ingenieur
wissenschaften
htw saar

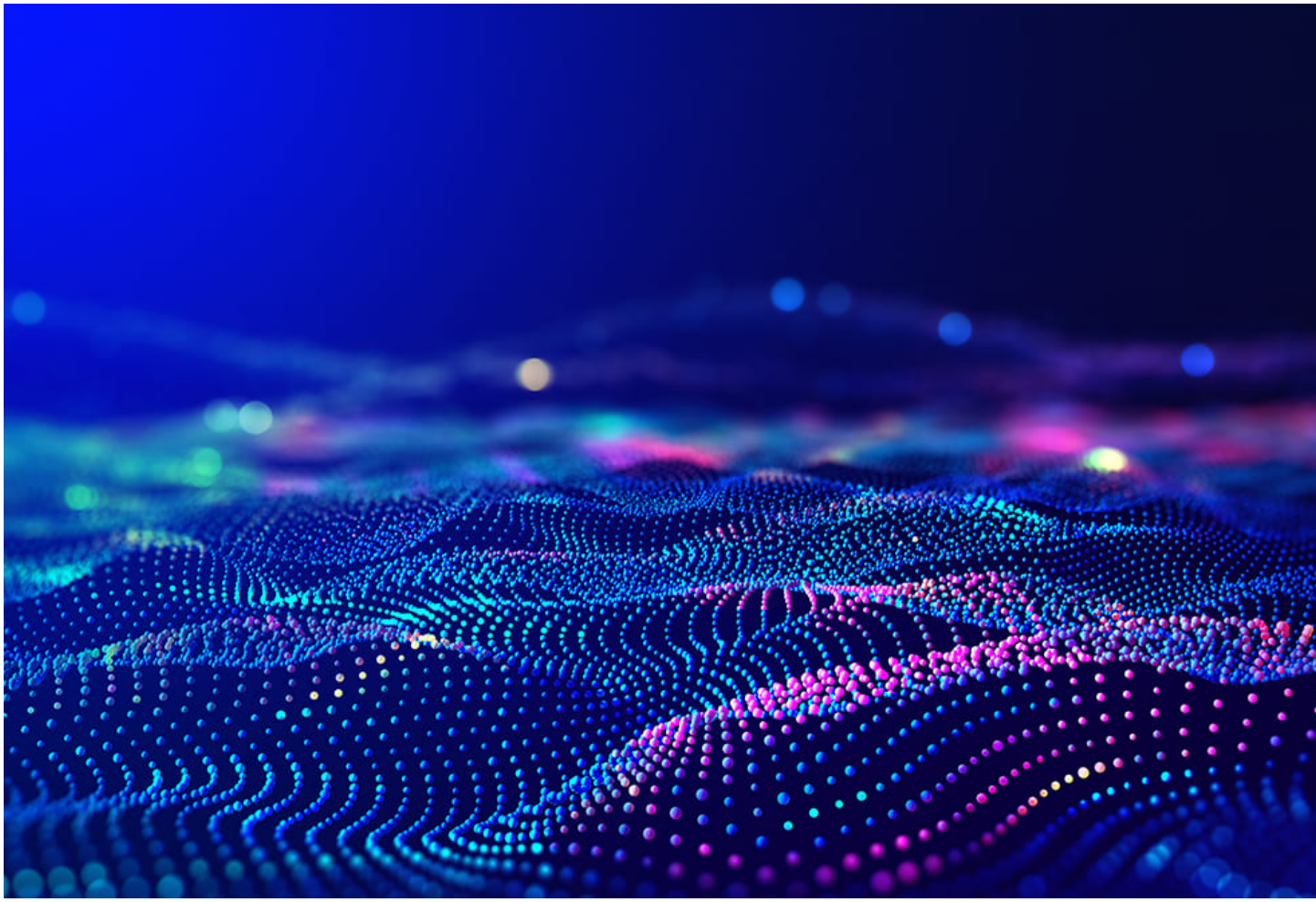
Hochschule für
Technik und Wirtschaft
des Saarlandes
University of
Applied Sciences

Revisiting the Curse of Dimensionality: Comparing Dimensionality Reduction Algorithms

by

Sosthene Donhou Fotso

Matriculation number: 3871878



Revisiting the Curse of Dimensionality: Comparing Dimensionality Reduction Algorithms

Sosthene Donhou Fotso
HTW Saar – Hochschule für Technik und Wirtschaft des Saarlandes
Seminar "Applied Computer Science Seminar"
Winter semester 2024

***Abstract*—The exponential growth of data in various domains has given rise to the challenge of dealing with high-dimensional data-sets. Dimensionality reduction techniques have emerged as powerful tools for effectively managing and extracting meaningful insights from such data. This paper presents a comparative analysis of popular dimensionality reduction techniques, aiming to shed light on their strengths, limitations, and applicability in different scenarios.**

The paper begins by providing a general overview on the curse of dimensionality as a problematic phenomenon in data analysis. It later on explores different dimensionality reduction techniques such as Principal Component Analysis (PCA [1]), Multidimensional Scaling (MDS [2]) and Density-Based Spatial Clustering of Applications with Noise (DB-SCAN) as useful algorithms in handling multi-dimensional data. Afterwords, the effectiveness of each of these algorithms is evaluated by applying them on two different data-sets (generated data with a low number of dimensions and empirical data with a with a high number of dimensions).

Finally, the paper concludes by summarizing the key findings, highlighting the most effective dimensionality reduction techniques for different types of high-dimensional data, and discussing potential future research directions.

Keywords: Dimensionality Reduction, Multidimensional Scaling (MDS), Principal Component Analysis (PCA), ensity- Based Spatial Clustering of Applications with Noise (DB-SCAN).

I. Introduction

In the vast landscape of data that permeates our daily lives, there exists an abundance of information. However, more data doesn't always translate to better insights; in fact, it often leads to a predicament known as the ***"curse of dimensionality"*** [3]. Imagine trying to make sense of a world where every interaction, image, or piece of text is described by an overwhelming number of features – it's akin to searching for a needle in a haystack. It is within this context that dimensionality reduction algorithms emerge as indispensable tools, providing a means to navigate the intricate landscapes of high-dimensional data.

One of the most prevalent usages of dimensionality reduction lies within the realm of text visualization. With the course of time, text visualization has indeed become an increasingly important field in the domain of information retrieval and visualization. As a result, statistical analysis tools and neuro-computation models have been widely used for dimensionality reduction. By reducing the number of features or variables in a text dataset for example, dimensionality reduction enhances the efficiency of text analysis by simplifying the data and improving algorithm performance [4]. As a result, the validity and effectiveness of these approaches or techniques largely depend on the specific data sets used and the semantics of the targeted application. To date, there has been very little evaluation to assess and compare dimensionality reduction techniques either numerically or empirically. In this vein, the focus of this paper is to investigate and eventually propose a mechanism for comparing and evaluating the effectiveness of different dimensionality reduction techniques. This shall be accomplished by studying some and possibly all these problems:

1) Which dimensionality reduction technique best preserves the interrelationships within a given data-set?
2) What is the sensitivity of the results to the number of output dimensions?
3) Can we automatically remove redundant or noise data while still preserving most information, and thus make dimensionality reduction more efficient?

A. Basics and terms

Before taking a deep dive into this topic, it shall be of utmost relevance to define key terminologies being employed such as Machine Learning, the Curse of Dimensionality as well as Dimensionality Reduction itself.

The data mining software framework being used ELKI (Environment for Developing KDD-Applications Supported by Index-Structures) shall also be briefly introduced.

1) Machine Learning: Machine learning is a type of artificial intelligence (AI) that allows software applications to learn from the data and become more accurate in predicting outcomes without human intervention [5]. It is a subset of AI that focuses on building systems that can learn from data, identify patterns, and make decisions with minimal human intervention. Machine learning algorithms enable the software to learn automatically from the data and improve over time.

2) The Curse of Dimensionality: As briefly acknowledged in the introductory words of this paper, the curse of dimensionality refers to the explosive

nature of increasing data dimensions and its resulting exponential increase in computational efforts required for its processing and/or analysis" [6]. This term was first mentioned by Richard E. Bellman [7], to explain the increase in volume of Euclidean space associated with adding extra dimensions in the field of dynamic programming. This phenomenon is increasingly observed nowadays in various domains such as data mining, data analysis and machine learning, just to name a few. An increase in the number of dimensions of a given data-set can theoretically add more information to it, thereby improving its quality [8]. The reality is, however, quite different. Increasing the dimensions (attributes) of the data-set practically increases noise (irrelevant data) and redundancy during analysis. This can be described with the following example:

If you take 100 random values from the one-dimensional space between 0 and 1, this area can be covered very well. However, if you have a 10-dimensional space with values between 0 and 1 and draws 100 random values again, then the coverage is much lower. However if you want a similar coverage as in one-dimensional space, you need $100^{10} = 10^{20}$ random values. This has result that the distance between the largest and the smallest distance within high-dimensional space approaches 0:

$$\lim_{n\to\infty} \left(\frac{dist_{max} - dist_{min}}{dist_{min}} \right) \to 0$$

Due to this peculiarity, there are problems with clustering with distance functions. To counteract this effect, one tries to remove irrelevant attributes from the data before clustering and thus reduce the number of dimensions.

3) Dimensionality Reduction: Dimensionality reduction is an attempt to transform a high-dimensional data-set into a low-dimensional space without losing relevant or valuable information for further data mining" [9]. This is done in different ways. For instance, an endeavor is undertaken to retain the maximum total variance using the smallest possible number of attributes, achieved either through direct filtering and the removal of weakly correlated attributes [10]. Principal Component Analysis (PCA) stands out as one of the most well-known representatives of this approach. However, it's important to note that each of these methods possesses its own set of strengths and weaknesses, a topic that will be thoroughly examined in the course of this exploration.

4) ELKI: "ELKI is an open source (AGPLv3) data mining software written in Java. The focus of ELKI is research in algorithms, with an emphasis on unsupervised methods in cluster analysis and outlier detection. In order to achieve high performance and scalability, ELKI offers data index structures such as the R*-tree that can provide major performance gains. ELKI is designed to be easy to extend for researchers and students in this domain, and welcomes contributions of additional methods. ELKI aims at providing a large collection of highly parameterizable algorithms, in order to allow easy and fair evaluation and bench-marking of algorithms." [11].

In the substantiation of this paper, CSV-Data-sets [12] retrieved from the ELKI-Project official GitHub Account shall be used.

II. Background

The commonly used methods for dimensionality reduction can be broadly classified into two main groups:

A. Feature Selection

Feature Selection is a dimensionality reduction approach in which the input variables to a given model are automatically reduced by using only relevant data and getting rid of noise data (data void of meaningful information). Some of its algorithms include correlation-based feature selection and Greedy-Forward-Selection (GFS) [13]. For textual data, an exemplary implementation of this approach is the common approach of removing frequently occurring "stop words" from the data-set.

B. Feature Extraction

Feature Extraction on the other hand presents itself as a dimensionality reduction technique which aims to reduce the number of features in a data-set by creating new features from the existing ones (and then discarding the original features). This new reduced set of features should then be able to summarize most of the information contained in the original set of features [14]. By so doing, a summarized version of the original features can be created from a combination of the original data-set. Some notorious feature selection algorithms include Principal Component Analysis (PCA) and Multidimensional Scaling (MDS).

In the context of this paper, PCA shall be used as one of the main analytic algorithms.

1) Principal Component Analysis (PCA): Principal Component Analysis (PCA) is a dimensionality reduction technique used in statistics and machine learning. It transforms high-dimensional data into a lower-dimensional representation while retaining as much original variance as possible. PCA identifies orthogonal axes, called principal components, that capture the most significant information in the data [15]. This is better illustrated through the following formula:

$$\begin{aligned} \text{PCA:} \quad & X = USV^T \\ \text{Principal Components:} \quad & Z = XV \\ \text{Explained Variance:} \quad & \text{Var}(Z_i) = \frac{\lambda_i}{\sum_{j=1}^{d} \lambda_j} \end{aligned}$$

X = original data matrix;
U,V = Orthogonal matrices from Singular Value Decomposition;
S = Diagonal matrix;
d = number of dimensions;
$\{\lambda_i\}$ = i-th eigenvalue of the covariance.

2) Multidimensional Scaling (MDS): Just like Principal Component Analysis, Multidimensional Scaling is also used in the context of dimensionality reduction for feature extraction. MDS can be referred to as being a statistical technique whose primary goal is to represent the pairwise dissimilarities between a set of objects in a lower-dimensional space. This approach is of particular relevance when trying to visualize and understand the structure or relationships within a dataset [16]. MDS aims to position each object in a lower-dimensional space (often two or three dimensions) in such a way that the pairwise distances or similarities between objects in the lower-dimensional space closely resemble those in the original higher-dimensional space [17].

C. Clustering

Data clustering, a foundational step in data analysis, involves grouping data points into distinct clusters to reveal underlying patterns and similarities. This organizational process enhances the interpretability and utility of datasets. Employing techniques such as k-Means and Density-Based Spatial Clustering of Applications with Noise (DB-SCAN), clustering aims to bring together similar data points within clusters [18]. k-Means minimizes within-cluster variance by assigning data points to the nearest centroid, while DB-SCAN identifies clusters based on data point density, distinguishing noise from meaningful clusters. This exploration delves into the strengths, limitations, and practical applications of these clustering methods, offering insights into their role in extracting valuable patterns from datasets.

III. Method

To compare the effectiveness of the techniques presented in the previous section, the cited algorithms shall be applied on two different data-sets: firstly, on generated data with a small number of dimensions and secondly on empirical data with a high number of dimensions. The following abbreviations are used in the graphical representations:

Plain - The data-set was not edited before the clustering
Euclid - The data-set undergone a euclidean transformation
PCA - The data-set was transformed using PCA

A. Generated Data-set

Generated data-sets from the ELKI-Project with clearly separated clusters shall be tested to verify correctness and make initial comparisons. Additionally, the quality of the clustering is measured by calculating the Silhouette-Coefficient. This coefficient measures how well-separated the clusters are and indicates the consistency of points within their assigned cluster compared to other clusters. It takes into account both the distance between data points within a cluster (intra-cluster distance) and the distance between data points of different clusters (inter-cluster distance). [19].

1) Small Dimensional Data: Figure 2 shows a data-set [20] with three dimensions in which four Gaussian-distributed clusters are clearly separated from each other. This should make it easy to find the clusters through clustering.

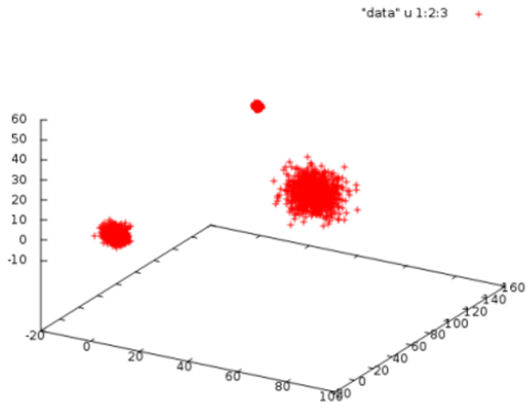


Figure 1. 3D-Visualization of a Gaussian distributed Cluster using DBSCAN and ELKI [12]

The results of the clustering in Figure 2 show the cluster scores after only one scaling (Euclid) or one rotation (PCA) of the data has taken place. It can be seen that even scaling the data can have a major impact, since there is hardly any difference in the results.

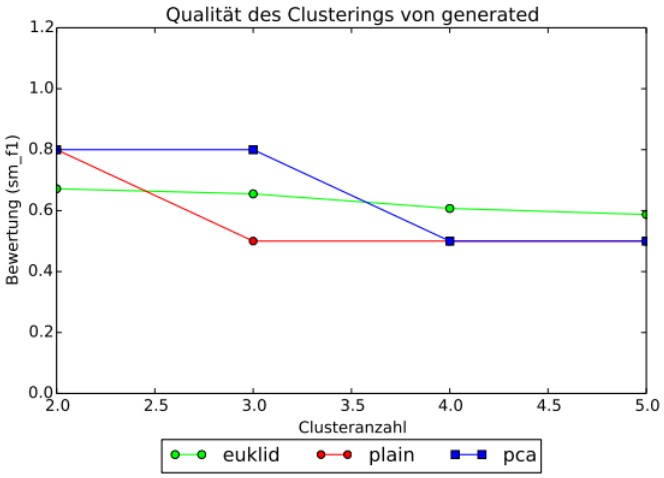


Figure 2. Quality of Clustering of generated Gaussian Distribution [12]

B. Empirical Data-set

Just like for small dimensional data, empirical data shall likewise be used here to demonstrate cluster differentiation.

1) High Dimensional Data: The mouse data-set comes from the ELKI developers and contains three clusters with 500 data sets as well as some noise in two dimensions in the shape of a mouse's head.

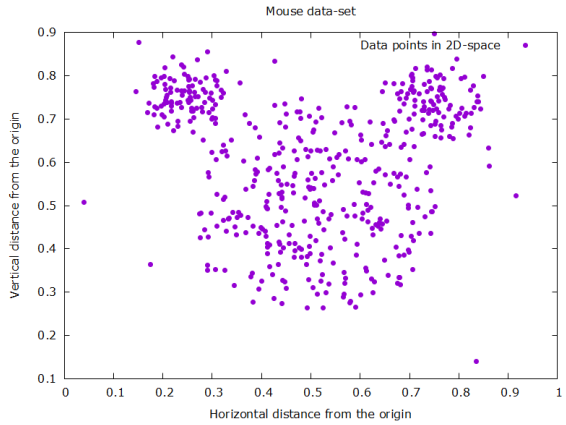


Figure 3. 2D-Visualization of clusters in the mouse data-set with GNU-Plot

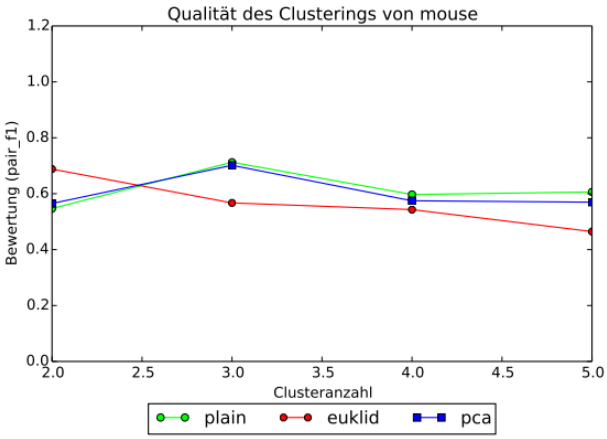


Figure 4. Quality of Clustering for the Mouse Data-set [12]

IV. Evaluation and Analysis

In order to achieve optimal clustering, one first has to think about the data, as the shape of the data in particular can have a major influence on clustering. In the case of low-dimensional data, it is helpful to plot the data in order to be clear about the shape. This is because clustering methods such as k-means, which optimize the distance squares, have their problems with non-spherical data.

For the purpose of data anaylsis, two metrics (F1-Score and Jaccard-index) shall be used as shown in the table below:

1) *F1-Score:* The F1-score, also known as the F1 measure or F1 value, is a metric commonly used in classification tasks, such as evaluating the performance of machine learning models. It combines precision and recall into a single numerical value, providing a balanced assessment of a model's overall accuracy [21].
Precision (P) is the ratio of true positive predictions to the total number of positive predictions. It measures the accuracy of positive predictions made by the model.
Recall (R) on the other-hand is also known as sensitivity or true positive rate, is the ratio of true positive predictions to the total number of actual positive instances. It quantifies the model's ability to capture all the positive instances.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (1)$$

2) *Jaccard Index:* The Jaccard index, also known as the Jaccard similarity coefficient or Jaccard coefficient, is a measure of similarity between two sets [22]. In the context of classification or clustering evaluation, the Jaccard index is often used to assess the similarity between the predicted set of labels and the true set of labels .

$$J(A,B) = \frac{\text{TP}}{\text{TP + FP + FN}} = \frac{\text{Size of Intersection}}{\text{Size of Union}} \quad (2)$$

Applying these metrics to the clustering algorithms K-Means and DBSCAN leads to the following results:

Algorithm	F1-Score	Jaccard
k-Means	0.34	0.78
DBSCAN	0.54	0.65

Table I
Results of k-Means and DBSCAN

Especially for data sets where the data is very close to each other and the variance is low, one should refrain from taking the step of dimension reduction, as this is always associated with a loss of variance. This has a correspondingly strong effect on such data sets.

Algorithm	F1-Score
Plain	0.61
PCA	0.59

Table II
Effects of PCA on the Mouse Data-set

V. Results and Discussion

In the context of this comprehensive exploration comparing dimensionality reduction algorithms, a distinct trend emerges, highlighting Principal Component Analysis (PCA) as a front-runner in achieving superior clustering outcomes. The discernible advantage of PCA is consistently notable across various datasets, showcasing its prowess in capturing inherent structures within high-dimensional data.

The Mouse dataset serves as a compelling case in point, where PCA not only outperforms alternative methods but also achieves a nearly flawless clustering outcome. This exceptional performance underscores PCA's adaptability and efficacy in scenarios where discerning intricate patterns is paramount.

Conversely, the comparative analysis reveals that other tested procedures exhibit varying degrees of success, lacking the pronounced clarity observed with PCA. The absence of a clear clustering optimum in these methods prompts a nuanced examination of their capabilities and limitations, emphasizing the intricate interplay between algorithmic approaches and dataset characteristics.

The significance of these findings extends beyond clustering efficacy, encompassing broader considerations such as interpretability, computational efficiency, and general applicability. The identification of PCA as a robust performer prompts a deeper exploration into its underlying mechanisms, providing valuable insights for practitioners navigating the complex landscape of dimensionality reduction.

As we delve into the nuances of each algorithm, understanding their strengths and limitations, this discussion forms a pivotal bridge between theoretical insights and practical considerations. It not only highlights the optimal clustering capabilities of PCA but also underscores the need for a tailored approach, recognizing that the most effective algorithm may vary depending on the specificities of the dataset in question.

VI. Related Works

Beyond the scope of algorithmic comparisons, dimensionality reduction can be further explored in many other domains. A lot of research papers have been published in related fields. Some related domains include:

1) *A Review of Dimensionality Reduction Techniques for Efficient Computation:* Herein, widely used feature extraction techniques such as PCA and feature selection techniques such as correlation, Linear Discriminant Analysis (LDA), forward selection, etc are extensively explored with reference to the usability in real-world situations. [23].
2) *Feature dimensionality reduction a review:* This work focuses on genetic algorithms and ant colony algorithms for dimensionality reduction [24].
3) *Dimensionality reduction: theoretical perspective on practical measures;* This compelling work aims to bridge the gap between theory and practice viewpoints of metric dimensionality reduction, laying the foundation for a theoretical study of more practically oriented analysis [25].

VII. Summary and Conclusion

A fundamental challenge in data processing lies in managing the vast amounts of data, as many algorithms are not optimized for high-dimensional datasets, leading to prolonged computing times or, in some cases, no results at all. To address this issue, methods for dimensionality reduction have been developed.

The dimensionality reduction methods discussed in this paper are integral to daily data processing, making a comparative analysis of their outcomes essential, particularly for applications in the realm of knowledge discovery.

The tested methods can be categorized into two groups: feature extraction and feature selection methods. It became evident that feature extraction, in particular, emerges as a potent tool for dimensionality reduction, consistently delivering superior results compared to feature selection in a majority of cases.

Notably, with Principal Component Analysis (PCA), a pure transformation of the data into a new basis, without explicit dimension reduction, often yields improved clustering results. Furthermore, PCA tends to enhance outcomes even when reducing the data, surpassing the performance of other methods. These observations underscore the efficacy of feature extraction techniques, especially PCA, in enhancing clustering and data reduction processes.

A. Open topics

While the obtained results and observations significantly contribute to the knowledge discovery process in the Computer Society and the Data Science Community, it is nevertheless imperative to acknowledge and address additional factors that play significant roles in the process of dimensionality reduction. The following topics are worth further investigation and scrutiny:

1) The impact of different clustering algorithms on the overall outcome
2) The optimal combination of techniques: Which optimal combination of dimensionality reduction algorithms will lead to the best possible transformation of high-dimensional data into low dimensional data?
3) What about scalability? How far are the today known and integrated approaches scalable to handle large-scale datasets? This could provide very interesting insights in fields such as Bio-informatics and Biomedical research on protein structures and DNA-Databases.

By delving into these topics, a more comprehensive understanding of the intertwined dynamics between dimensionality reduction and clustering algorithms can be achieved, thereby enhancing the process of knowledge discovery and innovation.

VIII. Acknowledgement

The inception and development of this paper can be traced back to the invaluable learning experiences encountered during the ***Seminar Applied Computer Science*** at HTW-Saar. I extend my sincere gratitude to **Professor Dr. Ing. Andre Miede**, whose guidance and mentorship have been instrumental throughout this academic journey. His unwavering support and guidance significantly enriched the quality of this work. His insightful suggestions and countless tips on effective research and paper writing have been integral to the refinement of my academic skills. His dedication to fostering a culture of scholarly excellence has left an indelible mark on my academic growth.

Furthermore, I am indebted to the technical reports series of HTW-Saar [26], which provided a foundational platform for exploring and articulating the concepts discussed in this paper. The seminar served as a crucible for intellectual growth and provided a stimulating environment for the exchange of ideas.

REFERENCES

[1] Y. Takane, *Constrained principal component analysis and related techniques*. Chapman and Hall/CRC, 2014.

[2] I. Borg, *Applied multidimensional scaling and unfolding*. Springer, 2018.

[3] N. K. Chandra, A. Canale, and D. B. Dunson, "Escaping the curse of dimensionality in bayesian model-based clustering," *Journal of Machine Learning Research*, vol. 24, no. 144, pp. 1–42, 2023.

[4] S. Ayesha, M. K. Hanif, and R. Talib, "Overview and comparative study of dimensionality reduction techniques for high dimensional data," *Information Fusion*, vol. 59, pp. 44–58, 2020.

[5] I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions," *SN computer science*, vol. 2, no. 3, p. 160, 2021.

[6] E. Keogh and A. Mueen, *Curse of Dimensionality*. Boston, MA: Springer US, 2017, pp. 314–315. [Online]. Available: https://doi.org/10.1007/978-1-4899-7687-1_192

[7] R. Bellman and R. E. Bellman, *Adaptive Control Processes: A Guided Tour*. Rand Corporation. Research studies. Princeton University Press, 1961.

[8] L. Samuel and A. M. Stuart, "The curse of dimensionality in operator learning," *arXiv*, 2023.

[9] K. Antosz, "Prediction model of product quality in production company: Based on pca and logistic regression," in *Flexible Automation and Intelligent Manufacturing: Establishing Bridges for More Sustainable Manufacturing Systems*, F. J. G. Silva, L. P. Ferreira, J. C. Sá, M. T. Pereira, and C. M. A. Pinto, Eds. Cham: Springer Nature Switzerland, 2024, pp. 425–432.

[10] . R. G. Meneghetti L., Demo N., "A dimensionality reduction approach for convolutional neural networks. applied intelligence," *Springer Link*, 2023.

[11] E. Schubert and A. Zimek, "Elki: A large open-source library for data analysis-elki release 0.7. 5" heidelberg"," *arXiv preprint arXiv:1902.03616*, 2019.

[12] E. Schubert, "Automatic Indexing for Similarity Search in ELKI," ser. Similarity Search and Applications - 15th International Conference, SISAP 2022, Bologna, Italy, October 5-7, 2022, Proceedings, 2022.

[13] W. Jia, M. Sun, J. Lian, and S. Hou, "Feature dimensionality reduction: a review," *Complex & Intelligent Systems*, vol. 8, no. 3, pp. 2663–2693, 2022.

[14] M. Suhaidi, R. A. Kadir, and S. Tiun, "A review of feature extraction methods on machine learning," *J. Inf. Syst. Technol. Manag*, vol. 6, no. 22, pp. 51–59, 2021.

[15] V. Charles, J. Aparicio, and J. Zhu, "The curse of dimensionality of decision-making units: A simple approach to increase the discriminatory power of data envelopment analysis," *European Journal of Operational Research*, vol. 279, no. 3, pp. 929–940, 2019.

[16] G. Dzemyda and M. Sabaliauskas, "Geometric multidimensional scaling: A new approach for data dimensionality reduction," *Applied Mathematics and Computation*, vol. 409, p. 125561, 2021.

[17] J. A. Lee, M. Verleysen *et al.*, *Nonlinear dimensionality reduction*. Springer, 2007, vol. 1.

[18] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Information Sciences*, vol. 622, pp. 178–210, 2023.

[19] I. D. Dinov, *Unsupervised Clustering*. Cham: Springer International Publishing, 2023, pp. 439–476. [Online]. Available: https://doi.org/10.1007/978-3-031-17483-4_8

[20] "Outlier scenario, gaussian 3d-cluster," https://github.com/elki-project/elki/blob/9292ec99c34af8b4ade763802d18d60df9856e26/data/synthetic/outlier-scenarios/3-gaussian-3d.csv#L4, accessed: August 2, 2023.

[21] F. E. Ayo, O. Folorunso, F. T. Ibharalu, and I. A. Osinuga, "Machine learning techniques for hate speech classification of twitter data: State-of-the-art, future challenges and research directions," *Computer Science Review*, vol. 38, p. 100311, 2020.

[22] Y. Lu, "Predicting research trends in artificial intelligence with gradient boosting decision trees and time-aware graph neural networks," in *2021 IEEE International Conference on Big Data (Big Data)*. IEEE, 2021, pp. 5809–5814.

[23] S. Velliangiri, S. Alagumuthukrishnan, and S. I. Thankumar joseph, "A review of dimensionality reduction techniques for efficient computation," *Procedia Computer Science*, vol. 165, pp. 104–111, 2019, 2nd International Conference on Recent Trends in Advanced Computing ICRTAC -DISRUP - TIV INNOVATION , 2019 November 11-12, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050920300879

[24] W. Jia, M. Sun, J. Lian *et al.*, "Feature dimensionality reduction: A review," *Complex Intelligent Systems*, vol. 8, pp. 2663–2693, 2022. [Online]. Available: https://doi.org/10.1007/s40747-021-00637-x

[25] Y. Bartal, N. Fandina, and O. Neiman, "Dimensionality reduction: theoretical perspective on practical measures," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/94f4ede62112b790c91d5e64fdb09cb8-Paper.pdf

[26] HTW Saar. (Year) Technical reports. Accessed: February 8, 2024. [Online]. Available: https://stl.htwsaar.de/forschung/technischeberichte/