# Abstract: Success Stories of CP in Data Mining

Tias Guns

July 18, 2016

The use of constraint programming technology in data mining is increasing. Especially in constraint-based pattern mining, where the goal is to find certain patterns in data, and constraint-based clustering, where the goal is to find a partition of the data.

The main trade-off in these lines of research is **efficiency versus generality**. Interestingly, in data mining, many highly efficient search methods are known, and a large body of research studies how to make them more generic. Oppositely, in constraint programming many generic search methods are known and a large body of research studies how to increase efficiency.

Just as one can make specific data mining systems more generic through effort and expertise, so can generic constraint solving systems be made as efficient, or more, than specialized mining systems. This is something that few people, especially in data mining, seem to realize.

This is a great opportunity for constraint programming though. The following four success stories aim to exemplify this.

**Success 1: constrained itemset mining.** In [4] we used a generic constraint solver to handle more constraint-based mining variants than the state-of-the-art. It was not very scalable, but outperformed specialized methods on complex tasks. In [9] a related branch-and-bound problem was addressed. An advanced bounds-consistent global propagator was introduced and the generic system outperformed the state-of-the-art in many cases. A special purpose algorithm with the same ideas outperformed both. At CP this year [7], a global constraint is proposed for a key component of all of the above, which improves scalability a lot.

Such advances are needed to make CP a viable framework for data miners, one that they will wish to use when trying to solve a new mining problem.

**Success 2: constrained sequence mining.** In [8], a generic CP approach to find constrained *sequences* in databases is proposed, but scalability was limited even when using global constraints. Soon after, [6] proposed a single global constraint that was far more scalable, and could outperform the state-of-the-art when considering regular expression constraints. Recently, we improved this global constraint to the point that a generic constraint solver outperforms all existing sequence mining methods [1].

**Success 3: constrained clustering.** In [2], CP is used for exact constrained clustering. In contrast to other approaches, it supports all common clustering constraints as well as multiple objective functions. Using global constraints for the objective functions (distance functions) achieves better efficiency then other exact techniques [3]. For the challenging problem solved heuristically by $k$-means, an iterative CP approach [5] can outperform other exact methods and it supports additional constraints.

**Outlook.** The above three data mining problems are problems that have constraints by nature, and where one is looking for all (enumeration) or the optimal solution. Both properties fit constraint programming really well.

Discovering and understanding potential benefits of CP for data mining is one issue, but explaining the benefits to data miners is another. One way that has been working well for me recently is to describe CP solvers as highly **optimized, modular depth-first search engines**. Highly optimized in their handling of backtracking and state, yet modular with the ability to add arbitrary global constraints.

While this is far from the declarative ideals of CP, it is close to current needs in data mining. It can also be a step towards achieving general-purpose approaches for both research fields.

## References

[1] John O. R. Aoga, Tias Guns, and Pierre Schaus. An efficient algorithm for mining frequent sequence with constraint programming. In *Machine Learning and Knowledge Discovery in Databases, European Conference, ECML PKDD 2016, Riva del Garda, Italy*, 2016.

[2] Thi-Bich-Hanh Dao, Khanh-Chuong Duong, and Christel Vrain. A declarative framework for constrained clustering. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III*, pages 419–434, 2013.

[3] Thi-Bich-Hanh Dao, Khanh-Chuong Duong, and Christel Vrain. A filtering algorithm for constrained clustering with within-cluster sum of dissimilarities criterion. In *2013 IEEE 25th International Conference on Tools with Artificial Intelligence, Herndon, VA, USA, November 4-6, 2013*, pages 1060–1067, 2013.

[4] Luc De Raedt, Tias Guns, and Siegfried Nijssen. Constraint programming for itemset mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-08)*, pages 204–212. ACM, 2008.

[5] Tias Guns, Thi-Bich-Hanh Dao, Christel Vrain, and Khanh-Chuong Duong. Repetitive branch-and-bound using constraint programming for constrained mss clustering. In *Proceedings of the 22nd Eureopean Conference on Artificial Intelligence, ECAI'2016*, 2016.

[6] Amina Kemmar, Samir Loudni, Yahia Lebbah, Patrice Boizumault, and Thierry Charnois. PREFIX-PROJECTION global constraint for sequential pattern mining. In *Principles and Practice of Constraint Programming - 21st International Conference, CP 2015, Cork, Ireland, August 31 - September 4, 2015, Proceedings*, pages 226–243, 2015.

[7] Nadjib Lazaar, Yahia Lebbah, Samir Loudni, Mehdi Maamar, and Valentin Lemire. A global constraint for closed itemset mining. In *Principles and Practice of Constraint Programming - 22nd International Conference, CP 2016, Toulouse*, 2016.

[8] Benjamin Négrevergne and Tias Guns. Constraint-based sequence mining using constraint programming. In *Integration of AI and OR Techniques in Constraint Programming - 12th International Conference, CPAIOR 2015, Barcelona, Spain, May 18-22, 2015, Proceedings*, pages 288–305, 2015.

[9] Siegfried Nijssen, Tias Guns, and Luc De Raedt. Correlated itemset mining in ROC space: A constraint programming approach. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-09)*, pages 647–656. ACM, 2009.