ISSN 2616-5775 Vol. 4, Issue 6: 78-81, DOI: 10.25236/AJCIS.2021.040613

Handwriting digit classification using PCA-transformed image features

Yingqiang Yuan

Washington State University 910 NE Providence CT Emerald Down C103 Pullman WA 99163

Abstract: *Machine learning model is capable of achieving an incredible great accuracy by training a large amount of data with many dimensions of features in it. To accomplish such a goal, we are currently in desperate need of powerful computers with CPUs (Computation Processing Unit) and GPUs (Graphic Processing Unit) that are competent of the workload of the training process of those machine learning models. Due to the problem that there are multiple disparate computationally intensive tasks that are lengthy period of time during training phase of machine learning, devices like smartphone and personal computer cannot achieve great efficiency during the training phase of machine learning. Instead of enhancing computational performance of devices and improve the machine learning algorithm, another more feasible way to bypass such problem is to apply dimensionality reduction methods on raw data in hope of reducing the number of features that are passing into the machine learning model during the training phase. By doing so, we can shrink the originally large number of features to several dimensions that is accomplishable for our device to compute and meanwhile, sufficient enough to have a model with acceptable accuracy trained. The purpose of this study to explore and investigate the potential benefits and disadvantages of dimensionality reduction on multiple distinct machine learning algorithm by comparing the performances of each machine learning algorithms when passing the features of raw datasets in and when passing the dataset with the features that survived after dimensionality reduction. The performance will be measured by completing a typical machine learning task: handwriting digit classification. Among all the dimensionalities methods that have been discovered, a traditional and conventional dimensionality reduction method, PCA (Principal Components Analysis), is selected. Such method has been proven successful and efficient in generating the set of data with lower dimension features and apply on an image classification task using several supervised machine learning methods, including KNN (k-nearest neighbors), SVM (Support vector machine) and CNN (Convolutional Neural Network).*

Keywords: *Dimensionality Reduction, PCA (Principle Components Analysis), KNN (k-nearest neighbors), SVM (support vector machine), CNN(Convolutional Neural Network)*

1. Data

A typical image classification dataset, MNIST database [1], is used in this study due to its quality. At the same time, this very data set selected is capable of getting a model with great accuracy trained under different algorithms. This dataset has total number of 60000 examples in training set and 10000 examples in testing set with 10 categories. Each example has 784 dimensions that are generated from a 28x28 gray image. This huge number of 784 dimensions could not be more perfect for this study. In order to achieve great efficiency for this study, a subset of MNIST database is extracted. First 2000 instances are obtained as the training set of this study, while instances from 2001st to 2500th are grouped as the testing set. This way of partition the dataset into training data and testing data is to exploit the effect of PCA (Principal Components Analysis), which is the goal of this study.

2. Feature Selection

Feature selection is the critical task that needs to be done on the data before the training of the machining learning model. Within this study, the task of feature selection is performed by PCA (Principal Components Analysis). The method uses an orthogonal transformation method to convert the dataset with raw features into the dataset with lower dimension features of linearly uncorrelated variables called principal components. Firstly, normalization is applied to balance each dimension:

ISSN 2616-5775 Vol. 4, Issue 6: 78-81, DOI: 10.25236/AJCIS.2021.040613

$x_j^{(i)} = \frac{x_j^{(i)} - \mu_j}{s_j}$. Secondly, covariance matrix is calculated: $Cov = \frac{1}{m}\sum_{i=1}^{m}(x^{(i)})(x^{(i)})^T = \frac{1}{m}X^TX$. Thirdly, SVD (Singular value decomposition) will be applied for the sake of obtaining eigenvectors (U) of the covariance matrix. $(U,S,V^T) = \mathbf{SVD}(Cov)$Then, the first k eigenvectors from U will be selected and distinguished as the principal components: $U_{pc} = (u^{(1)},u^{(2)},...,u^{(k)})$. Here the number of k will be selected differently for different machine learning algorithms. Finally, new features for the data are computed by $z^{(i)} = U_{pc}^T x^{(i)}$.

For the dataset that is being used for this study, we apply PCA (Principal components analysis) and compute top k principal components for k from 1 to 784. The detail of the variance of principle components is showed in Figure.1.

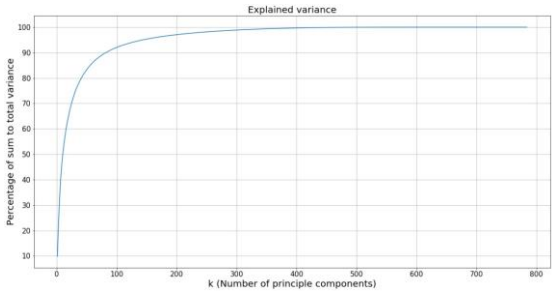

Figure 1. Explained variance of principle components

According to the result of explained variance, principal components analysis can explain most of variance, which is much more than 99% when k is larger than 400. Therefore, it is only appropriate to select different k ranging from 1 to 400 and generate a group of features with lower dimension transformed by corresponding principle components. These features can be used in the following classification tasks applying different machine learning algorithms.

3. Classification Methods

Three classification models generated by different machine learning algorithms, which are KNN (k-nearest neighbor**s)** [2], SVM **(**support vector machine) [3] and CNN **(**Convolutional Neural Network**)** [4]) with various diversified settings are trained on the training set of generated PCA features and tested on the remaining 500 examples. CNN learns convolutional kernels that can extract different types of features from raw image data. A typical architecture of Convolutional Neural Network is showed in Figure 2.

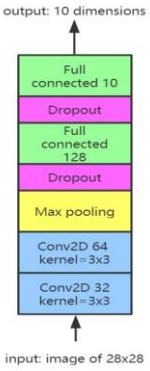

Figure 2. Network architecture of CNN

4. Results

For the model of KNN (k-nearest neighbor**s)**, different numbers of neighbors as: 1, 5, 20, 50 and 100 is selected to see the performance difference. Number of principle components are selected as: 1~10, 20, 30, …, 100, 200, 300 and 400. Instead of PCA features, the raw data is also trained for each number of neighbors as a baseline. The results of model accuracy on different testing conditions are

ISSN 2616-5775 Vol. 4, Issue 6: 78-81, DOI: 10.25236/AJCIS.2021.040613

showed in Figure 3.

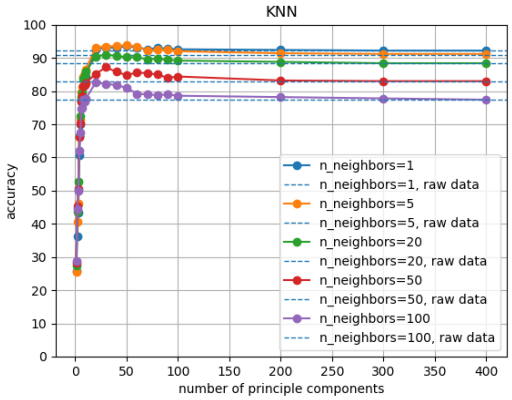

Figure 3. Model accuracy of KNN on testing set

For the model of SVM **(**support vector machine), two kernels are selected: linear kernel and RBF (Radial Basis Function) kernel. Number of principle components are selected as: 1~10, 20, 30, …, 100, 200, 300 and 400. Besides PCA features, the raw data is also trained for both Linear and RBF kernel as a baseline, respectively. The results of model accuracy on testing set are showed in Figure 4.

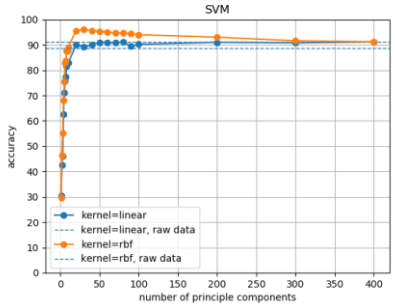

Figure 4. Model accuracy of SVM on testing set

For the model of CNN, since features of input image must be a perfect square number, we select number of principle components: 36, 49, 64, 81, 100, 121, 144, 169, 196, 225, 256, 289, 324, 361, 400. The raw image data (size 28 x 28) is also trained as a baseline. Each configuration spends 200 epochs training on the 2000 examples. The results of model accuracy on testing set are showed in Figure 5. The reason that here the number of principle components is selected from 36, which is 6*6, to 400 which is 20*20 is because it is the proper range for the number of dimensionality reduction,

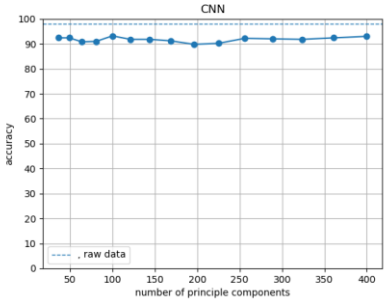

Figure 5. Model accuracy of CNN on testing set

5. Conclusions

This study successfully investigates how reduced features done by PCA (Principal Components Analysis) could affect different machine learning model performance in a typical machine learning task: image classification task. The results suggests that similar performance between PCA-transformed

ISSN 2616-5775 Vol. 4, Issue 6: 78-81, DOI: 10.25236/AJCIS.2021.040613

features and raw features, even though PCA-transformed features have less dimensions. For machine learning algorithms KNN (k-nearest neighbor**s)** ands **(**support vector machine), better performance when passing PCA-transformed features in compared to baseline also indicates that PCA-transformed features of low dimension can improve accuracy by removing dimensions with only noise and pointless information, to extent with certain level. However, local features of image may be discarded by PCA, which leads to worse performance for models like CNN. Besides, larger datasets for image classification should be tested on the conclusions raised in this study and later study.

References

[1] Yanbin Zheng,Hongxu Yun,Fu Wang,Yong Ding,Yongzhong Huang,Wenfen Liu. Defence Against Adversarial Attacks Using Clustering Algorithm[A]. ICPCSEE Steering Committee.Abstracts of the 5th International Conference of Pioneering Computer Scientists,Engineers and Educators(ICPCSEE 2019)Part I[C].ICPCSEE Steering Committee

[2] Aissam Jadli, Hain Mustapha, Chergui Adil. Handwritten Documents Validation Using Pattern Recognition and Transfer Learning[J]. International Journal of Web - Based Learning and Teaching Technologies, 2022, 17(5):1-13.

[3] Wang Yu-Chun, Chuang Chia-Min, Wu Chun-Kai, et al. Cross-language article linking with deep neural network based paragraph encoding[J]. Computer Speech & Language, 2022, 72