Project: Dimensionality Reduction Impact on Image Classification

Problem Statement:

Classification of images is widely used on the popular MNIST data set, however, what if we consider a different dataset that displays letters of different languages, explore their features and attempt to identify ways of improving classification using nonlinear dimensionality reduction (ISOMAP), linear dimensionality reduction (PCA), and explore the impact of such dimensionality reductions on classification improvements. The other question to explore is which classification technique will perform better with ISOMAP and in general, KNN or SVM?

Our goal is therefore to explore Chinese MNIST datasets for image classification and identify if nonlinear and/or linear dimensionality reduction improves classification

In summary, we have two main tasks that we want to accomplish:

1.Does nonlinear dimensionality reduction improves classification? How does ISOMAP compare to PCA on classification performance?

2. Which classification method will work better with ISOMAP, KNN or SVM?

Data Source:

The data source is provided by kaggle and includes 15,000 chinese characters as images. In addition, there's a csv file that consists of 5 columns. The columns are as follows: suite_id, sample_id, code, value, and character. Figure 1 is a table of the column definitions along with the unique values, and the number of unique values. For this project, we intend to use a subset of the 15,000 images and use the csv file as a supplement to guide and simplify the dataset if need be.

column name	suite_id	sample_id	code	value	character
definition	There are totally 100 suites, each created by a volunteer.	Each volunteer created 10 samples.	Each sample contains characters from 0 to 100M (totally 15 Chinenumber characters). This is a code used to identify	Numerical value of each character.	The actual Chinese character corresponding to one number.
# of unique values	100	10	15	100	15
values	1-100	1,2,3,4,5,6,7,8,9,10	1-15	1-100	九,十,百,千,万,一,二,三,四,六,七,八,零,五,null

Figure 1: Data Map of csv file containing identification values for the chinese characters

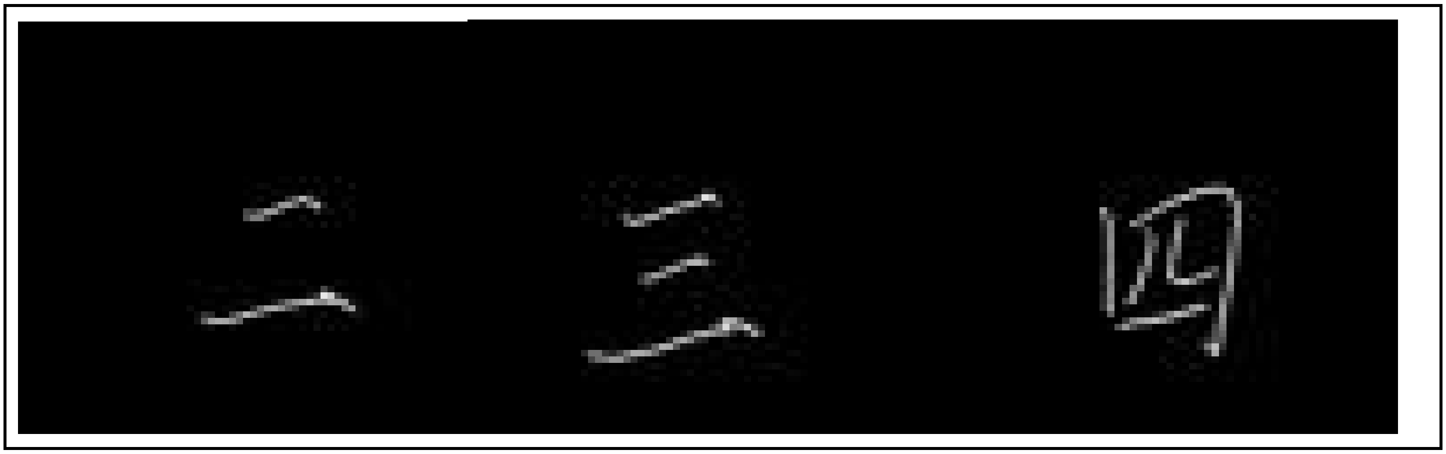
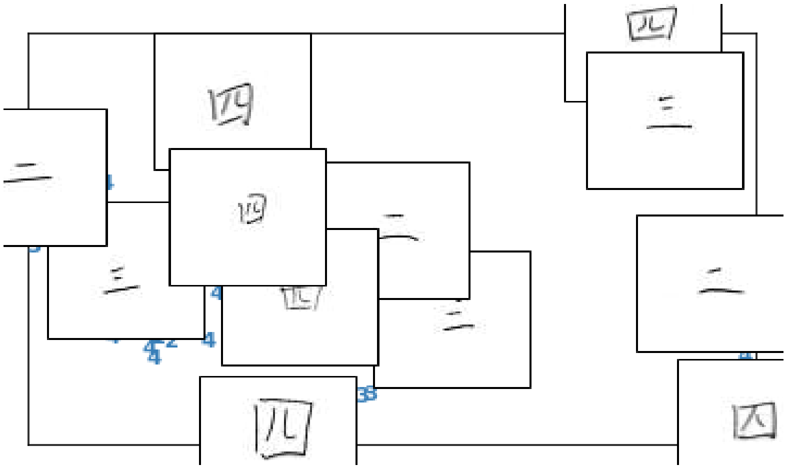


Figure 2: Chosen Chinese characters for analysis ordered from left to right numbers 2,3 and 4.

Methodology:

The images will be examined and a subset of the data set will be extracted consisting of an equal number of images for each of the four unique characters. The sample data will be split into a training and test data set through randomization with 80% being the training set and 20% being the test set. Two dimensionality reduction will be performed on the sample data which are PCA and ISOMAP. There will also be two classification models: KNN and SVM and the goal is to compare dimensionality reduction along with two types of classification models to see which combination yielded the highest accuracy score.

Recall, ISOMAP is a lower-dimensional embedding which maintains geodesic distances between all points. The scatter plot shows the location of the images along the embedded ISOMAP scatter plot. This shows a quite random placement of the images along the 2 dimensional ISOMAP embedding, it will be interesting to see upon tuning it, how it will perform in classification.



Evaluation and Results

There are a total of six different models, including two control groups. The purpose of the control groups is to establish a baseline for comparison with the two data dimensionality reduction techniques: PCA and ISOMAP. This will help determine whether reducing dimensions improve accuracy for the SVM and KNN models. Figure 4 below presents the Misclassification Rate, F-1 score, and the Micro-Average AUC. The F-1 score is the harmonic mean of the precision and recall of each binary classification of a multi-classification model, thus another metric of measuring the accuracy of classification models. It is used here as in this case, false negatives and false positives are more crucial as two of the three chinese characters selected are closely related and thus, could potentially lead to misclassification. The macro-average AUC is simply the mean accuracy scores of all three classes and the misclassification rate is the rate in which the model incorrectly classifies the sample data.

Method	Misclassification Rate	F-1 Score	Macro-Average AUC
SVM w/o Data Reduction	0.13	0.83 0.81 0.98	0.87
SVM w/ PCA	0.12	0.85 0.83 0.98	0.88
SVM w/ ISOMAP	0.59	0.36 0.35 0.49	0.40
KNN w/o Data Reduction	0.30	0.69 0.64 0.80	0.71
KNN w/ PCA	0.21	0.73 0.71 0.93	0.79
KNN w/ ISOMAP	0.38	0.55 0.50 0.81	0.62

Figure 4: Metrics for all six models presented for evaluation and comparison.

Support Vector Machine Models

Based on the results shown in Figure 4, SVM performed exceptionally well with PCA, showing a 88% micro-average accuracy score and relatively high f-1 scores. Various SVM models were performed, specifically tuning the parameters using the GridsearchCV method which will select the optimal combination based on the input parameters. In this case, both linear and kernel SVMs were evaluated along with auto and scaled parameters for gamma and various values for C for all three methods. In comparison with performing SVM without data reduction, there is a slight increase in the metrics used for evaluation. ISOMAP did not perform as well as its counterpart with a 41% accuracy score.

PCA was performed slightly differently. First, PCA was fit to the train data set to determine the number of components that make up for 95% of the variance, which was about 409 components. Then the train and test data set was re-fit using the optimal number of components to then fit into the SVM model. Figure 5 shows the relationship between the first and second principal components. It is apparent that there doesn't seem to be a pattern or cluster of any sort other than a dense focal point to the lower left corner. It is clear that if fitting the data using only the first two components, there will be crucial data lost.

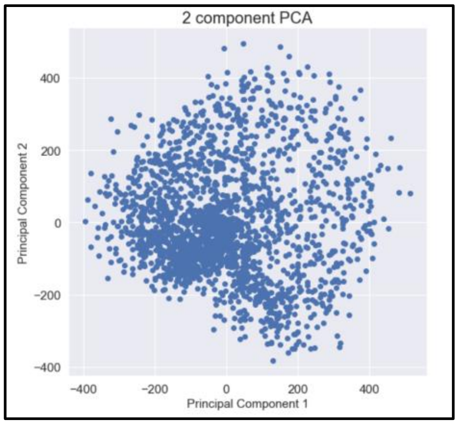
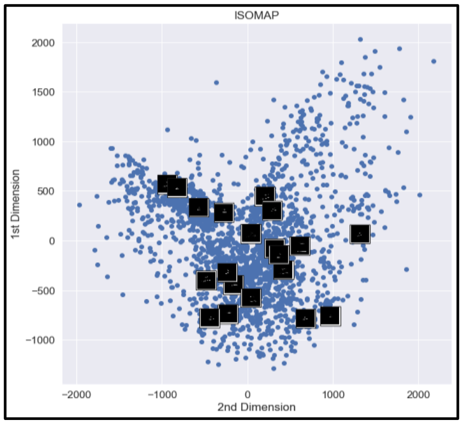


Figure 5&6: PCA Dimensions Plot (Above) ISOMAP Dimensionality Plot (Below)



ISOMAP yielded lower accuracy scores than PCA and the control group and the figure to the right (Figure 6) shows a graph of the first two components with the images embedded. Although the images are not clear, there is some evidence that there doesn't seem to be a clear indication of the location of the images and any clustering based on the similarities within the images selected.

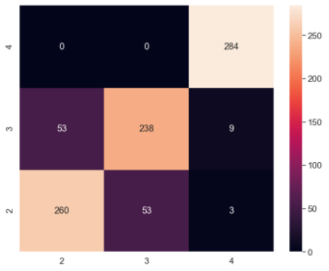
Another vital aspect of comparing models is to look at the confusion matrices in order to analyze the number of false negatives and false positives. Figure 7 shows the confusion matrices for each model along with its respective classification report. Classification for character 4 performed the best in all three models with a good amount of false negatives between character 2 and 3. This is not a surprise considering that character 2 and 3 are very similar. When taking into account false negatives and false positives, we can be confident that SVM with PCA performed the best.

SVM w/o Data Reduction (kernel with gamma = auto and c = 1)

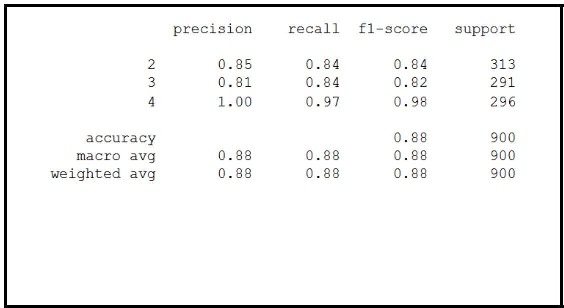
```
              precision    recall  f1-score   support

           2       0.82      0.83      0.83       313
           3       0.79      0.82      0.81       291
           4       1.00      0.96      0.98       296

    accuracy                           0.87       900
   macro avg       0.87      0.87      0.87       900
weighted avg       0.87      0.87      0.87       900
```



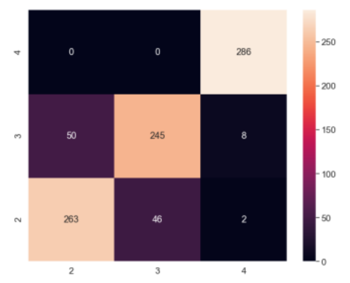
SVM with PCA (kernel with gamma = auto and c = 10)



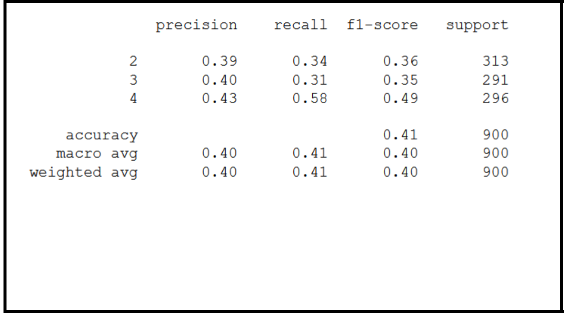
```
              precision    recall  f1-score   support

           2       0.85      0.84      0.84       313
           3       0.81      0.84      0.82       291
           4       1.00      0.97      0.98       296

    accuracy                           0.88       900
   macro avg       0.88      0.88      0.88       900
weighted avg       0.88      0.88      0.88       900
```



SVM with ISOMAP (kernel with gamma = auto and c = 1)



```
              precision    recall  f1-score   support

           2       0.39      0.34      0.36       313
           3       0.40      0.31      0.35       291
           4       0.43      0.58      0.49       296

    accuracy                           0.41       900
   macro avg       0.40      0.41      0.40       900
weighted avg       0.40      0.41      0.40       900
```

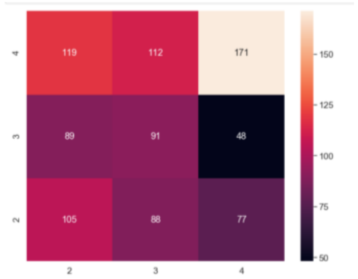


Figure 7: Classification Report and Confusion Matrices for the three SVM models.

K-Nearest Neighbor Models

As we can see from figure 4, we get similar results where The KNN model was presented with poorer classification results when presented with no PCA reduction. However, when combined with PCA, it performed better. The micro average and f1 score both compared higher lifts than no dimensionality reduction. The KNN model parameters were tuned using the GridSearchCV method, and data was scaled accordingly. The KNN model without PCA performed worse as we can see from the f1 score and accuracy scores. ISOMAP performed worse than no data reduction which was interesting as it goes to suggest that a strong nonlinearity between variables/features may not exist.

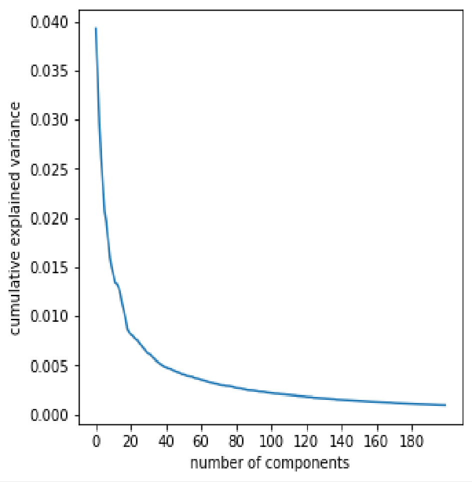
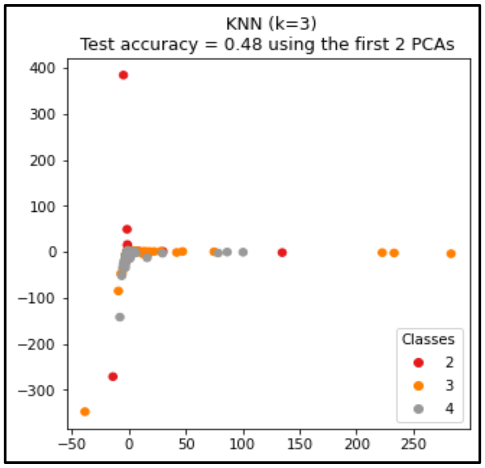


Figure 8 & 9: Explained variance for various # of components (top) & Testing Accuracy for KNN with PCA (bottom)



The PCA for KNN model was tuned differently, the first 50 PCs were chosen as the variation explained by the PCs declined significantly beyond 50 dimensions as we can see above in figure 8. The transformation was done on the training data initially and then when fitting the model with test data, we saw that improvement in model accuracy score and macro average AUC.

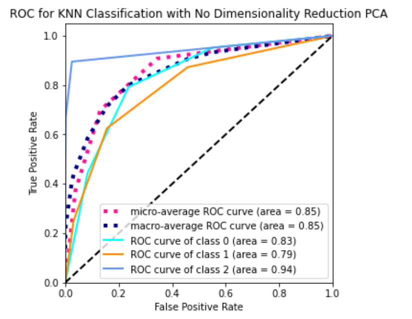
In figure 9, we are showing KNN with only 2 PCs, where the accuracy of the model was 0.48 which is very poor. Interestingly, we see that class 4 or number 4 is clustered and more distinct than class 2 or 3 which is intuitive given numbers 2 and 3 have very similar images. However,once we applied KNN with the tuned 50 PCs, the testing accuracy jumped to 0.79 as we can see in figure 10 below.

KNN w/o Data Reduction

```
              precision    recall  f1-score   support

           2       0.81      0.60      0.69       424
           3       0.63      0.66      0.64       277
           4       0.67      0.99      0.80       199

    accuracy                           0.70       900
   macro avg       0.70      0.75      0.71       900
weighted avg       0.72      0.70      0.70       900
```

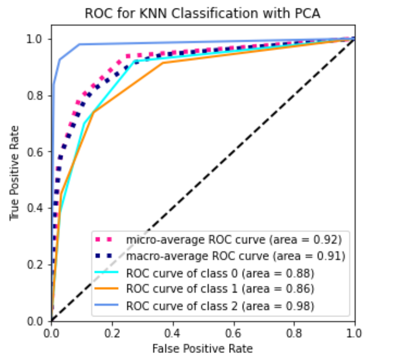


KNN with PCA

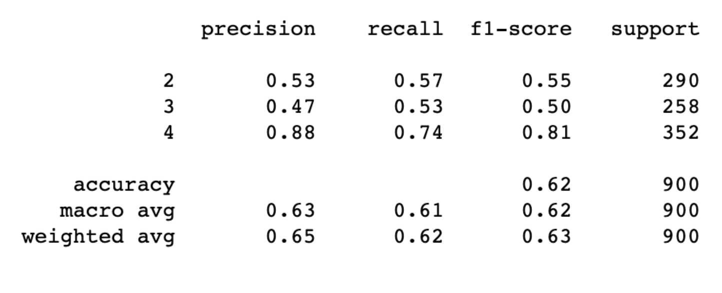
```
              precision    recall  f1-score   support

           2       0.73      0.73      0.73       311
           3       0.72      0.71      0.71       297
           4       0.93      0.94      0.94       292

    accuracy                           0.79       900
   macro avg       0.79      0.79      0.79       900
weighted avg       0.79      0.79      0.79       900
```



KNN with ISOMAP



```
              precision    recall  f1-score   support

           2       0.53      0.57      0.55       290
           3       0.47      0.53      0.50       258
           4       0.88      0.74      0.81       352

    accuracy                           0.62       900
   macro avg       0.63      0.61      0.62       900
weighted avg       0.65      0.62      0.63       900
```

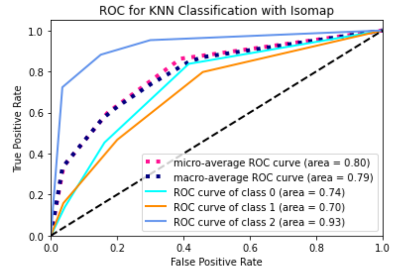


Figure 10: ROC Curves & Classification Reports for KNN models.

Similar to PCA, the ISOMAP transformation was done on the training data initially and then when fitting the model with test data, we saw a decline in model accuracy score and macro average AUC overall. The KNN model with ISOMAP was not able to classify the two numbers 2 and 3 any better than PCA even though we were hoping to see that ISOMAP may have captured non linear feature relationships that could improve on distinguishing images of digit two vs images of digit three. But the precision and recall for both numbers declined significantly compared to PCA which doesn't seem to support any non linear relationships.

Conclusion

By comparing different dimensions of data, PCA yielded the best results for both KNN and SVM. Taking a step back, we can also answer the question of which model performed best for classifying chinese characters and the conclusion is that SVM had higher accuracy scores in all two out of the three approaches. Theoretically, ISOMAP would expect to perform best for data reduction in non-linear relationships however, PCA performed better most likely because the data did not follow an S-shaped curve. This was the case with the input data which is exemplified by the poor performance of ISOMAP as a pre-processing step in both the KNN and SVM models.