A graph clustering problem with bounded number of clusters

Victor Il'ev, Svetlana Il'eva
Omsk State University, Omsk, Russia

We consider a version of the graph clustering problem, so called correlation clustering or graph approximation problem which is one of most visual formalizations of the clustering problem. The objective of the clustering problem is to partition of objects (data elements) into a family of subsets (i.e., clusters) such that objects within a cluster are more similar to one another than objects in different clusters. In the graph approximation problem one has to partition the vertices of a graph into clusters taking into consideration the edge structure of the graph: the goal is to minimize the number of edges between the clusters and the number of missing edges within the clusters. For statements and various interpretations of this problem, see [1–4].

We consider only *simple* graphs, i.e., the graphs without loops and multiple edges. A graph is called a *cluster graph* if each of its connected components is a complete graph. Denote by $\mathcal{M}_k(V)$ the set of all cluster graphs on a vertex set V consisting of exactly k nonempty connected components, $2 \leq k \leq |V|$. If $G_1 = (V, E_1)$ and $G_2 = (V, E_2)$ are graphs on the same vertex set V, then the *distance* between them is defined as $\rho(G_1, G_2) = |E_1 \setminus E_2| + |E_2 \setminus E_1|$.

The following version of the graph clustering problem is known as the *graph approximation problem* or *correlation clustering.*

Problem $\mathbf{A_k}$. Given a graph $G = (V, E)$ and an integer k, $2 \leq k \leq |V|$, find a graph $M^* \in \mathcal{M}_k(V)$ such that

$$\rho(G, M^*) = \min_{M \in \mathcal{M}_k(V)} \rho(G, M). \tag{1}$$

In machine learning clustering methods fall under the section of *unsupervised learning.* At the same time *semi-supervised* clustering methods use limited supervision. For example, relatively few objects are labeled (i.e., are assigned to clusters), whereas a large number of objects are unlabeled. This leads to the following version of the graph clustering problem.

Problem $\mathbf{A_k^+}$. Given a graph $G = (V, E)$, an integer k, $2 \leq k \leq |V|$, and a set $X = \{x_1, \ldots, x_k\} \subset V$ ($x_i \neq x_j$ unless $i = j$), find a graph $M^* \in \mathcal{M}_k(V)$ provided that minimum in (1) is taken over all cluster graphs $M \in \mathcal{M}_k(V)$ such that $x_i \in V_i$, $i = 1, \ldots, k$, where V_i is the vertex set of ith cluster (connected component) of the graph M.

Problem $\mathbf{A_k}$ is known to be *NP*-hard for any fixed integer $k \geqslant 2$ [3]. We prove that problem $\mathbf{A_k^+}$ is *NP*-hard for any fixed integer $k \geqslant 2$, and for $k = 2, 3$ we propose constant-factor approximation polynomial-time algorithms for problems $\mathbf{A_k}$ and $\mathbf{A_k^+}$.

Research of the first author was supported by RSF grant 15-11-10009.

References

[1] N. Bansal, A. Blum, S. Chawla, Correlation clustering. *Machine Learning,* **56** (2004) 89-113.

[2] V.P. Il'ev, G.Š. Fridman, On the problem of approximation by graphs with fixed number of components. *Doklady AN SSSR,* **264** (1982) 533-538 (in Russian). English transl. in *Soviet Math. Dokl.* **25** (1982) 666-670.

[3] R. Shamir, R. Sharan, D. Tsur, Cluster graph modification problems. *Discrete Appl. Math.* **144** (2004) 173-182.

[4] C.T. Zahn, Approximating symmetric relations by equivalence relations. *J. of the Society for Industrial and Applied Math.* **12** (1964) 840-847.