Assignment 3 : CS 7641

Muhammad Haris Masood
mmasood30@gatech.edu

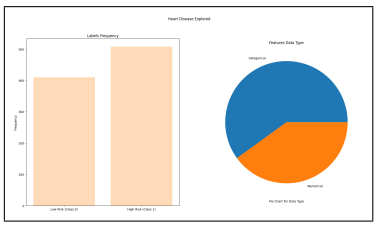

Figure 1. HEART Dataset Explored

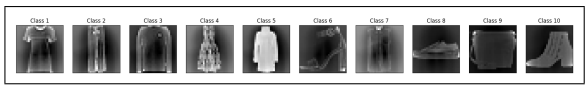

Figure 2. FMINST Dataset Visualized

1. Introduction and Data

1.1. Overview

My two chosen classification problems are classifying heart attack risk [4] (heart) and, a sampled version of the Fashion-MINST dataset (FMINST). For the heart dataset the features are several health markers for a patient, and the label identifies if there is a high or low risk of a heart attack. On the other hand, the FMINST dataset is a collection of 28×28 images that depict various fashion items. FMINST is a ten class multi-class dataset.

Figure 1 reveals the heart dataset's characteristics. Chosen for its balanced mix of categorical and numerical features, it showed potential for feature reduction and hinted at clusters. In contrast, the FMINST dataset, with equal samples across all 10 labels, offered ideal structure for clustering and dimensional reduction, alongside the ability to visualize intermediate results, making it a valuable complement to our heart dataset.

2. Hypothesis

1) Dimensionality reduction should greatly benefit the FMNIST dataset due to the prevalence of non-informative pixels.

2) EM clustering might struggle with the heart dataset due to its mixed nature and numerous categorical features.

3) Independent Component Analysis (ICA) should perform well on the FMINST dataset, producing meaningful results, based on our understanding from lecture material.

4) We expect that the optimal neural network (NN) architecture will be simpler for dimensionally reduced data, due to lower input dimensionality.

3. Experimentation Methodology

3.1. Preprocessing Data

For the heart dataset, we removed the Booking ID column. All categorical features were one-hot encoded. Both datasets were normalized.

3.2. Metrics

For the clustering experimentation, we utilized a variety of metrics to guide our cluster selection process. These metrics included Silhouette score, Within Cluster Distances (WCSS), BIC scores and AIC scores. For final evaluation of the clustering algorithm we relied on Adjusted Normalized Mutual Information Score (NMI) and Adjusted Rand Score (ARI). In addition to these metrics we utilized visual evaluation of the cluster plots. We believe that employing a range of diverse metrics for our clustering algorithms would lead to us arriving at the optimal result. Moreover, our choice of metrics was informed by their widespread usage in academic literature within the field, ensuring compatibility and comparability with existing research.

Important metrics used in our Dimensionality Reduction (DR) experimentation included Explained Variance (for PCA), Mean Absolute Kurtosis (for ICA), Pairwise Distance Error (for Random Projections), Trustworthiness (for Isomap) as well as Reconstruction error (for all). Similar to our metric selection process for clustering, we believed utilizing several metrics would be beneficial, and we utilized metrics that were use in literature for similar studies.

Due to the balanced nature of both the FMINST and heart dataset we selected accuracy as our metric for the Machine Learning section. We believe accuracy provides us a reliable metric to compare the results of our various datasets. Moreover, accuracy allows us to easily compare with results from our previous assignments/experimentation.

3.3. Procedure

Our paper progressed through four phases. In phase one, we applied explored two clustering methods, Gaussian Mixture Modeling (EM) and K-Modes, using our dataset. For K-Modes, numerical data required preprocessing via binning and one-hot encoding to align with the algorithm's categorical data requirement.To optimize the binning proces, we first determined the appropriate number of bins using a control dataset (MNIST). Our objective was to maintain spatial relationships before and after binning. We utilized t-SNE visualizations to observe these relationships graphically. Five bins were found to be sufficient, balancing effectiveness with manageable dimensionality. For the clustering itself each cluster value was repeated three to five times for each algorithm. The average scores for all runs was was used to select the optimal number of clusters. Once the optimal number of clusters were identified, the average NMI and ARI scores for three to fives runs was calculated. The same set of random seeds were utilized for all experiments to ensure reproducibility.

Phase two involved performing DR on both datasets. PCA, ICA, RP and Isomap were performed sequentially, all experiments were repeated three to five times. No additional preprocessing of the data was required at this step. For RP we chose Gaussian Random projection as sparse random projection appeared to have inconsistent performance.

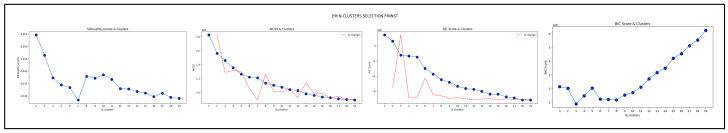

Figure 3. Various Metric Results & N-Clusters (EM FMINST)

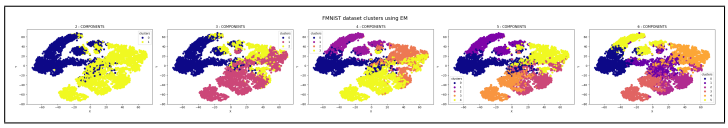

Figure 4. N-Clusters Visualization (EM FMINST)

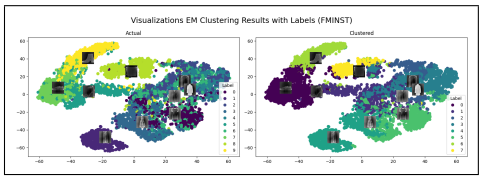

Figure 5. Comparing Actual & Final N-Clusters (EM FMINST)

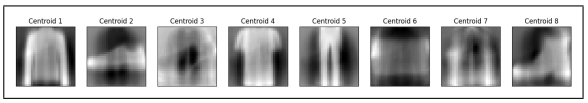

Figure 6. Visualizing Cluster Centroids (EM FMINST)

Phase three was to apply our initial two clustering algorithms on all eight of our DR'd datasets. Here again all data was converted to categorical one-hot encoded data prior to applying the K-Modes. Although we performed sixteen experiments our report only contains four experiments, namely, EM clustering after performing PCA and Isomap for both datasets.

Our final phase involved the Machine Learning. For this phase we selected the FMINST dataset and compared a simple sequential model's performance on standard normalized FMINST data, DR'd normalized FMINST data, and, normalized FMINST data but now containing the distance between that sample and each of the N cluster centroids (these were added in as N features).

4. Clustering

Table 1. Clustering Results

Dataset	Clustering Algorithm	NMI	ARI
FMINST	EM	0.52	0.38
	KModes	0.35	0.14
HEART	EM	0.1	0.05
	KModes	0.34	0.44

4.1. Gaussian Mixture Model (EM)

4.1.1 FMINST

To accurately determine the optimal number of clusters we tested several cluster sizes, results of each cluster evaluation can be observed in Figure 3. Our first graph helps visualize the trend in silhouette scores and cluster size, we note that as cluster size increases silhouette score decreases, this indicates a potentially poorer fit. A lower score could mean overlapping clusters or a decreasing density of points within the cluster, leading to higher average distance. One limitation of silhouette score is that the metric is specialized for assessing cluster quality when the clusters are convex in shape [8] due to this limitation we cannot solely rely on it. We do note high scores at clusters sizes 2,3,8 and 10. WCSS can be observed as our second subplot. As anticipated, WCSS decreases with an increase in the number of clusters. However, we utilize the elbow method (indicated by the red dotted line) to render this metric practical. Otherwise, the lowest score would occur when the number of clusters equals the number of samples, an inadequate solution. We note elbows present at cluster sizes of 2, 5, 8 and 13. The elbow method was also utilized for AIC, as relevant literature indicates that simply selecting the model with the minimum value can lead to overfit models [1]. We note elbows at 3 and 6. Our last subplot is that of BIC, however, we do not simply select the lowest BIC score as our ideal number of clusters, as there is research indicating that BIC may select overtly simple models in realistic situations [2]. We do note that a low BIC score is observed at values 3,6,7 and 8. Finally, we turn to Figure 4 for qualitative selection. We must note that all of these subplots are 2D projections of the data and likely have some error/bias inherent in them, however, we can still make some conclusions. We note that it does appear that the data agglomerate's into several components as result it is likely that the ideal number of clusters is greater than 2 or 3. Utilizing all our results above we conclude that the ideal cluster value is likely 8. Our reasoning is as follows, a cluster value of 8 was an elbow point for WCSS, generated a reasonable silhouette score, and resulted in a low BIC score. We also note that this value is inline with our intuitive conclusion based on a visual inspection of the graphed data.

Figures 5 and 6 help visualize our results, while Table 1 presents relevant metrics. We note that the clustering algorithm performed moderately well. Many of the images of the cluster centroids align with the original dataset labels, indicating good separation. Additionally, utilizing Figure 6 to compare actual labels with predicted clusters highlights many instances of correct cluster assignment. However, we do note that the algorithm was not perfect. It appears harder to differentiate samples were clustered together, such as Sandals/Sneakers/Boots (Class 6/Class 8/Class 10) and Pullovers/Coats (Class 3/Class 5). In terms of metrics, the NMI score indicates a moderate level of performance [7], whereas the ARI score suggests only marginally better than random performance, albeit still a poor result. It is interesting to note that the ARI score indicates a worse result than NMI, we believe this could be due to the resultant clusters being pure and unbalanced [9]

Based on our experimentation, it seems that the application of the GMM/EM algorithm on the FMINST dataset yielded less-than-exceptional results. We have several hypothesis for why this may be the case. We first note that the GMM algorithm assumes that the underlying distribution is a composition of Gaussian's, indeed this may not be the case for FMINST dataset. In our initial data exploration of the FMINST dataset, we conducted several tests to assess the presence of an underlying normal distribution. These included visualizing histogram plots and QQ plots. However, we found no strong evidence of normality in the underlying data. GMM's are also sensitive to initialization, and although we repeated each run many times, we may need to employ smarter initialization strategies to yield better results.

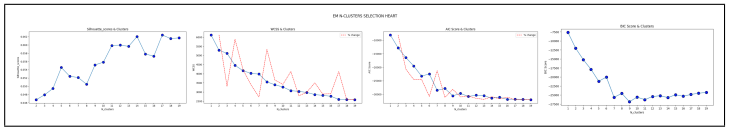

Figure 7. Various Metric Results & N-Clusters (EM HEART)

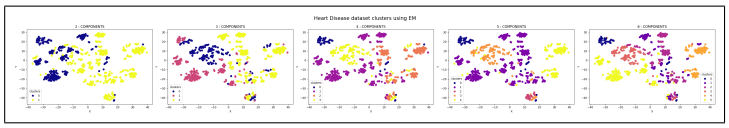

Figure 8. N-Clusters Visualization (EM HEART)

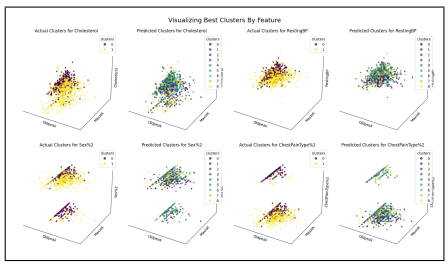

Figure 9. Sample Pair Plots (EM HEART)

4.1.2 HEART

A strategy similar to the one discussed in Section 4.1.1 was used to identify the ideal number of clusters. Figures 7 and 8 suggest the presence of multiple clusters in the sample data. Interestingly, the silhouette score shows a continuous increase with the number of clusters, possibly due to the categorical nature of the dataset, causing the data to appear more separated. Utilizing a holistic view of all metrics and visualizations we conclude that a cluster size of 9 is likely the ideal number. For this value, both the BIC and AIC scores are low, and although it is not an elbow point for the WCSS metric, further increases in the number of clusters lead to marginal decreases in WCSS.

Figure 9 helps visualize our results through a series of 3D pair plots. In each plot, the X and Y axes are fixed to the MaxHr and Oldpeak feature values. These features were selected as initial data exploration indicated that they were the two most predictive numerical features. This conclusion was drawn through testing with tree classifiers and by calculating correlation coefficients. The Z axis takes on the values of the other features of the dataset. Table 1 presents final evaluations. The visualizations and metric results show that that EM performed poorly on this dataset, NMI and ARI indicated performance slightly better than random chance, and, the pair plots indicated poor agreement between labels and assigned clusters.

The heart dataset primarily comprises of categorical features. We attribute the poor performance of GMM on this dataset to the discrepancy between the model's Gaussian assumption and the discrete nature of categorical data. Furthermore, our experimentation revealed a tendency towards favoring a high number of clusters, despite the binary nature of the data. This phenomenon could be attributed to the model's attempt to fit narrow Gaussian distributions to each of the numerous groupings formed due to the categorical nature of the dataset. Figure 8 provides visual evidence supporting this hypothesis.

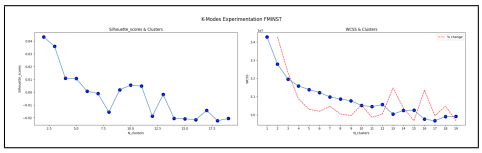

Figure 10. Various Metric Results & N-Clusters (KM CATEGORIZED FMINST)

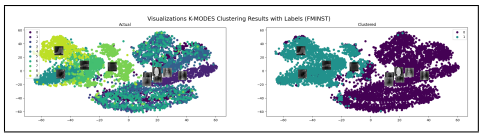

Figure 11. Visualizing K-Modes Clustering (KM CATEGORIZED FMINST)

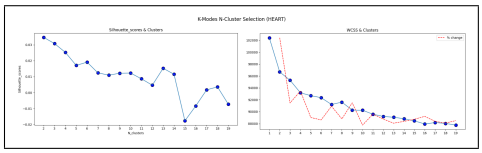

Figure 12. Cluster Selection (KM CATEGORIZED HEART)

4.2. K-Modes

4.2.1 FMINST

We visualize the metric evaluations for various cluster sizes for K-Modes on the FMINST dataset in Figure 10. Unlike EM, K-Modes is a non-parametric clustering method. Therefore, we cannot calculate BIC and AIC. However, we can rely on silhouette scores and WCSS to guide our search. We notice a similar trend as observed in section 4.1.1 for both silhouette score and WCSS, and, we believe the reasoning is the same.Based on our plots we identify the cluster size of 2 as ideal (WCSS elbow, high silhouette score)

Optimal clustering results can be visualized in Figure 11 and table 1. We note that that K-Modes performed significantly worse than EM. Both the NMI score and ARI score were significantly lower. Additionally, observing the grouping visualization highlights how the algorithm clustered all footwear together and all the tops (and jeans) together. We also note that the algorithm was split on bags, an amusing result. The clustering is logical; however, the granularity of the clustering was lost. We believe this was due to us having to convert the FMINST dataset into categorical data. We performed this conversion by binning and one-hot encoding. We now believe that this process led to significant information loss, resulting in the poor performance.

The K-Modes algorithm performed poorly on the FMINST dataset due to information loss during binning. Despite this, it still yielded meaningful results and could be valuable for downstream model training. This underscores the importance of choosing the right tool for the data; K-Modes is not suitable for numerical data.

4.2.2 HEART

We employed a similar elbow strategy as discussed in section 4.2.1 to determine the optimal number of clusters for the heart dataset. Figure 12 displays WCSS and silhouette scores. We observe a decrease in silhouette scores with an increase in the number of clusters. Moreover, a significant elbow is seen in the WCSS graph when transitioning from 1 to 2 clusters, indicating that 2 is an optimal choice.

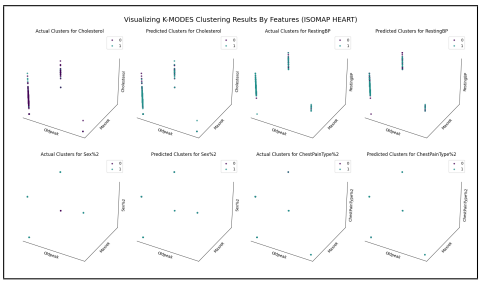

Figure 13. Results Visualized Pair Plot (KM CATEGORIZED HEART)

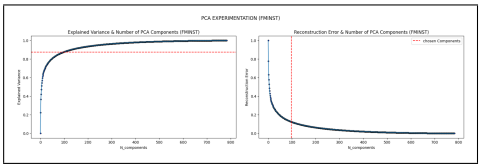

Figure 14. PCA Component Selection (FMINST)

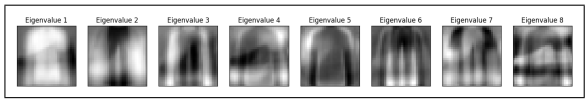

Figure 15. PCA Eigenvalues Visualized (FMINST)

We observe a significant improvement in clustering performance from both our quantitative evaluation (Table 1) and qualitative assessments (Figure 13), compared to the EM algorithm. The pair plots indicate a reasonable degree of similarity between the actual dataset labels and the assigned clusters. However, the results still indicate substantial room for improvement. We believe binning and one-hot encoding the numerical features of the dataset likely resulted in significant information loss, leading to poor clustering results.

K-Modes demonstrates its effectiveness on categorical data, producing notably superior clusters compared to EM. However, our binning strategy likely led to substantial information loss and poor performance. We propose a hybrid approach for mixed datasets, such as combining K-Means and K-Modes, to more effectively cluster mixed datasets.

5. Dimensionality Reduction

5.1. Principal Component Analysis

5.1.1 FMINST

We employed both a quantitative and qualitative approach to select our optimal number of components for PCA. For our quantitative approach we refer to relevant work done by Matan et al. [5] to calculate a hard threshold for singular values by utilizing the following equation $\lambda \geq 2.858 \cdot y_{\text{med}}$, where λ is the threshold for the singular values, and, y_{med} is the median singular value. Our qualitative approach involved analyzing explained variance and reconstruction error graphs presented in Figure 14. Combining both approaches we arrive at a value of 95 components, this respects the quantitative formula, and, it does appear that this value is close to elbow points within each of our plots.

The visualized components of our PCA object are shown in Figure 31. This visualization is particularly interesting as it provides evidence for the eigenfaces concept discussed in the lecture. We observe that PCA has generated an 'average' representation

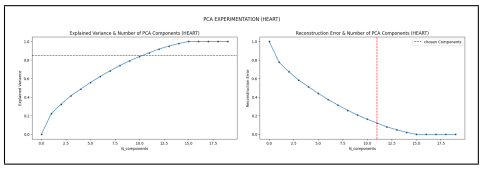

Figure 16. PCA Component Selection (HEART)

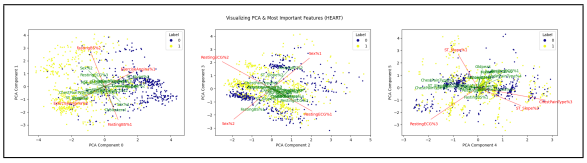

Figure 17. PCA Loading Plot (FMINST)

of fashion items. Additionally, it appears that PCA has identified brightness/luminance as a direction of variance. This does not mean that PCA was not useful, as it does appear that we have been able to successfully reduce our dimensionality by $\approx 88\%$ with minimal loss in information, but, it does appear that less than useful patterns may have been captured.

One unique benefit of PCA is the ability to get an understanding of collinearity in the dataset. We can evaluate the explained variance graph to perform a qualitative evaluation on the degree of collinearity, and, a qualitative evaluation of the singular values. Both the graph and our evaluation of singular values indicate a high degree of collinearity. There is a sharp elbow in the graph, and there is a significant drop-off in singular value magnitudes after the first few.

5.1.2 HEART

The quantitative strategy we applied in Section 5.1.1 could not be applied to our heart dataset due to its small sample size. Therefore, we turned to the literature and consulted metric plots in Figure 16 to guide our approach. We found that an explained variance ratio of 0.85 was a good threshold for our data [6] (components display even spread of explained variance). Utilizing this value, we were able to identify 11 as a good choice for the number of components. Our graphs provide further support for our findings, as this value results in a minimal reconstruction error.

Our loading plots for selected principal components can be observed in Figure 17. The loading plots highlight some very interesting conclusions. It appears that features representing angina, diabetes, sex, and ECG are very important in explaining the variability of the data. These features likely play a crucial role in the underlying pattern of the distribution and are likely very important for classification. PCA is a powerful tool that can also be employed for feature selection when paired with loading plots.

Applying PCA on the heart dataset not only aids in feature selection, but, yields very useful results, most importantly, by highlighting potentially important features. Additionally, we note that from the metric graphs it appears that reconstruction error is minimized, and, explained variance is maximized before the number of components equal the number of features. We believe this is due to how we pre-processed the data, the heart dataset is comprised of many binary categorical features that we one hot encoded, regardless, it is impressive that PCA was able to identify that we may have redundant features in our dataset.

It appears that PCA performed quite well on our dataset. We

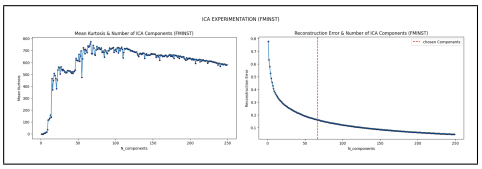

Figure 18. ICA Component Selection (FMINST)

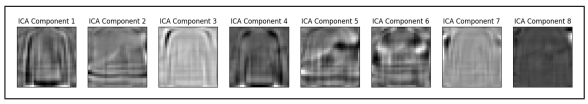

Figure 19. Ideal ICA Components Visualized (FMINST)

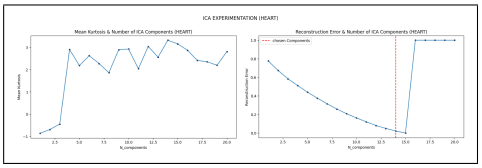

Figure 20. ICA Component Selection (HEART)

achieved a $\approx 50\%$ reduction in dimensionality with minimal information loss. Additionally, the loading plot reveals clear patterns and relationships. We believe this strong performance is likely due to the fact that our dataset is inherently simple, has linear relationships, and can be explained well with a linear technique. Similar to FMINST, we can utilize results from PCA to get an understanding of collinearity within the data. It appears there is virtually no elbow in the explained variance with number of components plot, additionally, we observed a gradual decrease in singular values, indicating significantly less collinearity than what observed in the FMINST dataset.

5.2. Independent Component Analysis

5.2.1 FMINST

To select the ideal number of components we utilized both kurtosis and reconstruction error. For kurtosis we calculate the mean absolute value of all the independent components produced after ICA. The results of our experimentation are presented in Figure 18. We observe that the maximum kurtosis value was achieve with 66 components. This value resulted in a reconstruction error similar to the reconstruction error we observed for our ideal PCA DR, as seen in Section 5.1.1.

We have visualized several ICA components in Figure 19. We note that ICA has produced components that identify key features, textures, and objects in our data, as opposed to creating an "average" image. ICA, by trying to find statistically independent components has successfully captured key differences between images. ICA has proved to be an effective technique for our image data, reducing dimensionality by $\approx 92\%$, and has effectively performed image segmentation, aiding in our understanding. Our findings confirm what was discussed in lecture, ICA has been able to outperform PCA for image data, providing more meaningful results and a greater reduction in dimensionality for the same loss in information (reconstruction error).

5.2.2 HEART

To select the ideal number of components we utilized methodology similar to that as discussed in section 5.2.1. We note that the ideal

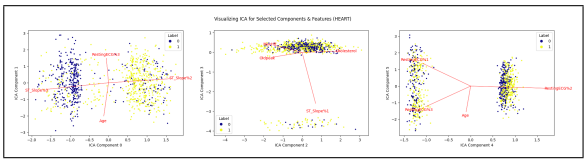

Figure 21. ICA Select Loading Plots (HEART)

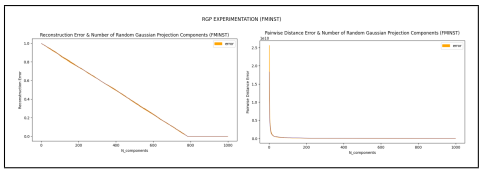

Figure 22. RGP Component Selection (FMINST)

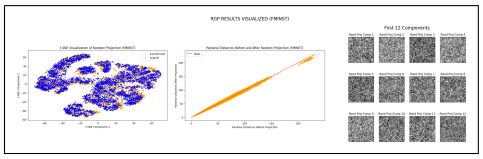

Figure 23. RGP Results Visualized (FMINST)

number of components for ICA for the heart dataset was identified as 14. The metric scores of our experimentation can be observed in Figure 20. It is interesting to note that the reconstruction error is 1.0 when we exceed 15 components. We believe the reasoning is similar to PCA, one-hot encoding binary features resulting in ICA calculating more independent components than 'true' features leading to large errors.

Figure 21 showcases the loading plot for ICA for select components along with the top 4 most important features for that pair. It is important to note that unlike PCA, the FASTICA method we used doesn't rank the components by any particular metric, as a result, their ordering is not meaningful. We can use the loading plot to make broader conclusions about the underlying patterns about the data, such as, it appears that the ST/HR slope is responsible for some of the underlying patterns within the data. Through our experimentation we identified ICA as an effective technique for DR, however, we note that for the heart dataset it appears to be slightly less efficient than PCA in terms of dimensionality reduction. For the same number of components we have a slightly higher reconstruction error. This may tie in with the heart dataset features having less collinearity than FMINST, and thus ICA requiring more components to properly represent the dataset. We tested this hypothesis with toy data and found our hypothesis held.

5.3. Random Projection

5.3.1 FMINST

We tried the elbow method with pairwise distance for RGP but failed to select an optimal number of components due to high reconstruction error. Consequently, we used a predefined threshold for reconstruction error to determine the number of components, basing our threshold value on the PCA and ICA experiments. We identified 666 as the ideal number of components for the RGP algorithm. The metric values of our experimentation can be observed in Figure 22, the high reconstruction error is of particular note. While difficult to discern, both the standard deviation (SD)/error for reconstruction and pairwise distance follows a similar trend: starting low, reaching a peak, and then gradually decreasing to

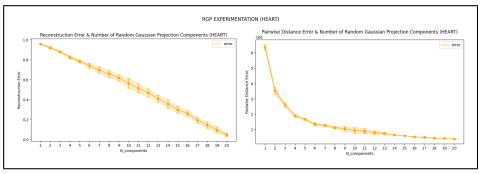

Figure 24. RGP Component Selection (HEART)

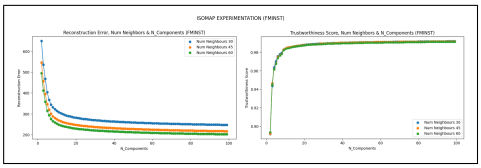

Figure 25. Isomap Component Selection (FMINST)

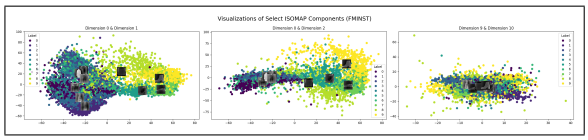

Figure 26. Isomap Results Visualized (FMINST)

zero. Surprisingly, the SD/error values remain close to the mean.

Several important metrics, and the components, of our ideal RGP projection are visualized in Figure 23. We note that random projection worked reasonably well. Both spatial and pairwise distance of the original dataset were preserved (albeit with small loss). The RGP components however do not provide any meaningful information. Compared to PCA and ICA, RGP provides less meaningful results, and, a smaller reduction in dimensionality, $\approx 15\%$. However, RGP was extremely cheap, $15\times$ faster than PCA and a $100\times$ faster than ICA.

5.3.2 HEART

A similar strategy to section 5.3.1 was employed and we identified 17 as the ideal number of components for the heart dataset. We observe the metric scores of our visualization in Figure 24, again note the high reconstruction error. Furthermore, the trend for SD/error is now clearer, showing a gradual increase followed by a decrease. It's worth noting that the SD/error is not notably large. Similar to FMINST, RGP provided a small reduction in dimensionality if we maintain the quality of information to match PCA/ICA. We again note that RGP was extremely cheap, $10\times$ faster than PCA and $40\times$ faster than ICA, however it does appear that the gains were substantially less than what we saw for the FMINST dataset. This underscores an important point: while RGP is advantageous for inexpensive dimensionality reduction, its computational advantage diminishes as the dimensionality/size of the original dataset decreases.

5.4. Manifold Technique - Isomap

5.4.1 FMINST

For Isomap, we needed to determine two parameters for dimensionality reduction: the number of components and the number of neighbors. We utilized reconstruction error and trustworthiness to guide our search, these are visualized in Figure 25. After evaluating the FMINST dataset, we found diminishing returns beyond 45 neighbors. It must be noted that Sklearn's implementation of

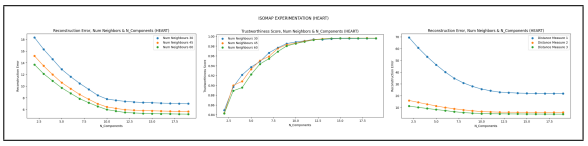

Figure 27. Isomap Component Selection (HEART)

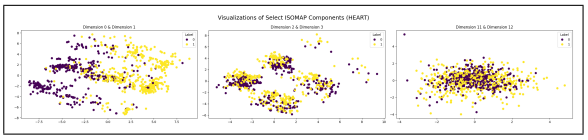

Figure 28. Isomap Select Components Visualized (HEART)

Isomap lacks mean squared error as its reconstruction error metric, thus we utilize the elbow method as opposed to setting a reconstruction threshold. We identified 20 as the optimal number of components.

Figure 26 displays select components of the DR'd dataset projected onto two dimensions. It's evident that the algorithm effectively preserved relationships between different images, positioning similar items closer together. Notice the proximity of sneakers, sandals, and toe boots, as well as the clustering of shirt, pullover, and coat images. An intriguing observation is that, akin to PCA, higher dimensions of the reduced dataset seem to convey less spatial information about the data, as it appears all images converge around the origin at higher dimensions. At this step we are unable to confirm if the algorithm performed better than the linear algorithms, but it does appear that important relationships between data were maintained, and, the dimensionality was successfully reduced by $\approx 97\%$, although the degree of information loss is uncertain. We can note that compared to the other DR techniques Isomap took significantly longer, $150\times$ PCA and $30\times$ that of ICA,

5.4.2 HEART

A similar strategy was applied to the heart dataset to select the ideal number of components and neighbors, as discussed in section 5.4.1. The results led us to identify 45 as the ideal number of neighbors and 13 as the ideal number of components. Figure 27 showcases the results of our experiments. Additionally, we tested the distance metrics' performance and noted that Euclidean distance performed the best, despite the dataset having a heavy categorical component. We believe this is likely due to the dataset still having many numerical features, which are too negatively impacted when Manhattan distance is used.

Figure 28 displays the visualization of select components post DR. Although more challenging to interpret than FMINST, it appears to preserve categorical separation, indicating the retention of important data relationships. However, Isomap achieved a significantly lower dimensionality reduction for the heart dataset compared to FMINST. This may stem from the dataset's small size, hindering the algorithm's ability to learn high-dimensional structure, or from the heart dataset's inherently complex high-dimensional structure, although our previous linear method experiments contradict this notion. We note that Isomap was significantly more expensive than the linear methods, the difference in cost being similar to that as discussed in the FMINST section, section 5.4.1.

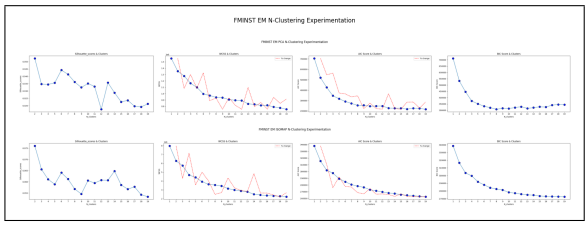

Figure 29. Clustering Experiments W/ DR'd FMINST

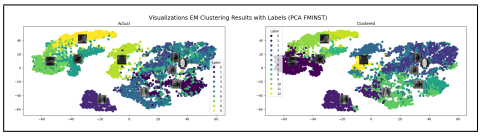

Figure 30. PCA DR'd FMINST EM Results Visualized

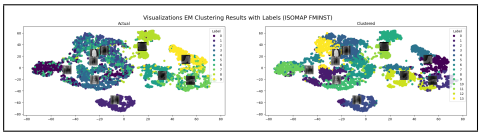

Figure 31. Isomap DR'd FMINST EM Results Visualized

6. EM/GMM Clustering on Dimensionality Reduced Datasets (PCA and Isomap)

6.1. Introduction

In this section, we assessed EM clustering on both datasets post-processing with the optimal PCA and Isomap algorithms. We omitted ICA and RP due to RP's poor performance and ICA's similarity to PCA, reserving the latter for the NN section due to it's effective image segmentation.

Dataset	DR Technique	NMI	ARI
FMISNT	PCA	0.56	0.40
	ISOMAP	0.57	0.42
HEART	PCA	0.15	0.08
	ISOMAP	0.18	0.14

Table 2. EM Clustering performance for dimensionality reduction techniques

6.2. FMISNT

6.2.1 PCA

We applied EM clustering on two different DR techniques. Our first DR reduction technique was the linear PCA method. We can observe the behaviour of silhouette score, WCSS, AIC and BIC in Figure 29. Utilizing techniques similar to what we discussed in section 4.1 we utilized the elbow method, as well as, the lowest scores for BIC to identify 13 as the optimal number of clusters for the PCA DR'd dataset. It is interesting to note that this value is higher than the original number of optimal clusters. We believe this is likely due to PCA creating new orthogonal projections of the data, resulting in more clusters being ideal as the data is more separable.

We can observe the performance, as measured by NMI and ARI, in Table 2. Additionally, we can visualize the actual clusterings in Figure 30. Comparing the clustering performance between pre- and post-PCA, we note a noticeable uptick in performance. We can visualize this improvement comparing our visualizations (Figure 5 and Figure 30), we note, better separation between the jeans and dress, and, the different types of tops.

It is likely that the clustering algorithm has benefited from the significant reduction in dimensionality. Additionally, by conducting this reduction with the amount of variability as its metric, it is possible that the redundancy has also decrease, leading to a less confused clustering algorithm. Additionally, PCA creates new data projections where better groupings may be observed, improving clustering.

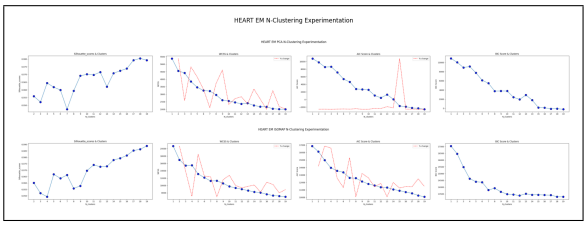

Figure 32. Clustering Experiments W/ DR'd HEART

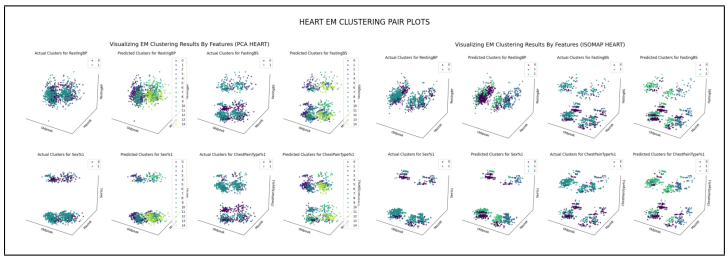

Figure 33. DR'd HEART Results Visualized (Pair-Plots)

6.2.2 Isomap

The ideal Isomap object was also utilized to DR the FMINST dataset. This reduced dataset also went through EM. We can note the effect of number of clusters on various metrics in Figure 29. Utilizing similar strategies as before, we identified 14 as the ideal number of clusters. Our ideal number of clusters is higher than the original dataset. We believe this may due to the projected data having better separability, resulting in more clusters being optimal.

The performance of the clustering algorithm on the Isomap projected data shows a reasonable improvement over the base case, and, a small improvement over the PCA projected dataset. This likely means that there exist high dimensional relationships within the data that the algorithm was able to take advantage of. If we compare the visualizations of the Isomap projected dataset (Figure 31), with that of the PCA projected dataset it does appear that there is some slightly better separation of image samples. Isomap resulting in slightly better clustering than PCA could be due to PCA identifying inconsequential features as important based on variability, luminance for example. This may have hampered PCA performance, but, as Isomap is more focused on attaining an understanding of the high dimensional structure of the data, it is less effected by such factors.

6.3. HEART

6.3.1 PCA

As with the other EM sections, we utilized a combination of elbow technique and lowest BIC scores to evaluate the ideal number of components. The metrics can be observed in Figure 32. The ideal number of components was found to be 15. This is higher than

what we calculated initially. We believe it is again a case of more separated data due to orthogonal projections.

Evaluating our results in Figure 33 and Table 2, we observe poor performance, slightly better than the original clustering. We attribute this poor performance to the dataset being mostly categorical. PCA is generally not well-suited to discrete-valued data, especially binary one-hot encoded data [3]. Another reason for the poor performance could be our choice of clustering algorithm. As mentioned earlier, EM does not handle discrete data well. Although we transformed our predominantly categorical dataset into a quasi-numerical format, it still retains many characteristics of its binary nature, which likely contributed to the poor performance.

6.3.2 Isomap

Analyzing the impact of the number of clusters on various metric values (as visualized in Figure 32) results in us identifying 7 as the optimum number of clusters. Surprisingly, this value is lower than the value we initially observed. This could be due to the Isomap algorithm simplifying the original dataset, producing a less complex projection. The performance of the EM clustering algorithm on the Isomap projected dataset is better than that on the base dataset and superior to the PCA projection dataset, but it still remains poor. We believe this is again attributed to the fact that the heart disease dataset is predominantly categorical, and the use of Euclidean distance (p=2) leads to suboptimal projections. However, it's important to note that the EM algorithm still outperformed both the base case and PCA. We attribute this to the algorithm's ability to capture some information about the high-dimensional structure of the data, which seems to be informative and consequently leads to better clustering results.

7. Neural Network Performance (FMINST)

Table 3. Neural Network Performance Comparison

Technique	Train Accuracy	Test Accuracy
Base Case	1.00	0.82
DR w/ ICA	1.00	0.85
DR w/ Isomap	1.00	0.78
EM Clusters Added	1.00	0.84
K-Modes Clusters Added	1.00	0.83

To gauge the effects of DR and clustering on a downstream learner, we utilized a NN to perform classification on our FMINST multi-class dataset. The performance of the classifiers is summarized in Table 3, we primarily measure accuracy. We excluded time measurements as we optimized the architecture for each dataset, rendering time comparisons less relevant.

The results of our experiments provide some very interesting results. We note that ICA performed the best out of all 6 cases, and, this follows what we have been taught in lecture, and, what we have gathered from our own research. ICA is able to successfully segment the different key features of image data, and, the projections allow for the learner to get a more rich understanding of the data, resulting in better performance.

We also note that inclusion of cluster distances from both EM and K-Modes improved learner performance. While the result for EM is somewhat expected, given our performance in the clustering section, it is surprising the K-Modes also lead to an improvement in model performance. It appears that even being able to cluster the images into two broad categories (shoes and tops + jeans) led to improved performance. We believe the performance is only marginally improved because there were many instances of incorrect cluster assignment, these likely confused the learner.

Finally, we note that the Isomap projected dataset performed poorly. In the Isomap 2D projection visualizations (Figure 26), we observed that the algorithm had difficulty distinguishing similar-looking images. There was significant overlap between shirts, coats, and pullovers. We believe this overlap made it harder for the classifier to separate the projected dataset, resulting in poor performance. One potential reason for the algorithm's poor performance could be the large variability in the image data, which may have made it difficult to accurately capture the geometric relationships.

8. Conclusion and Next Steps

In this paper we performed a deep dive for clustering and dimensionality reduction, at the start of the paper we outlined several initial hypothesis and now, we can assess their validity. Our hypothesis regarding FMINST was confirmed: most dimensionality reduction techniques significantly reduced dimensionality, except for RGP. This outcome suggests that the simplicity of the images allows for effective preservation of dataset information with minimal feature representation. Our second hypothesis concerning the heart dataset was verified: the application of EM/GMM resulted in suboptimal clusters. This outcome is likely attributable to the categorical or discrete nature of the dataset. Our third hypothesis, concerning the application of ICA on the FMINST dataset, was validated. ICA demonstrated excellent reduction in dimensionality, provided interpretable results, and yielded superior classification performance. Our final hypothesis expected neural networks trained on dimensionality-reduced data to perform equal to or better than the base model with simpler architecture. However, we found no clear correlation between model complexity and the dimensionality reduction degree. This may be because neural networks naturally ignore irrelevant features or because reduced datasets had fewer but more complex relationships between features, requiring more intricate models.

Although our experimentation was detailed and in-depth, not all our questions were answered, and new ones arose from our results. The reason why most DR'd datasets benefited from more clusters, the optimal method to handle mixed datasets, the lack of correlation between neural network complexity and degree of dimensionality reduction are all areas that require further exploration.

References

[1] Hamparsum Bozdogan. Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. *Psychometrika*, 52:345–370, 02 1987. 2

[2] Kenneth P Burnham and David R Anderson. *Model selection and multimodel inference: a practical information-theoretic approach*. Springer, 2002. 2

[3] Michael Collins, S. Dasgupta, and Robert E Schapire. A generalization of principal components analysis to the exponential family. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001. 8

[4] Fedesoriano. Heart failure prediction dataset, 2021. 1

[5] Matan Gavish and David L. Donoho. The optimal hard threshold for singular values is 4/sqrt(3), 2014. 4

[6] I. T. Jolliffe. *Choosing a Subset of Principal Components or Variables*, pages 92–114. Springer New York, New York, NY, 1986. 4

[7] Nezamoddin Kachouie and Meshal Shutaywi. Weighted mutual information for aggregated kernel clustering. *Entropy*, 22:351, 03 2020. 2

[8] Mehrnoosh Monshizadeh, Vikramajeet Khatri, Raimo Kantola, and Zheng Yan. A deep density based and self-determining clustering approach to label unknown traffic. *Journal of Network and Computer Applications*, 207:103513, 2022. 2

[9] Simone Romano, Nguyen Xuan Vinh, James Bailey, and Karin Verspoor. Adjusting for chance clustering comparison measures. *Journal of Machine Learning Research*, 17(134):1–32, 2016. 2