ROC with Cost Pareto Frontier Feature Selection Using Search Methods

Ryan Meekins¹ · Stephen Adams² · Kevin Farinholt¹ · Sherwood Polter³ · Peter A. Beling²

Received: 12 January 2020 / Revised: 28 July 2020 / Accepted: 25 August 2020 / Published online: 26 October 2020
© The Author(s) 2020

Abstract

Cyber-physical systems (CPS) are finding increasing application in many domains. CPS are composed of sensors, actuators, a central decision-making unit, and a network connecting all of these components. The design of CPS involves the selection of these hardware and software components, and this design process could be limited by a cost constraint. This study assumes that the central decision-making unit is a binary classifier, and casts the design problem as a feature selection problem for the binary classifier where each feature has an associated cost. Receiver operating characteristic (ROC) curves are a useful tool for comparing and selecting binary classifiers; however, ROC curves only consider the misclassification cost of the classifier and ignore other costs such as the cost of the features. The authors previously proposed a method called ROC Convex Hull with Cost (ROCCHC) that is used to select ROC optimal classifiers when cost is a factor. ROCCHC extends the widely used ROC Convex Hull (ROCCH) method by combining it with the Pareto analysis for cost optimization. This paper proposes using the ROCCHC analysis as the evaluation function for feature selection search methods without requiring an exhaustive search over the feature space. This analysis is performed on 6 real-world data sets, including a diagnostic cyber-physical system for hydraulic actuators. The ROCCHC analysis is demonstrated using sequential forward and backward search. The results are compared with the ROCCH selection method and a popular Pareto selection method that uses classification accuracy and feature cost.

Keywords Feature selection · Receiver operating characteristics · Pareto optimality

This article belongs to the Topical Collection: *Data-Enabled Discovery for Industrial Cyber-Physical Systems*
Guest Editor: Raju Gottumukkala

✉ Stephen Adams
sca2c@virginia.edu

Ryan Meekins
meekinsr@lunainc.com

Kevin Farinholt
farinholtk@lunainc.com

Sherwood Polter
sherwood.polter@navy.mil

Peter A. Beling
pb3a@virginia.edu

Introduction

Cyber-physical systems (CPS) [42] have combined the physical world with advanced computation through sensing and significant advances in artificial intelligence (AI) and machine learning. While CPS are relatively new, the term was first coined in 2006 [36], the potential advantages of these systems have been seen in numerous fields, including manufacturing [19, 21, 36], medicine [27], and transportation [44, 47]. Industrial CPS is a general term applied to any CPS functioning in an industrial setting [30].

Prognostics and health management (PHM) is one application area for CPS [28] in industrial settings. PHM is the engineering discipline of health-informed decision-making with respect to maintenance operations [22, 46]. Data-driven PHM approaches construct models for predicting the current and future health state of systems from collected data [29, 39].

While there are many areas of CPS research, one area that is lacking is the design of CPS with respect to cost. CPS are composed of sensors, actuators, a central decision-making

¹ Luna Innovations, Charlottesville, VA, USA

² Department of Engineering Systems and Environment, University of Virginia, Charlottesville, VA, USA

³ Naval Surface Warfare Center Philadelphia Division, Philadelphia, PA, USA

unit, and a network connecting all of these components. Each of these components has an associated financial cost, and these costs can grow as the size of the system increases. The central decision problem in the design of a CPS is the choice of hardware and software components. In many scenarios, this decision problem may be informed by some type of AI agent. There are numerous types of cost that can be associated with this agent. The most studied form of cost in AI is misclassification cost, i.e., the cost of classifying one class as another class. Cost-sensitive learning incorporates misclassification cost into a model's training process [31]. Another form of cost common in AI is the test cost or the cost of acquiring missing data [13]. Financial costs associated with the collection, storage, and processing of data are important considerations when designing an agent but not as widely studied in the field of AI.

Feature selection is a concept in machine learning where a subset of collected features are selected as inputs into an algorithm [1, 14, 17, 43, 50]. While feature selection's primary benefit is to eliminate noisy features and improve the performance of a model, feature selection can also help reduce the cost of a predictive model by limiting the data that must be collected, stored, and processed. In a physical sense, feature selection can help inform the sensor selection process, i.e., sensors that do not provide relevant information to the decision agent can be eliminated from the system. Generally, feature selection methods optimize the feature set based on some notion of performance. Cost-sensitive feature selection allows for the cost of each feature to be incorporated into the feature selection process [2, 4, 10, 20, 34, 35, 38, 52].

Feature selection methods are generally divided into three types: embedded techniques, filters, and wrappers. Embedded techniques simultaneously learn a model and select features for that model. Filters select features independent of a model given an objective function. Wrappers sequentially move through the feature space and evaluate sets of features given a specific model and an evaluation function. Wrappers are composed of three components: the model, the search algorithm, and the evaluation function. The model is any type of predictive model (both classification and regression). One of the strengths of wrappers is that the selected feature set is specific to a particular model. However, in some applications, this may be considered a weakness, e.g., when feature selection is used to gain knowledge about a system. The search algorithm sequentially selects sets of features for evaluation. Common search functions are forward selection, where the model starts with no features and sequentially adds them to the model, and backward selection, where the model begins with all possible features

and sequentially removes them from the model. The evaluation function is used to assess the relevance of each proposed feature set. Common evaluation functions for wrappers are classification accuracy, root mean squared error, and R -squared. A wrapper iteratively selects a feature set, trains a new model, and evaluates the selected feature set until a stopping criteria is met.

One of the most common problems in PHM settings is the determination of good or bad components. In machine learning terms, this can be cast as a binary classification problem. In most settings, the misclassification cost is assumed to be equal between the classes. When the misclassification cost is unknown, receiver operating characteristic (ROC) curves are a popular performance and comparison metric for binary classification models [40]. ROC curves plot the true positive rate of the classifier against the false positive rate and are popular evaluation tools when misclassification cost is unknown because the plot allows for comparison at any misclassification cost. The area under the ROC curve can be used as a point estimate for the performance of a binary classifier and used as the evaluation function for a wrapper feature selection algorithm. One disadvantage to the area under the curve metric is that it combines the performance at different points along the curve. The use of the area under the curve metric could lead to the scenario where a sub-optimal classifier is selected because the true misclassification cost is unknown during the selection process. A method for binary classifier selection that is robust to unknown misclassification cost is to find an ROC curve, and the corresponding classifier, that dominates all other ROC curves using the convex hull [9, 40].

This study proposes a method for cost-sensitive feature selection for binary classifiers when the misclassification cost is unknown. For this study, a classifier is defined by both the feature set and the type of classifier. A wrapper feature selection method is proposed that uses the previously developed ROC Convex Hull with Cost (ROCCHC) method [33] as the evaluation function. The ROCCHC method constructs three-dimensional ROC curves where the third dimension is the cost of the evaluated feature set. Using the proposed method, Pareto optimal feature sets with respect to performance and cost are returned. The proposed method extends the ROCCHC concept previously investigated in [33] and builds on the state-of-the-art of feature selection with cost by proposing the ROCCHC method as the selection method for wrapper feature selection algorithms. This study also extends the evaluation of ROCCHC.

The main contributions of this study are (1) the presentation of a new evaluation function for wrapper

feature selection algorithms that uses the ROCCHC method to incorporate feature set cost and (2) the demonstration of the proposed algorithm on common binary classification data sets and an industrial CPS data set. The proposed evaluation function is demonstrated using the common forward and backward search techniques, but it could be implemented with any search technique used for wrapper feature selection algorithms. This research provides efficient techniques for designing CPS when the central decision-making unit is a binary classifier and the selection of hardware components is influenced by the features of the classifier, for example the sensors in the system. The wrapper using ROCCHC is a general method that could be used for feature selection for any binary classifier where cost is a concern. However, it is especially suited for PHM activities in industrial CPS and is therefore primarily demonstrated on a condition-monitoring system for a hydraulic actuator. The method could be applied to numerous other domains including autonomous vehicles, medical CPS, and civil infrastructure.

The remainder of the paper is organized as follows. The “[Background](#)” section provides the necessary background on the hydraulic actuator problem that is the setting for our empirical tests, as well as background on current feature selection methods, the Pareto Front analysis, and the ROC Convex Hull (ROCCH) method. The “[Proposed Method](#)” section presents the proposed ROC Convex Hull with Cost feature set selection method. The “[Numerical Experiments and Results](#)” section discusses the numerical experiments and results. Finally, the “[Conclusion and Future Work](#)” section provides the conclusions and future work.

Background

In this section, background information on the hydraulic actuator problem, feature selection, the Pareto Front, and the ROCCH method are outlined.

Hydraulic Actuator Problem

Hydraulic actuators are used for many functions by the US Navy on both surface ships and submarines. The Navy is making a concerted effort to move from a time-based maintenance schedule to a condition-based system. However, there are some physical restrictions to the installation and use of condition-based PHM systems. Primarily, the power consumption of the installed PHM hardware (e.g., sensors, data processors, CPUs) should be minimized, and ideally these devices should be able to function for several years under battery power. Another concern is the reliability of any installed hardware and the

robustness to change of any algorithms used for estimating current and future health state.

Adams et al. [3] outline a hydraulic actuator test stand and an investigation into the power consumption of standard classification algorithms. The test stand has the capability to simulate common faults, and data is collected from acceleration, pressure, flow rate, and angular position sensors. In total, five common faults are simulated, each with varying degrees of severity. The training time for each classifier was used as a surrogate for power consumption. A hierarchical classification approach has been proposed to minimize resource consumption [6]. These studies use a relatively small feature set and do not focus on the feature selection problem. In a similar study, the low-power sensor node for collecting data on the hydraulic actuator is presented [16].

Feature selection is an important part of reducing the power consumption of this PHM system. Several feature selection and feature extraction methods have been compared for this hydraulic actuator system [5]. The originally collected hydraulic actuator data set was also used in developing the ROCCHC method [33]. However, this previous study only developed the ROCCHC method and did not include a selection method such as the wrapper approach proposed in the presented study. The actuator data set used in this paper is a newer data set that includes more engineered features that have led to increased predictive performance of machine learning models.

Feature Selection

Traditional feature selection methods attempt to choose a subset of features that increase classifier performance. As previously stated, feature selection methods can be classified into one of three broad categories: embedded techniques, filters, and wrappers. Each of these methods has strengths and weaknesses. The selection of the type of feature selection method is generally considered to be problem specific.

Embedded techniques select features while building the model. A common example of an embedded technique is the decision tree algorithm, which selects a new splitting feature from the full set by acting greedily with respect to an evaluation metric such as information gain. Embedded techniques are less common than filters and wrappers, but these techniques generally combine the benefits of both filters and wrappers. The drawback to embedded techniques is that they are usually integrated into the model training process so the machine learning model is pre-selected.

Filters use statistical merit metrics, such as information gain, to reduce or “filter” the feature set before choosing a model. One common filtering approach in regression

problems is to select features based on the correlation between the dependent and independent variables. Focus [8], Relief [23], and ReliefF [26] are other common filtering approaches. Filters are often computationally efficient and can be used independently of a model. However, due to the independence from the model, filters can often select feature sets that under perform.

Wrappers use an iterative search process of changing a model's feature set and then evaluating the new model's performance. This type of feature selection can often produce a near-optimal feature set for the selected model but are computationally expensive due to the iterative process. If the model is complex, training the model at each step of the feature selection algorithm will further increase the computational cost of the feature selection step. Numerous search methods and evaluation functions have been proposed for wrappers [7, 12, 24, 41, 45].

Algorithms from each of these three classes of feature selection methods have been developed to include feature cost. Ling et al. [32] developed a classification tree algorithm that includes both misclassification cost and feature cost. Zhu et al. [52] add feature cost to the random forest algorithm. Bolón-Canedo et al. [10] propose the cost-based feature selection framework, which can be used with any filtering evaluation function. Evaluation functions for wrappers have also been modified to include the cost of the features [25, 34]. However, there are an insufficient number of feature selection papers that perform a multiple objective optimization using Pareto optimality. Three recent examples include a particle swarm optimization wrapper [48], an artificial bee colony optimization wrapper [18], and an evolutionary algorithm for optimizing neural networks [51]. These all use a cost metric, either the number of features (using this metric assumes that the number of features is positively correlated with the cost and that the cost increases linearly with the number of feature; however, this assumption is dependent on the problem being investigated and not always the case, e.g., the relationship between the number of features and the cost is non-linear) or the model complexity, and accuracy or error rate in order to evaluate the feature sets or model parameters. However, accuracy has been deemed inadequate for real-world problems where the costs of misclassification and class distribution are unknown [40]. Therefore, our proposed method builds upon the ROC analysis because it assumes that the costs of misclassification and the class distributions are unknown.

Pareto Front

The Pareto Front method is a multiple objective analysis for evaluating competing alternatives. For the feature selection field, two common objectives are the feature set cost, usually based on the number of features, and the feature

Fig. 1 Pareto Front using two objectives, cost and accuracy

set performance, usually resulting from the accuracy of a model using a certain feature set such as in [18] and [49]. The Pareto Front then consists of feature sets that dominate other sets in terms of the cost and performance metrics. A sample Pareto Front is displayed in Fig. 1. This figure shows the cost and accuracy metrics resulting from various feature sets. Since we want to increase the accuracy and decrease the cost, the feature sets that have the highest accuracy and the lowest cost are on the Pareto Front, shown using the green dashed line. These feature sets dominate all others in terms of accuracy and cost. The final feature set would be chosen using tradeoffs between the objective and/or constraints.

Fig. 2 Example ROC curve

Fig. 3 Example ROCCH method

ROC Convex Hull

Feature selection aims to select a feature set that will minimize cost and maximize performance; however, this field has mainly focused on using accuracy as the sole performance metric, although it assumes that the real-world costs of misclassification and class distribution are known. All methods that break the classification performance into a single metric also make these assumptions, such as the area under the curve (AUC) metric. These metrics are inadequate because they fail to provide a sensitivity analysis on the models detection threshold (e.g., at what posterior probability to classify to a certain class). ROC analysis is a common tool used for evaluating binary models that does not require the real-world class distributions or misclassification costs.

The ROC analysis evaluates models using their false positive rate (FPR) and true positive rate (TPR). These are calculated as $FPR = \frac{\#FP}{\#N}$ and $TPR = \frac{\#TP}{\#P}$, where $\#FP$ is the

number of false positives, $\#N$ is the number of negatives, $\#TP$ is the number of true positives, and $\#P$ is the number of positives. ROC graphs show a model's TPR and FPR as the model's detection threshold is altered from assigning all observations as positives ("accept all") to assigning all observations as negatives ("reject all"). Two sample models, A and B, are shown in Fig. 2, along with the ROC curve of an arbitrary classifier.

When there are competing classifiers, the ROCCH method can be used to identify a set of classifiers that dominate in terms of the misclassification cost. It accomplishes this without requiring the actual costs of misclassification or the actual class distribution. This is accomplished using a convex hull in ROC space, where dominating classifiers will have operating points (FPR, TPR) on the convex hull [9, 40].

Feature set selection using the ROCCH method is shown in Fig. 3. This figure shows a ROC graph consisting of four feature sets, A, B, C, and D. The convex hull of ROC curves resulting from these feature sets are shown using the shaded region. Notice that feature sets C and D have operating points on the ROC convex hull and feature sets A and B do not. This means that either feature set C or D should be selected over A and B regardless of the misclassification cost. If the misclassification cost was known, a choice between feature sets C and D could be made. The ROCCH method allows for feature set selection when the misclassification cost is unknown but does not address the problem of feature set cost.

Proposed Method

In this section, the proposed feature selection method is outlined. The section begins by describing the previously developed ROCCHC and then describes how it is used as an evaluation function for a wrapper feature selection method.

Fig. 4 The ROC Convex Hull with Cost (ROCCHC) evaluation method

ROC Convex Hull with Cost

The ROCCHC method [33] extends the ROCCH analysis to aid in feature selection when there is additional cost metric, such as feature cost. The ROCCHC method adds this additional cost as a third dimension of the ROC graph and then identifies feature sets that are Pareto optimal by combining the ROCCH method with Pareto optimality. The feature set costs could be the financial cost of adding the sensors to the CPS system that are used to obtain the data or other types of cost such as the computational complexity of calculating these features on-board the CPS system.

This analysis assumes that the feature set cost is known for all feature sets, where this cost can be ordinal (i.e., 1st, 2nd, 3rd,...) or numeric. The ROCCHC method is shown visually in Fig. 4, which displays the ROC curves produced by feature sets A, B, C, and D, with feature set cost shown in parenthesis. Since there are two unique feature set costs of \$1 and \$2, two ROC convex hulls are computed and shown.

The ROC convex hull calculated for a feature set cost of \$1 is displayed in the left ROC graph of Fig. 4. This convex hull was computed while including the feature sets A and B because their feature set cost is less than or equal to \$1. This ROC convex hull selects feature set A. Feature set B is not selected because it has a worse performance and an equal cost to A, or is dominated by A in terms of the ROCCH method and Pareto optimality.

The ROC convex hull calculated at a feature set cost of \$2 is shown in the right ROC graph of Fig. 4. This convex hull was computed using feature sets (A, C, and D) because they all have a feature set cost of less than or equal to \$2 and were not previously dominated. This convex hull selects C and D because they dominate in terms of ROC performance. The ROCCHC optimal feature set would then include A, C, and D because they were all chosen in terms of the ROCCH method and their cost using Pareto optimality. Feature set A is ROCCHC optimal because it has a lower cost than C and D and out performs B, i.e., it was not dominated in terms of Pareto optimality.

A feature set can only be selected when computing the ROC convex hull while including all feature sets with a cost equal to or less than its feature set cost. This ensures that the ROCCH method selects feature sets that are Pareto optimal in terms of ROC graph and the feature set cost. The ROCCH feature sets can be in the ROCCHC set; however, some can be dominated if they overlap with ROC curves from a lower cost feature set. If there is correlation between features, the ROCCHC method would select the lower cost feature(s), as long as the higher cost feature(s) did not provide

additional ROC performance (e.g., are not on the ROC convex hull). If higher cost feature(s) provided additional ROC performance, they may be ROCCHC optimal if they are not dominated (in terms of cost and ROC performance) by other feature sets.

The steps to select ROCCHC optimal feature sets, provided a data set, are outlined in Algorithm 1. The inputs to the algorithm are the feature set costs and the ROC curves of all competing feature sets. The cost of all feature sets (including feature sets that contain correlated features) must be known or calculated before implementing the ROCCHC analysis. The algorithm outputs the feature sets in the ROCCHC set, α . The algorithm first determines the unique feature set costs and sorts them in ascending order. The algorithm then iterates over these unique feature set costs and obtains all of the feature sets with a cost less than or equal to the current unique feature set cost. The algorithm then takes a ROC convex hull of the ROC curves corresponding to all of these feature sets. A feature set is ROCCHC optimal if its cost equals the current unique cost of the current iteration and its ROC curve has at least one point on the ROC convex hull that has a higher ROC performance than all of the lower cost feature set.

Algorithm 1: ROCCHC Algorithm.

Input: cost of each feature set, ROC curve resulting from each feature set

Output: α : ROCCHC optimal feature set(s)

Obtain the unique costs of the feature sets in ascending order

for *feature set cost in the unique costs* **do**

 Obtain all feature sets with cost less than or equal to the current cost

$\alpha \leftarrow$ Select the feature sets with ROC curves on the ROCCH that are better than lower cost feature sets

end

ROCCHC Wrapper Feature Selection Method

This section describes how wrapper algorithms can use our proposed ROCCHC selection method as their feature set evaluation function. Any type of search wrapper could be used with the ROCCHC selection method; however, only forward and backward sequential search wrappers are demonstrated in this paper. The demonstration of a search wrapper algorithm with our proposed ROCCHC selection method is outlined in Algorithm 2 for forward and backward sequential selection.

Algorithm 2: ROCCHC sequential feature selection wrapper.

Input: data set

Output: S' : selected feature sets

Select starting feature set(s)

while starting or newly selected feature sets remain **do**

 for feature set in remaining feature sets **do**

 for feature in data set **do**

 Add feature if forward selection or remove
 feature if backward selection creating a
 new feature set

 end

 end

 Evaluate the feature sets to obtain ROC curves and
 feature set costs using a machine learning
 algorithm with a cross-validation method

 $S' \leftarrow$ select feature set(s) using the ROCCHC
 selection method, including the currently selected
 feature set(s), S' , and newly added feature set(s)

end

The search algorithm starts by obtaining a starting feature set. The starting set will include the full feature set for backward sequential selection and the empty feature set for forward sequential selection. The starting set can also include other randomized starting feature sets. The algorithm then enters a while loop that runs until there are no newly selected feature sets. Newly selected feature sets are feature sets that were just evaluated and were selected using the ROCCHC selection method. The algorithm iterates through each feature set in the starting feature sets or the newly selected feature sets (the remaining feature sets) and adds (forward) or removes (backward) each feature in the full feature set, thus creating a group of new feature sets to evaluate. The algorithm evaluates each feature set by creating a predictive model using a machine learning algorithm and computing a resulting ROC curve using a cross-validation method. The cost of each feature set is also computed using data set specific cost functions. The computed ROC curves and costs can then be used with the ROCCHC algorithm (1) to select the ROCCHC optimal feature sets.

To better illustrate how this multiple-objective selection is implemented when there can be multiple feature set selections at every step of the algorithm, Fig. 5 shows a detailed example search web using the sequential forward selection method. (A full decision space for the feature selection problem is shown in [24].) This figure shows a search for a problem that has four possible features. Each circle corresponds to a different feature set, where the feature set is represented using a bit string. The string shows a “0” for features not in the feature set and a “1” for features

Fig. 5 Sequential forward search method

in the feature set. Since this example has four features, there are 15 possible feature sets ($2^4 - 1$, the empty feature set is not counted as a feature set). The forward search web starts with no features on the left and generates four feature sets, each with one feature. The new feature sets are then evaluated and selected using the ROCCHC selection method. Feature sets that are not selected or are dominated are shown using a dotted pattern and the selected feature sets are shaded. This process is continued for features sets that are selected until none of the new feature sets are in the selected set. During evaluation of any feature sets, all of the currently selected feature sets are sent into the selection algorithm, which will determine whether they remain selected or are dominated by newer feature sets. At any step in this search process, multiple feature sets can be selected as ROCCHC optimal; the algorithm will branch off from all selected feature sets.

Numerical Experiments and Results

This section outlines the numerical experiments performed to compare the ROCCHC selection method to more common selection methods and demonstrate its use with wrapper feature selection techniques. The ROCCHC selection method was originally developed for the hydraulic actuator problem; however, in order to demonstrate its ability to generalize to other problems, the method is also demonstrated on five data sets that are publicly available on the UCI Machine Learning Repository [15].

UCI Data Sets

Each UCI data set with its number of features, number of possible feature sets, and number of samples is shown

in Table 1. Each data set involves a binary classification problem. The Hepatitis data set includes real feature costs in Canadian dollars attributed to the costs of medical tests. This data set's feature costs also include some group discounts due to shared common costs between particular tests or features. A custom feature set cost calculation function was created using this data set's specific feature cost information. For the other data sets, the number of features in each feature set was used as the feature set cost (e.g., if four features are in the feature set, the cost is 4) due to the lack of specific feature cost information.

Exhaustive Selection Comparison

To compare each feature selection method, Algorithm 3 was used to determine the selected feature sets for a given data set and selection method. The algorithm evaluates every feature set in an exhaustive set of feature combinations and then selects feature set(s) using a particular selection method. For this analysis, the common ROCCHC selection method and a Pareto-based accuracy and feature set cost method (Accuracy&Cost, where accuracy is a percentage calculated as the number of correct predictions over the total number of predictions) are compared with our proposed ROCCHC selection method. The exhaustive search was performed in order to demonstrate that if alternative selection methods (ROCCH and Accuracy&Cost) are used to select feature sets, then there may be a large number of feature sets that would not be selected even though they are ROCCHC optimal. In other words, using ROCCH and Accuracy&Cost would eliminate candidate feature sets that exist on the ROCCHC Pareto Front and should be considered as possible feature sets during system design.

All analyses were conducted in MATLAB 2018a. Prediction models created using the Classification and Regression Trees (CART) machine learning algorithm [11] using default hyperparameters along with 10-fold cross-validation for computing ROC curves and classification accuracy, although any machine learning algorithm and cross-validation method could be used with the evaluated feature selection methods.

Table 1 Data sets

Data set	Features	Feature sets	Samples
Adult	14	16,383	48,842
Breast cancer	9	511	569
German	20	1,048,575	1000
Hepatitis	19	524,287	155
Magic	10	1023	19,020

Algorithm 3: Algorithm to obtain all optimal feature sets for a given selection method, such as ROCCHC, ROCCH, or Accuracy&Cost.

Input: data set
Output: S^* : set of selected feature sets
for every combination of features do
 | Evaluate the feature set
end
 $S^* \leftarrow$ select the feature set(s) using a selection method

The exhaustive selection results are displayed in Table 2, which shows the number of feature sets each selection method chose. The ROCCHC method selected between 0.006 and 5.088% of the total feature combinations, the ROCCH selection method selected between 0.001 and 1.761% of the total feature combinations, and the Accuracy&Cost selection method selected between 0.0009 and 0.783% of the total feature combinations. As expected, the ROCCHC method selects more feature sets because there are more selection metrics (feature set cost, false positive rate, and true positive rate).

An example ROC graph of the selected ROCCHC feature sets is shown in Fig. 6 for the Magic data set. The color of each ROCCHC feature set's ROC curve corresponds to the feature set cost using the figure's colorbar. This figure shows that ROCCHC feature sets have a trend of increasing ROC performance with increasing feature set cost, as expected. ROC graphs for the other UCI data sets show a similar trend.

To show the overlap in the feature sets selected by each selection method, Venn Diagrams are displayed in Fig. 7. This figure includes a Venn Diagram for each data set, with circles for each selection method. Overlaps between the circles demonstrate the ratio of selected feature sets common to each selection method. This figure demonstrates that there are always overlaps in the selected feature sets between the three methods. In the Breast Cancer and Magic data sets, the ROCCHC's selection includes all of the ROCCH and Accuracy&Cost selected feature sets. For these five data sets, the ROCCHC selection included between 56 and 100% of the ROCCH feature sets and between 60

Table 2 Exhaustive ROCCHC, ROCCH, and Accuracy&Cost results. The number of selected feature sets is shown with the percentage of the total number of feature sets in parenthesis (%)

Data set	ROCCHC	ROCCH	Acc&Cost
Adult	52 (0.317)	27 (0.165)	5 (0.031)
Breast cancer	26 (5.088)	9 (1.761)	4 (0.783)
German	62 (0.006)	15 (0.001)	9 (0.0009)
Hepatitis	100 (0.019)	10 (0.002)	10 (0.002)
Magic	15 (1.466)	5 (0.489)	7 (0.684)

Fig. 6 ROCCHC selection results for the Magic data set

and 100% of the Accuracy&Cost selected feature sets. In other words, the ROCCH feature sets missed up to 44% of ROCCHC optimal feature sets and the Accuracy&Cost method missed up to 40% of ROCCHC optimal feature sets.

These Venn Diagrams also point out that the ROCCH and Accuracy&Cost selection methods selected feature sets that are not ROCCHC optimal (indicated by regions of ROCCH and Accuracy&Cost circles outside of ROCCHC circles). This informs us that not only are the ROCCH and Accuracy&Cost methods missing ROCCHC optimal feature sets, but they may select feature sets that are sub-optimal

(i.e., feature sets that are not Pareto optimal in terms of ROC performance and feature set cost or are not ROCCHC optimal).

Sequential Search Feature Selection

After comparing the exhaustive feature selection results for the three selection methods for the five UCI data sets, the next set of experiments aims to demonstrate that the ROCCHC selection method can be used with sequential selection techniques to improve efficiency. The forward and backward sequential search methods are evaluated, but any type of search technique could be used with the ROCCHC evaluation function. Each technique is evaluated based on how many feature sets are selected that were found using the exhaustive search method. These non-exhaustive search methods allow the ROCCHC selection method to be used with data sets with a high number of features where the exhaustive search technique is not feasible.

For these experiments, the starting feature sets were randomly chosen but remained the same for each selection method. Multiple starting feature sets were used so that the algorithms would continue until either all of the ROCCHC optimal feature sets were found or all of feature set combinations were evaluated (to ensure that the search would not just stop when there were no newly selected feature sets). Each feature set was evaluated using the same prediction models from the exhaustive experiments (CART machine learning algorithm with default hyperparameters with 10-fold cross-validation) and using the same metrics such as accuracy, feature set cost, and ROC performance. The common ROCCH and Accuracy&Cost selection

Fig. 7 Venn Diagrams showing the intersection between the ROCCHC, ROCCH, and Accuracy&Cost selected feature sets for each data set

methods were also used with 2 for a comparison of the selection methods when using a search algorithm.

The sequential search results are shown in Fig. 8. This figure shows the forward and backward results for each of the five data sets using the ROCCHC, ROCCH, and Accuracy&Cost selection methods. The x -axis of each plot is the percentage of feature sets evaluated, where 100% would be all feature combinations ($2^n - 1$, where n is the data set's number of features). The y -axis of each plot is the proportion of ROCCHC selected feature sets that the search has found, based on the exhaustive results for each data set. An optimal search method would find all of the ROCCHC optimal feature sets while evaluating a minimum number of feature sets, and thus produce a nearly vertical line starting at the plot's origin.

Figure 8 demonstrates that the ROCCH and Accuracy&Cost selection methods are not able to find all of the ROCCHC optimal feature sets, as expected from the exhaustive results. Therefore, a search algorithm is only as good as its selection criteria. Without knowing that a feature set is ROCCHC optimal, the ROCCH and Accuracy&Cost selection methods will miss these ROCCHC optimal feature sets. Furthermore, the ROCCHC method is often more efficient than the alternative methods, as ROCCHC tends to find more ROCCHC optimal feature sets while evaluating less

Table 3 Exhaustive ROCCHC, ROCCH, and Accuracy&Cost results for the Actuator data set

Data set	ROCCHC	ROCCH	Acc&Cost
Actuator	37 (0.113)	8 (0.024)	8 (0.024)

The number of selected feature sets is shown with the percentage of the total number of feature sets in parenthesis

possible feature sets. The proposed ROCCHC method is, therefore, a better selection criteria for wrapper algorithms involving binary problems, especially for real-world problems where class proportions and misclassification costs are unknown.

Actuator Results

The proposed ROCCHC selection method was also demonstrated on a data set from the hydraulic actuator test rig. This data set has 15 features, 32,767 possible feature sets, and 2269 samples. This data set is also a binary problem of trying to predict if the actuator is in a normal operation state or in a failed state. Prediction models were

Fig. 8 Search results for each data set

Fig. 9 Venn Diagrams showing the intersection between the ROCCHC, ROCCH, and Accuracy&Cost selected feature sets for the Actuator data set

again created using the CART machine learning algorithm with default hyperparameters with 10-fold cross-validation.

The exhaustive feature selection results for the Accuracy&Cost, ROCCH, and ROCCHC selection methods are shown in Table 3, which includes the number and percentage (%) of features sets selected by each method. The ROCCHC method selects the most feature sets with 0.113% and the ROCCH and Accuracy&Cost methods both select 0.024%. Although the ROCCHC method chooses almost 4 times as many feature sets, 37 feature sets is still a small number to evaluate for final selection manually. The Venn Diagram in Fig. 9 shows that all of the ROCCH and Accuracy&Cost feature sets are also ROCCHC optimal; however,

Fig. 10 ROCCHC selections for the Actuator data set

Fig. 11 Search results for the Actuator data set

both methods only choose 22% of the ROCCHC optimal feature sets or miss 78%. Again, this demonstrates that the Accuracy&Cost and ROCCH methods are missing optimal feature sets.

An ROC graph of the ROCCHC selected feature sets is also provided in Fig. 10 for the Actuator data set. This graph demonstrates that the highest cost ROCCHC feature set only contains 11 out of the 15 features (based on the max. feature set cost). This ROC graph also demonstrates that the ROC curves for the higher performing feature sets are very close to one another. The Actuator sequential search results are shown in Fig. 11, which demonstrates again that the ROCCH and Accuracy&Cost selection methods are not able to find all of the ROCCHC optimal feature sets using search methods, as expected from the exhaustive results.

Conclusion and Future Work

This study proposes the use of ROCCHC as an evaluation function for sequential feature selection methods. The proposed evaluation function is demonstrated using sequential forward and backward search techniques but could be used with any type of search function when implementing wrapper feature selection algorithms. The ROCCHC selection method was compared with two existing methods on common binary classification data sets and a hydraulic actuator data set. The Accuracy&Cost and ROCCH selection methods miss ROCCHC optimal feature sets for each of the data sets. However, for real-world problems, where actual class proportions and misclassification costs are unknown, these

ROCCHC optimal data sets should not be excluded during the design of CPS.

It was also demonstrated that the Accuracy&Cost and ROCCH selection methods are choosing feature sets that are not ROCCHC optimal. These sub-optimal feature sets are dominated in ROCCHC space in terms of ROC performance and feature set costs. These sub-optimal feature sets can only be discovered after using the ROCCHC selection method. Furthermore, the search results demonstrate that a search method can find all of the ROCCHC optimal feature sets without having to evaluate all of the possible feature sets.

The Venn Diagram plots presented in this paper are a useful tool for comparing feature set selection methods. They quickly show the intersection and proportional size of feature sets selected from the three analyzed selection methods. They showed that for some data sets, there may be Accuracy&Cost and ROCCH selected feature sets that are not ROCCHC optimal. We propose that these plots be used for comparing feature set selection methods in the future.

Future work will aim to perform further comparisons between other common selection methods, such as area under the ROC curve. Future work could also extend the ROCCHC selection method to multi-class problems. The ROCCHC method could also be used with more advanced search algorithms such as artificial bee colony optimization [18], genetic algorithms [37], and particle swarm optimization [49].

Funding This material is based upon work supported by the Naval Sea Systems Command under Contract N00024-17-C-4008.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. S. Adams, P.A. Beling, A survey of feature selection methods for Gaussian mixture models and hidden Markov models. *Artif. Intell. Rev.* 1–41 (2017)
2. S. Adams, P.A. Beling, R. Cogill, Feature selection for hidden Markov models and hidden semi-Markov models. *IEEE Access* **4**, 1642–1657 (2016)
3. S. Adams, P.A. Beling, K. Farinholt, N. Brown, S. Polter, Q. Dong, Condition based monitoring for a hydraulic actuator. in *Proceedings of the Annual Conference of the Prognostics and Health Management Society*, (2016)
4. S. Adams, R. Meekins, P.A. Beling, An empirical evaluation of techniques for feature selection with cost. in *2017 IEEE International conference on Data Mining Workshops (ICDMW)* (IEEE, 2017), pp. 834–841
5. S. Adams, R. Meekins, P.A. Beling, K. Farinholt, N. Brown, S. Polter, Q. Dong, A comparison of feature selection and feature extraction techniques for condition monitoring of a hydraulic actuator. in *Annual Conference of the Prognostics and Health Management Society, st. petersburg, FL*, (2017)
6. S. Adams, R. Meekins, P.A. Beling, K. Farinholt, N. Brown, S. Polter, Q. Dong, Hierarchical fault classification for resource constrained systems. *Mech. Syst. Signal Process.* **134**, 106266 (2019)
7. D.W. Aha, R.L. Bankert, A comparative evaluation of sequential feature selection algorithms. in *Learning from Data* (Springer, 1996), pp. 199–206
8. H. Almuallim, T.G. Dietterich, Learning with many irrelevant features. in *AAAI*, Vol. 91 (Citeseer, 1991), pp. 547–552
9. P. Beling, Z. Covaliu, R. Oliver, Optimal scoring cutoff policies and efficient frontiers. *J. Oper. Res. Soc.* **56**(9), 1016–1029 (2005)
10. V. Bolón-Canedo, I. Porto-díaz, N. Sánchez-Marono, A. Alonso-Betanzos, A framework for cost-based feature selection. *Pattern Recognit.* **47**(7), 2481–2489 (2014)
11. L. Breiman, J. Friedman, R. Olshen, C. Stone, Classification and regression trees. Wadsworth, Belmont, CA (1984)
12. R. Caruana, D. Freitag, Greedy attribute selection. in *Machine Learning Proceedings 1994* (Elsevier, 1994), pp. 28–36
13. X. Chai, L. Deng, Q. Yang, C.X. Ling, Test-cost sensitive naive Bayes classification. in *Fourth IEEE International Conference on Data Mining (ICDM'04)* (IEEE, 2004), pp. 51–58
14. M. Dash, H. Liu, Feature selection for classification. *Intell. Data Anal.* **1**(1–4), 131–156 (1997)
15. D. Dua, C. Graff, UCI machine learning repository. <http://archive.ics.uci.edu/ml> (2017)
16. K.M. Farinholt, A. Chaudhry, M. Kim, E. Thompson, N. Hipwell, R. Meekins, S. Adams, P. Beling, S. Polter, Developing health management strategies using power constrained hardware. in *Proceedings of the Annual Conference of the PHM Society*, Vol. 10, (2018)
17. I. Guyon, A. Elisseeff, An introduction to variable and feature selection. *J. Mach. Learn. Res.* **3**(Mar), 1157–1182 (2003)
18. E. Hancer, B. Xue, M. Zhang, D. Karaboga, B. Akay, Pareto front feature selection based on artificial bee colony optimization. *Inform. Sci.* **422**, 462–479 (2018)
19. M.M. Herterich, F. Uebernickel, W. Brenner, The impact of cyber-physical systems on industrial services in manufacturing. *Procedia Cirp* **30**, 323–328 (2015)
20. K. Iswandy, A. Koenig, Feature selection with acquisition cost for optimizing sensor system design. *Adv. Radio Sci.* **4**(C. 1), 135–141 (2006)
21. N. Jazdi, Cyber physical systems in the context of industry 4.0. in *2014 IEEE International Conference on Automation, Quality and Testing, Robotics* (IEEE, 2014), pp. 1–4
22. P.W. Kalgren, C.S. Byington, M.J. Roemer, M.J. Watson, Defining PHM, a lexical evolution of maintenance and logistics. in *2006 IEEE Autotestcon* (IEEE, 2006), pp. 353–358
23. K. Kira, L.A. Rendell et al., The feature selection problem: Traditional methods and a new algorithm. in *Aaai*, Vol. 2, (1992), pp. 129–134
24. R. Kohavi, G.H. John, Wrappers for feature subset selection. *Artif. Intell.* **97**(1–2), 273–324 (1997)
25. G. Kong, L. Jiang, C. Li, Beyond accuracy: learning selective Bayesian classifiers with minimal test cost. *Pattern Recogn. Lett.* **80**, 165–171 (2016)

26. I. Kononenko, Estimating attributes: analysis and extensions of relief. in *European Conference on Machine Learning* (Springer, 1994), pp. 171–182
27. I. Lee, O. Sokolsky, S. Chen, J. Hatcliff, E. Jee, B. Kim, A. King, M. Mullen-Fortino, S. Park, A. Roederer, et al., Challenges and research directions in medical cyber–physical systems. *Proc. IEEE* **100**(1), 75–90 (2011)
28. J. Lee, B. Bagheri, H.A. Kao, A cyber-physical systems architecture for industry 4.0-based manufacturing systems. *Manuf. Lett.* **3**, 18–23 (2015)
29. J. Lee, F. Wu, W. Zhao, M. Ghaffari, L. Liao, D. Siegel, Prognostics and health management design for rotary machinery systems—reviews, methodology and applications. *Mech. Syst. Signal Process.* **42**(1–2), 314–334 (2014)
30. P. Leitao, S. Karnouskos, L. Ribeiro, J. Lee, T. Strasser, A.W. Colombo, Smart agents in industrial cyber–physical systems. *Proc. IEEE* **104**(5), 1086–1101 (2016)
31. C.X. Ling, V.S. Sheng, Cost-sensitive learning. in *Encyclopedia of Machine Learning*, (2010), pp. 231–235
32. C.X. Ling, Q. Yang, J. Wang, S. Zhang, Decision trees with minimal costs. in *Proceedings of the Twenty-First International Conference on Machine Learning*, Vol. 69 (ACM, 2004)
33. R. Meekins, S. Adams, P.A. Beling, K. Farinholt, N. Hipwell, A. Chaudhry, S. Polter, Q. Dong, Cost-sensitive classifier selection when there is additional cost information. in *International Workshop on Cost-Sensitive Learning*, (2018), pp. 17–30
34. F. Min, Q. Hu, W. Zhu, Feature selection with test cost constraint. *Int. J. Approx. Reason.* **55**(1), 167–179 (2014)
35. F. Min, W. Zhu, Minimal cost attribute reduction through backtracking. in *Database Theory And Application, Bio-Science and Bio-Technology* (Springer, 2011), pp. 100–107
36. L. Monostori, B. Kádár, T. Bauernhansl, S. Kondoh, S. Kumara, G. Reinhart, O. Sauer, G. Schuh, W. Sihn, K. Ueda, Cyber-physical systems in manufacturing. *Cirp Annals* **65**(2), 621–641 (2016)
37. S. Oreski, G. Oreski, Genetic algorithm-based heuristic for feature selection in credit risk assessment. *Expert Syst. Appl.* **41**(4), 2052–2064 (2014). <https://doi.org/10.1016/J.ESWA.2013.09.004>
38. P. Paclík, R.P. Duin, G.M. van Kempen, R. Kohlus, On feature selection with measurement cost and grouped features. in *Joint IAPR international workshops on statistical techniques in pattern recognition (SPR) and structural and syntactic pattern recognition (SSPR)* (Springer, 2002), pp. 461–469
39. Y. Peng, M. Dong, M.J. Zuo, Current status of machine prognostics in condition-based maintenance: a review. *Int. J. Adv. Manuf. Technol.* **50**(1–4), 297–313 (2010)
40. F.J. Provost, T. Fawcett et al., Analysis and visualization of classifier performance: comparison under imprecise class and cost distributions. in *KDD*, Vol. 97, (1997), pp. 43–48
41. P. Pudil, J. Novovičová, J. Kittler, Floating search methods in feature selection. *Pattern Recognit. Lett.* **15**(11), 1119–1125 (1994)
42. R. Rajkumar, I. Lee, L. Sha, J. Stankovic, Cyber-physical systems: the next computing revolution. in *Design Automation Conference* (IEEE, 2010), pp. 731–736
43. Y. Saeys, I. Inza, P. Larrañaga, A review of feature selection techniques in bioinformatics. *bioinformatics* **23**(19), 2507–2517 (2007)
44. K. Sampigethaya, R. Poovendran, Aviation cyber–physical systems: foundations for future aircraft and air transport. *Proc. IEEE* **101**(8), 1834–1855 (2013)
45. P. Somol, P. Pudil, J. Novovičová, P. Paclík, Adaptive floating search methods in feature selection. *Pattern Recognit. Lett.* **20**(11–13), 1157–1163 (1999)
46. S. Uckun, K. Goebel, P.J. Lucas, Standardizing research methods for prognostics. in *2008 International Conference on Prognostics and Health Management* (IEEE, 2008), pp. 1–10
47. G. Xiong, F. Zhu, X. Liu, X. Dong, W. Huang, S. Chen, K. Zhao, Cyber-physical-social system in intelligent transportation. *IEEE/CAA J. Automatica Sinica* **2**(3), 320–333 (2015)
48. B. Xue, M. Zhang, W.N. Browne, Particle swarm optimization for feature selection in classification: a multi-objective approach. *IEEE Trans. Cybern.* **43**(6), 1656–1671 (2012)
49. B. Xue, M. Zhang, W.N. Browne, Particle swarm optimization for feature selection in classification: a multi-objective approach. *IEEE Trans. Cybern.* **43**(6), 1656–1671 (2013). <https://doi.org/10.1109/TSMCB.2012.2227469>. <http://ieeexplore.ieee.org/document/6381531/>
50. Y. Yang, J.O. Pedersen, A comparative study on feature selection in text categorization. in *Icml*, Vol. 97, (1997), p. 35
51. Y. Jin, B. Sendhoff, Pareto-based multiobjective machine learning: an overview and case studies. *IEEE Trans. Syst. Man Cybern. B* **38**(3), 397–415 (2008). <https://doi.org/10.1109/TSMCC.2008.919172>. <http://ieeexplore.ieee.org/document/4492360/>
52. Q. Zhou, H. Zhou, T. Li, Cost-sensitive feature selection using random forest: selecting low-cost subsets of informative features. *Knowl.-Based Syst.* **95**, 1–11 (2016)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.