Artificial Intelligence ••• (••••) •••–•••

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

Cost-optimal constrained correlation clustering via weighted partial Maximum Satisfiability ☆

Jeremias Berg, Matti Järvisalo *

Helsinki Institute for Information Technology HIIT, Department of Computer Science, University of Helsinki, Finland

ARTICLE INFO

Article history:
Received in revised form 18 June 2015
Accepted 3 July 2015
Available online xxxx

Keywords:
Boolean optimization
Boolean satisfiability
Maximum satisfiability
Correlation clustering
Cost-optimal clustering
Constrained clustering

ABSTRACT

Integration of the fields of constraint solving and data mining and machine learning has recently been identified within the AI community as an important research direction with high potential. This work contributes to this direction by providing a first study on the applicability of state-of-the-art Boolean optimization procedures to cost-optimal correlation clustering under constraints in a general similarity-based setting. We develop exact formulations of the correlation clustering task as Maximum Satisfiability (MaxSAT), the optimization version of the Boolean satisfiability (SAT) problem. For obtaining cost-optimal clusterings, we apply a state-of-the-art MaxSAT solver for solving the resulting MaxSAT instances optimally, resulting in cost-optimal clusterings. We experimentally evaluate the MaxSAT-based approaches to cost-optimal correlation clustering, both on the scalability of our method and the quality of the clusterings obtained. Furthermore, we show how the approach extends to *constrained* correlation clustering, where additional user knowledge is imposed as constraints on the optimal clusterings of interest. We show experimentally that added user knowledge allows clustering larger datasets, and at the same time tends to decrease the running time of our approach. We also investigate the effects of MaxSAT-level preprocessing, symmetry breaking, and the choice of the MaxSAT solver on the efficiency of the approach.

1. Introduction

Integration of the fields of constraint solving and data mining and machine learning has recently been identified within the AI community as an important research direction with high potential. This work contributes to this direction by studying the applicability of Boolean optimization to cost-optimal correlation clustering under constraints.

A common problem setting in data analysis is a set of data points together with some information regarding their pairwise similarities from which some interesting underlying structure needs to be discovered. One way of approaching the

☆ This work is supported by Academy of Finland (grants 276412, 284591, and 251170 Finnish Centre of Excellence in Computational Inference Research COIN), Doctoral Program in Computer Science DOCS, Research Funds of the University of Helsinki, and Finnish Funding Agency for Technology and Innovation (project D2I: From Data to Intelligence). The authors thank Jessica Davies for providing the MaxHS solver version used in the experiments. A preliminary version of this work appeared as [1] and was presented at the 2013 ICDM workshops. This article thoroughly revises and extends the earlier workshop paper considerable, for example by addressing the problem in a more general weighted setting, by introducing a third improved MaxSAT encoding, by extended experiments including comparisons with quadratic integer programming and several approximative algorithms, application of SAT-based preprocessing, symmetry breaking, and a MaxSAT solver comparison, as well as inclusion of full formal proofs and extended background and discussions.

* Corresponding author. Tel.: +358 50 3199 248; fax: +358 9 1915 1120.
E-mail addresses: jeremias.berg@cs.helsinki.fi (J. Berg), matti.jarvisalo@cs.helsinki.fi (M. Järvisalo).

http://dx.doi.org/10.1016/j.artint.2015.07.001
0004-3702/

problem is to attempt to divide the data into subgroups in a meaningful way, for example, so that data points in the same group are more similar to each other than to data points in other groups [2]. Discovering an optimal way of making such a division is in most settings computationally challenging and an active area of research [3]. A general term for problems of this kind is *clustering*: the groups the data is partitioned into are called *clusters*, and a partitioning of the dataset is called a *clustering* of the data.

In this work, we study the *correlation clustering* paradigm [4] in a general similarity-based setting. Correlation clustering is a well-studied [5–9] NP-hard problem. Given a labeled undirected graph with each edge labeled with either a positive or a negative label, the objective of correlation clustering is to cluster the nodes of the graph in a way which minimizes the number of positive edges between different clusters and negative edges within clusters. Taking a more general view to correlation clustering, we study the problem setting of *weighted* correlation clustering, in which each edge is associated with a weight (instead of merely a negative or positive label), indicating our confidence in that label. In the more general weighted case, the objective of correlation clustering is to minimize the sum of the weights of the positive edges between different clusters and the negative edges within clusters.

The correlation clustering paradigm is geared towards classifying data based on qualitative similarity information—as opposed to quantitative information—of pairs of data points. In contrast to other typical clustering paradigms, correlation clustering does not require the number of clusters as input. This makes it especially well-suited for settings in which the true number of clusters is unknown—which is often the case when dealing with real-world data. As a concrete example, consider the problem of clustering documents by topic without any prior knowledge on what those topics might be, based only on similarity information (edges) between pairs of different documents [4,10]. Indeed, correlation clustering has various applications in biosciences [11], social network analysis and information retrieval [12–14]. Furthermore, the related problem of *consensus clustering* [15], with recent applications in bioinformatics and in particular microarray data analysis [16–19], can also be naturally cast as correlation clustering.

Due to NP-hardness of correlation clustering [4], most algorithmic work on the problem has been heuristic, focusing on local search and approximative algorithms. While strong approximation algorithms have been proposed [4–6,9]—providing up to constant-factor approximations in restricted settings—these algorithms are unable to provide actual cost-optimal solutions in general. In this work, we take a different approach: we study the applicability of state-of-the-art Boolean optimization techniques to *cost-optimally* solving real-world instances of the correlation clustering problem. A baseline motivation for this work are the recent advances in applying constraint programming for developing generic approaches to common data analysis problems [20–25]. In a constraint programming based approach, the data analysis problem is stated in a declarative fashion within some constraint language, and then a generic solver for that language is used for solving the resulting instance.

Harnessing constraint solving for data analysis tasks has two key motivations. Firstly, declarative optimization systems allow for finding provably cost-optimal solutions. While heuristic approaches allow for scaling to very large datasets, quickly obtaining some hopefully meaningful clustering, the provably cost-optimal solutions obtained by the declarative approach can result in notably better clusterings which provide better insights into the data. This can be valuable especially when working on smaller scientific datasets which have taken years to collect [26]. Secondly, the declarative approach allows for easily integrating various types of additional constraints over the solution space at hand. This way, a user (domain data expert) may specify properties of solutions that are of interest to the user, without needing to extend available specialized algorithms in a non-trivial way to cope with such additional constraints. A constraint-based framework for clustering problems is well-suited for problem instances where some form of domain specific knowledge might be required in order to obtain meaningful clusterings. The paradigm for clustering problems of this type is known as *constrained clustering* [27–29]. Recently, Boolean satisfiability (SAT) [30] based approaches to solving constrained clustering within other clustering problems have been proposed [23,31], However, to the best of our knowledge the only work done on constrained correlation clustering is the linear programming based approach of [10]; this work is the first study on the applicability of Maximum Satisfiability (MaxSAT) [32], a well-known optimization version of SAT, to correlation clustering under constraints. The problem definition we study covers correlation clustering with additional constraints that, e.g., either force or forbid a pair of points from being assigned to the same cluster; known as must-link and cannot-link constraints [27].

1.1. Contributions

We present a novel and extensible MaxSAT-based approach to optimal correlation clustering. Using propositional logic as the declarative language, we formulate the correlation clustering task in an exact fashion as weighted partial MaxSAT [32] and apply a state-of-the-art MaxSAT solver to solve the resulting MaxSAT instance optimally. To our best knowledge this is the first practical approach to *exactly* solving correlation clustering for finding *cost-optimal* clusterings, i.e., optimal clusterings w.r.t. the actual objective function of the problem, for real-world datasets with hundreds of elements. In contrast, most of the previous work on correlation clustering has mainly focused on approximation algorithms and greedy local-search techniques which cannot in general find optimal clusterings.

At the core of the approach, we present three different MaxSAT formulations of correlation clustering, and provide formal proofs for their correctness. We experimentally evaluate our approach on real-world datasets and compare the approach to both two alternative exact approaches, based on linear and quadratic integer programming [5,33], and two approximation algorithms [5,34]. The results show that our approach can provide cost-optimal solutions and scales better than competing

Please cite this article in press as: J. Berg, M. Järvisalo, Cost-optimal constrained correlation clustering via weighted partial Maximum Satisfiability, Artificial Intelligence (2015), http://dx.doi.org/10.1016/j.artint.2015.07.001