Unsupervised Learning & Dimensionality Reduction

CS-7641 Assignment3

Zhengyang Chen GTID#903338861

1 Introduction

1.1 Summary

This passage is the analysis of experiments on unsupervised learning and dimensional reduction algorithms. The steps of experiments are mainly divided into 3 parts -The first task is to run the clustering task and try to figure out the out put, the second is to run the dimensionality reduction algorithm while compare and contrast the output and apply it to clustering. And the final part is to review the neural network algorithm and apply both clustering and dimensionality reduction on it as new features, and see the change of performance.

The main goal is to understand the unsupervised learning and feature selection and transformation, and see how them correspond to data characteristics.

1.2 Dataset selection

The dataset for this experiment is from previous experiments, the bank client dataset and the student grade dataset.

The bank client dataset consists of more than 10,000 piece of data in 18 features, some of them are related to each other while also contains several noise feature , and are labeled as binary classification task, it is interesting to see algorithm performance in large amount of data and also how they reveal the interaction as well.

The student grade dataset is from uci repository, with 32 features in 649 items, the final output of the dataset (final grade) is strongly related to the two period grades(G1 and G2), and other items are of less relation or just noises. It is interesting to see how the dimensionality reduction can deal with the amount of features, as well as figure out, or at least imply the most related items.

2 Clustering Algorithm

2.1 Introduction to Algorithm

2.1.1 K-Means

K-Means algorithm works by randomly select start points from the vector space defined by features to be the starting cluster center, and then examine the closest point by distance measure (usually euclidean), and assign the item to the cluster defined by center, then shift the center by calculating the mean of the cluster points to find the target space, after shifting it again calculate the distance and assign to new center, etc. At the end of iteration, we can find the clusters that describe some 'inner' relationship between data elements. It is unsupervised learning since it does not have the labels as before.

2.1.2 Expectation-Maximization Algorithm

The Expectation-Maximization is the core to implement Gaussian mixture clustering. Unlike calculating the distance to assign points rigidly to the cluster, the Gaussian Mixture clustering does not tell 'for sure' which cluster it belongs to, but calculate the possibility of that. Thus some point at the 'edge' may be of similar possibility(something between 0.4-0.6 or so) of being assigned to both clusters. It is also in two main iteration part- the E part evaluate the expectation of likelihood, and the M part finding the features that maximize it, after the iteration we can also figure out the clusters generated.

2.2 Compare and Contrast

2.2.1 Clustering distribution and visualization

To begin with the experiment, both of the dataset is set to binary classification task, while the student dataset has the possibility to produce multiple classifications, so let's do with k=2 to see whether the clustering have something that can line up with the classification

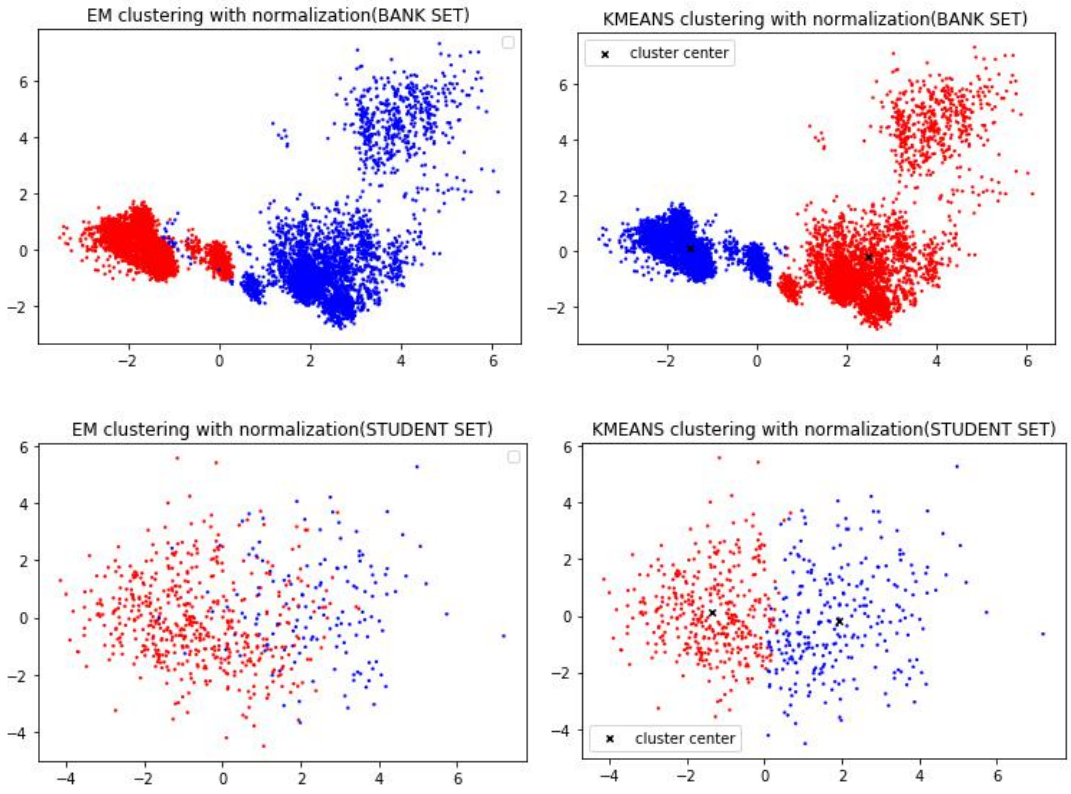


Figure 1

Figure 1 is provided by running the clustering algorithm on both data set and then do dimensionality reduction of PCA to project on the 2d figure. The first discover is that both clustering have the duplicated or overlaped zone, and that is the nature of dimensionality reduction, since both dataset have much more than 2 figures, while we project on 2 dimension, there should be loss by abandoning the features(see PCA part), and the student set seems to overlap more than the bank set, and that can imply the ratio of dimensionality reducted features distribution in both dataset - the first two reconstructed features have the most variance ratio in bank set, which can be proved in section 3-PCA part.

Another discover is that the EM clustering has more overlaped space than kMeans, to figure out the cause, we should refer to the mechanism of these two algorithm - the EM is calculating the expectation of likelihood, dealing with the probability while kMeans is of 'inflexible' assignment of data, thus kMeans is like shifting the boundary while shifting the center, but EM is trying to 'fit' the boundary. When in these two case where I stopped at limited iteration, EM can have the 'margin' projected into 2d space- more explicit than kMeans.

2.2.2 Clustering performance and scores comparison

At the very beginning we ve set the k value to 2,what if the number of cluster increases? Does the number of the cluster have some thing to do with the output? The following experiment is based on the performance measured by 3 scores- adjusted rand score, adjusted mutual info score, v measure score. These 3 scores are all used to describe the similarity of cluster labels and the dataset labels - the more score is, the more similar of the labels.

Figure 2 shows the result when we increase the k number, we can see that the student dataset experiments have the top score when k=2, that is definitely demonstrating the relationship between the cluster and labels, we can say that in some words the cluster is related to the label, and all the alike features have the tendency to gather and contribute to the output. In other words, the label can be considered as the result of the complex features. However in bank dataset, the score at k=3 is slightly higher than k=2, and the k-Means score is like the EM score, and both of them are quite low. Thus we can make the assumption that the natural distribution of the

feature in bank data set does not tightly linked to the output - maybe in this dataset one or few feature is dominant however other features tend to be unrelated or noise.

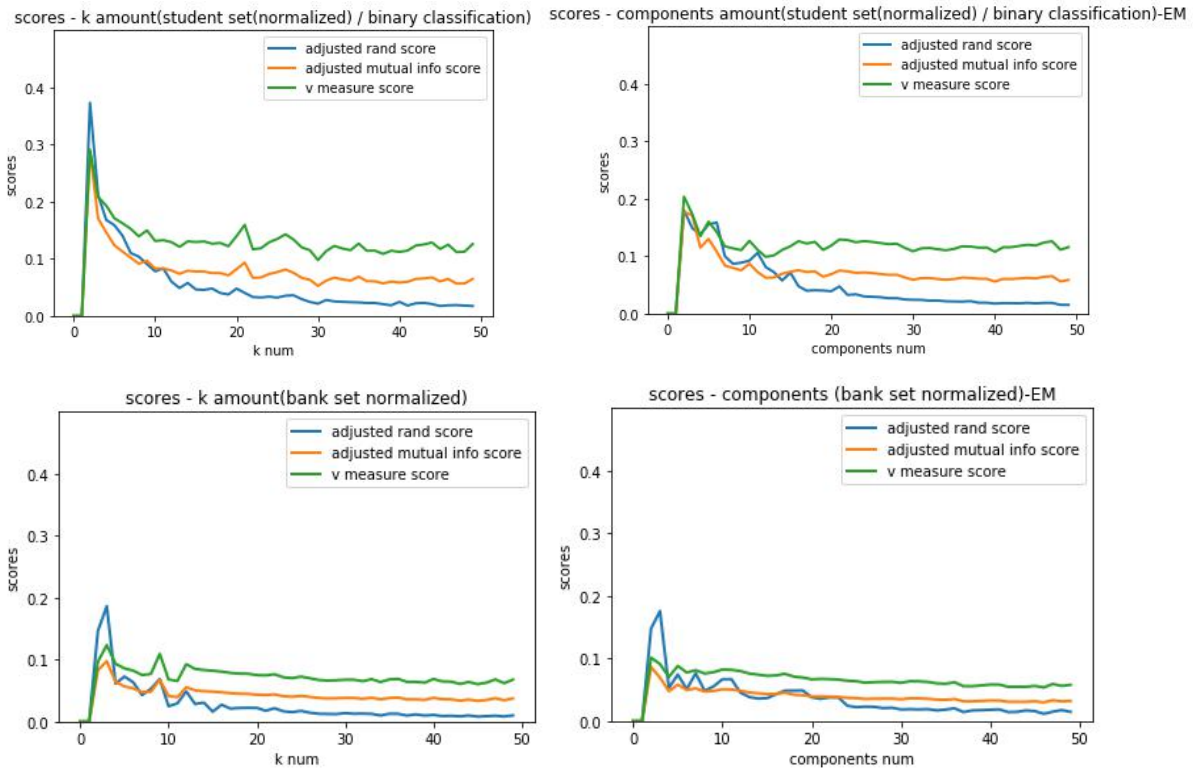


Figure2

Hence in the following experiment , we focus on student set, which seems to reveal more relationship between clusters and labels.

Another discovery is that in student dataset the kMeans tend to have higher scores over EM at k=2, it is abnormal since kMeans is a 'hard' clustering, but one possible assumption is that the distribution of data is linear-like, which make kMeans easy to find the discriminative boundary but confuse the EM calculation, and it lines up to the attribute of dataset with 2 most related features with significant variance

2.2.3 Normalized vs Un-normalized

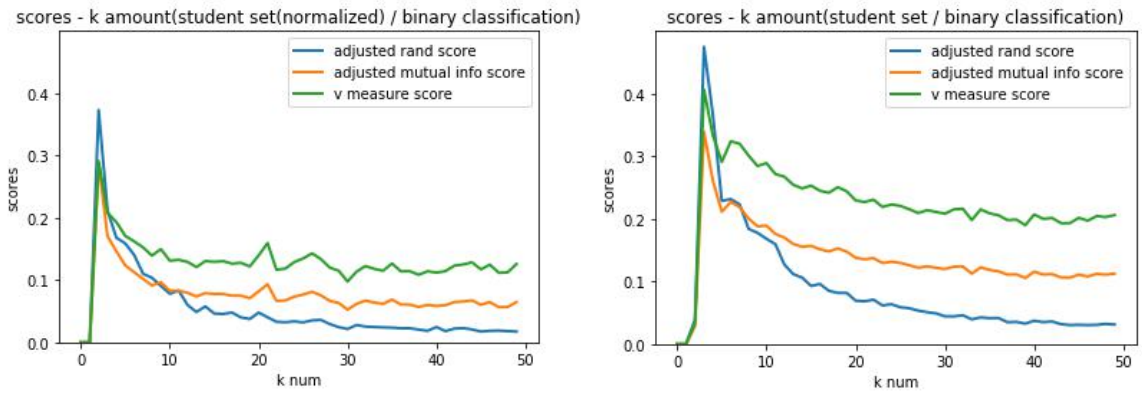
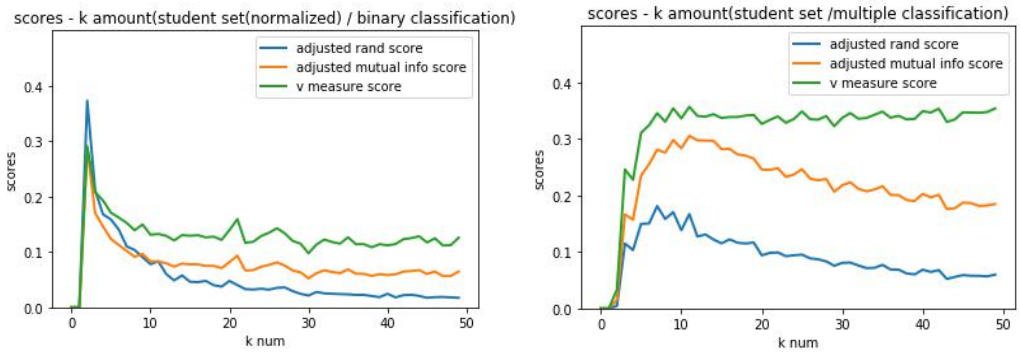


Figure 3

Another interesting discovery is that the normalized data seems to perform better than the one without normalization in student data set. Like the analysis above, we can make the explanation by describing the nature of the data set - look back on to the student set, we already have the knowledge that there're two feature most related to the output -the first and second period of grade. When we compare them with other features, we can

find out that compared with other binarized or labelized discrete features, these two feature are both on the highest mean and the largest variance - thus before normalization, the advantage of these two features enlarged.

2.2.4 Binary vs Multi-classification



3 Feature Transformation and Dimensionality Reduction

3.1 PCA

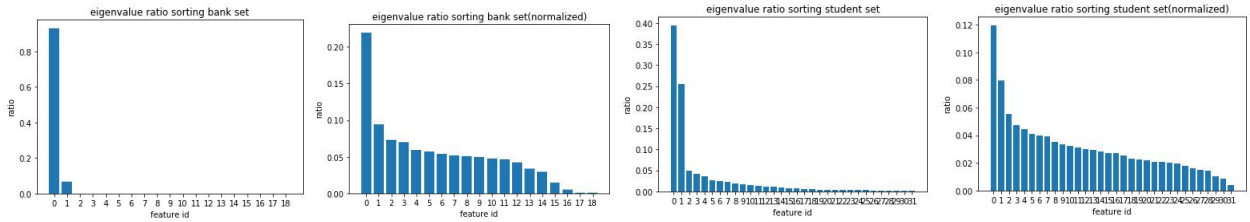


Figure 5

PCA is the method of dimensionality reduction by projecting them onto orthogonal axes that maximize the variance of data. The first axis is the projection with most variance while the last is least.

When we ran PCA on both dataset we can see that the distinct of ratio of reduced feature is more on the bank dataset, which can provide the explanation for the clustering results above in section2. Thus we can figure out that the first axes, or feature of reconstructed data from bank dataset is more dominant, and the variance in original space is less evenly distributed than student set.

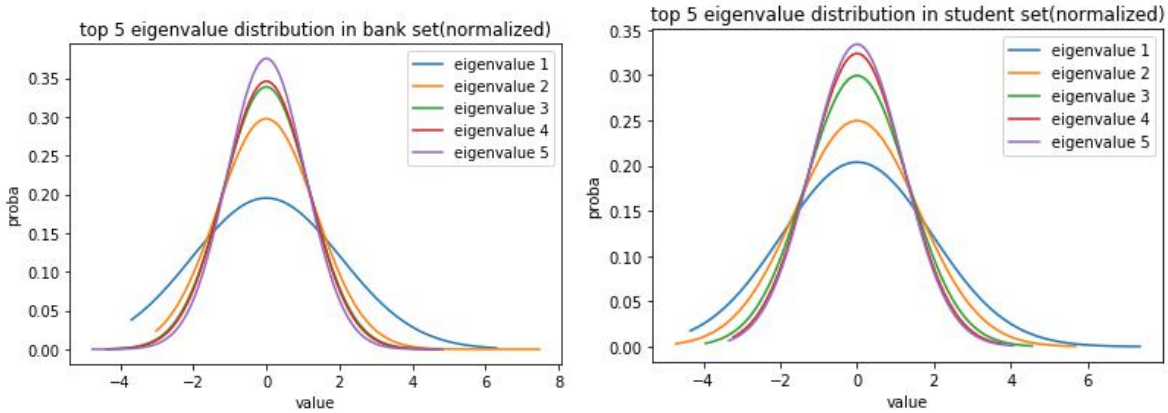


Figure 6

Figure 6 can also provide another aspect of view- by viewing the distribution of dataset, the sharper the curve is, the less variance it has. The conclusion is the same.

3.2 ICA

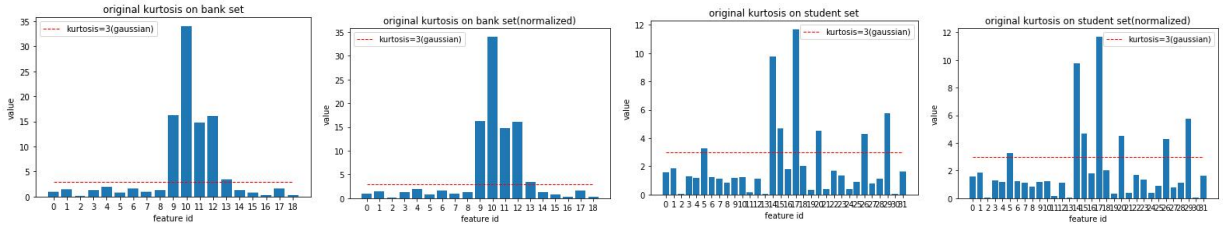


Figure7

ICA is something different from PCA, it is trying to find the 'independent' axis in the spaces, where the projection(in ICA in fact is the linear transformation) on the axis reconstruct the features, and each reconstructed features are independent from each other, or in other words, mutual orthogonal.

Figure7 indicates the original distribution of features in kurtosis, we can see that the scaling does not affect the distribution very much.

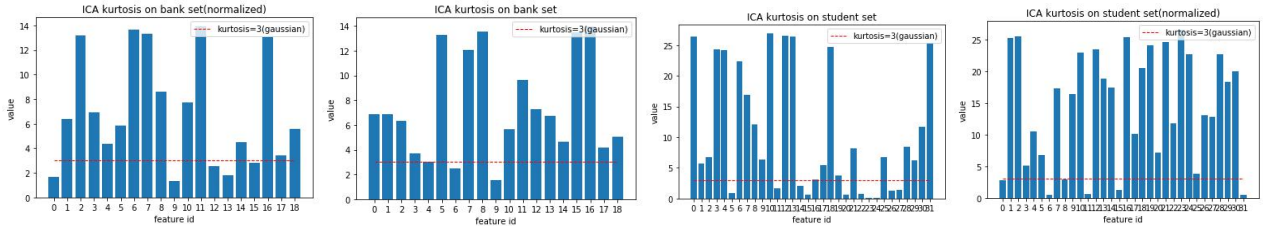


Figure8

However in figure8, the out put is not similar from each other before/after normalization, where it should be. As we analyse the origin data and after ICA calculation, we can figure out that some of the values is equal to 3- which indicates that it is gaussian- however it should not be, since ICA quickly get confused in Gaussian distribution.

3.3 Randomized Projection

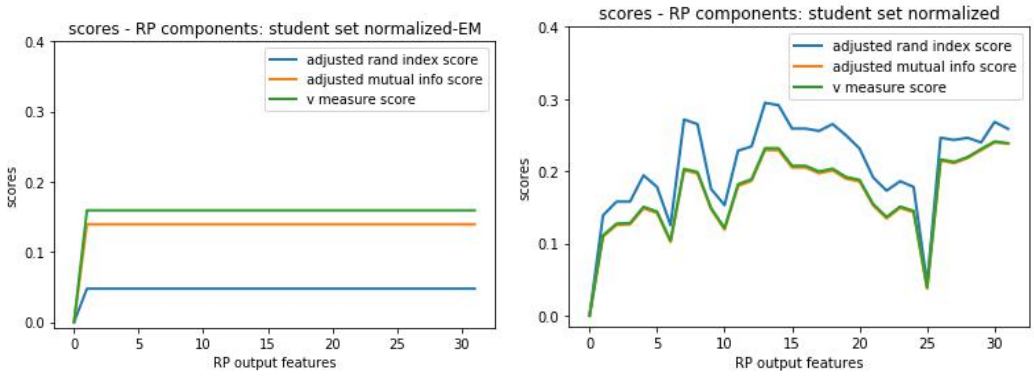


Figure9

Randomized projection is kind of a 'random' version of PCA, it lets the projection to axes, however the axis are randomized rather than choosing the max variance.

We can figure out from figure9 and figure 10 that if we make randomized projection on EM clusters the curve is something 'stablized', while the kMeans have the tend to zigzag when we increase output features, the explanation could be that EM is dealing with probability.

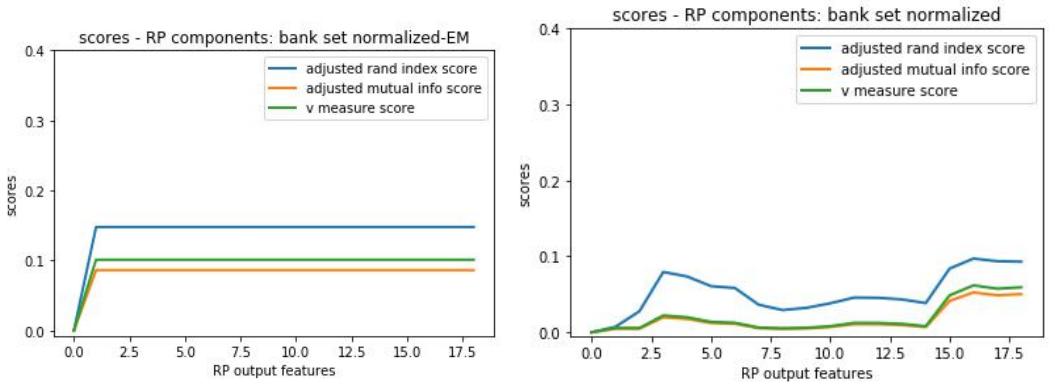


Figure10

We can also see from the diagram figure 11(right) that when we apply the RP algorithm on itself, it will eventually converge to a lower variance , however if we set the random state still, it will enlarge the variance.The explanation could be compared with SA algorithm, the randomized projection is just like the cooling process, and the first diagram is that if we define the random state, it will follow the same rate of variance change - unfortunately the first step is the variance going up - just like we do not cool , but remain heated in SA algorithm.

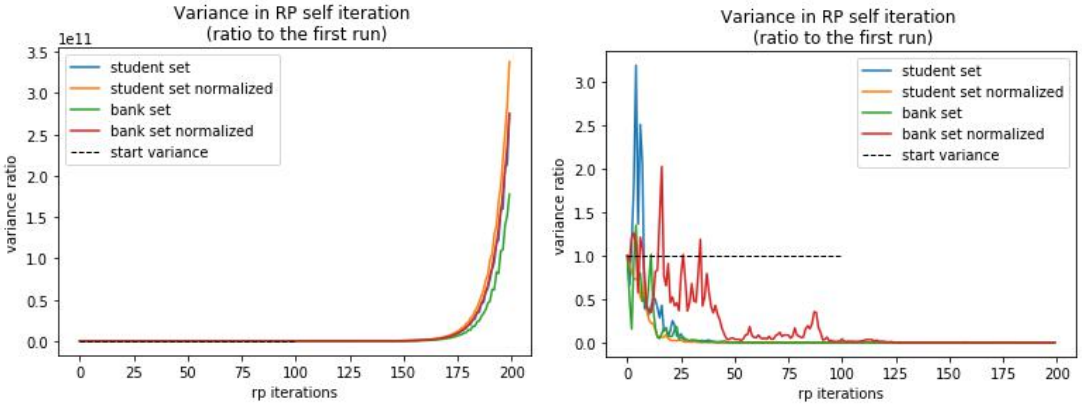


Figure 11

3.4 LDA

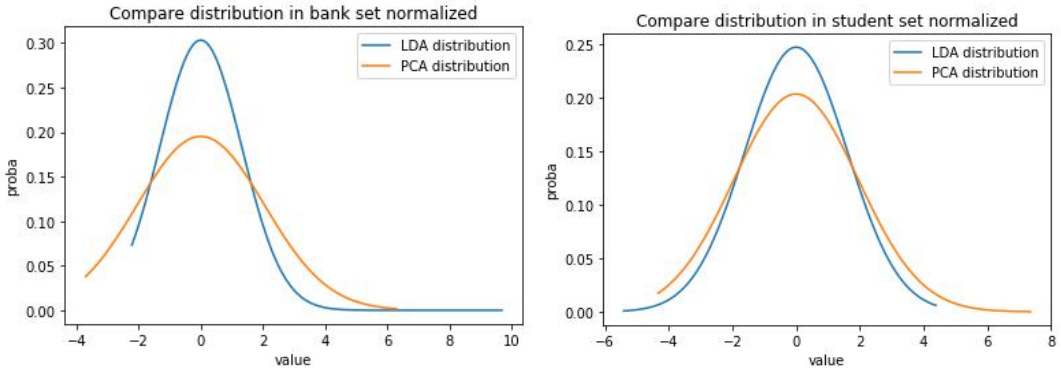


Figure 12

The final part comes to the LDA, it is a supervised algorithm that find a lower linear axis to discriminate the feature projection as the label indicates. The figure 12 shows that the LDA has the similar performance on the first projection as PCA, but it is more sharp- i.e. less variance. The reason can be explained that PCA is picking the most variance direction while LDA is of an axis of global distribution status, thus the mean of the variance is definitely lower than PCA.

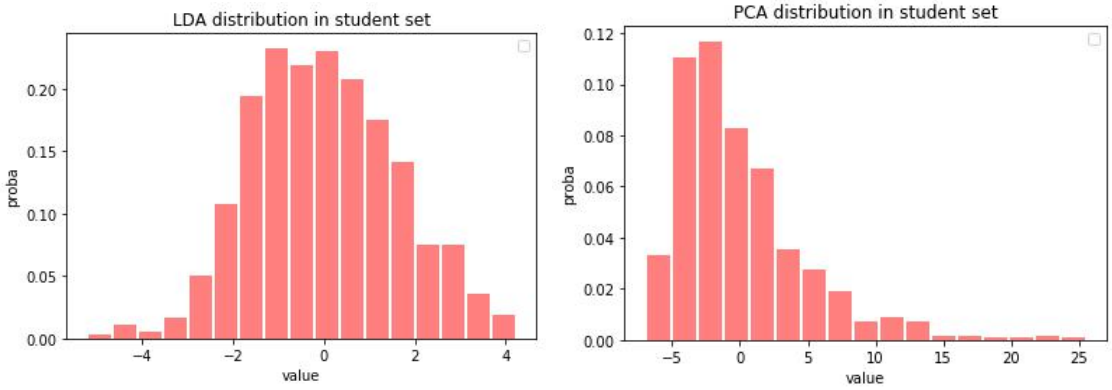


Figure 12

4 Collaboration of Dimensionality Reduction and Clustering

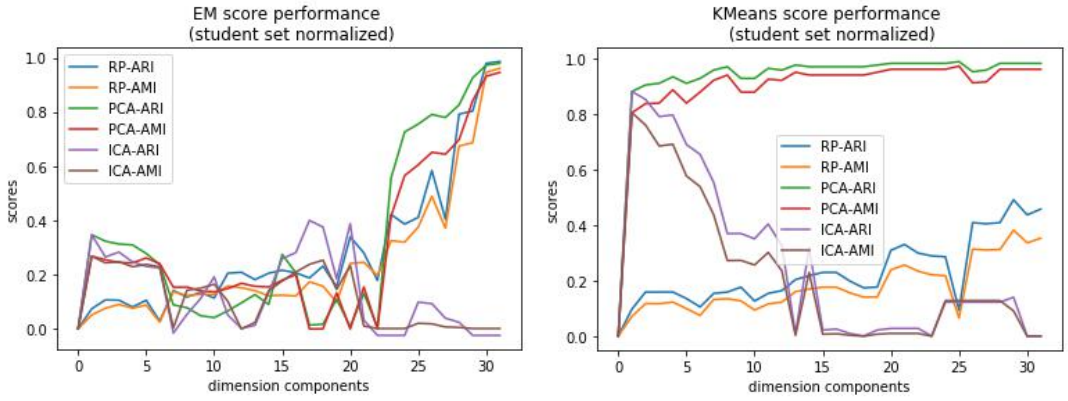


Figure 13

The performance of clustering may have something to do with the dimension reduction algorithm, when we make parameter of dimension components as x axis, and run EM and KMeans algorithm after dimension reduction, we can figure out the general tendency that - with the dimension increase, the performance of clustering incresed generally, it is because that the more dimension can indicates more information of distribution in the space, or saying, more authentic. However, the ICA seems quite different - it is achieving lower scores, that is because the natural or definition of ICA is to find the hidden variable behind, and if the reconstructed features, or components, is exceeding the real hidden variables, it get confused, just as the cocktail party problem, if 3 people are talking, we can not figure out 4.

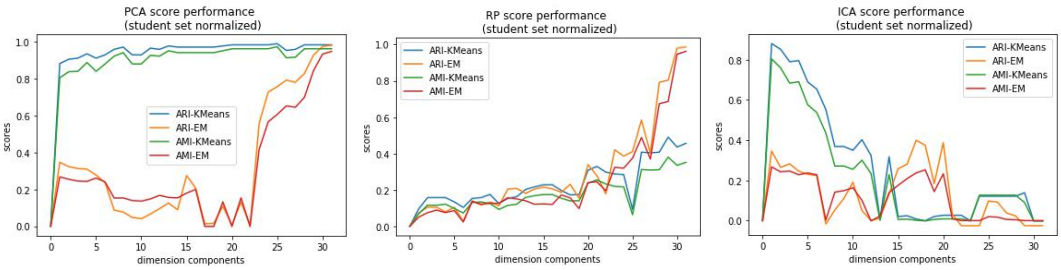


Figure 14

5 Application of Clustering and Dimensionality Reduction on Neural Network

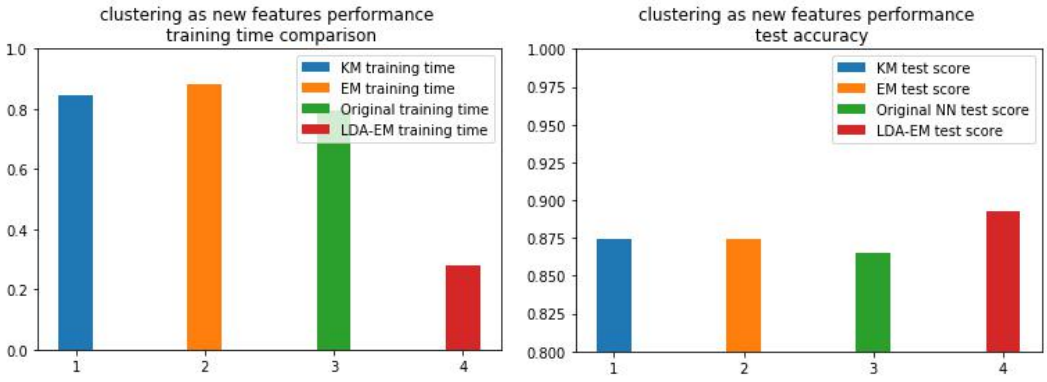


Figure 15

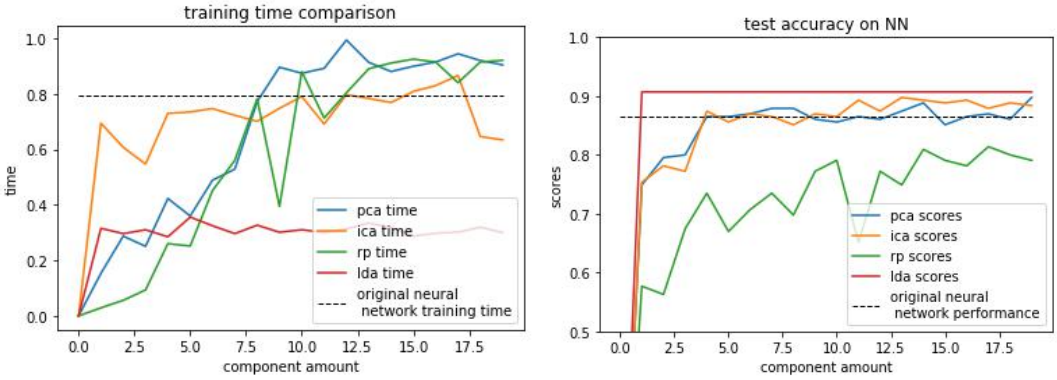


Figure 16

When we apply the clustering and dimension reduction algorithm back to the data set and rerun the neural network, we can figure out different outputs.The application of dimension reduction indicates that the testing accuracy can reach the same status or even above in neural network, and the training time is reduced(figure 16)

The reason is that in this case several dimension reduction algorithm do not only reduce the dimension, but also in some case act as a filter of unrelated items, and luckily in student set, the most related items are of most variance and mean, thus by dimension reduction the advantage is enlarged, and the unrelated items(usually lies in the discrete features) are filtered.

From figure 15 we can also see that the clustering result as new feature does not significantly increase performance ,even slightly increase the training time, but if we can apply them together with dimension reduction algorithm, the advantage are distinct - by doing EM together with LDA, the accuracy increased and the training time greatly declined.

6 Conclusion

In a nutshell, both EM and K-Means can help figure out the clusters to describe the relation inside the data in two steps of iteration - EM calculating the maximum expectation of likelihood while K-Means calculate the distance, then EM find the maximum and K-Means find the means. All these clustering algorithm is to 'learn' the model independently, i.e., without studying the labels. This is unlike the supervised learning that is label-dependent. However, the cluster sometimes does not always match the classification, and it can be also affected by the attributes of data -scales, variances, etc, so normalizing data is usually useful in the clustering task.

But it should not be considered a common step of normalizing, since the data attributes can collaborate and affect the output, if some related continuous values exists, like longitude and latitude, or math grades and physics grades, it is not always in need of scaling since it may cause distortion in data space. However if bunches

of unrelated features are of significantly different scale, then the standardizing should be useful to assign weights. And also the binary or discrete features can be confusing in the clustering task. So all saying goes in one sentence - It depends on the data.

As for dimensionality reduction, the task is to get rid of the curse of dimensionality when the dimension of data increases by constantly appending new features to it. The PCA is trying to find the best transformed axes by figuring out the direction of most variance, then second,third, etc, thus the sorted output is the composed data in an order of 'best' describe the most related features. Unlike physically projection to axes in space, the ICA is trying to find out the linear transformation that can transform the data to new feature space in which each component of the new feature space is of independent elements from others, the random projection is like the random version of PCA - to find the random axes for the data to project onto. The LDA, unlike the other 3, is of a 'supervised' dimensionality reduction - to study the labels and try to figure out the projection to components according to the classification numbers.

All these dimensionality reductions as pros and cons, PCA is a global algorithm and the reconstructed features are orthogonal to each other thus reduce the mutual relation effects, and it is also quite cheap to compute, however it can not figure out the real-related features that has less contribution to the variance.ICA is kind of local compared with PCA, and can quickly converge while figure out the independent components and is efficient to deal with the mixed data with independent 'cause' behind, however it can not deal with the original Gaussian distribution that contributes to the mixed data. RP is really fast to compute, however, the character of random make it hard to figure out good output in limited iteration. LDA is supervised thus can make use of the domain knowledge, and can be efficient when data distribution is more based on means rather than variance. However it has limitation to the dimension number (n_class-1), if we want more dimension in binary classification task it is helpless. It can also have the risk of overfitting.

So, all in all, the purpose of this passage is to conclude into one sentence - It depends on the data. And the task of clustering also targets to revealing the relationship between data elements. If the supervised learning is to learn from the data, then the unsupervised one is to study and explore it.