Reducing Data Storage Requirements for Machine Learning Algorithms using Principle Component Analysis

Saritha Kinkiri & Wim J.C. Melis

Faculty of Engineering and Science, University of Greenwich,
Chatham Maritime ME4 4TB, UK
k.sarithareddy20@gmail.com, Wim.J.C.Melis@greenwich.ac.uk

Abstract

While current computers have shown to be particular useful for arithmetic and logic implementations, their accuracy and efficiency for applications such as e.g. face, object and speech recognition, are not that impressive, especially when compared to what the human brain can do. Machine learning algorithms have been useful, especially for these type of applications, as they operate in a similar way to the human brain, by learning the data provided and storing it for future recognition. Until now, there has been a strong focus on developing the process of data storage and retrieval, merely neglecting the value of the provided information and the amount of data required to store. Hence, currently all information provided is stored, because it is difficult for the machine to decide which information needs to be stored. Consequently, large amounts of data are stored, which then affects the processing of the data. Thus, this paper investigates the opportunity to reduce data storage through the use of differentiation and combine it with an existing similarity detection algorithm. The differentiation is achieved through the use of, Principal Component Analysis (PCA), which not only reduces the data storage requirements by about 80%, but also improves the overall detection accuracy around 50 to nearly 80%.

Keywords: Machine Learning, Data Storage Efficiency, Principle Component Analysis (PCA)

Introduction

Human beings are very creative thinkers which helps them to overcome obstacles, however their operational abilities are restricted in time. For example, humans are only able to work for a limited number of hours per day [1], while machines can be programmed to operate continuously, and this is where machines have proven particularly useful to us, humans.

Replacing humans by machines generally reduces costs and can also improve time and quality of the job to be delivered. However, machines cannot yet deliver everything we humans can, and this applies particularly when human intelligence is required. It is in this context that machine learning is being developed and it has shown to be particular useful in contexts, such as: image, object and speech recognition along with several other applications [2]. More specifically, machine learning trains computers to recognize data patterns and adjusts itself when there are any changes, resulting in a system that more closely resembles human intelligence. With regards to machine learning, there are many different technologies available, which requires one to select one of the available algorithms.

Generally, these algorithms can be classified as either supervised or unsupervised. In supervised learning [3], the user stores the data in the machine to predict the output values based on previous experiences. The machine does not store any information that is not used and hence requires less storage space to be able to predict the output value [4]. In the case of unsupervised learning, the machine has to self-learn and check all possible solutions constantly. Therefore, it needs to store all required data to be able to compare any newly incoming data. Consequently, more space to store data is required.

The chosen algorithm also influences the storage principles used during training as well as the actual recognition phase. These algorithms are obviously challenged through the fact that they need to recognize e.g. the same face under various different lighting conditions, make-up, facial expressions, etc. In order to deal with these challenges, one either has to restrict oneself towards specific features, and/or add as much contextual information on top of that to deal with these various environmental factors. The extraction of features [5], is not always that straight forward, and often only works for particular types of objects, which means that more often than not large amounts of information are stored to cover as many cases as possible in order to achieve good recognition.

The large amount of data being stored, does not only affect the actual storage requirements, but also the processing, as during the recognition, one needs to work through all of this data to identify possible similarity before a final decision can be made. However, currently, there is a limited amount of work with regards to reducing the amount of data being stored, which is the main focus within this paper.

In reducing the required amount of storage, one could obviously improve the machine learning algorithm, although it would be essential to make a better judgement on which data is useful, and which is not. Considering that this is context dependent [6], this can be quite challenging [7] and so the approach used in this paper is to look at adding an extra stage before the machine learning algorithm, which identifies differences through Principle Component Analysis, leading to the full system as shown in Figure 1.

The remainder of the paper will look into the model as it was developed, and that more specifically for a face recognition example. The results of the various stages of the PCA, as well as the combination with the machine learning algorithm will then be discussed, after which the paper will conclude and present some future work.

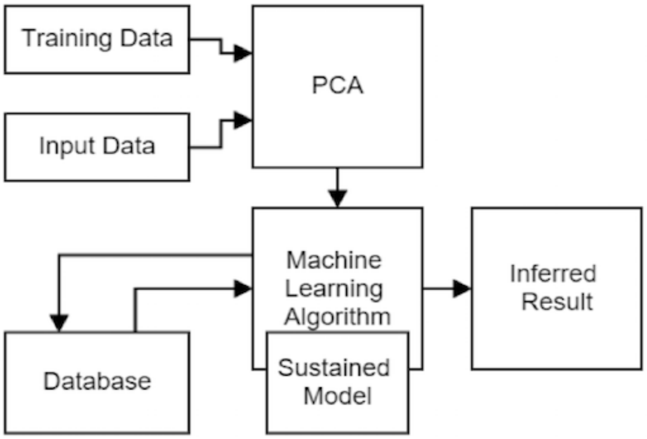

Figure 1: Learning System with PCA

Model Description

In order to reduce the storage requirements, and due to the importance of data being context dependent, it is important to be able to derive context as if one can identify context then it becomes possible to store the data with regards to the context in which you operate. This would then result in a reduction in data storage requirements, because the context is stored as generic information, while for each item only the deviation from this "central" context needs to be stored. To identify these contexts and improve on data storage efficiency, this paper investigates the use of Principle Component Analysis (PCA) to reduce the data storage requirements for a machine-learning algorithm that is used for face recognition. PCA summates large data sets by creating new vectors, called principle components that are a linear combination of the original data, which results in a reduction of the data's dimensionality. Therefore, PCA helps to reduce redundancy, filters noise in the data and compresses the data [8].

In order to achieve this compression, PCA takes data that is correlated, and identifies what one could call a "lowest common denominator". All other information is then stored as a difference from this "lowest common denominator" which significantly reduces the amount of information that needs to be stored.

To demonstrate a reduction in dimensionality that can be achieved through PCA, the algorithm was applied to 25 images from the AT&T face database [9] shown in Figure 2. The images were 112x92 each and converted into a single vector of 10304x1, this was then arranged into a matrix with each column of the matrix being an image. Consequently, the resulted matrix, *A*, has dimension of 10304 by 25.

The first step is to calculate the mean of the set of images, as shown in Figure 3. This mean image represents the common data shared by the set of images and is extracted by considering all input images.

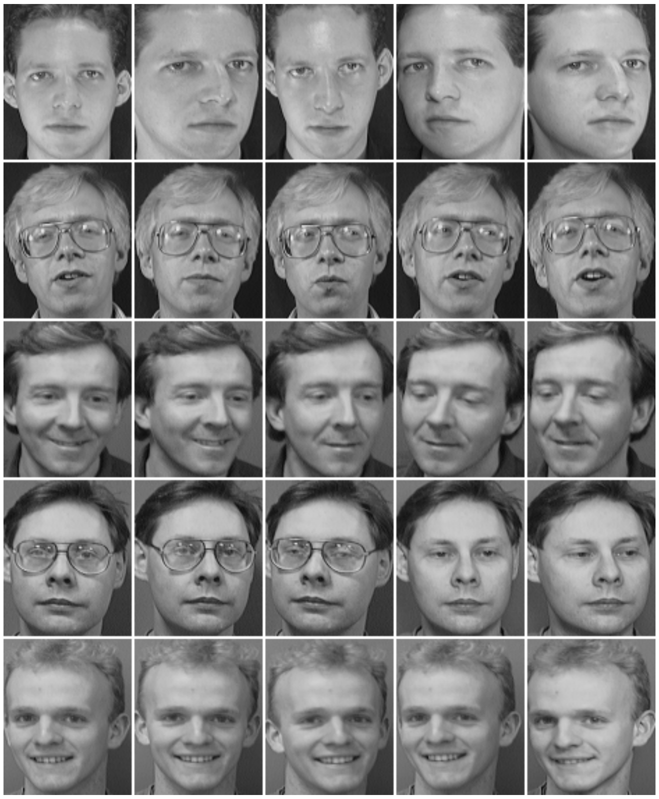

Figure 2: Training Images taken from AT&T Database [9]

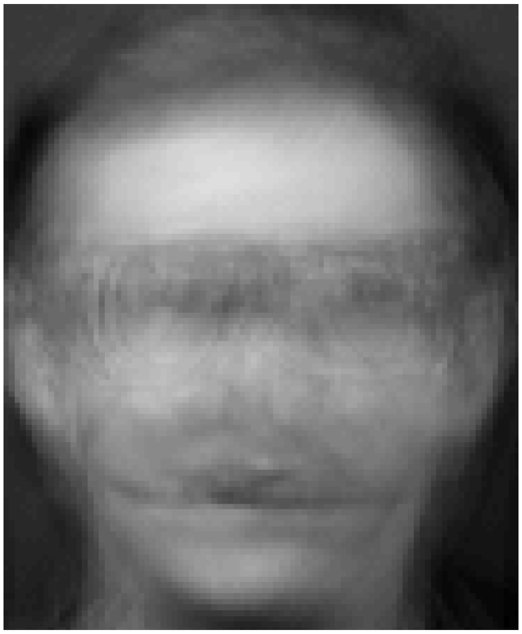

Figure 3: Mean Image

The next step involves the calculation of the Eigen faces via the calculation of the covariance matrix with reduced dimensionality [10]. This is achieved by multiplying the transposed of A with A, to achieve a 25 by 25 matrix. The reduction in dimensionality of the covariance matrix results in a much smaller number of Eigen vectors, namely 25 which corresponds with the number of images used as input.

Consequently, the Eigen vectors are only 25 elements per image. To obtain the 25 Eigen faces, the Eigen vectors are transformed back to the original dimension by multiplying them with the original input matrix *A*.

Figure 4 shows the obtained Eigen faces for the selected training images. At this point, any image of the training set can be represented by a weighted sum of the 25 Eigen Faces and the Mean Face, which results in a so-called weight matrix, and that for each image. This weight matrix is then used to reconstruct the image from the Eigen Faces, as shown in Figure 5 which shows an original image which was part of the training set and its corresponding reconstructed image.

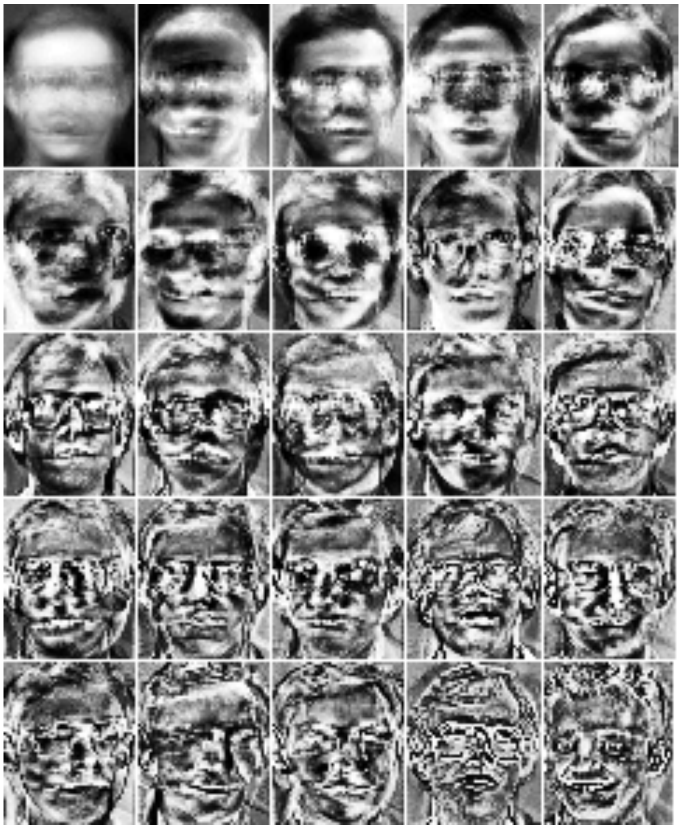

Figure 4: Eigen Faces

By using Euclidean distance calculation for a new input image in comparison to each of the images used in the original training set, one could identify the closest match. The graph shown in Figure 6 shows the Euclidean distance for the image of Figure 5 with each of the images from the training set (X axis). The values on the Y-axis are a combination of all the differences for the full image. It is clear from this graph that the set of images with the lowest Euclidean distance are the images from the same subject, which is in this case images 21-25, which correspond to the 5th subject.

To make the benefits of using PCA evident the trained system was also tested with an image that did not form part of the original image set. This does not only show the reconstruction accuracy, but also the ability to detect a particular subject. For this test, an image of subject one was chosen different from the ones used in the training set. This image is shown in Figure 7 on the left, and when using the standard PCA methods, the reconstructed image then became the one on the right hand side of Figure 7. When calculating the Euclidean distance for this reconstructed image versus the training images, then one achieves the results shown in Figure 8, which indicate that even though the reconstructed image is very sketchy, the closest match seems to be with Subject 1.

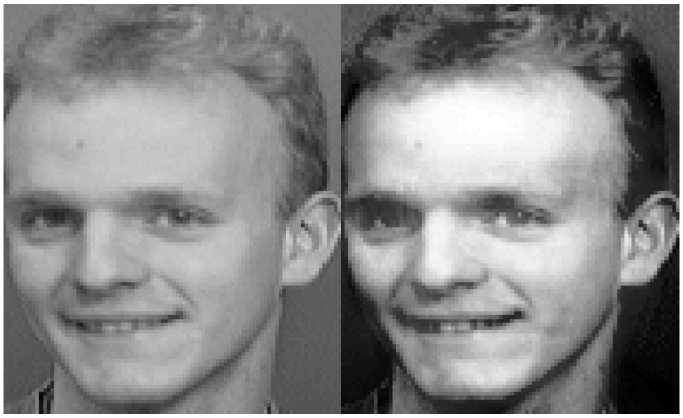

Figure 5: Original Image part of training set (Left), and Reconstructed Image(Right)

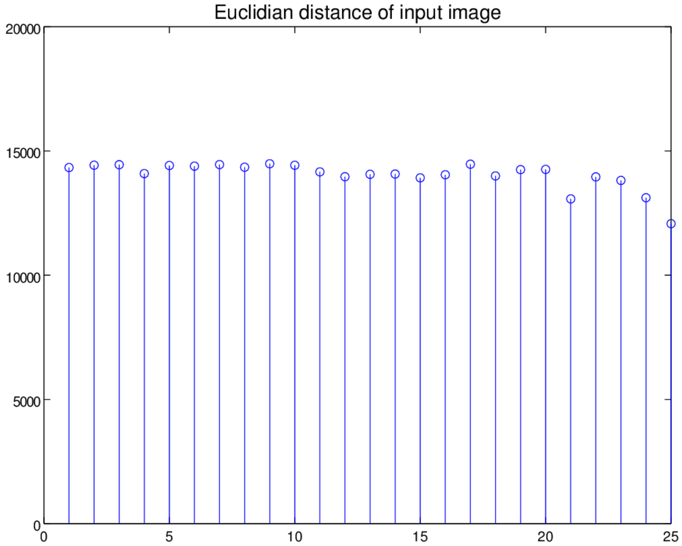

Figure 6: Euclidean Distance for the input Image shown in Figure 5, which is part of the training set versus allof the input images.

Based on the above application of PCA, one should be able to appreciate that there are significant savings with regards to storage requirements. Any input data going through the PCA model, as shown in Figure 1, will require less data to be stored due to the fact that the unique data is separated from the common information, where the latter is only stored once. Tests on the current data set show that 16% of the training data is unique, which corresponds to the Eigen Faces, on top of that one needs to store the mean Face and the Weight Matrices, which means that a total of only 20% of the original data set needs to be stored, resulting in a saving of 80%. This is in line with the results achieved in [5] which compares various

PCA-type algorithms and requires 23% of the original data to be stored. The minor difference with this previous ork lies in the use of different date sets.

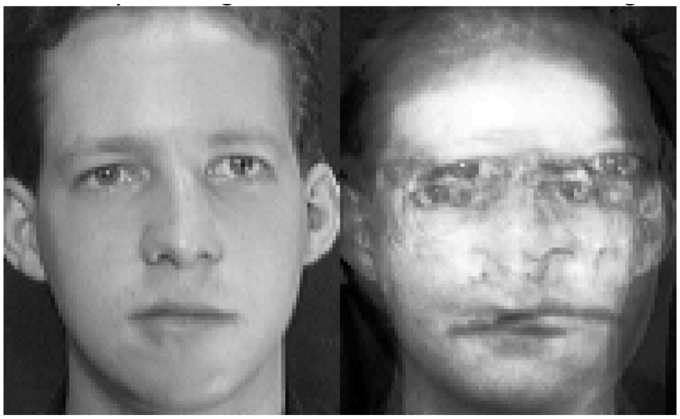

Figure 7: Original Image not part of training set (left), and reconstructed image (right).

When combining PCA with a supervised, decision tree based machine learning algorithm as standardly found within Matlab, and using various test sets of either training and/or test data, then the detection accuracy of these test sets achieves a detection accuracy of 77-79%, while if the machine learning algorithm is used on its own then the detection accuracy for the same test sets is only 45-53%.

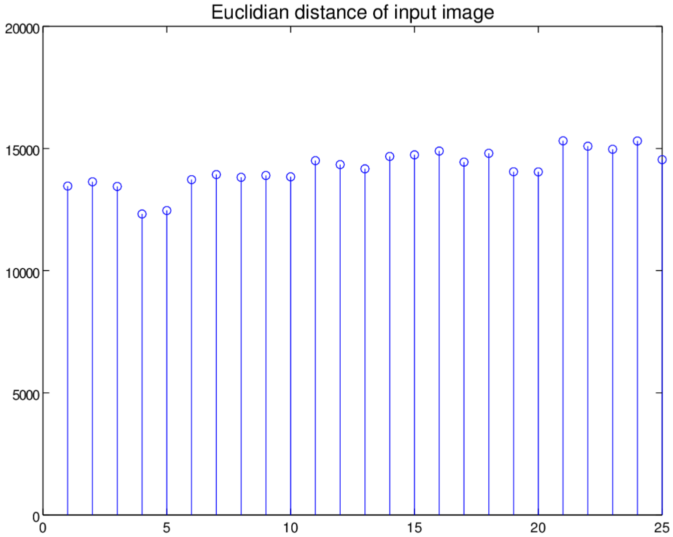

Figure 8: Euclidean Distance of Untrained Input Image with Trained Image

Conclusion

While most machine learning algorithms are built purely on the principle of similarity detecting, this paper has combined PCA with an existing machine learning algorithm, to identify the impact of using context/differences and to determine the impact of this combination with regards to detection accuracy and data storage requirements. The results show that the data storage requirements improve up to 80%, but also the detection accuracy improves by about 30%.

Hence, future work will focus on how to identify context automatically and how to integrate the "difference" principle, found in PCA, directly into machine learning algorithms.

References

[1] D. Shi, Z. DingDing and D. Wei, "The Study of Network Traffic Identification Based on Machine Learning Algorithm," in Computational Intelligence and Communication Networks (CICN), 2012 Fourth International Conference on, 2012.

[2] F. Burger, C. Buck, J. Pauli and W. Luther, "Image-based object classification of defects in steel using data-driven machine learning optimization," in Computer Vision Theory and Applications (VISAPP), 2014 International Conference on, 2014.

[3] A. Melo, E. German, M. Osorio, A. Ricardo, C. Orjuela and D. Alvaro, "Prediction of Spontaneous Termination of Atrial Fibrillation with Supervised Neural Networks," in Andean Region International Conference (ANDESCON), 2012 VI, 2012.

[4] L. Guerra, L. M. McGarry, V. Robles, C. Bielza, P. Larranaga and R. Yuste, "Comparison between supervised and unsupervised classifications of neuronal cell types: a case study," Developmental neurobiology, vol. 71, no. 1, pp. 71-82, 2011.

[5] M.-J. Yang, H.-R. Zheng, H.-Y. Wang, S. Mcclean and N. Harris, "Combining feature ranking with PCA: An application to gait analysis," in Machine Learning and Cybernetics (ICMLC), 2010 International Conference on, 2010.

[6] W. Hua, M. Cuiqin and Z. Lijuan, "A brief review of machine learning and its application," in Information Engineering and Computer Science, 2009. ICIECS 2009. International Conference on, 2009.

[7] G. J. Wolff, Unifying computing and cognition, CognitionResearch. org. uk, 2006.

[8] N. G. Chitaliya and A. Trivedi, "Feature Extraction Using Wavelet-PCA and Neural Network for Application of Object Classification \& Face Recognition," in Computer Engineering and Applications (ICCEA), 2010 Second International Conference on, 2010.

[9] AT&T Laboratories Cambridge, The Database of Faces, Available online at: http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html [Accessed on: 20/12/2015].

[10] K. I. Diamantaras and M. Strintsiz, "Noisy PCA theory and application in filter bank codec design," in Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on, 1997.