kNN and k-means Application in Written Digit Classification

Yu Jung Yeh (Sandy Yeh) and Yi Xiao
Department of Mathematics and Statistics, San Jose State University

Introduction

The primary goal of pattern classification is supervised or unsupervised classification. The most popular method is the k-nearest neighbors(kNN) classifier[1] and k-means classifier which the basic algorithms have already developed in pass 50 years[2]. Principal Component Analysis (PCA)[3] and Fisher Discriminant Analysis (FDA) [4] are two popular algorithms to achieve dimensionality reduction seek and exploit the inherent structure in the data, but in an unsupervised manner to classify data using less information. This can be useful to visualize data or to simplify data which can then be used in a supervised learning method. As the result, we show the most popular data classification algorithms and dimension reduction algorithms apply on MINIST written digit data.

Objective

- Compare the test error and key points (k) of kNN and local k-means classifier.
- Apply PCA and FDA on MINIST data.
- Apply weighted to give closer neighbors more voting weights.
- Find the optimal combination.

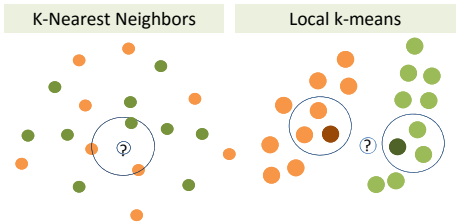

Figure 1. kNN and k-means classifier.
- K-nearest neighbors algorithm is a non-parametric method used for classification and regression.
- Local kmeans classifier is use the k nearest points in each class to represent the corresponding.

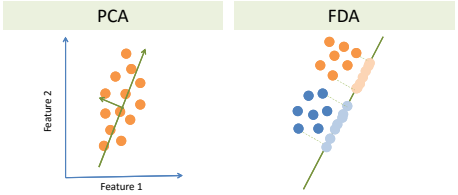

Figure 2. PCA and FDA dimensionality reduction.
- PCA is an unsupervised dimensionality reduction approach.
- FDA is a supervised dimensionality reduction approach

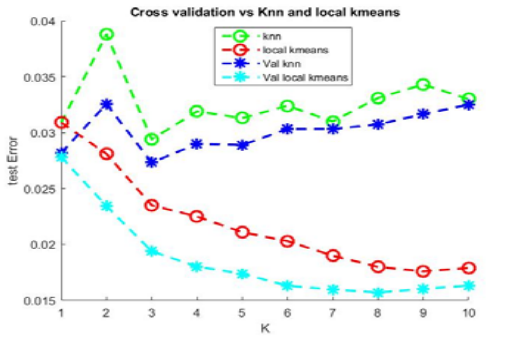

Figure 3. Cross Validation-kNN, kNN, Cross Validation-local kmeans and k-means Classifier.
- Determining the optimal key point (k) by the smallest test error.

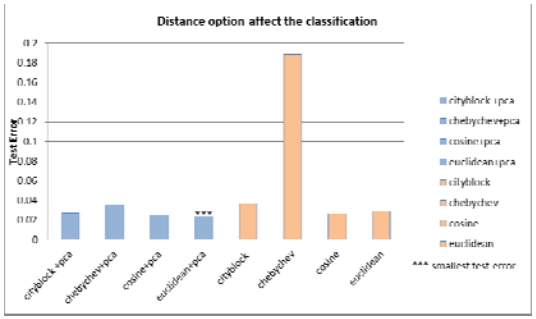

Figure 4. kNN classifier with different distance + PCA.
- The Euclidean distance gives a smallest test error.

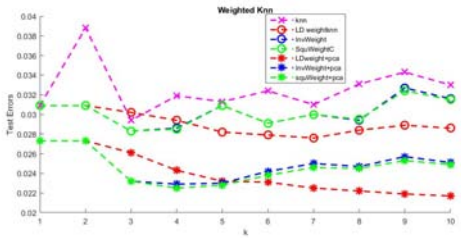

Figure 5. kNN classifier + Weighted.
- The closer neighbors more voting weights.
- The Common weights including linearly decaying weights, inverse weights, and squared inverse weights.
- Weighted kNN classifier gives smaller test error.

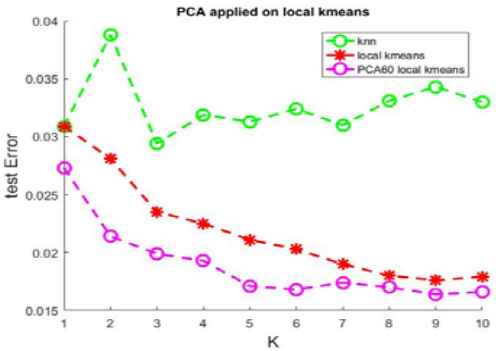

Figure 6. local kmeans classifier + PCA.
- The efficiency is increased and less time-consuming after apply PCA60 with k-means classifier.

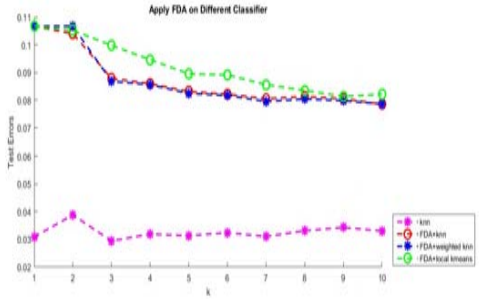

Figure 7. FDA + kNN/ weighted kNN/ local k-means Classifier.
- The efficiency is significantly increase after apply FDA but the accuracy is decreased.

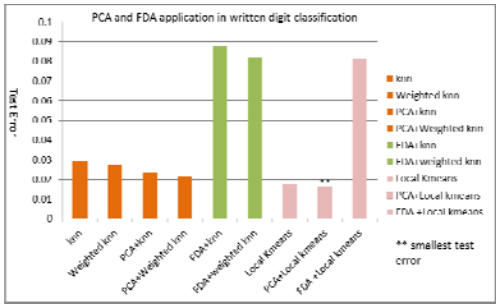

Figure 8. Results Summary.
- In kNN classifier, the PCA60 + weighted kNN has highest accuracy.
- In local k-means classifier, the PCA60 + local k-means has the highest accuracy.

Figure 9. Error Images of PCA + local kmeans application on MINIST data set.

Conclusion

- Both kNN and local k-means classifier are powerful tool in data classification.
- The k-means classifier has higher accuracy (smaller test error) but less efficiency (time consuming) then kNN classifier.
- In kNN classifier the Euclidean distance gives a smaller test error, especially, with PCA dimension reduction.
- Weighted kNN increase the accuracy especially linearly decaying weight in this study. The combination of weighted kNN + PCA can both increase efficiency and accuracy.
- PCA unsupervised dimensionality reduction approach is powerful tool on remove the 'noise' data and increase the efficiency and accuracy.
- Although FDA is such powerful tool on increase efficiency but the accuracy is sacrificed too much.
- Overall, the PCA + local kmeans combination would be suggested in this study.

Reference

1.T.Cover and P.Hart. Nearest neighbor pattern classification. IEEE Transactions in Information Theory, IT-13, 21-27, 1967.
2. S.Z. Selim and M.A. Ismail. K-Means-Type Algorithms: A Generalized Converagence Theorem and Characterization of Local Optimality. IEEE Transactions on Pattern Analysis and machine Intelligence. Vol6-1, 81-87, 1984.
3.J.A.K. Suykens, T. Van Gestel, J. Vandewalle, B. De Moor. A Support Vector machine Formulation to PCA analysis and its kernel Version. IEEE Transactions on Neural Networks. Vol 14-2, 447-450, 2003.
4. T.Connie, A.T.B. Jin, M.G.K.Ong, D. N. C. Ling. An Automated Palmprint Recognition System. Image and Vision Computing. Vol23-5, 501-515, 2005.

This project was supported by Dr. Guangliang Chen at San Jose State University