Technical notes on the `anticlust` package

Martin Papenberg

2019-02-22

This document explains the technical and algorithmical background of the `R` package `anticlust`. **It is still a work in progress.** The following topics are covered:

1. A formalization of the anticlustering problem
2. A description of the objective functions used to measure anticluster similarity
3. A documentation of the algorithms used to optimize the objective funtions

1 Problem formalization

A set of n d-dimensional data points $X = \{x_i\}$ $(i \in \{1, ..., n\})$ has to be partitioned into K clusters $C = \{c_k, k = 1, ..., K\}$, satisfying the following restrictions:

$$\bigcup_{k=1}^{K} c_k = X \tag{1}$$

$$S_k \cap S_j = \emptyset,\ \forall k, j \in \{1, ..., K\},\ k \neq j \tag{2}$$

$$|c_k| = |c_j|,\ \forall k, j \in \{1, ..., K\} \tag{3}$$

Restriction (1) ensures that each element from the underlying set X is assigned to an anticluster; restriction (2) ensures that each element is assigned to only one anticluster; restriction (3) ensures that each anticluster contains the same number of elements. It follows that $|c_k| = \frac{n}{K}$ $\forall k \in \{1, ..., K\}$. The objective is to select a partitioning that maximizes the similarity of the K anticlusters.

Note that restriction (3) is currently implemented for all methods in the `anticlust` package, but it is not an obligatory restriction for anticlustering in general.

2 Objective functions

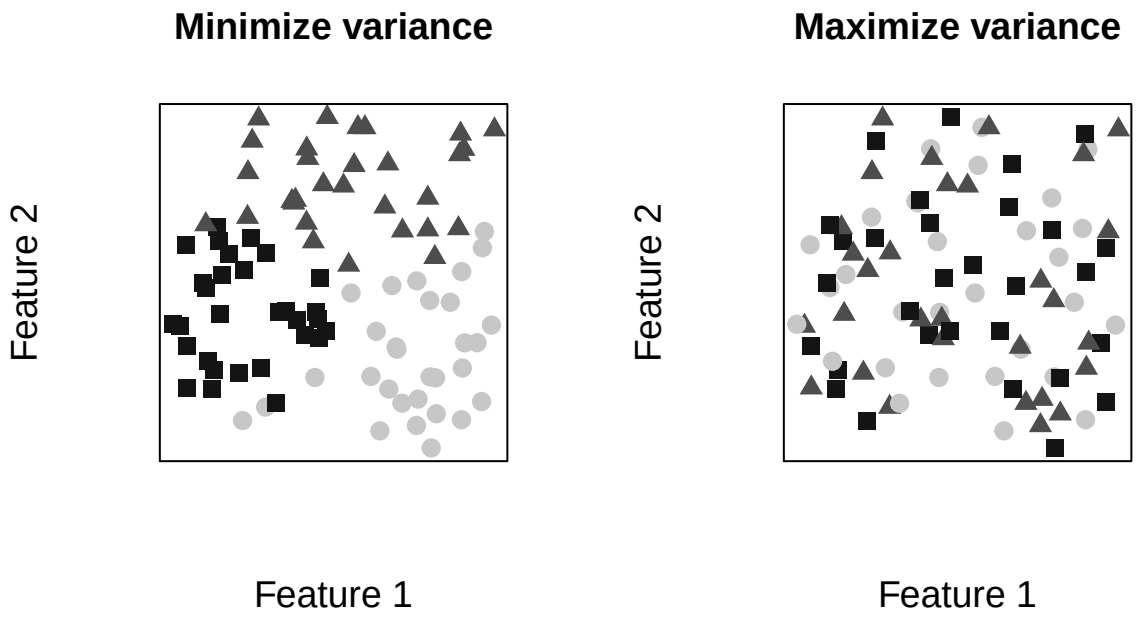
This section presents definitions of optimal set similarity as assumed in the `anticlust` package.

2.1 The variance objective

Späth (1986) and Valev (1998) independently proposed to maximize the variance criterion used in k-means clustering to create similar anticlusters. The variance criterion is given by sum of the squared errors between cluster centers (μ_k) and individual data points (Jain, 2010):

$$\sum_{k=1}^{K} \sum_{x_i \in c_k} ||x_i - \mu_k||^2 \tag{4}$$

The following plot graphically illustrates efforts to maximize and to minimize the variance criterion in a 2-dimensional feature space for three (anti)clusters, respectively. The partitions employ restrictions (1) - (3),



including the restriction of equal (anti)cluster sizes. Optimizing the variance objective is a computationally difficult problem that is usually tackled using heuristic methods (Jain, 2010; Späth, 1986; Valev, 1998).

2.2 The distance objective

In addition to the variance criterion, the `anticlust` package introduces another clustering objective to the anticlustering application. This objective has been used in the problem domain of cluster editing and is based on a measure of the pairwise dissimilarities of data point (Böcker & Baumbach, 2013; Rahmann et al., 2007).[1] In weighted cluster editing, the optimal objective is found when the sum of within-cluster dissimilarities is minimized; for the anticlustering application, the objective is maximized instead.

To formalize the cluster editing objective, we use variables x_{ij} to encode whether two data points x_i and x_j belong to the same anticluster c_k:

$$x_{ij} = \begin{cases} 1 & \text{if } x_i \in c_k \wedge x_j \in c_k \\ 0 & \text{otherwise} \end{cases}$$

Assume that d_{ij} represents a measure of the dissimilarity between two data points x_i and x_j, for example given as the euclidean distance. The cluster editing distance objective is then given as follows (Grötschel & Wakabayashi, 1989; Miyauchi & Sukegawa, 2015):

$$\sum_{1 \leq i < j \leq n} d_{ij}\, x_{ij} \tag{5}$$

Hence, the cluster editing objective is given as the sum of distances of elements within the same cluster. I refer to this objective function as the "distance objective" as opposed to the "variance objective" in (4).

Maximizing the distance objective corresponds to minimizing the average linkage distance between partitions. In hierarchical cluster algorithms, the average linkage distance is a quantification of the similarity of two

[1] Cluster editing has also been studied under different names such as correlation clustering (Bansal, Blum, & Chawla, 2004), clique partition problem (Grötschel & Wakabayashi, 1989), and transitivity clustering (Wittkop et al., 2010).

clusters (Bacon, 2001; Guha, Rastogi, & Shim, 1998). To appreciate the correspondence of the distance objective and the average linkage method, consider the total sum of paired distances of all elements. The total sum of all distances can be partitioned into within-cluster and between-cluster distances:

$$\sum_{1 \leq i < j \leq n} d_{ij} = \sum_{1 \leq i < j \leq n} d_{ij}\, x_{ij} + \sum_{1 \leq i < j \leq n} d_{ij}\, (1 - x_{ij}) \tag{6}$$

The total sum of distances is not influenced by the concrete partitioning x_{ij}. Hence, the following optimizations lead to the same partitioning x_{ij}:

$$\text{Maximize} \sum_{1 \leq i < j \leq n} d_{ij}\, x_{ij}$$

$$\text{Minimize} \sum_{1 \leq i < j \leq n} d_{ij}\, (1 - x_{ij})$$

In the special case of two partitions A and B ($K = 2$), the sum of the between-cluster distances can be written as follows:

$$\sum_{i \in A} \sum_{j \in B} d_{ij} \tag{7}$$

This formulation is very close to the average linkage objective that however also incorporates the cardinalities of the sets A and B (Guha et al., 1998):

$$\frac{1}{|A|\,|B|} \sum_{i \in A} \sum_{j \in B} d_{ij} \tag{8}$$

Given restriction (3) for the anticlustering problem, the partitions A and B are of equal size; therefore, (7) gives the same information with regard to the similarity of clusters as (8). The cluster editing objective is hence a generalization of the average linkage measure on more than two clusters.

3 Algorithmic approaches

Finding optimal data partitions usually corresponds to NP-complete problems (Arabie & Hubert, 1996; Bansal et al., 2004; Jain, 2010). For NP-complete problems, it is often infeasible to find the optimal objective, especially when n is large. To find the optimal solution for moderately large instances, `anticlust` employs integer linear programming. To process larger problem instances, `anticlust` uses heuristic methods based on repeated random sampling.

3.1 NP-completeness

In the following, I show that anticlustering using the distance criterion is NP-complete. First, the distance objective can be computed in polynomial time for a given partitioning C because the summation of all distance values d_{ij} is in $O(n^2)$.

Second, I show that if an efficient algorithm exists to solve distance anticlusting in polynomial time, it is also possible solve the NP-complete balanced number partitioning problem in polynomial time (Mertens, 2001). In the number partitioning problem, we have a list of positive integers $a_1, a_2, ..., a_n$ and try to find a subset $A \subset \{1, ..., n\}$ that minimizes the partition difference

$$E(A) = \left| \sum_{i \in A} a_i - \sum_{j \notin A} a_j \right| \tag{9}$$

In the balanced version of number partitioning, we can impose the restriction of $|A| = \frac{n}{2}$ – assuming that n is even – corresponding to restriction (3) of equal cluster sizes in anticlusting (Mertens, 2001).

To convert the number partitioning formulation into a formulation of distance anticlustering, we define d_{ij} as the absolute difference representing the "dissimilarity" of two numbers:

$$d_{ij} := |a_i - a_j| \tag{10}$$

We thus obtain

$$E(A) = \sum_{i \in A} \sum_{j \notin A} d_{ij} \tag{11}$$

Using variables $x_{ij} \in \{0, 1\}$ to represent whether two numbers belong to the same subset, i.e.,

$$x_{ij} = \begin{cases} 1 & \text{if } (x_i \in A \land x_j \in A) \lor (x_i \notin A \land x_j \notin A) \\ 0 & \text{otherwise} \end{cases}$$

we obtain $E(A)$ as the distance anticlustering objective:

$$E(A) = \sum_{1 \leq i < j \leq n} d_{ij}\, x_{ij} \tag{12}$$

Hence, anticlustering using the distance objective is equivalent to the balanced number partitioning problem in the special case where

a) $K = 2$
b) each element is described by an integer
c) d_{ij} is the absolute difference

Therefore, if a polynomial-time algorithm exists that solves distance anticlustering, we can solve the NP-complete balanced number partitioning in polynomial time. Hence, distance anticlustering is NP-complete.

3.2 Integer linear programming

The `anticlust` package uses integer linear programming to solve distance anticlustering exactly (Böcker, Briesemeister, & Klau, 2011; Grötschel & Wakabayashi, 1989). Despite the NP-complete nature of cluster editing, integer linear programming (ILP) has successfully been used to find optimal solutions even for relatively large cluster editing problem instances (Böcker et al., 2011; L. H. N. Lorena, Quiles, Carvalho, & Lorena, 2018).

The `anticlust` package uses an ILP based on formulations proposed by Grötschel and Wakabayashi (1989). The complete problem formulation used in the package `anticlust` is given as follows:

$$\text{Maximize} \sum_{1 \leq i < j \leq n} d_{ij}\, x_{ij} \tag{13}$$

$$-x_{ij} + x_{ik} + x_{jk} + y_k \leq 1, \qquad \forall\, 1 \leq i < j < k \leq n, \tag{14}$$

$$x_{ij} - x_{ik} + x_{jk} \leq 1, \qquad \forall\, 1 \leq i < j < k \leq n, \tag{15}$$

$$x_{ij} + x_{ik} - x_{jk} \leq 1, \qquad \forall\, 1 \leq i < j < k \leq n, \tag{16}$$

$$\sum_{1 \leq i < j \leq n} x_{ij} + \sum_{1 \leq k < i \leq n} x_{ki} = \frac{n}{K} - 1, \qquad \forall i \in \{1, ..., n\} \tag{17}$$

$$x_{ij} \in \{0, 1\},\ \forall\, 1 \leq i < j \leq n \tag{18}$$

The inequalities (14) - (16) are called triangular constraints and were developed by Grötschel and Wakabayashi (1989); they ensure that all within-cluster distances are used to compute the objective function and between-cluster distances are ignored. They also enforce the anticlustering restrictions (1) and (2), i.e., they ensure that each element is assigned to exactly one anticluster. The equality (17) enforces that each anticluster consists of the same number of elements $\frac{n}{K}$. This condition is ensured by enforcing that each element x_i is connected with $\frac{n}{K} - 1$ elements within the same anticluster. Constraint (18) ensures that the partitioning variables x_{ij} are binary.

3.2.1 Additional restrictions

TODO: Explain preclustering

3.3 The heuristic approach

- Simulated annealing
- Neighborhood of elements, preclustering

4 References

Arabie, P., & Hubert, L. J. (1996). An overview of combinatorial data analysis. In P. Arabie, L. J. Hubert, & G. De Soet (Eds.), *Clustering and classification* (pp. 5–63).

Bacon, D. R. (2001). An evaluation of cluster analytic approaches to initial model specification. *Structural Equation Modeling*, *8*(3), 397–429.

Bansal, N., Blum, A., & Chawla, S. (2004). Correlation clustering. *Machine Learning*, *56*(1-3), 89–113.

Böcker, S., & Baumbach, J. (2013). Cluster editing. In *Conference on Computability in Europe* (pp. 33–44). Springer.

Böcker, S., Briesemeister, S., & Klau, G. W. (2011). Exact algorithms for cluster editing: Evaluation and experiments. *Algorithmica*, *60*(2), 316–334.

Grötschel, M., & Wakabayashi, Y. (1989). A cutting plane algorithm for a clustering problem. *Mathematical Programming*, *45*(1-3), 59–96.

Guha, S., Rastogi, R., & Shim, K. (1998). CURE: An efficient clustering algorithm for large databases. In *ACM sigmod record* (Vol. 27, pp. 73–84). ACM.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, *31*(8), 651–666.

Lorena, L. H. N., Quiles, M. G., Carvalho, A. C. P. de L. F. de, & Lorena, L. A. N. (2018). Preprocessing technique for cluster editing via integer linear programming. In D.-S. Huang, V. Bevilacqua, P. Premaratne, & P. Gupta (Eds.), *Intelligent computing theories and application* (pp. 287–297). Cham: Springer International Publishing.

Mertens, S. (2001). A physicist's approach to number partitioning. *Theoretical Computer Science*, *265*(1-2), 79–108.

Miyauchi, A., & Sukegawa, N. (2015). Redundant constraints in the standard formulation for the clique partitioning problem. *Optimization Letters*, *9*(1), 199–207.

Rahmann, S., Wittkop, T., Baumbach, J., Martin, M., Truss, A., & Böcker, S. (2007). Exact and heuristic algorithms for weighted cluster editing. In *Computational systems bioinformatics: (Volume 6)* (pp. 391–401). World Scientific.

Späth, H. (1986). Anticlustering: Maximizing the variance criterion. *Control and Cybernetics*, *15*(2), 213–218.

Valev, V. (1998). Set partition principles revisited. In *Joint IAPR international workshops on statistical techniques in pattern recognition (SPR) and structural and syntactic pattern recognition (SSPR)* (pp. 875–881). Springer.

Wittkop, T., Emig, D., Lange, S., Rahmann, S., Albrecht, M., Morris, J. H., ... Baumbach, J. (2010). Partitioning biological data with transitivity clustering. *Nature Methods*, *7*(6), 419–420.