A survey on exact methods for minimum sum-of-squares clustering

Pierre Hansen[1], **and Daniel Aloise**[2]

[1] GERAD and HEC Montréal,
3000, chemin de la Côte-Sainte-Catherine, H3T 2A7 Montréal, Canada.
pierre.hansen@gerad.ca

[2] Département de mathématiques et génie industriel,
École Polytechnique de Montréal,
C.P. 6079, Succ. Centre-Ville , H3C 3A7 Montréal, Canada.
daniel.aloise@gerad.ca

Abstract:

Minimum sum-of-squares clustering (MSSC) consists in partitioning a given set of n entities into k clusters in order to minimize the sum of squared distances from the entities to the centroid of their cluster. Among many criteria used for cluster analysis, the minimum sum-of-squares is one of the most popular since it expresses both homogeneity and separation. A mathematical programming formulation of MSSC is as follows:

$$\begin{aligned}
&\min_{w,z} \quad \sum_{i=1}^{n}\sum_{j=1}^{k} w^{ij}\|X^i - z^j\|^2 \\
&\text{subject to} \\
&\sum_{j=1}^{k} w^{ij} = 1, \quad \forall i = 1,\ldots,n \\
&w^{ij} \in [0,1], \quad i = 1,\ldots,n; j = 1,\ldots,k.
\end{aligned}$$

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 $\mathcal{R}^s$ for $i = 1,\ldots,n$; k cluster centers must be located at unknown points $z^j \in \mathcal{R}^s$ 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.

Regarding computational complexity, minimum sum-of-squares clustering in the Euclidean metric for general values of k and s has recently been shown to be NP-hard [1]. We present a survey, with some new results, of the state-of-art and quite diverse exact methods for solving this problem.

The problem was formulated mathematically by Vinod [12] and Rao [9], but little was done towards its exact resolution in the general case until Koontz et al. [6] proposed a branch-and-bound algorithm which was posteriorly refined by Diehr [3]. However, these methods are confined to small data sets. It is important to remark that the hardness of a

MSSC instance is not directly measured by the number of entities, number of dimensions, number of clusters, etc. It also depends on the distribution of points. Consider an example of MSSC with n entities divided into k clusters which are each within a unit ball in $\mathcal{R}^s$. Assume these balls are pairwise at least n units apart. Then any reasonable branch-and-bound algorithm will confirm its optimality without branching as any misclassification more than doubles the objective function value. Note that n, k and s can be arbitrarily large. Recently, Brusco [2] proposed a repetitive branch-and-bound algorithm (RBBA) suited for instances where the entities must be partitioned into a small number of clusters. The method solves efficiently some benchmarks instances in the literature. By embedding the method into Variable Neighborhood Search [5], we were able to assess the potential gains in overall performance if better heuristic choices are made by the method.

The hardest task while devising branch-and-bound algorithms for the MSSC is to compute good lower bounds in a reasonable amount of time. Sherali and Desai [11] proposed to obtain such bounds by linearizing the model via the reformulation-linearization technique (RLT) [10]. We show that the RLT-model can be enforced by adding inequalities that actually break symmetry in the problem [8].

Peng and Xia [7] formulate the MSSC as a so-called 0-1 semi-definite programming (SDP). Then, the model is relaxed and used to calculate lower bounds for the problem in polynomial time. We aim to provide efficient tools for a new family of exact methods for the MSSC based on lower bounds originated from this 0-1 SDP formulation.

Regarding exact methods, we can still select from the literature, a column generation method proposed by du Merle et al. [4] which transfers the complexity of the MSSC to the resolution of the subproblem: an unconstrained hyperbolic problem in 0-1 variables with a quadratic numerator and linear denominator. This algorithm solved exactly, for the first time, fairly large data sets, including the Fisher's 150 iris. Yet, Xia and Peng [13] casted the MSSC as a concave minimization problem and adapted the Tuy's cut method to solve it. In the paper, good approximated results are reported for a version where the cutting plane algorithm is halted before global convergence. Regarding exact solutions only fairly small instances can be solved.

Bibliography

[1] D. Aloise, A. Deshpande, P. Hansen, and P. Popat. Np-hardness of euclidean sum-of-squares clustering. *Cahiers du GERAD*, G-2008-33, 2008.

[2] M.J. Brusco. A repetitive branch-and-bound procedure for minimum within-cluster sum of squares partitioning. *Psychometrika*, 71:347–363, 2006.

[3] G. Diehr. Evaluation of a branch and bound algorithm for clustering. *SIAM J. Sci. Statist.*, 6:268–284, 1985.

[4] O. du Merle, P. Hansen, B. Jaumard, and N. Mladenović. An interior point algorithm for minimum sum-of-squares clustering. *SIAM J. Sci. Comput.*, 21:1485–1505, 2000.

[5] P. Hansen and N. Mladenović. Variable neighborhood search: principles and applications. *European Journal of Operational Research*, 130:449–467, 2001.

[6] W.L.G. Koontz, P.M. Narendra, and K. Fukunaga. A branch and bound clustering algorithm. *IEEE Trans. Comput.*, C-24:908–915, 1975.

[7] 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.

[8] F. Plastria. Formulating logical implications in combinatorial optimisation. *European Journal of Operational Research*, 140:338–353, 2002.

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

[10] H.D. Sherali and W.P. Adams. Reformulation-linearization techniques for discrete optimization problems. In D.Z. Du and P.M. Pardalos, editors, *Handbook of combinatorial optimization 1*, pages 479–532. Kluwer, 1999.

[11] 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.

[12] H.D. Vinod. Integer programming and the theory of grouping. *J. Amer. Stat. Assoc.*, 64:506–519, 1969.

[13] Y. Xia and J. Peng. A cutting algorithm for the minimum sum-of-squared error clustering. In *Proceedings of the SIAM International Data Mining Conference*, 2005.