Predictive Analysis on Multivariate, Time Series datasets using Shapelets

Hemal Thakkar | hemal@stanford.edu | hemal.tt@gmail.com
Department of Computer Science, Stanford University

A: Motivation

This project is to implement current, innovative research on an industrial problem of Multivariate, Time Series Batch production dataset. Most research in predictive analysis for such problems mainly focuses on techniques like PCA and Dynamic Time Warping, thus ignoring most of the temporal variances in Time Series sequence. This project implements predictive techniques using shapelets (Temporal subsequences, representative of a class) on a real industrial dataset from chemical manufacturing. Along with predicting the quality of outcome of a batch the focus is also to build scalable algorithm and identifying landmarks on a running time series batch

B: Dataset

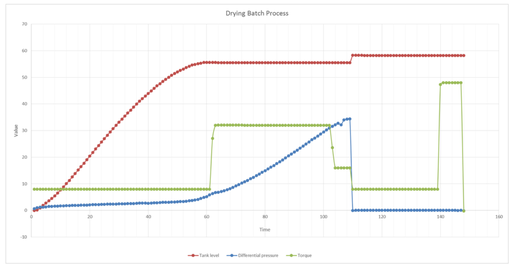
Domain: Chemical Manufacturing process
Independent variables: for each of the B=32 batches b_i i=1 to 33

Data	Description
$z_i \in R^k$	Initial chemical measurements at time t =0. k=11
x_i^n	n^{th} time series trajectory of batch i. Each batch contains n=10 time series trajectories

Output/Dependent variables:

Data	Description
$y_i \in R^j$	Final chemical measurements at time t = T (end of batch). j=11. Note n, j and k may represent different measurements
$Q_i \in \{1,-1\}$	Final quality of product. 1 indicates On specification, acceptable quality and -1 indicates off specification i.e. unacceptable

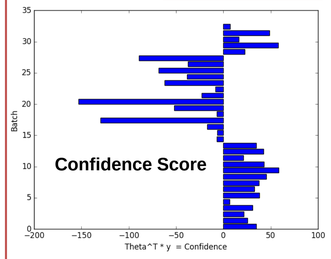
A typical Batch run showing 3 trajectories of x. z measured at t=0 and y, Q measured at t=T.



C: Learning Pipeline

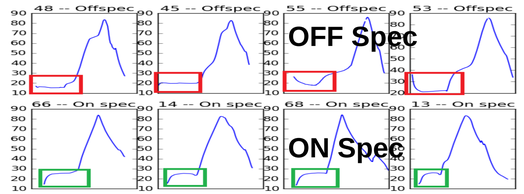
Derive Confidence Score

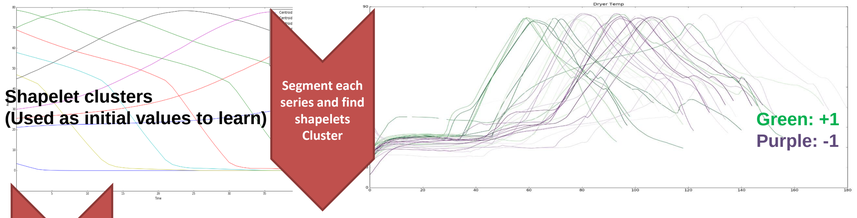
$\{y_i\}$ → Logistic Regression → Scores



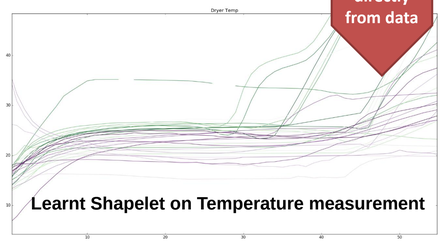
Shapelet discovery and Intuition

$\{x_i^n\}$ Scores → Visualize Shapelet and presence of landmarks





"Learn" Shapelets directly from data



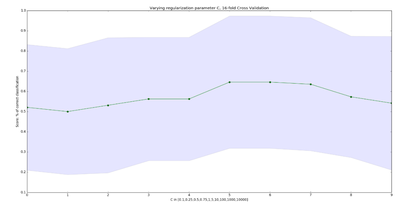
Learnt Shapelet on Temperature measurement

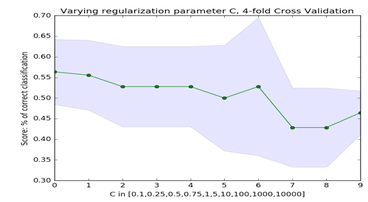
"Select" High Entropy shapelets → $\{z_i\}$

Estimate quality using supervised Learning → $\{\hat{q}_i\}$

Classify Batch using initial measurements z and shapelets

D: Cross Validation, Feature Selection and Parameter tuning





E: Summary

Technique	Test Accuracy	Comment
Supervised learning using z_i (Linear Regression)	0.56	Just Slightly better than random guessing, not a usable model
Classification learning using z_i (Logistic Regression)	0.64	Better model, but still low accuracy. Proves that initial condition do not fully determine the final quality measure.
Classification learning using z_i + shapelets (Logistic Regression)	0.75	Improved accuracy, with low generalization error

Other Accomplishments	
Improved visualizations and Landmark identification using shapelet on a running Batch (Online Visualization)	When to use Dynamic time Warping (DTW) vs Shapelet ? [See DTW distance Matrix (right)] DTW Matrix (Purple = Similar batches, Green = Dissimilar) shows cross dependency between classes. If these are linearly separable (Purple box across diagonals) use DTW else use Shapelet.

Dynamic Time Warp Distance Matrix
Time Warped distance between batches

F: References

1. Pattern Recognition and Classification for Multivariate Time Series
2. Multivariate Time Series Classification by Combining Trend-Based and Value-Based Approximations
3. Forecasting performance of multivariate time series models with full and reduced rank: an empirical examination
4. Detection and Characterization of Anomalies in Multivariate Time Series
5. Machine Learning Strategies for Time Series Prediction
6. Querying and mining of time series data: experimental comparison of representations and distance measures
7. Trouble-shooting of an industrial batch process using multivariate methods
8. Multivariate SPC Charts for Monitoring Batch Processes
9. Time Series Shapelets: A New Primitive for Data Mining
10. Learning Time-Series Shapelets
11. CS229 Notes Supervised Learning : Classification
12. CS229 Notes Regularization and model selection