Scenario-based Generalization bound for Anomaly Detection Support Vector Machine Ensembles

Roberto Rocchetta and Milan Petkovic

Department of Mathematics and Computer Science, Technical University of Eindhoven, The Netherlands.
E-mail: r.rocchetta@tue.nl, m.petkovic@tue.nl

Qi Gao

Department of Data Science, Philips Research, Eindhoven, The Netherlands.
E-mail: q.gao@philips.com

This work proposes an ensemble of robust Support-Vector-Machine (SVM) classifiers to monitor systems health-states given uncertain measurements from multiple sensors. Scenario optimization is a well-established theory to solve optimization problems in the presence of uncertainty and it is used to render a formal bound on the SVM misclassification probability. This probabilistic certificate of generalization (robustness) holds non-asymptotically and for any stationary random mechanism generating the data. A novel selection strategy is introduced seeking an ensemble design which maximize accuracy in the prediction and robustness given by Scenario theory. The framework is tested on a Prognostics and Health Management (PHM) challenge problem launched by the ARAMIS group in 2020 where a set of sensor measurements and labels are provided to abnormal operational states. The points of strength and the weaknesses of the proposed framework are presented and discussed.

Keywords: Scenario Optimization, Anomalies, Life-time, PHM, Robustness, Support-Vector-Machines.

1. Introduction

Prognostics and Health Management has recently become an important discipline for many engineering systems, products, and industry. Prognostics deals with several challenging topics such as detection and classification of anomalies, faults diagnostics, root causes analysis, and Remaining Useful Lifetime (RUL) estimation. Statistical data-driven tools, machine learning models, and ensembles of classifiers are particularly useful to tackle these challenges, Si et al. (2011).

Support-vector machines are supervised learning tools often used for anomalies detection and fault classification Han et al. (2019). Traditionally, SVM models are trained to maximize the classification accuracy while minimizing the risk of over-fitting the data. Accuracy and robustness are main concerns, especially when these models are used to monitor the health state of safety-critical components. Unfortunately, a long-lasting effectiveness can be hard to obtain due to the many uncertainties involved. Generalization errors (also known as out-of-sample errors) measure how accurately SVMs are able to predict uncertain outcome values for future data. Out-of-sample errors are often estimated on an empirical basis, for instance, partitioning the data in two sets, one for training the model and one to estimate the generalization error. A cross-validation method can be used to select a model design which maximizes the accuracy and generalization property. However, test-based methods reduces the body of data available for the training, i.e., a part of the labeled data is removed from the training set to estimate the model robustness empirically. Differently, generalization error bounds render formal guarantee on the out-of-sample error without the need for an empirical estimation, see e.g., Vapnik (1998). Generalization error bounds are mathematically derived and, therefore, partition of the labeled data is not required to estimate the classification error (at least in principle). An upper bound for SVM error rates was originally established in Vapnik (1998) while a lower bound can be derived from an argument of Haussler et al. (1994). A tighter bound was recently proposed by Hanneke and Kontorovich (2019) for agnostic cases.

Scenario optimization is a well-established technique to perform designs in presence of uncertainty and can be used to prescribe formal generalization bounds for the solution of scenario optimization programs, Campi and Garatti (2008). Scenario theory has been extensively studied for convex problems, e.g., Campi and Garatti (2011), and recently extended to non-convex cases, e.g., Campi et al. (2018). Only a few works applied Scenario theory to assess the robustness of classi-

Proceedings of the 30th European Safety and Reliability Conference and
the 15th Probabilistic Safety Assessment and Management Conference
Edited by Piero Baraldi, Francesco Di Maio and Enrico Zio
Published by Research Publishing, Singapore.
ISBN: 978-981-14-8593-0; doi:10.3850/978-981-14-8593-0

fication models. As example, Carè et al. (2018) proposed a classifier for medical applications and formal were provided for its specificity and sensitivity using Scenario theory. Recently, convex scenario programs with soft-constraints have been introduced by Garatti and Campi (2019) and later applied by Campi and Garatti (2020) to assess the robustness of SVM classifiers and other machine learning models. To the authors' knowledge, none of these works investigated Scenario theoretic generalization bounds for PHM approaches applied to composite systems.

In this work, we apply the results presented by Campi and Garatti (2020) to assess the robustness of the individual classifiers of an ensemble model specifically trained to detect abnormal components' operations and estimate the lifetime of composite systems. An upper bound on their misclassification probability is derived using Scenario theory. A novel model selection strategy is proposed seeking an ensemble model design with maximum prediction accuracy and good generalization properties. The ensemble approach is tested on a prognostics challenge problem proposed by the ARAMIS group in 2020. Similarly to Han et al. (2019), sensors measurements as explanatory input to tackle the prognostics challenge. SVM parameters are optimally selected by minimizing a metric of discrepancy between the ground truth and the predicted system lifetime while tuning the SVM ability to generalized given by the Scenario theoretic upper bounds on the misclassification probability.

2. PHM problem statement

This section summarizes the PHM challenge problem launched by the ARAMIS group, Cannarile et al. (2020). Let us consider a fleet of M parallel systems made of J identical components with mission time $T \in \mathbb{R}^+$. The $j = 1, .., J$ components in $m = 1, ..., M$ systems are subject to unknown degradation processes which may trigger anomalies in their operational states. An abnormal operation of a component does not imply its failure, but rather indicates harsher or difficult conditions. A label $y_t^{j,m} \in \{0, 1\}$ defines the health state of component j on system m at time t where $y_t^{j,m} = 1$ indicates an abnormal state. The time of the entry into an abnormal state is $\tau^{j,m} \in \mathbb{R}^+$. Notice that $y_t^{j,m} = 0$ for all $t < \tau^{j,m}$ and $y_t^{j,m} = 1$ for all $t \geq \tau^{j,m}$, i.e., during the mission time components are assumed non-reparable. A system fails when all its J components enter an abnormal state, T_f^m defines its time-to-failure and $T_{life}^m = \min\left\{T, T_f^m\right\}$ is the system life-time. If $T_{life}^m = T$ at least one component operated normally for the whole duration of the mission, i.e., $\tau^{j,m} > T$ for at least one j This is due to the parallel topology of the systems.

The participant to the challenge are asked to predict the time-to-anomaly $\tau^{j,m}$ and, to this end, a training data set of measurements from sensors and state labels is provided Saxena and Goebel (2008) as follows:

$$\mathcal{D}_N = \left\{ \left\{ \mathbf{x}_t^{(1:J),m}, y_t^{(1:J),m} \right\}_{t=0}^{T_{life}^m} \right\}_{m=1}^{M}$$

where the notation $(1 : J)$ is in place of $j = 1, ..., J$. $\mathcal{D}_N$ includes $N = M \times J \times \sum_{m=1}^{M} T_{life}^m$ vectors of sensors measurements defined by

$$\mathbf{x}_t^{j,m} = \left[s_t^{j,m,1}, ..., s_t^{j,m,K} \right] \tag{1}$$

where K is the number of sensors on each component and $s_t^{j,m,k}$ is the k^{th} sensor measurement at time t. The measurements $s_t^{j,m,k}$ depend on the (unknown) degradation level of the components and on the environmental-operational conditions. The vectors in $\mathcal{D}_N$ contains $N = 797456$ vectors $\mathbf{x}_t^{j,m}$ and corresponding labels. The data has been simulated from a fleet of $M = 200$ systems installing $J = 4$ parallel identical components each equipped with $K = 10$ sensors.

To test the predictor model, a test set $\mathcal{D}_{test}$ of unlabeled measurements from a fleet of $M_{test} = 50$ systems is given as follows:

$$\mathcal{D}_{test} = \left\{ \left\{ \mathbf{x}_t^{(1:J),m} \right\}_{t=0}^{T_{life}^m} \right\}_{m=1}^{M_{test}}$$

Note that, although the labels are not available in the test set, an information on the lifetime of each system is can be extracted $\mathcal{D}_{test}$. When $T_{life}^m < T$ a premature failure of all J components in the test system m occurred, i.e., $T_{life}^m = T_f^m$ and the failure time is given by $T_f^m = \max_{j \in \{1,...,J\}} \tau^{j,m}$.

3. Support Vector Machines

Support-Vector-Machine classifiers are used in this work to identify abnormal states $y_t^{j,m}$ from vectors of sensors measurements $\mathbf{x}_t^{j,m}$. To ease the notation, the vector of sensors measurements $\mathbf{x}_t^{j,m}$ are denoted by $x_i \in \mathcal{X} \subseteq \mathbb{R}^K$, the state labels $y_t^{j,m}$ by $y_i \in \{-1, +1\}$ and the data set with N signals and labels by $\mathcal{D}_N = \{x_i, y_i\}_{i=1}^{N}$. Normal state labels $y_t^{j,m} = 0$ are replaced by $y_i = -1$ without loss of generality. A standard linear 'hard-margin' SVM classifiers seeks a maximum width hyper-plane which separates the signal space $\mathcal{X}$ in two regions, one region for the class $y_i = -1$ and one for the class $y_i = 1$. The term 'hard-margin' indicates that all the samples x_i must fall in the region of $\mathcal{X}$ dedicated to the corresponding class y_i with no exceptions.

However, a linear separation of classes in $\mathcal{X}$ is often not obtainable in practice and this makes the standard linear 'hard-margin' training program unfeasible. To overcome this issue, the following non-linear 'soft-margin' SVMs optimization is considered:

$$\langle \omega^\star, b^\star, \zeta^\star \rangle = \arg \min_{\substack{b\in\mathbb{R},\omega\in\mathcal{U}\\ \zeta\in\mathbb{R}^N}} \{||\omega||^2 + \rho \sum_{i=1}^{N} \zeta_i : \quad (2)$$
$$1 - y_i(\omega\psi(x_i) - b) \leq \zeta_i,\ i = 1, ..., N$$
$$\zeta_i \geq 0,\ i = 1, ..., N\}$$

where ω and b are the parameters of the hyper-plane separating the classes, $|| \cdot ||$ is the vector norm operator, ζ is a vector of N non-negative slack variables, $\rho > 0$ is regularization parameter weighting the cost of margin violations and $\psi : \mathcal{X} \to \mathcal{U}$ is function mapping x from the physical space $\mathcal{X}$ to a Hilbert's space $\mathcal{U}$ (typically of higher dimension). Note that for a $\zeta_i = 0$, the original hard-margin constraint for sample i is satisfied, i.e., $1 - y(\omega\psi(x_i) - b) \leq 0$. Conversely, $\zeta_i > 0$ if sample i leads to a violation of the hard-margin constraint. Program (2) seeks a maximum width hyper-plane design $\langle \omega^\star, b^\star \rangle$ which separates the two classes in $\mathcal{U}$ while minimizing the cost of margin violations given by ζ. A linear classifier in $\mathcal{U}$ will map back to a non-linear classifier in $\mathcal{X}$.

3.1. *Misclassification probability*

The hyper-plane design $\langle \omega^\star, b^\star \rangle$ solution of (2) can be used to assign a label $\hat{y}$ to a new observations x by $\hat{y} = sign(\omega^\star\psi(x) - b^\star)$. If the true class y leads to a $y(\omega^\star\psi(x) - b^\star) > 0$, the observation x is classified correctly. A measure of the accuracy of the SVM model is the probability of misclassification, defined as follows:

$$P_f(\langle \omega^\star, b^\star \rangle) = \mathbb{P}[-y(\omega^\star\psi(x) - b^\star) > 0]$$

Similarly, the probability of margin violation is given by:

$$P_{MV}(\langle \omega^\star, b^\star \rangle) = \mathbb{P}[1 - y(\omega^\star\psi(x) - b^\star) > 0]$$

Note that a misclassification event also leads a violation of the margin constraints in program (2). However, the converse does not hold true. Hence, the probability of margin violation always upper bounds the miscalssification probability, i.e., $P_f \leq P_{MV}$.

In order to check weather the model generalizes well to unseen samples, an out-of-sample error estimate (sample-based misclassification probability) can be obtained as follows:

$$\hat{P}_f(\langle \omega^\star, b^\star \rangle) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{\hat{y} \neq y\}} \qquad (3)$$

where $\mathbf{1}_{\{\hat{y}_i \neq y_i\}}$ is the indicator function for a miscalssificaition event and n is the number of samples. Note that $\hat{P}_f$ only approximates the true (but unknown) probability P_f and the labels y_i are needed to compute (3). These may be unavailable for new observations x and, thus, the analysts is often forced to split the data set $\mathcal{D}_N$ in two sets. One set is used to optimized the model and one to test its generalization error given by $\hat{P}_f$. This procedure inevitably reduces the quantity of data available for the training, possibly leading to a sub-optimal performance.

Differently from traditional approaches, Scenario theory renders a formal bound on P_f and without the need for an empirical estimation, Campi and Garatti (2020). This bound is a probabilistic certificate of generalization (or robustness) on $\langle \omega^\star, b^\star \rangle$ which holds non asymptotically and independently from the underlying distribution of the samples. Scenario optimization programs with relaxed constraints and means to compute this bound will be presented in the next sections.

4. Scenario optimization theory

Consider a Data-Generating Mechanism (DGM), i.e., a probability space $(\Delta, \mathfrak{F}, \mathbb{P})$, where Δ is an event space equipped with a σ-algebra $\mathfrak{F}$ and a probability measure $\mathbb{P}$. In practice, the probability $\mathbb{P}$ is unknown and only a set of realization $\mathcal{D}_N = \{\delta_i\}_{i=1}^N$ is available. In Scenario theory the samples in $\mathcal{D}_N$ are called *scenarios* and are assumed to be Independent and Identically Distributed (IID) realizations of the unknown probability space. A convex scenario optimization program with relaxed constraints $\mathcal{SP}(\mathcal{D}_N)$ is defined as follows:

$$\langle d^\star, \zeta^\star \rangle = \arg \min_{d,\zeta} \{J(d) + \rho \sum_{i=1}^{N} \zeta_i : \qquad (4)$$
$$f(d, \delta_i) \leq \zeta_i,\ \delta_i \in \mathcal{D}_N$$
$$d \in \Theta, \zeta_i \geq 0,\ i = 1, ..., N\}$$

where d is a vector of controllable parameters (design variables) constrained in a closed convex space Θ, $f(d, \delta) : \Theta \times \Delta \to \mathbb{R}$ is any convex function in d and a scenario $\delta_i \in \mathcal{D}_N$ defines one of the N soft-constraints in $\mathcal{SP}(\mathcal{D}_N)$. The SVM training program (2) is equivalent to te scenario program (4) if the design is defined as $d = \langle \omega, b \rangle$, the uncertain factors by $\delta = (x, y)$, and the constraints $f(d, \delta_i) \leq \zeta_i$ is $1 - y_i(\omega \cdot \psi(x_i) - b) \leq \zeta_i$

for all $i = 1, .., N$. Notice that the uncertain factors δ enters as parameters in the constraints and, therefore, their dimension is inconsequential for the solution of $\mathcal{SP}(\mathcal{D}_N)$. Furthermore, note that if the cost of violation $\rho \to \infty$ the scenario program $\mathcal{SP}(\mathcal{D}_N)$ goes back to a standard from where the constraints are hard.

4.1. *Generalization error bound*

Scenario theory can be used to assess how well an optimal design $\langle \omega^\star, b^\star \rangle$ solution of (4) generalizes to future realizations from the DGM. Let us define a probability

$$V(d^\star) = \mathbb{P}\left[\delta \in \Delta : f(d^\star, \delta) > 0\right] \qquad (5)$$

$V(d^\star)$ is called violation probability and gives the likelihood that the optimized design $d^\star$ solution of (4) will fail to comply with a new constraint imposed by any new scenario $\delta \in \Delta$. Note that $V(d^\star)$ coincides with the probability of margin violation, that is, $V(d^\star) = P_{MV}(\langle \omega^\star, b^\star \rangle)$, when scenario program (4) defines a SVM soft-margin optimization as in (2).

Given a reliability parameter $\epsilon \in [0, 1]$, a design $d^\star$ is called ϵ-robust if $V(d^\star) \leq \epsilon$. The reliability $\epsilon(k)$ measures the ability of $d^\star$ to generalized to future data and it can be computed enumerating the support scenarios k in a support set $\mathcal{S}$. A set of support constraints (or support set) $\mathcal{S} \subseteq \mathcal{D}_N$ is a k-tuple $\mathcal{S} = \{\delta^{(i1)}, ..., \delta^{(ik)}\}$ for which the solutions of a scenario optimization is identical the solution of the same program with $\mathcal{S}$ in place of $\mathcal{D}_N$ Campi and Garatti (2008). For convex optimization programs, support constraints are also active constraints and, therefore, the support set is collection of scenarios

$$\mathcal{S} = \{\delta \in \mathcal{D}_N : f(d^\star, \delta \geq 0)\}$$

for which constraints in program (4) are active or violated in correspondence of the optimum $d^\star$. For the soft-margin program (2), $\mathcal{S}$ coincides with the set of support vectors. Theorem 4 in Garatti and Campi (2019) formally derives bounds on the violation probability $V(d^\star)$:

$$\mathbb{P}^N\left[\underline{\epsilon}(k) \leq V(d^\star) \leq \overline{\epsilon}(k)\right] \geq 1 - \beta \qquad (6)$$

where $[\underline{\epsilon}(k), \overline{\epsilon}(k)]$ are a lower and upper bounds on $V(d^\star)$, and β is a confidence parameter selected by the user. Given β and $k < N$, the bounds are obtained solving the following polynomial equation in t:

$$\mathcal{B}_N(t;k) = \frac{\beta}{2N}\sum_{i=k}^{N-1}\mathcal{B}_i(t;k) + \frac{\beta}{6N}\sum_{i=N+1}^{4N}\mathcal{B}_i(t;k) \qquad (7)$$

where $\mathcal{B}_N(t;k) = \binom{N}{k}t^{N-k}$. Equation (7) has two roots $[\underline{t}, \overline{t}]$ when $k = 0, 1, ..., N-1$ and bounds in (6) are obtained as $\underline{\epsilon}(k) = \max\{0, 1 - \overline{t}(k)\}$ and $\overline{\epsilon}(k) = 1 - \underline{t}(k)$. For the special case $k = N$ the following polynomial equation in t is considered:

$$1 = \frac{\beta}{6N}\sum_{i=N+1}^{4N}\mathcal{B}_i(t;k) \qquad (8)$$

equation (8) admit one solution, which is $\overline{t}(N)$. The corresponding upper bound is set to $\overline{\epsilon}(k) = 1$ whilst the prospective range for $V(d^\star)$ is $[\max\{0, 1 - \overline{t}(N)\}, 1]$.

For the theory to apply the samples in $\mathcal{D}_N$ are assumed independent and the optimization program (2) should admit a unique solution (existence and uniqueness assumptions). If more than one solution exists, uniqueness can be guaranteed by a tie-break rule, e.g., minimizing additional convex function in d. Furthermore, a technical assumption of non-accumulation is also required, i.e., for every $d \in \Theta$ the probability $\mathbb{P}[\delta : f(d, \delta) = 0]$ is null.

Non-accumulation is generally satisfied when δ has a probability density. Unfortunately, this might not be the case for the SVM program (2). For instance, given $\omega = 0$ and $b = \pm 1$ the probability $\mathbb{P}[(x, y) : 1 - y_i(\omega \cdot \psi(x_i) - b) = 0]$ becomes $\mathbb{P}[1 \pm y = 0] \neq 0$, hence leading to a degenerate situation. Nevertheless, Campi and Garatti (2020) presents a way to get around this difficulty and get the theory to work for a heated version of the problem. By a cooling procedure, one then finds rigorous results for SVM:

$$\mathbb{P}^N\left[P_{MV}(\langle \omega^\star, b^\star \rangle) \in [\underline{\epsilon}(k), \overline{\epsilon}(k)]\right] \geq 1 - 3\beta \quad (9)$$

where (9) gives bounds on the probability of margin violation. Since misclassification occurs more rarely than a constraints violation. Equation (9) can be used to prescribe a conservative upper bound on the misclassification probability:

$$\mathbb{P}^N[P_f(\langle \omega^\star, b^\star \rangle) \leq \overline{\epsilon}(k)] \geq 1 - 3\beta \qquad (10)$$

If $\omega^\star \neq 0$, the quantity k coincides with the number of support vectors, i.e., the number of scenarios for which $1 - y(\omega^\star \psi(x) - b^\star) \geq 0$. If $\omega^\star = 0$, k is the number of data points whose label belongs to the class with fewer elements. For a detailed discussion on the underlying assumptions and a formal description of the heating and cooling procedure the reader is reminded to Campi and Garatti (2020).

5. The proposed ensemble approach

An ensemble model for classification is a collection of several classifiers whose individual decisions are combined in some way to make a prediction. In this work, an ensemble of SVM models is design to monitor the health-state of composite systems. The SVMs detect malfunctions of individual components from sensor measurements and these predictions are combined to estimate the life-time of the composite system.

We propose an ensemble of J SVM classifiers used to monitor the health state of the J components in a system given sensors measurements, see Section 3. The time of entry into an abnormal state for component j is estimated from the ensemble model by

$$\hat{\tau}^{j,m} = \min_{t \in \{1,\dots,T^m_{life}\}} \{t : \hat{y}^{j,m}_t = 1\} \qquad (11)$$

where $\hat{\tau}^{j,m}$ it the time corresponding to the first predicted abnormal label, i.e., the smallest t such that $\hat{y}^{j,m}_t = 1$. The component-level predictions $\hat{\tau}^{j,m}, j = 1, ..., J$, are then combined in system life-time prediction obtained as follows:

$$\hat{T}^m_{life} = \min\{T, \max_{j \in \{1,..,J\}} \hat{\tau}^{j,m}\} \qquad (12)$$

where 12 is equal to T if one or more components operated normally for the whole mission time. Note that if all the J components fail, the predicted $\hat{T}^m_{life}$ coincides with the time of entry into an abnormal state of the last healthy component. This is due parallel structure of the system.

5.1. *Ensemble selection strategy*

An novel SVM parameters selection strategy is proposed to improve the accuracy and robustness of the ensemble. Traditionally, the sample-based estimate $\hat{P}_f$ and the ϵ-robustness given by Scenario theory are used to select the best classifiers to be embedded within the ensemble model. A classifier resulting in a small $\bar{\epsilon}(k)$ will generalize well to future scenarios.
The accuracy of the ensemble is given by

$$\gamma = \sqrt{\frac{\sum_{m=1}^{M_{test}} \left(\hat{T}^m_{life}(p) - T^m_{life}\right)^2}{M_{test}}} \qquad (13)$$

where γ is a discrepancy metric quantifying the error between the predicted system life-time $\hat{T}^m_{life}$ and the true system life-time T^m_{life} (derived from $\mathcal{D}_{test}$). The quantity p represents the set of SVM model parameters to be selected prior solving the soft-margin program (2). As example, p can include the functional form of the kernel $\psi(\cdot)$ and the cost margin violations ρ. The proposed procedure to select a robust ensemble model works as follows:

A: Parameters selection

1) Randomize a training set $\mathcal{D}_{N_0} \subset \mathcal{D}_N$ of size $N_0 < N$ and the SVM parameters $p = \{\psi, \rho\}$.
2) Optimize $\langle \omega^\star, b^\star \rangle$ via $\mathcal{SP}(\mathcal{D}_{N_0})$ and predict labels in $\mathcal{D}_N \setminus \mathcal{D}_{N_0}$.
3) Compute k, reliability $\epsilon(k)$ and $\hat{P}_f$.
4) Predict $\hat{\tau}^{j,m}$ and $\hat{T}^m_{life}$ by Eqs. (11)-(12) for the components and systems in $\mathcal{D}_{test}$. Compute γ given by equation (13).
5) Repeat steps A.1 - A.4 several times and select p leading to the lowest average $\hat{P}_f$, $\epsilon(k)$ and γ.

B: Finalize ensemble

1) Optimize $\langle \omega^\star, b^\star \rangle$ via program $\mathcal{SP}(\mathcal{D}_N)$ and for the selected p.
2) Compute number of support vectors k, reliability $\epsilon(k)$ and γ.

6. Analysis and results

6.1. *Data exploration*

The proposed ensemble approach is tested on the PHM challenge presented in Section 2. The data set $\mathcal{D}_N$ is heavily unbalanced and contains the 95.98% of normal states and only 32063 anomalies, i.e, the 4.02%. Among the 200 systems, only 19 have a $T^m_{life} < T$ (premature failures) whilst 181 completed the mission, i.e., $T^m_{life} = T$. Fourteen systems completed the mission without anomalies, 30 systems faced one anomaly, 85 dealt with two failures, and 52 completed the mission with only one component in a healthy state. Among the 800 components in the fleet, 389 have no recorded anomaly ($\tau^{j,m} = T^m_{life}$) and an anomaly occurs in average at $\mu(\tau^{j,m}) = 0.9186T$ and the standard deviation of the occurrence time is $\sigma(\tau^{j,m}) = 0.0448T$.

Figure 1 shows the measurements $s^{j,m,k}_t$ collected from the 10 sensors and the corresponding abnromal (red markers) or normal (blue markers) label. On the off-diagonal panels, we present correlations between pairs of sensors measurements using scatter plots. The panels on the diagonal present the conditional marginal distributions of $s^{j,m,k}_t$, $k = 1, ..., K$, given an abnormal operational state, i.e., $f_S(s^{j,m,k}_t | y^{j,m}_t = 1)$, and given a normal operational state, i.e., $f_S(s^{j,m,k}_t | y^{j,m}_t = 0)$. The marginal distributions differ the most for sensors 3, 4, 5 and 10. This may suggest a better separability of labels when the measurements from these sensors are used in the classification task.

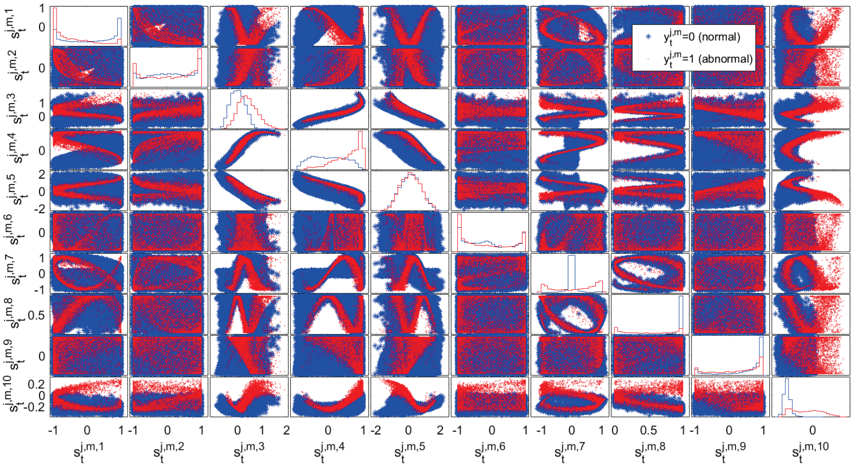

Fig. 1. The anomalous (red markers) and normal component states (blue markers) with respect to the 10 sensors measurements.

6.2. *Parameters selection and robustness assessment*

The first phase in the selection procedure is performed as described in Section 5.1. Multiple SVMs have been trained using the optimization program (2) constrained by random subsets of scenarios $\mathcal{D}_{N_0} \subset \mathcal{D}_N$ of increasing size N_0. Six kernel functions ψ are considered: Gaussian (radial basis function), polynomial (of order between 3 and 6) and linear. The scenarios $\mathcal{D}_N \setminus \mathcal{D}_{N_0}$ are used to estimate $\hat{P}_f$ whilst the robustness of the model is given by the upper bound $\overline{\epsilon}(k)$ presented in section (4). A confidence parameter $\beta = 10^{-6}$ is selected to compute the bound where a $\beta = 0$ means certainty For a given number of samples N_0 and confidence β, SVM models with less support vectors k produce better generalization bounds. Note that $\overline{\epsilon}(k)$ and $\hat{P}_f$ are inherently stochastic due to the uncertainty affecting the set of scenarios in $\mathcal{D}_{N_0}$.

This uncertainty is statistically quantified by sampling 10^3 realizations of the training set $\mathcal{D}_{N_0}$. Figure 2 presents the average values of $\hat{P}_f$ (right panel) and $\overline{\epsilon}(k)$ (left panel) for the six kernel functions and 8 sizes N_0. The Gaussian, linear and polynomial kernels are presented in black, red and and blue color, respectively. Round markers present the 10^3 realizations of $\overline{\epsilon}(k)$ and $\hat{P}_f$. SVMs trained on larger sets $\mathcal{D}_{N_0}$ are in average more accurate, i.e., lower $\hat{P}_f$, and robust, i.e., lower $\overline{\epsilon}$. The linear kernel result in the less accurate and less robust models. This can be explained looking at the distributions of the classes (see Figure 1) which are difficult to separate effectively with a linear classifier. The polynomial kernels led to the higher accuracy and robustness and a degree 3 provides the lowest $\overline{\epsilon}(k)$ for sizes $N_0 < 6000$. Interestingly, SVMs trained with the Gaussian kernel are the less robust for $N_0 < 6000$, but rapidly improve for larger sizes. This is probably due to the generally higher number of supports k.

6.3. *Ensemble performance*

The SVM models are used to label sensors measurements in the test data and to predict the time-to-anomaly $\hat{\tau}^{j,m}$ for the 200 components in $\mathcal{D}_{test}$. The predictions are combined to estimate the system life-time of the 50 test systems and the discrepancy metric γ measures the accuracy of the prediction given by the ensemble. Figure 3 presents the γ values for the six kernels in the panel on the right. The performance of the ensemble given by γ is compared to the performance of the individual classifiers in the ensemble in terms of their accuracy $\hat{P}_f$ and robustness ϵ (on the left panel). The averages values of the performance indicators and their variability due to the randomization of $\mathcal{D}_{10^4}$ are displayed by dashed lines and confidence boxes (including 50 % of the realizations), respectively. Ensembles models adopting polynomial SVM result in average, more accurate and robust to tackle this PHM problem. The lower discrepancy γ can be observed for an order 3. Similarly, the Gaussian kernel leads to relatively small discrepancies γ and with low uncertainty associated with. However, Gaussian kernel shows inferior performance of the individual classifiers. Hence, polynomial kernel is selected for the final training of the ensemble model.

The classifiers constituents of the ensemble are re-trained using the full set of measurements data $\mathcal{D}_N$ and polynomial kernels. Table 2 presents the robustness of the individual classifiers given

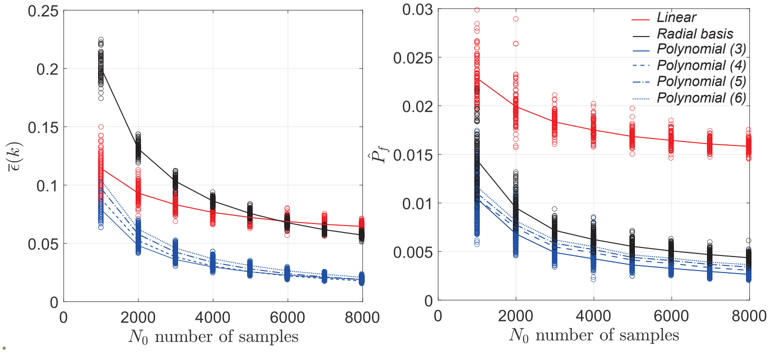

Fig. 2. Average $\hat{P}_f$ and $\epsilon(k)$ from 10^3 SVM models trained for data sets of increasing size N_0 and six kernels.

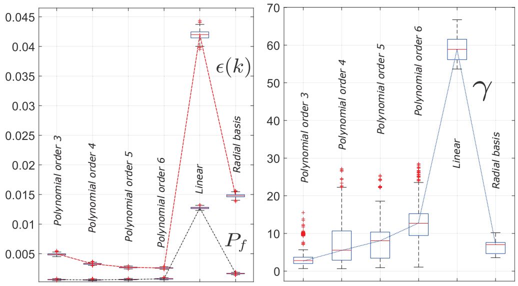

Fig. 3. Boxplot for $\epsilon(k)$, $\hat{P}_f$ and γ for 10^3 ensembles trained on a subset $\mathcal{D}_{N_0}$ of size $N_0 = 10^4$ and six kernels.

by $\bar{\epsilon}(k)$), the number of support vectors k and the ensemble accuracy given by γ. Polynomial degree 7 results in the best ensemble accuracy $\gamma = 0.38$. A polynomial kernel of degree 3 gives the best generalization $\bar{\epsilon}(343) = 5.736 \times 10^{-4}$, i.e., a probability of misclassification bounded by $\bar{\epsilon}(343)$. In other words, the SVM constituent of the ensemble model fail to classify correctly, at worst, 5.736 sensors signals out of 10000 new observations. Figure 4 displays the set of true system lifetimes T^m_{life}/T available for the 50 test systems and compares it to the set of prediction $\hat{T}^m_{life}/T$ provided by the ensemble model having $\gamma = 0.38$.

Table 1. Performance of ensembles of SVMs trained using the full set of $N = 797456$ scenarios and polynomial kernel of increasing order.

Order	k	$\bar{\epsilon}(k) \times 10^{-4}$	γ
3	343	5.736	0.7
4	579	9.096	2.14
5	491	7.855	1.84
6	581	9.124	1.94
7	549	8.674	0.38

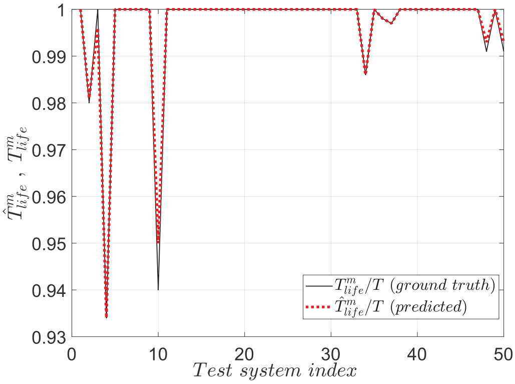

Fig. 4. Comparison between the predicted system lifetimes (red dashed line) and the available ground truth (black solid line).

6.4. *Discussion*

Support Vector Machines have great potential when used to tackle prognostic and health management problems. Uncertainty may affect the performance of the classifier and arises due to a lack of experimental evidence, dependency in the data, and poor modeling choices. Especially for safety-critical systems, uncertainty must be carefully addressed and quantified. The procedure introduced in this work provides a certificate of robustness, i.e., a probabilistic bound generalization error, which quantifies the uncertainty affecting a Support Vector Machine performance for future yet unseen samples.

The main advantages of the proposed approach can be summarized as follows: 1) The method is applicable to any kernel function which preserves the convexity of the training program. 2) Scenario-based certificates of generalization hold non-asymptotic, with minor assumptions on the DGM and provide a worst-case bound on the misclassification probability. 3) The dimension of the uncertain factors n_x has no effect on the computational burden of the SVM training program. 4) The bound $\overline{\epsilon}(k)$ is tighter for large data sets, i.e., it quantifies the value of information. 5) The bound $\overline{\epsilon}(k)$ worsen (slacken) for a higher number of support vectors k, i.e., this is a measures of the complexity of the SVM model and its tendency to overfit the data. 6) Traditionally, a data set of labeled measurements is required to estimate the out-of-sample performance of the model. The proposed method bound the out-of-sample performance and, therefore, allows exploiting all the available information without removing a subset of the labeled data for empirical error testing.

Some of the limitations of the proposed approach are as follows: 1) Dynamic effects and time-correlations between systems and signals may be not captured by the classifiers. 2) The generalization bound assumes IID samples and stationary DGMs. This is a common assumption for many other generalization error methods. 3) The bound ϵ gives no guarantees on the sensitivity and specificity individually. This is a major drawback for safety-critical applications where, e.g., the cost of a miss detection (failure to identify an anomaly) can be extremely high. Future extension of this work will investigate ways to amend these deficiencies.

7. Conclusions

This work introduces an ensembles model of Support Vector Machines classifiers equipped with a certificate of probabilistic generalization. Scenario optimization theory is used to equip the classifier design with a formally verifiable bound on its misclassification probability. This bound is a powerful certificate of generalization that holds non-asymptotically (for training sets of any size), irrespective of the random mechanism generating the data. This bound quantifies uncertainties which arise due to a lack of samples (bound gets tighter for larger data sets) and overly complex models prone to over-fitting. In fact, if a complex kernel is selected to answer a simple classification task, this likely leads to a high number of support vectors which, in turn, slacken the probabilistic guarantees on the probability of misclassification. This certificate is used within a novel ensemble selection procedure seeking ensemble designs which are, both, robustness and accurate in performing anomaly detections and system lifetime estimations. The ARAMIS challenge launched in 2020 is used to test the method and different classifiers are trained to identify abnormal component operations given a set of measurements from sensors. The uncertainty affecting the Scenario-based robustness bound, the out-of-sample accuracy, and the ensemble performance is analyzed with respect to six kernel functions and random subsets of the training data set.

Acknowledgement

The authors would like to acknowledge the gracious support of this work through the local authorities under grant agreement “ITEA-2018-17030-Daytime”.

References

Campi, M. and S. Garatti (2008). The exact feasibility of randomized solutions of uncertain convex programs. *SIAM Journal on Optimization 19*(3), 1211–1230.

Campi, M. C. and S. Garatti (2011, Feb). A sampling-and-discarding approach to chance-constrained optimization: Feasibility and optimality. *Journal of Optimization Theory and Applications 148*(2), 257–280.

Campi, M. C. and S. Garatti (2020). Scenario optimization with relaxation: a new tool for design and application to machine learning problems.

Campi, M. C., S. Garatti, and F. A. Ramponi (2018, Dec). A general scenario theory for nonconvex optimization and decision making. *IEEE Transactions on Automatic Control 63*(12), 4067–4078.

Cannarile, F., M. Compare, P. Baraldi, Z. Yang, and E. Zio (2020). The aramis challenge: Prognostics and health management in evolving environments. Technical report, Aramis Srl.

Carè, A., F. A. Ramponi, and M. C. Campi (2018). A new classification algorithm with guaranteed sensitivity and specificity for medical applications. *IEEE Control Systems Letters 2*(3), 393–398.

Garatti, S. and M. C. Campi (2019, Nov). Risk and complexity in scenario optimization. *Mathematical Programming*.

Han, H., X. Cui, Y. Fan, and H. Qing (2019). Least squares support vector machine (ls-svm)-based chiller fault diagnosis using fault indicative features. *Applied Thermal Engineering 154*, 540 – 547.

Hanneke, S. and A. Kontorovich (2019). Optimality of svm: Novel proofs and tighter bounds. *Theoretical Computer Science 796*, 99 – 113.

Haussler, D., N. Littlestone, and M. Warmuth (1994). Predicting 0, 1-functions on randomly drawn points. *Information and Computation 115*(2), 248 – 292.

Saxena, A. and K. Goebel (2008). Phm08 challenge data set. *NASA Ames Prognostics Data Repository, NASA Ames Research Center*.

Si, X.-S., W. Wang, C.-H. Hu, and D.-H. Zhou (2011). Remaining useful life estimation – a review on the statistical data driven approaches. *European Journal of Operational Research 213*(1), 1 – 14.

Vapnik, V. N. (1998). *Statistical Learning Theory*. Wiley-Interscience.