

Minimum Size Bases for Association Rules

José L. Balcázar

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

ECML PKDD – Antwerp 2008

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:

- ▶ they offer a clear, crisp notion of redundancy, namely **entailment**;
- ▶ we have algorithms to construct canonical axiomatizations (also called **bases**) of provably minimum size
(for us today, “size” is number of rules);
- ▶ we have a sound and complete syntactic, deductive **calculus** of entailment, namely 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 $X \implies Z$.

Association rules obey Reflexivity, but not Augmentation or Transitivity.

Entailment, I

From Implications to Associations

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 $X \implies Y$.

Entailment, I

From Implications to Associations

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 $X \implies Y$.

Standard Association Rules:

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

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

Redundancy, I

Existing Proposals for Association Rules

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

Redundancy, II

Main Results

All these notions are **equivalent!**

(We have proved also that there exists as well a syntactic deductive calculus that is sound and complete for the notion of entailment given by this redundancy.)

Redundancy, II

Main Results

All these notions are **equivalent!**

(We have proved also that there exists as well a syntactic deductive calculus that is sound and complete for the notion of entailment given by this redundancy.)

Rule basis:

- ▶ A **basis** is an irredundant set of rules that makes all the others redundant with respect to them.
- ▶ A proposal appearing in several previous references: $X \rightarrow Y$ is kept if its confidence is over the threshold, **but** properly smaller X or properly larger Y would drop below the threshold.

Redundancy, II

Main Results

All these notions are **equivalent!**

(We have proved also that there exists as well a syntactic deductive calculus that is sound and complete for the notion of entailment given by this redundancy.)

Rule basis:

- ▶ A **basis** is an irredundant set of rules that makes all the others redundant with respect to them.
- ▶ A proposal appearing in several previous references: $X \rightarrow Y$ is kept if its confidence is over the threshold, **but** properly smaller X or properly larger Y would drop below the threshold.

New Result: Absolute Optimality!

Every basis for this notion of redundancy has at least as many rules as this one.

Redundancy, III

The Closure-Based Approach

Further Reduction?

- ▶ Certainly desirable (after just a little testing).
- ▶ Needs a stronger redundancy notion.
- ▶ How to combine rules? **Key:** Transitive composition of an association rule with an implication maintains the confidence.
- ▶ Fix the set of implications \mathcal{B}_0 ; we use them through their associated closure operator cl .
- ▶ Two extensions of the previous proposals:
 - ▶ $c(X \rightarrow Y) \geq c(X' \rightarrow Y')$ in all datasets specified by \mathcal{B}_0 ;
 - ▶ $X' \subseteq cl(X)$ and $XY \subseteq cl(X'Y')$.

Redundancy, III

The Closure-Based Approach

Further Reduction?

- ▶ Certainly desirable (after just a little testing).
- ▶ Needs a stronger redundancy notion.
- ▶ How to combine rules? **Key:** Transitive composition of an association rule with an implication maintains the confidence.
- ▶ Fix the set of implications \mathcal{B}_0 ; we use them through their associated closure operator cl .
- ▶ Two extensions of the previous proposals:
 - ▶ $c(X \rightarrow Y) \geq c(X' \rightarrow Y')$ in all datasets specified by \mathcal{B}_0 ;
 - ▶ $X' \subseteq cl(X)$ and $XY \subseteq cl(X'Y')$.

Equivalent!

(And again we have a sound and complete deductive calculus.)

Redundancy, IV

The Minimum-Size Basis

Quite similar (but not identical) to the Zaki/Pasquier basis.
Generators can be substituted for closures with no essential change.

Basis \mathcal{B}^* :

$X \rightarrow Y$ is kept if its confidence is over the threshold, **but** properly smaller X or properly larger Y would drop below the threshold, **and this, only for closed** X and Y .

Redundancy, IV

The Minimum-Size Basis

Quite similar (but not identical) to the Zaki/Pasquier basis.
Generators can be substituted for closures with no essential change.

Basis \mathcal{B}^* :

$X \rightarrow Y$ is kept if its confidence is over the threshold, **but** properly smaller X or properly larger Y would drop below the threshold, **and this, only for closed** X and Y .

Absolute Optimum:

Every basis for this second notion of redundancy has at least as many rules as this one.

Note: Requires to be combined with a basis for the implications.
But this is well-studied (we use the GD basis here).

Empirical Comparisons

Examples of Minimum-Size Basis

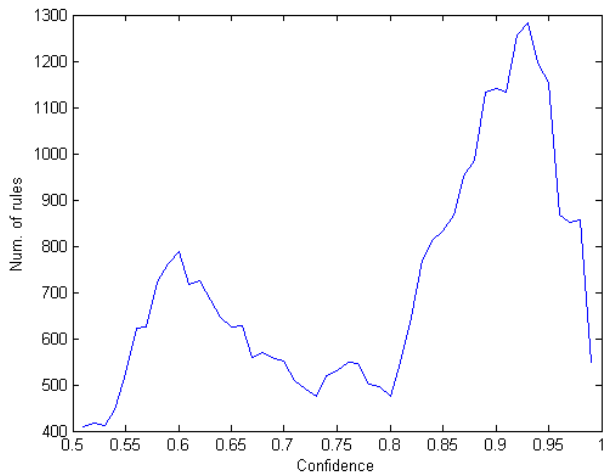
Dataset	Supp/Conf	Traditional	Closure-based	GD basis	B_γ^* basis	Total
Chess	80	552564	27711	5	226	231
Chess	70	8171198	152074	10	891	901
Connect	97	8092	1116	4	41	45
Connect	90	3640704	18848	14	222	236
Mushroom	40	7020	475	24	41	65
Mushroom	20	19191656	5741	170	158	328
Pumsb	95	1170	267	2	32	34
Pumsb	85	1408950	44483	9	1080	1089
Pumsb_star	60	2358	192	5	6	11
Pumsb_star	40	5659536	13479	47	82	129
T10I4D100K	0.5	2216	1231	0	585	585
T10I4D100K	0.1	431838	86902	214	4054	4268

Facts Worth Further Study

Distinguished Confidence Thresholds

The basis size is **not** monotonic!

Dataset **pumsb-star**, mined at support 20%: number of basis rules plotted against the confidence threshold:



Entailment, II

Towards a General Notion?

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

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

Entailment, II

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:

Full, exact, and somewhat complex characterization in terms of seven set-theoretic conditions generalizing the example above.

Conclusions

- ▶ We begin to understand the Logic of Association Rules:
 - ▶ 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;
 - ▶ the size of the optimal basis may show local minima.
- ▶ Forthcoming:
 - ▶ deductive calculi corresponding to these notions of entailment;
 - ▶ more careful consideration of support bounds;
 - ▶ work towards full entailment with several partial rules as premises: but the case of two partial premises is already near the current limit of human understanding.