Galois Connections for Mining Structured Objects

Gemma Casas-Garriga and Jose Luis Balcazar.

2004
Galois Connections for Mining Structured Objects

Gemma Casas-Garriga and José L. Balcazar
Departament de Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya
Jordi Girona Salgado 1-3, Barcelona
{gecas,balqui}@lsi.upc.es

ABSTRACT

In this paper we revisit, by means of examples, the application of Galois connections for the problem of mining structured objects. Starting from the most basic case, the mining of ordered data, we present some primary results obtained within this new framework: from the identification of plain subsequences called stable to the extraction of closed partial orders summarizing the data. Finally, we point out some further ideas to extent and complete this current proposal.

1 INTRODUCTION

Formal Concept Analysis is based on the mathematical theory of complete lattices; see e.g. (Ganter and Wille, 1998). The theory of concept (or Galois) lattices provides a natural graph setting that can be applied to a variety of fields in computer science, including tasks such as Learning or Knowledge Discovery. A formal treatment of Galois lattices begins by considering a binary relation \( R \subseteq O \times I \) between a set of objects \( O \) and a set of variables or items \( I \), called the formal context. We may think of a context as a set of bit vectors: each bit on corresponds to an object having that particular item. Formally, a concept of a context is a pair \( (I, O) \) of a set of items and a set of objects linked by a Galois connection.

The standard Galois connection for a binary relation \( R \) maps each family of objects to the set of those items that hold in all of them, and each set of items to the set of objects in which they hold. It is known that the compositions of these two operators forming a Galois connection provides two closure operators, one on the universe of objects and another on the universe of items. By closure of a set of objects with respect to \( R \), we mean a maximal set of objects sharing the same attributes of all \( o \in O \). Similarly, the closure of a set of items \( I \) picks up any other attributes that are common to all objects satisfying all \( i \in I \). Closed sets of objects (or items) are those coinciding with their closure. Given concepts \( (I, O) \), both \( I \) and \( O \) are closed set of items and objects respectively. So, from the set of closed sets of items and closed sets of objects which are linked by the Galois connection we can draw the lattice of concepts (see (Ganter and Wille, 1998)): the concepts found in \( R \)

*This work is supported in part by MCYT TIC 2002-04019-C03-01 (MOISES)
can be graphically organized in a Hasse diagram, i.e. a graph where each node is a concept and there is an edge between two nodes if and only if they are comparable and there is no other intermediate concept in the lattice. The ascending paths in the lattice represent the subclass/superclass relation.

This resulting graph-based system is very competitive in learning tasks because it provides a powerful and flexible representation that can be used in relational domains. In particular, the Galois lattice can be understood as a basic answer to the question of finding an appropriate classification of objects; for example, the work in (Carpineto and Romano, 1993) presents an approach to conceptual clustering and shows that this system, named GALOIS, can be used for class discovery and class prediction. Moreover, (Sahami, 1995) presents an induction algorithm that induces classification rules using a Galois lattice; the author shows that this algorithm is also capable of learning decision lists. Other works introduce this lattice-based method as a formalization of induction systems (Ngäfo and Njouwa, 1997), or as a methodology for a graph-based learning system (Gonzalez et al., 2001). Finally, in the field of Knowledge Discovery concept lattices are also of great importance: the closed set of items in each one of the concepts gathers all the characteristics of the well-known frequent itemsets of the datamining area. Also in datamining, concept lattices are proved to be an expressive modelling technique to show the structural relations implicit in the given binary data (see e.g. (Balcázar and Baixeries, 2003; Zaki, 2000)).

2 LATTICE THEORY FOR STRUCTURED OBJECTS

An important limitation of using the lattice method is the classical propositional description of the relation $R$. This is an important drawback specially in the field of Knowledge Discovery, where the objects to be mined can have a complex structure not always representable as one-valued contexts (for example, the mining of sequences, graphs, trees, molecules and so on). In this paper we contribute by describing a formal approach in terms of a Galois lattice to mine examples with an ordered structure. We are currently working towards enlarging the domain to provide ways to reason about other structured contexts.

<table>
<thead>
<tr>
<th>Seq id</th>
<th>Sequence</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$</td>
<td>$(A)(C)(D)(A)(A)(C)(A))$</td>
<td>$\alpha_1$</td>
<td>1,4,5,7</td>
<td>2,6</td>
<td>3</td>
</tr>
<tr>
<td>$s_2$</td>
<td>$(C)(B)(C)(A)(C))$</td>
<td>$\alpha_2$</td>
<td>4</td>
<td>2</td>
<td>1,3,5</td>
</tr>
<tr>
<td>$s_3$</td>
<td>$(C)(B)(A)(B)(C)(C)(A)(A))$</td>
<td>$\alpha_3$</td>
<td>3,7,8</td>
<td>2,4</td>
<td>1,5,6</td>
</tr>
<tr>
<td>$s_4$</td>
<td>$(A)(B)(C)(D)(C)(C)(A)(A))$</td>
<td>$\alpha_4$</td>
<td>1,7,8</td>
<td>2</td>
<td>3,5,6</td>
</tr>
</tbody>
</table>

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

(b) Context $\mathcal{X}$ for $\mathcal{D}$

Figure 1: Example of ordered data and its context

As a first step, we start by focusing on the most basic case, that is, when items keep an order in each one of the objects of the context. Formally, objects of the context are sequences of sets of items, that is, each input sequence $s$ is a list $s = \langle (I_1) \ldots (I_n) \rangle$ where each $I_j \subseteq \mathcal{I}$ is an itemset occurring before another itemset $I_j \subseteq \mathcal{I}$ if $i < j$. A simple example can be found...
Figure 2: Concept lattice for the ordered context

in data from figure 1(a), where each transaction is a sequence of single itemsets; the data in this example will be used throughout the paper to illustrate the results obtained with this framework. Formally, we can represent this kind of data with a new context (as it is showed in 1(b)): \( R \) is a ternary relation \( R \subseteq O \times I \times N \), where each tuple \( (o, i, t) \) indicates that the item \( i \) appears in the \( t \)-th position of the object \( o \) representing an input sequence \( s \). This is called an ordered context.

We already made available in (Casas-Garriga, 2003) the Galois connection for the ordered context, and the respective closure operator working on the universe of sets of sequences. In general, the closure of a set of sequences \( S \) includes all the maximal sequences that are present in all objects having all those sequences in \( S \). For example, from data in figure 1 we have that the closure of the set of sequences \( \{(A),\{(C)(C)\}\} \) is another set of sequences \( \{(A)(C)\}, \{(C)(C)(A)\} \) because all the maximal subsequences contained in those objects where \( \langle (A) \rangle \) and \( \langle (C)(C) \rangle \) belong together are \( \langle (A)(C) \rangle \) and \( \langle (C)(C)(A) \rangle \). Then, closed set of sequences are those coinciding with their closure. Similarly to any other Galois connection, it is possible to define a dual closure operator working on the universe of objects.

Each pair \((S, O)\) of a closed set of sequences \( S \) and the dual closed set of objects \( O \) (coinciding with the list of object identifiers where \( S \) is completely contained) are linked by the new Galois connection, and they form the new formal concept of this ordered context. Again, the Galois lattice is a Hasse diagram where each node is a closed set of sequences \( S \) labelled with the corresponding closed set of objects, \( O \). We depict graphically in figure 2 the corresponding lattice of formal concepts for data in figure 1. Specificity relation between nodes is defined by the proper ordering \( \preceq \): \( S \preceq S' \) if and only if \( \forall s \in S \exists s' \in S' \) s.t. \( s \subseteq s' \). An artificial top is usually added to the lattice, representing the set of sequences not belonging to any object, we note it by \( D \). The properties of the Galois connection will
ensure that this defined system is closed.

We also showed in (Casas-Garriga, 2003) that the closed set of sequences fitting in a formal concept are a particular case of so-called stable sequences, which enjoy several important properties. Stable sequences are those maximal sequences not extendable to others occurring in the same set of objects. For instance, $\langle (A)(C) \rangle$ is stable since it is the maximal one appearing in all the objects of the data; however, $\langle (C)(C) \rangle$ is not stable since it can be extended to $\langle (C)(C)(A) \rangle$ in all the objects where it is contained. The set of all stable sequences from data in figure 1 are represented in the table from figure 3. Algorithmically, these stable sequences can be mined by many existing algorithms, such as CloSpan (Yan et al., 2003) or BIDE (Wang and Han, 2003).

We observe that each individual sequence belonging to a closed set of sequences of the lattice is stable, and that all stable sequences belong to some node of the graph. Therefore, it is possible to construct the Galois lattice for the ordered context out of the raw stable sequences mined by algorithms such as CloSpan or BIDE among others. Note however, that the organization of stable sequences into our lattice is not a trivial task, since some stable sequences can belong to different closed sets of sequences; e.g. the stable sequence $\langle (C)(C)(A) \rangle$ in figure 3 belongs to two different nodes. We are still working towards the definition of an algorithmic solution for this problem.

Our next observation is that these groups of stable sequences belonging to the nodes of the lattice represent exactly the maximal paths of a partial order compatible with the closed set of objects linked to the same node. Broadly speaking, each node of the lattice returns a closed partial order as it is depicted in figure 4. For example, the partial order generated
from the bottom of this lattice will turn out to be compatible with all the input objects, as it corresponds to the object list associated with the node. In practice, this transformation implies that algorithms for mining stable sequences can transform their patterns into closed partial orders, and so, the mining operation of these tree-like structures directly from the data is clearly simplified.

3 Further Work

As a next step to complete the present proposal, we want to make available the convenient algorithm to construct our Galois lattice out of the raw stable sequences. As mentioned above, this is not a trivial task due to the non-bijective correspondance between nodes and stable sequences. Our idea is that the properties of the Galois connections will help to characterize a way of identifying those stable sequences that must be placed into several nodes of the lattice to keep the closure. The definition of this algorithm is an important task, since it will lead to the identification of those groups of stable sequences forming the final closed partial orders. Another observation after having the lattice constructed, is to realize that the resulting nodes provide a clusterization of input sequences by using stable sequences as features. In this line of research it can be interesting to compare our clustering system to other existing sequential clustering algorithms.

Other theoretical problems arising from this structure is how to characterize the tran-
formation of stable sequences into maximal paths of closed partial orders. The work in (Casas-Garriga and Balcázar, 2004) formalizes a similar transformation (but for a restricted case of not having repetition of items) in terms of the coproduct operation from category theory. Ongoing work is also to settle all the theoretical details for the general case.

Finally, further work must be the generalization of the Galois connection for more complex structured objects, such as graphs, molecules, trees, and so on. The difficulty arises when defining the specific Galois connection for these complex objects, and also when formalizing the comparison and generalization operation for the new language.

Acknowledgements:
The authors thank Pablo Díaz for helping with the implementations.

References