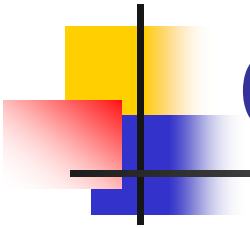
A fuzzy clustering method using Genetic Algorithm and Fuzzy Subtractive Clustering

Thanh Le, Tom Altman and Katheleen Gardiner

University of Colorado Denver

July 18, 2012



Overview

- Introduction
  - Fuzzy clustering using Fuzzy C-Means algorithm
  - Current genetic algorithms for fuzzy clustering
- Proposed method: fzGASCE
  - Genetic algorithm
  - Fuzzy subtractive clustering
  - Probability based fitness function
- Datasets:
  - Artificial datasets: Finite mixture model
  - Real datasets: UCI repository
- Experimental results
- Discussion

Fuzzy C-Means algorithm- FCM

- Objective function

$$J(X \mid U, V) = \sum_{i=1}^{n} \sum_{k=1}^{c} u_{ki}^{m} \left\| x_i - v_k \right\|^2 \rightarrow \min, \ \ m \geq 1$$

$$\sum_{k=1}^{c} u_{ki} = 1, \ \ i = 1..n$$

- Model parameters estimation:

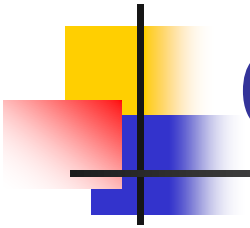
$$u_{ki} = \left(\frac{1}{\left\| x_i - v_k \right\|^2} \right)^{\frac{1}{m-1}} \Bigg/ \sum_{l=1}^{c} \left(\frac{1}{\left\| x_i - v_l \right\|^2} \right)^{\frac{1}{m-1}}$$

$$v_k = \sum_{i=1}^{n} u_{ki}^{m} x_i \Bigg/ \sum_{i=1}^{n} u_{ki}^{m}$$

FCM algorithm (contd.)

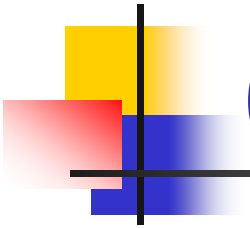
- Advantages
  - Model free
  - Rapid convergence
  - Multiple cluster assignment
- Shortcomings
  - Definition of the number of clusters
  - Fuzzy partition evaluation
  - Convergence to local optima
  - Defuzzification

Recent fuzzy clustering Genetic Algorithms (GA)



- Chromosome describes a clustering solution
- Fitness functions are based on cluster indices
- Random mutations
  - Genes to be replaced
  - Genes to replace

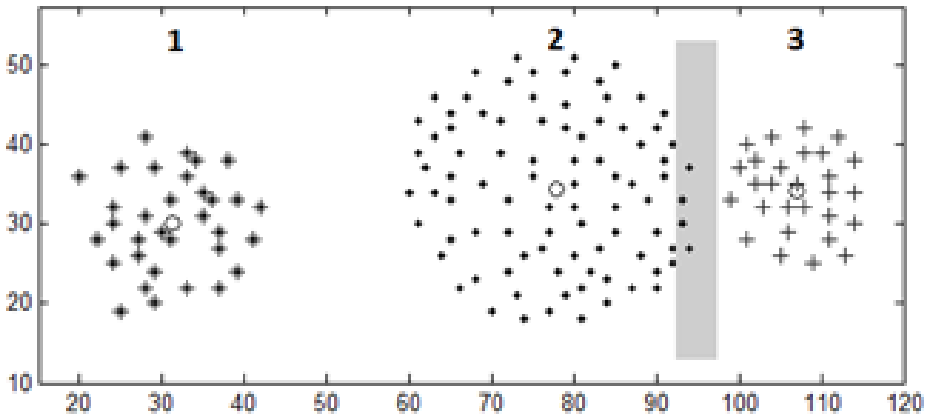
Recent fuzzy clustering GAs (contd.)



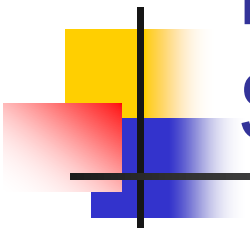
- Advantages
  - Search for the ‘best’ solution in the solution space
  - Can escape local optima
    - Cross-over operator
    - Mutation operator
  - Can determine the number of clusters using the ‘best’ solution

Recent fuzzy clustering GAs (contd.)

- Shortcomings
  - Problem with cluster indices
    - Scale between compactness and separartion
  - Random selection of genes to be replaced
  - Improper defuzzification



Fuzzy clustering using GA and Subtractive Clustering - fzGASCE



- Chromosome describes a clustering solution
- Data clustering using FCM
- Probability based fitness function
- Mutation gene selection using fuzzy Subtractive Clustering
- Defuzzification of fuzzy partition using probabilistic model

fzGASCE: the probabilistic model

- Bayesian validation method for fuzzy clustering - fzBLE (Le et al. , 2011)
  - Central limit theorem
  - Bayesian theory
- Possibility to probability transformation

  $\{u_{ki}\}_{i=1..n}$ - possibility distribution of X at v_k

  $\{p_{ki}\}_{i=1..n}$ - probability distribution of X at v_k,
- Create the probabilistic model at v_k using $\{p_{ki}\}_{i=1..n}$

Use of fzGASCE probabilistic model

- fzGASCE fitness function

  $fit(\{U,V\}) = Prob(X|\{U,V\})$

  - Address the problems with using cluster indices
  - Outperform cluster indices on artificial and real datasets (Le et al., 2011)
- Defuzzification of fuzzy partition

  $Prob(v^*|x_i) = \max\{Prob(v_k|x_i)\}$

  - Address the problems of maximum membership and spatial information methods (Le et al. 2012)

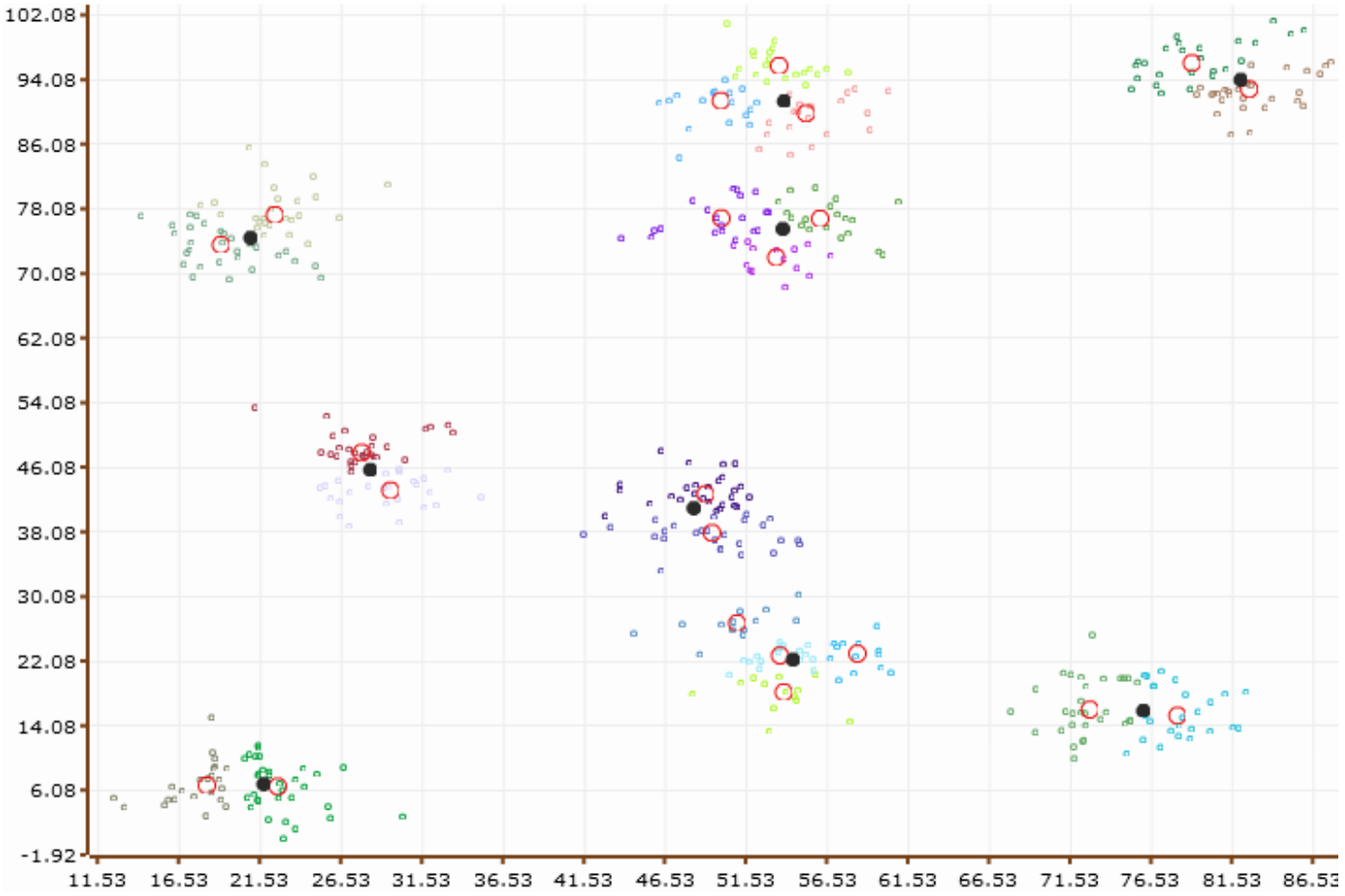
Application of fuzzy Subtractive Clustering (fzSC) in fzGASCE

- fzSC method (Le et al., 2011)
  - Fuzzy mathematics application
  - Histogram based density estimation
  - Data density using fuzzy partition
- Application of fzSC in fzGASCE
  - Order data points based on data densities
  - The most dense data points are used to replace mutated genes

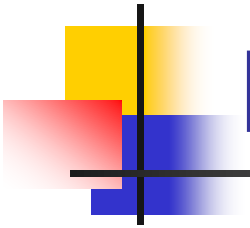
fzSC – an example on how it works

Red-circle: Cluster centers of fuzzy partition

Back-circle: The most dense data points found by fzSC



fzSC demonstration available online: http://demo.tinyray.com/fzsco



Datasets

- Artificial datasets
  - Datasets generated using finite mixture model
  - Non-uniform dataset
    - Clusters differ in size and density
- Real datasets
  - Iris
  - Wine
  - Glass

  These datasets are from UC Irvine Machine Learning

Performance measures

- Correctness ratio

$$COR = \frac{1}{N}\sum_{i=1}^{N} I(c - \hat{c})$$

where, N is the number of trials

- Error variance

$$EVAR = \frac{1}{N}\sqrt{(c - \hat{c})^2}$$

- Misclassification

$$EMIS = \frac{1}{N}\sum_{t=1}^{N}\frac{1}{n}\sum_{i=1}^{n} I(x_i^c - x_i^l)$$

Compare the cluster label of each data object with its actual class label

Uniform dataset – ASET1

Algorithm	COR	EVAR	EMIS
fzGASCE	**1.000**	0.000	0.000
fzGAE	0.640	0.500	0.000
PBMF	0.510	0.590	0.000
MPC	0.290	0.970	0.000
HPK	0.100	5.010	0.021
AGFCM	0.600	2.800	0.000
XB	0.490	1.450	0.000
FS	0.120	1.100	0.070
PC	0.230	1.040	0.000
ACVI	0.200	2.490	0.011

fzGAE is an immature version of fzGASCE, where the fzSC method is not used in the mutation operator

Uniform dataset – ASET2

Algorithm	COR	EVAR	EMIS
fzGASCE	**1.000**	0.000	0.000
fzGAE	0.710	0.380	0.000
PBMF	0.600	0.450	0.000
MPC	0.610	0.860	0.000
HPK	0.120	5.240	0.000
AGFCM	0.650	1.490	0.000
XB	0.640	0.430	0.000
FS	0.520	0.840	0.011
PC	0.620	0.890	0.000
ACVI	0.100	2.100	0.000

Non-uniform dataset – ASET4

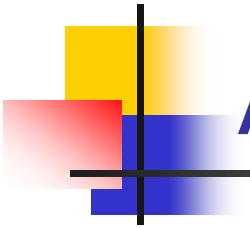
Algorithm	COR	EVAR	EMIS
fzGASCE	**1.000**	0.000	0.000
fzGAE	0.900	0.100	0.107
PBMF	0.700	0.300	0.107
MPC	0.050	0.960	0.107
HPK	0.000	5.770	-
AGFCM	0.000	8.470	-
XB	0.040	0.960	0.107
FS	0.020	3.480	0.107
PC	0.050	0.960	0.107
ACVI	0.080	0.920	0.107

Iris dataset

Algorithm	COR	EVAR	EMIS
fzGASCE	**1.000**	0.000	0.033
fzGAE	0.880	0.120	0.040
PBMF	0.860	0.140	0.040
MPC	0.040	0.970	0.160
HPK	0.000	5.720	-
AGFCM	0.000	8.120	-
XB	0.050	1.010	0.040
FS	0.390	0.780	0.154
PC	0.080	0.920	0.115
ACVI	0.150	0.850	0.040

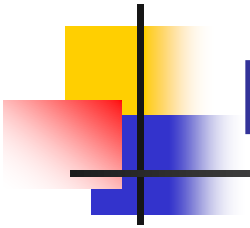
Wine dataset

Algorithm	COR	EVAR	EMIS
fzGASCE	**1.000**	0.000	0.213
fzGAE	0.860	0.140	0.303
PBMF	0.000	2.050	-
MPC	0.000	2.810	-
HPK	0.000	6.760	-
AGFCM	0.000	9.210	-
XB	0.270	1.010	0.303
FS	0.000	5.720	-
PC	0.110	0.920	0.303
ACVI	0.090	0.910	0.303



Advantages of fzGASCE

- Describe the data distribution using probabilistic model
- Apply the probabilistic model into fitness function and defuzzification
- Use of fzSC method with mutation operator to effectively escape local optima
- No parameters to be specified a priori



Future work

- Eliminate the oscillation during the convergence process when using fzSC to speed up fzGASCE
- Integrate with external distance measures to meet specific requirements of real-world applications.

Thank you!

Questions?

- We acknowledge the supports from
  - Vietnamese Ministry of Education and Training, the 322 scholarship program.
  - University of Colorado Denver, USA