Information Processing Letters 114 (2014) 360–364

Contents lists available at ScienceDirect

Information Processing Letters

www.elsevier.com/locate/ipl

Improved approximation algorithms for low-density instances of the Minimum Entropy Set Cover Problem

Cosmin Bonchiş [a,b], Gabriel Istrate [b,*]

[a] Department of Computer Science, West University of Timişoara, Bd. V. Pârvan 4, Timişoara, RO-300223, Romania
[b] e-Austria Research Institute, Bd. V. Pârvan 4, cam. 045 B, Timişoara, RO-300223, Romania

ARTICLE INFO

Article history:
Received 21 December 2012
Received in revised form 13 February 2014
Accepted 14 February 2014
Available online 20 February 2014
Communicated by J. Torán

Keywords:
Entropy
Set cover
Approximation algorithms

ABSTRACT

We study the approximability of instances of the *minimum entropy set cover* problem, parameterized by the average frequency of a random element in the covering sets. We analyze an algorithm combining a greedy approach with another one biased towards large sets. The algorithm is controlled by the percentage of elements to which we apply the biased approach. The optimal parameter choice leads to improved approximation guarantees when average element frequency is less than e.

1. Introduction

The *minimum entropy set cover problem* (MESC) [1] arose from a maximum likelihood approach to haplotype inference in computational biology (see also [2]). Halperin and Karp showed that the problem is NP-complete and provided an additive upper bound (equal to three) on the performance of the Greedy algorithm. This was later improved by Cardinal et al. [3], who showed a tight additive upper bound of $\log_2(e)$. Cardinal et al. [4] also studied several versions of this problem, notably minimum entropy graph coloring [5] and minimum entropy orientation [5], as well as a generalization to arbitrary objective functions [6]. Minimum entropy graph coloring has found applications to problems related to functional compression in information theory [7].

Minimum entropy set cover also lies behind a recently proposed family of measures of worst-case fairness in cost allocations in cooperative game theory [8]. This was accomplished by first studying [9] a minimum entropy version of the well-known *submodular set cover problem* [10,11]. Submodularity corresponds in the setting of cooperative game theory to *concavity* of the associated game, a property that guarantees many useful features of the game such as the non-emptiness of the core, membership of the Shapley value in the core, equivalence between group-strategyproofness and cross-monotonicity in mechanism design [12] and so on.

In this paper we impose an additional restriction on MESC: we parameterize its instances by f (formally defined below), the average number of sets that cover a random element. The previously studied minimum entropy orientation problem [5] corresponds to a special case of MESC with $f = 2$. With this additional restriction we provide approximation guarantees that often improve on those valid for the Greedy algorithm. To accomplish this goal we study the performance of an approximation algorithm BiasedGreedy(δ) parameterized by a constant $\delta \in [0, 1]$.

Our main result can be summarized in the following way: we give general upper bounds on the performance of our proposed algorithm. These bounds improve on the approximation guarantee of the greedy algorithm when average element frequency is less than the constant e.

* Corresponding author.
E-mail address: gabrielistrate@acm.org (G. Istrate).

http://dx.doi.org/10.1016/j.ipl.2014.02.006
0020-0190/

INPUT: An instance $(U, \mathcal{P})$ of MESC and
a real δ with $0 \leqslant \delta \leqslant 1$ where $|L| = \lceil \delta n \rceil$

Sort U by increasing element frequency in all subsets.
$L :=$ the set of first δn elements of U;
For all $e \in L$
choose $i_e \in [k]$ to maximize $|P_i|$ where $P_i \ni e$;
let $g(e) = i_e$;

Let $\mathcal{P}^H := \{P_1^H, P_2^H, \ldots, P_k^H\}$ where $P_i^H = P_i \setminus L$ for all $i \in [k]$
and $H = U \setminus L$

While (there exists $e \in H$)
choose $i_e \in [k]$ to maximize $|P_i^H|$ where $P_i^H \ni e$;
let $g(e) = i_e$;
erase e from all P_i^H;
$H := H \setminus \{e\}$;

OUTPUT: Cover g.

Algorithm 1. BiasedGreedy(δ).

The paper is structured as follows: in Section 2 we review basic notions and define algorithm BiasedGreedy. The main result is presented in Section 3. Its proof is given in Section 4. In Section 5 we present an application of our main result to the Minimum Entropy Graph Coloring problem.

2. Preliminaries

We will need the definition of Shannon entropy and the associated *Kullback–Leibler divergence* of two distributions P and Q: $D(P \parallel Q) = \sum_i p_i \log_2 \frac{p_i}{q_i}$. We recall that $D(P \parallel Q) \geqslant 0$ for all P and Q.

We are concerned with the following problem:

Definition 1. [Minimum Entropy Set Cover (MESC)]: Let $U = \{u_1, u_2, \ldots, u_n\}$ be an n-element ground set, for some $n \geqslant 1$, and let $\mathcal{P} = \{P_1, P_2, \ldots, P_k\}$ be a family of subsets of U which cover U. A *cover* is a function $g : U \to [k]$ such that for every $1 \leqslant i \leqslant n$, $u_i \in P_{g(u_i)}$ ("u_i is covered by set $P_{g(u_i)}$"). The entropy of cover g is defined by:

$$Ent(g) = -\sum_{i=1}^{k} \frac{|g^{-1}(i)|}{|U|} \log_2 \frac{|g^{-1}(i)|}{|U|}. \tag{2.1}$$

[Objective]: Find a cover g of minimum entropy.

Consider an instance $(U, \mathcal{P})$ as above. Define $f = \frac{\sum_{i=1}^{k} |P_i|}{|U|}$, the *average frequency of a random element in* U.

In the algorithm below we divide the elements of the ground set into *Light* and *Heavy* elements (denoted L and H in our algorithm), based on their frequency of occurrence in all given subsets. Parameter $0 \leqslant \delta \leqslant 1$ controls this division: the least frequent δn elements are deemed *Light*, while the rest are considered *Heavy*.

Let BG be the cover generated by the BiasedGreedy(δ) algorithm, defined below, and denote by $\flat = (\flat_i)_{i \in [k]}$, $\flat_i = \frac{|BG^{-1}(i)|}{n}$ the associated probability distribution. Algorithm BiasedGreedy(δ) also induces a probability distribution $q = (q_i)_{i \in [k]}$, over the Light elements, with $q_i = \frac{|Light \cap BG^{-1}(i)|}{|L|}$.

Informally, the algorithm will first cover Light elements in a *biased* manner, simultaneously covering each such element by a set of maximum cardinality containing it. Once this phase is complete all Light elements are deleted from all sets. The Heavy elements are handled in an incremental manner via a Greedy approach. The algorithm is presented in Algorithm 1.

One could apply this algorithm to the haplotype resolution problem of Halperin and Karp [1]. In this setting a partial haplotype is simply a string over $\{0, 1, *\}^k$, for some $k > 0$. A complete haplotype $h \in \{0, 1\}^k$ and a partial haplotype $h' \in \{0, 1, *\}^k$ are compatible if they are equal on all non-$*$ positions. One can model [1] haplotype resolution as searching for a minimum entropy cover: The ground set $U = \{h_1, h_2, \ldots, h_n\}$ contains partial haplotypes and $\mathcal{S} = \{S_h \mid h \in \{0, 1\}^k\}$ is the collection of subsets of U indexed by a complete haplotype h, where $S_h = \{h_i \in U \mid h$ compatible with $h_i\}$. Biological constraints arising from the specific database of complete haplotypes could make the haplotype reconstruction problem sparse: some of the partial haplotypes obtained by DNA sequencing could be compatible with very few complete haplotypes in the database. Then the *BG* algorithm could be more suitable than Greedy, by first covering all sparse (i.e. Light) partial haplotypes in parallel. It is an interesting issue (for computational biologists, beyond the scope of this paper) whether such a restriction occurs in practice.

3. Main result

Our main result gives a computable guarantee on the performance of algorithm BiasedGreedy:

Theorem 1. *Algorithm* BiasedGreedy(δ) *produces a cover* $BG : U \longmapsto [k]$ *satisfying*:

$$Ent(BG) \leqslant Ent(OPT) - (1-\delta)\log_2 \frac{(1-\delta)}{e} + \delta \log_2 f + \delta D(q \parallel \flat) + o(1). \tag{3.1}$$

Since distribution $\flat$ depends on δ it is rather difficult to optimize over constant δ in inequality (3.1). For the same reason it is **not** clear whether the bounds we give are sharp (it is an interesting open problem to give such bounds). However for $f < e$ choice $\delta = 1$ results in a better bound than the one given by the Greedy algorithm (BiasedGreedy(0)):

Corollary 1. *The Biased algorithm, defined as the BiasedGreedy algorithm with* $\delta = 1$, *produces a cover BI whose entropy satisfies* $Ent(BI) \leqslant Ent(OPT) + \log_2 f$.

Proof. When $\delta = 1$ all elements are light, hence distributions q and $\flat$ coincide and $D(q \| \flat) = 0$. □

4. Proof of the main result

Proof. Let OPT be the optimal solution. Denote $x_i = |OPT^{-1}(i)|$ and $y_i = |OPT^{-1}(i) \cap H|$ for all $1 \leqslant i \leqslant k$. By choice of δ, $\sum_{i=1}^{k} y_i = n - \lceil \delta n \rceil \leqslant (1-\delta)n$ while $\sum_{i=1}^{k} x_i = n$.