Improving Overlapping Clusters Obtained by a Pyramidal Clustering

Edwin Diday (Université Paris-Dauphine, France)

Francisco de A. T. de Carvalho, Danilo N. Queiroz
(CIn/UFPE-Brazil)

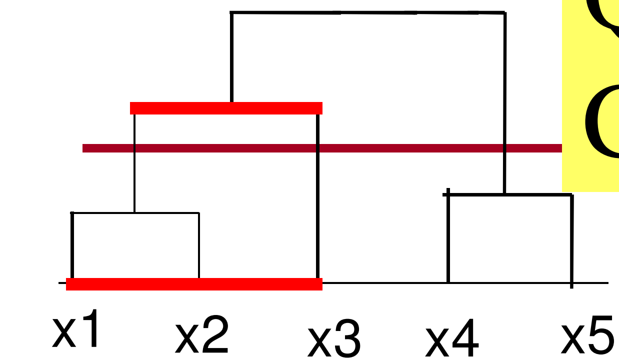
Outline

- Introduction
- Principles of the overlapping clustering algorithm
- Algorithm
- Convergence of the algorithm
- Applications
  - Fats and Oils Interval-Valued Data Set
  - Car Interval-Valued Data Set
- Final Remarks
- References

Introduction

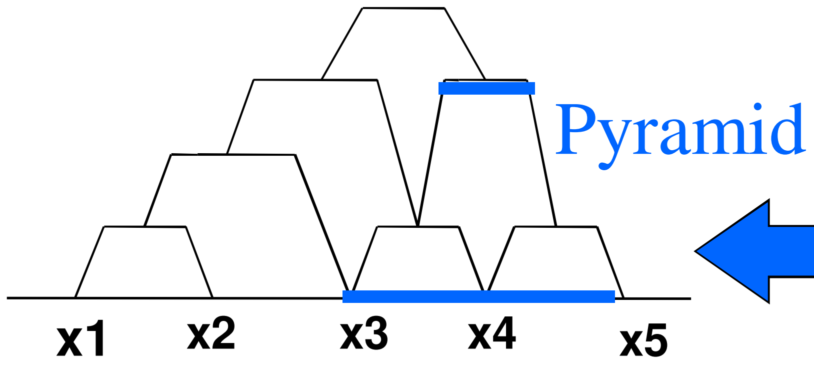
- Pyramidal clustering generalizes standard hierarchical clustering by giving overlapping clusters
- From a hierarchy any induced partition can be improved
- Question:
- How to improve an overlapping clustering induced by a standard or spatial pyramid?

QUALITY CONTROL OF CUTTING



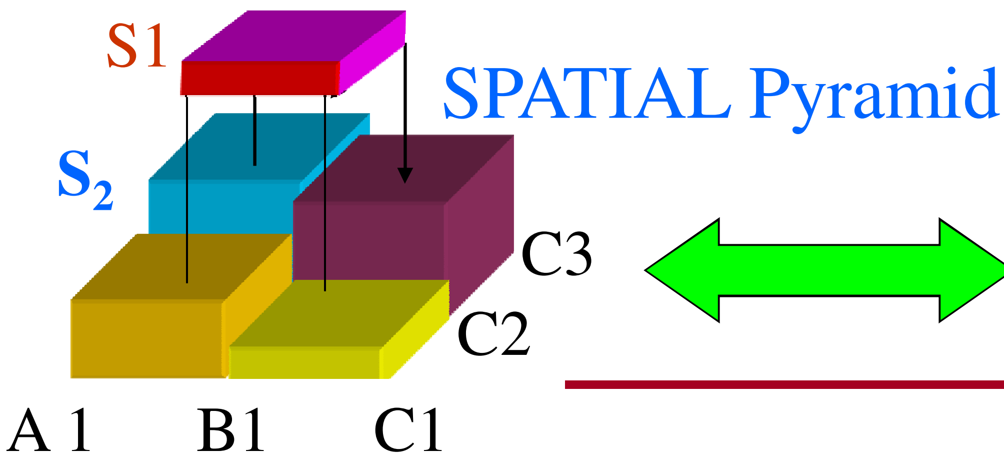
Ultrametric dissimilarity = U

W = |d - U |



Robinsonian dissimilarity = R

W = |d - R |



Yadidean dissimilarity = Y

W = |d - Y |

Principles of the Overlapping Clustering Algorithm-I

- $E = \{e_1, \ldots, e_n\}$: set of observations described by *p* variables $\{y_1, \ldots, y_p\}$;
- $R = (R_1, \ldots, R_K)$ is a covering of the set of observation *E*, i.e.,

$$E = \bigcup_{k=1}^{K} R_k$$

- Aim:
  - Obtain a covering of *E* into *K* overlapping clusters $R_1, \ldots, R_K$
  - Each cluster R_k (k=1,…,K) being represented by a prototype

Principles of the Overlapping Clustering Algorithm-II

- The overlapping clustering algorithm gives:
  - A list of overlapping clusters $R=(R_1, \ldots, R_K)$
  - The corresponding list of prototypes $G=(g_1, \ldots, g_K)$
  - by (locally) optimizing the following adequacy criterion:

$$W(R,G) = \sum_{k=1}^{K} \frac{1}{n_k} \sum_{e_i \in R_k} d(e_i, g_k)$$

  - d is a dissimilarity function; n_k is the cardinal of R_k
- The criterion takes more care to the observations which belong in small classes

Algorithm - I

- Initialization
  - Fixe the number K $(2 \leq K << n)$ of clusters; Set t=0;
  - Let a covering $R^{(0)} = (R_1^{(0)}, \ldots, R_K^{(0)})$ obtained randomly or given by a standard or a spatial pyramid
- Step 1: determination of the prototypes
  - Set t = t + 1; The covering $R^{(t-1)} = (R_1^{(t-1)}, \ldots, R_K^{(t-1)})$ is fixed
  - The prototypes $G^{(t)} = (g_1^{(t)}, \ldots, g_K^{(t)})$ are such that each $g_k^{(t)}$ $(k=1,\ldots,K)$ minimizes the sum of dissimilarities to all the observations of $R_k^{(t)}$

Algorithm -II

- Step 2: the insertion-deletion process is applied to $R^{(t-1)}$ in order to obtain $R^{(t)}$
  - each object *e'* is added to a cluster *C* (of cardinality n_c) belonging to *R* if

$$d(e', g_C) < (1/n_C)\sum_{e \in C} d(e, g_C)$$

  - If *e'* is added to a cluster *C* it can be deleted from its own cluster *C'* if

$$d(e', g_{C'}) > (1/n_{C'})\sum_{e \in C'} d(e, g_{C'})$$

- Stopping criterion: Repeat steps 1 and 2 until the criterion *W* converge

Convergence of the algorithm - I

- *Lemma 1:* A necessary and sufficient condition to obtain

$$I_C = \frac{1}{n_C} \sum_{e \in C} d(e, g_C) > I_{C \cup \{e'\}}$$

with

$$I_{C \cup \{e'\}} = \frac{1}{(n_C + 1)} \left(\sum_{e \in C} d(e, g_C) + d(e', g_C) \right)$$

is that

$$d(e', g_C) < \frac{1}{n_C} \sum_{e \in C} d(e, g_C)$$

Convergence of the algorithm - II

- *Lemma 2:* A necessary and sufficient condition for having

$$I_C = \frac{1}{n_C} \sum_{e \in C} d(e, g_C) > I_{C-\{e'\}}$$

with

$$I_{C-\{e'\}} = \frac{1}{(n_C - 1)} \left(\sum_{e \in C} d(e, g_C) - d(e', g_C) \right)$$

is that

$$d(e', g_C) > \frac{1}{n_C} \sum_{e \in C} d(e, g_C)$$

Convergence of the algorithm - III

- *Lemma 1 and 2* are used to proof the following:

- *Proposition*: the series

$$u_t = W(R^{(t)}, G^{(t)}) = \sum_{k=1}^{K} \frac{1}{n_k^{(t)}} \sum_{e_i \in R_k^{(t)}} d(e_i, g_k^{(t)})$$

decreases at each iteration and converges

Applications

- The covering of an interval-valued data sets will be obtained through a pyramidal clustering algorithm
- SODAS software: http://www.info.fundp.ac.be/asso/
- The overlapping clustering algorithm will start from this covering in order to obtain improved overlapping clusters
- Two interval-valued data sets will be considered
  - Fats and oils interval-valued data set (Ichino and Yaguchi, 1994)
  - Car interval-valued data set

Fats and Oils Data Set - I

- Eight objects
- Four interval-valued variables: *Specific Gravity, Freezing Point, Iodine Value* and *Saponification Value*
- All the variables were considered for clustering purposes
- Module HYPYR of the sodas software was applied on this data set, in order to obtain pyramidal classification
- From this pyramidal structure was obtained a covering of the fats and oils data set into 3 overlapping clusters

Fats and Oils Data Set - II

- Fats and oils data set a priori covering was as follows:
- Cluster 1
  - cotton seed oil; sesame oil; camellia oil; olive oil; beef tallow; hog fat
- Cluster 2
  - linseed oil; perilla oil; cotton seed oil; sesame oil; camellia oil; olive oil
- Cluster 3
  - perilla oil; cotton seed oil; sesame oil; camellia oil; olive oil; hog fat

Fats and Oils Data Set - III

- The overlapping clustering algorithm starts from this a priori covering
- The fats and oils data set final covering is as follows:
  - Cluster 1: beef tallow; hog fat
  - Cluster 2: linseed oil; cotton seed oil; sesame oil; camellia oil; olive oil
  - Cluster 3: perilla oil; cotton seed oil; sesame oil; camellia oil; olive oil
  - The starting and final values of the adequacy criterion W were, respectively, 11437.80 and 8130.63
  - The final covering was improved in comparison with the a priori covering concerning the clustering homogeneity expressed by the adequacy criterion

Car Interval-Valued Data Set - I

- 33 objects
- 8 interval-valued variables: *Price, Engine Capacity, Top Speed, Acceleration, Step, Length, Width* and *Height*
- All the variables were considered for clustering purposes
- Module HYPYR of the sodas software was applied on this data set, in order to obtain pyramidal classification
- From this pyramidal structure was obtained a covering of the fats and oils data set into 4 overlapping clusters

Car Interval-Valued Data Set - II

- Car data set a priori covering was as follows:
  - Cluster 1: Alfa 156/B; Skoda Octavia/B; Audi A3/U; Alfa 145/U; Rover 25/U; Focus/B; Lancia Y/U; Twingo/U; Nissan Micra/U; Skoda Fabia/U; Fiesta/U; Punto/U; Corsa/U
  - Cluster 2: Mercedes SL/S; Mercedes Classe S/L; Audi A8/L; Bmw serie 7/L
  - Cluster 3: Mercedes Classe S/L; Audi A8/L; Bmw serie 7/L; Mercedes Classe E/L; Audi A6/B; Bmw serie 5/L; Lancia K/L; Alfa 166/L; Rover 75/B; Passat/L; Mercedes Classe C/B; Bmw serie 3/B; Vectra/B; Alfa 156/B; Skoda Octavia/B; Audi A3/U; Alfa 145/U;
  - Cluster 4: Lamborghini/S; Aston Martin/S; Ferrari/S; Honda NSK/S; Maserati GT/S; Porsche/S; Mercedes SL/S

Car Interval-Valued Data Set - III

- The overlapping clustering algorithm starts from this a priori covering
- The car data set final covering is as follows:
  - Cluster 1: Punto/U; Corsa/U
  - Cluster 2: Mercedes SL/S
  - Cluster 3: Maserati GT/S; Audi A8/L; Mercedes Classe E/L; Audi A6/B; Bmw serie 5/L; Lancia K/L; Alfa 166/L; Rover 75/B; Passat/L; Mercedes Classe C/B; Bmw serie 3/B; Vectra/B; Alfa 156/B; Skoda Octavia/B; Audi A3/U; Alfa 145/U; Rover 25/U; Focus/B; Lancia Y/U; Twingo/U; Nissan Micra/U; Skoda Fabia/U; Fiesta/U; Punto/U; Corsa/U

Car Interval-Valued Data Set - IV

- The car data set final covering is as follows:
  - Cluster 4: Lamborghini/S; Aston Martin/S; Ferrari/S; Honda NSK/S; Porsche/S; Mercedes SL/S; Mercedes Classe S/L; Bmw serie 7/L
- The starting and final values of the adequacy criterion W were, respectively, 3.94383e + 010 and 2.40101e + 010
- The final covering was improved in comparison with the a priori covering concerning the clustering homogeneity expressed by the adequacy criterion

Final Remarks

- We give an overlapping clustering algorithm which is an extension of the K-means algorithm
- The aim is to improve overlapping clusters given by a pyramidal clustering algorithm
- The principe of the algorithm and the proof of its convergence are given
- Applications concerning interval-valued data sets showed the usefulness of this overlapping clustering

Bibliography

[1] BERTRAND, P. and JANOWITZ, M.F. (2002): Pyramids and weak hierarchies in the ordinal model for clustering. Discrete Applied Mathematics, 122(1-3), 55-81.

[2] BRITO, P. (1994): Order structure of symbolic assertion objects. IEEE Transactions on Knowledge and Data Engineering, 6 (5), 830-85.

[3] E. Diday and M. Noirhome, Symbolic Data Analysis and the SODAS Software, Wiley, 2008

[4] CLEUZIOU, G. (2008): An extended version of the k-means method for overlapping clustering. In: Proceedings of the Nineteenth International Conference on Pattern Recognition (ICPR 2008): 1-4.

[5] DIDAY, E. (2008): Spatial classication. Discrete Applied Mathematics, 156 (8), 1271-1294

[6] JOHNSON, S.C. (1967): Hierarchical clustering schemes. Psychometrika, 32, 241- 254.

Thank you