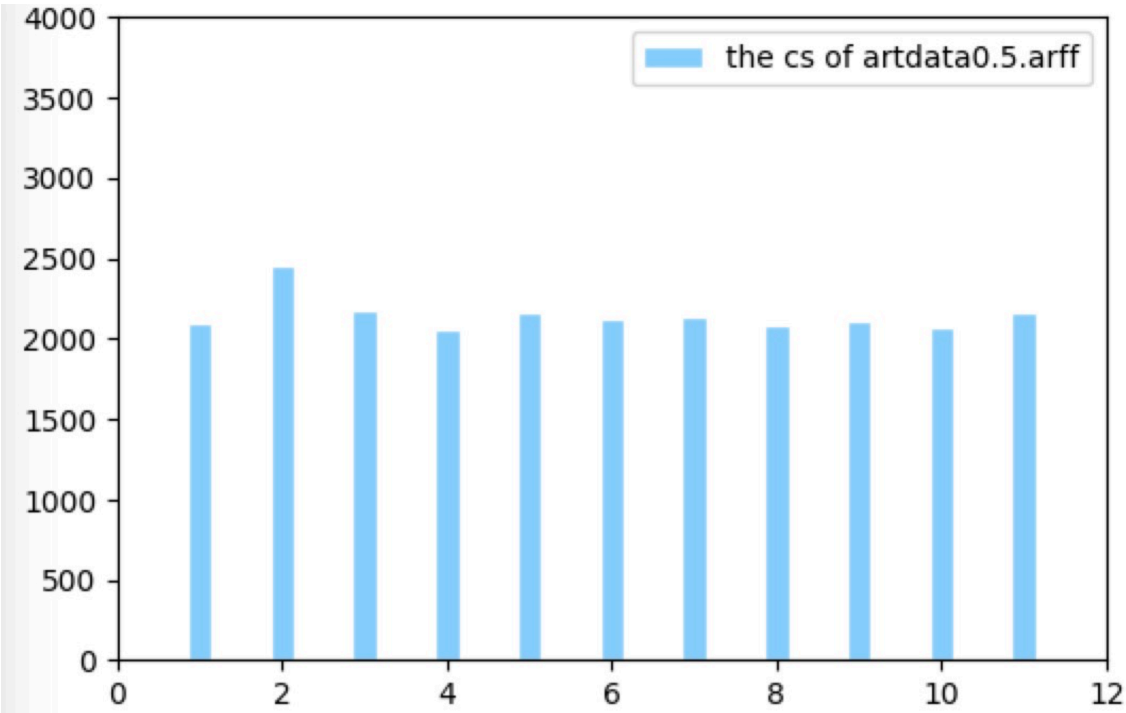
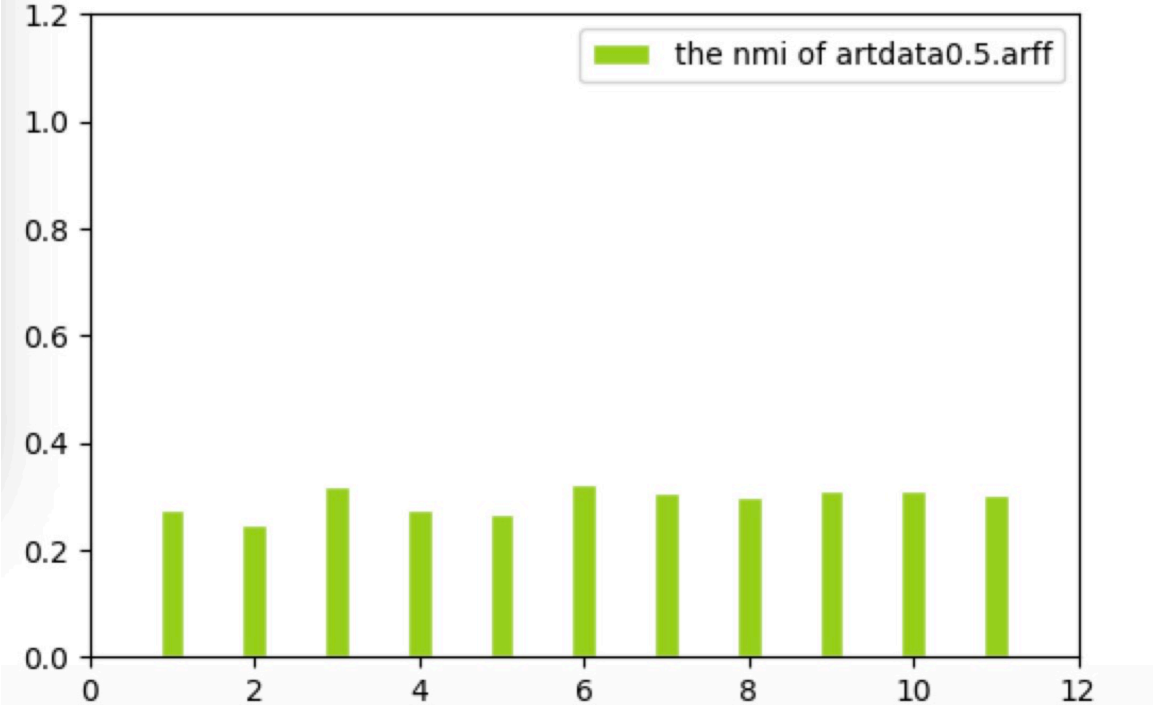
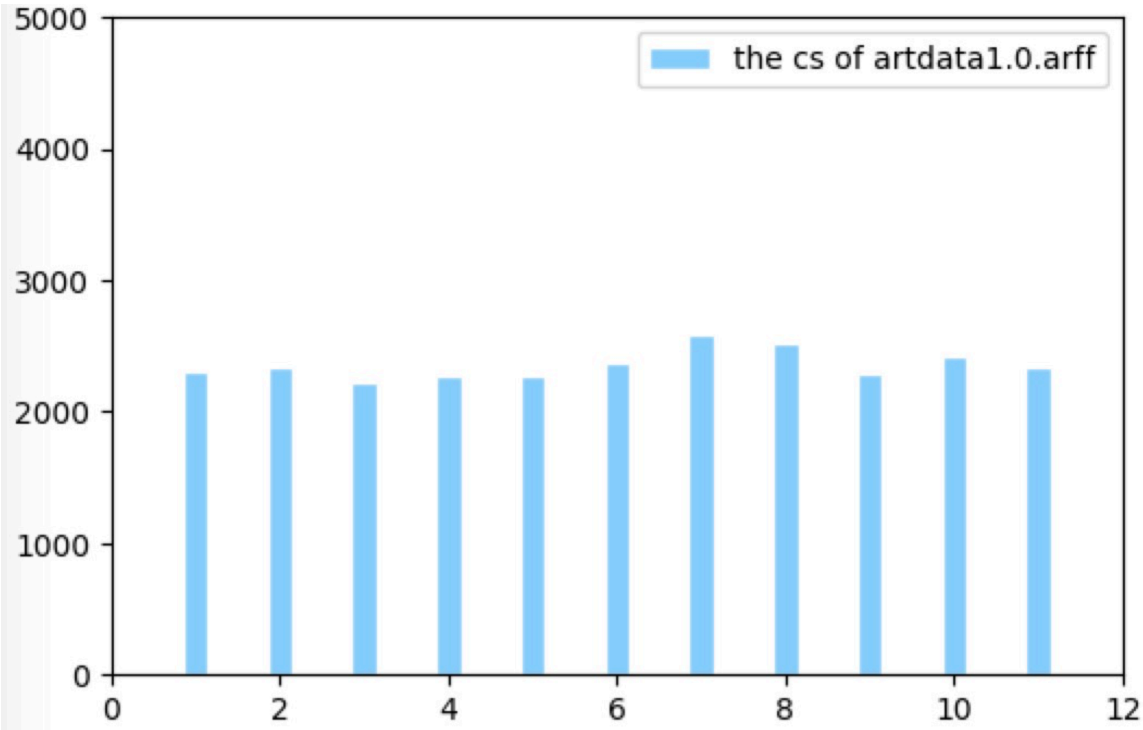
Project 3

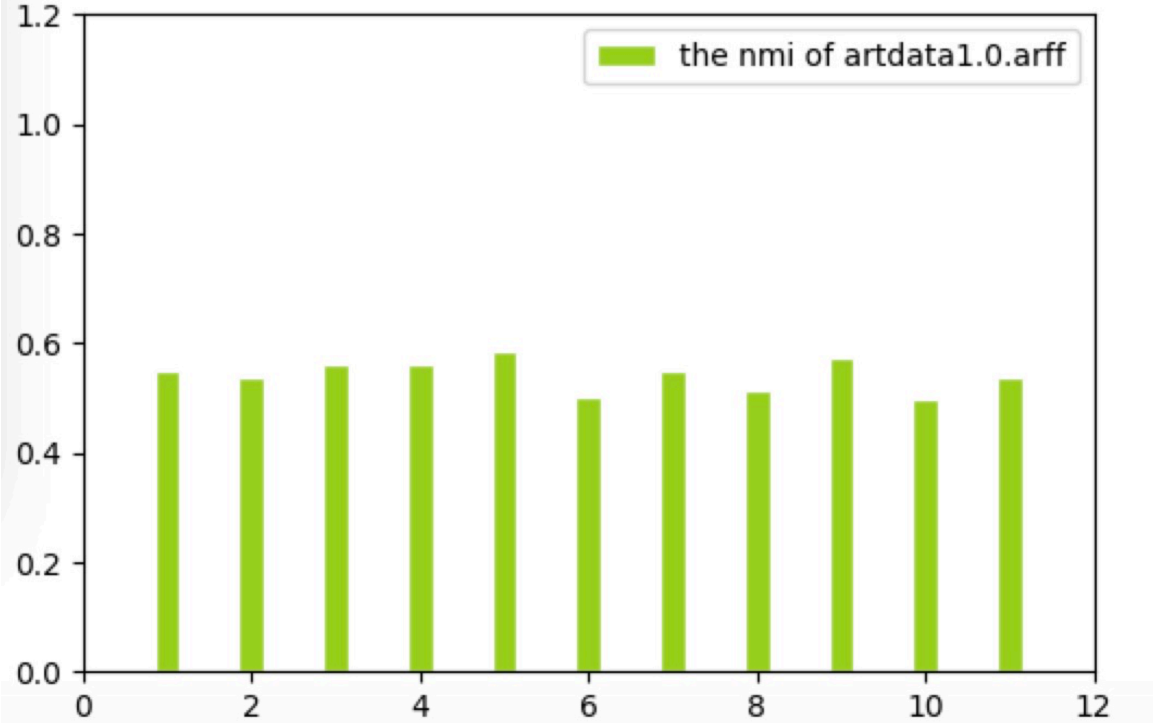
ZHEYI YI

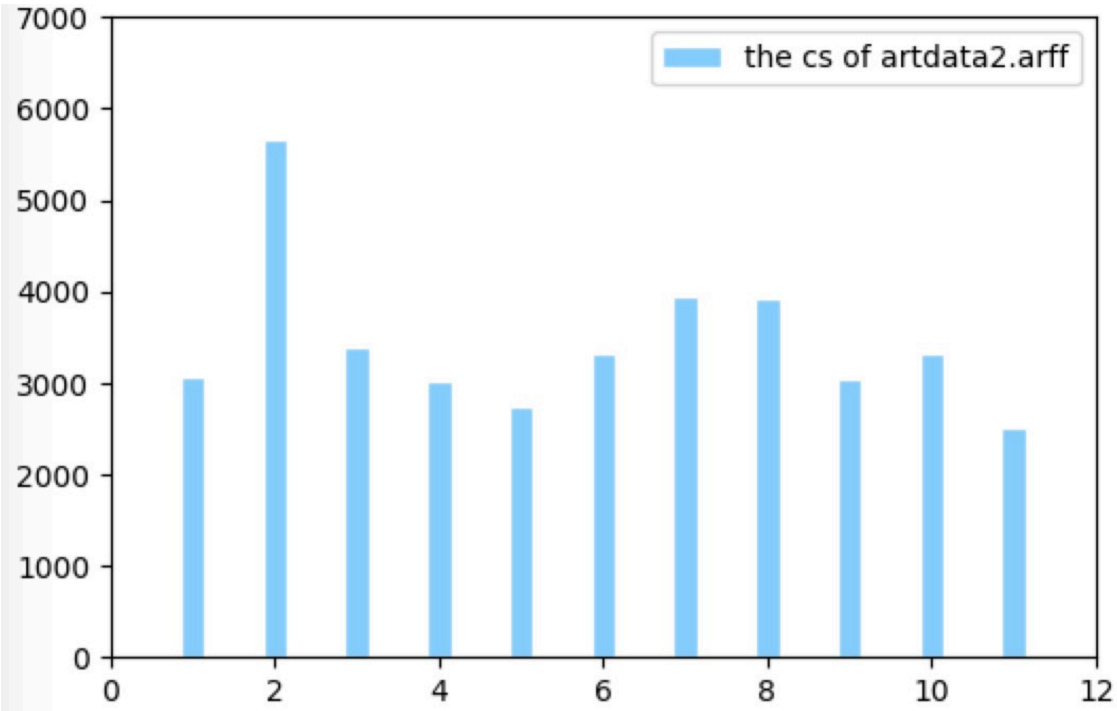
一. Calculate and plot the cs and nmi of 8 datasets.

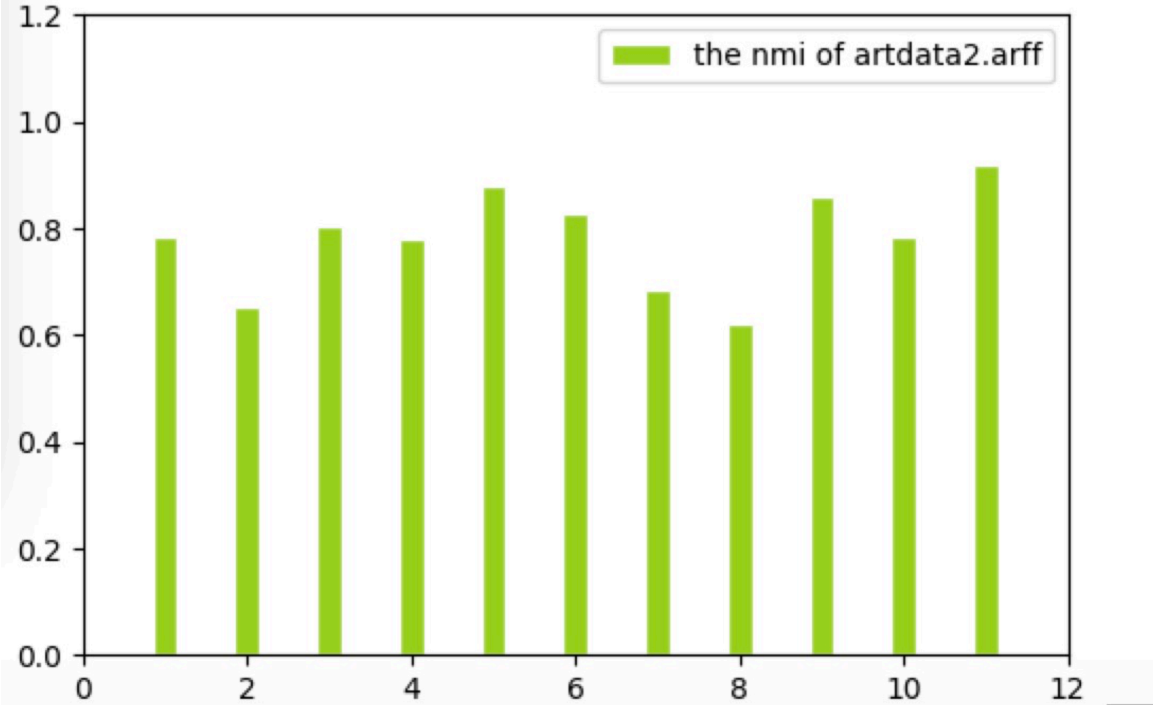

the cs of artdata0.5.arff
4000
3500
3000
2500
2000
1500
1000
500
0
0
2
4
6
8
10
12

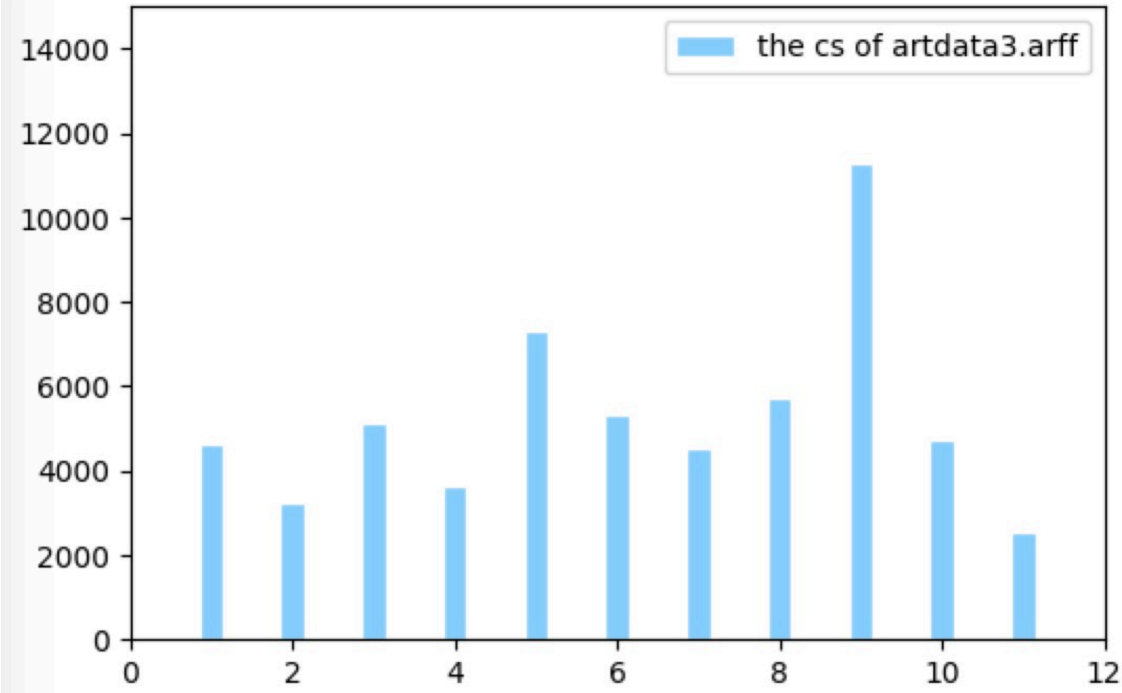

the nmi of artdata0.5.arff
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10
12

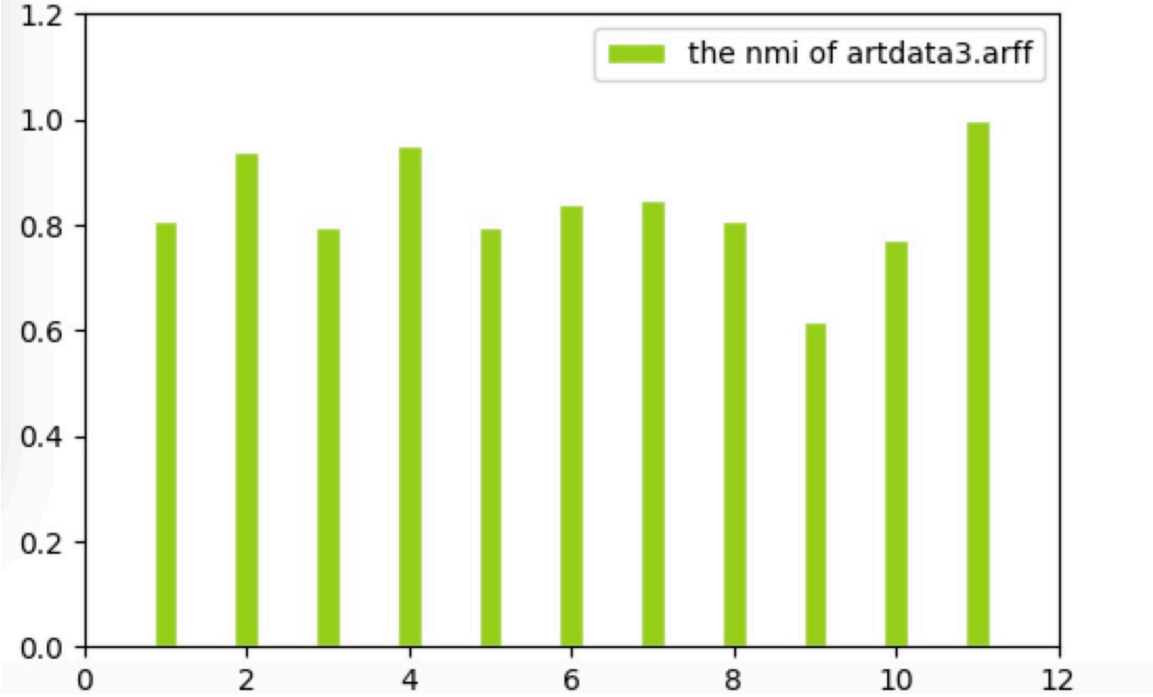

the cs of artdata1.0.arff
5000
4000
3000
2000
1000
0
0
2
4
6
8
10
12

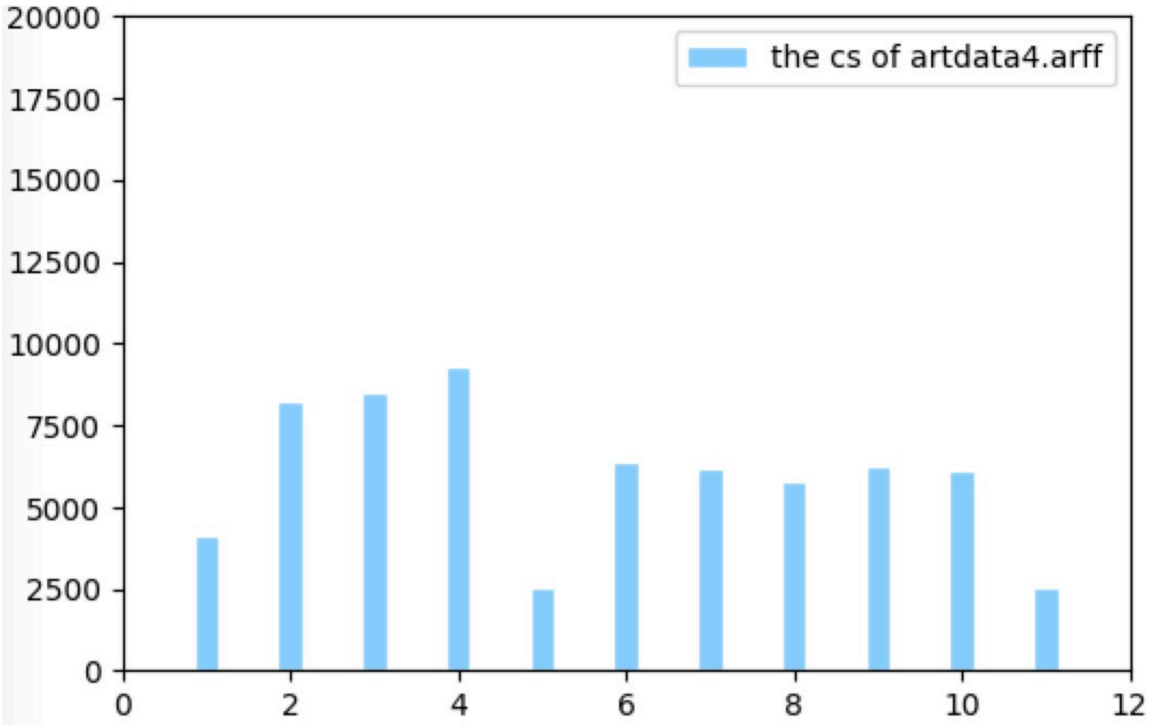

the nmi of artdata1.0.arff
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10
12

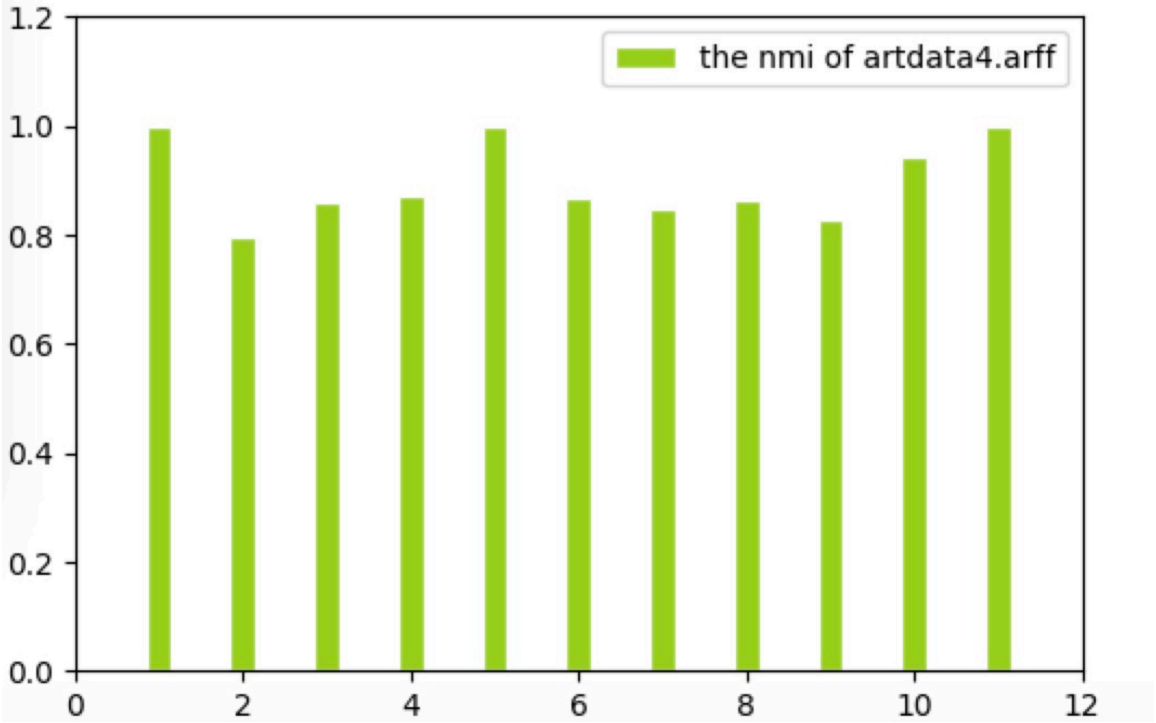

the cs of artdata2.arff
7000
6000
5000
4000
3000
2000
1000
0
0
2
4
6
8
10
12

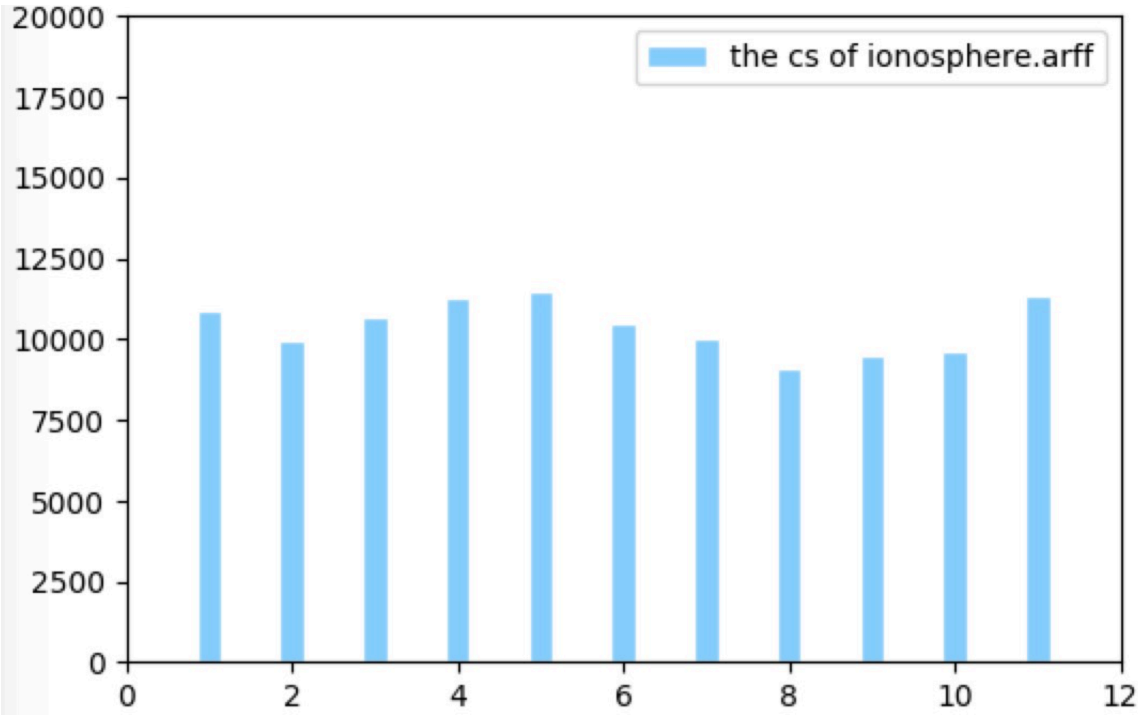

the nmi of artdata2.arff
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10
12

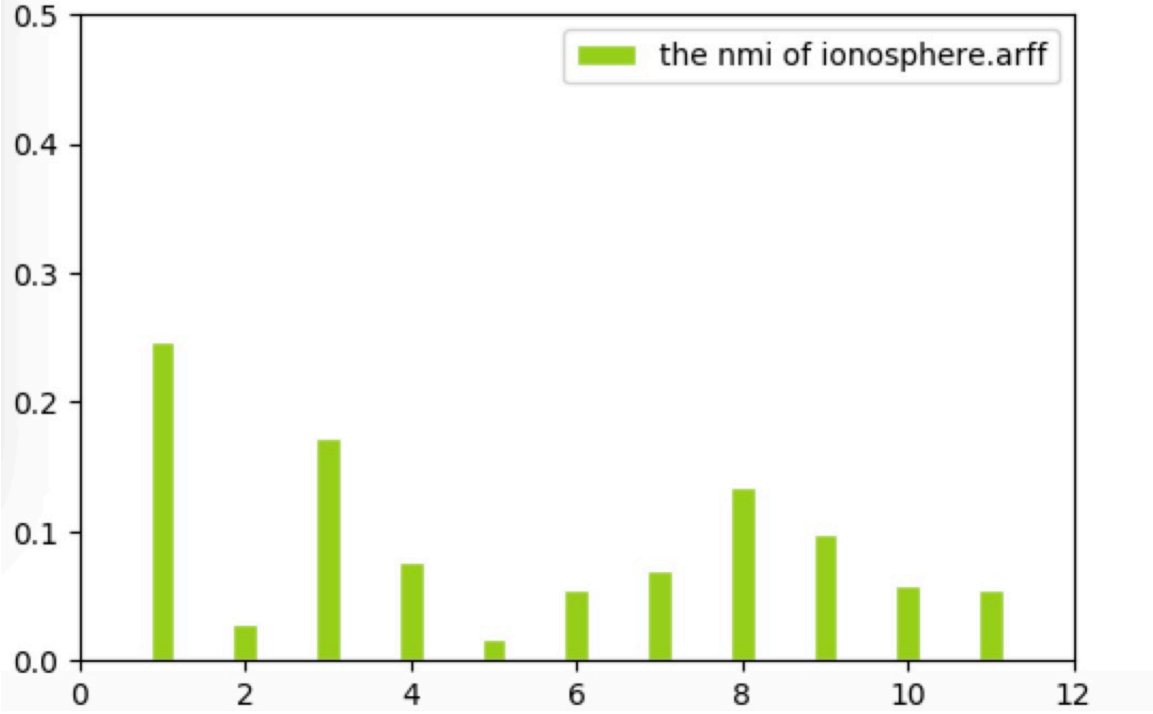

the cs of artdata3.arff
14000
12000
10000
8000
6000
4000
2000
0
0
2
4
6
8
10
12

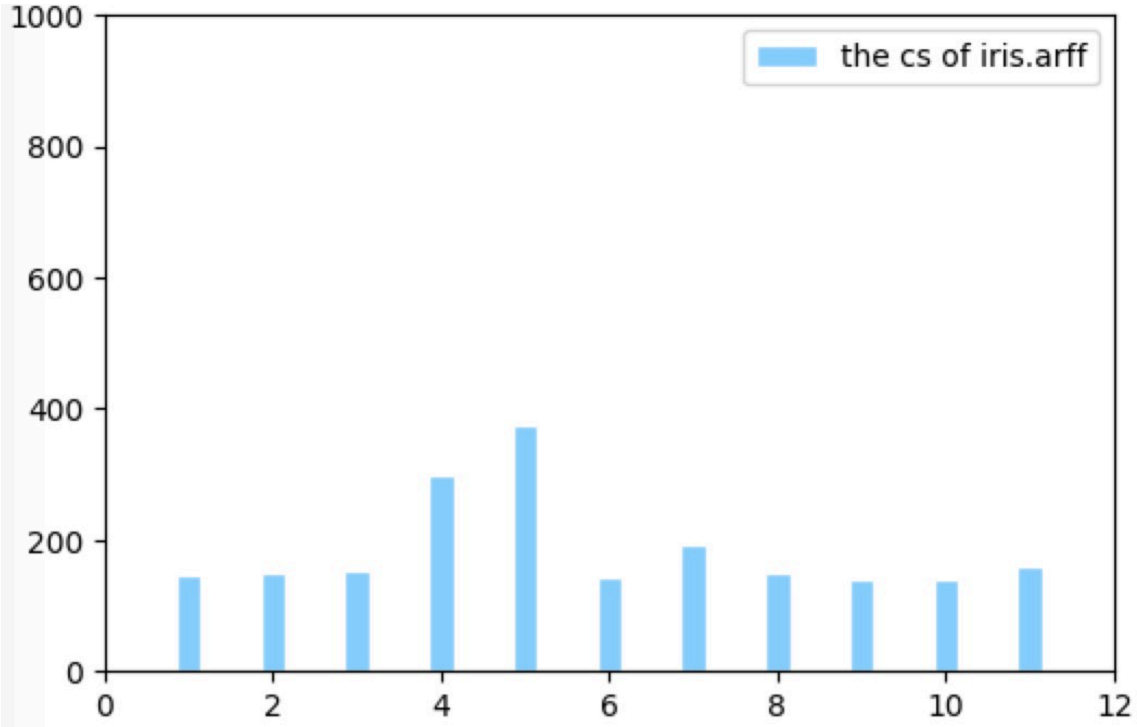

the nmi of artdata3.arff
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10
12

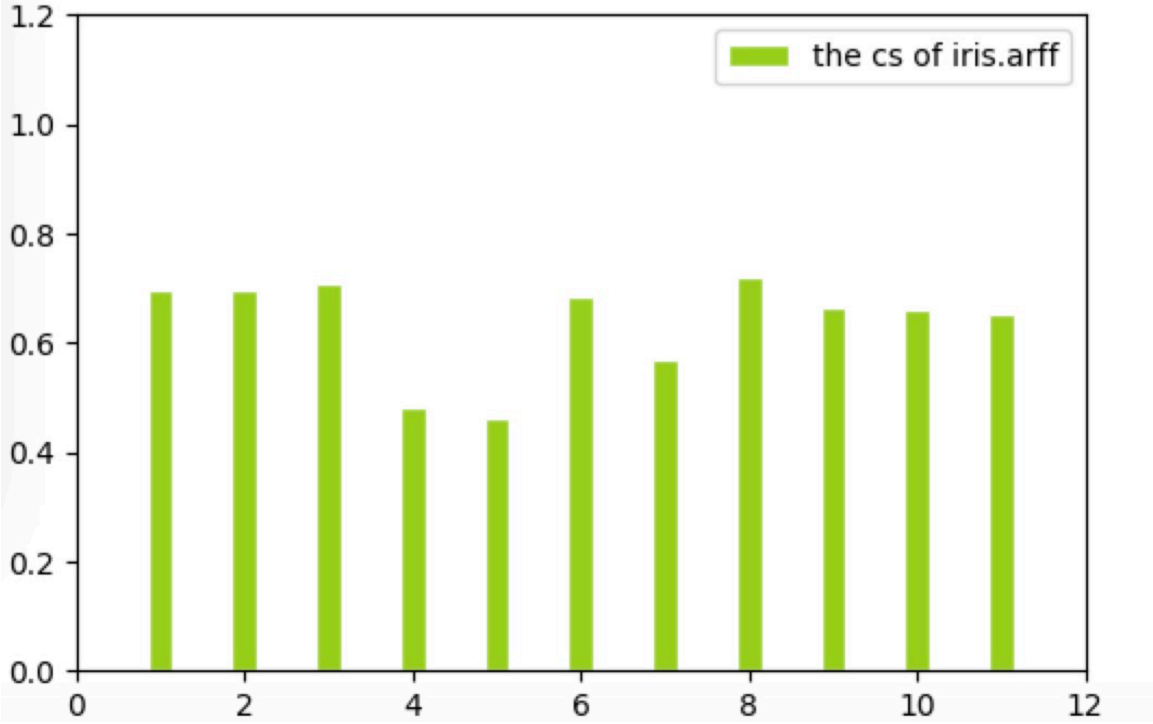

the cs of artdata4.arff
20000
17500
15000
12500
10000
7500
5000
2500
0
0
2
4
6
8
10
12

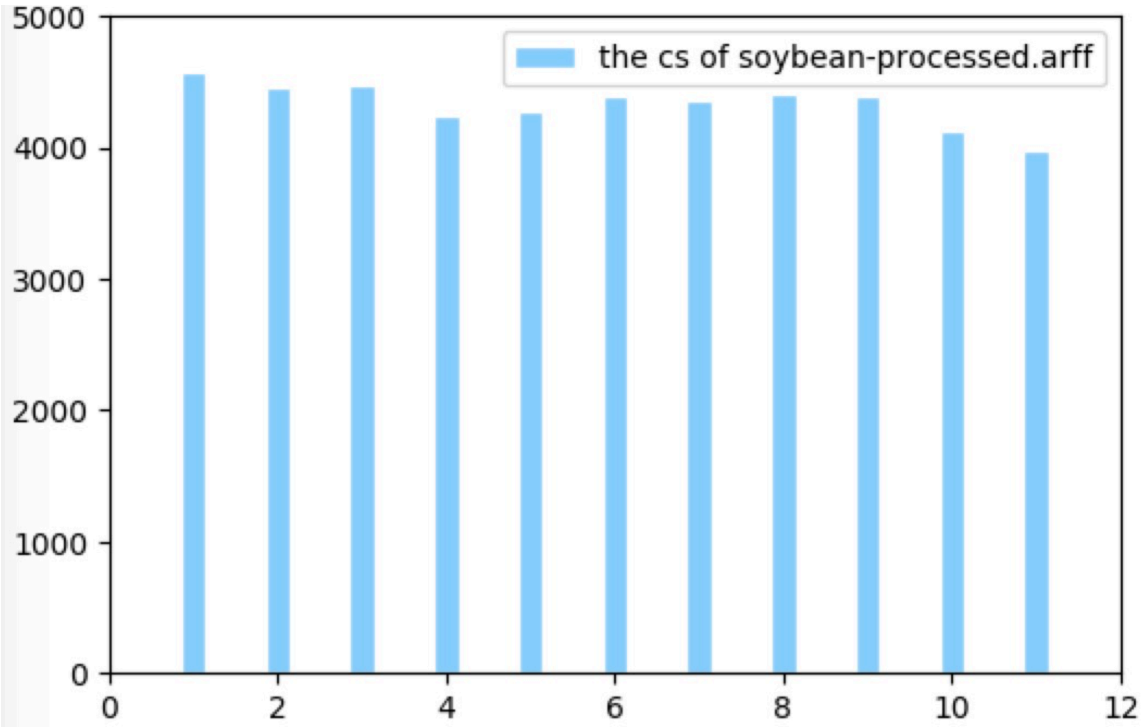

the nmi of artdata4.arff
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10
12

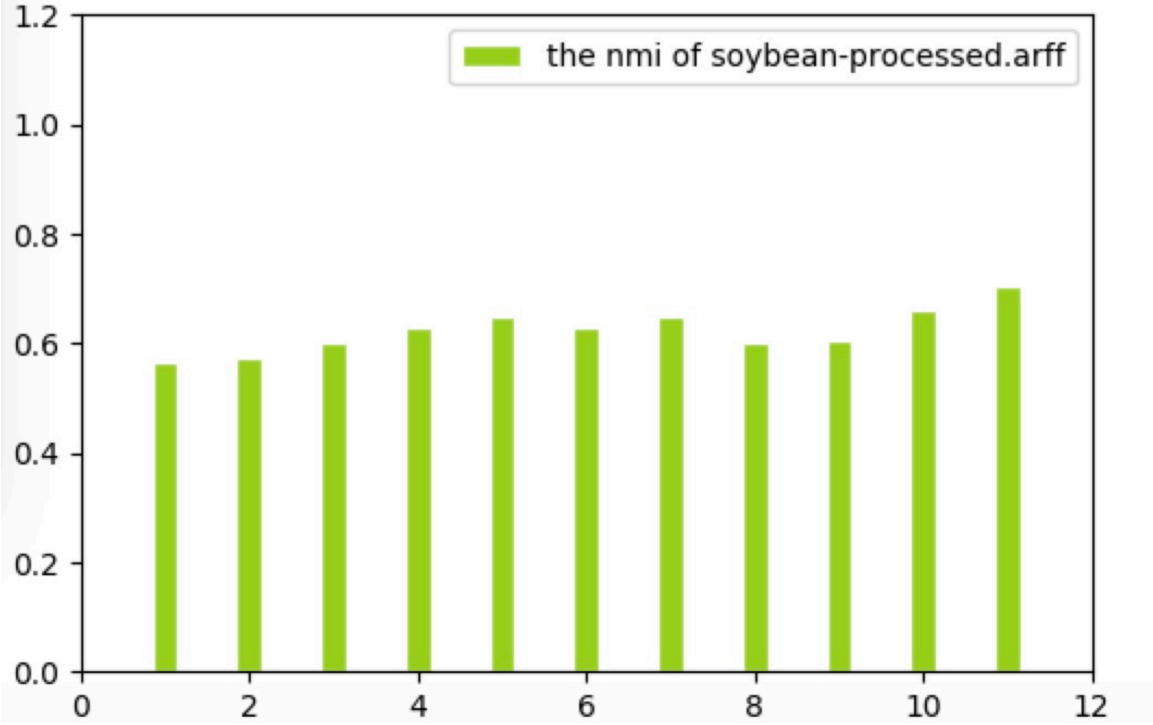

the cs of ionosphere.arff
20000
17500
15000
12500
10000
7500
5000
2500
0
0
2
4
6
8
10
12


the nmi of ionosphere.arff
0.5
0.4
0.3
0.2
0.1
0.0
0
2
4
6
8
10
12


the cs of iris.arff
1000
800
600
400
200
0
0
2
4
6
8
10
12


the cs of iris.arff
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10
12





Discussion:

1. Are CS and NMI stable across multiple random initializations:

From the above graphs, the answer is no, For example, we can see the the cs and nmi value changes across different initialization.

2. Are their quality judgements in agreement?

Yes, if we see the trend of CS and NMI for the same sample, we can see NMI is higher and its CS will lower.

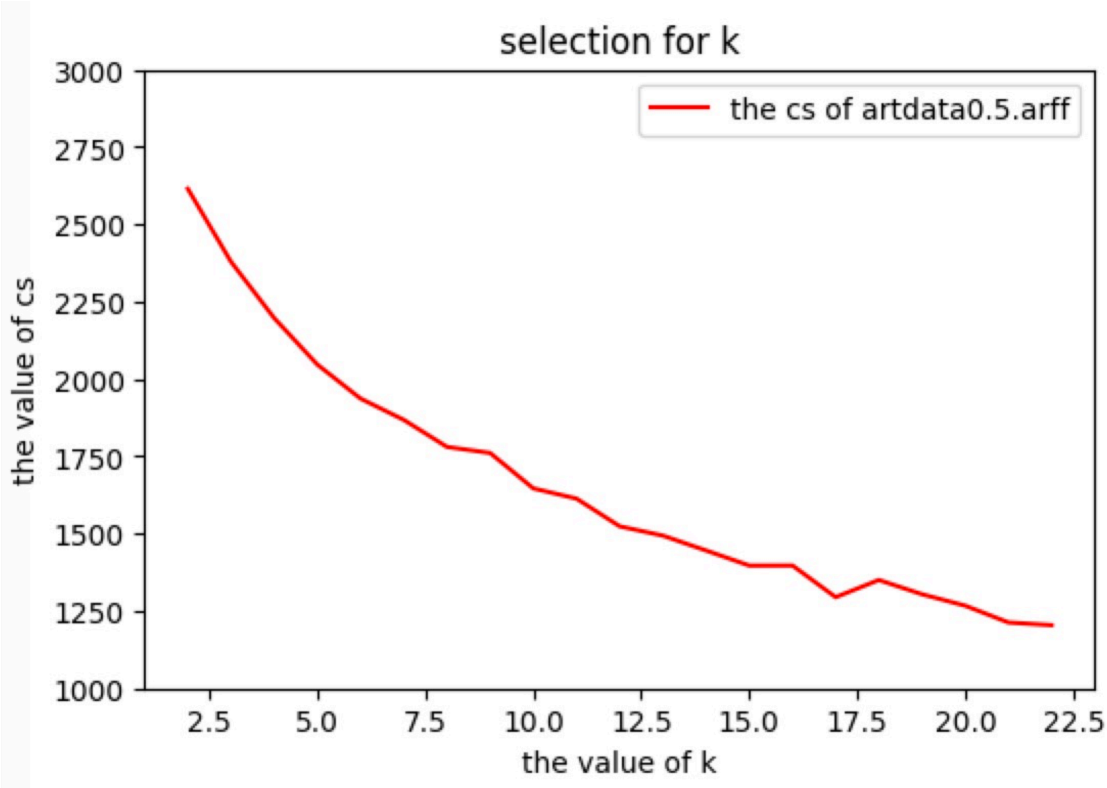
3. How does the smart initialization fare relative to the best random initialization.

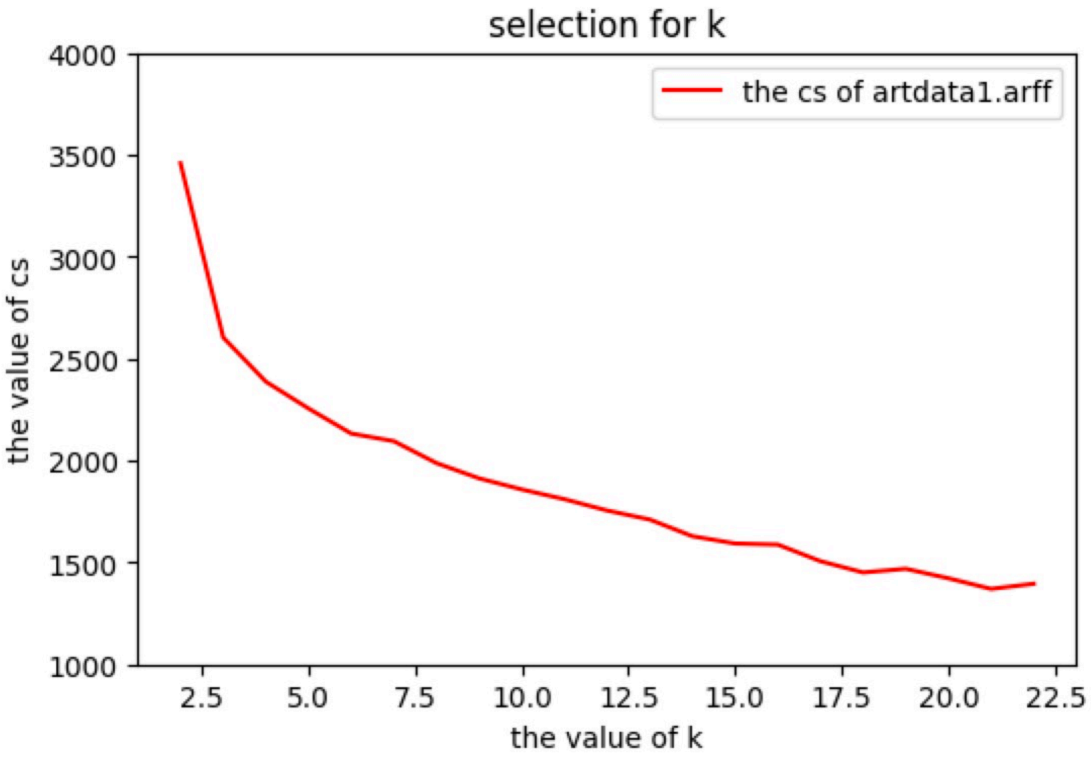
CS of smart initialization sometimes is higher than best random initialization and sometimes is lower than best random initialization. Compared random initialization, smart initialization cannot got the best result always but most of time, its result will be better than random initialization.

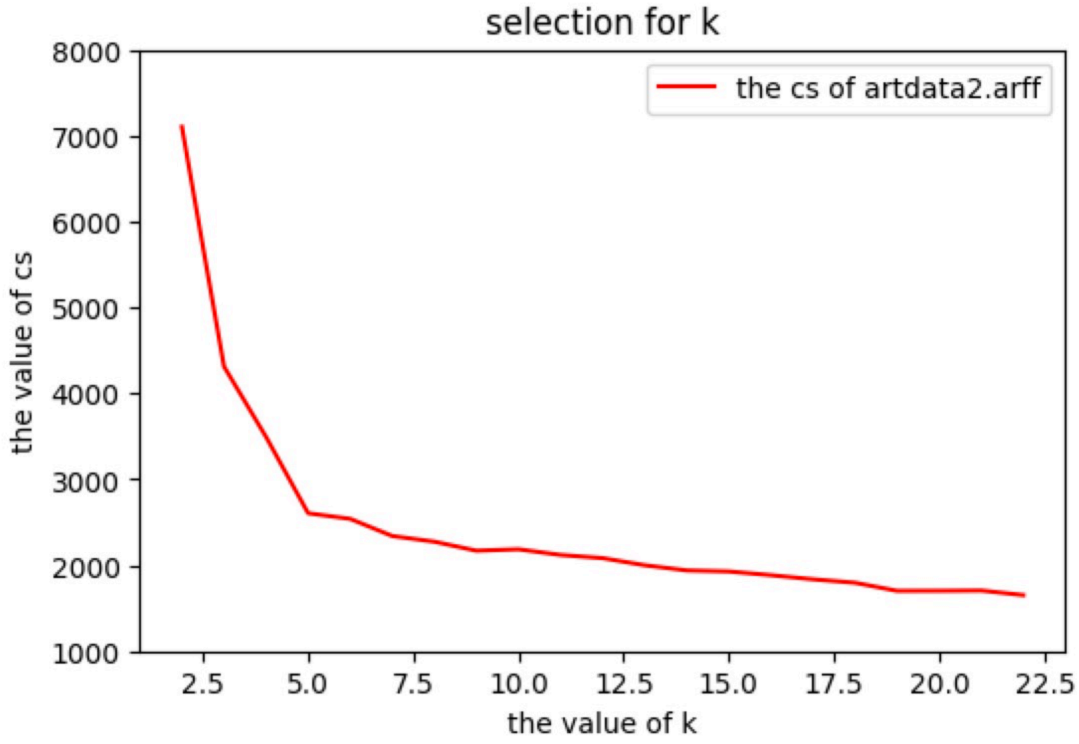
4. What other observations can you make from the results?

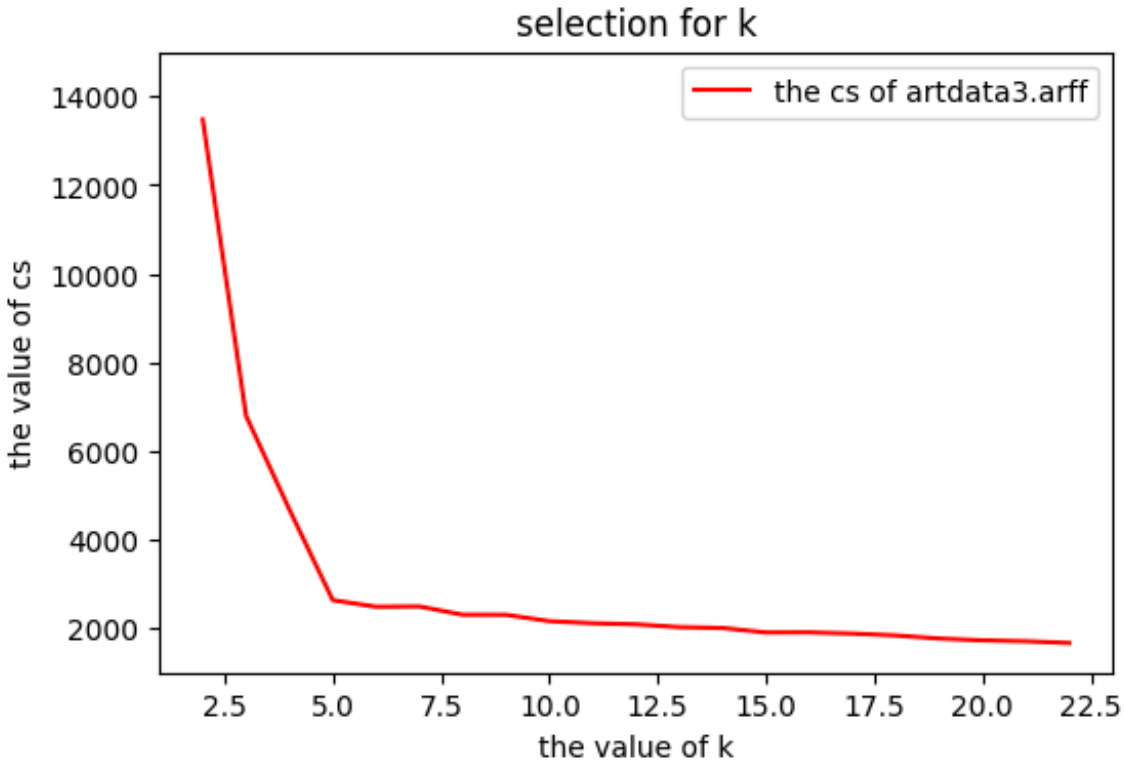
All these datasets were generated using 5 classes each corresponding to a Gaussian sphere, the main difference between these datasets relates to how close the spheres are to one another relative to their width. The closer they are, the harder it is to cluster the dataset. So compared with different artificial data, the variation of CS of the different initialization on the same sample increase with the sequence of artificial dataset(0.5,1,2,3,4) because the width of sphere to create data is smaller for 0.5 than that of 4.

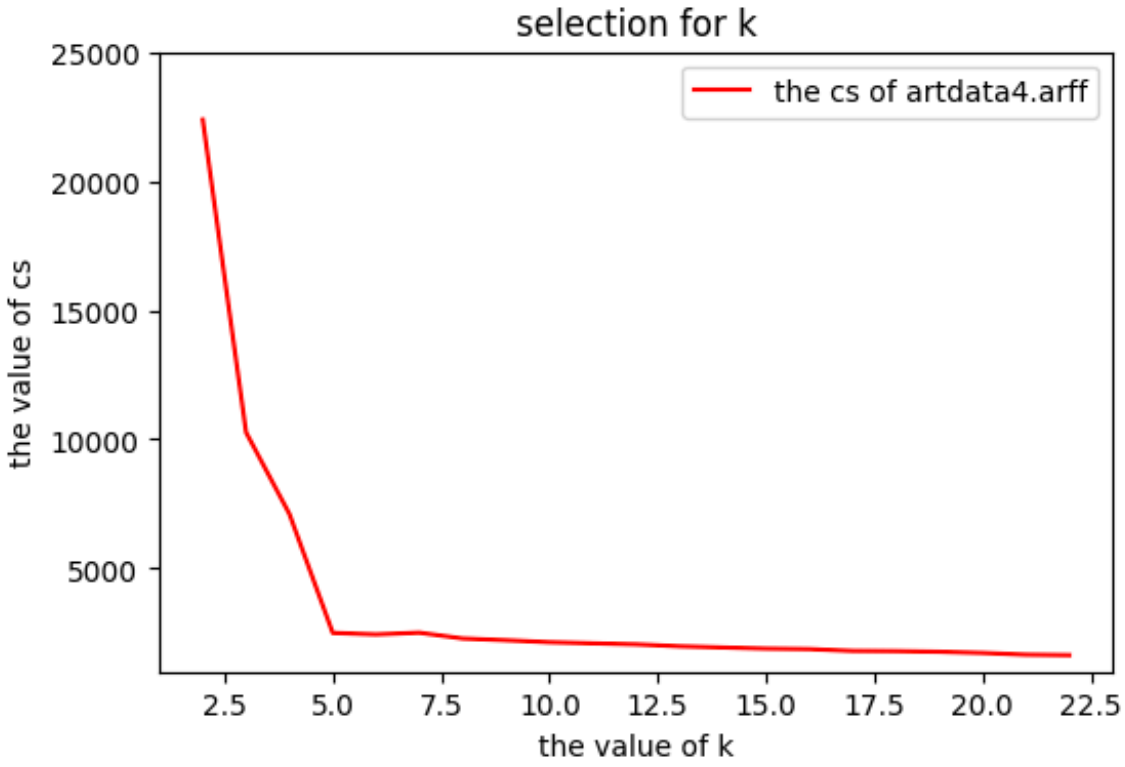
二. Calculate and plot CS value as a function of k(2 to 22).

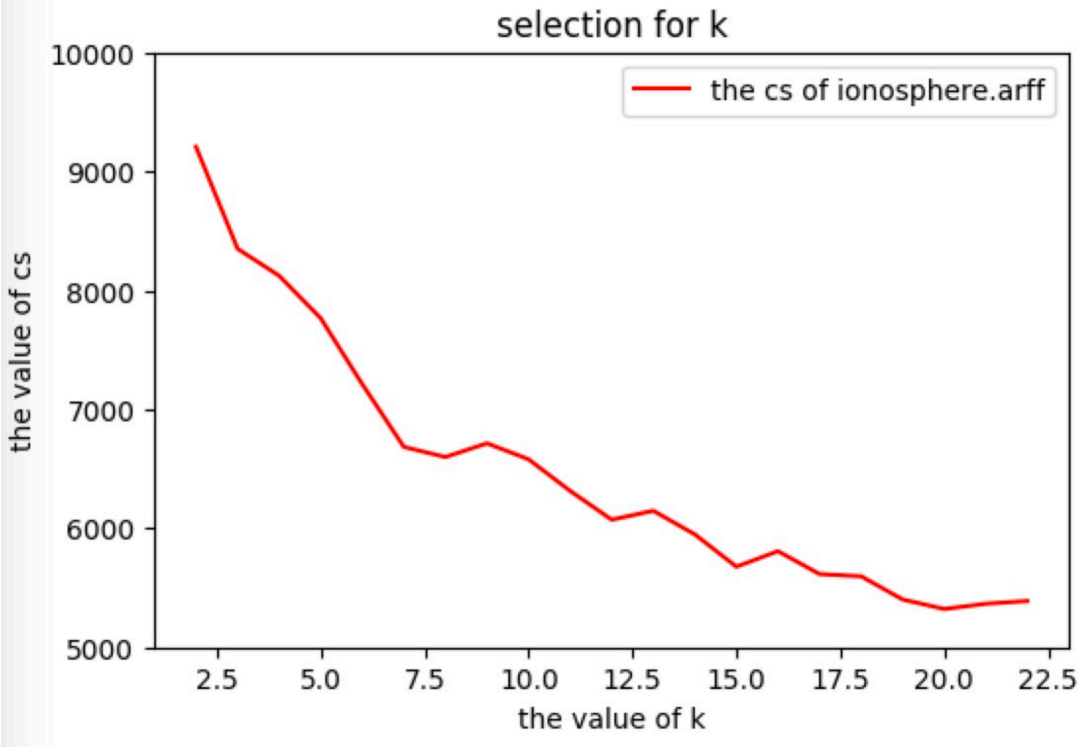

selection for k
the cs of artdata0.5.arff
the value of cs
the value of k
3000
2750
2500
2250
2000
1750
1500
1250
1000
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
22.5

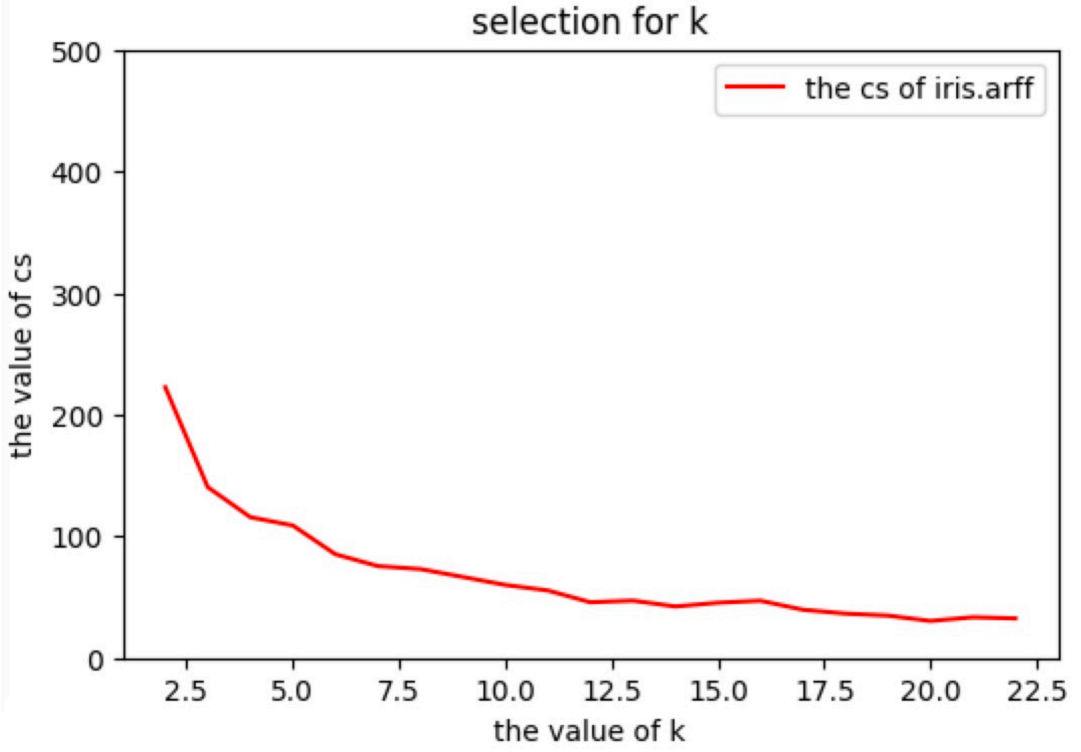

selection for k
the cs of artdata1.arff
the value of cs
the value of k
4000
3500
3000
2500
2000
1500
1000
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
22.5

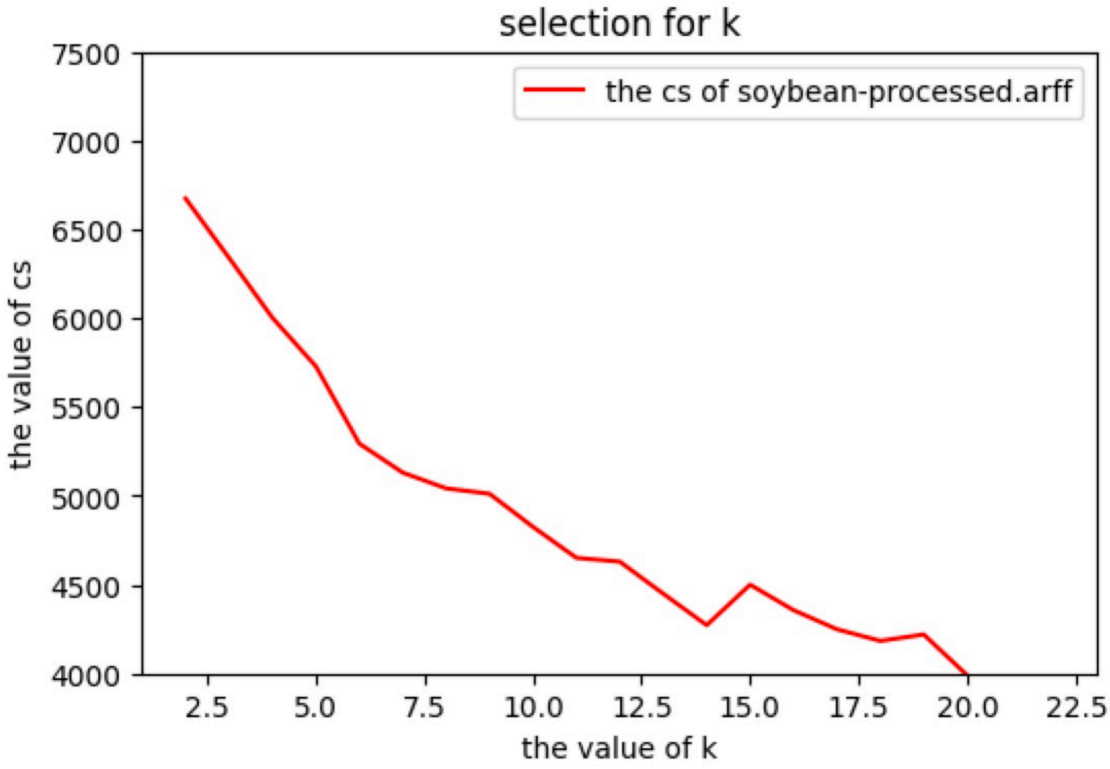

selection for k
the cs of artdata2.arff
the value of cs
the value of k
8000
7000
6000
5000
4000
3000
2000
1000
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
22.5


selection for k
the cs of artdata3.arff
14000
12000
10000
8000
6000
4000
2000
the value of cs
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
22.5
the value of k


selection for k
the cs of artdata4.arff
25000
20000
15000
10000
5000
the value of cs
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
22.5
the value of k


selection for k
the cs of ionosphere.arff
the value of cs
the value of k
10000
9000
8000
7000
6000
5000
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
22.5


selection for k
the cs of iris.arff
the value of cs
the value of k
500
400
300
200
100
0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
22.5



Discussion:

1, Is there visual evidence suggesting that CS can be used as a criterion for selecting k?

It depends, sometimes yes, sometimes no, it depends on data, if graph have knee criterion which mean we can use it to choose k.

2. What can you conclude from the relative performance on the artificial datasets?

The first artificial sample (0.5, 1) haven't knee criterion, others have the knee criterion. The reason why the first artificial sample haven't might be the data generated related to too close the spheres are relative to their width because all these datasets were generated using 5 classes each corresponding to a Gaussian sphere, the main difference between these datasets relates to how close the spheres are to one another relative to their width. The closer they are, the harder it is to cluster the dataset. The Gaussian distribution of sample (0.5, 1) is too close to cluster so there is no any knee criterion. Others' knee criterion have the same knee criterion when k is equal to 5 because all of data generated by five Gaussian sphere.

4. Are the results consistent across the datasets? If not, what observations can you make from the results?

No, Most artificial sample have knee criterion, but no any real dataset have knee criterion, it looks like this method is hard to be used in real world.