A Set Correlation Model for Partitional Clustering

Michael E. HOULE
NII

Nguyen Xuan VINH
National ICT Australia, University of New South Wales

Why

Traditional clustering-based data discovery tools often fail when the data sets are very large or have many attributes, or have performance parameters that are very difficult to tune in practice. We have developed a novel formulation for partitional data clustering that overcomes many of these difficulties, based on the generic Relevant Set Correlation (RSC) model.

What

The new model, GlobalRSC, resembles the famous K-means clustering heuristic, but with a shared-neighbor similarity measure instead of the Euclidean distance. Unlike K-means and most other clustering heuristics that can only work with real-valued data and distance measures taken from specific families, GlobalRSC has the advantage that it can work with any distance measure, and any data representation.

The Model

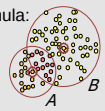
QUERY RELEVANT SETS

- How can we cluster when the nature of the data is hidden from us?
- Only assumption: oracle providing relevancy rankings for queries-by-example.
- $Q(q, k)$: ranked relevant set for query item q, of size k.
- Clusters to be patterned on query relevant set (QRS) $Q(q, k)$ for some q in S.

INTER-SET ASSOCIATION

- Pearson correlation formula:

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)}}$$



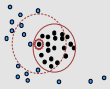
- Applying this to coordinate pairs of set vectors for $A, B \subset S$ gives set correlation between A and B :

$$R(A,B) = \frac{|S|}{\sqrt{(|S|-|A|)(|S|-|B|)}}\left(\frac{|A\cap B|}{\sqrt{|A|\cdot|B|}} - \frac{\sqrt{|A|\cdot|B|}}{|S|}\right)$$

CLUSTER COHERENCE

- Association criterion – if v belongs to cluster set C, then the items relevant to v should belong to C.
  - $R(C,Q(v,|C|))$ should be high.
- Measure of cluster coherence: the total correlation between a set and the (same-sized) relevant sets of its members.

$$\sum_{v\in A} R(A, Q(v,|A|))$$



CLUSTERING CRITERION

- Note: neighborhood size is variable, depending upon the cluster size.
- Set correlation is unbiased in terms of cluster sizes.
- Clustering criterion maximizes total cluster coherence.

$$\mathbf{R}(A) = \frac{1}{n}\sum_{i=1}^{K}\sum_{v\in A} R(A, Q(v,|A|))$$

COMPARISON WITH K-MEANS

K-Means:

- Clustering criterion measures average cluster coherence.

$$\mathbf{D}(A) = \frac{1}{n}\sum_{i=1}^{K}\sum_{v\in A} dist^2(A, c(A))$$

- Discovers globular/compact clusters.
- Works on numeric data.
- Works only with certain distance measures (such as the Euclidean distance, Bregman divergences).

GlobalRSC:

- Clustering criterion measures average cluster coherence.

$$\mathbf{R}(A) = \frac{1}{n}\sum_{i=1}^{K}\sum_{v\in A} R(A, Q(v,|A|))$$

- Discovers globular/compact clusters.
- Works with any data type.
- Any primary distance measure can be employed.

THE CLUSTERING HEURISTIC

- 2-phase hill-climbing heuristic similar to K-means: consider reassigning each data point to a new cluster, in such a way that the objective value is increased.
- Batch update: update cluster membership after all data points has been considered. Fast but does not ensure objective improvement.
- Incremental update: cluster membership is updated right away for each point. Slow but objective improvement is guaranteed, ensuring convergence.
- Complexity: $\tilde{O}\left(dn^2 + \frac{n^2}{K}\log n\right)$

GlobalRSC

Input: The data set S; the number of desired clusters K; (optionally) a hard initial clustering $\mathcal{A} = \{A_1, A_2, \ldots, A_K\}$ on S; minimum neighborhood set size m; neighborhood set buffer size b.

0. **Initialization:** If no initial clustering was provided, compute an initial clustering $\mathcal{A} = \{A_1, A_2, \ldots, A_K\}$ as follows:
   (a) Select K items of S uniformly at random as the seeds of the K clusters.
   (b) Assign each data item v to the cluster corresponding to the seed closest to v.
   (c) Build the neighborhood Q^v for each item $v \in A_i$, with size equal to $\min\{\lceil(1+b)|A_i|\rceil, m\}$.

1. **Batch phase:**
   (a) Calculate $\mathcal{R}(A_1), \mathcal{R}(A_2), \ldots, \mathcal{R}(A_K)$ and set the termination flag $halt \leftarrow$ FALSE.
   (b) Repeat until $halt =$ TRUE:
   i. Set $halt \leftarrow$ TRUE.
   ii. For each $v \in S$, build the *inverted neighborhood* I^v, where $r \in I^v$ if and only if $v \in Q^r$.
   iii. For each data item v currently in cluster A_i:
   A. Build the list $\mathfrak{C}$ of clusters (other than A_i) containing at least one item of Q^v.
   B. Tentatively reassign v to each of the A_j in $\mathfrak{C}$.
   C. Using the inverted neighborhood sets, calculate the index j for which the improvement value $\mathcal{R}(A_j \cup \{v\}) + \mathcal{R}(A_i \setminus \{v\}) - \mathcal{R}(A_j) - \mathcal{R}(A_i)$ is maximized.
   D. If the improvement value is positive, record (v, j) for future reassignment.
   iv. Tentatively apply all the recorded reassignments for this round, and let $\mathcal{A}' = \{A'_1, A'_2, \ldots, A'_K\}$ be the resulting clustering. Calculate $\mathcal{R}(\mathcal{A}')$. If $\mathcal{R}(\mathcal{A}') > \mathcal{R}(\mathcal{A})$, set $halt \leftarrow$ FALSE and $\mathcal{A} \leftarrow \mathcal{A}'$. Otherwise, proceed to the incremental phase.

2. **Incremental phase:**
   (a) Set $halt \leftarrow$ FALSE.
   (b) Repeat until $halt =$ TRUE:
   i. Set $halt \leftarrow$ TRUE.
   ii. For each data item v currently in cluster A_i:
   A. Build the list $\mathfrak{C}$ of clusters (other than A_i) that contribute items to Q^v.
   B. Tentatively reassign v to each of the A_j in $\mathfrak{C}$, and calculate the index j for which the improvement value $\mathcal{R}(A_j \cup \{v\}) + \mathcal{R}(A_i \setminus \{v\}) - \mathcal{R}(A_j) - \mathcal{R}(A_i)$ is maximized.
   C. If the improvement value is positive, reassign v to A_j immediately, adjust the values of $\mathcal{R}(A_j)$ and $\mathcal{R}(A_i)$, and set $halt \leftarrow$ FALSE.

High-Dimensional Data

- Neighborhood construction becomes expensive.
- Local optimization approaches become less efficient without a good initialization.

Scalability

- Initialization: using greedyRSC.
- Construction of neighborhoods using approximate similarity search structures, such as RCT or SASH.
- Restrict largest cluster size.

- Reuters Corpus Volume I (RCV1).
  - T1: 200,000 documents x 320648 features
  - T2: 20,000 documents x 320648 features

Data	Algorithm	ARI	AMI	EPrec	ERec	ECos	Loops	Time (s)
T2	CLUTO*	0.02±0.00	0.24±0.00	0.60±0.00	0.02±0.00	0.10±0.00	N/A	1201±14
	GreedyRSC	0.04±0.00	0.26±0.00	0.61±0.00	0.04±0.00	0.12±0.00	N/A	495±2
	GlobalRSC	0.09±0.00	0.29±0.00	0.64±0.01	0.07±0.00	0.16±0.00	17±4	1616±167
T1	GreedyRSC	0.01±0.00	0.25±0.00	0.71±0.00	0.01±0.00	0.06±0.00	N/A	10657±334
	GlobalRSC	0.02±0.00	0.26±0.00	0.72±0.00	0.02±0.00	0.08±0.00	18±5	19044±756

*: K-means and CLUTO run using the number of clusters K as determined by GreedyRSC

- 3 microarrays data sets with known class labels.
  - B1: 384 genes x 17 conditions
  - B2: 237 genes x 17 conditions
  - B3: 205 genes x 20 conditions

Data	Algorithm	ARI	AMI	EPrec	ERec	ECos	Loops	Time (s)
B1	K-means	0.42±0.05	0.48±0.02	0.55±0.02	0.56±0.04	0.55±0.03	11±4	0±0
	CLUTO	**0.50±0.00**	**0.52±0.00**	**0.59±0.00**	0.61±0.00	**0.60±0.00**	N/A	0±0
	GlobalRSC	0.48±0.01	0.50±0.01	0.56±0.01	**0.62±0.02**	0.58±0.01	11±3	2±1
B2	K-means	0.49±0.03	0.33±0.01	0.65±0.01	0.54±0.02	0.59±0.01	8±3	0±0
	CLUTO	**0.51±0.00**	**0.34±0.00**	0.65±0.00	**0.56±0.00**	**0.60±0.00**	N/A	0±0
	GlobalRSC	0.50±0.02	0.29±0.01	**0.69±0.01**	0.48±0.01	0.56±0.01	11±3	1±1
B3	K-means	0.83±0.09	0.78±0.08	0.92±0.02	0.83±0.09	0.86±0.06	7±4	0±0
	CLUTO	**0.96±0.00**	**0.91±0.00**	**0.95±0.00**	0.96±0.00	**0.96±0.00**	N/A	0±0
	GlobalRSC	0.92±0.04	0.84±0.07	0.92±0.05	**0.96±0.01**	0.93±0.03	3±0	0±1

- ALOI data sets, 110250 images of 1000 common objects.
  - I1: 13943 images x 641 features
  - I2: 110250 images x 641 features

Data	Algorithm	ARI	AMI	EPrec	ERec	ECos	Loops	Time (s)
I1	K-means	0.48±0.01	0.72±0.00	0.66±0.01	0.56±0.01	0.56±0.01	20±4	1278±347
	K-means*	0.49±0.01	0.74±0.00	0.61±0.01	0.59±0.01	0.55±0.01	23±5	993±416
	CLUTO	0.53±0.00	0.75±0.00	**0.67±0.00**	0.61±0.00	0.59±0.00	N/A	1015±51
	CLUTO*	0.53±0.00	0.75±0.00	0.62±0.00	0.62±0.00	0.58±0.00	N/A	847±9
	GreedyRSC	0.59±0.00	0.77±0.00	0.64±0.00	0.68±0.00	0.62±0.00	N/A	328±24
	GlobalRSC	**0.61±0.01**	**0.78±0.00**	0.66±0.00	**0.71±0.00**	**0.64±0.00**	17±3	417±29
I2	CLUTO	0.41±0.00	0.73±0.00	0.48±0.00	**0.52±0.00**	0.48±0.00	N/A	7893±16
	CLUTO*	0.48±0.00	0.69±0.00	0.80±0.00	0.35±0.00	0.50±0.00	N/A	27302±192
	GreedyRSC	0.56±0.00	0.72±0.00	0.82±0.00	0.42±0.00	0.56±0.00	N/A	4352±53
	GlobalRSC	**0.59±0.00**	**0.74±0.00**	**0.85±0.00**	0.46±0.00	**0.59±0.00**	28±3	5002±127

NII 連絡先：Michael E. HOULE（フール　マイケル）／ 国立情報学研究所　客員教授
TEL：03-4212-2538　FAX：03-3556-1916　Email：meh@nii.ac.jp