机器学习学习报告

周珊琳

上海电力大学
上海, 2019-07-17, 中国

第一节　算法概述

算法名称：BIRCH

算法原理：一般的层次聚类算法分为自顶向下的分裂层次聚类和自下而上的凝聚层次聚类。但是这些层次聚类算法都存在两个问题：无法退回并更正导致低质量的聚类结果；可扩展性较差，因为合并或分裂的决定需要检查和估算大量的对象或簇。因此，有一些改进的凝聚层次聚类算法解决了以上问题，比如：BIRCH、ROCK、CURE等。其中BIRCH算法是一种基于距离的层次聚类算法，其核心是聚类特征CF(Cluster Feature)和聚类特征树（CF-Tree），它们用于概括簇描述。这些结构可以帮助聚类方法在大型数据库中取得很好的速度和伸缩性，使得BIRCH算法对于增量和动态聚类也非常的有效。

CF聚类信息：

$$CF = (N, \overrightarrow{LS}, SS) \tag{1.1}$$

$$\overrightarrow{LS} = \sum_{i=1}^{N} \overrightarrow{X_i} \tag{1.2}$$

$$SS = \sum_{i=1}^{N} \overrightarrow{X_i}^2 \tag{1.3}$$

第二节　算法设计

2.1　算法流程

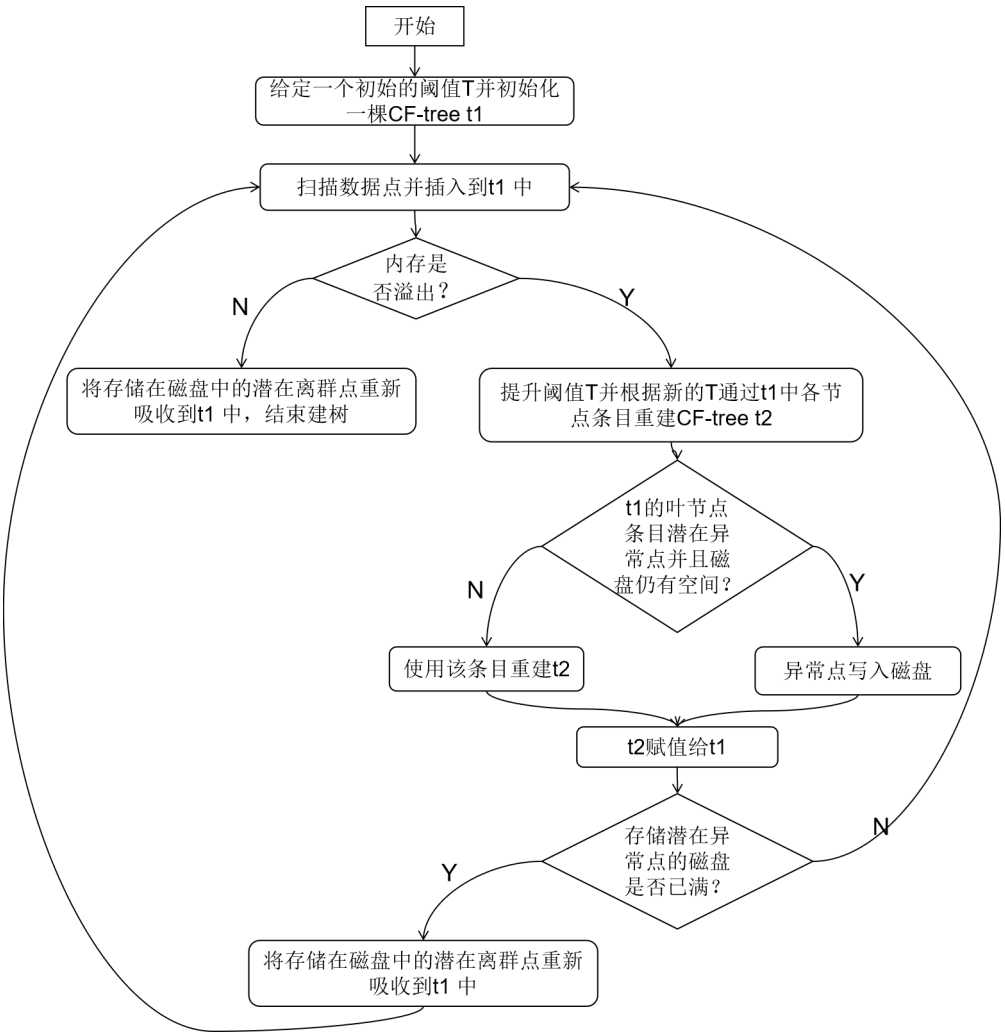

图 1: BIRCH算法流程

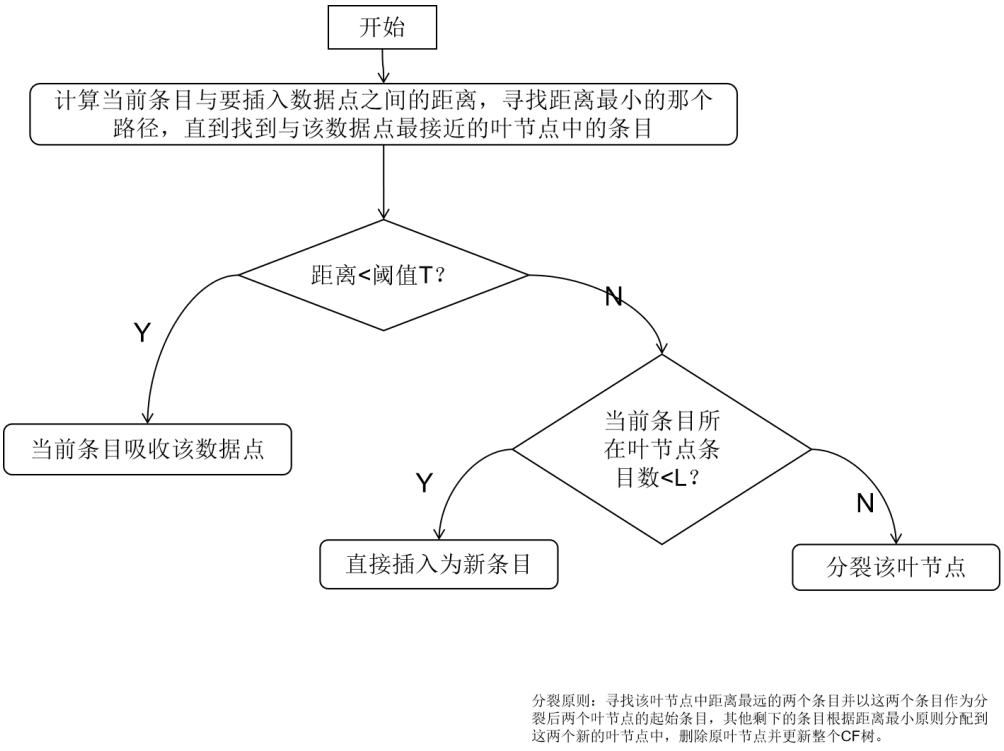

图 2: CF-tree构造流程

2.2 核心代码

源代码 1:

```
# -*- coding: utf-8 -*-
"""
Created on Thu Jul 18 16:34:12 2019

@author: zsl
"""

import numpy as np

import matplotlib.pyplot as plt
from sklearn.datasets import load_iris,load_wine
from sklearn.datasets.samples_generator import make_blobs
from sklearn import preprocessing
from sklearn.cluster import Birch

def process_data(data):
    min_max_scaler =preprocessing.MinMaxScaler()
    return min_max_scaler.fit_transform(dataset)


#iris


data=load_iris()
dataset =data.data
y = data.target
X=process_data(dataset)

#wine
data=load_wine()
dataset =data.data
y = data.target
X=process_data(dataset)




birch =Birch(n_clusters= None)


y_pred =birch.fit_predict(X)


plt.figure()

plt.subplot(2,2,1)

plt.scatter(X[:,0],X[:,1])

plt.title('DataSample')
```

```
51
52 plt.subplot(2,2,2)
53
54 plt.scatter(X[:,0], X[:, 1], c=y_pred)
55
56 plt.title('None')
57
58
59 birch =Birch(n_clusters =3)
60
61
62
63 y_pred =birch.fit_predict(X)
64
65 plt.subplot(2,2,3)
66
67 plt.scatter(X[:,0], X[:, 1], c=y_pred)
68
69 plt.title('n_clusters=3')
70
71 plt.show()
```

第三节　选用数据

iris行数：150 列数：5
列属性及取值:
1)萼片长度cm，数值型
2)萼片宽度cm，数值型
3)花瓣长度cm，数值型
4)花瓣宽度cm数值型
类别：
Iris Setosa
Iris Versicolour
Iris Virginica
wine，行数：178，列数：13
属性：
1) Alcohol
2) Malic acid
3) Ash
4) Alcalinity of ash
5) Magnesium
6) Total phenols
7) Flavanoids
8) Nonflavanoid phenols
9) Proanthocyanins
10) Color intensity
11) Hue
12) OD280/OD315 of diluted wines
13) Proline
类别：
Alcohol 1，2，3

第四节　实验结果展示

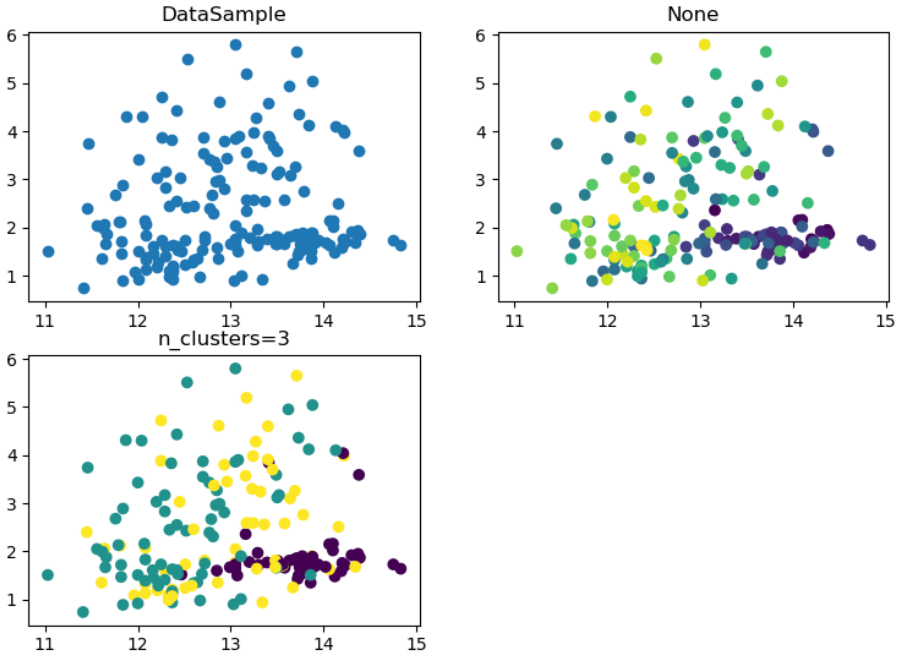

图 3: iris聚类对比图

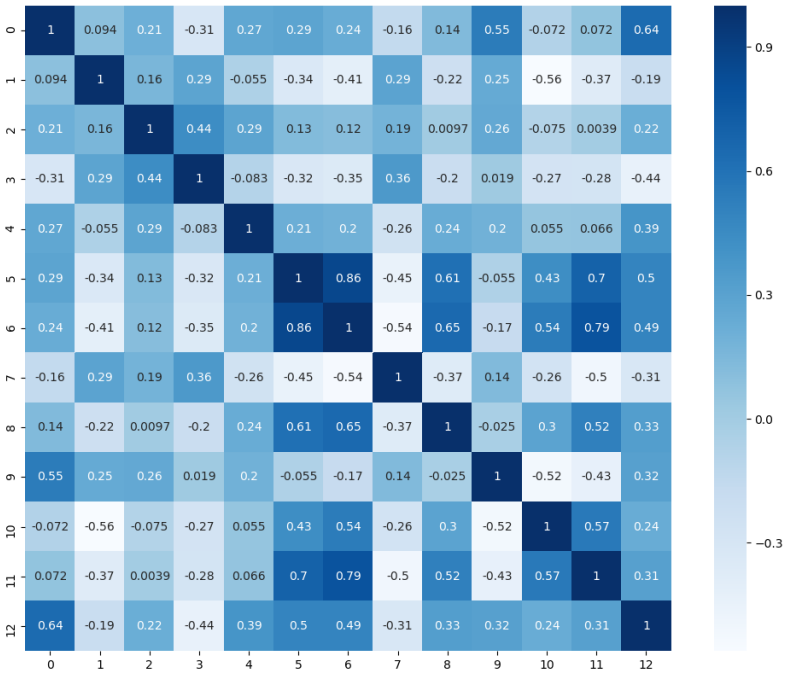

图 4: wine变量相关性热力图

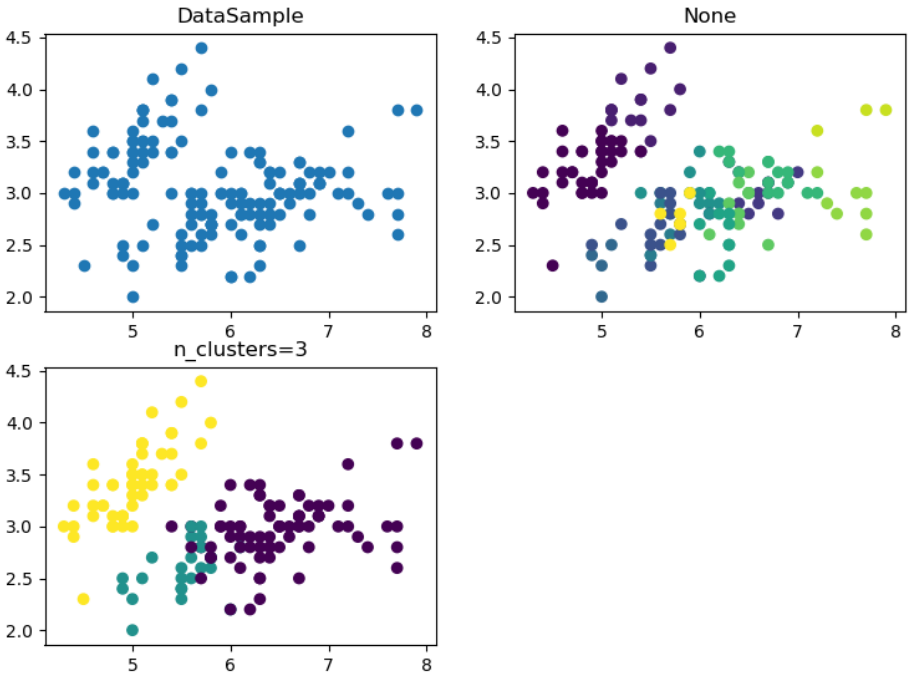

图 5: wine聚类对比图

第五节　实验分析和比较

birch需要数值型数据，对于car这个数据，不适合用这个算法。在使用wine之前，对wine进行了相关性分析，并对iris和wine数据进行了最大最小归一化处理，这样处理之后的数据，再使用聚类算法，效果明显更好。