CALIFORNIA STATE UNIVERSITY, LONG BEACH

College of Engineering
Department of Computer Engineering and Computer Science
Dr. Thinh V. Nguyen
Spring 2008
CECS-550/650: Pattern Recognition

PROJECT 2

Name: Foss, Shannon

Last, First

- **Dates:** Date assigned: Thursday April 10, 2008. Date due: Wednesday May 15, 2008. No late submission will be accepted. This project is worth 70% of the project grade.

- **Objectives**: The objectives of this project include: (1) to familiarize students with the basic clustering and data collection techniques, (2) to perform simulation of various types of clustering algorithms.

- **Project Description:**

A) (30 points) Collect data for clustering. The data are features related to world countries (excluding territories, colonies, and dependencies). The objective is to analyze the relationship among the world countries to understand their similarities or dissimilarities.

All world countries (195) must be included from Afghanistan to Zimbabwe. This should include Taiwan.
The required features are: geographical area (in square miles), population, and literacy rate. Additional features must be included for a total minimum of seven features (including the above three required features). Suggested additional features are: GDP per capita, infant mortality rate, life expectancy, trade exports and imports, etc.. When information is not available, it is acceptable to interpolate/extrapolate using any reasonable scientific methods. The features selected should be as independent as possible. For example, do not use population density because it can be derived from area and population. Feel free to generate any features that may be meaningful. For example, you may want to create a feature called World War (WW) 1 participants, or WW2 participants, etc. The feature values may be real numbers, integers, or boolean. However, for purpose of clustering, these values will be eventually coded in a consistent format (e.g., normalized real numbers).

Present the raw data in tables. Cite sources where applicable.

B) (30 points) Write a computer program to:

1) Perform data normalization to convert feature values into reasonable data set for clustering purposes. Use any reasonable techniques but you must explain and discuss the rational of the selected technique. For example, you may want to express all feature values to range from 0 to 1 where 1 represents the largest value. You may use scale factors as appropriate for the feature values. Discuss the effect of data normalization in the analysis (section C).

2) Perform clustering on the feature vectors using AT LEAST a), d) and g) in the following algorithms:
   a) Basic sequential algorithmic scheme (BSAS)

b) Modified BSAS
c) Two-Threshold sequential algorithmic scheme (TTSAS)
d) Generalized agglomerative scheme (GAS)
e) Matrix updating algorithmic scheme (MUAS): single link algorithm, complete link algorithm, WPGMA, UPGMA, UPGMC, WPGMC, Ward or minimum variance.
f) Generalize Divisive Scheme (GDS)
g) ISODATA or k-means algorithm.

If you implement more than the minimum a), d) and g), the additional work will receive extra credit.

For all of the above, use any reasonable refinement schemes as appropriate.

C) (40 points) Analyze the results. Evaluate the quality of the clustering. Discuss the effects of the features, the scale factors, or the data normalization. Provide tables or plot curves to show comparisons. For example, you may plot diagrams of the clusters using 2 features at a time or 2 most dominant features. Discuss any unexpected results.

The project report must include the above three sections (A, B, and C) clearly labeled and/or partitioned. For part B, no program listings should be included, but you must provide general algorithm descriptions, pseudo-code, or any implementation issues. Be concise but complete.

Attach these two sheets as the first two sheets of the project report.

- **Useful links**: The following links may be useful.

  http://www.infoplease.com/countries.html
  http://www.geohive.com/default1.aspx
  http://www.uis.unesco.org/ev.php?ID=2867_201&ID2=DO_TOPIC
  http://www.wto.org/english/res_e/statis_e/its2007_e/its07_toc_e.htm

END OF DOCUMENT

Shannon Foss
May 20, 2008
CECS 550
Project 2

Pattern Recognition:
Data Analysis and Clustering

This project was an attempt to collect relevant data about the countries of the world in order to cluster them together so that the relationships between them may be found and analyzed.

Project Description:

Part 1 – Data Collection:

The data for the 195 recognized world countries is collected to fully represent different aspects of the countries, such as geographic area, population, education, economic prosperity, overall health, government, society and other such demographics. This information needs to be as independent from each other as possible to get an even representation of the data, if one feature is too dependent on another (ex. Population and Birth rate), then the features will seemingly have an extra weight to them and the clustering will be more heavily reliant on that aspect. It is not always obvious when a feature is dependent on another, and so we must run calculations to be able to discover them.

Part 2 – Clustering:

In order to be able to compare the data that we have collected, it must first be normalized, this way all of the data is evaluated between the same range and is on equal bounds.

Clustering is performed after normalizing the data and this project will use several different algorithms to accomplish this. Afterwards, the results will be analyzed and compared to see which clustering algorithm performed the best.

Project Background:

Part 1 – Data Collection

The data was collected from several different sources: mainly from the CIA's World Factbook, and the United Nations Statistics Division's Common Database. These two sources had a great deal of topics to choose from and had information on nearly every country. Once in a while though, there was missing data for a country, and so the data was either found from another site, a previous year, or when desperate, a nearby or related country (ex. the literacy rate for the Solomon Islands could not be found, and so the average of three nearby island countries in the same area, Fiji, Vanuatu, and Papua New Guinea was used). Some countries, such as Myanmar and Kosovo, did not seem to have much if any data, but after some quick research about the country found that Kosovo had just recently split off of Serbia, and Myanmar is also better known as Burma. Other times, there were interesting features that could have been included but did not come near to having the data from the 195 countries, and so these features, unfortunately, had to be left out completely.

In the end I was able to get complete data listings for 16 different features. These include: geographical area, population, literacy rate, GDP, average life expectancy from birth, net migration, percent of GDP toward military spending, percentage of the population that has access to improved drinking water, number of patent applications submitted annually, satisfaction with life index, amount of chocolate imported annually in dollars, percentage of the population that is religious, if the country has a monarchy, if the country has the death penalty, if the country has created nuclear weapons, and the number of people with access to the internet.

These features were picked to represent different the aspects of countries that embody what type of people live there. Features such as the average life expectancy and the percentage of people with access to improved drinking water represent the health and quality of life of the people. Where as literacy rate, the number of patent applications submitted annually, and the number of people with access to the internet can represent the amount of education and creativity a country has. Features such as GDP, amount of chocolate imported annually, and the number of people with access to the internet might represent how a country is doing economically. If the country has a monarchy, if a country has the death penalty or how much of the GDP is going toward military spending might demonstrate what type of government a country has. Whereas features such as percentage of the population that is religious, satisfaction with life index, and the amount of chocolate imported annually would show what type of society that country has. It is clear to see that some features can represent multiple country aspects as well.

Once a representative set of features is compiled, we must determine how independent they are. Excel (which is where all of the data was compiled at) has a very nice function, CORREL(array1, array2), which returns the correlation coefficient between the two given arrays.

$$Correl(X,Y) = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}}$$

Equation 1: Correlation Function equation in Excel.

This function will determine, by percentage, if the two given features are dependent on each other.

Running all 16 features through the correlation function returned a very high percentage between GDP and the number of people with access to the internet, 97.1%, which makes sense, the internet is something you can only access with a computer, and most of the time a computer is a luxury item only available to people who have the funds. It is clear that one of these will have to be removed. GDP is obviously very important because it is a clear indicator of a nations wealth, but the number of internet users could represent several aspects of a country, not only of economic, but also social. Comparing both of these features to the other features will be required to eliminate one of them. GDP's next highest correlation is with the amount of chocolate imported, at 65.5%, which again, would seem to make sense, chocolate is a inexpensive luxury item. Internet users next highest correlation is with geographical area, at 68.2%, which only makes as much sense as population being correlated with geographical area; the more space, the more people can fit inside it. With these factors and including that GDP was a more complete data sample than internet users was, we can go ahead and exclude the data set for internet users.

The next highest feature correlation was between the percent of people with access to improved drinking water and average life expectancy, at 72.3%. Both of these are highly related to overall health. All of the correlations between these and the other features are extremely similar, so it seems like either of these features may be easily removed. Again, since the access to improved drinking water feature had a less complete data set we will remove it from the feature set.

Another feature correlation with a high percentage was between life expectancy and literacy rate, but since now life expectancy is the only health feature, and literacy rate is the main feature for education, we should keep both of these.

This process is continued until it is believed that we have a data set with a good representation of the countries. After eliminations 11 features were kept for testing: geographical area, population, literacy rate, GDP, life expectancy, net annual migration, percent of GDP spent toward military expenditures, number of annual patent applications, percentage of religious people, if the country has a monarchy, and if the country has the death penalty.

Part 2 – Clustering:

Before clustering can be performed, the data must be normalized, this is so the data can be relatable to each other. For example, if one feature set consists of Boolean values and another set consists of integers, then it can be difficult to compare them in a meaningful way. But, if we normalize all of the data so that all of the values are represented in values between 0 and 1, then it is possible to relate the features meaningfully.

After the data is normalized, clustering may be performed. The clustering will be executed by three different algorithms, and their results compared.

Basic Sequential Algorithmic Scheme (BSAS) is a sequential algorithm that makes a single pass through the data samples, and assigning each to a cluster based on its distance from the cluster. This algorithm is highly dependent on the ordering of the samples presented to the algorithm.

Generalized Agglomerative Scheme (GAS) is a hierarchical agglomerative algorithm that starts out with each sample assigned to its own class. In each successive iteration it condenses one cluster into another until a threshold is reached.

K-Means is an algorithm that clusters samples based on the distance between the data point and the cluster's mean. As points are added to a cluster, the mean changes and so does the distance between the samples and the cluster's mean. This algorithm iteratively reassigns samples until it converges on a clustering or a set limit is reached.

Algorithm Descriptions:

Normalization:

The algorithm for normalizing the data is fairly simple. First you must find the maximum and minimum values for each feature. Then for each data point, d in equation 2, you subtract the minimum value of that feature, and then divide by the difference between the maximum and minimum feature values.

$$\delta = \frac{d - d^{\min}}{d^{\max} - d^{\min}}$$

Equation 2: Normalization transform.

Distance Measure:

To find the distance between two clusters or a cluster and a point, we need a distance measure. One simple distance measure is Euclidian Distance. It takes the square root of the sum of the squares of differences between the feature vectors.

$$d = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

Equation 3: Euclidian Distance.

BSAS:

In BSAS, there is only a single run through the data, as such, the order in which you present the samples (countries) can affect the clustering. The first cluster is seeded with the first sample, and then the distance to each successive sample is calculated between the existing clusters and placed in the closest cluster. If there is no cluster within the threshold distance, then a new cluster is created up until a cluster limit is reached. This process repeats until all of the samples are placed into a cluster.

Algorithm 1: BSAS.

1. Loop through all samples.

2. If the shortest distance from a sample to a cluster C_k is greater than Θ and the amount of clusters is smaller than q, then assign the sample to a new cluster.
3. Else assign the sample to cluster C_k.

GAS:

The GAS algorithm is a relatively simple set of steps; you first assign all of your samples to their own clusters, and come up with a threshold value. Then, iteratively find the closest pair of clusters and compare it to the threshold value. If the distance between the two clusters is smaller than the threshold, then merge those two clusters. If it is not, then the algorithm is finished and your clustering is complete.

Algorithm 2: GAS.

1. Find the closest pair of clusters.
2. If the distance between the two clusters is smaller than the threshold value, then the two clusters are merged into one and the algorithm continues.
3. Else, if the distance between the two clusters is larger than the threshold value, then the clustering stops.

K-Means:

With the K-Means algorithm, you need to know how many clusters you want ahead of time. Also, you need to figure out a threshold value to determine when the algorithm has converged. Firstly, the given number of clusters are seeded with random values as their centers, these are the mean vectors that the points will be compared to. Each of the samples are placed into the closest cluster. After a sample has been assigned to a cluster, that cluster's mean vector is updated to reflect the addition. At the end of each iteration, the distance that the mean vectors have changed is compared to the threshold value. If this value is larger than the threshold, then another iteration is performed. Otherwise, clustering ends.

Algorithm 3: K-Means.

1. Initialize the given number of clusters with random seeds.
2. For each of the samples:
    a. Measure the distance between the sample and all of the clusters.
    b. Find the minimum distance.
    c. Assign the sample to that cluster.
    d. Update mean vector.
3. Calculate the distance all of the mean vectors have moved.
4. If this amount is greater than the threshold, repeat from 2.
5. Else end.

Results and Analysis:

To analyze the results of the clustering, we need to decide what a good clustering is. Since we cannot meaningfully represent more than three features in a clustering graph we need to find a way to see good clusters. This will probably only extend to picking a few countries that we would expect to see grouped together, such as: many European countries, the island countries in Oceania, and some South American countries. If we look for these groupings in our results we might be able to tell if it is a good clustering.

Results pre-feature selection (16 features, threshold = 0.5 and 0.25) –

For the BSAS clustering at a threshold of 0.5 and unlimited clustering, we can find some of the good clustering features we are looking for. In cluster #23 we see many of the Oceania island groups such as Fiji, The Solomon Islands, Papua New Guinea, and Vanuatu, but then we see that Samoa and Tonga were placed in cluster #13. In cluster #10 we see Belgium, Denmark, Luxembourg, Netherlands, Spain and

Sweden, a very good cluster. However, just a few clusters down we see that Norway, a country you might think fit in well with cluster #10 is in cluster #13 with two of the island countries.

In BSAS with a threshold of 0.25 and an unlimited clustering, many of the countries get too far spread out to have any kind of meaningful clustering. However, if the clustering is limited, it can force countries to be contained in a way that seems to defeat the purpose of having the threshold.

In the GAS clustering at a threshold of 0.5, it is noticeable that the same clustering of European countries from BSAS cluster #10 is now in cluster #16 of GAS, except now it does include Norway in the list. It also seems to have a good number of South American Countries in cluster #1. But it also seems to have spread out many individual countries into separate clusters, many of which could be grouped together.

With GAS at a threshold of 0.25, we again have the same problem as BSAS where the countries have become too far spread out within the clusters to have any kind of meaningful cluster.

The K-Means clustering at clusters=25 and a threshold of 0.5, we can see again that the European countries are grouped together in cluster #11, with a few extras. Many Middle Eastern countries have been grouped together in cluster #22, again with a few extras. These few extras in each group may be due to the restrictive number of clusters that is a part of K-Means, and so, many countries that may have started their own clusters have been forced into clusters that already existed.

K-Means with a threshold of 0.25 seems to have a better clustering than with the threshold of 0.5. Many of the clusters have a well-distributed number of countries, and they are often well matched.

While creating the code for K-Means, I ran into the problem that for number of clusters that were set at a high value, many of the clusters would be empty when the results came back. This is because as clusters pulled data points from other clusters they might empty out that cluster. When this happens the cluster stays empty for the rest of the clustering because it no longer has a varying mean vector that is near any other points to grab. However, because it is not restricted from creating empty clusters, it has the ability to pull whichever data points it needs to create a good clustering.

This new K-Means with clusters=25 and threshold of 0.5 is highly similar to the actual K-Means. It is difficult to say whether it is any better or worse, but it does have more countries that seem to be related in most clusters.

The new K-Means at threshold of 0.25 is, again, very similar to its K-Means counterpart, however it created a few larger clusters rather than the well distributed clusters it had.

Results after Feature Selection (11 features, threshold = 0.25) –

In the BSAS clustering with threshold at 0.25, the number of clusters dropped significantly, to 80 from the previous 127. Clusters such as #13 did well grouping Northern European countries, also cluster #9 did well grouping island nations. But as before, many of the countries were too well spread out among the clusters to have any meaningful grouping.

The GAS clustering at threshold=0.25, had a good clustering of Northern European countries, but in cluster #1 it grouped Eastern European countries with several South American and Oceania countries. This grouping came out strange because it had another grouping of Eastern Europe countries at cluster #66.

K-Means at a clustering of 10 and a threshold at 0.25 also provided the strange clustering of South American countries mixed in with Eastern Europe and Oceania countries. Cluster #6 returned a decent clustering of Europe, but even France and the U.K. were not together.

It seems as though the quality of clustering dropped somewhat when several features were removed in the feature selection process. Perhaps a feature that was redundant was amplifying another feature to create a higher weighted vector, which created a better clustering.

It also seems that some algorithms are more sensitive to the threshold value than others. For BSAS the higher the threshold value, the more countries that are placed into a cluster. This creates a better clustering than when the threshold is lower and the set number of clusters is higher because countries are not forced into cluster that they may not belong to. GAS has about the same sensitivities as BSAS but it is better at clustering the countries overall. K-Means, on the other hand, seems to do better with the slightly smaller threshold values.

Clustering Comparison:

To get a visual representation of the clustering, we must narrow down the features to just two. This way we can plot them along the axis and compare them against each other. These following graphs are of the Literacy Rate along the x-axis and GDP along the y-axis. Clustering was limited to 10 and the threshold was set at 0.25.

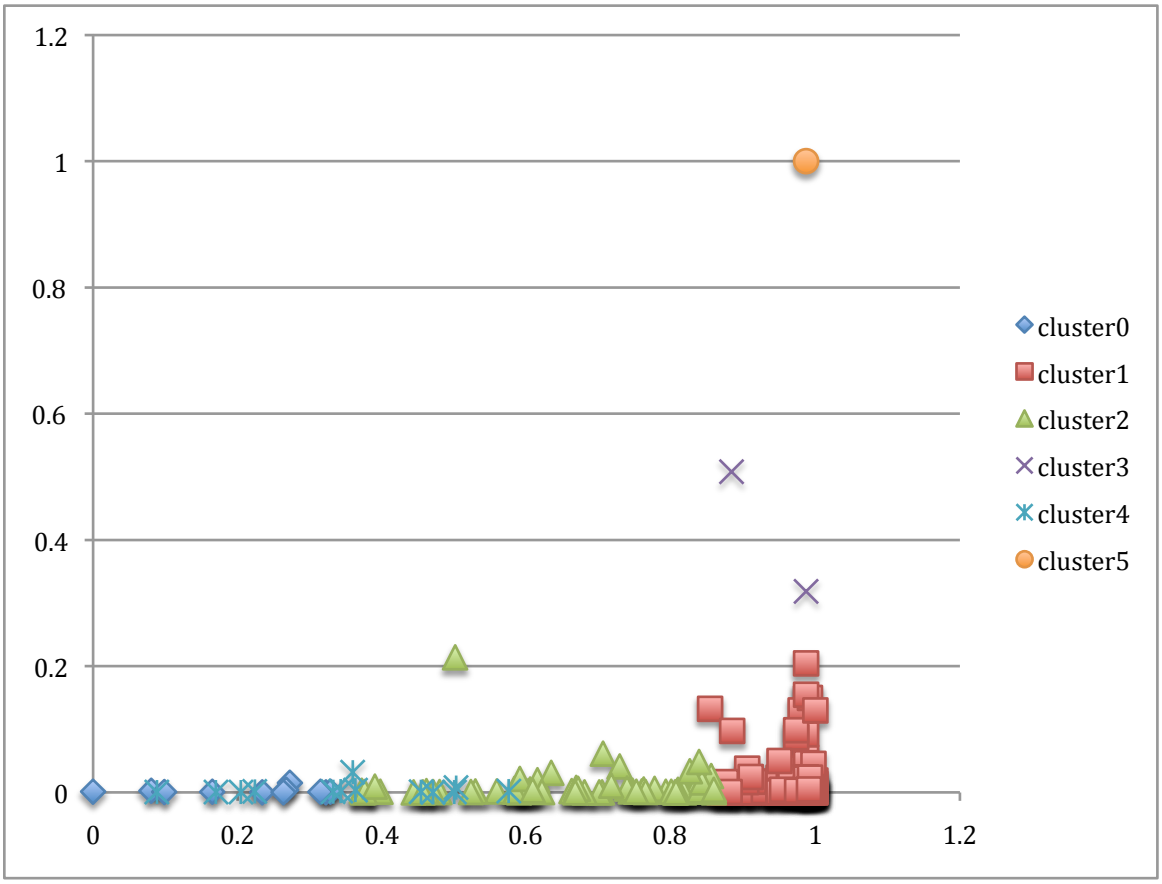

Figure 1: BSAS clustering on literacy rate and GDP.

As you can see in the BSAS point graph, there is a good deal of overlap between clusters 0, 2, and 4, whereas clusters 1, 3 and 5 all are nicely separated.

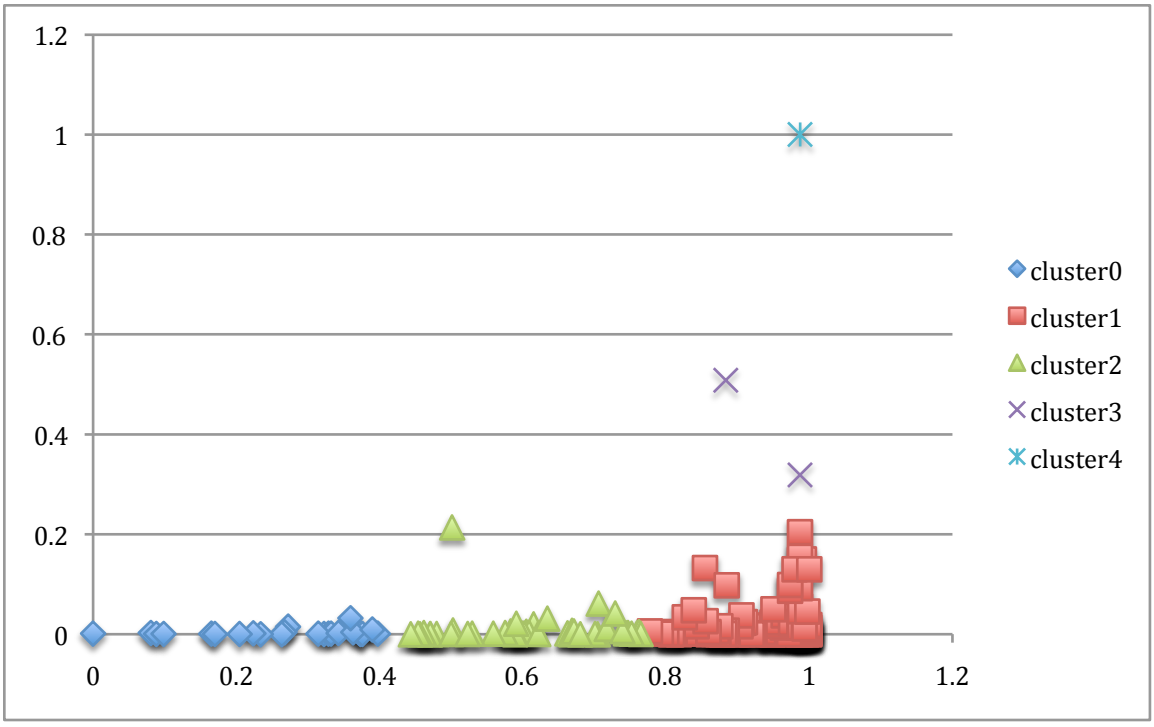

Figure 2: GAS clustering on literacy rate and GDP.

This graph, showing GAS, has much better separation between the clusters with no overlap. The clusters are elongated rather than compact as one would think.

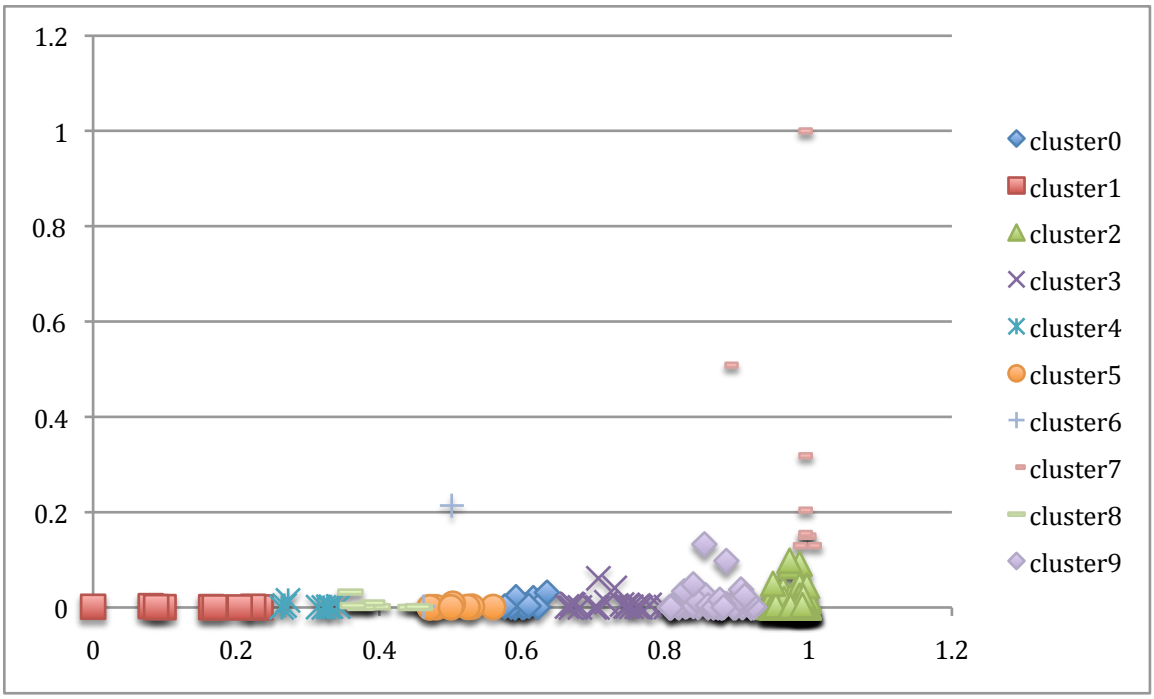

Figure 3: K-Means clustering on literacy rate and GDP.

K-Means with a max of 10 clusters makes very nice compact clusters that fit the data very well.

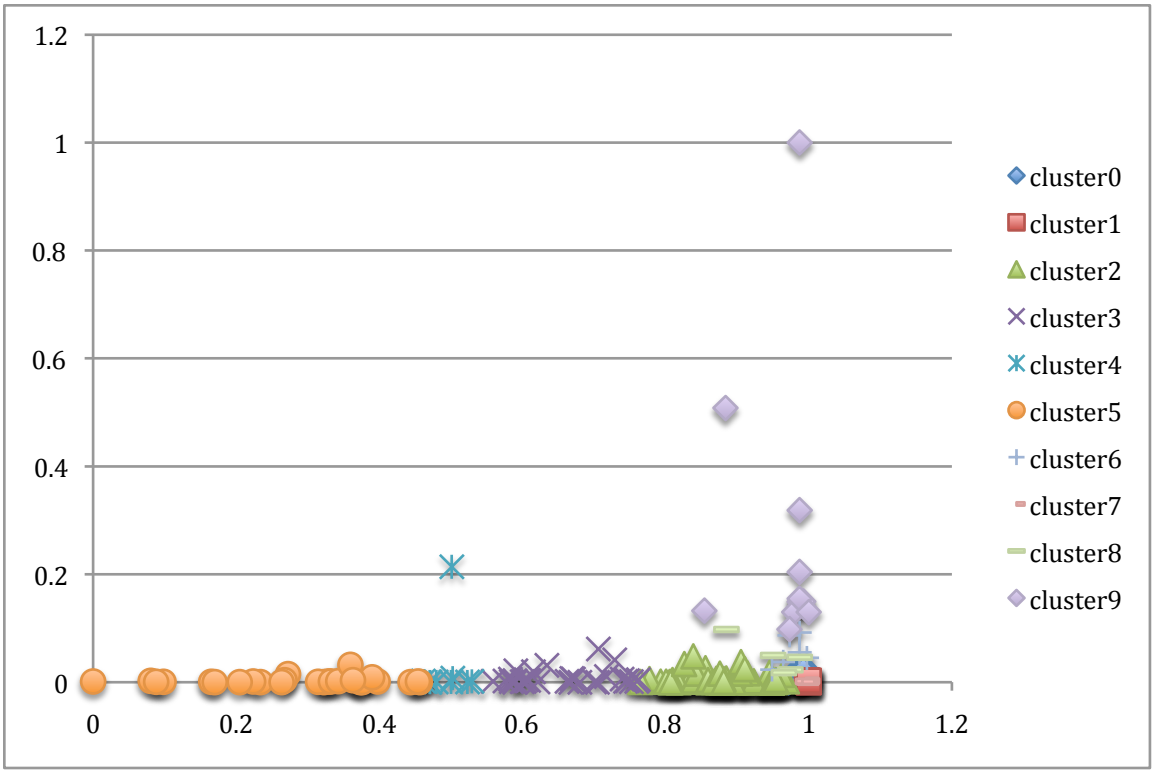

Figure 4: New K-Means clustering on literacy rate and GDP.

The data points in the New K-Means clustering are not as compact, they are quite scattered or elongated at times.

It seems like GAS and K-Means made the best clustering. They had well defined clusters that were tightly packed with data points.

Conclusion:

In most cases the quality of the clustering is highly dependent on what you put into it. Feature selection seems to be most important, it is the core of what the clustering consists of. Picking the correct features that convey the most information possible is what matters the most. Normalizing the data is critical in order to be able to relate the data to each other. Choosing the right algorithm, and threshold is also key to getting a proper clustering.

Future Ideas:

At one point during the project I came up with the idea of putting in a reverse engineering aspect in order to fill in the data that had been missing in the initial data collection. It should be based on how the country was clustered and with what countries it was clustered with. I may try to implement this at a later time.

Sources:

Geographical Area – CIA World Factbook - https://www.cia.gov/library/publications/the-world-factbook/fields/2147.html

Population – CIA World Factbook – https://www.cia.gov/library/publications/the-world-factbook/fields/2119.html

Literacy Rate – CIA World Factbook – https://www.cia.gov/library/publications/the-world-factbook/fields/2103.html

GDP (in dollars) – CIA World Factbook – https://www.cia.gov/library/publications/the-world-factbook/fields/2001.html

Life Expectancy (years at birth) – CIA World Factbook – https://www.cia.gov/library/publications/the-world-factbook/fields/2102.html

Military Expenditures (% GDP) – CIA World Factbook – https://www.cia.gov/library/publications/the-world-factbook/fields/2034.html

Number of Internet Users – CIA World Factbook - https://www.cia.gov/library/publications/the-world-factbook/fields/2153.html

Net Migration – UN - http://unstats.un.org/unsd/cdb/cdb_series_xrxx.asp?series_code=13640

Access to improved drinking water – UN - http://unstats.un.org/unsd/cdb/cdb_series_xrxx.asp?series_code=27910

Patent Applications – UN - http://unstats.un.org/unsd/cdb/cdb_series_xrxx.asp?series_code=28130

Religious Population (% population) – IP - http://www.infoplease.com/ipa/A0855613.html

Country has a Monarchy – IP - http://www.infoplease.com/ipa/A0775675.html

Country has the Death Penalty – IP - http://www.infoplease.com/ipa/A0777460.html

Country has Nuclear Weapons – IP - http://www.infoplease.com/ipa/A0762462.html

Amount of Chocolate Imported (dollars) – ITC - http://www.intracen.org/tradstat/sitc3-3d/ip073.htm

Satisfiability with Life Scale - http://www.le.ac.uk/users/aw57/world/sample.html