MATLAB聚类有效性评价指标（内部）

作者：凯鲁嘎吉 - 博客园 http://www.cnblogs.com/kailugaji/

外部评价指标(需要真实标签)，请看：MATLAB聚类有效性评价指标（外部），MATLAB聚类有效性评价指标（外部 成对度量）

MATLAB中有一个聚类内部评价指标的函数evalclusters()，即不需要知道聚类中数据的真实标签，也可以评价聚类的好坏。

若事先不知道数据的真实类别信息，使用内部评价指标；若数据的真实标签已知，用外部评价指标。

关于函数的官网，请看：https://www.mathworks.com/help/stats/evalclusters.html

1．evalclusters()函数用法

- eva= evalclusters(data,clust,'xxx');
- data：可以是归一化/标准化之后的。
- clust：聚类后的标签，参数可以是字符串或者矩阵，字符串了话就要写matlab规定的聚类算法，比如'kmeans'。如果你用的不是matlab自带的聚类算法去聚类的，而是想评估自己写的聚类算法性能，那么clust就填写你聚类结果的标签矩阵Label就行了，比如你的数据data是N*M的，N是样本个数，M是特征维数。那么标签矩阵Label就是N*1的矩阵，代表聚类后每个样本的标签，这样就可以输出评估结果了。
- 'xxx'：可以选择四种评价指标：'CalinskiHarabasz' | 'DaviesBouldin' | 'gap' | 'silhouette'
  - 'CalinskiHarabasz': Calinski-Harabasz index (CHI，越大越好)
  - 'DaviesBouldin': Davies-Bouldin index(用的最多，越小越好)
  - 'gap': gap statistic (越大越好)
  - 'silhouette': silhouette coefficient (SC，轮廓系数，越大越好)

2．MATLAB程序

实验数据来自：https://www.cnblogs.com/kailugaji/p/10861064.html#_label3_0_1_2，将生成的三维数据存为data.txt。

```
clear
clc
```

```
% 作者：凯鲁嘎吉 https://www.cnblogs.com/kailugaji/
% 用了SC与DBI这两个指标
% label：聚类后的标签，不是真实标签
data_load = dlmread("data.txt");
data = data_load(:, 1:end-1);
label = data_load(:, end); % 真实标签
label_2 = [ones(180, 1); 2.*ones(120, 1)]; % 瞎写的标签
%% 真实标签的评价结果
eva_DBI_1= evalclusters(data,label,'DaviesBouldin');
eva_SC_1= evalclusters(data,label,'silhouette');
DBI_1 = eva_DBI_1.CriterionValues; % 越小越好
SC_1 = eva_SC_1.CriterionValues; % 越大越好
%% 瞎写的标签的评价结果
eva_DBI_2= evalclusters(data,label_2,'DaviesBouldin');
eva_SC_2= evalclusters(data,label_2,'silhouette');
DBI_2 = eva_DBI_2.CriterionValues; % 越小越好
SC_2 = eva_SC_2.CriterionValues; % 越大越好
%% 显示评价结果
fprintf("DBI越小越好：DBI(真实标签): %f, DBI(坏标签): %f\n", DBI_1, DBI_2); % 越小越好
fprintf("SC越大越好：SC(真实标签): %f, SC(坏标签): %f\n", SC_1, SC_2); % 越大越好
```

3. 结果

```
DBI越小越好：DBI(真实标签): 0.892058, DBI(坏标签): 0.938236
SC越大越好：SC(真实标签): 0.599176, SC(坏标签): 0.574342
```

4. 参考

[1] https://www.mathworks.com/help/stats/evalclusters.html

[2] 不调用现成的matlab函数，而是别人动手写的评价指标函数：https://github.com/adanjoga/cvik-toolbox