Anomaly detection techniques for IVHM fault management

Varun Chandola, Tim Miller, Nishith Pathak, Lane Schwartz, Hanhuai Shan, Nisheeth Srivastava,
Stephen Wu, Junlin Zhou, Aleksandar Lazarevic(co-PI), Arindam Banerjee (co-PI) , William Schuler(co-PI),
Vipin Kumar(co-PI) and Jaideep Srivastava (PI)
University of Minnesota, Twin Cities; United Technologies Research Center

Handling multivariate and heterogeneous data

- **We have developed a new multivariate anomaly detection technique: WIN_{SS}- Subspace Based Anomaly Detection [1]**
  - Reduce each multivariate time series to a univariate time series
  - The "reduced" time series captures the evolution of time series
    - Use sliding windows to capture evolution
- **Hypothesis:**
  - Normal time series will have similar evolution pattern.
  - Anomalous time series will have different evolution pattern induced by the anomaly.
- Apply univariate time series anomaly detection technique ($WINC_{SVM}$) to detect anomalies

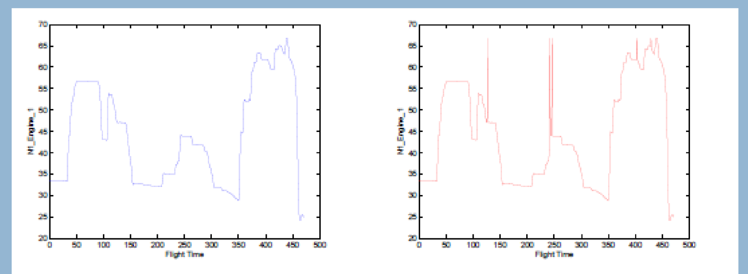
	WIN_{SS}	kNN	$WINC_{SVM}$
CMAPSS	0.75	0.50	0.50
EEG	0.66	0.17	0.43

- **Outperforms existing kNN multivariate methods and basic univariate schemes**

[1] **Varun Chandola**, *Anomaly Detection for Symbolic Sequences and Time Series Data*, PhD. Dissertation, Computer Science Department, University of Minnesota, September 2009

Handling heterogeneous sequences

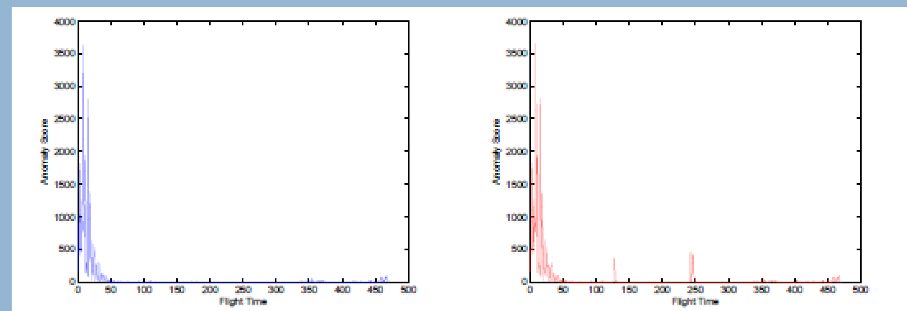
- Approach #1
  - Use MINDS, a density based anomaly detection technique, to assign anomalies to each heterogeneous multivariate observation of a test time series
  - Aggregate per observation scores to obtain overall anomaly score for the time series
  - Use a data driven distance measure (idf) to handle categorical attributes
  - Real flight data (CONEX) experiments found high anomaly scores for instants identified as anomalies by domain experts



(a) Normal Output (b) Faulty Output

Figure: Normal and faulty outputs. Faults injected at times 127, 242, 246, 403, 429.

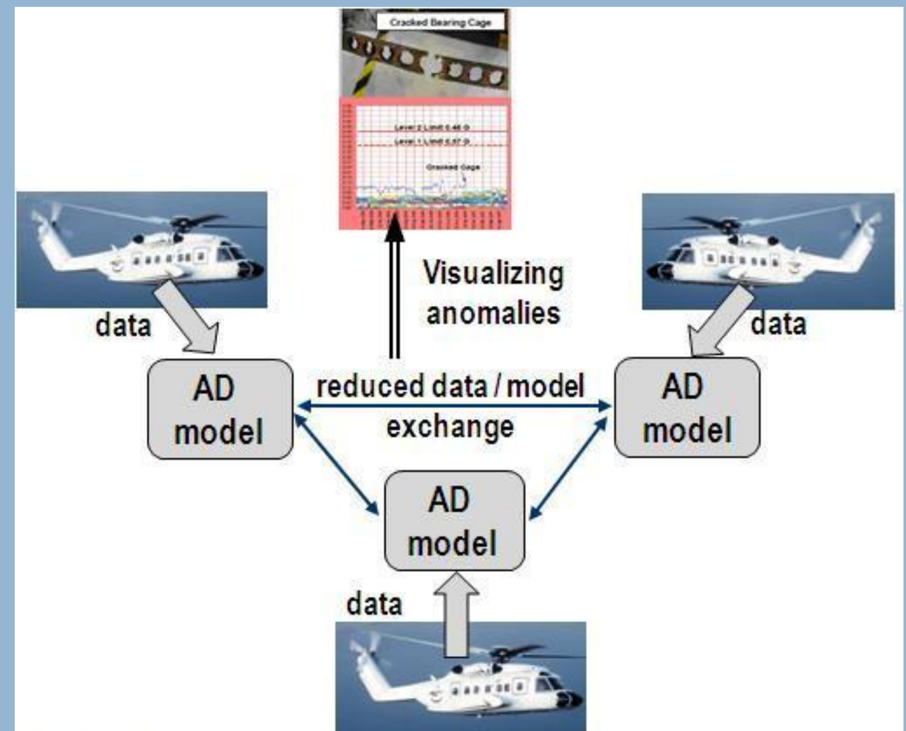
- Approach #2
  - A Linear Dynamical System (LDS) with discrete & continuous inputs, continuous outputs, and continuous hidden states is learnt using the Kalman Filter
  - The LDS can be used to learn a predictive model for continuous flight parameters, e.g., engine rpm, pitch, roll, etc., based on the pilot inputs, e.g., flap position, throttle position, elevator position, etc.
  - During testing, the anomaly score at any time is calculated as the *Mahalanobis Distance* between the predicted and the actual output
  - Experiments on real flight data show algorithm learns a stable EM model and makes accurate anomaly predictions



(a) Anomaly Scores for Normal Data (b) Anomaly Scores for Faulty Data

Figure: Anomaly scores for normal and faulty flights.

Handling distributed data



Data Sources
- Sikorsky S92 Flight Record Data (main and tail gearbox)
- ADAPT System Data (obtained from NASA)
- Other publicly available non-aviation data sets

Key accomplishments:

•Development of fast distributed anomaly techniques based on T^2 and Q statistics [2]
•Evaluation of several types of one-class anomaly detection algorithms
  •density based (Parzen density estimate, LOF)
  •clustering based methods
  •boundary based (unsupervised SVM)
  •reconstruction based methods (Minimal probability machine, auto-associative NNs, SOMs, minimum spanning trees)
•Development of new method for anomaly detection based on integrating clustering based methods and regression models
•Development of a novel method for combining anomaly detection models from distributed sources based on models' quality and diversity
•Development of a method for visualizing detected anomalies / faults and identifying variables most relevant to the fault

In unimodal scenario, only covariance matrices and mean vectors from individual data sets are exchanged

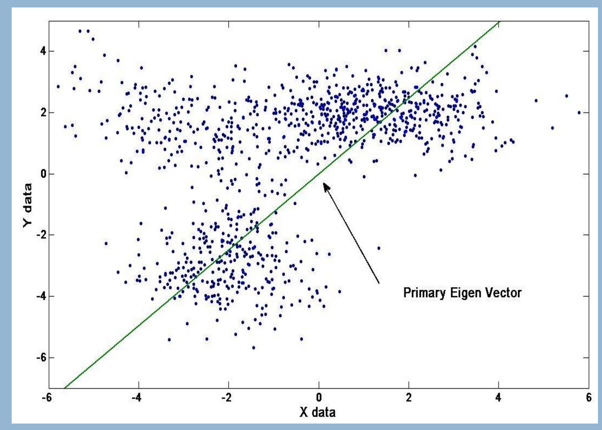
In multi-modal scenario, Gaussian mixture models (GMMs) are first identified and then corresponding covariance matrices and mean vectors are computed for each GMM mode

Unlike unimodal scenario, in multi-modal case, covariance matrices and mean vectors for all GMM modes identified at individual sites are exchanged among the sites

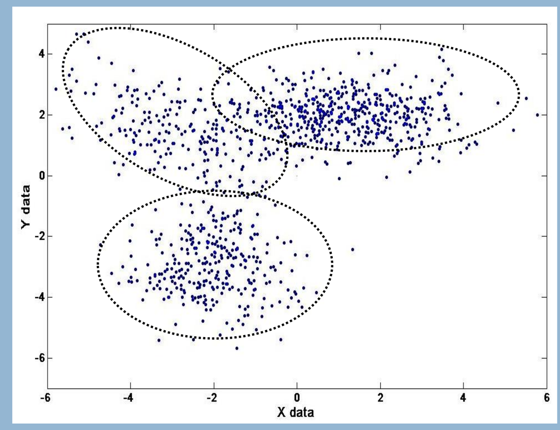
Limited communication overhead among distributed sites

Guaranteed same prediction performance as in centralized case

More accurate detection of anomalous flight records that unimodal models



Unimodal PCA Model



Gaussian Mixture Model

[2] **A. Lazarevic, N. Srivastava, A. Tewari, J. Isom, N. Oza, J. Srivastava**, Solving a prisoner's dilemma in distributed anomaly detection, *The Third International Workshop on Mining Multiple Information Sources, IEEE International Conference on Data Mining*, December 6, 2009, Miami, FL.

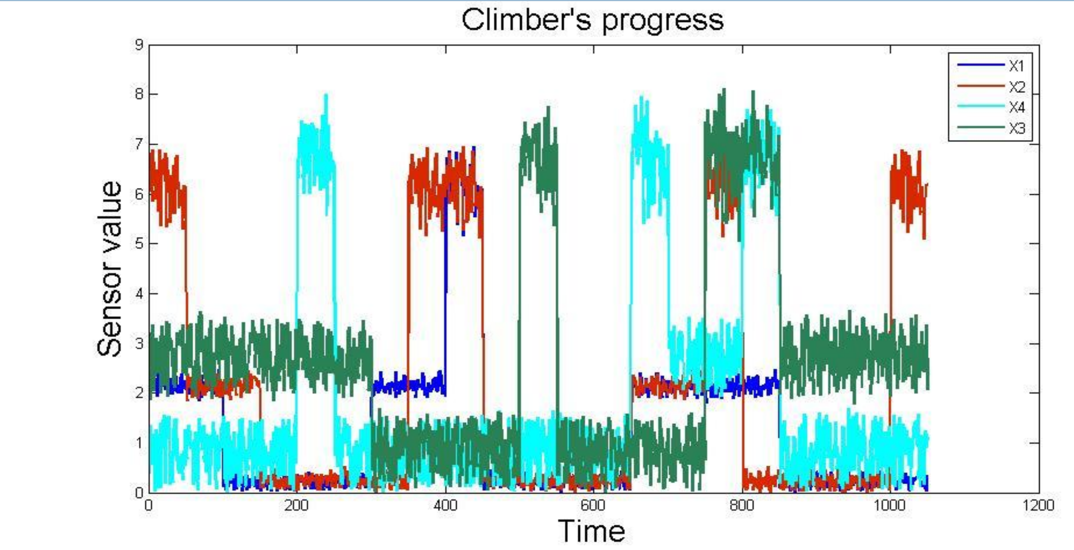
Handling systemic anomalies

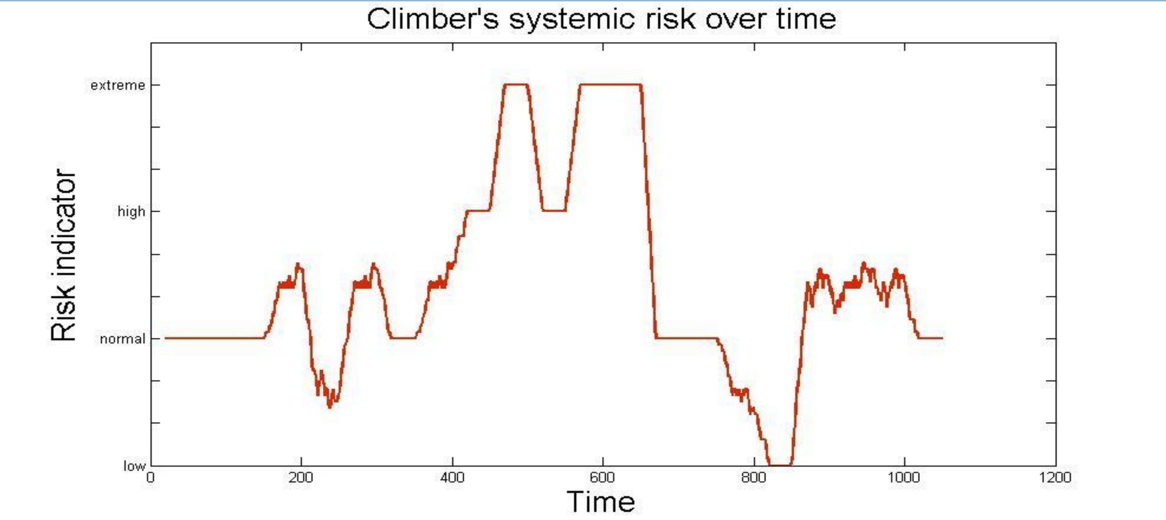
Systemic anomalies arise in situations where quantitative assessments of the performance of individual sub-system components of a complex system are available, system semantics are well-understood at the top-level and system performance is to be predicted and potential failures detected.

We have developed an algorithm [3] on a simulation of a rock-climber climbing with sensors attached to all four limbs; the algorithmic input comprises of four continuous-valued data streams X_i, i = 1…4.The task for the anomaly detection algorithm is to infer the safety of the climber given these inputs.

General algorithm proceeds in three steps
  Use EM algorithm to cluster
  Use z-test to find p-value of test samples and determine changes in system state
  Logically resolve new state to provide diagnostic information

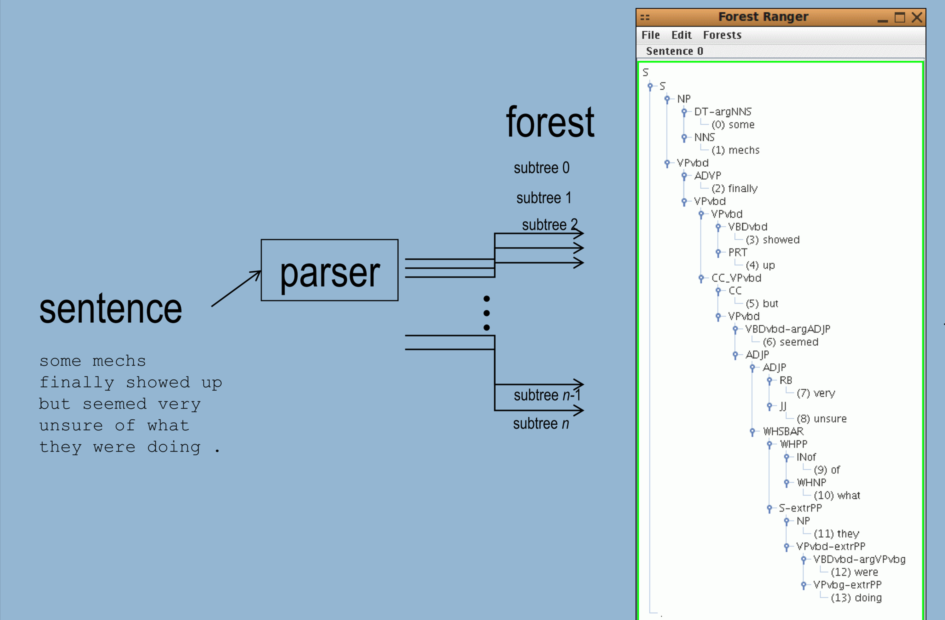
Positive results are also obtained on fault detection on the NASA ADAPT electrical dataset. Future work involves tests with flight data





[3] **N. Srivastava, A. Lazarevic, J. Srivastava**, Anomaly detection in complex systems, *NASA Conference on Intelligent Data Understanding*, October 14-16, 2009. Moffett Field, CA.

Annotating textual data



Developed an annotation tool [4]
  Accessible at http://sourceforge.net/projects/modelblocks/
  Forest Ranger v1.0
    Executable
    Source files with demo sentence
  Forest Ranger v1.1
    Tools for Semantic Knowledge Treebank (SNOWBANK) workflow

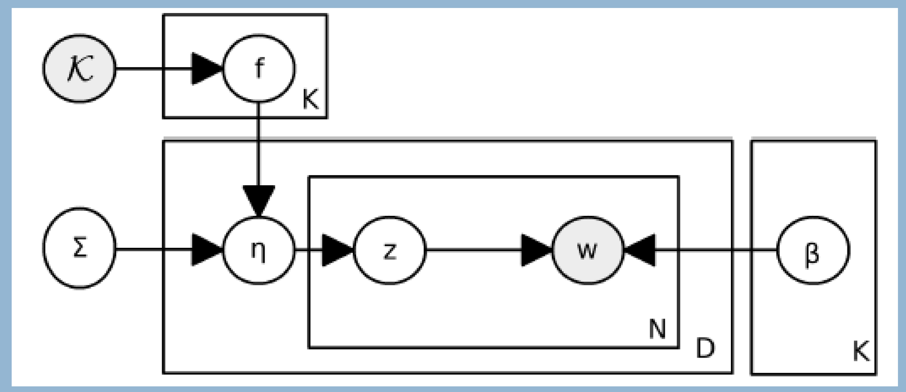
[4] **Schuler, William.** Positive Results for Parsing with a Bounded Stack using a Model-Based Right-Corner Transform. *Proc. of the North American Assoc. for Comp. Linguistics* (NAACL '09), Boulder, Colorado, 2009.

Clustering textual data

• Scalable implementation of fast D-LDA in C
• Good preliminary results on ASRS data – 58 categories over 66309 documents
• Sample results:

Catogory	FDLDA	Mariana	Docs
other spatial deviation.altitude heading rule deviation	0.99%	1%	57
other spatial deviation.controlled flight towards terrain	1.02 %	3%	607
aircraft equipment problem.critical	18.02%	18%	16081
non adherence.clearance	21.36%	20%	17713
non adherence.published procedure	30.45%	33%	24787
other anomaly.other	40.05%	43%	27066

• Proposed novel Gaussian Process Topic Models (GPTM) :
  incorporates kernel among documents into model
  mapping from document space into topic space
  models topic correlations and document correlations



Acknowledgements

The research presented here is supported in part by NASA contract number NNX08AC36A and NSF grant number CNS-0931931.The ideas and opinions presented, however, are solely the authors', and in no way express, either directly or through implication, official position(s) of the sponsoring organizations.