Doctopic: Algorithms, automata, complexity and games ARTICLE IN PRESS TCS:11464

Theoretical Computer Science ••• (••••) •••–•••

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

UPGMA and the normalized equidistant minimum evolution problem

Vincent Moulton [a], Andreas Spillner [b], Taoyang Wu [a,*]

[a] School of Computing Sciences, University of East Anglia, Norwich, NR4 7TJ, UK
[b] Department of Engineering and Natural Sciences, Merseburg University of Applied Sciences, Germany

ARTICLE INFO

Article history:
Received 12 April 2017
Received in revised form 22 January 2018
Accepted 23 January 2018
Available online xxxx
Communicated by V.Th. Paschos

Keywords:
UPGMA
Minimum evolution
Balanced minimum evolution
Hierarchical clustering

ABSTRACT

UPGMA (Unweighted Pair Group Method with Arithmetic Mean) is a widely used clustering method. Here we show that UPGMA is a greedy heuristic for the normalized equidistant minimum evolution (NEME) problem, that is, finding a rooted tree that minimizes the minimum evolution score relative to the dissimilarity matrix among all rooted trees with the same leaf-set in which all leaves have the same distance to the root. We prove that the NEME problem is NP-hard. In addition, we present some heuristic and approximation algorithms for solving the NEME problem, including a polynomial time algorithm that yields a binary, rooted tree whose NEME score is within $O(\log^2 n)$ of the optimum.

1. Introduction

Clustering (i.e. subdividing a dataset into smaller subgroups or clusters) is a fundamental task in data analysis, and has a wide range of applications (see, e.g. [1]). An important family of clustering methods aim to produce a clustering of a dataset in which the clusters form a hierarchy where the clusters nest within one another. Such hierarchies are typically represented by leaf-labeled tree structures known as dendrograms or rooted phylogenetic trees. Introduced in 1958 [2], average linkage analysis, usually referred to as UPGMA (Unweighted Pair Group Method with Arithmetic Mean), is arguably the most popular hierarchical clustering algorithm in use to date, and remains widely cited[1] and extremely popular (see, e.g. [3]). This is probably because UPGMA is conceptually easy to understand and fast in practice, an important consideration as big data sets are becoming the norm in many areas. UPGMA is commonly used in phylogenetics and taxonomy to build evolutionary trees [4, Chapter 11] as well as in related areas such as ecology [5] and metagenomics [6]. In addition, it is used as a general hierarchical clustering tool in bioinformatics and other areas including data mining and pattern recognition [7, Chapter 2].

UPGMA is a text-book algorithm that belongs to the family of agglomerative clustering methods that share the following common bottom-up scheme (cf. e.g. [4, p. 162]). They take as input a dissimilarity D on a set X, i.e. a real-valued, symmetric map on $X \times X$ which vanishes on the diagonal, and build a collection of clusters or subsets of X which correspond to a rooted tree with leaf-set X. To do this, at each step two clusters with the minimum inter-cluster dissimilarity are combined to create a new cluster, starting with the collection of clusters consisting of singleton subsets of X, and finishing when the cluster X is obtained. Different formulations of the inter-cluster dissimilarity, which specifies the dissimilarity of sets as a

* Corresponding author.
E-mail addresses: v.moulton@uea.ac.uk (V. Moulton), mail@andreas-spillner.de (A. Spillner), taoyang.wu@uea.ac.uk (T. Wu).

[1] According to Google Scholar, the method has been cited over 17,200 times during the period between 2011 and 2015.

https://doi.org/10.1016/j.tcs.2018.01.022
0304-3975/© 2018 Published by Elsevier B.V.

function of the dissimilarities observed on the members of the sets, lead to different heuristic criteria of the agglomerative methods. UPGMA, as the name average linkage analysis suggests, uses the mean dissimilarity across all pairs of elements that are contained within the two clusters. Formally, two clusters $A, B \subseteq X$ are selected for merging at each iteration step of UPGMA if the average

$$\frac{1}{|A||B|} \sum_{a \in A,\, b \in B} D(a, b)$$

is minimized over all possible pairs of clusters. Since the arithmetic mean is used, UPGMA is often more stable than linkage methods in which only a subset of the elements within the clusters are used (e.g. the single-linkage method).

UPGMA is commonly thought of as a method that greedily constructs a rooted phylogenetic tree that is closest to the input dissimilarity matrix in the least squares sense [8]. However, it is not guaranteed to do so, although it often does quite well in practice [4, p. 162]. In [9] it was shown that the related Neighbor-Joining [10] method for constructing unrooted phylogenetic trees from dissimilarity matrices can be thought of as a greedy heuristic that minimizes the so-called *balanced minimum evolution* score. Here we shall observe that (see Section 3), in a similar way, UPGMA is a greedy heuristic for computing a rooted phylogenetic tree that minimizes the so-called *minimum evolution* score [11] over all rooted phylogenetic trees on the same fixed leaf-set in which all leaves have the same distance in the tree to the root. We refer to this optimization problem as the *normalized equidistant minimum evolution* (NEME) problem, and expect that a better understanding of this problem will provide further insights into the behavior of the UPGMA algorithm.

Theoretical properties of discrete optimization problems arising in the construction of evolutionary trees have been studied for many years (for some earlier work see, e.g. [12–14]). Among these, the problems falling under the name of minimum evolution alone form a quite diverse family (see, e.g. [15]), in which the so-called balanced minimum evolution problem [16] is a particularly well-studied member. For this problem it was recently shown in [17] that for general $n \times n$-input dissimilarity matrices there exists a constant $c > 1$ such that no polynomial time algorithm can achieve an approximation factor of c^n unless P equals NP. We note that this hardness result does not rely on the often imposed restriction (see, e.g. [18,13]) that the edge lengths of the constructed tree must be integers. Moreover, in contrast to general input dissimilarity matrices, for inputs that are metrics (i.e. matrices that also satisfy the triangle inequality) a polynomial time algorithm with an approximation factor of 2 is presented in [17]. Interestingly, the proof of this approximation factor uses the fact that the balanced minimum evolution score of an unrooted tree can be interpreted as being the average length of a spanning cycle compatible with the structure of the tree [19].

Another recent, related direction of work considers the algebraic structure of the space of rooted phylogenetic trees induced by the UPGMA method (see, e.g. [8,20]). This algebraic structure is tightly linked with the property of consistency of a tree construction method, that is, those conditions under which the method is able to reconstruct a tree that has been used to generate the input dissimilarity matrix (see, e.g. [21]). In the context of our work, we are particularly interested in the consistency of methods that perform a local search of the space of all rooted phylogenetic trees on a fixed set of leaves (see, e.g. [16]). Again, balanced minimum evolution is the variant of minimum evolution for which some consistency results of this type are known [22,23].

After presenting some preliminaries in the next section, in Section 3 we begin by giving an explicit formula of the minimum evolution score of a rooted tree T as a linear combination of the input dissimilarities. This formula allows us to interpret the minimum evolution score of T in terms of the average length of a minimum spanning tree compatible with the set of clusters induced by T.

Using this observation, we explain how UPGMA can be regarded as a greedy heuristic for the NEME problem for binary rooted trees. In addition, we show that there are rooted phylogenetic trees with n leaves on which some input dissimilarity matrix has an optimal least squares fit while the NEME score of that tree for the same dissimilarity matrix is worse than the minimum NEME score by a factor in $\Omega(n^2)$. This highlights the fact that the NEME problem and searching for trees with minimum least squares fit are quite distinct problems.

Next, in Section 4, we explore solving the NEME problem by performing a local search of the space of binary rooted phylogenetic trees using so-called rooted nearest neighbor interchanges as the moves in the local search. We show that this approach is consistent. More specifically, for any input dissimilarity matrix that can be perfectly represented by a unique binary rooted phylogenetic tree with all leaves having the same distance from the root, we prove that the local search will arrive at this tree after a finite number of moves.

In Section 5 we show that the NEME problem is NP-hard even for $n \times n$ input distance matrices that satisfy the triangle inequality and only take on $O(\log n)$ different values. In light of this fact, in Section 6 we consider some approximation algorithms for solving the NEME problem. More specifically, we first show that the tree produced by UPGMA can have a score that is worse than the minimum score by a factor in $\Omega(n)$. Then, for dissimilarity matrices that satisfy the triangle inequality, we present a polynomial time algorithm that yields a binary rooted phylogenetic tree whose NEME score is within $O(\log^2 n)$ of the optimum. We conclude in Section 7 by mentioning two possible directions for future work.