Theoretical Computer Science 542 (2014) 71–82

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

On the parameterized complexity of consensus clustering ☆,☆☆

Martin Dörnfelder [a], Jiong Guo [a], Christian Komusiewicz [b,*], Mathias Weller [b]

[a] Universität des Saarlandes, Campus E 1.7, D-66123 Saarbrücken, Germany
[b] Institut für Softwaretechnik und Theoretische Informatik, Technische Universität Berlin, D-10587 Berlin, Germany

ARTICLE INFO

Article history:
Received 22 January 2013
Accepted 10 May 2014
Communicated by V.Th. Paschos

Keywords:
NP-hard problem
Data clustering
Search tree algorithm
Local search

ABSTRACT

Given a collection $\mathcal{C}$ of partitions of a base set S, the NP-hard Consensus Clustering problem asks for a partition of S which has a total Mirkin distance of at most t to the partitions in $\mathcal{C}$, where t is a nonnegative integer. We present a parameterized algorithm for Consensus Clustering with running time $O(4.24^k \cdot k^3 + |\mathcal{C}| \cdot |S|^2)$, where $k := t/|\mathcal{C}|$ is the average Mirkin distance of the solution partition to the partitions of $\mathcal{C}$. Furthermore, we strengthen previous hardness results for Consensus Clustering, showing that Consensus Clustering remains NP-hard even when all input partitions contain at most two subsets. Finally, we study a local search variant of Consensus Clustering, showing W[1]-hardness for the parameter "radius of the Mirkin-distance neighborhood". In the process, we also consider a local search variant of the related Cluster Editing problem, showing W[1]-hardness for the parameter "radius of the edge modification neighborhood".

1. Introduction

The NP-hard Consensus Clustering problem (also known as Cluster Ensemble [37] or Clustering Aggregation [17]) aims at reconciling the information that is contained in multiple clusterings of a base set S. More precisely, the input of a Consensus Clustering instance is a multi-set $\mathcal{C}$ of partitions of a base set S into subsets, also referred to as *clusters*, and the aim is to find a partition of S that is similar to $\mathcal{C}$. Herein, the similarity between two partitions is measured as follows. Two elements $a, b \in S$ are *co-clustered* in a partition C of S, if a and b are in the same cluster of C, and *anti-clustered*, if a and b are in different clusters of C. For two partitions C and C' of S and a pair of elements $a, b \in S$, let $\delta_{\{C,C'\}}(a, b) = 1$ if a and b are anti-clustered in C and co-clustered in C' or vice versa, and $\delta_{\{C,C'\}}(a, b) = 0$, otherwise. Then, the *Mirkin distance* $\operatorname{dist}(C, C') := \sum_{\{a,b\} \subseteq S} \delta_{\{C,C'\}}(a, b)$ between two partitions C and C' of S is the number of pairs $a, b \in S$ that are clustered "differently" by C and C'. The *total Mirkin distance* between a partition C and a multi-set $\mathcal{C}$ of partitions is defined as $\operatorname{dist}(C, \mathcal{C}) := \sum_{C' \in \mathcal{C}} \operatorname{dist}(C, C')$. Altogether, the Consensus Clustering problem is defined as follows.

☆ Supported by the DFG Excellence Cluster on Multimodal Computing and Interaction (MMCI) and DFG project DARE (NI 369/11).
☆☆ A preliminary version appeared in *Proceedings of the 22nd International Symposium on Algorithms and Computation*, pp. 624–633, volume 7074 of LNCS, Springer 2011.
* Corresponding author.
E-mail addresses: mdoernfe@mmci.uni-saarland.de (M. Dörnfelder), jguo@mmci.uni-saarland.de (J. Guo), christian.komusiewicz@tu-berlin.de (C. Komusiewicz), mathias.weller@tu-berlin.de (M. Weller).

http://dx.doi.org/10.1016/j.tcs.2014.05.002
0304-3975/© 2014 Published by Elsevier B.V.

Consensus Clustering
Input: A multi-set of partitions $\mathcal{C} = (C_1, \ldots, C_n)$ of a base set $S = \{1, 2, \ldots, m\}$ and an integer $t \geq 0$.
Question: Is there a partition C of S with $\mathrm{dist}(C, \mathcal{C}) \leq t$?

Consensus Clustering has a wide array of applications, for example in gene expression data analysis and classification [12,14,34], classification of electrocardiographic (ECG) test records [23], clustering categorical data [21], subtopic retrieval [8], detecting behavioral anomalies across multiple data sources [29], improving clustering robustness [14,23,37,39], and preserving privacy [17]. AbedAllah and Shimshoni [2] applied Consensus Clustering to the k-nearest neighbor classifier in machine learning, implementing heuristic data reduction techniques. The NP-hardness of Consensus Clustering was shown by Křivánek and Morávek [27] and Wakabayashi [38]. For $n = 2$, that is, with two input partitions, it is solvable in polynomial time: either input partition minimizes t. In contrast, already for $n = 3$ minimizing t is APX-hard [6]. The variant of Consensus Clustering where the output partition is required to have at most $d \geq 2$ subsets, d being a constant, is NP-hard for every $d \geq 2$ [7] but it admits a PTAS for minimizing t [7,9,22]. Various heuristics for Consensus Clustering have been experimentally evaluated [4,8,18,28,30,37]. Consensus Clustering is closely related to Cluster Editing [36], also known as Correlation Clustering [3].

So far, the study of the parameterized complexity [10,11,15,35] of Consensus Clustering seems to be neglected. One reason for this might be the lack of an obvious reasonable parameter for this problem: First, the assumption that the overall Mirkin distance of solutions is usually small is not realistic in practice: every element pair that is co-clustered in at least one partition and anti-clustered in at least one other partition contributes at least one to this parameter. Second, Consensus Clustering is trivially fixed-parameter tractable with respect to the number m of elements but m is also unlikely to take small values in real-world instances. Finally, Consensus Clustering is NP-hard for $n = 3$, ruling out fixed-parameter tractability with respect to n. Betzler et al. [5] considered the parameter “average Mirkin distance p between the input partitions”, that is, $p := \sum_{i \neq j} \mathrm{dist}(C_i, C_j)/(n(n-1))$, and presented a “partial kernelization” for this parameter. More precisely, they presented a set of polynomial-time data reduction rules whose application yields an instance with $|S| = m < 9p$ [5].[1] Then, checking all possible partitions of S gives an optimal solution, resulting in a fixed-parameter algorithm for the parameter p. The term “partial” refers to the fact that not the overall instance size is bounded but rather some “part” of the instance, in this case m. Since the Mirkin distance is a metric, the average Mirkin distance of solution partitions $k := t/n$ is at least $p/2$ [5]. Hence, the above also implies fixed-parameter tractability with respect to k. However, there are currently no efficient algorithms for parameter m (a brute-force check of all possible partitions of S leads to an impractical running time of roughly $2^{O(k \log k)} \mathrm{poly}(n, m)$).

Motivated by these observations, we study several parameterizations of Consensus Clustering. First, we complement the partial kernelization result by presenting a search tree algorithm with running time $O(4.24^k \cdot k^3 + nm^2)$. Second, we consider the parameter “maximal number of clusters in any input partition”. We show that Consensus Clustering remains NP-hard even if every input partition consists of at most two clusters, ruling out fixed-parameter tractability for this parameter. We also strengthen the hardness result of Bonizzoni et al. [7] by showing that, even if all input partitions contain at most two clusters, seeking a solution partition with at most two clusters remains NP-hard.

Finally, we consider Consensus Clustering under the local-search paradigm, which is one of the most popular approaches for solving NP-hard optimization problems. The basic idea is to improve a given solution by considering solutions in “close proximity” (with respect to some to-be-defined distance measure) to the given solution [1,33]. The combination of local search and parameterized complexity is relatively new. It has been initially considered for the Traveling Salesman problem by Marx [31], who showed W[1]-hardness for the local search variant using the k-exchange neighborhood (other neighborhoods were examined by Guo et al. [19]). On the positive side, Khuller et al. [24] showed that, the k-exchange neighborhood local search variant of the problem of finding a feedback edge set that is incident to a minimum number of vertices is fixed-parameter tractable with respect to k. Fellows et al. [13] considered local-search variants of graph problems and show that “local search versions of most graph problems are W[1]-hard or W[2]-hard on general graphs.” Further parameterized complexity results are known for local search variants of Boolean constraint satisfaction problems [26], Stable Marriage [32], Weighted Feedback Arc Set in Tournaments [16], and List Coloring [20]. In this work, we examine a canonical local search variant of Consensus Clustering, where, in addition to $\mathcal{C}$ and S, a partition C of S is given and the task is to decide whether there is a partition C' such that $\mathrm{dist}(C', \mathcal{C}) < \mathrm{dist}(C, \mathcal{C})$ and $\mathrm{dist}(C', C) \leq d$ for some integer $d \geq 0$. We show this problem to be W[1]-hard with respect to d. Moreover, our reduction can also be used to show W[1]-hardness of a natural local search variant of Cluster Editing.

Preliminaries Given a base set S and a multi-set $\mathcal{C}$ of partitions of S, let $n := |\mathcal{C}|$ and $m := |S|$. We use $\mathrm{co}(a, b)$ for $a, b \in S$ to denote the number of partitions in $\mathcal{C}$ where a and b are co-clustered and use $\mathrm{anti}(a, b)$ to denote the number of partitions where a and b are anti-clustered. Clearly, $n = \mathrm{co}(a, b) + \mathrm{anti}(a, b)$. For a partition C of S and elements $a, b \in S$, the function $\mathrm{dist}_C(a, b)$ is defined as the number of partitions in $\mathcal{C}$ in which a, b are clustered in a different way than in C. More precisely, if a and b are co-clustered in C, then $\mathrm{dist}_C(a, b) = \mathrm{anti}(a, b)$; otherwise, $\mathrm{dist}_C(a, b) = \mathrm{co}(a, b)$. Clearly, $\mathrm{dist}(C, \mathcal{C}) = \sum_{\{a,b\} \subseteq S} \mathrm{dist}_C(a, b)$.

[1] Subsequently, this was improved to a data reduction routine that yields an instance with $m < 16p/3$ [25].