Possibilistic Fuzzy Clustering Using Self Organizing Maps

Miguel Loureiro[1], Victor Lobo[1,2], Fernando Bação[1]
mloureiro@isegi.unl.pt, vlobo@isegi.unl.pt, bacao@isegi.unl.pt

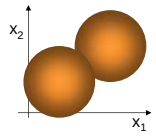
[1]Instituto Superior de Estatística e Gestão de Informação, Universidade Nova de Lisboa
[2]Escola Naval, Marinha Portuguesa

Summary

- Purpose of fuzzy clustering
- Some limitations of Fuzzy C-Means
- Self Organizing Map and U-Mat
- Proposed method
- Examples
- Conclusions

Purpose of fuzzy clustering

- Identify clusters of data
- Assign a membership value of each data pattern to each cluster
- Most well known algorithm:
  - FCM – Fuzzy C-Means [Dunn][Bezdek]

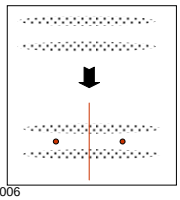


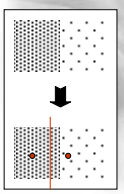
$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^m \left\| x_i - c_j \right\|^2 \quad , \quad 1 \le m < \infty$$

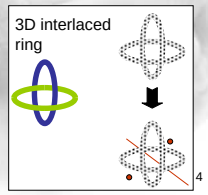
$$u_{ij} = \frac{1}{\sum_{k=1}^{C}\left(\frac{\left\| x_i - c_j \right\|}{\left\| x_i - c_k \right\|}\right)^{\frac{2}{m-1}}} \quad , \quad c_j = \frac{\sum_{i=1}^{N} u_{ij}^m \cdot x_i}{\sum_{i=1}^{N} u_{ij}^m}$$

Some limitations (1)

- Distance based criteria
  - Distance to centroid defines membership
  - Common in most clustering algorithms
  - Unable to identify some “natural” clusters
  - Examples:







Some limitations (2)

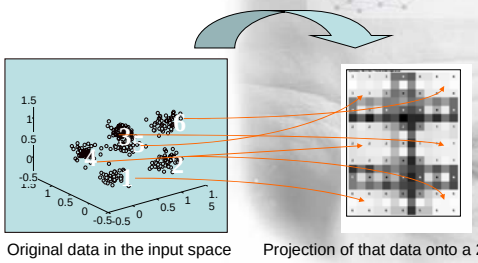
- Method to automatically assign membership
  - Usually “probabilistic”
    - Sum of memberships to all centroids = 1
    - Obtained from distance to each centroid
  - Inconvenient
    - Border patterns have 0.5 membership to both classes
      - (they should have a higher membership to both !)
    - Outliers have 0.5 membership to both classes
      - (they should have a smaller membership to both !)

Ideas to overcome limitations

- Distance based → Density based
  - Use variations in data density to define clusters
  - “linkage” based clustering has similar results
- Probabilistic → Possibilistic membership
  - Automatically assign memberships that do not add to 1
- Use a neural network (SOM/U-Mat) to define memberships and identify centroids

Self Organizing Map (SOM)

- Neural network proposed by Kohonen
- Performs a non-linear projection from $\Re^n$ to $\Re^2$



Original data in the input space

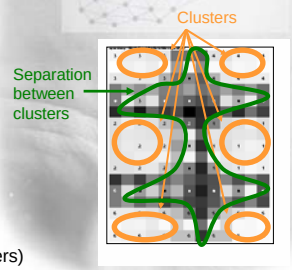
Projection of that data onto a 2D surface (each point is a neuron, or UNIT)

U-Matrix (U-Mat)

- U-Matrix maps distances (in the the input space) between neighboring units (in the output space)

Low values
⇒ Units are near
⇒ high data density
⇒ cluster

High values
⇒ Units are far apart
⇒ dispersed data
⇒ non-cluster
(or separation between clusters)



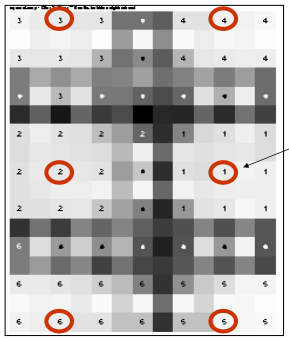
Proposal of a new fuzzy clusternig method - MFCM

- Definition of cluster centroids
  - By inspection of the U-Mat of the data
  - Advantages vs disadvantages…
- Definition of memberships to the centroids
  - Using the new proposed membership function MM, which is based on the U-Mat of the data

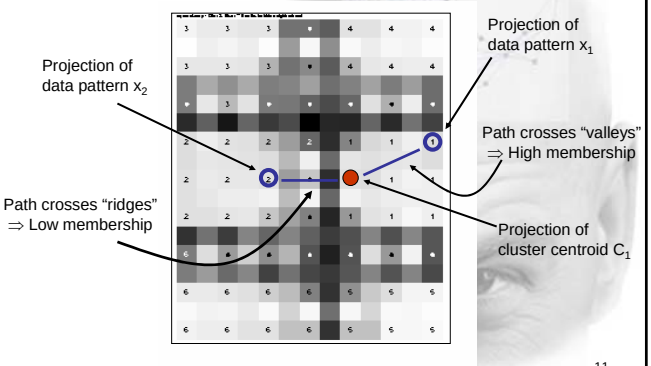
Basic Idea for centroid definition

•Choose:

1. As many centroids as "distinct areas"
2. The "most central", or "most typical" unit in each area



Basic Idea for Membership



Proposed membership method (MM)

- To compute the membership of each pattern to a given centroid, use the "height difference" measured in the U-MAT
- Basic idea:
  1. Obtain a SOM and U-Mat with all available data
  2. Map the data pattern on the U-Mat
  3. Map the cluster centroide on the U-Mat
  4. Compute the sum of differences in U-Mat values along the line that connects those two projections
  5. Compute membership as a function of that sum
- Basic consequence:
  - Membership is a function of variations in data density

Functions used

- Cost function along the U-Mat

$$D_u(x,c) = \min(\sum_{PB} \Delta U)$$

P(x)=Projection of data point x
P(c)=Projection of centroid c

ΔU = absolute diference between U-mat values of adjacent units

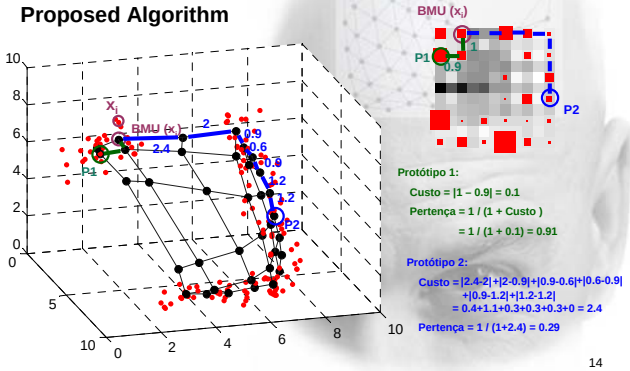
PB – path between points *P(x), P(c)*

  - This cost function IS NOT a distance in the U-MAT
  - The path is found using Simulated Annealing
- Membership of a data point to a centroid

$$M(x,c) = e^{-2\frac{D(x,c)}{\max(U)-\min(U)}}$$

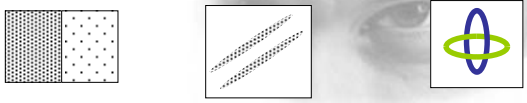
Example of cost computation

Proposed Algorithm



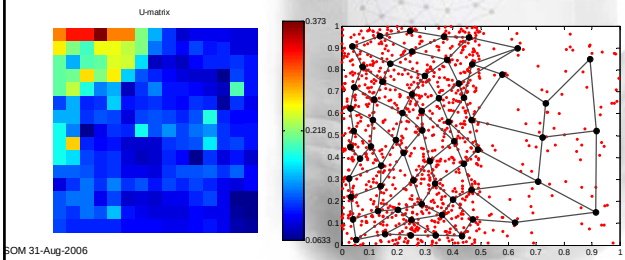
Examples

Implemented in Matlab/Octave
using SOMToolbox and
MagosToolbox



Results – Case1

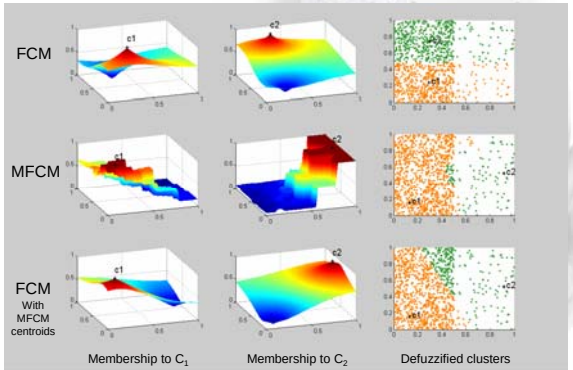
(different densities)



U-Matrix

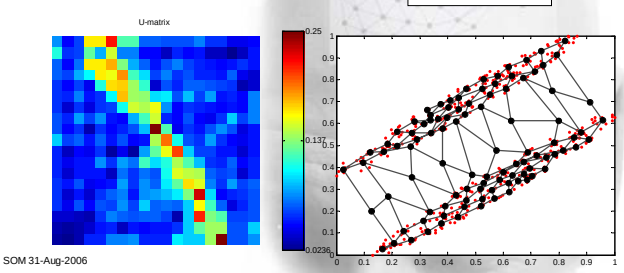
SOM units in the input space

Results – Case 1



Results – Case2

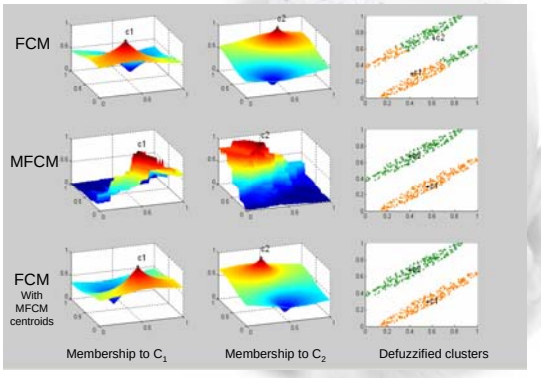
(long shapes)



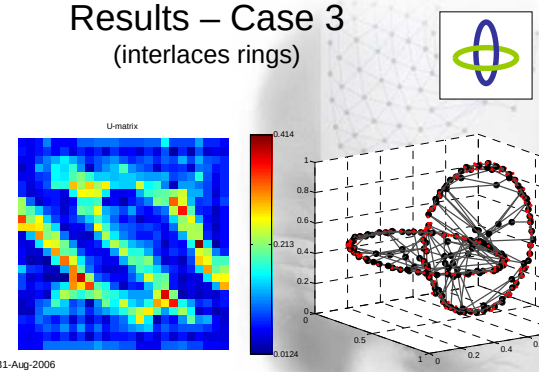
U-Matrix

SOM units in the input space

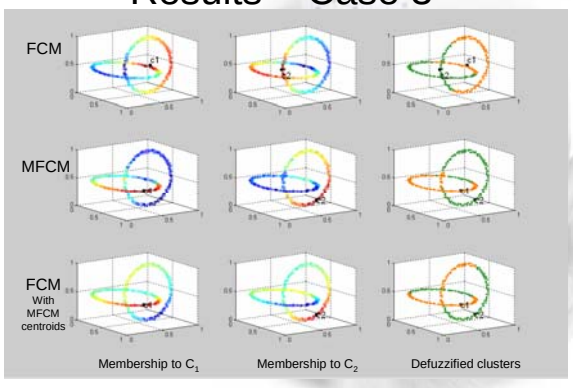
Results – Case 2



Results – Case 3
(interlaces rings)



Results – Case 3



Conclusions

- New method to compute fuzzy membership
- Clusters based on density variation
- Centroids chosen from a U-Mat
- Good results in difficult problems

Future work…

- Define a sound methodology for selecting centroids in the U-Mat
- Continue tests with simulated annealing procedure to guaranty its robustness
- Fine tune some parameters such as the decrease of the membership as a function of cost
- Improve the software implementation

Brought to you by…

Mining,
Aprendizagem,
Geografia,
Optimização e
Simulação