On Horn Axiomatizations for Sequential Data*

José L. Balcázar and Gemma Casas-Garriga

Departament de Llenguatges i Sistemes Informàtics,
Universitat Politècnica de Catalunya
{balqui, gcasas}@lsi.upc.es

Abstract. We propose a notion of deterministic association rules for ordered data. We prove that our proposed rules can be formally justified by a purely logical characterization, namely, a natural notion of empirical Horn approximation for ordered data which involves background Horn conditions; these ensure the consistency of the propositional theory obtained with the ordered context. The main proof resorts to a concept lattice model in the framework of Formal Concept Analysis, but adapted to ordered contexts. We also discuss a general method to mine these rules that can be easily incorporated into any algorithm for mining closed sequences, of which there are already some in the literature.

1 Introduction

According to a large number of sources, the field of Data Mining attempts at finding methods to extract from large masses of existing data, that was not gathered for that purpose, new, sound knowledge that allows to take actions with specific purposes. One natural way to interpret the last condition is to look for causal relationships, where the presence of some fact suggests that other facts follow from them. This is one of the reasons of the success of the association rules framework: in the presence of a community that tends to buy, say, sodas together with the less expensive spirits, a number of natural ideas to try to influence the behavior of the buyers and profit from the pattern easily come up.

However, association is not causality, even though it is frequently interpreted in that way (most of the times implicitly). As a token, one of the criticisms of the *lift* measure for the strength of association rules is its symmetry, which makes it impossible to "orient the rules", that is, disguise the association as causality. Along the same lines, criticisms of various sorts have been put forward for many other measures of the strength of implication such as confidence or correlation. The single case that would be beyond any such criticism is where the implication *always* holds. These cases have been named *deterministic association rules*, and are particularly interesting in domains coming from observations of scientific data, where underlying natural laws are actually causing the associations to appear in all cases [10].

An obvious criticism is that a single counterexample suffices to invalidate a deterministic association rule, and it could be due to data manipulation errors.

* This work is supported in part by MCYT TIC 2002-04019-C03-01 (MOISES).

T. Eiter and L. Libkin (Eds.): ICDT 2005, LNCS 3363, pp. 215–229, 2005.

However, this is not really an objection to the notion of deterministic association rules but simply a consideration that data cleaning techniques are necessary in any practical application of this notion; we come back to this point later on.

On the other hand, the central advantage of deterministic association rules is that they do not require to select, with little or no formal guidance, one single measure of strength of implication. Since they are pure standard implications, they can be studied in purely logical terms.

In fact, standard binary databases (as termed in data mining texts, even though they are rather just relations) of n attributes can be naturally viewed as sets of models (0/1 assignments to n propositional variables). Thus, from this perspective, association rules can be seen as propositional logic formulas capturing information contained in a set of models. Practically effective approaches to find such logical formulas have been proposed in the field of Knowledge Compilation ([3, 11]): among them, a prominent basic process is to "compile" the list of satisfying models, into a tractable set of Horn clauses ([8, 11]). Of course, it might happen that no Horn axiomatization exists for the given set of models; but then, a Horn approximation (the minimal Horn upper bound of the given theory, sometimes called the empirical Horn approximation) can always be computed.

In [2], the following is proved: if deterministic association rules are computed from data according to the published lattice-theoretic methods [9, 10, 14], the rules obtained axiomatize exactly the minimal Horn upper bound of the propositional theory given by the data. These lattice-theoretic methods are actually described in terms of *concept lattices* [7]; this framework allows also for the study of general association rules (see [15] and the references there) and functional dependencies (see e.g. [6]). Concept lattices are given by *closed* subsets of attributes and *closed* subsets of tuples, where all the tuples in a concept share the attributes of the same concept, and viceversa. The notion of closure can be defined in a number of equivalent ways.

However, mining closed sets of binary attributes is but the simplest closure-based data mining problem; our goal here is to extend these results into the case of ordered transactions [1]. In these applications, each input tuple no longer is a set of attributes, but rather a sequence of them. Standard examples, instead of typical market-basket data, are of a more structured sort, such as the sequence of actions on a single bank account. Recent work in [13] and [12] provides algorithmic solutions to discover closed sequential patterns, so that there exists indeed a notion of closure-based analysis for these sequences; but, so far, no notion of deterministic association rules for them. Our goal is to formulate a theory of associations for this ordered context, in such a way that

- it advances in the theory underlying the state of the art algorithms for closure-based analysis of sequences,
- it corresponds closely to the lattice-theoretic approach employed for the computation of deterministic association rules in the unordered case, and
- it allows for a precise logical characterization, similar in spirit to [2].

Our starting point is the model in [4], which formalizes a concept lattice of closed sets of sequences by means of a new Galois connection. Here, we con-

tribute with the proposal of notions of deterministic association rules for ordered contexts, and we validate formally the proposal by exhibiting a logical characterization of the deterministic association rules with order that parallels the existing one for unordered contexts. We also discuss the integration of the computation of these rules with existing algorithms to mine closed sequences.

2 Preliminaries

Let $\mathcal{I} = \{i_1, \ldots, i_n\}$ be a finite set of items. These will be our atomic objects. *Itemsets* are subsets $I_i \subseteq \mathcal{I}$. Since actually n is unbounded, we could alternatively have an infinite set of items from which, at every moment, only the finitely many ones appearing in a given dataset are relevant.

Sequences are ordered lists of itemsets. The set of all the possible sequences will be noted by $\mathcal{S}$. Here we are following the same framework for modeling sequences or temporal data tuples as in [1] or [13], whose closed sequential patterns (that will be later introduced) were formally characterized in our previous work [4], and which we seek in this paper to complement with adequate notions of association rules. Thus, our data consists of a database of ordered transactions that we model as a set of sequences, $\mathcal{D} = \{s_1, s_2, \ldots s_n\}$. Our notation for the component itemsets of a given sequence will be $s = \langle (I_1)(I_2)\ldots(I_n)\rangle$, meaning that itemset I_i occurs before itemset I_j for $i < j$.

An alternative view of our data, borrowed from Formal Concept Analysis, is in the form of an ordered context; *objects* of the context are sequences, *attributes* of the context are items, and the database becomes a ternary relation, subset of $\mathcal{O} \times \mathcal{I} \times \mathbb{N}$, in which each tuple $\langle o, i, t\rangle$ indicates that item i appears in the t-th element of the object o. A simple example of the described data and the associated context can be found in figure 1, where each object o_i of the formal context represents the corresponding input sequence (or ordered transaction) s_i. The context for a set of data $\mathcal{D}$ is relevant to this work to see objects $o_i \in \mathcal{O}$ and input sequences $s_i \in \mathcal{D}$ as equivalent.

Seq id	Sequence
s_1	$\langle (A)(B)(C)(D)\rangle$
s_2	$\langle (B)(C)(D)(A)\rangle$
s_3	$\langle (B)(C)(A)(D)\rangle$

(a) Collection of data $\mathcal{D}$

	A	B	C	D
o_1	1	2	3	4
o_2	4	1	2	3
o_3	3	1	2	4

(b) Context $\mathbb{K}$

Fig. 1. Example of ordered data $\mathcal{D}$ and its context $\mathbb{K}$

Sequence $s = \langle (I_1)\ldots(I_n)\rangle$ is a *subsequence* of sequence $s' = \langle (I'_1)\ldots(I'_m)\rangle$ if there exist integers $j_1 < j_2 \cdots < j_n$ such that $I_1 \subseteq I'_{j_1}, \ldots, I_n \subseteq I'_{j_n}$. We note this case by $s \subseteq s'$. For example, the sequence $\langle (A)(D)\rangle$ is a subsequence of the first and third sequences in figure 1.

The *intersection* of a set of sequences $s_1, \ldots, s_n \in \mathcal{S}$ is the set of maximal subsequences contained in all the s_i. Note that the intersection of a set of sequences, or even the intersection of two sequences, is not necessarily a single sequence. For example, the intersection of the two sequences $s = \langle (AD)(C)(B) \rangle$ and $s' = \langle (A)(B)(C) \rangle$ is the set of sequences $\{\langle (A)(C) \rangle, \langle (A)(B) \rangle\}$: both are contained in s and s', and among those having this property they are maximal; all other common subsequences are not maximal since they can be extended to one of these. The maximality condition discards redundant information since the presence of, e.g., $\langle (A)(B) \rangle$ in the intersection already informs of the presence of each of the itemsets (A) and (B).

We partially order also sets of sequences, as follows: $S \preceq S'$ if and only if $\forall s \in S \, \exists s' \in S' \, s \subseteq s'$.

2.1 Propositional Horn Logic

Assume a standard propositional logic language with propositional variables, noted by $\{v_i\}$. The number of variables is finite, and we note by $\mathcal{V}$ the set of all variables; but again, we could alternatively use an infinite set of variables provided that the propositional issues corresponding to a fixed dataset only involve finitely many of them (this is in fact the case of our application). A literal is either a propositional variable, called a positive literal, or its negation, called a negative literal. A clause is a disjunction of literals and can be seen simply as the set of the literals it contains. A clause is *Horn* if and only if it contains at most one positive literal. Horn clauses with a positive literal are called *definite*, and can be written as $H \rightarrow v$ where H is a conjunction of positive literals that were negative in the clause, whereas v is the single positive literal in the clause. Horn clauses without positive literals are called *nondefinite*, and can be written similarly as $H \rightarrow \Box$, where $\Box$ expresses unsatisfiability. A Horn formula is a conjunction of Horn clauses.

A *model* is a complete truth assignment, i.e. a mapping from the variables to $\{0, 1\}$. We note by $m(v)$ the value that the model m assigns to the variable v. The intersection of two models is the bitwise conjunction, returning another model. A model satisfies a formula if the formula evaluates to true in the model. The set of all models will be noted by $\mathcal{M}$.

A theory is a set of models. A theory is Horn if there is a Horn formula which axiomatizes it, in the sense that it is satisfied exactly by the models in the theory. When a theory contains another we say that the first is an upper bound for the second; for instance, by removing clauses from a Horn formula we get a larger or equal Horn theory. The following is known (see e.g. [8]):

Theorem 1. *Given a propositional theory T, there is exactly one minimal Horn theory containing it. Semantically, it contains all the models that are intersections of models of T. Syntactically, it can be described by the conjunction of all Horn clauses satisfied by all models from T.*

The theory obtained in this way is called sometimes the *empirical Horn approximation* of the original theory. Clearly, then, a theory T is Horn if and only

if it is actually *closed under intersection*, so that it coincides with its empirical Horn approximation. These concepts are a cornerstone of the area of research known as Knowledge Compilation [3].

2.2 Closures and Galois Connections

The framework introduced previously allows us to cast our reasoning in terms of closure operators. A closure operator Γ on a lattice, such as the one formed by the subsets of any fixed universe, is one that satisfies the three basic closure axioms: monotonicity, extensivity and idempotency. It follows from these properties that the intersection of closed sets is a closed set.

In the main case of interest for data mining, the universe will be our set of items $\mathcal{I}$. Then, closure operators give rise to closed sets of items, generators, and deterministic association rules. *Closed sets* are those sets of items that coincide with their closure, that is, $\Gamma(Z) = Z$ where $Z \subseteq \mathcal{I}$. When $\Gamma(G) = Z$ for a set G and G is minimal for that resulting Z, we say that G is a *generator* of Z. One way for constructing closure operators is by composition of two derivation operators forming a Galois connection [7]. Implications of the form $G \rightarrow Z$ where G is a generator of Z, turn out to be the particular case of association rules where no support condition is imposed but confidence is 1 (or 100%) [10], [9]. Such rules in this unordered context are sometimes called *deterministic association rules*.

It turns out that it is possible to exactly characterize this set of deterministic association rules in terms of propositional logic: we can associate a propositional variable to each item; then transactions become models, and each association rule becomes a conjunction of Horn clauses with the same left hand side. Then:

Theorem 2. *[2] Given a set of transactions, the conjunction of all the deterministic association rules defines exactly the empirical Horn approximation of the theory formed by the given tuples.*

So, the theorem determines that the empirical Horn approximation of the unordered data can be computed through the Formal Concept Analysis method of constructing deterministic association rules, that is, constructing the closed sets of attributes and identifying minimal generators for each closed set.

In this paper we want to find a notion of deterministic association rules for the more complex case of sequential data (ordered context), and of course we would like to support our proposal by proving a similar characterization.

3 Deterministic Association Rules in Ordered Contexts

Of course, the first task is to make available a closure operator that fits ordered data and specifies sensible results on practical cases. The most relevant existing contributions on mining closed sequential patterns are given by the algorithms CloSpan [13] or BIDE [12]. The extracted closed patterns by those algorithms are said to be stable in terms of support, which means that the closed patterns are maximal sequences in the set of objects where they are contained. For instance,

taking data from figure 1, we see that sequence $\langle(B)(D)\rangle$ is not a closed pattern since it can be extended to $\langle(B)(C)(D)\rangle$ in all the objects where it is contained. However, $\langle(B)(C)(D)\rangle$ or $\langle(A)(D)\rangle$ are closed (so, stable). We want to make sure that our theoretical notions fit appropriately these approaches. In fact, we do have already the closure operator set in place, through the Galois connection from [4], described below. There, two operators are defined in a formal context corresponding to sequences, and it is proved that they indeed enjoy the properties of a Galois connection so that their composition provides a closure operator.

Note that this task is nontrivial because it departs from the case of unordered transactions in the very definition of intersection. Whereas the intersection of two itemsets is another itemset, the intersection of two sequences (whether with or without the maximality condition we have imposed in the definition of intersection) does not in general result in a single sequence. So, the formal concept framework developed in [4] works with sets of sequences. Again, another difficulty arises, since ordering sets of sequences just by set inclusion does not give a Galois connection; using instead the ordering $S \preceq S'$ we have defined above does work, provided that the corresponding operators are defined adequately:

- For a set $O \subseteq \mathcal{O}$ of objects, $\phi(O) = \{s \in \mathcal{S} |\ s \textit{ maximal contained in } o,\ \forall o \in O\}$. This $\phi(O)$ is the set of maximal sequences common to all O, i.e.,$\phi(O)$ represents the intersection of the input sequences equivalent to O.
- For a set $S \subseteq \mathcal{S}$ of sequences, $\psi(S) = \{o \in \mathcal{O} |\ s \textit{ contained in } o,\ \forall s \in S\}$. This $\psi(S)$ is the set of objects containing all the sequences in S.

As mentioned, these two maps form a Galois connection (proved in [4]), and so, we can get the corresponding closure operator from their composition. We will call $\Delta = \phi \cdot \psi$ the closure operator on sets of sequences; thus, by definition, a set of sequences S is *closed* if and only if $\Delta(S) = S$. Similarly to any other Galois connection, we can also consider the dual operator Δ^{-1} that operates on sets of objects (although this dual operator is irrelevant for our present contribution).

It is proved that this operator Δ can characterize the closed sequences of CloSpan or BIDE as those sequences s that belong to the closure of $\{s\}$. Indeed, the instrumental property that connects the closure operator with the CloSpan sequences is the following:

Proposition 1. *[4] All sequences in a closed set are maximal in it w.r.t.* $\subseteq$.

Then it follows that $s \in \Delta(\{s\})$ if and only if s belongs to some closed set, and therefore the result of a mining task for closed sets under our Galois connection is the same as the result of the CloSpan or BIDE algorithm.

As described in the preliminaries (and exemplified by figure 1), given the data sequences $\mathcal{S}$ on items $\mathcal{I}$ we can construct the relation R which contains the same information as the individual components of each input sequence; thus, from R we obtain the collection of all formal concepts each corresponding to a closed set of sequences, and partially ordered by $\preceq$. As in any other Galois connections (see [7]), it gives immediately a lattice $\underline{\mathfrak{B}}(\mathcal{S}, \mathcal{I}, R)$ of formal concepts. For example, for the data in example 1(a), we depict graphically in figure 2 the corresponding lattice of closed sets of sequences. Together with each node S in

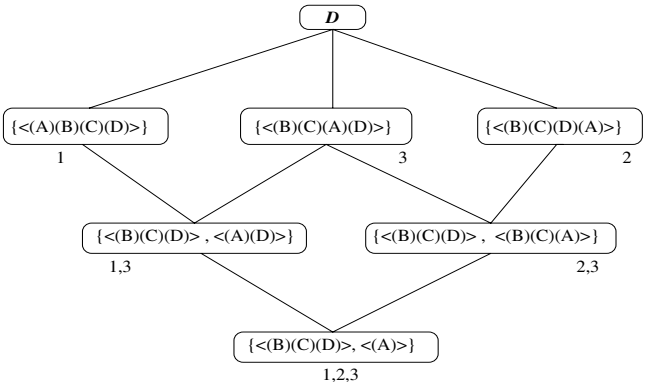

Fig. 2. Example of a concept lattice $\underline{\mathfrak{B}}(\mathcal{S}, \mathcal{I}, R)$

the lattice, we have added as a label the list of object identifiers where S is maximally contained (thus, as happens in general in Galois connections, these lists form a dual view of the same lattice that, in our case, is ordered by set-theoretic inclusion downwards). We also can see in the figure that, for each input sequence $s_i \in \mathcal{D}$, the set $\{s_i\}$ is a closed set; this always happens in general, also.

The set of sequences contained in all the input sequences will be called the *bottom or infimum* of the lattice; in most cases it will happen to be a trivial, somewhat artificial, element containing only the empty sequence. Similarly, we can also add an artificial set of sequences not contained in any input sequence, so that it forms the *top* of the concept lattice. In the example showed in figure 2, an artificial top not belonging to any object is added to the lattice and we note it by the set of input sequences $\mathcal{D}$ (i.e. we assume that $\mathcal{D} \not\preceq \{s_i\}$ for all $s_i \in \mathcal{D}$). This artificial top is not actually necessary in the model, and it was not originally presented in [4]; however, we add it to the lattice just to the effect of our later arguments. We say that a closed set of sequences S' is an *immediate predecessor* of another closed set of sequences S if $S' \preceq S$ and no closed set S'' exists in the lattice with $S' \preceq S'' \preceq S$. For example, in figure 2 $\{\langle(B)(C)(D)\rangle, \langle(A)\rangle\}$ is an immediate predecessor of two closed sets of sequences: $\{\langle(B)(C)(D)\rangle, \langle(B)(C)(A)\rangle\}$ and $\{\langle(B)(C)(D)\rangle, \langle(A)(D)\rangle\}$. Notice that the Galois connection presented in this section may be extended to other kind of structured data such as graphs or trees; we are currently working towards this formalization.

3.1 Generators of the Closed Set of Sequences

We say that a set of sequences G is a *generator* of S if we have that $\Delta(G) = S$. We say that a generator G is *minimal* if there is no other G' s.t. $G' \preceq G$ and $G \neq G'$, such that $\Delta(G') = S$. We will only consider minimal generators. These will be graphically added to the concept lattice model by dashed lines, as showed in figure 3. Minimal generators of the top of the lattice are not considered here, but, for the sake of illustration, it is easily seen that $\{\langle(C)(B)\rangle\}$ is among them.

We can define a family of deterministic association rules for sequences.

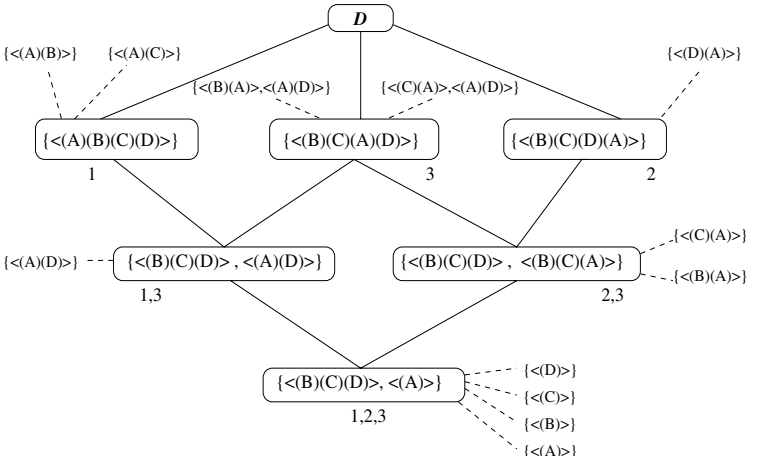

Fig. 3. Concept lattice $\underline{\mathfrak{B}}(\mathcal{S}, \mathcal{I}, R)$ with minimal generators

Definition 1. *A deterministic association rule with order is a pair* (G, S)*, usually denoted* $G \to S$*, where* $G, S \subseteq \mathcal{S}$ *and* $G \preceq S$ *s.t.* $\Delta(G) = S$*. We say that such a rule holds for a given set of sequences* $S' \subseteq \mathcal{S}$ *if either* $G \not\preceq S'$ *or* $S \preceq S'$*.*

The following lemmas characterize exactly the relation between the generators and their associated closed set of sequences, and will be useful to prove our main result characterizing deterministic association rules in ordered contexts by means of Horn logic.

Lemma 1. *Let* $\Delta(G) = S$*; then* $G \preceq S$ *and, for all closed sets of sequences* S' *s.t.* $S' \preceq S$ *and* $S' \neq S$*, we have that* $G \not\preceq S'$*.*

Proof. That $G \preceq \Delta(G)$ follows from the fact that Δ is a closure operator. We prove the following contrapositive of the rest: for closed sets S and S', if $\Delta(G) = S$ and $G \preceq S' \preceq S$ then $S' = S$. Indeed, by monotonicity of Δ, $\Delta(G) \preceq \Delta(S') \preceq \Delta(S)$ and, being S and S' closed, this translates into $S \preceq S' \preceq S$. Using here the fact that all sequences in all closed sets are maximal in them, it follows that $S = S'$. □

Actually, this is just a rephrasing of the well-known fact that closure operators assign to each set the *minimal* closed set that is above it; in the standard case (unordered data) the comparison is by set inclusion, but here the peculiarity is that the comparison is according to $G \preceq S$.

Lemma 2. *Let* $G \preceq S$ *where* S *is a closed set of sequences, and assume that, for all closed* S'*, if* $S' \preceq S$ *and* $S' \neq S$ *then* $G \not\preceq S'$*; then* G *contains at least one minimal generator of* S*.*

Proof. Consider all subsets of G for which the same property indicated for G still holds. Since they are a finite family, at least one of them is minimal in

the family (according to $\preceq$). Let G_{min} be this *minimal* subset of G that fufills the property (or, any of them if there are several): $G_{min} \preceq G \preceq S$, and for all closed $S' \preceq S$ s.t. $S' \neq S$, we have $G_{min} \not\preceq S'$. Then, the minimal closed set of sequences containing G_{min} is S, and so, $\Delta(G_{min}) = S$, being G_{min} one minimal generator contained in G. □

Due to the construction of the closure operator Δ, we can argue now that all the rules of our proposed form that can be derived from an input set of sequences $\mathcal{D}$ do hold for each of those input sequences; we could say that our implications with order have confidence 1 in our ordered data. Indeed, since $\{s_i\}$ is closed for each individual input sequence s_i of our database $\mathcal{D}$, we can consider any generator G and obtain, by monotonicity of Δ, $s_i \in \mathcal{D} \wedge G \preceq \{s_i\} \Rightarrow \Delta(G) \preceq \{s_i\}$; that is, the implication $G \rightarrow \Delta(G)$ holds for $\{s_i\}$.

4 Empirical Horn Approximation for Ordered Contexts

This section comes back to the propositional logic framework and Horn theories and introduces background knowledge to define the empirical Horn approximation for ordered contexts. To motivate our choices, let us briefly discuss a feature of the analysis in [2].

Indeed, the first step there, is to see each unordered transaction as a propositional model, and this is easy to obtain since actually it suffices to see the items as propositional variables. We can see this conceptual renaming as an isomorphism, or, even further, by using as propositional variables the very set of items, the translation is a mere identity function.

But this is no longer the case in our ordered contexts. Taking as propositional variables simply the items would not provide a sufficiently structured translation of our data sequences into propositional models. Thus, our next goal is to propose a more specific mapping that considers the ordered context. The resulting empirical Horn approximation of the ordered data will allow us to characterize the association rules defined in the previous section.

By way of example, consider figure 1, where the first object consists explicitly of the sequence $\langle (A)(B)(C)(D) \rangle$; however, it also contains implicitly all the subsequences $s' \subseteq \langle (A)(B)(C)(D) \rangle$. Thus, each input sequence can be also seen as a tuple of all those subsequences contained in it. Now we assign *one propositional variable to each subsequence* of each input sequence; and restrict the family of possible models by this background knowledge, thus discarding all models that would pretend to include a given sequence s but simultaneously discard some subsequence of s.

More precisely, let m be a model: we impose on it the constraints that if $m(x) = 1$ for a propositional variable x, then $m(y) = 1$ for all those variables y such that y represents a subsequence of the sequence represented by x. For instance, if a propositional variable x corresponds to the sequence $\langle (A)(B)(C) \rangle$, then a model m assigning 1 to x should also assign 1 to the variable representing $\langle (A)(B) \rangle$, and similarly with other subsequences.

We define more specifically the interpretation of variables as sequences by an *injective* function $\xi : \mathcal{S} \to \mathcal{V}$. For our convenience, we notationally extend this function with $\xi^{-1}(\Box) = \mathcal{D}$, where $\Box$ is the unsatisfiable boolean constant, and $\mathcal{D}$ is the notation for the set of sequences not belonging to any input sequence. Now, each input sequence s in the data corresponds to a model m_s: the one that sets to true exactly the variables $\xi(s')$ where $s' \subseteq s$; and we can find the empirical Horn approximation of the corresponding theory. It is important that the constraints we have imposed to the models, that when $s' \subseteq s$ then $\xi(s) \to \xi(s')$, are indeed Horn clauses, which we call *background Horn conditions*, and hold on all input models, so that they are imposed automatically unto the whole Horn approximation: the conjunction of all Horn clauses satisfied by all the models corresponding to input sequences. We call this conjunction the *empirical Horn approximation for ordered data*, and any model there can be mapped back into a set of sequences that is closed downwards under the subsequence relation.

4.1 Characterization

We are ready to present now the equivalence between the association rules extracted by the closure-based method presented in section 3, and the empirical Horn approximation for ordered data.

Theorem 3. *Given a set of input sequences S, the conjunction of all the deterministic association rules with order constructed as in section 3.1, seen as propositional formulas, and together with the background Horn conditions, axiomatizes exactly the empirical Horn approximation of the theory containing the set of models $M = \{m_s | s \in \mathcal{D}\} \subseteq \mathcal{M}$.*

Proof. We prove separately both directions for this theorem: 1/ that the deterministic association rules (that is, their corresponding propositional implications) are implied by the empirical Horn approximation; and 2/ that all the clauses in the empirical Horn approximation are implied by the conjunction of the (propositional implications corresponding to) deterministic association rules.

$\Rightarrow$/ Consider a deterministic association rule $G \to S$ s.t. $\Delta(G) = S$. By distribuitivity, we can rewrite the rule as a conjunction of different implications $G \to s_i$ where $S = \{s_1, \ldots, s_m\} \in 2^{\mathcal{S}}$. As explained after lemma 1, all the input sequences having as subsequences all the elements of G must have also s_i, so that the translation of $G \to s_i$ is a Horn clause that is true for all the given models in M and, by the theorems in the previous section, it belongs to the empirical Horn approximation. Likewise, the background Horn conditions are also satisfied by all models and thus hold in the empirical Horn approximation.

$\Leftarrow$/ Let $F \to v$ be an arbitrary Horn clause where F is a set of variables, and v is a single variable. Assume this clause to be true for all the given models $M = \{m_s | s \in \mathcal{D}\}$ that correspond to the input sequences; note that these follow the constraints mentioned above: if $m \in M$, and $m(x) = 1$ for a propositional variable x, then $m(y) = 1$ for all those variables y such that $\xi^{-1}(y) \subseteq \xi^{-1}(x)$. In order to show that $F \to v$ is a consequence of the rules found from the concept

lattice for S, we will find an association rule that, upon translation, and in the presence of the background Horn conditions, logically implies our Horn clause.

Looking at F as a set of variables, we can consider the set of corresponding sequences $S' = \{\xi^{-1}(v) | v \in F\}$; let $S'' = \Delta(S')$ be its closure. By previous lemmas 1 and 2, we know that S' will contain at least one minimal generator of S'', that is, $G \subseteq S'$ s.t. $\Delta(G) = S''$. Therefore, the rule $G \to S''$ will be one of the rules constructed by the FCA method.

On the other hand, we have assumed that the clause $F \to v$ holds for all the models M. By definition, it means that $S' \to \xi^{-1}(v)$ also holds in all the input sequences, in the sense that whenever $S' \preceq \{s\}$ for an input sequence s, also $\xi^{-1}(v) \subseteq s$; and this implies that $\{\xi^{-1}(v)\} \preceq \Delta(S') = S''$: so, for some sequence $s \in S''$ we have that $\xi^{-1}(v) \subseteq s$ or, equivalently, the Horn clause $\xi(s) \to v$ belongs to the background Horn conditions.

Finally, we have found that $G \to s$ is one of the rules composing $G \to S$, which is one of the association rules coming from the closure system. Since $G \subseteq S'$, the variables corresponding to sequences from G are all in F, and thus the clause $F' \to \xi(s)$ with $F' \subseteq F$ corresponds to one of the association rules. By subsumption, and one resolution step with $\xi(s) \to v$, we see that $F \to v$ follows indeed from the association rules plus the background Horn conditions. □

Note that this proof works also well when the Horn clause is nondefinite, that is, when considering $F \to \square$. In this case no model from M satisfies all the variables in F, so, $S' \not\preceq \{s_i\}$ for all $s_i \in \mathcal{D}$; indeed we have that $\Delta(S') = \mathcal{D}$ (top of the lattice not included in any input sequence).

Our characterization brings meaning to the deterministic association rules extracted by the lattice method of ordered data. We have seen that they exactly correspond to the empirical Horn approximation under the necessary background Horn conditions. Next step is then to discuss the algorithmic consequences of calculating these implication rules with order, and to propose specific algorithms.

5 Computing Rules in Ordered Contexts

As mentioned before and proved in [4], the closure operator Δ characterizes the closed patterns of CloSpan [13] (which are closed in the sense of not being extendable in support, thus stable) as those that belong to a closed set. This fact makes CloSpan a good candidate algorithm to construct the concepts of our lattice model. Recently, a more efficient algorithm, BIDE [12], has been presented; according to the authors, it outperforms CloSpan being more than an order of magnitude faster; however, the output patterns mined by CloSpan or BIDE are exactly the same. To the best of our knowledge, these two algorithms are the only contributions to the mining of closed sequences up to now. The output of either can be used to construct the concepts of our model, just by appropriately organizing them.

However, computing the deterministic association rules in the ordered data (equivalently, the empirical Horn approximation for the ordered context) we seem to need as well all the minimal generators, in order to output all rules $G \to S$

where S is closed and G is a minimal generator of S. Thus, an important next step to add to any current algorithm for closed sequences is then the calculation of minimal generators for each closed set. We want to compute these minimal generators by means of a general method, so that it can be plugged into any underlying algorithm of mining closed sequential patterns such as either CloSpan or BIDE. In this way, after computing the closed sets of sequences, the chosen algorithm can directly calculate the minimal generators as well, without incurring in inconvenient overheads for intersecting sequences of the database. In this section we show how to compute minimal generators of a closed set of sequences S as a sort of transversal of appropriately defined differences between S and all proper closed predecessors in the lattice.

The difficulty of this proposal will rely on the formalization of both steps: 1/ what it is exactly the difference between two sets of sequences, and 2/ how to properly define the appropriate variant of transversal. The motivation to look for such an approach is that it can be seen that the concept lattice we have obtained is isomorphic to a standard concept lattice for which such a method of computing rules does already exist [10]; note however that it is not immediate to carry over the isomorphism into the generators, so that we prefer to develop our method fully within the closure operator on sets of sequences.

For comparison purposes, we quote here a result that we found in [10] and that we would like to export here, whereby the minimal generators of a closed set in the unordered context obtained by a closure operator Γ are characterized (the original statement differs from ours but their equivalence is readily seen.)

Theorem 4. *Let Z be a closed set of items $Z = \Gamma(Z)$; the minimal generators of Z are found as the minimal transversal hypergraph of the hypergraph of the differences $Z - Z'$ where Z' are the proper closed subsets of Z in the unordered lattice.*

The transversal hypergraph consists of sets that intersect each and every of the given differences (called *faces* in [10], a term that comes from related matroid-theoretic facts). Also, it is not difficult to see that it suffices to state that the generator intersects the differences with $Z - Z'$ for the closed immediate subsets of Z. For instance, let $Z = \{a, b, c\}$ be a closed set of items, whose immediate closed predecessors in the lattice are $Z'_1 = \{a, b\}$ and $Z'_2 = \{a, c\}$; then, the minimal generators of Z can be found by transversing the hypergraph of differences $H = \{Z - Z'_1, Z - Z'_2\}$, that is, $H = \{\{c\}, \{b\}\}$. The minimal transversal of H is $\{c, b\}$, and so it is the minimal generator of Z.

We would like to have a similar result as theorem 4 for the minimal generators of the closed sets of sequences.

5.1 Computing Minimal Generators for Closed Set of Sequences

We preserve here the term *faces* for our appropriate formalization of the differences between one closed set and its proper closed predecessors (according to $\preceq$); for closed S, each face of S is $S - S'$, where $S' \preceq S$ is a proper closed predecessor of S, and the difference is defined as

$$S - S' = \{s | \{s\} \preceq S \text{ but } \{s\} \not\preceq S'\}$$

The main property now is:

Lemma 3. *Let S be a closed set of sequences and $G \preceq S$; then $\Delta(G) = S$ if and only if G intersects all the faces of S.*

Here by G intersecting a face $S - S'$ we understand set-theoretic intersection, that is, there must exist a common sequence in both. This corresponds to our notion of transversal for ordered data.

Proof. Assume first that G does not intersect the face $S - S'$, for some $S' \preceq S$; thus, no $s \in G$ fulfills the condition in the definition of the face. Since $G \preceq S$, for all such s, $\{s\} \preceq S$ as well, and this implies $\{s\} \preceq S'$, or actually $G \preceq S'$. Now, by monotonicity of Δ, from $G \preceq S' \preceq S$ and the fact that sequences in closed sets are maximal we obtain $S = S'$ just as in 1; and S' is not a proper predecessor so that $S - S'$ is not a face. Conversely, assume that G indeed intersects all the faces; from $G \preceq S$ and monotonicity again we have $\Delta(G) \preceq S$. Equality will follow as we need, if we prove that $\Delta(G)$ is not a proper predecessor. Indeed, by lemma 1, $G \preceq \Delta(G)$, so for all $s \in G$, $\{s\} \preceq \Delta(G)$, which negates the condition in the definition of $S - \Delta(G)$. Thus it can't happen that any s is both in G and in $S - \Delta(G)$, and this last difference cannot be a face because G intersects all of them. This implies that $\Delta(G)$ is not a proper predecessor. □

Again, we only need to consider immediate predecessors: if G intersects the faces corresponding to immediate predecessors, it must also intersect the other faces, which are larger. Additionally, we may be only interested in minimal generators (according to $\preceq$) since non-minimal generators only yield redundant association rules. It is not difficult to see that this can be enforced by using only those subsequences of sequences in S that are minimal in their respective face for the construction of the generators as in lemma 3.

For a more graphical example of our method, let $S = \{\langle (B)(C)(A)(D)\rangle\}$ be a closed set of sequences, as showed in the lattice of figure 2; the proper predecessors of S are the closed set of sequences $S'_1 = \{\langle (B)(C)(D)\rangle, \langle (A)(D)\rangle\}$, and $S'_2 = \{\langle (B)(C)(D)\rangle, \langle (B)(C)(A)\rangle\}$. The minimal new subsequences in S not contained in S'_1 are $F_1 = \{\langle (B)(A)\rangle, \langle (C)(A)\rangle\}$, and the minimal new subsequences in S not contained in S'_2 are $F_2 = \{\langle (A)(D)\rangle\}$. Now, to find the minimal generators of S we must minimally transverse these differences, which are indeed the two faces of S, obtaining two generators: $G_1 = \{\langle (A)(D)\rangle, \langle (B)(A)\rangle\}$ and $G_2 = \{\langle (A)(D)\rangle, \langle (C)(A)\rangle\}$, which are exactly the minimal generators of S (see figure 3).

6 Conclusions

We have proposed a notion of deterministic association rules in ordered data, building on the fact that such rules for unordered data can be formally justified as implications in a propositional logic framework; our extension provides a

way of mining facts where a set of subsequences implies another subsequence in the data, and proves that the mined rules can be formally justified as well by a purely logical characterization. We do that using the concept lattice model provided by the Galois connection and associated closure operator proposed in [4]: by means of minimal generators that imply a closed set of sequences of the concept lattice. Indeed, these deterministic association rules characterize exactly the natural notion of empirical Horn approximation for ordered data, which involves specifying a number of background Horn conditions that ensure consistency of the theory with the ordered context.

We have discussed as well the algorithmic consequences of deriving such implications with order. Since any current algorithm for mining closed sequences can be used for constructing the closed concepts of our lattice model, we just need to incorporate here the derivation of minimal generators. We consider the characterization of generators as transversals of faces, known in the unordered case, and we prove a parallel result in our ordered case. This provides a method that can be easily incorporated in any algorithm that constructs our closed sets in the appropriate order, such as the algorithms existing in fact for closed sequences, so that generators and association rules can be indeed inferred from just the system of closed sets. We are currently developing implementations of our methods to investigate their behavior in practice.

Other extensions of the basic itemset-based characterization are worth more research. A relevant property of the rules studied here is the need of absolute confidence; this can be inappropriate in two different ways. First, one may wish to take into account the possibility of small errors, such as miskeying, that make inapplicable a deterministic association rule; it is possible to adapt the case of itemsets to this consideration [15], which we consider a data cleaning problem rather than a data mining or relational problem. A second, inherently different case is the more usual application of association rules where more relaxed confidences are used. For this case, there is a large number of proposals of how to measure the strength of the implication; a survey and comparison, with appropriate references, is given in [5]. To our knowledge, there is no principled way to select one of them and know what one is actually doing through this choice; specific data mining software may allow only some of them, as a consequence mainly of research schools of their designers. In fact, most measures allow for examples of counterintuitive or misleading results.

We believe that it is possible to modify the definitions of Horn approximations so as to take into account the various forms of strength of implication, or at least some of them; so that, at the time of selecting one measure of strength of implication, we know more information about the specific bias we are introducing in the analysis, and maybe check the pertinence of such a bias against domain information that could be available to the data miner. This difficult but important extension of our work, which also will allow for consideration of sequential or more generally structured contexts, is to be pursued in the near future by the authors.

References

1. R. Agrawal and R. Srikant. Mining sequential patterns. In *Eleventh International Conference on Data Engineering*, pp. 3–14. IEEE Computer Society Press, 1995.
2. J.L. Balcázar and J. Baixeries. Discrete deterministic datamining as knowledge compilation. In *Workshop on Discrete Mathematics and Data Mining, in SIAM Int. Conf.*, 2003.
3. M. Cadoli. Knowledge compilation and approximation: Terminology, questions, references. In *AI/MATH-96, 4th. Int. Symposium on Artificial Intelligence and Mathematics*, 1996.
4. G. Casas-Garriga. Towards a formal framework for mining general patterns from structured data. In *Workshop Multi-relational Datamining, in KDD Int. Conf*, 2003.
5. G. Casas-Garriga. Statistical strategies to remove all the uniteresting association rules. In *Proc. 16th European Conf. on Artificial Intelligence*, pp. 430–435, 2004.
6. A. Day. The lattice theory of functional dependencies and normal decompositions. *Int. Journal of Algebra and Computation*, 2(4):409–431, 1992.
7. B. Ganter and R. Wille. *Formal Concept Analysis. Mathematical Foundations.* Springer, 1998.
8. H. Kautz, M. Kearns, and B. Selman. Horn approximations of empirical data. *Artificial Intelligence*, 74(1):129–145, 1995.
9. N. Pasquier, Y. Bastide, R. Taouil L., and Lakhal. Closed set based discovery of small covers for association rules. In *Proc. 15th Int. Conf. on Advanced Databases*, pp. 361–381, 1999.
10. J.L. Pfaltz and C.M. Taylor. Scientific knowledge discovery through iterative transformations of concept lattices. In *Workshop on Discrete Mathematics and Data Mining, in SIAM Int. Conf.*, pp. 65–74, 2002.
11. B. Selman and H. Kautz. Knowledge compilation and theory approximation. *Journal of the ACM*, 43(2):193–224, 1996.
12. J. Wang and J. Han. BIDE: Efficient mining of frequent closed sequences. In *Proc. 19th Int. Conference on Data Engineering*, pp. 79–90, 2003.
13. X. Yan, J. Han, and R. Afshar. Clospan: Mining closed sequential patterns in large datasets. In *Proc. Int. Conference SIAM Data Mining*, 2003.
14. M. Zaki. Generating non-redundant association rules. In *Proc. 6th Int. Conference on Knowledge Discovery and Data Mining*, pp. 34–43, 2000.
15. M. Zaki and M. Ogihara. Theoretical foundations of association rules. In *Workshop on Research Issues in Data Mining and Knowledge Discovery, in SIGMOD-DMKD Int. Conf.*, 1998.