Modelos de programación lineal entera para el problema de clustering con regiones hiper-rectangulares y outliers

Javier Marenco

Departamento de Computación, FCEyN, Universidad de Buenos Aires, Argentina
Instituto de Ciencias, Universidad Nacional de General Sarmiento, Argentina
jmarenco@dc.uba.ar

Dado un conjunto $\mathcal{X}$ de puntos en $\mathbb{R}^d$ y un entero k, el problema de *clustering con regiones hiper-rectangulares* consiste en determinar k hiper-rectángulos en $\mathbb{R}^d$ con el menor volumen posible de modo tal que cada punto de $\mathcal{X}$ esté incluido en al menos un hiper-rectángulo. Si además se especifica una cantidad p de posibles *outliers*, entonces se pueden tener hasta p puntos de $\mathcal{X}$ no incluidos en ningún hiper-rectángulo.

Las técnicas de *clustering* con hiper-rectángulos han sido propuestas como una alternativa de *clustering interpretable* [1], dado que es sencillo explicar los *clusters* obtenidos en función de sus límites. Existen métodos geométricos para este problema [2, 4], y también se han explorado alternativas basadas en programación lineal entera para variantes de este problema [3, 5]. En todos estos trabajos se asume $p = 0$.

En este trabajo estudiamos el problema de clustering con regiones hiper-rectangulares con una linealización de la función objetivo y para el caso $p > 0$. Es decir, se puede descartar una cantidad prefijada de puntos, que son declarados como *outliers*. Presentamos un modelo natural de programación lineal entera para este problema y estudiamos el poliedro asociado. Además, consideramos un esquema heurístico basado en generación de columnas, y presentamos experimentos computacionales para comparar los dos esquemas.

References

1. A. Bhatia, V. Garg, P. Haves y V. Pudi, *Explainable clustering using hyper-rectangles for building energy simulation data.* IOP Conference Series: Earth and Environmental Science 238 012068 (2019).
2. S. Lee y C. Chung, *Hyper-rectangle based segmentation and clustering of large video data sets.* Information Sciences 141 (1-2) (2002) 139–168.
3. V. Mago, N. Bhatia y S. Park, *Classification with axis-aligned rectangular boundaries.* Capítulo del libro: V. Mago and N. Bhatia (editores), "Cross-Disciplinary Applications of Artificial Intelligence and Pattern Recognition", Information Science Reference, 2012.
4. C. Ordóñez, E. Omiecinski, S. Navathe y N. Ezquerra, *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. S. Park y J. Kim, *Unsupervised clustering with axis-aligned rectangular regions.* Stanford University Technical Report (2009).