Feature Extraction with Description Logics Functional Subsumption

Rodrigo de Salvo Braz Dan Roth

Abstract

Most efficient machine learning algorithms rely on examples represented in a propositional form, such as feature vectors. However, most data in real world problems are more complex than that, being more conveniently described by sets of objects with attributes and relations between them. The most common approach to this problem is the extraction of propositional features by ad hoc methods, but this approach fails to regard feature extraction as a significant inference step of the process, making it difficult to think about it in a uniform way and reutilize ideas across different tasks, and to analyze the complexity of extraction and learning in an integrated fashion.

Inductive Logic Programs (ILP) and related solutions work directly with complex (first-order logic or similar) representations, but are much more costly than propositional algorithms due to huge search spaces, and in general do not take much advantage of the accumulated knowledge on propositional learning.

Propositionalization methods explore the middle ground by generating propositional features from complex data, distinguishing themselves from ad hoc generation by the use of a formal language, and from ILP by making use of background knowledge provided by humans rather than searching large spaces for useful features. The choice of language for describing abstract features to be generated in propositionalization methods is crucial, however, for one must use a language expressive enough to be useful while limiting that expressivity in convenient ways in order to keep the inference involved tractable.

Since the 1980's, Description Logics (DL) have been known for striking a good balance between expressivity and tractability, and were chosen by Cumby and Roth [CR03] in their Feature Description Logic framework. In this framework, DL descriptions are used to specify abstract features whose instances are to be identified in complex data, generating corresponding propositional features. At the core of this identification is the comparison of partial descriptions of objects according to their attributes.

In this paper we present an extension to this framework which allows us to use a general binary function for comparing those attributes instead of plain identity. This greatly improves expressivity since such a function can be used to describe varied forms of background knowledge not easily expressed by DLs. It also increases efficiency in the sense that it allows us to express certain concepts already expressable by DLs, but in a form that can be processed in less time. These advantages come at little cost, namely including a factor in the time complexity equal to the cost of calculating the function on two given attributes (typically a constant on the size of descriptions).

References

[CR03] Chad Cumby and Dan Roth. Learning with feature description logics. In S. Matwin and C. Sammut, editors, *Proceedings of the 12th International Conference on Inductive Logic Programming*, pages 32–47. Springer-Verlag, 2003.