

CONTINUOUS OPTIMIZATION APPROACHES FOR CLUSTERING VIA MINIMUM SUM OF SQUARES

Basak Akteke-Ozturk^a, Gerhard-Wilhelm Weber^a, Erik Kropat^b

^a Institute of Applied Mathematics
Middle East Technical University, 06531 Ankara, Turkey
E-Mail: bozturk@metu.edu.tr, gweber@metu.edu.tr

^b Institute of Applied Mathematics
University of Erlangen-Nuremberg, 91058 Erlangen, Germany
E-Mail: kropat@am.uni-erlangen.de

Abstract. In this paper, we survey the usage of semidefinite programming (SDP), and nonsmooth optimization approaches for solving the minimum sum of squares problem which is of fundamental importance in clustering. We point out that the main clustering idea of support vector clustering (SVC) method could be interpreted as a minimum sum of squares problem and explain the derivation of semidefinite programming and a nonsmooth optimization formulation for the minimum sum of squares problem. We compare the numerical results produced by the semidefinite formulation of minimum sum of squares with the results obtained from approaching it via nonsmooth optimization on two datasets.

Keywords: Support vector clustering, SDP, nonsmooth optimization, minimal sum of squares, k – means, relaxation.

1. Introduction

In recent years, methods of continuous optimization have shown to provide a conceptual framework for problems in clustering and classification that are related to the *minimum sum of squares clustering* (MSSC) problem. Various approaches based on *nonsmooth optimization* (Bagirov, 2001; Bagirov, 2003; Bagirov, 2006) and *semidefinite programming* as well as the associated *SDP and LP relaxation* (Peng, 2005; Peng, 2007) were proposed and successfully applied. In this article we demonstrate that the main idea of *support vector clustering* (SVC) can be reformulated as a minimum sum of squares clustering problem so that SVC could be integrated in this general framework. We present the reformulation of MSSC problem as a SDP problem and its relaxed linear and semidefinite forms. Then, we briefly summarize the nonsmooth optimization approach to MSSC. We compare the numerical results produced by SDP based algorithms with the algorithm of modified k-means based on the nonsmooth optimization approach.

2. Support Vector Clustering

The *support vector clustering method* (Benhur, 2001) provides a powerful clustering algorithm which is based on the support vector machines approach. In this method data points are mapped by means of a Gaussian kernel to a high dimensional feature space, where a minimal enclosing sphere has to be determined. This sphere, when mapped back to data space, can separate into several components, each enclosing a separate cluster of data points. In this paper, we are interested in the following part of the algorithm: “Minimizing the radius of the sphere in the feature space which encloses the images of all the data points”. For this, we consider a set $S = \{s_1, \dots, s_n\}$ of n given points $s_i \in \mathbb{R}^d$. Using $\Phi = (\Phi_1, \dots, \Phi_k)$ for transforming S to some high dimensional space \mathbb{R}^k , one looks for the smallest enclosing *sphere* of radius R . This sphere can be defined by a minimal number $R \geq 0$ satisfying the constraints $\|\Phi(s_i) - c\|_2^2 \leq R^2$ for $i = 1, \dots, n$, where $c = (c_1, \dots, c_k)$ is the center of the sphere. Now, we define the distance of the image $\Phi(s)$ in the feature space from the center of the sphere at each point s by $d^2(s) = \|\Phi(s) - c\|_2^2$. The idea of minimizing the radius R of the sphere can be formulated as the optimization problem

$$\begin{aligned} & \min_{R,c} R^2 \\ \text{subject to} & \sum_{j=1}^k (\Phi_j(s_i) - c_j)^2 \leq R^2 \quad \text{for all } i = 1, \dots, n \\ & R \geq 0 \end{aligned}$$

which can be restated as the *minimum sum of squares problem*

$$\min_c \sum_{i=1}^n \|v_i - c\|_2^2. \quad (\text{SVC})$$

Here, v_i is the vector whose j -th element is equal to $\Phi_j(s_i)$. In other words, the function Φ is used to map all the points in the input space to some points in the so-called feature (kernel) space. As a minimum sum of squares problem, (SVC) can be solved with the standard k -means clustering which has to be performed in the kernel space. In the next sections we provide a conceptual framework for the minimum sum of squares problem.

3. Minimum Sum of Squared Distances

The *minimum sum of squared distances clustering problem* is to partition the n points of our set $S = \{s_1, \dots, s_n\} \subseteq \mathbb{R}^d$ into k clusters $S_j \in \Sigma = \{S_1, \dots, S_k\}$ centered at c_j ($j = 1, \dots, k$) so that the total sum of squared Euclidean distances from each point s_i to its assigned cluster centroid c_j is minimized. The objective function of this minimization problem is

$$f(S, \Sigma) = \sum_{j=1}^k \sum_{i=1}^{|S_j|} \|s_i^{(j)} - c_j\|_2^2,$$

where $|S_j|$ is the number of points in S_j such that $\sum_{j=1}^k |S_j| = n$, and $s_i^{(j)}$ is the i -th point of S_j .

The classical k -means algorithm (McQueen, 1967) handles the clustering of a data set by minimizing the cost function in the form of squared distances:

K-means Clustering Algorithm

- (1) Choose k cluster centers randomly generated in a domain containing all the points of S .
- (2) Assign each point to the closest cluster center and obtain a k -partition of S .
- (3) Recompute the cluster centers for this partition.
- (4) If a convergence criterion is met or no more data points change the clusters, stop; otherwise go to step 2.

Below, we mention some optimization models for the minimum sum of squared distances clustering problem:

Bi-level program: Find the centers c_j of the clusters S_j such that the sum of squared Euclidean distances from each data point to its cluster centroid is minimal:

$$\min_{c_1, \dots, c_k} \sum_{i=1}^n \min\{\|s_i - c_1\|_2^2, \dots, \|s_i - c_k\|_2^2\}.$$

Mixed-integer program: Another way to model the minimum sum of squared distances clustering problem is based on the assignment matrix $X = [x_{ij}] \in \mathbb{R}^{n \times k}$ defined by $x_{ij} = 1$ if s_i is assigned to S_j and $x_{ij} = 0$ otherwise. By rewriting the center of the cluster S_j as $c_j = \sum_{l=1}^n x_{lj} s_l / \sum_{l=1}^n x_{lj}$ we obtain the *minimum sum of squares clustering problem*

$$\left. \begin{array}{l} \min_{x_{ij}} \quad \sum_{j=1}^k \sum_{i=1}^n x_{ij} \left\| s_i - \frac{\sum_{l=1}^n x_{lj} s_l}{\sum_{l=1}^n x_{lj}} \right\|_2^2 \\ \text{subject to} \quad \sum_{j=1}^k x_{ij} = 1 \quad (i = 1, \dots, n), \\ \quad \quad \quad \sum_{i=1}^n x_{ij} \geq 1 \quad (j = 1, \dots, k), \\ \quad \quad \quad x_{ij} \in \{0, 1\} \quad (i = 1, \dots, n; j = 1, \dots, k) \end{array} \right\} \quad (\text{MSSC})$$

This is a NP-hard mixed-integer optimization problem with discrete constraints and with a nonlinear and nonconvex objective function. Furthermore, (MSSC) is a global optimization problem with possibly many local minima. Most of the methods for “solving” (MSSC) are pure heuristics that can only locate a “good” local solution (Peng, 2007). In the next section we show that (MSSC) can be modelled by 0-1 SDP, which can be further relaxed to polynomially solvable linear programming and SDP by removing constraints.

4. Equivalence of Minimum Sum of Squared Distances to 0-1 SDP

(MSSC) can be modelled by 0-1 SDP that takes the general form

$$\left. \begin{array}{l} \min \quad \text{Trace}(WZ) \\ \text{subject to} \quad \text{Trace}(B_i Z) = b_i \text{ for } i = 1, \dots, m \\ \quad \quad \quad Z^2 = Z, Z = Z^T \end{array} \right\} \quad (0-1 \text{ SDP}_C)$$

where Z is the assignment matrix and W is an affinity matrix (Peng, 2007). For this we use $W_S = (s_1^T, \dots, s_n^T)^T$ and $Z = [z_{ij}] = X(X^T X)^{-1} X^T$. After some calculations we obtain the 0-1 SDP model for the minimum sum of squared distances problem:

$$\left. \begin{array}{l} \min \quad \text{Trace}(W_S W_S^T (I - Z)) \\ \text{subject to} \quad Ze = e, \text{Trace}(Z) = k \\ \quad \quad \quad Z \geq 0, Z = Z^T, Z^2 = Z. \end{array} \right\} \quad (0-1 \text{ SDP}_{(\text{MSSC})})$$

Here, we are interested in approximation methods based on SDP and LP relaxations of $(0-1 \text{ SDP}_{(\text{MSSC})})$ which, typically, are considerably easier to solve than the original problem.

5. Relaxations Based on SDP

We continue with some relaxations based on SDP for $(0-1 \text{ SDP}_{(\text{MSSC})})$. First we recall that in $(0-1 \text{ SDP}_{(\text{MSSC})})$ the argument Z is stipulated to be a projection matrix, i.e., $Z^2 = Z$, which implies that the matrix Z is a positive semidefinite matrix whose eigenvalues are either 0 or 1 (Peng, 2005). A straightforward relaxation is replacing the requirement $Z^2 = Z$ by the relaxed condition $I \succeq Z \succeq 0$. In $(0-1 \text{ SDP}_{(\text{MSSC})})$ we further assume that all the entries of Z are nonnegative, and the

sum of each row (or each column) of Z equals to 1. This implies that all the eigenvalues of Z are always less than 1. Hence, the constraint $I \succeq Z$ becomes unnecessary and can be dropped. Therefore, we obtain the following *SDP relaxation* for a minimum sum of squared distances problem (Peng, 2005):

$$\left. \begin{array}{ll} \min & \text{Trace}(W_S W_S^T (I - Z)) \\ \text{subject to} & Ze = e, \text{Trace}(Z) = k \\ & Z \geq 0, Z \succeq 0 \end{array} \right\} (0-1 \text{SDP}_C^1)$$

We note that $(0-1 \text{SDP}_C^1)$ is feasible and bounded from below. Furthermore, many existing optimization procedures such as interior-point methods can be successfully applied and an approximate solution to $(0-1 \text{SDP}_C^1)$ can be found in polynomial time (Peng, 2007).

Another relaxation to $(0-1 \text{SDP}_{(\text{MSSC})})$ can be obtained by dropping some constraints in $(0-1 \text{SDP}_C^1)$ (cf. Peng, 2005). For example, if we remove the nonnegative requirement on the elements of Z , then we obtain the relaxed SDP problem

$$\left. \begin{array}{ll} \min & \text{Trace}(W_S W_S^T (I - Z)) \\ \text{subject to} & Ze = e, \text{Trace}(Z) = k \\ & I \succeq Z \geq 0 \end{array} \right\} (0-1 \text{SDP}_C^2)$$

which can equivalently stated as

$$\left. \begin{array}{ll} \min & \text{Trace}(W_S W_S^T Z) \\ \text{subject to} & Ze = e, \text{Trace}(Z) = k \\ & I \succeq Z \geq 0 \end{array} \right\} (0-1 \text{SDP}_C^3)$$

and its solution can be found in (Peng, 2005).

6. Relaxations Based on LP

We can also state an LP relaxation for $(0-1 \text{SDP}_{(\text{MSSC})})$. Firstly, we observe that if both s_i and s_j , and s_j and s_k belong to the same clusters, then s_i and s_k are also lying in the same cluster. In this situation, from the definition of the matrix Z we can conclude that $Z_{ij} = Z_{jk} = Z_{ik} = Z_{ii} = Z_{jj} = Z_{kk}$. Such a relationship can be partially characterized by the inequality $Z_{ij} + Z_{ik} \leq Z_{ii} + Z_{jk}$. Correspondingly, we can define a metric polyhedron by $MET = \{Z \mid Z = [z_{ij}] : z_{ij} \leq z_{ii}, z_{ij} + z_{ik} \leq z_{ii} + z_{jk}\}$ and we obtain the relaxation (Peng, 2005)

$$\left. \begin{array}{ll} \min & \text{Trace}(W_S W_S^T (I - Z)) \\ \text{subject to} & Ze = e, \text{Trace}(Z) = k \\ & Z \geq 0 \\ & Z \in MET. \end{array} \right\} (0-1 \text{SDP}_{\text{LP}}^4)$$

7. Nonsmooth Optimization Approach for Minimum Sum of Squares

The problem (MSSC) can be reformulated as a clustering problem in terms of unconstrained nonsmooth and nonconvex optimization (Bock, 1974; Bagirov, 2001) as follows:

$$\min f(c_1, \dots, c_k) \quad ((c_1, \dots, c_k) \in \mathbb{R}^{d \times k}), \quad (\text{CP})$$

where

$$f(c_1, \dots, c_k) = \frac{1}{n} \sum_{i=1}^n \min_{j=1, \dots, k} \|s_i - c_j\|_2^2.$$

(CP) contains only continuous real variables but not the integer coefficients x_{ij} , whereas the problem (MSSC) reveals both integer and continuous variables. If the number of clusters is greater than one, i.e., $k > 1$, then the objective function of (CP) is non-convex and nonsmooth. This problem becomes *large-scale* if the number k of clusters and the number d of attributes are large. We also emphasize that the nonsmooth form of the objective function in (CP) makes it possible to accelerate the calculation of the objective function significantly by reducing the number of records in a data set. Some algorithms developed by improving k -means algorithm to overcome the difficulties in choosing the proper initial partition and in finding the global minimum are studied in (Bagirov, 2006; Bagirov, 2003; Bagirov, 2001).

8. Numerical Results

To compare the algorithms for the minimum sum of squares problem, we used one small size dataset and one medium sized dataset from UCI Machine Learning Repository (Asuncion, 2007).

- I) **Soybean data (small)** [47 instances, 35 normalized attributes, 4 clusters]: In (Peng, 2005) the numerical results obtained from the SDP approach to minimum sum of squares with a LP relaxation ($0-1\text{SDP}_{LP}^4$) is given. In Table 1 we compare them with the results we obtained with the modified k -means algorithm which solves a nonsmooth optimization subproblem for calculating the starting point for the k -th cluster center (Bagirov, 2003).

Table 1. Results from the algorithm based on LP relaxation of the SDP reformulation of k -means and the modified k -means algorithm for Soybean data (small).

k	LP relaxation		Modified k -means	
	Value of the Objective	CPU time	Value of the Objective	CPU time
2	404.4593	4.26	434.11081081	0.000
3	215.2593	1.51	240.45925926	0.000
4	205.9637	1.68	214.60000000	0.016

- II) **Spam E-mail Database** [4601 instances, 57 features, 2 clusters]: In Table 2, we compare the results of the approximate algorithm developed in (Peng, 2007) for solving ($0-1\text{SDP}_{LP}^2$) and the modified k -means algorithm (i.e. via nonsmooth optimization) for the Spam E-mail Database.

Table 2. Results from the algorithm based on LP relaxation of the SDP reformulation of k -means and the modified k -means algorithm for the Spam E-mail database.

k	SDP relaxation		Modified k -means	
	Value of the Objective	CPU time	Value of the Objective	CPU time
2	9.43479784e+08	0	9.43479784e+08	58.922

The results obtained from LP and SDP relations of ($0-1\text{SDP}_{(MSSC)}$) match with the global optimums for both data sets. Modified k -means method also finds the optimal result for Spam E-mail data. For Soybeans data, modified k -means algorithm becomes closer to the optimal solution as the number of clusters increases with a better CPU time than the LP relaxation. Probably for larger number of clusters, the modified k -means algorithm will give better results than those obtained by

LP relaxation. It should be noted that the number of local minimizers increases drastically as the number of clusters increases. In such situations, relaxation methods are not efficient.

9. Conclusions and Outlook

In this paper, we reformulated SVC as a minimum sum of squares problem and then explained 0-1 SDP model of the classical minimum sum of squared distances problem. This approach to SVC is quite new and the framework followed in this paper can be employed in solving other clustering problems. The LP and SDP relaxations which we stated for solving k -means type clustering problems can be used to attack the 0-1 SDP. We show that the nonsmooth optimization approach to minimum sum of squares is also quite promising. By combining k -means with a nonsmooth optimization subproblem, we see that this modified k -means method is able to find the optimal result or a good approximation to it. We point out that SVC allows a representation in form of the more general conic optimization and the newly introduced set-semidefinite optimization (Eichfelder, 2007) that is offering a further interesting avenue for (approximative) problem solution which has to be investigated in future work. The research of this paper is an element in our efforts on data mining (Akteke-Ozturk, 2007; Weber, 2007), e.g., for quality improvement in manufacturing. Indeed, our future studies will encompass clustering, classification and regression by the use of continuous optimization and computational statistics.

Acknowledgement: This work was supported by The Scientific and Technological Research Council of Turkey (TUBITAK) 105M138.

References

- Akteke-Ozturk, B.; Weber, G.-W.; Kayaligil, S. (2007) "Kalite iyilestirmede veri kümeleme: Dokum endustrisinde bir uygulama", proceedings of Yöneyem Arastirmasi ve Endüstri Muhendisligi 27. Ulusal Kongresi (YA/EM 2007), July 2-4, 2007, Izmir, Turkey, pp. 1207-1212.
- Asuncion, A; Newman, D.J. (2007) UCI Machine Learning Repository. Irvine, CA: University of California, Department of Information and Computer Science.
<http://www.ics.uci.edu/~mlern/MLRepository.html>
- Bagirov, A.M.; Yearwood J. (2006) "A new nonsmooth optimization algorithm for minimum sum-of-squares clustering problem", European Journal of Operational Research, Vol. 127, No. 2, pp. 578-596.
- Bagirov, A.M.; Rubinov, A.M.; Soukhoroukova, N.V.; and Yearwood, J. (2003) "Unsupervised and supervised data classification via nonsmooth and global optimization", TOP 11, 1, pp. 1-93.
- Bagirov, A.M.; Rubinov, A.M.; Yearwood J. (2001) "Using global optimization to improve classification for medical diagnosis and prognosis", Topics in Health Information Management 22, pp. 65-74.
- Ben-Hur, A.; Horn, D.; Siegelmann, H.T.; Vapnik, V. (2001) "Support vector clustering", *Journal of Machine Learning Research*, Vol. 2, pp. 125-137.
- Del Castillo, E., Montgomery, D.C., McCarville, D.R. (1996) "Modified desirability functions for multiple response optimization", *Journal of Quality Technology* 28, 3, pp. 337-345.
- Bock H.H. (1974), *Automatische Klassifikation*, Vandenhoeck and Ruprecht, Göttingen.
- Eichfelder, G.; Jahn, J. (2007) "Set-Semidefinite Optimization". Preprint No. 316, Preprint series of the Institute of Applied Mathematics, University Erlangen-Nürnberg.
- McQueen J. (1967) "Some methods for classification and analysis of multivariate observations", *Computer and Chemistry*, Vol. 4, pp. 257-272.
- Peng, J.M.; Wei, Y. (2007) "Approximating K-means-type clustering via semidefinite programming", *SIAM Journal of Optimization*, Vol. 18, Issue 1, pp. 186-205.
- Peng, J.M.; Xia, Y. (2005) "A new theoretical framework for k-means-type clustering", in Chu, W.; Lin, T.Y. (eds.), *Foundations and Advances in Data Mining*, Springer Verlag, pp. 79-96.
- Weber, G.-W.; Taylan, P.; Ozugur, S.; Akteke-Ozturk, B. (2007) "Statistical learning and optimization methods in data mining", in: *Recent Advances in Statistics*, eds.: Ayhan H.O., Batmaz, İ., Turkish Statistical Institute Press, Ankara, pp. 181-195.