Evaluation of Clustering and Dimensionality Reduction for High-Dimensional Datasets

Shawn Egan
Georgia Institute of Technology
segan3@gatech.edu

I. Introduction

Two clustering algorithms, k-means and expectation maximization (EM), are applied to two datasets to evaluate the effectiveness of their clustering capabilities. Three dimensionality reduction (DR) algorithms, principal component analysis (PCA), independent component analysis (ICA) and randomized projection (RP), are applied to the same datasets to evaluate ability to preserve information with less features. The clustering algorithms are then applied to the dimensionality-reduced datasets, and clusters are evaluated in the new feature space. The dimensionality-reduced datasets are then used as inputs for a simple Neural Network (NN) to investigate effectiveness of information preservation in the context of a classification task. Finally, the original clusters from k-means and EM are used as features along with the original datasets as input for a NN. All 5 NNs (3 DR and 2 clustering) are evaluated on network size, classification performance, and training time.

A. Datasets

The two datasets used for comparison of all algorithms are the Spotify and Customer datasets. The Spotify dataset includes data from 2023 on the top streamed songs from Spotify. The Customer dataset includes demographic and spending-habit data for customers of a particular business. Both raw datasets are cleaned and preprocessed to enable learning and DR. The only modifications are scaling and one-hot encoding. The resulting Spotify dataset has 30 features and 815 instances, while the Customer dataset has 25 features and 2208 instances.

The Spotify dataset uses a target variable of 'stream_category' based on number of streams for a given song. Five categories are defined for different streaming numbers, the same as in Assignment 1.

The Customer dataset uses a target variable of 'respondsToAds' based on whether or not a given customer has responded to any of the ad campaigns. The target variable is defined the same as in Assignment 1.

Hypothesis 1: *Dimensionality reduction will improve generalization performance (reduce bias) for the Neural Network classifier for the Customer classification problem.*

Justification

The purpose of dimensionality reduction is to make the dataset more manageable (lower feature space) while preserving predictive features and important underlying relationships in the data. Prior work investigating dimensionality reduction techniques for classification problems reveals that DR can result in better generalization performance [1], [2], [2].

Hypothesis 2: *Dimensionality reduction will enable the utilization of smaller Neural Networks for the Customer classification problem, leading to more efficient training times.*

Justification

Lowering the number of features for the dataset should allow smaller Neural Networks to provide predictive performance similar to the full dataset. Dimensionality reduction techniques are used on datasets of very high dimensionality, typically to improve computational efficiency [3].

Hypothesis 3: *The clusters generated from k-means and expectation maximization will provide more features for the Neural Network to use for classification, resulting in better classifier performance.*

Justification

The addition of unsupervised learning-predicted clusters will allow the Neural Network to utilize more information from the feature space, which will result in better performance on the Customer classification task [4].

II. Methods

A. Clustering

Clustering algorithms are implemented using scikit-learn's available implementations: the clustering package's KMeans and the mixture package's GaussianMixture (EM). Dimensionality reduction algorithms are implemented using scikit-learn's available implementations: the decomposition package's PCA and FastICA, and the random_projection package's GaussianRandomProjection. All NNs are implemented with the neural_network package's MLPClassifier.

Both clustering algorithms are run on each dataset for n=2 to n=10 clusters, and averaged over 10 random seeds.

Both clustering algorithms are evaluated on silhouette scores for each dataset. K-means is evaluated on inertia and mutual information for both datasets, and EM is evaluated on Akaike information criterion (AIC) and Bayesian information criterion (BIC).

B. Dimensionality Reduction

Principal components are generated for 10 seeds using PCA for each dataset, and optimal components are chosen based on a cumulative explained variance of greater than 90%.

Independent components are generated for 10 seeds using ICA for each dataset, and optimal components are chosen based on maximizing kurtosis for minimum number of components. A tradeoff point is selected where additional components have marginal kurtosis values.

Random projections are generated for 10 seeds for each dataset, and optimal number of projections are chosen based on pairwise distance preservation in the projection space (below 20% difference).

C. Clustering on Dimensionality Reduced Data

K-means and EM are run for all six dimensionality-reduced datasets (PCA, ICA, and RP for both Spotify and Customer data). Clustering is evaluated using the same criteria for the respective algorithm.

D. Dimensionality Reduction for NN Features

A Neural Network is hyper-parameter optimized on the full Customer dataset (25 features), with scoring parameter of positive-class recall. Positive-class recall is selected because it is the most important metric in whether a customer responds to ads or not: of all the people who do respond to ads, the goal is to capture as many of those people as possible. Parameters optimized are hidden layer size, activation function, and learning rate.

The dimensionality-reduced datasets are then used to train three separate Neural Networks, hyper-parameter optimized on the same three parameters. The hidden layer sizes tested for the three DR datasets are limited to the same size as the initial full dataset NN, and a topology of the same depth with each hidden layer having half the number of neurons. A topology of half-sized layers is chosen since the DR algorithms utilize roughly half the number of features as compared to the full dataset.

1) Performance Comparison

All four Neural Networks are evaluated on training time as well as positive-class recall.

E. Clusters as NN Features

The same Neural Network topology and hyper-parameters (for the full data) are used to train NNs utilizing features from the k-means clusters and the EM clusters. Clusters are one-hot encoded as individual binary features for the k-means clusters. Probability of cluster membership is used as the feature for the EM Neural Network.

1) Performance Comparison

All three NNs (full dataset, full data plus k-means clusters, and full data plus EM clusters) are evaluated on positive-class recall and training time.

III. Results

A. Clustering

Figs. 1 and 2 show silhouette scores, inertia, and normalized mutual information scores for k-means clustering on the Customer and Spotify datasets, respectively. Inertia is scaled by a factor of 100,000 to fit on the plot along with silhouette and mutual information.

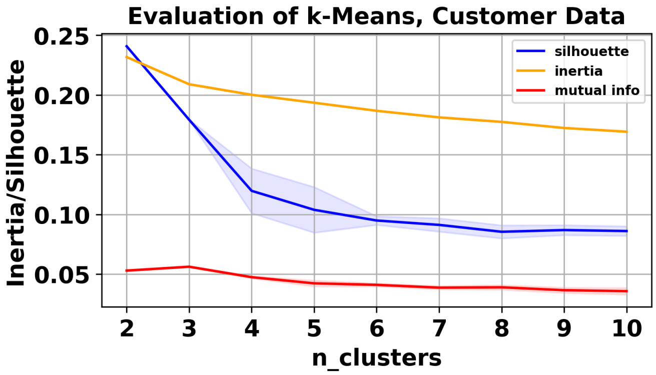

Fig. 1: Inertia, Silhouette Score, and Normalized Mutual Information for k-means clustering on the Customer dataset, 10-seed averages

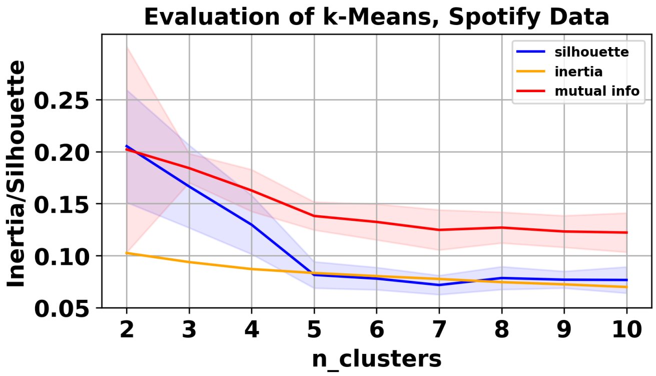

Fig. 2: Inertia, Silhouette Score, and Normalized Mutual Information for k-means clustering on the Spotify dataset, 10-seed averages

Figs. 3 and 4 show silhouette scores, AIC, and BIC scores for EM clustering on the Customer and Spotify datasets, respectively. AIC and BIC scores are scaled by a factor of 100,000 to fit on the plot along with the silhouette score.

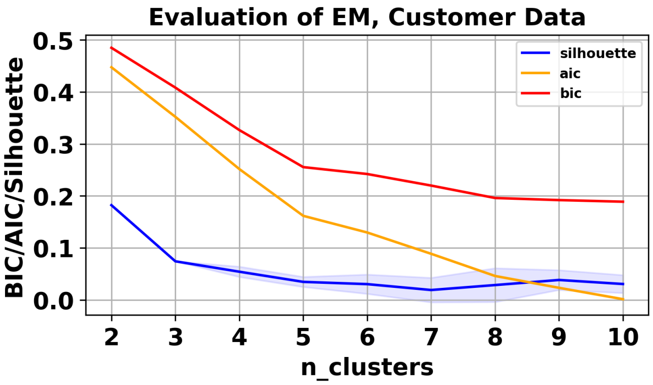

Fig. 3: Silhouette Score, AIC, and BIC for EM clustering on the Customer dataset, 10-seed averages

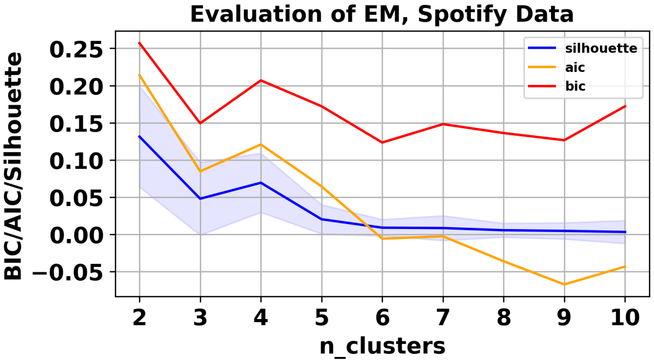

Fig. 4: Silhouette Score, AIC, and BIC for EM clustering on the Spotify dataset, 10-seed averages

B. DR

1) PCA

Figs. 5 and 6 show cumulative explained variance over all number of components, along with a red line indicating the number of components at which cumulative explained variance is greater than 90%.

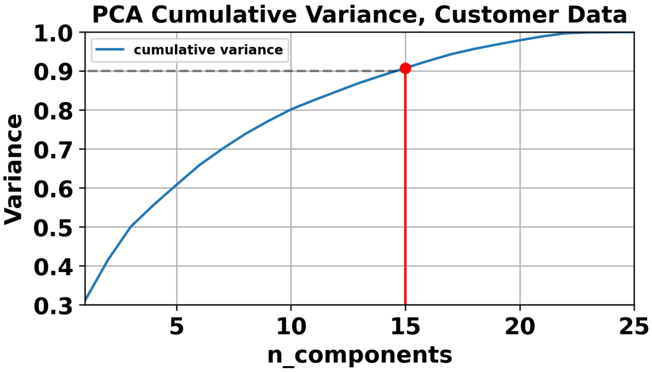

Fig. 5: Optimal number of components for PCA on Customer data

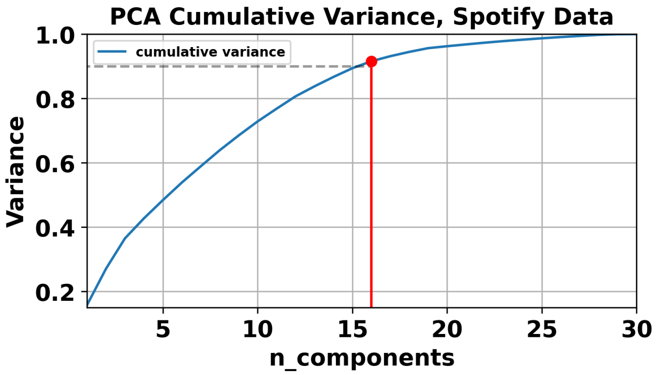

Fig. 6: Optimal number of components for PCA on Spotify data

2) ICA

Figs. 7 and 8 show (sorted) kurtosis scores over all components for the Customer and Spotify datasets, respectively. The red lines/dots indicate the point of diminishing returns for both datasets, where the components beyond that point are essentially Gaussian in nature and add no information.

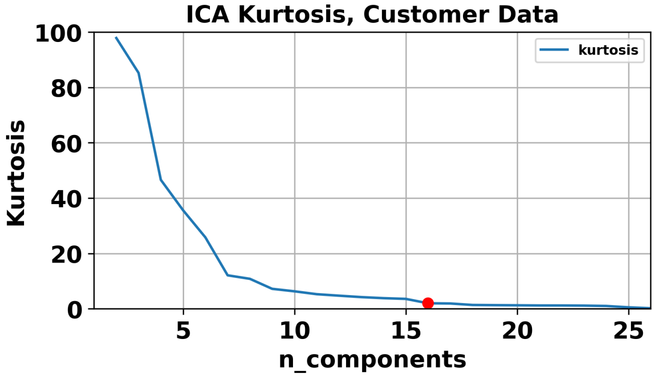

Fig. 7: Optimal number of components for ICA on Customer data

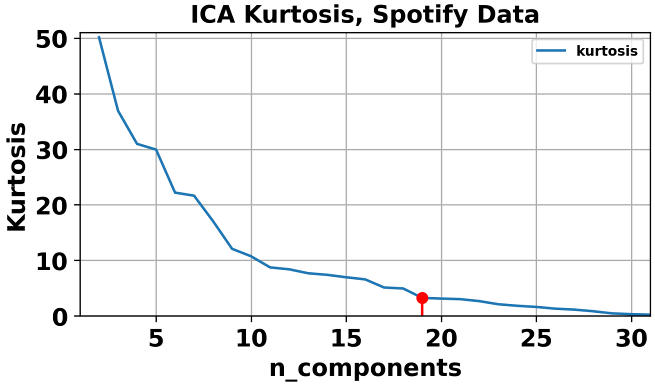

Fig. 8: Optimal number of components for ICA on Spotify data

3) RP

Figs. 9 and 10 show the distance preservation for RP over differing numbers of components for the Customer and Spotify data, respectively.

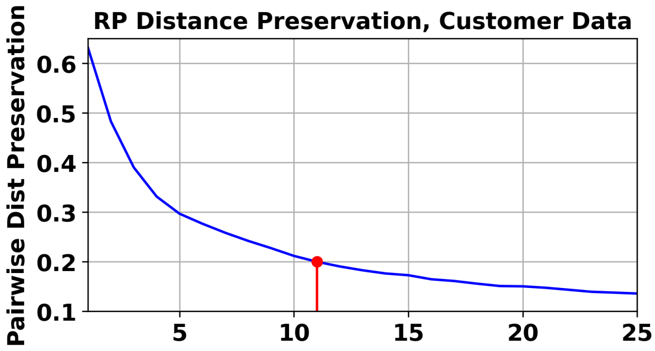

Fig. 9: Pairwise distance preservation for RP on the Customer data

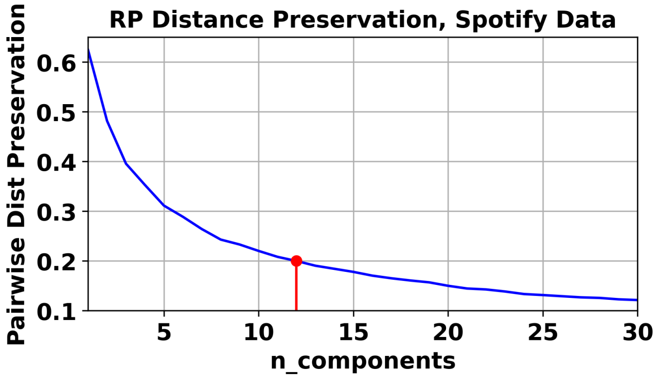

Fig. 10: Pairwise distance preservation for RP on the Spotify data

C. *Clustering on DR Data*

1) *k-Means*

Figs. 11 and 12 show the silhouette scores for the DR datasets on Customer data and Spotify data, respectively.

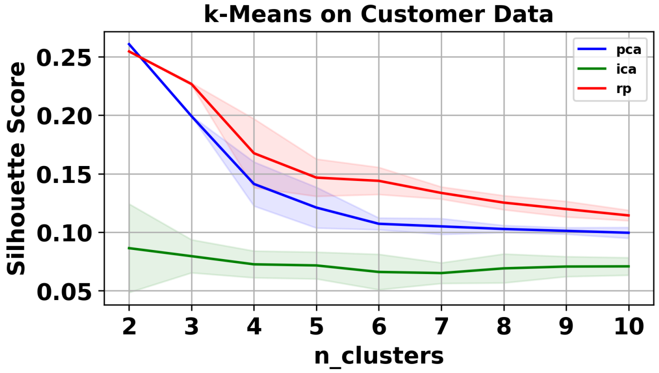

Fig. 11: Silhouette scores for PCA, ICA, and RP for k-means on Customer data

Figs.13 and 14 show a t-SNE pairplot of the first two principal components for both datasets, with n=4 clusters generated from k-means.

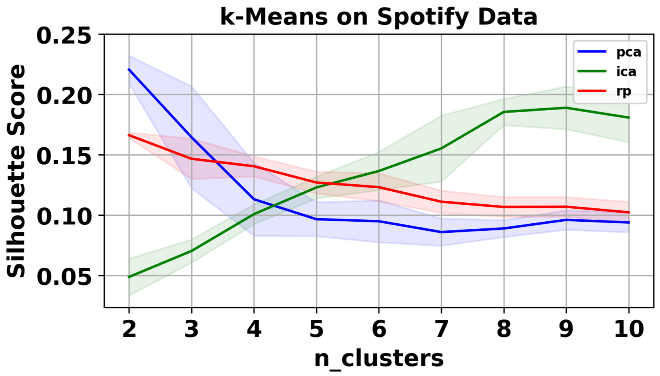

Fig. 12: Silhouette scores for PCA, ICA, and RP for k-means on Spotify data

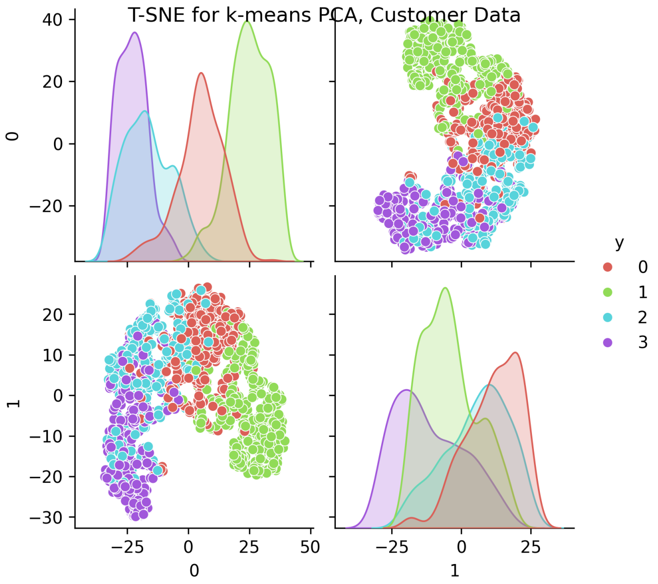

Fig. 13: Pairplot of t-SNE for 2 principal components of PCA, clusters generated from k-Means for n=4

2) *EM*

Figs.15 and 16 show a t-SNE pairplot of the first two principal components for both datasets, with n=4 clusters generated from EM.

Figs. 17 and 18 show AIC scores for the DR datasets on Customer and Spotify data, respectively.

Figs. 19 and 20 show BIC scores for the DR datasets on Customer and Spotify data, respectively.

D. *NN Comparison*

Table I shows NN hidden layers, learning rate, learning curve generation time, and positive-class recall for the full Customer dataset, the DR NNs, and the cluster-augmented NNs.

Fig. 21 show the Receiver Operating Characteristic (ROC) curves for all Neural Networks developed.

1) *Learning Curves for DR Data*

Fig. 22 show the learning curves for the full Customer dataset and all 3 DR NNs.

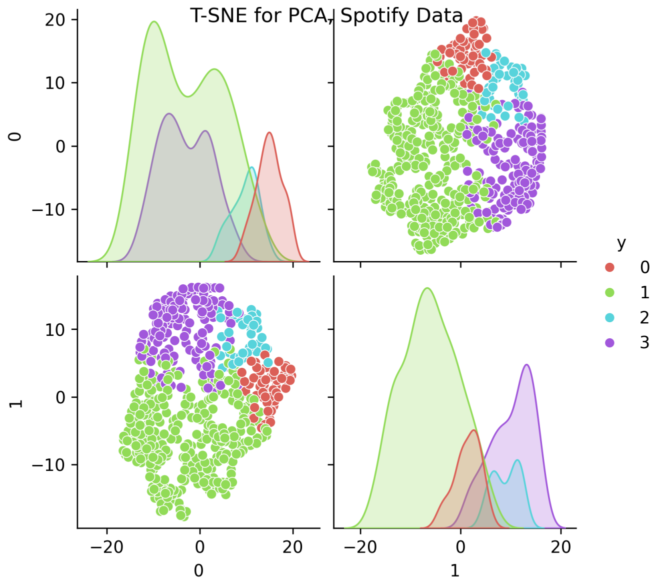

Fig. 14: Pairplot of t-SNE for 2 principal components of PCA, clusters generated from k-Means for n=4

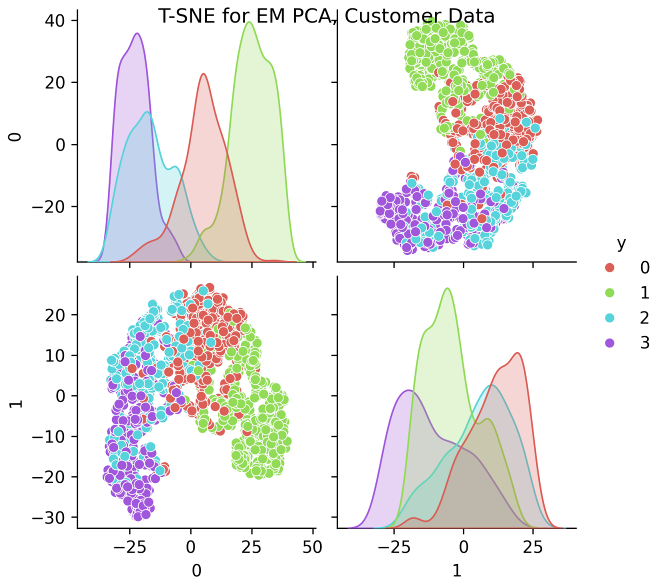

Fig. 15: Pairplot of t-SNE for 2 principal components of PCA, clusters generated from EM for n=4

TABLE I: NN Comparison for DR and Clustering Datasets

DR/Cluster	Layers	LR	Time (s)	Recall_1
None	(16,8,8,4)	2.0e-3	55.8	0.6080
PCA	(8,4,4,2)	6.3e-3	17.3	0.6533
ICA	(8,4,4,2)	1.1e-2	16.8	0.6281
RP	(16,8,8,4)	3.4e-2	12.2	0.5377
k-means	(16,8,8,4)	2.0e-3	49.7	0.6583
EM	(16,8,8,4)	2.0e-3	49.9	0.6482

Fig. 23 show the learning curves for the two cluster-augmented NNs.

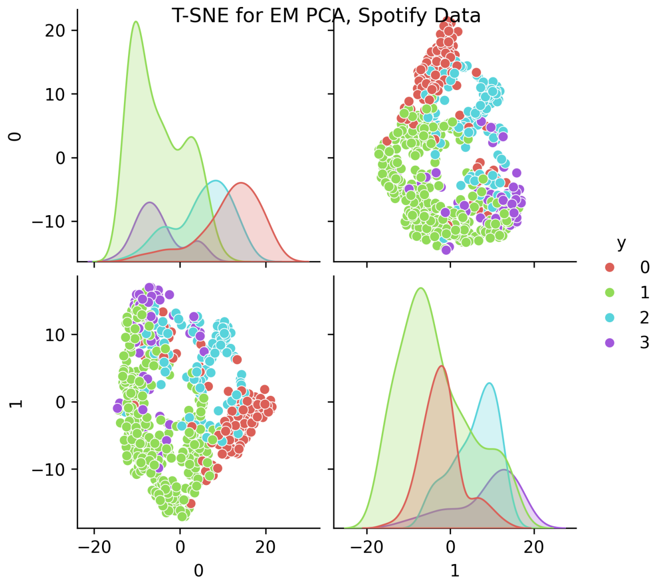

Fig. 16: Pairplot of t-SNE for 2 principal components of PCA, clusters generated from EM for n=4

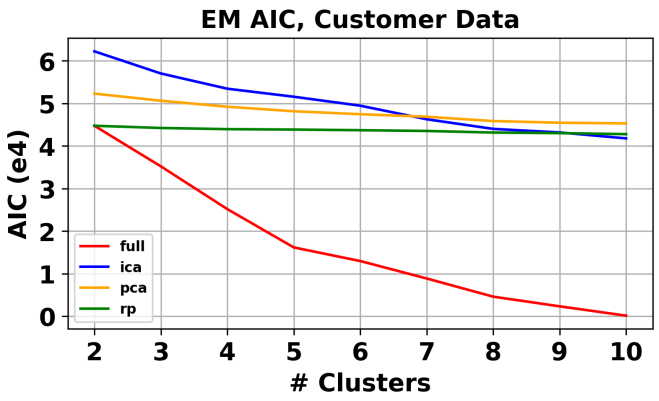

Fig. 17: AIC score for EM on Customer data

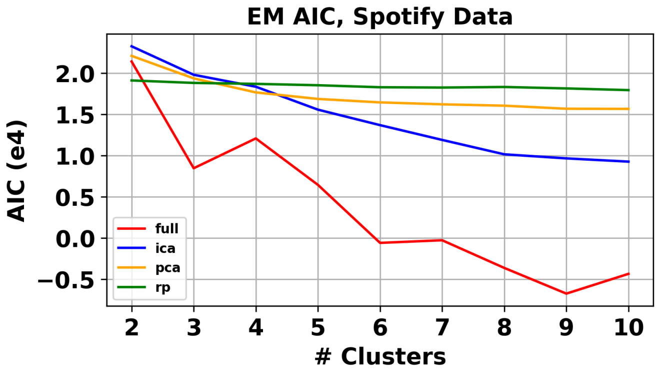

Fig. 18: AIC score for EM on Spotify data

2) *Learning Curves for Cluster Data*

IV. Discussion

A. Clustering

Figs. 1 and 2 show the decreasing performance of k-means for larger number of clusters. The Spotify

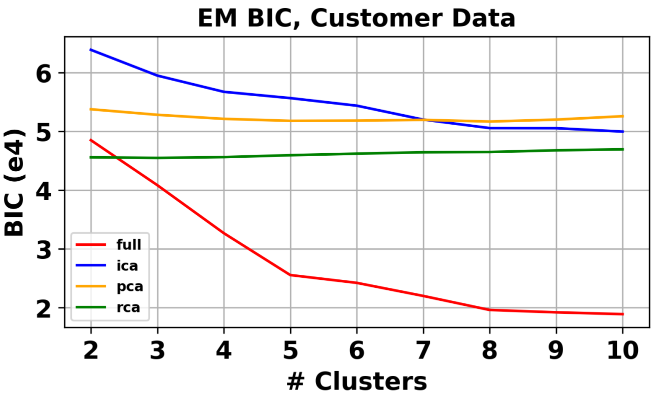

Fig. 19: BIC score for EM on Customer data

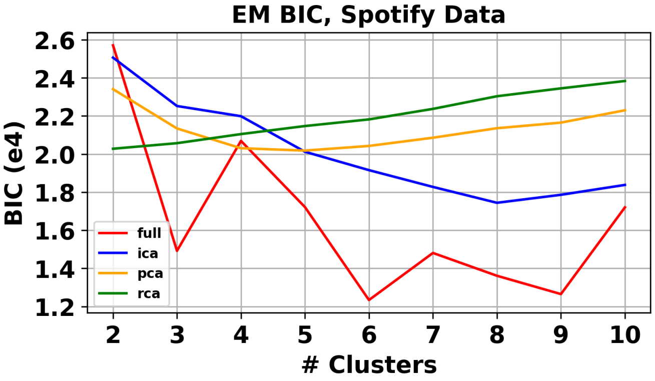

Fig. 20: BIC score for EM on Spotify data

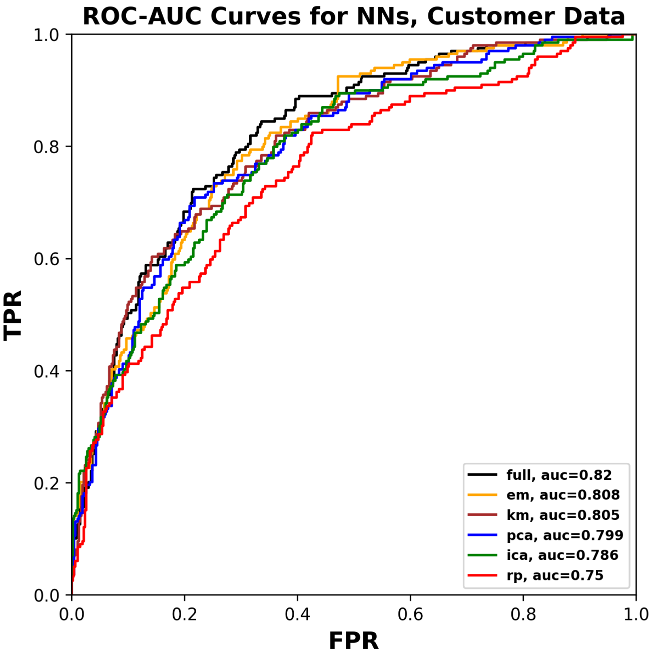

Fig. 21: ROC curves for all Neural Networks

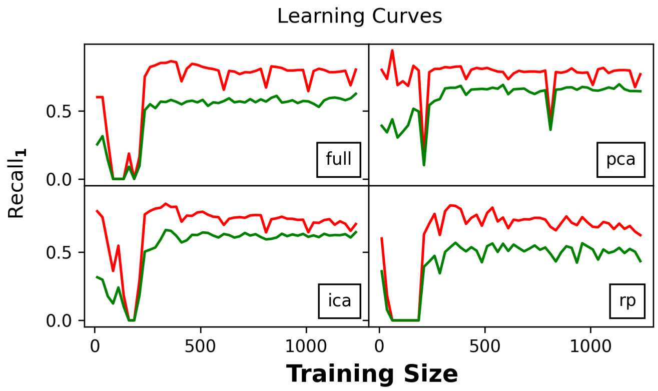

Fig. 22: Learning curves for full dataset and DR dataset NNs

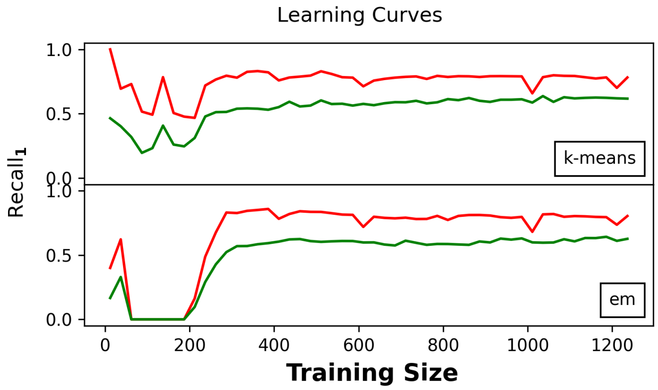

Fig. 23: Learning curves for clustering data

dataset has much more variability between seeds, evidenced by the plotted standard deviation windows for mutual information and silhouette score. Both datasets have relatively low silhouette scores, as 1 is the ’best’ value and indicates no overlapping of clusters. The Customer dataset has maximum silhouette for n=2 clusters (ss≈0.24) and minimum silhouette at n=10 clusters (ss≈0.09). The Spotify dataset has maximum silhouette for n=2 clusters (ss≈0.2) and minimum silhouette for n=10 clusters (ss≈0.08). The small silhouette scores for larger numbers of clusters indicates overlapping clusters, as is expected when data need to be assigned to more clusters. The silhouette score does not take on negative values, which is a good sign as negative values indicate bad cluster assignment. Optimal number of clusters for both datasets is somewhere between 3 and 5, where variance of the silhouette/mutual information score decrease and silhouette score is still high.

Inertia is seen to decrease slightly over the range of cluster values, which provides little information. Inertia is a measure of the distance of samples to cluster centers, which will naturally get lower with more clusters.

Mutual information follows the same trend as silhouette scores over the range of clusters for the Spotify data, providing little additional information. Mutual information gives almost no information for the Customer data.

Figs. 3 and 4 show tradeoffs between AIC and BIC for expectation maximization for both datasets. Lower values are better for both metrics. AIC assumes there is an unknown, high-dimensional reality which the features are approximating. Since AIC assumes a high-dimensional reality, it prefers complex models (higher number of clusters). BIC assumes there is a true model to predict (cluster) the data, and prefers simpler models (lower cluster number). The tradeoff of both information criterion is evident in both figures. BIC levels out beyond n=5 clusters, while AIC generally keeps decreasing.

B. DR

1) PCA

Figs. 5 and 6 show similar cumulative variance plots for PCA on both datasets. The Customer dataset is able to use a smaller number of components because the number of features is lower to begin with. Interestingly, the Spotify explained variance has more of a 'bowed' shape, indicating that the components generally hold more explained variance as compared to the Customer data.

2) ICA

Figs. 7 and 8 show the difference in ICA for the datasets. Kurtosis decreases more steeply for the Customer dataset, with component numbers greater than 10 having high Gaussianity. The Spotify figure shows a more steady decline in Kurtosis over the range of components, with components beyond 19 having high Gaussianity. Optimal component numbers were selected where Kurtosis no longer has any sharp decreases.

3) RP

Figs. 9 and 10 show similar pairwise distance preservation for both datasets. Less than 20% pairwise distance difference (compared to the full dataset) is reached slightly faster for the Customer data as there are already less features to start with.

C. Clustering on DR Data

1) K-means

Figs. 11 and 12 show k-means silhouette scores for all dimensionality reduction algorithms on the Customer and Spotify data.

The figure for the Customer data shows similar scores as the raw data for PCA and RP. ICA shows low silhouette scores over the range of clusters. This shows that PCA and RP do a relatively good job at preserving pairwise distances in the projection plane, but ICA fails to preserve distances and therefore k-means cannot find an effective clustering for the ICA data.

Interestingly, ICA performs better for larger numbers of clusters for the Spotify data, indicated by the increasing silhouette score as number of clusters increases. This suggests that ICA is able to develop components that 'spread out' in the projection plane, and therefore k-means can cluster well for larger numbers of clusters.

Figs. 13 and 14 show that PCA generates projections that k-means can effectively use to cluster data. The clusters are better defined for the Spotify data compared to the Customer data, but both plots show little separation between clusters, indicating that clustering may not be well suited for either dataset.

2) EM

Figs. 15 and 16 show that PCA generates projections that EM can effectively use to cluster data, though the clusters are less well-defined for the Spotify data as compared to k-means clustering (more overlapping instances). This suggests that the Spotify data may not provide meaningful distances for EM to cluster on.

Figs. 17 and 18 show AIC scores for EM on the dimensionality-reduced Customer data and Spotify data, respectively. The Customer data shows very little difference in performance for all dimensionality reduction techniques as compared to the full dataset. This suggests that EM may not provide optimal clusters for the Customer dataset.

Figs. 19 and 20 show BIC scores for EM on the dimensionality-reduced Customer and Spotify data, respectively. PCA and RP show little change for the Customer data, but ICA shows a consistent decrease in BIC (increase in performance).

PCA and RP show low optimal number of clusters for Spotify data, and ICA gives a larger number of optimal clusters. ICA shows good performance with respect to AIC and BIC for the Spotify data.

D. NN Comparison

Table I shows the results for all dimensionality-reduced and cluster-augmented Neural Networks, in addition to the full-data Neural Network.

1) DR

As expected, the PCA and ICA data require smaller hidden layer sizes for similar performance on the classification task. Interestingly, RP utilizes the full-sized NN, but requires significantly less training time, with the time being comparable to the other dimensionality reduced datasets. Most interestingly, though, is that PCA and ICA are able to outperform the full-dataset NN on the positive-class recall classification task. This shows that the dimensionality reduction, while removing some predictive capacity, actually generalizes better to unseen test data and therefore decreases bias. Additionally, the PCA and ICA datasets allow the utilization of a NN with half the number of hidden layers, resulting in a more than 70% reduction in training time while outperforming the full NN.

Though the RP NN underperformed with respect to positive-class recall, the training time is still an interesting result. The RP allowed an almost 80% reduction in training time for the same full-size NN topology, with some contribution coming from the utilization of a larger learning rate.

Fig. 22 show the learning curves for the full dataset along with all dimensionality-reduced datasets (green is validation score). All 4 NN are seen to learn quickly, reaching near-optimal performance with less than 250 training samples.

2) Clustering

The results table shows that the k-means and EM clusters provide good predictive ability. The addition of the k-means clusters as features provides an 8.3% boost to positive-class recall on unseen test data. The addition of the EM cluster probabilities as features provides 6.6% boost to positive-class recall on unseen test data. Neural network topology and learning rate are kept constant, utilizing the same parameters as the full NN. Training times are comparable for all three networks.

Fig. 23 show the learning curves for the k-means- and EM-augmented Neural Networks (green is validation score). Both networks are seen to learn quickly once training sizes reach 250 instances.

E. Conclusions

Hypothesis 1 was proven partially correct given the results observed. Dimensionality reduction (PCA and ICA) provided increased performance for the Customer classification task.

Hypothesis 2 was proven partially correct. The PCA and ICA NNs used NNs of half the full-size NN, requiring significantly less training time. The PCA and ICA NNs were better in every tested metric as compared to the full-sized NN on the full dataset. However, randomized projection did not use the smaller architecture, but was able to decrease the training time even more than PCA and ICA. Randomized projection did show decreased performance as compared to the NN on the full-dataset.

Fig. 21 shows a fuller picture of the performance of all neural networks. The ROC curve for the full NN with the full dataset shows more well-rounded classification performance as compared to all other NN classifiers (AUC=0.82). Randomized projection is still the worst performing augmentation of all dimensionality reduction and clustering techniques. Expectation maximization, k-means, PCA, and ICA all have AUCs roughly around 0.8, with ICA having the worst AUC of 0.786, a ∼4% reduction compared to the full neural network.

Hypothesis 3 was proven correct. Clusters generated from k-means, and cluster probabilities generated from expectation maximization provided more information to the NNs and allowed better positive-class recall performance on the Customer classification task.

The results in general show the power of dimensionality reduction and clustering. Dimensionality reduction or clustering alone may not provide the means to solve a classification problem, but when combined with the power of a Neural Network, both unsupervised tasks can increase classification performance and reduce Neural Network size, training time, and bias.

V. Resources

[1] Wang, W. and Carreira-Perpinan, M. A. ”The Role of Dimensionality Reduction in Classification.” *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 28. *Association for the Advancement of Artificial Intelligence*. Accessed: Mar. 23, 2025. https://cdn.aaai.org/ojs/8975/8975-13-12503-1-2-20201228.pdf

[2] Abdulhammed, R., Musafer, H., Alessa, A., Faezipour, M., and Abuzneid, A. ”Features Dimensionality Reduction Approaches for Machine Learning Based Network Intrusion Detection.” (2019) Electronics, 8(3), 322. doi: https://doi.org/10.3390/electronics8030322 https://www.mdpi.com/2079-9292/8/3/322

[3] Reddy, G. T., et al. ”Analysis of Dimensionality Reduction Techniques on Big Data,” *IEEE Access*, vol. 8, pp. 54776-54788, 2020, doi: 10.1109/ACCESS.2020.2980942. https://ieeexplore.ieee.org/document/9036908?denied=

[4] K. M. Faraoun and A. Boukelif, “Neural networks learning improvement using the K-means clustering algorithm to detect network intrusions”, *INFOCOMP Journal of Computer Science*, vol. 5, no. 3, pp. 28–36, Sep. 2006. https://infocomp.dcc.ufla.br/index.php/infocomp/article/view/140

[5] Nakamura, K. ”ML LaTeX Template”. *GitHub*. (2023). Accessed: Feb. 5, 2025. https://github.com/knakamura13/cs7641-ml-study-materials-2023/tree/main