MIXED INTEGER LINEAR PROGRAMMING FORMULATION FOR K-MEANS CLUSTER PROBLEM

Kolos Cs. Ágoston
Institute of Mathematics and Statistical Modeling, Corvinus University of Budapest,
Fővám tér 8., Budapest, 1098 Hungary
Email: kolos.agoston@uni-corvinus.hu
Marianna E.-Nagy
Corvinus Centre for Operations Research, Corvinus University of Budapest,
Fővám tér 8., Budapest, 1098 Hungary
Department of Differential Equations, Budapest University of Technology and Economics,
Egry József Street 1., Budapest, 1111 Hungary
Email: marianna.eisenberg-nagy@uni-corvinus.hu

Abstract: The minimum sum of clustering is the most used clustering method. The minimum sum of clustering is usually solved by the heuristic K-means algorithm which converges to a local optimum. Much effort was put into solving such kind of problem, but a mixed integer linear programming formulation (MILP) is still missing. In this paper, we formulate MILP models and solve them up to sample size 100. The advantage of MILP formulation is that users can extend the original problem with arbitrary linear constraints.

Keywords: Clustering, LP formulation, K-means

1 INTRODUCTION

Clustering is one of the most used methods in data science. Inside this area, K-means clustering is the most used approach which aims to minimize the within cluster sum of squares of distances. K-means clustering algorithm is a very fast method, but it is a heuristic algorithm without any guarantee for global optimum. In data science, it is said that the K-means algorithm is sensitive to the initial cluster centers, in optimization terminology the K-means algorithm converges to a local optimum. This phenomenon is well known, however, this method is implemented in most used statistical and data science softwares till nowadays, contrary to the fact that exact algorithms are known (see for instant du Merle et al. (1999), Peng and Wei (2007)).

Solving clustering problem using linear programming appeared early in the literature, see Vinod (1969), Rao (1971). Later, different types of clustering problems were solved using LP, see for instance Cornuejols et al. (1980), Kulkarni and Fathi (2007), Dorndorf and Pesch (1994), Gilpin et al. (2013), but the most frequently used minimum sum of squares clustering was less investigated. Du Merle at al. (1999) proposed an exact algorithm to solve the minimum sum of squares clustering problem, but this approach did not appear in statistical packages, probably due to the fact that it is a quite complicated algorithm.

We can also form the minimum sum of squares clustering problem as Semidefinite Programming (SDP) problem (see Peng and Wei (2007)). The drawback of this possibility is that SDP problems can be solved only for moderate size problems.

In this paper, we present Mixed Integer LP formulations for the minimum sum of squares problem. These formulations can be extended to problems with many types of constraints (for instance lower bound on the cardinality of clusters or must-link constraints: Bradley et al. (2000), Davidson and Ravi (2007)). The presented MILP model is based on formulation appeared in Awasthi et al. (2015).

2 MILP FORMULATION FOR MINIMUM SUM OF SQUARES CLUSTERING PROBLEM

We have N points in the n-dimensional space: $\mathcal{A} = \{a_1, \ldots, a_N\} \subset \mathbb{R}^n$. Our aim is to group these points into K clusters in a way that we minimize the within cluster sum of squared distances. Clusters of points are denoted by $\mathcal{A}_k$, $k = 1 \ldots K$. These sets form a partition of $\mathcal{A}$, since $\mathcal{A}_k \cap \mathcal{A}_\ell = \emptyset$, $\cup_{k=1}^{K} \mathcal{A}_k = \mathcal{A}$ and $\mathcal{A}_k \neq \emptyset$ for all $k = 1, \ldots, K$. Let $\mathcal{P}_\mathcal{A}$ denote the set of partitions of $\mathcal{A}$ into eaxctly K non-empty subsets. The center of cluster $\mathcal{A}_k$ is denoted by c_k, which is defined as the multidimensional mean. Sum of squared distances within the cluster $\mathcal{A}_k$ is given by the formula: $\sum_{a_i \in \mathcal{A}_k} d(a_i, c_k)^2$, where $d(a, b)$ is the Euclidean distance between points a and b (also called ℓ_2 norm). We can reformulate this sum of squares formula as $\frac{1}{|\mathcal{A}_k|} \sum_{a_i, a_j \in \mathcal{A}_k} d(a_i, a_j)^2$ (see du Merle et al. (1999), Awasthi et al. (2015)). The *minimum sum of squares clustering problem* is the following:

$$\min_{(\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_k) \in \mathcal{P}_\mathcal{A}} \sum_{k=1}^{K} \sum_{a_i, a_j \in \mathcal{A}_k} \frac{d(a_i, a_j)^2}{|\mathcal{A}_k|} \ .$$

In Awasthi et al. (2015), we can find a promising formulation:

$$\sum_{i,j} d(a_i, a_j)^2 \, z_{ij} \to \min \tag{1}$$

s.t.

$$\sum_{j=1}^{N} z_{ij} = 1 \qquad \forall \, i \in [N] \tag{2}$$

$$z_{ij} \leq z_{ii} \qquad \forall \, i, j \in [N] \tag{3}$$

$$\sum_{i=1}^{N} z_{ii} = K \tag{4}$$

$$z_{ij} \geq 0 \qquad \forall \, i, j \in [N] \tag{5}$$

$$z_{ij} \in \{0, 1/|\mathcal{A}_{t(j)}|\} \qquad \forall \, i, j \in [N]$$

where $\mathcal{A}_{t(j)}$ is the cluster which contains a_j, $|\mathcal{A}|$ is the cardinality of the set $\mathcal{A}$ and $[L] := \{1, \ldots, L\}$. Except for the last constraint, this is a linear model with nonnegative decision variables z_{ij} which indicates whether element i and j belongs to the same cluster or not. There are several problems with the last constraint, we do not know apriori the value of $t(j)$ and the cardinality of cluster $\mathcal{A}_{t(j)}$. However, it can be reformulated as $z_{ij}(z_{ij} - z_{ii}) = 0$, but it is still not a linear constraint.

Unfortunately (but not surprisingly), the optimal solution of the problem minimizing (1) subject to (2)-(5) does not give a 'legal' clustering. To ensure this, we need further constraints.

It is worth prescribing the symmetry of the variables z_{ij}, that is,

$$z_{ij} = z_{ji} \qquad \forall\, i, j \in [N]\ . \tag{6}$$

Another possible linear constraint is the 'triangle inequality':

$$z_{ij} + z_{i\ell} - z_{j\ell} \leq z_{ii} \qquad \forall\, i, j, \ell \in [N]\ . \tag{7}$$

Indeed, if both variables z_{ij} and $z_{i\ell}$ take positive value (meaning elements i and j are in the same cluster and also elements i and ℓ are in the same cluster), then variable $z_{j\ell}$ has to take positive value, and in this case, the values of all of the three variables have to equal variable z_{ii}. If both variables z_{ij} and $z_{i\ell}$ are 0, then the value of variable $z_{j\ell}$ is not constrained.

We refer to the model minimizing (1) subject to (2)-(7) as *MSSR: Minimum Sum of Squares Relaxation*. It still does not surely result in a 'legal' clustering structure, but as the numerical tests show, we already get an optimal clustering with this model in most of the cases. To get an exact model, we use binary variables. It can be done in different ways, we will discuss two of them.

First we introduce binary variable ζ_{ij}, which takes value 1, if elements i and j are in the same cluster, otherwise it takes value 0:

$$\zeta_{ij} \in \{0, 1\} \qquad \forall\, i, j \in [N]\ . \tag{8}$$

Values of variables z_{ij} and ζ_{ij} are not independent, hence we need constraints to ensure the relationship between them:

$$z_{ij} \leq \zeta_{ij} \qquad \forall\, i, j \in [N]\ . \tag{9}$$

$$z_{ii} - z_{ij} \leq 1 - \zeta_{ij} \qquad \forall\, i, j \in [N]\ . \tag{10}$$

Now the problem minimizing (1) subject to (2)-(10) gives an exact model for the K-means problem. However, we can add further constraints (cuts) to help the MILP solver to find an optimal solution faster. Two possibilities are considering the following constraints

$$(N - K + 1) z_{ij} \geq \zeta_{ij} \qquad \forall\, i, j \in [N]\ . \tag{11}$$

$$(N - K + 1)(z_{ii} - z_{ij}) \geq 1 - \zeta_{ij} \qquad \forall\, i, j \in [N]\ . \tag{12}$$

We reach the *MSS formulation*: minimizing (1) subject to (2)-(12). It is easy to see that if an optimal solution of MSSR is a 'legal' clustering, then all binary variables ζ_{ij} take integer values, i.e. branch and bound tree will only contain the root node.

In MSS formulation the number of binary variables can be quite large, its number increase quadratically as the number of elements increases. We tried another formulation, in which the number of binary variables is significantly less. Let γ_{ik} denote the binary variable which indicates if element i is assigned to cluster k:

$$\gamma_{ik} \in \{0, 1\} \qquad \forall\, i \in [N], k \in [K]\ . \tag{13}$$

Since every element belongs to exactly one cluster

$$\sum_{k=1}^{K} \gamma_{ik} = 1 \qquad \forall\, i \in [N]\ , \tag{14}$$

furthermore every cluster contains at least one element:

$$\sum_{i=1}^{N} \gamma_{ik} \geq 1 \qquad \forall\, k \in [K]\ . \tag{15}$$

We need to connect variables γ_{ik} to variables z_{ij}. If elements i and j are in different clusters, then z_{ij} has to be zero, therefore

$$z_{ij} \leq 1 + \gamma_{ik} - \gamma_{jk} \qquad \forall\, i \neq j \in [N]\ . \tag{16}$$

Now the problem minimizing (1) subject to (2)-(5) and (13)-(16) gives an exact model for the K-means problem. We call it *AMSS (Assignment-type Minimum Sum of Squares)* formulation. Let us again show some further constraints that can help the MILP solver. One possibility is to enforce i and j into different clusters if $z_{ij} = 0$:

$$\gamma_{ik} + \gamma_{jk} \leq 1 + (N - K + 1) z_{ij} \qquad \forall\, i, j \in [N], k \in [K]\ . \tag{17}$$

Furthermore, in a clustering problem, the essential result is a grouping, meaning which elements are in the same cluster and which are in different ones. The 'label' of the cluster is irrelevant. If we have K clusters the labels can be assigned in $K!$ way. We can break this symmetry by prescribing that the first element belongs to the first cluster:

$$\gamma_{1,1} = 1\ . \tag{18}$$

We could go further. If the second element belongs to the same cluster as the first element, it will be assigned to cluster 1 as well. Otherwise, let it be in the second cluster, so we have $\gamma_{2,k} = 0$, for $k \geq 3$. Similarly for the third element, $\gamma_{3,k} = 0$ for $k \geq 4$. Surprisingly, these constraints also slow down the process, it is not worth using all of them.

AMSS has signficantly less binary variables than MSS ($N \times K$ vs. $(N-1) \times (N-1)$). A further advantage of AMSS formulation is that more constraints can be formulated with the help of variables γ_{ik} than with the help of ζ_{ij}. On the other hand, it is not true that if the optimal solution of MSSR formulation gives a legal clustering, then all binary variables in the relaxation of AMSS take integer values.

3 NUMERICAL RESULTS

In order to test the above formulations, we generated uniformly distributed random points in the unit square. On these instances, we tested the MSSR, MSS and AMSS formulations. We used a desktop computer with 3.60 GHz Intel Pentium processor and 8 GB RAM. Operating system is Windows 10 Enterprise. We used Gurobi 9.1.1 solver with the default parameter settings for solving MILP problems.

instance (N,K)	#var. (bin.)	#conns.	# nonzero	#iter.	time (sec)	o.f. value
(50,2)	2,500 (0)	62,526	245,100	44,891	5.41	8.7706
(75,2)	5,625 (0)	210,976	832,650	47,323	29.55	14.4857
(100,2)	10,000 (0)	500,051	1,980,200	113,292	146.49	18.8850
(50,3)	2,500 (0)	62,526	245,100	32,493	2.30	4.8643
(75,3)	5,625 (0)	210,976	832,650	112,711	17.09	8.6184
(100,3)	10,000 (0)	500,051	1,980,200	271,365	110.24	11.8003
(50,5)	2,500 (0)	62,526	245,100	21,525	1.40	2.7192
(75,5)	5,625 (0)	210,976	832,650	74,264	9.46	4.3356
(100,5)	10,000 (0)	500,051	1,980,200	134,661	68.26	6.0120

Table 1: Essential information about MSSR formulation

instance (N,K)	#var. (bin.)	#const.	# nonzero	#iter.	time (sec)	o.f. value
(50,2)	3,725 (1,225)	67,426	257,350	13,057	4.86	8.7706
(75,2)	8,400 (2,775)	222,076	860,400	85,215	69.00	14.4857
(100,2)	14,950 (4,950)	519,851	2,029,700	199,582	547.48	18.8850
(50,3)	3,725 (1,225)	67,426	257,350	7,231	3.53	4.8643
(75,3)	8,400 (2,775)	222,076	860,400	51,086	45.91	8.6184
(100,3)	14,950 (4,950)	519,851	2,029,700	122,879	345.28	11.8003
(50,5)	3,725 (1,225)	67,426	257,350	8,950	3.58	2.7192
(75,5)	8,400 (2,775)	222,076	860,400	42,963	38.07	4.3356
(100,5)	14,950 (4,950)	519,851	2,029,700	199,143	1,064.93	6.0156

Table 2: Essential information about MSS formulation

instance (N,K)	#var. (bin.)	#const.	# nonzero	#iter.	time (sec)	o.f. value
(50,2)	2,600 (100)	70,029	267,451	559,853	306.13	8.7706
(75,2)	5,775 (150)	227,854	883,201	64,316	45.55	14.4857
(100,2)	10,200 (200)	530,054	2,070,101	155,495	270.05	18.8850
(50,3)	2,650 (150)	73,755	278,776	211,463	88.09	4.8643
(75,3)	5,850 (225)	236,255	908,476	81,300	56.23	8.6184
(100,3)	10,300 (300)	545,005	2,115,051	205,222	312.18	11.8003
(50,5)	2,750 (250)	81,207	301,226	182,144	72.10	2.7192
(75,5)	6,000 (375)	253,057	959,026	73,654	56.49	4.3356
(100,5)	10,500 (500)	574,907	2,204,951	766,669	16,128.02	6.0156

Table 3: Essential information about AMSS formulation

Some important statistics about the size of LP problems and some information about the solution process can be found in Table 1, Table 2 and Table 3.

As we can see in Table 1, 2 and 3, except the instance (100,5), the optimum solution of MSSR will result in a legal clustering structure, actually we do not need the integer variables. For all presented instances, the running times are less than 2.5 minutes for MSSR formulation. Not surprisingly, running times for MSS and AMSS formulations are higher, but still tolerable (except for the instance (100,5)). There is not a strict dominance between

MSS and AMSS formulations, MSS seems to have slightly better performance (mostly for instance (100,5)).

4 CONCLUSION

In this paper we investigated MILP formulations for the minimum sum of squares clustering problem. These formulations have higher running times than the well-known K-means algorithm, however for sample size at most 100 still tolerable. If in some application it is crucial to work with global optimum these formulations give a possibility for it. Furthermore, we are able to insert further (linear) constraints in the model.

References

Awasthi, P., Bandeira, A. S., Charikar, M., Krishnaswamy, R., Villar, S. and Ward, R. (2015). Relax, no need to round: integrality of clustering formulations. arXiv:1408.4045

Bradley, P. S., Bennett, K. P., and Demiriz, A. (2000). Constrained K-Means Clustering. https://www.microsoft.com/en-us/research/publication/constrained-k-means-clustering/

Cornuejols, G., Nemhauser, G. L. and Wolsey, L. A. (1980). A canonical representation of simple plant location-problems and its applications. SIAM Journal On Algebraic And Discrete Methods, vol(1), pp. 261–272.

Davidson, I. and Ravi, S. S. (2007). The complexity of non-hierarchical clustering with instance and cluster level constraints. Data Mining and Knowledge Discovery, vol. 14, pp. 25–61.

Dorndorf, U. and Pesch, E. (1994). Fast Clustering Algorithms, vol. 6, pp. 141–153.

Gilpin, S., Nijssen, S. and Davidson, I. N. (2012). Formalizing hierarchical clustering as integer linear programming. In Proceedings of the Twenty- Seventh AAAI Conference on Artificial Intelligence, July 14-18, 2013, Bellevue, Washington, USA, pp. 372–378.

du Merle, O., Hansen, P., Jaumard, B. and Mladenovic, N. (1999). An Interior Point Algorithm for Minimum Sum-of-Squares Clustering. SIAM Journal on Scientific Computing, vol(21), pp. 1485–1505.

Kulkarni, G. and Fathi, Y. (2007). Integer programming models for the q-mode problem. European Journal of Operational Research, vol. 182, pp 612-625.

Peng, J. and Wei, Y. (2007). Approximating K–means–type Clustering via Semidefinite Programming. SIAM Journal on Optimization, vol(18), pp. 186–205.

Rao, M. R. (1971). Cluster Analysis and Mathematical Programming. Journal of the American Statistical Association, vol(66): pp. 622–626.

Vinod H. D. (1969). Integer Programming and the Theory of Grouping. Journal of the American Statistical Association, vol(64): pp. 506–519.