SAND2021-11080C

Decision Analytics in Practice: Improving Data Analytics in Pulsed Power Environments Through Diagnostic and Subsystem Clustering

Andy Yu
Sandia National Laboratories
andyu@sandia.gov

Abstract

Modern day processes depend heavily on data-driven techniques that use large datasets clustered into relevant groups help them achieve higher efficiency, better utilization of the operation, and improved decision making. However, building these datasets and clustering by similar products is challenging in research environments that produce many novel and highly complex low-volume technologies. In this work, the author develops an algorithm that calculates the similarity between multiple low-volume products from a research environment using a real-world data set. The algorithm is applied to pulse power operations data, which routinely performs novel experiments for inertial confinement fusion, radiation effects, and nuclear stockpile stewardship. The author shows that the algorithm is successful in calculating similarity between experiments of varying complexity such that comparable shots can be used for further analysis. Furthermore, it has been able to identify experiments not traditionally seen as identical.

1. Introduction

Forecasting schedule, resource, and quality of work are crucial in operational settings. Modern day processes depend heavily on data-driven techniques to help them achieve higher efficiency, better utilization of the operation, and improved decision making regarding sequence of production and supporting processes (e.g. procurement, maintenance, etc.) [1]. These techniques often require large datasets clustered into relevant groups to confidently drive decision; in manufacturing environments, datasets are gathered over many cycles and grouped by similar products to model expected performance [2]. However, building these datasets is challenging in environments where complex, low-volume technologies produce less or even erroneous data (e.g. research operations) [3]. Furthermore, both small and imperfect groupings can impact the efficacy [4] and accuracy [5] of advanced analytical methods, like machine learning, inhibiting the use of state-of-the-art approaches to analyze these environments.

Research operations are an example of an environment that produces fewer overall data. There is a need for new approaches to improve operations and achieve higher efficiency to meet research mission needs. Employing analytic techniques inspired by both image processing and systems engineering techniques is one such approach that will be described in this paper as a practice-based contribution that can be considered in similarly data-constrained settings. This study shows how using analytics in research environments can improve understanding of complex, data-constrained operations. A feature vector for a pulsed power facility was developed and found similarities in scientifically novel fusion experiments.

2. Background

The contextual setting for this study is a nuclear fusion research facility. The U.S. Department of Energy's National Nuclear Security Administration has a relevant application of a data-constrained operation that often struggles to gather more or comprehensive datasets. The Z Facility (hereafter "Z") is the world's largest pulsed power facility that routinely supports research in inertial confinement fusion, radiation effects, and nuclear stockpile stewardship through the execution of experimental pulse power operations [6]. Z undergoes 140-160 operational cycles (hereby "shots") per year [7] with some shots containing multiple experiments and requiring multiple days to execute. Thus, compiling data from shots across one year only results in a relatively small dataset.

2.1 Anatomy of a Z Shot

Although each shot is unique and nuanced, most depend on 4 key components: (1) the Z facility itself, (2) an experiment target, (3) subsystems, and (4) diagnostics. The facility contains various components and systems that ensure electrical energy (pulsed power) is delivered to the target and protect the facility. The target is an assembly comprised of materials, geometries, and systems used for scientific inquiry. After the facility delivers power, the target will either

produce x-rays or neutrons. Subsystems can be fielded to modify targets (e.g. heating or cooling) before and during execution such that different behaviors are exhibited for study. Diagnostic systems measure the environments created and their effects, but various factors (e.g. radiation type, diagnostic survivability, etc.) limit which systems are fielded and what data can be captured. This has led to the proliferation of many heterogenous diagnostic systems with varying degrees of operational complexity, often fielded in different permutations and combinations [8]. To put this in perspective, a single Z shot routinely fields over 20 diagnostic and subsystems.

2.2 Data Collection Challenges at Z

Gathering data at Z is imperfect and challenging. Each data point (e.g. shot) is novel, sometimes requiring specialized planning, design work, hardware fabrication, and preparation of data collection equipment (diagnostics). Shots are also destructive in nature, with parts of the shot either vaporized or scattered throughout the test chamber. Coupled with the radiological and airborne hazardous, gathering postmortem data can be difficult. Such highly variable, limited, and imperfect datasets affect whether operational organizations can properly group similar shots and learn from previous shots. Failure to understand historical performance and potential correlation with future shots has led to operational errors being repeated. These errors affect quality of execution and associated delays cause a lower rate of shots. Z operational organizations have historically relied on non-analytic strategies to analyze and group similar shots; sources typically include tribal knowledge from subject matter experts, amongst others. The subjectively determined metric of similarity has resulted in grouping shots by 8 scientific programs [9] [10].

While this approach has been somewhat successful in grouping shots that require minimal coordination or resources, it fails to account for more complex, nuanced shots [11]. In the last 6 years, 503 out of 883 (57%) shots were classified as "complex" for Z operations. Moreover, the percentage of complex shots is expected to increase year over year [12]. Thus, there is pressure to develop an analytical approach in finding similar shots, including an analytically based definition for degree of similarity.

The author proposes the use of Shot boundary detection (SBD) [13] to find similar shots over a variety of complexities in data-constrained environments like Z operations. This technique, borrowed from image processing and systems engineering [14], would enable stakeholders to acquire larger and more appropriate datasets for analysis and find similarity in areas not previous under consideration.

3. Method

As described above, the majority of shots executed at the facility are complex and expected to increase in quantity. In order to determine analytical approaches to similarity, the author considered the anatomy of a given shot and determined common component categories. These categories were used to develop a feature vector, to serve as the basis to calculate a similarity metric, and to organize into clusters. Due to challenges in obtaining facility configuration and target data, only diagnostic and subsystems will serve as part of the feature vector.

The feature vector composed of diagnostics and subsystems can be used to calculate similarity between shots. The author developed an algorithm inspired by Euclidean distance to calculate the similarity index (SI) from the features. The SI is calculated for all shots to find the highest SI for further analysis.

The algorithm is divided into three parts that are repeated across multiple shots. Assume we want to know the similarity SI between our shot of interest H_i to historical shot H_{hist}. First, we calculate Δx_i by counting the number of similar features D_j shared between H_i and H_{hist}. Next, we calculate Δy_i by counting the total number of features in H_i minus Δx_i. Finally, square Δx_i and Δy_i before dividing and finding the square root. These steps are repeated for all H_{hist} and sorted to find the highest value *SI*.

$$\Delta x_i = \sum_j^n \left(D_j | H_i \cap D_j | H_{hist} \right)$$

$$\Delta y_i = \sum_j^n \left(D_j | H_i \right) - \Delta x_i$$

$$SI = \sqrt{\frac{\Delta x_i{}^2}{\Delta y_i{}^2}}$$

All *SI* can be calculated such that a symmetric matrix SI_{lookup} can be created to lookup any given shot H_m to any shot H_n.

$$SI_{lookup} = \begin{bmatrix} SI_{11} & S_{12} & \cdots & S_{1n} \\ SI_{21} & SI_{22} & \cdots & S_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ S_{m1} & S_{m2} & \dots & S_{mn} \end{bmatrix}$$

The author notes that acquiring the requisite data to feed into the algorithm in this setting was a nontrivial

problem, but not discussed within the scope of this paper.

4. Results and Discussion

Figure 1 depicts a symmetric heatmap of the SI_{lookup} for Z shots created in MATLAB. A heatmap was used such that the matrix can visually interpreted and shared with groups within Z Operations. Given that shots are executed based on facility needs and schedule rather than similarity, the shots were reordered based on values from SI_{lookup} to improve clustering. The reverse Cuthill Mckee algorithm was used based on ease and availability rather than efficacy. Each cell represents a similarity comparison of two shots, with H_i along the x-axis and H_{hist}, a transpose of H_i, along the y-axis. Similarity values are represented by a spectrum from blue at 0%, green 50%, and yellow 100%. The diagonal is 100% similar, which is expected in a symmetrical matrix where the diagonal is a shot compared against itself. Clusters of yellow and green indicate shots that have a higher similarity with each other (shot 10 has little to no similarity with others while shot 43-50 have very high similarity with each other). The matrix is 23.4% sparse (e.g. no similarity based on represented feature vector) as indicated by the clusters of blue.

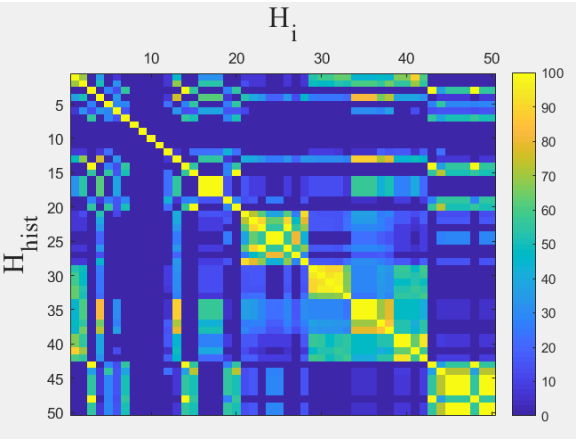

Figure 1: SI_{lookup} for 50 Z Shots

4.1 Insights from SI_{lookup}

The algorithm was able to identify similar shots not previously considered related. Despite the high number of combinations, useful and surprising patterns emerged using a clustering technique in our small dataset. Multiple (yellow) clusters in Figure 1 indicate that some shots reuse the same, or similar, sets of features. This is unexpected given the small number (883) of shots compared to the million potential combinations. Further analysis of these findings determined that those shots were not considered part of the same grouping using the historical methods described in Section 2.

A more important finding from this analysis is identifying critical overlap between different shots, represented by green clusters. These clusters suggest that many combinations are subsets or mixtures of other shots, which recharacterizes a complex and unique shot as a combination of several simple ones. This is consistent with operational practices of combining multiple experiments into one shot. The ability to separate these amalgamations is valuable in a data constrained environment where grouping similar products is not readily apparent. For reference, nearly 33.8% of the matrix has a SI between 25%-75%.

Additionally, stakeholders can adjust tolerances for acceptable similarity values to determine their sensitivity on statistical or machine learning tools to influence their goodness of fit. This type of sensitivity analysis can be used to refine acceptable values of similarity for different conditions/complexity levels.

4.2 Limitations and future work

Although some degree of sparsity was expected based on the type of shots executed, it has never been documented at Z. A sparsity of 23.4% is surprisingly low given the number of discoveries made at Z. This could be the result of a limited feature vector using only diagnostics and subsystems. Although the results show some similarity between shots, it does not account for target designs, facility configurations, or other potential variables that could increase the goodness of clustering.

Additionally, this paper only explores one clustering algorithm. There are other schemes that could improve clustering by taking advantage of the sparsity or additional interesting findings regarding similarity. Additional work is needed to compare the results across different clustering algorithms.

Future research can further the contributions of this study by exploring generalizability and goodness of small datasets like those collected from Z. Traditionally, imperfect datasets and outliers are excluded from analysis. However, the approach presented in this paper acknowledges that, despite initial impressions of dissimilarity, some datasets are actually identical to one another when viewed as combinations of smaller datasets. Research in selecting acceptable similarity values as a function of complexity, variability, and risk could help reduce over and/or underfitting of tools such as machine learning.

5. Conclusions

This paper presents a novel approach to analyzing fusion research operations through quantification of

similarity for pulsed power shots. The approach consists of creating a feature vector of diagnostics and subsystems, developing a similarity metric, and clustering for similar shots. The similarity algorithm has been successful in calculating similarity between shots of varying complexity such that comparable shots can be used for further analysis.

Not only can it identify identical shots formerly seen as dissimilar, but the timescale to perform this activity can also be compressed from several weeks to seconds. Furthermore, the outputs of this study were used to develop other analytical tools that ingest this information and estimate operational needs and risks for both future and hypothetical shots. A similar activity performed at Z that used to take several months of close coordination can now be executed within minutes. Other sectors (healthcare, space, and defense) that face similar financial, operational, political, and/or resource challenges in acquiring data could benefit from this approach.

6. Acknowledgments

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

7. References

[1] J. Harding, M. Shahbaz, Srinivas and A. Kusiak, "Data M A Review," *ASME. Journal of Manufacturing Science and I* 4, pp. 969-976, 2006.

[2] S. J. Qin, "Process data analytics in the era of big data," *American Institute of Chemical Engineers,* vol. 60, no. 9, pp. 3092-3100, 2014.

[3] F. A. Souza, R. Araujo and J. Mendes, "Review of soft sensor methods for regression applications," *Chemometrics and Intelligent Laboratory Systems,* vol. 152, pp. 67-79, 2016.

[4] J. G. A. Barbedo, "Impact of dataset size and variety on the effectiveness of deep learning and transfer learning for plant disease classification," *Computers and Electronics in Agriculture,* vol. 153, pp. 46-53, 2018.

[5] S. Dutta and E. Gros, "Evaluation of the impact of deep learning architectural components selection and dataset size on a medical imaging task," in *Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications*, Houston, Texas, 2018.

[6] D. Sinars, "Pulsed Power Science and Applications on Sandia's Z Machine," in *Euro-Asian Pulsed Power Conference*, Changsha, China, 2018.

[7] J. S. Lash, "The Sandia Z Machine: an overview of the world's most powerful pulsed power facility," in *SNL Connects Event with UT-Austin*, Austin, TX, 2017.

[8] R. G. McKee, "The Z Machine Supporting The Fundamental Science Program," in *Z Fundamental Science Program Workshop*, Albuquerque, NM, 2017.

[9] C. J. Bourdon, "Z Machine Overview," in *Z Fundamental Science Program Workshop*, Albuquerque, NM, 2019.

[10] J. Lash, "Z Update and Future," in *REHEDS ERB Meeting*, Albuquerque, NM, 2016.

[11] C. J. Bourdon, "Z Facility Strategy/Progress on Increasing Shot Rate," Radiation Effects and High Energy Density Sciences Research Foundation External Review, 15-17 May 2018.

[12] P. G. Lakshmi and S. Domnic, "Walsh–Hadamard transform kernel-based feature vector for shot boundary detection.," *IEEE Transactions on Image Processing,* vol. 23, no. 12, pp. 5187-5197, 2014.

[13] K. Sinha, S.-Y. Han and E. S. Suh, "Design structure matrix-based modularization approach for complex systems with multiple design constraints," *The International Council of Systems Engineering,* vol. 23, no. 2, pp. 211-220, 2019.

[14] C. J. Bourdon, "Diagnostics at the Z Facility: Current and Planned Diagnostic Capabilities," in *Z Fundamental Science Workshop*, Albuquerque, NM, 2015.