

Ontology elicitation: Structural Abstraction = Structuring + Abstraction + Multiple Ontologies

Adrian Silvescu and Vasant Honavar
Computer Science Dept, Iowa State University, Ames, IA 50011
{silvescu|honavar}@cs.iastate.edu

The overwhelming majority of Machine Learning problem settings rely on the existence of some primitive concepts and features (ontology) in terms of which the problem, including its goals, is phrased.

Issues: In unstructured domains such as text, protein sequences, temporal sequences, images, reinforcement learning tasks, ... an ontology is not always readily available or is largely incomplete. The incompleteness problem can appear in structured domains as well - and in general we can always benefit from further "structuring" the domain under study.

As a consequence there is a need for automatic methods that can "structure" the abovementioned domains and therefore provide learning algorithms with a useful feature/conceptual/relational representation. In fact, the lack of good methods for dealing with this problem is one of the main critiques against strong-AI [Haugeland, (1991)].

Main objectives: The design, analysis and development of automated and semi-automated methods that enable ontology elicitation ("structuring") / feature construction in unstructured and structured application domains. Furthermore, the development of validation methods for the proposed ontologies derived in the previous step relative to one or more predictive tasks. And also, examining ways to incorporate biases corresponding to one or more particular tasks at hand, in the process of deriving ontologies.

Initial Problem: The problem of identifying relevant features and ontologies in discrete sequence data is studied in two application domains: protein sequences and text data. An appropriate solution for this problem will result in improved performance for a variety of predictive task definable over these domains aside from providing more insight into their very structure. Some examples of predictive tasks are protein function prediction and protein-protein interaction in the first case and text categorization and information extraction in the second case respectively.

Proposed Initial Method: As an initial approach to solving this problem we use the following paradigm: "There are at least two essential "moves" that people make when attempting to make sense of/ "structure" a new application domain: Super-Structuring and Abstraction. Super-Structuring is the process of grouping and subsequently naming a set of entities that occur within "proximity" of each other, into a more complex structural unit. Abstraction, on the other hand, establishes that a set of entities belong to the same category, based on a certain notion of "similarity" among these entities, and subsequently names that category." More precisely, in the case of sequence data, given a set of sequences (text sentences, protein sequences, ...) one way to "operationalize" the above paradigm is the following:

```
until a limit criteria has been reached
  top_ka_abstractions = Abstract(sequence_data)
  top_ks_structures = SuperStructure(sequence_data)
  sequence_data = Annotate sequence_data with the new abstractions and structures
repeat
```

where Abstraction and SuperStructuring are "operationalized" as follows:

SuperStructure($S \rightarrow A * B$) - returns the topmost ks structures made out of two components that co-occur within a small distance of each other and are unlikely to occur by chance (as measured by the KL divergence between the probability of two components occurring together and the probability of them occurring together as independent events - $KL(P(A * B) | P(A)P(B))$).

Abstraction($S \rightarrow A | B$) - returns the topmost ka abstractions (clusters) of two entities (ranked according to the distance between the contexts in which the two abstracted entities appear (more exactly, we define the distance between the left contexts of the two entities, as the Jensen-Shannon distance between the probability distribution of entities that occur to the left of each entity similarly for the right contexts). Subsequently, in order to obtain a distance between two entities we sum the distances corresponding to the left and right context)

Example:

Step1 data: Mary loves John.
 Sue loves Curt.

Mary hates Curt.

Abstractions 1: A1 -> Mary | Sue because they have similar right contexts: loves.
 A2 -> John | Curt because they have similar left contexts: loves.

Step 2 data: [Mary, A1] loves [John, A2].
 [Sue, A1] loves [Curt, A2].
 [Mary, A1] hates [Curt, A2].

Abstractions 2: A3 -> loves | hates because of high similarity between their left and right contexts:
 This illustrates how abstraction begets more abstraction (A3 not possible on the raw data).

Step 3 data: [Mary, A1] [loves, A3] [John, A2].
 [Sue, A1] [loves, A3] [Curt, A2].
 [Mary, A1] [hates, A3][Curt, A2].

Structures 3: S1 -> A1 A3 because it occurs three times
 S2 -> A3 A2 because it occurs three times

This illustrates how abstraction begets structuring (S1 and S2 not possible on the raw data)

Structures 4: S3 -> S1 A2
 S4 -> A1 S2

This illustrates how structuring begets more structuring

Note that the abstractions and super-structures derived at one step beget more abstractions and super-structuring at subsequent steps, which cannot to be derived directly, based on the initial raw data. Note also that we allow multiple ontologies/points of view by allowing one entity to be abstracted or super-structured in more than one way (e.g. A3 in S1 and S2), unlike compression approaches such as [2].

Validation: In order to estimate the usefulness of the structures and abstractions derived we use them as additional features for a predictive task and measure the improvement in generalization accuracy versus the base case where we use only the primitive features (words only in the text domain example).

Experiments: Two application domains are examined: protein sequences and text. The two predictive tasks used for validation are protein function prediction and text categorization and the learning algorithm used is Naive Bayes. The preliminary experiments are directed towards assessing the improvements that can be gained in terms of generalization accuracy, by using as additional features the results of either abstraction alone, structuring alone, or the combination of the two. This assessment versus the base case where we use primitive features only, all of them in a “bag of features” setup.

Brief Literature review: Even though some of the work presented in [Jonyer et al., 2002], [Srikant and Agraval, 1995], [Maedche, 2002] shares some commonalities in spirit and even in form with the proposed method; none of them examines the interaction between super-structuring, abstraction and multiple ontologies in its full generality.

Future work:

- Validation on more predictive tasks (interaction site prediction and information extraction respectively) and on more than one task at a time (~ multitask learning) + incorporation of the task(s) bias into the feature / ontology derivation loop.
- Identifying the extensions of well separated concepts from the given structures and abstractions of two entities derived by the proposed method and further generalizing these concept extensions by learning corresponding intentional definitions and thus truly completing the ontology elicitation process.
- generalizing the approach in order to incorporate more operations (notably sub-structuring)
- generalizing the method from sequences to arbitrary topologies (e.g., in order to accommodate relational models)
- Evaluate and analyze the methods by defining a generative model based on a certain ontology and then assess how well the ontology is “retrieved” by the proposed method. The same for real domains where a partial ontology is available (e.g., WordNet for text).
- Exploring other application domains such as Metabolic Pathways, reinforcement learning, ...

[Haugeland, (1991) John Haugeland (ed.), Mind Design II, 1991.

[Jonyer et al., 2002] I. Jonyer, L. B. Holder, and D. J. Cook, "Concept Formation Using Graph Grammars", KDD Workshop on Multi-Relational Data Mining, 2002.

[Srikant and Agraval, 1995] R. Srikant, and R. Agraval, Mining generalized associations rules. In Proc of VLDB'95, 1995.

[Maedche, 2002] Alexander Maedche, Ontology learning for the semantic web, 2002.