Analysis of Dimensionality Reduction Techniques on ML Models

Sreenidhi K, Jayendra K, Roshan Reddy Y, Ranjit C
Group 16

CSE512 - Machine Learning

Abstract

- It gets very cumbersome to deal with huge datasets to train ML models.
- Datasets may have a huge number of important features with high collinearity.
- In such cases, it is usually suggested to compress the data such that it contains all the features in much lesser variables.
- We aim to do an exploratory data analysis of a dataset and observe the trends in the data, like the covariance between the different attributes of the data.
- We will then employ some dimensionality reduction techniques and study their effectiveness on different ML models.

Problem Statement

How effective are dimensionality reduction techniques on ML models?

We want to address,

- Does computational time come at a huge cost of accuracy?
- How well the techniques work for representing high dimensional data to low dimensions.

Related Works

- The authors[1] have employed a linear dimensionality reduction method called PCA as a preprocessing stage in machine learning and observed improved performance (of 6%) from complete features.
- Similar performance is observed with UMAP, a non-linear dimensionality reduction method. A major improvement is seen over the MNIST dataset using HDBSCAN clustering with accuracy improving to 77.65% from 27.65% [2]
- We want to extend the related works and explore the computational time saving vs accuracy trade-off.

1. Y. Zhang and Z. Zhao, ‘‘Fetal state assessment based on cardiotocography parameters using PCA and AdaBoost,’’ in Proc. 10th Int. Congr. Image Signal Process., Biomed. Eng. Informat. (CISP-BMEI), Oct. 2017, pp. 1–6.
2. Mebarka Allaoui, Mohammed Lamine Kherfi, and Abdelhakim Cheriet. “Considerably Improving Clustering Algorithms Using UMAP Dimensionality Reduction Technique: A Comparative Study”. In: Image and Signal Processing. Ed. by Abderrahim El Moataz et al. Cham: Springer International Publishing, 2020, pp. 317–325.

Experimental Setup

Dataset:

The 20 newsgroups dataset is a collection of around 18000 news posts divided into 20 newsgroups or categories.

Techniques used:

Dimensionality Reduction Techniques: Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Kernel PCA, Uniform Manifold Approximation and Projection (UMAP)

Machine Learning Algorithms Used: SVM, K Means Clustering, Random Forest

Dataset Preprocessing

Politics

Medicine

- We first perform the usual preprocessing techniques such as stop words removal, lemmatization, TF-IDF vectorization, normalization.
- We consider a frequency threshold of 0.1% i.e., when building the vocabulary, the words which have frequency lower than this will be ignored. This ensures having lesser number of features after vectorization to save memory (i.e., 114,000 to 9194 features).

Methodology

Feature engineering techniques such as normalization, and conversion of categorical data to numeric data.

Experimenting with the normalized dataset over different ML algorithms like SVM, K Means Clustering, and Random Forest.

Linear dimensionality reduction techniques like PCA and LDA used on the normalized dataset and repeat the experimentation over the ML algorithms.

Non-linear dimensionality reduction techniques like Kernel PCA, Uniform Manifold Approximation, and Projection (UMAP) on the normalized dataset and checking the same performance of ML algorithms.

Analyzing the results obtained by the ML algorithms with and without dimensionality reduction, the effect of dimensionality reduction on the performance of ML algorithms is investigated.

Performance of these classifiers is then evaluated on the metrics like Precision, Recall, F1-Score, Accuracy, and One-Shot Similarity Score[6].

Results

Dataset	Features	Accuracy	Computation Time (mins)
20 newsgroups	9286	69.6%	1.707
20 newsgroups + PCA	512	68.8%	0.728
20 newsgroups + LDA	17	57.6%	0.124
20 newsgroups + Kernel PCA	1024	68.8%	0.012
20 newsgroups + UMAP	32	60.64%± 1.49%	0.08

- Above results are calculated on Random Forest ML model.

Conclusion

We can see here that, upon applying dimensionality reduction techniques, the accuracy of the model reduces but the reduction in accuracy is not as significant as the time saved in execution. In the real world, we can have massive data where the model can take too long to execute. Dimensionality reduction can come in handy during that time.

We cannot always sacrifice the accuracy of the model. In cases such as cancer detection, we would like to be as accurate as possible because human lives are at stake!

References

[1] Jolliffe Ian T. and Cadima Jorge. “Principal component analysis: a review and recent developments”. In: The Royal Society 374 (2016).

[2] Alaa Tharwat et al. “Linear discriminant analysis: A detailed tutorial”. In: AI Commun. 30 (2017), pp. 169–190.

[3] Bernhard Sch¨olkopf, Alexander Smola, and Klaus Robert M¨uller. “Kernel principal component analysis”. English. In: Publisher Copyright: © Springer-Verlag Berlin Heidelberg 1997.; 7th International Conference on Artificial Neural Networks, ICANN 1997 ; Conference date: 08-10-1997 Through 10-10-1997. 1997, pp. 583–588.

[4] Mebarka Allaoui, Mohammed Lamine Kherfi, and Abdelhakim Cheriet. “Considerably Improving Clustering Algorithms Using UMAP Dimensionality Reduction Technique: A Comparative Study”. In: Image and Signal Processing. Ed. by Abderrahim El Moataz et al. Cham: Springer International Publishing, 2020, pp. 317–325.

[5] G. Thippa Reddy et al. “Analysis of Dimensionality Reduction Techniques on Big Data”. In: IEEE Access 8 (2020), pp. 54776–54788.

[6] Lior Wolf, Tal Hassner, and Yaniv Taigman. “The One-Shot Similarity Kernel”. In: Nov. 2009, pp. 897–902.