TFG:
Fuzzy C-Means and
clustering algorithm
a comparatrive stud

Intel·ligència Artificial

Victor Garcia Domingo

vgarciadomi@uoc.edu

Consultor: Joan M. Nuñez Do Rio

5 de juny de 2019

Universitat Oberta de Catalunya

Introduction (1/2)

- To find clustering in a dataset
  - Grouping similar observations
  - Distance function to group
- Two criteria to classify algorithms
  - Hierarchical, non hierarchical, mixture
  - Hard clustering and soft clustering
- Hard clustering
  - One observation belongs to one cluster
- Soft clustering
  - One observation can belong to more than one cluster

Universitat Oberta de Catalunya

Introduction (2/2)

- Brief history of clustering:
  - K-Means (KM)
    - Steinhaus (1956) based on the sum-of-squares criterion.
    - Well-separated clusters.
  - Fuzzy C-Means (FCM)
    - Bezdek (1973) based on KM.
    - Overlapping clusters.
  - Gustafson Kessel Fuzzy C-Means (GKFCM)
    - Gustafson and Kessel (1978) based on FCM.
    - Non-spherical clusters.
  - Possibilistic C-Means
    - Krishnapuram and Keller (1993)
    - Outliers and noise.
  - Suppressed-Fuzzy C-Means
    - Fan, Zhen and Xie (2003) based on FCM.
    - Computational efficiency.
  - Fuzzy C-Means++
    - Stetco, Zeng and Keane (2015) based on FCM.
    - Computational efficiency.

Universitat Oberta de Catalunya

Motivations

- Understand history of clustering
- Group data more efficiently
- How FCM improved KM and how soft algorithms improved FCM
- Better algorithm for each dataset

Universitat Oberta de Catalunya

Algorithms (1/2)

- K-Means
  - Minimize the objective function

$$J_{KM}(C_k) = \sum_i^C \sum_j^n D_{ij}^2$$

- Fuzzy C-Means
  - Minimize the objective function

$$J_{FCM}(X, U, V) = \sum_i^C \sum_j^n (u_{ij})^q D_{ij}^2$$

- Fuzzy C-Means++
  - Before FCM, find the best starting values for the centres of the clusters.
  - $\frac{dist^p}{sum(dist^p)}$

Algorithms (2/2)

- Suppressed-Fuzzy C-Means

$$u_{pj} = 1 - \alpha \sum_{i \neq p} u_{ij} = 1 - \alpha + \alpha u_{pj}$$

$$u_{ij} = \alpha u_{ij}, \quad i \neq p$$

- Gustafson Kessel Fuzzy C-Means
  - Adaptative distance norm by estimating the data covariance

$$A_i = [\rho_i \det(F_i)]^{1/n} F_i^{-1}$$

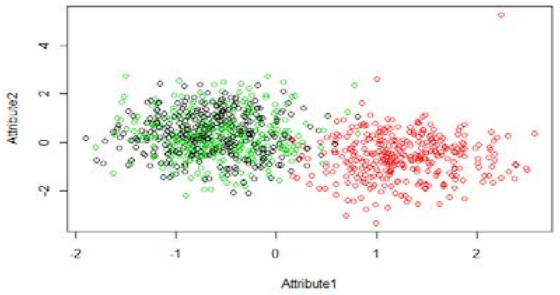
- Possibilistic C-Means

$$\eta_i = K \frac{\sum_{j=1}^{N} u_{ij}^m u_{ij}^2}{\sum_{j=1}^{N} u_{ij}^m}$$

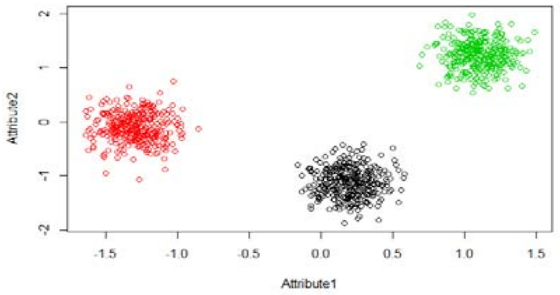
Universitat Oberta de Catalunya

Datasets (1/2)

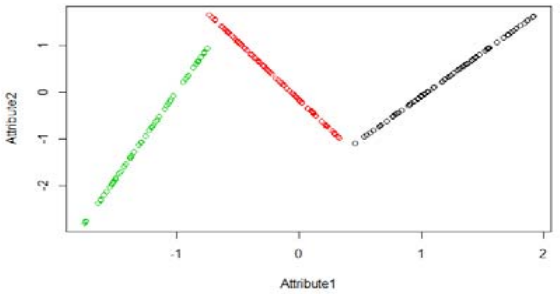
- SD1



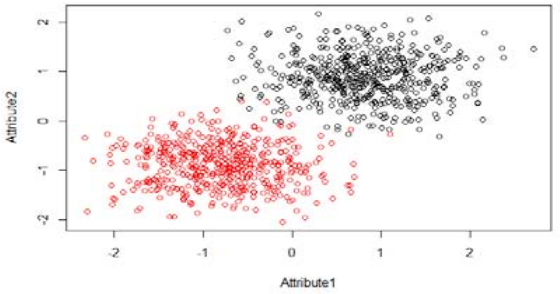
- SD2



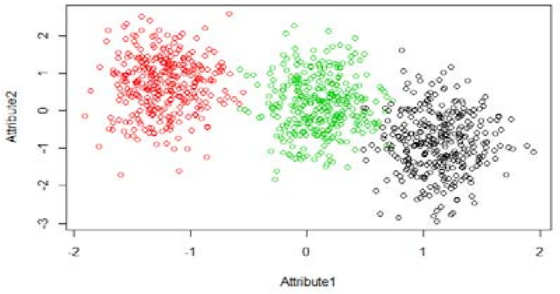
- SD3



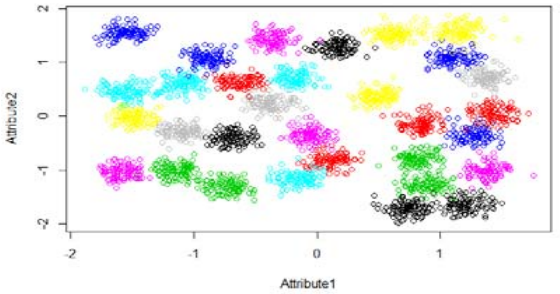
- SD4



- SD5

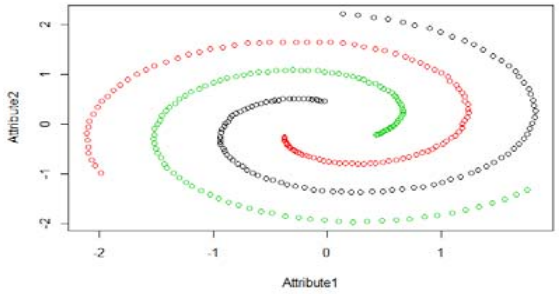


- SD6

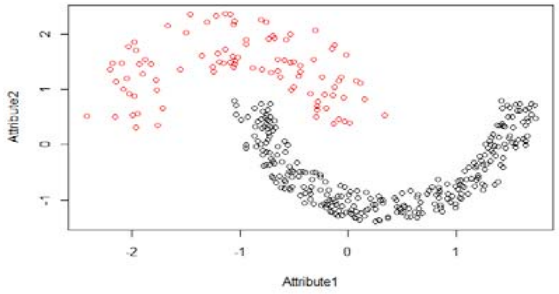


Datasets (2/2)

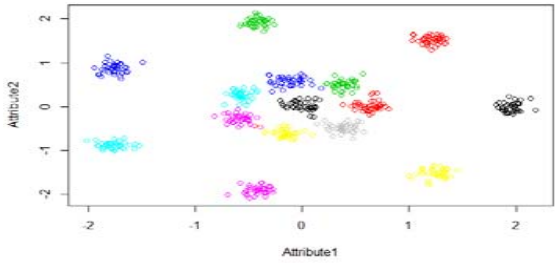
- SD7



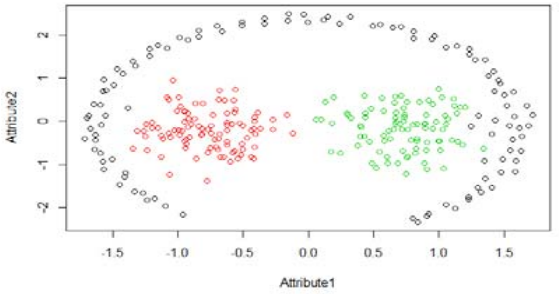
- SD8



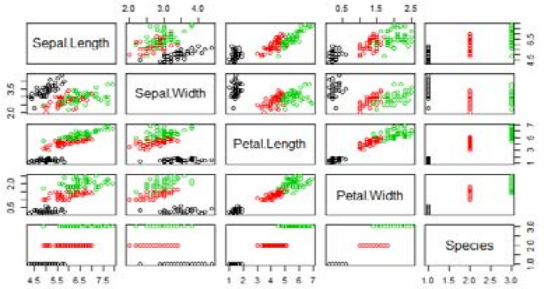
- SD9



- SD10



- RWD1



Universitat Oberta de Catalunya

Validation

- Internal validation
  - Computational efficiency
  - Performance
    - Xie-Beni

      $$[\sum_i \sum_{x \epsilon C_i} d^2(x, C_i)]/[n \cdot min_{i,j \neq i} d^2(x, C_i)]$$

    - S

      $$\{\frac{1}{NC}\sum_i\{\frac{1}{n_i}\sum_{x \epsilon C_i}\frac{b(x)-a(x)}{\max[b(x),a(x)]}\}$$

- External validation
  - Accuracy

Universitat Oberta de Catalunya

Experiments (1/6)

K-Means and Fuzzy C-Means

- Dataset SD1

Iterations K-Means	Iterations Fuzzy C-Means	Accuracy K-Means	Accuracy Fuzzy C-Means
2.89	63.68	29.8	36.5

- Dataset RWD1

Iterations K-Means	Iterations Fuzzy C-Means	Accuracy K-Means	Accuracy Fuzzy C-Means
2.14	47.17	29.5	28.7

- KM is more efficient in general
- FCM is more accurate when there is more overlapping of clusters

Universitat Oberta de Catalunya

Experiments (2/6)

Fuzzy C-Means, Fuzzy C-Means++ and Suppressed-Fuzzy C-Means

- Dataset SD1

Number of clusters (k)	It. FCM	It. FCM++	It. S-FCM	XB FCM	XB FCM++	XB S-FCM	Sil. FCM	Sil. FCM++	Sil. S-FCM
2	35.1	37.3	11.7	0.18	0.18	0.455	0.717	0.717	0.678
3	64.2	66.6	20	0.228	0.228	0.577	0.675	0.675	0.616
4	245.3	186.3	26.3	0.207	0.192	0.574	0.644	0.673	0.565
5	101.1	102.7	25.9	0.265	0.265	0.657	0.608	0.608	0.523

- Dataset RWD1

Number of clusters (k)	It. FCM	It. FCM++	It. S-FCM	XB FCM	XB FCM++	XB S-FCM	Sil. FCM	Sil. FCM++	Sil. S-FCM
2	23.4	22.8	10.5	0.113	0.113	0.253	0.807	0.807	0.792
3	45.7	48.4	9.8	0.222	0.222	0.59	0.729	0.729	0.69
4	68.9	73.9	13.4	0.272	0.272	0.835	0.669	0.668	0.596
5	107.5	73	14.3	0.367	0.359	1.68	0.609	0.658	0.524

- S-FCM is more efficie general
- FCM and FCM++ per better
- FCM++ behaves sam FCM

Experiments (3/6)

Fuzzy C-Means and Gustafson Kessel Fuzzy C-Means

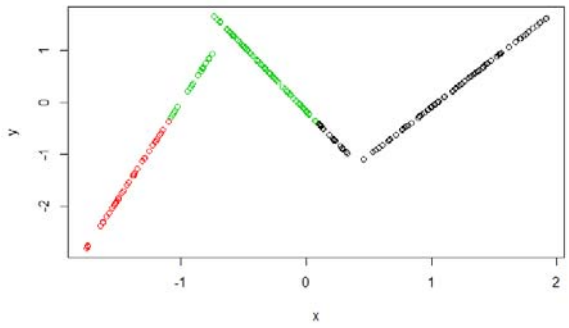
- Dataset SD3

Number of clusters (k)	It. FCM	It. GKFCM	XB FCM	XB GKFCM	Sil. FCM	Sil. GKFCM
2	71	161	0.208	0.388	0.713	0.492
3	113	13	0.179	0.416	0.73	0.497
4	45	14	0.085	0.359	0.814	0.397
5	684	14	0.222	1.088	0.752	0.155

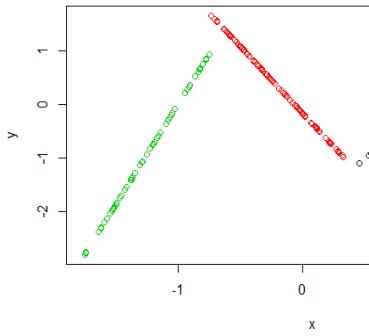
- Dataset SD7

Number of clusters (k)	It. FCM	It. GKFCM	XB FCM	XB GKFCM	Sil. FCM	Sil. GKFCM
2	1000	668	0.37	0.327	0.577	0.56
3	210	319	0.152	0.181	0.657	0.612
4	1000	1000	0.17	0.168	0.632	0.581
5	1000	356	0.154	0.496	0.619	0.353

- SD3:
  - FCM performs better and is more efficient
  - GKFCM is more accurate
- SD7:
  - GKFCM performs better
  - FCM and GKFCM similar efficiency



FCM



GKFCM

Experiments (4/6)

Fuzzy C-Means and Gustafson Kessel Fuzzy C-Means

- Dataset SD8

Number of clusters (k)	It. FCM	It. GKFCM	XB FCM	XB GKFCM	Sil. FCM	Sil. GKFCM
2	35	380	0.142	0.139	0.744	0.734
3	88	72	0.152	0.186	0.759	0.524
4	91	195	0.105	0.204	0.754	0.56
5	152	405	0.153	0.787	0.734	0.471

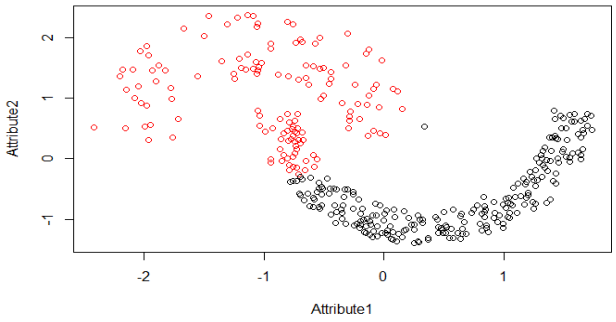
- Dataset SD9

Number of clusters (k)	It. FCM	It. GKFCM	XB FCM	XB GKFCM	Sil. FCM	Sil. GKFCM
15	70	818	1	0.844	0.818	0.85

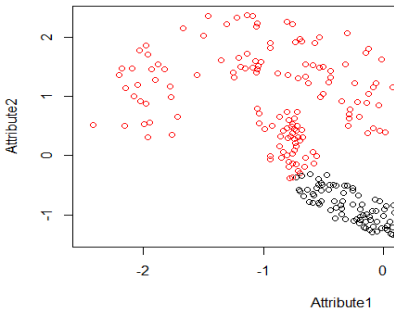
- Dataset SD10

	It. FCM	It. GKFCM	XB FCM	XB GKFCM	Sil. FCM	Sil. GKFCM
2	44	45	0.277	0.257	0.65	0.633
3	61	291	0.14	0.171	0.751	0.668
4	72	707	0.28	0.176	0.71	0.7
5	113	588	0.354	0.357	0.625	0.569

- SD8:
  - GKFCM performs better for the actual number of clusters
  - FCM and GKFCM same efficiency
  - Fail to find the actual clusters
- SD9:
  - GKFCM performs better
  - FCM is more efficient
- SD10:
  - GKFCM performs better
  - FCM is more efficient
  - Fail to find the actual clusters



SD8 - GKFCM



SD8 - FC

Universitat Oberta de Catalunya

Experiments (5/6)

Fuzzy C-Means and Possibilistic C-Means

- Dataset SD4

It. FCM	It. PCM	Accuracy FCM	Accuracy PCM
16	38	0,99	0,991

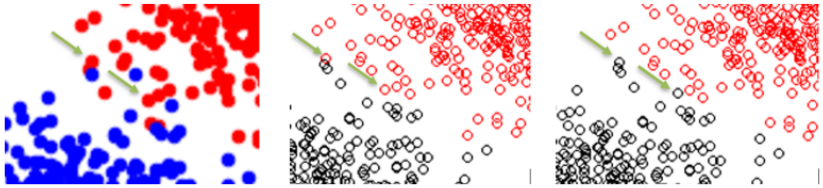
- Dataset SD5

It. FCM	It. PCM	Accuracy FCM	Accuracy PCM
49	194	88.7%	33.5%

- Dataset SD6

It. FCM	It. PCM	Accuracy FCM	Accuracy PCM
234	81	89.7%	87,7%

- SD4:
  - FCM is more efficient
  - PCM is slightly more accurate, but not significant
- SD5:
  - FCM is more efficient and accurate
- SD6:
  - PCM is more efficient
  - FCM and PCM are equally accurate



SD4 PCM FCM

Experiments (6/6)

Fuzzy C-Means and Possibilistic C-Means

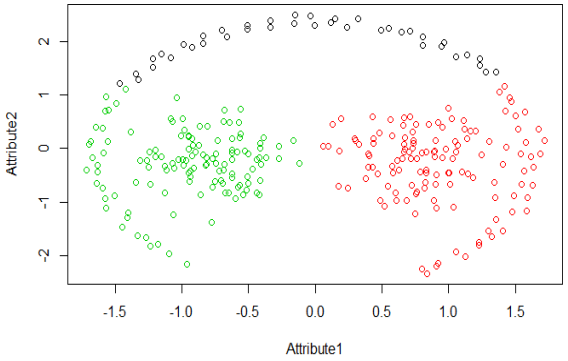
- Dataset SD9

It. FCM	It. PCM	Accuracy FCM	Accuracy PCM
32	56	99,67%	99,67%

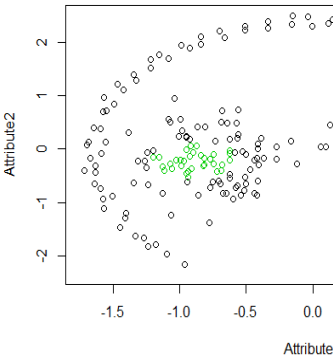
- Dataset SD10

It. FCM	It. PCM	Accuracy FCM	Accuracy PCM
56	96	76.3%	65%

- SD9:
  - FCM is more efficient
  - FCM and PCM same accuracy
- SD10:
  - FCM is more efficient and accurate
  - Both fail to find the correct clusters



SD10 - FCM



SD10 - P

Universitat Oberta de Catalunya

Discussion

- FCM is more accurate when there is more overlapping, but KM is more efficie
- S-FCM is more efficient than FCM and FCM++, but it performs worse
- FCM++ doesn´t have any advantage over FCM
- GKFCM performs better than FCM for non-spherical clusters but FCM is more efficient
- PCM does not show any advantage over FCM

Universitat Oberta de Catalunya

Conclusions

- History of clustering and main contributions
- Series of experiments for the validation of the improvements of the algorithms
- FCM for overlapping clusters
- KM for efficiency
- FCM and KM are quite solid despite the latest algorithms
- S-FCM more efficient for overlapping clusters
- GKFCM for some non-spherical clusters
- Further investigation is needed

Universitat Oberta de Catalunya

Thank you