Unsupervised Learning and Dimensionality Reduction

Yichao Zhang
yzhang3414
CS7641 - Machine Learning

Mar 22, 2020

Code Repository:
`https://github.com/yichigo/Unsupervised-Learning-and-Dimensionality-Reduction`

0. Introduction

In this report, there are mainly 4 parts of studies:
In the 1st chapter, we ran 2 clustering algorithms: K-Means, Expectation–Maximization(EM, we use Gaussian Mixture here) on 2 data sets.
In the 2nd chapter, we applied 4 dimensionality reduction algorithms: PCA, ICA, Randomized Projections, and t-SNE. on the 2 data sets, and then repeat both clustering algorithms(K-Means and EM) on each case.
In the 3rd chapter, we ran Neural Network classification models on the heart disease data after dimensionality reduction.
In the 4th chapter, we added the clustering label as a new feature to the heart disease data, and repeated the Neural Network classification models.

Heart Disease Data Set (Used in HW1)

It will be quite useful that heart disease can be predicted from some simple medical measurements. So that people can save a lot of time to treat the disease rather than waiting for the busy doctors and suffering the disease without any treatment. Our data set is from:
`https://www.kaggle.com/ronitf/heart-disease-uci`
there are 13 features: age, sex, cp (chest pain type), trestbps (resting blood pressure)....., and 1 target with value 0 or 1.

Gestures Sensors Data Set (Used in HW1)

If the machine learning algorithms can figure out people's gestures, then this powerful tool can be applied in many areas, such as recoding sign language, taking care of children or patients, etc. The data set is from:
`https://www.kaggle.com/kyr7plus/emg-4`
There are 8 x 8 = 64 features: 8 muscle readings, and each one has 8 sensors. The target indicates the gesture in 4 types: rock - 0, scissors - 1, paper - 2, ok - 3.

One-Hot Encoding

For the features which indicate discrete types, we use one-hot encoding to replace each feature column with multi-columns, and each column represent one type of that feature. To keep the independence, we drop the 1st type since it can be derived from the other types. We also applied the one-hot encoding on the new features from the clustering labels.

Normalize the Features Values

Since clustering algorithms like K-Means are sensitive the the scale of the feature values, we use MinMaxScaler to normalize the features among $[0, 1]$

1. K-Means and Expectation–Maximization(EM)

K-Means

K-Means split the data into k clusters. We randomly initialize k centers. Step 1: assign the data points to each cluster based on the distance to the center. Step 2: calculate the mean value of each cluster and use this value as the new center. We repeat these 2 steps until it converges.

Expectation–Maximization(EM)

Expectation–Maximization assume at the data points generated from a mixture of some distributions. For Gaussian Mixture model here, we assume the points are from a mixture of k Gaussian distributions. We randomly initialize k Gaussian functions, we optimize the parameters(center, size, weight) of these Gaussian functions to maximize the likelihood of the data points. Each data points can be assigned softly or hardly to each component based on the Gaussian function value on that data point.

Clustering on Heart Disease Data

We choose k = 4, since many features are binary, and if we plot these features in 2 dimension, their are 4 cases. Although we do not plot on binary features, they affect on dimensionality reductions in the following studies.

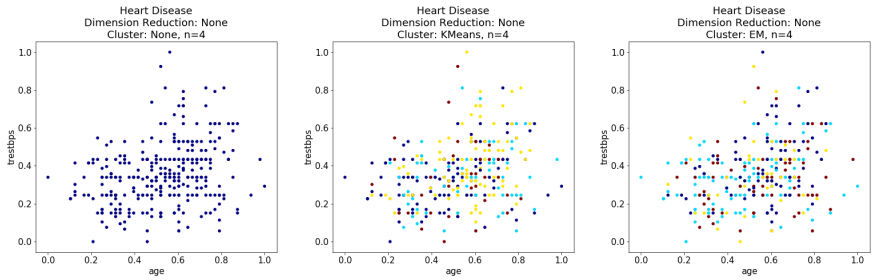

Figure 1: Scatter plot of heart disease data, with K-Means and EM Clustering results

As shown in Figure.1, both K-Means and EM worked bad on the original data. The adjusted rand scores are 0.225301 for K-Means and 0.204973 for EM.

Clustering on Gestures Sensors Data

We also choose k = 4 here, since there are 4 types of gestures, so we need at lease 4. But too many clusters are not good for visualization. Hopefully the clusters are useful in the predictive model.

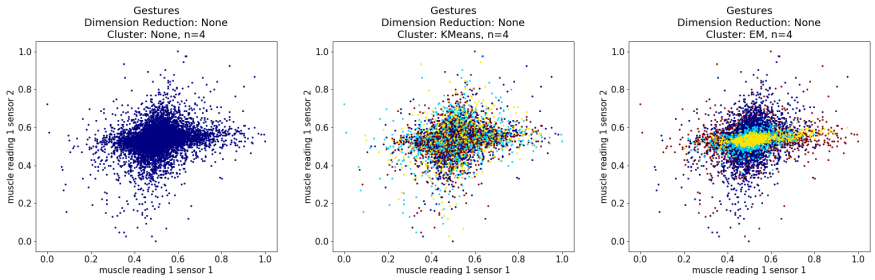

Figure 2: Scatter plot of gestures sensors data, with K-Means and EM Clustering results

As shown in Figure.2, K-Means worked bad, all the clusters are mixed to-

gether. But EM worked very good, it obviously split the original data set into 4 different Gaussian distributions: the yellow component in the center is the smallest, the light blue component in the same center has the middle size, the dark blue and dark red components are much larger. The adjusted rand score is only 0.0009 for K-Means and is 0.63 for EM.

2. Dimensionality Reduction and Clustering On Each Cases

To improve our clustering, we applied 4 dimensionality reduction below:

PCA find the direction which maximize the variance of the data. Then it remove this direction from the space, and find another direction which maximize the variance of the data. It repeat the process until it reaches the number of principle dimensions. Each principle directions are orthogonal.
ICA is very similar to PCA, the only difference is that it chooses the independent directions rather than orthogonal.
Randomized Projections (RP) the idea of Randomized Projection is that if points in a vector space are in a high dimension, they could be projected into a lower-dimensional space in some way that approximately keep the distances between the points.
T-distributed Stochastic Neighbor Embedding(t-SNE) embeds the points to a low-dimension space and keep the similarity of the data points, where the similarity can be calculated from Gaussian function based on the distance between points.

Dimensionality Reduction and Clustering on Heart Disease Data

Since there are few number (about 20) of features, we use 10 directions for PCA and ICA. We choose only 5 components for Randomized Projections and 3 components for t-SNE since they worked good and enough.
PCA has the best looking plots, since it focused more on the binary features. But these features may have nothing to do with the target, so the adjusted rand scores are very low: K-Means(0.25) and EM(0.2). The eigenvalues are $[0.286, 0.153, 0.102, 0.092, 0.072, 0.067, 0.051, 0.040, 0.031, 0.027]$

ICA worked good on both K-Means(0.51, the best) and EM(0.53). There are mainly 2 parts (left and right) in the scatter plot. The splitting on x

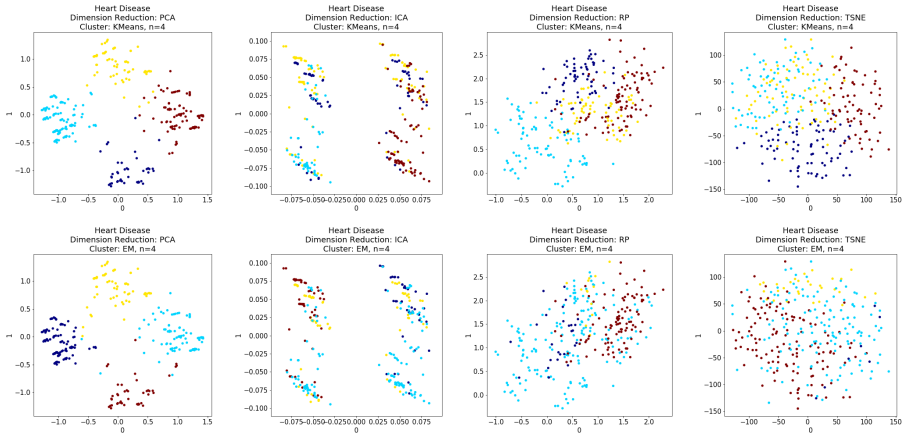

Figure 3: Dimensionality reduction (PCA, ICA, Randomized Projections, t-SNE) and clustering (K-Means, EM) on heart disease data

direction shows the domination of binary features, and the y direction is dominated by continuous features.

RP and t-SNE worked bad on K-Means($score < 0.2$), but did the best on the EM clustering, both scores are above 0.8.

When we re-run the RP, the clusters and variations changed, and the score changed from 0.6 to 0.9.

When we reproduced clustering experiments on the datasets projected onto the new spaces created by ICA, PCA, and RP, we get different results. Because K-Means and EM are sensitive to the initial state. Since we used a small data set here, K-Means was still robust, but EM changed a lot.

Dimensionality Reduction and Clustering on Gestures Sensors Data

Since there are 64 features, we decreased it by half and use 32 directions for PCA and ICA and RP. We choose only 3 components for t-SNE since it worked good and enough.

K-Means clustering (in the upper row of Figure.4) worked very bad ($score < 0.05$) on all the 4 dimensionality reductions. It is mainly because the points

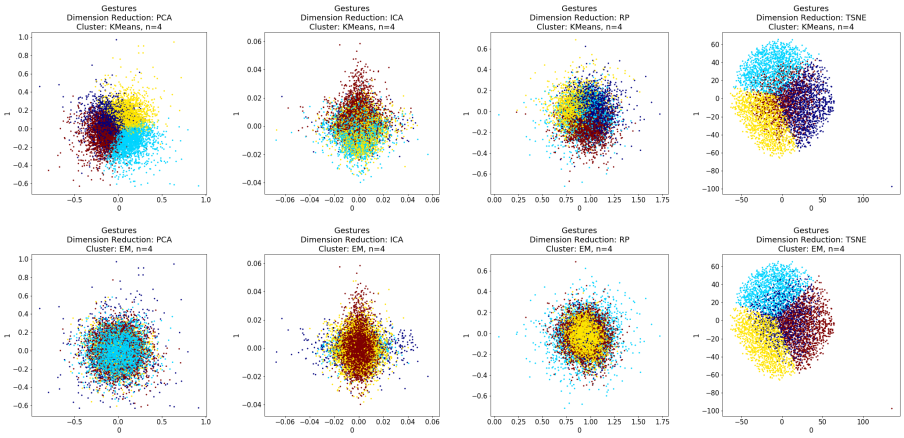

Figure 4: Dimensionality reduction (PCA, ICA, Randomized Projections, t-SNE) and clustering (K-Means, EM) on gestures sensors data

are together in a big single cloud, and K-Means split the cluster in some arbitrary directions, which are meaningless.

For EM Clustering:

PCA split the cluster into 4 similar Gaussian components in this plotting. It had best score(0.67) on EM, and caught the meaningful directions for target. The top 4 eigenvalues of PCA are: [0.06355192, 0.06068093, 0.05232303, 0.04682559,.....]

ICA split the points into 4 components, where the light blue and dark blue clusters spread more on 0th direction, while the red and yellow clusters spread more on 1st direction. They are not quite related to the target since the score is about 0.38, but not bad.

RP had a good score on EM 0.59, and we can clearly find different Gaussian components from the scatter plot. When we re-run the RP, the cluster and variance changed, and the score also changed from 0.4 to 0.7.

t-SNE is very slow on big data size, we use only 40% of the data. Althouth it has a good looking in the plot, it did not catch the meaningful features for the target (score = 0.03).

When we reproduced clustering experiments on the datasets projected onto the new spaces created by ICA, PCA, and RP, we get different results. Because K-Means and EM are sensitive to the initial state.

3. Neural Network After Dimensionality Reduction: Heart Disease Prediction

We optimized the Neural Network on the original data of heart disease by grid search. We chose a single hidden layer with 30 perceptrons(optimized from 10 to 300), identity activation(beat Relu, Tanh, Logistic), and Adam optimizer (beat 'lbfgs' and 'sgd').

Original Neural Network Performance

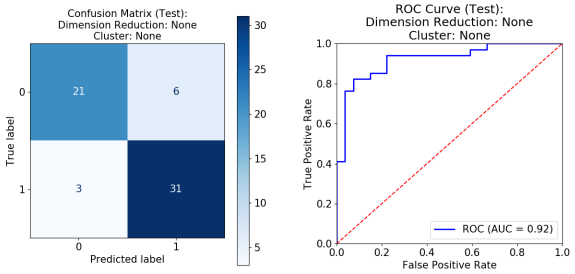

Figure 5: Confision Matrix and ROC curve for the original Neural Network model

The original neural network model's performance are shown in the Figure.5. It has 0.92 AUC score.

Neural Network after Dimensionality Reduction

By applying dimensionality reduction on the features, there is no improvement on the prediction performance (as shown in Figure.6). Data from PCA (AUC = 0.91) and ICA (AUC = 0.92) performed very similar with the previous data (AUC = 0.92). It shows that 10 components of new features are enough to describe the original features.

However, NN models on the data from RP (AUC = 0.85) and t-SNE(AUC = 0.82) are worse than before, the main reason is that both RP and t-SNE decrease the dimension of features a lot, and the simple linear neural network

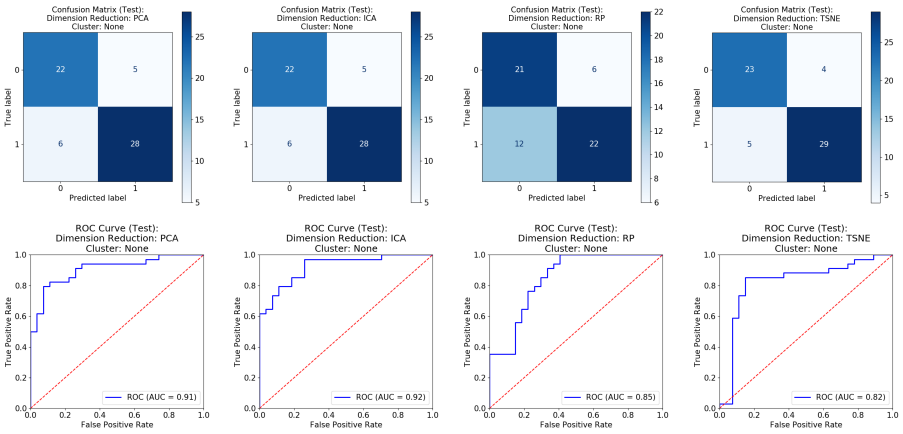

Figure 6: Confision Matrix and ROC curve for the heart disease prediction after dimensionality reduction (PCA, ICA, Randomized Projections, t-SNE)

cannot fit the target from such few features without knowing the clustering label information (which is non-linear).

The Neural Networks runs much faster on the data after dimensionality reduction, since we decreased the input dimensions, especially for RP and t-SNE.

4. Add Clustering Labels as New Features: Heart Disease Prediction)

Add K-Means label

Figure.7 shows the performance after adding the K-Means label as a new feature.

K-Means label of ICA really helped the ICA data for neural network. It increased the AUC score from 0.92 to 1.0 and perfectly predicted the heart disease. It also agreed with that ICA had the highest adjusted rand score (0.51) on the clustering performance.

However, data from PCA, RP and t-SNE had no improvement although they had good looking clustering scatter plots. The reason is that the K-Means algorithm split the points in some arbitrary way when the points distributed

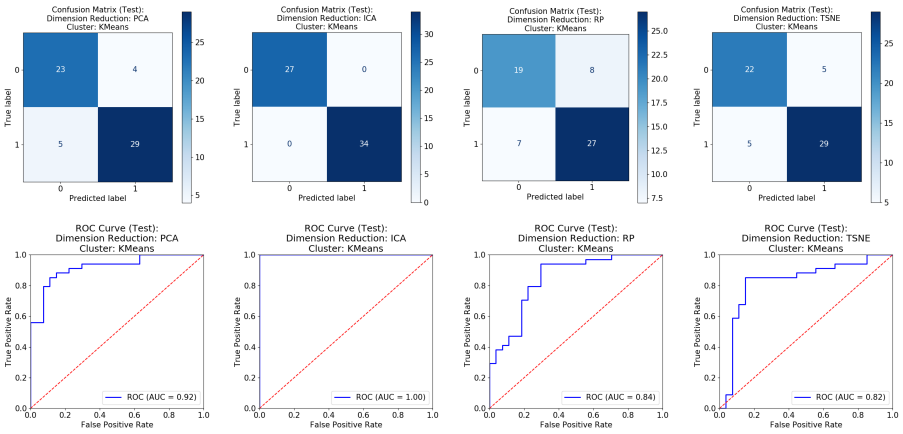

Figure 7: Confision Matrix and ROC curve of the heart disease prediction with clustering labels as new features. There are 4 dimensionality reductions (PCA, ICA, Randomized Projections, t-SNE) and 1 clustering algorithms (K-Means)

in a single big cloud. So in this case, the K-Means cluster label may not provide any meaningful information.

Add EM label

Figure.8 shows the performance after adding the EM label as a new feature.

EM labels are very helpful on the data from ICA, RP and t-SNE. The AUC score of neural network on these 3 data sets achieved 1.0 and perfectly predicted the heart disease.

However, data from PCA had no improvement. If we looking back the their EM clustering Figure.3 and their scores, we can find the reason:

Althouth EM clustering on PCA data split the points clearly into 4 clusters in the top 2 principle dimensions, these 2 dimensions may have nothing to do with the heart disease target. So the adjusted rand score is very low (0.22).

On the other hands, EM on the data of ICA, RP, and t-SNE may not split the points clearly in the scatter plot of top 2 dimensions, but their rand scores

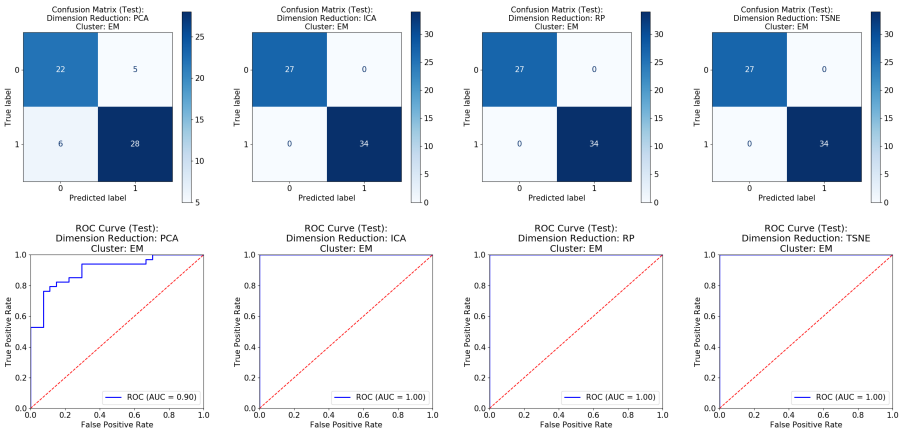

Figure 8: Confision Matrix and ROC curve of the heart disease prediction with clustering labels as new features. We have 4 dimensionality reduction (PCA, ICA, Randomized Projections, t-SNE) and 1 clustering algorithms (EM)

are much higher: ICA(0.53), RP(0.93), and t-SNE(0.83), Which means we caught the important features for heart disease prediction.

Since the added only 1 new feature, the Neural Networks still runs much faster on the data after dimensionality reduction, especially for RP and t-SNE.