A Comparison of Knee Strategies for Hierarchical Spatial Clustering

Brian J. Ross

Department of Computer Science
Brock University
St. Catharines, Ontario, Canada
bross@brocku.ca

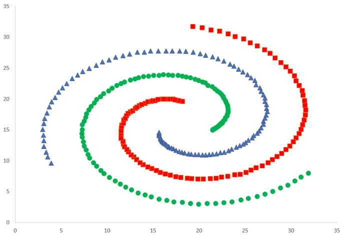
IEA-AIE 2018

Overview

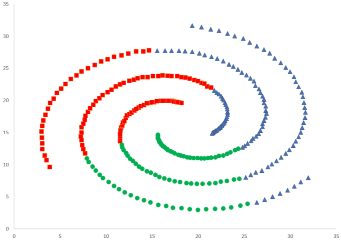
- Introduction
- Setup
  - Clustering algorithms
  - Knee heuristics
  - Data sets
- Results
- Conclusion

Introduction

- **Hierarchical clustering**: automatic grouping of data into sets with similar characteristics
  - Incrementally build clusters, from K clusters of 1 point each, to 1 cluster of all K points.
  - Dendogram represents incremental cluster creation by clustering algorithm.
  - Determine optimal clustering afterwords.
- **Clustering of 2D spatial data**: group planar points into sets
- **Computational limitations**
  - Clustering is in general NP-complete.
  - Optimality is often subjective and ill-defined.

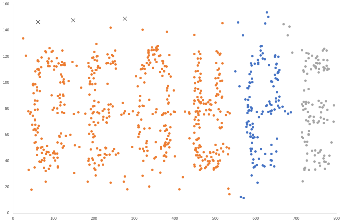


Spiral and single-linkage clustering, K=3.



Spiral and group average clustering, K=3.

Introduction



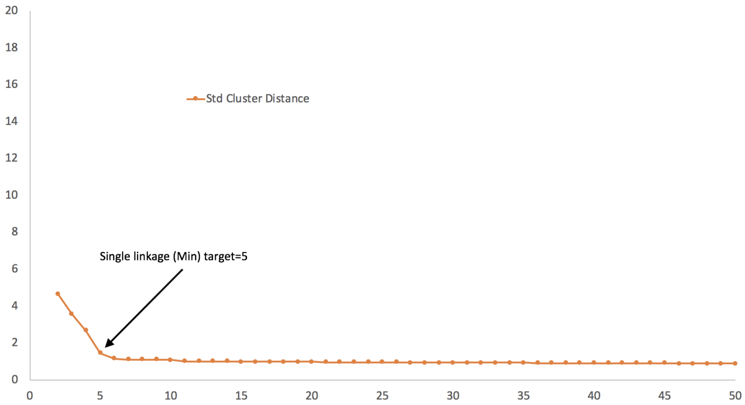
t5.8k and single-linkage, K=3.

t5.8k and group average, K=6.

Introduction

- **Knee**: heuristic for determining an optimal clustering
  - Conventional dendogram denotes distance measures used during incremental clustering merging.
  - Typically, the knee is a "bend" in the dendogram, that visually denotes the optimal clustering.
  - Knee shows point of *maximal marginal rate of return*. [Zhang *et al.* 2014]

Introduction



Aggregation dataset, standard dendogram, single-linkage clustering

Successful knee heuristics: max magnitude, max ratio, 2nd derivative

Introduction

- Issues
  - Different clustering algorithms.
  - Clustering is imperfect.
  - Optimal clustering is ill-defined.
  - Different datasets.
  - Different ways to characterize a knee.
  - Different ways to characterize *distance* in dendogram.
  - Knees don't always work, because they might not exist.

Introduction

- Goal: Comparative study...
  - Knee strategies (9)
  - 2D spatial datasets (16)
  - Clustering algorithms (2)
  - Dendogram distance measures (3)
  - $\Rightarrow$ Total 756 cases.
    (Not all datasets used for one clustering algorithm.)

- How do knee strategies compare, given the above parameters?

Setup: Hierarchical Clustering

Table 1. Hierarchical Clustering Algorithm

Input: $S_i = (x_i, y_i), i = 1, ..., K$
Output: Dendogram.

Initialization:
For i=1,...,K
    Add new cluster $C_{\{i\}}$
For all C_i, C_j $(i \neq j)$
    $Distance(C_i, C_j) \leftarrow DistanceMeasure(S_i, S_j)$

Cluster generation:
For i=1,...,K {
    Find C_p, C_q with minimum $d_{min} \leftarrow Distance(C_p, C_q)$.
    $Dendogram.Dist_i \leftarrow d_{min}$
    $Dendogram.Clust_i \leftarrow (C_p, C_q)$
    Remove C_p and C_q.
    Add new cluster $C_{p \cup q}$.
    Update *Distance* table:
        Remove all distances referring to C_p and C_q.
        For all clusters $C_w \neq C_{p \cup q}$
            $Distance(C_w, C_{p \cup q}) \leftarrow DistanceMeasure(C_w, C_{p \cup q})$
}

Return *Dendogram*.

Setup: Clustering Algorithms

1. Single-linkage: when clusters p and q merged, distance table for other clusters w revised...

$$\begin{array}{l} Distance(C_w, C_{p\cup q}) = \\ \quad minimum(Distance(C_w, C_p),\ Distance(C_w, C_q)) \end{array}$$

2. Group average:

$$\begin{array}{l} Distance(C_w, C_{p\cup q}) = \\ \quad average(Distance(C_w, C_p),\ Distance(C_w, C_q)) \end{array}$$

Setup: Dendogram Measures

1. Standard distance (Std): distance used by clustering algorithm.
2. Global average medoid distance (Avg Med):

$$AvgMed = \frac{\Sigma_{i=1}^{K} MD_i}{T}$$

   where MD_i is avg distance of medoid to other elements in cluster i, and T is total # clusters.
3. Global average centroid distnace (Avg Cent):

$$AvgCent = \frac{\Sigma_{i=1}^{K} CD_i}{T}$$

   where CD_i is avg distance of centroid to other elements in cluster i.

Setup: Knee Strategies

1. Magnitude: maximum $d_{i+1} - d_i$.
2. Ratio: maximum d_{i+1}/d_i.
3. Second derivative: maximum second derivative.
4. Minimum: minimum value.
5. L-method: [Salvador and Chan 2004] Fit 2 line segments to dendogram with min RMSE. Node at intersection is knee.
   - 6. L-method D: If N points on LHS line, then use next N points for RHS line.
   - 7. L-method S: If N points on LHS line, then evenly sample N points on dendogram for RHS line.

Setup: Knee Strategies (cont.)

- F score: Based on F test of one-way ANOVA, applied at each node of dendogram.
  8. F score A: The highest i in which:

     $$(f_{i+1} - f_i) > \delta^2_{1...i}$$

     where $\delta^2_{1..i}$ is the std dev of F scores 1 to i.
  9. F score B: The highest i in which:

     $$(f_{i+1} - f_i) > \delta^2_{1...k}$$

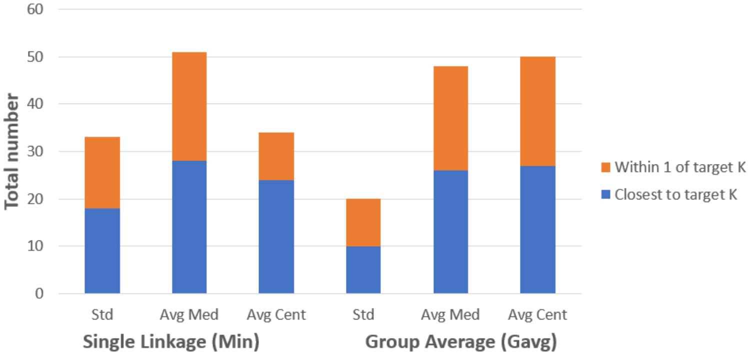
     where $\delta^2_{1..i}$ is the std dev of all F scores on dendogram.

Setup: 2D Spatial Datasets

	# Nodes		Target Cluster Size		
Name	Orig.	Reduced	Orig.	Min	Gavg
a1	3000	800	20	10	**20**
Aggregation	788	788	7	5	**7**
Birch3	10000	800	100	47	**69**
Compound	399	399	6	3	**6**
D31	3100	800	31	19	**31**
Flame	240	240	2	-	**2**
Jain	373	373	2	**2**	**2**
Pathbased	300	300	3	-	**3**
R15	600	600	15	11	**15**
RRR	54	54	3	**3**	**3**
Spiral	312	312	3	**3**	3
t4.8k	8000	800	6	-	**6**
t5.8k	8000	800	6	3	**6**
t7.10k	10000	800	9	-	**9**
t8.8k	8000	800	8	2	**8**
Unbalance	6500	800	8	**7**	**7**

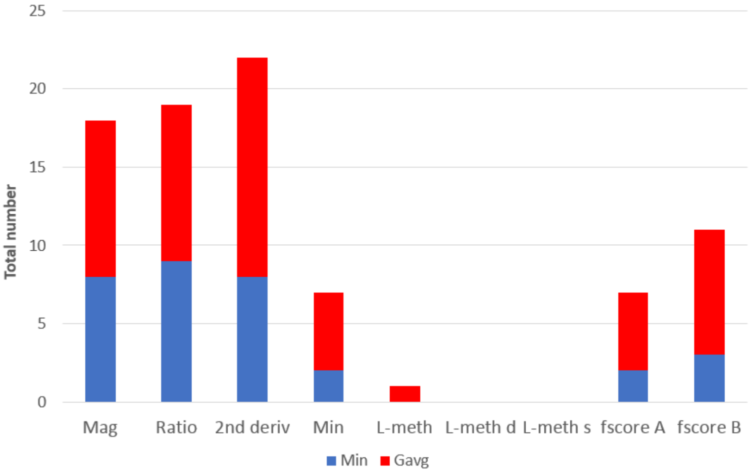
Results

- Knee performance wrt distance metric



Results

- Frequency that knee strategies were closest to target cluster size K



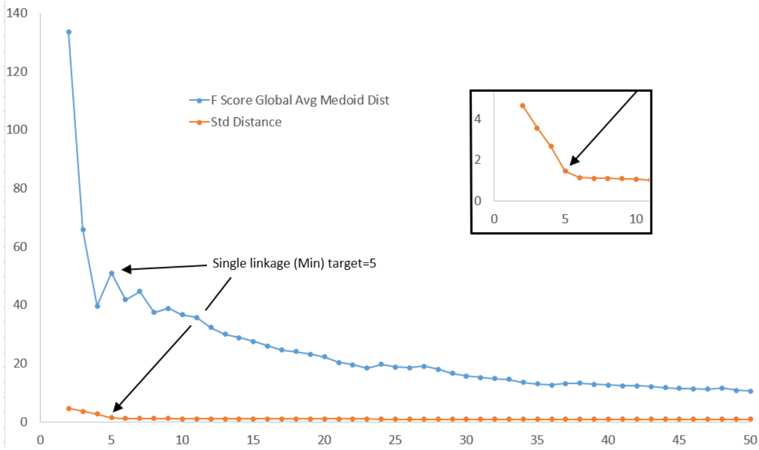
Results

- Details of knee performance: closest to target

	Std		Avg Med		Avg Cent		
Knee	Min	Gavg	Min	Gavg	Min	Gavg	Total
Mag	5	4	3	1	0	4	18
Ratio	5	2	3	3	1	5	19
2nd deriv	5	4	3	5	0	5	22
Min	0	0	1	2	1	3	7
L-meth	0	0	0	1	0	0	1
L-meth D	0	0	0	0	0	0	0
L-meth S	0	0	0	0	0	0	0
F score A	-	-	1	2	1	3	7
F score B	-	-	2	5	1	3	11

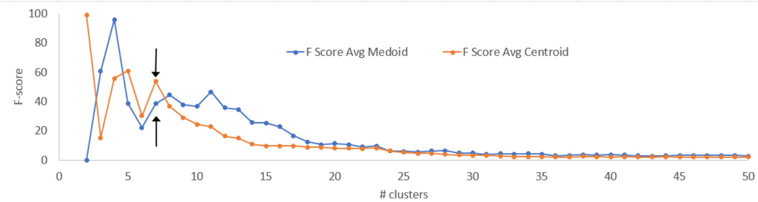
Results

- Comparing knees using different dendogram distances
  (Aggregation DS, single linkage clustering)



Results

- Knee found by F score A and B
  (Aggregation DS, group avg clustering)



- Knee shape is determined by *between group variance* term of ANOVA formula (see tech report).

Conclusion

- Knee detection is a heuristic. It is not guaranteed to work.
- Many factors for success: data set, clustering algorithm, distance measure, knee strategy.
- **Serendipity**.

Future work:

- Could consider more datasets, clustering algorithms, knee strategies. But results will be the same.
- Interesting idea: Use machine learning to discover new knee strategies for different families of datasets.
- Another idea: Use machine learning to identify families of datasets conducive to different clustering optimization strategies.
- Maybe knees are not necessary.