Alcom-FT Technical Report Series
ALCOMFT-TR-03-201

Discrete Deterministic Data Mining as Knowledge Compilation*

J.L. Balcázar
Departament de LSI
Universitat Politècnica de Catalunya
balqui@lsi.upc.es

J. Baixeries
Departament de LSI
Universitat Politècnica de Catalunya
jbaixer@lsi.upc.es

January 31, 2003

Abstract

In the subfield of data analysis known as Discrete Deterministic Data Mining, notions from Formal Concept Analysis have been used to design a procedure to construct association rules. Empirically, this method has proved effective in practice for certain relevant tasks. We mathematically prove that these methods, rephrased in a Propositional Logic framework, compute the so-called Empirical Horn Approximation, which is a major model of Knowledge Compilation and is well-known to produce very good practical results in several fields of Artificial Intelligence.

1 Introduction

There exists a large number of knowledge representation formalisms. One of the reasons is that the task of inferring new facts from the knowledge of previous facts is usually computationally hard. Thus, knowledge representation formalisms must strive for a balance between two desirable properties, namely, expressive richness and computational efficiency. Both aims are to be considered relative to the application context.

Two traditional avenues for research along this line correspond to weakening, respectively, the representation language or the inference methods. A third, not so deeply explored possibility is called Knowledge Compilation [6]; according to that view, knowledge is provided in a nonrestricted language and it becomes subsequently "compiled" into a restricted language which allows for fast deduction or abduction algorithms; thus, it can be seen as moving a substantial part of the computational task "offline" into the compilation process and, once this is done, short response times are obtained for whatever deductive queries come after. A price to be paid is that it could be that the restricted language allowing for fast queries might not be able to represent exactly the input knowledge given in the unrestricted language, and therefore approximations must be used. Examples of methodologies that can be seen as Knowledge Compilation techniques appear in [6] (see also [1] and the references there).

A prime, practically effective method of knowledge compilation starts off from theories expressed by extensionally providing the list of the satisfying models, and compiles the theory into the form of Horn clauses [6], [5]. This development is frequently phased in propositional languages (or in first order languages over finite domains, which to some extent is equivalent to propositional languages). Of course it is likely that no Horn axiomatization of the given theory exists; thus, Horn approximations to the theory have been identified, and it is known that there is always a single Horn theory that acts as minimal upper bound of any given theory. In contrast, there may

*This work is supported in part by EU ESPRIT IST-1999-14186 (ALCOM-FT), MCYT TIC2002-04019-C03-01 (Moises), and matching funds TIC2000-1970-CE, and CIRIT 2001SGR-00252.

be several nonequivalent maximal lower bounds. Horn theories are advantageous in that they allow for fast deduction and abduction algorithms [3], [5]. Specific examples of the usefulness of this approach are provided in [6].

Here we show that a recent proposal of Discrete Deterministic Data Mining, based on Formal Concept Analysis [4], and consisting of identifying a lattice of formal concepts, or closed sets (precise definitions are given later on), from the data and then constructing implications on the basis of the generators of the closed sets [7], [8], [9], corresponds exactly to that most common form of Knowledge Compilation. Indeed, we formally prove that the logical formulas obtained by this method (actually, a technically very slight variant) characterize a theory that is logically equivalent to the Empirical Horn Approximation of the data [5], i.e. the unique minimal Horn upper bound of the data tuples that would be obtained through knowledge compilation.

This formal connection provides a number of benefits; from a conceptual perspective, we obtain a better understanding of both the behavior of the DDDM method by Pfaltz, and of the combinatorics of propositional Horn clauses, which suggests that a crucial parameter for the algorithmics of Horn clauses is the number of intersections of models necessary to construct all the models of the theory out of the so-called characteristic models. From a more practical perspective, the properties of the models of a Horn theory allow us to identify the two key properties that will make the DDDM method by Pfaltz work efficiently: closure under intersection of models of the phenomenon at hand and enough data to capture all the characteristic models. Additionally, the characteristic models can be seen as a form of data summarization, in the sense that the outcome of the method will only depend on them, and all non-characteristic models that could be present in the data are irrelevant for such a DDDM task.

2 Pfaltz's DDDM Method

This process starts from a binary relation between objects and attributes. Each object is described simply by listing its attributes, so that we can see each object as either a tuple of attributes or as a subset of the set of all potential attributes in the relation. Of course, one can as well consider a dual relation by interchanging the roles of objects and attributes; we do not discuss this duality here, nor do we discuss the closely related Galois connection.

We only depart slightly from the original proposed method by Pfaltz [8] in one small technical point: we assume the existence of an attribute that all the objects have, and of an attribute that no object has; in case such attributes are not present in the data, artificial ones can be easily added. Actually, in most cases, there is no need of implementing them in the true dataset, and they are only necessary to the effect of our arguments. Note that these two attributes implement the constant boolean functions True and False over the objects. We denote by $\Box$ (as is customary for the constant false boolean function) the attribute that is not satisfied by any object.

From the tuples, a lattice of closed sets is constructed, on the basis of a closure operator on sets of attributes (or, alternatively, on sets of objects). The closure of a set of attributes G, which we denote $\Gamma(G)$ includes all attributes that are present in all objects having all attributes in G. Closed sets are those that coincide with their closure.

It is not difficult to see that, in the family of closed sets defined in this way, the set of all attributes is closed (via the attribute $\Box$), the set of attributes present in any individual object is also closed, and the intersection of closed sets is closed.

A subset G of a closed set F is a *generator* of F if given the closure operator Γ, $\Gamma(G) = F$, and there is no $G' \subset G$ such that $\Gamma(G') = F$ (here $G' \subset G$ denotes proper inclusion). Then, rules are constructed as $G \to F$ for each closed set F and each generator G of F; note that only the part $G \to F - G$ is relevant since $G \to G$ is tautologous. By the definition of closure, it follows that each of the rules obtained is true of each of the objects. A peculiar particular case is given by the generators of the set of all attributes, which has to be closed by definition. Our assumption that there is an attribute satisfied by no object implies that no object satisfies the set of all attributes. The generators G of this particular closed set (if it does not appear in the data) express, therefore, incompatibilities between attributes, in the sense that no object has been seen having simultaneously all the attributes in G.

The rules so obtained relate attributes in a spirit similar to the association rules obtained by Apriori or any other such algorithm, but they are deterministic in the sense that no nontrivial bound on the support is imposed (equivalently, nonzero support suffices) whereas an extreme bound of 1 is imposed on the confidence (equivalently, all tuples for which the left hand side holds must fulfill the right hand side as well). The advantages are, essentially, the ability to find connections that affect very few examples but, at the same time, reducing the combinatorial explosion that produces hundreds of thousands of rules according to the standard association notions, since in practical cases many closed sets turn out not to generate any (nontrivial) rule.

These support/confidence conditions make DDDM, taken literally, unsuitable for many social DM tasks, but appropriate for many scientific discovery tasks where cause-effect relations are deterministic. See [9] for a discussion of extensions of the method to deal with outliers.

Our discussion has been necessarily brief: see the references given for details and illustrative examples.

3 Propositional Logic Framework

The purpose of this section is essentially to fix notation, since it amounts to a trivial rephrasing of the previous section.

Each attribute in the relation can be seen as an atomic boolean function of the objects; thus, each corresponds to a propositional variable, and each object, by having it or not, assigns to it a boolean truth value. Therefore, each object is a tuple of boolean truth values: an assignment for the boolean variables, or, equivalently, a propositional model. A bitwise conjunction (variable by variable) of two models yields another model, which we call the intersection of the original two models (since it is indeed an intersection when we see each tuple as a subset of the attributes). A literal is either a boolean variable or its negation.

In such finite universes, a theory is syntactically a formula, and semantically a set of models: those satisfying the formula. If we see the formula as a conjunction of smaller formulas, possibly a number of them are redundant, in the sense that a smaller conjunction may define (or: axiomatize) the same semantical theory. This is so if all the deleted conjunctions follow from the remaining ones.

A Horn clause is a disjunction of literals, among which there is at most one positive literal. A definite Horn clause is a Horn clause having exactly one positive literal. A Horn formula is a conjunction of Horn clauses. A Horn theory is a theory that can be axiomatized by a Horn formula. We usually write definite Horn clauses as $G \to X$ for a propositional variable X (the positive literal) and a possibly empty set G of variables (the negative literals). We usually write nondefinite Horn clauses formed by all the negations of the variables of G as $G \to \square$, which is equivalent since $\square$ expresses unsatisfiability. Having $\square$ as an explicit unsatisfiable attribute allows us to treat later on both definite and nondefinite clauses in a unified way.

When a theory contains another we say that the first is an upper bound for the second. Note that by removing clauses from a Horn formula we get actually a larger or equal theory. The following is known:

Theorem 3.1 *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 is defined by the conjunction of all the Horn clauses that are satisfied by all models from T.*

We call that theory the Empirical Horn Approximation. Proofs can be tracked easily among the references in [6], [5]. A relevant consequence is that T is a Horn theory if and only if it is actually closed under intersection, and then it coincides with its empirical Horn approximation [2].

Of course, models that are intersection of other models are irrelevant in many cases. A characteristic model of a theory T is a model m that cannot be obtained by intersecting other members of $T - \{m\}$; then:

Theorem 3.2 *[5] Let T be a Horn theory. Then it coincides with the empirical Horn approximation of the set of characteristic models of T.*

Our kickoff point now is the following observation: the rules constructed by Pfaltz's DDDM method are conjunctions of Horn clauses, and the conjunction of all the rules is a conjunction of Horn clauses as well. Indeed, by distributivity, a rule of the form $G \to F$ can be decomposed into the conjunction of Horn clauses of the form $G \to X$ for each attribute $X \in F$. Those where $X \in G$ are trivially true but undisturbing. The others are actually the useful ones.

Before moving on, let us warn the reader of the two uses of the word "closure" that we are employing. On the one hand, for sets of attributes, some of them are deemed "closed", and the closure of a set G of attributes is another set of attributes, containing it, and being the smallest such closed set; note that, by definition, the set of all the attributes is closed. Thus, there are closed sets that contain G. Since the intersection of closed sets is closed, we obtain a closed set F when we intersect all closed sets that contain G, and it still contains it: thus the condition of "smallest" gives a well-defined notion of closure.

On the other hand, one power-set operation above, we consider the closure under intersection of sets of models. We will be repeatedly using the fact that the family of closed sets is indeed closed under intersection, as we just did.

4 Main Results

Let us start by pointing out some useful facts. Consider any closed set of attributes F.

Proposition 4.1 *If G is a generator of F then, for each proper closed subset $F' \subset F$, G is not included in F'; equivalently, G intersects $F - F'$: we have $G \cap (F - F') \neq \emptyset$. Moreover, G is minimal among the sets having this property, and all such minimal sets are generators of F.*

This is essentially theorem 2.1 in [8]. We have rephrased it slightly since it is stated there in a dual form, in terms of each such difference $F - F'$ being a "blocker" (a transversal, in terms of hypergraph theory) of all the generators.

We will also use the following technical tool.

Lemma 4.2 *Let $H \subseteq F$ where F is closed. Assume that for each proper closed subset $F' \subset F$, the following property holds: $H \cap (F - F') \neq \emptyset$. Then H contains at least one generator of F.*

Proof. Consider all subsets of H for which the property still holds. They are a finite family, so at least one of them (in general many more) is minimal in the family. Those of them for which the property holds minimally are generators, by the previous fact, thus there must be at least one generator included in H. □

We are ready to argue the first of our main results.

Theorem 4.3 *Given a set of tuples, the conjunction of all the deterministic association rules constructed by Pfaltz's DDDM method defines exactly the empirical Horn approximation of the theory formed by the given tuples.*

Proof. We have to prove that the conjunction of the deterministic association rules defines the same theory as the empirical Horn approximation. It suffices to prove that (a) all the deterministic association rules are implied by the empirical Horn approximation and (b) all the clauses in the empirical Horn approximation are implied by the conjunction of the deterministic association rules.

To prove (a), consider a deterministic association rule and decompose it into several Horn clauses by having one of them for each attribute in the consequent, as mentioned above. Each of these clauses holds for all the given tuples. Thus, being a Horn clause that is true for all the given models, by the theorems in the previous section it belongs to the empirical Horn approximation.

Proving (b) is the harder direction. We have to prove that any arbitrary Horn clause $H \mapsto A$ that holds for all the tuples is indeed a consequence of the rules found by the Formal Concept Analysis method. Note that, in our Horn clause, A is either $\square$ or a propositional variable, whereas H is a term, a conjunction of propositional variables, and can be seen as a set of attributes. The clause is trivially satisfied by a model m (seen as a set of attributes as described in the introduction) when $H \not\subseteq m$; for nondefinite clauses, this is the only way to satisfy them. For definite clauses, when $H \subseteq m$ so that the antecedent term is satisfied by the model m, satisfying the clause means including as well the attribute A.

Let F be the closure of H in the Formal Concept lattice on the set of tuples given. We already saw that it is the smallest closed set containing H. Thus, for every closed proper subset $F' \subset F$, we know that H is not included in F': $H \cap (F - F') \neq \emptyset$. By the previous lemma, H contains at least one generator of F, let it be $G \subseteq H$. We know now that the rule $G \mapsto F$ is one of the rules constructed by the Formal Concept Analysis method.

On the other hand, we were assuming that the Horn clause $H \mapsto A$ holds for all the tuples. Thus, each tuple including all the attributes in H must include as well the item A, and this means that A belongs to the closure of H, namely F. Therefore, one of the clauses composing the conjunction given by the rule found, $G \mapsto F$, is the clause $G \mapsto A$; and the rule $H \mapsto A$ is a logical consequence of $G \mapsto A$ since G is a subset of H. Therefore, $H \mapsto A$ is a logical consequence of one of the rules found from the given tuples by the Formal Concept Analysis method, and this completes the proof. $\square$

Note that the proof works as well when A is the attribute that no object satisfies: this corresponds to $A = \square$ and to the case where the clause is nondefinite.

As a consequence, we can precisely identify the characteristic models when we develop the incremental procedure proposed in [7], [8]. We find it instructive since it explains the role of the characteristic models as exactly those that make further progress in refining the concept lattice: the incremental method adds new concepts exactly upon processing a new characteristic model.

Theorem 4.4 *Given a theory T and a model m, the fact that this new model forces to create (at least) a new concept in the concept lattice induced by T is equivalent to m being a characteristic model of $T \cup \{m\}$.*

Proof. Let $T' = T \cup \{m\}$. We prove that m is a characteristic model of T' if and only if the concept lattices of T and T' differ.

- $\rightarrow$) Let us suppose that adding m to the set of models T adds a new concept (at least) in the concept lattice induced by T. This new closed set cannot be a model in T, since all models in T become closed sets, as previously seen. This new concept cannot be either an intersection of models in T, since the intersection of closed sets is a closed set. Consequently, m cannot belong to T nor to the closure of T under intersection, which means that m is a characteristic model of T'.

- $\leftarrow$) Let m be a characteristic model of T'. We want to show that the lattice induced by T' will contain at least one more closed itemset than that of T. Since m is a characteristic model of T', $m \notin \mathit{closure}(T' - \{m\})$, it follows that $m \notin \mathit{closure}(T)$. Will m create a new concept in the lattice induced by T? Let us suppose that it will not: it means that m is already in T, which means that either it is a model of T, or that it is an intersection of different closed itemsets in T, which is equivalent to belonging to the closure of T. Both cases contradict previous assumptions. $\square$

5 Conclusions

We have formally proved that the Pfaltz's FCA-based method for constructing deterministic association rules actually identifies an axiomatization of the empirical Horn approximation to the theory provided by the given data tuples.

There are evidences in the literature that this method is effective in practice. From the knowledge about Horn clauses, we are now in a position to recall in which cases this method will work.

The exact equivalence between the rules formed in a FCA analysis and a Horn Approximation of the theory allow us to see that if all input tuples are drawn from an underlying theory T which is closed under intersection, then, the rules obtained by the FCA method will define a smaller theory, if not all the characteristic models have yet appeared in the set of input tuples; otherwise, FCA rules will define exactly T. As long as more tuples are seen, and all the characteristic models get added to the set of input tuples, this smaller theory will converge to T.

Many natural applications of association rules are closed under intersection: for instance, there is no reason to believe that the market basket consisting of the intersection of two given ones is impossible. (Note however that the boolean recoding of a vertical fragment of the Mushroom database described in [9] as a target of DDDM analysis is not closed under intersection, since all models there have exactly nine attributes. Anyway, the intersection closure of these tuples does correspond to partial observations of mushrooms where some of the attributes are not considered at all, and therefore it does make sense that DDDM works so extremely well on those data.) Moreover, it is usual that the set of characteristic models of a Horn theory be reasonably small (with respect to the size of the whole theory), so that the chances of finding all of them in a large sample are high. This, together with our main results, explains the high practical effectivity of Pfaltz's FCA-based technique for several data mining applications.

References

[1] Cadoli M. *Knowledge Compilation and Approximation: Terminology, Questions, References.* In AI/MATH-96, 4th International Symposium on Artificial Intelligence and Mathematics, 1996.

[2] Dechter, R., Pearl, J. *Structure Identification in Relational Data.* Artificial Intelligence, 58, pages 237-270, 1991.

[3] Dowling, W.F. and Gallier, J.H. *Linear time algorithms for testing the satisfiability of propositional Horn formula.* Journal of Logic Programming, 3, pages 267-284.

[4] Ganter, B., Wille R. *Formal Concept Analysis. Mathematical Foundations.* Springer, 1999.

[5] Kautz H., Kearns M., Selman B., *Horn Approximations of Empirical Data.* Artificial Intelligence, 74/1, pages 129-145, 1995.

[6] Kautz H., Selman B. *Knowledge compilation using Horn approximations.* In Journal of the ACM 43, 2, pages 193-224, 1996

[7] Pfaltz, J.L. *Transformations of Concept Graphs: An Approach to Empirical Induction.* 2nd International Workshop on Graph Transformation and Visual Modeling Techniques. GTVM 2001, Satellite Workshop of ICALP 2001, Crete, Greece. Pages 320-326. July 2001.

[8] Pfaltz, J.L., Taylor, C.M. *Scientific Discovery through Iterative Transformations of Concept Lattices.* Workshop on Discrete Mathematics and Data Mining at 2nd SIAM Conference on Data Mining, Arlington. Pages 65-74. April 2002.

[9] Pfaltz, J.L., Taylor, C.M. *Closed Set Mining of Biological Data*, BIOKDD 2002. Workshop on Data Mining in Bioinformatics at KDD 2002, Edmonton, Alberta, July 2002, 43-48.