On Monotonic Tendency of Some Fuzzy Cluster Validity Indices for High-Dimensional Data

Fernanda Eustáquio

Tatiane Nogueira

October, 25 2018

Federal University of Bahia
Computer Science Department

Contents

Introduction

Introduction i

- High-dimensional data sets can have hundreds or thousands of dimensions:
  - E.g.: satellite image, text document and microarray.
- This type of data is commonly used in clustering techniques [12, 4, 13, 9, 17] because they can extract a structure of a data **without a previous information**;
- Clustering data sets with high-dimensional feature space is still a challenging problem [16];
  - **Hard** to analyze;
  - **Unfeasible** to **visualize** without a reduction of dimensions as feature transformation and feature selection techniques;

Fuzzy Clustering i

The introduction of fuzzy set theory into clustering captures the **uncertainty** and **imprecision** inherent to any data.

To believe that the boundaries between clusters of high-dimensional data are **ambiguous** and **not well-separated**, fuzzy clustering and specifically the Fuzzy c-Means (FCM) [7, 3] algorithm is suitable to clustering high-dimensional data.

FCM

- It is one of the most widely used fuzzy clustering models;
- Pseudo-partitioning algorithm;
- Assigns memberships degrees to an object in each cluster.

Cluster Validation

After a clustering algorithm extracts the structure of a data, its partition has to be validated for a cluster validity index (CVI).

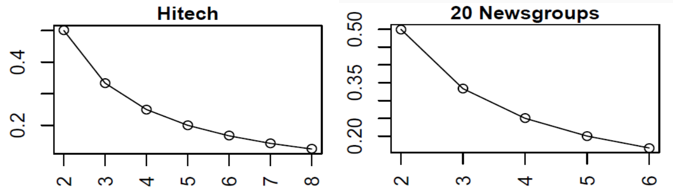
In most real applications, the number of clusters c is **unknown** and, for that, CVIs are used with the **relative evaluation criterion** to find the optimal value of c.

CVIs results must be **independently of any parameter** of a clustering algorithm:

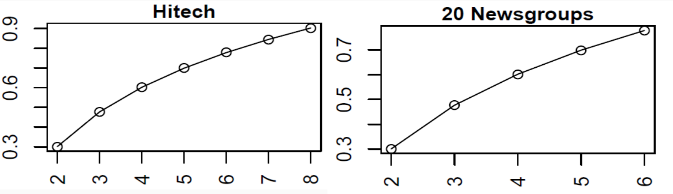
- Number of clusters c;
- Fuzzification factor m.

Monotonic tendency face to the number of clusters c i

$\uparrow c \downarrow PC$ (maximum) $\uparrow c \downarrow FS$ (minimum)



$\uparrow c \uparrow PE$ (minimum)



Monotonic tendency face to the number of clusters c ii

We made a hard exploratory investigation about:

- the monotonic tendency of the most common fuzzy CVIs (PC, PE, FS);
- the indices that were proposed to correct or reduce the monotonic tendency (MPC, IPC, NPE);

in a high-dimensional context.

What can be the major contribution in studying the monotonic tendency of fuzzy CVIs?

Monotonic tendency face to the number of clusters c iv

What would you think about an index that had these results?

Data set	c	$m = 1.5$	$m = 1.8$	$m = 2.0$	$m = 2.2$	$m = 2.5$
Data set 1	5	5	5	5	5	5
Data set 2	3	3	3	3	3	3
Data set 3	4	4	4	4	4	4
Data set 4	4	4	4	4	4	4
Data set 5	4	2	2	2	2	2
Data set 6	6	6	6	6	6	6
Data set 7	4	4	4	4	4	4

This index is amazing and Data set 5 should have some structure problem!

Monotonic tendency face to the number of clusters c v

In fact this is the **Fukuyama-Sugeno (FS)** index that validates FCM pseudo-partitions with $2 \leq c \leq C_{classes}$;

The **maximum** number of clusters was set equal to the number of clusters **expected**.

$\uparrow c \downarrow FS$ (Optimal c is found for the minimum value of FS)

Fuzzy Cluster Validity Indices

Notations

Notation	Means
n	number of objects to be clustered
c	number of clusters
x_k	one object to be clustered, $1 \leq k \leq n$
$A_i(x_k)$	membership degree of x_k in cluster i, $1 \leq i \leq c$
U	pseudo-partition fuzzy defined as $U = [A_i(x_k)]$
v_i	cluster center of cluster i
$\bar{v}$	$\sum_{i=1}^{c} v_i / c$ is the mean of the c cluster centers
m	fuzzification factor ($1 < m < \infty$) that determines the degree of fuzziness of membership degrees on clusters

Fuzzy CVIs Definitions

Optimal $c \to C_{min}$

Index	Definition	Optimal value
PC [2]	$\frac{1}{n}\sum_{i=1}^{c}\sum_{k=1}^{n} A_i^2(x_k)$	Maximum
MPC [6]	$1 - \frac{c}{c-1}(1 - PC)$	Maximum
IPC [15]	$100\left[\frac{PC(c-1)-PC(c)}{PC(c-1)} - \frac{PC(c)-PC(c+1)}{PC(c)}\right]$	Maximum
PE [1]	$-\frac{1}{n}\sum_{i=1}^{c}\sum_{k=1}^{n} A_i(x_k)\log_a(A_i(x_k))$	Minimum
NPE [8]	$\frac{nPE}{n-c}$	Minimum

Fuzzy CVIs Definitions

Optimal $c \to C_{max}$

Index	Definition	Optimal value
FS [10]	$\sum_{i=1}^{c}\sum_{k=1}^{n} A_i^m(x_k)(\lVert x_k - v_i\rVert^2 - \lVert v_i - \bar{v}\rVert^2)$	Minimum
MPO [11]	$\left(\frac{c+1}{c-1}\right)^{1/2} \frac{\sum_{k=1}^{n}\sum_{i=1}^{c} {A_i}^2(x_k)}{min_{1\leq i\leq c}\left\{\sum_{k=1}^{n} {A_i}^2(x_k)\right\}} - \frac{1}{n}\sum_{k=1}^{n}\left(\sum_{i=1}^{c-1}\sum_{j=i+1}^{c} O_{ijk}(c,U)^1\right)$	Maximum

[1] $O_{ijk}(c,U) = \begin{cases} 1 - \left|A_i(x_k) - A_j(x_k)\right| & \text{if} \quad A_{ijk} \geq T_o, i \neq j \\ 0 & \text{if} \quad otherwise \end{cases}$.

Experiments

Data sets i

Data set	p	n	$C_{classes}$	Domain
Sorlie	456	85	5	Breast Cancer
Christensen	1,413	217	3	N/A
CSTR	1,725	299	4	Scientific
Alon	2,000	62	2	Colon Cancer
Khan	2,308	63	4	SRBCT[2]
Su	5,565	102	4	N/A
Hitech	6,593	600	6	News articles
West	7,129	49	2	Breast Cancer
Subramanian	10,100	50	2	N/A
20Newsgroups	11,026	2,000	4	E-mails

[2] Small Round Blue Cell Tumors

Fuzzy c-Means Parameters

- Stop criterion: $E = 10^{-5}$;
- Fuzzification factor (weighting exponent):
  - Default value: $m = 2.0$;
  - $m = \{1.5, 1.8, 2.0, 2.2, 2.5\}$
- Number of clusters $C_{min} \leq c \leq C_{max}$:
  - Minimum number of clusters $C_{min} = 2$;
  - Maximum number of clusters C_{max}.

Maximum number of clusters i

- Three values of C_{max} to evaluate the indices that could have a monotonic tendency in selecting $c = C_{max}$:
  - $C_{max} = C_{classes}$: can make a biased index be erroneously well-evaluated;
  - $C_{max} = C_{classes} + C_{min}$: it is a definition to avoid the previous situation;
  - $C_{max} = 10$: common value, greater than all $C_{classes}$ of the data sets.

Dissimilarity measure

Dissimilarity between an object x_k to a cluster center v_i, $||x_k - v_i||$, was measured by Cosine

$$sim(x_k, v_i) = cos\theta = \frac{x_k \cdot v_i}{|x_k|\,|v_i|} \tag{1}$$

$$1 - sim(x_k, v_i)$$

Results

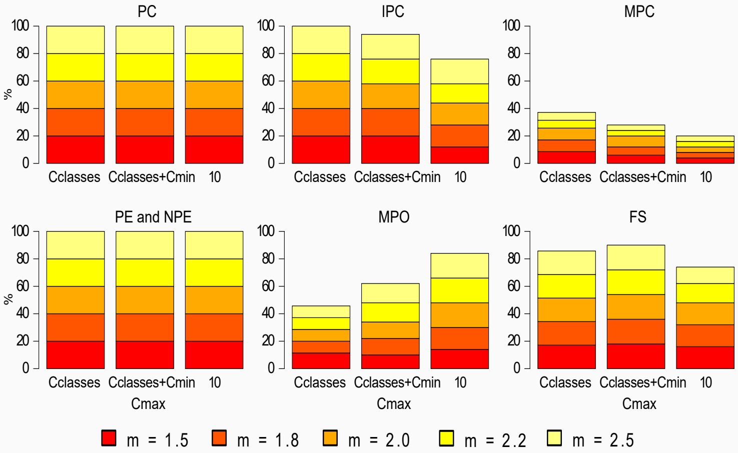
Results i

- PC, PE and NPE indices invariantly recognized $c = C_{min} = 2$;
- The FS index showed a monotonic tendency face to c, as described in [5]
  - For $C_{max} = \{C_{classes}, C_{classes} + C_{min}\}$: FS chose $c = C_{max}$ for all FCM partitions with exception of Su data set;

Data set	$m = 1.5$	$m = 1.8$	$m = 2.0$	$m = 2.2$	$m = 2.5$
Sorlie	10	10	10	10	9
Christensen	8	9	9	7	9
CSTR	10	10	10	10	10
Alon	10	10	10	10	10
Khan	10	10	10	10	8
Su	2	2	2	2	2
Hitech	10	10	10	10	10
West	10	10	10	8	10
Subramanian	10	10	10	10	10
20Newsgroups	10	10	10	10	10

Results ii

Percent of data sets, discriminate by the m values, in which each index selected C_{min} or C_{max} as the optimal number of clusters



Discussion

Discussion i

The PC and PE monotonic tendency can be explained by their limits when $m \to \infty$.

PC $\in [1/c; 1]$

$$\lim_{m \to \infty} PC = \frac{1}{c} \tag{2}$$

- $c = 2$ will maximize the value of $1/c$.

PE $[0, \log_a c]$

$$\lim_{m \to \infty} PE = \log_a c \tag{3}$$

- $c = 2$ will minimize the value of $\log_a c$.
- $m = 1.5$ is big enough to PC and PE select $c = 2$.

Discussion ii

- The PC and PE monotonic tendency appears to be more common when validating real data sets;
- In some researches:
  - All data sets have low dimensionality;
  - Some of the synthetic data selected $c \neq 2$;
  - For all real data PC and PE indicated $c = 2$.
- Real world data sets must contain some noise points that influence PC to select $c = 2$ (WU, 2008).

Discussion iii

NPE had values very close to the calculated by PE;

$$NPE = \frac{nPE}{n - c} \tag{4}$$

- $n - c$ is approximately equal to n;
- The number of clusters c is much smaller than the number of objects n for all data sets.

MPO was dependent of c for high-dimensional data

- Its term $(\frac{c+1}{c-1})^{1/2}$ used to avoid the monotonic tendency for the number of clusters did not work as expected.

Discussion iv

FS

- Same tendency of PC;
- It is sensitive to the fuzzification factor m and, because of that, it may be **unreliable** [14];
- FS results are **unexpected** and **unreasonable** because its tendency [5, 18].

FS tendency i

- FS tendency appears to be more common when validating high-dimensional data sets;
- In some researches the dimensionality of data sets is not greater than 296:
  - In (VALENTE et al., 2013; HUAPENG et al., 2016), FS selected $c \neq C_{max}$ for all data sets;
  - In the remaining researches, FS selected $c = C_{max}$ in the maximum 43% of FS validations;
  - In this work, FS selected $c = C_{max}$ in at least 83%.

FS tendency ii

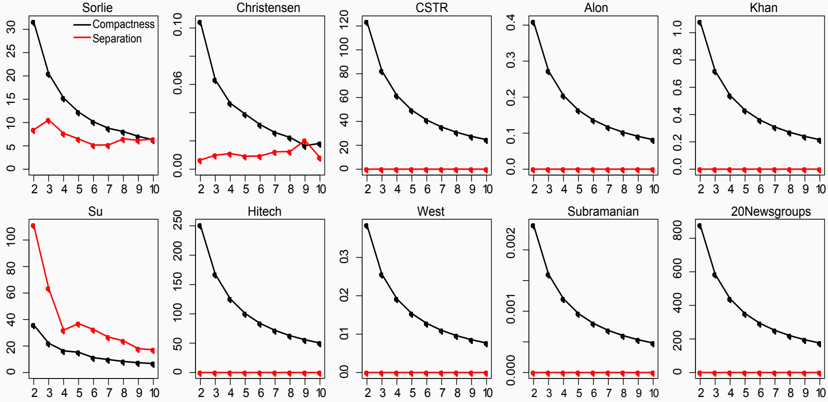
FS index works well when the clusters are **highly separable** [5]

$$FS = \sum_{i=1}^{c} \sum_{k=1}^{n} A_i^m(x_k)(\|x_k - v_i\|^2 - \|v_i - \bar{v}\|^2) \tag{5}$$

- The FS separation term is always significantly small and this can lead to the index favoring more clusters than the desired [5].

FS tendency iii

Compactness and separation of the FS index are shown for $m = 2.0$.



- FS worked well for Su that is a highly separable data set.

Conclusion

Conclusion

- The linear transformation made in PC by MPC works well in reducing the PC tendency until 80% for $C_{max} = 10$;
- MPC results were independent of any FCM parameter;
- FS may falsely appear to recognize the correct number of clusters when $C_{max} = C_{classes}$;
- C_{max} should be greater than $C_{classes}$.

Future work

To verify if other indices have some tendency in function of parameters of FCM or any other clustering algorithm that could disturb the clustering validation of high-dimensional data.

How to choose an appropriate index to validate fuzzy clustering of high-dimensional data?

Start eliminating those which have some tendency in function of any parameter is already a first big step!

Acknowledgment

I would like to thank the Coordination for the Improvement of Higher Education Personnel (CAPES) for granting the scholarship during the period of this master's degree at Federal University of Bahia.

The End

Thank you!

fernanda.eustaquio@ufba.br

tatiane.nogueira@ufba.br

J. C. Bezdek.
Cluster validity with fuzzy sets.
Journal of Cybernetics, 3(3):58–73, 1974.

J. C. Bezdek.
Numerical taxonomy with fuzzy sets.
Journal of Mathematical Biology, 1(1):57–71, 1974.

J. C. Bezdek.
Pattern Recognition with Fuzzy Objective Function Algorithms.
Kluwer Academic Publishers, Norwell, MA, USA, 1981.

N. V. Carvalho, S. O. Rezende, H. A. Camargo, and T. M. Nogueira.
Flexible document organization by mixing fuzzy and possibilistic clustering algorithms.
In *IEEE International Conference on Fuzzy Systems*, pages 790–797, 2016.

A. Chong, T. D. Gedeon, and L. T. Koczy.
A hybrid approach for solving the cluster validity problem.
In *International Conference on Digital Signal Processing Proceedings*, volume 2, pages 1207–1210, 2002.

R. N. Dave.
Validating fuzzy partitions obtained through c-shells clustering.

Pattern Recognition Letter, 17(6):613–623, 1996.

J. C. Dunn.
A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters.
Journal of Cybernetics, 3(3):32–57, 1973.

J. C. Dunn.
Indices of partition fuzziness and the detection of clusters in large data sets.
In *Fuzzy Automata and Decision Processes*, pages 271–284. Elsevier, New York, 1977.

References iv

F. Eustáquio, H. Camargo, S. Rezende, and T. Nogueira.
On fuzzy cluster validity indexes for high dimensional feature space.
In *Advances in Fuzzy Logic and Technology 2017: Proceedings of The 10th Conference of the European Society for Fuzzy Logic and Technology, 2017, Warsaw, Poland, Volume 2*, pages 12–23, 2018.

Y. Fukuyama and M. Sugeno.
A new method of choosing the number of clusters for fuzzy c-means method.
In *Fuzzy systems Symposium*, pages 247–250, 1989.

Y. Hu, C. Zuo, Y. Yang, and F. Qu.
A robust cluster validity index for fuzzy c-means clustering.
In *International Conference on Transportation, Mechanical, and Electrical Engineering*, pages 448–451, 2011.

References v

T. M. Nogueira.
Organização flexível de documentos.
PhD thesis, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2013.

T. M. Nogueira, S. O. Rezende, and H. A. Camargo.
Flexible document organization: Comparing fuzzy and possibilistic approaches.
In *2015 IEEE International Conference on Fuzzy Systems*, pages 1–8, 2015.

N. R. Pal and J. C. Bezdek.
On cluster validity for the fuzzy c-means model.
IEEE Transactions on Fuzzy Systems, 3(3):370–379, 1995.

C. sheng Li.
The improved partition coefficient.
International Conference on Advances in Engineering, 24:534–538, 2011.

M. Steinbach, L. Ertöz, and V. Kumar.
The challenges of clustering high dimensional data.
New Directions in Statistical Physics: Econophysics, Bioinformatics, and Pattern Recognition, pages 273–309, 2004.

R. Subhashini and V. J. S. Kumar.
Evaluating the performance of similarity measures used in document clustering and information retrieval.
In *International Conference on Integrated Intelligent Computing*, pages 27–31, 2010.

M.-S. Yang and K.-L. Wu.
A new validity index for fuzzy clustering.
In *IEEE International Conference on Fuzzy Systems*, volume 1, pages 89–92, 2001.