The p-median Problem with Order for Two-source Clustering*

Xenia Klimentova[1], Anton V. Ushakov[2], and Igor Vasilyev[2]

[1] INESC TEC, Campus da FEUP, Rua Dr. Roberto Frias, 378, 4200-465 Porto, Portugal,
xenia.klimentova@inesctec.pt
[2] Matrosov Institute for System Dynamics and Control Theory,
Lermontov 134, 664033, Irkutsk, Russia,
{aushakov, vil}@icc.ru.

Abstract. In this paper we present a hybrid approach to integrative clustering based on the p-median problem with clients' preferences. We formulate the problem of simultaneous clustering of a set of objects, characterized by two sets of features, as a bi-level p-median model. An exact approach involving a branch-and-cut method combined with the simulated annealing algorithm is used, that allows one to find a two-source clustering. The proposed approach is compared with some well-known mathematical optimisation based clustering techniques applied to the NCI-60 tumour cell line anticancer drug screen dataset. The results obtained demonstrate the applicability of our approach to find competitive integrative clusterings.

1 Introduction

Recent advances in microarray technologies give rise to collecting huge amount of high-dimensional omic data, generated simultaneously for the same biological samples. A huge number of techniques and approaches has been proposed to carefully combine and analyse continuous, discrete and categorical multisource data, mainly based on probabilistic and statistical modelling. Recent reviews [17, 20] cover a wide range of approaches to omic data integration. Modern techniques in statistics for integrative analyses of cancer data, including incorporating multiple heterogeneous genomic data types, are also reviewed in [19, 27].

This paper focuses on the cluster analysis, which is one of the main unsupervised learning methods. Clustering is an exploratory tool that consists in dividing a set of objects (biological samples, observations) into subsets (clusters) containing similar objects, while objects from different clusters are dissimilar. Despite a large number of general clustering algorithms having been proposed, there is a lack of such methods for

* The research of A. Ushakov and I. Vasilyev is supported by the Russian Science Foundation, research project No. 15-11-20015, the work of X. Klimentova is supported by the FCT — Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) within project SFRH/BPD/101134/2014.

In: A. Kononov et al. (eds.): DOOR 2016, Vladivostok, Russia, published at http://ceur-ws.org

efficient integrative clustering of multisource data. The simplest and obvious approach to integrative clustering consists in concatenating the feature vectors of the considered objects or samples followed by applying some well-known clustering algorithms to analyse the data obtained. Though this approach may be useful in some rare instances, it is rather inflexible due to its impossibility of capturing important features that are specific to each dataset and in the case of heterogeneous data.

We propose an approach to integrative clustering of two-source data which is based on bi-level integer linear programming and metaheuristic optimization. As opposed to previously developed integrative clustering methods, which are based on modelling each dataset using mixture models or post hoc combining multiple base clusterings, we propose a hybrid integer programming approach based on the bi-level p-median problem with clients' preferences. The approach is compared to those presented in [10], indicating that our approach can provide clustering solutions competitive with other distance-based integrative clustering algorithms. One of the first gene-drug integrative clustering approach was proposed in [21] and was based on a hierarchical clustering algorithm. Other mathematical optimization based approaches that have been applied to simultaneously clustering gene expression and drug activity profiles are the Soft Topographic Vector Quantization [8], relational k-means [11], genetic programming [3], and a consensus p-median approach [10].

The paper is structured as follows. The generic p-median clustering model is described in section 2. In section 3 an extension to the bi-level problem is presented. Finally, section 4 illustrates the comparison of different approaches and gives some concluding remarks.

2 The p-median clustering problem

The p-median problem, proposed by [13], consists in choosing p sites for locating facilities from the set of potential places in order to minimize the sum of weighted distances between customers and open facilities. This model can naturally be formulated in the form of the following combinatorial optimization problem:

$$\min_{S \subseteq I} \Big\{ \sum_{j \in J} \min_{i \in S} d_{ij} \,:\; |S| = p \Big\},$$

where $|I| = \{1, \ldots, m\}$ is a set of potential facility sites, $J = \{1, \ldots, l\}$ is a set of customers, and d_{ij} — the distance (or the cost of satisfying the demand) between customer $j \in J$ and the facility located at site $i \in I$.

The p-median problem is known to be NP-hard in the strong sense and now supposed to be well-studied. Apart from a number of applications, the p-median problem is a powerful tool for the cluster analysis. To transform the problem into a clustering tool, let us assume that $I = J$, i.e. the potential sites and customer locations coincide. In this case the p-median problem (also referred as the minimum-sum-of-stars clustering [16]), can be formulated on a simple digraph $G(I, A)$ with the node set I corresponding to the set of samples to be clustered and the arc set $A = \{(i, j) : i, j \in I;\ i \neq j\}$, each arc $(i, j) \in A$ has a weight (distance) $d_{ij} > 0$ measuring the dissimilarity between each pair of samples. Now the p-median problem is to select p nodes, those are called medians, in order to minimize the sum of the distances between each node and its nearest median. Any feasible solution to the problem consists of p stars with medians in the centres. Each star is a cluster and the median is a cluster representative (prototype).

The p-median model has several important advantages over centroid-based parametric clustering approaches, like k-means and its variations. First of all, modern hybrid heuristics and exact approaches are able to find optimal (global) or suboptimal solutions to the p-median instances of huge size (e.g. see [4, 12, 15]), while k-means or k-means type algorithms, like PAM, CLARA, CLARANS, k-medians, are heuristics that converge fast to a local optimum. Secondly, the classical k-means algorithm presupposes using the squared Euclidean distance as the measure of similarity, which not always results in a good clustering. For other types of parametric clustering problems, several algorithms using k-means type iterative relocation scheme have been proposed, e.g. k-medians in the case of Manhattan distance or the Linde-Buzo-Gray algorithm when the Itakura-Saito distance is considered [6].

The p-median problem can be formulated as an integer linear program. For each node $i \in I$, let us consider a binary variable y_i, which takes the value 1 if node i is a median (sample i is a cluster representative), and takes the value 0 otherwise. For each arc $(i,j) \in A$ let x_{ij} be the binary variable which is equal to 1 if node i is the closest median to node j (sample j is assigned to cluster i), and takes the value 0 otherwise. Let also $\delta^-(j) = \{i \in I |\ (i,j) \in A\}$ be a set of nodes of the graph G assigned to j with outgoing arcs, and let $\delta^+(i) = \{j \in I |\ (i,j) \in A\}$ be a set of nodes assigned to i with the arcs leaving node i. Thus, the p-median problem can be written as follows [5]:

$$\min_{(x,y)} \sum_{(i,j)\in A} d_{ij} x_{ij}, \tag{1}$$

$$\sum_{i\in\delta^-(j)} x_{ij} + y_j = 1, \qquad j \in I, \tag{2}$$

$$x_{ij} \le y_i, \qquad i \in I, j \in \delta^+(i), \tag{3}$$

$$\sum_{i\in I} y_i = p, \tag{4}$$

$$y_i \in \{0,1\}, \qquad i \in I, \tag{5}$$

$$x_{ij} \in \{0,1\}, \qquad (i,j) \in A. \tag{6}$$

Objective function (1) minimizes the overall total sum of distances between all nodes and the closest medians. Constraints (2) guarantee that either node j is a median or it is assigned to a median. Constraints (3) ensure that each node can only be assigned to medians. Constraint (4) enforces the number of medians to be p.

Several studies considered the p-median model for clustering various types of data, e.g. psychological data [18] or gene expression profiles and drug responses [10]. Moreover, p-median model is proved to be a powerful clustering tool that provides high quality clustering solutions outperforming those provided by CLARA and CLARANS [15] or k-means and k-means++ [23].

3 Bi-level p-median clustering

Suppose that we are given a set I of objects (cell lines) characterized by two feature vectors, i.e. representing gene expression and drug activity profiles. Here we propose an approach to two-source integrative clustering based on a bi-level version of the p-median problem and study its application to clustering cancer cell lines. Let two matrices $d_{ij} \geq 0$ and $g_{ij} \geq 0$ indicate the dissimilarity of a pair of cell lines $(i, j) \in A = \{(i, j) : \ i \in I, j \in I, i \neq j\}$ in the drug and gene space respectively. Then, let us consider the following bi-level p-median clustering problem:

$$\min_{y} \sum_{(i,j)\in A} d_{ij} x^*_{ij}(y), \tag{7}$$

$$\sum_{i\in I} y_i = p, \tag{8}$$

$$y_i \in \{0, 1\}, \qquad i \in I, \tag{9}$$

where $x^*_{ij}(y)$ is the optimal solution to the following lower-level problem:

$$\min_{x} \sum_{(i,j)\in A} g_{ij} x_{ij}, \tag{10}$$

$$\sum_{i\in\delta^-(j)} x_{ij} + y_j = 1, \qquad j \in I, \tag{11}$$

$$x_{ij} \leq y_i, \qquad i \in I, j \in \delta^+(i), \tag{12}$$

$$x_{ij} \in \{0, 1\}, \qquad (i, j) \in A. \tag{13}$$

On the upper level on seeks for p cell lines to be the cluster representatives such that the sum of dissimilarities of drug activity profiles between cell lines and its closest representatives is minimised. On the lower level all cell lines are assigned to the cluster representatives, selected at the first level, in order to minimise the sum of dissimilarities of gene expression profiles between cell lines belonging to the same cluster and its closest representatives. In other words, the decision about which cell lines will be the medians (cluster representatives) is made on the first level according to the matrix $\{d_{ij}\}$, while assigning remaining cell lines to clusters is performed on the second level taking into account the dissimilarity matrix $\{g_{ij}\}$, $(i, j) \in A$. Thus, if $g_{ij} \leq g_{kj}$ and both $i, k \in I$ are cluster representatives, then cell line j is assigned to cluster C_i.

Note that when all the columns of the lower-level matrix $\{g_{ij}\}$ are sorted in ascending order or when $d_{ij} = g_{ij}$, the problem (7)–(13) is a particular case of the p-median problem. Thus, it is NP-hard in the strong sense and does not belong to the class APX [2]. In general case the solution $x^*_{ij}(y)$ may be not unique, then one has to specify what kind of solution is supposed to be optimal to the clustering problem (7)–(13). Two extreme cases, i.e. cooperative and non-cooperative decision-making strategies, are often emphasised [1], depending on whether cell lines on the lower level are assigned to their closest representatives in order to respectively minimise or maximise the value of the upper-level objective. To avoid these cases for the sake of simplicity, we suppose that all elements of a column j of the matrix $\{g_{ij}\}$ are distinct [24].

The bi-level p-median clustering problem can be reduced to a single-level integer linear program [14, 7]. Let us denote $W_{ij} = \{k \in I : \; g_{ij} < g_{kj}\}$, $(i,j) \in A$, then the single-level problem is

$$\min_{(x,y)} \sum_{(i,j)\in A} d_{ij}x_{ij}, \tag{14}$$

$$y_i + \sum_{k\in W_{ij}} x_{kj} \le 1, \qquad i \in I, (i,j) \in A, \tag{15}$$

$$\sum_{i\in I} y_i = p, \tag{16}$$

$$\sum_{i\in \delta^-(j)} x_{ij} + y_j = 1, \qquad j \in I, \tag{17}$$

$$x_{ij} \le y_i, \qquad i \in I, j \in \delta^+(i), \tag{18}$$

$$y_i \in \{0,1\}, \qquad i \in I, \tag{19}$$

$$x_{ij} \in \{0,1\}, \qquad (i,j) \in A. \tag{20}$$

This formulation is identical to the p-median except constraints (15), which guarantee that if $i \in I$ is a cluster representative, then a cell line $j \in I$ is not assigned to more dissimilar (according to gene expression profiles) representatives from the set W_{ij}. Thus, x_{ij}, $(i,j) \in A$ is the optimal solution to the lower-level problem for any y_i, $i \in I$. Note that one can similarly consider the bi-level p-median clustering problem with the matrices $\{g_{ij}\}$ and $\{d_{ij}\}$ on the upper and lower levels respectively.

To find the optimal solution to the bi-level p-median clustering problem, we have developed an exact approach including a branch-and-cut algorithm and a metaheuristic to search for initial upper bounds of the optimal value, which is detailed in [24–26]. The method is based on the family facet-defining inequalities proposed in [26] for the simple-plant location problem with order. For some node j and subset $S \subseteq I$ let us denote by $b_j(S) \in S$ the nearest node from S for j, i.e. $W_{ij} \subset W_{b_j(S)j}$ for all $i \in S \setminus \{b_j(S)\}$.

Theorem 1 *For all $i, u, v \in I$ the inequalities*

$$\sum_{k\in W_{iu}} x_{ku} + \sum_{k\in U^t_{iu}} x_{kv} + y_t \le 1, \tag{21}$$

where $t = b_v(I \setminus W_{iu})$ and $U^t_{iu} = I \setminus (W_{iu} \cup \{t\})$, are valid for the polytope of the problem (14)-(20).

The cutting plane method for this family of inequalities was implemented using two computational tricks for reduction of the number of violated inequalities. On each iteration we add the inequalities corresponding to the most distant hyperplane from the current fractional solution, preventing the almost parallel inequalities from adding. To find an upper bound of the optimal solution the standard scheme of simulated annealing method was implemented, using flip-neighbourhood (see [26] for more detail).

Finally, as an exact method we have used Cut-and-Branch scheme, one of the effective methods tested in previous works for a similar problem. In root node of branching tree the cutting plane method is run and the new formulation obtained, as well as an upper bound provided by the simulated annealing, are used in order to solve the problem with branch-and-bound algorithm.

4 Clustering analysis of the NCI-60 dataset

The proposed approach was implemented as a program using C++ programming language. The MIP solver FICO Xpress callable library has been used as a branch-and-cut framework. We compare our results with those obtained in the paper [10] as well as with other integrative clustering techniques presented in that paper. The number of clusters p was equal to 9.

As a measure of dissimilarity between cell lines both in the drug and gene spaces, we apply one of the most widely used measure based on the Pearson correlation coefficient, i.e. $dist_{ij} = 1 - corr_{ij}$, $(i, j) \in A$ [9].

We report our results performing two-source cluster analysis of National Cancer Institute (NCI)-60 panel of human tumour cancer cell lines [22]. The dataset consists of 60 cell lines from 9 cancer tissues. To compare our approach, we use the same dataset as was previously considered in [10, 21]. It includes 1376 gene expression profiles and 1400 drug activity patterns. Thus, I is a set of $m = 60$ cell lines, and the dissimilarity matrices $\{g_{ij}\}$ and $\{d_{ij}\}$ are computed for each pair $(i, j) \in A$ as $dist_{ij}$, using the given gene expression and drug response data.

To evaluate the quality of a cluster solution we use the average Pearson correlation coefficient

$$P = \sum_{k=1}^{p} \frac{2}{m(|C_k| - 1)} \sum_{i,j \in \{1,\ldots,m\}: i<j} corr_{ij},$$

where C_k is a cluster k. Such coefficient was computed taking into account both the gene and drug spaces, thus providing P^G and P^D values respectively.

The results of the computational experiments are presented in Table 1, where the first column demonstrates the method applied, the second column contains the value of parameters μ and α setting the stepsize of the consensus p-median approach and weighted coefficients of the Soft Topographic Vector Quantization method respectively from [10]. The notations (d) and (g) indicate whether drug or gene dissimilarity matrix are used on the upper-level.

Note that the clustering results presented in [10] are obtained on the base of a leave-one-out cross validation, i.e. on the set $O \setminus \{i\}$ for each $\{1, \ldots, m\}$, and the confidence intervals as well as the mean are then estimated. The leave-one-out cross validation procedure provides a small effect size for most of the approaches both for correlations in the gene and drug space, thus we compare our results with the presented mean values.

Analysing the results obtained, we can conclude that our approach provides better integrative clustering solutions, than most of the methods under consideration, i.e. STVQ, p-Median, k-means, relational k-means, and probabilistic d-clustering. These results are better with respect to the cluster homogeneity in the drug space in all cases. They are also better in the gene space when STVQ with α less than or equal to 0.2 is considered. Concerning the consensus p-median clustering approach, our method provides competitive or little worse clustering solutions. Nevertheless the main advantage

		P^G	P^D
bi-level p-median (d)	–	0.5013	0.8349
bi-level p-median (g)	–	0.4879	0.8409
Consensus p-Median (d-g)	$\mu = 1.1$	0.4777	0.8553
	$\mu = 1.2$	0.5073	0.8595
	$\mu = 1.3$	0.5200	0.8522
	$\mu = 1.4$	0.5265	0.8497
	$\mu = 1.5$	0.5373	0.8401
	$\mu = 1.6$	0.5401	0.8357
	$\mu = 1.7$	0.5449	0.8349
	$\mu = 1.8$	0.5464	0.8334
Consensus p-Median (g-d)	$\mu = 1.1$	0.5054	0.8613
	$\mu = 1.2$	0.4586	0.8604
	$\mu = 1.3$	0.4232	0.8566
	$\mu = 1.4$	0.3735	0.8366
	$\mu = 1.5$	0.3689	0.8363
STVQ	$\alpha = 0.0$	0.5450	0.8200
	$\alpha = 0.1$	0.5300	0.8213
	$\alpha = 0.2$	0.5110	0.8217
	$\alpha = 0.3$	0.4960	0.8265
	$\alpha = 0.4$	0.4800	0.8289
	$\alpha = 0.5$	0.4770	0.8301
	$\alpha = 0.6$	0.4536	0.8303
	$\alpha = 0.7$	0.4298	0.8304
	$\alpha = 0.8$	0.4022	0.8306
	$\alpha = 0.9$	0.3713	0.8309
	$\alpha = 1.0$	0.3598	0.8310
p-Median	–	0.4596	0.8366
k-Means	–	0.4770	0.8301
Relational k-Means	–	0.4983	0.8240
Probabilistic D-Clustering	–	0.4122	0.7916

Table 1. Computational results for the Sherf et al. dataset

of the proposed approach is that one has not to solve a series of integer linear programs with different values of μ, which in the case of large problem instances may be of great importance. Moreover, our method can be a useful tool and provide competitive solutions when no prior information about the data structure is known. This is especially important in the field of unsupervised machine learning.

References

1. E. Alekseeva and A. Kochetov, Y. Plyasunov. An exact method for the discrete (r—p)-centroid problem. *J. Glob. Optim.*, 63(3):445–460, 2015.
2. E. Alekseeva, Yu. Kochetov, and A. Plyasunov. Complexity of local search for the p-median problem. *Eur. J. Oper. Res.*, 191(3):736–752, 2008.

3. F. Archetti, I. Giordani, and L. Vanneschi. Genetic programming for anticancer therapeutic response prediction using the nci-60 dataset. *Comput. Oper. Res.*, 37(8):1395–1405, 2010. Operations Research and Data Mining in Biological Systems.
4. P. Avella, M. Boccia, S. Salerno, and I. Vasilyev. An aggregation heuristic for large scale p-median problem. *Comput. Oper. Res.*, 39(7):1625–1632, 2012.
5. P. Avella, A. Sassano, and I. Vasilyev. Computational study of large-scale p-median problems. *Math. Program.*, 109(1):89–114, 2007.
6. A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with bregman divergences. *J. Mach. Learn. Res.*, 6:1705–1749, 2005.
7. L. Cánovas, S. García, M. Labbé, and A. Marín. A strengthened formulation for the simple plant location problem with order. *Oper. Res. Lett.*, 35(2):141–150, 2007.
8. J.-H. Chang, K.-B. Hwang, and B.-T. Zhang. *Methods of Microarray Data Analysis II: Papers from CAMDA' 01*, chapter Analysis of Gene Expression Profiles and Drug Activity Patterns by Clustering and Bayesian Network Learning, pages 169–184. Springer, Boston, 2002.
9. J. H. Do and D. K. Choi. Clustering approaches to identifying gene expression patterns from dna microarray data. *Mol. Cells*, 25(2):279–288, 2008.
10. E. Fersini, E. Messina, and F. Archetti. A p-median approach for predicting drug response in tumour cells. *BMC Bioinformatics*, 15(1):1–19, 2014.
11. E. Fersini, E. Messina, F. Archetti, and C. Manfredotti. Combining gene expression profiles and drug activity patterns analysis: A relational clustering approach. *J. Math. Mod. Alg.*, 9(3):275–289, 2010.
12. S. García, M. Labbé, and A. Marín. Solving large p-median problems with a radius formulation. *INFORMS J. Comput.*, 23(4):546–556, 2011.
13. S. L. Hakimi. Optimum distribution of switching centers in a communication network and some related graph theoretic problems. *Oper. Res.*, 13(3):462–475, 1965.
14. P. Hanjoul and D. Peeters. A facility location problem with clients' preference orderings. *Regional Sci. Urban Econom.*, 17(3):451–473, 1987.
15. P. Hansen, J. Brimberg, D. Urosević, and N. Mladenović. Solving large p-median clustering problems by primal-dual variable neighborhood search. *Data Min. Knowl. Discov.*, 19(3):351–375, 2009.
16. P. Hansen and B. Jaumard. Cluster analysis and mathematical programming. *Math. Program.*, 79(1-3):191–215, 1997.
17. E. R. Holzinger and M. D. Ritchie. Integrating heterogeneous high-throughput data for meta-dimensional pharmacogenomics and disease-related studies. *Pharmacogenomics*, 13(2):213–222, 2012.
18. H. F. Köhn, D. Steinley, and M. J. Brusco. The p-median model as a tool for clustering psychological data. *Psychol. Methods*, 15(1):87–95, 2010.
19. V. N. Kristensen, O. C. Lingjærde, H. G. Russnes, H. K. M. Vollan, A. Frigessi, and A.-L. Børresen-Dale. Principles and methods of integrative genomic analyses in cancer. *Nat. Rev. Cancer*, 14(5):299–313, 2014.
20. M. D. Ritchie, E. R. Holzinger, R. Li, S. A. Pendergrass, and D. Kim. Methods of integrating data to uncover genotype-phenotype interactions. *Nat. Rev. Genet.*, 16:85–97, 2015.
21. U. Scherf, D. T. Ross, M. Waltham, L. H. Smith, J. K. Lee, L. Tanabe, K. W. Kohn, W. C. Reinhold, T. G. Myers, D. T. Andrews, D. A. Scudiero, M. B. Eisen, E. A. Sausville, Y. Pommier, D. Botstein, P. O. Brown, and J. N. Weinstein. A gene expression database for the molecular pharmacology of cancer. *Nat. Genet.*, 24(3):236–244, 2000.
22. R. H. Shoemaker. The nci60 human tumour cell line anticancer drug screen. *Nat. Rev. Cancer*, 6(10):813–823, 2006.

23. A. V. Ushakov, I. L. Vasilyev, and T. V. Gruzdeva. A computational comparison of the p-median clustering and k-means. *International Journal of Artificial Intelligence*, 13(1):229–242, 2015.
24. I. Vasil'ev, X. Klimentova, and Yu. Kochetov. New lower bounds for the facility location problem with clients' preferences. *Comput. Math. Math. Phys.*, 49(6):1010–1020, 2009.
25. I. Vasilyev and X. Klimentova. The branch and cut method for the facility location problem with clients preferences. *J. Appl. Ind. Math.*, 4(3):441–454, 2010.
26. I. Vasilyev, X. Klimentova, and M. Boccia. Polyhedral study of simple plant location problem with order. *Oper. Res. Lett.*, 41(2):153–158, 2013.
27. Y. Wei. Integrative analyses of cancer data: A review from a statistical perspective. *Cancer Inform.*, pages 173–181, 05 2015.