Abstract:

In this report, two cluster algorithms (k-means and expectation maximization) are used to investigate two datasets (same as HW1). Four dimensionality reduction algorithms (Principal component analysis, Independent Components Analysis, Random Projection and Factor Analysis) are applied to two datasets. Then, clustering experiments are reproduced using the data after dimensionality reduction. Finally, neural networks are trained using the data with dimensionality recution and the data with both dimensionality reduction and cluster.

Datasets:

1. First data set is the ***Letter Recognition Data Set*** [1] from UCI databases. It has 20000 samples of black-or-white pixel displays as a capital letters in the 26 English alphabet. Each data has 16 primitive numerical features such as *width of box*, *total pixel,* etc. The dataset is featured as a balance set.
2. Second data set is the ***Car Evaluation Data Set*** [2] from UCI databases too. It has 1728 instance of car evaluation range from unacceptable, acceptable, good to very good. Six attributes deciding the value which includes buying cost, number of doors, safety level ,etc. The dataset is featured as an un-balance set. Detail information shown in reference.

Objectives:

1. ***Letter Recognition Data Set***: Letter recognition is a fundamental task in computer vison framework. However the role is quite crucial. Corresponding machine learning model could be applied to recognize the autopilot, robotics, object recognization and so on. It is also used to test the performance of these algorithms on balance data set.
2. ***Car Evaluation Data Set***: Evaluation problem could be widely applied to financial market. Current car evaluation model could be used in used car trading or further modified for other business, such as stock, real estate or other properties. The performance of cluster/dimensionality reduction algorithms are also showed for un-balance data set.

Environment:

Python version: 3.7

Package used: sklearn, Numpy, SciPy, Pandas, matplotlib and time.

Part 1. Clustering:

K-means clustering (KM)

K-means is a classic unsupervised learning which labels the data as clusters. The algorithm assigns data points to K clusters based on the distance to the cluster centroid and then re-calculate the centroid to update the cluster iteratively. Thus the data are labeled as different clusters with similar feature. In present report, Euclidean distance is used for clustering datapoints. Sum of squared distances of datapoints to their closest cluster center (SSE) is used to evaluate the number of clusters. Besides,

Silhouette score (how data points similar to its own cluster comparing to other clusters), homogeneity scoe (how each cluster contains only data of its own class), completeness (how all data points are members of a same cluster), and Normalized Mutual Information (NMI, normalized mutual Information between two clusterings) are plotted to give a deeper look of the results.

Expectation Maximization (EM)

The Expectation Maximization is an iterative algorithm which maximize the likelihood by adjusting the model parameters. It is an iterative way to approximate the maximum likelihood function. The algorithm iterates between expectation step that estimates the log-likelihood evaluated of current parameters and maximizing step that calculates the parameters for maximum log-likelihood based on expectation step. Besides Silhouette, homogeneity, completeness, Akaike information criterion (AIC) and Bayesian information criterion (BIC) are adopted for EM to evaluate the clustering results. Gaussian mixture models (GMM) in Sklearn is used for EM in this report.

Fig.1 shows the clustering results of K-means on the letter recognition data set. Elbow method is used to decide the number of clusters. However, as showed in the figure SSE vs. number of clusters, the curve is quite smooth. So it might suggest that the data are quite evenly distributed so no significant cluster are showed, which results in the difficulty of finding the elbow point for cluster. A flat low-value silhoutte line indicates K-means is not quite good at this dataset too. K=6 is used to approximate the elbow location here, the distribution of clustered data set is shown in Fig.1(c) also. For EM, change of AIC/BIC and silhouette score are both considered to yield the number of components, which is choosed to be 4 in this study. The data distribution could be seen in Fig.1(f).

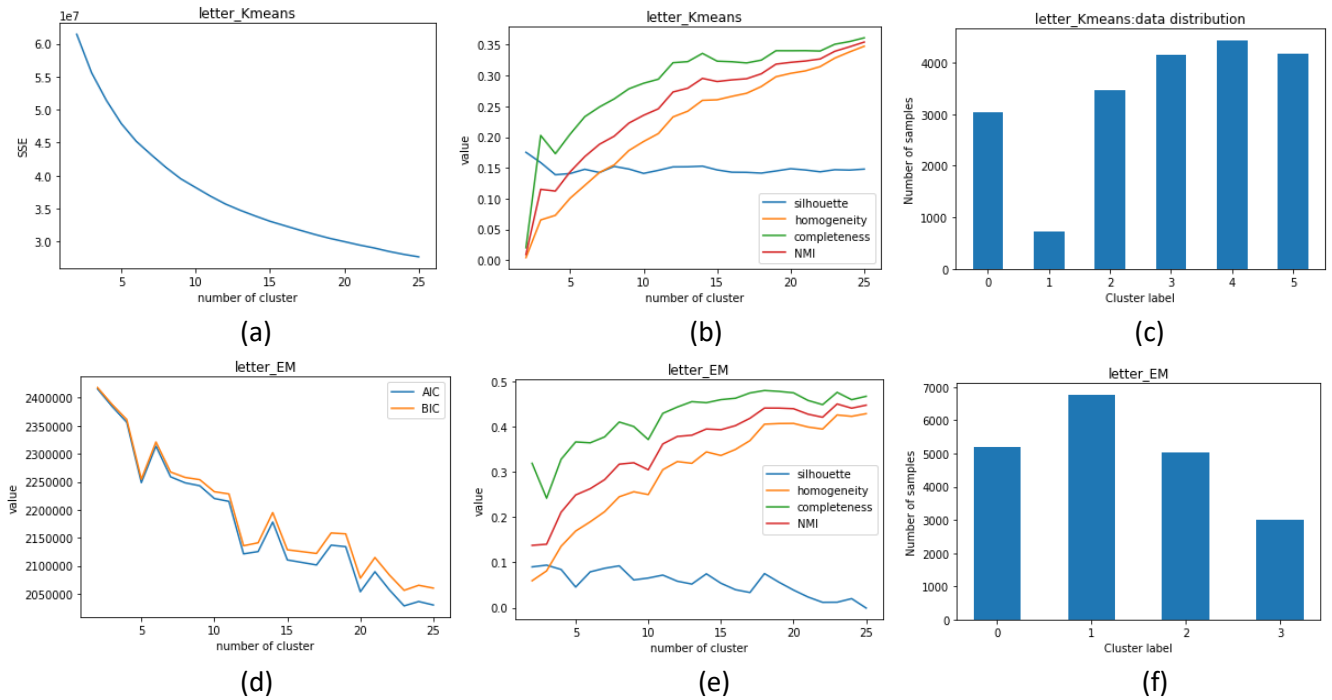


Fig. 1 Clustering for dataset 1 (letter recognition)

For data set 2 (car evaluation), both elbow method and silhouette score suggest that 4 cluster might be a reasonable number of K-means clusters. Clustered Data is visualized as Fig.1(c).

For EM, we select the number of component n=4 given the AIC/BIC curve and sihoutte score, which is same as number of cluster in K-means. Data distribution of different cluster label could be seen as Fig.2(f).

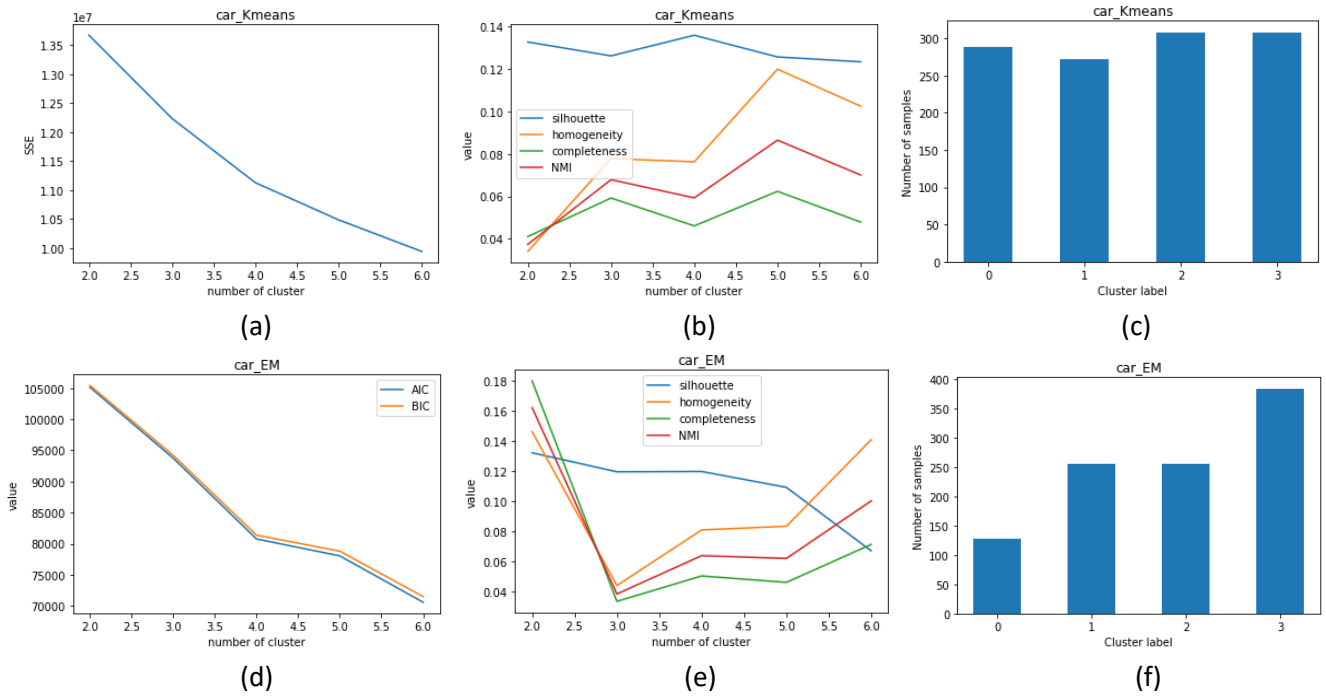


Fig. 2 K-means for dataset 2 (car evaluation)

Run time of K-means and EM is showed as Tab.1. These two cluster algorithms are basically at same level. EM costs a bit of more time when dealing with more data (dataset 1, 20000 data and 16 features).

Tab.1 Run time of clustering algorithms		
	K-means	EM
Dataset 1	222.7s	283.9s
Dataset 2	0.8s	0.4s

Part 2. Dimensionality Reduction:

Principal Component Analysis (PCA)

PCA is a wildely adopted dimension reduction method which finds the orthogonal eigenvectors of data that maxizing the variance. In present study, the components together to reach 70% cumulative explained variance ratio will be selected.

Independent Components Analysis (ICA)

Independent Components Analysis is a classic dimension reduction method for signal processing, which separates a composite signal to subcomponents. The subcomponents are assumed to be non-Gaussian and independent. A common application is the "cocktail party problem" that separates sound resourses of a mixed speech. Kurtosis is used to choose the number of independent components.

Randomized Projections (RP)

Random projection is a very simple alorithm used to reduce the dimensionality of data in Euclidean space. Higer dimensional data is randomly projected to lower subspace to reduce the dimension. To account for the randomness, each results involving RP will be calculated 3 times to give an average value. Sparse Random Projection (SRP) in skleran is used in this study. Reconstruction error is used to choose the number of dimension.

Singular Value Decomposition (SVD)

Singular-value decomposition (SVD) is a factorization of a matrix. By eigen-decomposition of the matrix, it reduces the dimension of matrix. Different from PCA, this method does not center the data before computing the SVD. Again, 70% cumulative variance ratio is used to choose the number of components. TruncatedSVD model in sklearn is used for computation.

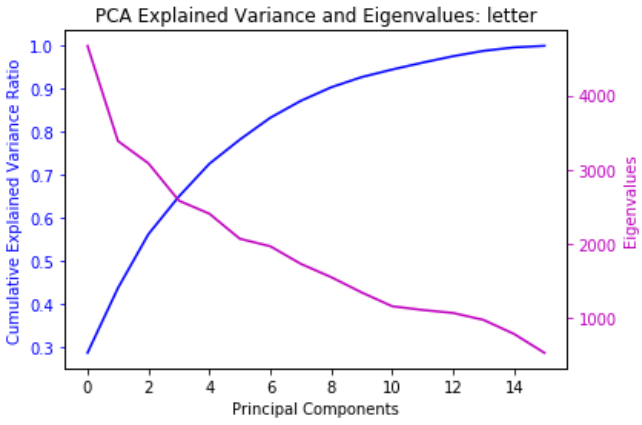
Dataset 1 (letter recognition):

For PCA, as shown in Fig.3(a), to reach 70% variance ratio, 4 principal components is choosed for dimensionality recution and will be used later for clustering.

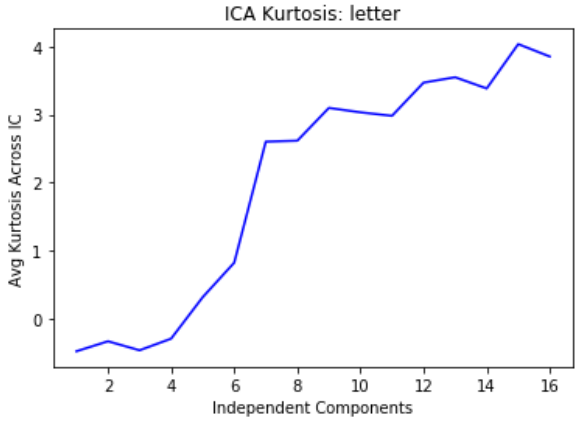
For ICA, as shown in Fig.3(b), more than 7 components results in slowly kurtosis growth. Thus 7 components is choosed for dimensionality recution and will be used later for clustering.

For SRP, as shown in Fig.3(c), considering the change rate of the reconstruction error, 5 dimension is choosed for dimensionality recution and will be used later for clustering.

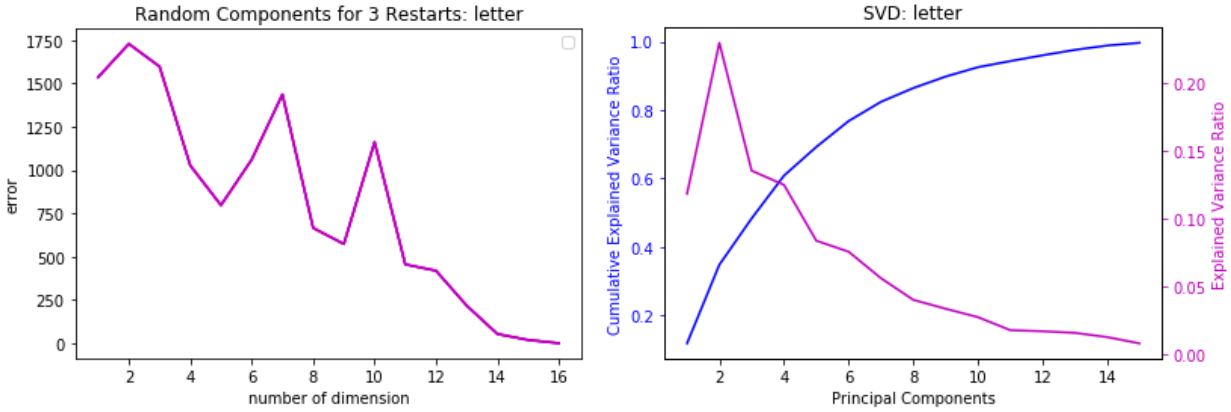
For SVD, as shown in Fig.3(d), to reach 70% variance ratio, 6 principal components is choosed for dimensionality recution and will be used later for clustering.



(a)



(b)



(c) (d)

Fig. 3 Dimensionality reduction for dataset 1 (letter recognition)

Run time of different algorithms for data set 1 (letter recognition) are showed as Tab.1. ICA is the most efficient algorithms among these four.

Tab.2 Run time of dimensionality reduction algorithms				
	PCA	ICA	SRP	SVD
Dataset 1	1.5	0.01	0.3	1.2

Dataset 2 (car evaluation):

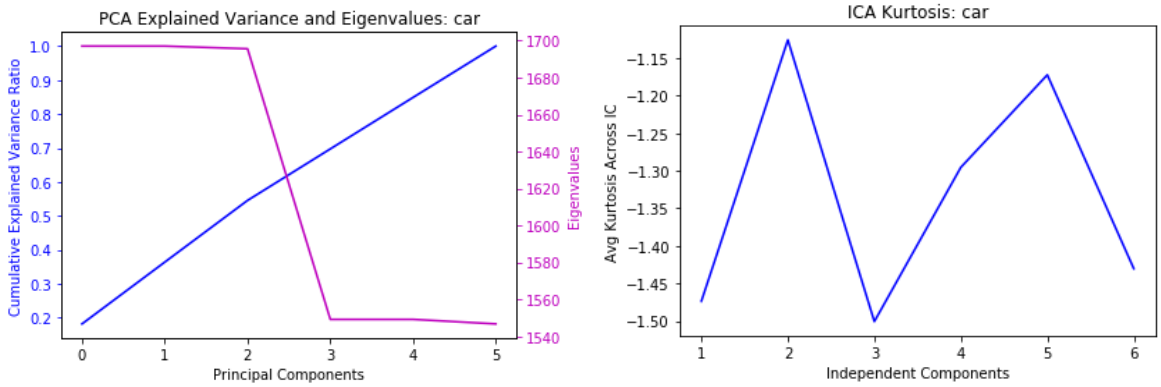
For PCA, as shown in Fig.4(a), to reach 70% variance ratio, 3 principal components is choosed for dimensionality recution and will be used later for clustering.

For ICA, as shown in Fig.4(b), 2 components gives highest kurtosis, thus is choosed for dimensionality recution and will be used later for clustering.

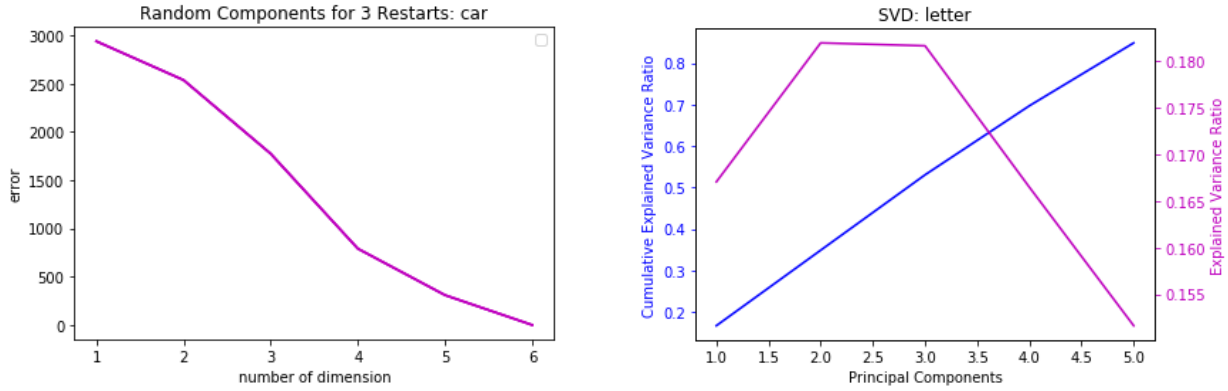
For SRP, as shown in Fig.4(c), considering the change rate of the reconstruction error, 4 dimension is choosed for dimensionality recution and will be used later for clustering.

For SVD, as shown in Fig.4(d), to reach 70% variance ratio, 4 principal components is choosed for dimensionality recution and will be used later for clustering.

Run time is not compared for this samll data set (2000 instances 6 features) because the run time is too short to make a fari comparison.



(a) (b)



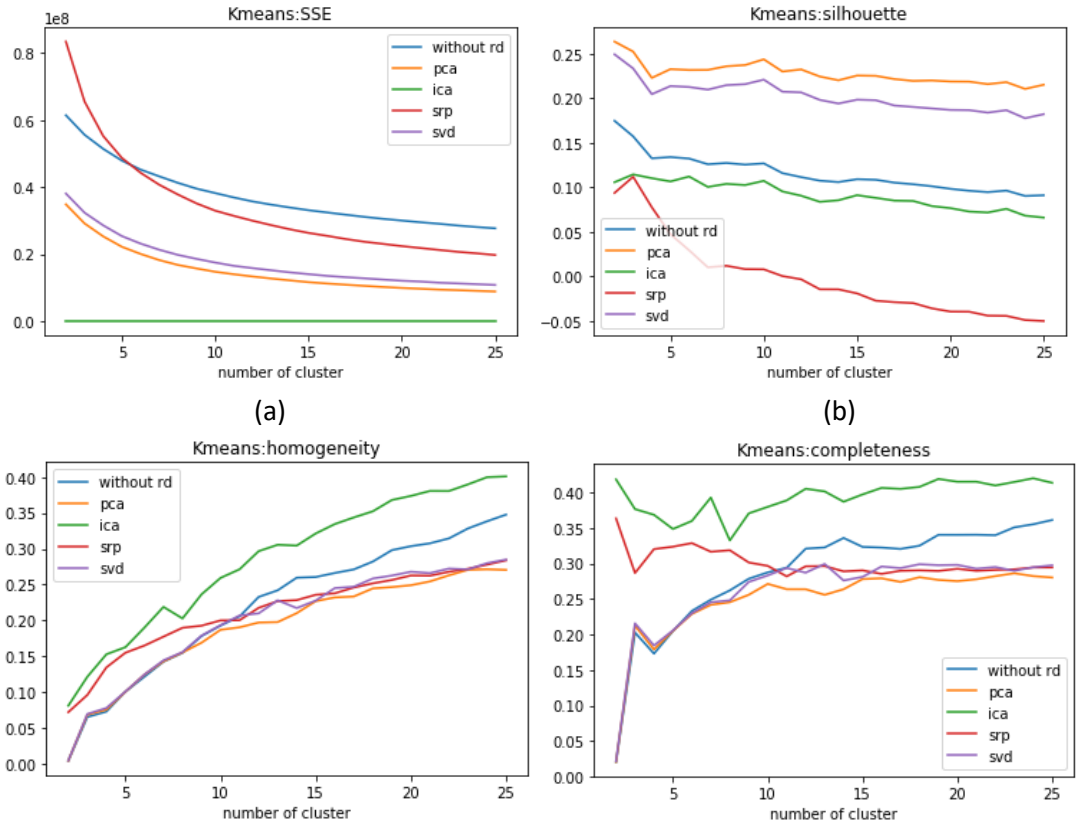
(c) (d)

Fig. 4 Dimensionality reduction for dataset 2 (car evaluation)

Part 3. Clustering with Dimensionality Reduction:

Data set 1: Letter Recognition

For K-means shown as in Fig.5, data is transformed by 4 dimensionality reduction algorithms results in lower SSE. However all SSE curvea are smooth, thus the elbow method is hard to apply. Notice k=10 yields highest sihouette score. Meanwhile the SSE, homogeneity and completeness curve become flatter after 10 cluster. Thus k=10 is choosed for k-means clustering.



(a) (b)

(c) (d)

Fig. 5 K-means with dimensionality reduction for dataset 1 (letter recognition)

For EM shown as in Fig.6, lower BIC value is given by four dimensionality algorithms. Notice k=13 yields highest sihouette score for all clusters. The homogeneity and completeness curve become flatter after 13 cluster. Thus n=13 is choosed for EM component.

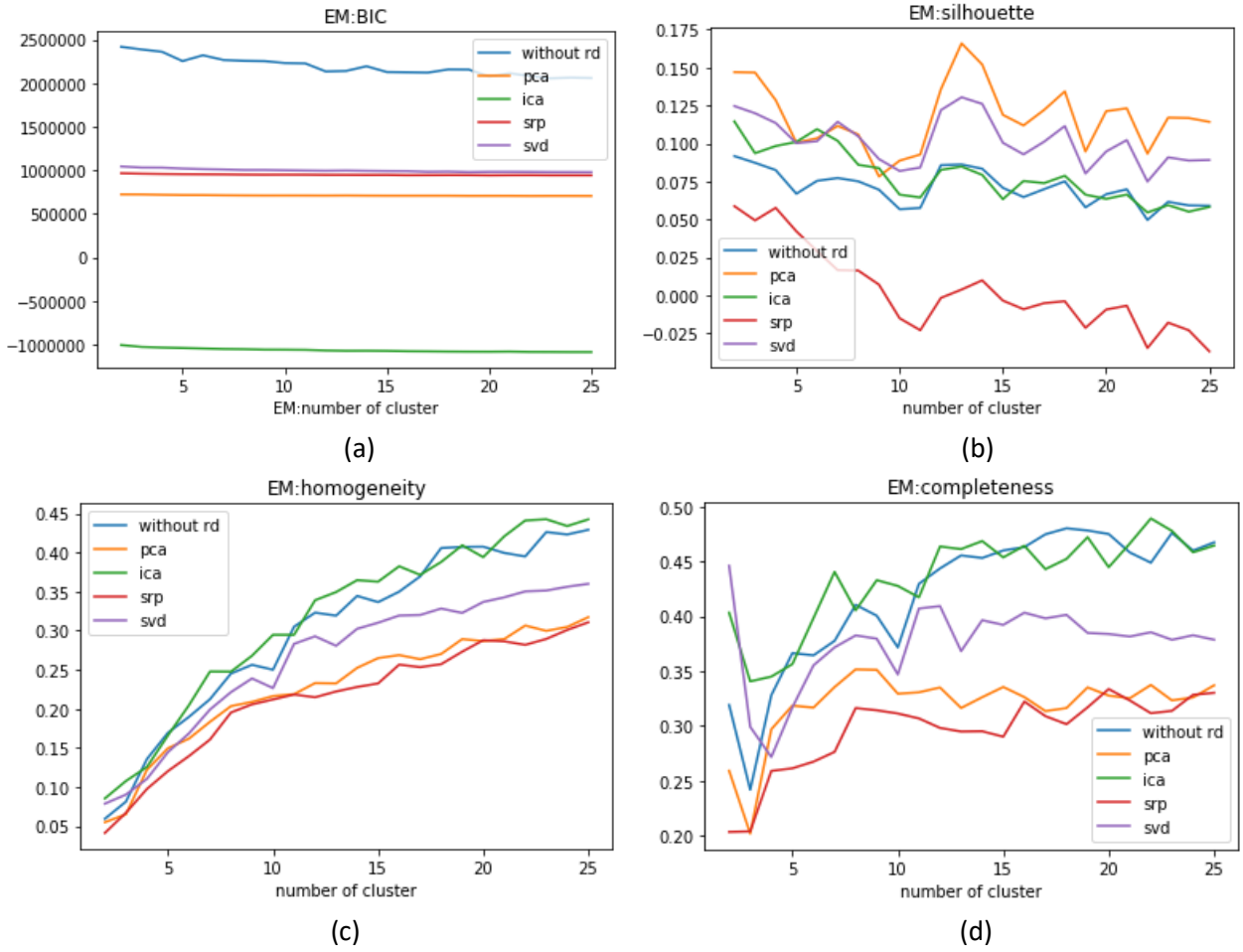


Fig. 6 EM with dimensionality reduction for dataset 1 (letter recognition)

Data set 2: Car Evaluation

Shown as Fig.7, K-means with PCA, ICA and SVD show higher sihouette, homogeneity and completeness than K-means without dimensionality reduction. It shows that PCA, ICA and SVD help clustering the data. However these three scores is lower for K-means with SRP, though 3 simulation are implemented to cancel the randomness. According to the results, k=4 could be an ideal cluster number.

For EM shown as in Fig.8, similar results could be drawn as K-means, that PCA, ICA and SVD help clustering the data, but the SRP is outperformed, even comparing with EM without dimemsionality reduction.

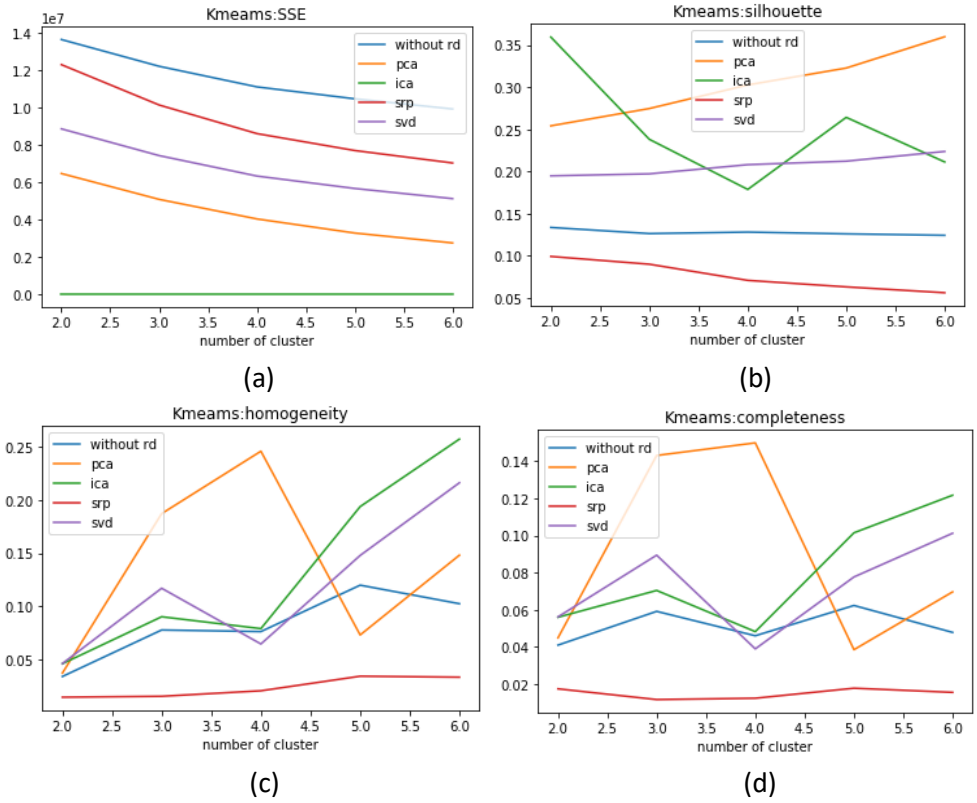


Fig. 7 K-means with dimensionality reduction for dataset 2 (car evaluation)

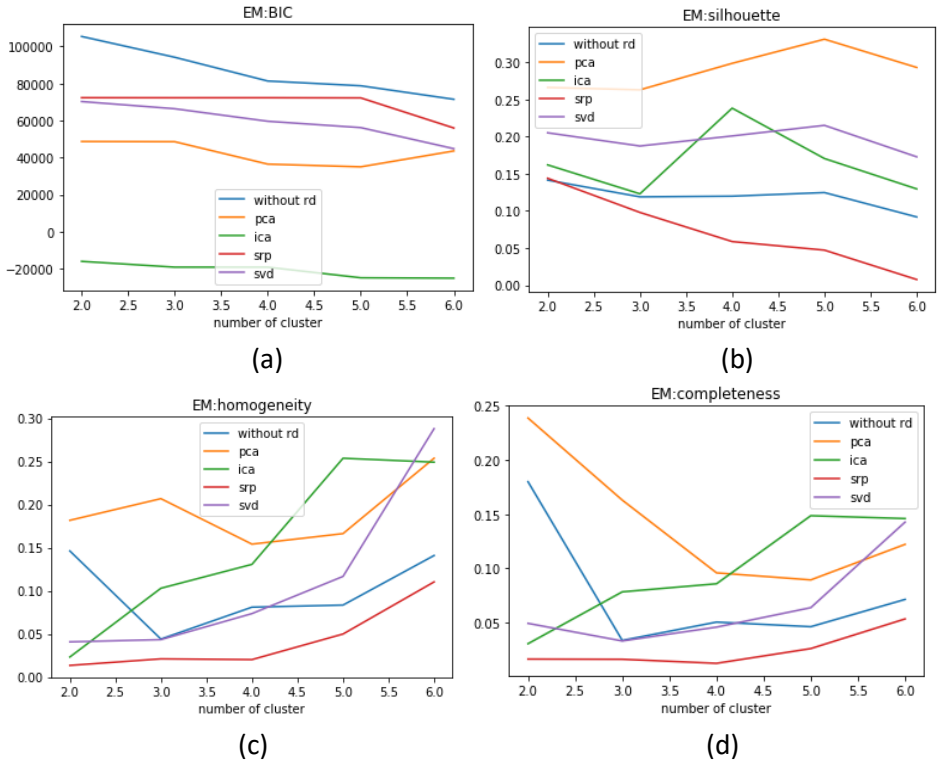


Fig. 8 EM with dimensionality reduction for dataset 2 (car evaluation)

Part 4. Neural Network with Dimensionality Reduction:

Data set 1: Letter Recognition

Comparing to dataset 2, dataset 1 with more instances (20000) and more attributes (16) is used in this part to train neural network. Same NN structure is used as HW1 (26 hidden layers). Data after dimensionality reduction with different algorithms are used to train the NN. As a result shows in Fig.9, it is reseanably that more number of components will results in higher trainning accuracy. However the accuracy grows slowly using more than 6 components, which roughly corresponding to the results in part 2 dimensionality reduction.

Fig.10 shows the NN without dimensionality reduction from HW1. By comparison we could conclude that NN with PCA gives a higher trainning accuracy.

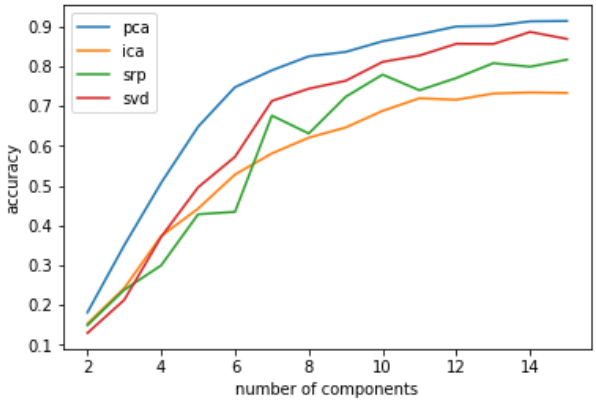


Fig. 9 NN with dimensionality reduction

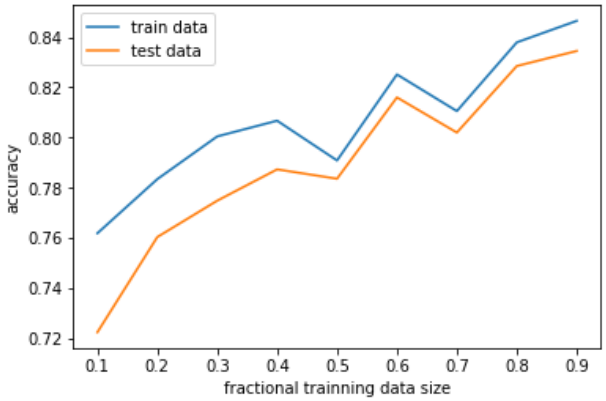


Fig. 10 Original NN from HW1

The trainning time are at same level for different NN. Because data size is same for different NN with same number of components.

Tab.3 Run time of NN with different dimensionality reduction for dataset 1				
	PCA	ICA	SRP	SVD
Run Time /s	149.7	151.1	181.1	145.3

Part 5. Neural Network with Cluster after Dimensionality Reduction:

Data set 1: Letter Recognition

In this apart, the data is dimension reduced by PCA, ICA, SRP and SVD. Then NN is trained with/without clustering algorithms. To make fair comparison, k=10 is choosed for both K-means and EM.

Shown as Fig.11, data with clustering shows high training/test accuracy, which is understandable that K-means/EM label similar data to same cluster, which makes NN easier to train.

Notice that NN with PCA yields highest accuracy. ICA and SVD are good too, worse than PCA but better than SRP. It shows PCA performs well whereas SRP is not so ideal, which coressponding to the results in part 3.

Comparing with original NN in Fig.10, with dimensionality reduction and clustering give higher training accuracy.

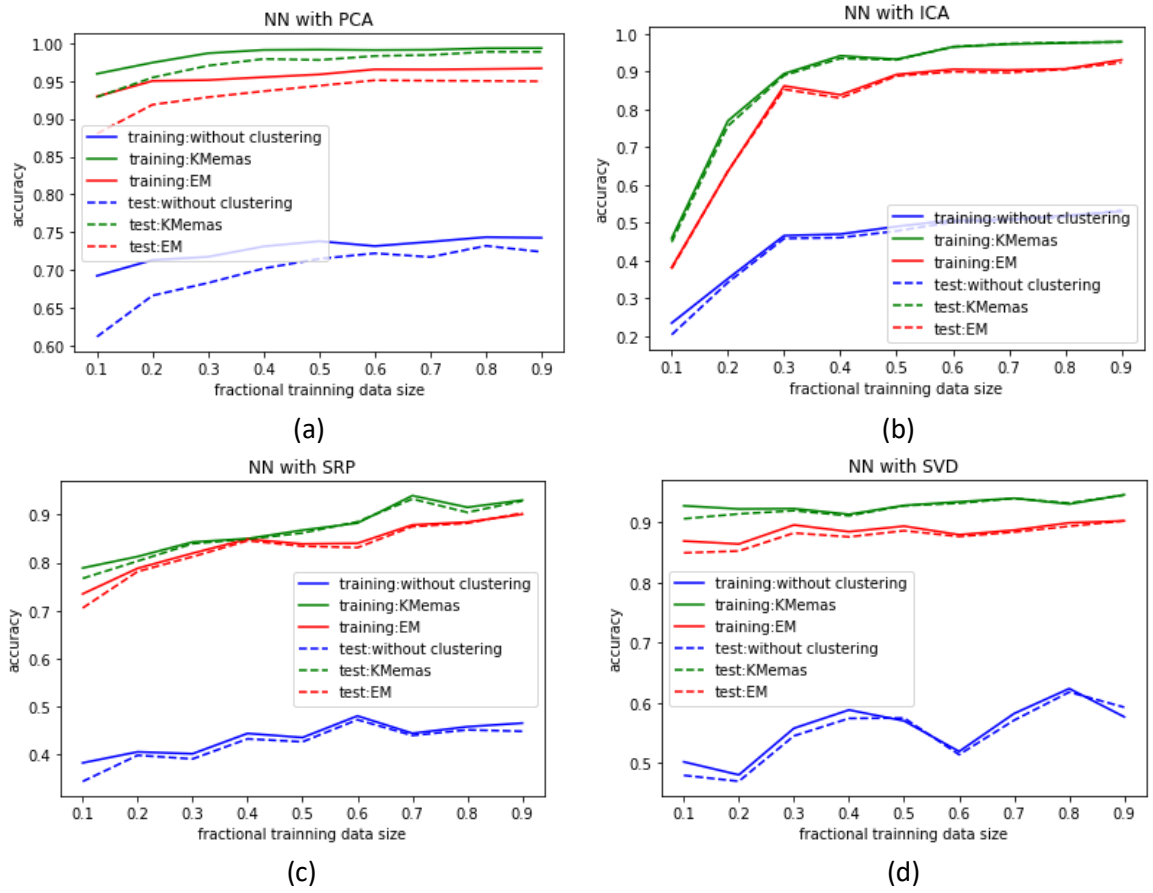


Fig. 11 NN with dimensionality reduction and clustering for dataset 1 (letter recognition)

Summary:

In this report we applied 2 clustering and 4 dimensionality reduction algorithms to 2 different datasets. NN is trained using the data after clustering and/or dimension reduction. Detail remarks could be concluded as:

Clustering: Though based on different algorithm, K-means and EM shows pretty similiar performance, regarding the accuracy and time cost. Besides, after clustering the data, NN could be trained with higher accuracy.

Dimensionality Reduction: Among all those four algrithms, PCA shows best performance helping cluster or training the NN. Whereas SRP is out performed by others.

Reference:

1. https://archive.ics.uci.edu/ml/machine-learning-databases/letter-recognition/

2. https://archive.ics.uci.edu/ml/machine-learning-databases/car/