

Characterizing Implications of Injective Partial Orders*

José L. Balcázar¹ and Gemma C. Garriga²

¹ Laboratori d'Algorísmica Relacional, Complexitat i Aprenentatge

Universitat Politècnica de Catalunya, Barcelona

balqui@lsi.upc.edu

² HIIT Basic Research Unit

Helsinki University of Technology

gemma.garriga@hut.fi

Abstract. It is known that implications in powerset-based closure systems correspond to Horn approximations in propositional logic frameworks. Here we focus on the problem of implications between injective partial orders. We set up the definitions that allow one to apply standard constructions of implications, and formally characterize the propositional theory obtained. We describe also some experimental applications of our development.

1 Introduction

One popular data representation formalism is given by binary relations, through which “objects”, or “models” or “transactions”, are described in terms of “attributes”, or “variables” or “items”. The concrete terms vary in function of the research community (concept lattices, propositional logic, relational databases...), and the standard notation changes accordingly [1,3]. A well-understood process of data analysis on these structures consists in finding pairs of sets of attributes for which the given data suggest some form of causality: each pair, frequently denoted $A \rightarrow B$, obtained along the data analysis process indicates that, due to the phenomenon where the data comes from, whenever a data object has all the attributes of A , it tends to have as well those of B . However, datasets available for data mining have, in general, no guarantee at all of being a correct sample of any phenomenon. Thus, the validity of such association rules is not to be taken for granted.

When the strength of the correlation is full, that is, all objects having A also have B , the rules obtained are Horn expressions, or implications, or deterministic association rules (depending again on the community), and the validity of the process is characterized by the question of whether the phenomenon at hand does allow a Horn axiomatization ([5,7]). In case it does not, it is known that the rules obtained correspond to a “best” Horn theory in a precise sense (the empirical Horn approximation). Additional parameters have been introduced for measuring the strength of the implication for other rules ([6]) or to focus on the rules holding for a certain support, that is, large enough ratio of the objects ([1]): we will do this last pruning in our empirical validations too.

* This work is supported in part by MCYT TIC (MOISES-TA TIN2005-08832-C03,Trangram TIC2004-07925-C03-02) and by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778.

Many data mining tasks proceed on the basis of structured data, instead of mere relational tuples. Here we study the extension of deterministic association rules to the same sort of data as in [4], and beyond it into partially ordered data.

2 Partial Orders and Implications

Our partial orders are formalized by labeled directed acyclic graphs. We assume that the labeling is injective; that is, the graph representing the partial order has no repeated labels. Formally, we fix our infinite set of labels L , and define our partial orders simply as pairs (V, E) where $V \subset L$ is the finite set of labeled vertices and $E \subseteq V \times V$ is antisymmetric, thus representing the reflexive and transitive closure of E . The set of all these partial orders is \mathcal{H} . Morphisms in \mathcal{H} are defined in the standard way.

Definition 1. We say that G is **more general** than H (denoted by $G \preceq H$), if and only if there is a morphism from G to H . Then H is also said to be **more specific** than G . (These are slight language abuses in that “or equal to” is left implicit.)

Definition 2. $G \cap H$ is the partial order having as vertices the intersection of the vertex sets of G and H , and where (e, e') is an edge of the intersection if and only if it is so in both G and H .

This operation is associative and commutative, so that we can express intersections of several partial orders. This notion corresponds to a meet operation with respect to the \preceq ordering, so that in fact we obtain a lower semilattice. We will also add an artificial element corresponding to the intersection of an empty set of partial orders (which would not be a partial order under our definition); we denote it by the unsatisfiable boolean constant \square , and it is “maximally specific” by convention. Therefore we obtain a lattice.

2.1 A Closure Operator

The analysis we want to attempt is made on the basis of a dataset consisting of N partial orders, identified by consecutive natural numbers from the interval $[1..N]$: $\mathcal{D} = \{G_i \mid 1 \leq i \leq N\}$. They are not necessarily different. We define the following two *derivation operators*: $\phi : 2^{[1..N]} \mapsto \mathcal{H}$ and $\psi : \mathcal{H} \mapsto 2^{[1..N]}$, as follows: $\phi(I) = \bigcap \{G_i \mid i \in I\}$ whereas for any partial order H , $\psi(H) = \{i \mid H \preceq G_i\}$. It is easy to check that they fulfill the property of the Galois connections: $H \preceq \phi(I) \iff I \subseteq \psi(H)$.

Therefore (see for instance [3]), we obtain a closure operator $\Delta = \phi \cdot \psi$ on partial orders, depending on the actual dataset $\{G_i\}$. The closure operator yields, in a fully standard manner, a notion of implication:

Definition 3. An *implication on partial orders* is a pair (G, H) , denoted $G \vdash H$, such that $\Delta(G) = H$.

Now we wish to characterize precisely the rules in a purely propositional way, in terms of Horn clauses. We assign *one propositional variable* \widehat{e} to each label $e \in L$, and *one propositional variable* $\widehat{ee'}$ to each pair of labels $e, e' \in L$, or edge of our graphs; and we express the elementary information that we want them to represent, in the form of the following five *background Horn axioms* (more precisely, axiom schemes):

1. $\widehat{ee'} \rightarrow \widehat{e}$
2. $\widehat{ee'} \rightarrow \widehat{e'}$
3. $\widehat{e} \rightarrow \widehat{ee}$
4. $\widehat{ee'} \wedge \widehat{e'e} \rightarrow \square$ (for different e, e')
5. $\widehat{ee'} \wedge \widehat{e'e''} \rightarrow \widehat{ee''}$

Note that indeed these axioms are Horn clauses, where the antisymmetry property is nondefinite, and written in clausal form as $\neg\widehat{ee'} \vee \neg\widehat{e'e}$. This is important in that, if such “background knowledge” cannot be expressed in that form, the correspondence between implications and Horn expressions would not hold.

Now, each input partial order $G_i \in \mathcal{D}$ in the data corresponds to a model m_i in a natural way, and we consider the set $M = \{m_i \mid 1 \leq i \leq N\}$, where each m_i corresponds to $G_i \in \mathcal{D}$. Similarly, each rule $G \vdash H$ obtained through the closure operator can be seen as a propositional implication: $\widehat{G} \rightarrow \widehat{H}$.

Theorem 1. *Given a set of input partial orders \mathcal{D} , the conjunction of all the implications constructed by the closure system, seen as propositional formulas, and together with the background Horn axioms, axiomatizes exactly the empirical Horn approximation of the theory containing the set of models M .*

The proof of this result is similar in structure to the main result proved in [2].

In the case of relational data, there are some standard methods to construct a set of axioms (or basis): one is based on pseudo-intents [3], and yields a minimal basis. We prefer here the instantaneous basis from [8] and [7], which has other advantages: we find it particularly intuitive in explaining the data analysis processes.

Lemma 1. *Let H be a closed partial order and $G \preceq H$; then $\Delta(G) = H$ if and only if \widehat{G} intersects all the faces of H .*

We omit the proof. This fact allows us to reduce the problem of constructing a basis to a hypergraph transversal problem. Faces are defined as:

$$H - H' = \{\widehat{e} \mid e \in V - V'\} \cup \{\widehat{ee'} \mid e, e' \in V', (e, e') \in E - E'\}$$

Then, by intersecting a face $H - H'$ we understand the set-theoretic intersection, that is, there must be a common edge or vertex in both.

2.2 Empirical Validation

We have developed a prototype implementation, and applied it to several datasets coming from real life. We found that the running times were negligible in comparison with the computation of the lattice, with real-time responses even with quite large lattices.

In general, real-life datasets do not offer the injectiveness condition; however, this is no big inconvenient since the very proposal of searching for implications, and the proposal of using our particular approach as well, are only based on the heuristic perception that this sort of analysis can provide useful explanations.

One interesting application is the analysis of the curricula of specific students of the Computer Engineering School of our university. There, a large number of elective

courses, with precedence recommendations of varying strength, give many different trajectories; we used a dataset corresponding to the courses registered along ten years, all corresponding to the same three variants of the curriculum (now superseded by a new one). Each time instant corresponds to a semester, and each of the 8793 transactions corresponds to a student and includes, for a number of consecutive semesters, the courses enrolled in each one. We restrict our analysis to the electives since compulsory courses follow a pre-established track. We tried supports of 5% and 7.5% on this dataset; respectively, this gave 1689 and 585 closures, and still the number of rules, which could be huge from such a lattice, remained very manageable: 502 and 94, respectively. About one sixth of them were redundancies, such as repeated rules or transitivity. Many of the others were consequence of the precedence recommendations imposed by the School. Examples of nontrivial rules found are: if Database Design is followed by Organizational Structures, then the same student has taken Economy 2; or: each student who took Economy 1 and also took Files and Databases followed by Informatics Projects Management, also did Database Design before Informatics Projects Management.

An even more interesting dataset was obtained from the abstracts of all the 706 research reports filed into the Pascal Network of Excellence (pascal-network.org) up to a specific moment in time. Closed partial orders were computed at frequency thresholds of 10%, 5%, and 2.1%; in this last case, 954 closures were obtained, but still the number of rules was a very manageable total of 70, including still some redundancy. Rules appearing include facts such as “kernel” and “support” implies “support vector”; also, if “selection” appears and “feature” appears at least twice then “feature selection” appears, and about 20 similar other rules involving “model”, “error”, and others.

References

1. Agrawal, R., Imielinski, T., Swami, A.N.: Mining association rules between sets of items in large databases. In: Buneman, P., Jajodia, S. (eds.) Proceedings of the 1993 ACM SIGMOD Int. Conference on Management of Data, pp. 207–216. ACM Press, New York (1993)
2. Balcázar, J.L., Garriga, G.C.: On Horn axiomatizations for sequential data. *Theoretical Computer Science* 371(3), 247–264 (2007)
3. Ganter, B., Wille, R.: *Formal Concept Analysis. Mathematical Foundations*. Springer, Heidelberg (1998)
4. Garriga, G.C., Balcázar, J.L.: Coproduct transformations on lattices of closed partial orders. In: Proceedings of 2nd. Int. Conference on Graph Transformation, pp. 336–351 (2004)
5. Kautz, H., Kearns, M., Selman, B.: Horn approximations of empirical data. *Artificial Intelligence* 74(1), 129–145 (1995)
6. Luxenburger, M.: Implications partielles dans un contexte. *Math. Inf. Sci. hum.* 29(113), 35–55 (1991)
7. Pfaltz, J.L., Taylor, C.M.: Scientific knowledge discovery through iterative transformations of concept lattices. In: SIAM Int. Workshop on Discrete Mathematics and Data Mining, pp. 65–74 (2002)
8. Wild, M.: A theory of finite closure spaces based on implications. *Advances in Mathematics* 108, 118–139 (1994)