Software Architecture for Automation

Project

Mathematical formulations for the *K* clusters with fixed cardinality problem and its implementations

Uğur Barış Öztürk - S277284

Olloshukur Atadjanov - S274478

Abstract

In this report we implement some of the mixed integer linear programming models proposed in the *"K clusters with fixed cardinality problem*" in order to observe the effect of increasing dimensionality on a combinatorial optimization problem.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

1. PROBLEM OVERIEW

K clusters with fixed cardinality problem (KCCP) is based on selecting M_k number of items for each disjoint cluster from a set with a cardinality of N by maximizing the total similarity among the items in the same cluster.

Notation used in the original paper will be same for the study and the fallowing formulation stated below:

i, j – items indexes ($i, j \in \{1, \ldots, N\}$),
k – cluster index ($k \in \{1, \ldots, K\}$),
N – number of items ($N \in \mathbb{N}$),
K – number of clusters ($K \in \mathbb{N}$, $K < N$),
M_k – number of items per cluster k($M_k \in \mathbb{N}$, $\sum_k M_k < N$),
s_{ij} – similarity between items i and j, element of a symmetric matrix with diagonal elements equal to zero ($0 \leqslant s_{ij} \leqslant 1$).
y_{ijk} – binary variable indicating whether items i and j are in the same cluster k ($=1$) or not ($=0$) ($i = 1, \ldots, N-1$; $j = i+1, \ldots, N$; $k = 1 \ldots, K$)

Selected formulations from the paper are F2 and F6 stated below and implemented on Xpress:

$$(\text{F2}) \max \sum_{k=1}^{K} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} s_{ij} y_{ijk}$$

$$s.\,t.: y_{ijk} \geqslant x_{ik} + x_{jk} - 1 \quad 1 \leqslant i < j \leqslant N;\ k = 1, \ldots, K$$

$$\sum_{k=1}^{K} x_{ik} \leqslant 1 \quad i = 1, \ldots, N$$

$$\sum_{i=1}^{N} x_{ik} = M_k \quad k = 1, \ldots, K$$

$$\sum_{i=1}^{j-1} y_{ijk} + \sum_{i=j+1}^{N} y_{jik} = (M_k - 1) x_{jk} \quad j = 1, \ldots, N;\ k = 1, \ldots, K$$

$$x_{ik} \in \{0, 1\} \quad i = 1, \ldots, N;\ k = 1, \ldots, K$$

$$0 \leqslant y_{ijk} \leqslant 1 \quad 1 \leqslant i < j \leqslant N;\ k = 1, \ldots, K.$$

Figure 1 Mathematical Model of F2

$$\left(\text{F6}\right) \max \sum_{k=1}^{K} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} s_{ij} y_{ijk}$$

$$s.\ t.: y_{ijk} \geqslant x_{ik} + x_{jk} - 1 \quad 1 \leqslant i < j \leqslant N;\ k = 1, \dots, K \quad (9)$$

$$y_{ijk} \leqslant x_{ik} \quad 1 \leqslant i < j \leqslant N;\ k = 1, \dots, K \quad (6)$$

$$y_{ijk} \leqslant x_{jk} \quad 1 \leqslant i < j \leqslant N;\ k = 1, \dots, K \quad (7)$$

$$\sum_{k=1}^{K} x_{ik} \leqslant 1 \quad i = 1, \dots, N \quad (2)$$

$$\sum_{i=1}^{N} x_{ik} = M_k \quad k = 1, \dots, K \quad (3)$$

$$\sum_{i=1}^{j-1} y_{ijk} + \sum_{i=j+1}^{N} y_{jik} = \left(M_k - 1\right) x_{jk} \quad j = 1, \dots, N;\ k = 1, \dots, K \quad (10)$$

$$x_{ik} \in \{0, 1\} \quad i = 1, \dots, N;\ k = 1, \dots, K \quad (4)$$

$$0 \leqslant y_{ijk} \leqslant 1 \quad 1 \leqslant i < j \leqslant N;\ k = 1, \dots, K. \quad (8)$$

Figure 2 Mathematical model of F6

Correctness of implemented solution on the Xpress tested with a several toy datasets that generated manually. Graph representation of the toy dataset is a fallows:

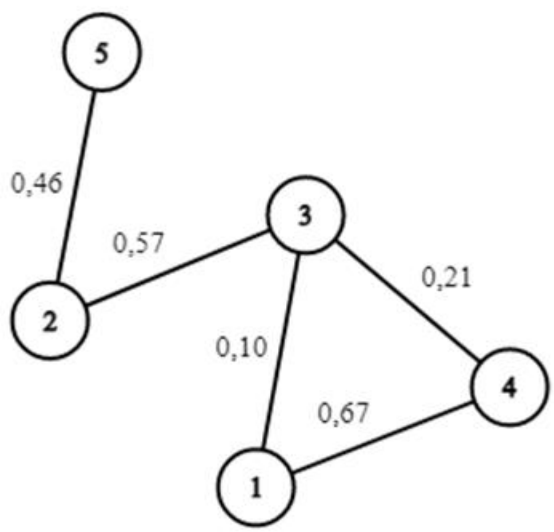

Figure 3 Graph representation of the toy instance

There are 5 items and 5 edges on the toy dataset.

Parameter settings for testing;

Number of Cluster: 2 <Scalar>

Number of items per cluster: [2,2] <Vector>

Solution of the problem is quite easy calculate manually and as follows:

Item 1 and 4 should be in the same cluster while item 2 and 3 together in the other. In this case value of the objective function is 0.57 + 0.67 = 1.24 since it is the sum of the similarity between the items in the same cluster.

Solutions

View last N solutions found by the Optimizer

	Column	Name	Obj=1.24
0/1	0	Xik(1,1)	1
0/1	1	Xik(1,2)	0
0/1	2	Xik(2,1)	0
0/1	3	Xik(2,2)	1
0/1	4	Xik(3,1)	0
0/1	5	Xik(3,2)	1
0/1	6	Xik(4,1)	1
0/1	7	Xik(4,2)	0
0/1	8	Xik(5,1)	0
0/1	9	Xik(5,2)	0

Figure 4 Solution of the toy instance

Solution of the optimizer shows that items 1 and 4 are in the cluster 1 and items 2 and 3 are in the cluster 2. Value of the objective function is found as 1.24.

2. TEST INSTANCES

The instances of the KCCP are downloaded from CEDRIS's library. A proper subset of the dataset in the paper generated by using a python script that attached with the project file. Both models are tested with the same instances.

In the original paper instances are defined with a graph by its density(d) and the number of items per cluster.

In this study there are 3 different values for density (0.25,0.5,0.75) 2 different values for M_k (10,20) in the case of number of the cluster is equal to 1 and 2 different values of M_k (5+5 , 2+8) for the case of number of cluster is equal to 2. For each fixed density and fixed M_k there are 3 different instances from corresponding to 3 different graphs each contain 40 nodes(N).

K	M_k	Graph density (d)	# Inst
1	10	(0.25 , 0.5 , 0.75)	9
1	20	(0.25 , 0.5 , 0.75)	9
2	5 , 5	(0.25 , 0.5 , 0.75)	9
2	2 , 8	(0.25 , 0.5 , 0.75)	9

Table 1 Instance parameters

Models are implemented with Xpress on 2 different devices. Tests for model F2 ran on a i5 processor with 2.5 GHz clock speed and 16 GB of RAM and tests for model F6 ran on a i5 processor with 1,6 GHz clock speed and 8 GB of RAM.

3. RESULTS & CONCLUSION

Instance	N	d	Mk	# Cluster	Sol. LP relaxation	Best Solution	Gap (%)	CPU Time(s)	Average Gap	Average Gap(Paper)
PB1_025	40	0,25	10	1	37,92	15,37	146,71	1,8		
PB2_025	40	0,25	10	1	39,23	16,3	140,67	4,2	148,53	153,7
PB3_025	40	0,25	10	1	40,15	15,55	158,20	3,4		
PB1_05	40	0,5	10	1	42,11	23,26	81,04	4,7		
PB2_05	40	0,5	10	1	41,79	23,82	75,44	5,4	81,94	78,3
PB3_05	40	0,5	10	1	42,66	22,53	89,35	8,5		
PB1_075	40	0,75	10	1	43,31	28,02	54,57	4,9		
PB2_075	40	0,75	10	1	43,24	29,06	48,80	4,1	50,14	48
PB3_075	40	0,75	10	1	43,57	29,63	47,05	4,7		
PB1_025	40	0,25	20	1	81,87	39,53	107,11	2,3		
PB2_025	40	0,25	20	1	100,77	42,99	134,40	8,5	125,24	124,8
PB3_025	40	0,25	20	1	95,86	40,93	134,20	5,8		
PB1_05	40	0,5	20	1	136,51	68,72	98,65	21,3		
PB2_05	40	0,5	20	1	135,98	73,78	84,30	25	93,89	92,5
PB3_05	40	0,5	20	1	138,68	69,79	98,71	37		
PB1_075	40	0,75	20	1	154,23	90,94	69,60	81,7		
PB2_075	40	0,75	20	1	155,59	93,35	66,67	43,5	66,63	62,9
PB3_075	40	0,75	20	1	155,92	95,29	63,63	63,4		
PB1_025	40	0,25	5,5	2	19,36	11,08	74,73	1,7		
PB2_025	40	0,25	5,5	2	19,46	12,09	60,96	4,1		
PB3_025	40	0,25	5,5	2	19,48	11,89	63,84	2,8	89,61	93,5
PB1_025	40	0,25	2,8	2	26,07	12,09	115,63	2,1		
PB2_025	40	0,25	2,8	2	26,65	12,95	105,79	2,1		
PB3_025	40	0,25	2,8	2	26,98	12,45	116,71	2,2		
PB1_05	40	0,5	5,5	2	19,64	15,2	29,21	4,2		
PB2_05	40	0,5	5,5	2	19,46	15,61	24,66	1,7		
PB3_05	40	0,5	5,5	2	19,8	14,79	33,87	4,3	45,16	46,3
PB1_05	40	0,5	2,8	2	27,75	17,46	58,93	2,3		
PB2_05	40	0,5	2,8	2	27,5	17,66	55,72	2,3		
PB3_05	40	0,5	2,8	2	28,2	16,73	68,56	3,8		
PB1_075	40	0,75	5,5	2	19,88	16,49	20,56	3,2		
PB2_075	40	0,75	5,5	2	19,72	16,77	17,59	2,1		
PB3_075	40	0,75	5,5	2	19,92	16,51	20,65	3	28,08	28,9
PB1_075	40	0,75	2,8	2	28,37	20,77	36,59	1,8		
PB2_075	40	0,75	2,8	2	28,24	20,43	38,23	2,7		
PB3_075	40	0,75	2,8	2	28,54	21,16	34,88	1,9		

Table 2 Result for model F2

Instance	N	d	Mk	# Cluster	Sol. LP relaxation	Best Solution	Gap (%)	CPU Time(s)	Average Gap	Average Gap(Paper)
PB1_025	40	0,25	10	1	21,01	15,37	36,69	7,7	48,40	47,1
PB2_025	40	0,25	10	1	25,2	16,3	54,60	13,1		
PB3_025	40	0,25	10	1	23,93	15,55	53,89	13,4		
PB1_05	40	0,5	10	1	33,91	23,26	45,79	14,1	46,76	40,9
PB2_05	40	0,5	10	1	33,83	23,82	42,02	19,3		
PB3_05	40	0,5	10	1	34,35	22,53	52,46	28,1		
PB1_075	40	0,75	10	1	35,9	27,05	32,72	21	29,77	27,7
PB2_075	40	0,75	10	1	37,55	29,06	29,22	16,8		
PB3_075	40	0,75	10	1	37,74	29,63	27,37	13,6		
PB1_025	40	0,25	20	1	43,481	39,53	9,99	13,8	20,86	20,9
PB2_025	40	0,25	20	1	54,67	42,99	27,17	30,9		
PB3_025	40	0,25	20	1	51,33	40,93	25,41	39,1		
PB1_05	40	0,5	20	1	93,17	68,72	35,58	58,6	37,14	35,9
PB2_05	40	0,5	20	1	100,725	73,78	36,52	86		
PB3_05	40	0,5	20	1	97,23	69,79	39,32	109,3		
PB1_075	40	0,75	20	1	118,575	87,72	35,17	341,8	33,69	31,9
PB2_075	40	0,75	20	1	124,315	93,35	33,17	150,1		
PB3_075	40	0,75	20	1	126,475	95,29	32,73	209,3		
PB1_025	40	0,25	5,5	2	14,01	11,08	26,44	5,4	45,51	40,5
PB2_025	40	0,25	5,5	2	15,69	12,09	29,78	10,9		
PB3_025	40	0,25	5,5	2	15,99	11,89	34,48	10,4		
PB1_025	40	0,25	2,8	2	16,69	12,09	38,05	5,8		
PB2_025	40	0,25	2,8	2	19,16	12,95	47,95	6,7		
PB3_025	40	0,25	2,8	2	18,53	12,45	48,84	9,6		
PB1_05	40	0,5	5,5	2	17,94	15,2	18,03	9,8	27,90	27,9
PB2_05	40	0,5	5,5	2	17,98	15,61	15,18	9,6		
PB3_05	40	0,5	5,5	2	18,18	14,79	22,92	13,1		
PB1_05	40	0,5	2,8	2	23,66	17,46	35,51	7,5		
PB2_05	40	0,5	2,8	2	23,49	17,66	33,01	9,6		
PB3_05	40	0,5	2,8	2	23,9	16,73	42,86	12,9		
PB1_075	40	0,75	5,5	2	18,3	16,24	12,68	6,8	17,50	17,8
PB2_075	40	0,75	5,5	2	18,62	16,77	11,03	5,6		
PB3_075	40	0,75	5,5	2	18,75	16,51	13,57	9,8		
PB1_075	40	0,75	2,8	2	24,541	19,88	23,45	7,1		
PB2_075	40	0,75	2,8	2	25,43	20,43	24,47	7,6		
PB3_075	40	0,75	2,8	2	25,59	21,16	20,94	7,9		

Table 3 Result for model F6

Result shows that dense graphs make problem computationally heavy since there are more relations between the nodes. We can confirm that average gap obtained during our test is quite close the those obtained in the paper. As it is mentioned in the paper constraint 6 and 7 which are added to F2 to obtain model F6 provided a significant reduction in the gaps for both cases of K=1 and K = 2. We are not able to directly compare the CPU times of model F2 and F6 since, different devices used for testing the models.

In terms of computational perspective, KCCP is NP-Hard. In order to get better feasible solutions exploiting some heuristic functions can be useful.

4. REFERENCES

1. Gonçalves, Graça Marques, and Lídia Lampreia Lourenço. "Mathematical formulations for the K clusters with fixed cardinality problem." *Computers & Industrial Engineering* 135 (2019): 593-600.