SYM-AM-21-060

EXCERPT FROM THE PROCEEDINGS OF THE EIGHTEENTH ANNUAL ACQUISITION RESEARCH SYMPOSIUM

Topological Data Analysis in Conjunction with Traditional Machine Learning Techniques to Predict Future MDAP PM Ratings

May 11–13, 2021

Published: May 10, 2021

Prepared for the Naval Postgraduate School, Monterey, CA 93943.

ACQUISITION RESEARCH PROGRAM
GRADUATE SCHOOL OF DEFENSE MANAGEMENT
NAVAL POSTGRADUATE SCHOOL

The research presented in this report was supported by the Acquisition Research Program of the Graduate School of Defense Management at the Naval Postgraduate School.

To request defense acquisition research, to become a research sponsor, or to print additional copies of reports, please contact any of the staff listed on the Acquisition Research Program website (www.acquisitionresearch.net).

ACQUISITION RESEARCH PROGRAM
GRADUATE SCHOOL OF DEFENSE MANAGEMENT
NAVAL POSTGRADUATE SCHOOL

Topological Data Analysis in Conjunction with Traditional Machine Learning Techniques to Predict Future MDAP PM Ratings

Brian B. Joseph—is a Senior Operations Research Analyst in the Office of the Under Secretary of Defense for Acquisition and Sustainment (OUSD (A&S)), Acquisition Data Analytics (ADA). In this role, Joseph serves as the Deputy Director, Data Analytics supporting the Principal Deputy to the Assistant Secretary of Defense, Acquisition Enablers regarding acquisition data, data science, metrics, and statistics. Prior to this, Joseph served as the Statistician, Security Policy & Oversight Division (SPOD), Counterintelligence and Security Directorate, Office of the Under Secretary of Defense for Intelligence. In this role, Joseph served as Principal Action Officer to the Director, SPOD regarding DoD personnel security investigations (PSI) metrics. Joseph earned his MS in Cost Estimation and Analysis from the Naval Postgraduate School, MS in Applied Statistics from the Rochester Institute of Technology, BS in Environmental Management from the University of Maryland University College, and an AS in Mathematics from Northern Virginia Community College. He is currently pursuing a PhD in Data Science from Northcentral University. [Brian.b.joseph.civ@mail.mil]

Trami Pham—is a Data Scientist and Senior Consultant at Booz Allen Hamilton, with experience in developing machine learning models using defense acquisition. Currently, Pham supports the Acquisition, Data, and Analytics Division of OUSD(A&S) by building predictive models to forecast program health. Pham earned her BA in Applied Mathematics from the University of California, Berkeley. [Trami.d.pham.ctr@mail.mil]

Christopher Hastings—is a Data Analyst and Consultant for Artlin Consulting, specializing in defense acquisitions. He works within the Acquisition, Data, and Analytics division of OUSD(A&S), providing analyses on Major Defense Acquisition Programs (MDAPs) and Middle Tier of Acquisition (MTA) programs. He earned his BA in mathematics and economics from the University of Virginia. [Christophe.p.hastings2.ctr@mail.mil]

Abstract

Topological data analysis (TDA) is an unconventional machine learning technique that is used to understand the underlying topology of data. The premise is that data has shape. The two methodologies used in TDA are persistent homology and the mapper algorithm. Traditional machine learning techniques include supervised unsupervised methods such as clustering, Bayesian networks, neural networks, support vector machines (SVM), and random forests. The goal of this study is to apply TDA methods in conjunction with traditional machine learning algorithms to Defense Acquisition Executive Summary (DAES) data to determine if TDA helps to improve prediction measures (accuracy, f-measure, sensitivity, and specificity) over using traditional methods only when predicting program manager ratings from Major Defense Acquisition Programs (MDAPs). We show that TDA when used in conjunction with traditional machine learning models at a local level of the DAES data improved the accuracy of predicting PM cost ratings of MDAPs at 80% of all nodes in training and testing as compared to implementing these models without TDA at the global level.

Keywords: Topological data analysis, machine learning, prediction measures

Background/Research/Business Need

The Data Analytics Division of Acquisition Enablers (AE) within the Office of the Under Secretary of Defense for Acquisition and Sustainment OUSD(A&S) has been developing machine learning model minimal viable products (MVP) to assist in prioritizing analysts' focus on which major defense acquisition programs (MDAPs) may become problematic. Human resources have been reduced in A&S to perform analytic tasks of determining problematic programs in the program assessment process in the Acquisition Data Analytics Division of the

AE Directorate since the reorganization of OUSD Acquisition, Technology, and Logistics into OUSD(A&S). As such, prioritizing problematic programs using machine learning models efficiently assists analysts in performing program assessment for executive leadership. There is anecdotal evidence that has shown that TDA, when used in conjunction with traditional machine learning models, improves overall accuracy of these machine learning models at localized sections of the data. SymphonyAI (2021) in a white paper discusses how traditional machine learning models use global optimization that assumes/guesses the shape of the data to derive parameters to approximate the dataset which often produces errors in some regions of the data. TDA in contrast creates separate models of the underlying data based on the output network topology that is responsible for different local sections of the data. This technique produces a better representation than a single globalized model. Therefore, we wanted to test whether this localized modeling methodology of TDA is more efficient and improves accuracy of predicting program manager ratings in DAES data.

Machine Learning

Machine learning is binned into unsupervised and supervised learning. Unsupervised learning uses methods such as clustering to segment data into smaller datasets and dimensionality reduction to make it easier to visualize data that are high dimensional (e.g., 25 or more features). Clustering models include hierarchical and K-Means. Supervised learning consists of regression and classification models. The classification models used to assist in the prioritization effort are neural networks, random forests and single tree models, and SVM.

Supervised Learning Classification Models

Random forests are an ensemble technique analogous to bagging trees. It works by collecting a bootstrapped sample of identical and independently distributed trees and conducting recursive partitioning on them. Classification is based on a majority vote of the aggregated trees. The beauty of this technique is that it obtains an estimate of the misclassification error and also performs random feature selection to estimate the relative importance of the explanatory variables (Friedman et al., 2009).

Support vector machines are large-margin powerful predictive models that can be utilized for classification or regression. They are a class of distance-based classifiers that attempt to use hard margins for stability in classification. They can be linear or nonlinear in form. The beauty and utility of SVM is the implementation of kernel methods that transform vectors from the input space and calculate their inner products in the feature space therefore bypassing the calculation of the function Φ in the input space, which would be untenable. This allows the SVM to perform classification of datasets in which the underlying boundaries of the classes are not readily clear. Some examples of kernels are the Gaussian radial basis, Laplace radial basis, and the hyperbolic tangent kernels. The use of kernels offers a rich model class to essentially tune the SVM (Clarke et al., 2009).

Neural Networks are extremely powerful classifiers as they can be tuned by many different parameters. They are also heavily nonlinear classification models. The sigmoid function ψ that defines the neural net may be modeled using the logistic, hyperbolic tangent, or heavy side step sigmoid functions. These sigmoid functions in conjunction with the size of the hidden layers offer ways to tune the neural network as a more robust classifier (Clarke et al., 2009).

TDA

TDA is an emerging and exciting form of unsupervised learning. Georges (2019) states that TDA is based on topology, a branch of mathematics that examines the notion of shape. TDA attempts to analyze highly complex data and draws on the notion that all data has a

fundamental shape and that shape has meaning. Figure 1 below is an illustration of some common shapes of data, which include regressions, clusters, flares, and loops.

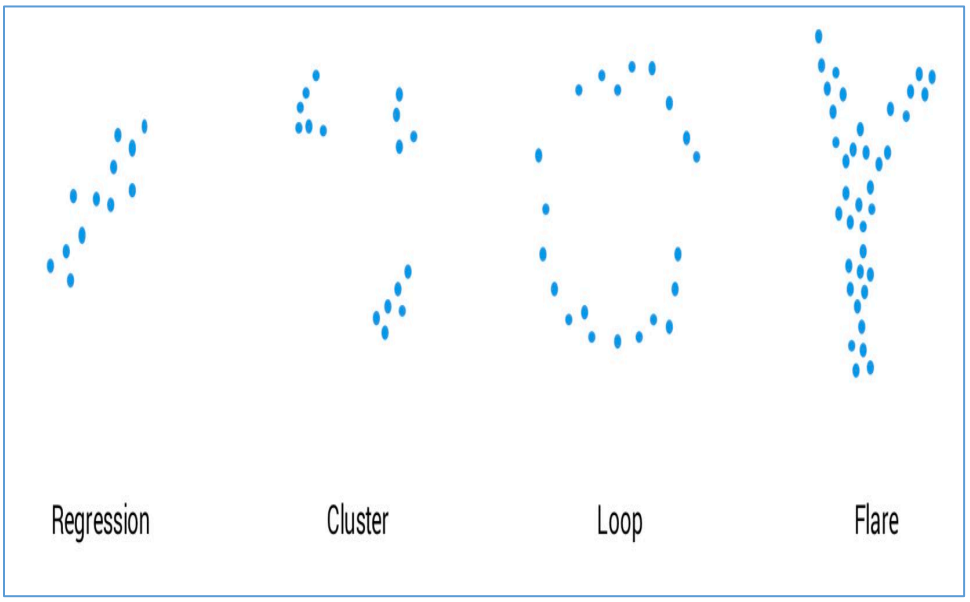

Figure 1 . Common Shapes of Data (Ayasdi, 2020)

The two methodologies used in TDA are persistent homology and the mapper algorithm. Persistent homology provides a framework and efficient algorithms to quantify the evolution of the topology of a family of nested topological spaces. Persistent diagrams are used to capture and visualize the birth and death of homological features over a specific period of time (Fasy et al., 2015). The mapper algorithm is a tool used to visualize the topology of the data under consideration. This method of TDA will be used for this research. The inputs to the algorithm are a point cloud of data, a filter function, a covering of a metric space, a clustering algorithm, and tuning parameters. Figure 2 depicts an illustration of the mapper algorithm and filter function.

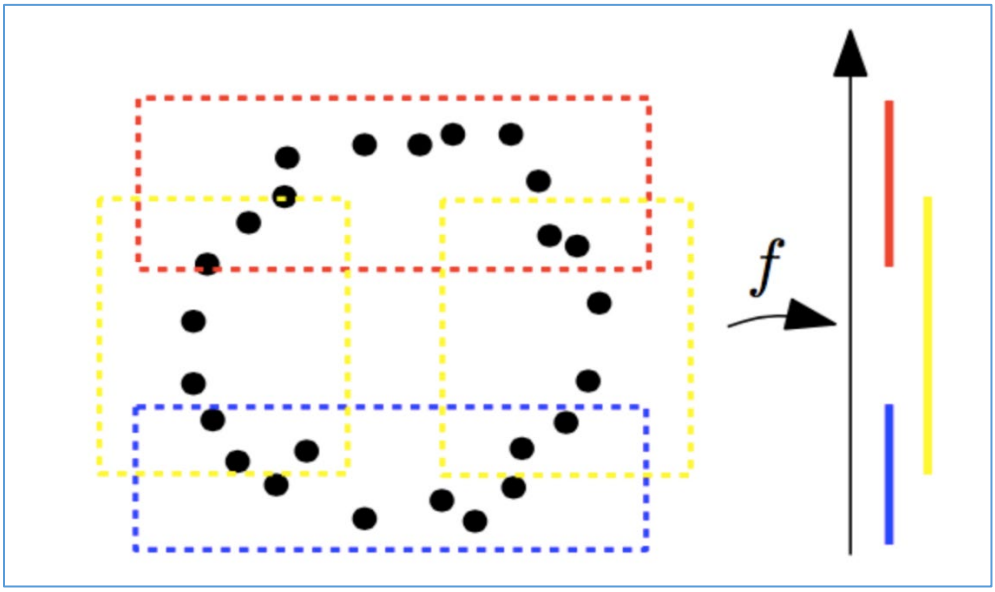

Figure 2. Mapper Algorithm and Filter Function (Chazal & Michel, 2016)

The output is a network graph that represents the topology of the data (Herring, 2018). Figure 3 below illustrates the steps to implement the mapper algorithm. It shows that the notional data to be mapped is a hand. Next, a filter function is identified. In this research, a kernel distance estimator will be used as the filter function. Third, determine the number of overlapping bins to map the input data. In this case, six bins are selected. Finally, create a network topology representation of the original dataset using nodes and edges (Lum et al., 2013). The nodes represent the clusters of local regions created by the binning. It is important to note that information from one node can be contained in another node as a result of overlapping bins. The edges connect clusters to display the overall topology.

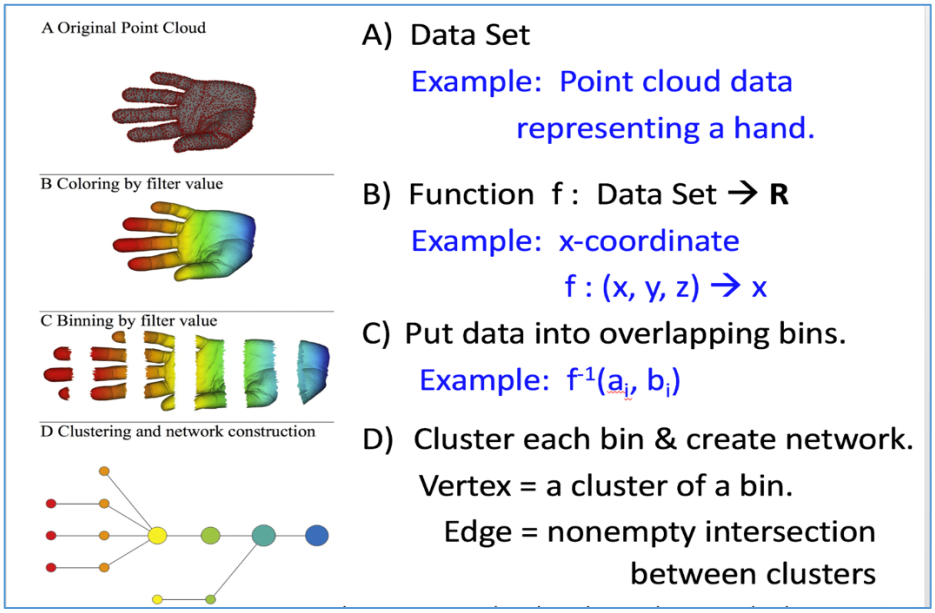

Figure 3. Implementation Steps of Mapper Algorithm (Lum et al., 2013)

Research Question

Can TDA in conjunction with traditional machine learning models improve the accuracy of the predictions of those machine learning models when used without TDA?

Hypothesis

H_0: Traditional machine learning algorithms (neural network, random forest, recursive partitioning, and SVM) have higher predictive accuracy when combined with TDA in at least 70% of nodes for training and testing sets.

H_a: Traditional machine learning algorithms (neural network, random forest, recursive partitioning, SVM) have higher predictive accuracy when not combined with TDA in at least 70% of nodes for training and testing sets.

Related Work

Chazal & Michel (2016) demonstrate how to use the mapper algorithm in R's TDAMapper package to construct topologies of any data set into network graphs, as well as how to label the categories of each node by a specific color to assist with understanding the data's topology better. Riihimaki et al. (2020) used a TDA classifier to determine if it provided better accuracy than a SVM classifier when modeling repeated measures data. The results of this experiment are that their TDA classifier outperformed the SVM classifier in accuracy 96.8%

to 68.7% respectively in one of three use cases. Kindelan et al. (2021) used persistent homology to build a TDA classifier that provided superior accuracies on eight separate data sets than traditional k-NN classifiers. Wu and Hargreaves (2020) implemented a TDA classification model on mixed data (numerical and categorical) using persistent homology of heart disease data. The results were that the TDA classification model performed better in accuracy than traditional state-of-the-art machine learning models such as decision trees, logistic regression, naïve Bayes, neural networks, single trees, and SVM in predicting heart disease. Joseph and Sconion (2020) used sentiment analysis to extract average sentiment of selected acquisition report executive summaries to determine if the average sentiment was highly correlated to be viable as a predictor feature/variable in predicting unit cost growth of MDAPs. Joseph and Hastings (2020) derived new schedule features/variables (months to threshold, difference from current to next DAES, difference from previous to current DAES, and previous milestone slips) from schedule milestone and APB schedule data gathered from DAES data to predict and understand the factors that may cause schedule slips in MDAPs.

Methodology

Four traditional machine learning classification models are initially applied to DAES data in order to predict future program manager cost ratings. PM cost ratings are the target variable and 10 other attributes (consisting of schedule, unit cost, and average sentiments of DAES executive summaries) are used as features for these models. The classification models used in this research are neural network, random forest, recursive partitioning (single tree based), and SVM. The accuracies of these models are recorded. Next, TDA is applied to the same DAES data using the mapper algorithm in R programming language to create a network topology of the data. This is an implementation of the localized modeling discussed above. The contents of each resulting network node of the TDA model are then modeled using the previous traditional machine learning classification models, and the resulting accuracies of each model are compared to the results of the globally optimized machine learning models when not used in conjunction with TDA to determine if accuracies improve more at the local node level over the global level of the DAES data. The null hypothesis is tested, and conclusion is drawn to answer the research question.

Data Collection and Preprocessing

Data for this research was collected from Defense Acquisition Management Information Repository (DAMIR) and the Defense Acquisition Visibility Environment databases. Unit cost, schedule, PM rating, and DAES executive summary data was extracted separately from the database and then joined by PNO, Schedule URI. Next, the data was cleansed to remove missing values. The next step was to remove unnecessary html tags from the executive summary and PM rating explanation text variables. The average sentiment variable was derived from previous research conducted by Joseph and Sconion (2020). Schedule slip features were derived from research conducted by Joseph and Hastings (2020). Further cleaning of text was done using R programming language's TM package to remove punctuations, stop words, conduct stemming, and convert all words to lower case to remove duplication during future text classification analysis. Average sentiment was extracted from DAES executive summaries using the sentimentR package and R programming language. The final dataset contained 10 feature variables, one target variable (PM cost rating), and 4,000 rows of non-missing entries of DAES data.

Analysis

Classification without TDA

Globally optimized supervised machine learning using the four classification models discussed above were implemented on the DAES data set with PM rating for cost as the target variable. Tables 1 and 2 show the confusion matrix outputs for the SVM model. The training set produced an accuracy of 79.3% while the test set provided a 73.7%. This is consistent with typical training and test sets. The accuracies of the training set are usually higher than those of the test set. The training accuracies for the random forest, recursive partitioning, and neural network models are 99.1%. 64.1%, and 60.6% respectively. The testing accuracies for the random forest, recursive partitioning, and neural network models are 98.3%, 62.6%, and 56.7% respectively.

Table 1. Confusion Matrix SVM Training

	Green	NoRating	Red	Yellow
Green	444	20	56	76
NoRating	63	618	17	19
Red	22	16	503	24
Yellow	130	16	92	551
Accuracy = 79.3%				

Table 2. Confusion Matrix SVM Testing

	Green	NoRating	Red	Yellow
Green	195	19	30	40
NoRating	55	293	17	10
Red	16	7	241	26
Yellow	75	11	44	254
Accuracy = 73.7%				

Classification with TDA

The TDA mapper algorithm was implemented on the data set in R programming language using the following parameters: a sample size of 4,000 rows of data with 10 features, a Euclidean distance similarity function, the kernel distance estimator (KDE) filter function, and bins with 10 intervals overlapping at 50%. Figure 4 illustrates the resulting network graphing output from the mapper algorithm in R programming language. Figure 4 also depicts that the network shape of the underlying original DAES data is a regression type. Other renderings were flare shaped in some iterations prior to this final rendering. The node numbers are from left to right.

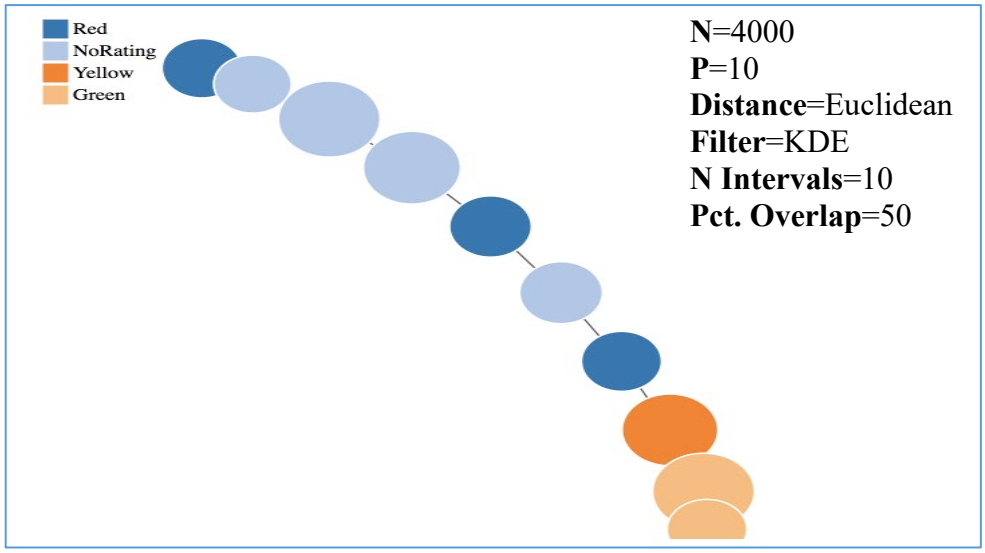

Figure 4. Network Topology Output of Mapper Algorithm in R of DAES Data (Shape Regression)

Figure 5 illustrates the number of rows of data assigned to each node from the mapper algorithm. There are 10 nodes because 10 bins were requested in the input of the mapper algorithm parameters. We notice that the sum of the row contents does not sum to the 4,000-sample size. This is due to the 50% overlap in the binning where some row IDs of one node may be included in other nodes. An extraction of that row ID information can give context to how each node can be described by an analyst and subject matter expert of the data.

Nodegroup	Nodesize	PM_Rating_Cost.maj.vertex	filter.kde
1	561	Red	0.001011081
2	529	NoRating	0.004631899
3	1028	NoRating	0.005935219
4	922	NoRating	0.007293477
5	607	Red	0.009355363
6	625	NoRating	0.011357024
7	575	Red	0.013117226
8	891	Yellow	0.015866479
9	1030	Green	0.016917012
10	570	Green	0.018434909

Figure 5. R output of TDA Mapper Network Graph Nodes

Tables 3 and 4 depict the confusion matrix and accuracy produced by implementing the SVM machine classification model on the contents of node 1 of the resulting TDA mapper algorithm network topology output data. Tables 5 and 6 depict the confusion matrix and accuracy produced by implementing the SVM classification model on the contents of node 10 of the resulting TDA mapper algorithm network topology output data. In both cases, the accuracy results of SVM when used with TDA at the local level is an improvement over the accuracy of the SVM model when implemented globally on the data set. The results of the other classification models can be found in Table 7.

Table 3. Confusion Matrix SVM TDA Training Node 1

	Green	NoRating	Red	Yellow
Green	54	1	3	2
NoRating	3	42	0	0
Red	8	0	179	10
Yellow	10	2	2	58
Accuracy = 89.0%				

Table 4. Confusion Matrix SVM TDA Testing Node 1

	Green	NoRating	Red	Yellow
Green	22	1	1	76
NoRating	2	27	1	19
Red	5	0	87	24
Yellow	3	3	0	551
Accuracy = 85.6%				

Table 5. Confusion Matrix SVM TDA Training Node 1

	Green	NoRating	Red	Yellow
Green	133	4	10	11
NoRating	1	18	0	0
Red	3	0	41	2
Yellow	11	0	15	131
Accuracy = 85.0%				

Table 6. Confusion Matrix SVM TDA Testing Node 10

	Green	NoRating	Red	Yellow
Green	64	3	6	11
NoRating	1	7	0	0
Red	1	0	16	0
Yellow	9	0	9	63
Accuracy = 78.9%				

Results

Table 7 shows the results of implementing machine learning classification models with and without TDA to predict future PM cost ratings. It can be seen that

- 80% of all training and testing models have improved accuracy when used in conjunction with TDA
- 85% of the training models from traditional machine learning methods produced improved accuracy when used in conjunction with TDA vice using the traditional methods independently
    - Random Forest model improved in 40% of the training nodes
    - All other models improved in 100% of the training nodes
- 75% of the testing models from traditional machine learning methods produced improved accuracy when used in conjunction with TDA
    - Random Forest model improved accuracy 0% of the TDA produced testing nodes
    - All other models improved accuracy 100% of the TDA produced training nodes
- Weaker learners improved in training and testing accuracy while the strongest learner (Random forest) decreased by 0.4%-6.2% accuracy in testing performance when used with TDA.
- There may be a point of diminishing returns on increased accuracy if traditional models already perform at 98% accuracy
    - Further research needed to unpack this phenomenon.

Table 7. Accuracy Results of Machine Learning With and Without TDA

Accuracy Results of Using TDA with Machine Learning Vs Machine Learning Only						
Node	**Accuracy**	**Sample Size**	**Recursive Partitioning**	**Support Vector Machine**	**Random Forest**	**Neural Network**
	Without TDA					
	Training	2,667	64.1	79.3	99.1	60.6
	Testing	1,333	62.6	73.7	98.3	56.7
	With TDA					
Node 1	Training	374	85.0	89.0	99.7	79.1
	Testing	187	83.4	85.6	96.3	80.7
Node 2	Training	353	87.2	92.4	98.0	84.3
	Testing	176	85.7	90.3	97.2	80.5
Node3	Training	685	88.5	88.3	98.1	83.1
	Testing	343	86.3	85.7	96.8	79.3
Nde 4	Training	615	87.5	84.9	98.7	86.8
	Testing	307	85.7	81.8	95.4	86.7
Node 5	Training	405	84.9	90.1	100.0	86.7
	Testing	202	76.2	82.2	92.1	83.7
Node 6	Training	417	89.9	89.2	99.8	92.1
	Testing	208	85.6	83.2	92.8	84.1
Node 7	Training	383	84.6	88.8	99.0	72.8
	Testing	192	81.8	87.0	94.3	70.8
Node 8	Training	594	84.0	84.7	97.8	79.6
	Testing	297	78.1	84.2	92.6	69.3
Node 9	Training	687	85.7	86.5	98.7	80.0
	Testing	343	81.9	76.7	94.2	67.9
Node 10	Training	380	86.1	85.0	100.0	88.6
	Testing	190	83.1	78.9	94.7	80.0
	Accuracy Increase With TDA Over Without TDA					
Node 1	Training	NA	20.9	9.7	0.6	18.5
	Testing	NA	20.8	11.9	-2.0	24.0
Node 2	Training	NA	23.1	13.1	-1.1	23.7
	Testing	NA	23.1	16.6	-1.1	23.8
Node3	Training	NA	24.4	9.0	-1.0	22.5
	Testing	NA	23.7	12.0	-1.5	22.6
Nde 4	Training	NA	23.4	5.6	-0.4	26.2
	Testing	NA	23.1	8.1	-2.9	30.0
Node 5	Training	NA	20.8	10.8	0.9	26.1
	Testing	NA	13.6	8.5	-6.2	27.0
Node 6	Training	NA	25.8	9.9	0.7	31.5
	Testing	NA	23.0	9.5	-5.5	27.4
Node 7	Training	NA	20.5	9.5	-0.1	12.2
	Testing	NA	19.2	13.3	-4.0	14.1
Node 8	Training	NA	19.9	5.4	-1.3	19.0
	Testing	NA	15.5	10.5	-5.7	12.6
Node 9	Training	NA	21.6	7.2	-0.4	19.4
	Testing	NA	19.3	3.0	-4.1	11.2
Node 10	Training	NA	22.0	5.7	0.9	28.0
	Testing	NA	20.5	5.2	-3.6	23.3

Conclusion and Recommendations

Based on the results of the analysis in 80% of training and testing cases, we can fail to reject the null hypothesis and conclude that traditional machine learning algorithms (recursive partitioning, SVM, and neural networks) have higher predictive accuracy when combined with TDA at least 70% of all nodes. The random forests improved accuracy 40% of time in training instances and is the only model that did not improve with TDA Mapper implementation in all cases, although it does at nodes 1, 5, 6, and 9 for the training set

Machine learning at the local network group level appears to improve classifier performance than if done solely at the global level in this use case and from literature on TDA. It is recommended that TDA be used in conjunction with traditional machine learning models when predicting targets for other acquisition-related use cases.

Continuing and Future Work

Based on the above research, my data analytics team in ADA (lead by Trami Pham) has implemented a random forest model with and without TDA to predict future PM ratings in the DoD Comptroller's Advana environment. This model has more feature variables than the MVP discussed above, so accuracy results are slightly different. Additionally, the team is working to implement long-short-term-memory (LSTM) neural network and SVM models in conjunction with

TDA. Table 8 depicts the results of the comparison of the random forest model with and without the use of TDA. The use of TDA has improved the accuracy of the random forest model by over 6.5% in all prediction periods.

Table 8. Prediction Accuracy comparison of A&S' Advanced Analytics MVP App in Advana

	30/60/90 Day Model Predictions		
	30 days/1 time step	60 days/ 2 time steps	90 days/3 time steps
Random Forest	90.9%	89.1%	90.0%
TDA + Random Forest	97.9%	97.4%	97.9%

Figure 6 is an illustration of TDA used in conjunction with a random forest classification algorithm implemented as part of the Advanced Analytics application housed in the OUSD(Comptroller) Advana environment and displays the network graph produced by the mapper algorithm. It is interactive so if one clicks on a node in the application, the contents of that node can be displayed. The confusion matrix, prediction accuracies, and other model prediction performance scores such as precision, recall, and f-measures are presented for each node.

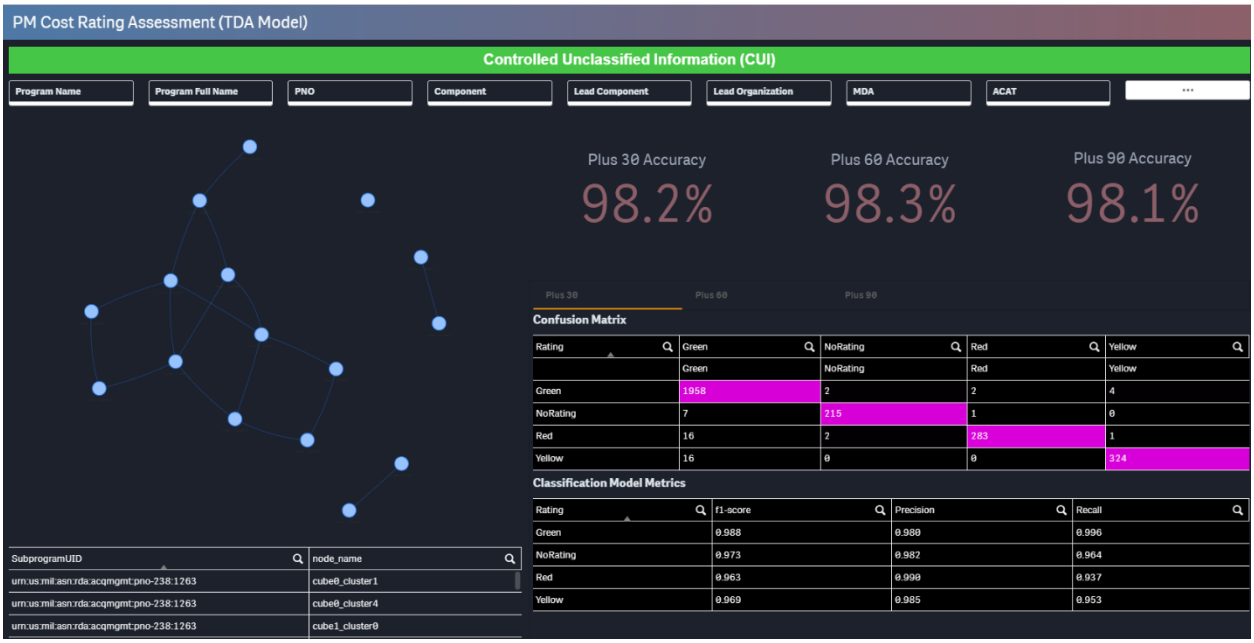

Figure 6. TDA With Random Forest Model Confusion Matrix and Network Graph Application in the USD(C) Advana Environment (Advana, 2021)

Figure 7 displays the predictions of future MDAP PM cost ratings in 30/60/90-day intervals for individual MDAPs that are currently reporting in the DAMIR/DAVE databases. As an example, it can be seen that the MQ-4 Triton is currently reporting a red PM cost rating but is predicted to turn green in 60 to 90 days. The analyst may decide based on current red and 30-day red predictions that this program may need some attention. Leadership, however, may determine that since the program is set to trend green in 60 to 90 days that it does not require attention. As another example, if programs are currently rated green and are projected to trend green over the 30/60/90-day time horizons, there is no need for the analyst or leadership to waste valuable time in conducting a program assessment for that MDAP. Better use of their time can be used prioritizing those programs that are green and yellow and trending to red.

Figure 7. Actual PM Cost Rating 30/60/90-Day Predictions for MDAPs in the Advanced Analytics Application of the Acquisition Analytics Dev Stream in Advana (Advana, 2021)

Finally, we are investigating the use of TDA to predict duration lengths in MTA programs. Besides improving the accuracy of machine learning models, we also plan to use the TDA to understand the relationships and topology of MTA program data.

References

Almgren, K., Kim, M., Lee, J. (2017). Extracting knowledge from the geometric shape of social network data using topological data analysis. *Entropy 2017*, *19*(7), 360. https://doi.org/10.3390/e19070360

Ayasdi. (2020). *AyasdiAI vs. open source TDA*. https://www.ayasdi.com/wp-content/uploads/2020/07/Ayasdi-TDA-vs.-Open-Source-TDA-06.03.20-1.pdf

Ayasdi. (2020). *Understanding Ayasdi: What we do, how we do it, why we do it* [White paper]. https://s3.amazonaws.com/cdn.ayasdi.com/wp-content/uploads/2018/04/04142230/UnderstandingAyasdi_WP_061617v011.pdf

Chazal, F., & Michel, B. (2016). *Mapper algorithm with the R-package TDAmapper*. http://bertrand.michel.perso.math.cnrs.fr/Enseignements/TDA/Mapper.html

Clarke B., Fokoue E., & Zhang, H. H. (2009). *Principles and theory for data mining and machine learning*. Springer.

Elyasi, N., & Moghadam, M. (2019). *An introduction to a new text classification and visualization for natural language processing using topological data analysis*. https://arxiv.org/pdf/1906.01726v1.pdf

Farrelly, C. (2018). *Topological data analysis for data professionals: Beyond Ayasdi*. https://www.kdnuggets.com/2018/01/topological-data-analysis.html

Fasy, B., et al. (2015). *Introduction to the R package TDA*. https://cran.r-project.org/web/packages/TDA/vignettes/article.pdf

Friedman J., Hastie T., & Tibshirani, R. (2009). *The elements of statistical learning. data mining, inference, and prediction* (2nd ed.). Springer.

Georges, A. (2019). *Topological based machine learning methods* [Dissertation]. https://escholarship.org/uc/item/4vr8963d

Guo, W., & Banerjee, A. G. (2017). Identification of key features using topological data analysis for accurate prediction of manufacturing system outputs. *Journal of Manufacturing Systems*.

Herring, A. W. (2018). *The Mapper algorithm*. https://jdc.math.uwo.ca/TDA/Herring-Mapper.pdf

Joseph, B., & Hastings, C. (2020). Using ANOVA and multinomial logistic regression to analyze defense acquisition executive summary (DAES) and acquisition program baseline (APB) milestone estimates to determine contributing factors to schedule slips. *Proceedings of the 17th Annual Acquisition Research Symposium*. 1–20.

Joseph, B., Hastings, C., & Houston, K. (2020). *Updates to selected analyses from the performance of the defense acquisition system series: 2019 SARs update*. https://www.acq.osd.mil/aap/assets/docs/PDAS%202019%20Excerpts_Final%20-cleared.pdf

Joseph, B., & Sconion, D. (2020). Using natural language processing, sentiment analysis, and text mining to determine if text in selected acquisition report executive summaries are highly correlated with major defense acquisition program (MDAP) unit costs and can be used as a variable to predict future MDAP costs. *Proceedings of the 17th Annual Acquisition Research Symposium*. 1–11. https://www.acq.osd.mil/aap/assets/docs/SYM-AM-20-061_Panel11_Joseph_Paper_4-13-2020.pdf

Kim, J., et al. (2015). *Tutorial on the R package TDA*. http://www.stat.cmu.edu/topstat/topstat_old/Talks/files/Jisu_150623_TDA_tutorial.pdf

Kindelan, R., Cerda, M., Frias, J., & Hitchsfeld, N. (2021). *Classification based on topological data analysis*. https://arxiv.org/pdf/2102.03709v1.pdf

Levine, P. (2018). *Annual weapons system assessment. GAO needs to step up its game*. https://www.ida.org/-/media/feature/publications/a/an/annual-weapons-acquisition-assessment---gao-needs-to-step-up-its-game/p-10692.ashx

Lum, P. Y., Singh, A., Lehman, T., Ishkanov, M., Vejdemo-Johansson, M., Allagapen, M., Carlsson, J., & Carlsson, G. (2013). Extracting insights from the shape of complex data using topology. *Scientific Reports*, *3*(1236). http://www.nature.com/srep/2013/130207/srep01236/full/srep01236.html

Riihimaki, H., Chacholski, W., Theorell, J., Hilbert, J., & Ramanujam, R. (2020). A topological data analysis based classification method for multiple measurements. *BMC Bioinformatics*, *21*(336). https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-020-03659-3

SymphonyAI. (2021). *Topological data analysis and machine learning: Better together*. https://www.symphonyai.com/wp-content/uploads/2020/07/SAI-Topological-Data-Analysis-and-Machine-Learning-Better-Together-vf.pdf

Wu, C., & Hargreaves, C. A. (2020). *Topological machine learning for mixed numeric and categorical data*. https://arxiv.org/pdf/2003.04584.pdf

Acquisition Research Program
Graduate School of Defense Management
Naval Postgraduate School
555 Dyer Road, Ingersoll Hall
Monterey, CA 93943

www.acquisitionresearch.net