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

Brian B. Joseph
Deputy Director, Data Analytics, OUSD(A&S)/ASD(A)/AE/ADA
brian.b.joseph.civ@mail.mil
https://www.acq.osd.mil/aap/#/da
For 18th Annual Acquisition Research Symposium
4/5/2021

Unclassified

Research Question

- Can TDA methods increase the predictive accuracy of traditional machine learning algorithms to improve initial MDAP PM Ratings for Cost?
  - 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.

Research Issue/Business Need

- ADA is developing machine learning models to assist in prioritizing which MDAPs need to be included in program assessments
- Literature and anecdotal evidence illustrate that TDA can improve the prediction accuracy of machine learning algorithms at the local level vice the global level
- Based on promising results of TDA usage, the purpose of this research is to determine the feasibility of integrating TDA with machine learning algorithms to understand the underlying topology of DAES data and to more accurately predict future MDAP PM ratings.
- Specifically, can TDA be used in conjunction with traditional machine learning methods to improve the performance of these models?

Unclassified

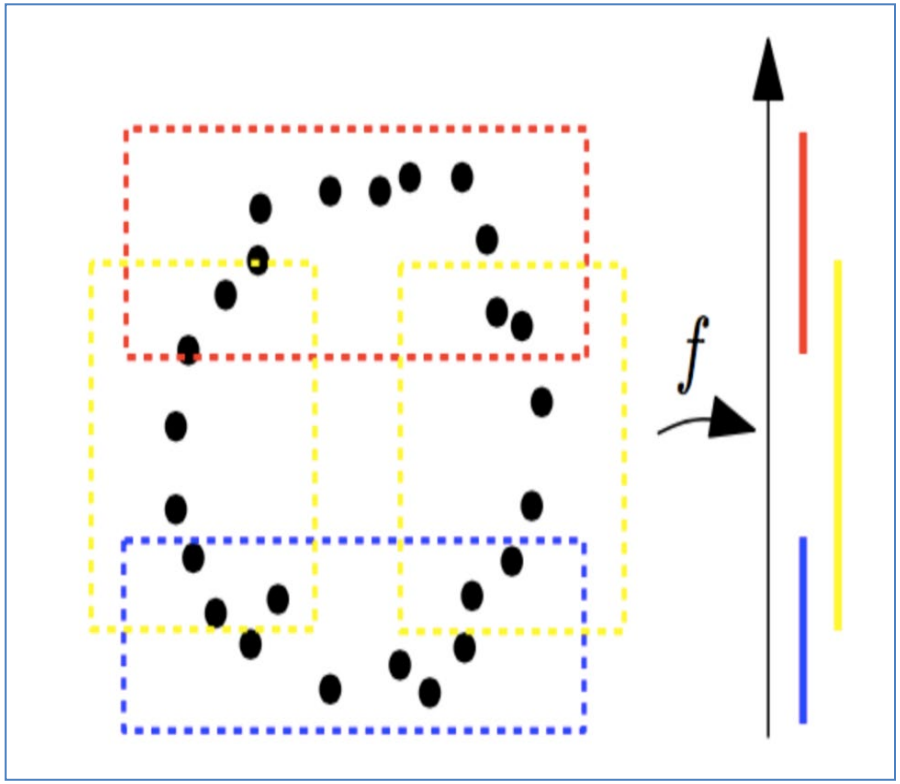
Topological Data Analysis

About TDA

Mathematical method used to study shape of data. Types of TDA Methods

- Persistent Homology
  - Captures births and deaths of of features over time
- Mapper Algorithm (Used for analysis in this briefing)
  - Graphically visualize topology using networks
  - components
    - Input data cloud
    - Filter function
    - Metric space
    - Clustering algorithm
    - Tuning parameters
      - Number intervals
      - Percent overlap
      - Distance measure

Figure of Mapper Algorithm Filter Function



Methodology

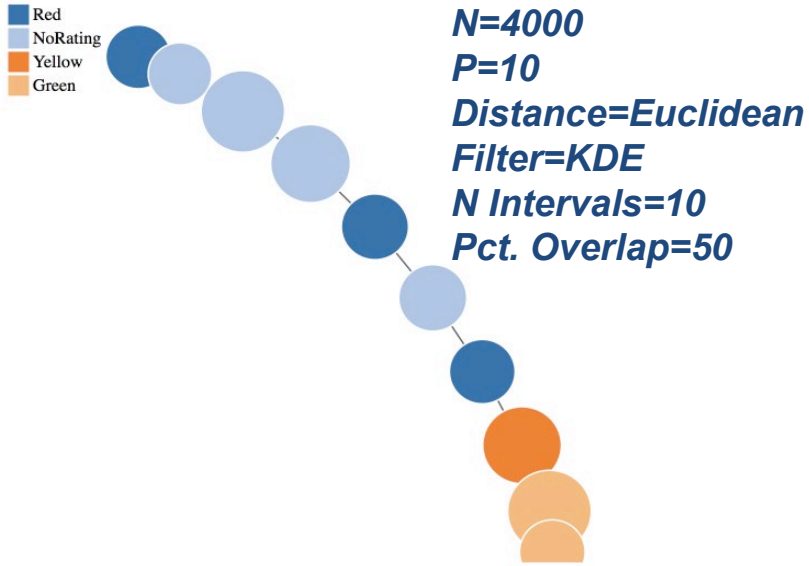
- Conduct traditional supervised machine learning on the DAES data set by setting the PM Rating Cost variable as the dependent (target) variable and set average sentiment, schedule variables, and unit cost variables as the independent variables for classification.
  - Measure Accuracy of each method without conjunction of TDA
- Use Mapper Algorithm in *R* to conduct TDA
  - Use Kernel Distance Estimator as filter function
  - Output TDA network graph
  - Use data from all 10 nodes of the output TDA network to conduct traditional machine learning techniques based on TDA implementation. Can be thought of as localized machine learning of rows of data at specified nodes.
  - Measure accuracy of each traditional method on the 10 nodes in conjunction with TDA

Analysis (TDA Mapper Algorithm Implemented)

Resulting Mapper Algorithm Node information

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

Resulting Mapper Algorithm Network Topology of DAES Input Data Cloud (Regression like shape topology)



N=4000
P=10
Distance=Euclidean
Filter=KDE
N Intervals=10
Pct. Overlap=50

Unclassified

Results

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

Results (Continued)

- 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.
- May be a point of diminishing returns on increased accuracy if models already perform at 98% accuracy
  - Further research to unpack

Conclusion & 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, support vector machine, and neural networks) have higher predictive accuracy when combined with TDA
- Random Forest algorithm only model that does not improve with TDA Mapper implementation in all cases
- Machine learning at the local network group level appears to improve classifier performance than if done solely at the global level.
- Use TDA in other acquisition use cases when implementing machine learning models