PCM Clustering based on Noise Level

Peixin Hou[1] Jiguang Yue[1] Hao Deng[2] Shuguang Liu[3]

[1]Department of Control Science and Engineering, Tongji University

[2]School of Physics and Electronics, Henan University

[3]Department of Hydraulic Engineering, Tongji University

FUZZ-IEEE, 2017

Outline

Introduction

How much information we should provide for the clustering algorithm in order to discover the natural (physical) clusters?

- information of the cluster number

Introduction

How much information we should provide for the clustering algorithm in order to discover the natural (physical) clusters?

- information of the cluster number
- information of the property (e.g., density or closeness) of clusters.

Our Main Contributions

- The proposed NPCM have two parameters, where m_{ini} is a possibly over-specified cluster number, and α characterizes the closeness of clusters in the clustering result.
- Both parameters are not required to be exactly specified.

The PCM Algorithm

The objective of possibilistic c-means (PCM) [1] is to minimize the following cost:

$$J(\mathbf{\Theta}, \mathbf{U}) = \sum_{j=1}^{c} J_j = \sum_{j=1}^{c} \left[\sum_{i=1}^{N} u_{ij} d_{ij}^2 + \gamma_j \sum_{i=1}^{N} f(u_{ij}) \right] \tag{1}$$

where $f(\cdot)$ can be chosen as:

$$f(u_{ij}) = u_{ij} \log u_{ij} - u_{ij}. \tag{2}$$

The PCM Algorithm

Minimizing $J(\mathbf{\Theta}, \mathbf{U})$ with respect to u_{ij} and $\boldsymbol{\theta}_j$ leads to the following two update equations:

$$u_{ij} = \exp\left(-\frac{d_{ij}^2}{\gamma_j}\right) \tag{3}$$

$$\boldsymbol{\theta}_j = \frac{\sum_{i=1}^{N} u_{ij}\mathbf{x}_i}{\sum_{i=1}^{N} u_{ij}} \tag{4}$$

Drawback of PCM

- γ_j is fixed after initialization, so its performance relies heavily on good initial partitions and parameters [2].

Improvement of Adaptive PCM [4]

- Adapt γ_j at each iteration.

Drawback of PCM

- γ_j is fixed after initialization, so its performance relies heavily on good initial partitions and parameters [2].
- The c dense regions found may be coincident [3].

Improvement of Adaptive PCM [4]

- Adapt γ_j at each iteration.
- The clusters with $\gamma_j = 0$ are eliminated.

The APCM Algorithm

The price of Adaptive PCM (APCM): introduce another parameter α to manually control the process of adaptive α_j.

$$\gamma_j = \frac{\hat{\eta}}{\alpha}\eta_j \tag{5}$$

$\hat{\eta}$ is initialized in the first iteration.
Let $A_j = \{\mathbf{x}_i | u_{ij} = \max_r u_{ir}\}$. η_j is updated at each iteration as:

$$\eta_j = \frac{1}{n_j} \sum_{\mathbf{x}_i \in A_j} ||\mathbf{x}_i - \boldsymbol{\mu}_j|| \tag{6}$$

where n_j and $\boldsymbol{\mu}_j$ are the number of points in A_j and the mean vector of points in A_j respectively.

What APCM Acutually Does

Introduce parameter α to correct the estimated bandwidth η_j, so that the corrected bandwidth (5) approximates the true one.

The UPCM Algorithm: an alternate Bandwidth Correction Method

In the UPCM paper [5], the bandwidth correction process is performed in a more fuzzy way by utilizing the conditional fuzzy set framework [6]:

$$\mu_{ij} = \exp\left(-\frac{d_{ij}^2}{\gamma_j}\right) \tag{7}$$

where $\gamma_j = \left(0.5\eta_j + 0.5\sqrt{\eta_j^2 + 4\sigma_v d_{ij}}\right)^2$ and σ_v controls the uncertainty of the estimated bandwidth.

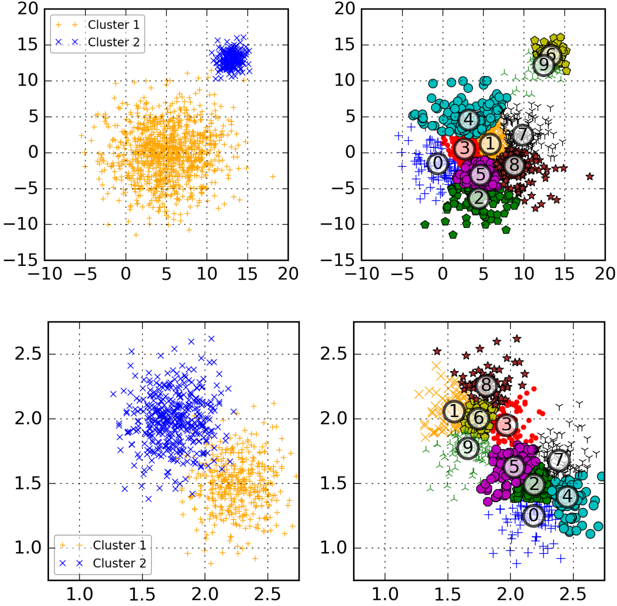
The UPCM Algorithm: an alternate Bandwidth Correction Method

UPCM also introduces the concept of *noise level* α of the data set in the update equation of prototypes:

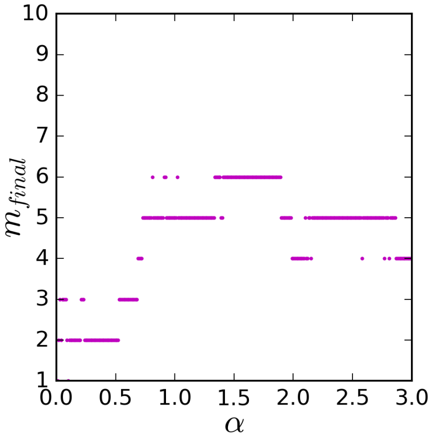
$$\boldsymbol{\theta}_j = \frac{\sum_{i=1}^{N} u_{ij}\mathbf{x}_i}{\sum_{i=1}^{N} u_{ij}} \quad \text{for } u_{ij} \geq \alpha. \tag{8}$$

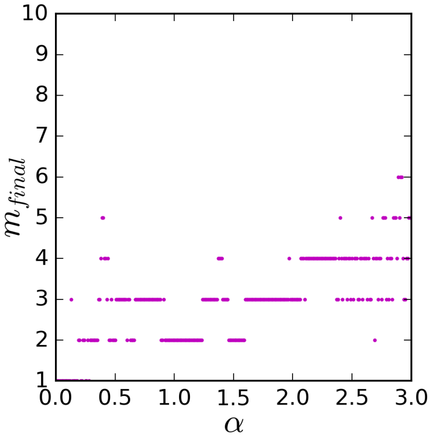
By setting an appropriate α, the influence of points in other clusters $\boldsymbol{\theta}_{i \neq j}$ on the $\boldsymbol{\theta}_j$ update is reduced.

Dataset 1 (upper) and Dataset 2 (lower) and their initializations



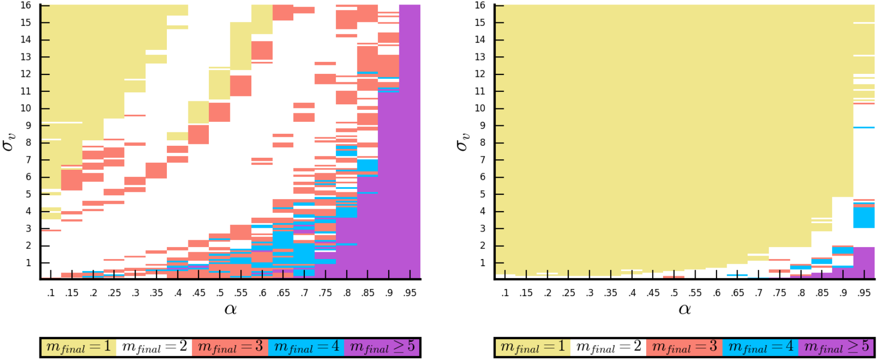
The Final Cluster Number of APCM





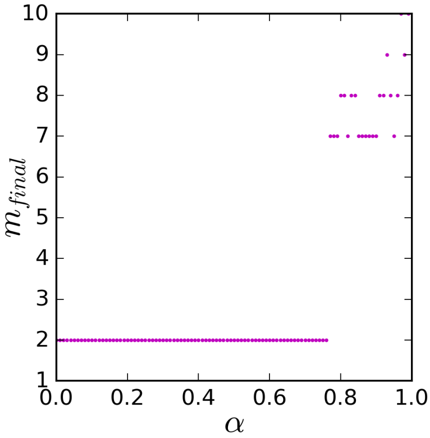
Result on Dataset 1 (left) and Dataset 2 (right)

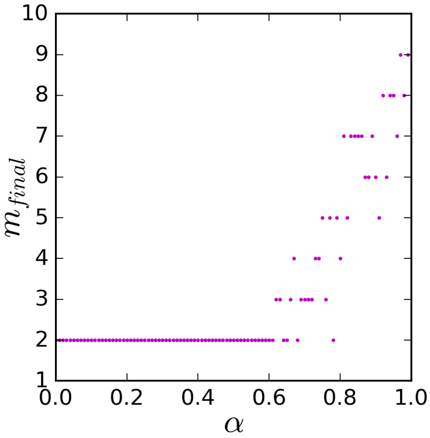
The Final Cluster Number of UPCM



Result on Dataset 1 (left) and Dataset 2 (right)

The Final Cluster Number of NPCM





Result on Dataset 1 (left) and Dataset 2 (right)

Motivation of This Paper

- UPCM has two hyperparameters, i.e., α and σ_v. It's possible to eliminate parameter σ_v

Motivation of This Paper

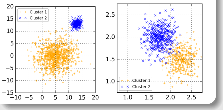
- UPCM has two hyperparameters, i.e., α and σ_v. It's possible to eliminate parameter σ_v
- Both APCM and UPCM suffer from the problem of background noise clusters, i.e., the background noise clusters are highly possible to become very large.

Algorithm Outline

Uncertainty Perspective

Elimination of σ_v

Clustering Problems



Motivations 1: to avoid small clusters being attracted to large clusters, and 2: to develop a more flexible bandwidth correction technique.

UPCM

- ❒ α: to reduce the influence of points from other clusters.
- ❒ σ_v: uncertainty modelling of the membership function.

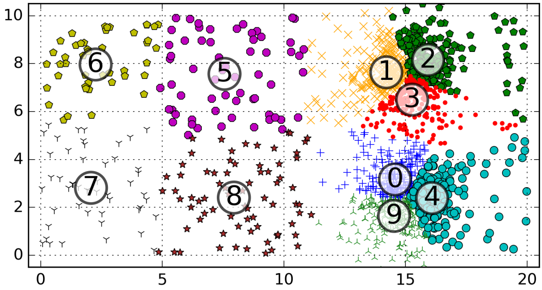
Stage 1: UPCM is an intermediate version of NPCM because UPCM naively implements the motivations and it's difficult to choose α and σ_v together.

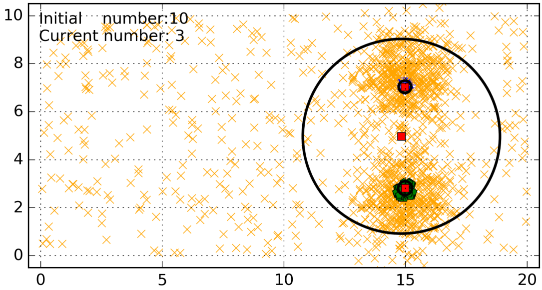
NPCM

- ❒ α: to characterize the clossness of clusters in the clustering result.

Stage 2: the final NPCM algorithm. After eliminating σ_v of UPCM, the α parameter has more sensible physical meaning and is much easier to set.

The Background Noise Problem





Eliminate Noise Clusters

Define the density of a cluster as:

$$\rho_j = \frac{n_j}{\eta_j^d} \tag{9}$$

where d is the dimension of $\mathbf{x}_i$. Let $\rho_0 = \max_j \rho_j$. Then the cluster C_j is considered as a noise cluster and is eliminated if $\rho_j < 0.1\rho_0$.

We can calculate the distance $d_{j\alpha}$ beyond which a point can't be used to contribute to the adaption of cluster C_j by letting

$$\exp\left(-\frac{(d_{j\alpha})^2}{\gamma_j}\right) = \alpha, \tag{10}$$

which leads to

$$d_{j\alpha} = \sqrt{-\ln\alpha}\left(\eta_j + \sqrt{-\ln\alpha}\sigma_v\right). \tag{11}$$

When there is no uncertainty in the estimated bandwidth, we get $d^0_{j\alpha} = \sqrt{-\ln\alpha}\eta_j$.

We fix the effect of σ_v as the correction of $d^0_{j\alpha}$ by considering the uncertainty of the estimated bandwidth

$$\frac{d_{j\alpha} - d^0_{j\alpha}}{d^0_{j\alpha}} = \frac{\sqrt{-\ln \alpha}\sigma_v}{\eta_j} = \beta, \tag{12}$$

which leads to

$$\sigma_v = \beta \frac{\eta_j}{\sqrt{-\ln \alpha}}. \tag{13}$$

where β characterizes the correction degree. In this paper, we choose $\beta = 0.2$.

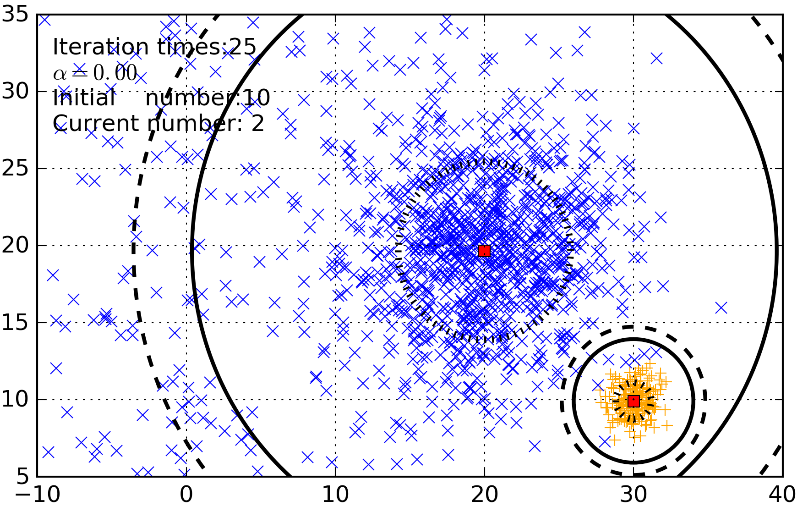
The update of the membership function is then modified according to (7) and (13) as:

$$\mu_{ij} = \exp\left(-\frac{d_{ij}^2}{\gamma_j}\right) \tag{14}$$

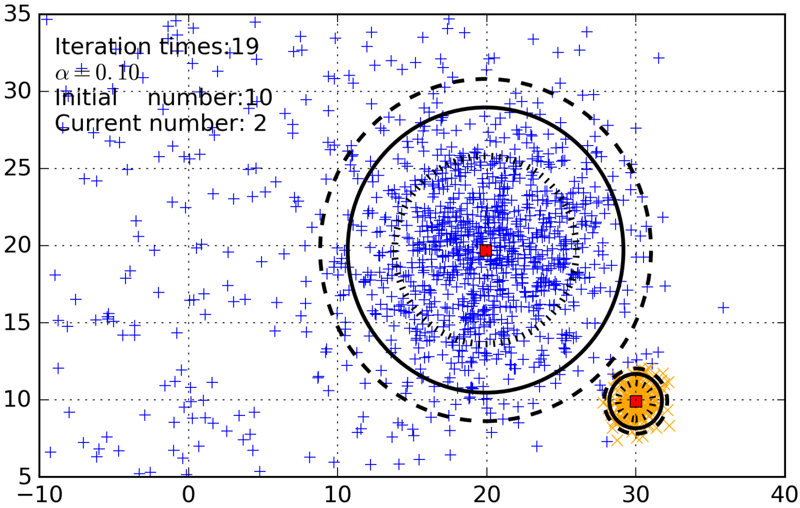
where $\gamma_j = \left(0.5\eta_j + 0.5\sqrt{\eta_j^2 + 0.8 d_{ij}\eta_j/\sqrt{-\ln\alpha}}\right)^2$ and $d_{ij} = ||\mathbf{x}_i - \boldsymbol{\theta}_j||$.

Outline

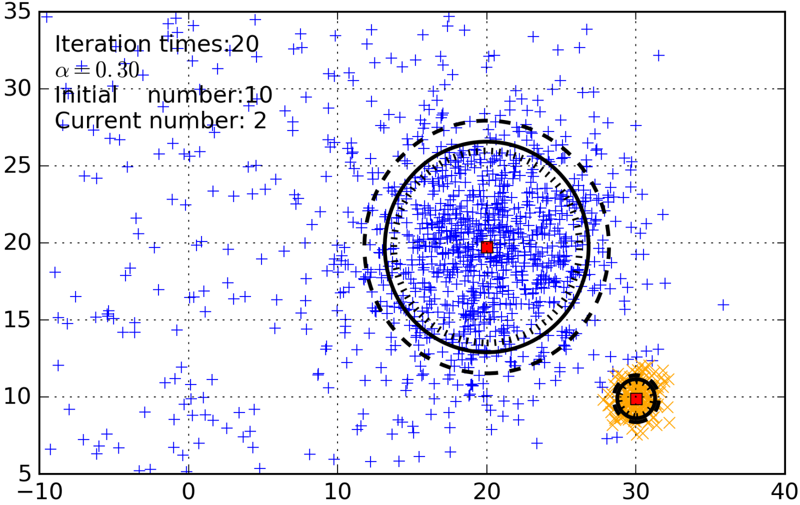
Experiment 1



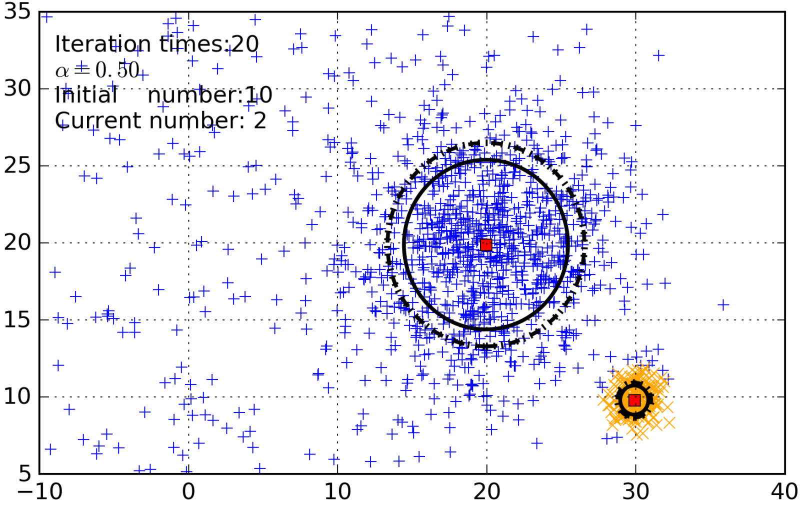
Experiment 1



Experiment 1



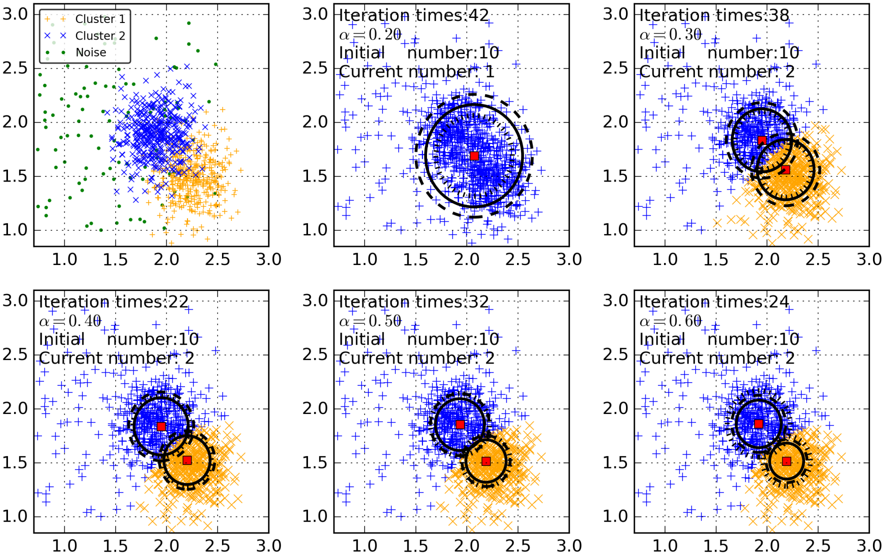
Experiment 1



Outline

1. Introduction
2. The PCM Based Algorithms
3. An Impression of the Algorithms' Performence
4. Algorithm Development
5. Experiments
   - Experiment 1
   - Experiment 2

Experiment 2



Summary

- NPCM addresses the background noise cluster problem encountered in APCM and UPCM.
- NPCM exploits the intuition from UPCM to eliminate the parameter σ_v basing on η_j.
- α is intuitive to choose and NPCM does not need strong prior information (e.g., the exact cluster number or the exact closeness of clusters) of the dataset.
- Outlook
  - Further researches can be conducted to reliably eliminate noise clusters in NPCM.
  - Nonlinear shape clusters.

[1] R. Krishnapuram and J. M. Keller, "A possibilistic approach to clustering," *IEEE Transactions on Fuzzy Systems*, vol. 1, no. 2, pp. 98–110, May 1993.

[2] O. Nasraoui and R. Krishnapuram, "An improved possibilistic C-Means algorithm with finite rejection and robust scale estimation," in *Fuzzy Information Processing Society, 1996. NAFIPS., 1996 Biennial Conference of the North American*, Jun. 1996, pp. 395–399.

[3] M. Barni, V. Cappellini, and A. Mecocci, "Comments on "A possibilistic approach to clustering"," *IEEE Transactions on Fuzzy Systems*, vol. 4, no. 3, pp. 393–396, Aug. 1996.

[4] S. D. Xenaki, K. D. Koutroumbas, and A. A. Rontogiannis, "A Novel Adaptive Possibilistic Clustering Algorithm," *IEEE Transactions on Fuzzy Systems*, vol. 24, no. 4, pp. 791–810, Aug. 2016.

References II

[5] P. Hou, H. Deng, J. Yue, and S. Liu, "PCM and APCM Revisited: An Uncertainty Perspective," *arXiv:1610.08624 [cs, stat]*, Oct. 2016.

[6] L. X. Wang, "A New Look at Type-2 Fuzzy Sets and Type-2 Fuzzy Logic Systems," *IEEE Transactions on Fuzzy Systems*, vol. PP, no. 99, pp. 1–1, 2016.