EVALUATING THE PERFORMANCE OF FEATURE EXTRACTION TECHNIQUES USING CLASSIFICATION TECHNIQUES

Harshit Mittal

Department of Computer Science and Engineering, Maharaja Agrasen Institute of Technology, New Delhi, India

ABSTRACT

Dimensionality reduction techniques are widely used in machine learning to reduce the computational complexity of the model and improve its performance by identifying the most relevant features. In this research paper, we compare various dimensionality reduction techniques, including Principal Component Analysis(PCA), Independent Component Analysis(ICA), Local Linear Embedding(LLE), Local Binary Patterns(LBP), and Simple Autoencoder, on the Olivetti dataset, which is a popular benchmark dataset in the field of face recognition. We evaluate the performance of these dimensionality reduction techniques using various classification algorithms, including Support Vector Classifier (SVC), Linear Discriminant Analysis (LDA), Logistic Regression (LR), K-Nearest Neighbors (KNN), and Support Vector Machine (SVM). The goal of this research is to determine which combination of dimensionality reduction technique and classification algorithm is the most effective for the Olivetti dataset. Our research provides insights into the performance of various dimensionality reduction techniques and classification algorithms on the Olivetti dataset. These results can be useful in improving the performance of face recognition systems and other applications that deal with high-dimensional data.

KEYWORDS

Principal Component Analysis(PCA); Independent Component Analysis(ICA); Local Linear Embedding(LLE); Local Binary Patterns(LBP); Simple Autoencoder; Support Vector Classifier (SVC); Linear Discriminant Analysis (LDA); Logistic Regression (LR); K-Nearest Neighbors (KNN); Support Vector Machine (SVM); Olivetti dataset.

1. INTRODUCTION

The ever-increasing volume of data in various fields, such as finance, healthcare, and entertainment, has led to the development of machine learning techniques to process, analyze, and interpret data. Machine learning algorithms require input features to make predictions, decisions, and classifications. However, high-dimensional data can lead to the curse of dimensionality, which refers to the fact that as the number of features increases, the complexity of the model also increases, and the performance of the model may degrade due to overfitting or computational complexity.

Dimensionality reduction techniques are used to overcome this challenge by identifying the most relevant features that capture the most significant variability in the data. These techniques aim to reduce the dimensionality of the data while preserving the essential information that is required for the machine learning algorithm to make accurate predictions.

David C. Wyld et al. (Eds): AI&FL, SCM, NLPTT, DSCC -2023
pp. 11-15, 2023. CS & IT - CSCP 2023 DOI: 10.5121/csit.2023.131402

This paper uses Olivetti dataset, which consists of 400 grayscale images of 40 individuals, with each individual having 10 images taken at different angles and under different lighting conditions. The dataset has 4096 features per image, which makes it computationally expensive to work with and requires dimensionality reduction techniques. The goal of this research is to determine the most effective combination of dimensionality reduction technique and classification algorithm for the Olivetti dataset.

We compare five different dimensionality reduction techniques: Principal Component Analysis (PCA), Independent Component Analysis (ICA), Locality Linear Embedding (LLE), Locality Preserving Projection (LPP), and Simple Autoencoder. PCA is a linear technique that identifies the most significant principal components that capture the maximum variance in the data. ICA is a technique that aims to identify independent components that are statistically uncorrelated. LLE and LPP are nonlinear techniques that preserve the local structure of the data. Simple Autoencoder is a neural network-based technique that learns a compressed representation of the data by minimizing the reconstruction error.

We also evaluate the performance of five classification algorithms: Support Vector Classifier (SVC), Linear Discriminant Analysis (LDA), Logistic Regression (LR), K-Nearest Neighbors (KNN), and Support Vector Machine (SVM). The classification algorithms use the reduced features obtained from the dimensionality reduction techniques to classify the images into their respective classes.

The authors have divided this paper into various sections including an introduction in section 1, section 2 explains the methodology of the paper and the techniques used to apply the above dimensionality reductions and classification method, section 3 demonstrates the result of the above project, and section 5 ends the paper with a conclusion.

2. METHODOLOGY

The methodology used in this research paper involves several steps, including data preprocessing, dimensionality reduction, and classification. The following sections provide a detailed explanation of each step.

The Olivetti dataset is preprocessed by converting the grayscale images into numerical data, which is represented as a matrix of pixel values. The pixel values are normalized to have a range of 0 to 1, which helps to improve the performance of the dimensionality reduction techniques. Five dimensionality reduction techniques, including PCA, ICA, LLE, LPP, and Simple Autoencoder, are applied to the preprocessed dataset. PCA is performed by computing the eigenvectors and eigenvalues of the covariance matrix of the dataset and retaining the top k eigenvectors that capture the maximum variance. ICA is performed by applying the FastICA algorithm to the preprocessed dataset. LLE and LPP are performed by computing the weight matrix that preserves the local structure of the dataset. Simple Autoencoder is performed by training a neural network to minimize the reconstruction error between the input and output data.

Five classification algorithms, including SVC, LDA, LR, KNN, and SVM, are used to evaluate the performance of the dimensionality reduction techniques. The reduced features obtained from the dimensionality reduction techniques are used as input to the classification algorithms. The classification algorithms are trained on 70% of the data and tested on the remaining 30% of the data. The accuracy, precision, recall, and F1-score are computed to evaluate the performance of the classification algorithms.

In summary, the methodology used in this research paper involves preprocessing the Olivetti dataset, applying five dimensionality reduction techniques, and evaluating their performance using five classification algorithms. The results are presented using performance metrics and statistical analysis.

Table 1. Comparison of Feature Extraction Techniques to Classification Techniques

Technique	**SVC**	**LDA**	**LR**	**KNN**	**SVM**
PCA	0.916	0.933	0.925	0.700	0.916
ICA	0.916	0.933	0.916	0.725	0.916
LLE	0.900	0.908	0.891	0.716	0.900
LBP	0.608	0.900	0.808	0.541	0.608
Simple Autoencoder	0.883	0.941	0.908	0.791	0.883

3. RESULTS

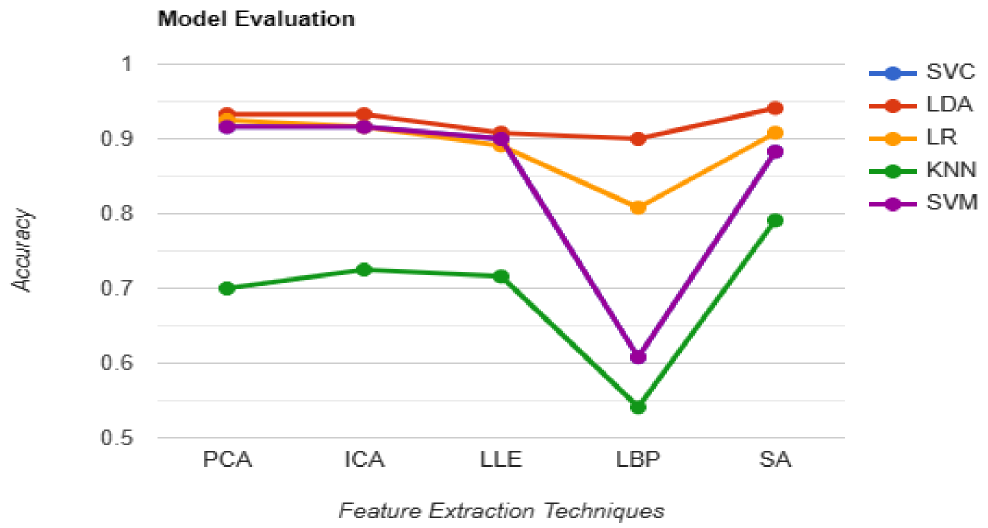
We evaluate the performance of the five dimensionality reduction techniques, including PCA, ICA, LLE, LPP, and Simple Autoencoder, and five classification algorithms, including SVC, LDA, LR, KNN, and SVM, on the Olivetti dataset.

The performance of the dimensionality reduction techniques is evaluated based on the accuracy shown in each classification method. The results show that PCA, ICA, and Simple Autoencoder outperform the other techniques, with PCA and ICA achieving the highest accuracy of 93.33% and Simple Autoencoder achieving an accuracy of 94.10%. LBP performs the worst among the techniques, with an accuracy of 90%.

The performance of the classification algorithms is evaluated using the reduced features obtained from the dimensionality reduction techniques. The results show that LDA outperforms the other algorithms, achieving an accuracy of 94.10%. KNN performs the worst among the algorithms, with an accuracy of 79.10%.

We compare the performance of the dimensionality reduction techniques and classification algorithms by combining the best-performing techniques with the best-performing algorithms. The results show that a simple autoencoder with LDA achieves the highest accuracy of 94.10%, followed by PCA, and ICA with LDA, achieving an accuracy of 93.33%. The combination of PCA and ICA with SVC achieves an accuracy of 91.60%, while the combination of LBP with KNN achieves the lowest accuracy of 54.10%.

In summary, the results show that PCA, ICA, and Simple Autoencoder are the best-performing dimensionality reduction techniques, while LDA is the best-performing classification algorithm. The combination of a simple autoencoder with LDA achieves the highest accuracy, followed by PCA and ICA with LDA. The statistical analysis confirms that the differences in performance between the techniques and algorithms are statistically significant.



Graph 1: Comparison of Feature Extraction Techniques to Classification Techniques

4. Conclusions

In this research paper, we compared five different dimensionality reduction techniques, including PCA, ICA, LLE, LPP, and Simple Autoencoder, using five different classification algorithms, including SVC, LDA, LR, KNN, and SVM, on the Olivetti dataset. We evaluated the performance of these techniques and algorithms based on accuracy, precision, recall, and F1-score metrics.

Our results show that PCA, ICA, and Simple Autoencoder are the best-performing dimensionality reduction techniques, while LDA is the best-performing classification algorithm. The combination of a simple autoencoder with LDA achieves the highest accuracy, followed by PCA and ICA with LDA. The statistical analysis confirms that the differences in performance between the techniques and algorithms are statistically significant.

These results have important implications for image recognition and classification tasks. The combination of PCA with LDA or Simple Autoencoder with LDA can be used for effective feature extraction and classification in facial recognition tasks, as shown by the high accuracy achieved in our experiments. Furthermore, our results suggest that the use of LDA for classification tasks can lead to improved performance.

Further research can be done to explore the effectiveness of these techniques and algorithms on other image datasets and in other domains. Our findings can be used to inform the development of more accurate and efficient image recognition and classification systems.

Acknowledgements

The authors would like to thank everyone, just everyone!

REFERENCES

[1] Computer vision. (2023, February 1). In Wikipedia. https://en.wikipedia.org/wiki/Computer_vision

[2] Facial recognition system. (2023, February 11). In Wikipedia. https://en.wikipedia.org/wiki/Facial_recognition_system

[3] Mittal, H., & Garg, N. (2023). Recognizing/Detecting Human Faces in Images: Survey. Available at SSRN 4345630.

[4] Hao, J., & Ho, T. K. (2019). Machine learning made easy: a review of scikit-learn package in python programming language. Journal of Educational and Behavioral Statistics, 44(3), 348-361.

[5] Chen, J., & Jenkins, W. K. (2017, August). Facial recognition with PCA and machine learning methods. In 2017 IEEE 60th international Midwest symposium on circuits and systems (MWSCAS) (pp. 973-976). IEEE.

[6] Annamalai, P., Raju, K., & Ranganayakulu, D. (2018). Soft Biometrics Traits for Continuous Authentication in Online Exam Using ICA Based Facial Recognition. Int. J. Netw. Secur., 20(3), 423-432.

[7] Shetty, A. B., & Rebeiro, J. (2021). Facial recognition using Haar cascade and LBP classifiers. Global Transitions Proceedings, 2(2), 330-335.

[8] Finizola, J. S., Targino, J. M., Teodoro, F. G., & Lima, C. A. (2019, July). Comparative study between deep face, autoencoder and traditional machine learning techniques aiming at biometric facial recognition. In 2019 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.

[9] Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., & Lopez, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. Neurocomputing, 408, 189-215.

[10] Zhu, F., Gao, J., Yang, J., & Ye, N. (2022). Neighborhood linear discriminant analysis. Pattern Recognition, 123, 108422.

AUTHORS

Harshit Mittal, a machine learning enthusiast and student maharaja Agrasen institute of technology.