): ratio of transactions that include  $X$  and  $Y$  to those that include  $X$ :  $s(XY)/s(X)$ .

# From Implications to Associations

## Standard Confidence-and-Support Setting

### Association Rules:

We aim at extending these advantages of implications into the usual notion of association rules based on **confidence** and **support** thresholds:

- ▶  $s(X \rightarrow Y)$  (support of the rule): number (or ratio) of transactions including the union  $XY$  of  $X$  and  $Y$ ;
- ▶  $c(X \rightarrow Y)$  (confidence of the rule): ratio of transactions that include  $X$  and  $Y$  to those that include  $X$ :  $s(XY)/s(X)$ .

### Armstrong Schemes do not hold anymore:

For a given, fixed, confidence threshold, association rules...

- ▶ obey Reflexivity,
- ▶ but do not obey Augmentation,
- ▶ nor Transitivity.

# Main Results, I

## Variants of Reflexivity and Augmentation

**Known:** Association rules do obey **variants** of these schemes:

Reflexivity (slightly generalized):

(R) If  $X \rightarrow Y$  and  $Z \subseteq Y$  then  $X \rightarrow Z$ .

# Main Results, I

## Variants of Reflexivity and Augmentation

**Known:** Association rules do obey **variants** of these schemes:

Reflexivity (slightly generalized):

(R) If  $X \rightarrow Y$  and  $Z \subseteq Y$  then  $X \rightarrow Z$ .

Right-hand-side Augmentation:

(rA) If  $X \rightarrow Y$  then  $X \rightarrow XY$ .

# Main Results, I

## Variants of Reflexivity and Augmentation

**Known:** Association rules do obey **variants** of these schemes:

Reflexivity (slightly generalized):

(R) If  $X \rightarrow Y$  and  $Z \subseteq Y$  then  $X \rightarrow Z$ .

Right-hand-side Augmentation:

(rA) If  $X \rightarrow Y$  then  $X \rightarrow XY$ .

Left-hand-side Augmentation:

(lA) If  $X \rightarrow YZ$  then  $XY \rightarrow Z$  (also  $XY \rightarrow YZ$ ).

# Main Results, I

## Variants of Reflexivity and Augmentation

**Known:** Association rules do obey **variants** of these schemes:

Reflexivity (slightly generalized):

(R) If  $X \rightarrow Y$  and  $Z \subseteq Y$  then  $X \rightarrow Z$ .

Right-hand-side Augmentation:

(rA) If  $X \rightarrow Y$  then  $X \rightarrow XY$ .

Left-hand-side Augmentation:

(lA) If  $X \rightarrow YZ$  then  $XY \rightarrow Z$  (also  $XY \rightarrow YZ$ ).

Soundness and Completeness:

Association rule  $X \rightarrow Y$  can be derived from rule  $X' \rightarrow Y'$  using these deduction schemes **if and only if**  $c(X \rightarrow Y) \geq c(X' \rightarrow Y')$ , **whatever the dataset**.

# Redundancy, I

## Alternative Notions

### Pick Your Choice:

Association rule  $X \rightarrow Y$  is redundant with respect to  $X' \rightarrow Y'$  if:

- ▶  $c(X \rightarrow Y) \geq c(X' \rightarrow Y')$  and  $s(X \rightarrow Y) \geq s(X' \rightarrow Y')$ , whatever the dataset (Aggarwal and Yu);
- ▶  $X' \rightarrow Y'$  **covers**  $X \rightarrow Y$ , in the following sense:  $X' \subseteq X$  and  $XY \subseteq X'Y'$  (Krzyszczewicz);
- ▶ (Aggarwal and Yu, essentially)
  - ▶ either  $X \subseteq X'$  and  $XY \subset X'Y'$ , e.g.  $A \rightarrow B$  versus  $A \rightarrow BC$  (Reflexivity or **strict redundancy**),
  - ▶ or  $X' \subset X$  and  $XY = X'Y'$ , e.g.  $A \rightarrow BC$  versus  $AB \rightarrow C$  (Left-Augmentation or **simple redundancy**),
  - ▶ or  $X = X'$  and  $XY = X'Y'$ , e.g.  $A \rightarrow B$  versus  $A \rightarrow AB$  (Right-Augmentation).

# Redundancy, I

## Alternative Notions

### Pick Your Choice:

Association rule  $X \rightarrow Y$  is redundant with respect to  $X' \rightarrow Y'$  if:

- ▶  $c(X \rightarrow Y) \geq c(X' \rightarrow Y')$  and  $s(X \rightarrow Y) \geq s(X' \rightarrow Y')$ , whatever the dataset (Aggarwal and Yu);
- ▶  $X' \rightarrow Y'$  **covers**  $X \rightarrow Y$ , in the following sense:  $X' \subseteq X$  and  $XY \subseteq X'Y'$  (Krzyszczewicz);
- ▶ (Aggarwal and Yu, essentially)
  - ▶ either  $X \subseteq X'$  and  $XY \subset X'Y'$ , e.g.  $A \rightarrow B$  versus  $A \rightarrow BC$  (Reflexivity or **strict redundancy**),
  - ▶ or  $X' \subset X$  and  $XY = X'Y'$ , e.g.  $A \rightarrow BC$  versus  $AB \rightarrow C$  (Left-Augmentation or **simple redundancy**),
  - ▶ or  $X = X'$  and  $XY = X'Y'$ , e.g.  $A \rightarrow B$  versus  $A \rightarrow AB$  (Right-Augmentation).

All these notions are **equivalent** to ours!

# Redundancy, II

## The Closure-Based Approach

### How useful is this redundancy notion?

- ▶ It is known how to construct **bases** (or: axiomatizations) of absolutely optimum size for this redundancy.
- ▶ Desirable smaller bases would need a stronger redundancy:  
Fix the set of implications  $\mathcal{B}_0$ ; we use them through their associated closure operator  $cl$ .

# Redundancy, II

## The Closure-Based Approach

### How useful is this redundancy notion?

- ▶ It is known how to construct **bases** (or: axiomatizations) of absolutely optimum size for this redundancy.
- ▶ Desirable smaller bases would need a stronger redundancy:  
Fix the set of implications  $\mathcal{B}_0$ ; we use them through their associated closure operator  $cl$ .

### Prior (Recent) Results:

- ▶ There are two **equivalent** extensions of the notion of redundancy in the presence of a closure operator:
  - ▶  $c(X \rightarrow Y) \geq c(X' \rightarrow Y')$  in all datasets specified by  $\mathcal{B}_0$ ;
  - ▶  $X' \subseteq cl(X)$  and  $XY \subseteq cl(X'Y')$ .  
(These conditions generalize some previous proposals.)
- ▶  $\mathcal{B}^*$  basis: of absolutely optimum size for this redundancy.  
(Needs to be combined with a basis for the implications.)

# Main Results, II

## Variants of Augmentation and Transitivity

Association rules do not obey Transitivity, but do obey composition with a **full implication** in a “Transitivity-like” form.

Right-hand-side Augmentation:

(rA) If  $X \rightarrow Y$  and  $X \implies Z$  then  $X \rightarrow YZ$ .

# Main Results, II

## Variants of Augmentation and Transitivity

Association rules do not obey Transitivity, but do obey composition with a **full implication** in a “Transitivity-like” form.

### Right-hand-side Augmentation:

(*rA*) If  $X \rightarrow Y$  and  $X \implies Z$  then  $X \rightarrow YZ$ .

### Left-hand-side Augmentation:

(*lA*) If  $X \rightarrow YZ$  then  $XY \rightarrow Z$  (also  $XY \rightarrow YZ$ ).

# Main Results, II

## Variants of Augmentation and Transitivity

Association rules do not obey Transitivity, but do obey composition with a **full implication** in a “Transitivity-like” form.

### Right-hand-side Augmentation:

(*rA*) If  $X \rightarrow Y$  and  $X \implies Z$  then  $X \rightarrow YZ$ .

### Left-hand-side Augmentation:

(*lA*) If  $X \rightarrow YZ$  then  $XY \rightarrow Z$  (also  $XY \rightarrow YZ$ ).

### Right-hand-side composition with an Implication:

(*rl*) If  $X \rightarrow Y$  and  $Y \implies Z$  then  $X \rightarrow Z$ .

(Has Reflexivity as particular case.)

# Main Results, II

## Variants of Augmentation and Transitivity

Association rules do not obey Transitivity, but do obey composition with a **full implication** in a “Transitivity-like” form.

### Right-hand-side Augmentation:

(*rA*) If  $X \rightarrow Y$  and  $X \implies Z$  then  $X \rightarrow YZ$ .

### Left-hand-side Augmentation:

(*lA*) If  $X \rightarrow YZ$  then  $XY \rightarrow Z$  (also  $XY \rightarrow YZ$ ).

### Right-hand-side composition with an Implication:

(*rl*) If  $X \rightarrow Y$  and  $Y \implies Z$  then  $X \rightarrow Z$ .

(Has Reflexivity as particular case.)

### Left-hand-side composition with an Implication:

(*lI*) If  $X \rightarrow Y$  and  $Z \subseteq X$  and  $Z \implies X$  then  $Z \rightarrow Y$ .

# Main Results, III

## Soundness and Completeness

Association rule  $X \rightarrow Y$  can be derived from rule  $X' \rightarrow Y'$  using these four deduction schemes ( $rA$ ), ( $rl$ ), ( $lA$ ), and ( $ll$ ) **if and only if**  $c(X \rightarrow Y) \geq c(X' \rightarrow Y')$  in all datasets specified by  $\mathcal{B}_0$ .

That is:

This simple deductive calculus is, again, **equivalent** to the notion of closure-based redundancy.

# Redundancy, III

## Towards a General Notion

Fix a confidence threshold and consider rules with respect to it.  
Do the following rules raise any suspicion of peculiarities?

- ▶  $A \rightarrow BC$
- ▶  $A \rightarrow BD$
- ▶  $ACD \rightarrow B$

# Redundancy, III

## Towards a General Notion

Fix a confidence threshold and consider rules with respect to it.  
Do the following rules raise any suspicion of peculiarities?

- ▶  $A \rightarrow BC$
- ▶  $A \rightarrow BD$
- ▶  $ACD \rightarrow B$

The first two actually entail the third one!

### Entailment From Two Premises:

Given full implications  $\mathcal{B}_0$ , rules  $X_1 \rightarrow Y_1$  and  $X_2 \rightarrow Y_2$  entail  $X_0 \rightarrow Y_0$  if, for any confidence threshold  $\gamma$ , **every dataset** that satisfies  $\mathcal{B}_0$  and assigns confidence at least  $\gamma$  to both rules  $X_1 \rightarrow Y_1$  and  $X_2 \rightarrow Y_2$  necessarily assigns confidence at least  $\gamma$  to  $X_0 \rightarrow Y_0$  also.

# Main Results, IV

## Exact Characterization of Entailment From Two Premises

Given full implications  $\mathcal{B}_0$  with corresponding closure operator  $cl$ :

$X_1 \rightarrow Y_1$  and  $X_2 \rightarrow Y_2$  entail  $X_0 \rightarrow Y_0$  if and only if either

1.  $Y_0 \subseteq cl(X_0)$ , or
2.  $c(X_0 \rightarrow Y_0) \geq c(X_1 \rightarrow Y_1)$  in all datasets specified by  $\mathcal{B}_0$ , or
3.  $c(X_0 \rightarrow Y_0) \geq c(X_2 \rightarrow Y_2)$  in all datasets specified by  $\mathcal{B}_0$ , or
4. the following seven conditions hold simultaneously:
  - 4.1  $X_1 \subseteq cl(X_0)$
  - 4.2  $X_2 \subseteq cl(X_0)$
  - 4.3  $X_1 \subseteq cl(X_2 Y_2)$
  - 4.4  $X_2 \subseteq cl(X_1 Y_1)$
  - 4.5  $X_0 \subseteq cl(X_1 Y_1 X_2 Y_2)$
  - 4.6  $Y_0 \subseteq cl(X_0 Y_1)$
  - 4.7  $Y_0 \subseteq cl(X_0 Y_2)$

# Main Results, IV

## Exact Characterization of Entailment From Two Premises

Given full implications  $\mathcal{B}_0$  with corresponding closure operator  $cl$ :

$X_1 \rightarrow Y_1$  and  $X_2 \rightarrow Y_2$  entail  $X_0 \rightarrow Y_0$  if and only if either

1.  $Y_0 \subseteq cl(X_0)$ , or
2.  $c(X_0 \rightarrow Y_0) \geq c(X_1 \rightarrow Y_1)$  in all datasets specified by  $\mathcal{B}_0$ , or
3.  $c(X_0 \rightarrow Y_0) \geq c(X_2 \rightarrow Y_2)$  in all datasets specified by  $\mathcal{B}_0$ , or
4. the following seven conditions hold simultaneously:
  - 4.1  $X_1 \subseteq cl(X_0)$
  - 4.2  $X_2 \subseteq cl(X_0)$
  - 4.3  $X_1 \subseteq cl(X_2 Y_2)$
  - 4.4  $X_2 \subseteq cl(X_1 Y_1)$
  - 4.5  $X_0 \subseteq cl(X_1 Y_1 X_2 Y_2)$
  - 4.6  $Y_0 \subseteq cl(X_0 Y_1)$
  - 4.7  $Y_0 \subseteq cl(X_0 Y_2)$

Deduction scheme form also available.

# Conclusions

We begin to understand the Logic of Association Rules:

- ▶ Recent:
  - ▶ a solid preliminary notion of redundancy as entailment;
  - ▶ a way of selecting bases of optimum size;
  - ▶ both can be refined to take into account closures.
- ▶ New contributions:
  - ▶ deductive calculi corresponding to these notions of entailment;
  - ▶ exact characterization of entailment from two partial rules as premises (optionally: with closures).
- ▶ Forthcoming:
  - ▶ more careful consideration of support bounds;
  - ▶ (improved) implementation of the current basis computation;
  - ▶ work towards full entailment with several partial rules as premises: potentially improved bases.