*Les Cahiers du GERAD* ISSN: 0711–2440

# On the Complexity of Minimum Sum-of-Squares Clustering

D. Aloise
P. Hansen

G–2007–50

July 2007

Les textes publiés dans la série des rapports de recherche HEC n'engagent que la responsabilité de leurs auteurs. La publication de ces rapports de recherche bénéficie d'une subvention du Fonds québécois de la recherche sur la nature et les technologies.

# On the Complexity of Minimum Sum-of-Squares Clustering

**Daniel Aloise**
*GERAD and Département de mathématiques et génie industriel*
*École Polytechnique de Montréal*
*C.P. 6079, Succ. Centre-ville*
*Montréal (Québec) Canada, H3C 3A7*
daniel.aloise@gerad.ca

**Pierre Hansen**
*GERAD and Méthodes quantitatives de gestion*
*HEC Montréal*
*3000, chemin de la Côte-Sainte-Catherine*
*Montréal (Québec) Canada, H3T 2A7*
pierre.hansen@gerad.ca

July 2007

*Les Cahiers du GERAD*

G–2007–50

## Abstract

To the best of our knowledge, the complexity of minimum sum-of-squares clustering is unknown. Yet, it has often been stated that this problem is NP-hard. We examine causes for such confusion and in the process show that a recent proof of Drineas et al. in *Machine Learning* 56, 9–33, 2004 is not valid and unlikely to be salvaged.

**Key Words:** clustering, sum-of-squares, complexity.

## Résumé

Au meilleur de notre connaisance, la complexité de la classification avec le critère de moindre somme des carrés des erreurs est inconnue. Cependant, il a été souvent affirmé que ce problème est NP-difficile. Nous examinons les causes de telles confusions et ce faisant montrons qu'une preuve récente de Drineas et al. publiée dans *Machine Learning* 56, 9–33, 2004 n'est pas correcte et qu'il est peu probable qu'elle puisse être réparée.

**Mots clés :** classification, somme de carrés des erreurs, complexité.

# 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 [27, 30, 40]. Homogeneity means that entities in the same cluster must be similar and separation that entities in different clusters must differ one from another.

One of the most used types of clustering is *partitioning*, where given a set $O = \{o_1, o_2, \ldots, o_n\}$ of $n$ entities, we look for a partition $P_k = \{C_1, C_2, \ldots, C_k\}$ of $O$ into $k$ clusters such that

- $C_j \neq \emptyset \quad j = 1, \ldots, k;$
- $C_{j_1} \bigcap C_{j_2} = \emptyset \quad j_1, j_2 = 1, \ldots, k$ and $j_1 \neq j_2$;
- $\bigcup\limits_{j=1}^{k} C_j = O.$

Many different criteria are used in the literature to express homogeneity and/or separation of the clusters to be found (see [24] for a survey). For instance, one may desire to maximize the *split* of a partition, i.e., the minimum dissimilarity between two entities assigned to two different clusters [8, 15], or to minimize the *diameter*, i.e., the largest dissimilarity between a pair of entities in the same cluster [23]. Among these criteria, the minimum sum of squared distances from each entity to the centroid of the cluster to which it belongs, or minimum sum-of-squares for short, is one of the most used, with the advantage that it expresses both homogeneity and separation (see [55], pages 60–61).

A mathematical programming formulation of the minimum sum-of-squares clustering (MSSC) problem is as follows:

$$\begin{array}{l} \min\limits_{w,z} \quad \sum\limits_{i=1}^{n} \sum\limits_{j=1}^{k} w^{ij} \|X^i - z^j\|^2 \\ \text{subject to} \\ \sum\limits_{j=1}^{k} w^{ij} = 1, \quad \forall i = 1, \ldots, n \\ w^{ij} \in [0,1], \quad i = 1, \ldots, n; j = 1, \ldots, k. \end{array}$$

The $n$ entities $\{o_1, o_2, \ldots, o_n\}$ to be clustered are at given points $X^i = (X^i_r, r = 1, \ldots, s)$ of $\mathbb{R}^s$ for $i = 1, \ldots, n$; $k$ cluster centers must be located at unknown points $z^j \in \mathbb{R}^s$ for $j = 1, \ldots, k$; the norm $\| \cdot \|$ denotes the Euclidean distance between the two points in its argument, in the $s$-dimensional space under consideration. The decision variables $w^{ij}$ express the assignment of the entity $o_i$ to the cluster $j$. We assume that the number of entities $n$ is greater than $k$, otherwise the problem is trivially solved by locating the cluster centers at each entity position.

The problem is also sometimes referred to in the literature as the *discrete clustering problem* or the *hard clustering problem.* Besides, it is well-known as the problem tackled by the classical $k$-means clustering algorithm [35]. From an initial partition, $k$-means proceeds by reassigning the entities to their closest centroids and updating their positions until stability is reached. Although it does not provide the global optimum solution, $k$-means is popular due to its simplicity and fast local optimum convergence observed in practice. Moreover, it takes advantage of some mathematical properties of the MSSC formulation.

If $z$ is fixed, the condition $w^{ij} \in [0,1]$ can be replaced by $w^{ij} \in \{0,1\}$, since in an optimal solution for the resultant problem each entity belongs to the cluster with the nearest centroid. This is exactly what $k$-means does after recomputing the centroids. Besides, for a fixed $w$, the gradient of the objective function requires that at an optimal solution

$$\sum_{i=1}^{n} w^{ij}(z_r^j - X_r^i) = 0, \quad \forall j, r,$$

i.e.,

$$z_r^j = \frac{\sum\limits_{i=1}^{n} w^{ij} X_r^i}{\sum\limits_{i=1}^{n} w^{ij}}, \quad \forall j, r.$$

In other words, it states that the optimal cluster centers are always at the centroids of the clusters. The $k$-means heuristic recomputes the centroids of the clusters whenever reassignments are performed, thereby improving the cost of the MSSC solution according to the optimality condition just mentioned. Because reassignments are performed only if profitable and the number of partitions is finite, we can conclude that $k$-means always converges to a local minimum.

## 2 Computational complexity

First of all, it is important to remark that the computational complexity of a clustering problem depends on the criterion used. For instance, split maximization is polynomially solvable [8] while diameter minimization is NP-hard [4, 23]. However, several authors wrongly state that partitioning is NP-hard regardless of the clustering criterion [5, 7, 17, 18, 21, 22, 28, 33, 45, 51, 53, 52, 56, 58].

To the best of our knowledge, the computational complexity of minimum sum-of-squares clustering in the Euclidean metric for general values of $k$ and $s$ is still unknown. However, several incorrect statements have been made about this problem being known to be NP-hard, many of them without providing a reference [14, 16, 34, 36, 46, 47, 49, 54, 60, 61].

Some confusion is also made [5, 9, 17, 18, 22, 32, 41] by referring to a paper of Garey et al. [20], which provides a NP-hardness proof for the *quantization* problem by a reduction from the exact covering problem by triples, which is known to be a NP-complete problem [19]. The quantization problem is defined in [20] as follows.

> "A source produces one sample of a random variable $X$ with equiprobable values in $\{1, 2, \ldots, n\}$.
> The encoder (quantizer) maps $X$ into a variable $Y$ with values in $\{1, 2, \ldots, k\}$. The decoder maps $Y$ into a decision variables $Z$ with values in $\{1, 2, \ldots, m\}$. If $X = i$ and $Z = j$ the resulting distortion is $d_{ij}$. All entries in the $n \times m$ matrix $[d_{ij}]$ are zeros or ones. The goal is to find an encoder function, $f : X \rightarrow Y$, and a decoder function, $g : Y \rightarrow Z$, such that the average distortion
>
> $$\frac{1}{n}\sum_{i=1}^{n} d_{ig(f(i)))}$$
>
> is as small as possible."

However, this is in fact a particular $k$-median problem (see e.g. [31] for a survey) where each cluster center is taken from a given finite set of fixed potential locations. This problem was already known to be NP-hard for general values of $k$ [29].

Other results due to Brücker [4] led to further confusion. This author proved that the partitioning problem is NP-hard for many different clustering criteria. In the classical book *Computers and Intractability* of Garey and Johnson [19], this paper is referenced in the following way:

> "[MS9] CLUSTERING
> INSTANCE: Finite set $X$, a distance $d(x, y) \in Z_0^+$ for each pair $x, y \in X$, and two positive integers $K$ and $B$.
> QUESTION: Is there a partition of $X$ into disjoint sets $X_1, X_2, \ldots, X_k$ such that, for $1 \leq i \leq k$ and all pairs $x, y \in X_i$, $d(x, y) \leq B$?
> Reference: **[Brücker, 1978]** Transformation from GRAPH 3-COLORABILITY.
> Comment: Remains NP-complete even for fixed $K = 3$ and all distances in $\{0, 1\}$. Solvable in polynomial time for $K = 2$. Variants in which we ask that the sum, over all $X_i$, of $\max\{d(x, y) : x, y \in X_i\}$ or of $\sum_{x,y \in X_i} d(x, y)$ be at most $B$ are similarly NP-complete (with the last one NP-complete even for $K = 2$)."

The problem described here is minimum diameter partitioning.

Despite the fact that nothing is mentioned about squared Euclidean distances in [4], many papers cited it to state that the MSSC is NP-hard [10, 37, 38, 39, 43, 44, 45, 57, 62], including one written by one of the present authors [26].

The papers [13, 28, 48, 56, 58] also cite Garey and Johnson's book without mentioning Brücker as a reference for MSSC to be NP-hard. This error may be due to the paragraph cited above or possibly to another one which refers to minimum sum-of-squares,

> “[SP19] MINIMUM SUM OF SQUARES
> INSTANCE: Finite set $A$, a size $s(a) \in Z^+$ for each $a \in A$, positive integers $K \leq |A|$ and $J$.
> QUESTION: Can $A$ be partitioned into $K$ disjoint sets $A_1, A_2, \ldots, A_K$ such that
> $$\sum_{i=1}^{K} \left( \sum_{a \in A_i} s(a) \right)^2 \leq J.\text{”}$$

Clearly, this last problem is different from MSSC.

It is worthwhile to mention that, after reformulating its objective function, minimum sum-of-squares clustering is actually proved NP-hard for a general metric space. First of all, from the Huygens' theorem (see e.g. Edwards and Cavalli-Sforza [12]), the MSSC objective function aforementioned can be rewritten as

$$\sum_{i=1}^{n} \sum_{j=1}^{k} w^{ij} \|X^i - z^j\|^2 = \sum_{j=1}^{k} \frac{\sum\limits_{i=1}^{n-1} \sum\limits_{\ell=i+1}^{n} w^{ij} w^{\ell j} \|X^i - X^\ell\|^2}{|C_j|},$$

where $|C_j|$ refers to the cardinality of the $j$-th cluster. This equivalence relation only makes sense for the Euclidean metric, since in its absence, the notion of centroids as cluster centers is meaningless. When clustering in a general metric space, the right-hand side formula is used to express the minimum sum-of-squares criterion. This is a strong assumption since the derivation of Huygens' result presupposes an Euclidean space.

Welch [59] examined a graph-theoretical proof of NP-hardness for the minimum diameter partitioning proposed in [23], and extended it to show NP-hardness of other clustering problems such as the MSSC. The proof that the minimum diameter partitioning is NP-hard is by reduction from the chromatic number of a graph, which asks if a graph $G = (V, E)$ can be colored with $M$ colors in such a way that both endpoints of an edge never has the same color. The polynomial transformation is performed as follows. The dissimilarity between each pair of entities $i$ and $j$ in $G$, denoted $d_{ij}$, is made equal to 1 if the edge $(i, j) \in E$, and equal to 0.5 otherwise. Thus, if the optimal partition into $M$ clusters has minimum diameter equal to 0.5, then $G$ is $M$-colorable, otherwise it is not. In [59], the chromatic number is reduced to the minimum sum-of-squares partitioning by the same transformation, except for making $d_{ij} = 0$ instead of 0.5 if the edge $(i, j)$ did not belong to $E$. Then, the minimum sum-of-squares is equal to 0 if the graph was $M$-colorable, and different from this value otherwise. However, this proof does not hold anymore in Euclidean metric spaces due to the fact that points belonging to $\mathbb{R}^s$ must necessarily obey triangle inequalities, i.e., it is not possible to have $d_{ij} = d_{j\ell} = 0$ and $d_{i\ell} = 1$.

For some $k$ and $s$ fixed, the Euclidean MSSC problem can in principle be solved in polynomial time. For instance, the number of bipartitions is polynomially bounded for fixed dimensions (see [25]) which leads to an $O(n^{s+1} \log n)$ algorithm. Moreover, if $s = 1$, Euclidean MSSC can be solved in $O(n^2 c)$ time using dynamic programming [2, 50].

## 3 An incorrect reduction from the $k$-section problem

Recently, Drineas et al. [11] proposed a NP-hardness proof for the Euclidean MSSC with $k = 2$ and general $s$ 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 already cited in [1, 3, 6, 42] as giving a proof that MSSC is NP-hard.

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

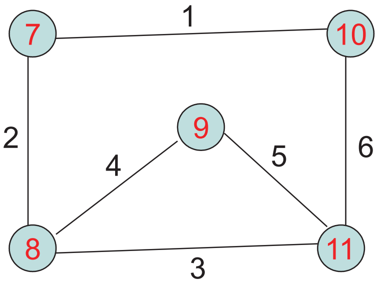
> "Let $G = (V, E)$ be the given graph with $n$ vertices 1,. . . ,n, with $n$ even. Let $d(i)$ be the degree of the i'th vertex. We will map each vertex of the graph to a point with $|E| + |V|$ coordinates. There will be one coordinate for each edge and one coordinate for each vertex. The vector $X^i$ 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 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 [11], we will show that this is always true for any Euclidean 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 $p$ and $q$, respectively. Regarding the last $|V|$ coordinates of the centroids, we have for $i = 1, \ldots, |V|$

$$z^P_{|E|+i} = \begin{cases} \frac{M}{p} & : \text{ if } i \in P \\ 0 & : \text{ otherwise} \end{cases}$$

$$z^Q_{|E|+i} = \begin{cases} \frac{M}{q} & : \text{ if } i \in Q \\ 0 & : \text{ otherwise} \end{cases}$$



$$X^7 = (1,1,0,0,0,0,\mathrm{M},0,0,0,0)^{\mathrm{T}}$$
$$X^8 = (0,1,1,1,0,0,0,\mathrm{M},0,0,0)^{\mathrm{T}}$$
$$X^9 = (0,0,0,1,1,0,0,0,\mathrm{M},0,0)^{\mathrm{T}}$$
$$X^{10} = (1,0,0,0,0,1,0,0,0,\mathrm{M},0)^{\mathrm{T}}$$
$$X^{11} = (0,0,1,0,1,1,0,0,0,0,\mathrm{M})^{\mathrm{T}}$$

Figure 1: Transformation of a graph into an Euclidean MSSC instance as defined by [11].

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

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

In Drineas et al. [11], 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 minimum whenever $p = q = n/2$. Thus, if $M$ is made sufficiently large, then balanced bipartitions have costs strictly smaller than unbalanced ones, since the contribution for the cost limited to the first $|E|$ coordinates is assuredly upper bounded. In fact, for $p = q$, this latter value is minimized when the solution of the Euclidean MSSC is the balanced bipartition that corresponds to the minimum bisection in the original graph (see Drineas et al. [11], page 16).

Unfortunately, after correcting the expression of the cost regarding the last $|V|$ coordinates, it does not depend any more on how the entities are distributed among the clusters. This implies that the proposed reduction from minimum bisection is invalid.

In order to fix the proof, the first idea that comes up to mind is to propose another polynomial transformation for which the balanced bipartitions have costs really strictly smaller than that of any unbalanced bipartition. However, the following proposition shows this cannot happen (the proof is presented in the appendix).

**Proposition 1** *For any instance of MSSC with $n > 2$ and $n$ even, there is always a balanced bipartition with cost greater or equal to the cost of a bipartition unbalanced by one element.*

Therefore, it is not possible to produce an Euclidean MSSC instance as needed. Moreover, the following corollary shows that obtaining a proof of NP-hardness for the Euclidean MSSC for $k$ greater than 2 from the minimum $k$-section is unlikely.

**Corollary 1** *Given $n \equiv 0 \pmod{k}$ distinct points in the $s$-dimensional Euclidean space, there is no MSSC instance with $n > k \geq 2$ such that the cost of balanced partitions are strictly smaller than the cost of unbalanced partitions.*

**Proof.** From proposition 1, it suffices to look at the sum-of-squares of the entities of two clusters to conclude that it is not possible to arbitrarily change the cluster assignment of an entity always obtaining a partition with smaller cost □

## 4 Concluding remarks

Many incorrect statements about MSSC being NP-hard have been made. We discussed their probable causes. Also, besides pointing out a mistake in the proof of Drineas et al [11] published in *Machine Learning* volume 56, we presented arguments to claim that a proof of NP-hardness of the Euclidean MSSC by reduction from the minimum $k$-section with $k \geq 2$ is unlikely to exist.

## A Appendix

**Proof.** Let $d_{ij}$ be equal to the squared Euclidean norm between entities $o_i$ and $o_j$, i.e., $\|X^i - X^j\|^2$ (in fact, this proof holds more generally for any dissimilarity matrix $D = (d_{ij})$ for which the following properties are satisfied: $d_{ij} \geq 0$, $d_{ii} = 0$ and $d_{ij} = d_{ji}$ for $i, j = 1, \ldots, n$.)

Assume by contradiction that our hypothesis is false, i.e., all balanced bipartitions have cost strictly smaller than bipartitions unbalanced by one element. Thus, if we consider a particular balanced bipartition with clusters $P$ and $Q$, and a particular unbalanced bipartition for which the entity $o_1$ leaves cluster $P$ and goes to cluster $Q$, we have that

$$\frac{\sum\limits_{o_i,o_j \in P} d_{ij}}{n/2} + \frac{\sum\limits_{o_i,o_j \in Q} d_{ij}}{n/2} - \frac{\sum\limits_{o_i,o_j \in P\setminus\{o_1\}} d_{ij}}{n/2-1} - \frac{\sum\limits_{o_i,o_j \in Q} d_{ij}}{n/2+1} - \frac{\sum\limits_{o_i \in Q} d_{1i}}{n/2+1} < 0,$$

By equalizing the denominators, we have that

$$\left(\frac{n^2}{4} - 1\right)\left(\sum_{o_i,o_j \in P\setminus\{o_1\}} d_{ij} + \sum_{o_i \in P} d_{1i} + \sum_{o_i,o_j \in Q} d_{ij}\right) \\ - \left(\frac{n^2}{4} + \frac{n}{2}\right) \sum_{o_i,o_j \in P\setminus\{o_1\}} d_{ij} - \left(\frac{n^2}{4} - \frac{n}{2}\right)\left(\sum_{o_i,o_j \in Q} d_{ij} + \sum_{o_i \in Q} d_{1i}\right) < 0$$

and

$$\left(\frac{n^2}{4}-1\right)\sum_{o_i\in P} d_{1i} - \left(\frac{n}{2}+1\right)\sum_{o_i,o_j\in P\setminus\{o_1\}} d_{ij} + \left(\frac{n}{2}-1\right)\sum_{o_i,o_j\in Q} d_{ij} - \left(\frac{n^2}{4}-\frac{n}{2}\right)\sum_{o_i\in Q} d_{1i} < 0. \quad (1)$$

Let us now consider the sum of all expressions of type (1) for balanced bipartitions on which the entity $o_1$ is initially assigned to cluster $P$ and then moved to cluster $Q$. There are $\begin{pmatrix} n-2 \\ n/2-2 \end{pmatrix}$ cases when $o_1$ and $o_i$ are assigned to cluster $P$, $\begin{pmatrix} n-2 \\ n/2-1 \end{pmatrix}$ cases when $o_1$ and $o_i$ are assigned to cluster $Q$, $\begin{pmatrix} n-3 \\ n/2-3 \end{pmatrix}$ cases when $o_i$ and $o_j$, $i \neq j \neq 1$, belong to cluster $P\setminus\{o_1\}$, and $\begin{pmatrix} n-3 \\ n/2-2 \end{pmatrix}$ additions cases $o_i$ and $o_j$, $i \neq j \neq 1$, belong to cluster $Q$. Thus, we have

$$\left[\begin{pmatrix} n-2 \\ \frac{n}{2}-2 \end{pmatrix}\left(\frac{n^2}{4}-1\right) - \begin{pmatrix} n-2 \\ \frac{n}{2}-1 \end{pmatrix}\left(\frac{n^2}{4}-\frac{n}{2}\right)\right]\sum_{i=2}^{n} d_{1i} + \left[\begin{pmatrix} n-3 \\ \frac{n}{2}-2 \end{pmatrix}\left(\frac{n}{2}-1\right) - \begin{pmatrix} n-3 \\ \frac{n}{2}-3 \end{pmatrix}\left(\frac{n}{2}+1\right)\right]\sum_{\substack{i<j \\ i\neq 1}} d_{ij} < 0. \quad (2)$$

By symmetry in (2), the following expressions are obtained.

$$\left[\begin{pmatrix} n-2 \\ \frac{n}{2}-2 \end{pmatrix}\left(\frac{n^2}{4}-1\right) - \begin{pmatrix} n-2 \\ \frac{n}{2}-1 \end{pmatrix}\left(\frac{n^2}{4}-\frac{n}{2}\right)\right]\sum_{\substack{i=1 \\ i\neq 2}}^{n} d_{2i} + \left[\begin{pmatrix} n-3 \\ \frac{n}{2}-2 \end{pmatrix}\left(\frac{n}{2}-1\right) - \begin{pmatrix} n-3 \\ \frac{n}{2}-3 \end{pmatrix}\left(\frac{n}{2}+1\right)\right]\sum_{\substack{i<j \\ i,j\neq 2}} d_{ij} < 0$$

$$\vdots$$

$$\left[\begin{pmatrix} n-2 \\ \frac{n}{2}-2 \end{pmatrix}\left(\frac{n^2}{4}-1\right) - \begin{pmatrix} n-2 \\ \frac{n}{2}-1 \end{pmatrix}\left(\frac{n^2}{4}-\frac{n}{2}\right)\right]\sum_{i=1}^{n-1} d_{ni} + \left[\begin{pmatrix} n-3 \\ \frac{n}{2}-2 \end{pmatrix}\left(\frac{n}{2}-1\right) - \begin{pmatrix} n-3 \\ \frac{n}{2}-3 \end{pmatrix}\left(\frac{n}{2}+1\right)\right]\sum_{\substack{i<j \\ j\neq n}} d_{ij} < 0.$$

By summing up all of them, we obtain in the left-hand side of the resulting expression $\frac{n(n-1)}{2}$ terms of the form

$$2\left[\begin{pmatrix} n-2 \\ \frac{n}{2}-2 \end{pmatrix}\left(\frac{n^2}{4}-1\right)-\begin{pmatrix} n-2 \\ \frac{n}{2}-1 \end{pmatrix}\left(\frac{n^2}{4}-\frac{n}{2}\right)\right] + (n-2)\left[\begin{pmatrix} n-3 \\ \frac{n}{2}-2 \end{pmatrix}\left(\frac{n}{2}-1\right)-\begin{pmatrix} n-3 \\ \frac{n}{2}-3 \end{pmatrix}\left(\frac{n}{2}+1\right)\right] d_{ij}$$

which appear for each pair of entities $o_i$ and $o_j$ of the instance.

After some algebraic manipulations, we conclude that each one of these terms is equal to zero which implies a contradiction

□

## References

[1] D. Arthur and S. Vassilvitskii. K-means++: the advantages of careful seeding. In *2007 ACM-SIAM Symposium on Discrete Algorithms (SODA'07)*, 2007.

[2] R. Bellman and S. Dreyfus. *Applied Dynamic Programming.* Princeton University Press, 1962.

[3] J. Beringer and E. Hüllermeier. Online clustering of parallel data streams. *Data & Knowledge Engineering*, 58:180–204, 2006.

[4] P. Brücker. On the complexity of clustering problems. *Lecture Notes in Economic and Mathematical Systems*, 157:45–54, 1978.

[5] H.-L. Chen, K.-T. Chuang, and M.-S. Chen. Labeling unclustered categorical data into clusters based on the important attribute values. In *Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM'05)*, 2005.

[6] R. Cilibrasi, L. van Iersel, S. Kelk, and J. Tromp. On the complexity of several haplotyping problems. *Lecture Notes in Computer Science*, 3692:128–139, 2005.

[7] F.A.B. da Silva, S. Carvalho, H. Senger, E.R. Hruschka, and C.R.G. de Farias. Running data mining applications on the grid: a bag-of-tasks approach. *Lecture Notes in Computer Science*, 3044:168–177, 2004.

[8] M. Delattre and P. Hansen. Bicriterion cluster analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-2(4):277–291, 1980.

[9] I.S. Dhillon and D.S. Modha. A data-clustering algorithm on distributed memory multiprocessors. *Lecture Notes in Artificial Intelligence*, 1759:245–260, 2002.

[10] J. Domingo-Ferrer and J. M. Mateo-Sanz. Practical data-oriented microaggregation for statistical disclosure control. *IEEE Transactions on Knowledge and Data Engineering*, 14:189–201, 2002.

[11] P. Drineas, A. Frieze, R. Kannan, S. Vempala, and V. Vinay. Clustering large graphs via the singular value decomposition. *Machine Learning*, 56:9–33, 2004.

[12] A.W. Edwards and L.L. Cavalli-Sforza. A method for cluster analysis. *Biometrics*, 21:362–375, 1965.

[13] D. Fasulo. An analysis of recent work on clustering algorithms. Technical Report UW-CSE-01-03-02, University of Washington, 1999.

[14] B. Fischer, V. Roth, and J.M. Buhmann. Clustering with the connectivity kernel. *Advances in Neural Information Processing Systems*, 16, 2004.

[15] K. Florek, J. Lukaszewicz, H. Perkal, H. Steinhaus, and S. Zubrzycki. Sur la liaison et la division des points d'un emsemble fini. *Colloquium Mathematicum*, 2:282–285, 1951.

[16] D. Fradkin, I.B. Muchnik, and S. Streltsov. Image compression in real-time multiprocessor systems using divisive k-means clustering. In *International Conference on Integration of Knowledge Intensive Multi-Agent Systems (KIMA'03)*, pages 506–511.

[17] P. Fränti and J. Kivijärvi. Randomised local search algorithm for the clustering problem. *Pattern Analysis & Applications*, 3:358–369, 2000.

[18] P. Fränti, O. Virmajoki, and T. Kaukoranta. Branch-and-bound technique for solving optimal clustering. In *International Conference on Pattern Recognition (ICPR'02)*, pages 232–235.

[19] M. Garey and D. Johnson. *Computers and Intractability.* W.H. Freeman and Company, New York, 1979.

[20] M.R. Garey, D.S. Johnson, and H.S. Witsenhausen. The complexity of the generalized lloyd-max problem. *IEEE Transactions on Information Theory*, IT-28:255–256, 1982.

[21] Y. Guan, A.A. Ghorbani, and N. Belacel. Y-means: a clustering method for intrusion detection. In *IEEE Canadian Conference on Electrical and Computer Engineering*, pages 1083–1086.

[22] Z. Güngör and A. Ünler. $k$-harmonic means data clustering with simulated annealing heuristic. *Applied Mathematics and Computation*, 184:199–209, 2007.

[23] P. Hansen and M. Delattre. Complete-link cluster analysis by graph coloring. *Journal of the American Statistical Association*, 73:397–403, 1978.

[24] P. Hansen and B. Jaumard. Cluster analysis and mathematical programming. *Mathematical Programming*, 79:191–215, 1997.

[25] P. Hansen, B. Jaumard, and N. Mladenović. Minimum sum of squares clustering in a low dimensional space. *Journal of Classification*, 15:37–55, 1998.

[26] P. Hansen and N. Mladenović. J-means: a new local search heuristic for minimum sum of squares clustering. *Pattern Recognition*, 34:405–413, 2001.

[27] J.A. Hartigan. *Clustering Algorithms.* Wiley, New York, 1975.

[28] Y. Jung, H. Park, D.-Z. Du, and B.L. Drake. A decision criterion for the optimal number of clusters in hierarchical clustering. *Journal of Global Optimization*, 25:91–111, 2003.

[29] O. Kariv and S.L. Hakimi. An algorithmic approach to network location problems; part 2. the $p$-medians. *SIAM Journal on Applied Mathematics*, 37:539–560, 1969.

[30] L. Kaufman and P.J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis.* Wiley, New York, 1990.

[31] M. Labbé, D. Petters, and J.F. Thisse. Location on networks. In M. Ball, T. Magnanti, C. Monma, and G. Nemhauser, editors, *Network Routing*, pages 551–624. North-Holland, 1995.

[32] Y. Li and S.M. Chung. Parallel bisecting $k$-means with prediction clustering algorithm. *The Journal of Supercomputing*, 39:19–37, 2007.

[33] P. Mangiameli, S.K. Chen, and D. West. A comparison of SOM neural network and hierarchical clustering methods. *European Journal of Operational Research*, 93:402–417, 1996.

[34] Y.M. Marzouk and A.F. Ghoniem. $k$-means clustering for optimal partitioning and dynamic load balancing of parallel hierarchical $n$-body simulations. *Journal of Computational Physics*, 207:493–528, 2005.

[35] J.B. McQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of* $5^{th}$ *Berkeley Symposium on Mathematical Statistics and Probability*, volume 2, pages 281–297, Berkeley, CA, 1967.

[36] M. Meilă. The uniqueness of a good optimum for k-means. *ACM International Conference Proceeding Series*, 148:625–632, 2006.

[37] P. Merz. Analysis of gene expression profiles: an application of memetic algorithms to the minimum sum-of-squares clustering problem. *Biosystems*, 72:99–109, 2003.

[38] P. Merz. An iterated local search for minimum sum-of-squares clustering. *Lecture Notes in Computer Science*, 2810:286–296, 2003.

[39] P. Merz and A. Zell. Clustering gene expression profiles with memetic algorithms. *Lecture Notes in Computer Science*, 2439:811–820, 2002.

[40] B. Mirkin. *Mathematical Classification and Clustering.* Kluwer, Dordrecht, The Netherlands, 1996.

[41] K. Niu, S.B. Zhang, and J.L. Chen. An initializing cluster centers algorithm based on pointer ring. In *Proceedings of the Sixth International Conference on Intelligent Systems Design and Applications (ISDA'06)*, 2006.

[42] R. Ostrovsky, Y. Rabani, L.J. Schulman, and C. Swamy. 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)*, 2006.

[43] J.A. Pacheco. A scatter search approach for the minimum sum-of-squares clustering problem. *Computers & Operations Research*, 32:1325–1335, 2005.

[44] J.A. Pacheco and O. Valencia. Design of hybrids for the minimum sum-of-squares clustering problem. *Computational Statistics & Data Analysis*, 43:235–248, 2003.

[45] S. Paterlini and T. Krink. Differential evolution and particle swarm optimisation in partitional clustering. *Computational Statistics and Data Analysis*, 50:1220–1247, 2006.

[46] J. Peng and Y. Wei. Approximating $k$-means-type clustering via semidefinite programming. *SIAM Journal on Optimization*, 18:186–205, 2007.

[47] J. Peng and Y. Xia. *A new theoretical framework for K-means-type clustering*, volume 180 of *Studies in Fuzziness and Soft Computing*. Springer, Berlin, 2005.

[48] Slobodan Petrovic and Gonzalo Álvarez. A method for clustering web attacks using edit distance. *CoRR*, cs.IR/0304007, 2003.

[49] V. Ramos and F. Muge. Map segmentation for colour cube genetic $k$-mean clustering. *Lecture Notes in Computer Science*, 1923:319–323, 2000.

[50] M.R. Rao. Cluster analysis and mathematical programming. *J. Amer. Statist. Assoc.*, 66:622–626, 1971.

[51] S.J. Redmond and C. Heneghan. A method for initialising $k$-means clustering algorithm using $kd$-trees. *Pattern Recognition Letters*, 28:965–973, 2007.

[52] W. Sheng and X. Liu. A genetic $k$-medoids clustering algorithm. *Journal of Heuristics*, 12:447–466, 2006.

[53] H.D. Sherali and J. Desai. A global optimization rlt-based approach for solving the hard clustering problem. *Journal of Global Optimization*, 32:281–306, 2005.

[54] M. Song and S. Rajasekaran. Fast k-means algorithms with constant approximation. *Lecture Notes in Computer Science*, 3827:1029–1038, 2005.

[55] H. Späth. *Cluster analysis algorithm for data reduction and classification of objects.* John Wiley & sons, New York, 1980.

[56] M. Teboulle. A unified continuous optimization framework for center-based clustering methods. *Journal of Machine Learning Research*, 8:65–102, 2007.

[57] H.M.M. ten Eikelder and A.A. van Erk. Unification of some least squares clustering methods. *Journal of Mathematical Modelling and Algorithms*, 3:105–122, 2004.

[58] J. Wang. A linear assignment clustering algorithm based on the least similar cluster representatives. *IEEE Transactions on systems, man, and cybernetics - part A: systems and humans*, 29:100–104, 1999.

[59] W.J. Welch. Algorithmic complexity: three np-hard problems in computational statistics. *Journal of Statistical Computation and Simulation*, 15:17–25, 1982.

[60] F.-X. Wu, W.J. Zhang, and A.J. Kusalik. A genetic k-means clustering algorithm applied to gene expression data. *Lecture Notes in Artificial Intelligence*, 2671:520–526, 2003.

[61] M. Xu and P. Fränti. Delta-MSE dissimilarity in suboptimal k-means clustering. In *Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)*, 2004.

[62] W. Zhou, C. Zhou, Y. Huang, and Y. Wang. Analysis of gene expression data: application of quantum-inspired evolutionary algorithm to minimum sum-of-squares clustering. *Lecture Notes in Artificial Intelligence*, 3642:383–391, 2005.