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

A term paper as part of the lecture

”Applied Computer Science Seminar”

at Saarland University of Applied Sciences

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 report presents a comparative analysis of popular dimensionality reduction techniques, aiming to shed light on their strengths, limitations, and applicability in different scenarios.**

The report 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 report 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

With the course of time, text visualization has 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. 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 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. 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: "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" [3]. This term was first mentioned by Richard E. Bellman [4], 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 [5]. The reality is, however, quite different. Increasing the dimensions (attributes) of the data-set practically increases noise 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" [6]. This is done in different ways. For example, an attempt is made to preserve the total variance as much as possible with as small as possible number of attributes or, for example, you remove weakly correlated attributes [7]. One of the best-known representatives is principal component analysis (PCA). But all these methods have their strengths and weaknesses, which will be dealt with in 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." [8].

In the substantiation of this paper, CSV-Data-sets [9] retrieved from the ELKI-Project official GitHub Account shall be used.

II. Background

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). 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. 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 [10]. 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 [11]. 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 [12].

C. Clustering

Data clustering refers to the process of dividing data points into diverse groups or clusters such that data points with the most similarities find themselves within the same cluster. This is achieved through the implementation of

miscellaneous techniques such as k-Means and Density-Based Spatial Clustering of Applications with Noise (DBSCAN).

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). [13].

1) Small Dimensional Data: Figure 2 shows a data-set [14] 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.

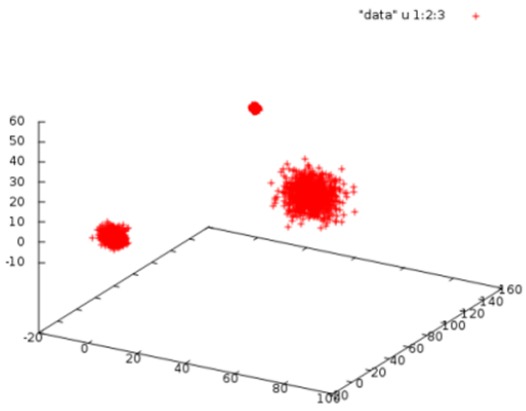

Figure 1. 3D-Visualization of a Gaussian distributed Cluster using DBSCAN and ELKI [9]

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.

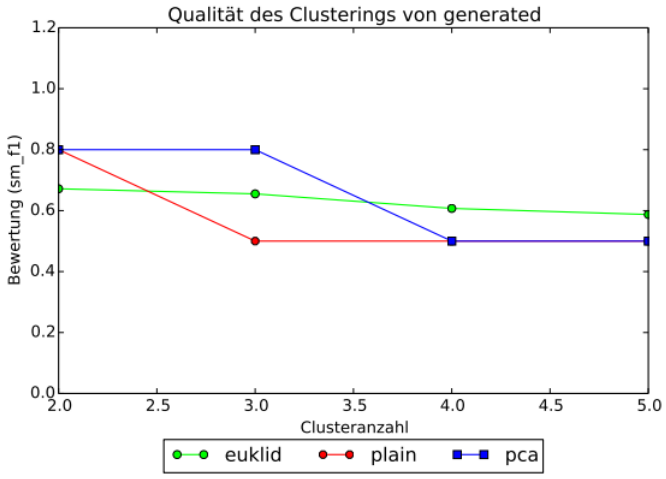

Figure 2. Quality of Clustering of generated Gaussian Distribution [9]

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.

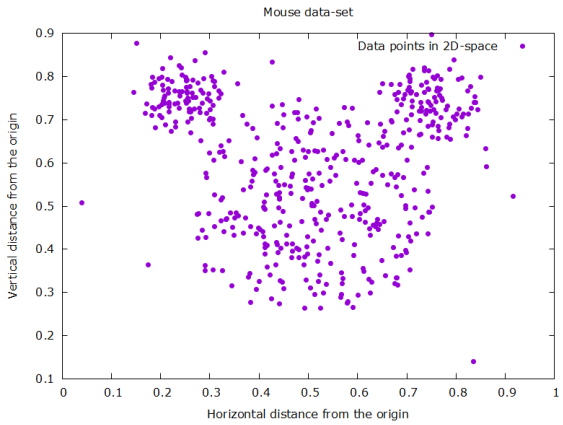

Figure 3. 2D-Visualization of clusters in the mouse data-set with GNU-Plot

IV. Evaluation and Analysis

In order to achieve optimal clustering, you first have 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.

Especially for data sets where the data is very close to each other and the variance is low, you 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.

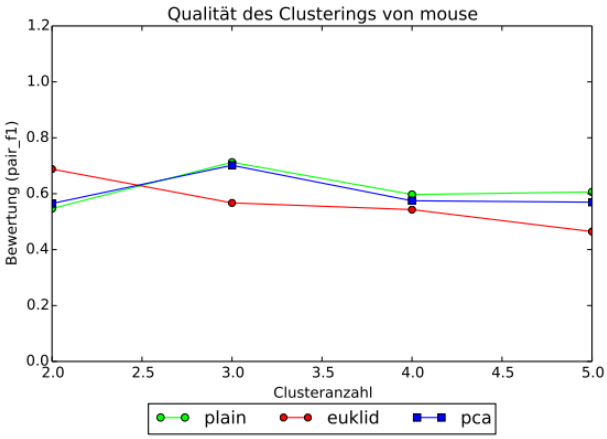

Figure 4. Quality of Clustering for the Mouse Data-set [9]

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

Table I
RESULTS OF K-MEANS AND DBSCAN

V. RESULTS AND DISCUSSION

Overall, it can be seen that the best clustering can be better identified with PCA and is usually more pronounced than with the other tested procedures. This is particularly evident in the Mouse data-set, where almost perfect clustering according to PCA was possible, whereas in all other methods there was no clear maximum.

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 such as:

1) *A Review of Dimensionality Reduction Techniques for Efficient Computation:* This paper presents widely used feature extraction techniques such as EMD, PCA, and feature selection techniques such as correlation, LDA, forward selection, etc [15].

2) *Feature dimensionality reduction a review:* This work focuses on genetic algorithms and ant colony algorithms for dimensionality reduction [16].

3) *Dimensionality reduction: theoretical perspective on practical measures;* This 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 [17].

VII. SUMMARY AND CONCLUSION

A basic problem in data processing is the large amount of data, since many Algorithms are not designed for high-dimensional data and you have to put up with a long computing time or even get no results at all. To counteract this, methods for dimensional reduction have been developed.

Algorithm	F1-Score
Plain	0.61
PCA	0.59

Table II
EFFECTS OF PCA ON THE MOUSE DATA-SET

The methods for dimension reduction presented in this paper are used every day and therefore a comparison of the results of these methods is necessary. interesting for work in the field of knowledge discovery.

The tested methods can be divided into two groups, the feature extraction and the feature selection methods. It turned out that Feature-Extraction in particular is a powerful tool for dimension reduction, that it is largely of cases delivered a better result than feature selection.

Especially with PCA, it is striking that a pure transformation of the data into a new basis without dimension reduction already an improvement of the clustering often Brings. In addition, PCA usually also brings an improvement in the case of a reduction in data of the result compared to the other methods.

A. Open topics

While the obtained results and observations significantly contribute to the knowledge discovery process in the Computer Society, 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 reductions algorithm 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 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 [18], 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] 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

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

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

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

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

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

[13] 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

[14] "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.

[15] 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

[16] 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

[17] 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

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