使用 Pycluster 包进行聚类分析实例

Esri 中国 卢萌

“方以类聚，物以群分，吉凶生矣“
——周易·系辞上

人类在几千年前就认识到了所谓的聚类和分类，是用来认知和描述万事万物之间关系的主要方法。一个没读什么书小贩，也知道将不同色泽和品相的水果分开，可以卖不同的价格。所以不论是否受过高等教育，聚类和分类的思想都根深蒂固的在人类的思想中。

聚类是数据挖掘的主要手段之一，对于探索数据之间的规律有着重要的作用。但是很多想做数据分析的同学限于没有时间去写各种算法，只能停留在算法描述上面。

所以这里给大家介绍一个很好很强大的开源包：Pycluster 包。

Pycluster 包是东京大学医学研究所，人类基因研究中心的米歇尔德勋（Michiel de Hoon），星矢井本（Seiya Imoto），宫野悟（Satoru Miyano）等人编写的开源算法工具包，提供了 C/C++、python 和 Perl 三个版本，因为本人主要玩的 python，所以这里主要讲其中的 Pycluster 包，其他的内容，可以下载详细文档（本文中的代码、数据和文档，在最下面的链结中有，我放的是百度云盘）。

Pycluster 封装了基于划分的算法中的两个最经典的算法 K-means 和 k-medoids，以及基于层次的算法，主要还是说了 k-means 和 k-medoids 算法，算法的实行描述我就不详细说了，网上资料大把多。

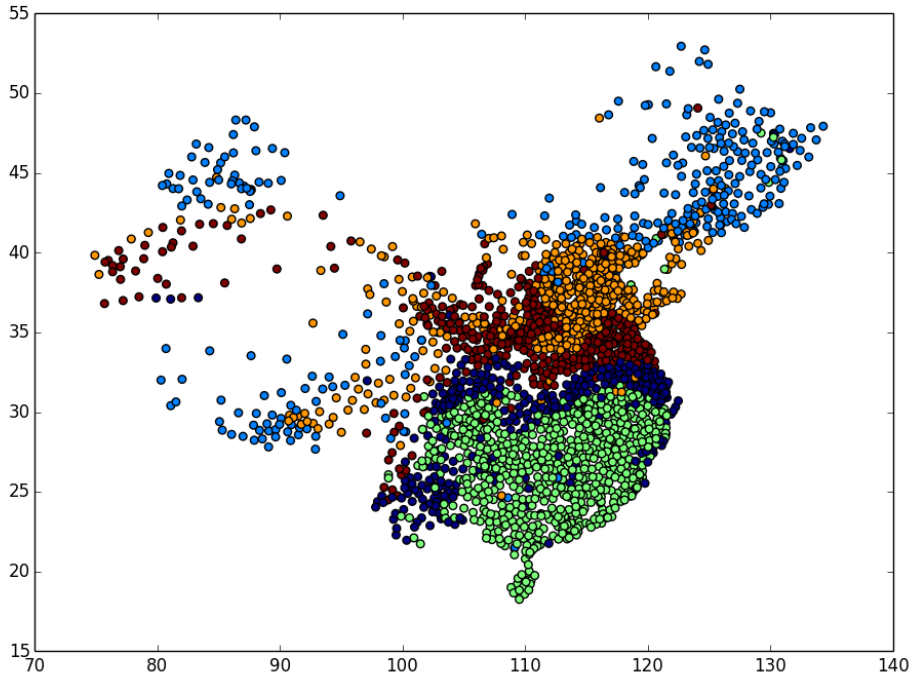
下面解析一下整个包实现的代码以及各种参数说明：其中斜体是我写的注释。

```
# -*- coding:utf-8 -*-
'''
Created on 2015-6-3

@author: godxia
'''
import Pycluster as pc
import numpy as np
```

```
import matplotlib.pylab as pl

def myCKDemo(filename,n):
    #以下两个语句是获取数据，用于聚类分析的数据位于第3 和第4 列（从0 开始计算）
    data = np.loadtxt(filename, delimiter = "," ,usecols=(3,4))
    #第8 和第9 列，保存了城市的经纬度坐标，用于最后画散点图
    xy = np.loadtxt(filename, delimiter = "," ,usecols=(8,9))
    #clustermap 是聚类之后的集合,记录每一组数据的类别 id
    clustermap = pc.kcluster(data, n)[0]
    #centroids 是分组聚类之后的聚类中心坐标
    centroids = pc.clustercentroids(data, clusterid=clustermap)[0]
    #m 是距离矩阵
    m = pc.distancematrix(data)

    #mass 用来记录各类的点的数目
    mass = np.zeros(n)
    for c in clustermap:
        mass[c] += 1



    #sil 是轮廓系统矩阵，用于记录每个簇的大小
    sil = np.zeros(n*len(data))
    sil.shape = ( len(data), n )

    for i in range( 0, len(data) ):
        for j in range( i+1, len(data) ):
            d = m[j][i]
            sil[i, clustermap[j] ] += d
            sil[j, clustermap[i] ] += d

    for i in range(0,len(data)):
        sil[i,:] /= mass

    #s 轮廓系数是一个用来评估聚类效果的参数
    #值在-1 —— 1 之间，值越大，表示效果越好。
    #小于0，说明与其簇内元素的平均距离小于最近的其他簇，表示聚类效果不好。
    #趋近与1，说明聚类效果比较好。
    s=0
    for i in range( 0, len(data) ):
        c = clustermap[i]
        a = sil[i,c]
        b = min(sil[i,range(0,c)+range(c+1,n)])
        si = (b-a)/max(b,a)
        s+=si
```

```
        print n, s/len(data)

        #使用 matplotlib 画出散点图。
        fig, ax = pl.subplots()
        #cmap 是用于区分不同类别的颜色
        cmap = pl.get_cmap('jet', n)
        cmap.set_under('gray')
        #xy 是经纬度，主要为了通过经纬度来画出不同城市在地理上的位置
        x = [list(d)[0] for d in xy]
        y = [list(d)[1] for d in xy]
        cax = ax.scatter(x, y, c=clustermap, s=30, cmap=cmap, vmin=0, vmax=n)
        pl.show()

if __name__ == '__main__':
        #filename 是数据 c2.txt 所在的路径，改成自己机器上的路径即可
        filename = r"e:\c2.txt"
        #n 是预设分成几类。
        n = 5
        myCKDemo(filename,n)
```

最终计算的结果如下：（当然，你运行之后显示的颜色可能和我这里不同，因为不同的类别颜色画出来的时候，是随机的）。



可以很明显的看出长江气候带，秦岭-淮河气候带、天山南北麓气候带等，通

过温度属性进行聚类，空间位置不参与计算，能够很明显的划分出中国的气候带分布情况，说明聚类本身对于数据之间的关系模式的探索效果是非常显著。

最后，如果你复制的代码无法运行，可能是因为格式问题，因为 python 是严格缩进的，你可以下载我下面共享的源代码文件。注意要先安装 Pycluster 包和 matplotlib 包。

其中：c2.txt 是数据，cluster.pdf 是官方文档，kclusterDemo.py 是我的 python 代码源文件（我用的 utf-8 编码的），Pycluster-1.52.win32-py2.7.exe 是我使用的 Pycluster 包的版本，我用的 python2.7 如果你用的其他版本，可以去在以下地址下载对应你的 python 版本的包：

http://bonsai.hgc.jp/~mdehoon/software/cluster/software.htm#pycluster

最后，云盘下载地址如下：

链接：http://pan.baidu.com/s/1npMkm 密码：0b2l