**Diagnosing Wind Turbine Faults Using Machine Learning Techniques Applied to Operational Data**

**Abstract**

* Component failures incur significant downtime 🡪 loss of revenue
* Continuously monitoring turbine health will make it possible to detect incipient faults and schedule maintenance as needed
* CMS (Condition Monitoring Systems)
* By performing complex analysis of existing data from the turbine’s SCADA system, valuable insights into turbine performance can be obtained at a much lower cost.
* Data is obtained from the SCADA system of a turbine in the South-East of Ireland. Fault and alarm data is filtered and analyzed in conjunction with the power curve to identify periods of nominal and fault operation.
* Classification techniques are applied to recognize fault and fault-free operation by taking into account other data from SCADA such as temperature, pitch, and rotor data.
  + Data is extended to allow prediction and diagnosis in advance of specific faults

**Section I – Introduction**

* Wind turbines experience highly irregular loads due to varied and turbulent wind conditions
  + Components undergo high stress throughout their lifetime
* Operations and maintenance account for up to 30% of the cost of generation of wind power.
* Remotely monitoring component health is even more important in the wind industry
  + Wind turbines are deployed to operate autonomously in remote sites
    - Periodic visual inspections can be impractical.
* Condition-based Maintenance (CBM)
  + Strategy whereby the condition of the equipment is actively monitored to detect impending or incipient faults
  + Can save up to 20-25% of maintenance costs vs. scheduled maintenance of wind turbines.
  + Allows for prognostic analysis
    - Where remaining useful life (RUL) of a component is estimated.
* Condition Monitoring Systems (CMS)
  + Typically consists of vibration-sensors in combination with optical strain gauges or oil particle counters
  + Data is then sent to a central data processing platform where it is analyzed and if an incipient fault is detected, an alarm is raised.
* CBM and prognostic technologies have no been taken up extensively by the wind industry
  + Capital cost of retrofitting sensors, data collection/analysis is quite high, about 13K euros per turbine
  + Commercially not as successful
  + False alarms 🡪 very costly due to the downtime and manual inspections needed
* Effort in applying CM techniques to wind turbines by analyzing data collected by the SCADA system.
* SCADA
  + Data recorded in 10 minute intervals to reduce transmitted data bandwidth and storage
  + Data includes:
    - Active and Reactive power
    - Generator current and voltages
    - Anemometer measured wind speed
    - Generator shaft speed
    - Generator, gearbox and nacelle temperature
* Statistical Analyses on various trends within data can be used to detect when the turbine is entering a time of sub-optimal performance or if a fault is developing
* Paper Outline
  + Section II – describe how a wind turbine’s power curve and other SCADA data can be used for fault detection through performance monitoring, and give a brief review of methods used in the past.
  + Section III – Describe the turbine site and data we use
  + Section IV – Describe the model used for detecting, diagnosing and predicting faults, and the results obtained.
  + Section V – Evaluate the performance of our model against previous methods used in the literature
    - Accuracy and effectiveness of predicting faults

**Section II – Review of SCADA Based CM Systems**

* Wind Turbine Failure Modes
  + Failure Mode Effects Analysis (FMEA)
  + Shows frequency of different failure modes on turbine components and sub-assemblies and their contribution to down time
  + Biggest contribution to the overall failure rate was the POWER SYSTEM
    - Translates to just below a 40% contribution to overall downtime on the turbines surveyed
    - Data comes from the EU FP7 ReliaWind Project
* Power Curve
  + Relationship between power and wind speed
    - Shows turbine power output as a function of hub height wind speed
  + Important metric of determining wind turbine performance
  + Performance of a turbine under different wind speeds can be related to three key points on this graph
    - – the cutting speed; min useful wind speed at which the turbine begins to generate power
    - – the rate speed, the speed at which maximu rate generator output is obtained
    - – the cutout speed, the maximum speed at which the turbine can produce power.
  + Comparing the generated power of a turbine at a given wind speed to the supplied power curve is important to checking turbine performance
  + Obtain data when turbine is in fault-free operation
* Review of SCADA-based CM systems
  + Modeling power curve
  + Modeling performance monitoring
    - Using wind speed trended against power output, rotor speed and blade pitch angle
    - Good performance metric for turbine
* Wider spectrum of SCADA parameters 🡪 better fault classification and limited fault prediction
  + NN, boosting trees, SVM and standard classification regression trees

**Section III – Data**

* Data from study comes from 3 MW direct-drive turbine
  + Supplies power to a major biomedical device manufacturing plant
* Three separate datasets taken from turbine SCADA system
  + Operational
  + Status
  + Warning
* Data covers an 11 month period
* Operational Data
  + Control system monitors the following
    - Instantaneous parameters: wind speed, ambient temperature
    - Power characteristics: real and reactive power and various currents and voltages in the electrical equipment
    - Temperature of components: generator bearing and rotor
  + Avg Min and Max of these values taken over 10 min period
  + Data was used to train the classifiers and labelled according to various filters
* Status Data
  + Number of normal operating states for the turbine
  + Number of statuses for when the turbine is in abnormal or faulty operation
  + Split into two data sets
    - WEC status data
      * Corresponds to status messages directly related to the turbine itself
    - RTU status data
      * Corresponds to power control data at the point of connection to the grid
      * Deals exclusively with active or reactive power set-points
* Warning Data
  + Corresponds to general information about the turbine
  + Sometimes correspond to a potentially developing fault on the turbine
  + Warning messages can be mostly ignored in this analysis
* Data Labeling
  + Three levels of classification
    - Fault/no fault: classified as faulty or fault-free
    - Fault Diagnosis: classified as a specific fault or fault-free
    - Fault Prediction: Attempt to classify data as a specific fault up to one hour in advance of the fault occurring
    - Achieved by splitting the initial operational data into labelled “no-fault”, “all-faults”, “specific faults” and “fault prediction” datasets
    - No Fault Dataset
      * Needed consisting of nominal fault-free operation
      * Three different filters applied to the full set of 10-min operational data
        + First, WEC status codes corresponding to nominal operation were selected
        + These are

0:0 Turbine in Operation

2:1 Wind Speed too Low

3:12 Storm Wind Speed

* + - * + Operational data with timestamps that corresponded to at least 30 minutes AFTER these statuses came into effect and 120 minutes BEFORE they changed were chosen
      * Time bands were found empirically and eliminate any transients that may arise from going from fault-free to faulty operation or vice-versa
      * Operational data corresponding to RTU statuses where power output was being curtailed were filtered out
      * Times corresponding to a single specific warning message were filtered out
        + To give a clearer distinction for fault classification
      * To verify that only data from when the turbine was in nominal operation was included in the “no-fault” dataset, the power curve of the filtered data was plotted to check that it conformed to the nominal shape as seen in Fig 2a
      * Algorithm developed to filter out power curve anomalies used to highlight the data points which were outside the estimated bounds of nominal operation
    - All Faults Dataset
      * List of frequently occurring faults was made
      * Status messages with codes corresponding to the faults were selected
      * Time band of 600s before the start and after the end of these turbine states were used to match up the associated 10-minute operational data.
      * Fault frequency refers to the specific instances of each fault rather than the number of data points of operational data associated with it
      * Feeding faults refer to faults in the power feeder cables of the turbine
      * Excitation error refers to problems with generator excitation system
      * Mains failure refers to problems with mains electricity supply to the turbine
      * Malfunction air cooling refers to problems win the turbine
      * Generator heating faults refer to the generator overheating
    - Specific Faul Datasets
      * Similar to all faults
    - Fault Prediction Datasets
      * Time band around which faults were classified was extended by varying degrees
        + 10, 20, 30, 60, 120, 360 minutes before a specific fault
        + This means that operational data points leading up to a specific fault were also included in that fault class

**Section IV – Methodology**

* SVM
  + Trained using scikit-Learn LibSVM implementation
  + Randomly shuffled, split into training-testing 80-20%
  + Only a subset of 30 specific features were chosen to be included for training purposes
  + A number of original features corresponded to sensors on the turbine which were broken (frozen or incorrect values)
  + The average and standard deviation of the 12 inverter temperatures
  + 29 features being used to train the SVMs, scaled individually to unit norm
  + Randomized grid search was then performed over a number of hyperparameters used to train each SVM to find the ones which yielded the best results
    - Verified using 10-fold cross validation
  + Scoring metric used for cross validation was a mean of the weighted precision and recall
  + Hyperparameters searched over were *C*, gamma, and the kernel used
    - Three kernels
      * Simple linear
      * Radial-basis
      * Polynomial
  + Data was heavily imbalanced, there were on the other of 10^2 more no-fault samples than fault-class samples
    - Two approaches to mitigating this effect
      * A class weight, *c.w.*, added to the minority class when calculating *C* for that class.
      * A number of different class weights were added to the set of hyperparameters being searched over for this approach

**A black text with black text

Description automatically generated**

* + - * Second approach:
        + Selected a balanced set of data to train on
        + After splitting training and testing data, training was further split to include the same number of fault-free instances as fault instances
        + Test data was not altered in any way so as to preserve the imbalance seen in the real world
  + Scoring metrics to evaluate final performance on the test set
    - Specificity
    - Precision
    - Recall Score
    - F1 Score
  + Overall accuracy was not used as a metric due to the massive imbalance in the data sets
  + Calculating Equations

A white background with black text

Description automatically generated

**Section V – Results & Discussion**

* Hyperparameter search for the case of a balanced training set
* Results of performing this search on the full imbalanced training set, including different values of *c.w.*

A table with numbers and text

Description automatically generated

A table with numbers and lines

Description automatically generated