Beijing Forest Studio
北京理工大学信息系统及安全对抗实验中心

聚类知识及其初始化问题

秦枭喃　　硕士

2019年08月25日

内容提要

- 背景简介
- 基本概念
- 算法原理
- 优劣分析
- 应用总结
- 参考文献

预期收获

- 预期收获
  - 1. 了解聚类算法的基础知识
  - 2. 了解概率生成模型的相关知识
  - 3. 理解概率模型下对初始化参数设置敏感问题的解决办法

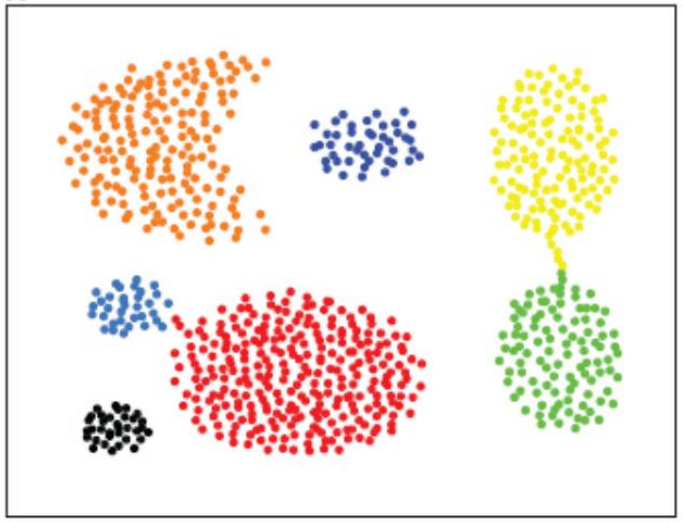
背景简介

- 聚类
  - 聚类分析是现代科学研究的常用方法，也是从大数据中挖掘价值的一项重要技术，目前已经广泛的应用在模式识别，机器学习，图像分析和生物医学等领域。

- 聚类模型的分类
  - 基于区域划分的聚类：K-means，K-medoids
  - 基于密度的聚类：DBSCAN
  - 基于层次的聚类：AGNES
  - 基于网格的聚类方法：STING，CLIQUE
  - 基于概率模型的聚类：GMM，TMM
  - 基于神经网络的聚类：SOM，深度嵌入式聚类

基本概念

- 聚类
  - 定义：通过依据指定的相似度准则将数据划分为类似对象的簇，用以探索数据中的隐藏特性以用于决策或特定应用的方法。
  - 作用：一个是可作为一个单独过程，用于找寻数据内在的分布结构，另一个是作为分类等其他学习任务的前驱过程。



基本概念

- 基本问题
  - 距离度量、性能度量
- 距离度量
  - 闵可夫斯基距离（Minkowski distance）

$$\mathrm{dist}_{\mathrm{mk}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left(\sum_{u=1}^{n} |x_{iu} - x_{ju}|^p \right)^{\frac{1}{p}}$$

  - 当p=2时，闵可夫斯基距离即欧式距离

$$\mathrm{dist}_{\mathrm{ed}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = ||\boldsymbol{x}_i - \boldsymbol{x}_j||_2 = \sqrt{\sum_{u=1}^{n} |x_{iu} - x_{ju}|^2} .$$

  - 当p=1时，闵可夫斯基距离即曼哈顿距离

$$\mathrm{dist}_{\mathrm{man}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = ||\boldsymbol{x}_i - \boldsymbol{x}_j||_1 = \sum_{u=1}^{n} |x_{iu} - x_{ju}| .$$

适用于属性为连续属性和有序的离散属性。

基本概念

- 距离度量
  - VDM(Value Difference Metric)

$$\mathrm{VDM}_p(a,b) = \sum_{i=1}^{k} \left| \frac{m_{u,a,i}}{m_{u,a}} - \frac{m_{u,b,i}}{m_{u,b}} \right|^p$$

适用于属性为无序的离散属性。

k为样本簇数，$m_{u,a}$表示在属性u为a的样本数，$m_{u,a,i}$表示在第i个样本簇中属性u为a的样本数。

样本	样本A	样本B	样本C	样本D	样本E
属性u	x	x	y	x	y
		簇1		簇2	

P=1

K=2

$m_{u,x}$=3, $m_{u,y} = 2$

$m_{u,x,1} = 2, m_{u,x,2} = 1$

$m_{u,y,1} = 1, m_{u,y,2} = 1$

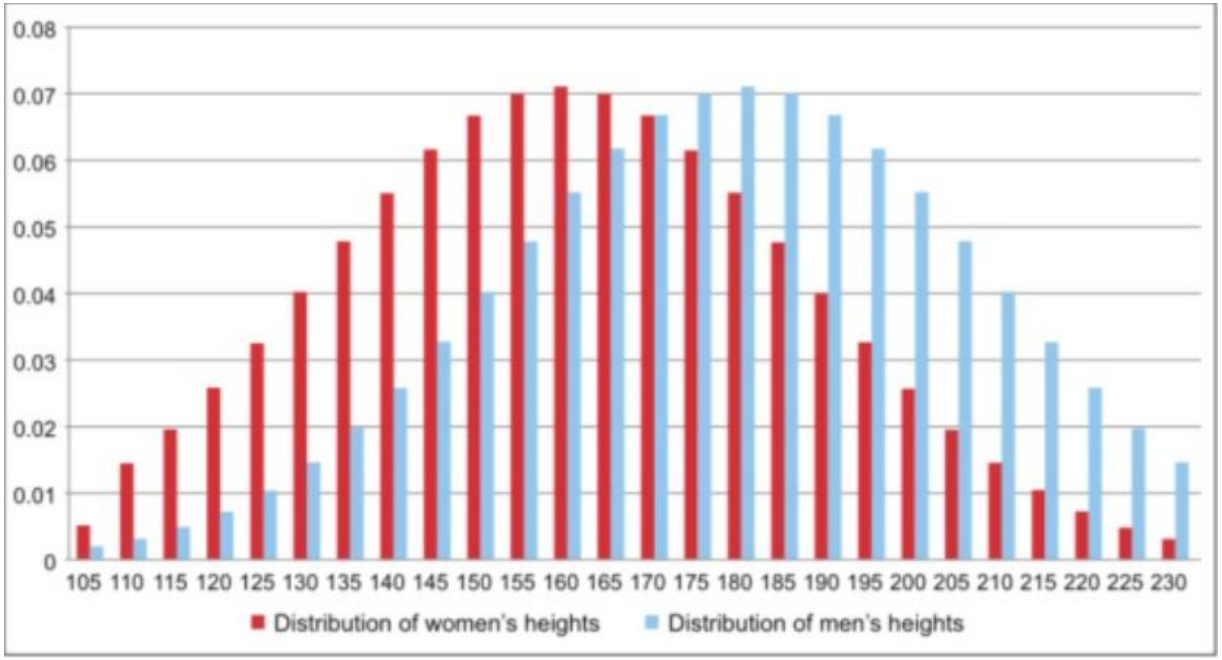
计算属性u中“x”和“y”的VDM:|2/3-1/2|+|1/3-1/2|=1/3

基本概念

- 性能度量
  - 目标：聚类结果“簇内相似度”高且“簇间相似度”低。
  - 聚类性能度量分为两类，一类是“外部指标”，另一类是“内部指标”。
    - “外部指标”指聚类结果与某个“参考模型”进行比较
      - Jaccard系数（Jaccard Coefficient，简称JC）
      - FM指数（Fowlkes and Mallows Index，简称FMI）
      - Rand指数（Rand Index，简称RI）
    - “内部指标”指直接考察聚类结果而不利用任何参考模型
      - DB指数（Davies-Bouldin Index，简称DBI）
      - Dunn指数（Dunn Index，简称DI）

基本概念

- 基于概率模型的聚类
  - 基于概率模型聚类方法是通过假设数据集遵循某种设定概率分布以进行统计建模。常用的模型有：高斯混合模型，混合拉普拉斯分布模型，混合t分布模型。



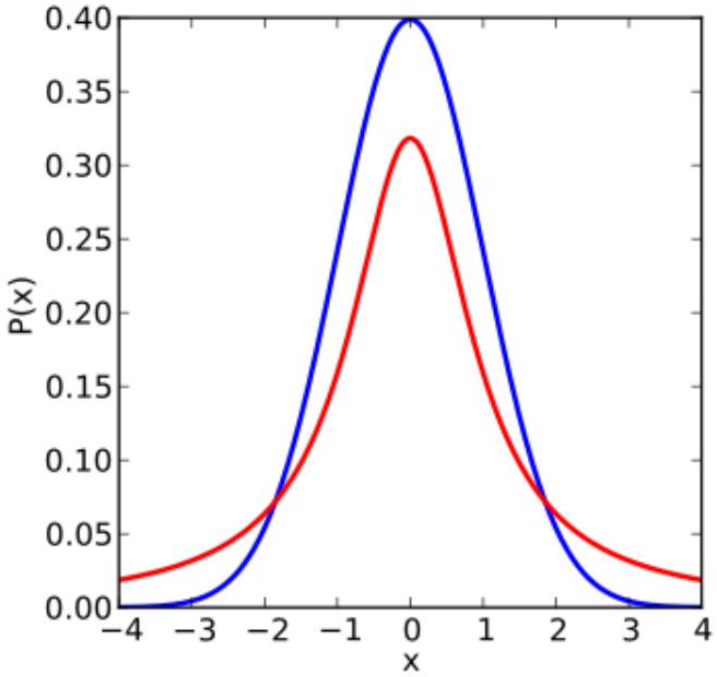
高斯混合模型及求解算法-王殿元-2019-02-17 19_00_00.pptx

基本概念

- 高斯混合模型存在的问题
  - 对离群点敏感
    - 由于高斯混合模型是薄尾分布。
  - 需要先验性的设置聚类簇数
  - 对初始参数设置敏感 ★
    - 由于EM算法为非凸优化算法
    - 初始参数为随机设定
      - 多次随机初始化取最好结果
      - 引入其他聚类算法做初始化：Kmeans++，密度测量
      - 并不需要考虑参数初始化：贪婪法，合并法，熵惩罚算法

基本概念

- 混合t分布模型（TMM）
  - 定义
    - 混合t分布模型（TMM）是混合概率模型中的一种，可以看做是高斯分布和Gamma分布混合体，由于其分布特性为重尾分布，能够对异常值有一定的稳健性，已经在多个领域得到应用。



基本概念

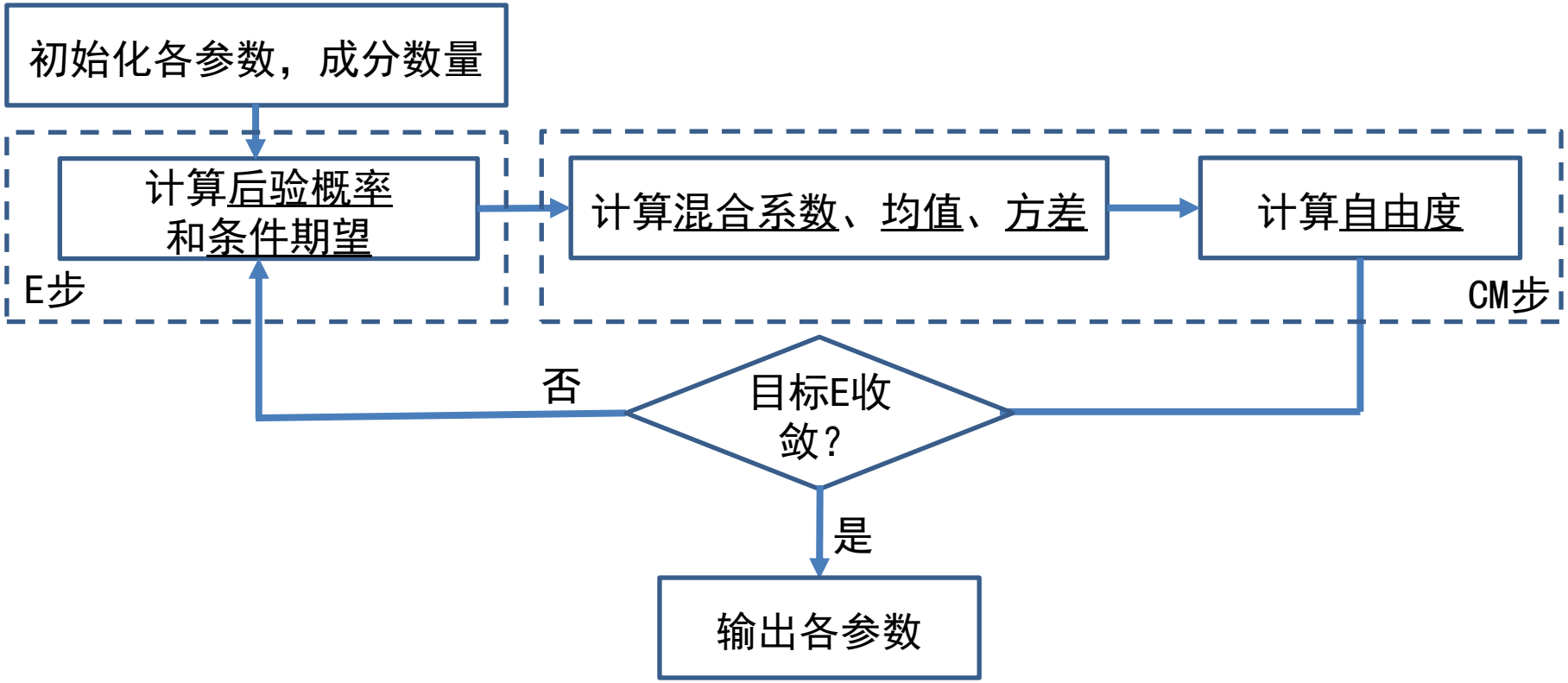
- 混合t分布模型（TMM）
  - 目标优化函数

$$E_0 = \sum_{i=1}^{n} \log f(x_i; \Theta_j) = \sum_{i=1}^{n} \log \sum_{j=1}^{G} \pi_j t(x_i; \mu_j, \Sigma_j, \nu_j)$$

  - 优化参数
    - 参数集包括四个部分：混合系数π，均值μ，协方差Σ，自由度ν
    - 采用的优化方法为ECM算法，迭代的过程分为E步和CM步
      - E步：固定混合系数、均值、方差和自由度，更新计算后验概率和权重
      - CM步：固定后验概率和权重，更新计算混合系数、均值、

基本概念

- 混合t分布模型（TMM）
  - 优化步骤



基于t分布的熵惩罚最大期望算法

T	目标	解决混合t分布对初始参数设定敏感
I	输入	带有无标注数据
P	处理	基于t分布的熵惩罚最大期望算法（EPEM）
O	输出	各成分的参数：均值、协方差和自由度

P	问题	EM算法只能保证参数估计到稳定点，使模型对初始参数设置敏感
C	条件	数据符合混合t分布
D	难点	调节混合系数的最小门限阈值以获得指定簇数
L	水平	国内先进

算法原理

- 基于t分布的熵惩罚的最大期望算法
  - 目标：为解决混合t分布模型对初始参数设置敏感的问题。
  - 方法：通过在目标函数中E中加入混合系数的惩罚项，使得在优化目标函数的迭代过程中鼓励各成分间样本的竞争，在此过程中不断的删除小于设定阈值的成分以获得最终的成分数量和初始化参数。

原目标函数：$\mathrm{E}_0 = \sum_{i=1}^{\mathrm{n}} \log \sum_{j=1}^{G} \pi_j t(x_i; \mu_j, \Sigma_j, \nu_j)$

加入熵惩罚：

- $\mathrm{E}_1 = \sum_{i=1}^{n} \log \sum_{j=1}^{G} \pi_j t\left(x_i; \mu_j, \Sigma_j, \nu_j\right) - \left(-\beta \sum_{i=1}^{n} \sum_{j=1}^{\mathrm{G}} \pi_j \mathrm{In}\left(\pi_j\right)\right)$

算法原理

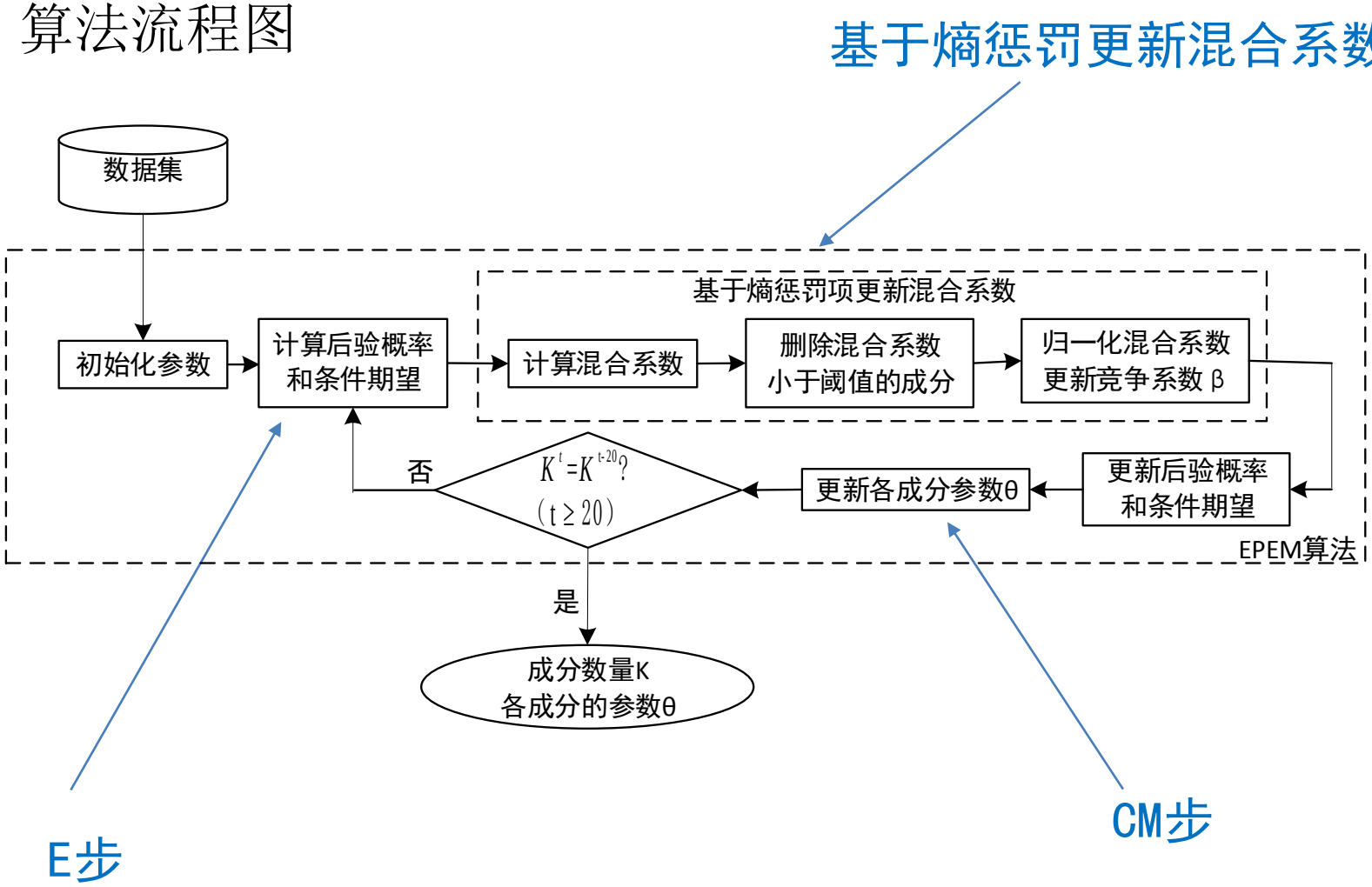
- 基于熵惩罚的最大期望EM算法
  - 目标函数

$$E_1 = \sum_{i=1}^{n} \log \sum_{j=1}^{G} \pi_j t(x_i; \mu_j, \Sigma_j, \nu_j) - (-\beta \sum_{i=1}^{n} \sum_{j=1}^{G} \pi_j \ln(\pi_j))$$

- $-\ln(\pi_j)$表示一个样本属于第j类成分的信息量
- $\sum_{j=1}^{G} -\pi_j \ln(\pi_j)$ 表示一个样本的平均信息量，*也就是信息熵*
  - *当混合系数都相同时，对应的信息熵将取得最大值，混合系数处于最不确定的状态。*
- $\sum_{i=1}^{n} \sum_{j=1}^{G} -\pi_j \ln(\pi_j)$ 表示所有样本的信息熵
- β为控制竞争程度的系数，成为竞争系数

算法原理

- 算法流程图


基于熵惩罚更新混合系数
数据集
初始化参数
计算后验概率
和条件期望
基于熵惩罚项更新混合系数
计算混合系数
删除混合系数
小于阈值的成分
归一化混合系数
更新竞争系数 β
否
$K^t=K^{t-20}$?
$(t \geq 20)$
更新各成分参数θ
更新后验概率
和条件期望
EPEM算法
是
成分数量K
各成分的参数θ
E步
CM步

算法原理

- 合成数据集实验
  - 原理：
    - 引入竞争机制，促使聚类成分在不断的减少并趋于稳定。通过调节混合系数的门限阈值可以获得指定的聚类簇数和各成分参数。
    - 在基于熵惩罚的迭代更新混合系数时，$\sum_{j=1}^{G}\pi_j \ln(\pi_j)$作为混合系数的加权求和，若$\ln(\pi_j^{(k)})$比其加权求和小，则$\pi_j^{(k+1)}$将小于$\pi_j^{(k)}$。因此混合系数小的成分，将越来越小，而混合系数大的成分将越来越大。
  - 初始化参数：β=1, k=50, iteration=100, $\mu=\frac{1}{k}, \Sigma=I_d$, ν=100
  - 合成数据集设置：

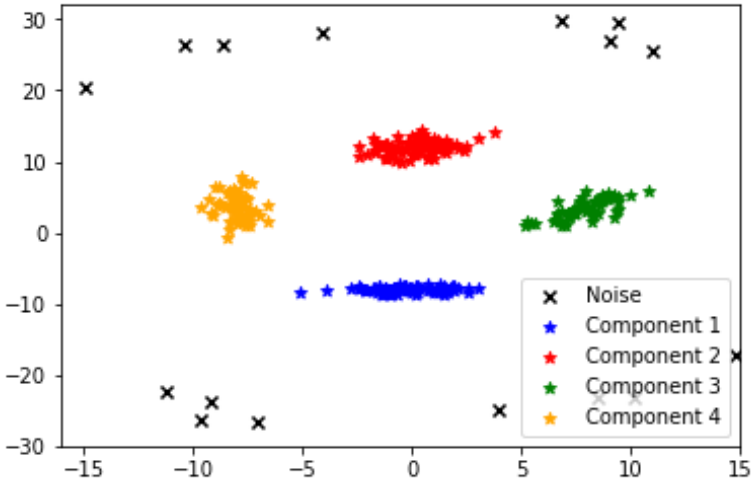
    其数据总量为396，各成分的参数如下：

    $N_1=80, N_2=80, N_3=60, N_4=60, N_{noise}=16$

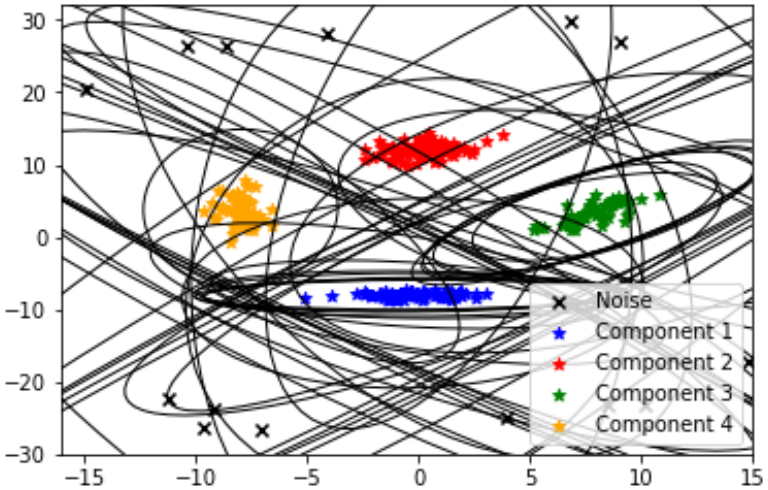
    $\mu_1=[0,-8]^T, \mu_2=[0,12]^T, \mu_3=[8,3.5]^T, \mu_4=[-8,3.5]^T$

    $[3 \quad 0.1 \qquad [0.7 \quad 1.1 \qquad [0.8 \quad 0.1 \qquad [0.5 \quad 0]$

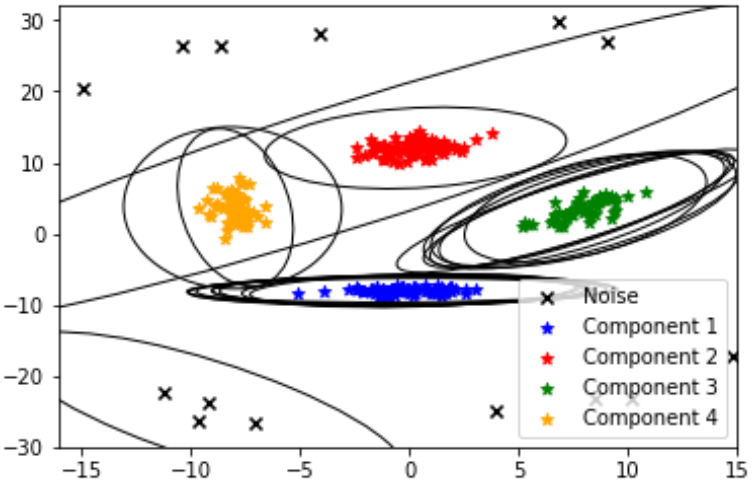
算法原理



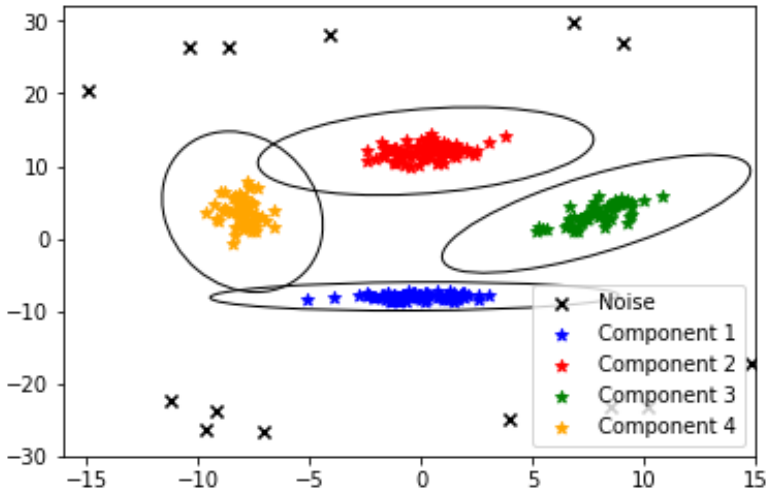
合成数据集



迭代次数5，簇数40



迭代次数10，簇数22



迭代次数15，簇数04

优劣分析

- 横向对比
  - 相比于Kmeans等区域划分算法，该算法采用的是概率密度模型，适用于拟合具有重叠簇的数据。
- 纵向对比
  - 该EPEM算法能够有效解决对初始参数的设置敏感的问题，其聚类结果更加稳定。相较于用Kmeans++做初始化的方法，效果更好。
  - 该算法是基于混合t分布模型构建的，能够对离群点具有较好的稳健性。

应用总结

- 算法的应用领域
  - 基于用户位置信息的商业选址
    - 如百度与万达进行合作，通过定位用户的位置，结合万达的商户信息，向用户推送位置营销服务，提升商户效益。
  - 群体用户画像分析
    - 群体用户画像分析旨在依据不同的评估维度和模型算法，通过聚类方式将具有相同特征的用户划归成同一个族群，进而发现核心的、规模较大的用户群，从而在设计推荐系统时考虑优先满足核心用户群的需求，进一步在不存在冲突的情况下尽量满足次要用户群的需求。
- 未来的发展
  - 目前聚类模型众多，如何将算法模型与实际应用结合，创造商业价值是未来发展的重点。

参考文献

【1】Miin-Shen Yang,Chien-Yo Lai, Chih-Ying Lin. A robust EM clustering algorithm for Gaussian mixture models[C]. Pattern Recognition, 2012:3950-3961.

【2】 https://my.oschina.net/ydsakyclguozi/blog/2992433

谢谢！

大成若缺，其用不弊。大盈若冲，其用不穷。大直若屈。大巧若拙。大辩若讷。静胜躁，寒胜热。清静为天下正。