Dr. Jack K. Rasmus-Vorrath

8/7/18

MSDS_7335: Machine Learning

Final Project Report : Time Series Clustering of Sensor Data for Radiation Profiling

The WHY:

- New technologies designed to regulate exposure of interior spaces to light and heat radiation use on-site sensor data to conduct control-structured decision making in response to real-time changes in radiation levels.
    - Foreknowledge of typical radiation profiles informs, facilitates, and expedites appropriate command signals to hardware components controlling interior exposure, ensuring timely responses to continuously changing conditions.
- Unsupervised and supervised ML can be used to develop and validate site-specific radiation profiles to improve application control logic.
    - Site-specific differences between buildings outfitted with light- and heat-sensing hardware warrant the use of both unsupervised and supervised approaches. Unsupervised approaches enable discovery of relevant radiation profiles of which domain experts may be unaware, augmenting the set of classes whose detection is at stake when applying supervised methods.

The WHAT:

- Over the course of 7 months, from August 2017 to February 2018, approximately 245K records of minute-level time series data have been gathered using a rooftop hardware unit with 13 photo sensors, placed atop a high-rise commercial building on Wall St. in New York City.
    - The raw values collected by the sensors are in Lux, which range from 0-800 Watts per squared-meter after conversion. These raw

values are preprocessed according to the required input format of the machine learning methods applied.

- Preprocessing procedures include: Min-Max Scaling, Mean-Variance Scaling, and Piecewise Aggregate Approximation
- As this use case is primarily concerned with classification of radiation profiles, the outputs of the applied machine learning methods are the identified time series cluster centroids in the unsupervised setting, and assigned labels for such centroids in the supervised setting.

The HOW:

- The project workflow consisted of: cleaning and preprocessing, exploratory data analysis, model building, fitting, and predicting, and visualizing and evaluating the results.
    - In addition to the above-mentioned preprocessing techniques, the effects of SMOTE and ADASYN oversampling procedures were explored as applied to the most performant supervised classification method.
    - As part of the validation and evaluation procedure, model building was also regularly preceded by the use of random permutation cross-validation in splitting the data into training and testing sets.
    - Exploratory data analysis included the use of the (soft-) dynamic time-warping metric in applying the K-means and Barycenter Averaging algorithms
    - Supervised algorithms included K-Nearest-Neighbors, a specialized time-series implementation of the former, called K-Neighbors, as well as a Shapelet Classifier.
        - The above algorithms were applied with a range of various distance metrics, including: Euclidean, Cosine, Minkowski, DTW, and Soft-DTW. The computation of nearest neighbors applied the Ball-Tree, KD-Tree, and Brute Force approaches.

        - Performance metrics used during evaluation included: (Subset) Accuracy, Precision, Recall, F-1, Adjusted Mutual Information, Adjusted Rand Index, Normalized Mutual Information, Homogeneity, Completeness, and the V-Measure
    - Unsupervised algorithms included K-Means++, a specialized time-series implementation of the former, called Time Series K-Means, K-Shape, as well as an LSTM Sequence-to-Sequence Autoencoder, which was used to perform unsupervised representation learning during a pretraining module, providing the previously-mentioned unsupervised approaches with a lower-dimensional projection of features from which they could more readily identify and discriminate between classes of radiation profiles.
        - The above algorithms applied distance metrics including: Euclidean, Cosine, Minkowski, DTW, and Soft-DTW.
        - Performance metrics used during evaluation included: Inertia, Silhouette, and Calinski-Harabaz

Preprocessing:

- Data cleaning begins with grouping by calendar date, as the object of analysis primarily concerns radiation profiles occurring on scale less than or equal to a single day.
    - Lux are converted to Watts per squared-meter.
    - The maximum of the 13 sensor values available at each minute-level interval is subset from the rest, as only the maximum is used in driving the control logic of the on-site hardware.
    - Night-time rows of maximum values with less than 1 Watt per squared-meter are filtered out.
    - Values are preliminarily scaled from 0 to 1 using a Min-Max Scaler.

- To facilitate comparison across radiation profiles, the minimum row length amongst the date-grouped list of data frames is used as the baseline for Piecewise Aggregate Approximation, which divides time series into a number of segments equal to the desired number of time steps (the minimum row length), before replacing each segment by the mean of its data points.
- After preprocessing, the date-grouped data frames are vertically stacked according to calendar date, resulting in a single data frame with 171 rows (days) and 514 columns (time-steps).
- For use in the supervised learning setting, ground truth classes of radiation profiles are hand-labeled using available domain knowledge of conditions typically occurring on-site. The classes assigned to each row include: (1) Sunny, (2) Cloudy, (3) Partially Cloudy, (4) Mixed Sunny and Partially Cloudy, (5) Sunny with Occlusion, (6) Partially Cloudy with Occlusion, and (7) Cloudy with Sun-spike.

Exploratory Data Analysis:

- In a first-pass visualization of possible cluster centroids occurring on a one-week time frame, (Soft-) DTW K-Means and Barycenter Averaging were applied.
- Using both the DTW and Soft-DTW distance metrics, Barycenter Averages were produced for each week for the full length (171 days) of the concatenated data frame.
    - As evident in the included interactive notebook (.IPYNB and .HTML), the hand-labeled classes were readily identifiable in the weekly Barycenters produced, e.g.: Partially Cloudy in week 3, Sunny in week 5, Mixed Sunny and Partially Cloudy in week 6, Sunny with Occlusion in week 13, Cloudy with Sun-spike in week 14, Partially Cloudy with Occlusion in week 16, and Cloudy in week 23.

Unsupervised (1): K-Means++ | Time Series K-Means Approach

- For comparison, the DTW and Soft-DTW distance metrics were used in addition to the default Euclidean metric used by K-Means++ during cluster assignment and Barycenter computation.
- A value of K=7 (the number of hand-labeled classes) was used as a baseline in exploring possible sets of cluster centroids.
    - Evaluation of the results entailed use of the Minkowski and Cosine metrics in addition to Euclidean and (Soft-) DTW.
        - The Inertia score of the Time Series K-means approach was significantly lower (1.073) than that of Euclidean K-Means++ (1764.23), demonstrating the advantage of using a distance metric capable of accounting for differences in phase when comparing otherwise similar time series.
        - This unique property of DTW metrics may also explain the fact that the Silhouette Score for the K-Means++ algorithm (0.22) was slightly higher than that of the Time Series K-Means implementation (0.19), insofar as it affects the difference between the mean intra-cluster distance and nearest-cluster distance for each sample.
        - The same applies to the slightly higher Calinski-Harabaz Score for K-Means++ (42.56) compared to the Time Series K-Means implementation (41.03), insofar as it represents the ratio between within- and between-cluster dispersion.
        - Finding similarity between time series whose phase is misaligned will impact dispersion properties (Silhouette, and Calinski-Harabaz), even while reducing the sum of the distances of samples to their closest cluster center (Inertia).

Supervised (1): KNN | Time Series K-Neighbors Approach

- As mentioned above, hand-labeled classes were used as an approximation to ground truth.
- The elbow-method was used to identify a best value for the number of nearest neighbors used during model fitting. Consistent with recent

research in the field of time series data-mining, a value of K=1 or 2 were shown to reduce classification error rate.

- Several combinations of distance metrics (Euclidean, Cosine, Minkowski) and algorithms for computing nearest neighbors (KD-Tree, Ball-Tree, and Brute Force) were applied in using the KNN approach before identifying Minkowski and Ball-Tree as the most performant.
- For comparison the DTW metric was used for the Time Series implementation of K-Neighbors.
- Random permutation cross-validation was used in splitting the data into training and testing sets before model fitting.
    - Evaluation of the results showed roughly comparable performance of the standard and time series implementation of K-Neighbors.
        - For the standard KNN approach, applying the most performant combination (Ball-Tree and Minkowski):
            - (Subset) Accuracy was 0.73
            - Precision was 0.76
            - Recall was 0.73
            - F1-Score was 0.72
            - NMI was 0.63
            - AMI was 0.50
            - Adjusted-Rand Index was 0.49
            - Homogeneity was 0.60
            - Completeness was 0.66
            - V-Measure was 0.63
        - For the Time Series implementation of K-Neighbors, using the DTW distance metric:

- (Subset) Accuracy was 0.70
- Precision was 0.78
- Recall was 0.70
- F1-Score was 0.65
- NMI was 0.64
- AMI was 0.48
- Adjusted-Rand Index was 0.45
- Homogeneity was 0.58
- Completeness was 0.70
- V-Measure was 0.63

- Interestingly, the metrics indicating the greatest difference in performance when comparing standard KNN to the time series implementation of K-Neighbors were Precision, Recall, and the resulting F1-Score. In particular, use of the DTW metric resulted in more misclassifications of K-Neighbors in distinguishing between the Sunny with Occlusion and Partially Cloudy with Occlusion classes, suggesting that the phase invariant properties of DTW may encumber a model tasked with discriminating between two classes whose differences primarily consist in misalignments of this kind.

Unsupervised (2): K-Shape Approach

- In considering the possibility that differences between radiation profiles may consist on the scale of overall shape, the K-Shape algorithm was applied.
- K-Shape preserves the shapes of compared sequences using a parameter-free, scale- and shift-invariant distance metric based on cross-correlation values of timesteps.

        - Use of a normalized cross-correlation measure enables K-Shape to handle distortions in amplitude and phase in computing cluster centroids
- Mean-variance scaling of the data was performed prior to the clustering procedure, and the DTW distance metric was applied during model fitting.
    - Evaluation of results entailed use of the Euclidean, Cosine, Minkowski, DTW, and Soft-DTW metrics.
        - The best Inertia score (0.03) was lower than both the standard K-Means++ (1764.23) and Time Series K-Means implementations (1.073), indicating that class differences between time series may occur on a scale larger or longer than that of the bounded window of DTW distances used by a time series implementation of K-Means lacking the scale- and shift-invariant properties of K-Shape.
        - The same may apply in interpreting the slightly better Silhouette Score (0.26), compared with that of the K-Means++ algorithm (0.22) and the time series implementation of K-Means (0.19). Scale- and shift-invariance directly impacts the calculation of the difference between the mean intra-cluster distance and nearest-cluster distance for each sample.
        - Interestingly, the Calinski-Harabaz Score of K-Shape (26.52) was slightly lower than that of K-Means++ (42.56) and the time series implementation of K-Means (41.03). In a scale- and shift-invariant setting, even if the mean intra-cluster distance is relatively large with respect to the mean nearest-cluster distance, sufficiently large between-cluster dispersion may result in a lower Calinski-Harabaz score.

Supervised (2): Shapelet Classification Approach

- The performance of K-Shape relative to that of K-Means raised the question of whether inter-class differences between time series occur

on the scale of sub-sequences, between that of overall structure and that of a bounded window of nearby pointwise distances.

- The power of shapelet classifiers consists in their ability to learn maximally discriminative time series sub-sequences.
- Distances between series and shapelets represent shaplelet-transformed classification features whose segregation properties are used in optimizing the objective function
- Before model fitting, Min-Max scaling (from 0 to 1) was performed and random permutation cross-validation was used in splitting the data into training and testing sets
    - Several different configurations of optimizers and baseline segment lengths were applied before identifying the most performant, which used the RMSProp optimizer and a baseline length of 14% of the total number of timesteps
    - Evaluation of results showed the Shapelet classifier to outperform the KNN and Time Series K-Neighbors approach on all metrics:
        - (Subset) Accuracy was 0.79
        - Precision was 0.86
        - Recall was 0.79
        - F1-Score was 0.80
        - NMI was 0.73
        - AMI was 0.64
        - Adjusted-Rand Index was 0.61
        - Homogeneity was 0.74
        - Completeness was 0.71
        - V-Measure was 0.73
    - Unlike the time series implementation of K-Neighbors, the Shapelet classifier had no difficulty in distinguishing between the

Sunny with Occlusion and Partially Cloudy with Occlusion classes, as short term variations in sensor values of the latter radiation profile facilitate the distinctions made between the candidates patterns of shapelet matching.

- By the same token, the Shapelet classifier had greatest difficulty in distinguishing between the Cloudy and Partially Cloudy classes, as these exhibit very similar shapelet patterns regardless of length, differentiating themselves only on the basis of how relatively high or low the sensor values of these patterns climb over the course of the entire day.

- To facilitate the learning process for this best-performing supervised classification method, additional runs of the Shapelet modeling procedure applied the Synthetic Minority Oversampling Technique and the Adaptive Synthetic Sampling approach to oversample underrepresented classes.
- By applying its strategy of using a nearest-neighbors rule in weighting minority class instances to adaptively shift the classification boundary toward more difficult examples, the ADASYN approach outperformed the SMOTE method.
- Although neither oversampling strategy improved the overall (Subset) Accuracy, Precision, Recall, F-1 Score, or Adjusted-Rand Index, ADASYN did result in marginal improvements to the following performance metrics:
    - NMI was 0.74
    - AMI was 0.65
    - Homogeneity was 0.75
    - Completeness was 0.73
    - V-Measure was 0.74
  - Synthesizing difficult examples of the underrepresented ‘Sunny’ and ‘Cloudy with Sun-spike’ classes improved their

Recall metrics from 0.75 (imbalanced) and 0.67 (imbalanced) to 1.00 (ADASYN oversampled) and 1.00 (ADASYN oversampled), respectively, impacting the Homogeneity and Completeness metrics of the ADASYN oversampled Shapelet classifier, in turn.

Unsupervised (3): LSTM Sequence-to-Sequence Autoencoder

- Recent research has shown the performance gains achieved when applying unsupervised representation learning during a pretraining module to provide both supervised and unsupervised classifiers with features from which they can more readily learn.
- In this regard, a uniquely powerful property of Autoencoders consists in their ability to project data to a fixed-length lower dimensional space.
- Unlike densely-connected Stacked Autoencoders and Restricted Boltzmann Machines, recurrent sequence-to-sequence (Seq-2-Seq) autoencoders explicitly account for sequentiality, which makes them especially amenable to the time series clustering use case.
- The encoded representations learned during pretraining were used to reconstruct the features input to the other unsupervised approaches applied (K-Means++, Time Series K-Means, and K-Shape).
    - In using the LSTM Autoencoded data, these three unsupervised approaches applied the same cluster centroid computation algorithms and distance metrics as were used previously with unencoded data.
        - Evaluation of the results of using the recurrent autoencoded unsupervised clustering approach noted improvements in all three of the performance metrics used:
            - Inertia was 0.0076 (encoded K-Shape), compared with 0.029 (unencoded K-Shape), 1.073 (unencoded DTW Time Series K-Means), and 1764.23 (unencoded K-Means++), indicating a further reduction in the sum

of distances of samples to their closest cluster center when using autoencoded input features.

- The same effect was identifiable in the Silhouette Score, which had a significantly higher value of 0.59 (encoded K-Means++), compared with 0.22 (unencoded K-Means++), 0.26 (unencoded K-Shape), and 0.19 (unencoded Soft-DTW Time Series K-Means).
- Performance gains were equally visible in the Calinski-Harabaz Score, which was 116.24 (encoded K-Means++), compared with 42.56 (unencoded K-Means++), 26.52 (unencoded K-Shape), and 41.03 (unencoded Soft-DTW Time Series K-means).

- All three algorithms registered significant performance gains in Silhouette and Calinski-Harabaz Score when using autoencoded features, suggesting that a lower-dimensional projection of the data enhances the inter- and intra-class cluster dispersion properties, facilitating the learning process.

Results Summary:

- Amongst supervised methods, the Shapelet Classifier stood out as the most performant across all metrics.
    - Furthermore, applying ADASYN oversampling improved AMI, NMI, Homogeneity, Completeness, and V-Measure scores.
- Amongst unsupervised methods, using the LSTM Sequence-to-Sequence Autoencoder to perform unsupervised representation learning in a pretraining module significantly improved the Silhouette and Calinski-Harabaz metrics of all algorithms to which it was applied.
- A table summarizing best performance for all the machine learning methods applied is included below:

	Inertia	Silhouette	Calinski-Harabaz	Accuracy	Precision	Recall	F1-Score	AMI	Adj_Rand	NMI	Homogeneity	Completeness	V_Measure
KNN	--	--	--	0.73	0.76	0.73	0.72	0.5	0.49	0.63	0.6	0.66	0.63
TS_K-Neighbors	--	--	--	0.7	0.78	0.7	0.65	0.48	0.45	0.64	0.58	0.7	0.63
Shapelet Classifier	--	--	--	0.79	0.86	0.79	0.8	0.65	0.61	0.74	0.75	0.73	0.74
K-Means++	1764.23	0.22	42.56	--	--	--	--	--	--	--	--	--	--
TS_K-Means	1.073	0.19	41.03	--	--	--	--	--	--	--	--	--	--
K-Shape	0.029	0.26	26.52	--	--	--	--	--	--	--	--	--	--
LSTM_AE_K-Means++	1898.12	0.59	116.24	--	--	--	--	--	--	--	--	--	--
LSTM_AE_TS_K-Means	2.96	0.47	105.5	--	--	--	--	--	--	--	--	--	--
LSTM_AE_K-Shape	0.0076	0.48	74.57	--	--	--	--	--	--	--	--	--	--

Key Insights:

- Results of the highly performant Shapelet Classifier suggest that sub-sequence patterns are more informative than the information gleaned from overall shape or a bounded window of pointwise distances in discriminating between different radiation profiles.
    - As mentioned in Grabocka, et. al. (2014), interpretable shapelets have been used to enable early classification of time series
    - This quality is readily applicable to the radiation-level forecasting use case, and can be leveraged to improve application control logic and expedite timely commands to changing conditions on a site-specific basis.
- While supervised approaches were robust enough to learn from a relatively small data number of records with hand-labeled classes (171 days), it is notable that the classes identified by unsupervised approaches did not completely align with those identified using domain knowledge.
    - In particular, use of the highly performant LSTM Sequence-to-Sequence representation learning strategy distinguished between two forms of ‘Mixed Sunny and Partially Cloudy’ classes, one of which has maximum sensor values that peak early (Sunny-to-

Partially Cloudy), the other of which peaks late in the day (Partially Cloudy-to-Sunny).

- Moreover, in choosing cluster centroids, all unsupervised approaches deprioritized differentiating between the 'Sunny' and 'Sunny with Occlusion' classes in favor of other quantitatively distinct patterns, less readily identified when applying domain knowledge unassisted by machine learning.

References

- Kasetty, S., Stafford, C., Walker, G., Wang, X. and Keogh, E. (2008). Real-Time Classification of Streaming Sensor Data. In: *Tools with Artificial Intelligence, ICTAI '08*. IEEE.
- Lin, J., Williamson, S., Borne, K., and DeBarr, D. (2012). Pattern Recognition in Time Series. In: *Advances in Machine Learning and Data Mining for Astronomy*. New York: Chapman & Hall, CRC Press, 1-28.
- Mitsa, T. (2010). *Temporal Data Mining*. 1st ed. New York: Chapman & Hall, CRC Press.
- Nishida, T. and Mohammad, Y. (2015). Mining Time-Series Data. In: T. Nishida and Y. Mohammad, ed., *Data Mining for Social Robotics: Toward Autonomously Social Robots*. New York: Springer, 35-83.
- Seto, S., Zhang, W. and Zhou, Y. (2015). Multivariate Time Series Classification Using Dynamic Time Warping Template Selection for Human Activity Recognition. In: *2015 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE Xplore.
- Spiegel, S., Gaebler, J., Lommatzsch, A., De Luca, E. and Albayrak, S. (2011). Pattern recognition and classification for multivariate time series. In: *Fifth International Workshop on Knowledge Discovery from Sensor Data*. New York: Association for Computing Machinery.

- Keogh, E., Chakrabarti, K., Pazzani, M. et al. Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases. In: *Knowledge and Information Systems* (2001) 3: 263. https://doi.org/10.1007/PL00011669.
- Keogh, E. and Ratanamahatana, C. Exact indexing of dynamic time warping. In: *Knowledge and Information Systems*(2005) 7: 358. https://doi.org/10.1007/s10115-004-0154-9.
- Petitjean F., Ketterlin, A., and Gancarski, P. (2011). A global averaging method for dynamic time warping, with applications to clustering. In: *Pattern Recognition*. 44(3). 678-693.
- Blondel, M. and Cuturi, M. (2018). Soft-DTW: a Differentiable Loss Function for Time-Series. In: *Proceedings of the 34th International Conference on Machine Learning.* Sydney, Australia. PMLR 70*,* arXiv:1703.01541v2.
- Rousseeuw, P. (1986). Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. In: *Journal of Computational and Applied Mathematics*. 20. 53-65.
- Paparrizos J., and Gravano, L. (2015). K-Shape: Efficient and Accurate Clustering of Time Series. In: *The ACM SIGMOD Record*. 45(1). https://doi.org/10.1145/2949741.2949758.
- Grabocka, J., Schilling, N., Wistuba, M., and Schmidt-Thieme, L. (2014). Learning Time-Series Shapelets. In: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 392-401.
- Chawla, N., Bowyer, K., Hall, L., and Kegelmeyer, W. (2002). SMOTE: Synthetic Minority Over-sampling Technique. In: *Journal of Artificial Intelligence Research*. 16. 321-357.
- He, H., Bai, Y., Garcia, E., and Li, S. (2008). ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. In: *2008 IEEE International Joint Conference on Neural networks*. https://doi.org/10.1109/IJCNN.2008.4633969.
- Amiriparian, S., Freitag, M., Cummins, N., and Schuller, B. (2017). Sequence to Sequence Autoencoders for Unsupervised Representation Learning from

Audio. In: *Detection and Classification of Acoustic Scenes and Events.* Munich, Germany.

- Arthur D. and Vassilvitskii S. (2007). K-Means++: the Advantages of Careful Seeding. In: *SODA '07 Proceedings of the 18th annual ACM-SIAM Symposium on Discrete Algorithms*. New Orleans, USA. 1027-1035.
- Caliński T. and Harabasz J. (1974). A Dendrite Method for Cluster Analysis, In: *Communications in Statistics*, 3:1, 1-27, https://doi.org/10.1080/03610927408827101.
- Cho, K., van M., B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar. 1724-1734, arXiv:1406.1078v3.