Proceedings of the 2024 Winter Simulation Conference
H. Lam, E. Azar, D. Batur, S. Gao, W. Xie, S. R. Hunter, and M. D. Rossetti, eds.

AGGLOMERATIVE CLUSTERING OF SIMULATION OUTPUT DISTRIBUTIONS USING REGULARIZED WASSERSTEIN DISTANCE

Mohammadmahdi Ghasemloo[1] and David J. Eckman[1]

[1]Dept. of Industrial and Systems Engineering, Texas A&M University, College Station, TX, USA

ABSTRACT

We investigate the use of clustering methods on data produced by a stochastic simulator, with applications in anomaly detection, pre-optimization, and online monitoring. We introduce an agglomerative clustering algorithm that clusters multivariate empirical distributions using the regularized Wasserstein distance and apply the proposed methodology on a call-center model.

1 INTRODUCTION

Outputs of a simulation model typically correspond to key performance indicators (KPIs) of interest to the decision maker, e.g., profit, throughput, or service level. For stochastic simulation models, simulating a given scenario generates outputs that vary from replication to replication, thus each scenario has an associated probability distribution describing the stochastic behavior of its outputs. Common tools for analyzing simulation output data include summary statistics (e.g., sample means, variances, and covariances) and visualization (e.g., histograms and boxplots). For problems with multiple KPIs, the multivariate empirical distribution produced by the data contains valuable information about system performance, but can be difficult to analyze and plot. To reveal important patterns and relationships that cannot be detected by conventional data analysis methods, we propose clustering the empirical distributions of simulated scenarios. Traditional clustering methods often fail to capture the distributional characteristics due to their reliance on simple distance metrics. We introduce an agglomerative clustering approach for simulation outputs using regularized Wasserstein distance, which enhances both accuracy and computational efficiency by leveraging the theory of optimal transport. We discuss the use cases and versatility of clustering simulation output distributions.

Anomaly Detection In simulation experiments, anomalies can be artificial or systemic. Artificial anomalies are associated with logic or coding errors, while systemic anomalies arise from inherent features of the system. When using hierarchical clustering algorithms, anomalous output distributions can be identified by examining dendrograms, inter-cluster distances, and cluster sizes. When an anomaly is detected, the simulation code is first scrutinized to determine if the anomaly is artificial; if it is not, further investigation might be conducted on the marginal distributions, correlation matrices, and input variables.

Pre-Optimization In practical scenarios, multiple KPIs often present tradeoffs, making it difficult for the decision maker to define good versus bad performance a priori. Clustering output distributions and obtaining barycenters allows a decision maker to compare a manageable number of distributions, which can be further pared down through closer examination. Clustering analysis can also aid in identifying which metrics should appear as objectives or constraints in simulation-optimization problems and setting achievable thresholds based on performance outcomes. By examining inputs associated with promising clusters, decision makers can identify regions in the input space from which to initiate optimization searches.

Online Monitoring This application addresses how simulation outputs are influenced by state variables, which evolve over time and are observable but not directly controllable. An online monitoring framework is proposed, where clustering is performed offline and state variables are tracked online (in real time), with classification algorithms being used to predict the cluster to which an observed state's output distribution belongs and thus anticipate changes in system performance. If the classification algorithm

struggles to assign a state to a cluster, it suggests impending changes in system performance, prompting possible intervention.

2 METHODOLOGY

To measure the dissimilarity between distributions, we employ the regularized Wasserstein distance. The regularized optimal transportation map between two empirical distributions μ and μ' having probability vectors p_μ and $p_{\mu'}$, respectively, denoted by $\boldsymbol{\gamma}^*_\lambda$, is defined as follows:

$$\boldsymbol{\gamma}^*_\lambda := \underset{\boldsymbol{\gamma}_\lambda \in \Pi(p_\mu, p_{\mu'})}{\operatorname{argmin}} \langle \mathbf{D}, \boldsymbol{\gamma}_\lambda \rangle - \lambda E(\boldsymbol{\gamma}_\lambda), \tag{1}$$

where λ is the regularization parameter, $\mathbf{D} \in \mathbb{R}^{M_\mu \times M_{\mu'}}$ is a cost matrix consisting of the pairwise distances between points in the supports of μ and μ', and $\langle \cdot, \cdot \rangle$ denotes the summation of the element-wise product of two matrices. The map $\boldsymbol{\gamma}^*_\lambda$ is calculated using an iterative procedure introduced in Benamou et al. (2015), and the clustering algorithm uses $\langle \mathbf{D}, \boldsymbol{\gamma}^*_\lambda \rangle$ as the distance between distributions.

Agglomerative clustering is a hierarchical method that begins by treating each instance as an individual cluster and successively merges the closest pairs based on a specified distance metric, allowing clusters to form organically from the data. This method is advantageous for several reasons. It excels in situations where the optimal number of clusters is unknown, unlike k-means clustering which requires a predefined number. Additionally, centroid-based methods such as k-means are sensitive to outliers, which can significantly skew the centroid's location and distort the clustering process. In contrast, the complete-linkage approach in agglomerative clustering considers the maximum distance between points in distinct clusters, making it more robust to outliers and resulting in tighter, more spherical clusters. Furthermore, agglomerative clustering provides a valuable output in the form of a dendrogram. This dendrogram depicts the distances between clusters at each stage of the merging process, which can help to comprehend relationships between instances and determine an appropriate number of clusters. Unlike the k-means algorithm, agglomerative clustering does not require repeated calculations of cluster centroids.

After clustering the distributions, we propose using the regularized Wasserstein barycenter to summarize each cluster. The regularized Wasserstein barycenter is a discrete distribution that minimizes the average regularized Wasserstein distance between itself and each distribution in the cluster, effectively acting as an "average" of distributions. The barycenter is computed using an iterative algorithm introduced in Benamou et al. (2015).

3 NUMERICAL EXPERIMENTS

We tested our method on a call-center simulation model that simulates customer arrivals, call routing, and services, incorporating variability in arrival times and service times. Our clustering approach successfully identifies distinct patterns in the simulation outputs, aiding in the determination of optimal staffing levels to achieve acceptable performance across five performance metrics, including mean waiting time of the customers and the mean overwork time of the operators. Additionally, we demonstrate the effectiveness of the online monitoring framework by tracking queue lengths in real time and predicting the performance metrics over the next hour. Future research directions include clustering simulation output distributions in a streaming-data setting and clustering simulation sample paths, which can provide deeper insights into dynamic system behavior.

REFERENCES

Benamou, J.-D., G. Carlier, M. Cuturi, L. Nenna and G. Peyré. 2015. "Iterative Bregman Projections for Regularized Transportation Problems". *SIAM Journal on Scientific Computing* 37(2):A1111–A1138.