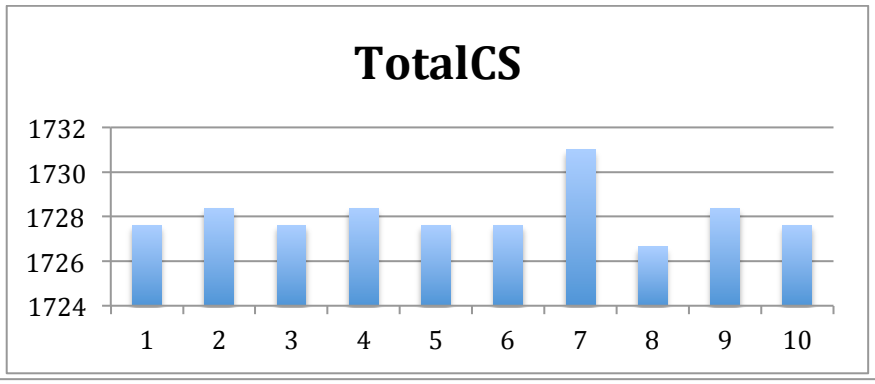
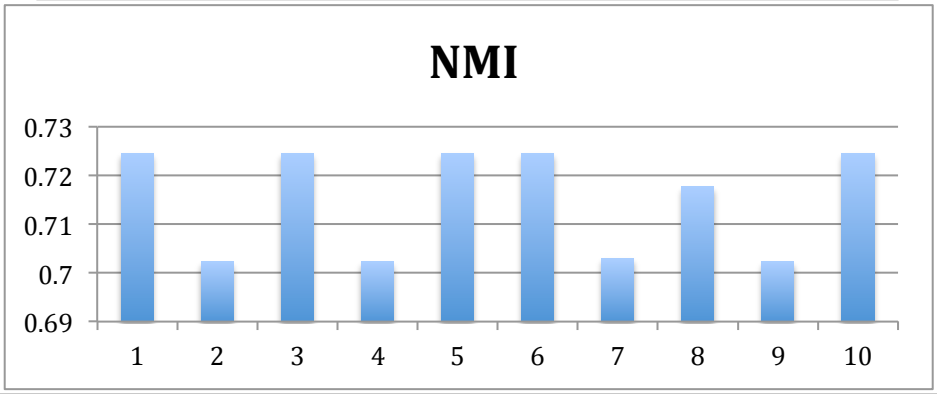
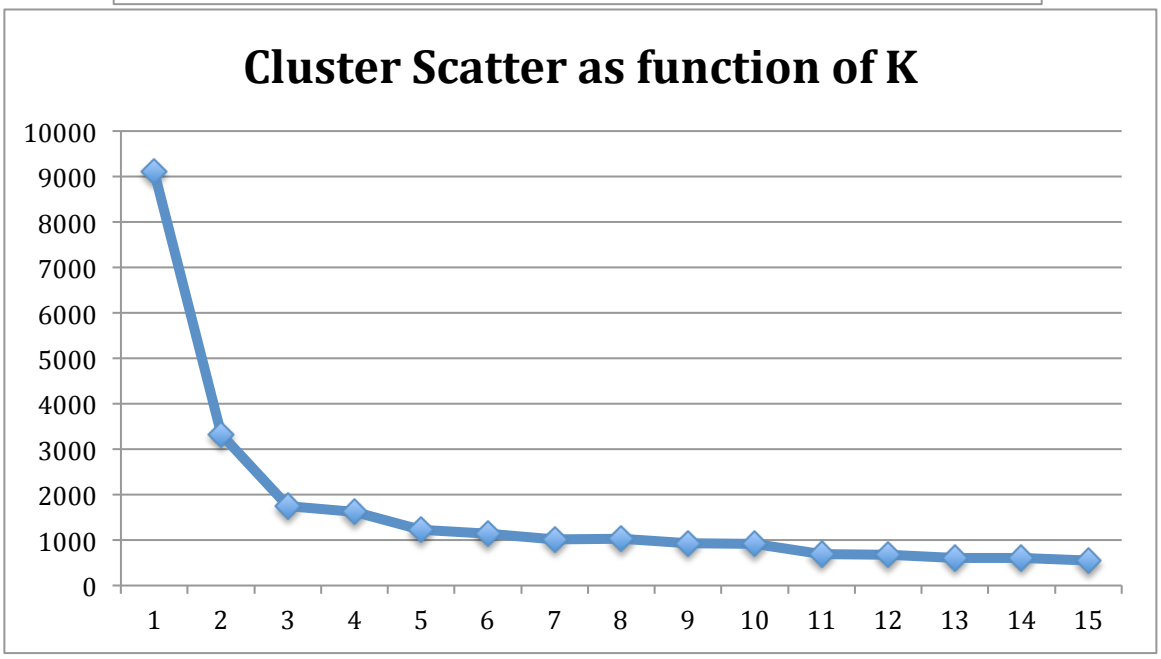
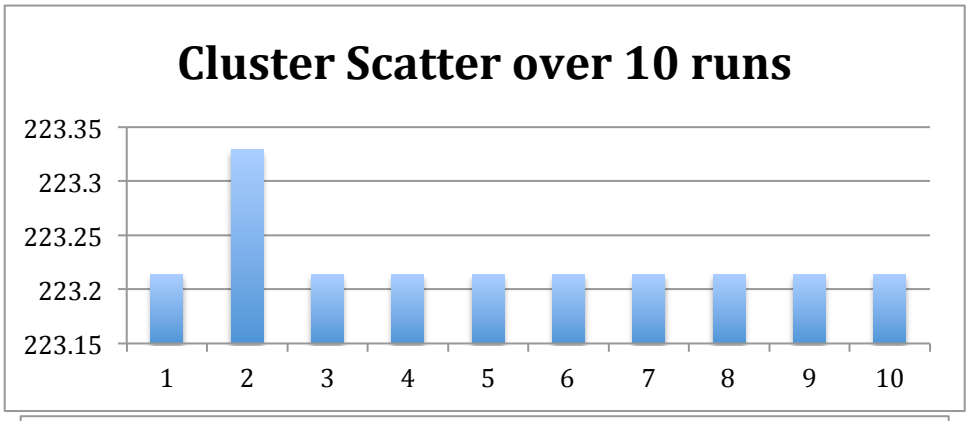
SEEDS.ARFF

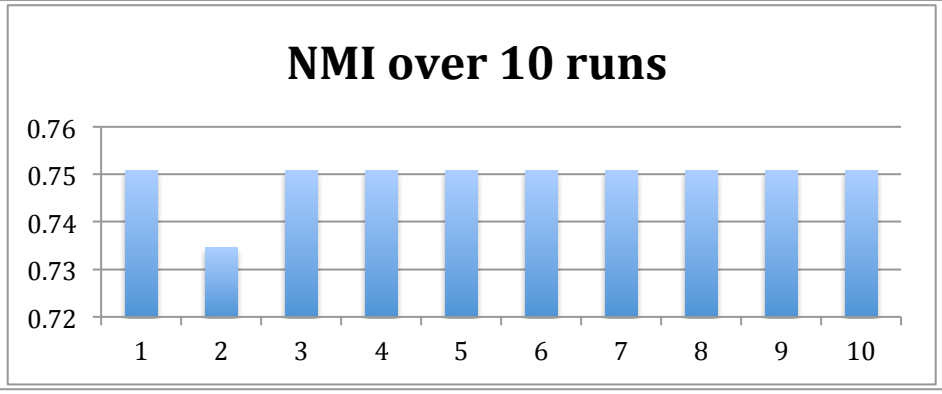

TotalCS
1732
1730
1728
1726
1724
1 2 3 4 5 6 7 8 9 10

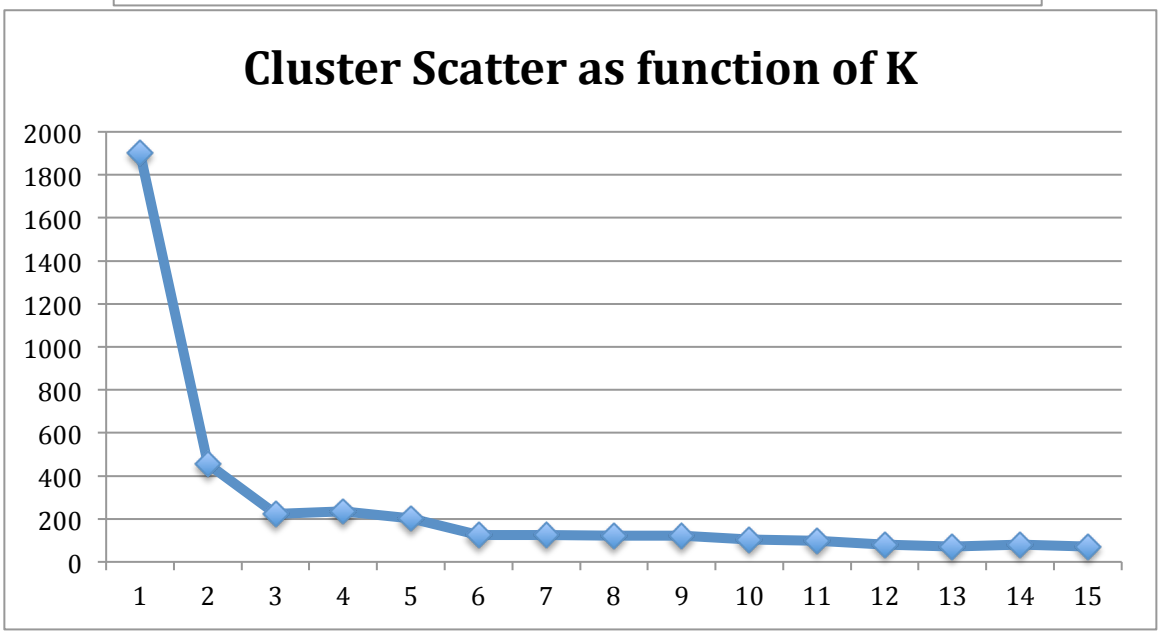

NMI
0.73
0.72
0.71
0.7
0.69
1 2 3 4 5 6 7 8 9 10


Cluster Scatter as function of K
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

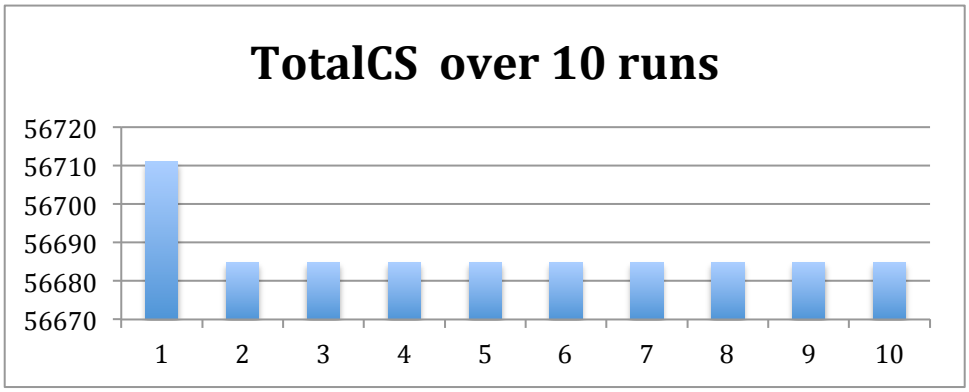
IRIS.ARFF

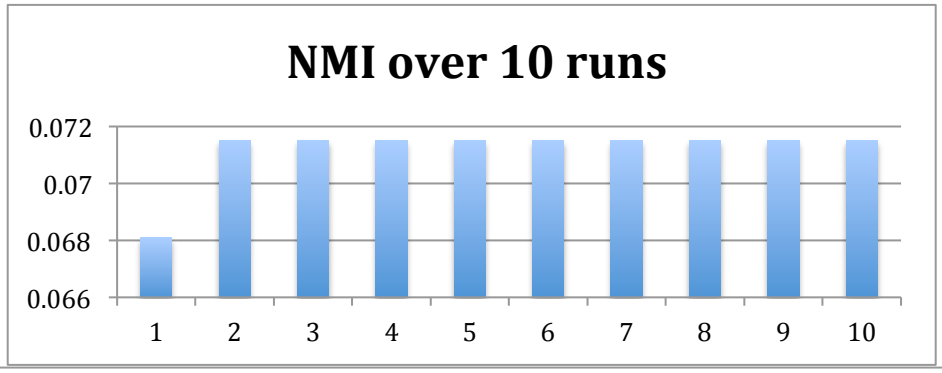

Cluster Scatter over 10 runs
223.35
223.3
223.25
223.2
223.15
1
2
3
4
5
6
7
8
9
10

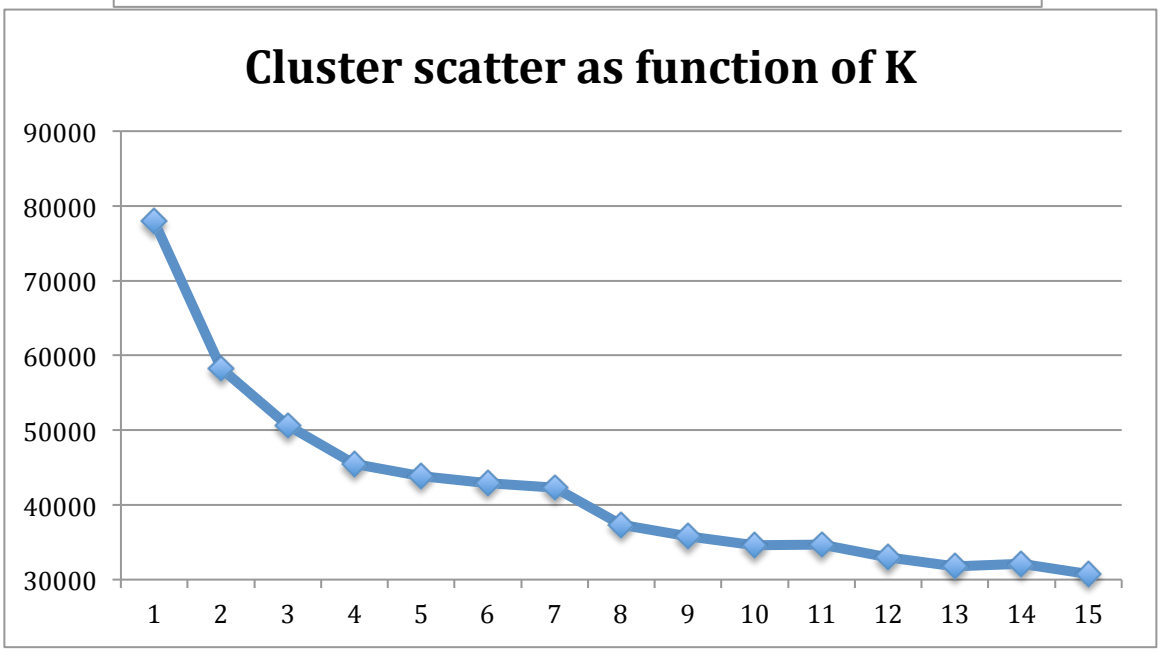

NMI over 10 runs
0.76
0.75
0.74
0.73
0.72
1
2
3
4
5
6
7
8
9
10


Cluster Scatter as function of K
2000
1800
1600
1400
1200
1000
800
600
400
200
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

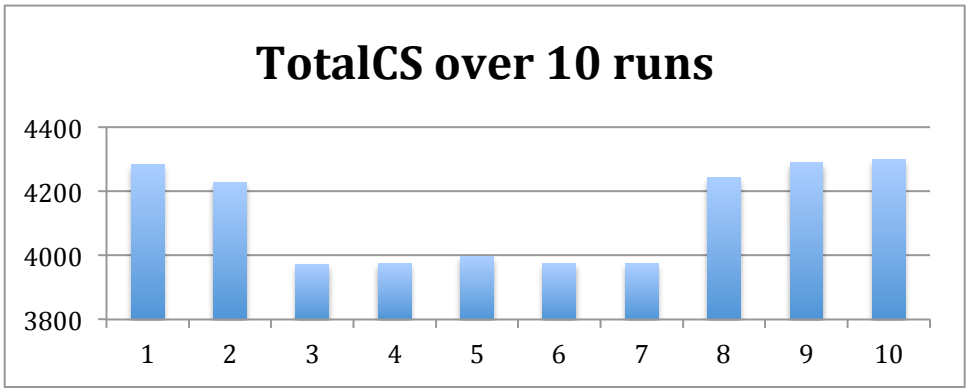
IONOSPHERE.ARFF

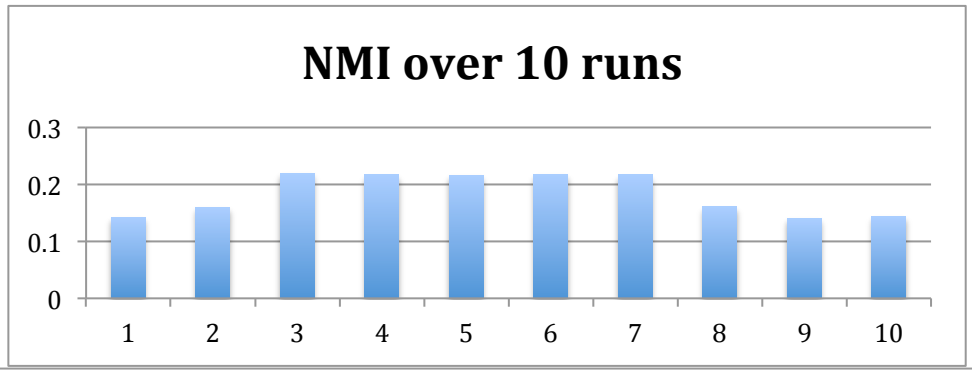


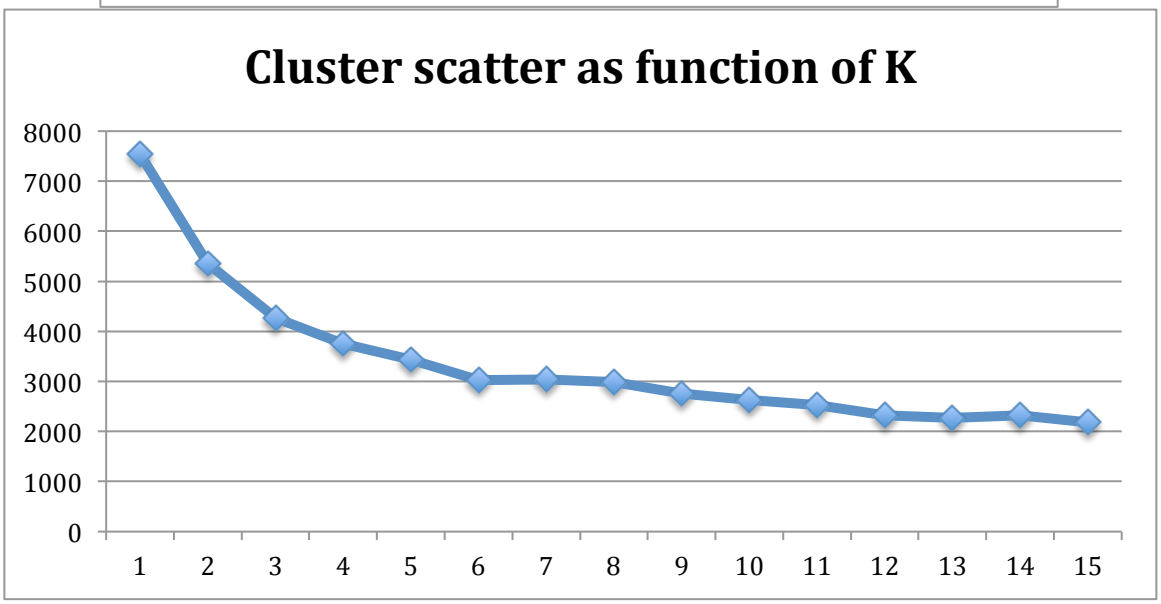




ARTDATA.ARFF







Expermient Results:

Selecting K:

In this experiment, we wanted to see how the value of k affects the CS of the clusters. For all datasets, we see a really nice curve that shows the decrease of CS as k improves, as we expect it to. Then they all plateau eventually, and the values of CS start leveling out. We could use these lines to find the value of k that gives us a good CS without sacrificing performance (aka keeping k small). Iris.arff is a good example of this. We see a sharp decline in CS up until 3 and then it generally levels out. Therefore, picking k =3 or possibly k = 6 would be a good choice.

Sensitivity to initialization:

I was more surprised by the values I obtained from this experiment. We were testing the affects of initializations on CS and NMI, but what I saw from most of the datasets is that regardless of the initialization, the values eventually settled down and remained almost constant. I expected a bigger variation, but the values are very similar. The only one that shows any variation is the artData.arff. This is the dataset that was artificially generated and includes a high level of noise. Therefore, we see that the variation was much higher for NMI and CS if the data is noisy. Otherwise, the NMI and CS should remain relatively stable for multiple initializations.