Possibilistic Fuzzy Clustering Using Self Organizing Maps

Miguel Loureiro¹, Victor Lobo¹,², Fernando Baçãao³
mloureiro@isegi.unl.pt, vlobo@isegi.unl.pt, bacao@isegi.unl.pt
¹Instituto Superior de Estatística e Gestão de Informação, Universidade Nova de Lisboa
²Escola 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]

Some limitations (1)
• Distance based criteria
  – Distance to centroid defines membership
  – Common in most clustering algorithms
  – Unable to identify some “natural” clusters
  – Examples:
    3D interlaced ring

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 $\mathbb{R}^n$ to $\mathbb{R}^2$

![Image of SOM](SCRA2006)

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)

![Image of U-Matrix](SCRA2006)

Proposal of a new fuzzy clustering 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

![Image of MFCM](SCRA2006)

Basic Idea for centroid definition
- Choose:
  1. As many centroids as “distinct areas”
  2. The “most central”, or “most typical” unit in each area

![Image of centroid definition](SCRA2006)

Basic Idea for Membership
- 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 centroid 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

![Image of membership](SCRA2006)
Functions used

- Cost function along the U-Mat
  \[ D_u(x,c) = \min \left( \sum_{PB} \Delta U \right) \]
  - 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
  \[ M(x,c) = e^{-\frac{D(x,c)}{\max(U)-\min(U)}} \]

Example of cost computation

Suppose \( P(x) \) is the projection of data point \( x \), \( P(c) \) is the projection of centroid \( c \), and \( \Delta U \) is the absolute difference between U-mat values of adjacent units.

**Proposed Algorithm**

1. **BMU (xi)**
2. **xi**
3. **BMU (xi)**

\[ \text{Cost} = |1 - 0.9| = 0.1 \]
\[ \text{Pertença} = \frac{1}{1 + \text{Cost}} = \frac{1}{1 + 0.1} = 0.91 \]

\[ \text{Cost} = 0.4 + 1.1 + 0.3 + 0.3 + 0.3 + 0 = 2.4 \]
\[ \text{Pertença} = \frac{1}{1 + 2.4} = 0.29 \]

Examples

Implemented in Matlab/Octave using SOMToolbox and MagosToolbox

Results – Case 1

(different densities)

Results – Case 2

(long shapes)
Results – Case 2

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