Mach Learn (2009) 75: 245–248
DOI 10.1007/s10994-009-5103-0

NP-hardness of Euclidean sum-of-squares clustering

Daniel Aloise · Amit Deshpande · Pierre Hansen · Preyas Popat

Received: 20 July 2007 / Revised: 23 October 2008 / Accepted: 5 January 2009 /
Published online: 24 January 2009

Abstract A recent proof of NP-hardness of Euclidean sum-of-squares clustering, due to Drineas et al. (Mach. Learn. 56:9–33, 2004), is not valid. An alternate short proof is provided.

Keywords Clustering · Sum-of-squares · Complexity

1 Introduction

Clustering is a powerful tool for automated analysis of data. It addresses the following general problem: given a set of entities, find subsets, or clusters, which are homogeneous and/or well separated. Many different criteria are used in the literature to express homogeneity and/or separation of the clusters to be found (see Hansen and Jaumard 1997 for a survey). One key criterion is the minimum sum of squared Euclidean distances from each entity to the centroid of the cluster to which it belongs, which expresses both homogeneity and separation. Note that due to Huygens' theorem this is equivalent to the sum over all clusters of the sum of all squared distances between pairs of entities within that cluster divided by

Editor: Nina Mishra.

D. Aloise (✉)
École Polytechnique de Montréal, Montreal, H3C 3A7, Canada
e-mail: daniel.aloise@gerad.ca

A. Deshpande
Microsoft Research India, Bangalore 560 080, India
e-mail: amitdesh@microsoft.com

P. Hansen
GERAD and HEC Montréal, Montreal, H3T 2A7, Canada
e-mail: pierre.hansen@gerad.ca

P. Popat
Chennai Mathematical Institute, Siruseri 603103, India
e-mail: preyas@cmi.ac.in

its cardinality. Partitioning into k clusters with this objective is known as minimum sum-of-squares clustering (MSSC). This problem is tackled by the classical k-means heuristic (MacQueen 1967) and numerous other algorithms.

The MSSC problem in general dimension for $k \geq 2$ was often referred to in the literature as NP-hard without a correct reference (see Aloise and Hansen 2007 for a detailed discussion). In particular, as shown in Sect. 2, a proof of Drineas et al. (2004) is invalid. An alternate short proof, due to the second and fourth authors (Deshpande and Popat 2008), is given in Sect. 3. Note that another longer proof was obtained independently, and almost at the same time, by Dasgupta (2008). Moreover, a proof which is essentially the same as ours was obtained independently and more recently by Kanade et al. (2008).

2 An incorrect reduction from the k-section problem

Drineas et al. (2004) propose a NP-hardness proof for the MSSC with $k = 2$ and general dimension by a reduction from the minimum bisection problem, whose objective is to partition a graph into two equal-sized parts so as to minimize the number of edges going between the two parts. The authors state that a proof for $k > 2$ is similar via a reduction to the minimum k-section problem. The paper is cited in Arthur and Vassilvitskii (2007), Beringer and Hüllermeier (2006), Cilibrasi et al. (2005), Ostrovsky et al. (2006) as giving a proof that MSSC is NP-hard.

The polynomial transformation for performing the reduction from the bisection problem is described as follows:

> "Let $G = (V, E)$ be the given graph with n vertices $1, \ldots, n$, with n even. Let $d(i)$ be the degree of the ith vertex. We will map each vertex of the graph to a point with $|E| + |V|$ coordinates. There will be one coordinate for each edge and one coordinate for each vertex. The vector X^i for a vertex i is defined as $X^i(e) = 1$ if e is adjacent to i and 0 if e is not adjacent to i; in addition $X^i(i) = M$ and $X^i(j) = 0$ for all $j \neq i$."

Figure 1 illustrates an example of such a transformation for a given graph. It can be checked in the example that all partitions with non-empty clusters have the same cost value regarding the last $|V|$ coordinates. Correcting an error in the proof presented in Drineas et al. (2004), we will show that this is always true for any MSSC instance constructed by the proposed transformation.

Let us consider a bipartition of the entities into two clusters P and Q whose cardinalities are denoted by p and q, respectively. Regarding the last $|V|$ coordinates of the centroids $z^P, z^Q \in \mathbb{R}^{|E|+|V|}$, we have for $i = 1, \ldots, |V|$

$$z^P_{|E|+i} = \begin{cases} \frac{M}{p} & \text{if } i \in P, \\ 0 & \text{otherwise,} \end{cases} \qquad z^Q_{|E|+i} = \begin{cases} \frac{M}{q} & \text{if } i \in Q, \\ 0 & \text{otherwise.} \end{cases}$$

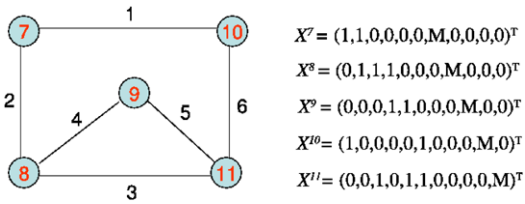

Fig. 1 Transformation of a graph into an MSSC instance as defined in Drineas et al. (2004)

Therefore, the sum of squared distances of each entity to its centroid, limited to the last $|V|$ coordinates, is equal to

$$\begin{aligned} &p\left(M-\frac{M}{p}\right)^2+q\left(M-\frac{M}{q}\right)^2+\mathbf{p(p-1)}\left(\mathbf{0}-\frac{\mathbf{M}}{\mathbf{p}}\right)^{\mathbf{2}}+\mathbf{q(q-1)}\left(\mathbf{0}-\frac{\mathbf{M}}{\mathbf{q}}\right)^{\mathbf{2}} \\ &\quad = nM^2-4M^2+M^2\left(\frac{1}{p}+\frac{1}{q}\right)+2M^2-M^2\left(\frac{1}{p}+\frac{1}{q}\right) \\ &\quad = (n-2)M^2. \end{aligned}$$

In Drineas et al. (2004), the authors forget to add the squared distances of the null components to the centroids, which are indicated in boldface in the expression. If they are not taken into consideration, then the sum-of-squares limited to the last $|V|$ coordinates is equal to

$$nM^2+M^2\left(\frac{1}{p}+\frac{1}{q}\right)-4M^2,$$

which is minimized whenever $p=q=n/2$. Thus, if M is made sufficiently large, balanced bipartitions have costs strictly smaller than unbalanced ones, since the contribution for the cost limited to the first $|E|$ coordinates is upper bounded. In fact, for $p=q$, this last value is minimized when the solution of MSSC is the balanced bipartition that corresponds to the minimum bisection in the original graph (see Drineas et al. 2004, p. 16). Unfortunately, after correcting the expression of the cost regarding the last $|V|$ coordinates, there is no dependence on the cardinalities of the clusters. This implies that the proposed reduction from minimum bisection is invalid.

3 A new proof by reduction from the densest cut problem

Nevertheless, there is a similar (valid) reduction that shows that the problem is in fact NP-hard.

Theorem 1 *MSSC in general dimension is NP-hard for* $k=2$.

Proof The reduction is from the densest cut problem, whose objective is to maximize for a given graph $G=(V,E)$ the ratio $|E(P,Q)|/|P|\cdot|Q|$ over all bipartitions (P,Q) of the vertices in G, where $E(P,Q)$ denotes the edge set of the cut. The problem is equivalent to the sparsest cut problem on the complement graph, which was shown to be NP-hard in Matula and Shahrokhi (1990).

Given a graph G with no parallel edges, let us define a $|V|$ by $|E|$ matrix M as follows. An entry (v,e) in M is equal to 0, if edge $e\in E$ is not incident to vertex $v\in V$. Otherwise, it is $+1$ for one endpoint of e and -1 for the other. It does not matter which endpoint corresponds to $+1$ and which to -1. Thus, each column of M has exactly one entry equal to $+1$ and exactly one entry equal to -1.

Now, let us suppose that the rows of M are points in $\mathbb{R}^{|E|}$ and compute the value of the MSSC criterion for a bipartition into two clusters P and Q, with $|P|=p$, $|Q|=q$ and $p+q=n$. The centroid of cluster P has in its e-th coordinate a value equal to either $+1/p$

or $-1/p$ if $e \in E(P, Q)$, or 0 otherwise. The same holds for the coordinates of the centroid of cluster Q. Then, by computing the total cost of the bipartition, we have that

$$\begin{aligned}
&\sum_{e \in E} \text{cost of } P \text{ due to the } e\text{-th coordinate} + \text{cost of } Q \text{ due to the } e\text{-th coordinate} \\
&= \sum_{e \in E(P,Q)} (p-1)\frac{1}{p^2} + \left(1 - \frac{1}{p}\right)^2 + (q-1)\frac{1}{q^2} + \left(1 - \frac{1}{q}\right)^2 + \sum_{e \notin E(P,Q)} 2 \\
&= \left(2 - \frac{1}{p} - \frac{1}{q}\right)|E(P, Q)| + 2|E(P, P)| + 2|E(Q, Q)| \\
&= 2|E| - \frac{n}{p \cdot q}|E(P, Q)|,
\end{aligned}$$

by using $p + q = n$. The MSSC for $k = 2$ minimizes the above, which means that it maximizes $|E(P, Q)|/p \cdot q$ and hence finds the densest cut in the given graph G. □

References

Aloise, D., & Hansen, P. (2007). *On the complexity of minimum sum-of-squares clustering*. Cahiers du GERAD, G-2007-50, July 2007, available online at http://www.gerad.ca.

Arthur, D., & Vassilvitskii, S. (2007). K-means++: the advantages of careful seeding. In *2007 ACM-SIAM symposium on discrete algorithms (SODA'07)*.

Beringer, J., & Hüllermeier, E. (2006). Online clustering of parallel data streams. *Data & Knowledge Engineering*, *58*, 180–204.

Cilibrasi, R., van Iersel, L., Kelk, S., & Tromp, J. (2005). On the complexity of several haplotyping problems. *Lecture Notes in Computer Science*, *3692*, 128–139.

Dasgupta, S. (2008). *The hardness of k-means clustering* (Technical Report CS2008-0916). University of California, 17 January 2008.

Deshpande, A., & Popat, P. (2008). Email sent to Ravi Kannan et al. and transmitted by Nina Mishra to the first and third authors. 22 January 2008.

Drineas, P., Frieze, A., Kannan, R., Vempala, S., & Vinay, V. (2004). Clustering large graphs via the singular value decomposition. *Machine Learning*, *56*, 9–33.

Hansen, P., & Jaumard, B. (1997). Cluster analysis and mathematical programming. *Mathematical Programming*, *79*, 191–215.

Kanade, G., Nimbhorkar, P., & Varadarajan, K. (2008). *On the NP-hardness of the 2-means problem*. Manuscript of 14 February 2008.

Matula, D., & Shahrokhi, F. (1990). Sparsest cuts and bottlenecks in graphs. *Discrete Applied Mathematics*, *27*, 113–123.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of 5th Berkeley symposium on mathematical statistics and probability* (Vol. 2, pp. 281–297), Berkeley, CA.

Ostrovsky, R., Rabani, Y., Schulman, L. J., & Swamy, C. (2006). The effectiveness of Lloyd-type methods for the k-means problem. In *Proceedings of the 47th annual IEEE symposium on foundations of computer science (FOCS'06)*.