Contents

Table of Figures

PART 1: Data examination and preprocessing

(1) Review the data and pre-process it, if necessary. Any features that should not be used?

Since we do not have labels (classes) for the data (that is the task for this case study), we limited this part to data scaling. This step was done for two reasons: a) the plan is to use some distance metrics to find out how similar exemplars are and, b) in some cases, the dynamic range (max - min value) amongst the features is significantly different and therefore, they will dominate any distance calculation when clustering the data. To this effect, the data was scaled about its mean in units of the standard deviation.

Only the physical features were used per the professor's suggestion.

PART 2: MBSAS Clustering

(2a) Apply clustering algorithm to estimate the number of intrinsic clusters in the data. Identify several subsets of features focused on different observed characteristics of the GRBs. For example, time dependent features or energy related features. Compute a separability measure to assess the quality of the clustering estimate across the classes.

The MBSAS clustering algorithm requires us to define two parameters: the threshold of dissimilarity (theta) and the maximum allowable number of clusters (q). Since the number of clusters in this data is unknown, we computed the number of clusters formed for different numbers of dissimilarity (q vs theta):

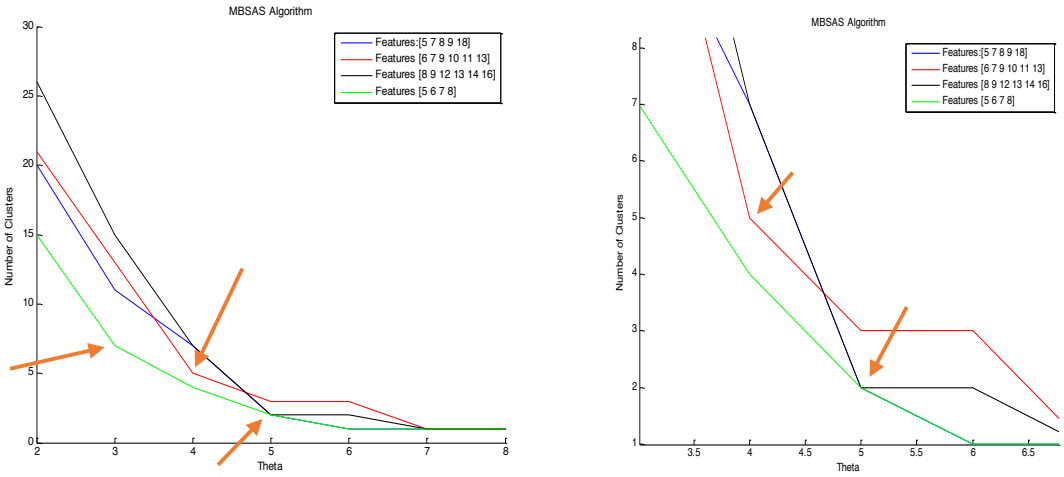

Figure 1 MBSAS algorithm. Left: number of clusters vs dissimilarity value. Right: Zoomed in version.

Let us denote the utilization of features 5, 7, 8,9,18 as case 1, features 6, 7, 9,10,11,13 as case 2, features 8,9,12,13,14,16 as case 3, and features 5,6,7,and 8 as case 4. It can be seen that for case 1, the number of clusters seems to be 2. For case 2 it is 5, for case 3 it is 2, and for case 4 it is 6. In essence, we looked for flat regions (i.e. case 1 and 3) or drastic changes in slope (i.e. case 2 and 4)

The averaged J3 value that measures the class separability in each case was:

Case 1: 0.030107 dB
Case 2: 146.3072 dB
Case 3: 0.83053 dB
Case 4: 2.3567 dB

Note that we use the logarithm scale since the range between them was very large. Based on this number we can conclude that case 2, which "hypothesizes" that there are 5 clusters, had the largest class separability. Case 4 (6 clusters) had the second largest J3 value. On the other hand, case 1 and 2 had a smaller J3 value. This results seem to suggest that the number of inherent clusters in the data may be somewhere between 3 and 7. We will have to see what other methods "tell" us.

The reasoning behind our feature selection was due to their histogram shape, especially feature #8. It can be seen some bimodality, which could be associated to sensible clusters.

PART 3: k-Means Clustering

(2b) Apply clustering algorithm to estimate the number of intrinsic clusters in the data. Identify several subsets of features focused on different observed characteristics of the GRBs. For example, time dependent features or energy related features. Compute a separability measure to assess the quality of the clustering estimate across the classes.

The following plot shows the average J3 value across different number of clusters for each feature set:

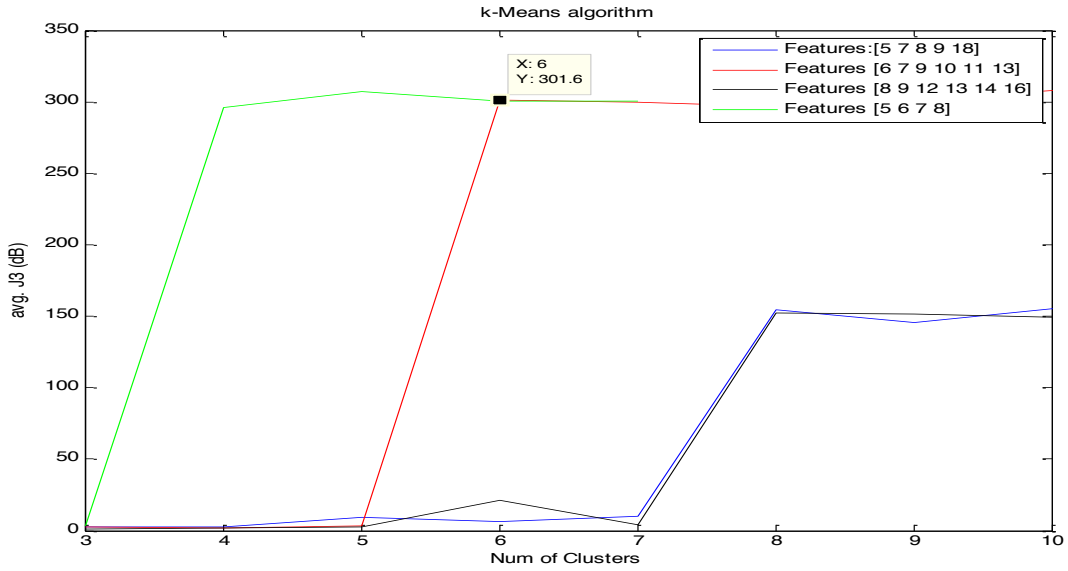

Figure 2 k-Means average J3 values

It can be seen that case 1 and 3 agree to some extend: they both suggest that there are 8 clusters. On the other hand, case 1 and 3 suggest that there are 4 and 5 clusters, respectively. Note that when we use the term "suggest", we mean that average J3 value is maximized near that number of clusters.

PART 4: Fuzzy C-Means Clustering

(2c) Apply clustering algorithm to estimate the number of intrinsic clusters in the data. Identify several subsets of features focused on different observed characteristics of the GRBs. For example, time dependent features or energy related features. Compute a separability measure to assess the quality of the clustering estimate across the classes.

The following plot shows the different J3 values across different number of clusters for each feature set:

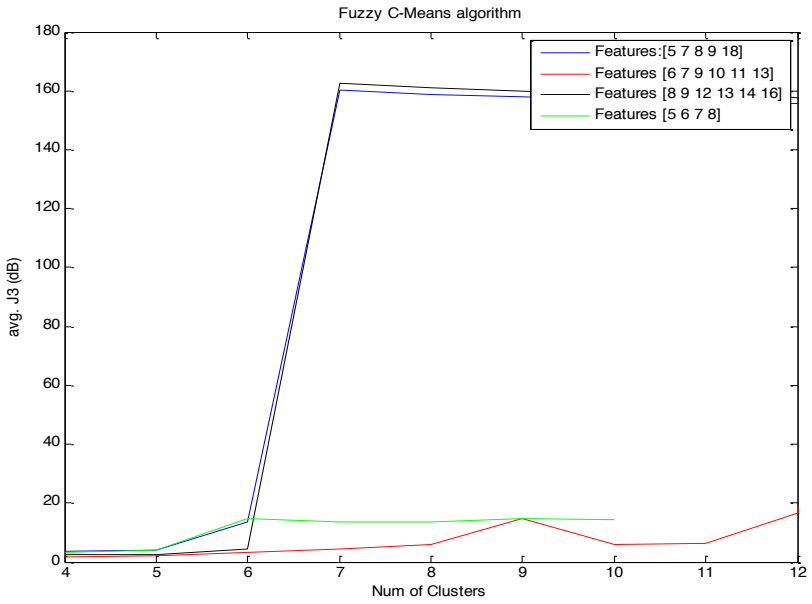

Figure 3 Fuzzy C-Means average J3 values

Again, this clustering algorithm "suggests" that there may be 6 or 7 clusters in the data. Case 2 (red), suggests that there may be 9 clusters. Again, we looked for drastic changes in the slope and local maximums.

PART 5: Results Summary & Comparison

(2c) Apply clustering algorithm to estimate the number of intrinsic clusters in the data. Identify several subsets of features focused on different observed characteristics of the GRBs. For example, time dependent features or energy related features.

The following table summarizes the number of clusters that each algorithm suggests based on the averaged J3 value.

Case # (features)	MBSAS (avg. J3 value in dB)	k-Means(avg. J3 value in dB)	Fuzzy C-Means(avg. J3 value in dB)
Case 1 (5,7,8,9,18)	2 (0.03)	8(154)	7 (160)
Case 2 (6,7,9,10,11,13)	5 (146)	6 (301)	9 (14)
Case 3 (8,9,12,13,14,16)	2 (0.83)	8 (152)	7 (162)
Case 4 (5,6,7,8)	6 (2.35)	4 (295)	6 (14)

Figure 4 Table with results

MBSAS Cluster distribution:	k-Means Cluster distribution:	k-Means(avg. J3 value in dB)
Case #1 Cluster#1:96 Cluster#2:3 Case #2 Cluster#1:48 Cluster#2:3 Cluster#3:3 Cluster#4:2 Cluster#5:43 Case #3 Cluster#1:83 Cluster#2:16 Case #4 Cluster#1:21 Cluster#2:16 Cluster#3:31 Cluster#4:5 Cluster#5:2 Cluster#6:12	Case #1 Cluster#1:3 Cluster#2:2 Cluster#3:22 Cluster#4:33 Cluster#5:6 Cluster#6:7 Cluster#7:19 Cluster#8:7 Case #2 Cluster#1:17 Cluster#2:5 Cluster#3:27 Cluster#4:11 Cluster#5:14 Cluster#6:25 Case #3 Cluster#1:10 Cluster#2:2 Cluster#3:16 Cluster#4:3 Cluster#5:14 Cluster#6:17 Cluster#7:11 Cluster#8:26 Case #4 Cluster#1:12 Cluster#2:17 Cluster#3:20 Cluster#4:50	Case #1 Cluster#1:29 Cluster#2:2 Cluster#3:14 Cluster#4:2 Cluster#5:8 Cluster#6:23 Cluster#7:21 Case #2 Cluster#1:7 Cluster#2:13 Cluster#3:2 Cluster#4:26 Cluster#5:5 Cluster#6:17 Cluster#7:7 Cluster#8:2 Cluster#9:20 Case #3 Cluster#1:29 Cluster#2:2 Cluster#3:14 Cluster#4:2 Cluster#5:8 Cluster#6:23 Cluster#7:21 Case #4 Cluster#1:23 Cluster#2:30 Cluster#3:20 Cluster#4:16 Cluster#5:2 Cluster#6:8

Figure 5 Cluster Distribution for each algorithm and case

Based on figure 4, the case that maximizes the average J3 value is case 2 with the k-Means algorithm, which suggests there are 6 clusters. However, case 4 using Fuzzy C-means also produces one of the

lowest average J3 value. When the clustering algorithms produce 5 clusters, the average J3 value decreases. Similarly, when the number of clusters is 7, the value decreases as well.

The most important observation is that when the number of clusters is 7 or 8, regardless of the clustering algorithm, a similar J3 value is computed. This leads us to believe that there are 7 or 8 inherent clusters in the data. After inspecting the clutter distribution of 7 vs. 8, we look for consistency among the cases and algorithms. This leads us to believe that there are 8 clusters in the data.

Of the three methods, k-Means seems to provide the most consistent results