An integer programming approach for the hyper-rectangular clustering problem with axis-parallel clusters and outliers

Javier Marenco

Sciences Institute, National University of General Sarmiento,
J. M. Gutiérrez 1150 (B1636GSX) Buenos Aires, Argentina,
jmarenco@campus.ungs.edu.ar

Abstract. We present a mixed integer programming formulation for the problem of clustering a set of points in $\mathbb{R}^d$ with axis-parallel clusters, while allowing to discard a pre-specified number of points, thus declared to be outliers. We identify a family of valid inequalities separable in polynomial time, we prove that some of them induce facets of the associated polytope, and we show that the dynamic addition of cuts coming from this family is effective in practice.

Keywords: clustering, integer programming, cutting planes

1 Introduction

Given a set $\mathcal{X} = \{x^1, \ldots, x^n\}$ of points in $\mathbb{R}^d$ and an integer p, the *hyper-rectangular clustering problem with axis-parallel clusters* consists in determining the p "smallest" axis-parallel hyper-rectangles $\mathbb{R}^d$ such that each point in $\mathcal{X}$ is included in at least one such hyper-rectangle. If we also specify a number q of possible *outliers*, then up to q points may not be included in any hyper-rectangle.

Hyper-rectangular clustering has been proposed as an alternative of *explainable clustering* [1], since it is straightforward to describe the obtained clusters by their bounds. Previous works for this problem [2, 4, 3, 5] ask all points to be clustered, i.e., $q = 0$ is assumed therein. In this work we tackle the case $q > 0$, namely the clustering may discard up to a pre-specified number q of points, which are thus declared to be outliers.

We consider the following mixed integer program minimizing the total cluster *span*. For $i \in [n] := \{1, \ldots, n\}$ and $c \in [p] := \{1, \ldots, p\}$, we consider the binary variable z_{ic} representing whether x^i is assigned to the cluster c or not. Also, for $c \in [p]$ and $t \in [d] := \{1, \ldots, d\}$, the real variables $l_c^t, r_c^t \in \mathbb{R}$ represent the lower and upper bound, respectively, of the cluster c in the axis t. For $t \in [d]$, define $X_t := \{x_t : x \in \mathcal{X}\}$, $\min_t := \min(X_t)$, and $\max_t := \max(X_t)$. In this setting, we can formulate the problem as follows.

$$\min \sum_{c=1}^{p} \sum_{t=1}^{d} r_c^t - l_c^t$$

$$\text{s.t.} \sum_{c=1}^{p} z_{ic} \leq 1 \qquad \forall i \in [n], \tag{1}$$

$$l_c^t + (\max_t - x_t^i) z_{ic} \leq \max_t \qquad \forall i \in [n], c \in [p], t \in [d], \tag{2}$$

$$r_c^t + (\min_t - x_t^i) z_{ic} \geq \min_t \qquad \forall i \in [n], c \in [p], t \in [d], \tag{3}$$

$$l_c^t \leq r_c^t \qquad \forall c \in [p], t \in [d], \tag{4}$$

$$\sum_{c=1}^{p} \sum_{i=1}^{n} z_{ic} \geq n - q, \tag{5}$$

$$z_{ic} \in \{0, 1\} \qquad \forall i \in [n], c \in [p], \tag{6}$$

$$\min_t \ \leq \ l_c^t, \ r_c^t \leq \max_t \qquad \forall c \in [p], t \in [d]. \tag{7}$$

The objective function asks to minimize the sum of all cluster spans along all axes. Constraints (1) ask every point to be assigned to at most one cluster, and a point is considered to be an outlier if it is assigned to no cluster. Constraints (2)-(3) bind the variables, whereas constraints (4) avoid bound crossings for empty clusters. Constraints (5) specify that at most q outliers can be selected, and constraints (6)-(7) specify the variable domains. We define $\mathcal{P}(P, p, q)$ to be the convex hull of all vectors $(z, l, r) \in \mathbb{R}^{np+2pd}$ satisfying (1)-(7).

Theorem 1. *Fix $c \in [c]$ and $t \in [d]$, and let $\alpha, \beta \geq 0$. The inequality*

$$\alpha \, r_c^t - \beta \, l_c^t \ \geq \ \sum_{i=1}^{n} \gamma_i z_{ic} - \delta \tag{8}$$

is valid for $\mathcal{P}(P, p, q)$ if and only if (a) $\alpha \, x_1 - \beta \, x_2 \geq \sum_{x_1 \leq x_t^i \leq x_2} \gamma_i - \delta$ for every $x_1, x_2 \in X_t$, (b) $\delta \geq \beta \max_t - \alpha \max_t$, and (c) $\delta \geq \beta \min_t - \alpha \min_t$.

The family of valid inequalities identified by Theorem 1 includes facet-defining inequalities, as the following result shows.

Theorem 2. *Assume $x_t^1 \leq x_t^2 \leq \cdots \leq x_t^n$ and $x_t^1 < x_t^n$. Let $c \in [p]$, $t \in [d]$, and $s \in [n]$. Fix $\alpha, \beta \geq 0$ and let $\delta := \min\{(\beta - \alpha) x_t^n, (\beta - \alpha) x_t^1\}$. If $\alpha x_t^i + \delta \geq \beta x_t^{i+1}$ for $i = 1, \ldots, n-1$, then the inequality (8) defines a facet of $\mathcal{P}(P, p, q)$, with $\gamma_s := (\alpha - \beta) x_t^s + \delta$, $\gamma_i := \beta(x_t^{i+1} - x_t^i)$ for $i = 1, \ldots, s-1$, and $\gamma_i := \alpha(x_t^i - x_t^{i-1})$ for $i = s+1, \ldots, n$.*

Theorem 1 shows that it suffices to check $O(n^2)$ conditions in order to guarantee validity of the inequality (8). This allows for a polynomial separation procedure for these inequalities, via linear programming. Given a fractional solution (z^*, l^*, r^*), we consider the following formulation.

$$\max \ \ \beta \, (l^*)_c^t - \alpha \, (r^*)_c^t + \sum_{i=1}^{n} (z^*)_{ic} \, \gamma_i - \delta$$

$$\sum_{x_1 \leq x_t^i \leq x_2} \gamma_i \leq x_1 \alpha - x_2 \beta + \delta \qquad \forall x_1, x_2 \in X_t \tag{9}$$

$$\beta \max_t - \alpha \max_t \leq \delta \tag{10}$$
$$\beta \min_t - \alpha \max_t \leq \delta \tag{11}$$
$$\alpha + \beta = n + 1 \tag{12}$$
$$\alpha, \beta \geq 0$$
$$\gamma_i \geq 0 \qquad \forall i \in [n]$$

The objective function asks to maximize the cut depth. Constraints (9)-(11) enforce the validity conditions (a)-(c) specified by Theorem 1, whereas constraint (12) normalizes the coefficients. If the optimal value of this linear program is positive, then the optimal solution provides the coefficients of a violated inequality (8) and viceversa. Since the fractional solution (z^*, l^*, r^*) only participates in the objective function, we can set up one such linear program for each axis $t \in [d]$, and warm-start the resolution by updating the coefficients in the objective function every time a new fractional solution must be separated.

We have implemented a branch and cut procedure for the hyper-rectangular clustering problem with axis-parallel clusters and outliers within the framework provided by Cplex 12.4. The dynamic addition of cuts coming from the separation procedure specified above allows to trim the overall running time for instances up to 80 points, due to a dramatic improvement in the total number of nodes in the enumeration tree. Also, the dynamic addition of cuts helps to solve with optimality larger instances than with out-of-the-box solvers. As future work, it is important to consider better separation strategies, since the separation overhead degrades the overall performance for large instances.

An undesirable property of the formulation (1)-(7) is the presence of symmetry among the clusters. However, the addition of straightforward symmetry-breaking constraints does not seem to improve running times in our experiments, having in fact the opposite effect. Also, preliminary experiments with a simple column-generation-based procedure over an extended formulation do not seem promising either. The exploration of effective symmetry-breaking techniques for this formulation is left as future work.

References

1. Bhatia, A., Garg, V., Haves, P., Pudi, V.: Explainable clustering using hyper-rectangles for building energy simulation data. IOP Conference Series: Earth and Environmental Science 238 012–068 (2019).
2. Lee, S., Chung, C.: Hyper-rectangle based segmentation and clustering of large video data sets. Information Sciences 141 (1-2) 139–168 (2002).
3. Mago, V., Bhatia, N., Park, S.: Classification with axis-aligned rectangular boundaries. In: Mago, V., Bhatia, N. (eds.) "Cross-Disciplinary Applications of Artificial Intelligence and Pattern Recognition", Information Science Reference (2012).
4. Ordóñez, C., Omiecinski, E., Navathe, S., Ezquerra, N.: A clustering algorithm to discover low and high density hyper-rectangles in subspaces of multidimensional data. Georgia Institute of Technology Technical Report GIT-CC-99-20 (1999).
5. Park, S., Kim, J.: Unsupervised clustering with axis-aligned rectangular regions. Stanford University Technical Report (2009).