Nick Handelman (nrh8883)

March 21, 2012

CMPS 509

Program 1: Clustering Report

**Part 1: Derivation of Which Metric to Implement**

My CLID is "nrh8883" so I will let n0 = 8, n1 = 8, n2 = 8 and n3 = 3.

This gives n = (8\*8\*8\*3) mod 3 = 0.

Also, the number 8883 mod 3 gives 0.

Therefore, I will implement option 0: Calinski-Harabasz (CH) Index.

**Part 2: Source Code**

The source code can be found on the following pages.

**Part 3: Numerical Results**

The following are the numbers derived from the CH Index using the given data sets with clusters K = 1, 2, 3, 4, 5, 6, 7, 8 and a line plot of those numbers.

|  |  |
| --- | --- |
| K - clusters | Calinski-Harabasz Index |
| 1 | Undefined |
| 2 | 1487.702637 |
| 3 | 1436.425049 |
| 4 | 2362.874268 |
| 5 | 1398.365967 |
| 6 | 1586.669678 |
| 7 | 1722.878662 |
| 8 | 1287.726807 |

**Part 4: Analysis and Discussion of Results**

The CH Index is a normalized ratio of within-cluster scatter to between-cluster scatter. Throughout my analysis and discussion, I will refer to the formula for the CH Index and its variables. They are as follows:

N: number of data samples

K: number of clusters

B: between-cluster scatter

W: within-cluster scatter

The goal of the equation is to maximize the CH Index value, as seen through the influence of variables B and W. In the numerator, we want B to be as large as possible, to maximize the CH Index. This makes sense because clusters that are spread out represent the data better, and this spread will cause B to be greater. However, in the denominator, we want W to be as small as possible to maximize the CH index. Again, this makes sense because tightly bound clusters will have samples that are closer to the cluster center, causing W to be smaller.

In the case of one cluster, K = 1, it is obvious from the CH Index formula that the CH Index value is undefined, due to a division by zero. However, since it may be the case for some data set that one cluster is in fact the best way to represent the data, the Calinski-Harabasz Index falls short in this respect. If the data indicate that one cluster may be the best representation (perhaps a data set gives CH Index values that are all roughly equal for the tested clusters) then I would recommend using a different index to determine if one cluster offers the best representation.

Since our goal is to maximize the CH Index value, we can see from the data that four clusters offers the best representation of the data, whereas eight clusters offers the worst representation of the data. Further analysis reveals that the minimum (eight clusters) and 2nd-largest (seven clusters) CH Index values have a difference of about 600, whereas the 2nd-largest (seven clusters) and largest (four clusters) CH Index values also have a difference of about 600. This analysis serves to further prove that four clusters offers the best representation of the data, and does so much better than its competitors, in particular due to it being numerically approximately 33 percent "better" than its nearest competitor and 100 percent "better" than its worst competitor.