**Randomized Optimization**

**Zhijian Li**

**1 Dataset**

I will be using two datasets, both from assignment 1.

As a recap:

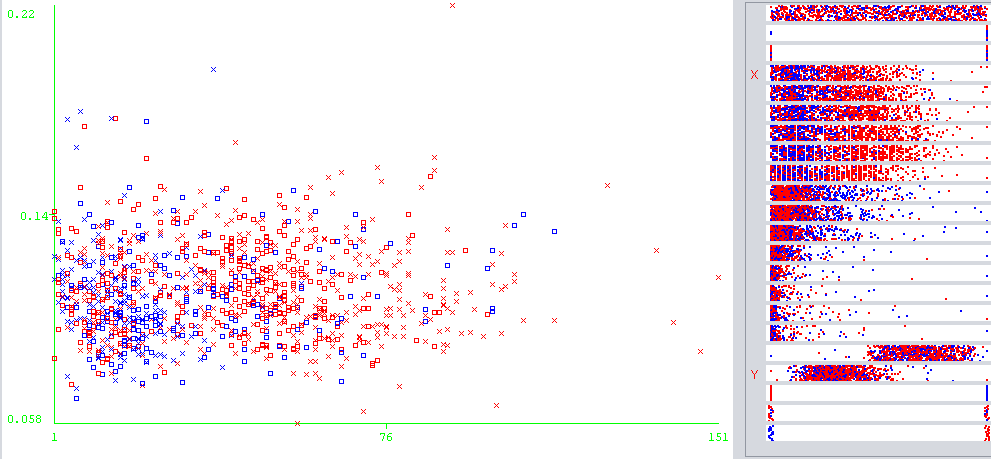
*Diabetic Retinopathy Data Set (DR dataset):* This dataset contains data was taken from the Messidor image set, a collection of DR examinations. The goal is to predict whether the image contains signs of diabetic retinopathy. The features are extracted from the images and is either a detected lesion, a descriptive feature of an anatomical part or an image-level descriptor. The dataset contains 1151 instances and 20 attributes.

*Mammographic Mass Data Set (MM dataset):* This dataset contains data from mammograms. The goal is to predict whether the patient has malignant breast cancer. The features contain BI-RADS assessments along with related measurements of the patient. The dataset contains 961 instances and 6 attributes.

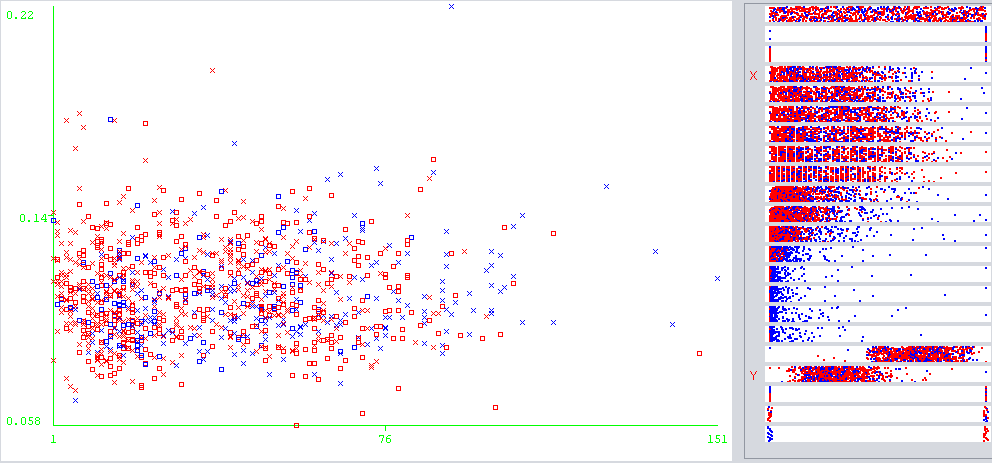
**2 Clustering**

I performed two clustering algorithms, K-Means and EM on both datasets. I used Weka’s Classes to Cluster tool to evaluate how well each algorithm sorted the data based off the initial classification problem, severity of breast cancer (MM) and existence of diabetes (DR). I used Euclidean distance for both algorithms. Because both my datasets are binary classification problems, I set each algorithm to generate two clusters. I will be analyzing clustering between two attributes as a graph and the percent of incorrectly clustered instances.

**DR**



K-Means Clustering for DR



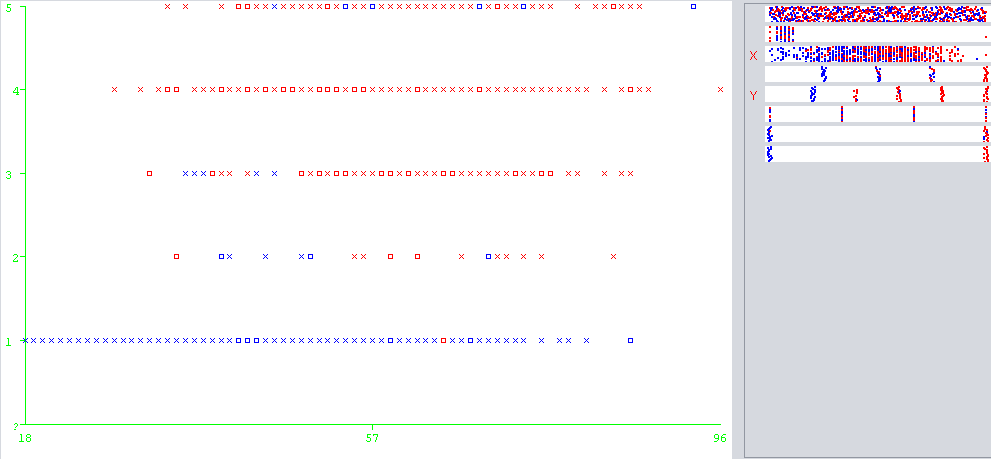
EM Clustering for DR

|  |  |  |
| --- | --- | --- |
| **Algorithm** | **Time Taken** | **% Incorrectly Clustered** |
| K-Means | 0.02 | 47.0026 |
| EM | 0.15 | 41.0947 |

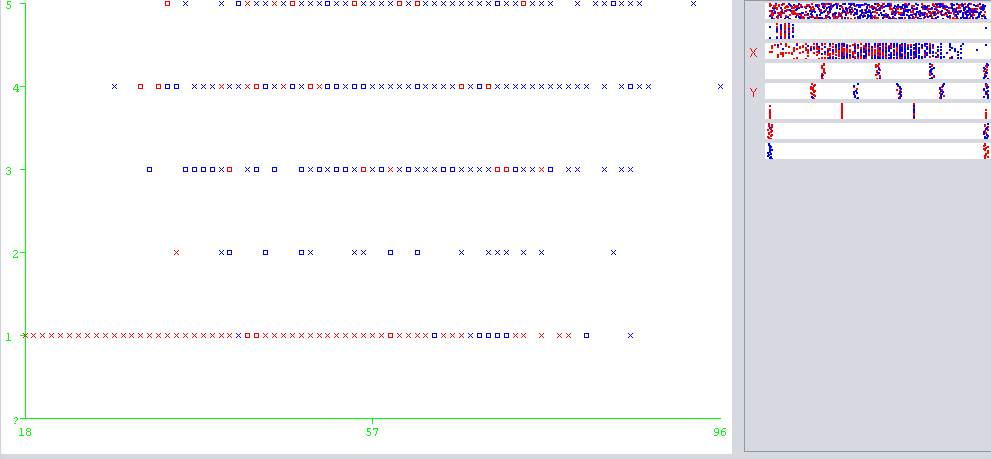
For DR, both algorithms performed relatively poorly, but EM performed better than K-Means. The graphs were taken directly from Weka, with the y-axis plotting Diameter of Optic Disc and the x-axis plotting MA Detection 1. The blue represents the cluster representing healthy patients and the red represents patients with DR, DR’s dataset contained various MA Detections, from 1 to 6, all with similar graphs, so I decided to stick to MA Detection 1. I chose these two attributes because they had the best graphs, all other pairs of attributes produced graphs that were very messy, with almost no differentiation between the clusters and the axes.

As seen from the graphs, there wasn’t much correlation between these two attributes and which cluster they belonged too. This was the case for every single pair of attributes and suggests that the classification problem is not linearly separable in two dimensions. This makes sense as the problem itself is a hard problem pertaining to medical diagnosis and due to the number of attributes the dataset contains. The difference is that with higher values of MA detection, K-Means classified almost always classified patients to have DR while for EM, it was more evenly split throughout.

**MM**



K-Means Clustering for MM



EM Clustering for MM

|  |  |  |
| --- | --- | --- |
| **Algorithm** | **Time Taken** | **% Incorrectly Clustered** |
| K-Means | 0.01 | 21.7482 |
| EM | 0.02 | 26.9511 |

For MM, the accuracy was much better. The y-axis represents the mass margin and the x-axis represents the age of the patient. Similarly, with DR, the blue represents a benign tumor while the red represents a malignant tumor. Despite all the data values being numerical, in reality, each number from 1 to 5 represents the different categories of mass margin (circumscribed=1 microlobulated=2 obscured=3 ill-defined=4 spiculated=5). I chose these two attributes because they highlighted the differences between the algorithms. K-Means clusters were very apparent from the graph, with benign tumors having a mass margin of mostly 1 only with no differentiation between the age of the patient. On the other hand, EM results were almost the opposite, with almost all benign tumors having a mass margins strictly greater than 1 and no differentiation between age.

Unlike the DR dataset, the MM dataset only had 6 attributes total, so there was a much higher correlation between pairs of attributes. The higher accuracy between these two datasets also suggests this. With this being the case, there was a much cleaner split between attributes and their clusters and ultimately resulted in a higher accuracy. Not only was the MM dataset a much easier classification problem, we also know now that the data was easily separable.

**2 Dimensionality Reduction**

**Principal Components**

I ran PCA on both datasets, changing the percentage of variance covered. For both datasets, PCA produced almost similar number of attributes, but this does not necessarily mean the number of attributes were reduced. For the DR dataset, the number of selected attributes decreased from 20 to 9 with 0.95 variance covered, a decrease of 11 attributes. For the MM dataset, this was not the case. While there were still 9 attributes selected, the MM dataset had 6 to begin with.

|  |  |
| --- | --- |
| **Variance Covered** | **# Selected Attributes** |
| 0.95 | 9 |
| 0.85 | 6 |
| 0.75 | 4 |
| 0.5 | 2 |

PCA on DR dataset

For the DR dataset, from the generated attributes, it is very clear which attributes from the original dataset are important and which are not. The attributes that were ranked the most important all contained linear combinations of the MA detection and MA detection exudates attributes. The ranking then went on to the diameter of the optic disc and the distance to the macula, with the quality of the assessment and results of the prescreening having much less importance. This makes sense because the most important attributes are the actual features extracted from the DR images. It is also important to note that the prescreening results have little affect for the diagnosis.

|  |  |
| --- | --- |
| **Variance Covered** | **# Selected Attributes** |
| 0.95 | 9 |
| 0.85 | 8 |
| 0.75 | 7 |
| 0.5 | 4 |

PCA on the MM dataset

For the MM dataset, there was no reduction in the number of attributes. This is because the PCA algorithm takes each of the different nominal values each attribute can take on and treats it as a separate attribute. This changed the dataset because instead of one shape and margin attribute, there were four shape attributes and 5 margin attributes, one for each value. The attributes that ranked the most important consisted of linear combinations of margin and shape. Then came age, density and assessment. This shows that the important attributes were shape and margin with the rest of the attributes having less of an effect on the outcome. This is interesting because, even with 6 attributes, PCA tells us that 2 of the attributes are very dominant indicators, while the rest of the attributes are less important.

**Clustering after PCA**

After PCA, the clustering algorithms showed little change for the DR dataset and showed an increase in the percent of incorrectly clustered instances for the MM dataset. For the DR dataset, this makes sense, because while PCA generates new attributes, it preserves the space between instances, thus it should not have a big impact on clustering, despite a reduction in attributes. This is not the case for the MM dataset. The MM dataset gained attributes, and this can be clearly seen in the increase in error.

|  |  |  |
| --- | --- | --- |
| **Variance** | **Algorithm** | **% Incorrectly Clustered** |
| 0.95 | K-Means | 45.1781 |
| 0.85 | K-Means | 45.0912 |
| 0.75 | K-Means | 45.0043 |
| 0.5 | K-Means | 45.0043 |
| 0.95 | EM | 49.3484 |
| 0.85 | EM | 49.2615 |
| 0.75 | EM | 46.5682 |
| 0.5 | EM | 40.6603 |

Clustering for the DR dataset after PCA

|  |  |  |
| --- | --- | --- |
| **Variance** | **Algorithm** | **% Incorrectly Clustered** |
| 0.95 | K-Means | 39.0129 |
| 0.85 | K-Means | 39.0219 |
| 0.75 | K-Means | 39.0219 |
| 0.5 | K-Means | 39.0219 |
| 0.95 | EM | 43.1842 |
| 0.85 | EM | 43.1842 |
| 0.75 | EM | 43.0801 |
| 0.5 | EM | 43.1842 |

Clustering for the MM dataset after PCA

*DR Dataset*

PCA improved K-means by about four percent while EM’s performance varied. This is because of how PCA works. Essentially, PCA not only reduced the number of attributes, it essentially gave each attribute a stronger or weaker rank depending on how important it was. This affects the Euclidean distance between data points because it essentially scaled all the distances by importance, so outliers have less of an impact.

*MM Dataset*

There was a drastic change before and after PCA. Both K-Means and EM performed worse with PCA, which can be explained by the increase in the number of attributes. This can also be because of how PCA creates new attributes that are linear combinations of old attributes, increasing the complexity and making it harder to cluster.

ICA

-ICA

-clustering on ICA

Random Projection

-RP

-clustering on RP

Infogain

-IG

-clustering on IG

NN on one dataset (DR)

-PCA, ICA, RP, IG

-4 graphs, each with 3 lines (K-means, EM, no clustering) using parameters for each Dim Reduc (PCA - .95, .85, .75, etc.)

References:

1 <https://archive.ics.uci.edu/ml/datasets/Mammographic+Mass>

2 <https://archive.ics.uci.edu/ml/datasets/Diabetic+Retinopathy+Debrecen+Data+Set>

3 <https://github.com/pushkar/ABAGAIL>