James Liu

CS 4641, Spring 2014, Isbell

Unsupervised Learning and Dimensionality Reduction Assignment

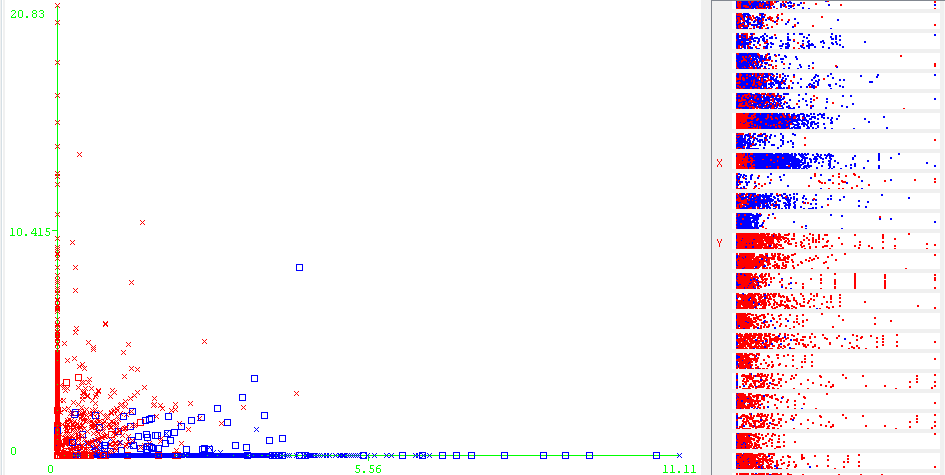
**Dataset Description**

The two datasets used in this assignment are the same two used in assignment 1: spambase and OTHER DATASET. spambase is a dataset of word and character frequencies in emails and whether or not the email was considered spam by the user. It is the exact kind of dataset one would expect great improvement from dimensionality reduction algorithms: it has a large number of attributes, 58 in total, that are discrete and highly kurtotic in distribution, and many of them may be redundant/irrelevant. The classification of whether it is spam or not is directly dependent on a small subset of attributes, each of which provides a significant amount of information on the classification of the instance. OTHER DATASET DESCPRITION. spambase has 4601 instances. OTHER DATASET INSTANCE COUNT.

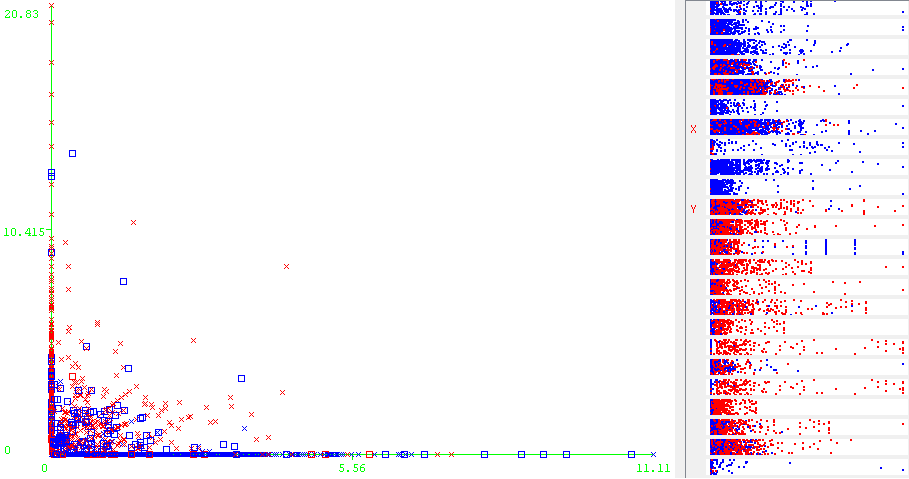
**Initial Clustering Experiments and Results**

For the purposes of this analysis and to maintain consistency for the experiments done, both k-means and expectation maximization used Euclidean distance measure as a similarity measure between instances. EUCLIDEAN DISTANCE JUSTIFICATION.

For the purposes of testing k-means and EM, I chose to use a k equal to the number of possible values the nominal class attribute for each dataset. This will test to see how well the clusters created by k-Means and EM matches the given classes. The clusters were compared to classes by use of Weka’s “classes to cluster” evaluation, which simply looks for the best matched class for each cluster, which does not run well with a large number of clusters/classes. Thus why I chose to use datasets with binary classifications. I had also tried datasets like a large number of classes, which took too long to test using this method and was also difficult to visualize the result due to the increased number of clusters. This test method’s only result is how many of the best fit cluster match the given class as a percentage of the database, which will be used as the main performance metric for this analysis.



**Figure 1.** The k-Means results plotting instances based on the frequency of the word “your” (X axis), and the frequency of the word “hp” (Y axis). Red are those in cluster 1, blue are those in cluster 2. Note the clear distinction between both clusters and little overlap. On the left is the relative distribution, and cluster assignment of instances relative to certain attributes. Note the distribution of each cluster relative to said attributes.



**Figure 2.** The results of expectation maximization plotted in the same format as the one for k-means above. Unlike k-means, there is no clear distinction between either cluster in this plot or the attribute-class distributions to the right, particularly the attribute labeled with “X”.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset | Clustering Algorithm | Log Likelyhood | Time Elapsed | % Incorrectly Clustered |
| spambase | k-Means | - | 0.39s | 20.0826% |
| spambase | EM | -28.58234 | 4.51s | 21.7996% |
| OTHER DATASET |  |  |  |  |
| OTHER DATASET |  |  |  |  |

Spambase’s results are unexpected. I initially thought that due to the high skew of the distributions for most of the dataset’s distributions that it would make it difficult to differentiate, as most of the values for all most of the attributes for both classes were close to 0.

**Dimensionality Reduction Experiments w/Clustering and Results**

**Principle Component Analysis**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Dataset | Clustering Algorithm | Variance Covered | # of result attributes | Log Likelihood | Time Elapsed | % Incorrectly Clustered |
| spambase | k-Means | 0.25 | 6 | - | 0.08s | 19.0828% |
| spambase | k-Means | 0.5 | 17 | - | 0.28s | 17.9092% |
| spambase | k-Means | 0.75 | 32 | - | 0.42s | 18.5829% |
| spambase | k-Means | 0.95 | 49 | - | 0.81s | 20.7564% |
| spambase | EM | 0.25 | 6 | -7.52479 | 0.85s | 48.1852% |
| spambase | EM | 0.5 | 17 | -21.31495 | 4.59s | 47.3375% |
| spambase | EM | 0.75 | 32 | -37.4544 | 7.82s | 43.1645% |
| spambase | EM | 0.95 | 49 | -55.24697 | 10.63s | 42.2734% |
| OTHER | k-Means | 0.25 |  | - |  |  |
| OTHER | k-Means | 0.5 |  | - |  |  |
| OTHER | k-Means | 0.75 |  | - |  |  |
| OTHER | k-Means | 0.95 |  | - |  |  |
| OTHER | EM | 0.25 |  |  |  |  |
| OTHER | EM | 0.5 |  |  |  |  |
| OTHER | EM | 0.75 |  |  |  |  |
| OTHER | EM | 0.95 |  |  |  |  |

**Random Projection**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dataset | Clustering Algorithm | # final attributes | Log Likelihood | Time Elapsed | % Incorrectly Clustered |
| spambase | k-Means | 10 | - | 0.22s | 42.1213% |
| spambase | k-Means | 20 | - | 1.8s | 17.9092% |
| spambase | k-Means | 30 | - | 0.42s | 18.5829% |
| spambase | k-Means | 40 | - | 0.81s | 20.7564% |
| spambase | EM | 10 | -68.76933 | 3.2s | 29.5805% |
| spambase | EM | 20 | - | 4.59s | 47.3375% |
| spambase | EM | 30 | -37.4544 | 7.82s | 43.1645% |
| spambase | EM | 40 | -55.24697 | 10.63s | 42.2734% |
| OTHER | k-Means |  | - |  |  |
| OTHER | k-Means |  | - |  |  |
| OTHER | k-Means |  | - |  |  |
| OTHER | k-Means |  | - |  |  |
| OTHER | EM |  |  |  |  |
| OTHER | EM |  |  |  |  |
| OTHER | EM |  |  |  |  |
| OTHER | EM |  |  |  |  |

**Neural Network Training with Dimensionally Reduced Datasets with and without Clusters as Attribute**