TDT4300 Datavarehus og datagruvedrift - Spring 2013
Assignment 5: Cluster Validation

Magnus L Kirø

April 14, 2013

Evaluation

From the graph in figure 2 2 we can see the SSE value for the first file. We clearly see that the error drops significantly at k=3 (the x-axis represents k-1). This indicates that a clustering with three centroids give a good clustering of the data in the data set. Similarly we can see in figure 6 6 that the best results are provided at k=2. The lower the SSE value the better the clustering is. With fewer clusters we get faster computation time, which is positive. Thus a lower value of k will give quicker results. Althought maybe not the best results.

SSE measures the cohesion of a cluster, og the entire dataset. This differs from SSB which measures the independence of each cluster. The silhouette value is a combination of SSB and SSE.

The measures show the correctness of the clustering and what degree of clustering should be used in this case.

Graphs and tables

The two files are: "iris.arff" and "segment-challenge.arff". Graph descriptions are in the image caption. The x axis represents K-1.

	A	B	C	D
1	sse	ssb	sum	silhouette
2	0.124394829	2.922957422	3.047352251	-0.5
3	0.199383619	3.215203761	3.41458738	-0.5
4	0.132794091	3.281793289	3.41458738	-0.5
5	0.195175321	3.710131971	3.905307293	-0.5
6	0.174363332	3.707565124	3.881928455	-0.5
7	0.160018347	3.721910108	3.881928455	-0.5
8	0.135503361	3.760571518	3.896074879	-0.5
9				
10	sse	ssb	sum	silhouette
11	35.49110245	340.0397052	375.5308076	-0.5
12	87.14983191	960.0918015	1047.241633	-0.5
13	85.43342589	1081.994366	1167.427792	-0.5
14	65.39035546	1218.414528	1283.804883	-0.5
15	59.3045657	1224.500317	1283.804883	-0.5
16	55.5959823	1228.208901	1283.804883	-0.5
17	53.41053099	1230.394352	1283.804883	-0.5

Figure 1: Table containing all results from the two files. The top one is the first file.

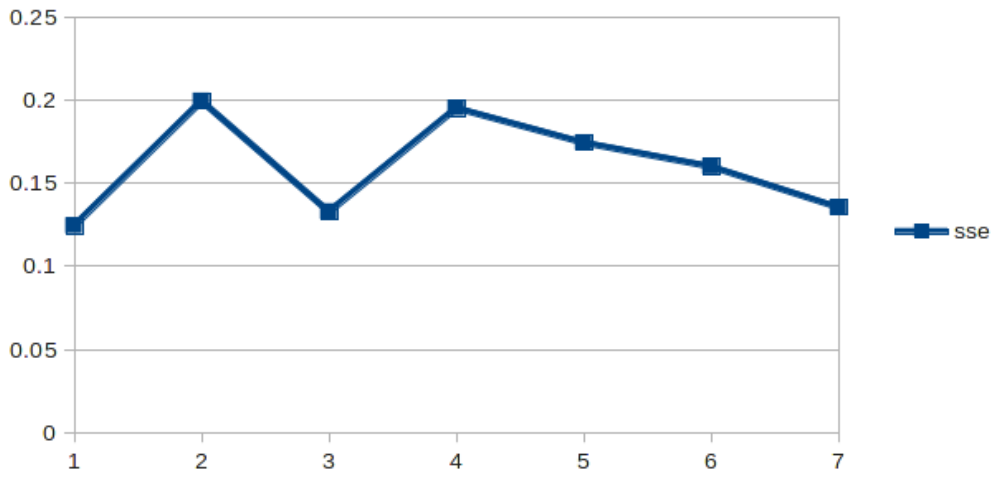

Figure 2: SSE for the first file.

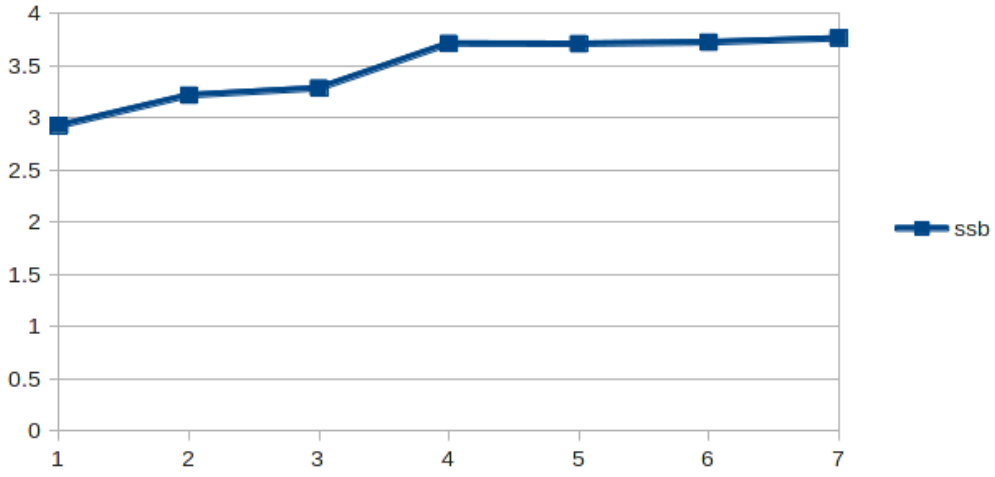

Figure 3: Ssb for the first file.

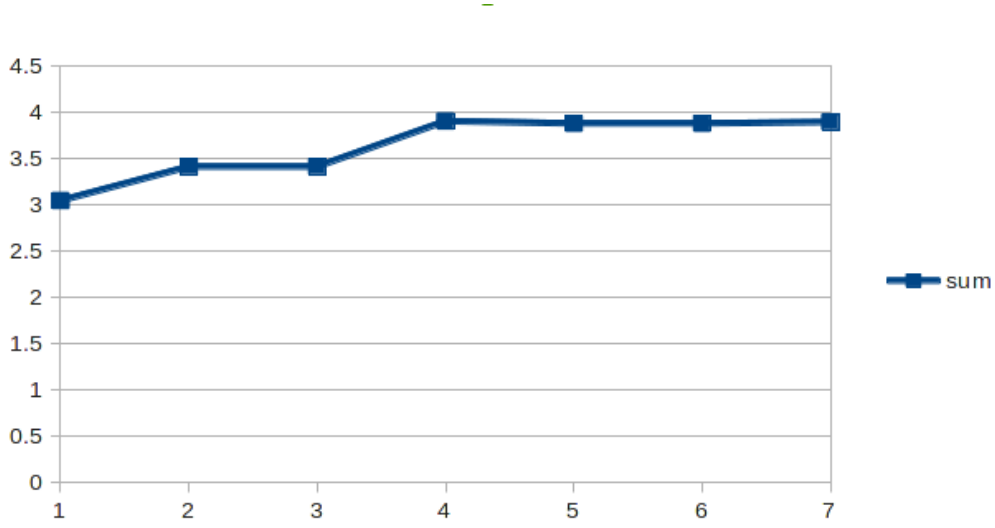

Figure 4: Sum for the first file.

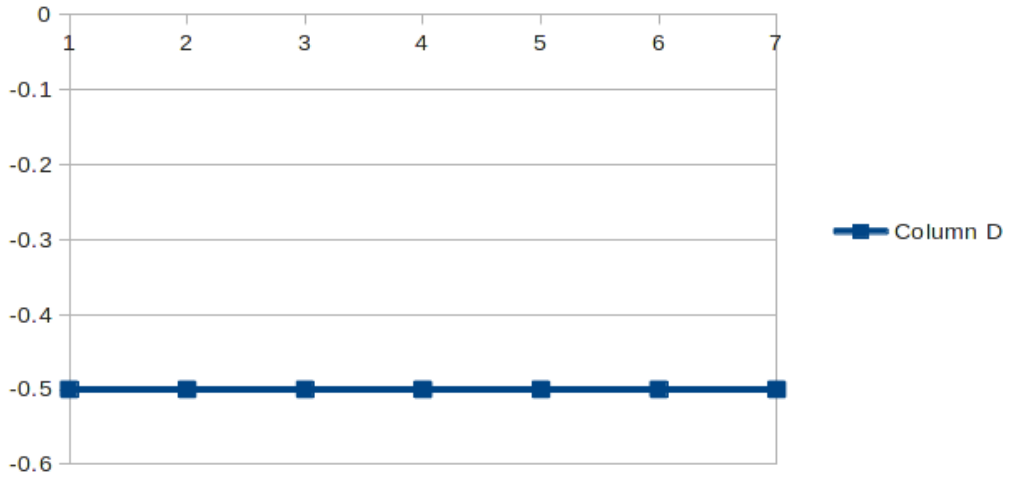

Figure 5: Shows the silhouette values for bot files. They were the same consistently. There are probably some error in the code but I could not find it.

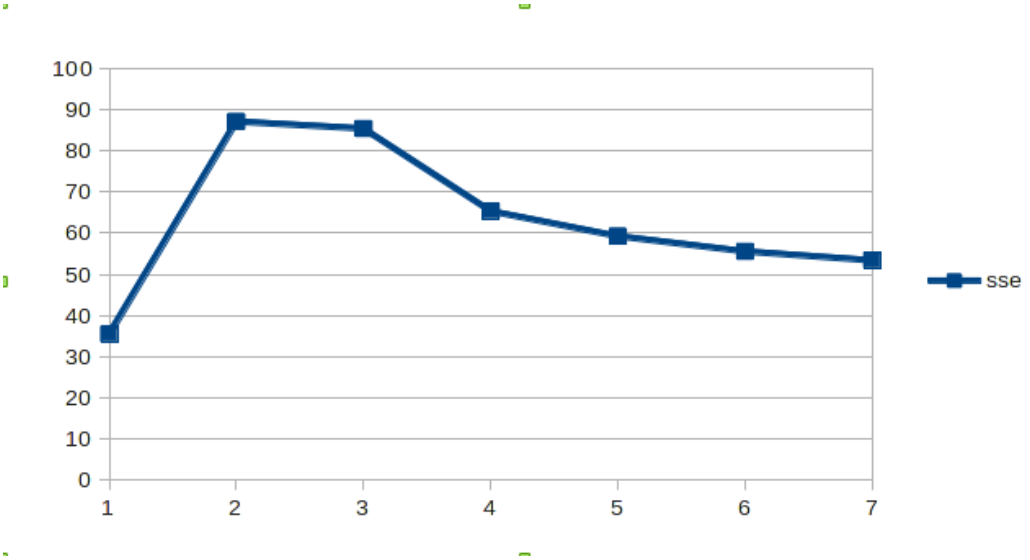

Figure 6: Shows the sse values for the second file

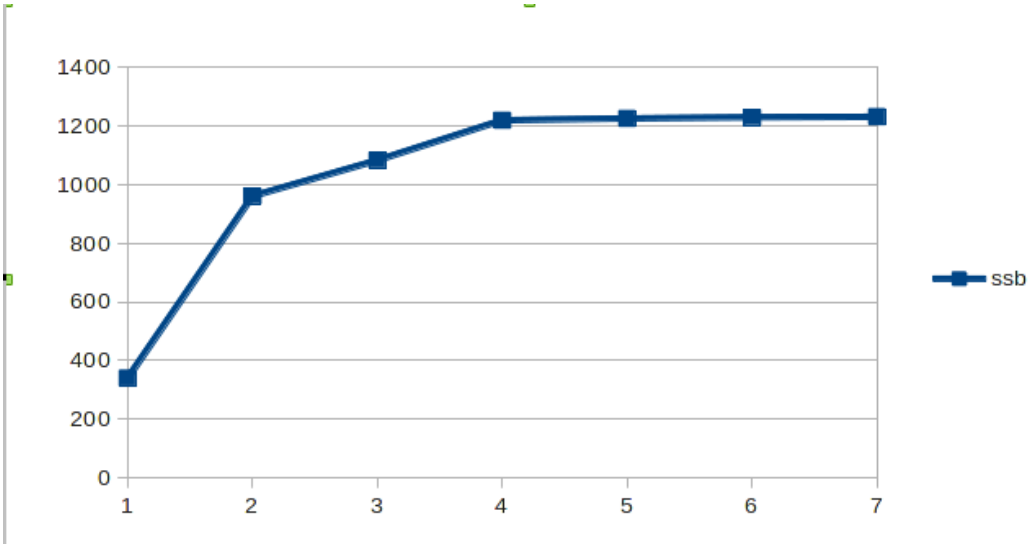

Figure 7: Shows the ssb values for the second file.

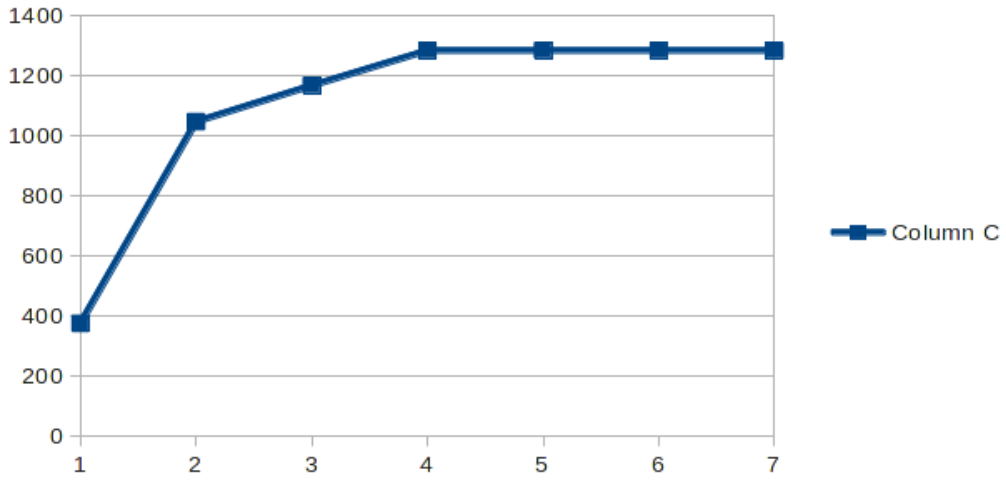

Figure 8: Shows the sum of the scond file.