Comparative Analysis of Unsupervised Learning and Dimensionality Reduction on 2 Problem Domains

1st Zixin Feng
Georgia Institute of Technology
Atlanta, USA
zfeng305@gatech.edu

***Abstract*—In this paper, I utilized two unsupervised learning clustering algorithms—K-means and Expectation Maximization (EM)—and three linear dimensionality reduction algorithms—Randomized Projections (RP), Principal Component Analysis (PCA), and Independent Component Analysis (ICA)—on two datasets: the Iris Dataset and the Hospital Dataset. The study includes a comparative analysis of these five algorithms, highlighting their respective pros and cons, attributes, and features based on test results.**

I. DATASET RECAP

A. Iris Dataset

The Iris dataset includes measurements of properties like sepal length, sepal width from samples of three species of Iris flowers. It's a balanced dataset with an equal number of samples from each species. One species is distinctly different from the other two in these measurements. The species are labeled as 0, 1, and 2 for analysis. All these measurements are important as they have a strong relationship (correlation above 0.4) with the species label. However, the dataset is relatively small, containing only 120 data points.

B. Hospital Dataset – Patient Treatment

The Hospital dataset contains medical records from a private hospital in Indonesia, including important metrics like hematocrit and hemoglobin levels. Patients are divided into two groups: those currently receiving treatment and those who have finished treatment. Unlike the Iris dataset, this dataset is larger, with over 4000 records. However, only two metrics in the dataset have a strong relationship (correlation above 0.2) with whether a patient is receiving or has completed treatment. Additionally, the dataset is heavily imbalanced, which is common in medical datasets, with more records of patients who are not currently receiving treatment compared to those who are.

II. MAJOR HYPOTHESIS

- K-means is a simpler and faster algorithm that handles large datasets very well, while Expectation Maximization (EM) is computationally expensive due to the need to update a series of hidden variables in a probability matrix.
- K-means handles outliers in a more robust manner, whereas Expectation Maximization is very sensitive to outliers.
- Expectation Maximization is highly dependent on initialization and can be costly if restarts are needed.
- The biggest advantage of Randomized Projections (RP) is its fast and simple nature, effectively handling simple problems with balanced data points such as the Iris dataset.
- Including clustering features in a neural network learner might not necessarily lead to better prediction results for a linearly separable dataset like the Iris dataset. This can be caused by the addition of extra dimensions or overfitting.

III. CLUSTERING ALGORITHMS

A. Introduction to Algorithms and Evaluation Metrics

K-means is an unsupervised machine learning algorithm that focuses on following major steps:

- Pick K centers at random.
- Assign each center's closest points.
- Recompute the center by averaging.
- Repeat until convergence.

Expectation Maximization is an unsupervised machine learning algorithm that focuses on following major steps:

- initializing cluster centers and their shapes.
- Expectation: Calculate the probability that each point belongs to each cluster based on its current positions. Probability is repreented by hidden variables.
- Maximization: Update cluster centers and their shapes
- Repeat until convergence or reaching maximum iteration.

I applied K-means and Expectation Maximization on Iris and Hospital Dataset in an unsupervised fashion. To evaluate the performance of Iris Dataset, I used 2 metrics:

- Rand: The Rand Index compares cluster as a whole and it "considers all pairs of data points and counts how often they are either in the same cluster or in different clusters in both the true and predicted clusterings".(C.D. Manning, 2008). This metric requires ground truth labels and compare the clustering results to these labels.

- Silhouette Score: Silhouette Score measures how similar a point is to its own cluster compared to other clusters. It evaluates the clustering based on the data itself without requiring ground truth labels.

B. Performance of Algorithms and Analysis

The top right graph depicts the Rand Index of 2 algorithms versus number of clusters. The top right graph depicts the time of 2 algorithm versus number of clusters. The bottom right graph depicts the Silhouette Score of 2 algorithms versus number of clusters and the bottom left graph is the learning curve which depicts Rand Index versus count of data points.

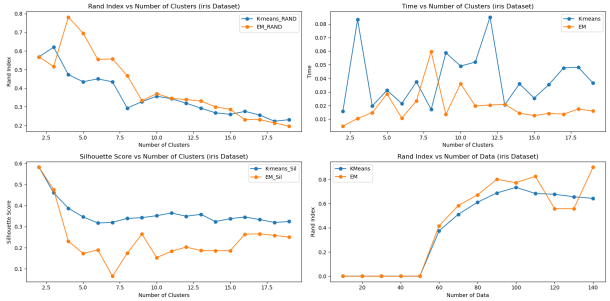

Fig. 1. Clustering Performance on Iris Dataset

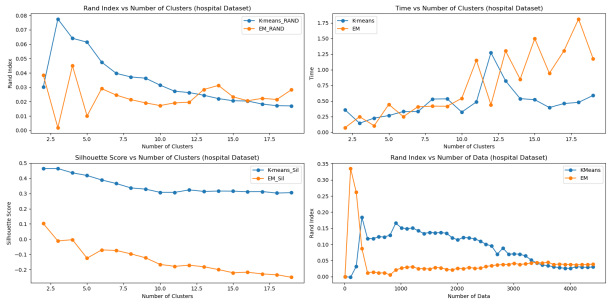

Fig. 2. Clustering Performance on Hospital Dataset

There are many interesting findings:

- In Iris dataset, K-means achieves the best Rand Index when the number of cluster is 3 and Expectation Maximization performs the best when the number of cluster is 4. This is almost consistent with the dataset description – there are 3 major species in the Iris dataset.
- In both datasets, when cluster size increases, the Rand Index gets lower. A lower Rand Index when cluster size increases doesn't necessarily indicate poorer clustering quality outright. It reflects the challenge of achieving agreement in cluster assignments across larger, potentially more dispersed clusters. When cluster size is super large, many points might just partition by chance and therefore ends up with a higher error.

  As cluster size increases, the Silhouette score curve tends to remain relatively flat and does not decrease significantly. This stability can be attributed to clustering algorithms effectively maintaining reasonable inter-cluster distances. This characteristic ensures that clusters remain internally cohesive and sufficiently separated from each other, thus supporting stable Silhouette scores across varying cluster sizes.
- Expectation Maximization outperforms K-means in Iris dataset as obviously shown in the Rand Index Score graph when cluster size is less than 10. This is related to the fact that Iris dataset is very balanced. There is a larger chance for a good initialization to be made and data points following Gaussian distribution.
- K-means outperforms Expectation Maximization in the Hospital Dataset when cluster size is less than 12. This is related to the fact that Hospital Dataset is super large, unbalanced, and super noisy and Expectation Maximization can't handle dataset like that very well.
- In the Hospital Dataset, K-means demonstrates strong performance as indicated by a high Silhouette Score but significantly poorer performance according to the Rand Index. This discrepancy may stem from many dataset features having low correlation with the target variable. While these features appear as noise when considering true labels, they are treated equally alongside highly correlated features in unsupervised clustering. This approach can add value by capturing patterns that align with the data's inherent structure rather than its predictive relationship to a specific outcome.
- In the learning curve of the Iris Dataset, both algorithms begin to converge when the dataset size reaches 50, with slight improvement as the dataset grows larger, eventually leveling off. However, in the Hospital Dataset's learning curve, the Expectation Maximization algorithm shows a sudden increase near a dataset size of 100, followed by a dramatic decline. This pattern is attributed to the dataset's severe imbalance and high variance. Initially, the top 100 data points primarily comprise sick cases, facilitating straightforward partitioning. As the dataset size increases, more healthy cases and outliers are introduced, making it progressively challenging to achieve effective clustering. Additionally, the jump in performance at a dataset size of 100 might be caused by an exceptionally good initialization, followed by a return to a less effective base initialization in subsequent iterations.
- K-means is significantly faster than Expectation Maximization, which is consistent with our initial hypothesis.

IV. FEATURE TRANSFORMATION ALGORITHMS

A. Introduction to Algorithms

All three algorithms are used to the feature transformation of unsupervised machine learning.

- PCA: PCA finds orthogonal components that explain the most variance. It's forced to find global features.
- ICA: ICA seeks components that are statistically independent. It's a good edge detector.
- Random Projection: Random Projection randomly projects high-dimensional data into a lower-dimensional space.

B. Performance Evaluation

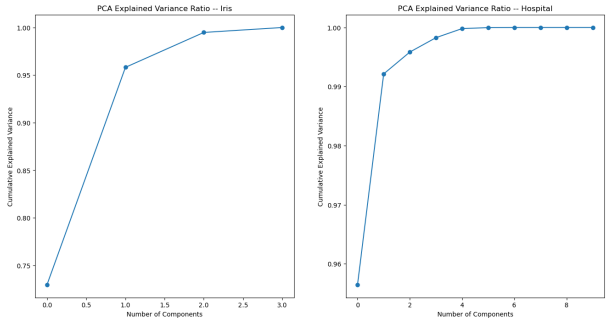

Fig. 3. PCA performance on both datasets

For PCA, as shown in Figure 3, I plotted the cumulative explained variance versus the number of components. The number of components required to retain 95 percent of the variance in the Iris dataset is 2, while for the Hospital dataset it is 1. Both curves exhibit a steep initial increase, indicating that a few principal components effectively summarize the original datasets. The curves then gradually flatten, which is because most of the dataset's variance has been explained by the initial few principal components, and additional components contribute less to the explained variance.

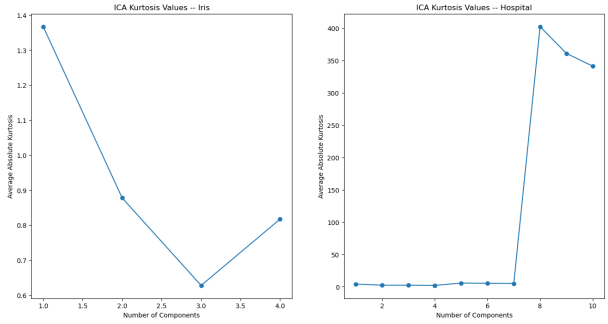

Fig. 4. ICA performance on both datasets

For ICA, as shown in Figure 4, I plotted Kurtosis versus the number of components. High kurtosis values represent more informative components from more independent features.

The Hospital Dataset ICA plot shows an expected pattern, starting with high kurtosis values that gradually decrease. This indicates that the initial components are highly independent and informative. As the number of components increases beyond a certain point, additional components do not add extra value, leading to a decrease in kurtosis. We will select 8 as number of ICA components for Iris Dataset for future studies

The Iris Dataset ICA plot first decreases and then increases. This might be because the initial components extracted have super high kurtosis, explaining the high start. Subsequently, components with lower kurtosis are extracted, leading to the decreasing trend. Finally, the curve increases again as more informative components are identified. We will select 1 as number of ICA components for Iris Dataset for future studies.

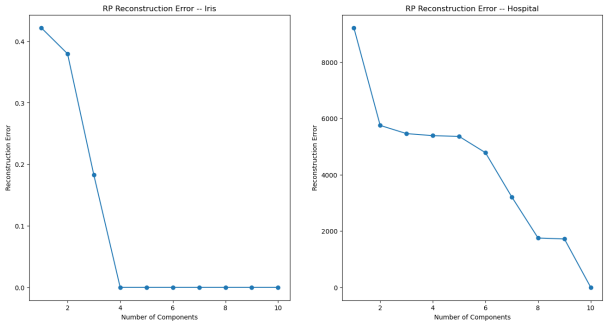

Fig. 5. RP performance on both datasets

For RP, as shown in Figure 5, I plotted the reconstruction error versus the number of components for both datasets. Both curves exhibit a decreasing trend that eventually flattens out. In the Iris dataset, the decreasing trend stops at 4 components, where the error is close to 0. In the Hospital dataset, this trend stops at 10 components, also resulting in an error close to 0. The reconstruction error in the Iris dataset drops much quicker and steeper because Iris Dataset has clear, separable classes, allowing a small number of components to capture most of the variance in the data. We will select 4 as the number of RP components for Iris Dataset and 10 for that of Hospital Dataset in future studies.

V. Apply Clustering on Transformed Dataset

A. Introduction

In this section, I applied two clustering algorithms to six transformed datasets using the three algorithms introduced in the previous section for dimension reduction. Initially, I present the performance of K-means and Expectation Maximization individually on Iris datasets transformed by the three algorithms. Subsequently, I provide a table summarizing the Rand Index values for all combinations.

B. Performance and Analysis

I this section, I will post 6 graphs, each of which compares the performance of K-means and Expectation Maximization in different combinations of datasets and feature transformation methods. The top right graph depicts the Rand Index of 2 algorithms versus number of clusters. The top right graph depicts the time of 2 algorithm versus number of clusters. The bottom right graph depicts the Silhouette Score of 2 algorithms versus number of clusters and the bottom left graph is the learning curve which depicts Rand Index versus count of data points.

There are several interesting findings:

- ICA achieving the highest Rand Index for K-means clustering on the Iris Dataset, with the order of effectiveness being ICA ¿ Random Projection (RP) ¿ PCA, can be explained by the characteristics of each feature transformation method. The Iris Dataset benefits from features that are more independent rather than correlated, which is exactly what ICA extracts. Additionally, the Iris Dataset

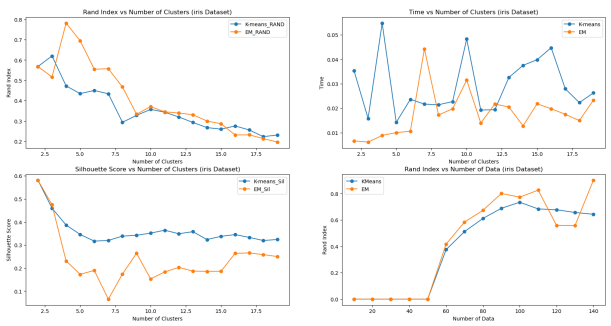

Fig. 6. K-means and Expectation Maximization performance on Iris datasets transformed by PCA

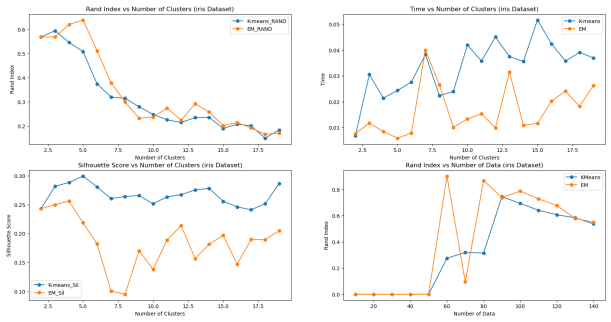

Fig. 7. K-means and Expectation Maximization performance on Iris datasets transformed by ICA

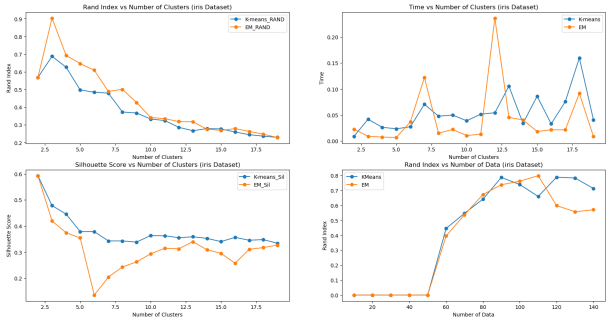

Fig. 8. K-means and Expectation Maximization performance on Iris datasets transformed by RP

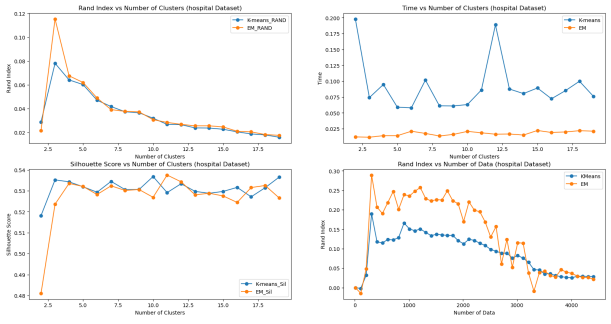

Fig. 9. K-means and Expectation Maximization performance on Hospital datasets transformed by PCA

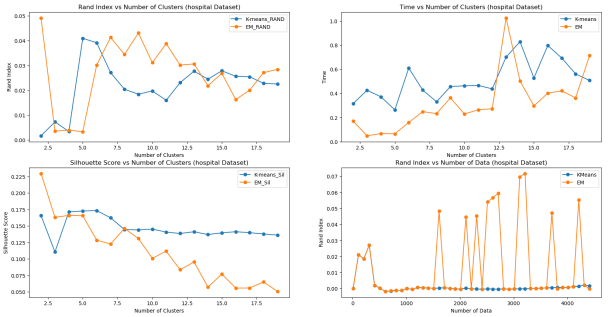

Fig. 10. K-means and Expectation Maximization performance on hospital datasets transformed by ICA

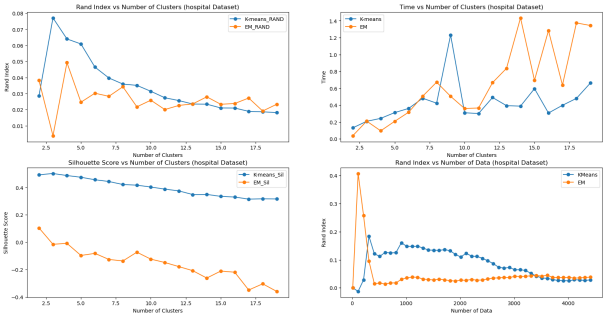

Fig. 11. K-means and Expectation Maximization performance on hospital datasets transformed by RP

Rand Value	Combination
0.620	K-means+Iris+RAW
0.620	K-means+Iris+Pca
0.802	K-means+Iris+Ica
0.744	K-means+Iris+RP
0.780	Expectation Maximization+Iris+RAW
0.598	Expectation Maximization+Iris+Pca
0.757	Expectation Maximization+Iris+Ica
0.904	Expectation Maximization+Iris+RP
0.0774	K-means+Hospital+RAW
0.0780	K-means+Hospital+Pca
0.041	K-means+Hospital+Ica
0.077	K-means+Hospital+RP
0.045	Expectation Maximization+Hospital+RAW
0.115	Expectation Maximization+Hospital+Pca
0.050	Expectation Maximization+Hospital+Ica
0.049	Expectation Maximization+Hospital+RP

TABLE I
COMPARISON OF THE HIGHEST RAND INDEX IN DIFFERENT ALGORITHMS

is low-dimensional. While RP and PCA are effective in reducing dimensionality and capturing variance, they may not optimize for the specific clustering structure as effectively as ICA does in this instance.

- Only RP surpasses the raw data in enhancing Expectation Maximization clustering on the Iris Dataset. RP transforms the data in a manner that better aligns with the Gaussian assumption of the Expectation Maximization algorithm. While PCA and ICA excel in other contexts, their transformations do not notably improve Expectation Maximization clustering performance in this specific case.

- ICA performs poorly for K-means clustering on the Hospital Dataset, and the other two algorithms also do not significantly enhance overall clustering performance. The Hospital Dataset includes many features that are irrelevant for clustering, acting as noise. ICA may amplify these irrelevant features, thereby contributing to the poor clustering performance observed.
- PCA performs exceptionally well for Expectation Maximization clustering on the Hospital Dataset, while the other two algorithms do not significantly enhance the overall clustering performance. This could be attributed to the fact that the few informative features for clustering in the Hospital Dataset also possess the highest variance, thereby being identified as the principal components by PCA.

VI. Performance of NN with Feature Transformation

In this section, I re-run my neural network learner from Assignment 1 on the Iris Dataset with each of the dimension reduction algorithms applied. For ease of comparison, I will use a learning rate of 0.05 and a hidden layer configuration of (5,), which were determined to be optimal parameter values in Assignment 1.

A. Performance

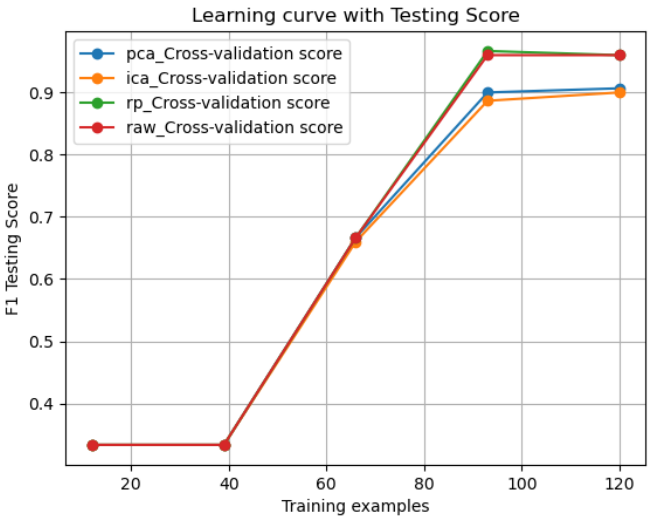

Fig. 12. Learning Curve for Iris Dataset with Testing Score

B. Analysis

There are several interesting findings:

- RP is the fastest among 3 dimension reduction algorithms, which verifies my hypothesis that biggest advantage of RP is its fast and simple nature and it can effectively handle simple problems with balanced data points. Figure 14 and 15 show the above phenomenon. This is because RP randomly picks direction, reduces the dimension of the dataset efficiently, leading to faster training and testing times.

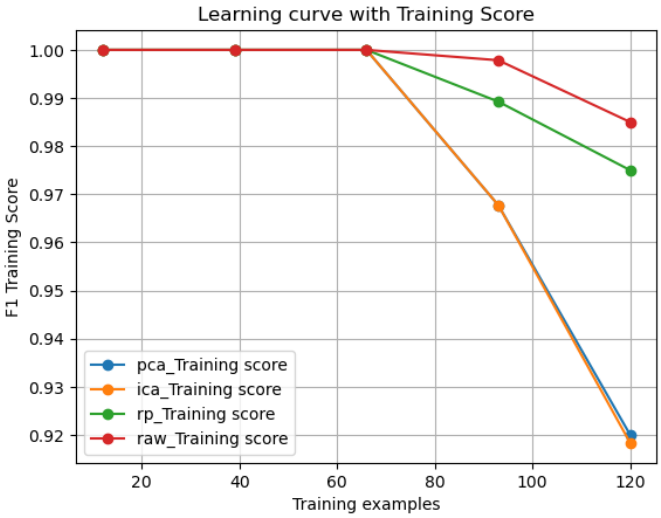

Fig. 13. Learning Curve for Iris Dataset using Train Score

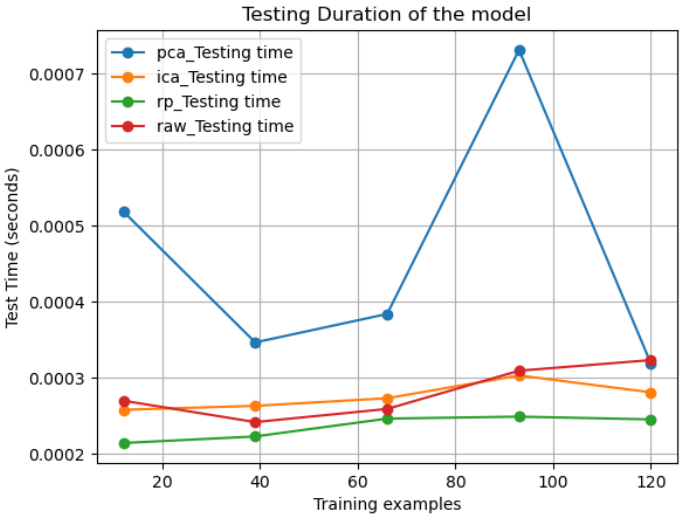

Fig. 14. Test Time for Iris Dataset

- In A1, the final testing result is 0.97. For PCA, it is 0.91; for ICA, it is 0.9; and for RP, it is 0.963. RP preserves pairwise distances between data points to some extent. However, some information may be lost during feature transformation, causing both training and accuracy scores to be lower compared to the A1.
- RP f1 results nearly catch up to A1 as shown in Figure 12. This might be because the Iris dataset is linearly separable, and RP makes data points more separable, simplifying subsequent supervised classification. In contrast, the other two algorithms might slightly disrupt the linear separability through their feature transformations.
- The variance of training and testing scores is very small, as shown in the green dark area of Figure 16. This is

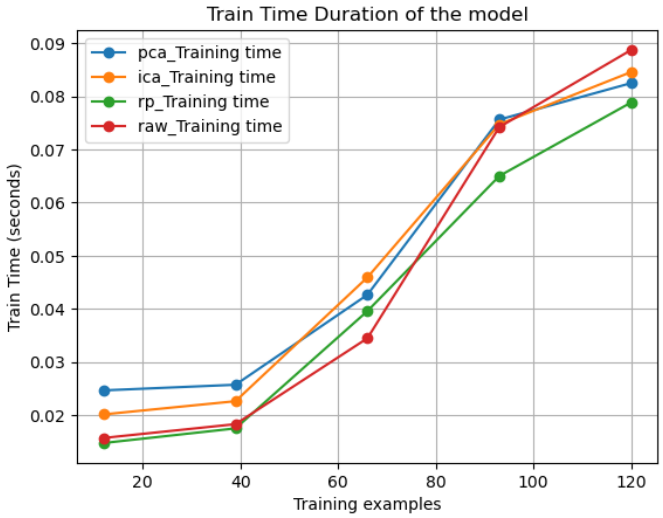

Fig. 15. Train Time for Iris Dataset

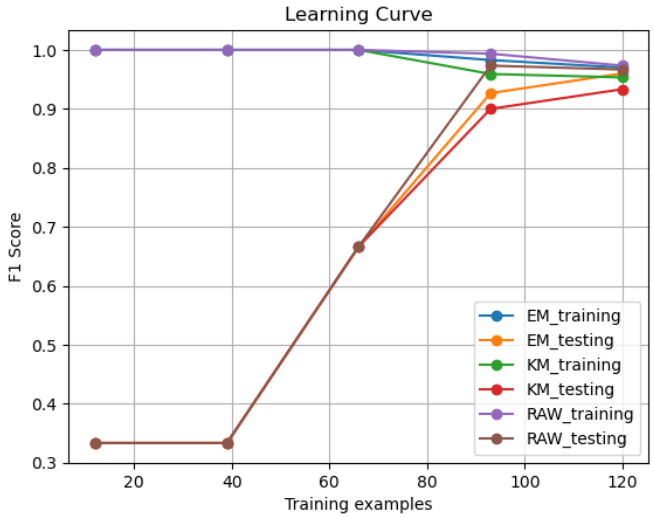

Fig. 17. Learning Curve with Training Score

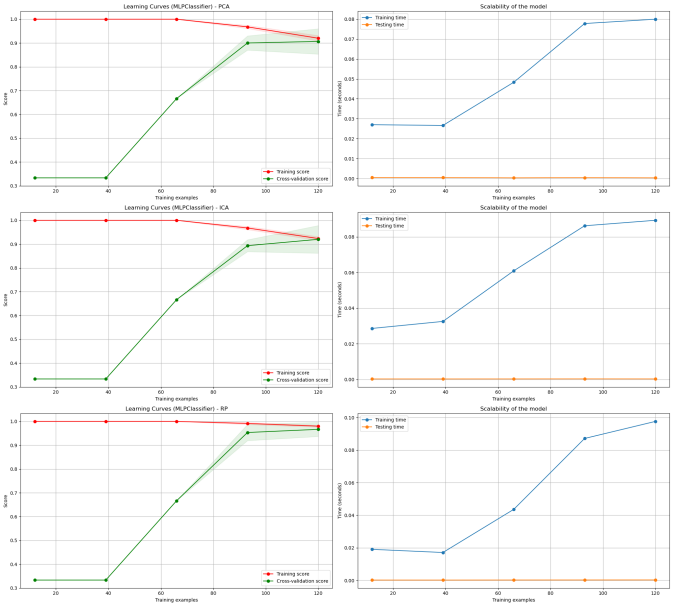

Fig. 16. Plot dimension reduction algorithms separately

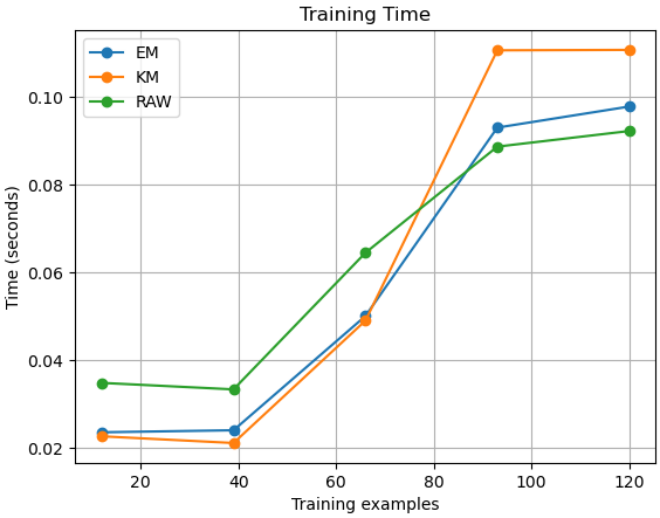

Fig. 18. Training Time

because Iris dataset is highly balanced.

In short summary, RP is the best among 3 dimensionality reduction algorithms because it gives the highest testing score within lowest time.

VII. Apply Clustering on Iris Dataset

In this section, I incorporate the clusters generated in Step 1 as new features in the Iris dataset and rerun my neural network learner on this augmented data. For consistency and comparison, I will use a learning rate of 0.05 and a hidden layer configuration of (5,) from Assignment 1.

A. Performance and Analysis

- As shown in Figure 17 and 18, for both training and testing learning curve, incorporating the clusters as new features in the Iris dataset results in a worse F1 score for my neural network learner. The F1 accuracy is Original learning ¿ K-means ¿ Expectation Maximization. This might because our original Iris dataset is already linearly separable and contains enough strong signal for partition. Clustering features only add duplicated and redundant information. I used to think this might also because of over-fitting brought by extra dimensional. However, our training and testing score both has Original learning ¿ K-means ¿ Expectation Maximization, and there is no significant gap between training and testing score for both clustering methods, over-fitting is less likely.
- The original MLP learner also spent less time in training

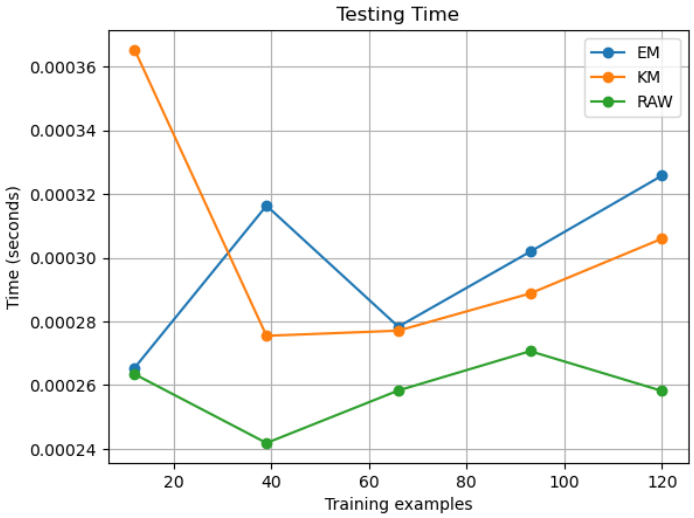

Fig. 19. Testing Time

and testing compared to those with clustering features. This might be related to the extra dimension added, which validates my hypothesis that including clustering feature could complicate the problem without providing useful information to the learner.

- K-means takes significantly longer time in both training and testing compared to the other 2. The K-means clustering introduces non-linear features based on the distance to cluster centroids. These non-linear features might increase the computational complexity of the neural network's operations.
- Is there a way to make K-means and Expectation Maximization faster? We could probably reduce number of clusters. Then training and testing time will be shorter.

VIII. Conclusion

Through this study, we have tested several important hypotheses introduced at the beginning of this paper by conducting a series of experiments.

We learned that K-means is a simpler and faster algorithm that handles large datasets efficiently, whereas Expectation Maximization is computationally expensive due to the need to update a series of hidden variables in a probability matrix. This explains why K-means has a much shorter clock time compared to Expectation Maximization when clustering the hospital dataset.

Additionally, we verified that K-means handles outliers more robustly, whereas Expectation Maximization is very sensitive to outliers. This is why K-means performed much better than Expectation Maximization on the Hospital dataset, while both algorithms performed equally well on the Iris dataset.

We also found that Random Projections is fast and simple, effectively handling simple problems with balanced data points, such as the Iris dataset. However, for more complicated datasets, like the Hospital dataset, RP does not necessarily perform better than the other two feature transformation algorithms (PCA and ICA).

Moreover, we discovered that including clustering features in a neural network learner does not necessarily lead to better prediction results for a linearly separable dataset like the Iris dataset. In the case of the Iris dataset, this is primarily due to the addition of extra dimensions rather than over-fitting.

References

[1] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval, Cambridge University Press, 2008, ch. 16.5.4, p. 451.