**Homework 4 Report**

Zach Brueske, Spencer Hanson, Auston Schendzielos

This table shows how the K value and the Distance Measure used effected the clusters that were generated and uses WSS to define the quality of the clusters for the Iris data set.

|  |  |
| --- | --- |
| Distance Measure| K Value | WSS |
| Euclidean| K = 3 | 160.2465037 |
| Euclidean| K = 6 | 12.1700177 |
| Euclidean| K = 9 | 9.179020357 |
| Manhattan| K = 3 | 79.14342874 |
| Manhattan| K = 6 | 42.22612894 |
| Manhattan| K = 9 | 28.86413933 |

This table shows how the K value and the Distance Measure used effected the clusters that were generated and uses WSS and BSS to define the quality of the clusters for the Iris data set.

|  |  |  |  |
| --- | --- | --- | --- |
| Distance Measure| K Value | WSS | BSS | BSS/WSS |
| Euclidean| K = 3 | 160.2465037 | 5.59E+13 | 2.51E+13 |
| Euclidean| K = 6 | 12.1700177 | 5.50E+13 | 2.75E+12 |
| Euclidean| K = 9 | 9.179020357 | 1.01E+13 | 4.19E+11 |
| Manhattan| K = 3 | 79.14342874 | 5.28E+08 | 8232015.058 |
| Manhattan| K = 6 | 42.22612894 | 1.25E+08 | 2790788.048 |
| Manhattan| K = 9 | 28.86413933 | 3.93E+07 | 1078025.896 |

This table shows how the K value and the Distance Measure used effected the clusters that were generated and uses the Total Weighted Entropy of the clusters to define the quality of the clusters for the Iris data set.

|  |  |
| --- | --- |
| Distance Measure| K Value | Total Weighted Entropy |
| Euclidean| K = 3 | 0.468908165 |
| Euclidean| K = 6 | 0.21062696 |
| Euclidean| K = 9 | 0.107830001 |
| Manhattan| K = 3 | 0.257048963 |
| Manhattan| K = 6 | 0.074175605 |
| Manhattan| K = 9 | 0.067504937 |

This table shows how the K value and the Distance Measure used effected the clusters that were generated and uses WSS to define the quality of the clusters for the AllGenes data set.

|  |  |
| --- | --- |
| Distance Measure| K Value | WSS |
| Euclidean| K = 2 | 8.45E+18 |
| Euclidean| K = 4 | 6.43E+18 |
| Euclidean| K = 6 | 5.11E+18 |
| Manhattan| K = 2 | 2.17E+11 |
| Manhattan| K = 4 | 1.87E+11 |
| Manhattan| K = 6 | 1.67E+11 |

This table shows how the K value and the Distance Measure used effected the clusters that were generated and uses WSS and BSS to define the quality of the clusters for the AllGenes data set.

|  |  |  |  |
| --- | --- | --- | --- |
| Distance Measure| K Value | WSS | BSS | BSS/WSS |
| Euclidean| K = 2 | 8.45E+18 | 1.78E+29 | 2.97E+09 |
| Euclidean| K = 4 | 6.43E+18 | 6.58E+27 | 3.75E+08 |
| Euclidean| K = 6 | 5.11E+18 | 5.06E+26 | 2.47E+08 |
| Manhattan| K = 2 | 2.17E+11 | 4.73E+14 | 346513.298 |
| Manhattan| K = 4 | 1.87E+11 | 1.23E+16 | 11632.17883 |
| Manhattan| K = 6 | 1.67E+11 | 2.42E+15 | 8276.412872 |

This table shows how the K value and the Distance Measure used effected the clusters that were generated and uses the Total Weighted Entropy of the clusters to define the quality of the clusters for the AllGenes data set.

|  |  |
| --- | --- |
| Distance Measure| K Value | Total Weighted Entropy |
| Euclidean| K = 2 | 0.766012495 |
| Euclidean| K = 4 | 0.632436756 |
| Euclidean| K = 6 | 0.703546198 |
| Manhattan| K = 2 | 0.875826268 |
| Manhattan| K = 4 | 0.753596759 |
| Manhattan| K = 6 | 0.624235617 |

This table shows how the K value and the Distance Measure used effected the clusters that were generated and uses WSS to define the quality of the clusters for the SigGenes data set.

|  |  |
| --- | --- |
| Distance Measure| K Value | WSS |
| Euclidean| K = 2 | 4.46E+17 |
| Euclidean| K = 4 | 3.20E+17 |
| Euclidean| K = 6 | 2.80E+17 |
| Manhattan| K = 2 | 1.21E+10 |
| Manhattan| K = 4 | 9.28E+09 |
| Manhattan| K = 6 | 7.31E+09 |

This table shows how the K value and the Distance Measure used effected the clusters that were generated and uses WSS and BSS to define the quality of the clusters for the SigGenes data set.

|  |  |  |  |
| --- | --- | --- | --- |
| Distance Measure| K Value | WSS | BSS | BSS/WSS |
| Euclidean| K = 2 | 4.46E+17 | 1.24E+26 | 1.07E+09 |
| Euclidean| K = 4 | 3.20E+17 | 1.07E+25 | 2.27E+07 |
| Euclidean| K = 6 | 2.80E+17 | 1.85E+24 | 9556007.859 |
| Manhattan| K = 2 | 1.21E+10 | 5.31E+14 | 49540.98415 |
| Manhattan| K = 4 | 9.28E+09 | 1.38E+14 | 15317.77891 |
| Manhattan| K = 6 | 7.31E+09 | 2.78E+13 | 7840.649947 |

This table shows how the K value and the Distance Measure used effected the clusters that were generated and uses the Total Weighted Entropy of the clusters to define the quality of the clusters for the SigGenes data set.

|  |  |
| --- | --- |
| Distance Measure| K Value | Total Weighted Entropy |
| Euclidean| K = 2 | 0.641985103 |
| Euclidean| K = 4 | 0.63945851 |
| Euclidean| K = 6 | 0.69784149 |
| Manhattan| K = 2 | 0.698794421 |
| Manhattan| K = 4 | 0.648197503 |
| Manhattan| K = 6 | 0.544331013 |

For the Iris data set the best measure and k combination for WSS was k = 9 using the Euclidean distance measure (5.762023569). For BSS/WSS the best combination was k = 9 using the Euclidean distance again (4.19E+11). The combination with the best weighted entropy was k = 3 using the Manhattan distance measure (0.467775706). We got the best information gain from the combination k = 6 and using Euclidean distances (0.933542586).

For the AllGenes data set the best measure and k combination for WSS was k = 6 using the Manhattan distance measure (1.68E+11). For BSS/WSS the best combination was k = 4 using the Euclidean distance (3.75E+08). The combination with the best weighted entropy was k = 2 using the Manhattan distance measure (0.886410013). We got the best information gain from the combination k = 6 and using the Manhattan distances (0.315226092).

For the SigGenes data set the best measure and k combination for WSS was k = 2 using the Manhattan distance measure (1.20E+10). For BSS/WSS the best combination was k = 4 using the Euclidean distance (2.27E+07). The combination with the best weighted entropy was a tie between

k = 2 using the Manhattan distance measure and k = 2 using the Euclidean distance measure (0.705412755). We got the best information gain from the combination k = 4 and using Manhattan distances (0.397281389).

The three validity measures appear to agree the best when K is equal to double the amount of class labels. (K=4, K=6, etc.) This doesn’t give us the most optimal values for each measure of validity, but it seems to give a good number for each measure.

The Significant Genes data set seemed to cluster better using the K-Means algorithm. The average information gain for SigGenes is .3457, while the average information gain for AllGenes is .2169. The average WSS for SigGenes is 1.87E+17, and the average WSS for AllGenes is 3.50E+18. Finally, the average weighted entropy for SigGenes is .6542, and the average weighted entropy for AllGenes is .7830. Even though the weighted entropy was higher for AllGenes, this could be due to reasons such as outliers or the fact that there are a significant amount more genes in the AllGenes data set, thus skewing the effect.