Granular Signature of Data

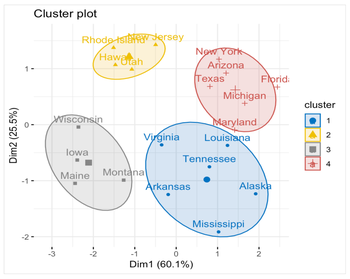
Jiahao Li[1], Witold Pedrycz[1]

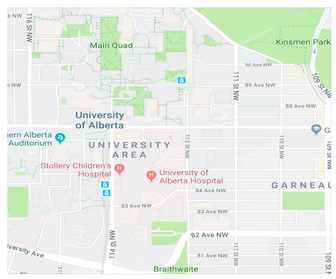
[1]Department of Electrical and Computer Engineering University of Alberta, Edmonton, AB, Canada

Introduction

Introduction

- The structure of data can be revealed in an unsupervised mode of learning, involving Fuzzy C-means Clustering. The representatives of data, produced by the clustering method, are referred to as prototypes. And the prototype would be further refined by generalizing them to information granules. This project involves detailed algorithmic development of Fuzzy C-means and formation as well as optimization of the granular signature of the real-world dataset. A graphical user interface is also developed for visualizing the performance of developed algorithm.
- Information granules form the abstract representation of the pattern for a dataset and intuitively playing an important role in human cognitive. The information granules is constructed by measuring the similarity, such as spatial or temporal similarity, among each entries in the dataset and can be interpreted under various levels of resolutions.
- The experiment shows that information granules can extract the pattern of the dataset by using Fuzzy C-means clustering and the Principle of Justifiable Granularity. The optimized information granules can give the symbolic representations for helping users to understand the data.





Fuzzy C-means Clustering

Fuzzy C-means

Fuzzy C-means, FCM, is a clustering method which allows a data point belongs to more than one clusters. This method is to minimize the following objective function.

$$J = \sum_{i=1}^{N}\sum_{j=1}^{c} u_{ij}^{m}\|x_i - c_j\|^2, \quad 1 < m < \infty$$

The clustering results is constructed by an iterative method as following.

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

The similarity between data points is measured by Weighted Euclidean Distance. Weighted Euclidean distance can equalize the effect of each feature on similarity measurement.

$$\|x - u\|^2 = \sum_{i=1}^{d} \frac{(x_i - u_i)^2}{\sigma_i^2}$$

Formation of Information Granules

The Principle of Justifiable Granularity

Information granules involve the data points which is covered by the 'radius' of those granules. Constructing information granules is maximizing the objective function to get optimal coverage value, represented by ρ_{opt}, where d is the dimensionality of data.

$$V = cov * sp$$

$$cov = \sum_{x_k : \|x_k - v_i\|^2 \le d\rho} u_{ik}$$

$$sp = (1 - \rho)^{\beta}$$

$$\rho_{opt} = argMax_{\rho}\, V$$

The similarity between clusters is measured by

$$sim(A_i, A_j) = \frac{1}{p}\sum_{k \in R_{ij}}\left(1 - \frac{|i_k - j_k|}{c_k}\right),$$

where R_{ij} is a set of features for which A_i and A_j overlap, p is the number of features and c_k is the number of prototypes projected on k-th features.

Experimental Results

Experimental Results

The objective of this project is to apply Fuzzy C-Means Clustering and construct information granules on high-dimensional data analysis. And, in this project, we have applied our developed algorithm on Boston Housing datasets from UCI Machine Learning Repository.

Boston Housing dataset records the house price values in suburb of Boston with 13 others attribute values, such as per capital crime rate and pupil-teacher ratio. The experimental results with c=6 are presented following.

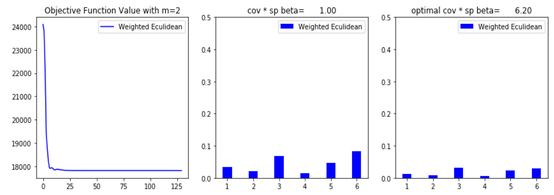


Figure 1 Objective function value of FCM and Granularity value for information granules

Info. Granules (m=2, beta=1)	1	2	3	4	5	6
V	26.46	18.95	47.61	12.92	37.99	56.55
ρ_{opt}	0.35	0.45	0.275	0.425	0.375	0.25

Table 1 Granularity value and optimal coverage for clusters

House Prices (mean)	1	2	3	4	5	6
Beta=1	24.29	22.78	22.64	22.16	20.04	23.79
Opt. Beta	26.12	25.41	22.82	19.76	16.3	24.76

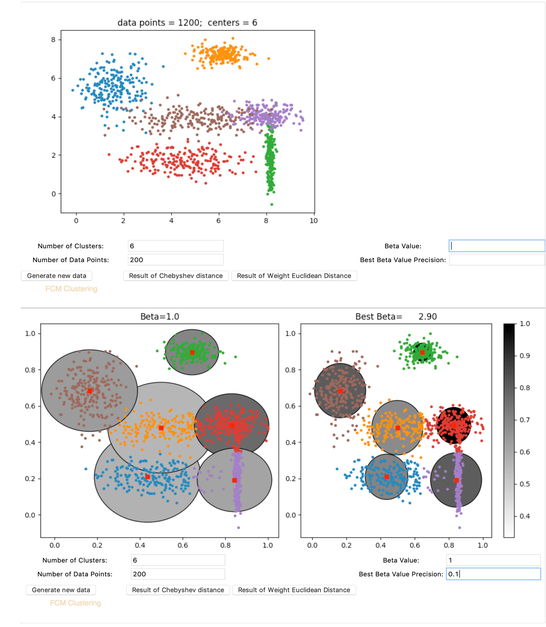
Table 2 The mean value of house price before and after optimization for information granules

User-centric Interafce

Graphical User Interface

The Graphical User Interface is developed to visualize the validity of our developing algorithm for Fuzzy C-means Clustering and the construction of Information Granules.

This GUI is developed on Python's de-facto standard GUI package, Tkinter. This is an user-centric interface, which allows user to generate synthetic data and select the tunable parameters of the algorithms manually.



Conclusion

Conclusion

- The Fuzzy C-mean clustering can generate effective data for the construction of information granules
- The information granules can extract the pattern and knowledges from the dataset
- The Principle of Justifiable Granularity can decrease the similarity among different information granules and differentiate the representation of each information granule.