Extended abstract*

\(k\)-Pattern Set Mining under Constraints

Tias Guns, Siegfried Nijssen, Luc De Raedt
Department of Computer Science, K.U. Leuven, Belgium.
{firstname.lastname}@cs.kuleuven.be

October 4, 2010

The problem of local pattern mining can be formalised as that of finding the set of patterns \(\text{Th}(\mathcal{L}, p, D) = \{\pi \in \mathcal{L} | p(\pi, D) \text{ is true}\}\), that is, the set of all patterns \(\pi \in \mathcal{L}\) that satisfy a constraint \(p\) with respect to a database \(D\). Numerous approaches to pattern mining have been developed to effectively find the patterns adhering to a set of constraints. Despite the popularity of local pattern mining, there are still unsolved problems with this approach.

First, the interaction between the mining and the use of patterns is not well understood. The result of a pattern mining operation can almost never be used directly and needs to be post-processed in order to become useful. The reasons are that the generated set of local patterns is typically too large, and that the patterns are too hard to interpret in the global context as they interact with one another. This leads to a typical step-wise procedure in which pattern mining only forms an intermediate step in the knowledge discovery process. In the first step, the patterns adhering to some constraints are exhaustively searched for. In the second step, some patterns are selected and combined in a heuristic way. One example is associative classification, where systems such as CBA and CMAR build a classifier from a set of association rules. This raises the question as to whether it is possible to develop a single step mining approach that solves the global pattern set mining problem directly, and avoids having to generate patterns that will never be used anyway.

Secondly, there is a multitude of different global pattern mining methods and approaches, but a general solution is missing. Each combination (or system) is often tailored towards one specific task, such as concept-learning, conceptual clustering, redescription mining, tiling, etc. While these combinations are all—to some extent—based on principles of pattern mining, there exists—to the best of the authors’ knowledge—no integrated problem specification, approach or system that is able to tackle all of these tasks in a uniform way. Despite a lot of progress on specific aspects of pattern mining, this hinders our overall understanding of pattern mining. It also raises the question as to whether it is possible to provide a more general approach to data mining that can be instantiated to a wide variety of tasks such as the ones listed above. The answer to this question cannot be expected to be a highly optimized system that is competitive with much more specialized approaches on standard data mining tasks. For many types of data analysis today, run time efficiency is not a major bottleneck anymore. On the other hand, developing novel algorithms and systems to accommodate for new constraints and to tackle novel data mining tasks is still a time consuming task and requires the involvement of highly specialized data mining experts. A general purpose data mining framework and system, in which new problems can be specified in a declarative manner, would alleviate these problems and would be an adequate answer to the above question. It is precisely the goal of this paper to contribute towards such a framework by using constraint programming principles and techniques.

The key contribution of this paper is that we positively answer these two questions. We answer the first question by introducing the problem of mining sets of \(k\) patterns under constraints, that is, \(k\)-pattern sets. A \(k\)-pattern set thus consists of \(k\) patterns and the task is to find those \(k\)-pattern sets that satisfy the constraints. Constraints can be specified both at the local and at the global level that is, both at the level of individual patterns and at the level of the pattern set as a whole. In this paper we focus on patterns in

the form of itemsets, which can be regarded as conjunctions of items. A set of patterns is often interpreted as a disjunction of conjunctions, and hence corresponds to a boolean formula in disjunctive normal form (DNF). This paper studies the mining of pattern sets consisting of exactly \( k \) patterns, thus the resulting pattern sets can be conceived as \( k \)-term DNF formulas.

We also provide a positive answer to the second question by showing how instances of the tasks of concept learning, conceptual clustering, tiling, and redescription mining can all be formulated within a constraint programming (CP) framework. Constraint programming (CP) is a generic framework for solving combinatorial and optimization problems under constraints. It has been used successfully in numerous applications including constraint based mining of individual patterns. The key power of constraint programming lies in its declarative approach towards problem solving: in constraint programming, users model a problem by specifying a set of constraints, and the CP system is responsible for solving the problem. In this way, the specification of the problem is separated from the search strategy. This has the advantage that different problems can be specified by merely changing the declarative problem specification in terms of constraints. It is this approach that we will pursue in the present paper. By casting data mining problems in constraint programming models, we will show that a wide variety of problem settings can be addressed in a uniform way.

An important part of this work is the extension of our earlier work on constraint programming for local pattern mining towards the mining of entire pattern sets. Our earlier work showed that appropriate constraint programming models can lead to practical solutions for computationally hard problems, in particular for the case of correlated itemset mining in . The challenge is here to identify the modelling choices that make the approach also for pattern set mining feasible. Previous studies made a distinction between constraints on the local level (on individual patterns) and the global level (on patterns sets as a whole). We will identify two special cases, namely, local look-ahead constraints and global pairwise constraints. Such constraints are more powerful than their regular counterparts, making them crucial to reduce the search space to manageable size. This categorisation also gives us new insights into the relationship between local and global constraints.

The resulting framework for pattern set mining is very flexible and allows us to specify a wide range of problems for a wide range of data mining tasks. Examples of constraints are that patterns must be frequent, that the pairwise overlap between patterns is small, that the pattern set occurs frequently in a set of positive examples and is highly infrequent in a set of negative examples, or that the coverage of the overall pattern set in a data set is maximised.

The experiments show clearly the generality and flexibility of the constraint programming approach to \( k \)-pattern set mining. They show that the constraint programming approach can find solutions for many \( k \)-pattern set mining problems of interest, especially in those cases where the solutions are heavily constrained. On the other hand, because CP is based on exhaustive search and guarantees to find a global optimum, solving problems with few, weak constraints is often not tractable. In practice, this means that even though many problems can be modelled, the solution strategy of existing CP systems is only feasible for certain types of models. Nevertheless, also in such cases the formulation as a constraint programming model provides insights that may be used for developing more heuristic declarative approaches to the \( k \)-pattern set mining problem (e.g., using local search), and hence this work may contribute to the longer term vision of developing general purpose declarative data mining tools.