

Identificação e Classificação de Dígitos Manuscritos com Auxílio de PCA

Lucas A. A. Ramos Ponte¹
 ITA/Casa dos Ventos, São Paulo, SP
 Régis M. A. da Costa²
 ITA/Casa dos Ventos, Fortaleza, CE

O reconhecimento de dígitos manuscritos é um desafio crucial na visão computacional, com aplicações importantes em serviços postais e processamento de documentos. Este relatório propõe uma abordagem usando Análise de Componentes Principais (PCA) para extrair características essenciais das imagens de dígitos, visando aprimorar a eficiência e precisão dos modelos de aprendizado de máquina. O conjunto de dados analisado inclui 5620 imagens de dígitos manuscritos, cada uma em 8×8 pixels em tons de cinza (Figura 1), utilizadas para avaliar algoritmos de reconhecimento. As imagens foram linearizadas para facilitar o processamento dos dados e a aplicação de técnicas analíticas.

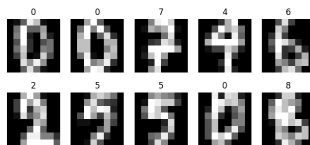


Figura 1: Amostra de imagens de dígitos manuscritos contidas no conjunto de dados em análise. Fonte: dos autores

O PCA é então aplicado para identificar componentes principais descorrelacionadas, ordenadas por sua contribuição para a variância total, reduzindo a dimensionalidade e preservando a informação das características [2]. Portanto, com a redução da quantidade de dados, preservando a informação, auxilia no desempenho de modelos de aprendizado de máquina.

Dessa forma, de posse da matriz de imagens $\mathbf{X} \in \mathbb{R}^{5620 \times 64}$ seguiu-se o algoritmo de PCA [1] dada a decomposição em valores singulares da matriz de covariância, seguindo as equações (1) e (2).

$$\mathbf{K} = \frac{\mathbf{X}_{norm}^T \cdot \mathbf{X}_{norm}}{5620 - 1}, \text{ onde } \mathbf{X}_{norm} = \mathbf{X} - \bar{\mathbf{X}}. \quad (1)$$

$$\mathbf{Y} = \mathbf{K} \cdot \mathbf{V} \Rightarrow \mathbf{Y} = \mathbf{U} \cdot \mathbf{\Sigma}, \quad (2)$$

A contribuição de cada componente principal pode ser expressa pela variância dos dados originais explicada por cada uma delas: $var_{exp}(y_i) = \frac{\sigma_i}{\sum_{j=1}^n \sigma_j}$, onde σ representa os valores singulares calculados anteriormente, correspondentes aos valores da diagonal de $\mathbf{\Sigma}$. Dessa forma, escolhe-se o número de componentes que explique uma parcela satisfatória do conjunto de dados. No caso em questão, 20 componentes principais explicam 90% da variância das imagens originais (Figura 2).

¹lucas.ramos@casadosventos.com.br

²regis.miller@casadosventos.com.br

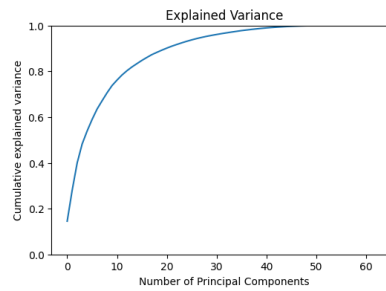


Figura 2: Variância explicada utilizando determinado número de componentes principais. Fonte: dos autores

Utilizou-se o modelo de classificação *K-Nearest Neighbors* (KNN) [3] e a redução de 64 componentes para apenas 20 demonstrou um desempenho excelente, indicando que os 64 componentes não eram totalmente necessários para obter resultados satisfatórios. Comparou-se o erro relativo da matriz $\left(e_k = \frac{\|A - A_k\|_2}{\|A\|_2} = \frac{\sigma_{k+1}}{\sigma_1} \right)$ e a acurácia do modelo variando o número de componentes utilizadas. Para isso, variou-se de 1 a 64 componentes e visualizou-se essa relação.

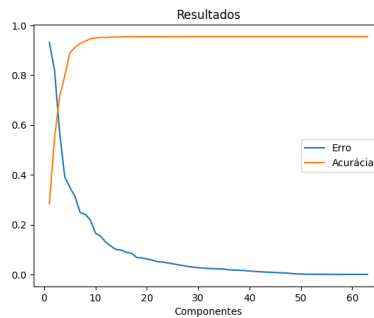


Figura 3: Gráfico de erro e acurácia por número de componentes principais. Fonte: dos autores

Portanto, obteve-se uma alta acurácia de 96,09% no reconhecimento dos dígitos manuscritos, sendo possível atingi-la com economia de informação através da redução da dimensionalidade por componentes principais do PCA. A redução de 64 componentes para as 20 primeiras que apresentaram 90% da variância explicada dos dados obteve precisão praticamente idêntica a sem redução de dimensionalidade e o erro entre as matrizes de dados não ultrapassou 10%. Este resultado mostra como o PCA é uma ótima ferramenta para economia do custo computacional da inteligência artificial e armazenamento de imagens.

Referências

- [1] I. T. Jolliffe. **Principal component analysis for special types of data**. Springer, 2002.
- [2] S. K. Joshi e S. Machchhar. “An evolution and evaluation of dimensionality reduction techniques—A comparative study”. Em: **2014 IEEE International Conference on Computational Intelligence and Computing Research**. IEEE, 2014, pp. 1–5.
- [3] J. VanderPlas. **Python data science handbook: Essential tools for working with data**. O’Reilly Media, Inc., 2016.