SAND2019-5963PE

Multiple Instance Learning

Sandia National Laboratories
Livermore, CA

Fulton Wang (fulwang@sandia.gov)
Ali Pinar (apinar@sandia.gov)

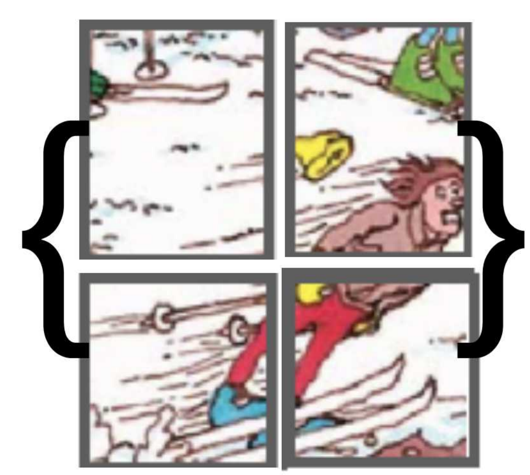
Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Sandia National Laboratories

Scenario

- Training Data is Weakly Labelled
- Consists of bags of instances
- Unobserved: instance labels
- Observed: whether each bag has some positive instance

Contains a Waldo instance



Has no Waldo instance

Sandia National Laboratories

Goals

- Learn instance classifier

Is  a Waldo instance?

- Learn bag classifier

Does {  } have a Waldo instance?

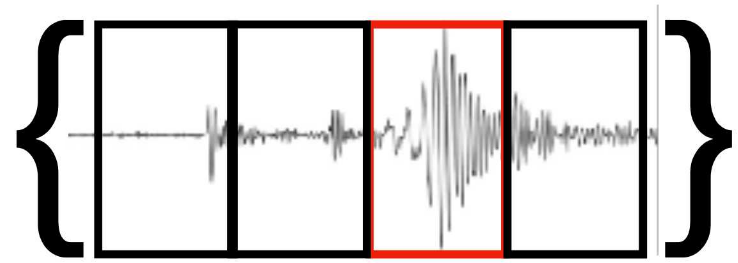
- Do both: **interpretability**

Does {  } have a Waldo instance?

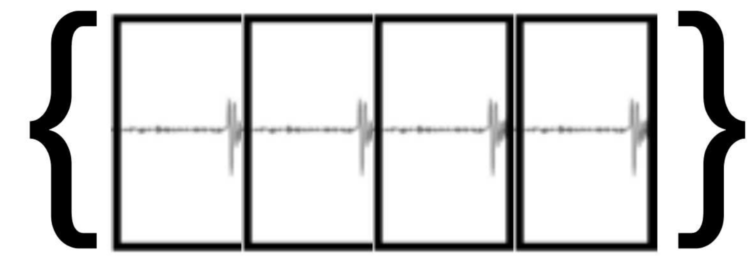
If yes, where is Waldo?

ADAPD Problem: Seismic Event Detection

- Training data:



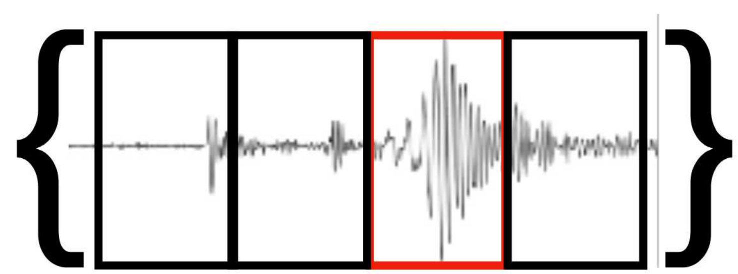
Contains frequency with event



Contains no frequency with event

- Goal:

Does



contain frequency w/ event?

If yes, which frequency is the event?

ADAPD Needs

- Sparse labels: labeled data is, will be, hard to find.
    - Consistently labeled data, is will, be even harder.
- Multi-phenomenology: labels can come from a different data source (e.g., predict chemical release from activity logs)
- We observe collection of events, not isolated events
    - Traditional ML: data point <u>is</u> a red/blue event
    - Practical problems: data point <u>contains</u> a red/blue event.

Sandia National Laboratories

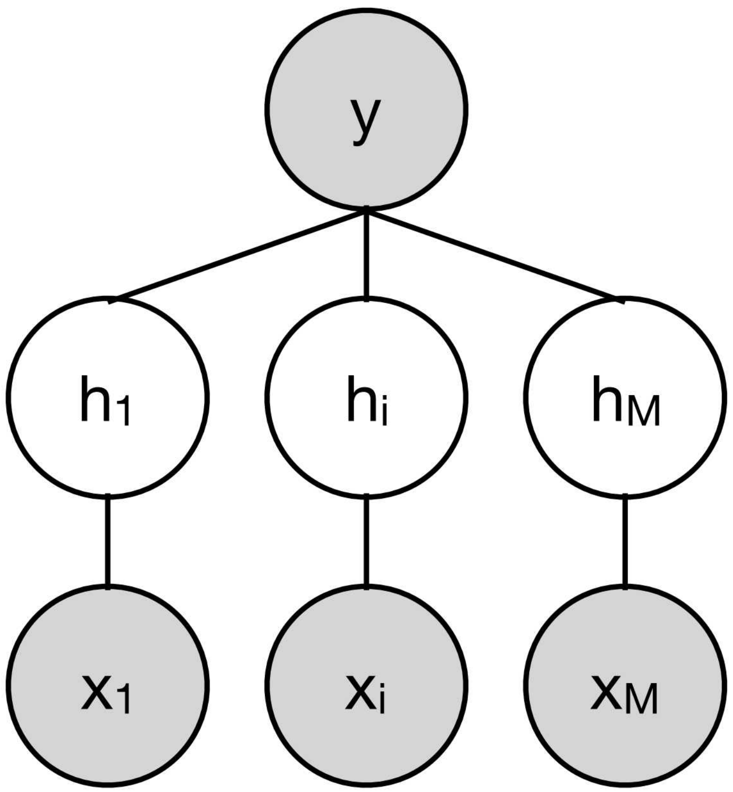
Why is multiple instance learning hard?

- Increased flexibility of multi-instance learning comes at the cost of increased complexity of algorithms.
  - Loose information leads to a larger search space constrained at two levels (bags and instances).
  - Labels can be correlated.
    - Instance labels may be structured as sequences - if one frequency is event, then more likely neighboring frequency also is event.
  - Variance in bag sizes pose an additional challenge.
    - More activities in a day does not mean increased likelihood of a rare event.
- This is an emerging field without established methods and associated software.
  - We are at the leading edge.

Sandia National Laboratories

How to do multiple instance learning?

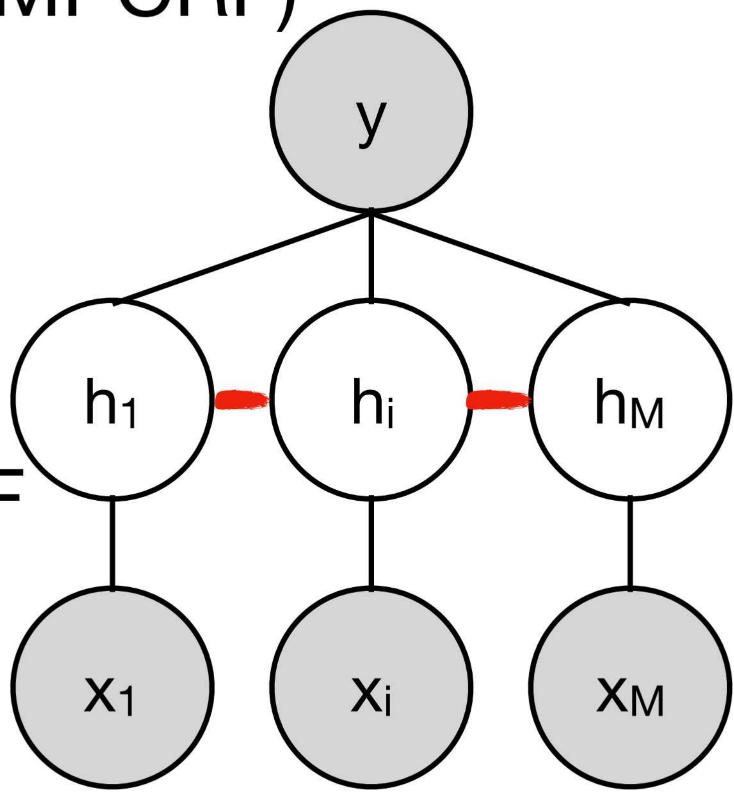
- Use probabilistic **latent** variable model (MI-logreg)
- Variables for each bag:
  - Bag label y (0/1)
  - Instance feature vectors $x_1..x_M$
  - Latent instance labels $h_1..h_M$ (0/1)



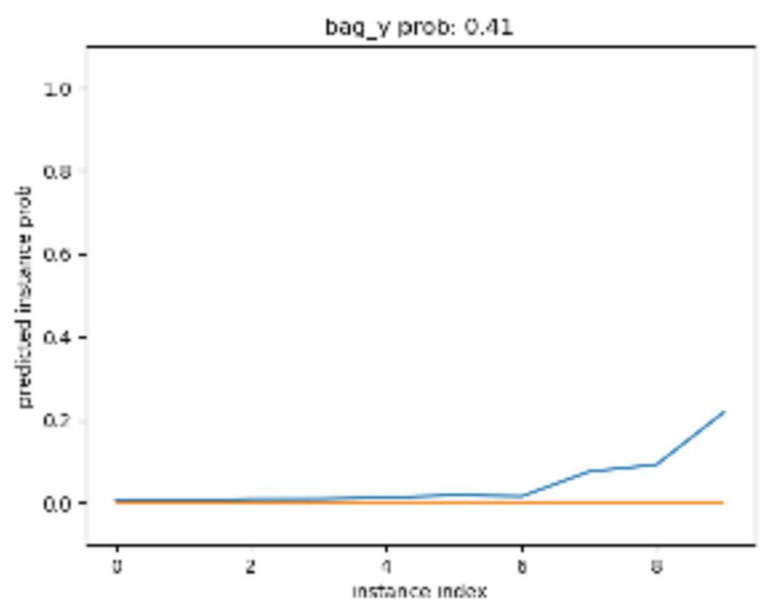
- Multiple instance assumption: $y = 1$ if some $h_i = 1$
- Model $h_i|x_i$ independently using logistic regression
  - $P(h_i=1|x_i) = \text{sigmoid}(B^T x_i)$ where B is regression parameter

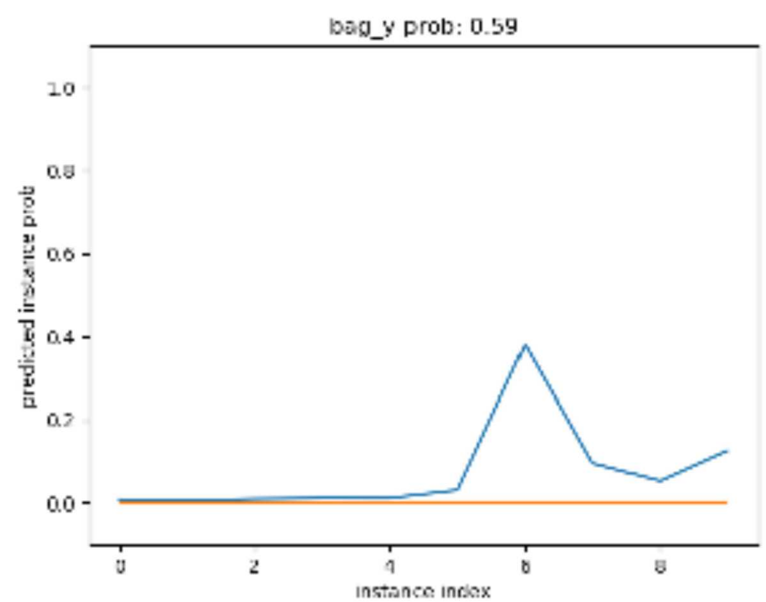
Status

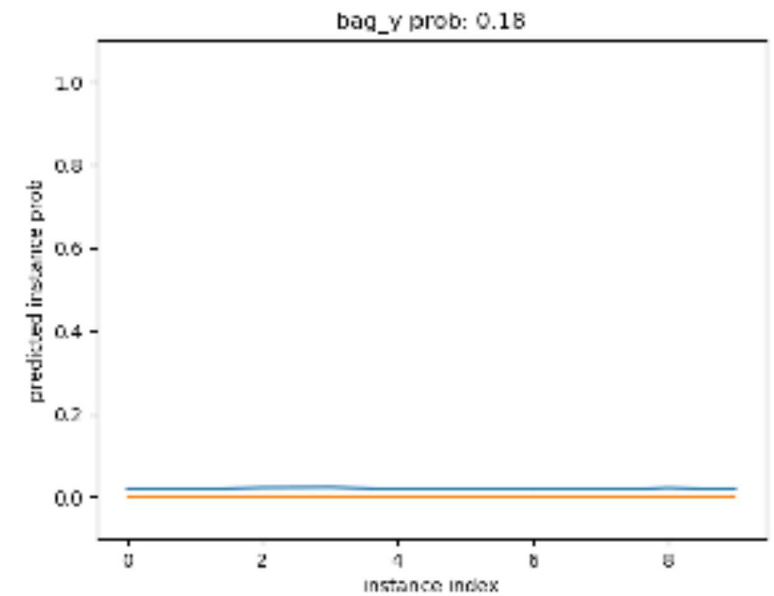
- Implemented existing method, initial results on seismic and cybersecurity data. Focus: sequence-structured instances.
- Extended method to account for instance label dependencies using conditional random field (MI-CRF) instead of logistic regression
  - MI-Logreg: model $h_i|x_i$ independently
  - MI-CRF: jointly model $h_1..h_M\,|x_1..x_M$ with CRF
- Wrongly assuming independence can lead to false positives under positive dependence (example: suppose labels are always equal)



Multi-instance learning is interpretable by design







Instances ordered by frequency on x-axis.
Blue line indicates probability a frequency contains event $P(h_i=1|x_i)$
Orange line is 0/1 prediction of h_i

Sandia National Laboratories

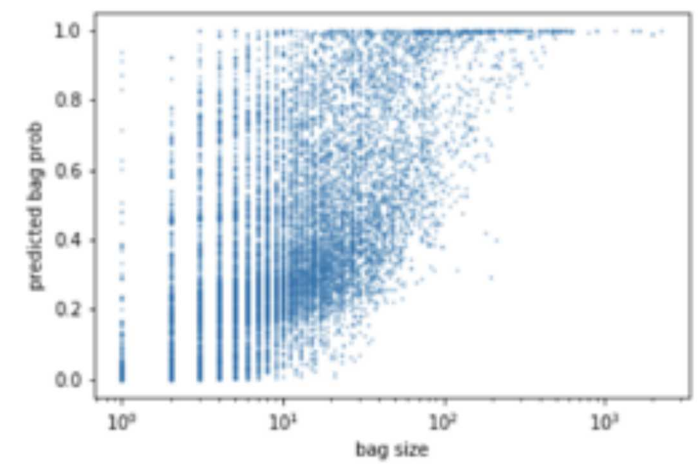
Performance on earthquake detection

	AUC	FP
MI-CRF	.75(.02)	.16(.02)
MI-Logreg	.77(.01)	.24(.01)
Logreg	.71(.01)	.16(.01)
RF	.77(.01)	.16(.02)

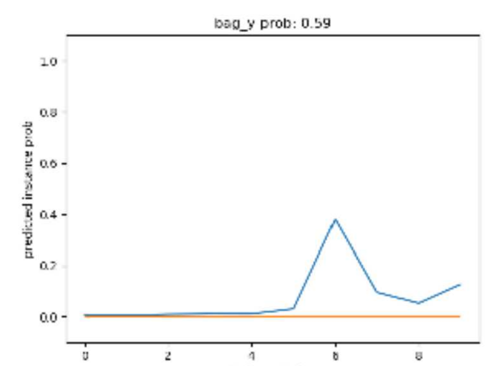
- Labelled raw signal from LYNM decomposed into contributions from 9 non-overlapping frequency ranges. Each bag has 9 instances. ~5000 bags.
- Logreg and RF (random forest) are vanilla bag classifiers which concatenates the 9 instance feature vectors to form bag feature vector.
- No instance labels, so all metrics are bag-level.
- MI-Logreg higher AUC than Logreg (better model, less parameters)
- MI-CRF lower FP than MI-Logreg b/c model dependences. Both not calibrated.
- RF higher AUC than Logreg. Future work: MI-RF (can't use gradient descent)

Future work

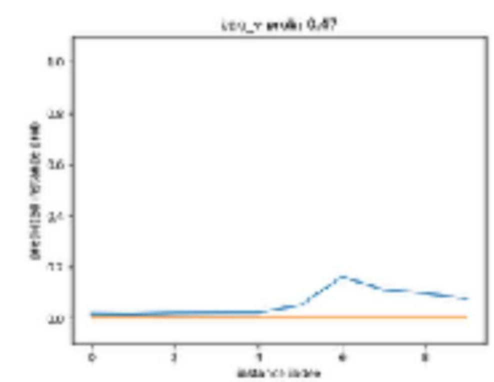
- Accounting for bag size variability
  - Example: time of event known
    with differing uncertainty
  - With MI-logreg,
    larger bag -> higher bag positive probability.
  - CRF addresses this issue for sequence data only.
- Improving interpretability
  - Want fewer predicted positive instances
- Incorporating nonlinear models;
  improving calibration
- Other ADAPD applications:
  - Text data: A document is a bag. Can we identify the suspicious paragraph?
  - Multi-phenomenology: Fuse data sets
  - Graph data: can we identify the patterns to search for?



Larger bag -> higher bag probability in cybersecurity application



More interpretable. Know which instance to examine



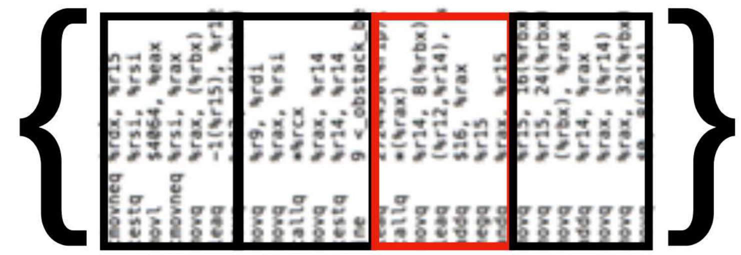
Less interpretable

Sandia National Laboratories

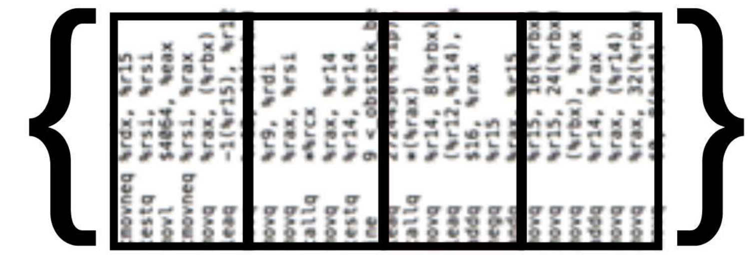
Backup

Malware Classification

- Training data:

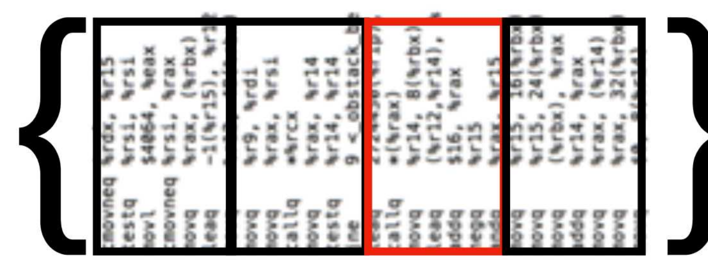


Contains some malicious segment



Contains no malicious segments

- Goal:

Is  malware?

If so, which segments are malicious?

References

- Solving the Multiple-Instance Problem with Axis Parallel Rectangles, Dieterrich, JAIR 1997: Introduced MI learning problem - know whether molecule binds to protein, but not which of its conformations
- Joint Multi-label Multi-instance Learning for Image Classification, Zha, CVPR 2008: CRF for image classification.
- Efficient Multi-Instance Learning for Activity Recognition from Time Series Data, Guan, ICML 2016: Generative model, does not actually use multiple instance labelling assumption
- Discriminative probabilistic framework for generalized multi-instance learning, Pham, ICASSP 2018: Extension of MI-Logreg method. Allows for more general bag label model, i.e. bag positive if # positive instances > non-zero threshold

Sandia National Laboratories