

Assignment 3:

Unsupervised Learning

Steven Nord
snord3@gatech.edu

1 Datasets

Two datasets were explored for comparing clustering and dimension reduction algorithms (the Breast Cancer dataset and the Steel Plates

Table 1		
Datasets	Breast Cancer	Steel Plate Faults
Samples	699	1,941
Num. of Features	9	27
Ranges	1-10	Varied
Target Class	Benign, Malignant	K_Scratch, Bumps, Dirtiness, Z_Scratch, Pastry, Stains, Other

dataset). These datasets posed interesting results since they possessed different qualities (**Table 1**) and will be highlighted throughout the paper. Feature scaling was applied to the datasets so features with large ranges would not disproportionality influence each component. Then, training and test sets were split 75/25 for the experiments highlighted below.

2 Clustering

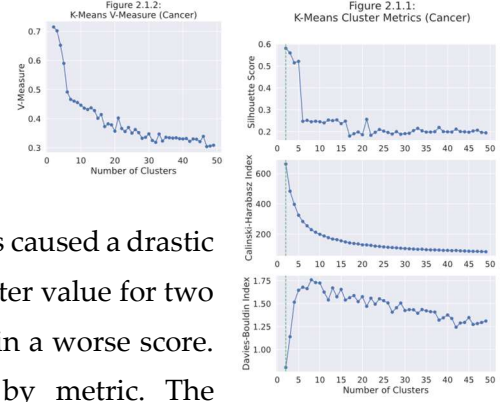
K-Means and Expectation Maximization clustering algorithms were performed on both dataset's training set. The first step for each algorithm was to determine the optimal number of clusters (k). This was done by measuring the Silhouette Coefficients, Calinski-Harabasz Index (CHI), and Davies-Bouldin Index (DBI). Each run of k was performed 10 times with different initializations to reduce the opportunity for algorithms to get stuck in a local optimum.

Each metric measures the within-cluster distance to the between-cluster distance, but with slightly different methodologies. Silhouette and CHI indicated the optimal number of clusters when their values are highest, while Davies-Bouldin was optimal when its value was low. Since each has its strengths and weaknesses, the optimal number of clusters was voted on as a collective group. The target class was discarded for this process as to not make this into a supervised learning experiment.

Since the ground truth labels were available, V-measure was used to assess how each clustering translates back to the target classes. V-measure takes into account both Homogeneity (the ratio of a cluster pertaining to a single target class) and Completeness (the ratio of a target class being contained in a single cluster).

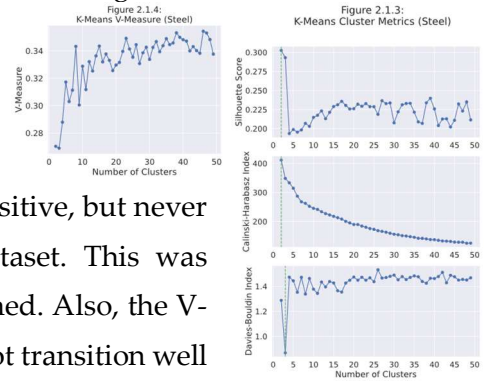
2.1 K-Means

Figure 2.1.1 shows each metric versus the number of clusters on the Cancer dataset. From the Silhouette plot alone, two through five would be reasonable selections for the number of clusters. Going from two to three clusters caused a drastic decrease in CHI. On the other hand, DBI maintained a better value for two or three clusters, but going to four clusters would result in a worse score. The green dashed line indicates the optimal chose by metric. The unanimous choice across all metrics was two clusters. The resulting clusters have a well-defined inter versus intra-cluster relation based on Silhouette being closer to one as opposed to negative one.



Once two clusters were selected, the resulting V-measure was 0.7151. **Figure 2.1.2** shows choosing a different number of clusters would have resulted in a lower score. This number of clusters aligned nicely with the fact that the Cancer dataset only had two target classes.

Figure 2.1.3 shows the clustering metrics from K-means on the Steel dataset. Silhouette and CHI agreed on the choice for the optimal number of clusters while DBI would have preferred three clusters. The Silhouette scores remained positive, but never reached the same level as they did for the Cancer dataset. This was informative as it described the clusters to be less well defined. Also, the V-measures shown in **Figure 2.1.4** indicate two clusters did not transition well



to the target classes. This was not much of a surprise given the dataset has seven target classes, so the Homogeneity of two clusters could not possibly consist of a single class. 10-15 clusters would have been more ideal numbers of clusters, since this range still consisted of relatively high Silhouette and CHI score while also having a relatively low DBI. Since target class was not permitted for the selection of k, the optimal k for the Steel dataset was set to two.

When comparing the V-measures for the two datasets, the trajectory for each was completely opposite. As more clusters were added, Cancer data seemed to be impacted negatively unlike when additional clusters added benefit for the Steel dataset. This depicted the complexity involved in the Steel dataset as compared to the simplistic nature of the Cancer dataset.

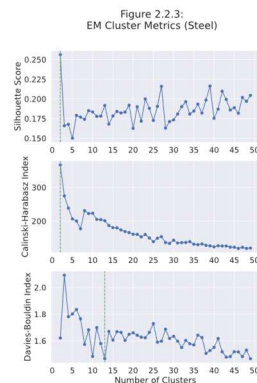
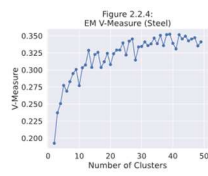
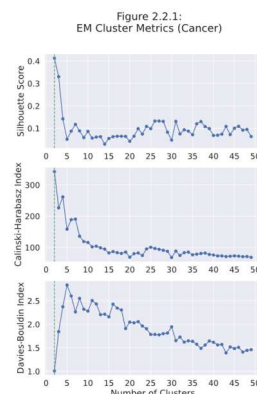
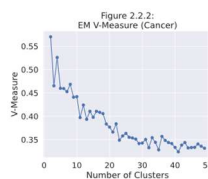
2.2 Expectation Maximization (EM)

Expectation Maximization resulted in the same number of clusters for the Cancer dataset, which was two clusters.

From **Figure 2.2.1**, it was clear all metrics agreed on two clusters. Interestingly though, EM did not generate clusters that were as compact and distant as K-Means. EM assumed the data was comprised of a mixture of Gaussian distributions, so this assumption did not hold for the Cancer dataset as 56% of the features had kurtosis greater than three (the kurtosis of a Gaussian Distribution). Unfortunately, the V-measure took a hit as well from the switch to EM (**Figure 2.2.2**).

For the Steel dataset, the Silhouette and CHI plots again preferred two clusters, but DBI would have selected 13 as the number of clusters (**Figure 2.2.3**). The vote went to the majority, but this actually resulted in a lower v-measure score than if 13 clusters would have been chosen.

EM also took a small performance hit compared to K-means for the Steel dataset, but not to the same degree as the Cancer. The 52% of the features distributions for the Steel dataset were supportive of the Gaussian distribution assumption.

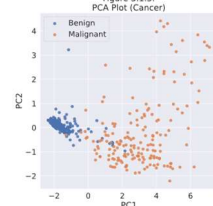
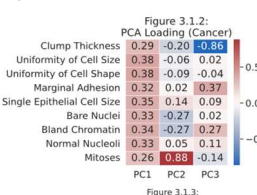
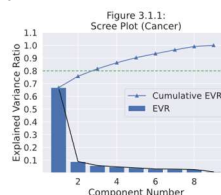


3 Dimension Reduction

Dimension reduction was performed on both datasets using four separate analyses (Principal Component Analysis, Impendent Component Analysis, Random Projects, and Linear Discriminant Analysis).

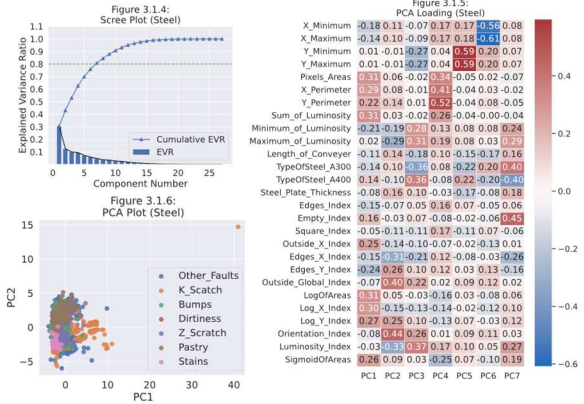
3.1 Principal Component Analysis (PCA)

The number of components was determined by extracting enough of the data to explain 80% of the variance. **Figure 3.1.1** shows 80% of the original variance in the Cancer data can be explained with the first three principal components. The elbow methodology would have suggested to go with two components, but three were chosen since they explained more than 80%. **Figure 3.1.2** lists the features with the loading score for each component. Mitoses and Clump Thickness had a strong influence on components two and three respectively, while component



one was a blend of all the features since the loading scores ranged from 0.23 to 0.3808. **Figure 3.1.3** shows the result from PCA for components one and two. The plot displays a slight separation between Benign and Malignant labels.

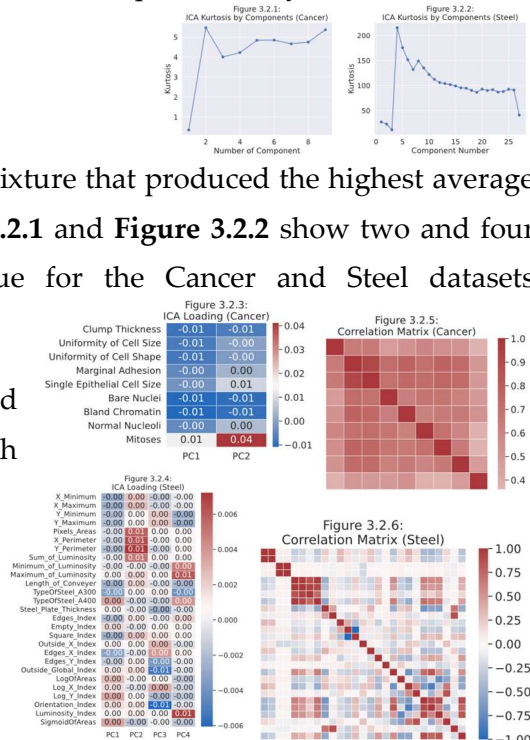
The scree plot for the Steel dataset can be seen in **Figure 3.1.4**. For this dataset, the first component did not explain as much of the original variance as it did in the Cancer dataset. Also, from **Figure 3.1.5** the max loading scores were not as high which indicated components were not highly influenced by any given feature. This was due to the Steel dataset having more features to consider than the Cancer dataset. The number of components needed to reach the 80% threshold was 7 components. **Figure 3.1.6** shows the Steel dataset in a 2-dimensional space, but not much of the variation can be made out from the plot. PC1 was able to provide some insight into K_Scarch versus other labels. PC2 begins to draw a distinction between Pastry and Stains, but according to the scree plot, the first two components only account for 43.1% of the variance.



3.2 Independent Component Analysis (ICA)

The number of components chosen for ICA was the mixture that produced the highest average kurtosis across all the resulting components. **Figure 3.2.1** and **Figure 3.2.2** show two and four components produced the maximum kurtosis value for the Cancer and Steel datasets respectively.

The loading scores for both datasets (**Figure 3.2.3** and **Figure 3.2.4**) were all relatively insignificant for each component. The ICA algorithms, whitened strategies, function to approximate the neg-entropy, and various other ICA parameters were tweaked to improve the loading scores with little changes observed. When investigating more into the data, **Figure 3.2.5** and **Figure 3.2.6** show the ICA primary assumption did not hold for either dataset. The assumption was the sources of variation in the data were



independent of one another. The correlation matrix, especially for the Cancer dataset, showed a moderately high level of feature correlation.

3.3 Random Projection (RP)

The process for selecting the number of components for RP was based on the reconstruction error. Each component set was run five times to account for the randomness involved in RP.

The mean reconstruct error along with stand deviation is shown in **Figure 3.3.1** and **Figure 3.3.2** for each dataset. There was no larger drop-off in the error from one to any other component size, so the ultimate selection was set to match the selection from PCA (Cancer = 3 and Steel = 7). The rationale was to allow for the two techniques to be more comparable.

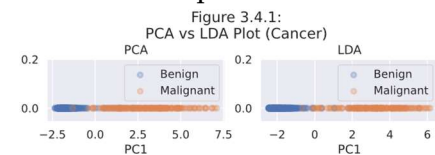
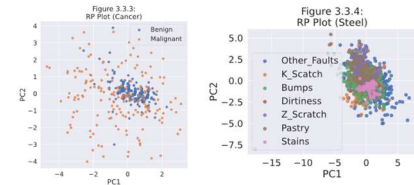
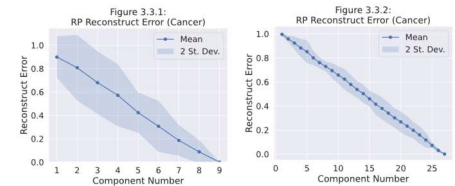
Figure 3.3.3 shows a 2-dimensional space of the first two components for the Cancer dataset reproduced by RP. The distinction between classes was not as well defined as it was in PCA. The RP for the Steel dataset seems comparable to the results from PCA with a little more scatter between Pastry and Stains (**Figure 3.3.4**).

If the datasets are comprised of a very high dimensional space, there is a stronger case for using RP. For these datasets of 9 and 27 features and fewer than several thousand samples, this was not a powerful consideration. 100 fits where run for both PCA and RP; RP fit the data 27% faster than PCA for the Steel dataset. A reduction of that size would go a long way for datasets that are much larger. With the Cancer dataset being in an even lower dimensional space there was no considerable difference in fit time for PCA and RP.

3.4 Linear Discrimination Analysis (LDA)

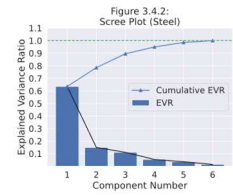
Since both datasets consisted of a target class, LDA was explored for the last dimension reduction technique. LDA used the target class as input and focused on projecting the data onto a space that maximized the separation between classes. The motivation was to compare this supervised dimension reduction method to the other unsupervised methods.

LDA restricts the max number of components to one less than the number of classes in the data. Since the Cancer dataset only had two classes it limited the choices to only one component. **Figure**

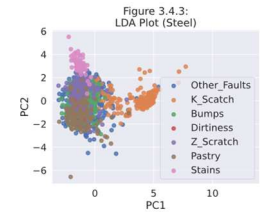


3.4.1 compares PCA and LDA for the first component. Since LDA focused on maximizing the separation between the classes, there was more whitespace in the LDA plot compared to PCA.

Steel had seven classes so it was limited to six as the maximum number of components for LDA. **Figure 3.4.2** shows the scree plot for the Steel dataset. The elbow approach would have suggested to choose two and the 80% threshold used for PCA would have picked three. Since including all the variation still resulted in a component size smaller than the seven selected for PCA, all six were used moving forward to maximize information for the Neural Networks.



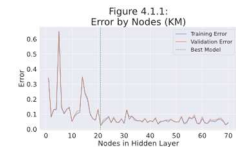
LDA knowledge of the “ground truth” created more split between K_Scratch, Pastry, and Stains (**Figure 3.4.3**) than can be observed from the PCA plot (**Figure 3.1.6**).



4 Neural Networks with Clustering

A neural network was evaluated on the initial features along with the clustering information previously gathered. The Base ANN was an optimized neural network including only the original features for the dataset. The hypothesis was the additional information from the two clustering algorithms would provide additional insight into the target class predictions. This part of the experiment was performed on only the Cancer dataset. Hyperparameter tuning was performed to optimize the model to allow comparison to be on level playing fields. Tuning of the nodes will be discussed in the sections below, but tuning for learning rate and activation function were also considered. The latter were chosen based on lowest validation error across Gridsearch results. Hyperparameters for optimal models along with the initial baseline neural network is provided in **Table 2**.

4.1 K-Means (KM)

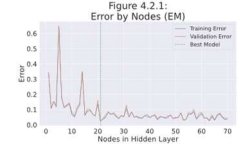


The number of nodes for this model was selected from observing the data from **Figure 4.1.1**. The first 20 nodes were not considered since they generated larger errors. There could be a case made for several values beyond 20, but the optimal model was selected to have 21 nodes since including more added further complexity.

The conclusion was the additional cluster feature did not improve on the base model. The training score went up slightly, but the added dimension did not translate into a higher test score. However, according to the learning curve (**Table 2**) it appears increasing the dataset could result

in improvements for this model. Also, the model took a tad longer to train since there was an extra feature included.

4.2 Expectation Maximization (EM)

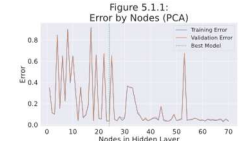


There was an uncanny resemblance between **Figure 4.1.1** (KM) and **Figure 4.2.1** (EM). The rationale behind this was that both KM and EM added a single feature to the data because they both chose two clusters from Section 2.1 and 2.2 respectively. Even further, they agreed on 86.13% of the cluster assignments so there was very minimal information change between these models. This model also preferred 21 nodes as indicated by the green dashed line in **Figure 4.2.1**.

However, the 13.87% difference in cluster assignment did negatively impact the accuracy of the model. This was not surprising given the V-measure score for EM (Section 2.2) was 0.15 lower than the score for KM (Section 2.1). The extra complexity resulted in a slightly longer training time as well when compared to the Base ANN. In conclusion, this model did not improve the accuracy or the training time versus the Base ANN.

5 Neural Networks with Dimension Reduction

The next experiment was to perform a neural network with just the components created from the dimension reduction techniques discussed previously. The hypothesis was that performance would not take much of a hit while reducing the training time since there would be less dimensions for the network to have to process. Hyperparameter tuning was performed the same way for neural networks with clustering. Hyperparameters for optimal models along with the initial baseline neural network are provided in **Table 2**.



5.1 Principal Component Analysis (PCA)

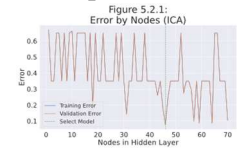
The errors by node can be seen in **Figure 5.1.1** for the PCA neural network model. The validation error was lowest for 24 nodes and was relatively stable beyond this point. The optimal model was set to include 24 nodes and the results can be observed in **Table 2**.

The optimal model was comparable to Base NN while only including three components. It is important to recall that three components were chosen because they explained 80% of the variance in the original data. This was clearly enough as it did not drastically impact the accuracy of the model and did much better as it reduced the training time by over 60%. This incentive was minimal since this dataset was not large; therefore, training time was already relatively

insignificant. The learning curve for this model indicated that adding more data would not improve the accuracy prediction which makes the training time incentive negligible. For other datasets it would be obvious how the training time could be a key motivation for using PCA to reduce training time without sacrificing much accuracy. Ultimately, this reduction technique on the Cancer dataset was an improvement on the Base NN because the scores were comparable and the reduction in training time was noteworthy.

5.2 Independent Component Analysis (ICA)

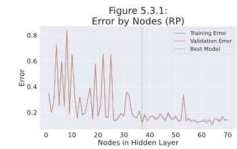
Figure 5.2.1 shows ICA for the Cancer dataset struggled with making accurate predictions since most of the error rates were over 30%. This aligned with the discovery made back in Section 3.2 when the loading scores were all relatively small. The conclusion drawn was that the Cancer dataset was not a strong candidate for ICA due to its dependent features. The optimal number of nodes for this model was set at 46, but as seen in **Table 2** this did not render strong results.



The optimal model did significantly worse than the base model only generating a test accuracy of 66.08%. The training time was drastically cut even compared to PCA. However, since the accuracy score was so much lower than the Base NN, ICA would not be considered an improvement for the model.

5.3 Random Projections (RP)

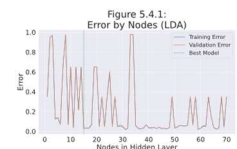
The number of nodes was set to 37 for the neural network built from the reduced data created by RP. **Figure 5.3.1** does not show any major change in the error beyond this point.



The optimal model did not perform well. The accuracy was low at just under 90% for both training and test sets. This neural network did not do as well as the PCA NN because the class distinction created from the RP projection (**Figure 3.3.3**) was less clear than the projection space created for PCA (**Figure 3.1.3**). Again, the reduced dimensional space allowed the model to train in a shorter amount of time, but PCA training time was not much longer while still maintaining high accuracy scores. In conclusion, RP would not be an improvement to the Base ANN model.

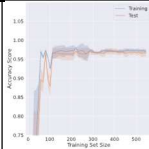
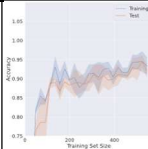
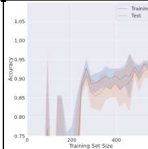
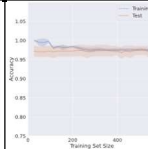
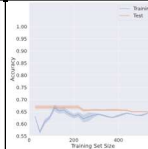
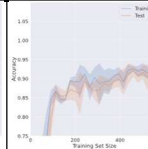
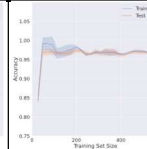
5.4 Linear Discriminant Analysis (LDA)

Figure 5.4.1 shows the errors by number of nodes with the ultimate selection of 15 nodes indicated by the green dashed line. Several other node sizes performed just as well,



but 15 was the smallest value and therefore reduced the complexity of the model. This model had the fewest features to train on and resulted in the lowest nodes value as well.

The optimal LDA model performed very well. The accuracy matched the Base ANN given the dimensions were reduced by 88.89%. This reduction resulted in a training time that was less than 17% of the time needed for the Base ANN. It was also less than half the amount of time needed to train the PCA neural network. In conclusion, LDA objective to project the data onto a space that maximized the separation between target classes translated well for predicting benign and malignant cells.

Table 2: Optimal Results Across Neural Networks (Cancer)							
	ANN (Base)	K-Means	EM	PCA	ICA	RP	LDA
# Clusters/Comp	N/A	2	2	3	2	3	1
# of Features	9	10	10	3	2	3	1
Optimal Parameters	Activation: ReLU Nodes: 25 Learning Rate: 0.01	Activation: ReLU Nodes: 21 Learning Rate: 0.002	Activation: ReLU Nodes: 21 Learning Rate: 0.002	Activation: ReLU Nodes: 24 Learning Rate: 0.01	Activation: ReLU Nodes: 46 Learning Rate: 0.001	Activation: ReLU Nodes: 37 Learning Rate: 0.01	Activation: ReLU Nodes: 15 Learning Rate: 0.01
Accuracy	Test: 97.66% Train: 97.46%	Test: 97.08% Train: 97.66%	Test: 95.32% Train: 96.68%	Test: 97.08% Train: 97.07%	Test: 66.08% Train: 65.04%	Test: 89.47% Train: 88.28%	Test: 97.66% Train: 97.85%
Learning Curve							
Training Time	0.273 seconds	0.293 seconds	0.284 seconds	0.101 seconds	0.0756 seconds	0.092 seconds	0.046 seconds
Testing Time	0.003 seconds	0.003 seconds	0.004 seconds	<0.001 seconds	<0.001 seconds	<0.001 seconds	<0.001 seconds

6 Clustering on Dimension Reduced Data

K-Means and Expectation Maximization clustering were performed on the dimensionally reduced datasets from the techniques discussed in Section 3. Results across all the runs are displayed in **Table 3** and **Table 4** for the Cancer and Steel datasets respectively. The tables highlight the number of clusters that each combination would have selected based on the three different metrics mentioned previously. The selected number of clusters was determined by a majority vote and if no majority then Silhouette clustering was selected. The tables also display whether the selected number of clusters matched the number of clusters from the maximum V-Measure score.

6.1 Cancer Dataset

The number of clusters that resulted in the highest V-Measure was two for all combinations. This aligns with the dataset since there were two target classes. Every combination, except ICA,

selected this as the appropriate number of clusters. Consistently throughout this experiment, ICA showed it was not a suitable dimension reduction technique for the Cancer dataset. It was not as surprising that LDA produced the highest V-Measure score since it utilized the target classes when formulating its clusters.

Table 3 (Cancer)								
Dim. Reduction	PCA	PCA	ICA	ICA	RP	RP	LDA	LDA
Clustering	KM	EM	KM	EM	KM	EM	KM	EM
Number of Clusters								
Silhouette	2	2	4	4	2	2	2	2
CDI	2	2	4	12	2	12	14	13
DBI	7	9	8	9	3	3	5	3
Majority	2	2	4	4	2	2	2	2
VM Majority	0.723	0.659	0.640	0.580	0.319	0.479	0.769	0.795
VM Maximum	0.723	0.659	0.758	0.676	0.319	0.479	0.769	0.795
VM Clusters	2	2	2	2	2	2	2	2

6.2 Steel Dataset

The number of clusters really varied based on the combination of techniques used. Even when LDA utilized the target class for insight into the clusters, it only produced two or three clusters.

Table 4 (Steel)								
Dim. Reduction	PCA	PCA	ICA	ICA	RP	RP	LDA	LDA
Clustering	KM	EM	KM	EM	KM	EM	KM	EM
Number of Clusters								
Silhouette	2	2	8	2	2	2	2	2
CDI	2	2	9	9	2	2	14	14
DBI	4	3	2	4	3	5	7	7
Majority	2	2	8	2	2	2	2	2
VM Majority	0.268	0.294	0.311	0.269	0.202	0.296	0.323	0.313
VM Maximum	0.331	0.341	0.311	0.309	0.257	0.316	0.323	0.316
VM Clusters	8	8	5	14	14	10	2	3

This was much less than the seven target classes included in the Steel dataset. The unsupervised techniques seemed to cluster the data into more groups so this could lead to some insight into how the original target classes were determined. There may be evidence for samples to be misclassified or not enough valuable information to properly separate the classes for this dataset.

7 References

1. Hyvärinen, Aapo and Oja, Erkki (2000). Independent Component Analysis: Algorithms and Applications. HUT, Finland.
http://mlsp.cs.cmu.edu/courses/fall2012/lectures/ICA_Hyvarinen.pdf