If constraint-based mining is the answer: what is the constraint? (invited talk)

Jean-François Boulicaut
University of Lyon, CNRS
INSA-Lyon, LIRIS UMR5205, 69621 Villeurbanne, France
jean-francois.boulicaut@insa-lyon.fr

Abstract

Constraint-based mining has been proven to be extremely useful. It has been applied not only to many pattern discovery settings (e.g., for sequential pattern mining) but also, recently, on classification and clustering tasks (see, e.g., [1]). It appears as a key technology for an inductive database perspective on knowledge discovery in databases (KDD) [5, 3], and constraint-based mining is indeed an answer to important data mining issues (e.g., for supporting a priori relevancy and subjective interestingness but also to achieve computational feasibility). However, few authors study the nature of constraints and their semantics. Considering several examples of non trivial KDD processes, we discuss the Hows, Whys, and Whens of constraints in a broader context than [2]. Our thesis is that most of the typical data mining methods are constraint-based techniques and that it is worth studying and designing them as such. In many cases, we exploit constraints that are not really explicit (e.g., the objective function optimization of a clustering for a given similarity measure) and/or constraints whose operational semantics are relaxed w.r.t. their declarative counterparts (e.g., the optimization constraint is not enforced because of some local optimization heuristics). We think that is important to explicit every primitive constraint and the operators that combine them because this constitutes the declarative semantics of the constraints and thus the mining queries. Then, a well-studied challenge is to design some operational semantics like correct and complete solvers and/or relaxation schemes for more or less complex constraints. Designing complete solvers has been extensively studied in useful but yet limited settings (see, e.g., the algorithms for exploiting combinations of monotonic and anti-monotonic primitives [4]). It is however clear that many relevant constraints lack from such nice properties. On another hand, understanding constraint relaxation strategies remains fairly open, certainly because of its intrinsically heuristic nature. Interestingly, the recent approaches that suggest global pattern or model construction based on local patterns enable to revisit the relaxation issue thanks to constraint back propagation possibilities. This can be discussed within a case study on constrained co-clustering [6]. This talk is based on results from two IST FET Open European projects dedicated to inductive databases, namely the CINQ *(2001-2004) and its follow-up* IQ *(2005-2008). This research is partly funded by ANR Bingo2 (2008-2010)*

References

[1] S. Basu, I. Davidson, and K. Wagstaff. *Constrained Clustering: Advances in Algorithms, Theory and Applicationsa*. Chapman & Hall/CRC Press, Data Mining and Knowledge Discovery Series, July 2008.

[2] R. J. Bayardo. The hows, whys, and whens of constraints in itemset and rule discovery. In *Constraint-Based Mining and Inductive Databases*, pages 1–13, 2005.

[3] J.-F. Boulicaut, L. De Raedt, and H. Mannila, editors. *Constraint-Based Mining and Inductive Databases*, volume 3848 of *LNCS*. Springer, 2005. 400 pages.

[4] L. De Raedt, M. Jaeger, S. D. Lee, and H. Mannila. A theory of inductive query answering. In *Proceedings IEEE ICDM'02*, pages 123–130, 2002.

[5] T. Imielinski and H. Mannila. A database perspective on knowledge discovery. *CACM*, 39(11):58–64, 1996.

[6] R. G. Pensa, C. Robardet, and J.-F. Boulicaut. *Constrained Clustering: Advances in Algorithms, Theory and Applications*, chapter Constraint-driven Co-Clustering of 0/1 Data, pages 123–148. Chapman & Hall/CRC Press, July 2008.

978-0-7695-3503-6/08 $25.00 © 2008 IEEE
DOI 10.1109/ICDMW.2008.96