http://bipolar.ibspan.waw.pl
https://www.kamilkmita.com

The non-linear impact of the scaling factor α on the outcomes of Semi-Supervised Fuzzy C-Means

Kamil Kmita Katarzyna Kaczmarek-Majer Olgierd Hryniewicz

Systems Research Institute Polish Academy of Sciences, Warsaw, Poland

International Symposium on Fuzzy Sets
19 May 2023, Rzeszów, Poland

Fuzzy clustering - finding good $c-$partitions

BIPOLAR

Clustering: partitioning data set X into c clusters that contain observations **similar** to each other and dissimilar to the rest of the data,

Fuzzy clustering: uses a soft assignment of each observation to each cluster (**a membership degree** u_{jk}) that is grounded in fuzzy set theory.

Fuzzy c-partition space[1]

Let X be any finite set, c a number of clusters $2 \leq c < N$, W_{Nc} a set of real matrices of $N \times c$ dimension. Then a **fuzzy** $c-$**partition space** for X is the set

$$M_{fc} = \left\{ U \in W_{Nc} \mid u_{jk} \in [0,1]; \quad \sum_{k=1}^{c} u_{jk} = 1 \, \forall j; \quad 0 < \sum_{j=1}^{N} u_{jk} < n \, \forall k \right\} \tag{1}$$

[1] James C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms.* Springer US, Boston, MA, 1981

An illustrative example of a fuzzy 2-partition

$X = \{x_1, x_2, x_3\}$, $x_j \in R^p$.

$j = 1, \ldots, 3;\ N = 3.$

$k \in \{1, 2\};\ c = 2.$

A possible fuzzy 2−partition:

$$U = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \end{array} \begin{array}{c} \begin{array}{cc} k=1 & k=2 \end{array} \\ \begin{pmatrix} 0.98 & 0.02 \\ 0.6 & 0.4 \\ 0.06 & 0.94 \end{pmatrix} \end{array}$$

Observation x_1 belongs strongly to cluster 1, observation x_3 belongs strongly to cluster 2, while observation x_2 seems to be a "hybrid": it belongs to both clusters to similar degree.

Struggling with imagining a "hybrid"?

A classical example from [Bez81]:

- x_1: a `peach`,
- x_3: a `plum`,
- x_2: a `nectarine`, **supposedly** a hybrid of a `peach` and a `plum`.

Supposedly...

Struggling with imagining a "hybrid"?

BIPOLAR

A classical example from [Bez81]:

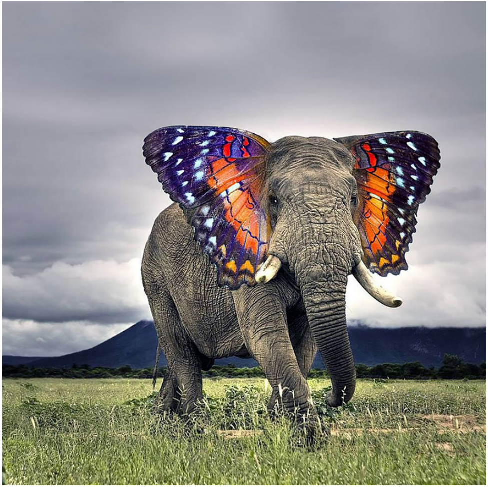
- x_1: a `peach`,
- x_3: a `plum`,
- x_2: a `nectarine`, **supposedly** a hybrid of a `peach` and a `plum`.

Supposedly... because it turns out to be a controversial topic, e.g. `http://www.bctreefruits.com/fruits/other-fruits/detail/0/Nectarines/` state "There is some misconception that nectarines are a cross between a peach and a plum, **but this is not the case. They're simply a fuzzless peach**."

Unreal, but proper hybrid

- x_1: a butterfly,
- x_3: an elephant,
- x_2: a butterphant

Unreal, but proper hybrid



- x_1: a butterfly,
- x_3: an elephant,
- x_2: a butterphant

Figure: A butterphant. Source: https://www.boredpanda.com/animals-hybrids-photoshop/?media_id=321587

Semi-supervised fuzzy clustering

BIPOLAR

- Semi-Supervised Learning (SSL)[2]: labels $y_j \in Y$ are available for a part of observations M out of all N observations ($M < N$),
- an arbitrary 1-1 mapping must be established between clusters (columns of U) and classes (columns of F).

$$U = \begin{array}{c c} & \begin{array}{cc} k=1 & k=2 \end{array} \\ \begin{array}{c} x_1 \\ x_2 \\ x_3 \end{array} & \begin{bmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \\ u_{31} & u_{32} \end{bmatrix} \end{array} \qquad F = \begin{array}{c c c} & \begin{array}{cc} k=1 & k=2 \end{array} & \text{s(i)} \\ \begin{array}{c} x_1 \\ x_2 \\ x_3 \end{array} & \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix} & \begin{array}{c} s(1)=1 \\ \\ s(3)=2 \end{array} \end{array}$$

Function $s(i)$ retrieves the index of the class (a column in F) associated with i-th supervised observation.

[2] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien, editors. *Semi-Supervised Learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, Mass, 2006

Semi-Supervised Fuzzy C-Means (SSFCMeans) model

BIPOLAR

Objective function J based on [PW97][3] introducing *partial supervision*

$$J = \sum_{k=1}^{c}\sum_{j=1}^{N} u_{jk}^2 \cdot d^2(x_j, v_k) + \alpha \sum_{k=1}^{c}\sum_{j=1}^{N} \underbrace{(u_{jk} - b_j f_{jk})^2}_{\text{penalization}} \cdot d^2(x_j, v_k).$$

- $u_{jk} \in [0,1]$ is a membership degree
- $d_{jk} = d(x_j, v_k)$ is a Euclidean distance between jth observation and kth prototype v_k (k-th cluster is associated with its prototype $v_k \in R^p$),

---

- $F = [f_{jk}]$ is a matrix introducing partial supervision with binary entries $f_{jk} \in \{0,1\}$,
- $b_j \in \{0,1\}$ is an indicator variable equal to 1 iff x_j is labeled,
- $\alpha \geq 0$ **is a scaling factor that weighs the strength of partial supervision**.

[3] W. Pedrycz and J. Waletzky. Fuzzy clustering with partial supervision. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 27(5):787–795, October 1997

Semi-Supervised Fuzzy C-Means

Finding optimal c-partitions

Notation:

- $X = [x_j]$, $x_j \in R^p$
- $U \in M_{fc}$: a memberships matrix,
- $V \in W_{cp}$: a prototypes matrix ($V = [v_k]$),
- Θ: a set of hyper-parameters.

Task:

$$(U^\star, V^\star) = \underset{U,V}{\arg\min} \quad J(U, V; X, \Theta), \tag{2}$$

where objective function J quantifies a notion of similarity between observations and prototypes (*typically, using a distance function such as e.g. Euclidean distance*).

Optimal $\hat{U}$

An iterative optimization algorithm is frequently performed. Optimal $\hat{U} = [\hat{u}_{jk}]$ matrix is obtained by considering first-order necessary conditions of a global minimizer, leading to

$$\hat{u}_{jk} = \frac{1}{1+\alpha} \cdot \left(\frac{1 + \alpha \cdot \left(1 - b_j \sum_{s=1}^{c} f_{js}\right)}{\sum_{s=1}^{c} \left(d_{jk}^2 / d_{js}^2\right)} + \alpha f_{jk} b_j \right). \tag{3}$$

In a case of a supervised observation i and its membership degree to the supervised cluster $s(i)$

$$\hat{u}_{i,s(i)} = \frac{1}{1+\alpha} \cdot \frac{1}{\sum_{s=1}^{c} \left(d_{ik}^2 / d_{is}^2\right)} + \frac{\alpha}{1+\alpha}. \tag{4}$$

Interpretations of the scaling factor α

objective function	$\sum_{k=1}^{c}\sum_{j=1}^{N} u_{jk}^2 d_{jk}^2 + \boldsymbol{\alpha} \sum_{k=1}^{c}\sum_{j=1}^{N} \underbrace{(u_{jk} - b_j f_{jk})^2}_{\text{penalization}} d_{jk}^2.$
optimal membership $\hat{u}_{i,s(i)}$	$\frac{1}{1+\alpha} \cdot \frac{1}{\sum_{s=1}^{c} \left(d_{ik}^2 / d_{is}^2\right)} + \frac{\alpha}{\underbrace{1+\alpha}_{\text{ALB}}}$

- [PW97, p. 788] "a scaling factor whose role is **to maintain a balance** between the supervised and unsupervised component",
- "The scaling factor α quantifies the impact of partial supervision as $\mathrm{IPS}(\alpha) = \frac{\alpha}{1+\alpha}$, and establishes an Absolute Lower Bound for a membership of a supervised observation to the supervised cluster $u_{i,s(i)} > \mathrm{IPS}(\alpha)$"[4].

[4] K. Kmita, K. Kaczmarek-Majer, O. Hryniewicz, Explainable Impact of Partial Supervision in Semi-Supervised Fuzzy Clustering, *manuscript under review*

The functional form of IPS

What if we are unhappy with the functional form $\text{IPS}(\alpha) = \frac{\alpha}{1+\alpha}$?

The form of IPS function is a result of[5]:

- the iterative optimization algorithm,
- functional form of the objective function J,
- the Langrage multipliers technique,
- the constraint $\sum_{k=1}^{c} u_{jk} = 1 \; \forall j$.

[5] K. Kmita, K. Kaczmarek-Majer, O. Hryniewicz, Explainable Impact of Partial Supervision in Semi-Supervised Fuzzy Clustering, *manuscript under review*

The functional form of IPS

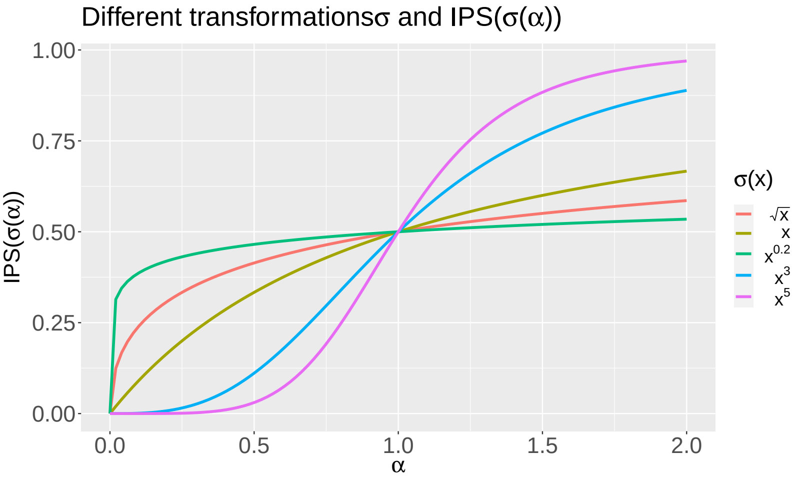
What if we are unhappy with the functional form $\mathrm{IPS}(\alpha) = \frac{\alpha}{1+\alpha}$?

The form of IPS function is a result of[5]:

- the iterative optimization algorithm,
- functional form of the objective function J,
- the Langrage multipliers technique,
- the constraint $\sum_{k=1}^{c} u_{jk} = 1 \; \forall j$.

But we could simply consider transformations $\sigma(\alpha)$, sustaining all of the above!

[5] K. Kmita, K. Kaczmarek-Majer, O. Hryniewicz, Explainable Impact of Partial Supervision in Semi-Supervised Fuzzy Clustering, *manuscript under review*


Different transformationsσ and IPS(σ(α))
IPS(σ(α))
1.00
0.75
0.50
0.25
0.00
0.0
0.5
1.0
1.5
2.0
α
σ(x)
√x
x
x^0.2
x^3
x^5

Experiments on real-life data

BIPOLAR

Data for this work were collected from patients diagnosed with bipolar disorder within a prospective observational study[6] carried out by the Institute for Psychiatry and Neurology and Systems Research Institute, Polish Academy of Sciences in Warsaw, Poland in years 2017-2018.

- $N = 1295$ summaries of phone calls (indexed by $j = 1, \ldots, N$),
- each phone call's summary $x_j \in R^5$. The 5 selected variables include physical descriptors of speech (e.g. jitter),
- $M = 261$ phone calls are treated as supervised (indexed by $i = 1, \ldots, M$),
- $Y = \{\texttt{depression}, \texttt{mixed}, \texttt{euthymia}, \texttt{dysfunction}\}$,
-

class	depression	mixed	euthymia	dysfunction
#	58	55	85	63

[6] The study obtained the consent of the Bioethical Commission at the District Medical Chamber in Warsaw (agreement no. KB/1094/17)

Results of SSFCMeans models by α - single observation

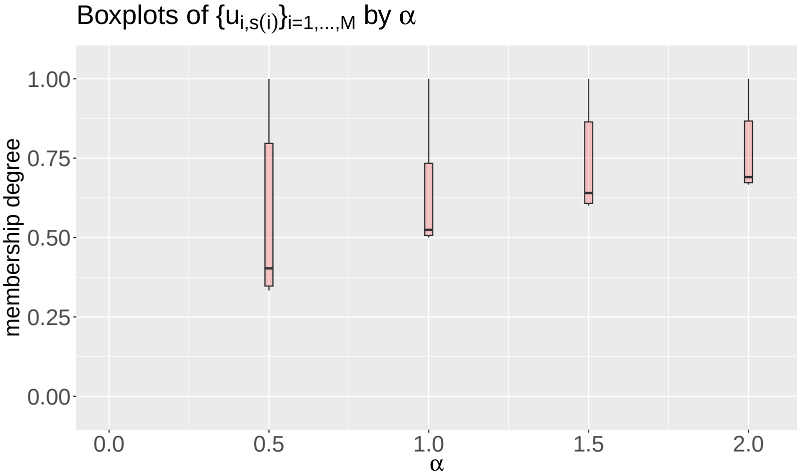
BIPOLAR

Results for a given observation $x_{i=3}$ for 4 SSFCMeans(α) models, $\alpha \in \{0.5, 1., 1.5, 2.\}$.

i	alpha	IPS	y_i	depression	mixed	euthymia	dysfunction
3	0.50	0.33	depression	0.95	0.01	0.03	0.01
3	1.00	0.50	depression	0.90	0.05	0.02	0.03
3	1.50	0.60	depression	0.90	0.03	0.05	0.02
3	2.00	0.67	depression	0.91	0.05	0.03	0.02

The blue color marks $u_{i,s(i)}$: a membership of a supervised observation $i = 3$ to the supervised cluster $s(i) = 1$ (a **supervised membership**).

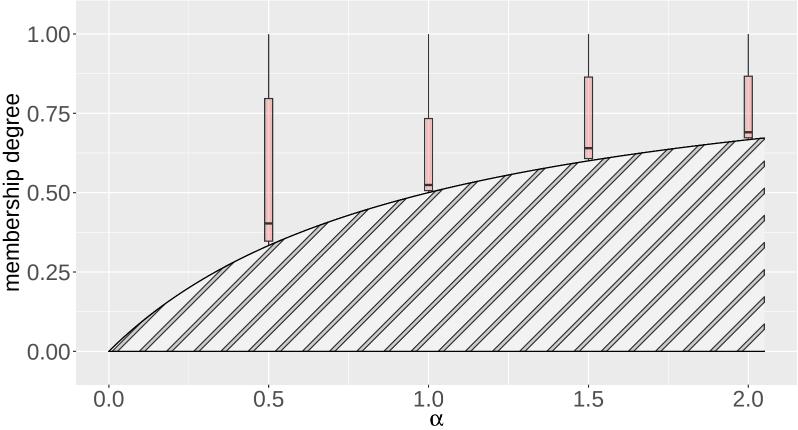
Results of SSFCMeans models by α - a summary



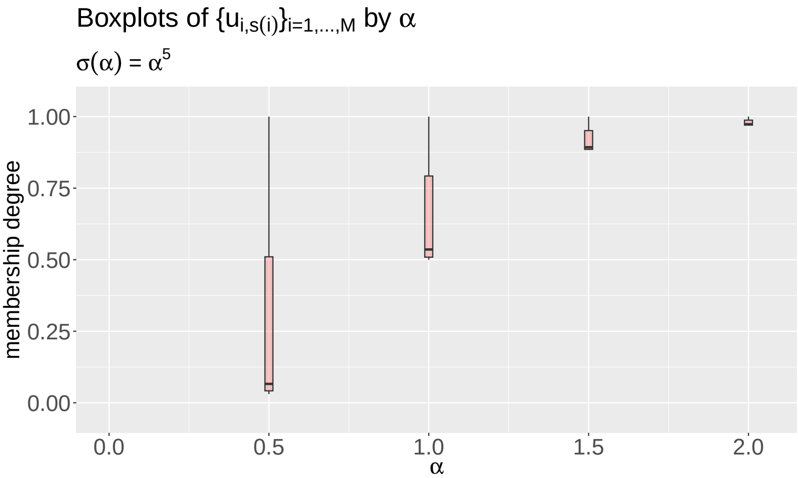
Results of SSFCMeans models by α - a summary

BIPOLAR

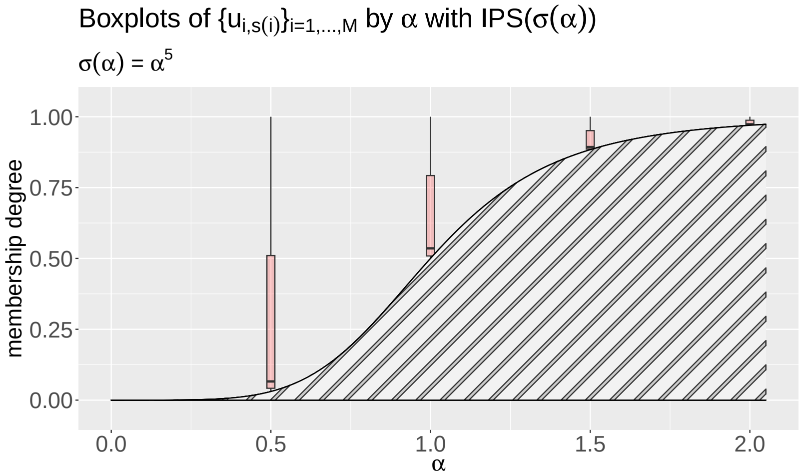
Boxplots of $\{u_{i,s(i)}\}_{i=1,\ldots,M}$ by α with IPS(α)



Results for SSFCMeans models by α: $\sigma(\alpha) = \alpha^5$



Results for SSFCMeans models by α: $\sigma(\alpha) = \alpha^5$



Conclusions

- We have shown the differences in interpretation of the scaling factor α in SSCMeans;
- the impact of α on the estimated $\hat{u}_{i,s(i)}$ is non-linear and scales as $\mathrm{IPS}(\alpha) = \frac{\alpha}{1+\alpha}$;
- it is hard to change the functional form of IPS, but one can adjust it by considering transformations $\sigma(\alpha)$.

Why it matters?

1. explainability, using $\hat{u}_{i,s(i)}$ in advanced procedures building on SSFCMeans,
2. the way α enters the objective function matters for the optimal prototypes as well (further research directions).

Thank you for your attention!

kmita@ibspan.waw.pl
https://kamilkmita.com
http://bipolar.ibspan.waw.pl
https://github.com/ITPsychiatry/bipolar

Kamil Kmita and Katarzyna Kaczmarek-Majer received funding from Small Grants Scheme (NOR/SGS/BIPOLAR/0239/2020-00) within the research project: ”Bipolar disorder prediction with sensor-based semi-supervised Learning (BIPOLAR)”.

Bibliography I

[Bez81] James C. Bezdek.
Pattern Recognition with Fuzzy Objective Function Algorithms.
Springer US, Boston, MA, 1981.

[CSZ06] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien, editors.
Semi-Supervised Learning.
Adaptive Computation and Machine Learning. MIT Press, Cambridge, Mass, 2006.

[PW97] W. Pedrycz and J. Waletzky.
Fuzzy clustering with partial supervision.
IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 27(5):787–795, October 1997.