

Similar Elements and Metric Labeling on Complete Graphs

Pedro F. Felzenszwalb
Brown University
Providence, RI, USA
pff@brown.edu

March 28, 2018

We consider a problem that involves finding similar elements in a collection of sets. The problem is motivated by applications in machine learning and pattern recognition (see, e.g. [3]). Intuitively we would like to discover something in common among a collection of sets, even when the sets have empty intersection. A solution involves selecting an element from each set such that the selected elements are close to each other under an appropriate metric. We formulate an optimization problem that captures this notion and give an efficient approximation algorithm that finds a solution within a factor of 2 of the optimal solution.

The similar elements problem is a special case of the metric labeling problem defined in [2] and we also give an efficient 2-approximation algorithm for the metric labeling problem on *complete graphs*. Metric labeling on complete graphs generalizes the similar elements problem to include costs for selecting elements in each set.

The algorithms described here are similar to the “center star” method for multiple sequence alignment described in [1].

Beyond producing solutions with good theoretical guarantees, the algorithms described here are also practical. A version of the algorithm for the similar elements problem has been implemented and used to find objects in a collection of photographs [4].

1 Similar Elements

Let X be a (possibly infinite) set and d be a metric on X . Let S_1, \dots, S_n be n finite subsets of X . The goal of the *similar elements* problem is to select an element from each set S_i such that the selected elements are close to each other under the metric d . One motivation is for discovering something in common among the sets S_1, \dots, S_n even when they have empty intersection.

We formalize the problem as the minimization of the sum of pairwise distances among selected elements. Let $x = (x_1, \dots, x_n)$ with $x_i \in S_i$. Define the similar elements objective as,

$$c(x) = \sum_{1 \leq i, j \leq n} d(x_i, x_j). \quad (1)$$

Let $x^* = \operatorname{argmin}_x c(x)$ be an optimal solution for the similar elements problem.

Optimizing $c(x)$ appears to be difficult, but we can define easier problems if we ignore some of the pairwise distances in the objective. In particular we define n different “star-graph” objective

functions as follows. For each $1 \leq r \leq n$ define the objective $c^r(x)$ to account only for the terms in $c(x)$ involving x_r ,

$$c^r(x) = \sum_{j \neq r} d(x_r, x_j). \quad (2)$$

Let $x^r = \operatorname{argmin}_x c^r(x)$ be an optimal solution for the optimization problem defined by $c^r(x)$. We can compute x^r efficiently using a simple form of dynamic programming, by first computing x_r^r and then computing x_j^r for $j \neq r$.

$$x_r^r = \operatorname{argmin}_{x_r \in S_r} \sum_{j \neq r} \min_{x_j \in S_j} d(x_r, x_j), \quad (3)$$

$$x_j^r = \operatorname{argmin}_{x_j \in S_j} d(x_r^r, x_j). \quad (4)$$

Each of the n “star-graph” objective functions leads to a possible solution. We then select from among the solutions x^1, \dots, x^n as follows,

$$\hat{r} = \operatorname{argmin}_{1 \leq r \leq n} c^r(x^r), \quad (5)$$

$$\hat{x} = x^{\hat{r}}. \quad (6)$$

Theorem 1. *The algorithm described above finds a 2-approximate solution for the similar elements problem. That is,*

$$c(\hat{x}) \leq 2c(x^*).$$

Proof. First note that,

$$c(x) = \sum_{r=1}^n c^r(x).$$

Since the minimum of a set of values is at most the average, and x^r minimizes $c^r(x)$,

$$\min_{1 \leq r \leq n} c^r(x^r) \leq \frac{1}{n} \sum_{r=1}^n c^r(x^r) \leq \frac{1}{n} \sum_{r=1}^n c^r(x^*) = \frac{1}{n} c(x^*).$$

By the triangle inequality we have

$$c(x) = \sum_{1 \leq i, j \leq n} d(x_i, x_j) \leq \sum_{1 \leq i, j \leq n} (d(x_i, x_r) + d(x_r, x_j)) = 2n \sum_{l=1}^n d(x_r, x_l) = 2nc^r(x).$$

Therefore

$$c(\hat{x}) \leq 2nc^{\hat{r}}(\hat{x}) = 2n \min_{1 \leq r \leq n} c^r(x^r) \leq 2c(x^*).$$

□

To analyze the running time of the algorithm we assume the distances $d(p, q)$ between pairs of elements in $S = S_1 \cup \dots \cup S_n$ are either pre-computed and given as part of the input, or they can each be computed in $O(1)$ time.

Let $k = \max_{1 \leq i \leq n} |S_i|$. The first stage of the algorithm involves n optimization problems that can be solved in $O(nk^2)$ time each. The second stage of the algorithm involves selecting one of the n solutions, and takes $O(n^2)$ time.

Remark 2. If each of the sets S_1, \dots, S_n has size at most k the running time of the approximation algorithm for the similar elements problem is $O(n^2k^2)$.

The bottleneck of the algorithm is the evaluation of the minimizations over $x_j \in S_j$ in (3) and (4). This computation is equivalent to a nearest-neighbor computation, where we want to find a point from a set $S \subseteq X$ that is closest to a query point $q \in X$. When the nearest-neighbor computation can be done efficiently (with an appropriate data structure) the running time of the similar elements approximation algorithm can be reduced.

2 Metric Labeling on Complete Graphs

Let $G = (V, E)$ be an undirected simple graph on n nodes $V = \{1, \dots, n\}$. Let L be a finite set of labels with $|L| = k$ and d be a metric on L . For $i \in V$ let m_i be a non-negative function mapping labels to real values. The *unweighted metric labeling* problem on G is to find a labeling $x = (x_1, \dots, x_n) \in L^n$ minimizing

$$c(x) = \sum_{i \in V} m_i(x_i) + \sum_{\{i,j\} \in E} d(x_i, x_j). \quad (7)$$

Let $x^* = \operatorname{argmin}_x c(x)$. This optimization problem can be solved in polynomial time using dynamic programming if G is a tree. Here we consider the case when G is the *complete graph* and give an efficient 2-approximation algorithm based on the solution of several metric labeling problems on star graphs.

For each $r \in V$ define a different objective function, $c^r(x)$, corresponding to a metric labeling problem on a star graph with vertex set V rooted at r ,

$$c^r(x) = \sum_{i \in V} \frac{m_i(x_i)}{n} + \sum_{j \in V \setminus \{r\}} \frac{d(x_r, x_j)}{2}. \quad (8)$$

Let $x^r = \operatorname{argmin}_x c^r(x)$. We can solve this optimization problem in $O(nk^2)$ time using a simple form of dynamic programming. First compute an optimal label for the root vertex using one step of dynamic programming,

$$x_r^r = \operatorname{argmin}_{x_r \in L} \left(\frac{m_r(x_r)}{n} + \sum_{j \in V \setminus \{r\}} \min_{x_j \in L} \left(\frac{m_j(x_j)}{n} + \frac{d(x_r, x_j)}{2} \right) \right). \quad (9)$$

Then compute x_j^r for $j \in V \setminus \{r\}$,

$$x_j^r = \operatorname{argmin}_{x_j \in L} \left(\frac{m_j(x_j)}{n} + \frac{d(x_r^r, x_j)}{2} \right). \quad (10)$$

Optimizing each $c^r(x)$ separately leads to n possible solutions x^1, \dots, x^n , and we select one of them as follows,

$$\hat{r} = \operatorname{argmin}_{r \in V} c^r(x^r), \quad (11)$$

$$\hat{x} = x^{\hat{r}}. \quad (12)$$

Theorem 3. *The algorithm described above finds a 2-approximate solution for the metric labeling problem on a complete graph. That is,*

$$c(\hat{x}) \leq 2c(x^*).$$

Proof. First note that,

$$c(x) = \sum_{r=1}^n c^r(x).$$

Since the minimum of a set of values is at most the average, and x^r minimizes $c^r(x)$,

$$\min_{1 \leq r \leq n} c^r(x^r) \leq \frac{1}{n} \sum_{r=1}^n c^r(x^r) \leq \frac{1}{n} \sum_{r=1}^n c^r(x^*) = \frac{1}{n} c(x^*).$$

Since d is a metric and m_i is non-negative,

$$\begin{aligned} c(x) &= \sum_{i \in V} m_i(x_i) + \sum_{\{i,j\} \in E} d(x_i, x_j) \\ &= \sum_{i \in V} m_i(x_i) + \sum_{(i,j) \in V^2} \frac{d(x_i, x_j)}{2} \\ &\leq \sum_{i \in V} m_i(x_i) + \sum_{(i,j) \in V^2} \left(\frac{d(x_i, x_r)}{2} + \frac{d(x_r, x_j)}{2} \right) \\ &= \sum_{i \in V} m_i(x_i) + 2n \sum_{l \in V \setminus \{r\}} \frac{d(x_r, x_l)}{2} \\ &\leq 2n \sum_{i \in V} \frac{m_i(x_i)}{n} + 2n \sum_{l \in V \setminus \{r\}} \frac{d(x_r, x_l)}{2} \\ &= 2nc^r(x). \end{aligned}$$

Therefore

$$c(\hat{x}) \leq 2nc^{\hat{r}}(\hat{x}) = 2n \min_{1 \leq r \leq n} c^r(x^r) \leq 2c(x^*).$$

□

The first stage of the algorithm involves n optimization problems that can be solved in $O(nk^2)$ time each. The second stage involves selecting one of the n solutions, and takes $O(n^2)$ time.

Remark 4. *The running time of the approximation algorithm for the metric labeling problem on complete graphs is $O(n^2k^2)$.*

Acknowledgments

We thank Caroline Klivans, Sarah Sachs, Anna Grim, Robert Kleinberg and Yang Yuan for helpful discussions about the contents of this report. This material is based upon work supported by the National Science Foundation under Grant No. 1447413.

References

- [1] Dan Gusfield. Efficient methods for multiple sequence alignment with guaranteed error bounds. *Bulletin of Mathematical Biology*, 55(1):141–154, 1993.
- [2] Jon Kleinberg and Eva Tardos. Approximation algorithms for classification problems with pairwise relationships: Metric labeling and markov random fields. *Journal of the ACM*, 49(5):616–639, 2002.
- [3] Oded Maron and Aparna Lakshmi Ratan. Multiple-instance learning for natural scene classification. In *International Conference on Machine Learning*, volume 98, pages 341–349, 1998.
- [4] Sarah Sachs. Similar-part approximation using invariant feature descriptors. Undergraduate Honors Thesis, Brown University, 2016.