Clustering and Dimensionality Reduction

Overview

Unsupervised learning is an area of machine learning concerned with developing inferences from unlabeled data, and dimensionality reduction is another area of machine learning concerned with the process of reducing feature sets used in analysis. When both are used together, one can uncover meaningful insights. This paper will analyze and discuss the results achieved when unsupervised algorithms - such as k-means clustering and expectation maximization - and dimensionality reduction algorithms - such as principle component analysis, independent component analysis, randomized projections, and recursive feature elimination - are applied to the Wisconsin Breast Cancer Dataset and the Abalone Dataset.

Data

Wisconsin Breast Cancer Data: The Wisconsin Breast Cancer Dataset is from the UCI Data Repository. The data includes 30 attributes that are used to describe patient tumor cells. The attributes include different measurements, such as size and density, for each tumor. Below are some reasons why I found the data to be interesting:

- **Numerical Data:** The data collected for this dataset are all numerical. It will be interesting to compare approach and results with the Abalone Dataset which contains mixed data (categorical and numerical).
- **Additional Noise:** Additional noise was introduced into the dataset to see how the different algorithms handle it. The noise constitutes for roughly 10% of the data. The noise will manifest itself as irreducible error.
- **Application:** I have seen loved ones battle cancer and witnessed its effects on family and friends. I thought it would be interesting to see how analytics can be used in the fight against cancer.

Abalone Data: Abalone Dataset is also from the UCI Data Repository. The data includes 8 attributes (10 after one hot encoding) that are used to identify the number of rings of an abalone. The attributes include different measurements, such as diameter and height, for each abalone. Below are some reasons why I found the data to be interesting:

- **Mixed Data:** The data collected for this dataset are mixed (categorical and numerical). It will be interesting to compare approach and results with the Wisconsin Breast Cancer Dataset which contains only numerical data.
- **Many Classifications:** The original dataset has 29 classifications, which was distilled down to 3 classes. It will be interesting to compare the results to a binary dataset such as the Wisconsin Breast Cancer Dataset

Analytical Approach

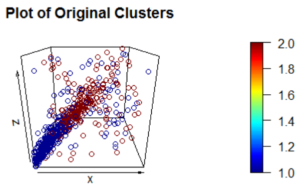
In developing the analytical approach for this project, certain key decisions were made regarding how the experiments were to be performed, ensuring proper analysis. Below is a summary of those key decisions.

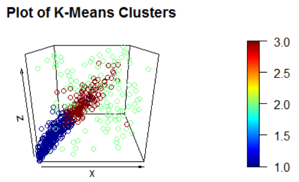
- **One Hot Encoding:** Many of the algorithms (both clustering and dimensionality reduction) require that the data be mapped to a Euclidean Space. It makes it difficult for these algorithms to handle categorical data in this space without performing one hot encoding first.
- **Euclidean Distance:** The Euclidean distance was selected as the measure of distance for the k-means clustering algorithm. The Euclidean distance is often used as the de-facto standard of measure, when there is a lack of domain knowledge to help guide the decision.
- **Cluster Evaluation:** V-Measure will be used to rate the overall performance of clustering algorithms, since it takes into account both completeness and homogeneity.
- **Reducing Randomness:** As the name suggests, randomized projects introduces a level of randomness into the results. To reduce the level of randomness, projections will be evaluated over 25 trials.

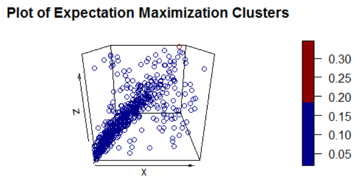
Clustering

For the initial clustering of the Wisconsin Breast Cancer Dataset and the Abalone Dataset, both the k-means and expectation maximization algorithm was applied. An elbow graph and silhouette score was plotted over a range of clusters, to determine the best k for the k-means algorithm. For brevity, neither graph is included in the report, but can be reproduced with the code. For expectation maximization, the BIC score was plotted over a range of clusters and Gaussian Mixture Models to determine the best k. For brevity, the graph is not included in the report but can be reproduced with the code.

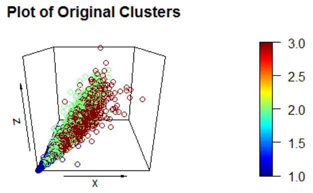
Wisconsin Breast Cancer Data: A plot of the elbow graph and the silhouette scores identified the optimal number of clusters, for the k-means algorithm, as 3. A plot of those clusters can be seen below. A review of the clusters also show that a majority of the data fits into an ellipsoid. Because k-means does not take into account variance in the data, but focuses on distance from a centroid (effectively creating spherical clusters), k-means splits the ellipsoid into two clusters which closely aligns with the sample classes. What is more interesting is that k-means was also able to capture the injected noise data in a separate cluster. When the noise is removed from the data, the k-means algorithm scores a 90% accuracy when compared to the original sample classes. On the other hand, the expectation maximization algorithm took into account variance and was able to use an ellipsoid to cluster the data into one cluster. Although the k-means algorithm was able to correctly identify the ground truth classes of the data, the expectation maximization was able to identify and capture the natural shape of the data properly. A graph of the expectation maximization classification can be seen below. Overall, the k-means algorithm did a better job of capturing the ground truth classes and scored a V-Measure of 0.37, whereas the expectation maximization scored a V-Measure of 0.

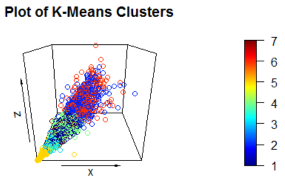


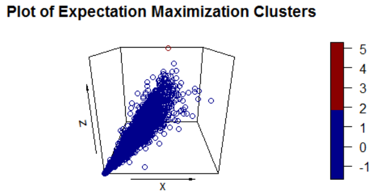




Abalone Data: The k-means algorithm was able to identify 7 clusters within the Abalone Dataset. A plot of those clusters can be seen below. The plot also shows that the data covers an ellipsoidal area. Because the algorithm does not take into account variance it was not able to capture the true shape of the data. Instead it separated the ellipsoid into 7 clusters. The expectation maximization algorithm was able to capture the natural shape of the data and identified one cluster for the ellipsoidal cluster. Overall the k-means algorithm was able to align better with the ground truth classes and scored a V-Measure of 0.17. The expectation maximization algorithm did a better job of capturing the natural shape of the data, but did a poor job of aligning with the ground truth classes – thus scoring a V-Measure of 0.



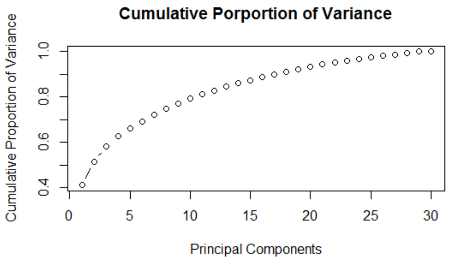


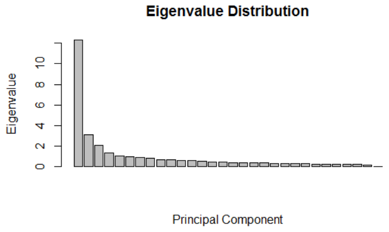


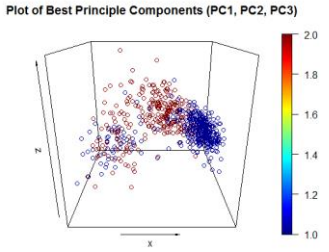
Dimensionality Reduction

Wisconsin Breast Cancer Data: Principle component analysis, independent component analysis, randomized projections, and recursive feature selection were applied to the Wisconsin Breast Cancer Dataset to effectively reduce the data's feature set. You can find a summary of results for each algorithm below.

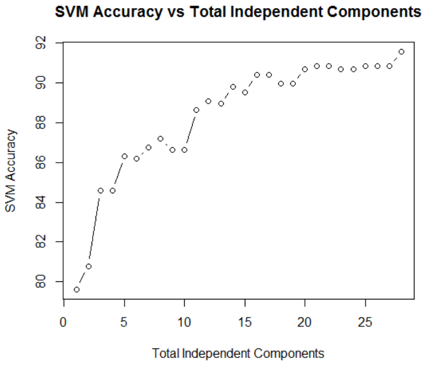
Principal component analysis converted the Wisconsin Breast Cancer Dataset into 30 orthogonal components. The first component accounts for over 40% of the variance in the data, which is nearly 4 times more than any other component. This is because the first component aligns with the major axis of the data's natural ellipsoidal shape, capturing majority of the data's variance. To see the variance explained by each component and their respective eigenvalue, view the "Cumulative Portion of Variance" and "Eigenvalue Distribution" charts below. Graphing the ground truth classes of the dataset versus the first three principal components shows roughly three clusters – a dense cluster of benign tumors, another dense cluster of malignant tumors, and a sparse cluster of mixed tumors (essentially a cluster of noise). The visual is consistent with the results from the k-means clustering algorithm in the previous section. You can view the visual below in the chart labeled "Plot of Best Principle Components (PC1, PC2, PC3)". Using 80% of "variance explained" as the cutoff, the first 11 principal components will be used going forward.

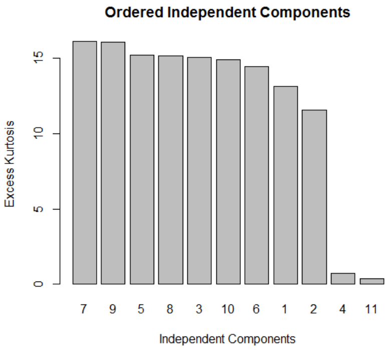

Cumulative Porportion of Variance

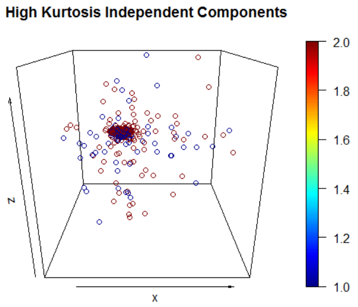

Eigenvalue Distribution

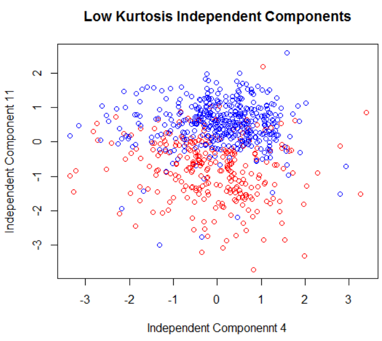

Plot of Best Principle Components (PC1, PC2, PC3)

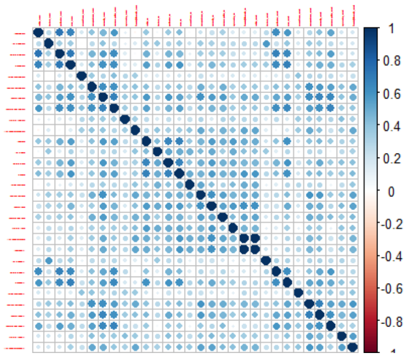
Independent component analysis was used to project the Wisconsin Breast Cancer Dataset onto non-Gaussian, statistically independent components. To choose an optimal value of independent components, a range of components were used to train and test an SVM model. A review of the "SVM Accuracy vs. Total Independent Components" chart below will show that after 11 components, each additional component only adds a marginal amount of accuracy – making 11 the ideal number of independent components. The kurtosis over those 11 components reveal that 9 of the components have a high kurtosis and the other 2 have a low kurtosis. When doing independent component analysis it is preferred to have components with high kurtosis. Despite this fact, a plot of the high kurtosis and low kurtosis components versus the ground truth classes show that the low kurtosis components do a better job of projecting the data. This indicates that the data failed to meet the key assumption that the true sources be independent of one another. You can view the plot of the high and low kurtosis components versus the ground truth classes below. The feature correlation plot below also supports the notion that the features are not independent – showing high correlations between features. Going forward the 9 high kurtosis independent components will be used in the analysis, and the 2 low kurtosis independent components will be used as a means of contrast.


SVM Accuracy vs Total Independent Components

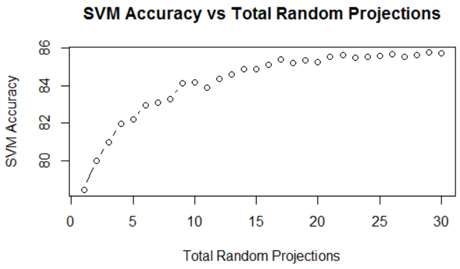

Ordered Independent Components

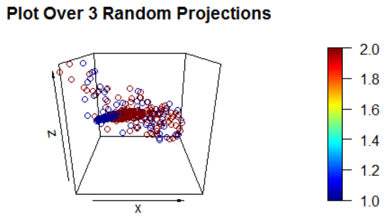

High Kurtosis Independent Components



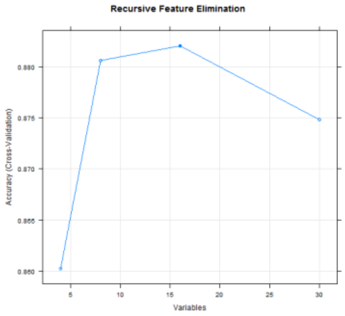


Random projections was used to project the Wisconsin Breast Cancer Dataset along random components. To find an optimal number of random projections, a range of random projections was used to train and test a SVM model. It turns out that at 9 random projections the SVM gains marginal accuracy for each additional projection, making 9 the optimal number of projections. You can view the results below with the "SVM Accuracy vs Total Random Projections". A visual of the ground truth classes versus random components show that the random projections maintained the ellipsoidal structure of the data. Although the projections are produced at random, the algorithm maintains the pairwise distance between any two samples. Over several random projections, the data was displaced but the structural integrity and interrelationship between samples was never compromised.



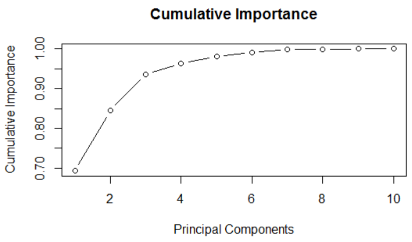


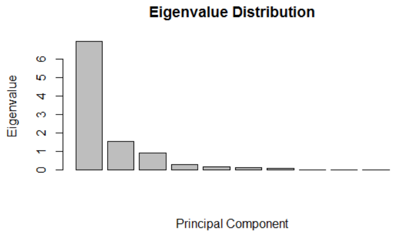
Recursive Feature Elimination was applied to the Wisconsin Breast Cancer Dataset to reduce the feature set by selecting only relevant features. This approach is unlike the other dimensionality reduction algorithms that project data into a new feature space to reduce the feature set. The algorithm tested several subsets, and found that only 16 of the 30 features were relevant in classifying the data. You can view the results of the different subsets tested in the "Recursive Feature Elimination" Chart.

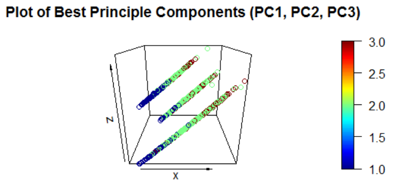


Abalone Data: Principle component analysis, independent component analysis, randomized projections, and recursive feature selection was also applied to the Abalone Dataset. You can find a summary of results for each algorithm below.

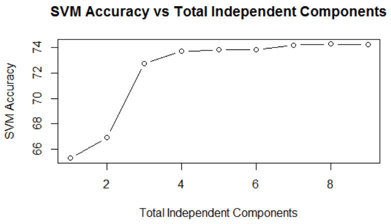
Principal component analysis was used to project the Abalone Data onto 10 orthogonal components. Analysis of the cumulative proportion of variance and the distribution of eigenvalues show that two components account for more than 80% of variance in the data. These two components will be chosen as the focus of analysis going forward. You can view the relevant plots below. A plot of the true classes versus principal components show three ellipsoidal – indicating that the algorithm captured the variance of the ellipsoidal's major axis and the variance between the 3 dummy variables created by hot encoding the "sex" variable. You can view these plots below.

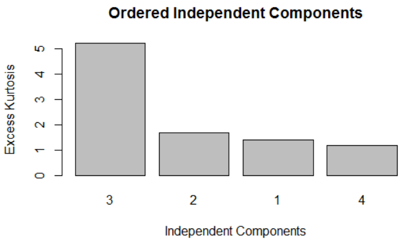


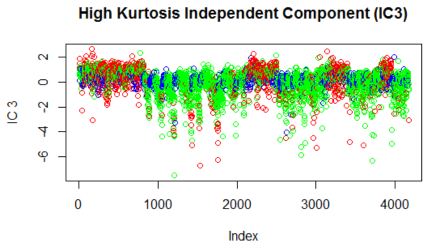


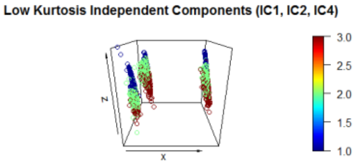


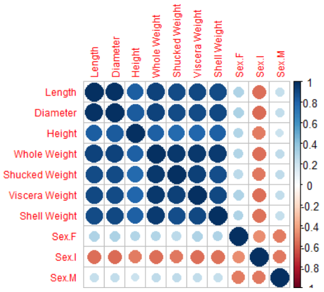
Independent component analysis was used to project the Abalone Dataset onto non-Gaussian, statistically independent components. To pick the optimal number of independent components, the SVM model was trained and tested over a wide range of principal components. A plot of the SVM's accuracy shows that the optimal number of independent components is 4. The plot can be seen below in the chart labeled "SVM Accuracy vs Total Independent Components". It turns out that only 1 component has a high kurtosis, and the other three have low kurtosis. The low kurtosis components do a better job of displaying the data than the high kurtosis component, indicating that the features in the dataset are not statistically independent. To confirm, the correlation between features was plotted. Plots of the low kurtosis components, high kurtosis components, and feature correlations can be seen below. The most interesting observation from the analysis is the fact that independent component 1 was able to capture the three dummy variables that were created as a result of hot encoding the "sex" variable (Male, Female, and Infant). You can see the plot below.

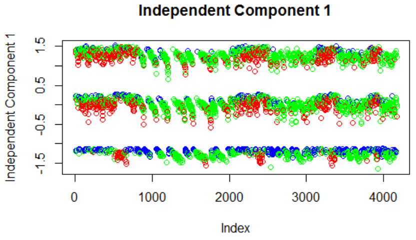




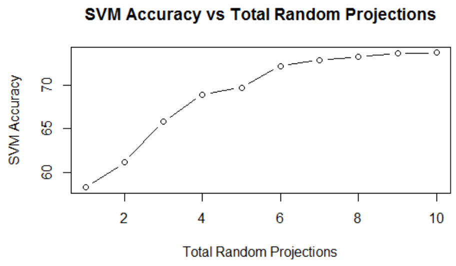


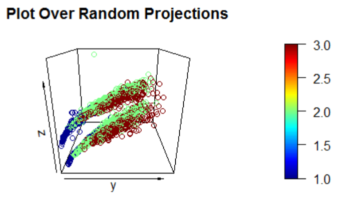




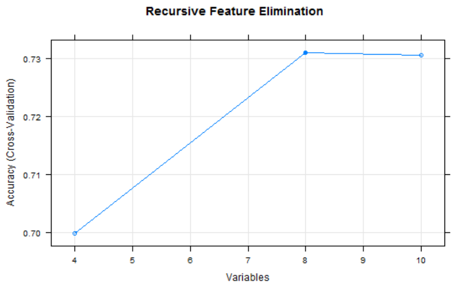


Randomized projections was used to project the Abalone Dataset onto random components. To find an optimal number of random projections, a range of projections was used to train and test a SVM model. A plot of the SVM's test accuracy identifies 6 as the optimal number of projections. You can see the plot below. Similar to the randomized projections used in the Breast Cancer Dataset, the algorithm was able to maintain the ellipsoidal structure of the data. Despite the number of trials conducted, the structure is maintained over the 6 randomized projections.





Recursive Feature Elimination was applied to the Abalone Dataset to reduce the feature set by selecting only relevant features. The algorithm tested several subsets, and found that only 8 features were relevant in classifying an abalone by ring size. You can view the results below.



Clustering and Dimensionality Reduction

Wisconsin Breast Cancer Data: There are a few insights that can be gleamed when the k-means algorithm is used in combination with different dimensionality reduction algorithms and applied to the Wisconsin Breast Cancer Dataset. First, the dimensionality reduction algorithms, in many cases, improved the speed of the k-means algorithm. Reducing the number of features, reduces the amount of data the k-means algorithm has to process. Second, all algorithms, with the exception of independent component analysis, improved k-means performance. Removing unimportant or irrelevant features improves k-means ability to process relevant information as it analyzes the feature space. Third, the 9 high kurtosis independent components, drastically underperformed during k-means. Further proof that the features are not independent of one another. In contrast, the 2 low kurtosis independent components had a v-measure score that was more comparable to the baseline dataset. Fourth, the recursive feature elimination algorithm performed far better than the other dimensionality reduction algorithms. The 14 features that the recursive feature elimination algorithm removed, contained irrelevant or redundant information. By removing those fields, the dataset only contained features that were only relevant for clustering the data. Whereas the other dimensionality reduction algorithms used all features (both relevant and irrelevant) to project into a lower dimensions. Below you can find a summary of results.

K-Means					
Algorithm	**Number of Clusters**	**Homogeneity**	**Completeness**	**V-Measure**	**Time**
Full/Baseline Dataset (30 Components)	3	0.29	0.22	0.25	0.01 sec
PCA (11 Components)	3	0.46	0.31	0.37	0.009 sec
ICA (9 High Kurtosis Components)	6	0.02	0.02	0.02	0.01 sec
ICA (2 Low Kurtosis Components)	3	0.30	0.20	0.24	0.009 sec
RP (9 Projections)	3	0.28	0.22	0.24	0.007 sec
RFE (16 Features)	3	0.52	0.35	0.42	0.02 sec

Insights were also gleamed from using the expectation maximization algorithm in combo with the different dimensionality reduction algorithms and applying it to the Wisconsin Breast Cancer Dataset. First, the application of a dimensionality reduction algorithm improved the performance of the expectation maximization algorithm. In all cases, the expectation maximization algorithm was able to better identify the structure of the data and develop several clusters (as opposed to the one cluster created in the baseline data). Second, the low kurtosis independent components, with 80% less data, outperformed the high kurtosis independent components, reconfirming that the features are not independent. Third, the expectation maximization algorithm performed best when recursive feature elimination was applied to the data. The recursive feature elimination algorithm was able to just focus on relevant features, whereas the other algorithms included irrelevant information in their projections. Below you can find a summary of results.

Expectation Maximization					
Algorithm	**Number of Clusters**	**Homogeneity**	**Completeness**	**V-Measure**	**Time**
Full/Baseline Dataset (30 Components)	1	0	1	0	4.65 sec
PCA (11 Components)	5	0.43	0.18	0.25	6.5 sec
ICA (9 High Kurtosis Components)	4	0.40	0.20	0.27	2.59 sec
ICA (2 Low Kurtosis Components)	2	0.30	0.30	0.30	2.11 sec
RP (9 Projections)	5	0.50	0.21	0.30	4.14 sec
RFE (16 Features)	5	0.53	0.23	0.32	6.48 sec

Abalone Dataset: There are a few insights that can be gleamed when the k-means algorithm is used in combination with the different dimensionality reduction algorithms and applied to the Abalone Dataset. First, the removal of irrelevant information, redundant information, or noise by a dimension reduction algorithm improves the clustering performance of the k-means algorithm. Second, the application of a dimension reduction algorithm allowed k-means to find a different, but effective, set of clusters. By focusing on features or components that are most relevant, allows the k-means algorithm to analyze the data from a different perspective. Third, the low kurtosis components out-performed the high kurtosis components, reconfirming the lack of independence across features. You can view a table of the results below.

K-Means					
Algorithm	**Number of Clusters**	**Homogeneity**	**Completeness**	**V-Measure**	**Time**
Full/Baseline Dataset (10 Components)	7	0.24	0.12	0.16	0.04 sec
PCA (2 Components)	5	0.23	0.14	0.18	0.02 sec
ICA (1 High Kurtosis Components)	4	0.06	0.05	0.06	0.02 sec
ICA (3 Low Kurtosis Components)	9	0.29	0.14	0.19	0.02 sec
RP (6 Projections)	6	0.22	0.12	0.16	0.02 sec
RFE (8 Features)	6	0.26	0.14	0.18	0.05 sec

The application of the expectation maximization algorithm in combo with the different dimensionality reduction algorithms on the Abalone Dataset reconfirmed many of the insights that were uncovered when the k-means algorithm was applied. First, the expectation maximization improved its v-measure score after applying dimensionality reduction. Second, the removal of irrelevant information in the data allowed the expectation maximization to find the true structure of the relevant data – resulting in different clusters. Third, the low kurtosis components performed better than the high kurtosis because the data does not satisfy the assumption of independence. You can view the table of results below.

Expectation Maximization					
Algorithm	**Number of Clusters**	**Homogeneity**	**Completeness**	**V-Measure**	**Time**
Full/Baseline Dataset (10 Components)	1	0	1	0	4.65 sec
PCA (2 Components)	7	0.21	0.13	0.16	1.00 sec
ICA (1 High Kurtosis Components)	3	0.02	0.03	0.03	0.50 sec
ICA (3 Low Kurtosis Components)	7	0.20	0.11	0.14	2.89 sec
RP (6 Projections)	7	0.23	0.12	0.16	3.37 sec
RFE (8 Features)	5	0.18	0.12	0.15	3.46 sec

Dimensionality Reduction and ANN

How does dimensionality reduction affect supervised learning? To gain insight into the question dimensionality reduction was applied to the Wisconsin Breast Cancer Dataset and then the data used to train and test an artificial neural network. Reducing the dimensions of the data improved the timing of the artificial neural network by at least 66%. What is even more surprising is that the reduced columns produced similar results compared to the “full dataset”. Independent Component Analysis performed the worse, because of the data’s inability to meet the necessary assumption of independent sources. Recursive feature elimination performed the best because the algorithm’s ability to detect and disregard irrelevant features, whereas the other algorithms leveraged all features (relevant and irrelevant features) to project the data into a lower dimension. You can see the results below.

Algorithm	Time	Accuracy
Full/Baseline Dataset (30 Components)	92.52 secs	87.80 %
PCA (11 Components)	13.89 secs	86.83 %
ICA (9 High Kurtosis Components)	21.11 secs	74.15 %
ICA (2 Low Kurtosis Components)	23.55 secs	76.59 %
RP (9 Projections)	27.78 secs	84.88 %
RFE (16 Features)	16.67 secs	87.88 %

Clustering and ANN

How does adding the clustering results to the dataset affect classification? To find out, k-means and expectation maximization was applied to the Wisconsin Breast Cancer Dataset, the results added to the data as an additional feature, and then the data was used to train and test an artificial neural network. Adding the clustering results from k-means and expectation maximization improved the artificial neural network’s training time by approximately 40% and 96% respectively. The addition of the clustering information to the dataset had little effect on the accuracy of the artificial neural network; the results were comparable to that of the baseline dataset.

Algorithm	Time	Accuracy
Full/Baseline Dataset (30 Features)	95.58 secs	87.80 %
Full/Baseline Dataset + K-Means Clusters (31 Features)	38.79 secs	84.31%
Full/Baseline Dataset + EM Clusters (31 Features)	2.36 secs	85.85%

How does using clustering as a form of dimensionality reduction algorithm affect classification algorithms? To address the question, k-means and expectation maximization was applied to the Wisconsin Breast Cancer Dataset, and the results was used to train and test an artificial neural network. It turns out that the results from the classification algorithm (in this case the artificial neural network) is only as good as the homogeneity of the clusters you feed it. Clusters with a high homogeneity, can simply be assigned a class with a high level of accuracy. The K-means clusters produced clusters that aligned closely to the ground truth classes, resulting in the artificial neural network to score a high accuracy. On the other hand, the expectation maximization algorithm placed the data in one cluster, resulting in poor accuracy.

Algorithm	Homogeneity Score	Time	Accuracy
Full/Baseline Dataset (30 Features)	NA	95.58 secs	87.80 %
K-Means Clusters Only (1 Features)	0.45	1.05 secs	84.39%
EM Clusters Only (1 Features)	0	2.59 secs	60.00 %

Summaries

K-Means: The k-means algorithm is relatively fast at finding clusters when compared to the expectation maximization algorithm. Its speed is partly due to the fact that the algorithm only considers the distance and not the variance of the data points. This also comes at a cost. Because it does not take into account the variance of the data, it is less effective at capturing the true structure of more complex datasets. It should also be noted that, performing dimensionality reduction before performing k-means improves the algorithm's execution time and performance. To improve k-means performance directly, one could experiment with different measures of distance.

Expectation Maximization: Expectation maximization takes into account variance of the data and therefore does a great job of determining the true structure of the data. This comes at the cost of execution time. It should also be noted that, performing dimensionality reduction before performing k-means improves the algorithm's execution time and performance. To directly improve the algorithm's performance, one could consider different Gaussian models (i.e. diagonal multivariate mixtures with spherical equal volumes).

Principal Component Analysis: Principal component analysis maps data to orthogonal uncorrelated components with respect to variance in the data. The higher the correlation between variance and relevant information within the data, the better principal component analysis is at capturing that relevant information.

Independent Component Analysis: Independent component analysis maps the data to statistically independent, non-Gaussian components. The algorithm assumes that the data is formed from independent, non-Gaussian sources. If these assumptions are not met, as in the Wisconsin Breast Cancer Dataset and the Abalone Dataset, then the algorithm performs less than ideal.

Random Projections: Random projections maps the data to random components. Because it is not analytically intensive to produce random components, the algorithm tends to be faster than most dimensionality reduction algorithms. Despite the randomness of the components, the true structure of the data tends to hold up very well.

Recursive Feature Elimination: Recursive feature elimination identifies irrelevant, redundant, or noisy features and eliminates them. When noise or irrelevant data is contained within select features, recursive feature elimination has a larger impact on clustering algorithms. On the other hand, if all features are relevant the algorithm has a hard time reducing the feature set.