Unsupervised Learning Model Comparison

Zhilan Deng

***Abstract*-** This paper first applied k-means clustering and expectation maximization on the two datasets: income prediction and accent prediction, and then applied dimension reduction (PCA, ICA, Randomized Projection and Multidimensional scaling (MDS)) on these two datasets followed by applying clustering algorithms again on datasets after dimension reduction. Lastly, this paper will apply neural network analysis on accent prediction dataset after dimension reduction and after clustering to compare the results with original result from homework1.

PART 1 CLUSTERING

I. Introduction

Briefing of Datasets

The two datasets used in this homework are the same as the datasets used in homework1: the income dataset and accent dataset. The **income dataset** has 18326 data points and 12 variables with a label marking data points with income larger than 50K as 1 and income smaller than 50K as 0. The 12 variables are demographic attributes of each data point such as education, age, occupation, sex and race. Different from homework1 where the variables can be input as category variables, the algorithms in this homework can work with numeric variables only, so this paper altered the original category variable to several numeric binary variables, for example, the sex variable with values of Male and Female is transformed to two variables: is Female and is Male. The accent dataset has 211 data points and 12 variables with a label marking the accents from : US, ES, FR, GE, IT and UK. The input variables are the signals extracted from Mel-Frequency Cepstral Coefficients (MFCCs), all variables are quantitative variables.

Briefing of Datasets

The clustering algorithms applied in this part will be: (1) k-means clustering and (2) expectation maximization.

K-means clustering is an algorithm looking for a fixed number of clusters in the dataset and group similar data points into the same group. The algorithm will stop when the clusters are stabilized, or the maximum number of iterations has been reached.

Expectation maximization considers the clustering problem as an optimization problem and is trying to find the maximum likelihood clustering with the presence of latent variables.

The details of each algorithm will be described in the following section.

II. Clustering Algorithms

K-means Clustering

K-means clustering starts with k random clusters, it first select k random places as the cluster centroid and assign the remaining data points to the closest centroid, then this algorithm compute the new k centroids and reassign the remaining data points to the new centroid. The algorithm repeats these steps until the centroids are stabilized or the max number of iterations ahs been reached.

To start with k-means clustering, we need to first find the best k for each dataset and we will apply elbow method to find the best k, the results are shown in figure 1. The best k value of both datasets are 4, so we will use k = 4 in k-means clustering algorithms for these two datasets.

The results of k-means clustering on both datasets with k = 4 are shown in figure 2. This paper experiments three different distance matrices for k-means clustering and shows the performance of different distance matrices under different maximum number of iterations and number of random starts.

To compare different clustering algorithms, this paper used adjusted accuracy[1] which evaluates the quality of clusters.

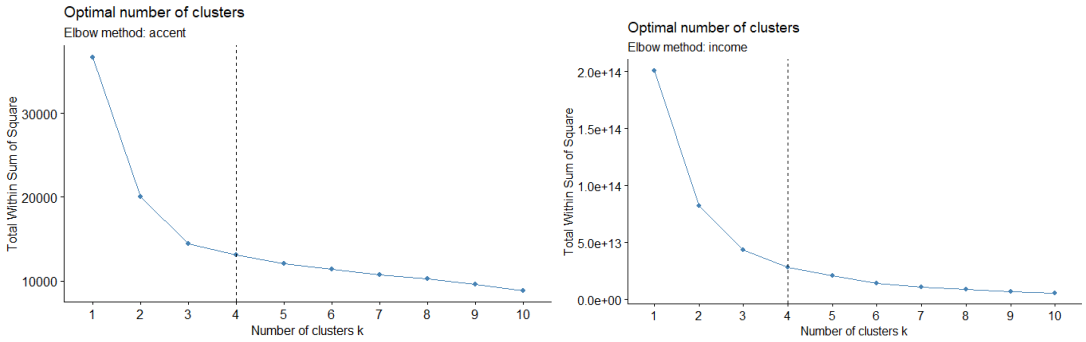
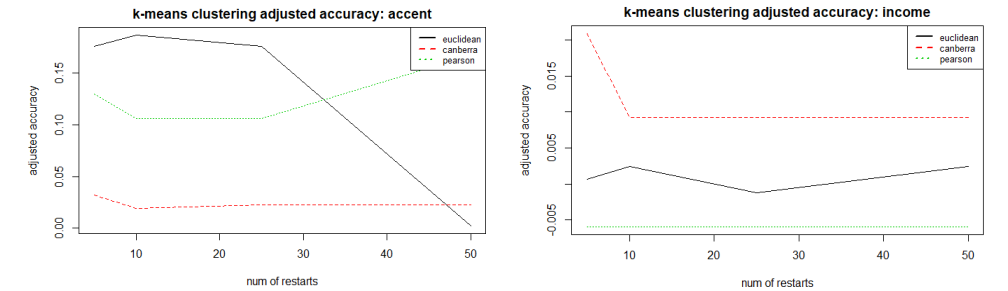
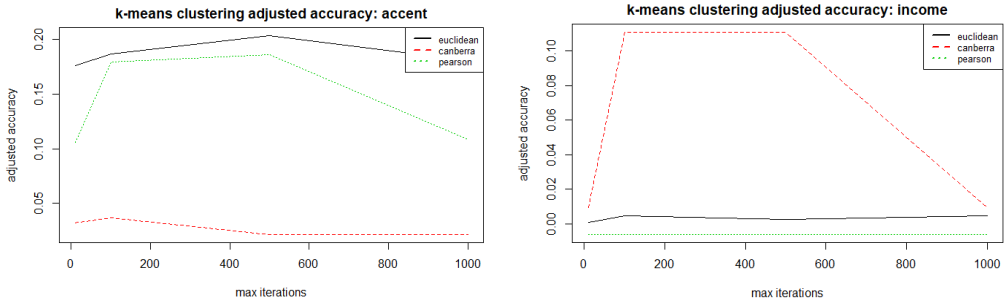


Figure 1. Elbow method for accent dataset and income dataset.



(a)



(b)

Figure 2: k-means cluster performance of different distance method under different num of restarts (a, b) and different maximum number of iterations (c, d).

From figure 2, it is obvious that the best distance algorithm for these two datasets are different, Euclidean distance works the best for accent dataset since it is quantitative and Canberra works the best for income dataset because income dataset is mix of quantitative and categorical variables and these two distance matrices works better in different structures of the datasets.

It is also interesting to see that the performance of k-means clustering is not always increasing while the number of restarts or the number of maximum number of iterations increases. In this experiments, the best number of restarts and best maximum number of iterations for accent dataset is 10 and 500 respectively, and for income dataset is 5 and 100 respectively.

EXPECTATION MAXIMIZATION (EM)

Expectation maximization consists two main steps: (1) estimate the probability of each point belongs to each cluster and (2) re-estimate the parameter vector of the probability distribution of each class.

The result of expectation maximization is shown in figure 3. The EM function automatically tuned the hyperparameters and return the best set of hyperparameters.

EM clusters the accent dataset to 5 clusters which is almost the same with the original 6 labels and successfully clusters income dataset into two classes which is the same with original labels.

The best models used in different dataset is different. The best model for accent clustering is VVE which assumes that the mixture components share the same orientation matrix, and the best model for income clustering if VEV which assumes that the mixture components share the same shape but not the same orientation matrix. The difference is also because of the different structure of the two datasets.

```
Mclust VVE (ellipsoidal, equal orientation) model with 5 components:

Clustering table:
  1   2   3   4   5
 14  57  14 105  21
```

```
Mclust VEV (ellipsoidal, equal shape) model with 2 components:

Clustering table:
    1     2
```

Figure 3. the result of expectation maximization: (a) accent clustering, (b) income clustering

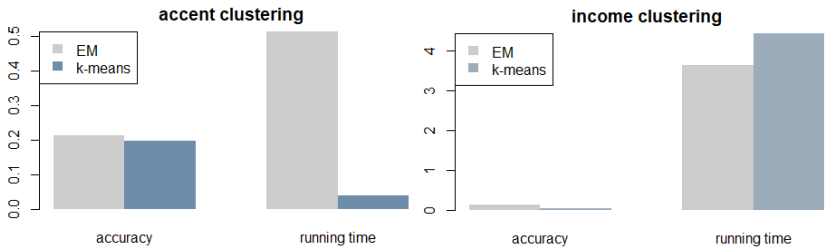
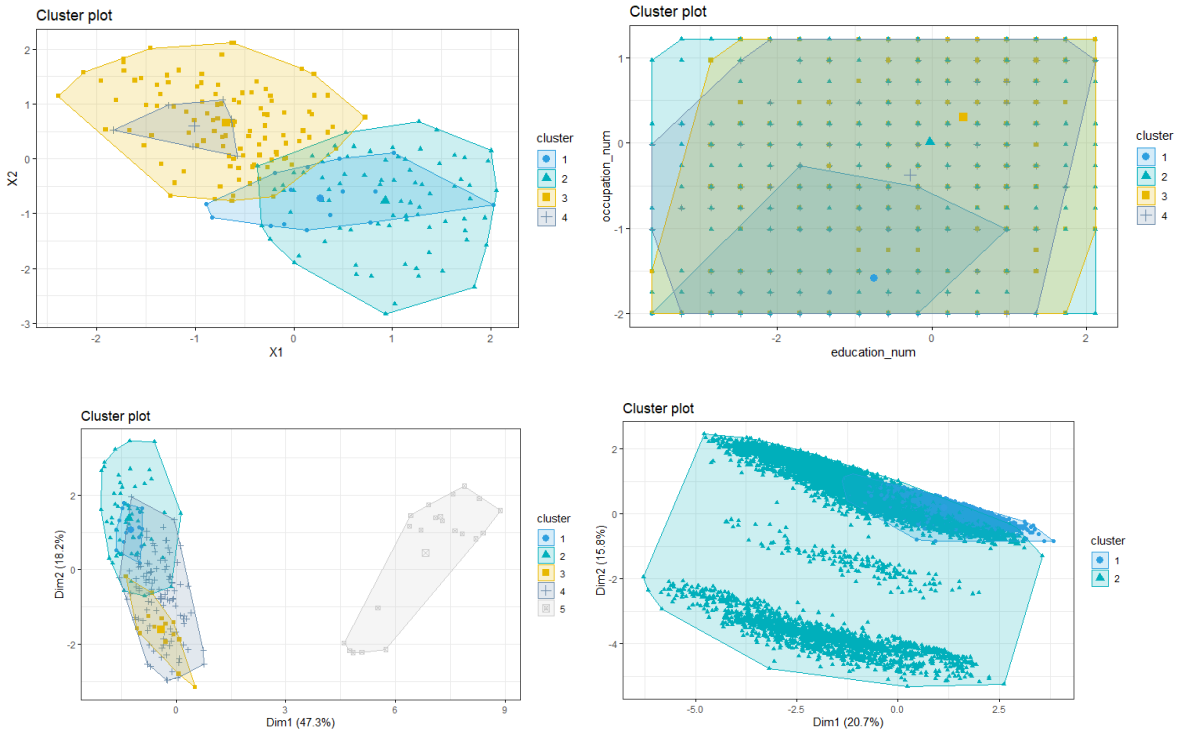


Figure 4. The running time and accuracy of k-means clustering and expectation maximization in two dataset



(b)

Figure 5. The output clusters in (a) k-means clustering and (b) EM.

```
        age      In.gov     fngwgt education_num in.marriage occupation_num  Own.child
1 0.2985160 0.23224044 0.1495905     0.5823770 0.369535519      0.5049391 0.13934426
2 0.3361795 0.00000000 0.1190095     0.6337562 0.652376365      0.6508633 0.03363841
3 0.0574167 0.04950495 0.1220183     0.5476788 0.002640264      0.3387154 0.84554455
4 0.3592003 1.00000000 0.1170340     0.7070406 0.644504749      0.7263682 0.03934871
    iswhite     isAsian   isBlack    isMale hours_perweek
1 0.0000000 0.000000000 0.9829235 0.5409836     0.3966000
2 0.9738279 0.017822736 0.0000000 0.7768144     0.4365079
3 0.9874587 0.001320132 0.0000000 0.4640264     0.2972385
4 0.9710538 0.021257350 0.0000000 0.6978743     0.4213441
```

Figure 6. scaled cluster centroids in income dataset

DISCUSSION

The accuracy and running time of k-means clustering and EM with the best parameters are shown in figure 4. For accent dataset, the accuracy of EM is higher than that of k-means but the running time is also much higher than k-means clustering. For income dataset, the accuracy of EM is also higher than that of k-means clustering, however, the running time of EM is shorter than that of k-means clustering, this might because for EM, we input the best number of clusters which is 2 but in k-means, the best number of clusters is 4 so it took more time in k-means to stabilize the 4 clusters while the EM only needs to find 2.

The output clusters of k-means clustering and EM for two datasets are shown in figure 5. For visualization, this paper only selects two variables to output the clusters. The clusters in accent datasets are more visible compared to the clusters in income datasets, and the clusters produced by EM are more visible compared to that produced by k-means clustering, although both of them will need more variables to separate the data points into different clusters. Since the variables in accent dataset are just signals extracted from MFCC, this is little to explain in accent clusters, so this paper will just focus on the clusters in income dataset. The cluster centroids are shown in figure 6. In the 4 clusters, it is interesting to see that in the 4 clusters, if the data point is White and if the data point is Black is compliment, which means that this two races are in very different clusters. Some other interesting variables with big difference between the 4 clusters are if the data point owns a child or not, if the data point works in government or not, and if the data point in marriage or not.

Compared the two algorithms, for small dataset with quantitative variables like the accent dataset, k-means clustering might be a good choice since the accuracy is almost as good as EM and the running time is much shorter. But for large dataset with mix of quantitative variables and category variables like the income dataset, if the number of clusters is small, EM would be a better choice to have both higher accuracy and shorter running time.

PART 2 DIMENSION REDUCTION

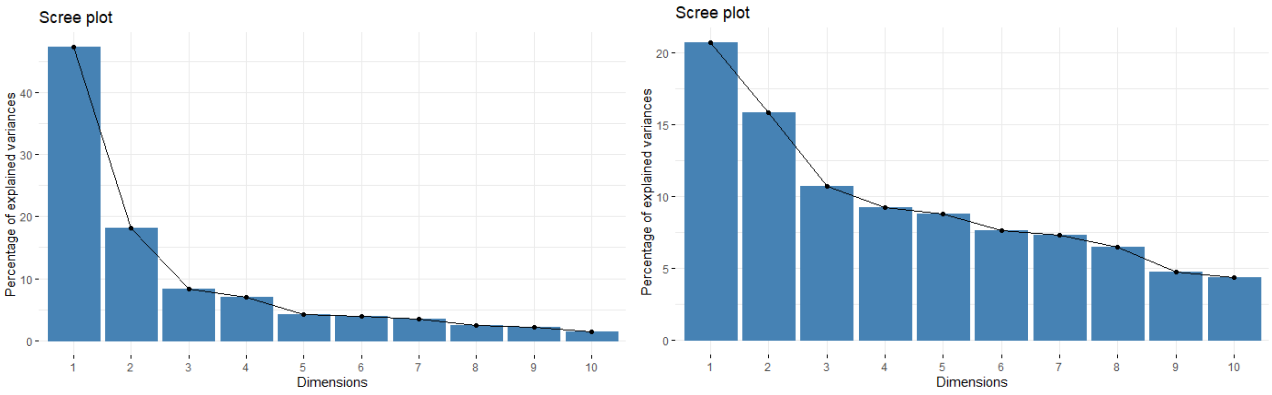
I. INTRODUCTION

This part will apply 4 dimension reduction algorithms on both dataset and rerun clustering algorithms on the datasets after dimension reduction. The 4 dimension reduction algorithms are: (1) PCA, (2) ICA, (3) Randomized Projection, (4) MDS.

II. DIMENSION REDUCTION ALGORITHMS

PCA

PCA looks for the most variance and reconstruct the input variables into principal components which explains the most variance in the data. The components generated by PCA is shown in figure 7. For accent dataset, the first three components have explained almost 80% of the dataset, and the difference between the explained variance in different component is obvious while that is not much obvious in income dataset. For the purpose of dimension reduction, PCA works better for accent dataset which is because the variables in accent datasets are all quantitative while some of the variables in the income dataset are categorical.



(a) (b)

Figure 7. The principle components generated from PCA for (a) accent dataset and (b) income dataset

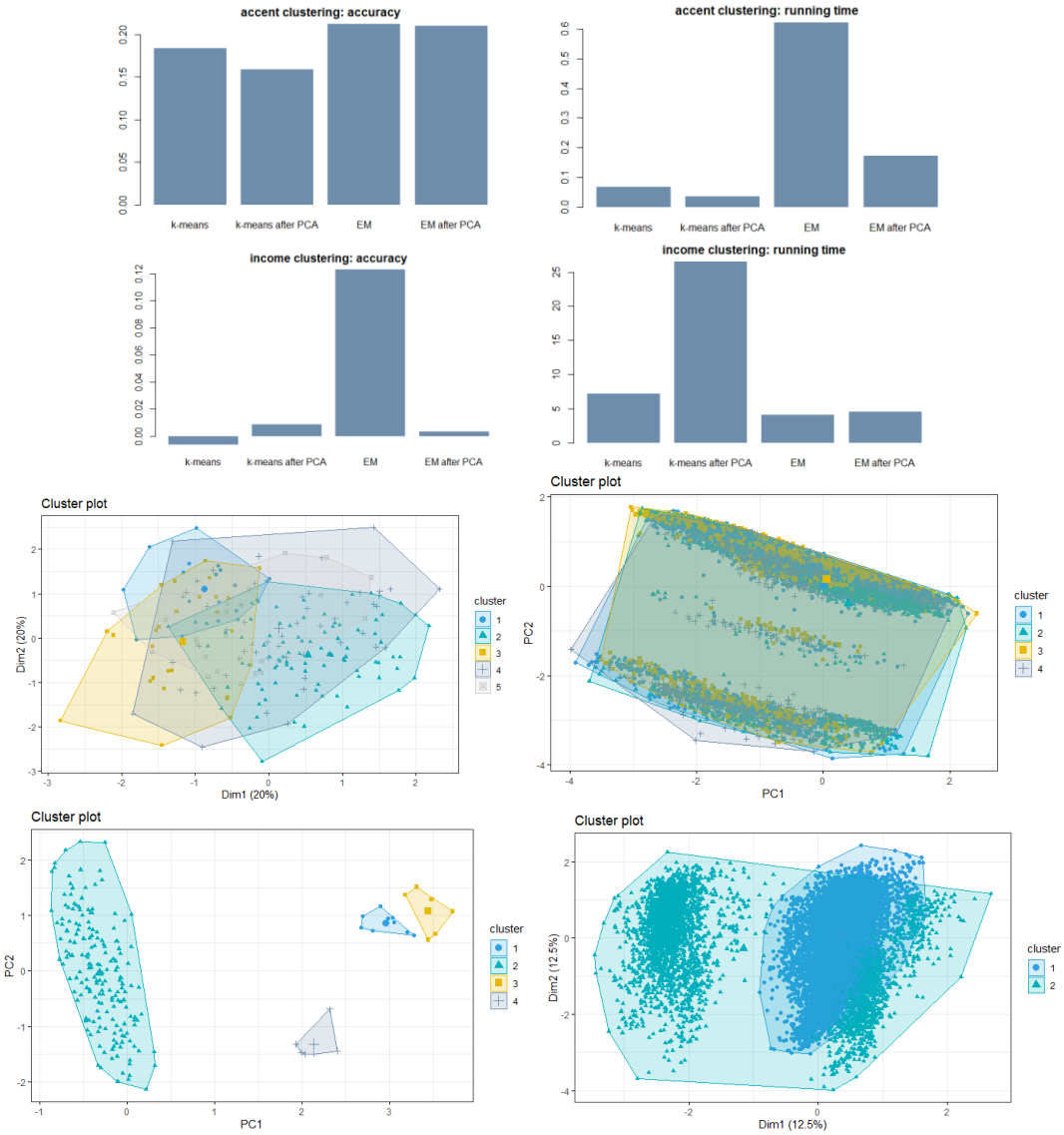


Figure 8. the running time and accuracy before and after PCA in two datasets.

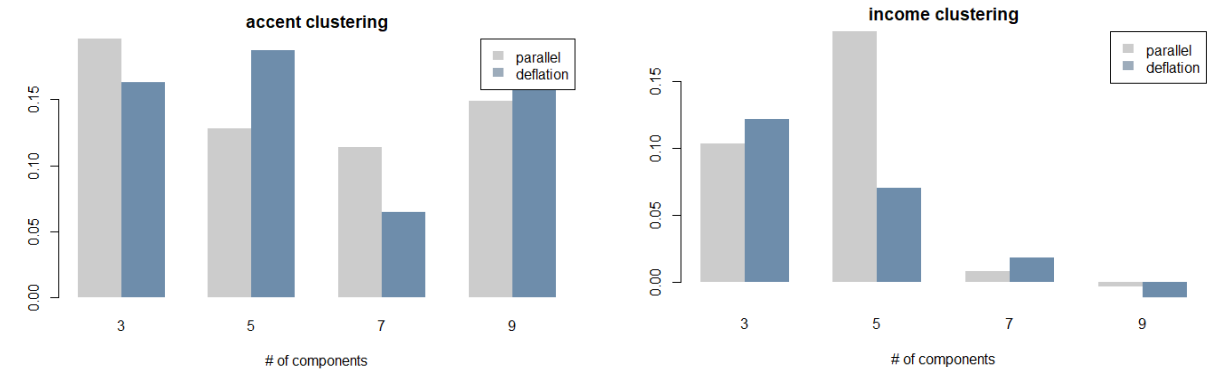
To rerun clustering after PCA, this paper selects the number of principal components which explains more than 85% of the variance, for accent dataset, first 5 components are selected and for income dataset first 8 components are selected. The accuracy and running time comparison between datasets before and after PCA is shown in figure 8. For accent dataset, PCA output the similar accuracy while decreases the running time substantially, especially in EM, however for income dataset, the accuracy decreases and the running time increases after implementing PCA. The performance also confirmed with what showed in the principal components generated from the two datasets that PCA does not work well in dataset with a mixture of quantitative and qualitative variables but works well in dataset with quantitative variables only. From the new generated clusters, we can also see that the clusters separates more obvious in the components generated from PCA.

ICA

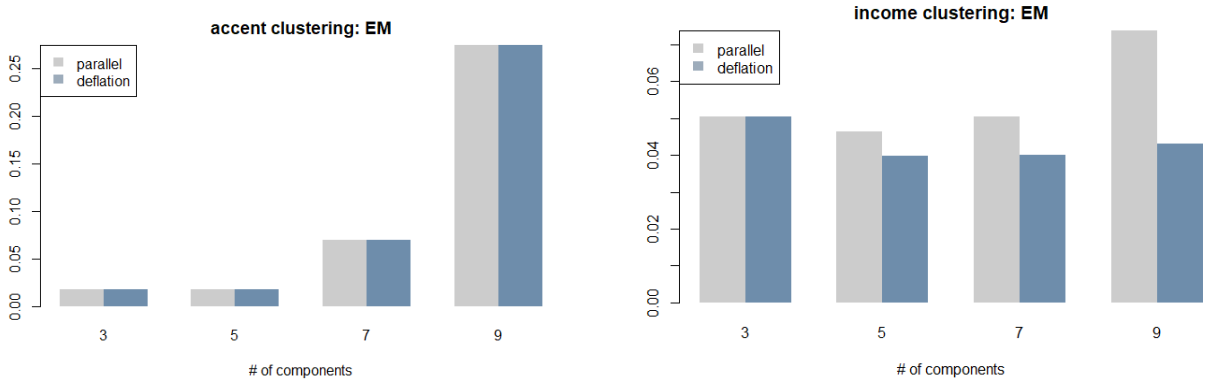
ICA looks for the most independence and separated the mix of variables to independent variables. The accuracies of ICA under different parameters are shown in figure 9. For accent dataset, the best parameter is parallel with 3 components for k-means and parallel with 9 components for EM. For income dataset, the best parameter is parallel with 5 components for k-means and parallel for 9 components for EM.

Using the best parameter, the running time and accuracy in k-means and EM are shown in figure 10. For accent dataset, the accuracy decreases but still almost as good as the accuracy before ICA while the running time decreases after the implementation of ICA. For income dataset, unlike the performance of PCA, ICA substantially improves the model in the performance of k-means algorithm, but also increases the running time in k-means clustering algorithm. .

In general, it is worth to implement ICA in accent dataset to reduce the compute complexity, but would not be an ideal algorithm for income dataset.



(a)



(b)

Figure 9. the adjusted accuracy of clustering after ICA in different algorithm types and different number of components in (a) k-means clustering and (b) EM.

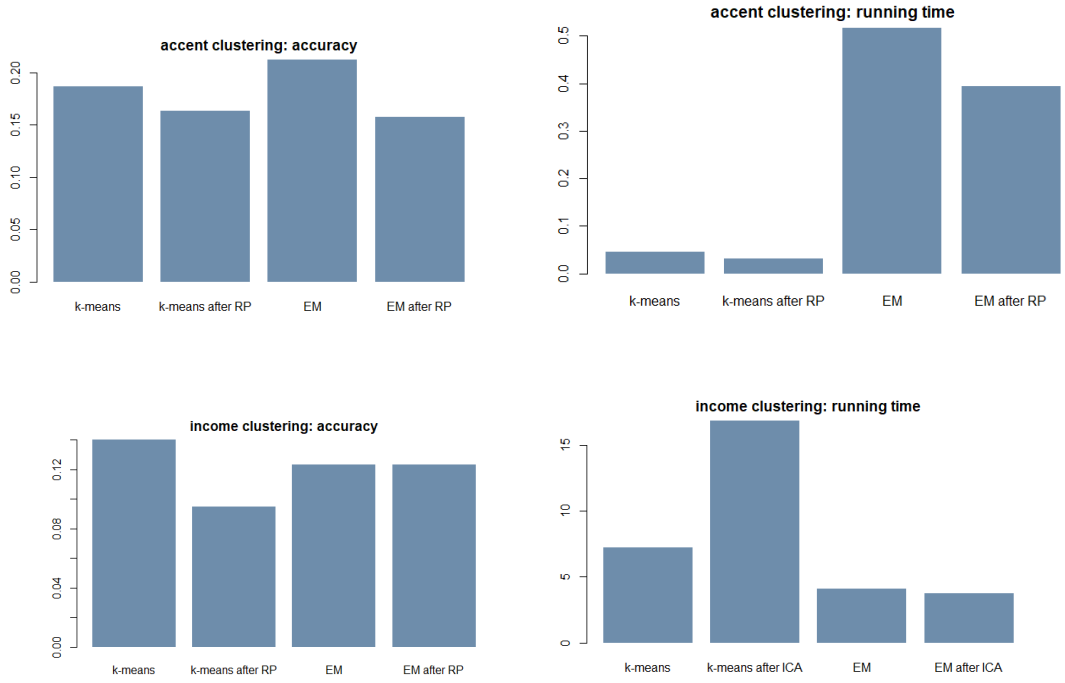


Figure 10. The accuracy and running time before and after ICA.

RANDOMIZED PROJECTIONS

Randomized projections decreases the number of dimensions of the original dataset by projecting the data in high dimensional space into the low dimensional space. For the experiment in this paper, randomized projection does not work in income dataset because of some structure of the dataset, so this paper will only apply this algorithm on accent dataset and the result is shown in figure 11. For accent dataset, although the running time decreases substantially because of the reduction of dimensions, but the accuracy also decreases substantially. So according to the experiment in this paper, randomized projections is not an ideal algorithm.

UNIFORM MANIFOLD APPROXIMATION AND PROJECTION (UMAP)

Uniform Manifold Approximation and Projection (UMAP) produce a low dimensional representation of high dimensional data that preserves relevant structure. It searches for a low dimensional projection of the data that has the closest possible equivalent fuzzy topological structure to the original high dimensional data. The results of umap are shown in figure 12.

The result shows that after implementing UMAP, the accuracy of the two clustering methods on accent dataset increases and the running time decreases substantially. But for income dataset, again, the accuracy decreases while the running time increases or stay the same. The different result shows that UMAP is more suitable for quantitative dataset other than categorical dataset or a mixture of both.

DISCUSSION

From what experimented above, it shows that most of the dimension reduction algorithms are more suitable for quantitative dataset, such as the accent dataset in this paper but have more limitations on the qualitative dataset, such as the income dataset in this paper.

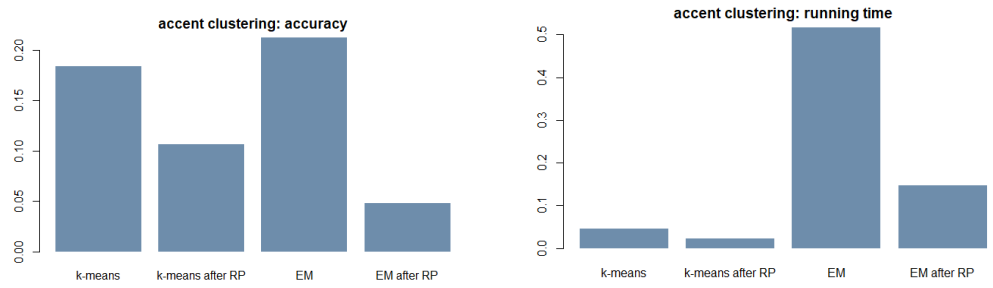


Figure 11. the accuracy and running time of accent dataset before and after the implementation of randomized projections.

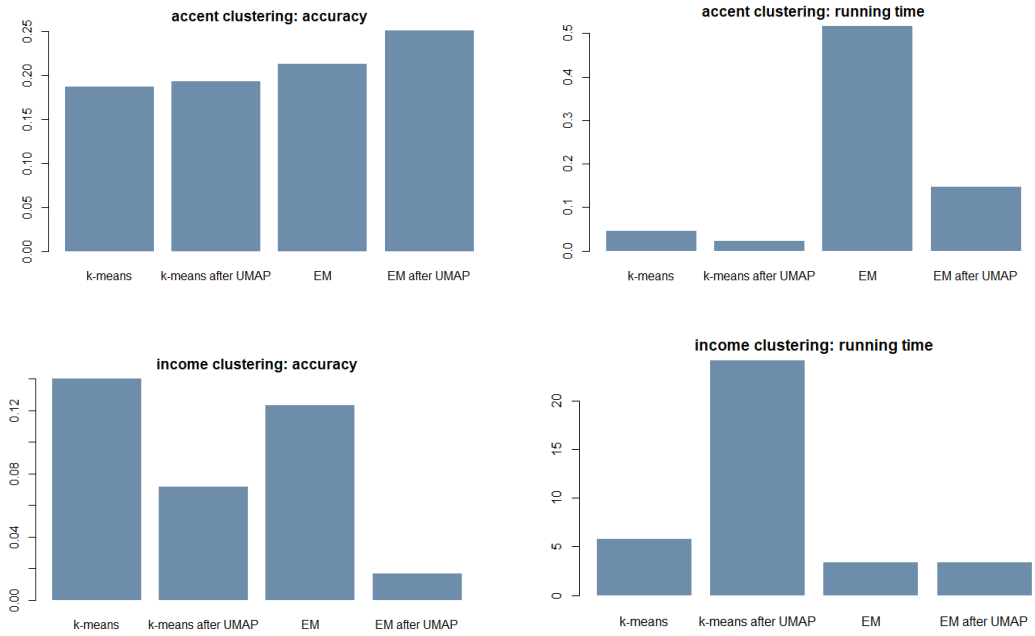


Figure12. the accuracy and running time of accent dataset before and after the implementation of umap.

PART 3 RERUN NEURAL NETWORK ON DATASETS AFTER DIMENSTION REDUCTION

I. Introduction

This part will first rerun NN on datasets after dimension reduction and then use the clusters the data point belongs to as the dependent variable and then rerun NN on the newly generated dataset. the dataset used in this part is the accent dataset.

II. NN After Dimension Reduction

The result of NN after Dimension Reduction is shown in figure 13. It shows that the training time are decreases after dimension reduction, but surprisingly the query time increases compared to original ANN models. And for accent dataset, PCA performs the best while has the almost as good as accuracy to the original ANN model and less training time.

III. NN After Dimension Reduction And Clustering

After replacing the original labels with the clusters, all four randomized reduction method have the accuracy of classification of 1 through neural network. The reason might be by reduce the dimension of the original datasets, we loss some information and the noise in the dataset, so the clusters generated from the dataset after dimension reduction would be almost noise-free, so when we try to classify the noise-free dataset, the result would be clean and the accuracy is 1.

The above result shows that while dimension reduction algorithm reduces the running time of the following algorithms, but it also introduces the risk of removing useful information which might be helpful in building a robust model.

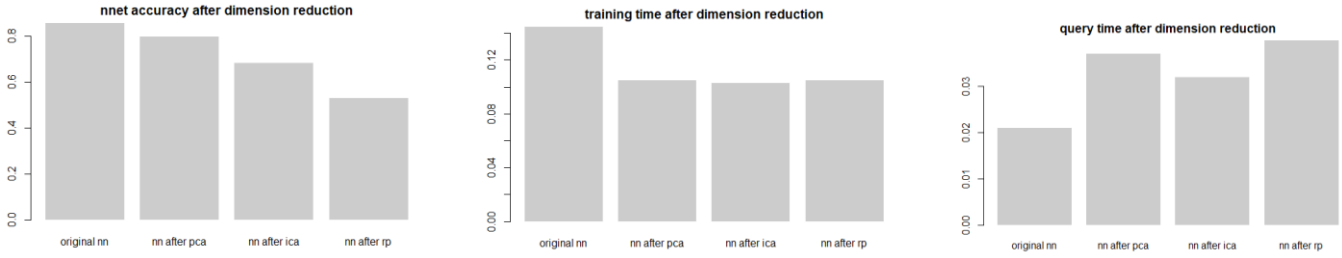


Figure 13. the performance of neural network after dimension reduction method.

References

[1] https://rpubs.com/Saskia/520216

[2] http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/118-principal-component-analysis-in-r-prcomp-vs-princomp/

[3] https://www.rdocumentation.org/packages/fastICA/versions/1.2-2/topics/fastICA