2016 23rd International Conference on Pattern Recognition (ICPR)
Cancún Center, Cancún, México, December 4-8, 2016

Wind turbine fault prediction using soft label SVM

Rui Zhao
ECSE Dept., RPI
Troy, NY, USA
zhaor@rpi.edu

Md Ridwan Al Iqbal
Computer Science Dept., RPI
Troy, NY, USA
iqbalm@rpi.edu

Kristin P. Bennett
Mathematical Science Dept., RPI
Troy, NY, USA
bennek@rpi.edu

Qiang Ji
ECSE Dept., RPI
Troy, NY, USA
jiq@rpi.edu

***Abstract*—In this paper, we address the problem of predicting wind turbine electrical subsystem fault using time series data obtained from multiple sensors on wind turbine. While considering this as a time series classification problem, we are facing with the challenge that there is no explicit label information regarding the temporal location and duration of symptoms of the fault. Besides, significant data variation caused by both external and internal factors make the identification of change point non-trivial. To address these challenges, we propose a soft label SVM method where the probability of fault instead of binary label is used to train classifier to handle the uncertainty in label information. The probability is determined using temporal information of fault instances. We consider this as a weakly supervised learning problem. To handle large variation within data, we perform customized normalization on different sensor data based on their physical meanings and relationships. Finally, we evaluate our method on 38 different forced outage instances. The experiment on real SCADA data obtained from wind turbines show promising results where we can predict the triggering of fault 18 hours beforehand with an average AUC value 0.91.**

I. Introduction

Wind energy as a major renewable natural resource has been harvested for electricity since early 1900s. As the increasing need for clean energy and the advancement of power generation equipment in particular, wind turbine, there is a fast growth of wind energy consumption since early 2000s. At the end of 2014, the global total electricity produced by wind energy is 369.6 GW. The average production growth over the last 10 years (2005-2014) is about 23%. The projection of global market consumption need by 2019 is 666.1 GW with average growth rate at about 12.7% per year [1].

The increasing needs of wind energy result a fast growth in installation of wind turbines, whose reliability become a critical issue for the long term profit. There are several factors contributing to the high maintenance and repair cost of wind turbines. First of all, modern wind turbine by itself is a complex system which consists of different components such as pitch system, gearbox, generator, yaw system, tower, control system, *etc.* There are different failure modes caused by different subsystems with different possible root causes [2]. Secondly, the location of wind farm is often remote to residential area and even offshore due to the requirement of wind condition. Should failure happened, a turbine must be shut down and specialized engineers need to be dispatched to perform diagnosis. Additional time will be needed for component replacement, resulting economic loss.

To increase the reliability of power generation and make quick response to turbine fault, condition monitoring system (CMS) [3] is usually installed on the turbine. Real-time sensor readings and other operation records are collected and stored in central database through supervisory control and data acquisition (SCADA) system. Data collection starts when the turbine is deployed to work. The typical sampling rate is at the granularity of seconds or sub-seconds. A smoothed stream of data is obtained by averaging values over a period (*e.g.* 10 minutes) and stored in central database. Massive amount of operation data provides a rich resource to analyze the dynamic behavior and operating condition of wind turbine. This is where the machine learning techniques can be of potentially significant help.

In particular, the prognostics of fault based on historical operation data will give more time for the monitoring personnel to schedule maintenance and avoid catastrophic component failure which cause irreversible damage to the turbine and significant economic loss. Nevertheless, the scenario of this application brings a couple of challenges. First, the timecourse of fault development is unknown. In other words, the onset and duration of the signature of a fault is unknown. Second, the environment condition is constantly changing, resulting a non-stationary sensor signals. Third, the fault is usually specified to the level of subsystem of wind turbine. However, multiple root causes with distinct signatures may exist.

In this paper, we are trying to address these challenges by formalizing a time series classification problem with uncertain labels. A review of related work is provided in section II. Then we formally define the problem and describe proposed soft label SVM model in section III and IV respectively. Data processing and experimental evaluation are discussed in section V. Finally, we make conclusion and point out future work in section VI.

II. Related work

In a thorough review on condition monitoring system, Hameed *et al.* [4] divided methods of fault detection and prediction into different categories depending on their focus during analysis process. The major categories include physical model based approach, signal analysis based approach and artificial intelligence based approach. Our approach falls into the third category with a combination of signal analysis method for feature extraction purpose. One popular way of condition monitoring uses SCADA data collected by wind

978-1-5090-4846-5/16/$31.00 ©2016 IEEE

turbine controller due to its cost-effectiveness [2]. We also use SCADA data as our resource for fault prediction. There is a line of work on fault detection using SCADA data [5], [6], [7], [8]. Wang *et al.* [9] summarized different machine learning algorithms for SCADA data analysis on different types of fault. However, most of these works are conducted in a fully supervised fashion where empirical labels differentiating normal and abnormal is utilized during the training process. In our work, we do not assume the availability of such labels on data. Instead, the probabilities associated with different classes are used in our approach, which exploit the uncertainty in the data.

Alternatively, identifying abnormality from data can also be approached by an unsupervised fashion. A closely related problem is time series segmentation whose goal is to divide time series into disjoint statistically consistent parts. Time series segmentation has been studied in a variety of disciplines including speech signal segmentation [10], context recognition [11], change point detection [12] and subsequence clustering [13]. A typical approach is to construct a cost function for the segmentation on original or some high level representation of the time series. Then the segmentation algorithm determines the optimal start and end point for each segment by minimizing the cost function.

More recently, Hoai and De la Torre [14] extended the maximum margin clustering framework to temporal data for joint segmentation and separation of different temporal clusters in time series. However, segmentation is not sufficient for our problem. First of all, obtaining a meaningful segmentation with multivariate time series is not trivial. More importantly, direct segmentation ignores the temporal order of the data points in time series where it is more likely to observe abnormality as time approaches to the fault event. Our approach incorporates the uncertainty of label information into the learning process yet still can take advantage of the discriminative capability of max-margin framework which is usually constructed in a fully supervised setting. Therefore, our approach can be considered as a weakly-supervised learning method.

III. Problem Statement

A. Definition and goal

We now formally define the time series classification problem as follows. Suppose we have a multi-dimensional time series $\mathbf{x}^L = \{x_0, ..., x_{L-1}\},\ x_i \in \mathbb{R}^d,\ \forall i = 0, ..., L-1$, where d is the dimension of each sample and L is the total number of samples. Let $\mathbf{x}_n^l = \{x_n, ..., x_{n+l-1}\}$ $(0 \leq n \leq L-l,\ 1 \leq l \leq L)$ denote a subsequence consists of l consecutive samples with the first sample be x_n. Let y_n^l be an integer variable indicating the label of subsequence $\mathbf{x}_n^l$. In the application of wind turbine fault prediction, we define three operation status of turbine. The first status is called normal when the turbine's operation satisfies its design specification with minimum risk of fault. The second status is called pre-fault when the turbine's operation satisfies design specification with developing fault which eventually trip the turbine. Noticing that under the first two status, turbines are operating. The third status is called forced outage where turbine is tripped due to fault and loses power generation. When turbine is tripped, everything stops working. Thus, the data falls into this category will not be considered for prognosis purpose. Let $y_n^l = -1$ corresponds to the normal status and $y_n^l = 1$ corresponds to the pre-fault status. The goal is to learn a mapping function $f : \mathbb{R}^d \to \mathbb{R}$ with $y_n^l = \text{sign}(f(\mathbf{x}_n^l))$ for each length l subsequence in $\mathbf{x}^L$. Figure 1 illustrates a stream of data with shaded color indicating the developing process of a fault.

B. Assumptions

We assume that the finite time series $\mathbf{x}^L$ does not contain any samples with forced outage status. In addition, the forced outage status proceeds immediately after the last entry x_{L-1} in $\mathbf{x}^L$. In other words, $\mathbf{x}^L$ covers the temporal scope prior to a forced outage event. During training, we do not have the labels for each subsequence $\mathbf{x}_n^l$. We assume that the probability of being in pre-fault status is non-decreasing as index of sample gets larger i.e.

$$P(y_n^l = 1|\mathbf{x}_n^l) \geq P(y_m^l = 1|\mathbf{x}_m^l),\ \forall n > m \tag{1}$$

The motivation of this assumption is that as time gets closer to a forced outage event, it is more likely for the turbine operating with some developing fault. We explore a few different probability assignment functions and compare their performance in the experiment. To evaluate the performance of proposed model, we assume the label of testing time series $\mathbf{x}^L$ using the following empirical criteria.

$$y_n^l = \begin{cases} -1, & \text{if } n < n^- \\ 1, & \text{if } n > n^+ \end{cases} \tag{2}$$

where $0 < n^- < n^+ < L - l$. In our experiment, n^-, n^+, L, l are pre-defined constants. For the rest of this paper, we omit L, l in the superscript when the meaning of notation is clear.

IV. Methods

A. Support vector machines

Support vector machines (SVM) [15] has been widely used in pattern recognition problems [16]. In particular, for classification problem, SVM maps data into a high dimensional feature space and seeks for a hyperplane in feature space which separates data and maximizes the distance between decision boundary to the closest data points. Cortes and Vapnik [17] introduced soft margin idea which allows for mislabeled sample. We briefly review the formulation of soft margin SVM and introduce our extension in the next section.

Given a training set of N data points $\{y_n, \mathbf{x}_n\}_{n=1}^N$ where $\mathbf{x}_n \in \mathbb{R}^D$ is n^{th} sample and $y_n \in \mathbb{Z}$ is the associated label. SVM is aimed at constructing a classifier of the form

$$y_n \equiv y(\mathbf{x}_n) = \text{sign}(\mathbf{w}^T \phi(\mathbf{x}_n) + b) \tag{3}$$

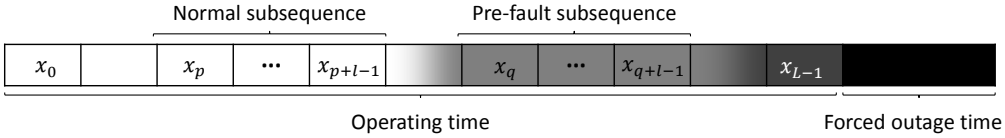

Fig. 1. Illustration of a time series covering different wind turbine status. Example of normal subsequence and pre-fault subsequence are provided. The shade becomes darker as time gets closer to forced outage event, indicating an increasing severity of fault development.

where $\mathbf{w}, b$ are the model parameters and ϕ is the feature mapping function with proper dimension. The model parameters are learned by solving the following primal problem.

$$\begin{aligned} \min_{\mathbf{w},b} \ & \frac{1}{2}||\mathbf{w}||^2 + \gamma \sum_{n=1}^{N} \xi_n \\ \text{s.t. } & y_n(\mathbf{w}^T\phi(\mathbf{x}_n) + b) \geq 1 - \xi_n, \\ & \xi_n \geq 0, \ n = 1, ...N \end{aligned} \tag{4}$$

where γ is a parameter controls the trade-off between slack variables ξ_n and the margin, which is inverse proportional to the norm of $\mathbf{w}$.

By introducing Lagrangian multipliers α_n, we obtain the following Lagrangian dual problem

$$\begin{aligned} \max_{\alpha_n} & \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{m=1}^{N} \sum_{n=1}^{N} \alpha_m \alpha_n y_m y_n k(\mathbf{x}_m, \mathbf{x}_n) \\ \text{s.t. } & \sum_{n=1}^{N} \alpha_n y_n = 0, \\ & 0 \leq \alpha_n \leq \gamma, \ n = 1, ...N \end{aligned} \tag{5}$$

where $k(\mathbf{x}_m, \mathbf{x}_n) = \phi(\mathbf{x}_m)^T\phi(\mathbf{x}_n)$ is the kernel function. The use of kernel function can avoid explicit computation of $\phi(\mathbf{x}_n)$. The objective function in dual problem is convex to α_n and can be solved using quadratic programming. Then the model parameters $\mathbf{w}, b$ can be recovered from the optimality condition $\mathbf{w} = \sum_{n=1}^{N} \alpha_n y_n \phi(\mathbf{x}_n)$ and the fact that $y_n(\mathbf{w}^T\phi(\mathbf{x}_n) + b) = 1, \ \forall n \in \mathcal{S}$, where $\mathcal{S}$ is the set of support vectors. During testing, given a new sample $\mathbf{x}$, we obtain the class label as follows.

$$y(\mathbf{x}) = \text{sign}(\sum_{n=1}^{N} \alpha_n y_n k(\mathbf{x}_n, \mathbf{x}) + b) \tag{6}$$

B. Soft label SVM

In standard SVM, the learning of classifier requires the label y_n for each sample $\mathbf{x}_n$. To handle the situation where only the probability of label is available, we propose the soft label SVM. To be specific, given a training set $\{u_n^+, \ u_n^-, \mathbf{x}_n\}_{n=1}^N$, where

$$\begin{aligned} P(y_n = 1|\mathbf{x}_n) &= u_n^+ \\ P(y_n = -1|\mathbf{x}_n) &= u_n^- \end{aligned}$$

We consider binary labels only, thus $u_n^+ + u_n^- = 1$. The goal of learning is the same as standard SVM where we construct a classifier in the form of Eq.(3). Now the primal problem can be written as

$$\begin{aligned} \min_{\mathbf{w}} & \frac{1}{2}||\mathbf{w}||^2 + \gamma \sum_{n=1}^{N} (u_n^+ \xi_n^+ + u_n^- \xi_n^-) \\ \text{s.t. } & \mathbf{w}^T\phi(\mathbf{x}_n) + b \geq 1 - \xi_n^+, \\ & -\mathbf{w}^T\phi(\mathbf{x}_n) - b \geq 1 - \xi_n^-, \\ & \xi_n^+, \xi_n^- \geq 0, \ n = 1, ...N \end{aligned} \tag{7}$$

where ξ_n^+, ξ_n^- are slack variables. Essentially, we introduce two slack variables for each sample to characterize a weighted soft margin. In the extreme case where either $u_n^+ = 1$ or $u_n^- = 1$. The summation involves the n^{th} sample in the objective function reduces to only one term. Eq.(7) reduces to Eq.(4) *i.e.* the standard SVM with soft margin.

We now derive the dual problem of Eq.(7). Introducing Lagrangian multipliers $\alpha_n^+, \alpha_n^-, \mu_n^+, \mu_n^-$ associated with constraints on $\mathbf{x}_n, \xi_n^+, \xi_n^-$. We have the following Lagrangian function.

$$\begin{aligned} \mathcal{L}(\mathbf{w}, b, \alpha) = & \frac{1}{2}||\mathbf{w}||^2 + \gamma \sum_{n=1}^{N} (u_n^+ \xi_n^+ + u_n^- \xi_n^-) \\ & - \sum_{n=1}^{N} \alpha_n^+ (f(\mathbf{x}_n) - 1 + \xi_n^+) - \sum_{n=1}^{N} \mu_n^+ \xi_n^+ \\ & - \sum_{n=1}^{N} \alpha_n^- (-f(\mathbf{x}_n) - 1 + \xi_n^-) - \sum_{n=1}^{N} \mu_n^- \xi_n^- \end{aligned} \tag{8}$$

where $f(\mathbf{x}_n) = \mathbf{w}^T\phi(\mathbf{x}_n) + b$. The optimal solution satisfies the Karush-Kuhn-Tucker (KKT) condition [18].

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mathbf{w}} &= \mathbf{w} - \sum_{n=1}^{N} (\alpha_n^+ - \alpha_n^-)\phi(\mathbf{x}_n) = 0 \\ \frac{\partial \mathcal{L}}{\partial b} &= -\sum_{n=1}^{N} (\alpha_n^+ - \alpha_n^-) = 0 \\ \frac{\partial \mathcal{L}}{\partial \xi_n^+} &= \gamma u_n^+ - \alpha_n^+ - \mu_n^+ = 0 \\ \frac{\partial \mathcal{L}}{\partial \xi_n^-} &= \gamma u_n^- - \alpha_n^- - \mu_n^- = 0 \\ \alpha_n^+(f(\mathbf{x}_n) - 1 + \xi_n^+) &= \alpha_n^-(-f(\mathbf{x}_n) - 1 + \xi_n^-) = 0 \\ \mu_n^+ \xi_n^+ &= \mu_n^- \xi_n^- = 0 \\ f(\mathbf{x}_n) - 1 + \xi_n^+ \geq 0, \ & -f(\mathbf{x}_n) - 1 + \xi_n^- \geq 0 \\ \xi_n^+, \xi_n^-, \alpha_n^+, \alpha_n^-, \mu_n^+, \mu_n^- &\geq 0 \\ n = 1, ...N. & \end{aligned} \tag{9}$$

Using the equalities and inequalities specified by Eq.9), we have the following Lagrangian dual problem

$$\max_{\alpha_n^+,\alpha_n^-} \mathcal{Q}(\alpha_n^+,\alpha_n^-) = \sum_{n=1}^{N}(\alpha_n^+ + \alpha_n^-) \quad (10)$$
$$-\frac{1}{2}\sum_{m=1}^{N}\sum_{n=1}^{N}(\alpha_m^+ - \alpha_m^-)(\alpha_n^+ - \alpha_n^-)k(\mathbf{x}_m,\mathbf{x}_n)$$
$$\text{s.t. } 0 \le \alpha_n^+ \le \gamma u_n^+,\ 0 \le \alpha_n^- \le \gamma u_n^-$$
$$\sum_{n=1}^{N}(\alpha_n^+ - \alpha_n^-) = 0,\ n = 1,...N$$

where $k(\mathbf{x}_m,\mathbf{x}_n) = \phi(\mathbf{x}_m)^T\phi(\mathbf{x}_n)$ is the kernel function.

$\mathcal{Q}$ is a quadratic function of α_n^+, α_n^-. We can solve Eq.(10) using quadratic programming. Then the parameter $\mathbf{w}$ can be computed using Eq.(9). To compute b, from KKT condition we know when $0 < \alpha_n^+ < \gamma u_n^+$ and $\alpha_n^- = \gamma u_n^-$, $f(\mathbf{x}_n) = 1$. Define $\Omega^+ = \{\mathbf{x}_n : f(\mathbf{x}_n) = 1\}$. Similarly, when $0 < \alpha_n^- < \gamma u_n^-$ and $\alpha_n^+ = \gamma u_n^+$, $f(\mathbf{x}_n) = -1$. Define $\Omega^- = \{\mathbf{x}_n : f(\mathbf{x}_n) = -1\}$. Though any point in set Ω^+ or Ω^- can be used to compute b. We compute a numerically more stable estimation as follows.

$$b = \Big[\sum_{n\in\Omega^+}(1 - \sum_{m=1}^{N}(\alpha_m^+ - \alpha_m^-)k(\mathbf{x}_m,\mathbf{x}_n)) \quad (11)$$
$$-\sum_{n\in\Omega^-}(1 + \sum_{m=1}^{N}(\alpha_m^+ - \alpha_m^-)k(\mathbf{x}_m,\mathbf{x}_n))\Big]/(|\Omega^+| + |\Omega^-|)$$

where $|\mathcal{S}|$ is the cardinality of set $\mathcal{S}$. After learning the model, the prediction value for new point $\mathbf{x}$ is given by

$$f(\mathbf{x}) = \mathbf{w}^T\phi(\mathbf{x}) + b = \sum_{n=1}^{N}(\alpha_n^+ - \alpha_n^-)k(\mathbf{x},\mathbf{x}_n) + b \quad (12)$$

Again, we use the kernel trick here to avoid explicit computation of feature mapping. For classification purpose. The label is obtained by $y = \text{sign}(f(\mathbf{x}))$.

C. Generalized formulation of soft label SVM

The objective function in primal problem defined by Eq.(7) contains an L2-norm regularization term on model parameter $\mathbf{w}$ and hinge loss function of slack variables $\xi_n^+, \xi_n^-, n = 1, ..., N$. We can write the primal problem in a more general form as follows.

$$\min_{\mathbf{w}} \frac{1}{2}||\mathbf{w}||^p + \gamma\sum_{c=1}^{2}\sum_{n=1}^{N} u_{nc}E[f(\mathbf{x}_n), y_c] \quad (13)$$

where $p > 0$ is the order of regularization and $u_{nc} = P(y_n = y_c|\mathbf{x}_n), y_c = (-1)^c$ is the soft label for each sample. E is a loss function which can take different forms. In the experiment, we try three different loss functions:

- Hinge loss: $E[f(\mathbf{x}_n), y_c] = \max(0, 1 - y_c f(\mathbf{x}_n))$.
- Squared hinge loss: $E[f(\mathbf{x}_n), y_c] = \max(0, 1 - y_c f(\mathbf{x}_n))^2$.
- Squared loss: $E[f(\mathbf{x}_n), y_c] = (1 - y_c f(\mathbf{x}_n))^2$.

When regularization order $p = 2$, the solution for the 2nd and 3rd loss functions can be obtained in a similar way as the 1st one described in section IV-B. When $p = 1$, the dual problem to Eq.(13) is no longer quadratic to the dual variables. We adopt ADMM [19] framework to solve Eq.(13) directly. We omit the derivation due to page limit.

V. Experiment

A. Data description

The SCADA data we used are collected from 125 1.6MW wind turbines at one wind farm over the time period from June 2013 to May 2014. We identified 38 forced outage events with faults related to electrical subsystem. The total accumulate hours of down time is 2305 hours. The maximum and minimum down time are 544.7 hours and 1.7 hours respectively. There are 55 sensor values collected via SCADA measuring different quantities such as power generation, tower vibration, temperature, wind speed, blade angle, *etc.* The sensors operate with a sampling rate at sub-second level. Then an averaged value over 10 minutes period is collected and stored via SCADA system, resulting 144 samples per day. We select 12 days of data prior to the beginning of each forced outage event, resulting a dataset with 456 days of turbine operating time.

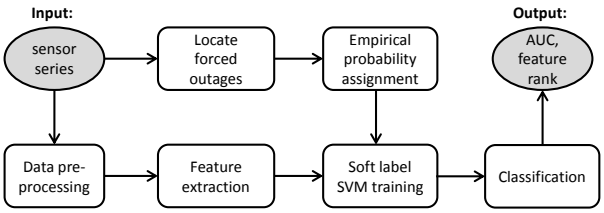

Fig. 2. An overview of the experiment process.

B. Data processing

1) Pre-processing: The goal of pre-processing is to clean data used for analysis. The sensor recordings may contain duplicate or incomplete values. We exclude channels that are mostly constant or only contain a few (< 10) distinct values. We also exclude channels with substantial portion ($> 50\%$) of incomplete values. Resulting 39 channels belonging to four different categories, where we use the same taxonomy proposed by [20], 1) wind parameters, 2) energy conversion parameters, 3) vibration parameters and 4) temperature parameters. The resulting dataset contains 39 multivariate time series with total number of samples about 2.56×10^6.

2) Feature extraction: As we described in section III-A, we classify each fixed length subsequence from a complete sequence. We choose the fixed length to be 6 hours. Then each subsequence contains 36 samples, each sample is 39 dimensional vector. One challenge of this problem is the within class variation caused by both external factors such as environment change and internal factors such as component degradation. We are only interested at the pattern that results from developing fault which is usually an internal factor. To suppress the variation caused by irrelevant factors, we perform a customized normalization on some selected channels

based on their physical meanings and relationship with other channels. Specifically, we subtract ambient temperature from all the temperature sensors. We divide blade angle by wind speed and divide generator speed by rotor speed. For tri-phase voltage and current, we subtract mean value from each phase of voltage and current respectively.

After normalization, we extract two types of feature to further capture the spatial and temporal dependency among different channels. For spatial dependencies, we compute the sample covariance of the 39 dimensional vectors over $N = 36$ samples as $\Sigma = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})^T$, where $\bar{\mathbf{x}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n$. Then we choose the upper triangle entries due to symmetry of Σ, resulting 780 dimensional vector. For temporal dependencies, we fit a simple autoregressive model [21] for each channel defined by following regression equation.

$$\mathbf{x}_{n+1} = A\mathbf{x}_n + \epsilon \tag{14}$$

where ϵ is error term and A is a diagonal matrix whose diagonal entries are first order autoregressive coefficients for each channel. We compute A using the least square estimation as follows, $A^* = \arg\min_{A \text{ diagonal}} \sum_{n=1}^{N-1} ||\mathbf{x}_{n+1} - A\mathbf{x}_n||^2$. This results additional 39 dimensional feature vector. An extension to this is using full matrix A or higher order autoregressive model. Finally, we concatenate the original signals, normalized signals, spatial and temporal feature vectors into one vector as the representation of each subsequence. The dimension of resulting feature vector is 3123.

C. Experiment setting

Each complete sequence contains 1728 multivariate samples (12 days prior to forced outage). We use sliding window to obtain subsequences. Based on the assumption we introduced in section III-B, we empirically choose a couple of different probability assignment functions $P(y_n|\mathbf{x}_n)$. We decide the split of training and testing data of subsequences as follows. For training data, we select a temporal region covering both high and low probability of being at pre-fault status. For testing data, we select two separate temporal regions. The one closer to forced outage is labeled as pre-fault. The one further away from forced outage is labeled as normal. Figure 3 illustrates different probability assignment functions as well as division of training and testing data.

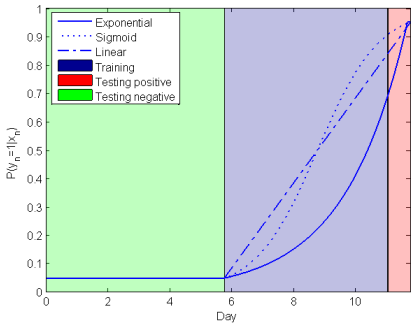

Fig. 3. Division of time series data into training and testing set based on their temporal location. Three different probability assignment functions which satisfy the non-decreasing assumption described in section III-B are plotted.

We fix the normal region of testing data to be 6th-12th day prior to forced outage while vary the pre-fault region of testing data which we call prediction horizon. The training data region is between prediction horizon and 6th day prior to forced outage. We also vary the choice of sliding window length. Due to the multiple possible root causes of the fault, the signatures of different outages may be different. Therefore, we use data from the same forced outage event for training and testing. We use linear feature mapping so $\phi(\mathbf{x}) = \mathbf{x}$. The overall experiment flow is summarized by Figure 2.

D. Classification results

Since we perform classification for each outage event, we use average area under curve (AUC) as evaluation metric, which is computed from ROC of each classification result. In the first experiment, we vary the sliding window length from 3h-12h and prediction horizon from 12h-48h. The proposed soft label SVM with squared hinge loss is used for classification. The results is shown in Figure 4. The best average AUC is achieved when window length is 6h and prediction horizon is 18h. The prediction ability gradually degrades as window length increases or prediction horizon increases.

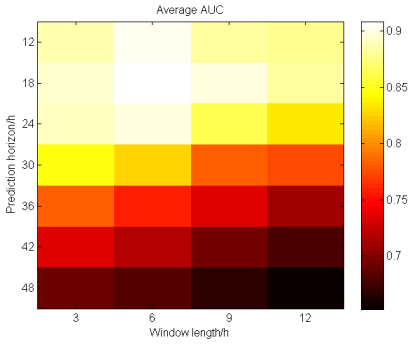

Fig. 4. Average AUC under different sliding window size and prediction horizon with value indicated by color.

For the second experiment, we fix window length at 6h and prediction horizon at 18h. We compare the performance of proposed soft label SVM under different loss functions and regularization. For baseline model, we use standard SVM with training data equally split into positive and negative class. The positive half is the one closer to the outage event. The results are listed in Table I. For all three probability assignment function considered, the proposed soft label SVM outperforms standard SVM whose results are listed in the column 'Hard label'. The best results is obtained under exponential case with average improvement over different objective functions be 12.51%. As shown in Table I, we also compare our method with another popular classification method namely K nearest neighbor (K-NN). We use hard label for training data and the best average result is reported when varying K from 1 to 100 with stride of 5.

E. Feature selection results

The L1-norm regularized objective function used in previous experiment can be also used for feature selection due to the

TABLE I
CLASSIFICATION RESULTS OF DIFFERENT METHODS. H: HINGE LOSS, SH: SQUARED HINGE LOSS, S: SQUARED LOSS, p: REGULARIZATION ORDER.

Method		AUC mean ± std.			
Ours		Hard label	Linear	Sigmoid	Exponential
Loss	p				
H	2	0.757±0.194	0.790±0.184	0.785±0.186	0.873±0.138
SH	2	0.757±0.194	0.904±0.121	0.898±0.135	**0.909±0.135**
S	2	0.766±0.186	0.899±0.126	0.894±0.139	0.905±0.137
SH	1	0.788±0.177	0.879±0.130	0.876±0.148	0.882±0.157
SH-PCA	2	0.758±0.203	0.846±0.177	0.842±0.178	0.879±0.160
K-NN		0.728±0.220			

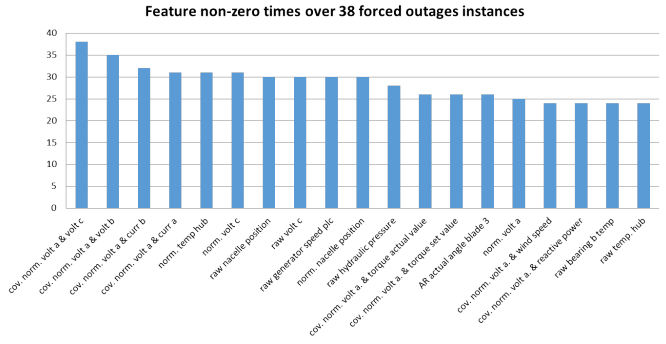

Fig. 5. Top-ranked feature based on the number of counts when its corresponding coefficient learned by L1-norm regularized soft label SVM being non-zero among 38 outages.

sparse solution on parameter $\mathbf{w}$. For each one of the 3123 features, we count the number of times each feature has non-zero coefficient. Figure 5 shows the top-ranked features by such counts. As we can see, the top-ranked features are related to electrical subsystem, which indicates current and voltage sensors are more informative in terms of predicting the electrical subsystem fault. As a comparison, we also perform PCA on original features retaining 95% energy and use squared hinge loss with L2-norm regularization. For both cases, the proposed soft label SVM achieves better results than the hard label counterpart.

VI. CONCLUSION

In this work, we addressed the wind turbine forced outage prediction problem under the challenge of unknown labels, which indicates whether the turbine is running normally as designed or not. We extended standard SVM to handle probability of label as input, which we called soft label SVM. The classifier is learned using multivariate SCADA data and the classification result is used as the prediction of the normality condition of turbine. Given 6-hour window of data, we can predict the happening of one type of electrical subsystem fault 18 hours ahead of time. The average AUC over 38 forced outage instances obtained by this prediction is 0.909 with standard deviation 0.135. The generalization across different outages remain challenging partially due to different root causes of the fault and significant data variation caused by external factors such as weather condition. By learning a L1-norm regularized version of soft label SVM. We identified a set of features that is related to the electrical subsystem fault. One limitation of current approach is the manually chosen probability values. To relax such assumption, we can either incorporate probability learning in the model or exploit only ordinal labels. We also plan to generalize across different forced outage instances in the future.

ACKNOWLEDGMENT

This work was supported in part by a contract from General Electric on wind turbine fault diagnosis.

REFERENCES

[1] G. W. E. Council, "Global wind report 2014," *Brussels, Belgium*, 2015.
[2] K. Kim, G. Parthasarathy, O. Uluyol, W. Foslien, S. Sheng, and P. Fleming, "Use of scada data for failure detection in wind turbines," in *International Conference on Energy Sustainability*. ASME, 2011, pp. 2071–2079.
[3] R. Hyers, J. McGowan, K. Sullivan, J. Manwell, and B. Syrett, "Condition monitoring and prognosis of utility scale wind turbines," *Energy Materials*, vol. 1, no. 3, pp. 187–203, 2006.
[4] Z. Hameed, Y. Hong, Y. Cho, S. Ahn, and C. Song, "Condition monitoring and fault detection of wind turbines and related algorithms: A review," *Renewable and Sustainable energy reviews*, vol. 13, no. 1, pp. 1–39, 2009.
[5] M. Schlechtingen, I. F. Santos, and S. Achiche, "Wind turbine condition monitoring based on scada data using normal behavior models. part 1: System description," *Applied Soft Computing*, vol. 13, no. 1, pp. 259–270, 2013.
[6] A. Zaher, S. McArthur, D. Infield, and Y. Patel, "Online wind turbine fault detection through automated scada data analysis," *Wind Energy*, vol. 12, no. 6, pp. 574–593, 2009.
[7] J. L. Godwin and P. Matthews, "Classification and detection of wind turbine pitch faults through scada data analysis," *IJPHM Special Issue on Wind Turbine PHM*, p. 90, 2013.
[8] O. Uluyol and G. Parthasarathy, "Multi-turbine associative model for wind turbine performance monitoring," in *Proceedings of the Annual Conference of the Prognostics and Health Management Society*, 2012.
[9] K.-S. Wang, V. S. Sharma, and Z.-Y. Zhang, "Scada data based condition monitoring of wind turbines," *Advances in Manufacturing*, vol. 2, no. 1, pp. 61–69, 2014.
[10] T. Svendsen and F. K. Soong, "On the automatic segmentation of speech signals," in *ICASSP*, vol. 12. IEEE, 1987, pp. 77–80.
[11] J. Himberg, K. Korpiaho, H. Mannila, J. Tikanmaki, and H. T. Toivonen, "Time series segmentation for context recognition in mobile devices," in *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*. IEEE, 2001, pp. 203–210.
[12] S. Liu, M. Yamada, N. Collier, and M. Sugiyama, "Change-point detection in time-series data by relative density-ratio estimation," *Neural Networks*, vol. 43, pp. 72–83, 2013.
[13] E. Keogh and J. Lin, "Clustering of time-series subsequences is meaningless: implications for previous and future research," *Knowledge and information systems*, vol. 8, no. 2, pp. 154–177, 2005.
[14] M. Hoai and F. De la Torre, "Maximum margin temporal clustering," in *Proceedings of international conference on artificial intelligence and statistics*, 2012, pp. 1–9.
[15] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*. ACM, 1992, pp. 144–152.
[16] C. M. Bishop *et al.*, *Pattern recognition and machine learning*. springer New York, 2006, vol. 4, no. 4.
[17] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.
[18] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.
[19] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.
[20] A. Kusiak and W. Li, "The prediction and diagnosis of wind turbine faults," *Renewable Energy*, vol. 36, no. 1, pp. 16–23, 2011.
[21] G. E. Box, G. M. Jenkins, and G. C. Reinsel, *Time series analysis: forecasting and control*. John Wiley & Sons, 2013.