A Taxonomy for Data Reduction Techniques

Vitor Fernandes[1], Jorge Bernardino[1,2], Vasco Pereira[2], and Bruno Cabral[2]

[1] Polytechnic Institute of Coimbra, Coimbra Institute of Engineering (ISEC), Rua Pedro Nunes - Quinta da Nora, 3030-199 Coimbra, Portugal
[2] University of Coimbra, CISUC, Department of Informatics Engineering, Coimbra, Portugal

Abstract. Data plays a critical role in various aspects of human life, but the exponential growth in data production poses challenges to its effective processing and management while maintaining its integrity. Researchers have been actively seeking solutions to address this problem, which has led to the emergence of data reduction techniques. The analysis of existing data reduction techniques identified a gap in the lack of a taxonomy that specifically addresses data loss. This led us to propose a new taxonomy that serves as a structured framework for categorizing these techniques, providing clarity and organization. It consists of two levels, with the first level focusing on the degree of data loss associated with the techniques and the second level examining the algorithmic characteristics, such as data loss, data handling capacity, and processing efficiency. With this taxonomy, researchers and practitioners will be able to better understand and evaluate data reduction techniques and make informed decisions for effective data management and processing.

Keywords: Data Reduction, Data Reduction Techniques, Data Loss, Data Compression.

1 Introduction

Data production has reached unprecedented levels, largely due to the proliferation of IoT and Big Data environments. These technological advances have contributed significantly to the exponential growth in data generation. According to IDC [1], the number of interconnected IoT devices worldwide is estimated to exceed 50 billion by 2025. This vast network of devices is expected to generate an impressive volume of approximately 79.4 zettabytes (ZB) of data. As the world becomes increasingly interconnected and data-driven, exploiting the potential of this vast amount of data will play a critical role in driving innovation and enabling progress across multiple sectors.

Data reduction techniques are essential for optimizing storage, processing, bandwidth consumption and analysis in Big Data environments. Data reduction involves reducing the size or complexity of data while preserving its essential characteristics and minimizing information loss [2]. In Big Data [3], where large amounts of complex data are generated at massive rates, these techniques play an important role in managing and extracting value from the data. By eliminating redundancy, removing irrelevant or less important data, and compressing the data size, data reduction enables more efficient storage and analysis without sacrificing important information. These techniques are essential to overcoming the challenges posed by the volume and complexity of Big Data, enabling organizations to derive meaningful insights and make informed decisions. With the rapid growth of data across all industries, the adoption of effective data reduction techniques is becoming increasingly important to unlock the full potential of data analytics.

Some of the major challenges in data reduction are the processing effort, the time required, and the energy consumption [2]. As datasets continue to grow, efficient algorithms, and strategies are needed to enable timely processing and analysis. Researchers are exploring new ways to improve computational efficiency and develop faster data reduction methods without compromising accuracy. Another challenge is the potential data loss associated with some techniques. Researchers are trying to mitigate data loss in data reduction techniques and ensure that the reduced dataset retains its meaningful insights and value. By addressing these challenges, researchers aim to improve the effectiveness of data reduction techniques and provide robust solutions for handling large and complex datasets.

In this paper, we propose a new taxonomy that provides a systematic categorization of data reduction techniques. This taxonomy allows researchers to approach the problem in a more structured way and taking into account the specifics of each set of techniques. Additionally, it enables users or solution developers to balance the benefits and drawbacks of each set of strategies for different application scenarios.

2 Proposed Taxonomy

Data reduction techniques have emerged as a method of minimizing the size of data while preserving its value to the business. In addition, data reduction plays a critical role in improving the processing of data streams by significantly reducing the size of large original datasets.

Data reduction techniques cover a range of approaches that can vary in their data loss characteristics. For some organizations, the amount of data loss can be a critical factor in their algorithm selection and decision-making process. Given the above, we can have two types of data reduction techniques: lossless data reduction techniques and lossy data reduction techniques.

In the case of lossless data reduction techniques [4–6], the focus is on identifying redundancies, patterns, and other inherent characteristics within the data to eliminate or minimize unnecessary and repetitive information. While achieving a reduction in data size without any loss is a remarkable achievement, these algorithms are often dependent on the type of data they are analyzing. They tend to perform better on repetitive single-sensor data, such as temperature readings. For large-scale datasets these techniques tend to require a significant processing time to complete compression, which can be a significant challenge in real-time environments. All these lossless algorithms compress data and can decode the reduced algorithm back to its original size. Some examples are Delta Encoding, LZ77, LZ78, Huffman coding, and Run Length Encoding [7, 8].

When faced with the challenge of processing large amounts of data in a short period of time or dealing with highly complex datasets, the use of lossy data reduction algorithms becomes a viable option. These algorithms include a variety of methods for reducing the size of the data by selectively discarding or approximating information from the original dataset. These techniques have several advantages, including faster processing speeds, and the versatility to accommodate different data formats. Typically, the performance of these algorithms is measured by the accuracy of the compressed data set compared to the original data.

In lossy data reduction techniques [9–12], we have techniques that focus on compressing data by discarding some details that are considered less relevant or less noticeable to human perception. These typesof techniques achieve higher compression ratios than other methods by selectively removing data that has minimal impact on the overall perception or analysis of the data. Some examples of these techniques are Transform Encoding, Discrete Cosine Transform, or Fractal Compression [13–15]. In addition to lossy compression techniques, numerosity data reduction techniques [9–11]are used to reduce the amount of data by capturing the overall trend or pattern of the data. These techniques aim to represent the data in a concise and summarized form while preserving the essential characteristics and patterns. Unlike lossy compression, numerosity data reduction techniques do not intentionally discard data details, but rather summarize them to make them more manageable and efficient.

There are two main types of numerosity data reduction techniques, parametric and non-parametric. Parametric techniques [16] rely on pre-existing data models, such as Linear Regression and Log-Linear, to estimate the data and reduce its quantity. Non-parametric techniques [17, 18], on the other hand, focus on creating compressed versions of the data while preserving essential characteristics and patterns. Examples of non-parametric techniques include Sampling, Clustering, and Data Aggregation [10, 11, 17, 19].

The choice between lossy compression and numerosity data reduction techniques depends on the desired trade-off between data size reduction and retained accuracy, and on the nature of the data.

After conducting an extensive literature review, a notable gap in existing research was identified, which relates to the lack of a taxonomy that specifically addresses data loss in data reduction techniques [2, 20, 21]. Surprisingly, none of the reviewed references provided a comprehensive taxonomy that focused directly on the degree of data loss incurred by different algorithms. This gap led to the recognition of the need for a taxonomy that prioritizes data loss as the primary criterion, followed by algorithmic features.

The proposed taxonomy, shown in Figure 1, provides a structured framework for categorizing data reduction techniques. It is divided into two main levels: the first level focuses on the degree of data loss after applying the data reduction algorithm, and the second level focuses on algorithmic characteristics. On the first level, we get lossy and lossless data reduction techniques, followed by the second level, which distinguishes the algorithmic differences explained earlier. This taxonomy provides a methodological approach to organizing and understanding the various data reduction techniques available. The visual representation of the taxonomy makes it easier for researchers and practitioners to navigate and understand the different categories and subcategories of techniques. Ultimately, this taxonomy improves the clarity and accessibility of data reduction techniques, facilitating their evaluation and selection for specific data management and analysis purposes.

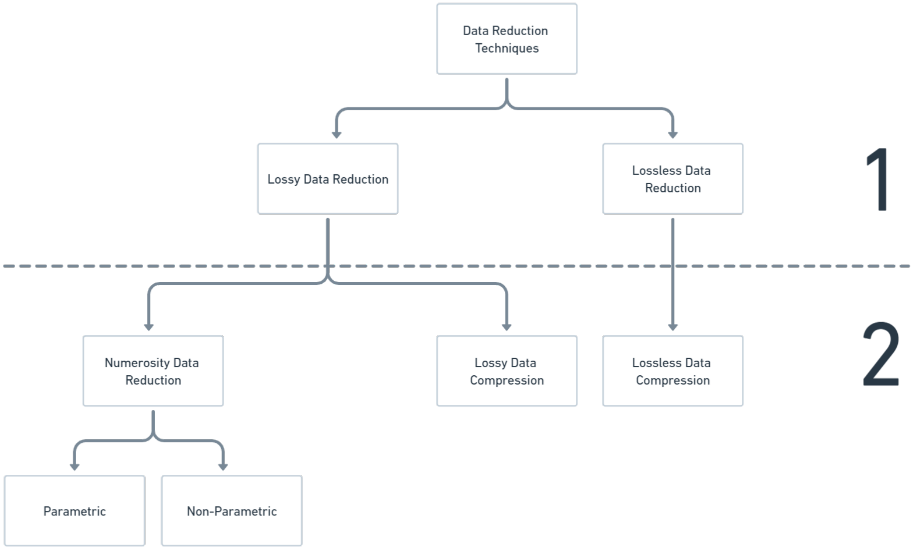

Fig. 1. Proposed data reduction techniques taxonomy.

The development of such a taxonomy would have several significant benefits. First, it would provide a clear and structured framework for classifying and categorizing data reduction techniques based on their impact on data loss. This taxonomy would enable researchers and practitioners to quickly evaluate and compare different techniques in terms of their impact on data loss, thereby facilitating informed decision-making. Second, the taxonomy would facilitate the analysis and understanding of different algorithms, allowing researchers to gain insight into their performance and limitations without having to delve into extensive technical details.

Table 1 provides a useful tool for data analysis and understanding by presenting various data reduction techniques and their respective taxonomies. By systematically organizing the techniques and highlighting their strengths and limitations, users can quickly determine the most appropriate approach for their unique data set and research objectives. This helps users make informed decisions and gain more insightful information from their analyses.

Table 1. Data reduction techniques classification.

Data reduction classification	Data reduction techniques
Lossless Data Compression Techniques	Huffman Encoding, Run-Length Encoding, Lempel-Ziv Compression (LZ77, LZ78, LZW)
Lossy Data Compression Techniques	Discrete Cosine Transform, Wavelet Compression, Cartesian Perceptual Compression, Fractal Compression
Parametric Numerosity Data Reduction Techniques	Linear Regression, Random Forest, Support Vector Machines, Gaussian Regression, Log-Linear Models
Non-Parametric Numerosity Data Reduction Techniques	K-Means, DBSCAN, Mean-Shift, OPTICS, Simple Random Sampling, Cluster Sampling

3 Conclusions and future work

In conclusion, this research highlights the critical considerations for data reduction algorithms in Big Data environments. In addition, performance requirements play a critical role, requiring careful selection of algorithms to meet application-specific needs.

The taxonomy presented is a valuable tool for efficiently identifying suitable algorithms that satisfy multiple requirements, such as data loss, algorithm processing speed, and data type. It provides some examples and comparisons between different data reduction techniques, giving researchers with a basis for further investigation.

By going deeper into each class and analyzing the performance of different techniques on different types of data, researchers can better understand the capabilities and limitations of algorithms. This research contributes to the field by facilitating informed decision-making about data reduction strategies in IoT networks.

In summary, this study provides valuable insights and guidance for future research, enabling advances in data handling and processing in the Big Data domain. By considering the taxonomy and performing in-depth analysis, researchers can make informed decisions and drive progress in this rapidly evolving field. As future work, we intend to experimentally evaluate the data reduction techniques.

References

1. Siddiqui ST, Khan MR, Khan Z, Rana N, Khan H, Alam MI (2023) Significance of Internet-of-Things Edge and Fog Computing in Education Sector.

2023 International Conference on Smart Computing and Application (ICSCA) 1–6. https://doi.org/10.1109/ICSCA57840.2023.10087582

2. Rani R, Khurana M, Kumar A, Kumar N (2022) Big data dimensionality reduction techniques in IoT: review, applications and open research challenges. Cluster Comput 25:4027–4049. https://doi.org/10.1007/S10586-022-03634-Y/FIGURES/1
3. Sagiroglu S, Sinanc D (2013) Big data: A review. Proceedings of the 2013 International Conference on Collaboration Technologies and Systems, CTS 2013 42–47. https://doi.org/10.1109/CTS.2013.6567202
4. Singh S, Devgon R (2019) Analysis of encryption and lossless compression techniques for secure data transmission. 2019 IEEE 4th International Conference on Computer and Communication Systems, ICCCS 2019 120–124. https://doi.org/10.1109/CCOMS.2019.8821637
5. Gia TN, Qingqing L, Pena Queralta J, Tenhunen H, Zou Z, Westerlund T (2019) Lossless Compression Techniques in Edge Computing for Mission-Critical Applications in the IoT. 2019 12th International Conference on Mobile Computing and Ubiquitous Network, ICMU 2019. https://doi.org/10.23919/ICMU48249.2019.9006647
6. Hanumanthaiah A, Gopinath A, Arun C, Hariharan B, Murugan R (2019) Comparison of Lossless Data Compression Techniques in Low-Cost Low-Power (LCLP) IoT Systems. Proceedings of the 2019 International Symposium on Embedded Computing and System Design, ISED 2019 63–67. https://doi.org/10.1109/ISED48680.2019.9096229
7. Nasif A, Othman ZA, Sani NS (2021) The Deep Learning Solutions on Lossless Compression Methods for Alleviating Data Load on IoT Nodes in Smart Cities. Sensors 2021, Vol 21, Page 4223 21:4223. https://doi.org/10.3390/S21124223
8. Jindal R, Kumar N, Patidar S (2022) IoT streamed data handling model using delta encoding. International Journal of Communication Systems 35:e5243. https://doi.org/10.1002/DAC.5243
9. Dias GM, Bellalta B, Oechsner S (2016) A Survey About Prediction-Based Data Reduction in Wireless Sensor Networks. ACM Computing Surveys (CSUR) 49:. https://doi.org/10.1145/2996356
10. Dias GM, Bellalta B, Oechsner S (2017) The impact of dual prediction schemes on the reduction of the number of transmissions in sensor networks. Comput Commun 112:58–72. https://doi.org/10.1016/J.COMCOM.2017.08.002
11. Reddy GT, Reddy MPK, Lakshmanna K, Kaluri R, Rajput DS, Srivastava G, Baker T (2020) Analysis of Dimensionality Reduction Techniques on Big Data. IEEE Access 8:54776–54788. https://doi.org/10.1109/ACCESS.2020.2980942
12. Abdulzahra SA, Al-Qurabat AKM, Idrees AK (2020) Data Reduction Based on Compression Technique for Big Data in IoT. 2020 International Conference on Emerging Smart Computing and Informatics, ESCI 2020 103–108. https://doi.org/10.1109/ESCI48226.2020.9167636
13. Agarwal M, Gupta V, Goel A, Dhiman N (2022) Near Lossless Image Compression Using Discrete Cosine Transformation and Principal Component Analysis. AIP Conf Proc 2481:. https://doi.org/10.1063/5.0104371/2826499

14. Ince IF, Bulut F, Kilic I, Yildirim ME, Ince OF (2022) Low dynamic range discrete cosine transform (LDR-DCT) for high-performance JPEG image compression. Visual Computer 38:1845–1870. https://doi.org/10.1007/S00371-022-02418-0/FIGURES/3
15. Pinto AC, Maciel MD, Pinho MS, Medeiros RR, Motta S F, Moraes AO (2022) Evaluation of lossy compression algorithms using discrete cosine transform for sounding rocket vibration data. Meas Sci Technol 34:015117. https://doi.org/10.1088/1361-6501/AC97FE
16. Sharanyaa S, Renjith PN, Ramesh K (2020) Classification of parkinson's disease using speech attributes with parametric and nonparametric machine learning techniques. Proceedings of the 3rd International Conference on Intelligent Sustainable Systems, ICISS 2020 437–442. https://doi.org/10.1109/ICISS49785.2020.9316078
17. Harb H, Jaoude CA (2018) Combining compression and clustering techniques to handle big data collected in sensor networks. 2018 IEEE Middle East and North Africa Communications Conference, MENACOMM 2018 1–6. https://doi.org/10.1109/MENACOMM.2018.8371009
18. Cui Z, Jing X, Zhao P, Zhang W, Chen J (2021) A New Subspace Clustering Strategy for AI-Based Data Analysis in IoT System. IEEE Internet Things J 8:12540–12549. https://doi.org/10.1109/JIOT.2021.3056578
19. Muhammad Habib, Liew CS, Abbas A, Jayaraman PP, Wah TY, Khan SU (2016) Big Data Reduction Methods: A Survey. Data Sci Eng 1:265–284. https://doi.org/10.1007/S41019-016-0022-0/FIGURES/2
20. Abdulwahab HM, Ajitha S, Saif MAN (2022) Feature selection techniques in the context of big data: taxonomy and analysis. Applied Intelligence 2022 52:12 52:13568–13613. https://doi.org/10.1007/S10489-021-03118-3
21. Ray P, Reddy SS, Banerjee T (2021) Various dimension reduction techniques for high dimensional data analysis: a review. Artif Intell Rev 54:3473–3515. https://doi.org/10.1007/S10462-020-09928-0/FIGURES/27