Comp 135 hw3

Fanying Ye

Due: November 5 2015

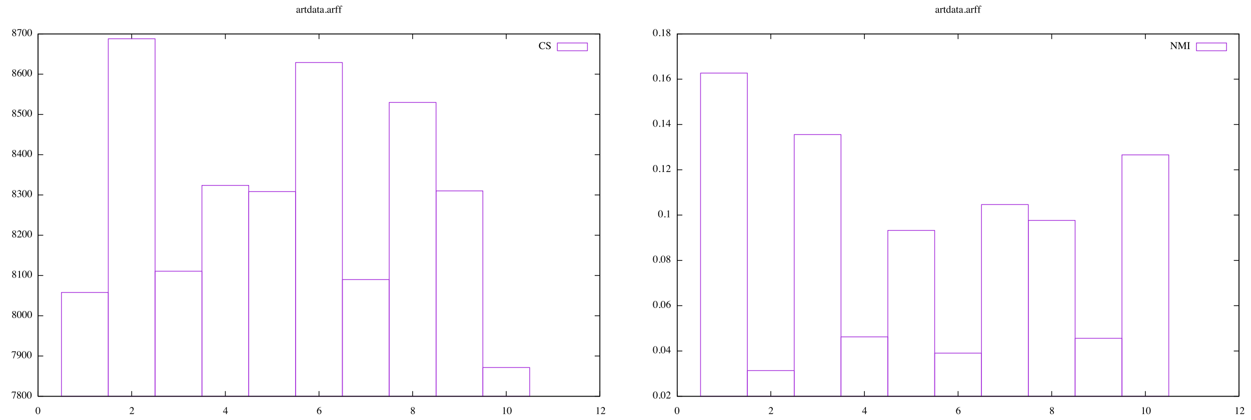
1 Introduction

In this assignment I conducted experiments to implement and test properties of the k-means algorithm. I evaluate the sensitivity of k-means to initialization, and calculate CS under different number of clusters to check whether CS can be used as a criterion for seleting k.

2 Plots and Conclusion

2.1 Sensitivity of k-means to initialization

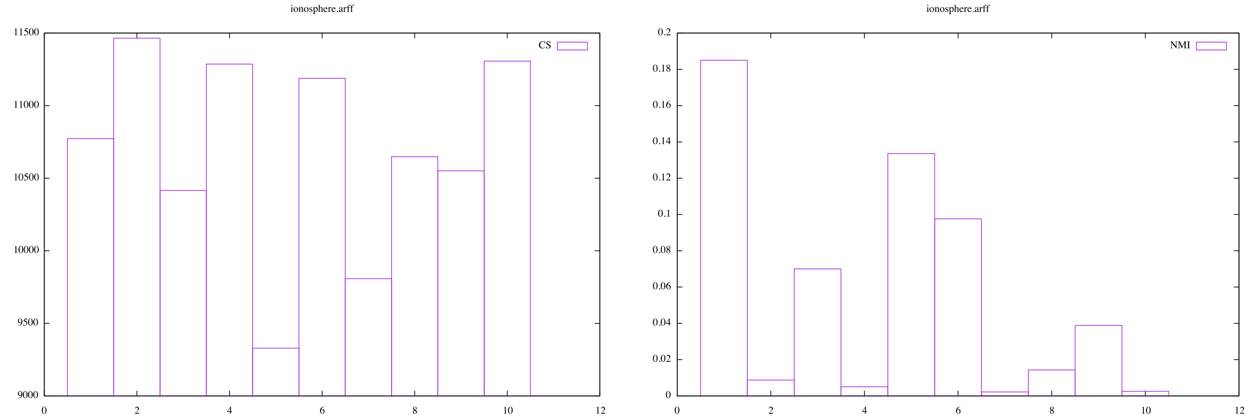
Questions:
Are CS and NMI stable across multiple initializations? Are their quality judgements in agreement? What other observations can you make from the results?
Answers:
According to the plots, we can see that the both CS and NMI are not quite stable with different random initialization, which proves that k-means algorithm is quite sensitive to starting points. This happens because we pick k cluster center arbitrarily, once the initial value of the selection is not good , it may result in ineffective clustering results.
Besides, we can conclude from the plot that the quality judgement of CS and NMI are often in agreement while not always. In another word, the quality judgement of CS and NMI are sometimes different. For example, in the result of data set ionosphere.arff, the biggest NMI appears in the second try, while the smallest CS appears in the seventh try; in the result of data set iris.arff, the biggest NMI appears in the tenth try, while the smallest CS appears in the third try. This happens because CS criteria taking into account of only cluster scatter, while NMI compare the clustering result to the labels on the labeled dataset.As a result, when the cluster is dense and have clear distinction between classes, CS and NMI tend to have judgment in agreement.



(a) CS of artdata.arff

(b) NMI of artdata.arff

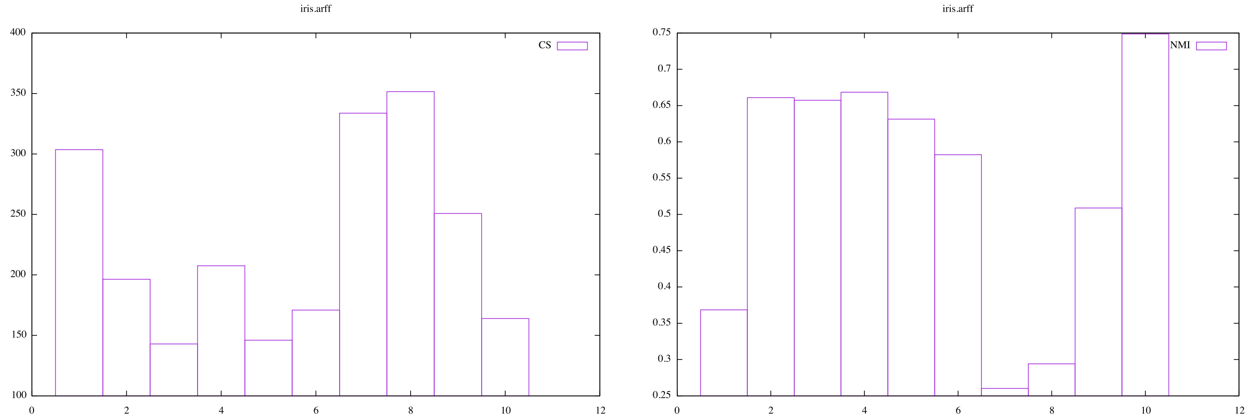
Figure 1: CS and NMI of with 10 different initialization of artdata.arff



(a) CS of ionosphere.arff

(b) NMI of ionosphere.arff

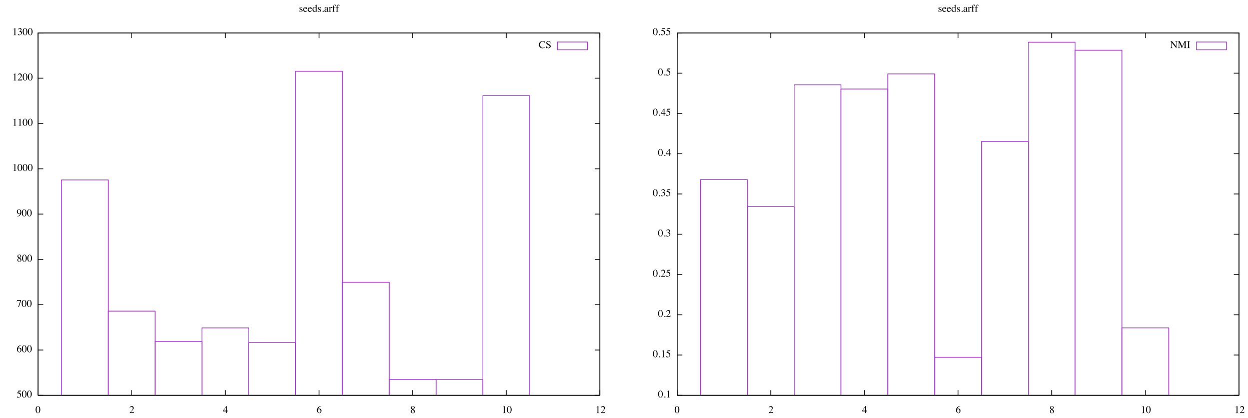
Figure 2: CS and NMI of with 10 different initialization of ionosphere.arff



(a) CS of iris.arff

(b) NMI of iris.arff

Figure 3: CS and NMI of with 10 different initialization of iris.arff



(a) CS of seeds.arff

(b) NMI of seeds.arff

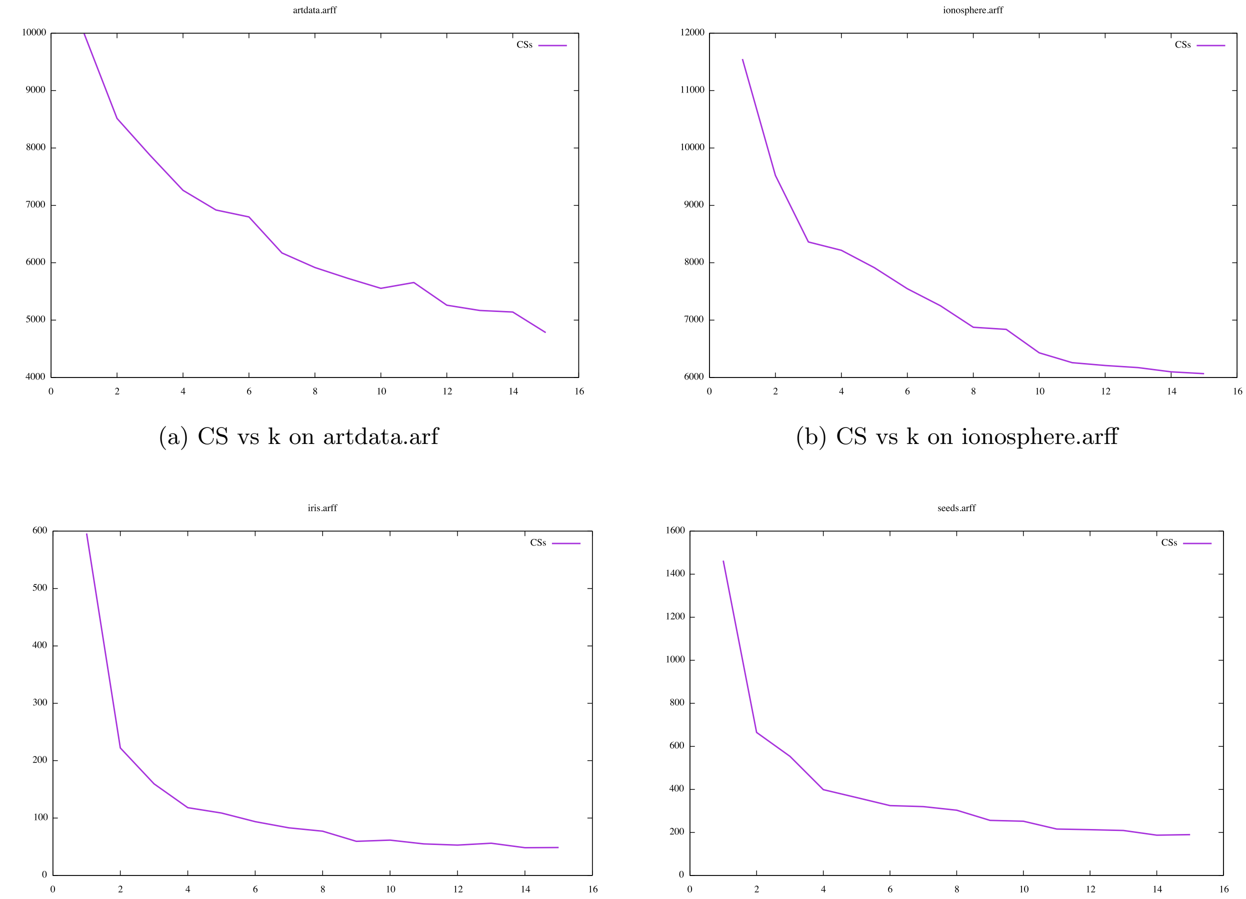
Figure 4: CS and NMI of with 10 different initialization of seeds.arff

2.2 Selecting k

Conclusion:
We hope to see big drop in criterion until we get the right k and moderate drop after that.

We know that artdata has 3 classes; ionosphere has 2 classes; iris has 3 classed and seeds has 3 classes. According to the plot, we can conclude that CS performs well in the iris and seeds dataset, while poor in artdata and ionoshere. This happens because data in ionosphere dataset are high dimension and data in artdata are noisy.

Thus, we can conclude that CS can be used as a criterion for selekting k when the cluster is dense and have clear distinction between classes. Otherwise, it can lead to higher k than expected.



(a) CS vs k on artdata.arff

(b) CS vs k on ionosphere.arff

(c) CS vs k on iris.arff

(d) CS vs k on seeds.arff

Figure 5: CS as a function of k under different data set