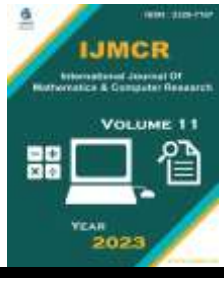
International Journal of Mathematics and Computer Research

ISSN: 2320-7167

Volume 11 Issue 09 September 2023, Page no. 3713-3715

Index Copernicus ICV: 57.55, Impact Factor: 7.362

DOI: 10.47191/ijmcr/v11i9.02



Data Dimensionality Reduction Algorithms for Machine Learning

Marina Popolizio[1], Alberto Amato[1], Rita Dario[2], Vincenzo Di Lecce[1]

[1]Politecnico di Bari – DEI, via Re David 200 – 70125 Bari - Italy

[2]COVID Emergency Task Force, Giovanni XXIII Children Hospital, Policlinico di Bari, 70126 Bari - Italy

ARTICLE INFO

Published Online:
05 September 2023

Corresponding Author:
Marina Popolizio

ABSTRACT

The rise of machine learning yields remarkable outcomes across fields. Phrases like big data, AI, and cloud computing are becoming commonplace. Yet, data abundance doesn't assure success. Numerous works address data preprocessing for information discovery. This study assesses three techniques on unsupervised clustering, spotlighting SPQR's novel application. Findings stress preprocessing's data impact, urging caution against oversimplified datamining solutions.

KEYWORDS: data analysis; PCA; SPQR; SVD; FCM; Clustering Silhouette

I. INTRODUCTION

In AI, data quality directly influences learning performance. However, assessing the suitability of data for machine learning is not straightforward, as its quality may not be evident. This challenge affects both small datasets, common in the medical field, and large datasets prevalent in environmental studies.

According to experts, the critical quality attributes for machine learning are completeness, correctness and variety. In essence, data quality refers to data meeting user/activity requirements and suitability for the intended purpose.

Many machine learning algorithms aim to amplify knowledge extraction. They strive to reduce the dataset size while retaining information content.

Various dimension reduction techniques exist, like Isomap, locally linear embedding (LLE), Hessian LLE, and kernel Principal Component Analysis (PCA). These methods unveil the geometric structure of high-dimensional data. These techniques aim to pare down variables while minimizing information loss. Hence, the reduced dataset can often substitute the original without substantial information loss.

This study evaluated three dimensionality reduction algorithms: PCA, singular value decomposition (SVD), and Semi-Pivoted QR approximation (SPQR). The aim was to verify whether these algorithms could yield comparable results to using the full dataset as model input. Comparisons were conducted utilizing fuzzy clustering and silhouette analysis. The outcomes revealed a significant influence of the employed preprocessing techniques on the effectiveness of machine learning algorithms. In particular, using the SPQR method always gave the best possible outcomes.

II. REVIEW OF SVD, PCA AND SPQR

An interesting overview on feature selection methods can be found in [1]. Here we just recall the main aspects of the three methods we aim to compare, with special emphasis on the SPQR method that, at least according to the authors' experience, appears rarely used in this context.

The SVD is a fundamental mathematical technique used in linear algebra and data analysis [2]. It involves breaking down a matrix into three separate matrices, which represent the orthogonal transformation of the original matrix. It is very effective to extract underlying patterns, relationships, and features from data by decomposing it into its constituent parts.

PCA is a widely used statistical technique to simplify complex datasets by transforming them into a new coordinate system where the data's variance is maximized along the new axes [3]. This process helps reveal the most significant patterns and relationships within the data. Mathematically, PCA involves computing the covariance matrix of the original data and then finding its eigenvectors and eigenvalues, by resorting to its SVD. Both PCA and SVD alter the initial feature space into spaces defined by eigenvectors, making the evaluation of feature contributions and the handling of new data points more intricate, unlike SPQR which always treats the original data.

The SPQR method is a numerical technique used for matrix factorization and computation, particularly in the context of

large sparse matrices [4,5]. It builds upon the QR decomposition, which breaks down a matrix into an orthogonal matrix (Q) and an upper triangular matrix (R). In the SPQR algorithm, the "semi-pivoting" aspect refers to a strategic approach to column selection and sorting during the factorization process, which helps to preserve numerical stability and improve computational efficiency, particularly when dealing with sparse ill-conditioned matrices.

For these reasons, SPQR has found application in various fields, including scientific computing, engineering simulations and machine learning, where efficient handling of sparse matrices is essential for both accuracy and performance. It is often used as a preprocessing step in data analysis and machine learning algorithms to reduce the dimensionality of data while retaining its essential features, promoting faster and more accurate computations.

III. SILHOUETTE METHOD AND FCM

The purpose of this article is to observe the effect of the three data preprocessing techniques described above during data clustering.

Clustering is a data analysis technique that involves grouping similar data points together based on their shared characteristics. The primary objective of clustering is to discover underlying patterns and structures within a dataset, organizing the data into distinct clusters where points within the same cluster are more alike than those in different clusters. This technique aids in uncovering insights, simplifying data representation, and enabling further analysis by highlighting intrinsic relationships among data points. Clustering finds applications in various fields, such as customer segmentation, image recognition, and anomaly detection, contributing to better data understanding and decision-making. Specifically, we use the fuzzy clustering method for our tests, thus allowing for the possibility of a data point belonging to multiple clusters.

The silhouette parameter is commonly used to assess the clustering performance. This method involves evaluating the similarity of each object in its cluster (tightness) and to the other clusters (separation) comparatively. For a given data point *y* this parameter is calculated as follows:

$$s(y)=(b(y)-t(y))/\max\{b(y),t(y)\} \qquad (1)$$

where *t(y)* is the mean distance between point *y* and all other points in the same cluster, and *b(y)* is the smallest mean distance between *y* and any other cluster. The silhouette score ranges from -1 to 1. In particular, a high positive silhouette score indicates that the data point is well-matched to its own cluster and distant from neighboring clusters, implying a good clustering assignment. For our tests we use values between 0,7 and 0,9.

An in-depth analysis of this parameter is reported in [8].

IV. EXPERIMENTS AND RESULTS

In this study, four openly accessible UCI website databases [9] were examined:

- 60,000 instances and 171 features from APS Failure at Scania Trucks Dataset [9].
- 13,611 instances and 16 (visual) features from Dry Bean Dataset [6].
- 2,000 instances and 649 (numeral handwrite) features from Multiple Features Dataset [9].
- 4,746 instances and 21 features (Wikipedia contributors women) from Gender Gap in Spanish WP Dataset [7].

For each database we consider the silhouette analysis of the fuzzy clustering applied to the original database and to the database preprocessed by using the PCA, SVD and SPQR methods.

The clustering analysis was conducted using the FCM algorithm, varying the number of clusters from 10 to 50. For each database, the percentages of points with a silhouette level greater than 0.7, 0.8, and 0.9 are reported. These analyses were carried out 3 times on each database using:

- The raw data (namely, the original data stored in each database);
- The dataset reduced using the PCA algorithm, losing less than 2% of the total variation in the dataset;
- The dataset reduced using the SPQR algorithm, setting the same number of features used with PCA;
- The dataset reduced using the SVD algorithm, losing less than 2% of the total variation in the dataset.

The following tables show the mean results obtained in each test.

Table 1 - Mean clustering performance (percentage) for 10 clusters

Silhoutte	Raw data	PCA	SVD	SPQR
0,7	35,11	61,94	50,71	**77,70**
0,8	24,26	46,99	42,37	**68,65**
0,9	6,73	23,61	26,23	**49,90**

Table 2 - Mean clustering performance (percentage) for 20 clusters

Silhoutte	Raw data	PCA	SVD	SPQR
0,7	34,18	64,09	45,51	**69,45**
0,8	23,50	51,56	39,66	**58,34**
0,9	10,88	32,70	27,30	**34,86**

Table 3 - Mean clustering performance (percentage) for 30 clusters

Silhoutte	Raw data	PCA	SVD	SPQR
0,7	33,48	61,93	49,57	**67,53**
0,8	23,97	50,28	42,99	**55,49**
0,9	9,75	29,92	29,75	**34,87**

Table 4 - Mean clustering performance (percentage) for 40 clusters

Silhoutte	Raw data	PCA	SVD	SPQR
0,7	39,19	58,84	46,10	**68,11**
0,8	31,20	46,56	37,99	**56,55**
0,9	21,74	23,15	21,43	**32,78**

Table 5 - Mean clustering performance (percentage) for 50 clusters

Silhoutte	Raw data	PCA	SVD	SPQR
0,7	41,63	58,81	47,09	**68,26**
0,8	34,03	45,89	39,02	**54,93**
0,9	21,68	23,50	24,20	**31,20**

The SPQR always outperforms all other methods, with percentages that are much larger than those relative to other preprocessing methods and very impressive improvements with respect to the use of original data.

V. CONCLUSION

This study introduces the application of the SPQR algorithm as a preprocessing technique for machine learning. The authors utilized the Fuzzy C-Means (FCM) unsupervised learning method on four publicly available databases. Unlike prior research, they explored SPQR's novel use in this context. Comparative analysis involved established PCA and SVD methods under uniform conditions, evaluated through the silhouette parameter. The findings highlight the following:

- Dataset dimensionality reduction has potential for improving classification algorithm performance. Both computational and classification enhancements were observed.
- PCA and SVD transform the original feature space into eigenvector-defined spaces, complicating feature contribution assessment and handling new points.
- Notably, SPQR offers advantages over PCA and SVD. It doesn't alter the original dataset; instead, it adjusts feature positions, addressing earlier limitations.
- Experiment results consistently favored SPQR's application as a preprocessing technique over PCA and SVD, indicating superior performance.

REFERENCES

1. S. Visalakshi and V. Radha, "A literature review of feature selection techniques and applications: Review of feature selection in data mining," 2014 IEEE International Conference on Computational Intelligence and Computing Research, 2014, pp. 1-6, doi: 10.1109/ICCIC.2014.7238499.
2. G.H. Golub and C.F. Van Loan, *Matrix Computations*, 3rd, Johns Hopkins, 1996, ISBN 978-0-8018-5414-9.
3. I.T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments". *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 374 (2065): 20150202*
4. G.W. Stewart, "Four algorithms for the efficient computation of truncated pivoted QR approximations to a sparse matrix". Numer. Math. 83, 313–323 (1999)
5. M. Popolizio, A. Amato, V. Piuri and V. Di Lecce, "Improving Classification Performance Using The Semi-Pivoted QR approximation algorithm", 7-8 January 2022 2nd FICR International Conference on Rising Threats in Expert Applications and Solutions
6. M. Koklu and I.A. Ozkan "Multiclass Classification of Dry Beans Using Computer Vision and Machine Learning Techniques". Comput. Electron. Agric. 2020, 174, 105507.
7. J. Minguilln, J. Meneses, E. Aibar, N. Ferran-Ferrer and S. Fãbregues "Exploring the gender gap in the Spanish Wikipedia: Differences in engagement and editing practices". PLoS ONE 2021, 16, e0246702.
8. P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis", Journal of Computational and Applied Mathematics, Volume 20, 1987, Pages 53-65, ISSN 0377-0427.
9. D. Dua and C. Graff (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science