IndyCar Anomaly Event Detection

Jiayu Li
Xinquan Wu

IndyCar

Many events can happen during the racing period, such as pit stops, crashes, mechanical breakdown, and drivers ranking changes.

Use the data to detect some interesting event which can be treat as anomaly

Anomaly detection algorithms

anomaly detection is the identification of rare items, events or observations which raise suspicions by differing significantly from the majority of the data.

intrusion detection

fraud detection

fault detection

system health monitoring

Scoreboard

The NAB scores are normalized such that the maximum possible is 100.0 (i.e. the perfect detector), and a baseline of 0.0 is determined by the "null" detector (which makes no detections).

Detector	Standard Profile	Reward Low FP	Reward Low FN
Perfect	100.0	100.0	100.0
Numenta HTM*	70.5-69.7	62.6-61.7	75.2-74.2
CAD OSE†	69.9	67.0	73.2
earthgecko Skyline	58.2	46.2	63.9
KNN CAD†	58.0	43.4	64.8
Relative Entropy	54.6	47.6	58.8
Random Cut Forest ****	51.7	38.4	59.7
Twitter ADVec v1.0.0	47.1	33.6	53.5
Windowed Gaussian	39.6	20.9	47.4
Etsy Skyline	35.7	27.1	44.5
Bayesian Changepoint**	17.7	3.2	32.2
EXPoSE	16.4	3.2	26.9
Random***	11.0	1.2	19.5
Null	0.0	0.0	0.0

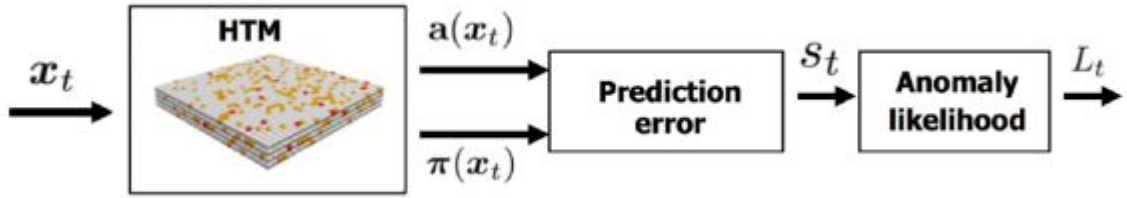
Hierarchical temporal memory

Hierarchical Temporal Memory (HTM) is a machine learning technology that aims to capture the structural and algorithmic properties of the neocortex.

Inspired by the pyramidal cells in neocortex layers

Thousands of synapses

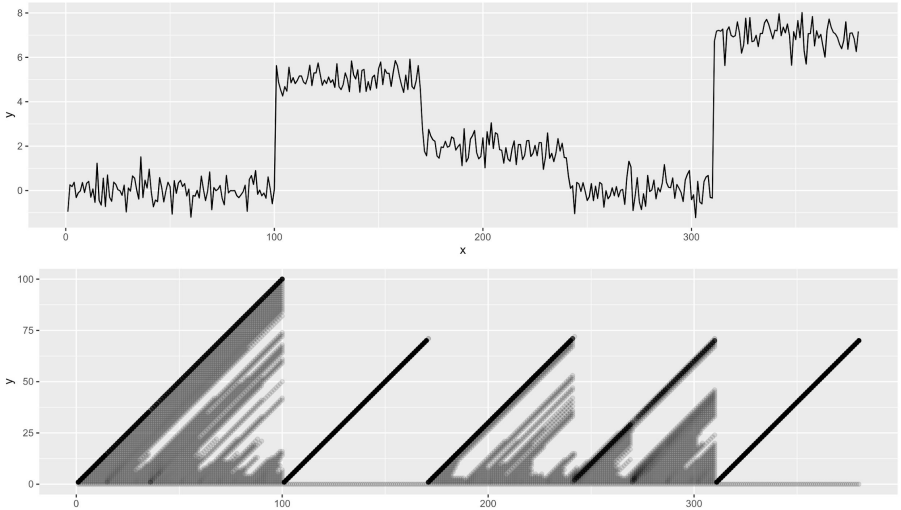
Learns by modeling the growth of new synapses



Bayesian Online Changepoint Detection

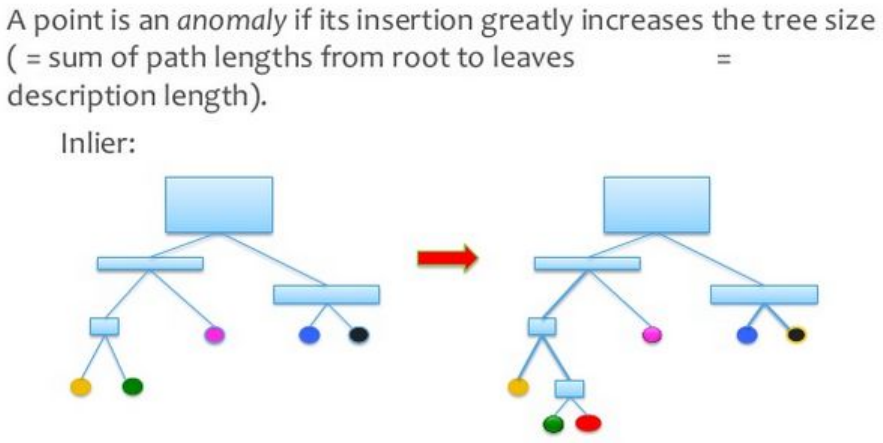
For each record at step x in a data stream, the probability that the current record is part of a stream of length n for all n <= x.

For a given record, if the maximimum of all the probabilities corresponds to a stream length of zero, the record represents a changepoint in the data stream. These probabilities are used to calculate anomaly scores for NAB results.



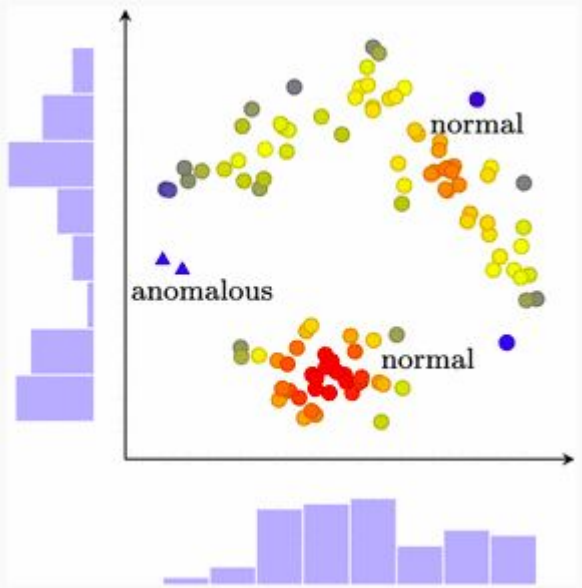
Random Cut Forest

A tree is an ordered way of storing numerical data. To create a tree, you randomly subdivide the data points until you isolate the point you're testing to determine whether it's an anomaly. Each time you subdivide the data points, it creates a new level of the tree. The fewer times you need to subdivide the data points before you isolate the target data point the more likely it is that the data point is an anomaly for that sample of data.



EXPoSE

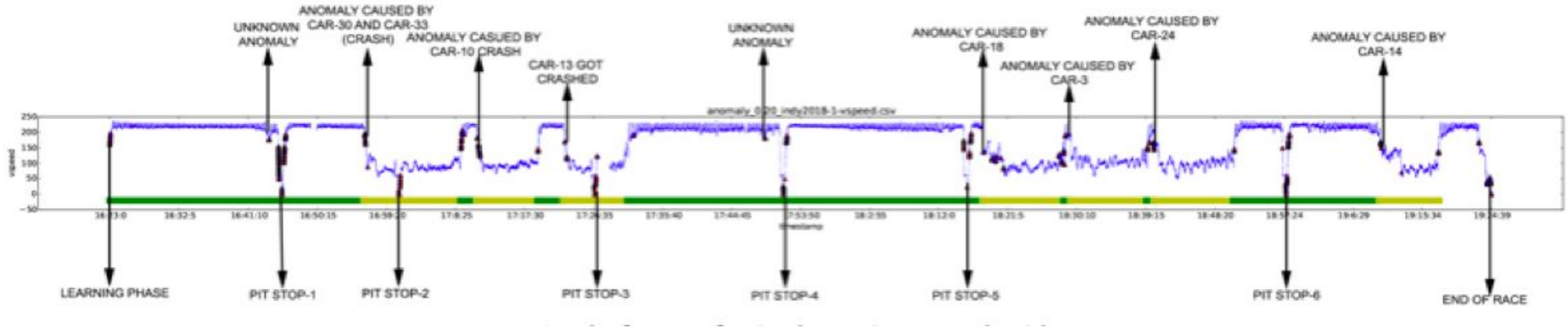
EXPoSE (Expected Similarity Estimation for Lage-Scale Batch and Streaming Anomaly Detection) calculates the likelihood of a data point being normal by using the inner product of its feature map with kernel embedding of previous data points. This measures the similarity of a data point to previous points without assuming an underlying data distribution.



Example of two instances (*triangle*) which are different from the distribution of normal data (*circle*) along with histograms of the marginal distributions

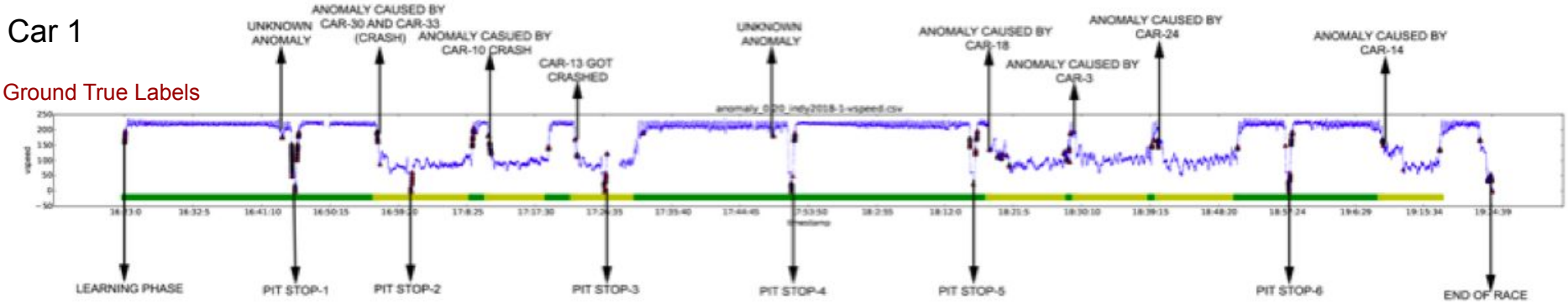
Results of Anomaly Detection

Compare results for 4 different algorithms with ground true labels of car 1

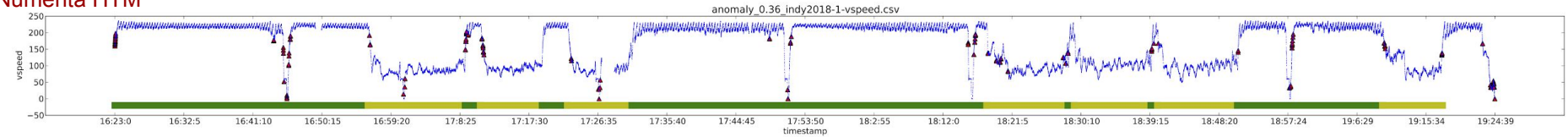


Car 1

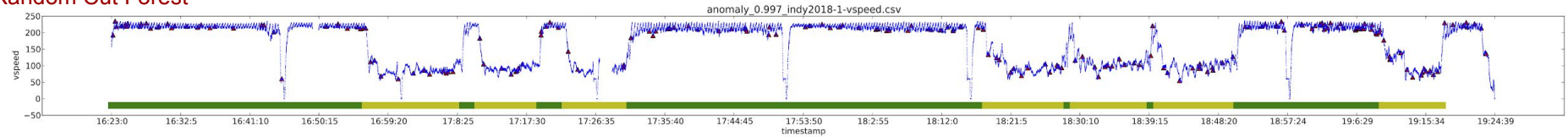
Ground True Labels



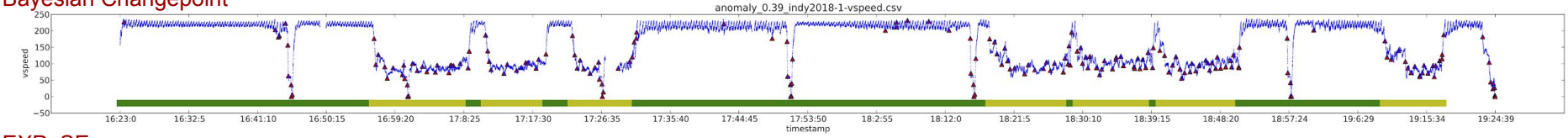
Numenta HTM



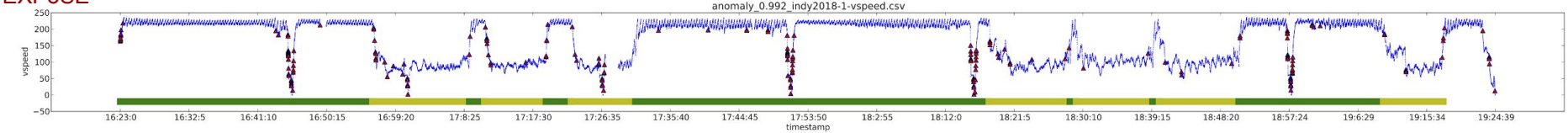
Random Cut Forest



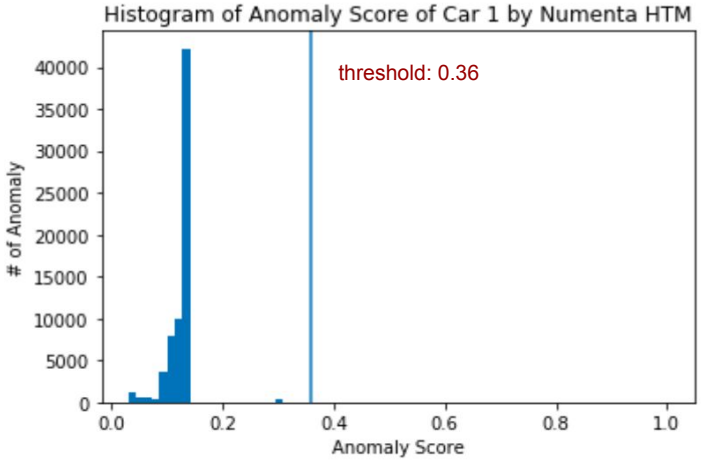
Bayesian Changepoint

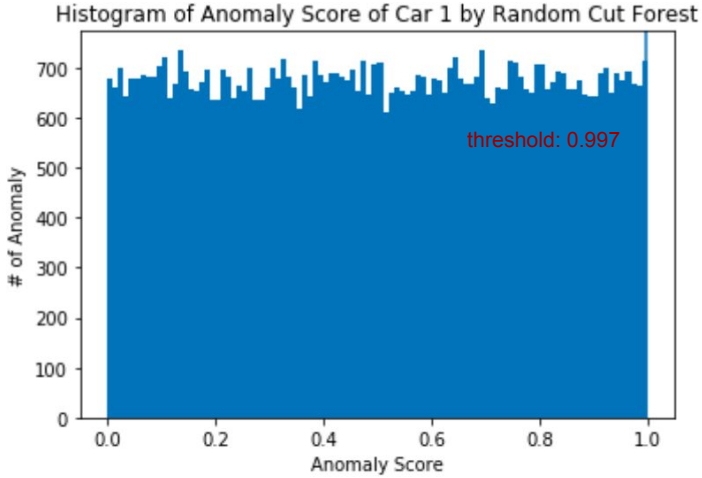


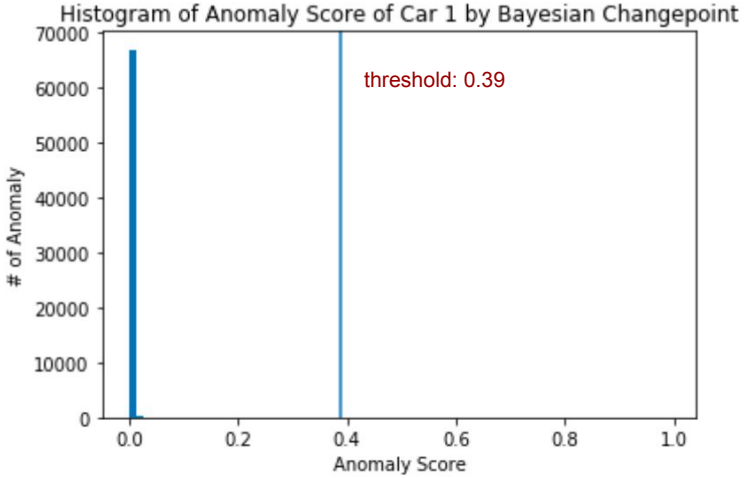
EXPoSE

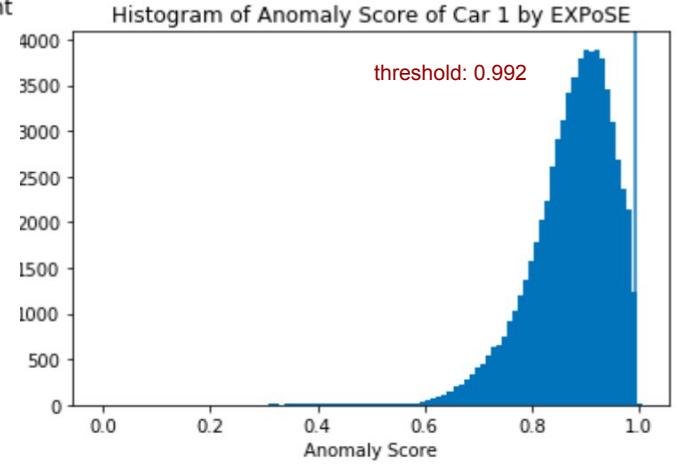


Histogram of anomaly score of car 1 by 4 different algorithms

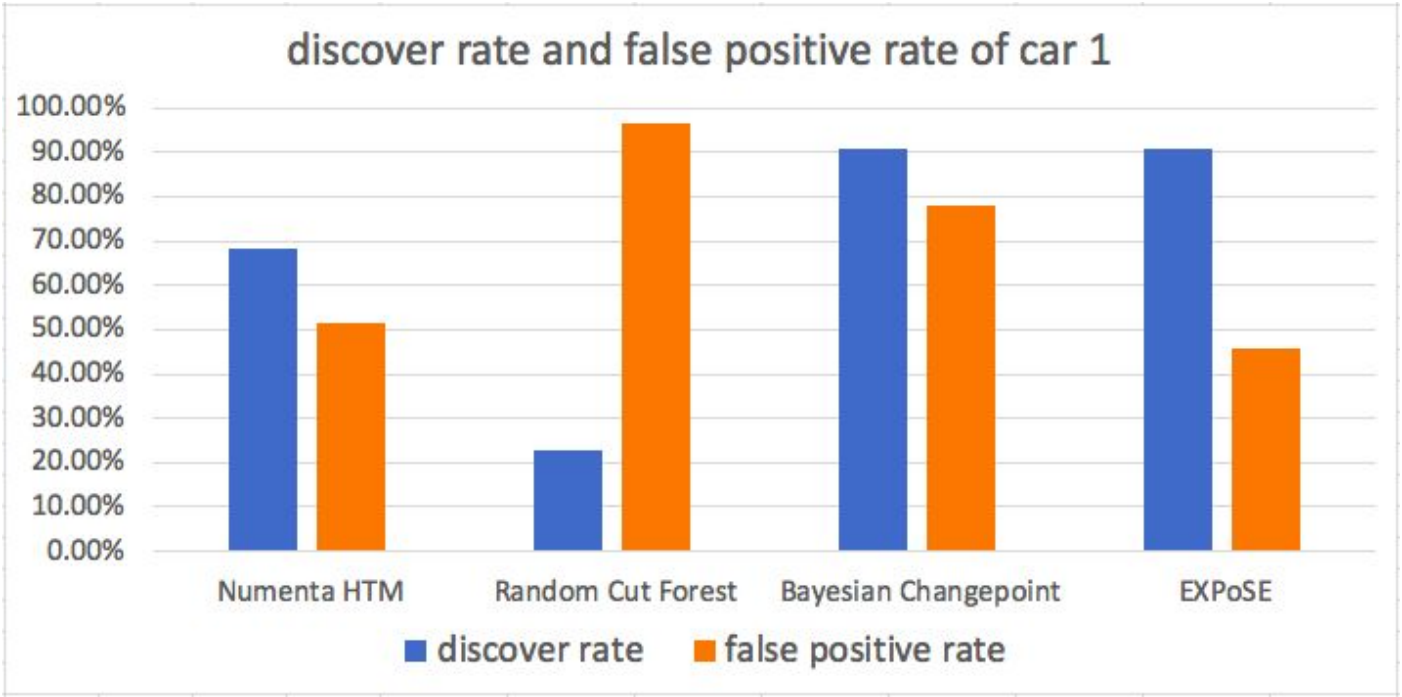








discover rate and false positive



Discover Rate = # of intervels discovered / total # of labeled intervals

False Positive Rate = # of anomaly outside intervals / total # of anomaly

Conclusion

- We compare 4 different algorithms with Ground True Labels:
  - Numenta HTM, Random Cut Forest, Bayesian Changepoint, and EXPoSE
- Random Cut Forest has very low discover rate
- Bayesian Changepoint has both high discover rate and high false positive
- Numenta HTM and EXPoSE achieved good discover rate
  - However, EXPoSE has higher false positive rate