Machine Learning Algorithms for Anomaly Detection

Stella White

stella17w/Pattern-Recognition

12/3/2024

Overview

- Goal: Compare the following four different machine learning (ML) algorithms on their accuracy in detecting data anomalies like outliers
  - Support Vector Machine (SVM)
  - Random Forest
  - k-Nearest Neighbor
  - K-Means Clustering

Background

- Task: Determine the best method for anomaly detection
- Baseline Method:
  - Compare Accuracy and Values to “Evaluation of Machine Learning-based Anomaly Detection Algorithms on an Industrial Modbus/TCP Data Set” paper
  - In this paper, SVM and Random Forest tied for the best, followed by k-nearest neighbour, and then k-means clustering
- Approach:
  - Test these algorithms on detecting outliers, and then detecting differences in waveform patterns
- Differences:
  - Different dataset
  - Added additional result values (Precision, Recall, AUC-ROC, AUC-PR, False positives, and False Negatives)

Support Vector Machines (SVM)

- Method: Create a divider (hyperplane) between two groups in such a way that each instance has the most possible distance from the divider
- For the divider, SVM calculates the normal vector of the separator hyperplane and the offset from the hyperplane
- When applying SVM, the set of instances divided by a linear kernel or the set will be mapped into a higher dimensional feature space if can’t

Random Forest

- Method: Builds collection of decision trees and combines all trees' results
- Decision trees consists of a root node, internal nodes, split nodes, and leaf nodes
- Each node corresponds to a class predicted
- Each tree is trained using a different, completely random subset of the data and a different completely random subset of the features
- In this project, 2000 decision trees was used
- Disadvantages are robot to noise and overfitting

$$F(p) = \frac{1}{N} \sum_{i=1}^{N} f_i(p)$$

Equation 1: Random Forest [2]

K-Nearest Neighbor

- Method: determines the set of the k nearest neighbours by calculating the Euclidean distance in an n-dimensional feature space
- In this project, 2 nearest neighbours was used

$$D = \sqrt{\sum_{i=1}^{n}(x_i - w_i)^2}$$

Equation 2: K-Nearest Neighbor [1]

K-Means Clustering

- Method: Classifies sample to closest cluster based on Euclidean distance from the center of a cluster
- In this project, 2 clusters was used
- Unsupervised method: doesn't need prior information

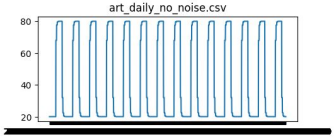
$$E = \sum_{j=1}^{k} \sum_{i_l \in C_j} |i_l - w_j|^2$$

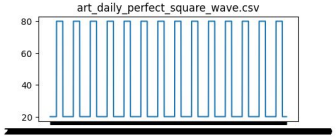
$$j \in \{1, \ldots, k\}, l \in \{1, \ldots, n\}$$

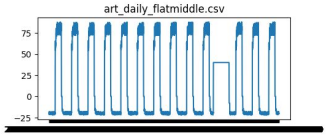
Equation 3: K-Means Clustering [1]

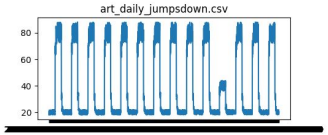
Dataset

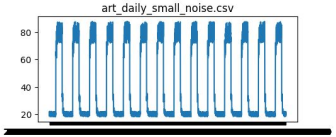
- Two public datasets from Numenta Anomaly Benchmark
- There are artificially-generated datasets of datastreams of one feature plus time with and without anomalies
- I plan on using 80% of each clean and outlier dataset for training and the other 20% for testing.
- I split the 4 datasets with an anomaly in pattern with 2 for training and 2 for testing
- The data was normalized and split into windows of a certain size before put into the algorithms

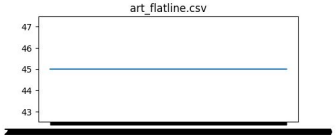


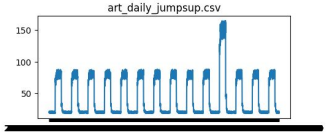


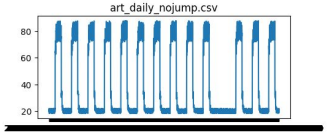


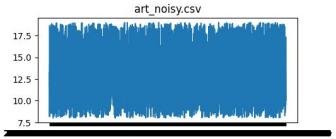


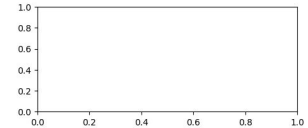


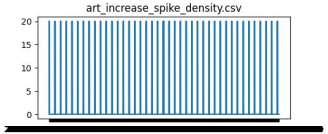


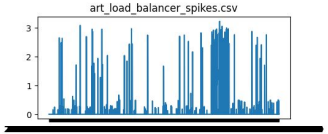




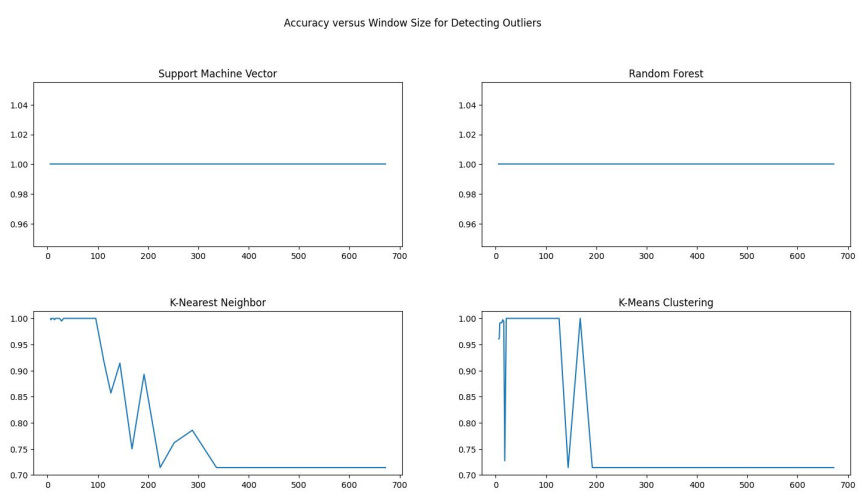








Experimental Results



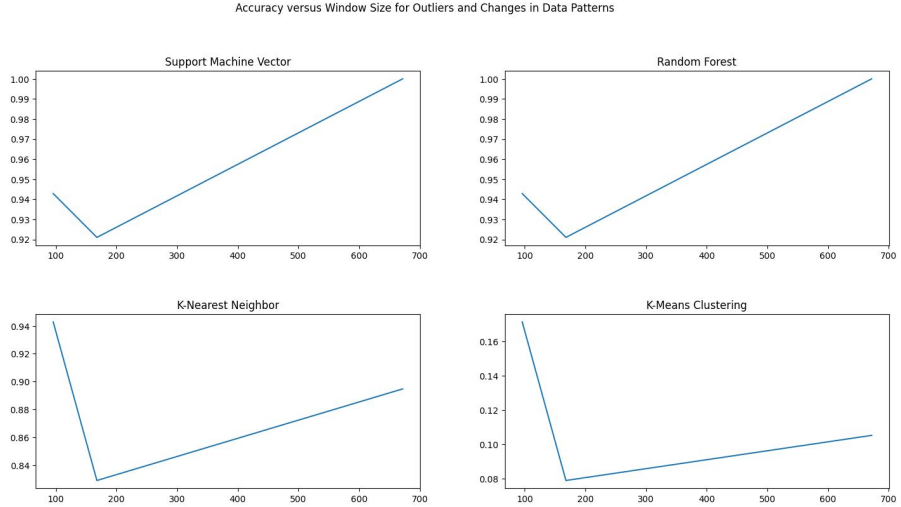


Table 6: *Accuracy* and *F1-score* of *SVM*

Dataset	*Accuracy*	*F1-score*
DS1	1,0	1,0
DS2	1,0	1,0
DS3	0,999 936	0,999 968

Table 10: *Accuracy* and *F1-score* of *Random Forest*

Dataset	*Accuracy*	*F1-score*
DS1	1,0	1,0
DS2	0,999 701	0,999 851
DS3	0,999 973	0,999 986

Table 14: *Accuracy* and *F1-score* of *k-nearest Neighbour*

Dataset	*Accuracy*	*F1-score*
DS1	0,997 097	0,998 527
DS2	0,999 118	0,999 559
DS3	0,999 412	0,999 706

Table 18: *Accuracy* and *F1-score* of *k-means Clustering*

Dataset	*Accuracy*	*F1-score*
DS1	0,981 018	0,990 383
DS2	0,556 242	0,714 853
DS3	0,633 624	0,775 728

Tables 6, 10, 14, 18 are from [1]

Experimental Results

	Support Machine Vector	Random Forest	K-Nearest Neighbor	K-Means Clustering
Best Window Size	672	672	96	96
Accuracy	1.0	1.0	0.9429	0.1714
Precision	1.0	1.0	1.0	0.1714
Recall	1.0	1.0	0.6667	1.0
F1-score	1.0	1.0	0.8	0.2927
AUC-ROC	1.0	1.0	0.8333	0.5
AUC-PR	1.0	1.0	0.7238	0.1714
False Positives	0.0	0.0	0.0	1.0
False Negatives	0.0	0.0	0.3333	0.0

Conclusion

- Outlier Detection:
  - All methods could detect if outliers were presented within a certain window size
- Window Size:
  - Random Forest and SVM wasn't impacted for outliers, but performed best on bigger window sizes for detecting differences in patterns
  - K-nearest Neighbor and K-means clustering did the best on smaller window sizes for both
- Comparison of Overall Results:
  - Random Forest and SVM had perfect values and matched the high values from the baseline
  - K-nearest Neighbor performance values were slightly less than the baseline values
  - K-means Clustering did much worse than the baseline
- Improvements:
  - Used a bigger dataset that includes many different types of data anomalies
  - Add more clusters to K-means clustering for different types of data anomalies
- Future Work:
  - Compare Random Forest and SVM on a deeper level
  - Compare Traditional SVM versus One-Class SVM

References

[1] Simon Duque Anton, Suneetha Kanoor, Daniel Fraunholz, and Hans Dieter Schotten. 2018. Evaluation of Machine Learning-based Anomaly Detection Algorithms on an Industrial Modbus/TCP Data Set. In Proceedings of the 13th International Conference on Availability, Reliability and Security (ARES '18). Association for Computing Machinery, New York, NY, USA, Article 41, 1–9. https://doi-org.portal.lib.fit.edu/10.1145/3230833.3232818

[2] Priyanka More, Dharmesh Dhabliya, Jambi Ratna Raja Kumar, Supriya Arvind Bhosale, Aarti S. Gaikwad, and Sonu V. Khapekar. 2024. Utilizing Machine Learning Approaches for Anomaly Detection in Industrial Control Systems. In Proceedings of the 5th International Conference on Information Management & Machine Intelligence (ICIMMI '23). Association for Computing Machinery, New York, NY, USA, Article 101, 1–7. https://doi-org.portal.lib.fit.edu/10.1145/3647444.3647928