Research Proposal:

Similarity-Preserved Dimensionality Reduction Techniques for Incomplete Data

Runze ZHAO 120090715

Supervised by Prof. Wenye Li

香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

SME Quantitative Finance

Chinese University of Hongkong, Shenzhen

DATE SUBMITTED

[Revised November 19, 2023]

1 Introduction

Dimensionality reduction (DR) [1, 2] plays a pivotal role in simplifying complex datasets, thereby enabling more efficient and insightful data analysis. Techniques like Principal Component Analysis (PCA) [3, 4], Multidimensional Scaling (MDS) [5] and t-Distributed Stochastic Neighbor Embedding (t-SNE) [6] are central to reducing variables without significant information loss. However, a significant challenge in this domain is the presence of **incomplete data** with missing values. Traditional dimension reduction methods often falter when confronted with such datasets, leading to skewed or unreliable results. This issue is particularly acute in fields where data integrity is critical, such as healthcare and finance. Therefore, there is a growing need for dimension reduction techniques specifically tailored for incomplete datasets.

This project aims to explore and evaluate various approaches to performing dimensionality reduction in the context of incomplete data, with an emphasis on developing methodologies that are robust and effective in managing the intricacies of missing data points.

2 Research Values

Motivation:

1) Dimensionality reduction techniques are commonly used to reduce the computational costs for large-scale, high-dimensional data, such as similarity search and information retrieval.
2) Most existing works perform dimensionality reduction on the assumption of complete data.

Research Questions:

1) How to design similarity-preserved dimensionality reduction techniques that effectively handle incomplete data?
2) How to effectively retrieve information from incomplete datasets?

Research Significance:

1) Learning low-dimensional representations for incomplete data is crucial yet under-explored. This research will significantly contribute to reducing computational burdens associated with large-scale incomplete datasets.
2) Creating high-quality representations of incomplete data in lower dimensions is essential for real-world applications, such as information retrieval, similarity search, and clustering tasks.

3 Related Work

Most DR techniques assume data completeness, leaving a gap in approaches for incomplete datasets. Traditional methods typically rely on imputation, filling missing values with estimated ones. However, the "imputation-reduction" strategy often introduces noise and degrades performance. **This motivates us to design a novel imputation-free method combined with dimensionality reduction on incomplete data.**

3.1 Dimensionality Reduction Techniques on Complete Data

Dimensionality reduction can be broadly categorized into Linear Dimensional Reduction (LDR) and Nonlinear Dimensional Reduction (NLDR). LDR techniques are fundamental in reducing data complexity while preserving maximum variance. Principal Component Analysis (PCA) [3, 4], a cornerstone of LDR, linearly transforms data to a lower-dimensional space while retaining as much variability as possible. Multidimensional Scaling (MDS) [5] is designed to preserve the distances between data points, ensuring spatial relationships are maintained in the reduced space. Linear Discriminant Analysis (LDA) [7] focuses on maximizing the separation between different classes, enhancing the discriminative power of the reduced features.

On the other hand, NLDR methods cater to complex, nonlinear data structures where LDR techniques may not suffice. Kernel PCA (KPCA) [8], an extension of PCA, adapts it to nonlinear dimensions, allowing for a more nuanced data transformation. Locally Linear Embedding (LLE) [9] emphasizes preserving the local structure of data, making it ideal for datasets where neighborhood relationships are critical. t-Distributed Stochastic Neighbor Embedding (t-SNE) [6] excels in high-dimensional data visualization by reducing the dimensionality while maintaining relative distances between points. Uniform Manifold Approximation and Projection (UMAP) [10] is another significant contribution, offering a scalable solution for complex dimensionality reduction tasks.

3.2 Imputation Methods on Incomplete Data

Addressing incomplete data is a fundamental aspect of data preprocessing. Imputation methods, categorized into statistical imputation and matrix completion, are employed to fill missing values in datasets. Statistical imputation replaces missing values with statistical estimates [11, 12], utilizing approaches like Zero, Mean, and Mode imputation, where missing entries are filled with basic statistical measures. The kNN imputation method [13] leverages the concept of nearest neighbors,

estimating missing values based on the proximity to other data points. Linear regression models [14] offer another approach, predicting missing values by analyzing existing relationships in the data.

Matrix completion techniques, on the other hand, reconstruct entire data matrices, offering a more comprehensive solution for handling missing data. Singular Value Thresholding (SVT) [15] strategically focuses on significant singular values for matrix approximation. Polynomial Matrix Completion (PMC) [16] and Non-linear Matrix Completion (NLMC) [16] further advance these methodologies by incorporating polynomial and nonlinear functions, respectively, to manage intricate data structures more effectively.

Through these dimensionality reduction and imputation methods, the preprocessing stage of data analysis is significantly enhanced, providing robust techniques for simplifying and completing datasets for subsequent analysis.

4 Experimental Evaluation

4.1 Evaluation Datasets

This research will utilize a diverse array of datasets to ensure comprehensive evaluation across different data types:

- Text datasets:
  - BBC [17]: This dataset includes 2,225 text files spanning five topical areas, originally published on the BBC news website. It offers a rich source of textual data for analysis.
  - CNAE-9 [18]: Comprising 1,080 documents, this dataset features free-text business descriptions of Brazilian companies, categorized into nine distinct classes, providing a varied business context.
  - Sports Article [19]: This collection of 1,000 sports articles, classified as objective or subjective using Amazon Mechanical Turk, includes raw texts, extracted features, and source URLs, offering a unique dataset for content analysis.
- Image datasets:
  - MNIST [20]: A classic dataset of grayscale images of handwritten digits, widely used for benchmarking image processing systems.
  - Fashion MNIST [21]: Comprising grayscale images of 10 different types of fashion items,

this dataset provides a modern twist on image classification challenges.

  - COIL 20 [22]: Featuring grayscale images of 20 objects captured under 72 different rotations, offering a comprehensive set for object recognition tasks.

- Other datasets:
  - Waveform [23]: This dataset includes samples from three wave classes with 21-dimensional attributes, integrated with noise, presenting a challenge in signal classification.

4.2 Evaluation Metrics

Our evaluation will encompass both qualitative and quantitative results, focusing on aspects like clustering accuracy and visualization quality. The following metrics will be employed:

- Metrics for Similarity Preservation:
  - **Relative Error Analysis:** Assessing the discrepancy between the true similarity matrix and the estimated similarity matrix.
  - **Local Relationship Assessment:** Examining the congruence of local relationships within the true and estimated similarity matrices.
- Metrics for Similarity Search:
  - **k-nearest Neighbors Accuracy:** This metric evaluates how effectively the dimensional reduction preserves cluster information in the original data space.
- Metrics for Clustering:
  - **Normalized Mutual Information (NMI):** A widely recognized metric for assessing the performance of clustering algorithms.
  - **Silhouette Coefficient (SC):** Ranges from [-1,1], measuring the clarity and separation of clusters in the reduced dimensional space.

5 Research Route

1. Conduct a comprehensive literature review on the following three parts:
   - Dimensionality reduction techniques for complete data
   - Dimensionality reduction techniques for incomplete data
   - Imputation methods for incomplete data

2. Evaluate the performance of baselines:
   - Evaluate the performance of dimensionality reduction techniques designed for incomplete data
   - Evaluate the performance of the "imputation-reduction" strategy
3. Design a new imputation-free dimensionality reduction method
   - Obtain a high-quality similarity matrix for incomplete data by solving the optimization problem
   - Design a new similarity-preserved strategy for dimensionality reduction
4. Perform a theoretical analysis of the proposed method
   - Analyze a theoretical bound for the constructed similarity matrix
   - Analyze the computational complexity of the proposed method

6 Research Plan

The project is scheduled to spans from November 13 to February 2, covering a duration of 12 weeks, to catch the submission deadline for the International Conference on Machine Learning (**ICML'2024**). The following table outlines the detailed schedule:

Task	Week1-2	Week3-4	Week5-6	Week7-8	Week9-10	Week11-12
Topic Selection and Preliminary Research	X					
In-depth Research on former algorithms	X	X				
Outline and Thesis Development		X	X	X		
Experiment: test the workability of proposed algorithm			X	X	X	
Drafting the Essay				X	X	
revision and peer review					X	X

References

[1] G Thippa Reddy, M Praveen Kumar Reddy, Kuruva Lakshmanna, Rajesh Kaluri, Dharmendra Singh Rajput, Gautam Srivastava, and Thar Baker. Analysis of dimensionality reduction techniques on big data. *IEEE Access*, 8:54776–54788, 2020.

[2] Xuan Huang, Lei Wu, and Yinsong Ye. A review on dimensionality reduction techniques. *International Journal of Pattern Recognition and Artificial Intelligence*, 33(10):1950017, 2019.

[3] Ian T Jolliffe and Jorge Cadima. Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

[4] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[5] John W Sammon. A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, 100(5):401–409, 1969.

[6] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11), 2008.

[7] Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936.

[8] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

[9] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[10] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv Preprint arXiv:1802.03426*, 2018.

[11] Susanne Rässler, Donald B Rubin, and Elizabeth R Zell. Imputation. *Wiley Interdisciplinary Reviews: Computational Statistics*, 5(1):20–29, 2013.

[12] Ton De Waal, Jeroen Pannekoek, and Sander Scholtus. *Handbook of Statistical Data Editing and Imputation*, volume 563. John Wiley Sons, 2011.

[13] Lorenzo Beretta and Alessandro Santaniello. Nearest neighbor imputation algorithms: A

critical evaluation. *BMC Medical Informatics and Decision Making*, 16(3):197–208, 2016.

[14] Zhongheng Zhang. Missing data imputation: Focusing on single imputation. *Annals of Translational Medicine*, 4(1), 2016.

[15] Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

[16] Jicong Fan, Yuqian Zhang, and Madeleine Udell. Polynomial matrix completion for missing data imputation and transductive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3842–3849, 2020.

[17] Derek Greene and Pádraig Cunningham. Practical solutions to the problem of diagonal dominance in kernel document clustering. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 377–384, 2006.

[18] Patrick Ciarelli and Elias Oliveira. CNAE-9. UCI Machine Learning Repository, 2012. DOI: https://doi.org/10.24432/C51G7P.

[19] Yara Rizk and Mariette Awad. Sports Articles for Objectivity Analysis. UCI Machine Learning Repository, 2018. DOI: https://doi.org/10.24432/C5801R.

[20] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[21] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: A novel image dataset for benchmarking machine learning algorithms. *arXiv Preprint arXiv:1708.07747*, 2017.

[22] Sameer A Nene, Shree K Nayar, Hiroshi Murase, et al. Columbia object image library (coil-20). 1996.

[23] Leo Breiman. *Classification and Regression Trees*. Routledge, 2017.