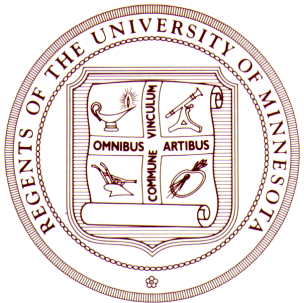
Anomaly Detection from Sensor Data: Streaming, Heterogeneous, Distributed

Jaideep Srivastava
University of Minnesota, Twin Cities
srivasta@cs.umn.edu

Presented by: Arindam Banerjee

ASIAS Tools and Technology Symposium

July 27-28, 2009



United Technologies Research Center

Project Team

- **University of Minnesota**
  - Jaideep Srivastava, PI
  - Vipin Kumar, Co-PI
  - William Schuler, Co-PI
  - Arindam Banerjee, Co-PI
  - Student Researchers
    - **Varun Chandola, Deepti Cheboli, Kuo-Wei Hsu, Tim Miller, Nishith Pathak, Lane Scwartz, Hanhuai Shan, Nisheeth Srivastava, Stephen Wu, Junlin Zhou**
- **UTRC**
  - Aleksander Lazarevic, Co-PI
  - Ashutosh Tewari

United Technologies Research Center

Talk Outline

- Anomaly Detection from Temporal Data
  - Discrete/symbolic sequences
  - Time Series
- Distributed Anomaly Detection
  - Homogenous data, multiple sources
- Conclusions & Future Work

United Technologies Research Center

Anomaly detection from Temporal Data

United Technologies Research Center

Objective of Research

- Detect anomalies in large databases of discrete sequences, univariate time series, multivariate time series, and heterogeneous sequence data.
- Understand the relation between the proposed as well as existing techniques and the nature of the data.
- **Relevance to the IVHM goal:**
  - Demonstrate automated anomaly detection in an offline mode on large heterogeneous datasets from multiple aircraft.
  - Generation of simulated data for testing of detection, diagnosis, and prognosis of anomalies on continuous, discrete
  - Implement and benchmark improved algorithms for fault diagnosis in offline mode on large heterogeneous data sets (continuous, discrete, and text) from multi-aircraft data systems.

Key Accomplishments

- Developed a framework to understand the nature of symbolic sequences in the context of anomaly detection*
- Investigated anomaly detection techniques that detect anomalies in univariate time series**
  - Developed several variants of existing techniques
  - Evaluated several techniques on publicly available data sets
  - Results connect the strengths and weaknesses of each technique to the nature of the time series data
- Developed a package of anomaly detection techniques for discrete sequences and time series data

* Understanding Anomaly Detection Techniques for Symbolic Sequences – Varun Chandola, Varun Mithal, and Vipin Kumar, Computer Science Technical Report (TR 09-001), 2009.

** Detecting Anomalies in a Time Series Database – Varun Chandola, Deepthi Cheboli, and Vipin Kumar, Computer Science Technical Report (TR 09-004), 2009.

Understanding Anomaly Detection Techniques for Symbolic Sequences

- **Problem:** Which anomaly detection technique is best suited for a given data set (of discrete sequences)?
  - A follow up analysis of our published experimental study*.
- **General Approach**:**
  - Identify characteristics to differentiate between normal and anomalous sequences.
  - For a given data set, measure the separability between the normal and anomalous sequences using the discriminating characteristics.
  - High separability => Good performance.
- **Advantages:**
  - Canonical characteristics to understand a wide variety of techniques.
  - Estimating optimal parameter settings.

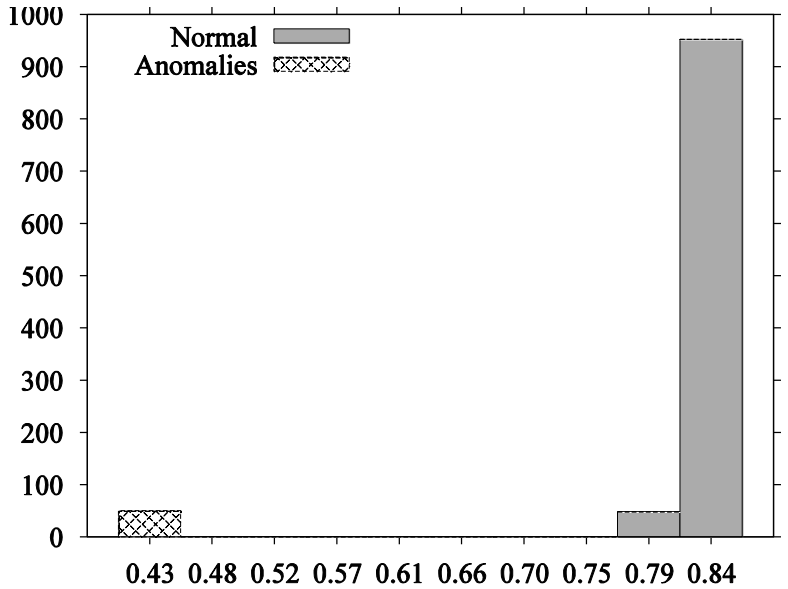
* A Comparative Evaluation of Anomaly Detection Techniques for Symbolic Sequences, Varun Chandola, Varun Mithal, and Vipin Kumar, In Proceedings of IEEE International Conference on Data Mining, December 2008.

** A similar approach was proposed in the context of anomaly detection for categorical data in – "A framework for analyzing categorical data, Varun Chandola, Shyam Boriah, and Vipin Kumar, To Appear in Proceedings of SIAM Data Mining (SDM) conference, April 2009.

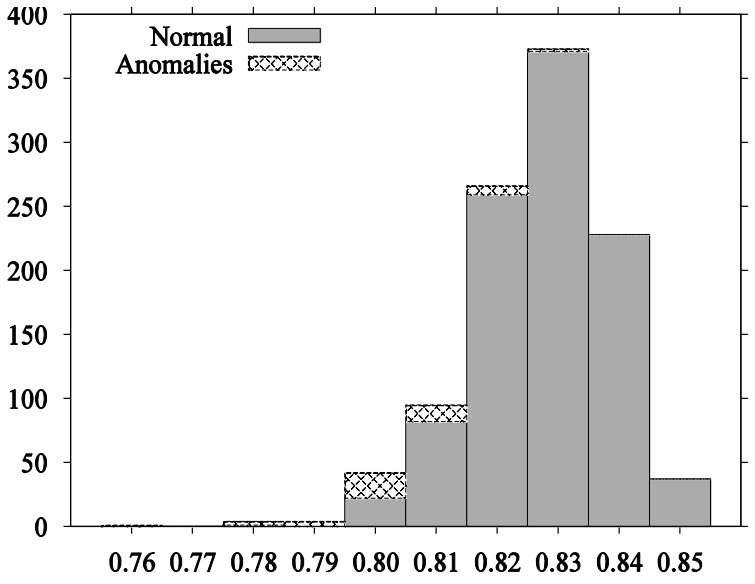
United Technologies Research Center

Understanding Kernel Based Techniques

- **Discriminating Characteristic**: Average similarity of normal and anomalous test sequences to training sequences
- Relates to CLUSTER and kNN.



CLUSTER – 100 %*
kNN – 100 %



CLUSTER – 64 %
kNN – 68 %

* Precision on anomaly class

United Technologies Research Center

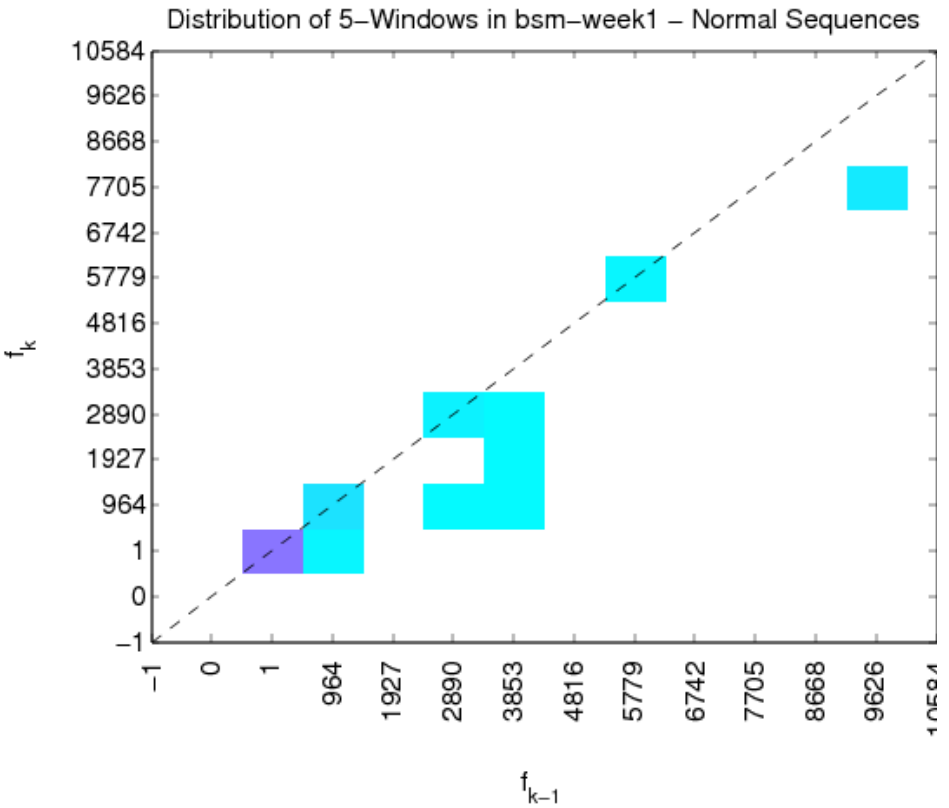
Understanding Window Based Techniques

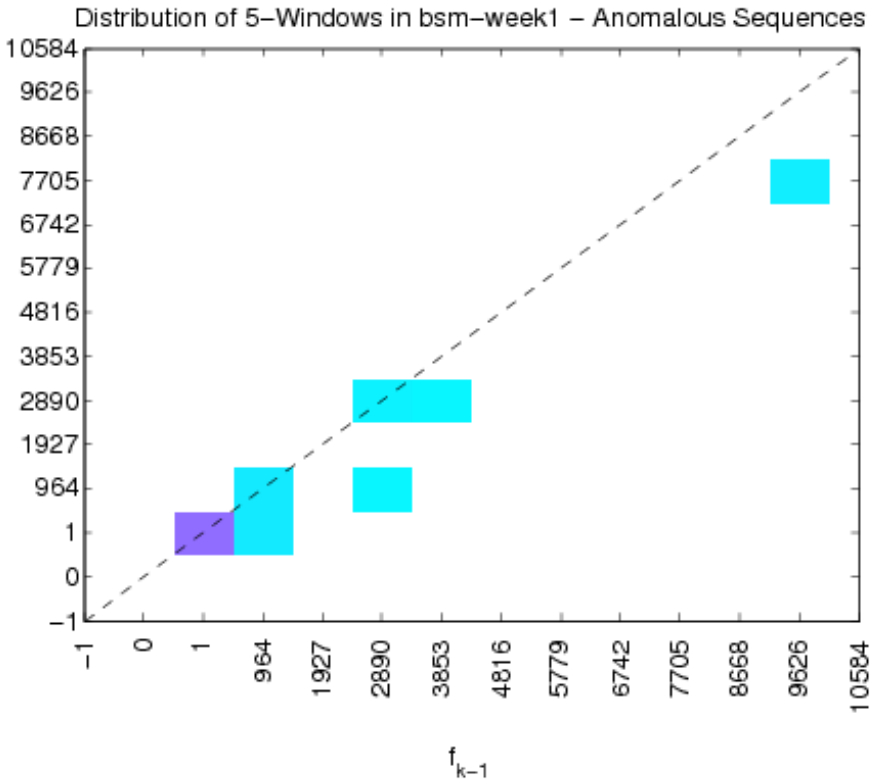
- **Discriminating Characteristic:**
  - Frequency of k length windows
  - Frequency of (k-1) length prefixes

- Novel visualization of discrete sequences: *frequency profiles*.
  - Frequency of each k-window and its (k-1) length prefix.
  - 2-D histogram of the frequency tuples.

United Technologies Research Center

Frequency Profiles for Discrete Sequences

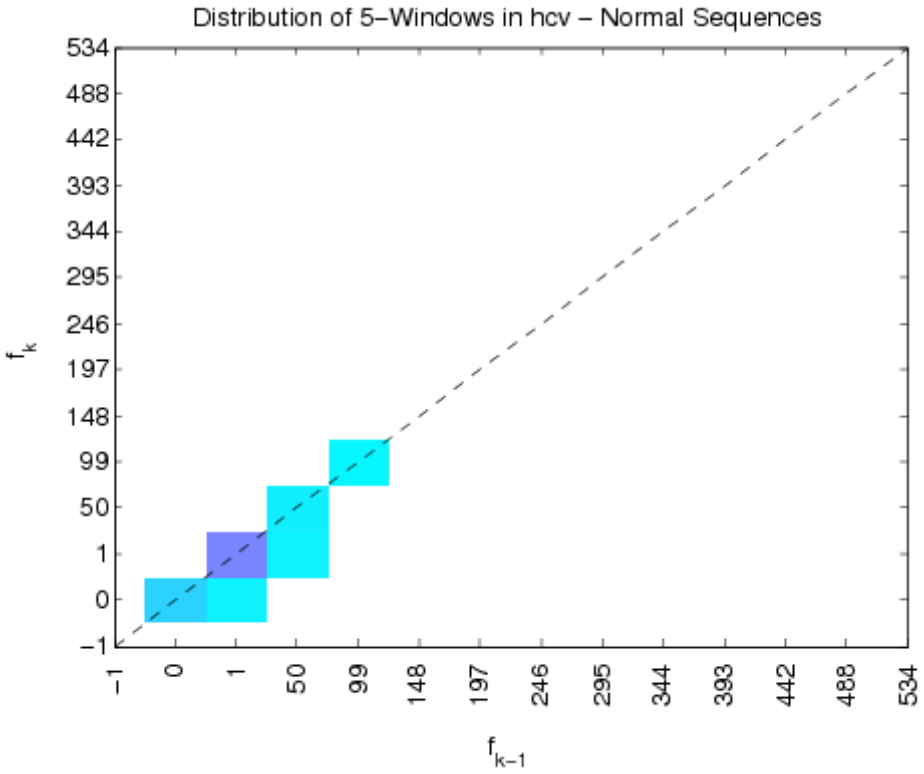


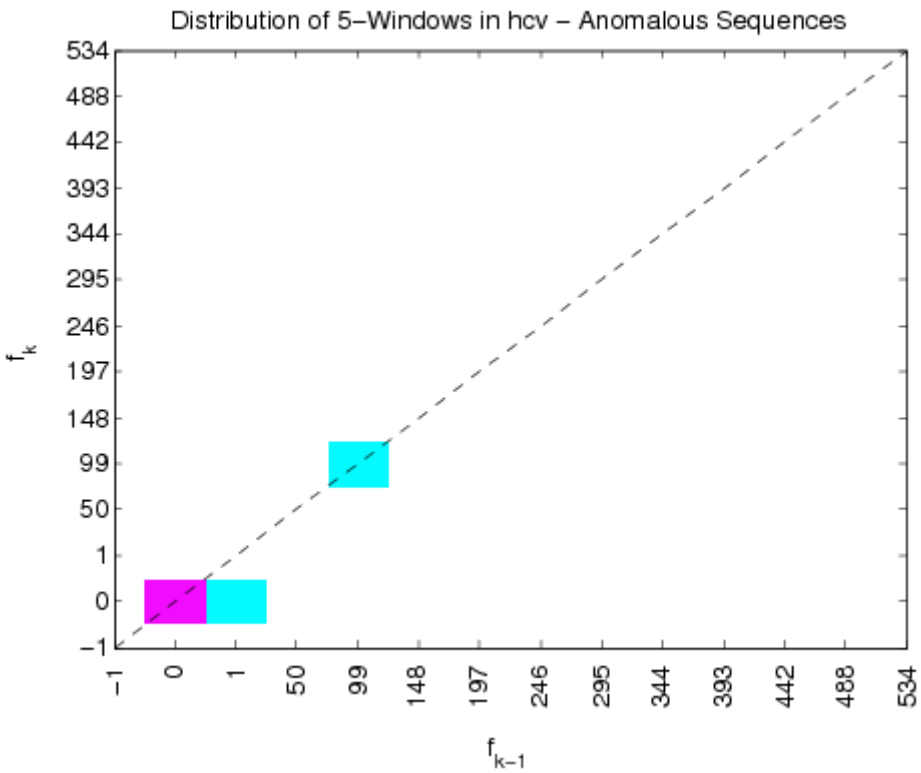


- The two profiles are similar.
- Performance of history based techniques is poor, tStide – 20 %, fsaz – 50 %

United Technologies Research Center

Frequency Profiles for Discrete Sequences





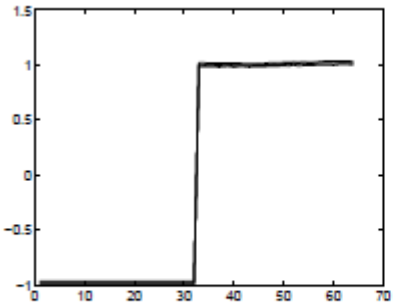
- The two profiles are significantly different.
- Performance of history based techniques is good, tStide – 90 %, fsaz – 92 %

United Technologies Research Center

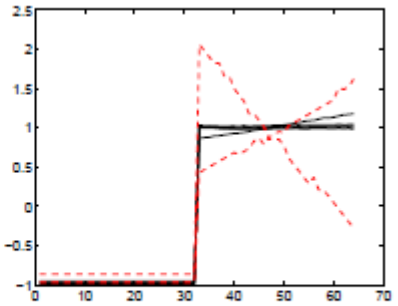
Detecting Anomalies in a Time Series Database

- **Problem:** Assign anomaly score to a test time series (univariate) with respect to a training data base of normal time series.

- Evaluated a suite of anomaly detection techniques for this task.
  - Kernel based
    - Using different distance/similarity measures
  - Window based
  - Predictive model based
    - Auto Regressive
    - Support Vector Regression
  - State based
    - Box model based*



(a) Reference



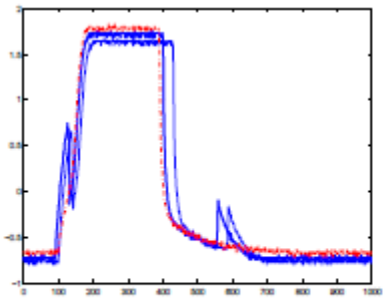
(b) Test

*P. K. Chan and M. V. Mahoney. Modeling multiple time series for anomaly detection. ICDM, p. 90–97, 2005.

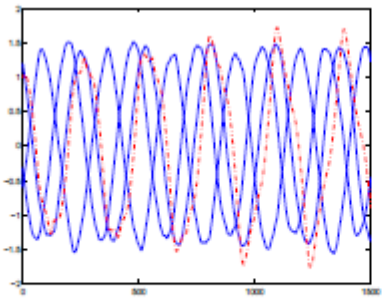
United Technologies Research Center

Hypothesis

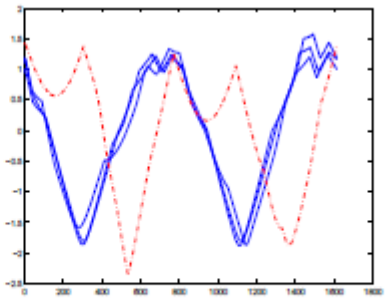
- Time series data sets have different characteristics.
- A technique shown to perform well for one type of data is not guaranteed to perform well on a different type of data.



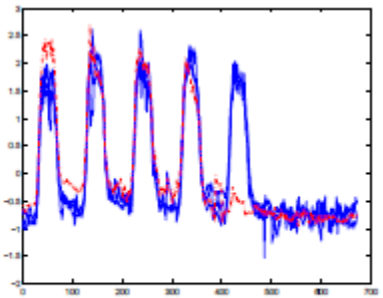
(a) Valve



(b) Motor



(c) Shapes



(d) Power

United Technologies Research Center

Datasets Used

			|Y|						
Name	L	$\|X\|$	$\|Y_N\|$	$\|Y_A\|$	λ	A	P	S	l
disk1	64	10	500	50	0.09	s	×	✓	–
disk2	64	10	500	50	0.09	s	×	✓	–
disk3	64	10	500	50	0.09	s	×	✓	–
motor1	1500	10	10	10	0.50	p	✓	×	250
motor2	1500	10	10	10	0.50	p	✓	×	250
motor3	1500	10	10	10	0.50	p	✓	×	250
motor4	1500	10	10	10	0.50	p	✓	×	250
power	672	11	33	8	0.19	s	✓	✓	96
valve1	1000	4	4	8	0.67	s	×	✓	300
shape1	1614	10	10	10	0.50	p	×	×	–
shape2	1614	30	30	10	0.25	p	×	×	–
chfdb1	2500	250	250	25	0.09	s	✓	×	250
chfdb1	2500	250	250	25	0.09	s	✓	×	250
ltstdb1	2500	250	250	25	0.09	s	✓	×	250
ltstdb2	2500	250	250	25	0.09	s	✓	×	250
edb1	250	500	500	50	0.09	p	×	×	50
edb2	250	500	500	50	0.09	p	×	×	50
mitdb1	360	500	500	50	0.09	p	×	×	50
mitdb2	360	500	500	50	0.09	p	×	×	50

*** All data sets are available for download at www.cs.umn.edu/~chandola/timeseries**

Results

Data	KNNC	KNND	WINC	WIND	STIDE	SVR	AR	FSAz	BOX
disk1	0.88	0.26	0.98	0.09	0.32	0.09	0.74	0.08	0.94
disk2	0.96	1.00	0.96	0.09	1.00	1.00	0.40	1.00	0.92
disk3	0.96	0.96	1.00	0.52	0.96	0.98	0.48	0.96	0.94
motor1	0.60	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.80
motor2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.80
motor3	0.90	1.00	1.00	1.00	1.00	1.00	1.00	0.90	0.90
motor4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.90
power	0.88	0.88	0.88	0.62	0.75	0.62	0.50	0.62	0.50
valve1	0.75	1.00	0.75	0.75	0.75	0.75	0.75	1.00	0.62
shape1	1.00	1.00	1.00	0.50	0.80	1.00	1.00	0.80	0.80
shape2	0.80	0.90	0.50	0.25	0.20	0.30	0.00	0.40	0.70
chfdb1	0.20	0.36	0.40	0.72	0.24	0.48	0.20	0.52	0.72
chfdb2	0.40	0.12	0.32	0.24	0.04	0.04	0.00	0.16	0.04
ltstdb1	0.52	0.12	0.56	0.40	0.16	0.04	0.00	0.16	0.60
ltstdb2	0.44	0.28	0.28	0.20	0.32	0.04	0.20	0.24	0.08
edb1	0.74	0.76	0.78	0.74	0.56	0.74	0.02	0.74	0.66
edb2	0.30	0.30	0.16	0.14	0.12	0.36	0.00	0.22	0.10
mitdb1	0.78	0.70	0.90	0.66	0.32	0.70	0.00	0.38	0.18
mitdb2	0.94	0.86	0.94	0.84	0.56	0.90	0.02	0.62	0.18
Avg	0.74	0.71	0.76	0.57	0.58	0.63	0.44	0.62	0.60
time (s)	8	291	73	609	2	7	1	18	857

United Technologies Research Center

Conclusions

- Window based and KNN based techniques generally perform very well for most data sets.
  - Advantage of KNN based techniques : Faster than WIN
  - Advantage of Window based techniques : Can be used for online anomaly detection.
- Predictive and State based models do not seem to perform well.
- Performance of techniques on data in discretized domain is inferior to its continuous counter part.
- Nature of Data Vs Technique
  - Non periodic time series : KNN with distance measure DTW
  - Periodic : Window based techniques.
  - Window based techniques perform poorly when compared to KNN if the data is from multi-modal distribution.

SQUAD Package

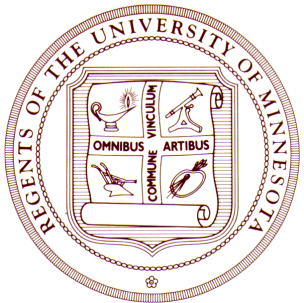
- Developed a SeQUence Anomaly Detection (SQUAD) package for detecting anomalies in symbolic sequences and time series data.
- The package is available as a GNU installation package from http://www.cs.umn.edu/~chandola/squad/squad.php
- The package consists of seqlib library that contains routines for reading and writing sequences/time series data and to compute similarity/distance.
- Written in C++

Future Plan

- Extend the techniques to handle multivariate time series and multivariate heterogeneous sequences
  - Define similarity/distance measures to handle multivariate time series and hetereogenous sequences*.
- Develop novel techniques to handle multivariate time series
  - Linear dynamical systems based.
  - Covariance structure monitoring based.

* Similarity Measures for Categorical Data: A Comparative Evaluation, Shyam Boriah, Varun Chandola, and Vipin Kumar, SDM 2008, April 2008.

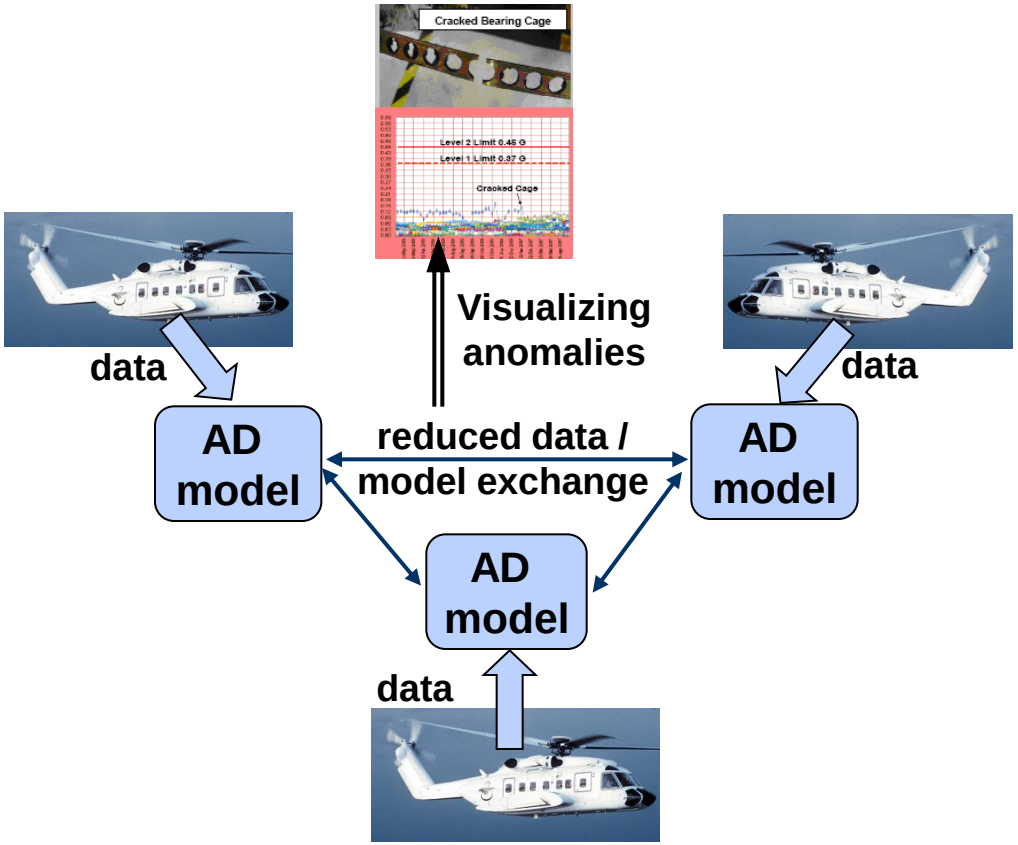
Distributed Anomaly Detection



United Technologies
Research Center

Objective of Research

Identify anomalous events or trends from multiple, homogeneous data sources



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
- 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

Hypothesis

Relevance to IVHM goals, Benefits and Risks

Relevance to the IVHM goal:

Efficient identification and visualization of anomalous events across multiple, homogeneous data sources can be successfully associated with detecting multiple faults/failures, their diagnosis, and allow prognostic and mitigation decisions.

- Anomaly/Trend/Change Detection
- Distributed Anomaly Detection
- Visualizing Anomalous Events

IVHM System

HUMS data (condition & health indicators from all aircraft and their flights)

Main Gearbox / Tail GearBox

- Detect any abnormal events, short-term and long-term temporal trends that lead to faults from multiple aircraft
- Illustrate detected anomalies from hi-dimensional space in a simple and understandable manner

Major Benefits to IVHM:

- Early detection of failures (faults) in the aircraft and improve the aircraft safety
- Reduce the maintenance cost through Condition Based Maintenance (CBM)

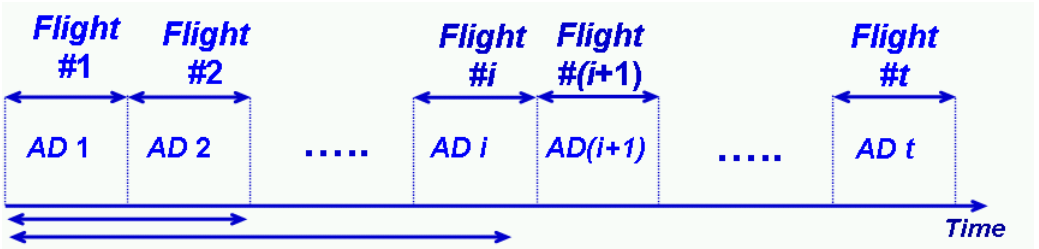
Critical Risk Items:

- Preprocessed HUMS data may not be sufficient to capture anomalous data records
- Flight sequences are short and of varying lengths (average length close to 15).

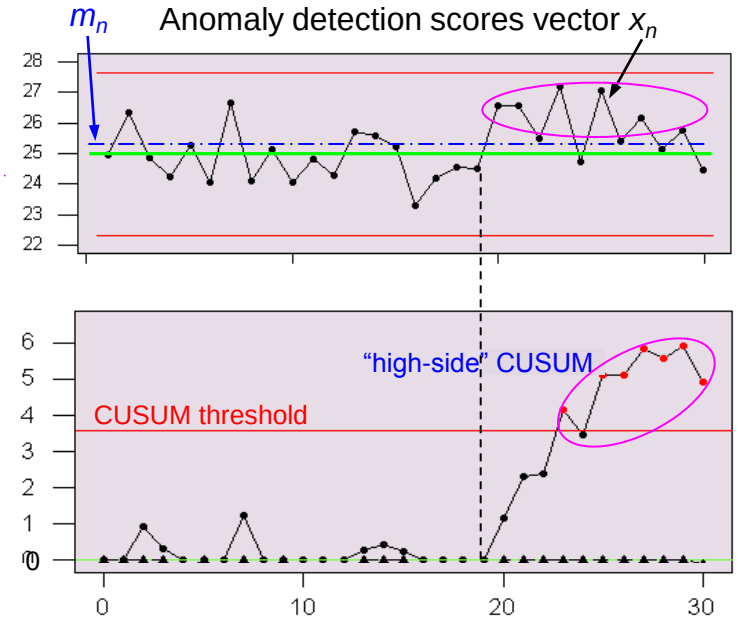
Fast Anomaly Detection (AD) Framework

Integrating T^2 and Q statistics with CUSUM approach

- Data preprocessing
  - Reduce dimensionality of flight data using Principal Component Analysis or diffusion maps
- Perform multivariate distance and density based AD after each flight:
  - T2 statistics based anomaly detection
  - Q statistics based anomaly detection



- Generate AD scores (scores proportional to probability of data records being anomalies)
- Perform "high-side" CUSUM on AD scores
  - $S_0 = 0$; $S_{n+1} = \max(0, S_n + x_n - m_n)$
  - where x_n is AD vector, m_n assigned weights
- Identify variables most relevant to the detected fault



T^2 statistics based Anomaly Detection Methods

Successfully used in practice to detect faults from multivariate process data[1]

- X is a given data set of *n* data records and *m* process variables
- T^2 statistics for normalized X is defined as: $$T^2 = n \cdot X^T \cdot \Sigma^{-1} \cdot X$$

  where Σ is sample covariance matrix defined as: $$\Sigma = \frac{1}{\sqrt{n-1}} X^T \cdot X$$
- An eigenvalue decomposition of the matrix Σ is: $$\Sigma = V \cdot \Lambda \cdot V^T$$

  where V is eigenvactor matrix, and Λ is diagonal eigenvalue matrix
- If we only retain a eigenvectors corresponding to a largest singular values, we can stack them into a m x a matrix P

$$T^2 = \frac{n}{\sqrt{n-1}} \cdot X^T \cdot P \cdot \Lambda_a^+ \cdot P^T \cdot X$$

- Threshold for T^2 statistics is: $$T_\alpha^2 = \frac{a(n-1)(n+1)}{n(n-a)} \cdot F_\alpha(a, n-a)$$

Q statistics based Anomaly Detection Methods

- Unlike T^2 statistics, Q statistics focuses on *m-a* smallest singular values[1]
- Q statistics is computed as:

$$Q = e^T \cdot e$$

- e is the residual vector computed as: $e = (I - P \cdot P^T) \cdot x$

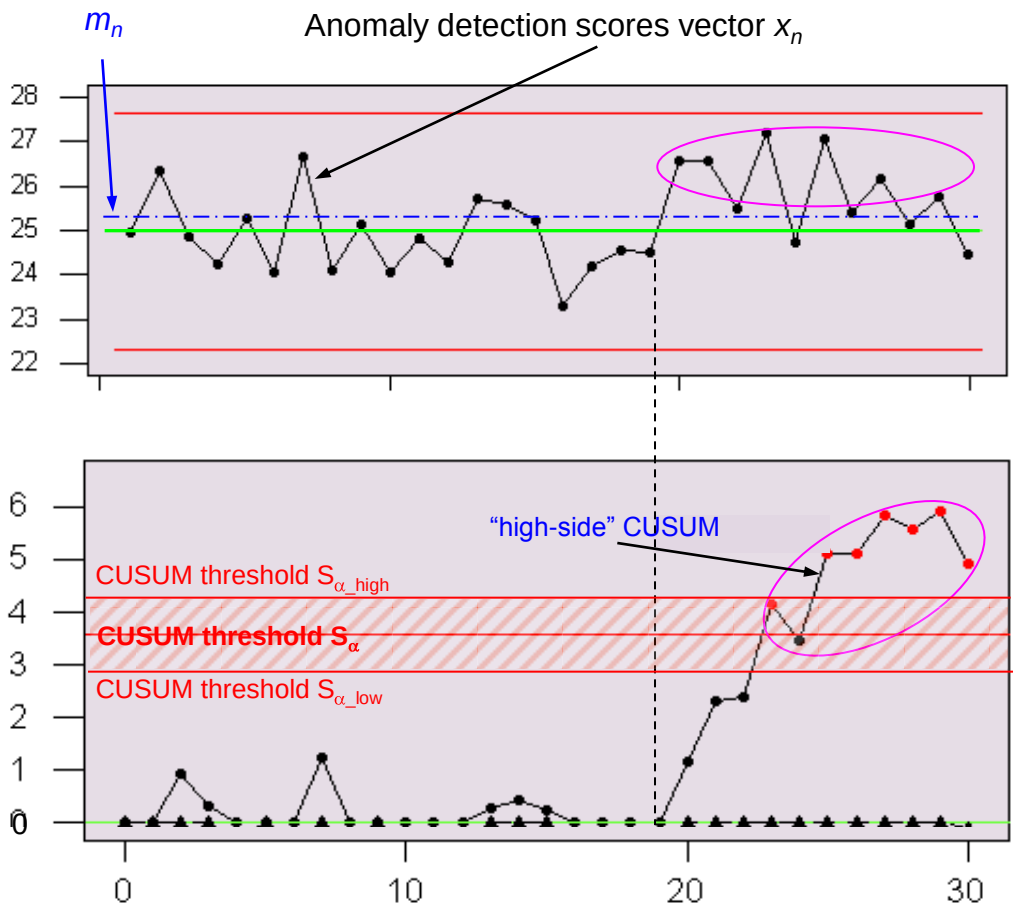
- Threshold for Q statistics is:

$$Q_\alpha = \theta_1 \left[\frac{h_0 c_\alpha \cdot \sqrt{2\theta_2}}{\theta_1} + 1 + \frac{\theta_2 h_0 (h0 - 1)}{\theta_1^2} \right]^{1/h_0}$$

$$\theta_i = \sum_{j=a+1}^{n} \sigma_j^{2i}, \; h_0 = 1 - \frac{2\theta_1 \theta_3}{3 \cdot \theta_2^2}$$

CUSUM Persistence Anomaly Detection

- CUSUM is a sequential analysis technique that is typically used for change detection
- Perform "high-side" CUSUM on AD scores to detect changes only in the positive direction:
  - $S_0 = 0$
    $S_{n+1} = \max(0, S_n + x_n - m_n)$
  - where x_n is AD vector, m_n are assigned weights
- CUSUM Adaptations
  - hysteresis procedure - interval around the threshold line $[S_{\alpha\_low}, S_{\alpha\_high}]$ is introduced such that the CUSUM curve has to pass not only S_{α_low} but also S_{α_high}.



  - Exponential decrease of CUSUM curve after period of inactivity is implemented ($m_{n+1} = a \cdot m_n$), where $a > 1$ (typically chosen between 1.1 and 1.5)

Fast Distributed Anomaly Detection (AD) Framework

Distributed Integrated T^2 and Q statistics with CUSUM approach

- Assume 2 data sites with data sets X_1 and X_2
- For merged data $X_{nxm} = X_1 \cup X_2$, T^2 statistics is defined as: $T^2 = n \cdot X^T \cdot \Sigma^{-1} \cdot X$
- T^2 statistics for X_1 and X_2 are: $T_1^2 = n_1 \cdot X_1^T \cdot \Sigma_1^{-1} \cdot X_1$ and $T_2^2 = n_2 \cdot X_2^T \cdot \Sigma_2^{-1} \cdot X_2$
- Sample covariance matrix Σ for data X_{nxm} ($n = n_1 + n_2$) is:

$$\Sigma = \frac{1}{n-1}(X-\mu)^T \cdot (X-\mu) = \begin{bmatrix} \sigma_{1,1} & \dots & \sigma_{1,m} \\ \dots & \dots & \dots \\ \sigma_{n,1} & \dots & \sigma_{n,m} \end{bmatrix}, \quad \sigma_{i,j} = \frac{1}{n-1}\sum_{k=1}^{n}(x_i(k)-\mu_i)\cdot(x_j(k)-\mu_j) = \frac{1}{n-1}\sum_{k=1}^{n}x_i(k)\cdot x_j(k) - \frac{n}{n-1}\mu_i \cdot \mu_j$$

- Sample covariance matrices for X_1 and X_2 are defined as:

$$\Sigma_1 = \frac{1}{n_1-1}(X_1-\mu_1)^T \cdot (X_1-\mu_1) = \begin{bmatrix} \sigma^1_{1,1} & \dots & \sigma^1_{1,m} \\ \dots & \dots & \dots \\ \sigma^1_{n_1,1} & \dots & \sigma^1_{n_1,m} \end{bmatrix}, \quad \sigma^1_{i,j} = \frac{1}{n_1-1}\sum_{k=1}^{n_1}(x^1{}_i(k)-\mu^1{}_i)\cdot(x^1{}_j(k)-\mu^1{}_j) = \frac{1}{n_1-1}\sum_{k=1}^{n_1}x^1{}_i(k)\cdot x^1{}_j(k) - \frac{n_1}{n_1-1}\mu^1{}_i \cdot \mu^1{}_j$$

$$\Sigma_2 = \frac{1}{n_2-1}(X_2-\mu_2)^T \cdot (X_2-\mu_2) = \begin{bmatrix} \sigma^2{}_{1,1} & \dots & \sigma^2{}_{1,m} \\ \dots & \dots & \dots \\ \sigma^2{}_{n_1,1} & \dots & \sigma^2{}_{n_1,m} \end{bmatrix}, \quad \sigma^2{}_{i,j} = \frac{1}{n_2-1}\sum_{k=1}^{n_2}(x^2{}_i(k)-\mu^2{}_i)\cdot(x^2{}_j(k)-\mu^2{}_j) = \frac{1}{n_2-1}\sum_{k=1}^{n_2}x^2{}_i(k)\cdot x^2{}_j(k) - \frac{n_2}{n_2-1}\mu^2{}_i \cdot \mu^2{}_j$$

Fast Distributed Anomaly Detection (AD) Framework

Distributed Integrated T^2 and Q statistics with CUSUM approach

- How to estimate T^2 statistics without merging X_1 and X_2 ?
- Element of covariance matrix Σ for data X is defined as:

$$\sigma_{i,j} = \frac{1}{n-1}\sum_{k=1}^{n} x_i(k)\cdot x_j(k) - \frac{n}{n-1}\mu_i\cdot\mu_j = \frac{1}{n_1+n_2-1}\left(\sum_{k=1}^{n_1} x_i(k)\cdot x_j(k) + \sum_{k=n_1+1}^{n_1+n_2} x_i(k)\cdot x_j(k)\right) - \frac{n_1+n_2}{n_1+n_2-1}\mu_i\cdot\mu_j$$

$$\sigma_{i,j} = \frac{1}{n_1+n_2-1}\left[(n_1-1)\sigma_{1,i,j} + n_1\cdot\mu_{1,i}\cdot\mu_{1,j} + (n_2-1)\sigma_{2,i,j} + n_2\cdot\mu_{2,i}\cdot\mu_{2,j}\right] - \frac{n_1+n_2}{n_1+n_2-1}\mu_i\cdot\mu_j$$

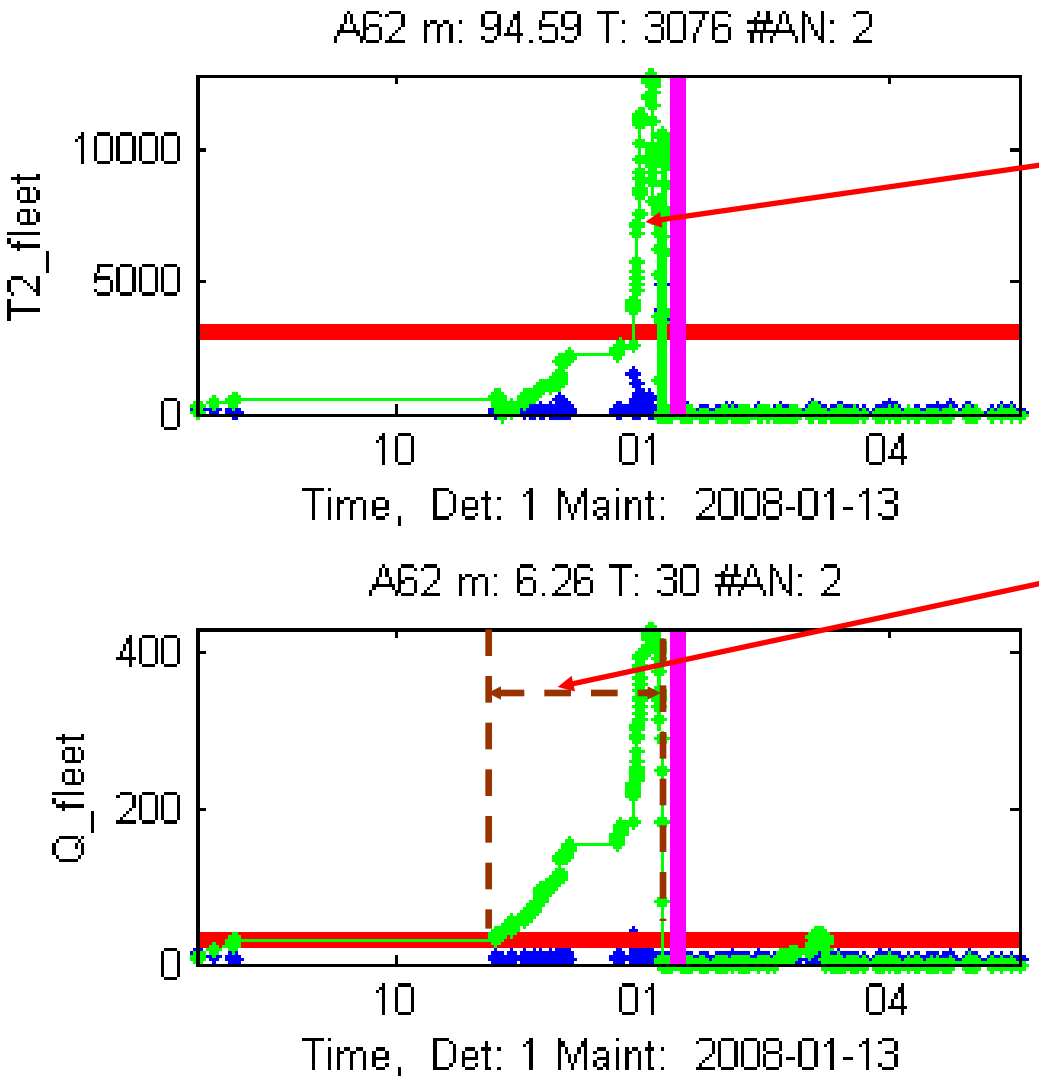
- and $\mu_i = \dfrac{n_1\mu_{1,i} + n_2\mu_{2,i}}{n_1+n_2}$
- Covariance matrix Σ and mean can be expressed as:

$$\Sigma = \frac{1}{n_1+n_2-1}\left[(n_1-1)\Sigma_1 + n_1\cdot\mu_1{}^T\cdot\mu_1 + (n_2-1)\Sigma_2 + n_2\cdot\mu_2{}^T\cdot\mu\right] - \frac{n_1+n_2}{n_1+n_2-1}\mu^T\cdot\mu, \qquad \mu = \frac{n_1\mu_1 + n_2\mu_2}{n_1+n_2}$$

- **Exact T^2 & Q statistics** for dataset X can be computed* by exchanging only covariance matrices and mean vectors from individual data sets X_1 and X_2

Experimental results using aircraft and fleet level models

All replacement events are detected from data with replacement information



Replacement event on aircraft #62 on Jan 13, 2008

pred - how many days before the replacement event we detect this event as anomaly

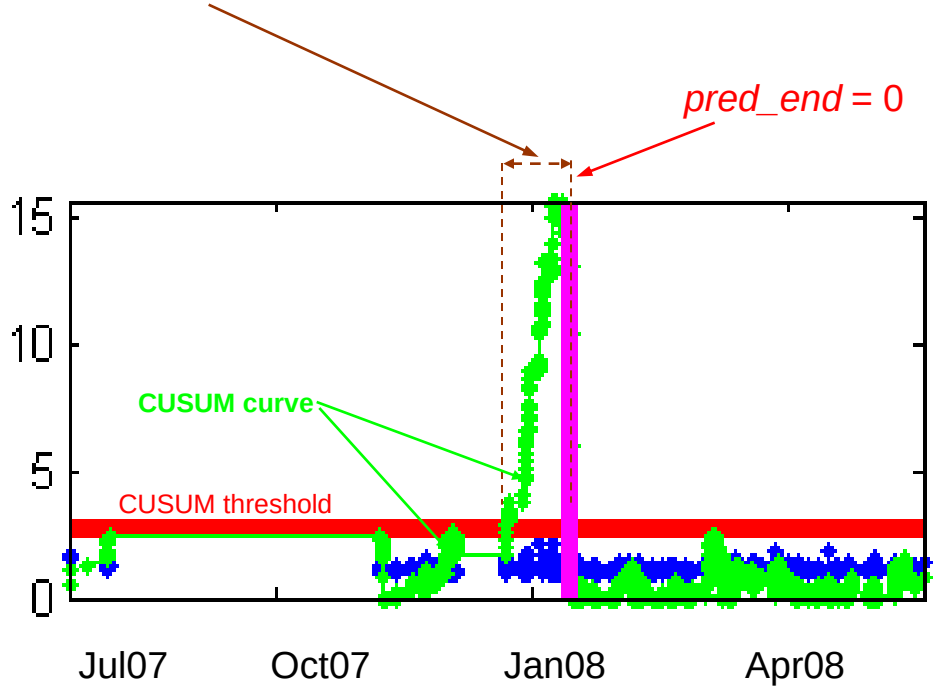
For all aircraft for which we had information about replacement events, T^2 statistics based and incremental density based (LOF) AD algorithms were able to detect all of them in advance!

United Technologies Research Center

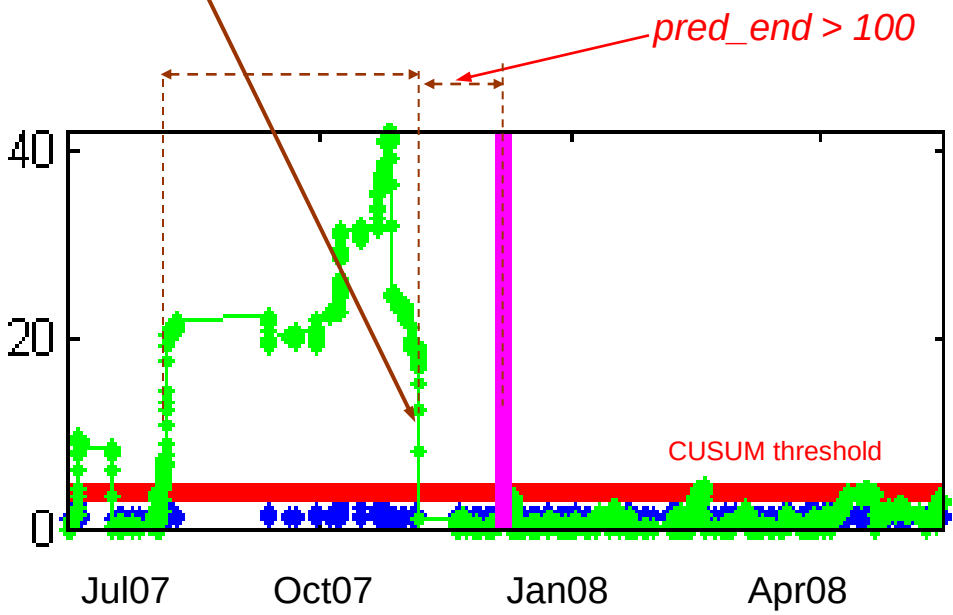
Proposed Anomaly Detection (AD) Metrics

Timeliness of detecting a maintenance event

an_days - Total number of days (data records) when the CUSUM curve is above threshold



If CUSUM curve falls more than 100 records before actual maintenance, we consider that we **DID NOT** detect the maintenance event!

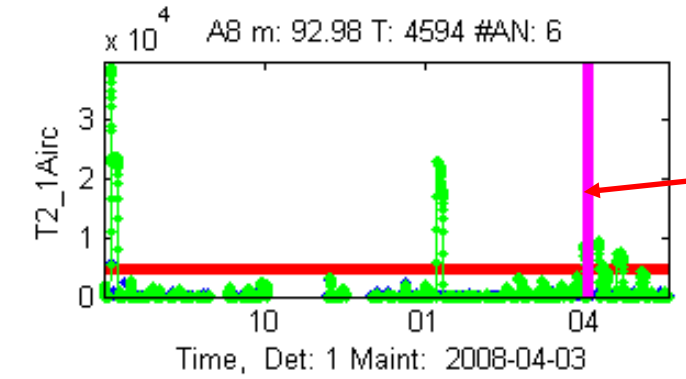


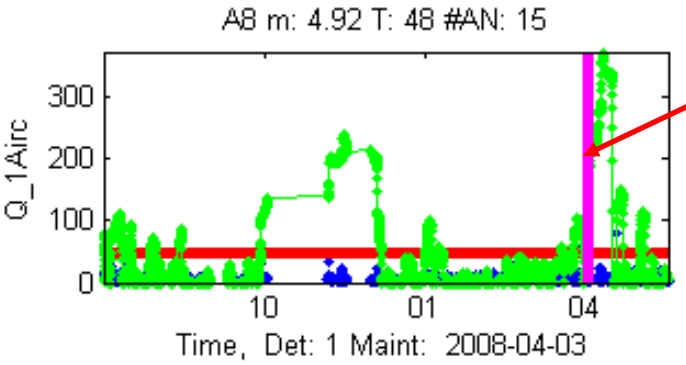
Goodness of early detection of maintenance event is defined as:

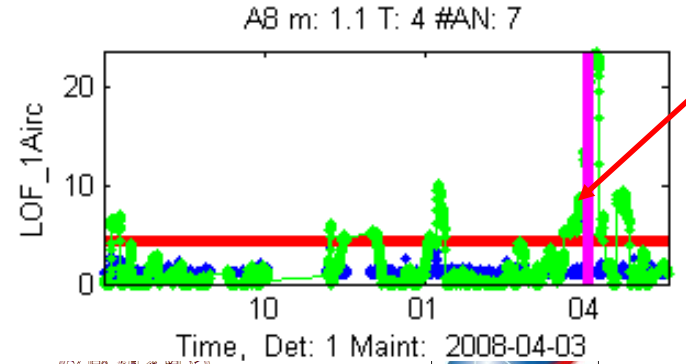
$$dr\_time = \begin{cases} \dfrac{an\_days(most\ \ recent\ \ detected\ \ anomaly)}{pred\_end + \varepsilon}, & if\ \ pred\_end < 100\ \ data\ \ records \\ 0 & otherwise \end{cases}$$

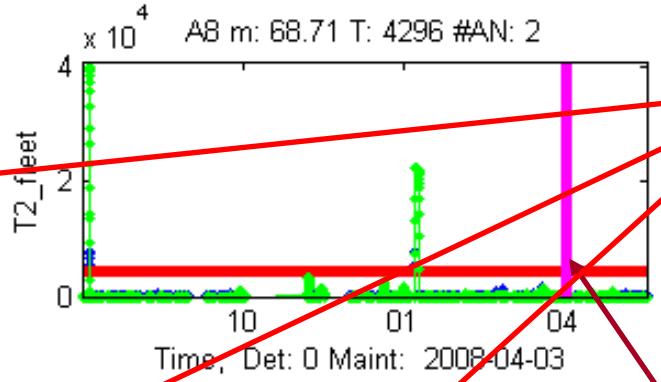
Experimental results using aircraft and fleet level models

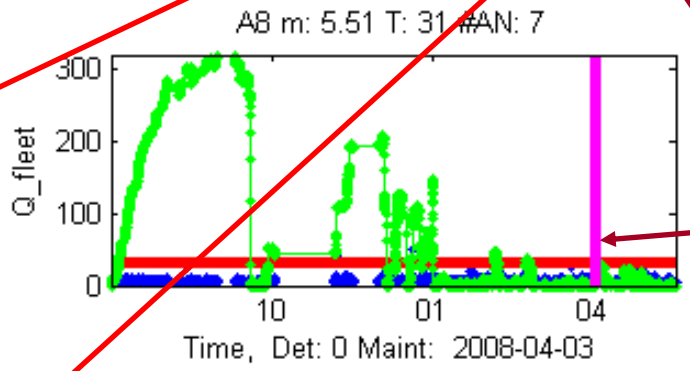
TGB replacement event was detected in advance for aircraft #8

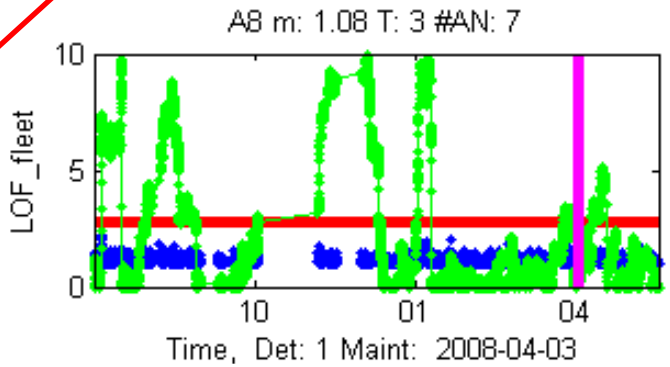












Detected replacement event on the aircraft #8:

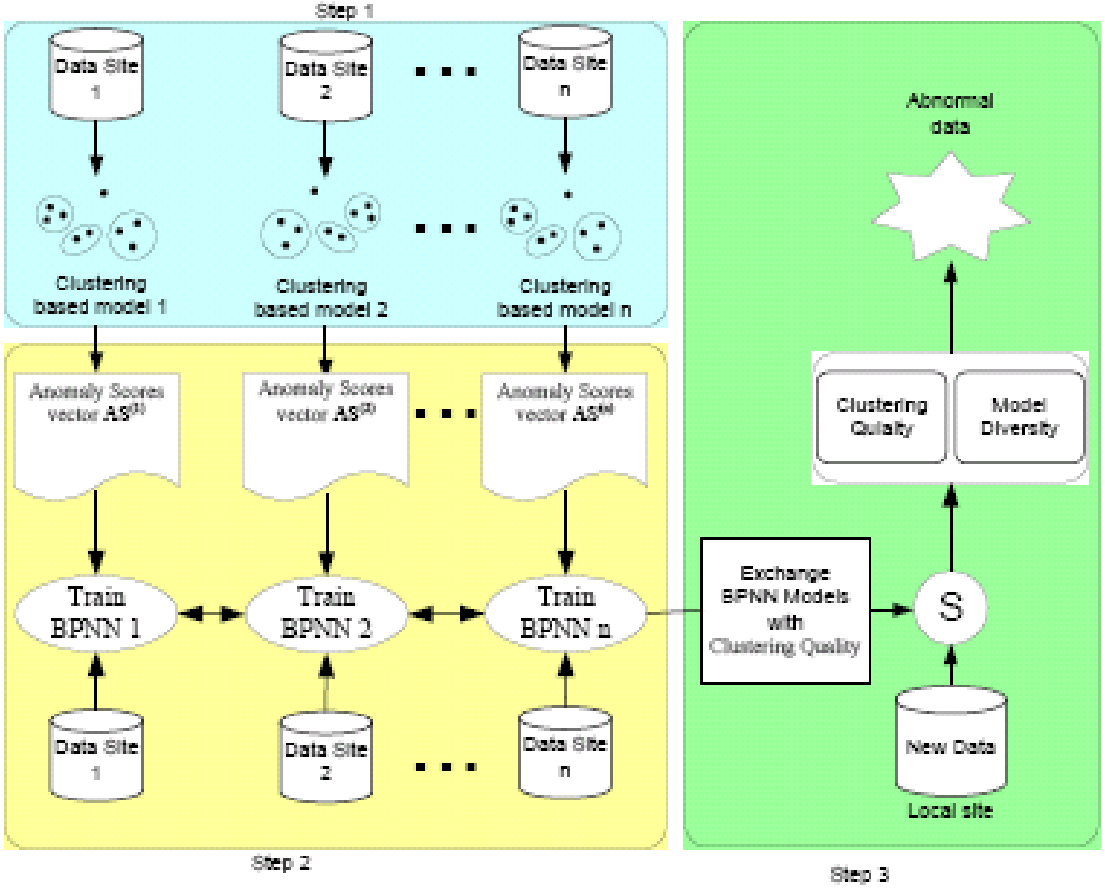
April 3, 2008

Missed detection of replacement event

United Technologies Research Center

Privacy-Preserving Combining Anomaly Detection Methods

Quality and diversity based combining



- Main idea:
  - Perform clustering and identify modes of normal behavior*
  - Compute anomaly detection score as a Mahalanobis distance to the closest cluster
  - Build regression local models (BPNNs) to learn anomaly detection score from each data set
  - Combine local modes to detect global anomalies by using both quality and diversity

United Technologies Research Center

Methodology

- Combine local models' results by model quality and diversity
  - Quality - The performance of anomaly detection is related to the clustering quality of the uniform model
    - Silhouette index (SI) - reflecting the compactness and separation of clusters
    - Davies-Bouldin (DB) - Average similarity between each cluster
    - Dunn index (DI) - How similar the objects are within each cluster and how well the objects of different clusters are separated
    - Calinski-Harabasz (CH) - centroid intra-cluster and inter-cluster distances
  - Diversity- Diversity plays a significant role in combining prediction models, higher diversity leads to higher predict accuracy.
    - Adjusted Rand index (AR)
    - Jaccard index (JI)
    - Fowlkes-Mallows index (FM)

United Technologies Research Center

Experiment results

- Set up
  - Data set:
    - Synthetic
    - KDDCUP 1999
    - Mammography
    - Rooftop
    - Satimage
    - NASA data
    - Sikorsky data
  - Data distributed into five (ten for KDD data) local sites
- Measures
  - F-value used for Anomaly detection performance
  - Clustering quality used for local model quality
  - Agreement on test data used for local model diversity

Experiment results

F-MEAUSURE COMPARISON FOR COMBINATION MODEL AND GLOBAL MODEL ON ALL DATA SETS

Dataset	Quality / Diversity / Model	Silhouette index			Davies-Bouldin			Calinski-Harabasz			Dunn index		
		AR	JA	FM	AR	JA	FM	AR	JA	FM	AR	JA	FM
Synthetic	CoM	0.9843	**0.9873**	0.9867	0.9885	0.9836	0.9836	0.9861	0.9836	0.9861	0.9824	0.983	0.985
	GlM	**0.987(DBSCAN)**			0.973(SOM)						0.976(K-means)		
KDD	CoM	0.9963	0.9965	0.9963	0.9968	0.9968	**0.9970**	0.9963	0.9968	0.9968	0.9963	0.9968	0.9965
	GlM	**0.99667 (DBSCAN)**			0.99632 (SOM)						0.99489 (K-means)		
Mg	CoM	**0.9795**	0.9723	0.9783	0.9717	0.9759	0.9686	0.9767	0.9677	0.9669	0.9791	0.9739	0.9783
	GlM	0.97949(DBSCAN)			**0.98033(SOM)**						0.97932(K-means)		
Rooftop	CoM	**0.9656**	0.9653	0.9653	0.9648	0.9650	0.9650	0.9651	0.9650	0.9705	0.9624	0.9625	0.962
	GlM	**0.97663(DBSCAN)**			0.96836(SOM)						0.96283(K-means)		
Satimage	CoM	0.9196	0.9289	0.933	0.9333	**0.9368**	0.9272	0.9325	0.9338	0.9285	0.9196	0.9289	0.933
	GlM	**0.93294(DBSCAN)**			0.9271(SOM)						0.9306(K-means)		
NASA	CoM	0.65	0.7373	0.66	0.6326	0.65	0.632	**0.7655**	0.6294	0.6764	0.6326	0.6532	0.6567
	GlM	**0.70518(DBSCAN)**			0.70368(SOM)						0.69214(K-means)		

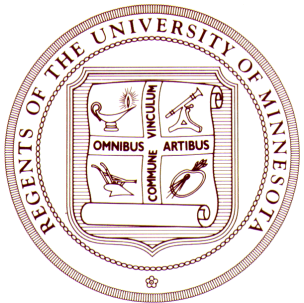
Legend: KDD = KDDCUP 1999, Mg = Mammo-graphy , CoM = Combined Model(The model combined by distributed models), GlM = Global Model(The model built by collecting all the distributed data sets, the global model is not available in most cases, here we build it just for performance evaluation), AR = Adjusted Rand index, JA = Jaccard index, FM = Fowlkes-Mallows index

Next Steps, Issues, Concerns, Risks

- Next Steps:
  - Demonstrate capability of fast distributed anomaly detection algorithms on appropriate very large data sets (10GB per site)
- Issues, Concerns, Risks:
  - Selection of appropriate large data set (10GB per site)
  - Relevant data sets are not readily available
  - How to verify the performance of anomaly detection algorithms in the absence of ground truth data

Thank You!

Questions/Comments: srivasta@cs.umn.edu



United Technologies
Research Center