Title

Robust evaluation of time series classification algorithms for structural health monitoring

Permalink
https://escholarship.org/uc/item/1vp1z993

ISBN
9780819499905

Authors
Harvey, Dustin Y
Worden, Keith
Todd, Michael D

Publication Date
2014-03-09

DOI
10.1117/12.2044790

Peer reviewed

Robust evaluation of time series classification algorithms for structural health monitoring

Dustin Y. Harvey[a], Keith Worden[b], Michael D. Todd[a]

[a]Department of Structural Engineering, University of California, San Diego, 9500 Gilman Dr., La Jolla, CA USA 92093
[b]Department of Mechanical Engineering, University of Sheffield, Mappin Street, Sheffield, S1 3JD, United Kingdom

ABSTRACT

Structural health monitoring (SHM) systems provide real-time damage and performance information for civil, aerospace, and mechanical infrastructure through analysis of structural response measurements. The supervised learning methodology for data-driven SHM involves computation of low-dimensional, damage-sensitive features from raw measurement data that are then used in conjunction with machine learning algorithms to detect, classify, and quantify damage states. However, these systems often suffer from performance degradation in real-world applications due to varying operational and environmental conditions. Probabilistic approaches to robust SHM system design suffer from incomplete knowledge of all conditions a system will experience over its lifetime. Info-gap decision theory enables non-probabilistic evaluation of the robustness of competing models and systems in a variety of decision making applications. Previous work employed info-gap models to handle feature uncertainty when selecting various components of a supervised learning system, namely features from a pre-selected family and classifiers. In this work, the info-gap framework is extended to robust feature design and classifier selection for general time series classification through an efficient, interval arithmetic implementation of an info-gap data model. Experimental results are presented for a damage type classification problem on a ball bearing in a rotating machine. The info-gap framework in conjunction with an evolutionary feature design system allows for fully automated design of a time series classifier to meet performance requirements under maximum allowable uncertainty.

Keywords: Structural health monitoring, time series classification, feature extraction, info-gap decision theory, robustness

1. INTRODUCTION

1.1 Structural health monitoring and time series classification

Structural health monitoring (SHM) systems collect response measurements from civil, mechanical, and aerospace structures in order to analyze the data to detect, classify, and estimate the extent of damage present. Response measurements typically consist of time series or spectral measurements of acceleration, velocity, force, or strain. As such, SHM can be viewed as a specialized field of time series classification. Time series classification methods extend standard machine learning to directly address numeric sequence data input[1,2,3]. In SHM, model-based approaches to feature design are typically employed where the available knowledge of the structure, loading, and operating conditions drives the development and verification of a structural model. In the model-based approach, parameters of the model or error measures within the modeling and prediction processes are used as features in conjunction with standard classification or regression algorithms from machine learning to estimate the damage state. In contrast, general time series classification methods typically require no knowledge of the data-generating system instead relying on instance-based learning and a multitude of possible distance measures. However, the distance-based approach fails when transformation of the input space is required to identify features relevant for a given task. The authors previously proposed a genetic programming based system, Autofead, for data-based, automated feature extraction algorithm development[4,5]. Autofead provides a feature-based time series classification solution with no required knowledge of the data-generating system. This approach is widely applicable to SHM problems involving complex structures or failure modes which are difficult to model or to reduce the required development time and effort for an SHM system.

1.2 Classifier evaluation criteria

Regardless of the overall time series classification approach adopted within an SHM system, a quantitative measure is required to select the best features, parameter values, classifier choices, and other aspects of the analysis. While sampling methods and other best practices from machine learning are necessary to avoid estimation errors, they are often insufficient to provide a reliable estimate of the performance of a real-world SHM system. In SHM, performance estimation is complicated by the severe lack of available data in most applications. During development, data is typically only available for a limited number of damage scenarios and a small subset of the possible operational and environmental conditions the system may experience in usage. Due to these constraints, estimating the generalization error to select a robust SHM solution for data outside the space covered in the training dataset can be very difficult.

Typically SHM researchers have relied on simple classification accuracy to measure performance or perhaps a receiver operating characteristic (ROC) analysis as suggested by Provost, et al[6]. Many data normalization approaches have been proposed to deal with issues such as temperature fluctuation, but the efficacy of these methods is difficult to evaluate due to the same uncertainties and lack of data that make them needed. Probabilistic analysis is only suitable when sufficient knowledge is available regarding the source and character of uncertainties. Therefore, an alternative method is necessary to robustly evaluate solutions to account for uncertain knowledge of the conditions under which an SHM system will be required to operate. Such a measure will assist in selecting the most reliable time series classification method for SHM systems from a set of candidate solutions.

1.3 Paper overview

The rest of the paper is organized as follows. Section 2 describes the experimental dataset used for this study and summarizes the process for finding candidate solutions through Autofead. In section 3, an info-gap decision theory[7] (IGDT) based methodology is proposed and demonstrated for evaluating the robustness of candidate solutions. Conclusions from this work and suggestions for future development are detailed in section 4.

2. CONDITION-BASED MONITORING EXPERIMENT

2.1 Experimental configuration

A condition-based monitoring experiment was designed to perform detection and classification of bearing damage in rotating machinery. The experimental dataset was generated on the Machinery Fault Simulator from Spectraquest, Inc. shown in Figure 1. The motor drives a gearbox through a set of drive belts and a 91.4 cm shaft supported by three Rexnord ER12K ball bearings. Damage is introduced to the bearing farthest from the motor by replacing the healthy bearing with a pre-damaged specimen. The damage cases include a bearing with ball spalling and a bearing with an outer race fault. Time series response measurements are collected from the accelerometer located on top of the housing of the bearing of interest.

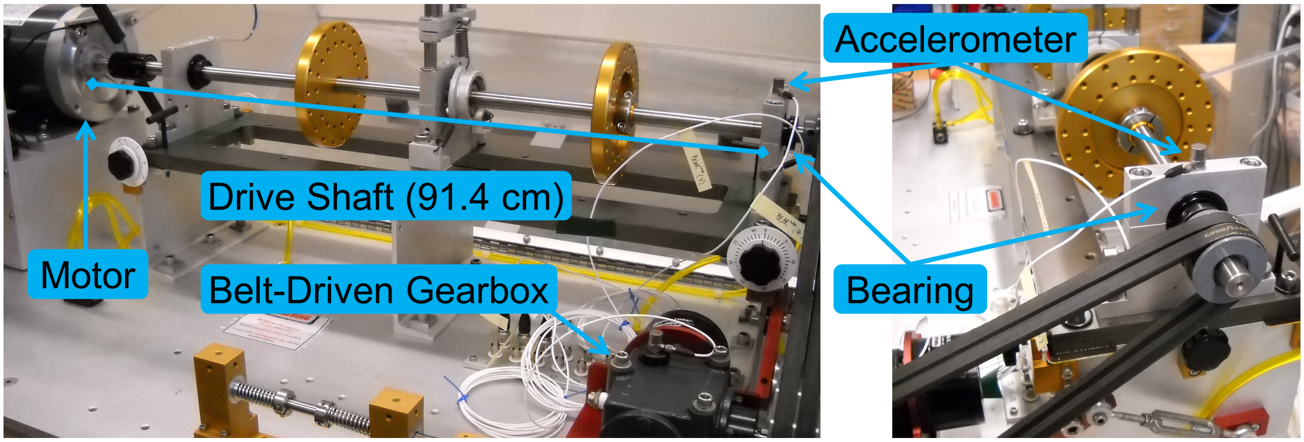

Figure 1. Condition-based monitoring experiment configuration.

Each response measurement consists of 540 samples collected at a sampling rate of 2.56 kHz. Responses are collected at a steady-state shaft speed of 1,000 rpm (16.7 Hz). The loading of the gearbox through the drive belts and the presence of an outer race fault on one bearing specimen produces an asymmetry to the system along the shaft axis. Therefore, the system is disassembled and reassembled 8 times per bearing specimen to mitigate experimental errors. 5,120 response measurements were collected for each of the three damage conditions. Ten percent of the 15,360 instances were used to generate candidate solutions with the rest held out as a test dataset. The objective of the experiment is to design a time series classifier that best identifies the three classes: healthy, ball spalling, and outer race fault. Figure 2 shows example measurements for each class.

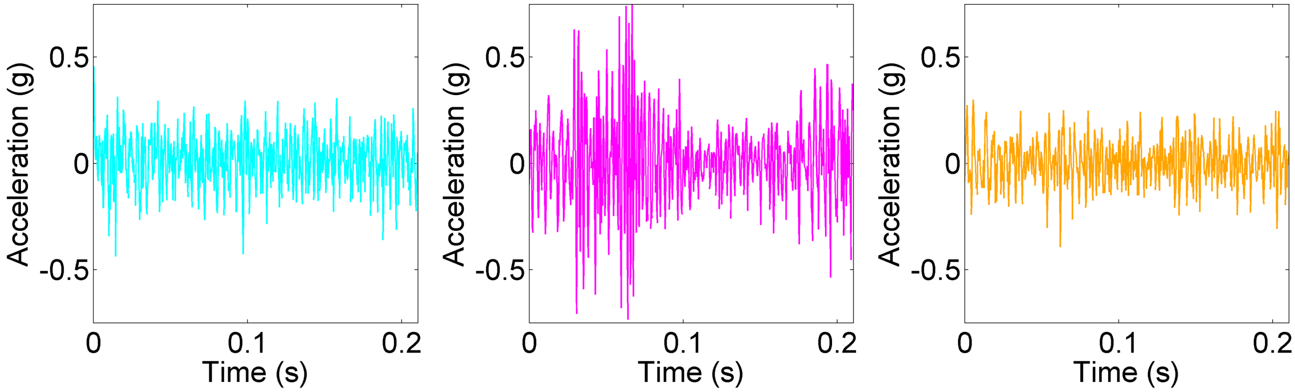

Figure 2. Example time series from healthy condition (left), ball spalling damage (center), and outer race damage (right).

2.2 Evolving candidate solutions

Autofead is a specialized genetic programming variant that evolves time series classification algorithms directly from training data. Solutions consist of a set of 1 or more feature algorithms that compose the solution's feature vector and selection of an appropriate classifier for nominal output or regression method for numeric output. Algorithms consist of a sequence of functions from the Autofead function library which operate on a single numeric sequence input such as a time series. For further details of the Autofead method, the reader is directed to Harvey and Todd[4,5]. Table 1 summarizes the configuration used for this study. For ease of visualization and implementation, solutions were restricted to two-dimensional feature spaces and a limited function library.

Table 1. Koza tableau for Autofead candidate solution generation.

Parameter	Setting
Objective	Design features and select classifier to detect and classify bearing damage
Solution structure	Classifier selection and set of features each consisting of a sequence of functions from function library
Function library	*Absolute value, center, center and scale, control chart, cube, cumulative summation, demean, difference, exponential, Hanning window, inverse, keep beginning, keep end, log10, normalize, select, set maximum value, set minimum value, sigmoid, sliding windows, slope fit, sort order, sorted bisection select, square, sum, transpose windows*
Pattern recognition	Classifier choice of Gaussian Naïve-Bayes, linear discriminant analysis, logistic regression, or CART decision tree
Fitness	Quadratic loss
Individual size limits	Maximum of 2 features per individual and up to 15 functions and 8 parameters per feature
Population initialization	Ramped to maximum of 2 features per individual and up to 5 functions and 3 parameters per feature
Population size	Total population size is 2,000 individuals. Parent pool truncated to 1,000 individuals. 100 offspring produced per search iteration.
Selection	Tournament (tournament size 2)
Genetic operators	Crossover, 75%; mutate, 10%; reproduce, 5%; add feature, 5%; remove feature, 5%
Termination	20,000 total individuals
Search repetitions	475 (9.5 million individuals)

Solutions within Autofead are ranked according to fitness as measured by quadratic loss through a cross-validation sampling procedure. The minimum fitness solution, here named solution A, uses a Gaussian Naïve-Bayes classifier. Figure 2 depicts the feature space and decision boundaries for solution A. Feature A1 uses the algorithm *set maximum value* (to 0.09), *sigmoid*, *difference*, *square*, and *sum* to separate the healthy and outer race fault conditions. The algorithm for feature A2 is *set maximum value* (to 0.38), *difference*, *difference*, *normalize*, *difference*, *absolute value*, and *sum* which tends to produce a smaller feature value for the ball spalling condition than the other classes. From Figure 2, it is evident that the two features together provide a well separated feature space with only a few misclassifications.

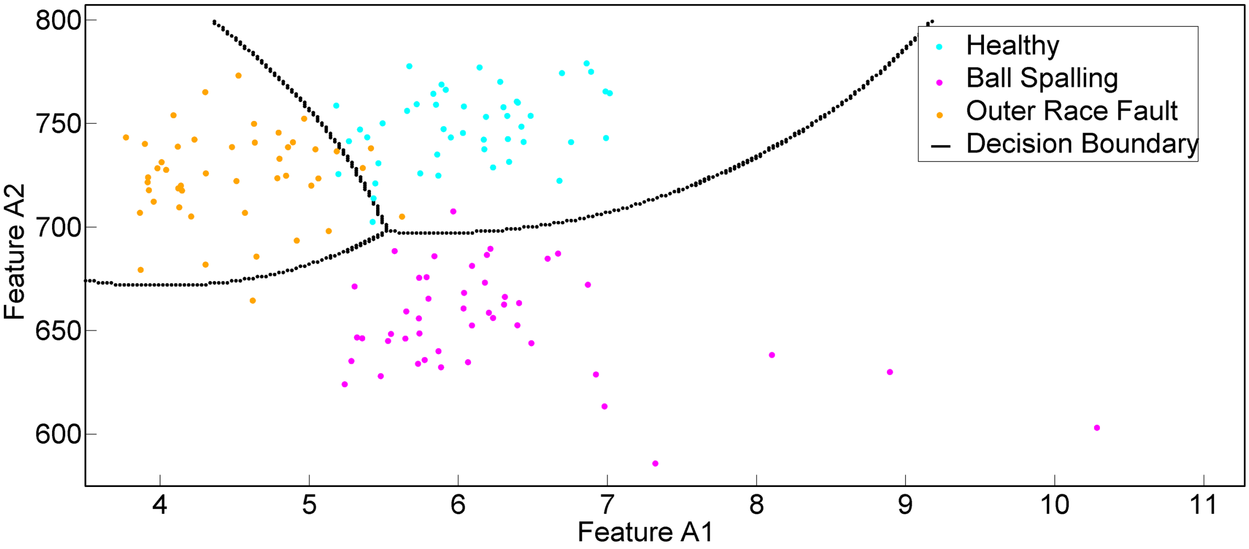

Figure 2. Feature space and decision boundaries for minimum fitness Autofead solution, solution A.

2.3 Candidate solution performance

Although solution A appears to perform very well, nine and a half million candidate solutions were generated, of which 27 are within 15% percent difference of the maximum fitness solution. The fitness of a solution is not a reliable estimate of expected performance as it is essentially an in-sample measure on the training dataset used within Autofead. A better estimate of the generalization performance is found by computing classification accuracy on an independent dataset. Figure 3 compares fitness and classification accuracy on the test dataset for the 100 minimum fitness solutions. The correlation between the two performance measures is only 0.45 although in this case the minimum fitness solution A, highlighted in red, also has the maximum classification accuracy on the test data set. The test data classification accuracy is preferred over fitness since it is an out-of-sample measure; however, the test data is still collected from a very limited set of conditions such as shaft speeds, amounts of damage, temperatures, etc. which leads us to search for an alternative measure to select the most robust solution.

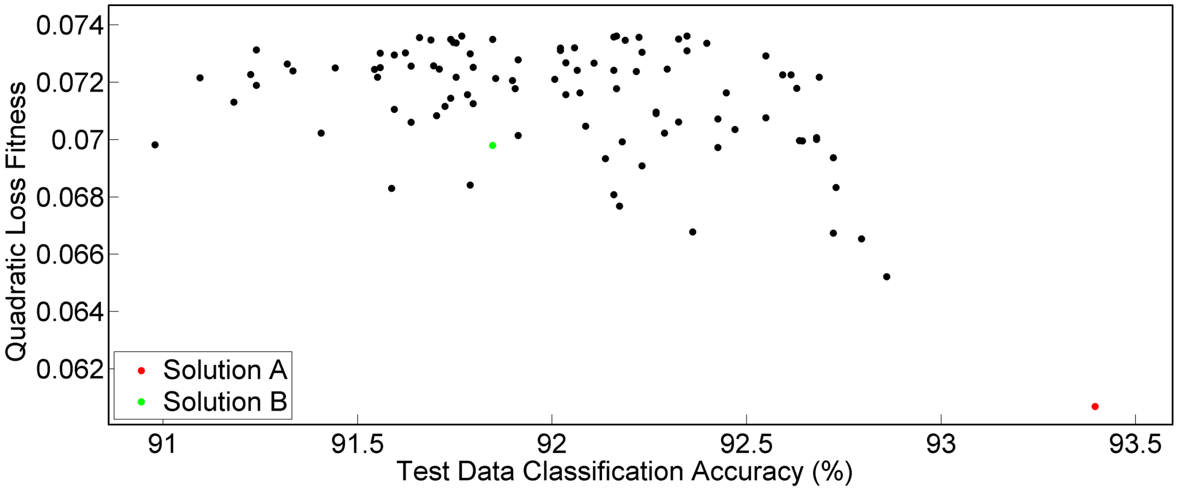

Figure 3. Comparison of 100 minimum fitness Autofead solutions to classification accuracy on independent test dataset.

3. ROBUSTNESS EVALUATION

3.1 Methodology

The goal of this work is to develop a method to robustly evaluate time series classifiers in the presence of multiple unknown sources of uncertainty that may affect an SHM system in real-life deployment. Probabilistic methods are poorly suited for this task as they require pre-specification of probabilities and repeated sampling to obtain good estimates of performance probabilities. IGDT provides a non-probabilistic framework to quantitatively evaluate the robustness and opportunity of making distinct choices in the presence of uncertainty[7]. Recently, Stull proposed the use of IGDT to account for the uncertainties in many aspects of SHM system design[8]. Pierce applied IGDT to the training of neural networks for classification and avoided sampling from the uncertainty model through an interval arithmetic implementation of IGDT for machine learning applications[9,10]. In this work, a similar interval arithmetic IGDT analysis is proposed to account for uncertainty within raw time series response measurements in the context of the SHM problem presented in section 2.

3.2 Info-Gap Model

The uncertainty model selected is an instantaneous energy-bound model on an independent test dataset in the time domain[7]. The model consists of a variable but equal-width interval applied to each time series sample in the test data. The interval radius, r, is dependent on the variable uncertainty level, α, and normalized by the standard deviation of the test dataset, σ, such that

$$r = \alpha\sigma. \tag{1}$$

Figure 4 depicts the first 50 samples of a single time series under no uncertainty, very small uncertainty, and a noticeable amount of uncertainty. The uncertainty model states that each nominal sample, x, may occur anywhere within $x - r \leq x \leq x + r$, the intervals shown in red.

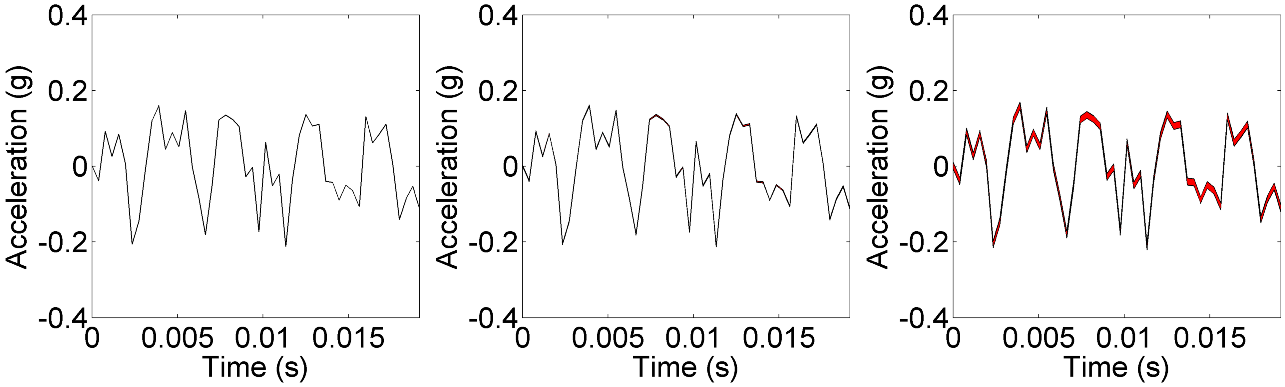

Figure 4. First fifty samples of example time series under envelope-bound uncertainty model for three levels of uncertainty, α = 0 (left), α = 0.014 (center), and α = 0.065 (right).

3.3 Uncertainty propagation

The selection of an envelope-bound info-gap model allows for the time series intervals to be efficiently propagated through each feature extraction algorithm using interval arithmetic. The Autofead functions were implemented in MATLAB using the interval arithmetic package INTLAB[11] to generate feature intervals for increasing uncertainty levels. The envelope-bound time series info-gap model results in a rectangular feature interval box in the two-dimensional feature spaces. Boxes that reside in a single decision region result in a single class prediction while boxes that intersect one or more decision boundaries produce multiple class prediction possibilities. Direct identification of class prediction possibilities from the feature interval boxes and decision boundaries reduces the required interval arithmetic operations and simplifies the use of complex classifiers. This approach is equivalent to the threshold-based analysis of class posteriors applied in[9,10]. Once class prediction probabilities are found for test dataset instance, a simple worst-case, best-case analysis is performed to determine the classification accuracy interval.

Figure 5 depicts the resulting feature intervals for two candidate solutions with the same uncertainty level, α = 0.014. The dot inside each feature interval box depicts the feature value under no uncertainty. For solution A, every feature interval overlaps at least one decision boundary at the level of uncertainty shown resulting in 0% worst-case classification accuracy. In contrast, the worst-case performance of solution B is only reduced to 85.5% from a nominal level of 91.8% with no uncertainty on the test data. Solution B achieves improved robustness primarily through the use of *sorted bisection select* at the end of each feature as opposed to *sum* in features A1 and A2 which accumulates uncertainty from all the summed samples. Solution B's feature algorithms are *center and scale*, *sliding windows* (21 sample windows), then *sorted bisection select* (minimum value) for B1 and *difference* four times then *sorted bisection select* (86th percentile) for B2. Interestingly, feature B1 divides the time series into 8.20 ms segments using *sliding windows*. The segment length is on the order of the characteristic vibration periods for the bearing geometry, and longer segment lengths degrade the performance significantly. Therefore, the algorithm may be exploiting the relative vibration frequency characteristics for the ball spalling and outer race defects to separate the two damage classes.

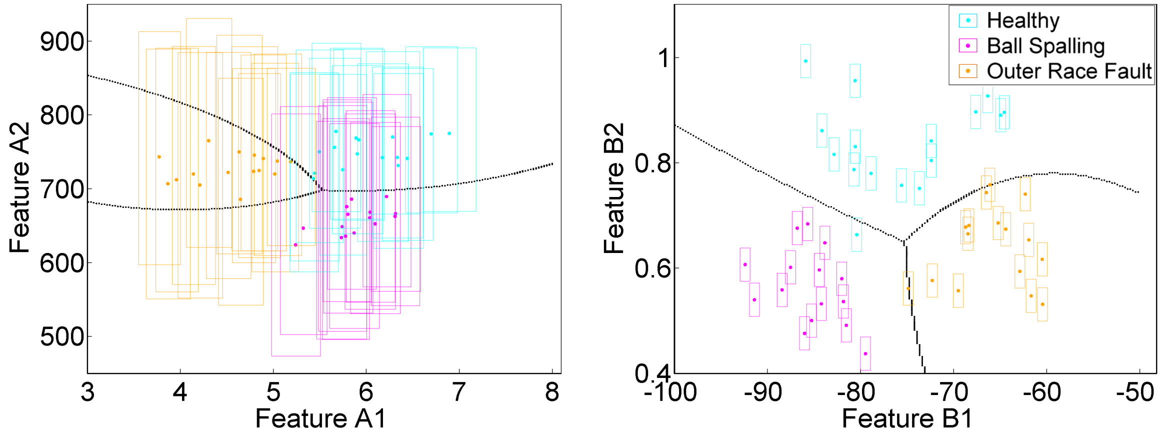

Figure 5. Example feature intervals for solution A (left) and solution B (right) under uncertainty level, $\alpha = 0.014$.

3.4 Robustness and opportunity

After propagating the time series info-gap uncertainty model through to the feature space and class prediction possibilities, robustness and opportunity curves are generated by considering the worst-case and best-case class prediction scenarios for increasing levels of uncertainty. Robustness of a candidate solution is measured as the largest uncertainty level which, worst-case, meets the required minimum classification accuracy. Opportunity represents the smallest level of uncertainty that could produce windfall of higher maximum classification accuracy. Windfall, here, describes achieving better-than-expected performance as uncertainty increases. Robustness and opportunity curves for the one hundred minimum fitness Autofead solutions are depicted in Figure 6. Solution A has poor robustness but good opportunity, while solution B is the most robust but provides the worst opportunity.

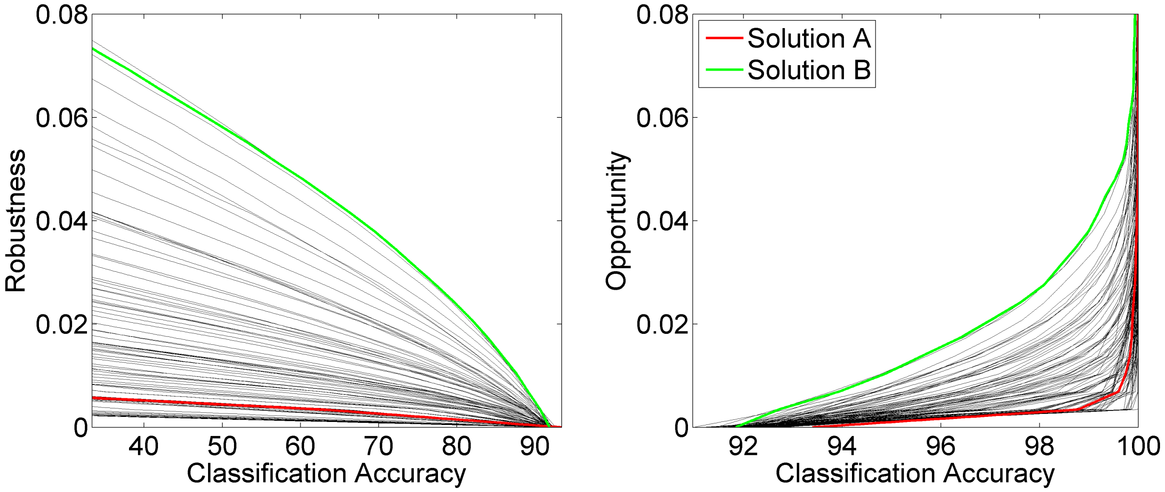

Figure 6. Robustness (left) and opportunity (right) of 100 minimum fitness Autofead solutions.

3.5 Preference reversal

In general, a single solution may not have the highest robustness or provide the best opportunity over the entire performance space. Change in decision preference as performance requirements vary is called preference reversal. Table 2 depicts the preference reversals in the top ten most robust solutions as the required minimum classification accuracy increases. Solution B is the most robust for required minimum classification accuracy from 60-89%.

Table 2. Top ten robust solutions for various minimum classification accuracy requirements.

Classification accuracy performance requirement	**Solutions in order of robustness**									
	1st	**2nd**	**3rd**	**4th**	**5th**	**6th**	**7th**	**8th**	**9th**	**10th**
50%	C	**B**	E	D	F	H	I	J	K	G
65%	**B**	C	E	D	F	I	H	K	J	G
80%	**B**	C	E	D	F	I	H	K	J	G
85%	**B**	E	C	D	F	I	H	J	K	G
90%	D	**B**	E	F	G	I	J	C	H	L

4. CONCLUSION

This work proposes and demonstrates experimental results for an info-gap decision theory based robustness analysis of time series classification algorithms. The approach is particularly suited to the design of SHM analysis processes to account for the typically restrictive set of damage scenarios and operational and environmental conditions represented within available training data. The genetic programming system Autofead was used to generate candidate solutions to detect and classify bearing damage in a rotating machine. An envelope-bound info-gap uncertainty model is applied directly to the time series in the test dataset then propagated through feature extraction and classification to generate class prediction possibilities. The interval arithmetic IGDT implementation provides a fast, quantitative robustness measure to select among candidate solutions in the presence of uncertainty and guarantee required minimum performance. In comparison to the fitness measure provided by Autofead or classification accuracy estimated on an independent test set, the proposed robustness measure should provide an improved evaluation criterion to estimate performance of real-world systems.

ACKNOWLEDGEMENTS

This work was supported by the Department of Defense (DoD) through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program.

REFERENCES

[1] Xing, Z., Pei, J., and Keogh, E., “A brief survey on sequence classification,” ACM SIGKDD Explorations Newsletter 12(1), 40–48 (2010).

[2] Fu, T., “A review on time series data mining,” Engineering Applications of Artificial Intelligence 24(1), 164–181 (2011).

[3] Wang, X., Mueen, A., Ding, H., Trajcevski, G., Scheuermann, P., and Keogh, E., “Experimental comparison of representation methods and distance measures for time series data,” Data Mining and Knowledge Discovery 26(2), 275–309 (2013).

[4] Harvey, D.Y., and Todd, M.D., “Automated extraction of damage features through genetic programming,” Proc. SPIE Smart Structures and Materials+ Nondestructive Evaluation and Health Monitoring, 86950J (2013).

[5] Harvey, D.Y., and Todd, M.D., “Automated Near-Optimal Feature Extraction Using Genetic Programming with Application to Structural Health Monitoring Problems,” Proc. International Workshop on Structural Health Monitoring, (2013).

[6] Provost, F.J., Fawcett, T., and Kohavi, R., “The case against accuracy estimation for comparing induction algorithms.,” Proc. International Conference on Machine Learning, 445–453 (1998).

[7] Ben-Haim, Y., [Info-gap decision theory: decisions under severe uncertainty], Academic Press (2001).

[8] Stull, C.J., Hemez, F.M., and Farrar, C.R., “On assessing the robustness of structural health monitoring technologies,” Proc. International Modal Analysis Conference, 1–11 (2012).

[9] Pierce, S.G., Worden, K., and Manson, G., “A novel information-gap technique to assess reliability of neural network-based damage detection,” Journal of Sound and Vibration 293(1), 96–111 (2006).

[10] Pierce, S.G., Ben-Haim, Y., Worden, K., and Manson, G., “Evaluation of neural network robust reliability using information-gap theory,” Neural Networks, IEEE Transactions on 17(6), 1349–1361 (2006).

[11] Rump, S.M., “Developments in reliable computing,” INTLAB-INTerval LABoratory, Kluwer Academic Publishers, Dordrecht, NL (1999).