



研讨 / 分享 / 感悟

仓颉社区中的灵感碰撞 社区先行者的干货分享

仓颉模糊聚类库FCM

主讲人：杨海龙

- 模糊集
- 模糊聚类
- 仓颉版模糊聚类库
- 应用场景



模糊集



Lotfi A. Zadeh

•Lütfi Rəhim oğlu

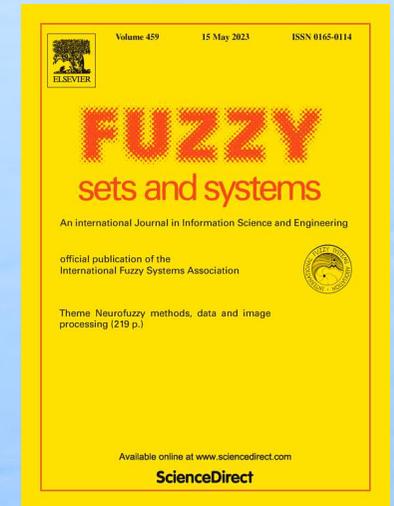
Ələsgərzadə

لطفی علی عسکرزاده

拉特飞·扎德 (Lotfi Zadeh, 1921年2月-2017年9月6日), 男, 生于现阿塞拜疆的首都巴库。1965年, 美国加利福尼亚大学控制论专家扎德(L. A. Zadeh)教授在《信息与控制》杂志上发表了一篇开创性论文《模糊集合》(Fuzzy Sets), 这标志着模糊数学的诞生. 扎德是世界公认的系统理论及其应用领域最有贡献的人之一, 被誉为“模糊集之父”。

在人类社会和各个科学领域中, 人们所遇到的各种量大体上可以分成两大类: 确定性的与不确定性的, 而不确定性又可分为随机性和模糊性. 人们正是用三种数学来分别研究客观世界中不同的量, 即:

- ✓ 第一类是**确定性数学模型**. 这类模型研究的对象具有确定性, 对象之间具有必然的关系, 最典型的就是用微分法、微分方程、差分方程所建立的数学模型.
- ✓ 第二类是**随机性数学模型**. 这类模型研究的对象具有随机性, 对象之间具有偶然的的关系, 如用概率分布方法、马尔可夫 (Markov) 链所建立的数学模型.
- ✓ 第三类是**模糊性数学模型**. 这类模型所研究的对象与对象之间的关系具有模糊性.



WORKSHOP

随机性与模糊性

随机性与模糊性是数学中遇到的两个概念，两者既有相同点也有区别。

相同点：

1. 都研究不确定现象；
2. 不确定性的度量（隶属度与概率）均在 $[0, 1]$ 取值。

不同点：

1. 随机性：**与必然性相对**。所属领域：随机数学。**指事件本身的定义是确定的，但事件是否发生不确定。**

例如：抛一枚质地均匀的硬币，正常情况下，要么正面朝上，要么反面朝上，只有这两种结果，所以这是结果本身是精确的，只是哪种结果会发生，这是不确定的。因此，这种情况下，认定抛硬币具有随机性。

2. 模糊性：**与精确性相对**。所属领域：模糊数学。**指对事件本身的认知不确定。**常见词汇：偏大、稍大、偏小、很小...

例如：听德云社相声专场，好多人都在抢票，快抢不上了！好多人抢票，多少算多？100人？1000人？10000人？“好多人”这个词具有模糊性。

因此，总结一下，随机性与模糊性都是用来描述不确定性，但随机性是事件是否发生，我们无法预见，而模糊性则是带有很强的主观色彩，无法用一个准确的标准去衡量。



WORKSHOP

模糊聚类

硬聚类 (hard clustering) 是指把数据点划分到确切的某一聚类中，如K-均值聚类。

模糊聚类 (Fuzzy clustering, 属于**软聚类**, Soft clustering) 中，数据点则可能归属于不止一个聚类中。这些聚类与数据点通过一个成员水平（实际上类似于模糊集合中**隶属度**的概念）联系起来。成员水平显示了数据点与某一聚类之间的联系有多强。模糊聚类就是计算这些成员水平，按照成员水平来决定数据点属于哪一个或哪些聚类的过程。

模糊C-均值算法 (FCM, Fuzzy c-means) 是应用最为广泛的模糊聚类算法之一。

非模糊聚类（硬聚类）会将数据分到不同类别中，即每个数据**仅属于一个确定的类别**。模糊聚类会将数据点分到**多个可能的类别**中。

例如，一个苹果可以是红的或绿的（硬聚类）；一个苹果可以是红的和绿的（模糊聚类）。这个苹果可能是某种程度的红同时另一种程度的绿。与苹果是绿的而非红的 (green=1, red=0) 相比，苹果可以既绿又红 (green=0.5, red=0.5)。模糊聚类一个样本的**隶属度和为1**的，但是模糊的特殊性在于表示一个样本与所有类别都有关系。



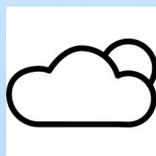
模糊聚类算法FCM, Fuzzy C-means

隶属度

例如：预测明天的天气：
晴0.9，阴0.1。隶属度是
(0.9,0.1)，也就是
明天隶属于晴天（类）的可能性
是0.9，隶属于阴天（类）是0.1



0.9



0.1

FCM是简单有效的聚类算法。

James C. Bezdek
FCM: THE FUZZY c -MEANS CLUSTERING ALGORITHM

JAMES C. BEZDEK
Mathematics Department, Utah State University, Logan, UT 84322, U.S.A.

ROBERT EHRlich
Geology Department, University of South Carolina, Columbia, SC 29208, U.S.A.

WILLIAM FULL
Geology Department, Wichita State University, Wichita, KS 67208, U.S.A.

(Received 6 May 1982; revised 16 May 1983)

1. Initialize $U=[u_{ij}]$ matrix, $U^{(0)}$ 1.初始化隶属度矩阵U
2. At k -step: calculate the centers vectors $C^{(k)}=[c_j]$ with $U^{(k)}$

$$c_j = \frac{\sum_{i=1}^N u_{ij}^m \cdot x_i}{\sum_{i=1}^N u_{ij}^m}$$

2.计算类中心

3. Update $U^{(k)}, U^{(k+1)}$ 3. 计算隶属度矩阵U

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

4. If $\|U^{(k+1)} - U^{(k)}\| < \varepsilon$ then STOP; otherwise return to step 2.

仓颉版模糊聚类库

<https://gitee.com/HW-PLLab/fcm>  欢迎star

我和课题组张启贤 (<https://gitee.com/CrazyMR>) 同学合作开发仓颉FCM库，欢迎试用和star!

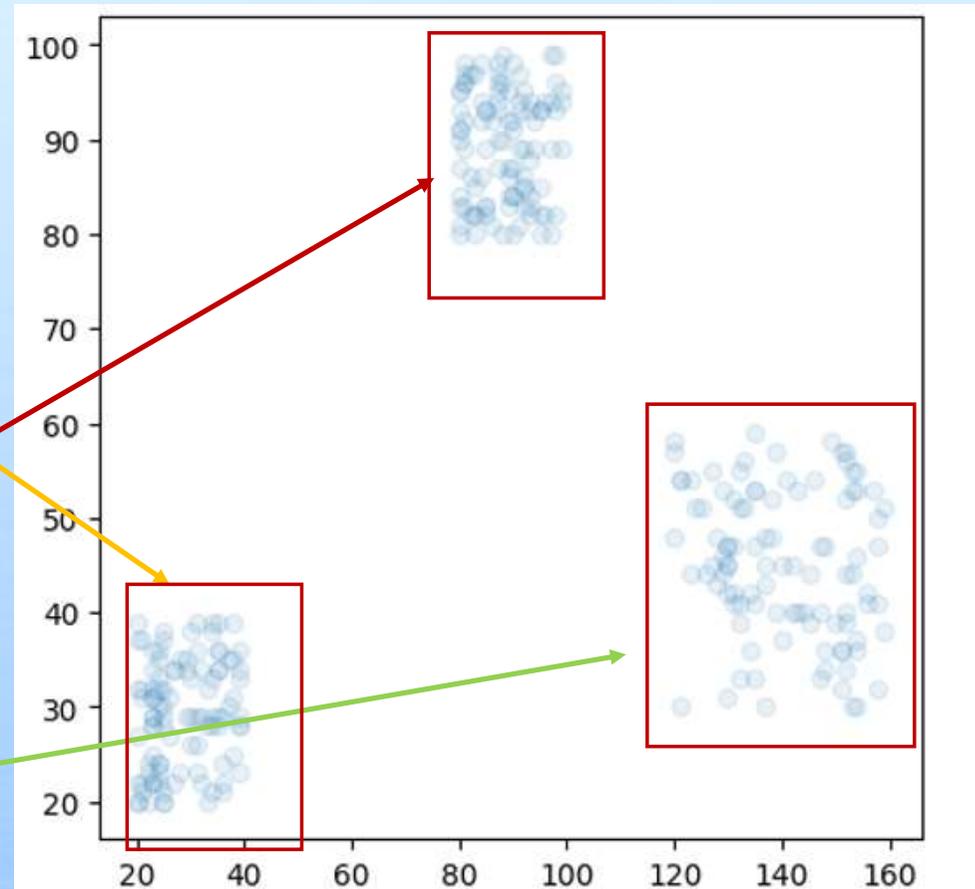
```
var fcm:FCM = FCM("RGB", 3, data, 2.0, 1000)
var centers = fcm.cluster()
var sample_with_label = fcm.defuzzify()
```

fcm库中主要是一个FCM的类，包含聚类cluster和去模糊defuzzify成员函数。

- ✓ `FCM("RGB", 3, data, 2.0, 1000)`
 - ✓ RGB是模糊聚类器的名称，
 - ✓ 3是想要聚成几类。
 - ✓ 2.0是模糊因子m，一般设置为2。
 - ✓ 1000是聚类迭代次数。
- ✓ cluster是在输入数据集没有标签的情况下无监督的学习，最后得到聚类中心。
- ✓ defuzzify是将模糊隶属度经过去模糊化得到清晰的标签值。

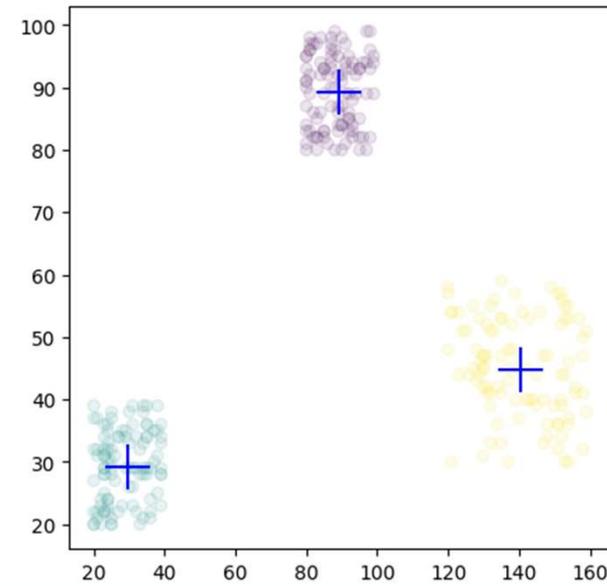
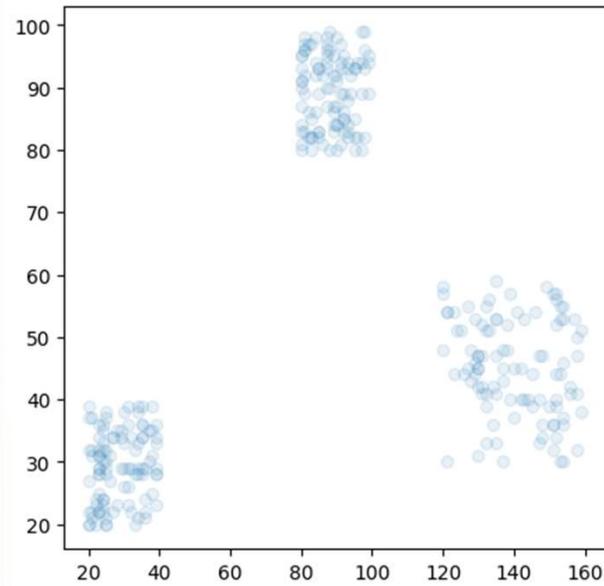
FCM库测试-准备数据集

```
public func prepareData(){
    var data: Array<Array<Float64>> =
    Array<Array<Float64>>(300,
    item:Array<Float64>(2, item:0.0))
    for (i in 0..100){
        let x = getData(20,40)
        let y = getData(20,40)
        data[i]= [x,y]
    }
    for (i in 100..200) {
        let x = getData(80,100)
        let y = getData(80,100)
        data[i]= [x,y]
    }
    for (i in 200..300) {
        let x = getData(120,160)
        let y = getData(30,60)
        data[i]= [x,y]
    }
    return data
}
```



FCM库测试

```
main(){  
    println("prepare sample data...")  
    var data:Array<Array<Float64>> = prepareData()  
    println("FCM clustering...")  
    var fcm:FCM = FCM("RGB", 3, data, 2.0, 1000)  
    var centers = fcm.cluster()  
    println("final centers:${centers}")  
    var sample_with_label = fcm.defuzzify()  
    for(i in 0..sample_with_label.size){  
        println("sample:[${sample_with_label[i][0]}  
            ${sample_with_label[i][1]}],  
            label:${sample_with_label[i][2]}")  
    }  
}
```

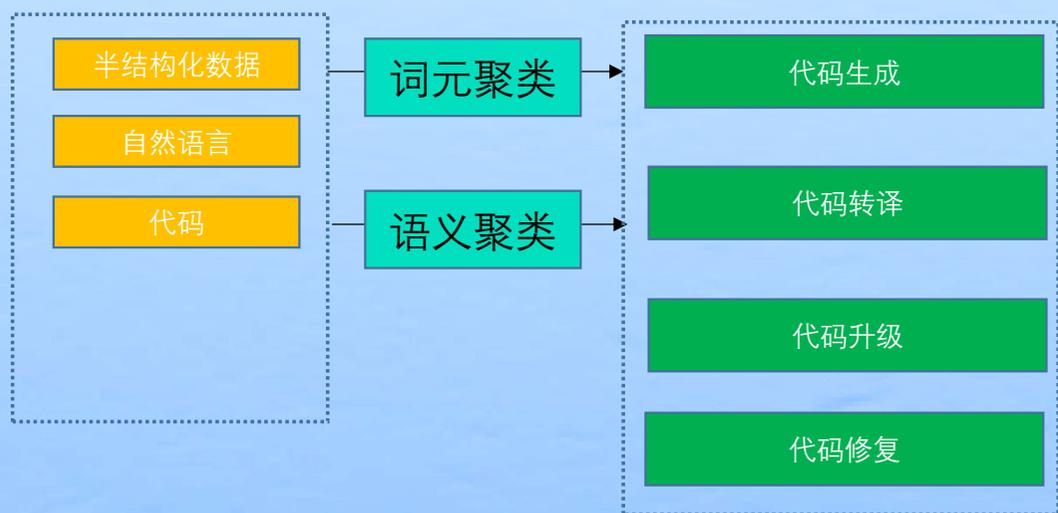


```
yang@fc25aebb7afd:~/gitee/fcm$ cpm build  
cpm build success  
yang@fc25aebb7afd:~/gitee/fcm$ ./build/bin/main  
prepare sample data...  
FCM clustering...  
init cluster_centers:[[27.000000, 21.000000], [21.000000, 34.000000], [25.000000, 36.000000]]  
name      :RGB  
cluster_num:3  
cluster_dim:2  
sample_num :3  
sample_dim :2  
final centers:[[28.022047, 28.925423], [89.965421, 90.600334], [140.205543, 44.306268]]  
sample:[27.000000 , 21.000000], label:0  
sample:[21.000000 , 34.000000], label:0  
sample:[25.000000 , 36.000000], label:0  
sample:[39.000000 , 26.000000], label:0  
sample:[36.000000 , 32.000000], label:0  
sample:[25.000000 , 25.000000], label:0  
sample:[29.000000 , 27.000000], label:0
```

FCM应用场景

模糊聚类的典型应用场景：

- 1、图像分割，像素点作为特征聚类。
- 2、为模糊系统建模提供前件。
- 3、金融与医疗的大数据挖掘与分析。
- 4、新冠疫情人群聚类分析。
- 5、在股票投资中的应用



CangJie Evaluation Dataset for Code Generation Model

<https://gitee.com/HW-PLLab/cje>

WORKSHOP



研讨 / 分享 / 感悟

仓颉社区中的灵感碰撞 社区先行者的干货分享

Thanks!



添加“编程语言小助手”，
获取仓颉相关资讯!

