MC-ANN: A Mixture Clustering-Based Attention Neural Network for Time Series Forecasting

Yanhong Li, David C. Anastasiu

Santa Clara University

BayLearn Bay Area Machine Learning Symposium

Problem Description

Problem: solving a demanding univariate time series forecasting challenge, dealing with a non-stationary series characterized by significant variance and containing extreme events.

$$[x_1, x_2, \ldots, x_T] \in \mathbb{R}^T \rightarrow [x_{T+1}, \ldots, x_{T+H}], \in \mathbb{R}^H$$

x_1 to x_T : the input sequence

x_{T+1} to x_{T+H} : the output sequence

*In our research: for reservoir water level problem, T = 15 * 24 = 1440, H = 3 * 24 = 72.*

Challenges:

- Long-range dependencies.
- Rare but important extreme values.

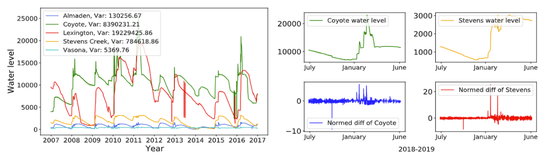
Goal:

- An end-to-end model concurrently learns extreme and normal prediction functions.
- Long sequence forecasting (predicted length = 72).

Dataset:

- 9 reservoirs in California. Prediction: 3 days, hourly, 72 points. Over 31 years of sensor data, 276,226 values.
- Reservoirs: multipurpose, vast bodies of water. Function: flood control, navigation, irrigation, energy production, human safety and welfare.

Extreme Events



Sensor	Min	Max	Variance	Skewness	Kurtosis
Original data					
Almaden	-0.10	1714.08	1714.08	-0.27	-0.76
Coyote	2319.42	27421.25	13304899.11	1.30	3.33
Lexington	867.88	20109.10	21374524.46	0.50	-0.36
Stevens Creek	93.92	3229.98	713429.64	-0.20	-0.93
Vasona	140.15	619.48	5044.46	-1.05	0.67
First-order difference data					
Almaden	-408.51	417.11	16.13	0.10	3541.97
Coyote	-10262.42	11782.86	5865.46	6.10	8985.19
Lexington	-10852.95	7841.56	4122.16	-18.42	11104.88
Stevens Creek	-1395.84	1092.97	49.75	-22.78	21075.51
Vasona	-90.29	74.39	2.45	3.41	438.38

High skewness and kurtosis scores indicate that there is significant deviation from a normal distribution in our data!

Motivation

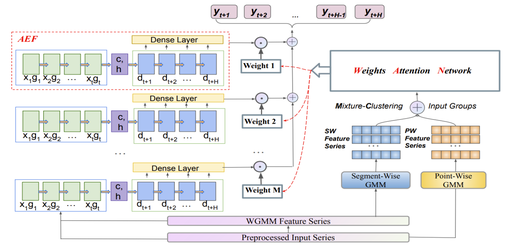
Achieving the best overall prediction performance, without sacrificing either the quality of normal or of extreme predictions.

Root Mean Square Error ***(RMSE)***

Mean Absolute Percentage Error ***(MAPE)***

Proposed Framework

- **End-to-End Mixture Clustering-Based Model**

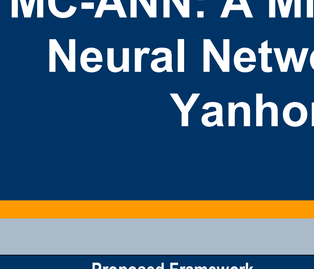


MC-ANN learns the time series data distribution, as a mixture of Gaussian distributions on both point-wise and segment-wise levels, consisting of two parts:

1)Grouped Auto-Encoder based Forecaster (AEF) and

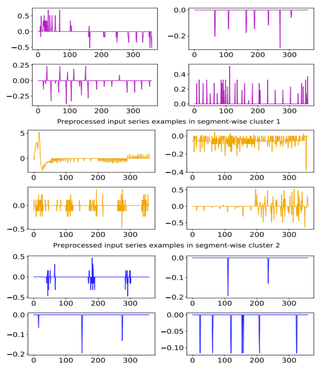
2) GMMmixture clustering-based learnable Weights Attention Network (WAN) for disentangling extreme values from normal ones.

PW: Point-Wise Statistical Clustering Features



Point-wise clustering feature series: initial water levels (top). Values after preprocessing and the WGMM feature series (middle). PW features series for a GMM model with M = 3 components (bottom).

SW: Segment-Wise Statistical Clustering Features

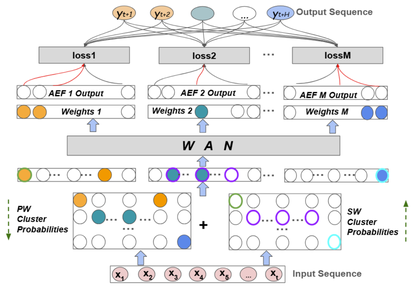


To capture long temporal patterns.

Segment-Wise Clustering (SW GMM) Feature Series is to extend the Gaussian Mixture Modeling process from individual time points to contiguous sequences.

Enable the discovery of repeated temporal motifs or shape patterns across the time series.

Attention Weights-guided Backpropagation



(WAN) employs an attention mechanism to enhance predictive accuracy through adaptive loss computation.

These attention weights are then used to weight the prediction outputs from the AEF.

Allowing the network to adaptively emphasize or de-emphasize different.

Auto-regularized Loss Function

$$\mathcal{L}_i = \hat{AEF}_i \odot Weights_i, \ \forall\, i \in \{1, 2, \ldots, M\},$$

$$\mathcal{L} = RMSE\left(\sum_{i=1}^{M} \mathcal{L}_i, y\right).$$

Motivations:

- Feed the combined probability matrix derived from both pointwise and segment-wise clusters to WAN to generate loss weights.
- These weight vectors are then applied to the outputs of the AEF, contributing to the loss computation and influencing the final prediction.

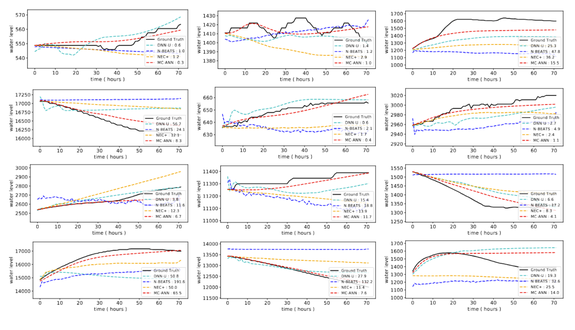
Baselines

- NEC+
- iTransformer, which applies the attention and feedforward network on the inverted dimensions
- Informer
- Fedformer
- Attention-LSTM
- N-BEATS, renowned for its superior performance on several benchmark datasets, DNN-U [41], a state-of-th
- NLinear
- Dlinear
- DNN-U

Impact of WAN and GMM Clustering Inputs on RMSE

Dataset	Type	no_WAN	no_PW	$PW_{factor} = 0.4$
Coyote	rolling	51.530	47.75	**45.373**
Coyote	3-day	472.578	435.032	**433.571**
Stevens Creek	rolling	9.051	9.470	**7.382**
Stevens Creek	3-day	80.089	**65.064**	71.303
Vasona	rolling	6.805	6.541	**5.137**
Vasona	3-day	19.792	19.032	**18.474**

Effects of Proposed Methods



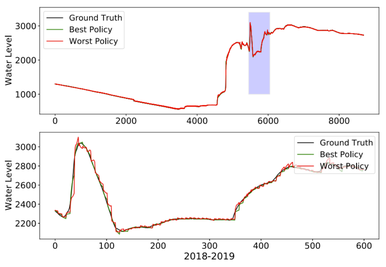
EFFECTIVENESS COMPARISONS AGAINST STATE-OF-THE-ART METHODS ON ROLLING 8-HOUR PREDICTION.

Datasets	Metric	MC-ANN	NEC+	iTransformer	Informer	NLinear	DLinear	NBEATS	DNN-U	A-LSTM
Almaden	RMSE	**7.412**	10.580	65.683	211.241	16.199	23.009	23.229	9.676	18.040
	MAPE	**0.002**	**0.002**	0.016	0.204	0.005	0.006	0.009	0.004	0.016
Coyote	RMSE	**45.373**	64.590	755.083	7437.162	385.546	346.678	159.024	117.000	1282.913
	MAPE	**0.002**	**0.002**	0.020	0.653	0.013	0.012	0.006	0.004	0.126
Lexington	RMSE	**255.739**	303.510	1600.916	9565.245	645.141	798.529	468.829	318.024	660.354
	MAPE	**0.003**	**0.003**	0.048	0.773	0.023	0.024	0.011	0.005	0.068
Stevens Creek	RMSE	**7.382**	14.977	48.256	714.468	27.380	48.692	34.998	13.363	117.497
	MAPE	**0.002**	**0.002**	0.0136	0.589	0.005	0.012	0.008	0.006	0.104
Vasona	RMSE	**5.137**	5.775	15.269	19.580	7.045	12.544	10.572	11.370	23.587
	MAPE	**0.004**	**0.004**	0.020	0.028	0.006	0.013	0.011	0.016	0.049

EFFECTIVENESS COMPARISONS AGAINST STATE-OF-THE-ART METHODS ON 3-DAYS PREDICTION.

Datasets	Metric	MC-ANN	NEC+	iTransformer	Informer	NLinear	DLinear	NBEATS	DNN-U	A-LSTM
Almaden	RMSE	**53.539**	58.117	59.272	217.641	60.516	64.596	64.764	58.648	57.649
	MAPE	**0.014**	**0.014**	0.015	0.162	0.017	0.021	0.018	0.017	0.021
Coyote	RMSE	433.571	466.276	608.228	8507.417	631.056	730.057	535.886	**417.24**	1338.622
	MAPE	**0.011**	**0.011**	0.015	0.619	0.019	0.022	0.013	0.012	0.128
Lexington	RMSE	**774.209**	794.842	926.294	11878.486	1019.081	1082.898	931.356	832.329	1050.5
	MAPE	0.015	**0.014**	0.020	0.930	0.030	0.033	0.023	0.018	0.078
Stevens Creek	RMSE	**71.303**	91.810	93.042	1052.549	90.084	99.598	89.918	76.090	156.591
	MAPE	0.013	**0.011**	0.015	0.888	0.014	0.024	0.015	0.016	0.128
Vasona	RMSE	**18.474**	20.893	18.264	22.051	20.157	20.021	21.405	20.683	32.245
	MAPE	**0.018**	0.020	0.020	0.031	0.020	0.024	0.023	0.027	0.062

A Whole Year Rolling P rediction Example



Acknowledgements

Research supported by a Supermicro GPU SuperServer SYS-420GP-TNAR+ node contributed by Supermicro and NVIDIA, integrated into the Santa Clara University HPC

Contact Information: danastasiu@scu.edu