Stochastic Constraint Propagation for Mining Probabilistic Networks *

Anna Louise D. Latour[1][0000−0002−5802−8271], Behrouz Babaki[2][0000−0002−0512−4323], and Siegfried Nijssen[3]

[1] Leiden University, Leiden, The Netherlands, a.l.d.latour@liacs.leidenuniv.nl
[2] Polytechnique Montréal, Montreal, Canada, behrouz.babaki@polymtl.ca
[3] UCLouvain, Louvain-la-Neuve, Belgium, siegfried.nijssen@uclouvain.be

Abstract. A number of data mining problems on probabilistic networks can be modelled as Stochastic Constraint Optimisation and Satisfaction Problems, i.e., problems that involve objectives or constraints with a stochastic component. Earlier methods for solving these problems used Ordered Binary Decision Diagrams (OBDDs) to represent constraints on probability distributions, which were decomposed into sets of smaller constraints and solved by Constraint Programming (CP) or Mixed Integer Programming (MIP) solvers. For the specific case of monotonic distributions, we propose an alternative method: a new propagator for a global OBDD-based constraint. We show that this propagator is efficient and maintains domain consistency. We experimentally evaluate this global constraint in comparison to existing decomposition-based approaches. As test cases we use problems from the data mining literature.

This is an extended abstract of an earlier publication at IJCAI 2019 [3].

Making decisions under uncertainty is an important problem in business, governance and science. Examples are found in the fields of planning and scheduling, but also occur naturally in fields like data mining and bioinformatics.

Many of these problems can be formulated on *probabilistic networks*. Examples include signalling regulatory networks representing stochastic interactions between proteins and genes [4] and social networks [2] where we are uncertain about how likely people are to adopt ideas from others. We model this by associating probabilities with edges or nodes in the network.

We study a general class of problems, which we call *stochastic constraint optimisation or satisfaction problems on monotonic distributions* (SCPMDs). SCPMDs have the following characteristics: (1) they involve (Boolean) random variables and decision variables; (2) they can be formulated on probabilistic networks and involve the calculation of a probability or an expectation on such networks; (3) the probabilities and expectations are higher if more decision variables are selected to be *true* (monotonicity); and (4) constraints limit this selection. While (3) seems limiting, problems with this characteristic are plentiful.

Consider the following example of a *viral marketing problem* [2]. We are given a social network of people (vertices) that have stochastic relationships (edges). We want to use word-of-mouth advertisement to turn friends of our customers into new customers (a stochastic process). We start this viral campaign by distributing at most k free product samples to members of the network. What is the k-sized set S of most influential nodes in this network? Note: adding extra nodes to S cannot decrease the expected number of eventual customers (*monotonicity*).

SCPMDs like this one are NP-hard because they involve two computationally expensive tasks. We need perform probabilistic inference, which can be reduced to a #P-complete counting problem [5]. Additionally, solving SCPMDs involves traversing a search space that grows exponentially with the size of the network.

The first contribution of this work is that we show that some relations between variables are lost in the decomposition process. Consequently, this method prunes the search space inadequately and thus lacks efficiency. Specifically, it does not guarantee *generalised arc consistency* (GAC).

As our second contribution we address this flaw by introducing a *global* constraint for SCOPs whose underlying probability distributions are *monotonic*. We propose a constraint propagation algorithm for this *stochastic constraint on monotonic distributions* (SCMD), which preserves relations between variables and guarantees GAC. As in our earlier approach, we represent the probability distributions as OBDDs. We use the concept of *derivatives* of propositional formulas [1], and exploit the the structure of the OBDDs and the fact that they represent monotonic distributions, to ensure that our propagator maintains GAC. We use this SCMD propagator to develop a generic method for programming, modelling and solving SCPMDs *exactly* (our third contribution).

We demonstrate the effectiveness of this method by evaluating the running time of our new algorithm on problems from the datamining literature, comparing its performance to that of our earlier CP-based and MIP-based methods. In these experiments, our new approach outperforms the CP method, and performs complementory to the MIP method. However: the running times of our new, global constraint scale much better with problem size than this MIP method.

References

1. Darwiche, A.: On the tractable counting of theory models and its application to truth maintenance and belief revision. Journal of Applied Non-Classical Logics **11**(1-2), 11–34 (2001)
2. Kempe, D., Kleinberg, J.M., Tardos, É.: Maximizing the spread of influence through a social network. In: KDD. pp. 137–146. ACM (2003)
3. Latour, A.L.D., Babaki, B., Nijssen, S.: Stochastic constraint propagation for mining probabilistic networks. In: IJCAI. pp. 1137–1145. ijcai.org (2019)
4. Ourfali, O., Shlomi, T., Ideker, T., Ruppin, E., Sharan, R.: SPINE: A framework for signaling-regulatory pathway inference from cause-effect experiments. In: ISMB/ECCB (Supplement of Bioinformatics). pp. 359–366 (2007)
5. Roth, D.: On the hardness of approximate reasoning. Artif. Intell. **82**(1-2), 273–302 (1996)