Face Verification using SVM and Dimension Reduction using PCA

1st Harold Mansilla
Department of Physical Sciences and Mathematics
University of the Philippines, Manila
Manila, Philippines
hrmansilla@up.edu.ph

2nd Virgilio M. Mendoza III
Department of Physical Sciences and Mathematics
University of the Philippines, Manila
Manila, Philippines
vmmendoza1@up.edu.ph

***Abstract*—This paper discusses PCA, a dimension reduction tool, and its effects on an SVM classifier, specificallly, the effect of the number of principle components used during classification. The results obtained from the study indicate that increasing the number of principle components and thereby increasing the number of total variance accounted for, increases the accuracy of the classifier, albeit a small amount.**

***Index Terms*—PCA, SVM, face verification, image classification, dimension reduction**

I. Introduction

A. Face Verification

Suppose a system, given a picture of a face of a person, must determine whether or not the said person belongs to a predefined group of people who are allowed to do actions within the system. One way to implement this is to have a list of authorized users' images and upon encountering a user, analyze image of the user's face and determine whether the face matches one of the images of the authorized users. This problem is known as *face verification*.

B. Principal Component Analysis (PCA)

Principal Component Analysis, PCA, is a simple, non-parametric method for extracting relevant information from confusing data sets; that is, it is a dimension-reduction tool that can be used to reduce a large set of variables to a smaller set of components that contains most of the information. It is also capable of quantifying the importance of each component by using the measurement of variance along each component. [3].

C. Support Vector Machine (SVM)

The support vector algorithm finds a hyperplane in an n-dimensional space, where n is the number of features, that directly classifies the data points [4]. Data points that fall on either side of the hyperplane can be said to belong to different classes. Data points that are close to the hyperplane that can influence its position and orientation are called support vectors. Using these points, a hyperplane who separates the classes with the maximum margin between classes is formed.

II. Data Set

The data set used in this paper is from the Labeled Faces in the Wild, a database containing 13,000 images of faces from the web [2]. Each face in the data set has been labeled with the name of the person in the picture. In particular, the set used from the database is the ”funneled” images.

III. Dimension Reduction

To implement dimension reduction on the data set, the PCA of `sklean` was used [1]. An important parameter for PCA is `n_components` which will determine the number of principle components to be used by PCA. As mentioned earlier, PCA is a tool that allows us to quantify the importance of each component by measuring its variance. By using the `explained_variance_ratio` in PCA, it can be seen which components encompass majority of the variance of the data set. Referring to the values yielded by this, the parameter `n_components` was set to 150 to encompass about 90% of the variance of the data set. Increasing the components past 300 only yielded in an increase in variance by more or less 0.001%. This increase was deemed negligible for the cost of increasing the principle components.

IV. Results

Here the effects of changing the number principal components on the accuracy of the SVM classifier is discussed. For the sake of this study, the `svm` from `sklearn` will be used. The parameters for the classifier will be `gamma` = 0.001 and `kernel` = rbf.

TABLE I
Accuracy, precision and F1-scores of the SVM classifier with varying n_components

n_components	Total Variance	Accuracy	Precision	F1-score
1	15.9512%	56%	57%	53%
50	77.1162%	61%	63%	58%
100	85.4328%	60%	62%	55%
150	89.6392%	59%	63%	53%
200	92.2349%	60%	63%	53%

The table above shows the result of the study. The classifier gained an accuracy of $60 \pm 1\%$ with the exception of `n_components` = 1 earning 56%. The same is true for precision with a score of 62% and 63% with `n_components` = 1 being at 57%. Lastly, the F1-scores of were 53% and 55% with `n_components` = 50 gaining 58%. From these results, it can be said that the effect of dimension reduction has a small impact on the overall metrics of the classifier.

References

[1] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research, vol. 12, pp. 28252830, 2011.

[2] Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. University of Massachusetts, Amherst, Technical Report 07-49, October, 2007.

[3] Shlens, J. (2014). A tutorial on principal component analysis. arXiv preprint arXiv:1404.1100.

[4] R. Gandhi, Support vector machineintroduction to machine learning algorithms, 06 2018.