
This item was submitted to [Loughborough's Research Repository](#) by the author.
Items in Figshare are protected by copyright, with all rights reserved, unless otherwise indicated.

All versus one: An empirical comparison on retrained and incremental machine learning for modeling performance of adaptable software

PLEASE CITE THE PUBLISHED VERSION

PUBLISHER

IEEE

VERSION

AM (Accepted Manuscript)

PUBLISHER STATEMENT

© 2019 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

LICENCE

All Rights Reserved

REPOSITORY RECORD

Chen, Tao. 2019. "All Versus One: An Empirical Comparison on Retrained and Incremental Machine Learning for Modeling Performance of Adaptable Software". figshare. <https://hdl.handle.net/2134/9876320.v1>.

All Versus One: An Empirical Comparison on Retrained and Incremental Machine Learning for Modeling Performance of Adaptable Software

Tao Chen

Department of Computing and Technology

Nottingham Trent University, UK

txc919@gmail.com

Abstract—Given the ever-increasing complexity of adaptable software systems and their commonly hidden internal information (e.g., software runs in the public cloud), machine learning based performance modeling has gained momentum for evaluating, understanding and predicting software performance, which facilitates better informed self-adaptations. As performance data accumulates during the run of the software, updating the performance models becomes necessary. To this end, there are two conventional modeling methods: the retrained modeling that always discard the old model and retrain a new one using *all* available data; or the incremental modeling that retains the existing model and tunes it using *one* newly arrival data sample. Generally, literature on machine learning based performance modeling for adaptable software chooses either of those methods according to a general belief, but they provide insufficient evidences or references to justify their choice. This paper is the first to report on a comprehensive empirical study that examines both modeling methods under distinct domains of adaptable software, 5 performance indicators, 8 learning algorithms and settings, covering a total of 1,360 different conditions. Our findings challenge the general belief, which is shown to be only partially correct, and reveal some of the important, statistically significant factors that are often overlooked in existing work, providing evidence-based insights on the choice.

Index Terms—Performance modeling, self-adaptive system, machine learning, software runtime

I. INTRODUCTION

Predicting the performance of adaptable software systems can serve as the powerful foundation for reasoning in various tasks, e.g., anomaly detection [1], resource provisioning [2] and self-adaptation at runtime [3]. The key challenge of this is to build an effective performance model that takes the relevant features¹ of the adaptable software as inputs (e.g., number of threads, used cache and utilized memory) and predict a value of the performance indicator, such as latency, throughput and reliability.

While different classic approaches (e.g., analytical model [4][5] and simulation [6]) exist for modeling the performance of adaptable software, machine learning based performance modelings have been gaining momentum due to the following reasons: (i) the ever-increasing complexity of modern adaptable software made the exploitation of classic

approaches for modeling performance difficult, as they are often restricted to certain features and scenarios that are obtained via heavy human analysis. (ii) The effectiveness of classic approaches is highly depending on their assumptions about the internal structure of the adaptable software being modeled, and the environment. However, many modern environments, such as cloud-based systems, virtualized and multi-tenant software, intentionally hide such information to promote ease of use, which further reduce the reliability of those approaches. In contrast, without heavy human intervention, machine learning based modeling takes a black box manner that relies on observing the system’s actual behaviors under certain conditions, in order to infer a statistical model for the concerned performance indicator [7].

One fundamental to effective application of machine learning in performance modeling is the data, which determines the levels of knowledge that a model can learn and generalize. However, many real world scenarios do not have sufficient data, or the available data do not adequately represent what the adaptable software is likely to behave in changing and uncertain environments. Therefore, modeling software performance at runtime with evolving data stream has been increasingly important [8] [9]. Machine learning based performance modeling at runtime has the advantage that the model can be updated using the most up-to-date data samples, which inherently improves the effectiveness of the model.

For modeling performance at runtime, the problem that a software engineer would face is: how to update the model when using a learning algorithm² under evolving data? Literature from the Software Engineering and Machine Learning communities take two predominate *modeling methods* to achieve this: (i) either completely retraining the model by learning a new data sample in conjunction with the historical ones (i.e., the retrained modeling), or (ii) simply tuning the existing model using a new data sample as it arrives (i.e., the incremental modeling). The choice between those two methods does not change the interpretation of the model, but they make fundamentally different assumptions about how a model is learned and hence they lead to different variants of a learning

¹This should not be confused with the functionality of a software; it merely refers to the quantifiable properties of the software being measured.

²A learning algorithm refers to a particular model structure and the related algorithmic procedure that builds the model.

algorithm [10] [11]. As a result, an inappropriate choice can have serious impacts to the accuracy and training time of the model which could violate the requirements. However, existing papers take either of those modeling methods without providing sufficient evidence or references to justify the choice. Instead, a general belief is often implied:

Incremental modeling is chosen for faster training [12] [13] [14] [8] while the retrained modeling is chosen when higher accuracy is preferred [15] [16] [17] [18] [9] [19] [20] [7] [14]. The choice is a trade-off between accuracy and training time.

In this paper, we address such an important lack of understanding in machine learning based performance modeling through comprehensive empirical study on three real-world adaptable software under various settings, leading to a total of 1,360 different conditions (as explained in Table III). We show that the general belief is flawed and inaccurate. Particularly, we answer the following research questions:

- **RQ1:** Does the retrained version of a given learning algorithm always make more accurate model than its incremental counterparts when modeling adaptable software?

No it does not, the incremental modeling can achieve statistically better accuracy under certain learning algorithms, the adaptable software and the fluctuations of the obtained data, which is clearly contradict to what the general belief claims.

- **RQ2:** Does the incremental version of a given learning algorithm constantly leads to faster training than its retrained counterparts when modeling adaptable software?

Yes it does, as the general belief stated. However, the gain on training time may be practically trivial.

- **RQ3:** When choosing modeling methods considering different learning algorithms, do the trade-offs between accuracy and training time for modeling performance of adaptable software always needed?

Trade-off is indeed required, in which the incremental modeling could train faster but with worse accuracy. However, this is not always the case—it is possible that the incremental modeling achieves the best for both properties. Therefore, the general belief is inaccurate.

- **RQ4:** How the modeling methods can be affected by the runtime fluctuations of the adaptable software, i.e., the number of concept drifts and the deviations in the data?

The errors of both modeling methods exhibit considerably positive monotonic correlations to the number of drifts, and non-trivial negative monotonic correlations to the deviations of data. We did not observe clear correlations of their training time to the number of concept drift and data deviations in general. The only exception is the strong correlation between training time of incremental modeling and the number of concept drift.

In the following, we present the background, motivation and prior work in Section II. We then specify our empirical study methodology in Section III. The results are demonstrated and analyzed in Section IV, followed by discussions on the key lessons learned in Section V. Finally, we discuss threats to validity and conclusions in Section VI and VII, respectively.

II. BACKGROUND, RELATED WORK AND MOTIVATION

A. Learning Performance Model for Adaptable Software

Software performance model refers to a correlation function that takes the features of interest in adaptable software as inputs and outputs the expected performance quality. In general, those features fall into one of the following categories:

- **Environmental features affecting the performance.** These features refer to those that cannot be controlled by the software itself, which causes dynamics and uncertainties, e.g., the workload, order of requests, size of jobs.
- **Adaptable features affecting the performance.** These features refer to those that can be adjusted, at design time or runtime, to influence the adaptable software, e.g., cache mode, used cache size and number of threads.

Given the time-varying environment in which the software operates, the performance model of software is inherently temporal and could take the historical data points as input features, which can be formally defined as:

$$Q(t+1) = f\left(\begin{bmatrix} c_1(t+1) & c_1(t) & \dots & c_1(2) \\ c_2(t+1) & c_2(t) & \dots & c_2(2) \\ \vdots & \vdots & \ddots & \vdots \\ c_n(t+1) & c_n(t) & \dots & c_n(2) \end{bmatrix}, \begin{bmatrix} e_1(t) & e_1(t-1) & \dots & e_1(1) \\ e_2(t) & e_2(t-1) & \dots & e_2(1) \\ \vdots & \vdots & \ddots & \vdots \\ e_m(t) & e_m(t-1) & \dots & e_m(1) \end{bmatrix}\right) \quad (1)$$

where $c_1(t+1), \dots, c_n(t+1)$ are the values of adaptable features for time point $t+1$; $e_1(t), \dots, e_m(t)$ are the values of environment features observed at time point t . Those features can be either continuous, e.g., CPU, memory and size of thread pool, or discrete, e.g., cache mode and compression method. f is the actual function learned by a machine learning algorithm. $Q(t+1)$ is the expected performance indicator given the values of adaptable features up to $t+1$ and under the values of environment features up to t .

Since the most recent data points are often more relevant than the order ones [8] [9], the generic performance model can be simplified as:

$$Q(t+1) = f\left(\begin{bmatrix} c_1(t+1) \\ c_2(t+1) \\ \vdots \\ c_n(t+1) \end{bmatrix}, \begin{bmatrix} e_1(t) \\ e_2(t) \\ \vdots \\ e_m(t) \end{bmatrix}\right) \quad (2)$$

The performance modeling is often regarded as a *regression* problem as the performance indicators are usually continuous values, e.g., throughput. However, it is not difficult to build the model as *classification*, where the output is a discrete value, e.g., for security levels. In this work, we focus on the performance indicators that demands regression as they are arguably the most popular ones in existing work [21][22].

In machine learning terminology, the process of building this model is called *training*. The ultimate goal of a learning algorithm is to minimize the error for the data samples in the training, while preserving the ability to generalize the concept learned in the training stage to predict the performance on given inputs, which may not be seen during training.

B. The Retrained and Incremental Modeling

Modeling the performance of adaptable software via machine learning often require the model to learn whenever newly observed data sample becomes available as the software runs. However, the problem that a software engineer would face is: how to update the model when using machine learning under evolving data? According to the literature from both the Software Engineering and the Machine Learning community, there are two predominate *modeling methods* to achieve this:

Retrained modeling: retrained modeling is similar to the traditional offline learning, where the old model is discarded and a new model is retrained using whatever data that is available, i.e., the new data samples and all the historical ones. The good side of retrained modeling is that it is able to capture the interrelation between different data samples given the fact that they are always learned in conjunction with each others. Henceforth, it is expected to be more accurate as it has more knowledge about the correlation of data. Note that it is possible to retrain the model only upon a batch of samples have been obtained, in which case the size of batch is customizable. However, different sizes may impose different impacts on the accuracy of the model and may also be highly sensitive to different contexts [11]. Therefore, in this work, we use the extreme case as a representative, where the model is retrained upon the arrival of one new sample, together with all the historical data.

Incremental modeling: incremental modeling follows the online learning paradigm, which is truly incremental in the sense that instead of replacing the entire model, its internal structure is tuned using the new data sample. In other words, it learns each new data sample in isolation as they arrive. The good side of incremental modeling is the likely small computation effort. However, the fact that each data sample is learned individually may ignore some joint correlations that can only be discovered when data samples are learned in conjunction with each others, which may affect the accuracy.

Algorithm 1 Retrained MLP

```

Require: collected data  $D_{set} = \emptyset$ 
1: while new data  $D_{new}$  arrives do
2:   discard the old model
3:   initialize a new model  $M$ 
4:    $D_{set} := D_{set} \cup D_{new}$ 
5:   for  $i = 1$  to  $epoch_{max}$  do
6:     for sample  $d$  in  $D_{set}$  do
7:       predict output on  $d$ 
8:       get the error  $e$ 
9:       get  $\Delta w_h$  (hidden)
10:      get  $\Delta w_i$  (input)
11:     end for
12:     update all weights in  $M$ 
13:   end for
14: end while

```

Algorithm 2 Incremental MLP

```

1: initialize a new model  $M$ 
2: while a new data sample  $d$  arrives do
3:   for  $i = 1$  to  $epoch_{max}$  do
4:     predict output on  $d$ 
5:     get the error  $e$ 
6:     get  $\Delta w_h$  (hidden)
7:     get  $\Delta w_i$  (input)
8:     update all weights in  $M$ 
9:   end for
10: end while

```

In algorithm 1 and 2, we used the three layered Multi-Layer Perceptron (MLP) [23], trained by back-propagation [24], to

algorithmically illustrate the difference of the two modeling methods. Here the key distinction between them is the order in which the weights are updated. In retrained modeling, the weights are updated once all data samples are presented to the learning algorithm; subsequently, the updated model is used in the next iteration, which ends when a fix number of epochs has been reached. On contrary, in incremental modeling, the weights are updated w.r.t. each single data sample for a fix number of epochs. Even if we assume identical number of data samples for both modeling methods, the different order of weights updating would create different intermediate model, which serves as the base for updating weights in the next iterations, leading to diverse finally trained model. As a result, deciding on which modeling method to follow is non-trivial for performance modeling on adaptable software.

C. Prior Retrained Performance Modeling

To build machine learning based performance models under evolving data stream, a large amount of research has relied on retrained modeling. Among others, Kundu et al. [15][16] have relied on Multi-Layer Perceptron (MLP) [10] and Support Vector Machine (SVM) [25] to model the performance of cloud-based and service-oriented software. Their models are built in the retrained manner, where certain amount of historical data is used to train the MLP model at design time, then at runtime, such a model is retrained whenever new data sample is available. Similarly, Siegmund et al. [20], Sieber et al. [17] and Gerostathopoulos et al. [26] use Linear Regression (LR) [27] to build the performance model at runtime, but again, the model is retrained completely instead of being tuned when significant outliers are detected or as new data is collected. Another notable effort of retrained modeling based on the Decision Tree (DT) family (e.g., M5 decision tree [28]), such as FUSION [18] and Guo et al. [19], where the performance model is discarded and rebuilt using all the available data when the adaptable software collects new information. A general framework for modeling performance of adaptable software using the retrained method, which is agnostic to the learning algorithm, were proposed by Ghahremani et al. [29].

Didona et al. [7] propose to model performance through a hybrid of analytical and machine learning based modeling where the model is still updated in a retrained manner, but such a training is guided or combined with domain knowledge. To more accurately model the performance of adaptable software in the cloud, Chen and Bahsoon [9][30] also build an ensemble of machine learning models using retrained modeling method.

D. Prior Incremental Performance Modeling

The other direction of effort on performance modeling assumes truly incremental modeling. For example, incremental modeling has been used in relatively simpler learning algorithms, e.g., linear regression (e.g., in [12][14]) and ARMA (e.g., in [13]), when modeling performance under changing environment of an adaptable software. The linear nature of those models make incremental modeling much more straightforward and can be tuned using Recursive Least Squares (RLS)

TABLE I: The Selected Performance Modeling Studies

Ref.	Method	Algorithm	Why Chose the Modeling Method?
[15]	Retrained	MLP,SVM	Great accuracy; acceptable overhead.
[16]	Retrained	MLP	Improved accuracy.
[19]	Retrained	DT	High accuracy.
[20]	Retrained	LR	High accuracy.
[12]	Incremental	LR	Negligible training time.
[8]	Incremental	MLP	Small overhead.
[17]	Retrained	LR	High accuracy to handle outliers.
[18]	Retrained	DT	High accuracy.
[9]	Retrained	LR,DT,MLP	Great accuracy; acceptable overhead.
[7]	Retrained	DT,MLP	High accuracy to extract information.
[14]	Incremental	LR	Fast model updates.
[30]	Retrained	LR,MLP	Great accuracy; acceptable overhead.
[13]	Incremental	ARMA	Negligible training time.

filter [31]. Complex models can also be tuned in an incremental manner: Lama et al. [8] model software performance via a fuzzy based MLP, which is tuned incrementally using back-propagation and RLS filter when new data is available. Recently, Jamshidi et al. [32] apply transfer learning to model the performance of adaptable software, where the transferred model is also incremental, such that it is continually updated as new samples arrives.

E. Prior Comparative Study

In the machine learning community, Read et al.’s study [11] is the most close work that aims to understand the differences between retrained and incremental modeling when handling evolving data, which has similar motivation to ours. However, their work is fundamentally different to this paper in the following aspects: (i) Read et al. focus on classification problems only where our work emphasizes on regression problems. (ii) Read et al. consider learning algorithms that can only be used in a retrained manner as well as those that can be applicable only in an incremental way. In contrast, our work compares an identical learning algorithm under both modeling methods, which eliminates the bias introduced by the nature of different learning algorithms. (iii) We particularly focus on data collected from running adaptable software while Read et al. use data set from other domains, which are irrelevant to software performance modeling. This is important as the data collected adaptable software can exhibit patterns that are difficult to observe in data sets from the other domains [9].

F. Why A Modeling Method was Chosen?

The discussed related work above is the result of a systematic literature review conducted in 2017, in which we searched on Google Scholar using the keywords of “Performance Modeling” AND “Machine Learning” AND (“Software Engineering” OR “Self-Adaptive System”) AND (“Incremental Model” OR “Retrained Model”). We did not only cover papers in the software engineering domain, but also work from the system engineering community. We then applied the exclusion criteria, e.g., published in the last decade, explicitly or implicitly state which modeling method was chosen. As a result, we have identified 13 most notable papers, which, together with why a modeling method was chosen, are listed in Table I. The key observation from the reviewed papers is that, regardless if incremental or retrained modeling is followed, the choice of modeling method comes with insufficient evidence

TABLE II: The Studied Learning Algorithms

Learning Algorithm	Characteristics	Setting
Linear Regression (LR) [27]	linear, interpretable	$\alpha = 0.1, \lambda = 1$
Decision Tree (DT) [33]	nonlinear, interpretable	
Support Vector Machine (SVM) [25]	nonlinear with kernel, convex, black-box	$\alpha = 0.01, \lambda = 10^{-5}$
Multi-Layer Perceptron (MLP) [10]	nonlinear, black-box	Sigmoid, $\alpha = 0.6, epoch = 5000$
Bagging-LR (Ba-LR) [34]	linear, weighted ensemble	$n=5$
Bagging-DT (Ba-DT) [34]	nonlinear, weighted ensemble	$n = 5$
Boosting-LR (Bo-LR) [35]	linear, sequential ensemble	$n = 5$
Boosting-DT (Bo-DT) [35]	nonlinear, sequential ensemble	$n = 5$

or references of justification. Often, the choices are derived according to the general belief presented in Section I.

The results of the review have raised serious concerns that question the validity of the general belief, which in turn, motivates this work to provide insights for choosing modeling method for adaptable software under evolving data stream, especially at runtime.

III. THE EMPIRICAL STUDY METHODOLOGY

A. The Machine Learning Algorithms Studied

A list of 8 widely-used machine learning algorithms, their characteristics and settings³ have been shown in Table II. For each algorithm, we study their incremental version and the retrained counterparts. We have omitted those lazy learning algorithms (e.g., k NN), which do not differ in the sense of retrained and incremental modeling, as they do not build model in the training stage but delay the learning during prediction.

As shown in Table II, to enrich the generality of our study, we have considered different learning algorithms with diverse characteristics. It is worth noting that each learning algorithm is designed for general propose regardless if it is used as incremental or retrained modeling. However, to fit with the context of different modeling methods, they can be tailored during the training process as shown in Section II-B. Further, the studied machine learning algorithms contain 4 single learners and 4 ensembles, in which more than one base learners are used to learn the correlations and their outputs are combined. We have considered two most common ensembles, namely Bagging [34] and Boosting [35]. In this way, our empirical study covers not only the single learning algorithm, but also the algorithms in which more than one learners are combined. We have used the implementation of those algorithms in the WEKA [36] and MOA [37] frameworks.

B. The Subject Adaptable Software Systems

We conducted an empirical study on three running adaptable software from different domains, each of which represents a common category of real-world software, i.e., web-based, service-oriented and cloud-based. They are explained as below and their details are summarized in Table III:

S-RUBiS: RUBiS [39] is a well-known software benchmark (with 26 different services). As shown in Figure 1, to realize an adaptable complex software system, we extended

³The settings of these parameters were identified recurring to the standard methodology used to tune a learning algorithms. The chosen ones appear to be adequate overall for all the adaptable software studied.

TABLE III: Subject Adaptable Software Systems and Their Characteristics

	<i>S-RUBiS</i>	<i>ASOS</i>	<i>C-VARD</i>
# Adaptable Features	10, e.g., the number of threads, cache mode, if enable compression.	10, the number of parallel and redundant instances of a service for each abstract service. (1 to 10)	2, the utilized CPU and used memory.
# Environmental Features	26, number of requests for each services, e.g., <i>BrowseCategory</i> , <i>Browse</i> .	20, the status of each end-user, i.e., indicating whether s/he is requesting the workflow or not.	1, the hour of a day.
# Performance Indicators	2, response time and energy consumptions.	2, reliability (the percentage of end-users who are served within a time limit) and the mean throughput of all end-users.	1, latency.
# Cases	72, a combination of 6 workload traces (<i>FIFA98</i> [38], <i>stable</i> , as in Figure 2a), 2 workload patterns (<i>read-only</i> and <i>read-write</i>) and 6 workload frequencies (amplified by a factor).	10, different matrices of randomly extracted services and end-users. (as in Figure 2b)	6, instance types in Figure 2c

The 1,360 conditions are derived from: 8 learning algorithms \times (2 indicators \times (72 cases + 10 cases) + 1 indicator \times 6 cases).

RUBiS, denoted as S-RUBiS, using various adaptable real-world software, e.g., Tomcat [40]. The feature inputs and the performance indicators are discussed in Table III. We run the software on a dedicated server, and we used Xen as the hypervisor to create a virtualized environment. While S-RUBiS is running in a configurable guest Virtual Machine (VM), we exploited a multi-objective optimization framework [3] as the adaptation engine in root domain to adapt RUBiS at runtime. To create a realistic workload within the capacity of our testbed, we vary the number of clients according to the different workload traces, as shown in Figure 2a. Those workload traces can generate up to 600 parallel requests, which are sufficiently significant for our experiments. The workload is produced by another machine using the client emulator provided in the original RUBiS. Notably, through different workload patterns, traces and frequencies, we obtained a total of 72 cases for each pair of performance indicator and learning algorithm on S-RUBiS. For each case, the validation data is collected under a sampling interval of 120s for 102 intervals. By the end of each interval, an adaptation is performed.

S-RUBiS emulates the adaptable software systems with diverse runtime behaviors, in which the data is likely to be highly fluctuated, imposing extra difficulty to the learning.

ASOS: Adaptive Service-Oriented Software (ASOS) [41][42] is an adaptable software composed from a set of services. As shown in Figure 1, we have considered a workflow of 10 abstract services, each of which can be composed by different numbers of parallel and redundant instances of a pre-located concrete service. The feature inputs and the quality indicator of performance are discussed in Table III. Since it is difficult to emulate the actual running software system over the Internet, in this work, we configured the performance of services and the actual end-users by randomly extracting the data from the realistic WS-DREAM [43], which is a readily available dataset that contains a performance matrix for 4,500 services and 142 end-users from different countries. ASOS runs on the same physical machine as S-RUBiS. As the end-users exhibits randomized changes, i.e., whether an end-user requests ASOS or not at a given time point is completely random, we adapted the ASOS following the Monte Carlo method, and collected the data samples. Notably, through using different services-users matrices (as in Figure 2b), we obtained a total of 10 cases for ASOS. In this way, the associated service of each abstract service and the users varies across different

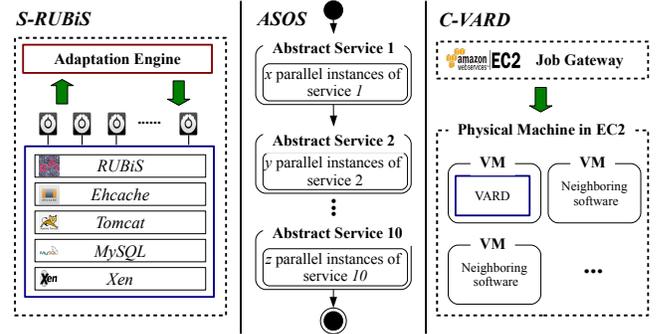


Fig. 1: Architecture of the subject adaptable software systems.

cases. For each case, the validation data is collected under a sampling interval of 10s for the total of 64 intervals. By the end of each interval, an adaptation is performed.

This adaptable software represents those that have different but stable runtime behaviors, where the data is likely to follow certain patterns with little emergent fluctuations.

C-VARD: C-VARD refers to the cloud-based VARD application [44], which is a corpus linguistics analysis software. We run C-VARD on Amazon EC2 using different VM instance types and a job of 2MB texts were submitted periodically, as shown in Figure 1. The feature inputs and the quality indicator of performance are discussed in Table III. Given the widely acknowledged issue of quality/performance interference in the cloud, running the same software (under the same environment conditions) can have a variety of performance depending the time of a day and day of a week, as the non-observable neighboring software and VMs running in the public cloud can influence the performance of C-VARD [9]. Therefore, we run C-VARD at different time on a day for one week on the same workload, and collected the samples. Those time slots of a day serve as the environmental feature in this study and they have been proven to be the cause of performance variance [45]. Through different VM instance types, as shown in Figure 2c, we obtained a total of 6 cases for C-VARD. For each case, the validation data is collected under a sampling interval of 600s for the total of 900 intervals. Since the cases determine all the possible adaptations and they are setup at design time, no further adaptation is needed at runtime.

This is a subject adaptable software that runs under an uncontrollable public domain, i.e., Amazon EC2. It serves as example of the software systems involves hidden and non-observable features that can influence their behaviors, i.e., through the VM interference [9].

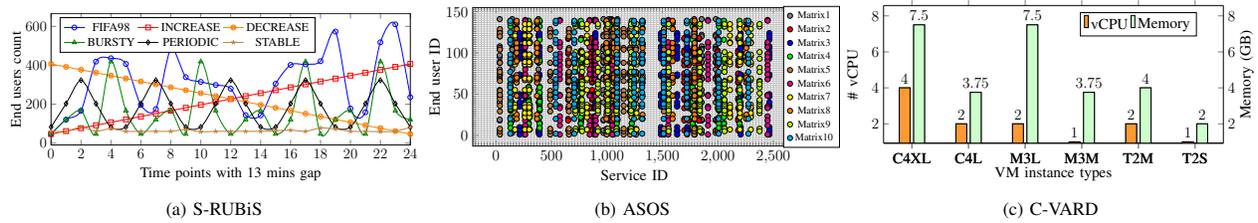


Fig. 2: (a) The different workload traces. (b) The different extracted services-users matrices (from WS-DREAM). Each point is a user-service pair, which contains data of throughput and reliability of a concrete service when the user is served. Each matrix is a set of 200 user-service pairs, i.e., 10 services and 20 users each. (c) The different VM instance types.

C. Analysis of the Fluctuation in Subject Software Systems

To analyze the fluctuation of the adaptable software, we use the following criteria to represents the changes at runtime:

Concept Drift: the concept drift [46] refers to the statistical properties of the target performance indicator, which the model is trying to predict, change over time in unforeseen ways. In general, for real-world software and data as what we studied in this work, there is no exact understanding about when the concept drift occurs. Therefore, we leverage ADWIN [47], a well-known drift detector, to measure the number of drifts in the data stream. Since we can only count the number of drifts not the extents of drifts, we apply another metric below.

Relative Standard Deviations (RSD): RSD measures the extents of change in the data stream by calculating the ratio between standard deviations and mean. This includes the data about the performance indicators and the related features of the software that can be used to train a model. The normalized nature of RSD allows us to report the mean value of the RSD, denoted as mRSD, for the features and performance indicators under all cases. A larger mRSD often imply that the overall extent of concept drifts is also more significant.

In Figure 3, we report on the average percentage of detected drifts over all the intervals and the mRSD for predicting each performance indicator for all cases. As we can see, the percentage of drifts on S-RUBiS and ASOS do not differ much, while the C-VARD exhibits smaller percentage. When we comparing the mRSD, we can observe that the S-RUBiS is clearly more fluctuated than the ASOS, which implies that although they have similar percentage of drifts, the extents of changes in the S-RUBiS is much greater than that of the ASOS. On both S-RUBiS and ASOS, it is obvious that one performance indicator is more fluctuated than another: e.g., response time is more fluctuated than energy consumption while the reliability is more fluctuated than throughput. As for C-VARD, even through it tends to be the most stable adaptable software (and thus easier to be modeled accurately), it involves hidden information that cannot be exploited in the modeling, which can negatively affect the model accuracy.

D. The Comparison Procedure and Metrics

To ensure generality, we investigated a wide range of combinations on scenarios and cases, which are defined as:

— **Scenario:** A scenario refers to each pair of learning algorithm and performance indicator of a software, e.g., using LR to predict the throughput of ASOS.

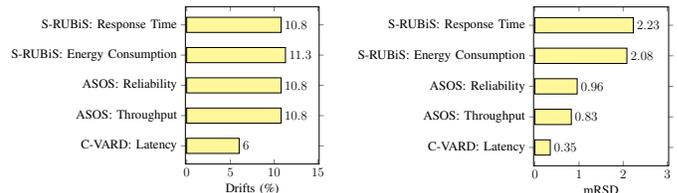


Fig. 3: The fluctuations of the subject adaptable software’s performance indicators in terms of the percentage of concept drifts and the mean RSD.

— **Case:** A case denotes a specific environmental dynamic and setups in a given scenario, e.g., for S-RUBiS, read-only pattern for FIFA98 workload trace with a heavy frequency.

Under each scenario-case pair, the models following both retrained and incremental modelings were pre-trained using the samples collected under random environment changes and cases. Next, we apply the pre-trained models to learn and predict the performance over all the runtime intervals, as specified in Section III-B. To mitigate the bias, we repeat 10 runs for each case under a scenario. Since we are interested in modeling at runtime, three metrics are particularly important:

Accuracy (Error): We measure the accuracy of the model as the adaptable software runs and as the model evolves⁴. At each time point t , a model is firstly updated by the data samples up to $t-1$ ($t-2$ for environment features). Then in the validation phase, the model takes the adaptable features at t and the environment features at $t-1$ to predict the performance at t , which is then compared with the ground truth at t . Given a scenario, we adopt Mean Absolute Error (MAE) to show the accuracy over all the intervals and repeated runs of a case, as it can additionally reflect the practicality of the error in the original scale. Suppose $y_{k,t}$ and $\hat{y}_{k,t}$ are the predicted and actual performance of the k th run at time t respectively; the MAE over n intervals and m repeated runs is:

$$MAE = \frac{1}{m \times n} \times \sum_{k=1}^m \sum_{t=1}^n |y_{k,t} - \hat{y}_{k,t}| \quad (3)$$

Training Time: We collected the time taken for training, and analyzed the Mean Training Time (MTT) over all the time intervals and repeated runs of a case.

Robustness: By analyzing the variance of the accuracy and training time, we aim to understand the robustness of

⁴Unlike classic validation in machine learning, the validation with changing model is done against each new data sample in the data stream [11].

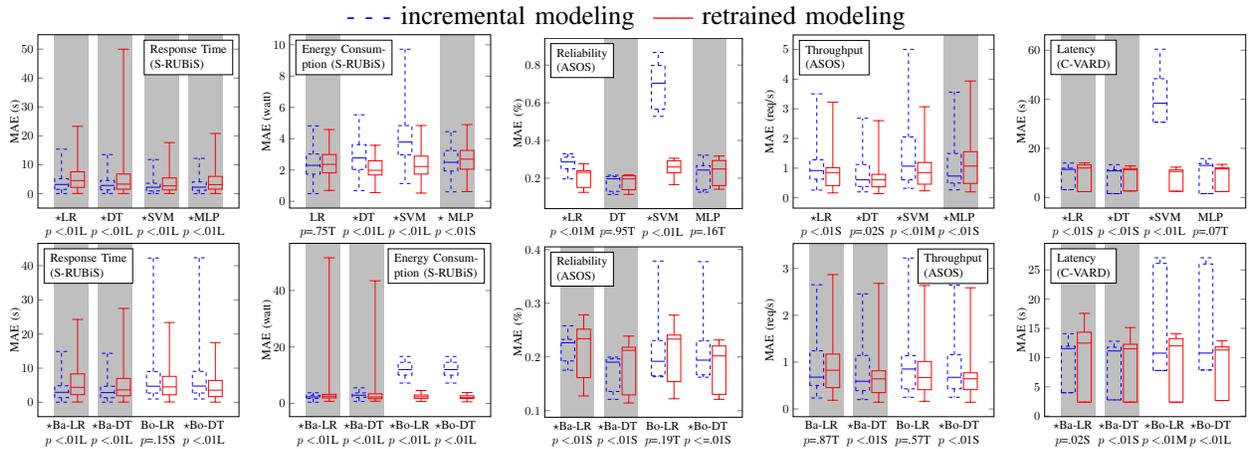


Fig. 4: The boxplot of mean absolute error. The scenarios for which the incremental modeling has better average on all cases are highlighted in gray. The \star means there is a statistically significant difference of the accuracy between incremental and retrained modeling, i.e., $p < .05$. T=trivial effect size; S=small effect size; M=medium effect size; L=large effect size.

each modeling method on both metrics (i.e., the smaller the variation, the better the robustness), when learning under abnormal and noisy data samples of the adaptable software.

Given that different features and performance have distinct scales, during training, it is necessary to normalize them into the same range using the upper and lower bounds. This can ensure good numeric stability, which in turn, significantly improves the prediction accuracy [30] [9]. Practically, the fact that the upper and lower bounds are unknown in prior implies that all data samples needs to be stored, and rescaled when the upper and lower bounds change. This has no implication on the retrained modeling method as the old model is discarded anyway, but it can affect the incremental modeling since the existing model becomes invalid once the samples are rescaled. Therefore in this work, whenever the data samples are rescaled, we recreate the incremental models by feeding the data sample one by one. However, for most of the scenarios, it is expected that the frequency of rescaling is low. Indeed, we only observed trivial amounts of rescaling in our study.

IV. RESULTS

We now discuss the results for all the research questions⁵:

A. RQ1: Accuracy

To answer **RQ1**, we study the model error of incremental and retrained modeling on each *scenario*, i.e., a given performance attribute and a learning algorithm. The reported results are extracted from the data set in which each data point represents the MAE of a case (over all intervals and 10 repeated runs) under a given scenario. We performed *Wilcoxon Signed-Rank test* for all comparisons of the results and repeated runs, as our data does not follow Gaussian distributions. We set $\alpha = .05$, which means that, if the test produces a p value that is smaller than $.05$, then we can reject the null hypothesis H_0 , which states that the given two

modeling methods cannot be distinguished statistically. We follow [48] to measure and classify the effect size.

As shown in Figure 4 top row, for the single learning algorithms, retrained and incremental modeling yield very competitive, but differentiable accuracy. Notably, the incremental one being slightly better as it outperforms the retrained modeling for 11 out of the 20 scenarios, and there are 8 improvements tends to be statistically significant with non-trivial effect sizes. The retrained modeling, on the other hand, dominates the incremental one for 9 scenarios, 8 of which are statistically significant with non-trivial effect sizes. However, even with the same learning algorithm, the better one can vary depending on the scenario, e.g., for LR, incremental modeling is better for 3 out of the 5 performance indicators while the retrained one is better for the rest. Taking a deeper look w.r.t. the learning algorithms, we found that the retrained and incremental modeling have similar effect on LR and DT, but they lead to considerably diverse result while working on SVM and MLP: overall, the retrained modeling is clearly better in accuracy on SVM while the incremental one yields much better results on MLP. For the ensembles in Figure 4 bottom, the accuracy between retrained and incremental modeling are clear: the former outperforms the latter on Boosting while it is the opposed on Bagging, where most comparisons are statistically significant with non-trivial effect sizes.

The variance of accuracy between incremental and retrained modeling has been similar for most of the learning algorithms, except for SVM and the two Boosting ensembles (Bo-RL and Bo-DT): the incremental SVM, Bo-RL and Bo-DT have generally larger variance than their retrained counterparts. This implies that the retrained modeling is more robust overall.

The two modeling methods have also led to interesting discovery with regards to the three different adaptable software systems under single learners: for highly fluctuated cases such as S-RUBiS, incremental modeling tends to offer better accuracy because it is more sensitive to the changes in data. For relatively stable cases of ASOS, the benefit of retrained modeling is more obvious since it is more important to learn

⁵All experiment data and results can be accessed at: <https://github.com/taochen/all-versus-one>

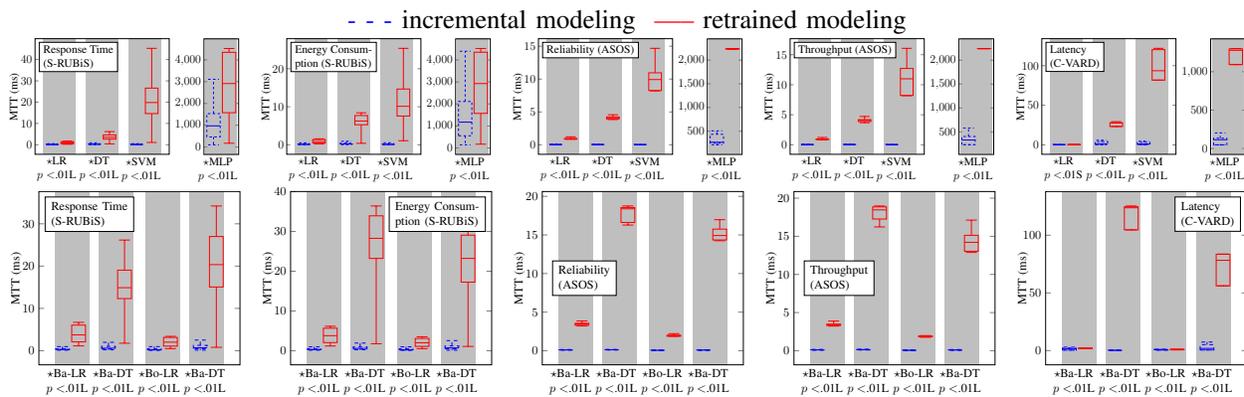


Fig. 5: The boxplot of mean training time. The scenarios for which the incremental modeling has better average on all cases are highlighted in gray. The \star means there is a statistically significant difference of the training time between incremental and retrained modeling, i.e., $p < .05$. T=trivial effect size; S=small effect size; M=medium effect size; L=large effect size.

the correlation between data samples in such case. As for C-VARD where there are hidden features, both modeling methods perform similarly, but the incremental one suffers larger variance when using SVM. However, we cannot rule out the implication of the actual performance indicator: as mentioned in Section III-C that the response time is more sensitive to runtime changes than the energy consumption on S-RUBiS while the reliability is more fluctuated than the throughput on ASOS. This is reflected in Figure 4, where the incremental modeling is generally more accurate than the retrained one for more fluctuated performance indicator.

For **RQ1**, we obtained the following findings:

Finding 1: The retrained version of a given learning algorithm does not always lead to higher accuracy than its incremental counterpart. In fact, the winner on accuracy can be considerably affected by the actual learning algorithm, i.e., incremental modeling is better with MLP while the retrained one is better with SVM, and the characteristics of subject adaptable software, i.e., the incremental modeling is more accurate for highly fluctuated adaptable software while the retrained one is better for stable software.

Finding 2: Overall, the retrained modeling tends to be more robust accuracy than that of the incremental modeling. This would affect the choice for adaptable software where the stability is more important than having greater accuracy.

Finding 3: For ensemble learning algorithms, the incremental modeling has consistently better accuracy on Bagging while the retrained one shows less error on Boosting.

B. RQ2: Training Time

To answer **RQ2**, we investigate the training time of incremental and retrained modeling on the scenarios and cases, similar to the previous section. We used the same statistical test and measure of effect size as mentioned before. As shown in Figure 5, for all scenarios and cases, the incremental modeling poses much less training time than the retrained modeling (including rescaling). The differences are statistically significant and with large effect sizes, even for simple learning algorithms like LR. For different learning algorithms, the reduction ranges

from one fold (e.g., for MLP) up to two order of magnitude (e.g., for SVM). The reduction is also significant for ensemble learners: up to three order of magnitude for Ba-DT.

In addition, the training time of retrained modeling vary depending on the subject adaptable software while the incremental one is more robust. For example, the incremental DT reveals limited variance while the retrained DT is more variable, and the deviation is higher on the highly fluctuated S-RUBiS. Yet, we did not observe obvious distinctions for different performance indicators on the same adaptable software.

However, although the incremental modeling has statistically shorter training time than the retrained one, it is not always practically meaningful: we see that for most of the learning algorithms, the differences are as of milliseconds, which may be negligible for non-critical systems. For MLP, we observed statistically and practically meaningful advantage of the incremental modeling on training time as of seconds.

For **RQ2**, we have the following findings:

Finding 4: Although the incremental modeling has statistically shorter training time than that of the retrained one (from 15% to three order of magnitude), the practical improvement may be trivial depending on the learning algorithms, e.g., for MLP, this can be practically important but may be negligible for other learning algorithms.

Finding 5: Training time of incremental modeling is more robust while that of the retrained one varies depending on the subject adaptable software: more stable software system can lead to robust training time while fluctuated ones can impose varied training time. This would affect the choice for adaptable software where any single spike of high training time can cause serious consequence.

C. RQ3: Trade-off Analysis

Next, to answer **RQ3**, we conduct trade-off analysis between accuracy and training time for both modeling methods. As shown in Figure 6, each point represents the *average value* of MAE and MTT for a learning algorithm (using either incremental or retrained modeling) under all the cases, when

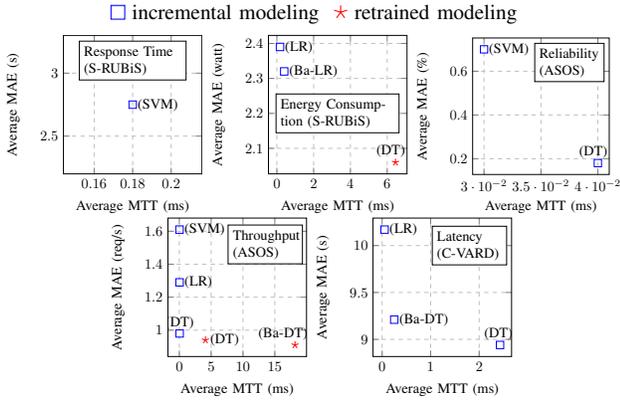


Fig. 6: The trade-off between accuracy and training time.

predicting a particular performance indicator. We show only the Pareto-optimal points as the others are dominated anyway.

Clearly, for 3 out of 5 performance indicators, the Pareto front contains only the incremental modeling as the non-dominated points, suggesting that it is better in terms of both accuracy and training time. In addition, we found that even for the same learning algorithm, e.g., DT, the incremental and retrained modeling can yield different positions on the trade-off surface. When considering the subject adaptable software of different characteristics, we observe that the retrained modelings tend to offer greater accuracy when the performance indicator of an adaptable software is less fluctuated, e.g., the energy consumption and throughput. As a result, the Pareto fronts for modeling energy consumption and throughput contain both modeling methods, implying that when considering all the learning algorithms studied, the incremental modeling could exhibit shorter training time but worse accuracy while the retrained modeling tends to impose longer training time but lead to better accuracy. This arises a trade-off between accuracy and training time when deciding whether incremental modeling or the retrained one is more suitable.

For **RQ3**, we obtained the following findings:

Finding 6: With all the learning algorithms studied, the incremental modeling yields better accuracy and training time for 3 out of the 5 performance indicators considered. For the remaining two indicators, there is a trade-off when considering all the learning algorithms studied: the incremental modeling could exhibit shorter training time but worse accuracy. Conversely, the retrained modeling tends to impose longer training time but lead to better accuracy. This means that it is possible for the incremental modeling to achieve the best on both accuracy and training time.

Finding 7: Even for the same learning algorithm, the decision of using incremental or retrained modeling can be a trade-off, see for example the DT on throughput.

D. RQ4: Correlations to Runtime Fluctuations

To analyze the correlations of accuracy and training time to the number of detected drifts and mRSD, we exploit *Spearman correlation* r_s . Specifically, r_s ranges from -1 to 1 where $-1 < r_s < 0$ means there is a negative monotonic correlation

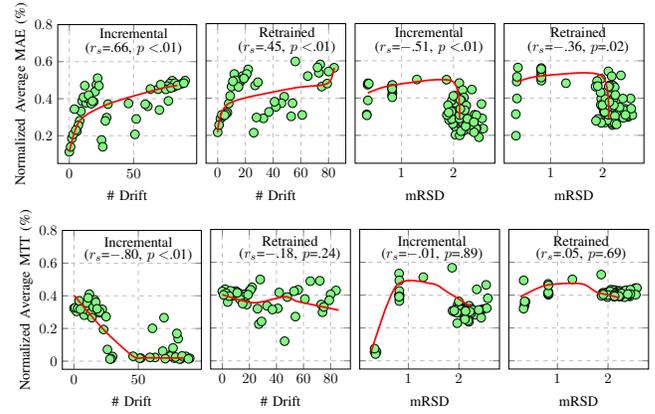


Fig. 7: The correlations of accuracy and training time to the runtime fluctuations.

while $0 < r_s < 1$ implies that there is a positive and monotonic correlation; the greater the absolute value of r_s , the stronger the correlation. $r_s = 0$ suggests that the two random variables have no monotonic correlation. We follow the guidance provided by [49] to measure the meaningfulness of r_s and we used the *Fisher transformation* to determine whether the r_s is statistically different from 0 under $\alpha = .05$.

Figure 7 report the correlations of accuracy and training time to the runtime fluctuations (i.e., number of drift and mRSD), together with the related r_s and its statistical significance (p value), for both the incremental and retrained modeling. Each point represents the *normalized average value* of the MAE (or MTT) for all cases and scenarios. To simplify the exposition, those points that have the same number of drifts (or mRSD) are merged and shown using their mean values.

As we can see from Figure 7 top, the error of incremental and retrained modeling exhibit at least moderate (strong for incremental modeling) positive monotonic correlation to the number of drifts, leading to r_s value of 0.66 and 0.45 respectively. The correlations are statistically significant as $p < .05$. This result suggests that the more number of drifts present in the data, the more likely to negatively affect the model accuracy on both modeling methods. In particular, the incremental modeling is more sensitive to the number of drifts, meaning that, as the number of drifts increases, its accuracy degrades quicker than that of the retrained one. As for mRSD, surprisingly, the errors of both incremental and retrained modeling possess similar pattern: they both negatively correlated with the mRSD, meaning the larger the extents of deviations in the data, the better the accuracy. Although the incremental modeling has merely a moderate correlation ($r_s = -0.51$) and the retrained one has a weak correlation ($r_s = -0.36$), the r_s values can be statistically distinguished from $r_s = 0$. They are weaker than the correlations between accuracy and the number of drifts, though. Relatively, the incremental modeling improves its accuracy quicker as the mRSD becomes larger. This explains why the incremental modeling tends to be more accurate than the retrained one on S-RUBiS, but worse on ASOS: because the mRSD of S-RUBiS is higher than that of ASOS; while they have similar

percentage of drifts.

Next, in Figure 7 bottom, we noted that the training time of incremental modeling has a remarkably strong negative correlation ($r_s = -0.80$) to the number of drifts which is statistically significant, implying that the more concept drifts, the less the training time. In other cases, we observed negative correlations, but they are rather weak and it is not significant statistically. As for mRSD, we discovered no noticeable correlation of training time to the mRSD, meaning that they are likely to change arbitrarily for both modeling methods.

For **RQ4**, we obtained the following findings:

Finding 8: For both the incremental and retrained modeling, their errors exhibit considerably positive monotonic correlations to the number of drifts, and non-trivial negative monotonic correlations to the deviations (mRSD) of data. Relatively, the accuracy of incremental modeling worse off faster when the number of drifts increase; and improve quicker when the mRSD becomes larger.

Finding 9: For the incremental modeling, its training time has strong negative monotonic correlations to the number of drifts while the correlation between the training time of retrained modeling and the number of drifts is arbitrary. There is also no clear relationship between the training time of both modeling methods and the deviations (mRSD) of data, or such a relationship is rather arbitrary.

V. LESSONS LEARNED

Lesson 1: The original belief has flaws and is inaccurate. Findings 1 - 3 are clear contradictions to the general belief when a learning algorithm is considered, such that the retrained modeling do not always lead to better accuracy than its incremental counterpart. Our findings have revealed some patterns when choosing the method, for example, the incremental modeling is more accurate for highly fluctuated adaptable software while the retrained one is better for stable software. The retrained modeling also exhibits more robust accuracy overall. Despite that the incremental modeling is always trained faster with better robustness than its retrained counterpart (Finding 4 and 5), which is consistent with the belief, the distinction may be practically insignificant, e.g., they differ only in milliseconds.

Lesson 2: Trade-off between accuracy and training time exists, but not always. When considering all learning algorithms, trade-off is needed based on preferences, but not always. The findings (Finding 6 and 7) reveal that it is possible for the incremental modeling to perform better on both accuracy and training time; This is partially comply with the general belief.

Lesson 3: Runtime fluctuation (i.e., number of drifts and deviations of data) could indeed impose non-trivial monotonic impacts on the accuracy, but limited on training time of both modeling methods. Our empirical findings (Finding 8 and 9) reveal that, in contrast to the retrained modeling, the accuracy of incremental modeling exhibits generally stronger, monotonic correlations to the number of drifts

and deviations of data, causing it to degrades faster with higher number of drifts and to improve quicker when the data deviations increases. The correlation between training time of both modeling methods and mRSD tends to be arbitrary, with the only clear correlation such that the incremental modeling trains faster as the number of drifts increases.

These results provide greater insights for the software engineers to choose the modeling methods. For example, by analyzing the collected data, one can infer that the incremental modeling is more promising in terms of accuracy when the number of drifts becomes smaller while the data deviations increases, which also come with an increased training time.

VI. THREATS TO VALIDITY

Due to the large variability of adaptable software, our findings may not be always applicable to all contexts. However, this kind of threats to external validity is not uncommon in empirical software engineering. We have attempted to mitigate this by conducting evaluations on three diverse and real-world adaptable software systems, 5 performance indicators, 8 machine learning algorithms and various settings, leading to a combinatorial total of 1,360 different conditions.

Threats to internal validity may be related to the parameters and stochastic nature of the algorithms. Indeed, the setting parameters of the learning algorithms may not necessarily be optimal, they were tailored using step-wise, trail and error method. But, the same strategy was used for every learning algorithm and thus this serves as a fair comparison between them. To improve the stability of results, all cases were repeated 10 runs, each of which with up to 900 data points. The resulted traces were observed to be stable. We have also reported their statistical significance and effect sizes.

VII. CONCLUSION

In this paper, we investigate an important lack of understanding in machine learning based performance modeling of adaptable software by means of empirical study and evaluations. To this end, we empirically compare both retrained and incremental modeling method on three diverse and real-world adaptable software systems, under a combinatorial total of 1,360 different conditions. Our finding reveals that the general belief is inaccurate, and shows some of the important, statistically significant factors that are often overlooked in existing work. The lessons learned would provide greater insights on the choice of modeling methods when machine learning is used to build runtime performance model.

Drawing on the findings from this work, we hope to stimulate a fruitful future research of a more systematic methodology on choosing between incremental and retrained modeling. In particular, the results serve as the foundation to create automatic tools that assist the software engineers on deciding the choice for a given adaptable software and case.

REFERENCES

- [1] O. Ibidunmoye, A. Rezaie, and E. Elmroth, "Adaptive anomaly detection in performance metric streams," *IEEE Transactions on Network and Service Management*, vol. 15, no. 1, pp. 217–231, March 2018.

- [2] T. Chen and R. Bahsoon, "Self-adaptive trade-off decision making for autoscaling cloud-based services," *IEEE Transactions on Services Computing*, vol. 10, no. 4, pp. 618–632, July 2017.
- [3] T. Chen, K. Li, R. Bahsoon, and X. Yao, "Femosaa: Feature-guided and knee-driven multi-objective optimization for self-adaptive software," *ACM Trans. Softw. Eng. Methodol.*, vol. 27, no. 2, pp. 5:1–5:50, Jun. 2018. [Online]. Available: <http://doi.acm.org/10.1145/3204459>
- [4] G. Bolch, S. Greiner, H. De Meer, and K. S. Trivedi, *Queueing networks and Markov chains: modeling and performance evaluation with computer science applications*. John Wiley & Sons, 2006.
- [5] N. Roy, A. Dubey, A. Gokhale, and L. Dowdy, "A capacity planning process for performance assurance of component-based distributed systems," in *Proceedings of the 2Nd ACM/SPEC International Conference on Performance Engineering*, ser. ICPE '11. New York, NY, USA: ACM, 2011, pp. 259–270. [Online]. Available: <http://doi.acm.org/10.1145/1958746.1958784>
- [6] F. Fittkau, S. Frey, and W. Hasselbring, "Cdosim: Simulating cloud deployment options for software migration support," in *Maintenance and Evolution of Service-Oriented and Cloud-Based Systems (MESOCA), 2012 IEEE 6th International Workshop on the*. IEEE, 2012, pp. 37–46.
- [7] D. Didona, F. Quaglia, P. Romano, and E. Torre, "Enhancing performance prediction robustness by combining analytical modeling and machine learning," in *Proceedings of the 6th ACM/SPEC international conference on performance engineering*. ACM, 2015, pp. 145–156.
- [8] P. Lama, Y. Guo, C. Jiang, and X. Zhou, "Autonomic performance and power control for co-located web applications in virtualized datacenters," *IEEE Transactions on Parallel and Distributed Systems*, vol. 27, no. 5, pp. 1289–1302, 2016.
- [9] T. Chen and R. Bahsoon, "Self-adaptive and online qos modeling for cloud-based software services," *IEEE Transactions on Software Engineering*, vol. 43, no. 5, pp. 453–475, 2017.
- [10] D. R. Wilson and T. R. Martinez, "The general inefficiency of batch training for gradient descent learning," *Neural Networks*, vol. 16, no. 10, pp. 1429–1451, 2003.
- [11] J. Read, A. Bifet, B. Pfahringer, and G. Holmes, "Batch-incremental versus instance-incremental learning in dynamic and evolving data," in *International Symposium on Intelligent Data Analysis*. Springer, 2012, pp. 313–323.
- [12] S. K. Tesfatsion, E. Wadbro, and J. Torndsson, "Autonomic resource management for optimized power and performance in multi-tenant clouds," in *Autonomic Computing (ICAC), 2016 IEEE International Conference on*. IEEE, 2016, pp. 85–94.
- [13] Q. Zhu and G. Agrawal, "Resource provisioning with budget constraints for adaptive applications in cloud environments," *IEEE Transactions on Services Computing*, vol. 5, no. 4, pp. 497–511, 2012.
- [14] J. Chen, G. Soundararajan, S. Ghanbari, Z. Jamil, M. D. Wang, A. B. Hashemi, and C. Amza, "Chorus: an interactive approach to incremental modeling and validation in clouds," in *Proceedings of 24th Annual International Conference on Computer Science and Software Engineering*. IBM Corp., 2014, pp. 61–74.
- [15] S. Kundu, R. Rangaswami, A. Gulati, M. Zhao, and K. Dutta, "Modeling virtualized applications using machine learning techniques," in *ACM Sigplan Notices*, vol. 47, no. 7. ACM, 2012, pp. 3–14.
- [16] G. Kousiouris, A. Menychtas, D. Kyriazis, K. Konstanteli, S. V. Gogouvtis, G. Katsaros, and T. A. Varvarigou, "Parametric design and performance analysis of a decoupled service-oriented prediction framework based on embedded numerical software," *IEEE Transactions on Services Computing*, vol. 6, no. 4, pp. 511–524, 2013.
- [17] C. Sieber, A. Obermair, and W. Kellerer, "Online learning and adaptation of network hypervisor performance models," in *Integrated Network and Service Management (IM), 2017 IFIP/IEEE Symposium on*. IEEE, 2017, pp. 1204–1212.
- [18] N. Esfahani, A. Elkhodary, and S. Malek, "A learning-based framework for engineering feature-oriented self-adaptive software systems," *IEEE transactions on software engineering*, vol. 39, no. 11, pp. 1467–1493, 2013.
- [19] J. Guo, K. Czarnecki, S. Apely, N. Siegmundy, and A. Wasowski, "Variability-aware performance prediction: A statistical learning approach," in *Proceedings of the 28th IEEE/ACM International Conference on Automated Software Engineering*. IEEE Press, 2013, pp. 301–311.
- [20] N. Siegmund, A. Grebhahn, S. Apel, and C. Kästner, "Performance-influence models for highly configurable systems," in *Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering*. ACM, 2015, pp. 284–294.
- [21] T. Chen, R. Bahsoon, and X. Yao, "A survey and taxonomy of self-aware and self-adaptive cloud autoscaling systems," *ACM Comput. Surv.*, vol. 51, no. 3, pp. 61:1–61:40, Jun. 2018. [Online]. Available: <http://doi.acm.org/10.1145/3190507>
- [22] T. Chen, R. Bahsoon, S. Wang, and X. Yao, "To adapt or not to adapt?: Technical debt and learning driven self-adaptation for managing runtime performance," in *Proceedings of the 2018 ACM/SPEC International Conference on Performance Engineering*, ser. ICPE '18. New York, NY, USA: ACM, 2018, pp. 48–55. [Online]. Available: <http://doi.acm.org/10.1145/3184407.3184413>
- [23] W. S. Sarle, "Neural networks and statistical models," 1994.
- [24] D. E. Rumelhart, G. E. Hinton, R. J. Williams *et al.*, "Learning representations by back-propagating errors," *Cognitive modeling*, vol. 5, no. 3, p. 1, 1988.
- [25] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.
- [26] I. Gerostathopoulos, C. Prehofer, and T. Bures, "Adapting a system with noisy outputs with statistical guarantees," in *Proceedings of the 13th International Conference on Software Engineering for Adaptive and Self-Managing Systems*, ser. SEAMS '18. New York, NY, USA: ACM, 2018, pp. 58–68. [Online]. Available: <http://doi.acm.org/10.1145/3194133.3194152>
- [27] D. A. Freedman, *Statistical models: theory and practice*. cambridge university press, 2009.
- [28] J. R. Quinlan *et al.*, "Learning with continuous classes," in *5th Australian joint conference on artificial intelligence*, vol. 92. Singapore, 1992, pp. 343–348.
- [29] S. Ghahremani, C. M. Adriano, and H. Giese, "Training prediction models for rule-based self-adaptive systems," in *2018 IEEE International Conference on Autonomic Computing (ICAC)*, Sep. 2018, pp. 187–192.
- [30] T. Chen and R. Bahsoon, "Self-adaptive and sensitivity-aware qos modeling for the cloud," in *Proceedings of the 8th International Symposium on Software Engineering for Adaptive and Self-Managing Systems*, 2013, pp. 43–52.
- [31] M. H. Hayes, "9.4: Recursive least squares," *Statistical Digital Signal Processing and Modeling*, p. 541, 1996.
- [32] P. Jamshidi, M. Velez, C. Kstner, N. Siegmund, and P. Kawthekar, "Transfer learning for improving model predictions in highly configurable software," in *2017 IEEE/ACM 12th International Symposium on Software Engineering for Adaptive and Self-Managing Systems (SEAMS)*, May 2017, pp. 31–41.
- [33] L. Rokach and O. Maimon, *Data mining with decision trees: theory and applications*. World scientific, 2014.
- [34] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.
- [35] Y. Freund and R. E. Schapire, "A desicion-theoretic generalization of on-line learning and an application to boosting," in *European conference on computational learning theory*. Springer, 1995, pp. 23–37.
- [36] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, Nov. 2009. [Online]. Available: <http://doi.acm.org/10.1145/1656274.1656278>

- [37] A. Bifet, G. Holmes, R. Kirkby, and B. Pfahringer, "Moa: Massive online analysis," *Journal of Machine Learning Research*, vol. 11, no. May, pp. 1601–1604, 2010.
- [38] M. Arlitt and T. Jin, "A workload characterization study of the 1998 world cup web site," *IEEE network*, vol. 14, no. 3, pp. 30–37, 2000.
- [39] R. University, "Rice university bidding systems," <http://rubis.ow2.org/>, 2009.
- [40] A. S. Foundation, "Apache tomcat," <http://tomcat.apache.org/>, 1999.
- [41] H. Wada, J. Suzuki, Y. Yamano, and K. Oba, "E³: A multiobjective optimization framework for sla-aware service composition," *IEEE Transactions on Services Computing*, vol. 5, no. 3, pp. 358–372, 2012.
- [42] T. Chen, M. Li, and X. Yao, "On the effects of seeding strategies: A case for search-based multi-objective service composition," in *Proceedings of the Genetic and Evolutionary Computation Conference*, ser. GECCO '18. New York, NY, USA: ACM, 2018, pp. 1419–1426. [Online]. Available: <http://doi.acm.org/10.1145/3205455.3205513>
- [43] Z. Zheng, Y. Zhang, and M. R. Lyu, "Investigating qos of real-world web services," *IEEE Transactions on Services Computing*, vol. 7, no. 1, pp. 32–39, 2014.
- [44] A. Baron and P. Rayson, "Vard2: A tool for dealing with spelling variation in historical corpora," in *Postgraduate conference in corpus linguistics*, 2008.
- [45] F. Samreen, Y. Elkhatib, M. Rowe, and G. S. Blair, "Daleel: Simplifying cloud instance selection using machine learning," in *Network Operations and Management Symposium (NOMS), 2016 IEEE/IFIP*. IEEE, 2016, pp. 557–563.
- [46] G. Widmer and M. Kubat, "Learning in the presence of concept drift and hidden contexts," *Machine learning*, vol. 23, no. 1, pp. 69–101, 1996.
- [47] A. Bifet and R. Gavaldà, "Learning from time-changing data with adaptive windowing," in *Proceedings of the 2007 SIAM International Conference on Data Mining*. SIAM, 2007, pp. 443–448.
- [48] V. B. Kampenes, T. Dybå, J. E. Hannay, and D. I. Sjøberg, "A systematic review of effect size in software engineering experiments," *Information and Software Technology*, vol. 49, no. 11, pp. 1073–1086, 2007.
- [49] K. L. Wuensch, "Straightforward statistics for the behavioral sciences," *Journal of the American Statistical Association*, vol. 91, no. 436, pp. 1750–1752, 1996.