Efficient Data Reduction Technique by selecting Top-K discriminative Features using Principal Component Analysis for Efficient Light-Weight AI Models

Student: Faith Nwokoma

Principal Investigator: Dr. Cajetan Akujuobi

Prairie View A&M University, Prairie View, TX 77446

Abstract

Feature selection is of great important for applications where dimensionality reduction, analysis, and pattern discovery are to be deployed. This need is perhaps more for systems with limited computing resources like IoT networks. In this paper, we considered time series datasets and propose a unsupervised learning technique to identify the top-k discriminative features. The technique used Principal Component Analysis (PCA) statistical foundation to deduce the relative importance of the principal components of the dataset with its coefficients along the principal directions, consequently uncovering the ranks of the features. We use multiple benchmark datasets for various experiments evaluate the performance of the proposed method in terms of its ability for feature selection and and its capacity to minimize the original by evaluating the data reconstruction error. Our proposed method compared with other existing methods, results verify increased efficiency and accuracy.

Background

The explosion of big data based upon technological advances presents its own challenges that are not sufficiently solved by the existing data reduction, analysis and feature selection methodologies. This is also the challenge of limited computing resources for some edge computing systems like the IoT networks and many other edge devices. The additional presence of noise in high dimensional datasets makes it more difficult to uncover significant patterns in data, and this affects the quality of the systems.

This makes feature selection and feature extraction important preprocessing steps for improved accuracy and efficiency in uncovering patterns and trends in a dataset and for reduction of data size useful for computing systems with limited computing resources. This leads to improved efficiency of the overall system.

Methods

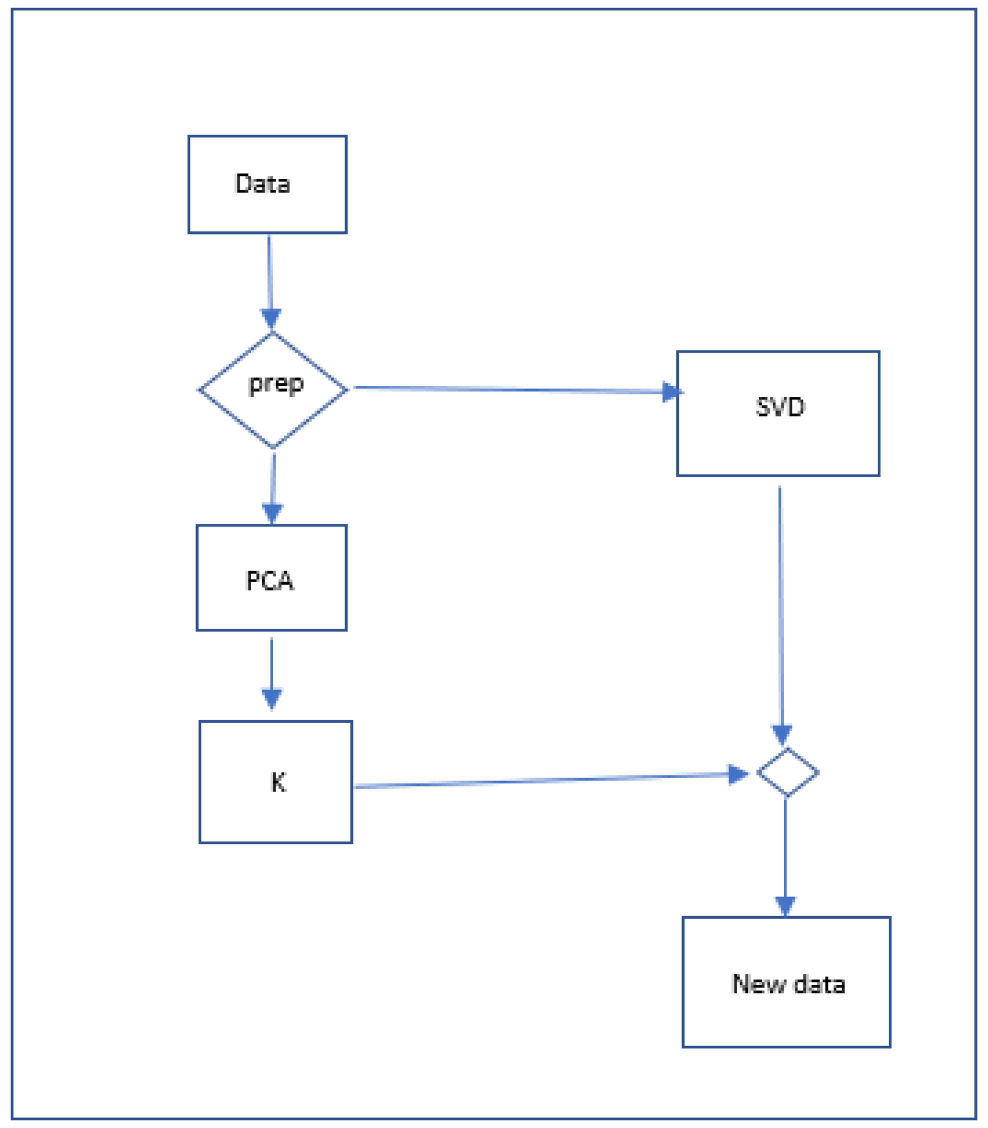

Fig.1 System diagram.

Algorithm 1 - Uncover the number k of PCs to retain

Input: $A \in R^{nxm}$, θ (cumulative variance explained)
Output: k, the number of principal components to retain.
begin
1: Uncover fraction of total explained variance
2: f(k) $\leftarrow \Sigma_{z=1}^{k} \lambda_z / \Sigma_{z=1}^{r} \lambda_z$ for all z = {1, …, r}
3: Choose the smallest k so that f(k) $\geqslant \theta$ and retain that number of k eigenvectors to keep explained variance θ in the new embedding.
4: return k
end

The (rank-k) weighted score of the i-th column of A is then computed as $wS_i^{(k)} = |\Sigma_{j=1}^{k} w_j t_{i,j}|$.

Algorithm 2 - Weighted Scores (WS)

Input: $A \in R^{n \times m}$, θ (cumulative variance explained)
Output: $S_r \in R^{n \times k}$and has the top k most representative ranked features of A.
begin
1: Compute the Singular Value Decomposition
$[U, S, V^T] \leftarrow SVD(A)$
2: Compute the proportion of variance carried by each component
For j $\leftarrow$ 1 to r
$\lambda_j \leftarrow (s_j^2/(n-1))$, where $s_j \in S$
end for
3: Identify the number k of principal components to retain
k $\leftarrow$ Algorithm1(A, θ)
4: $M \leftarrow V_k$
5: Build the weighted matrix $[wV_k]$
For j $\leftarrow$ 1 to k
$w_j \leftarrow \lambda_j / \Sigma_{j=1}^{r} \lambda_j$
$[wV_k]_{*,j} \leftarrow w_j * [M]_{*,j}$
end for
6: Compute the eighted score for each variable
$wS_i^{(k)} = |\Sigma_{j=1}^{k} w_j t_{i,j}|$, for all i = {1, 2, ..., m}.
7: Sort the variables according to their weights:
w$S_1^{(k)} \geq$ ... w$S_i^{(k)} \geq$ ... $\geq$ w$S_m^{(k)}$
end

Results

Dataset	SVD Reconstruction Error	Test Accuracy	Inference _time (sec)	N_samples	N_features	Best_N_Components
Arrhythmia	0.715595	0.604396	0.012478	452	279	13
Ionosphere	0.364809	0.971831	0.013278	351	34	15
madelon	0.978347	0.785000	0.020385	2000	502	5
Gissette	0.828217	0.971972	0.035977	5999	5000	89
IoT Intrusion	0.068727	0.974825	0.671137	80037	115	29

Table 1: Result Table

Fig 2: Reconstruction Error and Number of Samples Trained on

Conclusion

In this paper we propose an effective feature reduction technique that uses Principal Component Analysis and Singular Value Decomposition. It leverages the statistics of the principal components to identify the features that retain the maximum variability of the data, helping to reduce the reconstruction error. Our experiments conducted on various public datasets shows that while our method picks the topmost representative k features validated by the accuracy, the reconstruction error values shows that not much information is lost in the transformation process and the performance of the model on real IoT data shows that our algorithm performed well. The number of Principal components to retain should also be careful decided as we discovered at the number of components, the higher the reconstruction error and lower the accuracy. However, the margin of interest lies at the region where significant increase in the number of PCs, has little effect on the accuracy and reconstruction error. This makes the case for industry wide adoption of the process.

References

[1] R. Kavitha and E. Kannan, "An efficient framework for heart disease classification using feature extraction and feature selection technique in data mining," 2016 International Conference on Emerging Trends in Engineering, Technology and Science (ICETETS), Pudukkottai, India, 2016, pp. 1-5.

[2] M. F. I. Ibrahim and A. A. Al-Jumaily, "PCA indexing based feature learning and feature selection," 2016 8th Cairo International Biomedical Engineering Conference (CIBEC), Cairo, Egypt, 2016, pp. 68-71.

[3] Y. Li, K. Shi, F. Qiao and H. Luo, "A Feature Subset Selection Method Based on the Combination of PCA and Improved GA," 2020 2nd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), Taiyuan, China, 2020, pp. 191-194.

[4] C. Yumeng and F. Yinglan, "Research on PCA Data Dimension Reduction Algorithm Based on Entropy Weight Method," 2020 2nd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), Taiyuan, China, 2020, pp. 392-396.

[5] H. Shah and K. Verma, "Voltage stability monitoring by different ANN architectures using PCA based feature selection," 2016 IEEE 7th Power India International Conference (PIICON), Bikaner, India, 2016, pp. 1-6.

[6] A. P. Kale and S. Sonavane, "PF-FELM: A Robust PCA Feature Selection for Fuzzy Extreme Learning Machine," in IEEE Journal of Selected Topics in Signal Processing, vol. 12, no. 6, pp. 1303-1312, Dec. 2018.