## CS 7641 Machine Learning Assignment 3

X. Sheldon Gu
xgu60@gatech.edu

### The problems and datasets (also used in Assignment 1)

**1.** Phishing websites attempt to obtain valuable information (e.g. user name, passwords for bank accounts or credit card accounts). Once the criminals behind phishing websites get the information, they either directly withdraw the money from victims' accounts or sell the information to others for profits. The phishing websites can be very similar to the real one, and easily fool people who lack of experience. If we can find unique features of phishing websites and use some algorithm to identify them, we may build some Apps or extensions for web browsers to directly filter phishing websites.

Phishing Websites Data Set from UCI Machine Learning Repository was used. (https://archive.ics.uci.edu/ml/data sets/Phishing+Websites)[1]. The data set has 30 attributes and total 11055 instances (about half the websites are phishing websites).

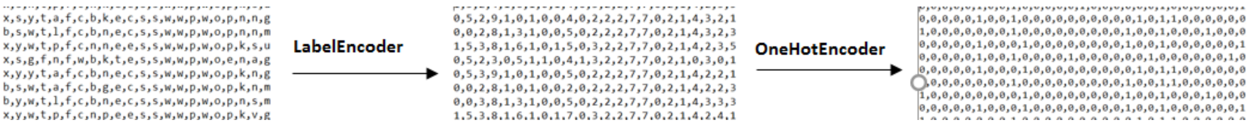
**2.** We all occasionally read sad news that some people and their family members died from eating poisonous mushrooms. The poisonous mushrooms look so close to other edible mushrooms and that leads confusions to even experienced person and causes a series of tragedies. It causes great loss not only to the family but also to the whole society. The goal is to develop a method to identify poisonous mushrooms by features.

Mushroom Data Set from UCI Machine Learning Repository was used. (https://archive.ics.uci.edu/ml/data sets/Mushroom)[2]. The original data set contains 8124 samples of different mushrooms with 22 features (e.g. cap-shape, odor, gill-attachment). About half of these mushrooms are edible versus half of them are poisonous. The data set is quite balance in general.

### The methodology

These two datasets all have categorical features. Scikit implemented LabelEncoder was used to transform categorical features into numerical features, and OneHotEncoder was used for feature binarization (Figure 1).

**Figure 1. the illustration of data preprocessing of mushroom dataset using scikit implemented modules.**



```
x,s,y,t,a,f,c,b,k,e,c,s,s,w,w,p,w,o,p,n,n,g
b,s,w,t,l,f,c,b,n,e,c,s,s,w,w,p,w,o,p,n,n,m
x,y,w,t,p,f,c,n,n,e,e,s,s,w,w,p,w,o,p,k,s,u
x,s,g,f,n,f,w,b,k,t,e,s,s,w,w,p,w,o,e,n,a,g
x,y,y,t,a,f,c,b,n,e,c,s,s,w,w,p,w,o,p,k,n,g
b,s,w,t,a,f,c,b,g,e,c,s,s,w,w,p,w,o,p,k,n,m
b,y,w,t,l,f,c,b,n,e,c,s,s,w,w,p,w,o,p,n,s,m
x,y,w,t,p,f,c,n,p,e,e,s,s,w,w,p,w,o,p,k,v,g
```

```
0,5,2,9,1,0,1,0,0,4,0,2,2,2,7,7,0,2,1,4,3,2,1
0,0,2,8,1,3,1,0,0,5,0,2,2,2,7,7,0,2,1,4,3,2,3
1,5,3,8,1,6,1,0,1,5,0,3,2,2,7,7,0,2,1,4,2,3,5
0,5,2,3,0,5,1,1,0,4,1,3,2,2,7,7,0,2,1,0,3,0,1
0,5,3,9,1,0,1,0,0,5,0,2,2,2,7,7,0,2,1,4,2,2,1
0,0,2,8,1,0,1,0,0,2,0,2,2,2,7,7,0,2,1,4,2,2,3
0,0,3,8,1,3,1,0,0,5,0,2,2,2,7,7,0,2,1,4,3,3,3
1,5,3,8,1,6,1,0,1,7,0,3,2,2,7,7,0,2,1,4,2,4,1
```

```
0,0,0,0,0,1,0,0,1,0,0,0,0,0,0,0,0,0,1,0,1,1,0,0,0,0,0,0
1,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,1,0,0,1,0,0,0,1,0,0,0
0,0,0,0,0,1,0,0,0,1,0,0,0,0,0,0,0,1,0,0,1,0,0,0,0,0,0,1
0,0,0,0,0,1,0,0,1,0,0,0,0,1,0,0,0,0,0,0,1,0,0,0,0,0,1,0
0,0,0,0,0,1,0,0,0,1,0,0,0,0,0,0,0,0,1,0,1,1,0,0,0,0,0,0
[illegible],0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,1,0,0,1,1,0,0,0,0,0,0
1,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,1,0,0,1,0,0,0,1,0,0,0
0,0,0,0,0,1,0,0,0,1,0,0,0,0,0,0,0,1,0,0,1,0,0,0,0,0,0,1
```

Seven other scikit implemented algorithms were used to accomplish this assignment. K-Means and GaussianMixture(EM) were used for clustering; PCA, FastICA, GRP and LDA were used for dimension reductions; Multi-layer Perceptron classifier (MLPClassifier) was used for supervised learning.

All plots were plotted by using python library: matplotlib. (Python files used for preprocessing data, clustering, dimension reductions, neural network training, and plotting are included in the attachment.)

### *1. Run the clustering algorithms on the data sets and describe what you see.*

Two clustering algorithms (K-Means, and Expectation Maximization) were used to generate clusters from phishing websites dataset (Figure 2A) and mushroom dataset (Figure 2B). The original labels of both datasets were removed before the clustering, and later used to evaluate the performance of clustering algorithms. There are multiple clustering performance evaluation metrics, and here five of them are presented. They are Adjusted Rand index (ARS), a function that measures the similarity of predicted labels versus original labels; Adjusted Mutual Information score (AMIS), a function that measures agreement between labels; Homogeneity score (HS), Completeness score (CS) and V-measure score (VMS). Each metric has its own advantages and drawbacks, and the combination of them delivers more accurate performance evaluation.

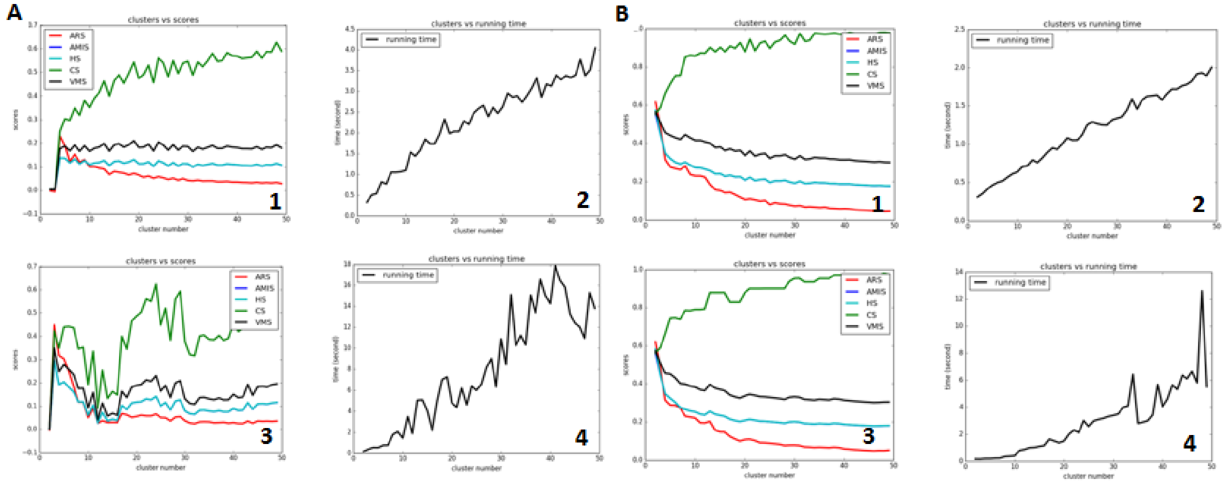


**Figure 2. performance evaluation of clustering algorithms. A:** phishing websites dataset, **1,** scores vs cluster numbers using k-means algorithm, **2,** runtime vs cluster numbers using k-means algorithm, **3,** scores vs cluster numbers using GaussianMixture, **4,** runtime vs cluster numbers using GaussianMixture. **B:** mushroom dataset, **1,** scores vs cluster numbers using k-means algorithm, **2,** runtime vs cluster numbers using k-means algorithm, **3,** scores vs cluster numbers using GaussianMixture, **4,** runtime vs cluster numbers using GaussianMixture.

For phishing dataset (Figure 2A), the performance of both algorithms increase when the cluster number increase from 2 to 4, then four of the metrics scores decrease and plateaued except Completeness score (CS). The running time linearly increases with the cluster number, and the Expectation Maximization algorithm costs much more time compared with K-Means algorithm. For mushroom dataset (Figure 2B), the performance of both algorithms decrease with the increment of cluster number. When the cluster number equals 2, four metrics have best scores except CS score. The running time linearly increases as the cluster number increases.

The clustering result is a surprise to me. Mushroom dataset should have much more clusters (each mushroom species can form a cluster since mushroom samples belong to same species have almost identical features) and phishing website dataset should have much less clusters. The mushroom dataset has 120 binary features, and the phishing websites has 70 binary features. It is hard to visualize data in a 120 dimensional space or a 70 dimensional space, and the dimension reduction algorithms will reduce the dimensions for a better visualization.

## 2. Apply the dimensionality reduction algorithms to the two datasets and describe what you see.

To visualize the datasets, four dimensional reduction algorithms were applied. There are Principal Component Analysis (PCA), Individual Component Analysis (FastICA), Gaussian Randomized Projection (GRP), and Linear Discriminant Analysis (LDA). These four algorithms use different approached to generate new components and reduce the original dimensions. LDA delivers best separation for both phishing website dataset (Figure 3 A4) and mushroom dataset (Figure 3 B4). It is easy to understand since LDA is a "supervised" dimensional reduction algorithm, the true labels were provided during the dimensional reduction. PCA and ICA deliver similar results (Figure 3 A1,2 for phishing websites dataset and Figure 3 B1,2 for mushroom dataset). Since PCA and ICA are different algorithms and PCA chooses components with biggest variance while ICA chooses independent features, the results in Figure 3 means the first two components generated by PCA and ICA coincidently have biggest variance and are independent. The separations by GRP (Figure 3 A3, B3) are much worse, obviously two randomized projections are not enough for better separation of either phishing websites dataset or mushroom dataset.

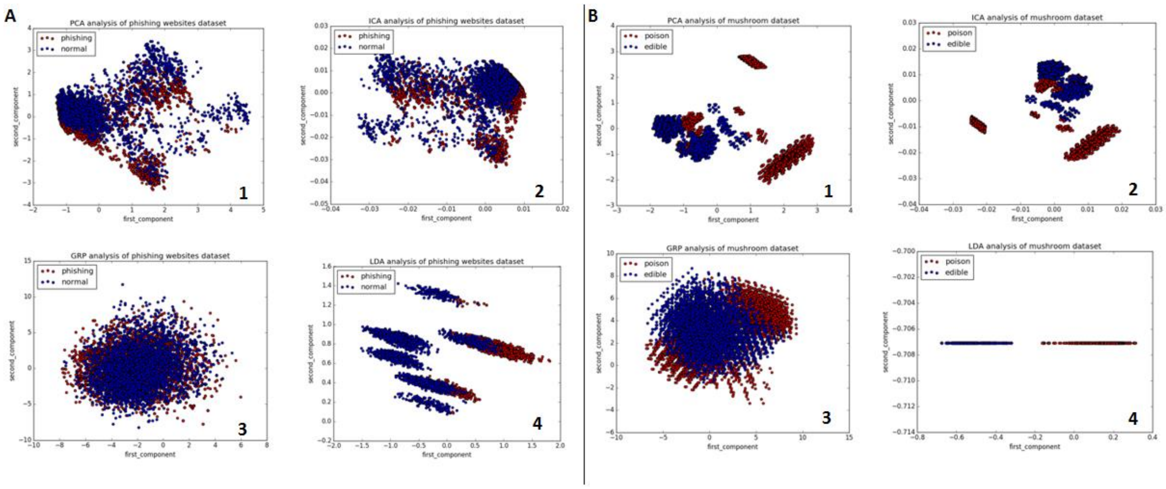


**Figure 3. datasets visualization in 2-d spaces.** Four dimensional reduction algorithms (PCA, FastICA, GRP, LDA) were applied to phishing websites dataset (**A**) and mushroom dataset (**B**). The reduction of features (70->2 for phishing dataset and 120->2 for mushroom dataset) facilitate the direct visualization of the original data. Data points are labeled into different colors (red vs blue) based on their original labels as illustrated in figures. **1,** dimensional reduction by Principal Component Analysis (PCA). **2,** dimensional reduction by Individual Component Analysis (FastICA), **3,** dimensional reduction by Gaussian Randomized Projection (GRP), **4,** dimensional reduction by Linear Discriminant Analysis (LDA).

The 2-d plots deliver direct visualization for these two datasets providing better understanding of the data. The data points from phishing websites are more scattered, which represent the diversity of websites. Different websites carry their unique combination of features. Both PCA and ICA plots show there is a better separation for phishing websites if choosing 3 or 4 cluster centers, which is consistent with results in Figure 2A. The data points from mushroom dataset are much more aggregated, that is because mushroom samples belong to same species have almost identical features. Different species are also aggregated into a few big classes and the poisonous and edible mushrooms are well separated even in the 2-d space. That also explained why it achieved good metrics scores when only use two cluster centers (Figure 2B).

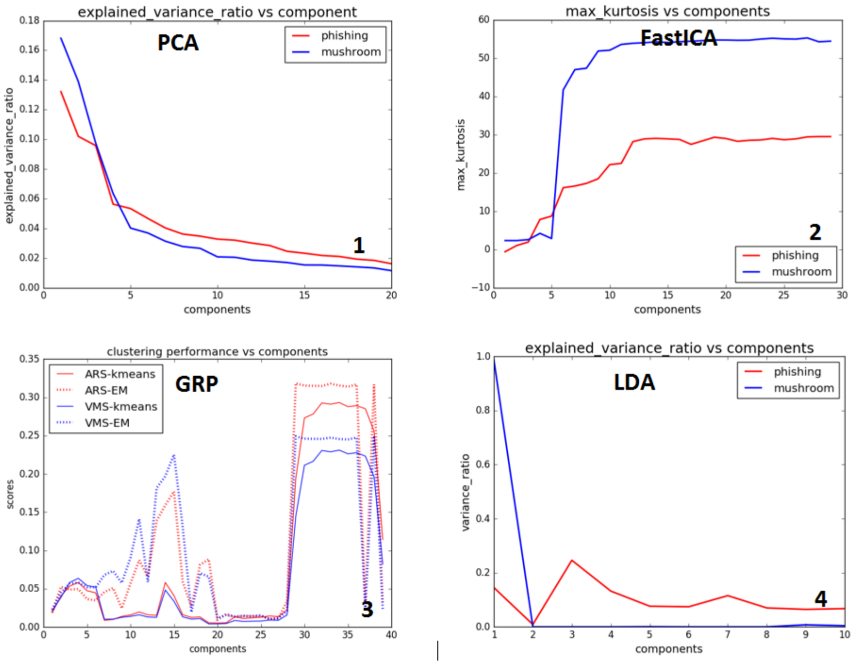


**Figure 4. the detailed analysis of new components obtained through different dimensional reduction algorithms. 1**, the explained_variance_ratio of new components after dimensional reduction using PCA. **2**, the max-kurtosis with the increment of new components after dimensional reduction using FastICA. **3**, the clustering performance (ARS) with the increment of new components after dimensional reduction using GRP. **4**, the explained_variance_ratio of new components after dimensional reduction using LDA.

2-d or 3-d plots can provide direct visualization for data, however, most times only choose two or three components cannot represent the original data well. The detailed analysis of new generated components from dimensional reduction using PCA, ICA, GRP and LDA is presented in Figure 4. Figure 4.1 shows explained_variance_ratio of components generated by PCA for both phishing websites dataset and mushroom dataset. The first component has the highest variance ratio (which is consistent with the definition of PCA) and the value decreased quickly with the number of components. When the component number larger than 5, the variance ratios decrease much slower and almost plateaued. Figure 4.2 shows max-kurtosis when FastICA produced components increase. When the component number is low (<5), the kurtosis of each is small. When the component number increases, new components with much larger kurtosis show up (means these new components may not as independent as the previous components). The performance of GRP is measured by the metrics of clustering (Figure 4.3, detailed discussion in Part 3), and when component number reaches 30, best performance is achieved. Figure 4.4 shows explained_variance_ratio of each component generated by LDA, and obviously LDA only needs small number of components.

### 3. *Reproduce your clustering experiments, but on the data after you've run dimensionality reduction on it.*

Clustering experiments were reproduced on phishing websites datasets (Figure 5.1) and mushroom dataset (Figure 5.2) after using different dimensional reduction algorithms. Since K-Means and GaussianMixture produce similar results in Figure 2, only K-Means is used in this experiment. The clustering performance metric ARS is plotted after using different components generated by PCA (red line), FastICA (blue line), GRP (cyan line) and LDA (green line).

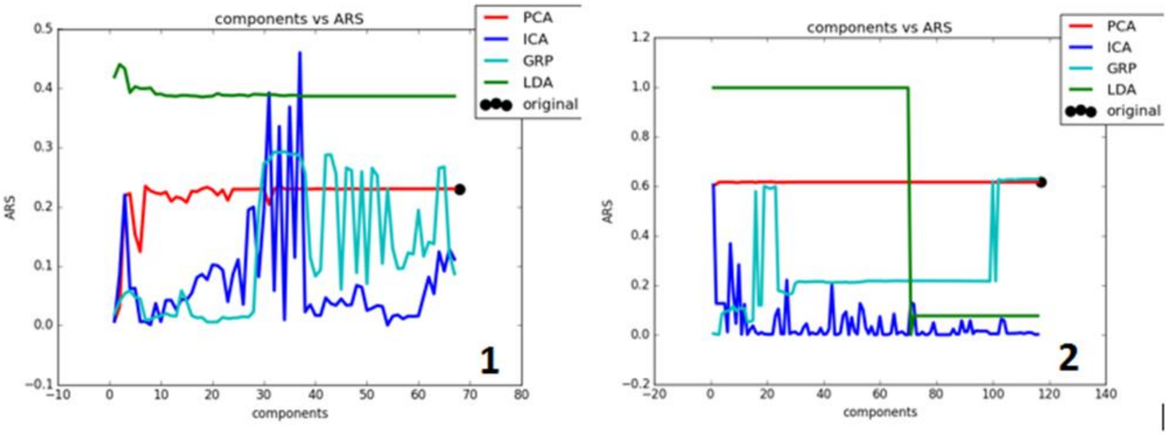


**Figure 5. the performance evaluation of clustering algorithm (K-Means) with increment of components obtained through dimensional reduction using PCA (red), FastICA (blue), GRP (cyan) and LDA (green). 1,** phishing websites dataset. **2,** mushroom dataset. The original ARS scores of both datasets are presented in figures as black dots. The phishing websites dataset chooses four cluster centers, while mushroom dataset only chooses two cluster centers. The selection of clusters is based on results shown in Figure 2.

New dataset generated after LDA dimensional reduction shows best clustering performance with highest ARS much better even compared with the original score using all original features. That maybe due to the “supervised” dimensional reduction process. The clustering performance on PCA dataset increases with the components, but quickly plateaued, which is consistent with the results shown in Figure 4.1 that larger number components have relatedly low variance ratios. The clustering performance on FastICA dataset shows big variations. For phishing websites dataset, the ARS increase with components but quickly decreased. For mushroom dataset, the ARS has it maximum value at the beginning and decreases with the increment of components. It shows big fluctuations on both datasets, especially on phishing websites dataset. The clustering performance on GRP datasets increases with components until enough projects have been reached. The phishing websites dataset needs more than 30 randomized projections, while mushroom datasets only needs less than 20. Another interesting observation is that clustering performance not always increase with the components. Sometimes more components lead to poor performance.

4. ***Apply the dimensionality reduction algorithms to one of your datasets from assignment #1 (if you've reused the datasets from assignment #1 to do experiments 1-3 above then you've already done this) and rerun your neural network learner on the newly projected data.***

To generate accurate learning curves, it is essential to keep the testing data untouched to avoid the mistake “peek in the future”. In my previous assignments, I used 70% original data as training dataset, while the left 30% for testing. In this assignment, four dimensional reduction algorithms (PCA, FastICA, GRP and LDA) with different component numbers (2, 4, 8, 12, 16, 32) were applied to 70% of original data to fit the model, and the left 30% of original data were transformed using the pre-fit model. Even though the true labels were isolated from original dataset for PCA, FastICA, and GRP, it is still better not to use the testing data to fit the models. And it is especially important for LDA, which uses true labels to fit the model. After dimensional reduction, 24 new datasets were generated. Then different proportions (10%-100%) of training data sets were used for training and accuracies of predictions were examined on the same training proportions to get training curves (Figure 6A1-4) and on the testing datasets to get the testing curves (Figure 6B1-4). The original training and testing curves are also plotted as solid black lines for an easy comparison.

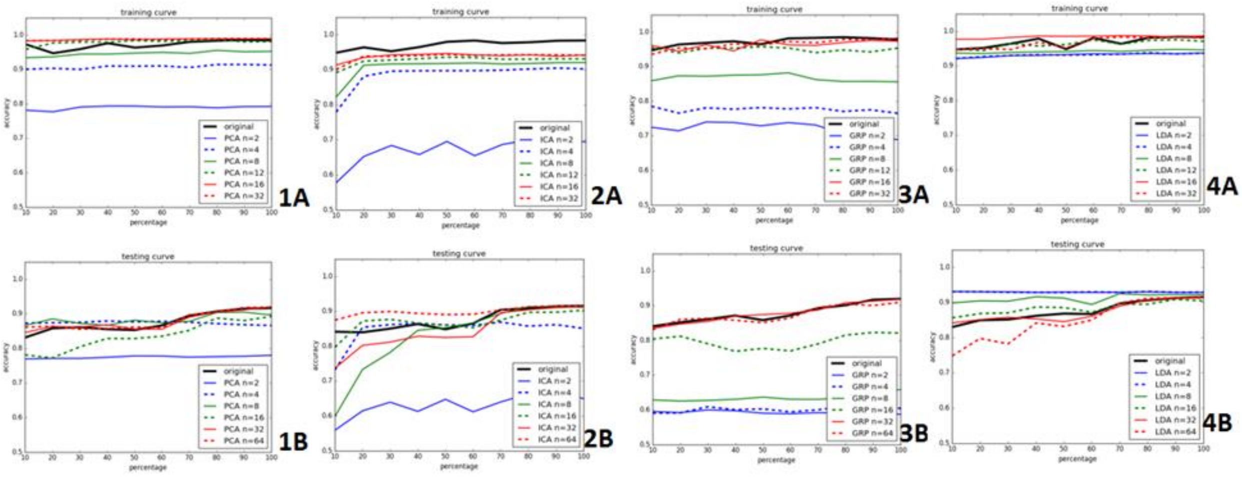


**Figure 6. the learning curves (training and testing) of scikit Multi-layer Perceptron classifier (MLPClassifier) on phishing websites dataset after dimensional reduction using different algorithms. 1,** learning (A) and testing (B) curves after dimensions have been reduce to 2, 4, 8, 16, 32, 64 by PCA. **2,** learning and testing curves after dimensional reduction by FastICA. 3, learning and testing curves after dimensional reduction by GRP**. 4,** learning and testing curves after dimensional reduction by LDA. The learning and testing curves using original features are presented in plots as solid black lines.

For the learning curves for datasets generated by PCA (Figure 6 1A, 1B) when only choose the first two components, both training and testing curves are much lower than the original curves meaning poor prediction and low accuracies. When the datasets have more components, the learning and testing curves are much closer or even better than the original curves. It seems 4 or 8 components are enough to generate same accuracies as original learning curves. These results are consistent with the clustering performance results shown in Figure 5.

The learning curves on datasets generated by FastICA (Figure 6 2A, 2B) are similar to PCA curves. When more components are used, both learning and testing curves are closer and closer to the original curves, but with big variations. Same fluctuations are also seen in Figure 5.

The learning curves on datasets generated by GRP (Figure 6 3A, 3B) show the same pattern, but unlike PCA or FastICA, the approach to original learning curves is much slower and more components are needed. For PCA or FastICA, only use 4 or 8 components are enough to reproduce the same accuracies as original learning curves, while GRP may need 32 components to reproduce the same learning curves.

The learning curves on datasets generated by LDA (Figure 6 4A, 4B) are much better than the original learning curve, especially for the testing curve on dataset using only two components. Since LDA is a "supervised" dimensional reduction algorithm, better training and testing curves are expected. However, there is a big question mark for the real prediction accuracy. Since the generation of training data already used 70% of true labels, even only use 10% of training data for training, actually 100% training data were used.

5. ***Apply the clustering algorithms to the same dataset to which you just applied the dimensionality reduction algorithms (you've probably already done this), treating the clusters as if they were new features. In other words, treat the clustering algorithms as if they were dimensionality reduction algorithms. Again, rerun your neural network learner on the newly projected data.***

This part of experiment is quite complicate and thus a detailed description of my experiment may be essential. The original phishing websites dataset was used, and after applying four dimensional reduction algorithms (PCA, FastICA, GRP and LDA) to reduce final components to 10, 30, 30, 10, respectively, to generate four new datasets. The choosing of specific numbers of components for each algorithm is based on their clustering performance shown in Figure 5 and learning curves shown in Figure 6. For these four new datasets, the clustering algorithm k-means was applied to each of them, where k equals 2, 4, 8, 12 and 16. The predicted labels (total 5 sets of data) are used as new features. Five new datasets are generated by using only feature, or two features or … all five features (1f, 2f, ... 5f respectively). These datasets with category features then go through data preprocess step to be transformed into binary features as described at the beginning of this assignment.

For these new datasets (total twenty datasets), 70% of each dataset were used as training dataset, while the left 30% were used for testing. Then different proportions (10%-100%) of training data sets were used for training and accuracies of predictions were examined on the same training proportions to get training curves (Figure 7A1-4) and on the testing datasets to get the testing curves (Figure 7B1-4). The original training and testing curves are also plotted as solid black lines for an easy comparison.

These learning curves (Figure 7 1-4) are quite similar to each other. The training and testing curves form datasets having more features are always closer to the original learning curve. The differences are how many features needed to reproduce similar accuracies as the original learning curve. Another observation is that unlike normal learning curve that accuracies increase with training percentages, the new learning curves are more flat, that there are minor differences on the training or testing accuracies when 10% or 50% or 100% training data was used. one possible explanation for this may be due to the data process, during dimensional reduction and clustering, the whole information of the dataset was averaged and dispersed into new dataset, thus there is no learning process (more information gain with the increment of sample numbers).

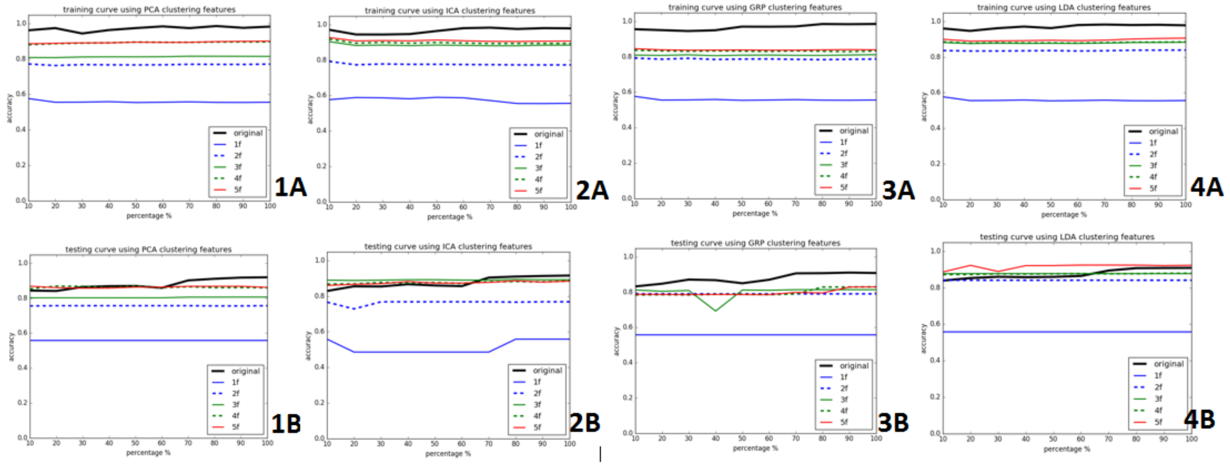


**Figure 7. the learning curves (training and testing) of scikit Multi-layer Perceptron classifier (MLPClassifier) on phishing websites dataset using new features generated through clustering on dimensional reduced data.** After dimensional reduction using PCA, FastICA, GRP and LDA, K-Means clustering algorithm was applied on each new dataset using 2, 4, 8, 12, 16 cluster centers. The predicted labels are used as new features. Five new datasets (1f – 5f) are generated by use only one feature (1f) or two features (2f) or … all five features (5f). The five new datasets also pass through data preprocess steps to be converted into binary features to generate the learning curves.

**Reference:**

[1] Mohammad, Rami, McCluskey, T.L. and Thabtah, Fadi (2012) An Assessment of Features Related to Phishing Websites using an Automated Technique. In: International Conferece For Internet Technology And Secured Transactions. ICITST 2012 . IEEE, London, UK, pp. 492-497. ISBN 978-1-4673-5325-0

[2] Mushroom records drawn from The Audubon Society Field Guide to North American Mushrooms (1981). G. H. Lincoff (Pres.), New York: Alfred A. Knopf