PHM Data Challenge 2015

The PHM Data Challenge is a competition open to all potential conference attendees. This year the challenge is focused on fault detection and prognostics, a common problem in industrial plant monitoring. Participants will be scored on their ability to detect plant faults from a set of potential faults and additionally to precisely localize faults in time.

This is a fully open competition in which collaboration is encouraged. The teams may be composed of any combination of students, researchers, and industry professionals. The results will be evaluated by the Data Challenge Committee and all teams will be ranked. The top scoring teams will be invited to present at a special session of the conference and first and second place finishers will be recognized at the Conference Banquet event.

Data Challenge Chairs

Justinian Rosca, Siemens

Nicholas Williard, Schlumberger

Neil Eklund, Schlumberger

Zhen Song, Siemens

Teams

The team judged to have the first and second best scores will be awarded prizes of $600 and $400 respectively contingent upon:

- Having at least one member of the team attend the PHM 2015 Conference
- Presenting the analysis results and technique employed at a special session within the Conference program
- Submitting a peer-reviewed Conference paper. (Submission of the challenge special session papers is outside the regular paper submission process and follows its own modified schedule.)

The top entries will also be encouraged to submit a journal-quality paper to the International Journal of Prognostics and Health Management (ijPHM).

The organizers of the competition reserve the right to both modify these rules and disqualify any team for any practices it deems inconsistent with fair and open practices.

Registration

Teams may register by contacting the Competition organizers (justinian.rosca@siemens.com) with their name(s) and a team alias under which the scores would be posted.

Please note: In the spirit of fair competition, we allow only one account per team. Please do not register multiple times under different user names, under fictitious names, or using anonymous accounts. Competition organizers reserve the right to delete multiple entries from the same person (or team) and/or to disqualify those who are trying to "game" the system or using fictitious identities.

Key Dates

Key Conference Dates	
Competition Open	5 June 2015
Competition Closed	29 Aug 2015
Preliminary Winners Announced	6 Sep 2015
Winners Announced	11 Oct 2015
Winning Papers Due	14 Oct 2015
PHM Conference Dates	**19 Oct - 24 Oct 2015**

Data

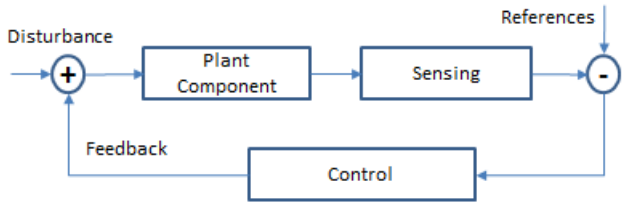
This year's prognostics topic focuses on the operation of a plant and the capability to detect plant failure events in advance. Data given represents: (a) time series of sensor measurements and control reference signals for each of a number of control components of the plant (e.g. 6 components); (b) time series data representing additional measurements of a fixed number of plant zones over the same period of time (e.g. 3 zones), where a zone may cover one or more plant components; (c) plant fault events, each characterized by a start time, an end time, and a failure code. Each plant is specific through its number of components and the number of zones. However each plant logs faults from the same fixed set of faults. Only faults of type 1-5 are of interest, while code 6 represents all other faults not in focus. The frequency of measurements is approximately one sample every 15 minutes, and the time series data spans a period of approximately three to four years. The task is to predict future failure events of types 1-5 and the time of their occurrence from past data.

For example, the data set for *Plant #1* is given by a collection of three [.csv] files: *plant-1a.csv*, *plant-1b.csv*, *plant-1c.csv*. Each of the (a), (b), and (c) files contains information as described above. More precisely the columns of each of the (a), (b), and (c) [.csv] file are:

a. Plant measurements per component: Component number "*m*", time "*t*", sensors "*S1*"-"*S4*", and control references "*R1*"-"*R4*"
b. Additional plant measurements per zone in the plant: zone number "*n*", time "*t*", sensors "*E1*" and "*E2*"
c. Faults: Start time "*t1*", end time "*t2*", and fault code "*F*".

Additionally, the following physical plant model information is provided:

- Each plant component is controlled by a feedback loop system as represented in the figure below; plant components are disjoint
- Each zone measures cumulative energy consumed (*E1*) and instantaneous power (*E2*) in disjoint sections of the plant covering one or more components
- Faults are independent of one another. Also, a fault *F* is independent of data outside of a three hour window of time before the fault start time.



The participants will have access to the following data sets:

1. *Training data* for approximately thirty plants (e.g. see Plant 1 described above).
2. *Test data* for approximately ten plants. For each test plant, the first half of the fault file data is complete, while the second half is incomplete. Some faults of type 1-5 are missing.

Submission and Scoring

Each team is permitted one submission a week on the test data. A correct submission will be represented by a zip archive [.zip] file with failure result files on the test data, whose name is the team alias. If the team alias is "*eagles*", then the filename will be "*eagles.zip*". For each test fault file *plant-#c.csv*, the submission should include a corresponding file named *plant-#c-new.csv* stating predicted missing faults in the test data. Each line of the [.csv] file should indicate a new predicted fault in the same format as the type (c) training data file *plant-#c.csv*. Thus each line should specify the start time of the fault, the end time of the fault, and the fault type. Each submission will receive a score, as described below. Each team will be given the location to submit entries when they register.

The scoring system rewards correct detection (true positives or *tp* below) of a fault within a one hour tolerance of the time it actually occurs for each fault *j*, at any sampling time. It penalizes false positives (*fp;* other faults incorrectly classified as fault *j*) and false negatives (*fn;* fault *j* incorrectly classified as another fault). The overall prediction score of a plant test file is computed with a formula dependent on the table of confusion scores for the first *N=5* fault codes:

$$Score = \frac{1}{N}\sum_{j=1}^{N}(4*tp_j - 10*fp_j - fn_j)$$

The overall prediction score of a submission will sum the scores for all test plants.

The Data Challenge committee will make available a validation set of data consisting of approximately ten plants, to be released before the final submission, based on which the final standing will be decided. The submission with the highest score on the validation data will be the winner of the data challenge competition.

The Data Challenge committee will regularly evaluate submissions on the unlabeled portion of the test data, will report to each team the performance of their intermediate submissions, and will post the intermediate standing.

Notes

The data provided for the challenge is grounded in the real world and it is raw. It has inconsistencies, which should be taken into account, for example:

1. The sampling interval is roughly 15 minutes, however variations may occur due to logging delays. For example, line 1 in *plant_1a.csv* indicates a time stamp of 8/18//2009 18:12, and ideally the hour stamp should be 18:00.
2. Some data may be missing.
3. Data intervals with non-increasing values of *E1* are invalid, and should be disregarded. An example could be seen in *plant_25b.csv* starting in row number 16253.

Questions

Questions about the data challenge will be answered on the challenge forum at *http://www.phmsociety.org/forum/583*