

AutoWeka4MCPS-AVATAR: Accelerating Automated Machine Learning Pipeline Composition and Optimisation

Tien-Dung Nguyen^{a,*}, Bogdan Gabrys^a, Katarzyna Musial^a

^a*Advanced Analytics Institute, University of Technology Sydney, Australia*

Abstract

Automated machine learning pipeline (ML) composition and optimisation aim at automating the process of finding the most promising ML pipelines within allocated resources (i.e., time, CPU and memory). Existing methods, such as Bayesian-based and genetic-based optimisation, which are implemented in Auto-Weka, Auto-sklearn and TPOT, evaluate pipelines by executing them. Therefore, the pipeline composition and optimisation of these methods frequently require a tremendous amount of time that prevents them from exploring complex pipelines to find better predictive models. To further explore this research challenge, we have conducted experiments showing that many of the generated pipelines are invalid in the first place, and attempting to execute them is a waste of time and resources. To address this issue, we propose a novel method to evaluate the validity of ML pipelines, without their execution, using a surrogate model (AVATAR). The AVATAR generates a knowledge base by automatically learning the capabilities and effects of ML algorithms on datasets' characteristics. This knowledge base is used for a simplified mapping from an original ML pipeline to a surrogate model which is a Petri net based pipeline. Instead of executing the original ML pipeline to evaluate its validity, the AVATAR evaluates its surrogate model constructed by capabilities and effects of the ML pipeline components and input/output simplified mappings. Evaluating this surrogate model is less resource-intensive than the execution of the original pipeline. As a result, the AVATAR enables the pipeline composition and optimisation methods to evaluate more pipelines by quickly rejecting invalid pipelines. We integrate the AVATAR into the sequential model-based algorithm configuration (SMAC). Our experiments show that when SMAC employs AVATAR, it finds better solutions than on its own. This is down to the fact that the AVATAR can evaluate more pipelines within the same time budget and allocated resources.

Keywords: automated machine learning, pipeline composition and optimisation, machine learning pipeline evaluation

*Corresponding author

Email addresses: tiendung.nguyen-2@student.uts.edu.au (Tien-Dung Nguyen), bogdan.gabrys@uts.edu.au (Bogdan Gabrys), katarzyna.musial-gabrys@uts.edu.au (Katarzyna Musial)

1. Introduction

Automated machine learning (AutoML) aims at automating the process of data analysis, from the collection and integration of data, through composition and optimisation of ML pipelines, to deployment and maintenance of generated solutions (Kadlec and Gabrys, 2009; Salvador et al., 2018; Zöller and Huber, 2019). Although many existing studies proposed methods to tackle the problem of pipeline composition and optimisation (Salvador et al., 2016a,b, 2018; Olson and Moore, 2016; Feurer et al., 2015; Mohr et al., 2018a; Gil et al., 2018; de Sá et al., 2017; Tsakonas and Gabrys, 2012), one of the drawbacks of the majority of these methods is constructing more invalid pipelines than valid ones. There are two main approaches to deal with this issue.

In the first class of approaches, the pipelines' structures, which define the executed order of the pipeline components, use fixed templates (Kadlec and Gabrys, 2009; Salvador et al., 2018; Feurer et al., 2015). Although using fixed structures can reduce the number of invalid pipelines during the composition and optimisation, these approaches limit the exploration of promising pipelines which may have different structures than the ones predefined in the templates.

In the second class of approaches, there have been several attempts to reduce the randomness of pipeline construction by generating specific ML pipeline structures based on constraints. These constraints can be represented by using context-free grammars (de Sá et al., 2017; Tsakonas and Gabrys, 2012) or AI planning to guide the construction of pipelines (Mohr et al., 2018a; Gil et al., 2018). Nevertheless, all of these methods evaluate the validity of a pipeline by executing them (which we will refer to as a T-method). After executing a pipeline, if the result is an executable predictive model, the T-method evaluates the pipeline to be valid; otherwise, it is invalid. If a pipeline is complex, the number of preprocessing/predictor components within the pipeline can be high, or if the size of the dataset is large, the evaluation of the pipeline can be computationally expensive. Consequently, the optimisation process will require a significant time budget to find well-performing pipelines.

To address the drawback of generating invalid pipeline, we propose the AVATAR to evaluate ML pipelines using their surrogate models. The AVATAR transforms a pipeline to its surrogate model and evaluates it instead of executing the original pipeline. In the proposed approach, the evaluation of the surrogate models requires a knowledge base which is generated from a variety of synthetic datasets automatically.

AutoWeka for multi-component predictive systems (AutoWeka4MCPS) is an AutoML tool which can compose and optimise complex ML pipelines. AutoWeka4MCPS limits the generation of invalid pipelines by using a fixed pipeline template. This template includes the following components in their respective order: missing value handling → outlier removal → transformation → dimensionality reduction → sampling → predictor → meta-predictor. While the found solutions must include at least a predictor, the other data transformation steps are optional. However, if they are included in the

evaluated candidate solution, they have to follow the prescribed order. Apart from introducing the fundamental concepts for the AVATAR, we integrate the AVATAR with AutoWeka4MCPS and the SMAC optimisation to demonstrate the effectiveness of the AVATAR to reduce search space by quickly rejecting invalid pipelines.

To this end, this paper¹ has three main contributions:

- We conduct experiments on current state-of-the-art AutoML tools to show that the construction of a large number of invalid pipelines during the pipeline composition and optimisation may result in overall bad performance of the obtained ML pipeline.
- We propose the AVATAR to accelerate the automatic pipeline composition and optimisation by evaluating pipelines using a surrogate model. The AVATAR generates a knowledge base by learning the capabilities and effects of ML components automatically. The AVATAR uses this knowledge base to perform the simplified mapping from an ML pipeline component to its surrogate model, then finds the pipeline validity by evaluating the surrogate model.
- We perform extensive experiments to show that the combination of the multiple configuration initialisation and the AVATAR can better stabilise the convergence of SMAC in comparison with the single configuration initialisation. In order to do that, We customise SMAC to initialise with multiple configurations.

This paper is divided into five sections. After the Introduction, Section 2 reviews previous approaches to representing and evaluating ML pipelines in the context of AutoML. Section 3 describes the AVATAR which we propose to evaluate ML pipelines. Section 4 presents experiments to motivate our research and prove the efficiency of the proposed method. Finally, Section 5 concludes this study and points in the direction of the future research in the AutoML area with an emphasis on composition and optimisation of complex ML pipelines.

2. Related Work

In Section 4, through our experiments, we show that the generation of invalid pipelines has a negative effect that wastes time of the ML pipeline composition and optimisation. In this section, we review previous approaches that aimed at reducing this negative impact. Particularly, we have reviewed the current state-of-the-art pipeline composition and optimisation methods which are implemented as the AutoML tools such as AutoWeka4MCPS (Salvador et al., 2018), ML-Plan (Mohr et al., 2018a), P4ML

¹This study is a significant extension of our previous work (Nguyen et al., 2020), a conference paper presented at the Eighteenth International Symposium on Intelligent Data Analysis

(Gil et al., 2018), TPOT (Olson and Moore, 2016) and Auto-sklearn (Feurer et al., 2015). These approaches can be classified into two main types:

Fixed pipeline structure template: The main idea of these approaches is to use fixed pipeline structure templates designed based on the experience of experts. A template consists of component types connected in a fixed order. A component type has ML algorithms with their hyperparameters. When generating a pipeline, there are two main steps. Firstly, a number of component types are selected. Secondly, an algorithm and its hyperparameters are selected for each component type. These approaches are implemented in AutoWeka4MCPS and Auto-sklearn.

- AutoWeka4MCPS implements an automatic pipeline composition and optimisation method of multicomponent predictive systems (MCPS) to deal with the problem of combined algorithm selection and hyperparameter optimisation (CASH) (Salvador et al., 2018). AutoWeka4MCPS (Salvador et al., 2018) is developed on top of Auto-Weka 0.5 (Thornton et al., 2013). While Auto-Weka 0.5 only supports single-component pipeline (i.e. predictors/meta-predictors), AutoWeka4MCPS extends this pipeline structure with preprocessing components. The fixed pipeline structure template of AutoWeka4MCPS is *missing value handling* \rightarrow *outlier removal* \rightarrow *transformation* \rightarrow *dimensionality reduction* \rightarrow *sampling* \rightarrow *predictor* \rightarrow *meta-predictor*. The ML algorithms for these component types are from Weka libraries. The sequential model-based optimisation (SMAC) is used to solve the CASH problem with the extension of the pipeline structure up to seven components.
- Auto-sklearn also implements an automatic pipeline composition and optimisation method to deal with the CASH problem (Feurer et al., 2015). The fixed pipeline structure template of Auto-sklearn is *one hot encoding* \rightarrow *missing value imputation* \rightarrow *rescaling* \rightarrow *class balancing* \rightarrow *feature preprocessing* \rightarrow *predictor* \rightarrow *meta-predictor*. Auto-sklearn selects one algorithm for *feature preprocessing* component to perform matrix decomposition, univariate feature selection, classification-based feature selection, feature clustering, kernel approximations, polynomial feature expansion, feature embeddings, or sparse representation transformation. The ML algorithms for these component types are from scikit-learn libraries. Similar to AutoWeka4MCPS, Auto-sklearn employs SMAC to solve the CASH problem. Being different from AutoWeka4MCPS, Auto-sklearn employs a meta-learner that uses datasets’s meta-features (e.g., min, max, mean, standard deviation, class entropy, ratio of categorical, numerical and missing values) to find promising initialised configurations, whereas AutoWeka4MCPS uses random initialised configurations.

Knowledge base to limit pipeline structures: These approaches are implemented in ML-Plan, P4ML and TPOT. The main idea of these approaches is to use constraints in forms of context-free grammars, primitive taxonomy or ad-hoc configurations. These knowledge bases document structures

with generic components (e.g., preprocessing \rightarrow predictor) to a multilevel dependency of these generic components (e.g., EMImputation \rightarrow Standardise \rightarrow J48). Although these approaches allow more flexible pipeline structure than using fixed pipeline structure template, the knowledge bases are also designed based on the experience of experts that may not include all possibilities of valid pipelines.

- TPOT is a pipeline optimisation tool (Olson and Moore, 2016) which implements a genetic-programming-based optimisation method (Deb et al., 2002). TPOT generates pipelines which have data preprocessing components (e.g., data cleaning, feature selection, feature construction, and feature preprocessing), classification/regression and ensemble component. These components are from scikit-learn libraries. Theoretically, the pipeline can have an infinite number of components. However, the authors empirically limited the length of the pipeline to four components. The knowledge base of TPOT for generating pipelines is a Python file. According to TPOT’s authors, there are approximately 20% invalid pipelines generated during the evolutionary stages. While allowing a degree of randomness of mutation and cross-over, which imitates natural evolution, can help TPOT to explore better configurations, evaluations of these invalid pipelines significantly increase the complexity for the composition and optimisation.
- ML-Plan is a pipeline optimisation tool (Mohr et al., 2018b) which employs an AI planning approach, so-called the hierarchical task network (HTN), for pipeline composition and optimisation. This method generates pipelines by performing the depth-first search on an HTN search graph. The knowledge base of ML-Plan for generating pipelines is an ad hoc configuration file to describe HTN search graph. Because of using HTN search space as a knowledge base designed by experts for guiding pipeline composition, this search space may not have better configurations which are not known by the experts yet. Therefore, it reduces the chance to find more promising pipeline structures for a given dataset. On the other hand, this approach can guarantee to generate valid pipelines; therefore, it does not waste optimisation time for evaluating invalid pipelines as in the case of TPOT.
- P4ML is another pipeline optimisation tool which uses the AI planning approach (Gil et al., 2018). The knowledge base of P4ML for generating pipelines is a primitive taxonomy. This taxonomy describes the dependencies between primitives (i.e., ML algorithms). P4ML constructs pipelines sequentially within 5 phases. Phase 1 uses primitives for data preprocessing. Phase 2 identifies potential classifiers. Phase 3 validates the pipeline’s requirement. Phase 4 optimises hyperparameters of the pipeline. Phase 5 creates ensembles of pipelines. However, the planning-based optimisation algorithm is not presented in the study. It also lacks a detailed description of a pipeline structure.

Although AutoWeka4MCPS, ML-Plan, P4ML, TPOT and Auto-sklearn evaluate pipelines by executing them, these methods have strategies to limit the generation of invalid pipelines. Auto-sklearn and AutoWeka4MCPS use a fixed pipeline structure template. TPOT, ML-Plan and P4ML use a knowledge base to limit pipeline structures, which are designed manually, to guide the construction of pipelines. Although these approaches can reduce the number of invalid pipelines, our experiments showed that the time wasted for evaluation of the invalid pipelines is significant. Moreover, using fixed templates or a knowledge base reduces search spaces of potential better performing pipelines previously not known by experts, which is a drawback during pipeline composition and optimisation. Our proposed method can automatically generate a knowledge base and reuse this knowledge base to quickly evaluate the validity of pipelines using a surrogate model. This knowledge base can be extended further with new ML components using the same methodology as the one described in the next section when automatically generating the knowledge base in the first place.

3. AVATAR – Evaluation of ML Pipelines Using Surrogate Models

To address the challenges connected with the time-consuming ML pipelines composition, optimisation and evaluation, we propose the AVATAR². Its goal is to speed up the process by evaluating the surrogate pipelines which do not require the execution of the pipeline using the training, validation or test data. The main idea of the AVATAR is to extend the representation of MCPS introduced in (Salvador et al., 2017). The MCPS (i.e., ML pipelines), which are generated by AutoWeka4MCPS, are represented using Petri nets (Salvador et al., 2017). A Petri net is a mathematical modelling language used to represent pipelines (Salvador et al., 2018) as well as data service compositions (Tan et al., 2010). The main idea of Petri nets is to represent transitions of states of a system. Therefore, a Petri net can be used to describe the semantic as well as execution states (i.e., how data is transformed after going through each ML component) of a pipeline. The AVATAR replaces functions of Petri nets’ transitions, which in Salvador et al. (2017) represent full ML components, by functions to calculate transformations of datasets’ characteristics.

The AVATAR uses a surrogate model in the form of a Petri net. This surrogate pipeline keeps the structure of the original pipeline, replaces the datasets in the form of data matrices (i.e., components’ input/output simplified mappings) by the matrices of dataset-characteristics, and the ML components by transition functions to calculate the output from the input tokens (i.e., the matrices of dataset-characteristics). Because of the simplicity of the surrogate pipelines in terms of the size of the tokens and the simplicity of the transition functions, the evaluation of these pipelines is substantially less expensive than the original ones.

²<https://github.com/UTS-AAI/AVATAR>

3.1. The Validity of Pipelines

We define the validity of a pipeline as follows. Given a pipeline p , a dataset d and allocated resources r (i.e., CPU and memory), if we can execute the pipeline successfully and produce an output model, this pipeline is valid. Otherwise, it is not valid.

$$validity_p = exec(p, d, r) \quad (1)$$

By this definition, there are three factors (i.e., the pipeline, the dataset and the allocated resources) that impact the validity of a pipeline.

- Pipeline and dataset - these two factors directly relate to each other. Generating a valid pipeline requires to select ML components so that the output of the previous ML component is compatible with the next ML component. This compatibility depends on the characteristics of the data. For example, let's assume that we have a dataset with missing values and two pipelines. The first pipeline is *IndependentComponents* \rightarrow *J48*. The second pipeline is *EMImputation* \rightarrow *IndependentComponents* \rightarrow *J48*. The first pipeline is invalid because the component *IndependentComponents* can not handle a dataset with missing values. The second pipeline is valid because the component *EMImputation* fills in missing values. Therefore, the output of *EMImputation* is another dataset without missing values, and *IndependentComponents* can work with this dataset.
- Allocated resources - relate to the execution environment of the pipeline. The execution may require more resources (i.e., CPU and memory) than allocated ones due to the high complexity of ML components and large datasets. As a result, this execution may fail, leading to an invalid pipeline due to allocated resources.

Because the validity of pipelines depends on the characteristics of the data, we call these characteristics as dataset-characteristics. They can be changed depending on the transformations of a given dataset by ML components. Table 1 describes the dataset-characteristics used for this knowledge base. We select these dataset-characteristics because the validity of a pipeline for a given dataset depends on these characteristics. These dataset-characteristics are extracted from the capabilities of Weka algorithms³.

3.2. The AVATAR Architecture

Figure 1 shows the overall architecture of the AVATAR. The arrows represent data flows. There are two main flows in the architecture.

³<http://weka.sourceforge.net/doc.dev/weka/core/Capabilities.html>

Table 1: Descriptions of the data-characteristics.

Data-characteristics	Description
BINARY_CLASS	a dataset has binary classes
NUMERIC_CLASS	a dataset has numeric classes
DATE_CLASS	a dataset has date classes
MISSING_CLASS.VALUES	a dataset has missing values in classes
NOMINAL_CLASS	a dataset has nominal classes
SYMBOLIC_CLASS	a dataset has symbolic data in classes
STRING_CLASS	a dataset has string classes
UNARY_CLASS	a dataset has unary classes
BINARY_ATTRIBUTES	a dataset has binary attributes
DATE_ATTRIBUTES	a dataset has date attributes
EMPTY_NOMINAL_ATTRIBUTES	a dataset has an empty column
MISSING_VALUES	a dataset has missing values in attributes
NOMINAL_ATTRIBUTES	a dataset has nominal attributes
NUMERIC_ATTRIBUTES	a dataset has numeric attributes
UNARY_ATTRIBUTES	a dataset has unary attributes
PREDICTIVE_MODEL	a predictive model generated by a predictor

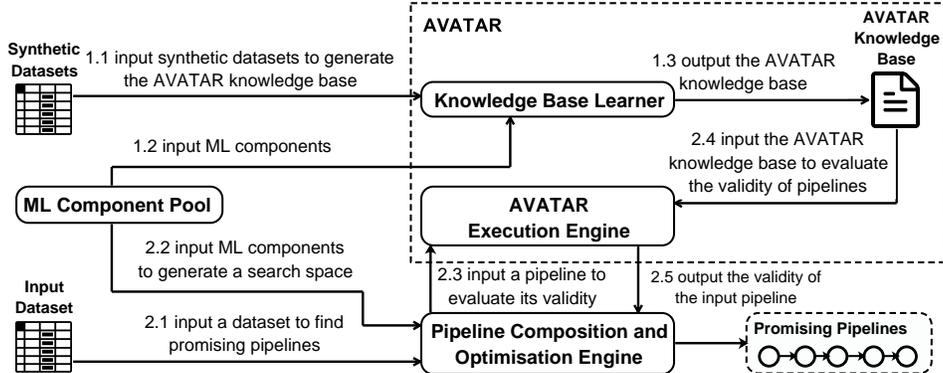


Figure 1: The overall architecture of the AVATAR

- *Learning the AVATAR knowledge base automatically (i.e., this data flow is denoted 1.x):* The Knowledge Base Learner employs synthetic datasets to learn the compatibility of each ML component in *ML Component Pool* (i.e., the set of ML components for the pipeline composition and optimisation) with characteristics of data, and the transformations of these characteristics after being executed by this component. This knowledge is stored in the *AVATAR Knowledge Base*.
- *Evaluating the validity of pipelines (i.e., this data flow is denoted 2.x):* The AVATAR Execution Engine employs the *AVATAR Knowledge Base* to evaluate the validity of pipelines. Given an input

dataset and the *ML Component Pool*, the *pipeline Composition and Optimisation Engine* (e.g., SMAC) finds the most promising pipelines. Before evaluating the quality of pipelines by executing them, the *Pipeline Composition and Optimisation Engine* employs the *AVATAR Execution Engine* to quickly evaluate the validity of these pipelines.

3.3. Learning the AVATAR knowledge base automatically

First of all, we discuss how to create *Synthetic Datasets*, machine learning algorithms in the *ML Component Pool*, and the structure of AVATAR Knowledge Base. After that, we present the algorithm, which is the core of the *Knowledge Base Learner*, to generate the AVATAR Knowledge Base from the *Synthetic Datasets* and *ML Component Pool* automatically.

3.3.1. Synthetic Datasets

To generate the *AVATAR Knowledge Base*, we propose to use synthetic datasets⁴ to isolate dataset-characteristics in each dataset to evaluate which and how dataset-characteristics impact on the capabilities and effects of ML components. Real-world datasets usually have many dataset-characteristics that make them not suitable for our purpose. To generate the AVATAR Knowledge Base, one task of the AVATAR is to evaluate the changes of a dataset before and after passing through an ML component. If the dataset has many dataset-characteristics, it is difficult to understand which dataset-characteristics cause these changes. We minimise the number of available dataset-characteristics in each synthetic dataset so that the knowledge base can be applicable in a variety of pipelines and datasets. For example, to learn the dataset-characteristic *MISSING_VALUES*, we create two datasets for classification and regression problems respectively. One dataset has one missing value attribute and one numeric class for the regression problem. Another dataset has one missing value attribute and one nominal class for the classification problem. Note that our approach can be extended with other dataset-characteristics by creating synthetic datasets containing these dataset-characteristics and numeric/nominal classes using the above methodology.

3.3.2. ML Component Pool

The *ML Component Pool* is a set of ML components. These components are data preprocessing algorithms and predictors/meta-predictors. Table 2 presents ML components that are used in this study. We use Weka libraries for the implementation of the *ML Component Pool*. The *ML Component Pool* plays two roles in the AVATAR architecture. Firstly, the *ML Component Pool* is used for the Knowledge Base Learner to learn how ML components transform dataset-characteristics and store the

⁴<https://github.com/UTS-AAI/AVATAR/tree/master/synthetic-datasets>

knowledge into the *AVATAR Knowledge Base*. Secondly, the *ML Component Pool* is used by the *Pipeline Composition and Optimisation Engine* to build search spaces.

3.3.3. The AVATAR Knowledge Base

The purpose of the *AVATAR Knowledge Base*⁵ is to describe the logic of transition functions of the surrogate pipelines. The logic includes the capabilities and effects of ML components (i.e., pipeline components). The capabilities and effects are constructed by dataset-characteristics. Listing 1 shows a segment of the AVATAR knowledge that describes the capabilities and effects of *EMImputation*.

The capabilities are used to verify whether an algorithm can be applied to a given dataset without making any changes to either of them. For example, whether the linear regression algorithm can be used on data with missing values and numeric attributes or not? The capabilities for a given algorithm are represented as a list of dataset-characteristics. The value of each capability-related dataset-characteristic is either 0 (i.e., the algorithm cannot work with the dataset which has this dataset-characteristic) or 1 (i.e., the algorithm can work with the dataset which has this dataset-characteristic). Based on the capabilities, we can determine which components of a pipeline (i.e., ML algorithms) are not able to process specific dataset-characteristics of a dataset. For example, *EMImputation* is compatible with missing values. Therefore, we have *MISSING_VALUES = 1* in the capabilities.

The effects describe data transformations. Similar to the capabilities, the effects are represented by a list of dataset-characteristics. Each effect-related dataset-characteristic can have three values, 0 (i.e., does not transform this dataset-characteristic), 1 (i.e., transforms one or more attributes/classes to this dataset-characteristic), or -1 (i.e., disables the effect of this dataset-characteristic on one or more attributes/classes). For example, *EMImputation* fills in all missing values of attributes. As a result, the output dataset of *EMImputation* does not have missing values. Therefore, we have *MISSING_VALUES = -1* in the effects.

⁵https://github.com/UTS-AAI/AVATAR/blob/master/avatar-knowledge-base/avatar_knowledge_base.json

Listing 1: A segment of the AVATAR Knowledge Base

```

...
  "componentId" : "weka.filters.unsupervised.attribute.EMImputation",
  "componentName" : "EMImputation",
  "listOfCapabilities" : [ {
    "mLComponentCapability" : "NOMINAL_CLASS",
    "value" : 0
  }, {
    "mLComponentCapability" : "NUMERIC_CLASS",
    "value" : 1
  }, {
    "mLComponentCapability" : "MISSING_VALUES",
    "value" : 1
  }, {
    "mLComponentCapability" : "NOMINAL_ATTRIBUTES",
    "value" : 0
  }, {
    "mLComponentCapability" : "NUMERIC_ATTRIBUTES",
    "value" : 1
  }
  ],
  "listOfEffects" : [ {
    "mLComponentCapability" : "NOMINAL_CLASS",
    "value" : 0
  }, {
    "mLComponentCapability" : "NUMERIC_CLASS",
    "value" : 0
  }, {
    "mLComponentCapability" : "MISSING_VALUES",
    "value" : -1
  }, {
    "mLComponentCapability" : "NOMINAL_ATTRIBUTES",
    "value" : 0
  }, {
    "mLComponentCapability" : "NUMERIC_ATTRIBUTES",
    "value" : 0
  }
  ],
...

```

3.3.4. The Knowledge Base Learner

The *Knowledge Base Learner* implements an algorithm to generate the *AVATAR Knowledge Base* from the *ML Component Pool* and *Synthetic Datasets*. Figure 2 presents the algorithm to generate the *AVATAR Knowledge Base*. This algorithm has four main stages:

1. Initialisation: The first stage initialises all dataset-characteristics available in the lists of capabil-

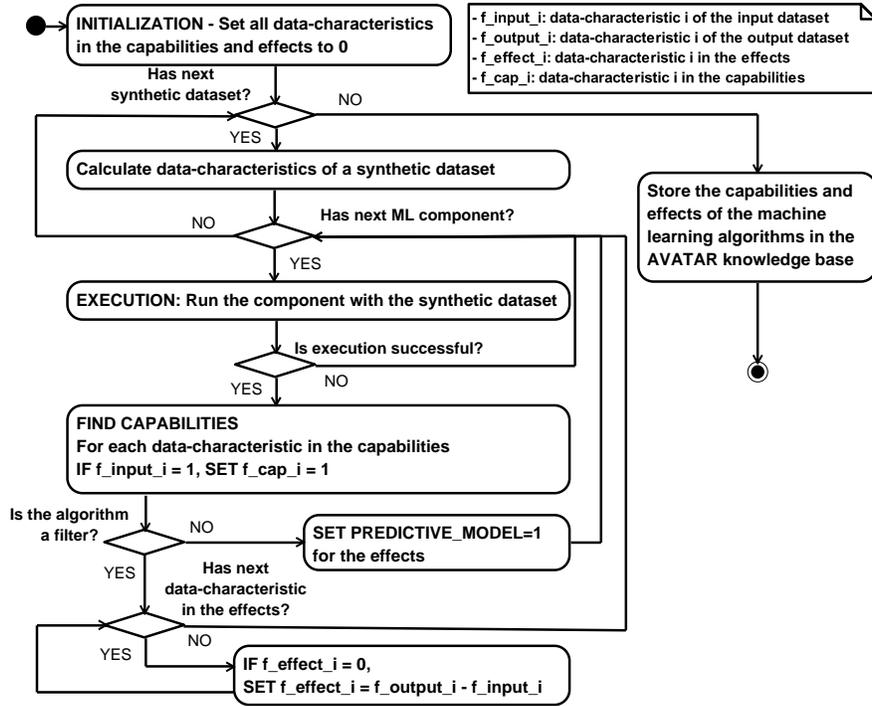


Figure 2: Algorithm to generate the knowledge base for evaluating surrogate pipelines.

- ities and effects to 0.
2. Execution: Run ML components with every synthetic dataset and get outputs (i.e., output datasets or predictive models).
 3. Find capabilities: If the execution is successful, the algorithm calculates the values of dataset-characteristics of the input dataset. After that, the algorithm sets the active dataset-characteristics (i.e., the dataset-characteristics have the value 1) of the input dataset for the ones in the capabilities. For example, the algorithm executes *EMImputation* with a dataset successfully. The algorithm calculates the values of dataset-characteristics of the input dataset. One of the dataset-characteristics is `MISSING_VALUE_ATTRIBUTE` which equals 1. The algorithm sets `MISSING_VALUE_ATTRIBUTE` to 1 in the capabilities. It means that *EMImputation* can work with a dataset containing missing values.
 4. Find effects: If an ML component is a predictor/meta-predictor, we set `PREDICTIVE_MODEL` for its effects. If the ML component is a data preprocessing component and its current value is a default value (the default value is 0 by the initialisation), we set this effect-related dataset-characteristic to the value of a difference between the value of this dataset-characteristic of the output and input dataset. For example, the algorithm executes *EMImputation* with an input dataset successfully. This input dataset has missing values (i.e., the token of the input dataset has `MISSING_VALUE_ATTRIBUTE=1`), and the output dataset does not have a missing value

(i.e., the token of the output dataset has MISSING_VALUE_ATTRIBUTE=0) because the missing values are filled in by *EMImputation*. The value of MISSING_VALUE_ATTRIBUTE in the effect of *EMImputation* equals $0-1=-1$.

The *AVATAR Knowledge Base* can be extended with the knowledge for others ML components by adding the effects and capabilities of these components manually, or automatically by using the *Knowledge Base Learner*. The *AVATAR Knowledge Base* can also be extended with other dataset-characteristics by creating synthetic datasets to isolate these dataset-characteristics and functions to calculate the values of these dataset-characteristics, then using the *Knowledge Base Learner* to learn these new dataset-characteristics for all ML Components in the *ML Component Pool*.

3.4. Evaluating the validity of pipelines

The AVATAR Execution Engine receives requests from the Pipeline Composition and Optimisation Engine to evaluate the validity of the pipelines. First of all, it maps the ML pipeline to its surrogate pipeline. After that, it evaluates this surrogate pipeline by firing all transitions. We claim that a Petri net is the most promising method to represent a surrogate pipeline. The reason is that it is fast to verify the validity of a Petri net based on simplified ML pipeline.

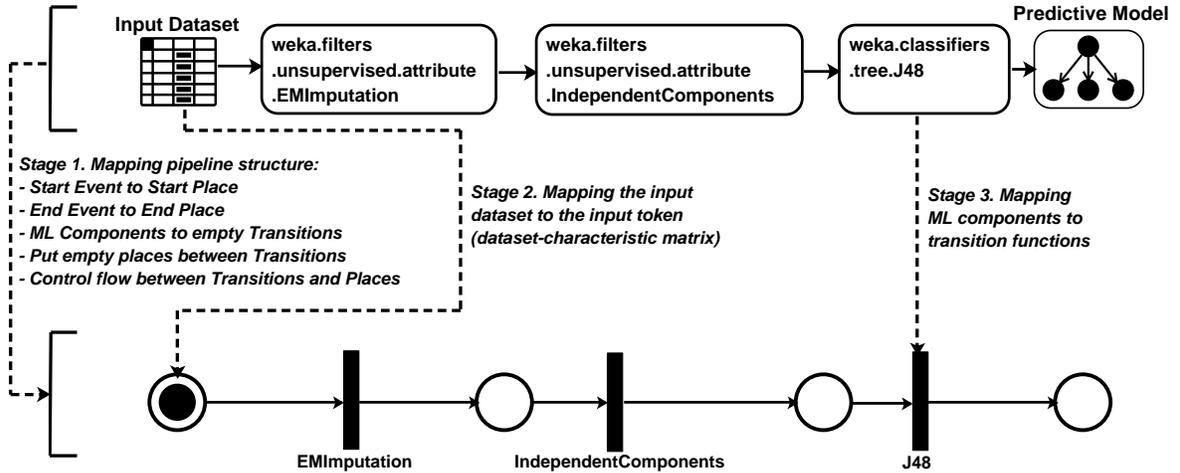


Figure 3: Mapping a ML pipeline to its surrogate model.

3.4.1. Mapping an ML pipeline to its surrogate model

The AVATAR maps an ML pipeline to a Petri net pipeline in three steps as presented in Figure 3.

1. The structure of the original ML pipeline is mapped to the respective structure of the Petri net surrogate pipeline. The start and end events are mapped to the start and end places, respectively.

The components are mapped to empty transitions. Empty places (i.e., places without tokens) are put between all transitions. Finally, all flows are mapped to arcs.

2. The values of dataset-characteristics are calculated from the input dataset to form a dataset-characteristic matrix which is the input token in the start place of the surrogate pipeline.
3. The transition functions are mapped from the original components. In this stage, only the corresponding component information (i.e., component name, identifier and parameters) is mapped to the transition function.

3.4.2. Evaluating a surrogate model

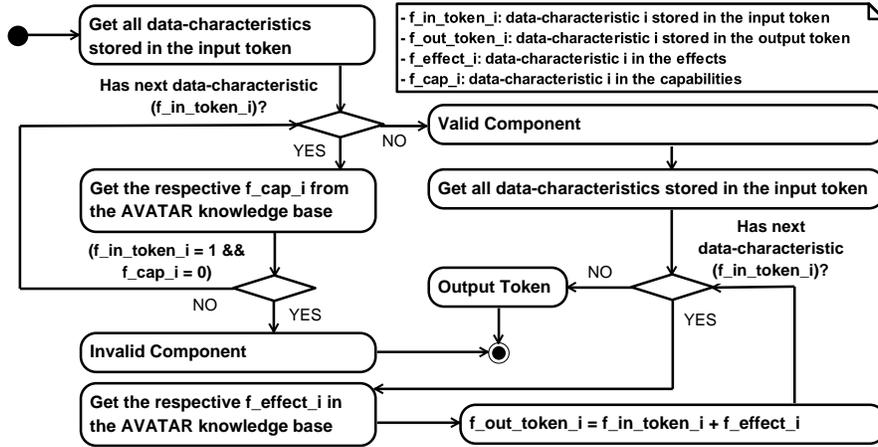


Figure 4: Algorithm for firing a transition of the surrogate model.

The evaluation of a surrogate model executes a Petri net pipeline. This execution starts by firing each transition of the Petri net pipeline and transforming the input token. As shown in Figure 4, firing a transition consists of two tasks: (i) the evaluation of the capabilities of each component and (ii) the calculation of the output token. The first task verifies the validity of the component using the following rules. If the value of a dataset-characteristic stored in the input token ($f_{in.token.i}$) is 1 and the corresponding dataset-characteristic in the component’s capabilities ($f_{cap.i}$) is 0, this component is invalid. Otherwise, this component is valid. If a component is invalid, the surrogate pipeline is evaluated as invalid. For example, if the input dataset has missing values in a class (i.e., $MISSING_CLASS_VALUES=1$) and the component can handle missing values in a class (i.e., $MISSING_CLASS_VALUES=1$), the component is valid. If the input dataset has missing values in a class (i.e., $MISSING_CLASS_VALUES=1$) and the component cannot handle missing values in a class (i.e., $MISSING_CLASS_VALUES=0$), the component is invalid. If the input dataset does not have missing values in a class (i.e., $MISSING_CLASS_VALUES=0$), the component is always valid.

The second task calculates each dataset-characteristic stored in the output token ($f_{out.token.i}$) in the next place from the input token by adding the value of a dataset-characteristic stored in the input

token ($f_{in_token_i}$) and the respective dataset-characteristic in the component's effects (f_{effect_i}). For example, it is assumed that an effect-related dataset-characteristic X is `MISSING_CLASS_VALUES`. If the input dataset has missing values in a class (i.e., `MISSING_CLASS_VALUES` = 1) and the component can remove/replace missing values in a class (i.e., `MISSING_CLASS_VALUES` = -1), so that the output dataset will not have missing values in a class (the dataset-characteristic of the output token is $1-1=0$). If the input dataset has missing values in a class (i.e., the dataset-characteristic X of the input token is 1) and the algorithm of the component cannot remove/replace missing values in a class (i.e., `MISSING_CLASS_VALUES` = 0), the output dataset still has missing values in a class (i.e., the dataset-characteristic X of the output token is $1+0=1$). If the input dataset has no missing values in a class (i.e., the dataset-characteristic X of the input token is 0), the output dataset also has no missing value in a class (i.e., $0-1=-1$). In this case, because the values of dataset-characteristics of a dataset are either 0 or 1, the minimum value 0 is set to `MISSING_CLASS_VALUES`.

3.5. An example to illustrate the pipeline evaluation of the AVATAR

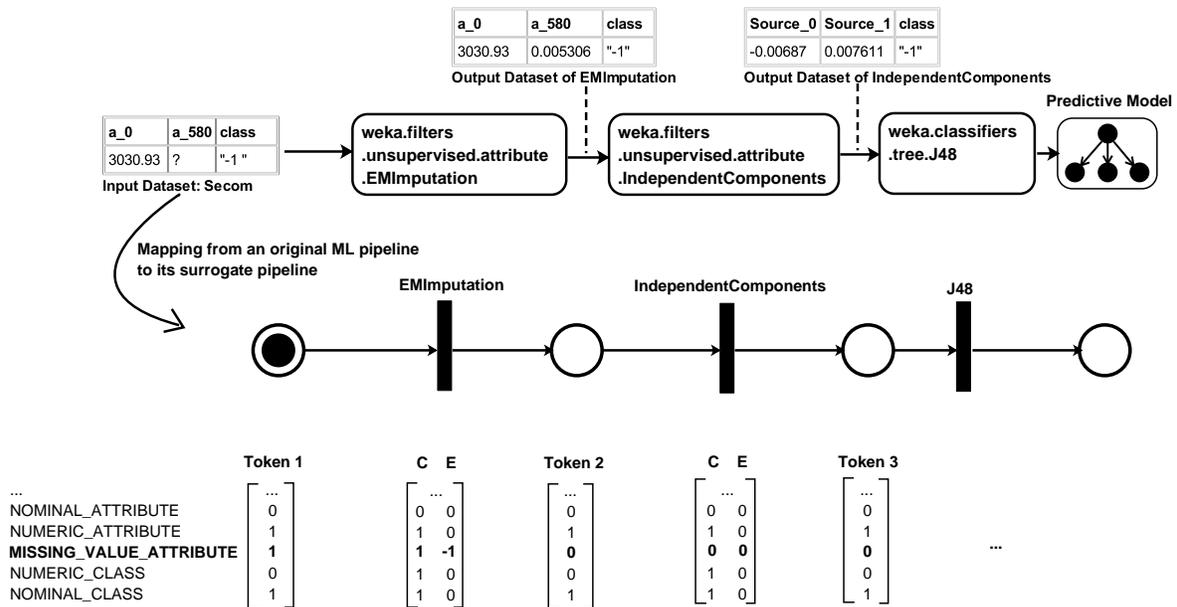


Figure 5: An example of evaluating a ML pipeline using the AVATAR

Figure 5 shows an example of how the AVATAR works. The original pipeline is an ML pipeline including three components: EMImputation, IndependentComponents and J48. The input dataset is *secom* which has numeric attributes, missing values in these numeric attributes and nominal values in the class/target attribute. We show the first row of two selected attributes and the target column of this dataset. We also show the first row of the transformed datasets after data preprocessing steps.

Firstly, the AVATAR maps the ML pipeline to the Petri net based pipeline. The input token for the

Petri net based pipeline is *token 1*. The values of `NOMINAL_ATTRIBUTE` and `NUMERIC_CLASS` equal 0. It means that the input data does not have nominal attribute and numeric class. The values of `NUMERIC_ATTRIBUTE`, `MISSING_VALUE_ATTRIBUTE` and `NOMINAL_CLASS` equal 1. It means that the input data have numeric attribute, missing value attribute and nominal class. The capability (C) and effect (E) of each ML component is retrieved from the AVATAR Knowledge Base.

Secondly, executing the Petri net based pipeline will fire all transitions. For example, token 1 has `MISSING_VALUE_ATTRIBUTE=1`. EMImputation has C of `MISSING_VALUE_ATTRIBUTE=1`. It means that EMImputation can work with missing value attributes. EMImputation has E of `MISSING_VALUE_ATTRIBUTE=-1`. It means that after passing token 1 to EMImputation, the value of `MISSING_VALUE_ATTRIBUTE` is $1 + (-1) = 0$ (i.e., the missing values are removed).

If IndependentComponents is put in front of EMImputation, we will have an invalid pipeline. This is because the IndependentComponents has C of `MISSING_VALUE_ATTRIBUTE=0`, which means it cannot handle missing value attributes.

4. Experiments and Discussion

To show the capabilities of the AVATAR, we have conducted four sets of experiments. The first set of experiments investigates the time required (wasted) to evaluate invalid pipelines which are generated during the traditional ML pipeline composition and optimisation process. After that, we have conducted the second set of experiments to compare the pipeline evaluation time of the AVATAR and the traditional approaches requiring pipelines executions to validate pipelines. The third set of experiments compares the quality (i.e., classification error) of the most promising pipelines found by SMAC with and without the AVATAR. Finally, the fourth set of experiments proposes and investigates the convergence speed of SMAC using the AVATAR and multiple initialisations of pipeline configurations during a single search for the best ML pipeline. We also provide a supplementary document⁶ with more details and experimental results which, due to the document size constraints, have not been included in this paper.

Table 3 summarises characteristics of datasets⁷ used for experiments. We have selected these datasets as they were used in previous studies (Salvador et al., 2018; Olson and Moore, 2016; Feurer et al., 2015) and therefore allow for easier comparative analysis. We have run all the experiments on AWS EC2 *t3a.medium* virtual machines which have 2 vCPU and 4GB memory.

⁶https://github.com/UTS-AAI/autoweka/blob/master/autoweka4mcps/doc/supplementary_document.pdf

⁷<https://archive.ics.uci.edu>

4.1. Experiments to investigate the time required to evaluate invalid pipelines

4.1.1. Experiment settings

The AutoML tools used for the experiments are AutoWeka4MCPS (Salvador et al., 2018) and Auto-sklearn (Feurer et al., 2015). These tools have been selected because of their abilities to construct and optimise hyperparameters of complex ML pipelines, and they have been empirically proven to be effective in a number of previous studies (Salvador et al., 2018; Feurer et al., 2015; Balaji and Allen, 2018). However, these previous experiments did not investigate the time needed for the evaluation of invalid pipelines and its impact on the pipeline composition and optimisation process. This is the goal of our first set of experiments.

To investigate this issue, we use five iterations (Iter) for the first set of experiments. Each iteration uses a different seed number. We set the time budget to 2 hours and the memory to 1GB per iteration. We evaluate the pipelines produced by the AutoML tools using three criteria: (1) the number of invalid/valid pipelines generated in each iteration, (2) the total evaluation time of invalid/ valid pipelines (seconds) and (3) the wasted evaluation time (%). The wasted evaluation time is calculated by the percentage of the total evaluation time of invalid pipelines over the total optimisation time. This criterion represents the percentage of the wasted time used to evaluate invalid pipelines.

4.1.2. Experiment results

Table 2: The list of ML Components.

ML Component	Data Preprocessing	Predictor/meta-predictor
weka.filters.unsupervised.attribute.ReplaceMissingValues	✓	
weka.filters.unsupervised.attribute.EMImputation	✓	
weka.filters.supervised.instance.ClassBalancer	✓	
weka.filters.supervised.instance.SpreadSubsample	✓	
weka.filters.unsupervised.attribute.PrincipalComponents	✓	
weka.filters.unsupervised.instance.Resample	✓	
weka.filters.unsupervised.instance.ReservoirSample	✓	
weka.filters.unsupervised.instance.RemovePercentage	✓	
weka.filters.unsupervised.instance.PeriodicSampling	✓	
weka.filters.supervised.attribute.AttributeSelection	✓	
weka.filters.unsupervised.attribute.Center	✓	
weka.filters.unsupervised.attribute.Standardize	✓	
weka.filters.unsupervised.attribute.Normalize	✓	
weka.filters.unsupervised.attribute.IndependentComponents	✓	
weka.filters.unsupervised.attribute.Discretize	✓	
weka.filters.unsupervised.attribute.NominalToBinary	✓	
weka.filters.unsupervised.attribute.NumericToBinary	✓	
weka.filters.unsupervised.attribute.NumericToNominal	✓	
weka.filters.unsupervised.attribute.StringToNominal	✓	
weka.classifiers.bayes.NaiveBayes		✓
weka.classifiers.bayes.NaiveBayesUpdateable		✓
weka.classifiers.functions.LinearRegression		✓
weka.classifiers.functions.Logistic		✓
weka.classifiers.functions.SimpleLogistic		✓
weka.classifiers.functions.GaussianProcesses		✓
weka.classifiers.functions.MultilayerPerceptron		✓
weka.classifiers.functions.SimpleLinearRegression		✓
weka.classifiers.functions.SMOreg		✓
weka.classifiers.functions.SMO		✓
weka.classifiers.lazy.IBk		✓
weka.classifiers.lazy.KStar		✓
weka.classifiers.lazy.LWL		✓
weka.classifiers.meta.AdaBoostM1		✓
weka.classifiers.meta.AttributeSelectedClassifier		✓
weka.classifiers.meta.Bagging		✓
weka.classifiers.meta.ClassificationViaRegression		✓
weka.classifiers.meta.LogitBoost		✓
weka.classifiers.meta.RandomCommittee		✓
weka.classifiers.meta.RandomSubSpace		✓
weka.classifiers.rules.DecisionTable		✓
weka.classifiers.rules.JRip		✓
weka.classifiers.rules.OneR		✓
weka.classifiers.rules.PART		✓
weka.classifiers.rules.ZeroR		✓
weka.classifiers.trees.DecisionStump		✓
weka.classifiers.trees.J48		✓
weka.classifiers.trees.LMT		✓
weka.classifiers.trees.RandomForest		✓
weka.classifiers.trees.RandomTree		✓
weka.classifiers.trees.REPTree		✓

Table 3: Summary of datasets' characteristics: the number of numeric attributes, nominal attributes, the number of distinct classes, instances in training and testing sets.

Dataset	Numeric	Nominal	No of distinct classes	Train	Test
abalone	7	1	28	2,924	1,253
adult	6	8	2	32,561	16,281
amazon	10,000	0	50	1,050	450
car	0	6	4	1,210	518
cifar10small	3,072	0	10	10,000	10,000
convex	784	0	2	8,000	50,000
dexter	20,000	0	2	420	180
dorothea	100,000	0	2	805	345
gcredit	7	13	2	700	300
gisette	5,000	0	2	4,900	2,100
kddcup	192	38	2	35,000	15,000
krvskp	0	36	2	2,238	958
madelon	500	0	2	1,820	780
mnist	784	0	10	12,000	50,000
secom	590	0	2	1,097	470
semeion	256	0	10	1,116	477
shuttle	9	0	7	43,500	14,500
waveform	40	0	3	3,500	1,500
winequality	11	0	11	3,429	1,469
yeast	8	0	10	1,039	445

Table 4: Negative impacts of invalid pipelines using AutoWeka4MCPS. (1): the number of invalid/ valid pipelines, (2): the total evaluation time of invalid/ valid pipelines (s), (3): the wasted evaluation time (%).

Dataset	Criteria	Iter 1	Iter 2	Iter 3	Iter 4	Iter 5
abalone	(1)	9/54	25/66	16/53	29/97	3/13
	(2)	16/7,284	8,639/322	7,571/286	3,361/2,351	8,070/74
	(3)	0.21	96.40	96.36	58.84	99.10
adult	(1)	5/25	4/23	22/23	8/25	4/19
	(2)	3,982/4,308	12/8,188	5,493/805	7,321/356	3,656/6,229
	(3)	48.04	0.14	87.22	95.36	36.98
amazon	(1)	12/23	6/4	26/0	17/19	1/0
	(2)	5,413/3,056	7,662/66	4,928/0	7,701/1,689	3,603/0
	(3)	63.92	99.67	100.00	82.01	100.00
car	(1)	22/94	26/123	31/133	71/263	32/128
	(2)	4,353/1,416	8,239/224	4,206/5,039	2,745/3,614	13,092/1,443
	(3)	75.48	97.35	45.49	43.17	90.07
cifar10small	(1)	2/6	11/30	16/0	1/6	10/17
	(2)	56/9,700	6,376/854	6,590/0	1,474/4,563	2,006/7,602
	(3)	0.57	88.19	100.00	24.41	20.88
convex	(1)	5/24	4/40	5/2	2/15	10/15
	(2)	3,625/4,309	5,640/3,787	4,124/2,129	2,229/4,968	6,622/3,364
	(3)	45.69	59.83	65.95	30.97	66.31
dexter	(1)	14/55	3/18	4/0	0/4	6/13
	(2)	1,827/4,246	3,604/3,852	7,205/0	0/8,796	7,210/89
	(3)	30.08	48.34	100.00	0.00	98.78
dorothea	(1)	5/17	9/3	4/0	1/0	5/16
	(2)	3,627/1,477	231/7,261	7,208/0	3,602/0	3,639/3,511
	(3)	71.05	3.09	100.00	100.0	50.89
gcredit	(1)	75/314	52/209	86/378	24/165	139/706
	(2)	479/5,513	5,233/941	1,387/5,908	3,363/1,975	1,827/2,664
	(3)	8.00	84.77	19.01	63.00	40.68
gisette	(1)	5/22	3/16	5/6	2/5	11/28
	(2)	930/7,853	638/6,218	5,239/3,522	4,119/6,116	4,533/3,185
	(3)	10.59	9.31	59.80	40.24	58.74
kddcup	(1)	8/18	46/22	11/32	4/0	5/2
	(2)	4,294/2,870	3,778/3,739	3,927/6,250	7,309/0	71/7,219
	(3)	59.94	50.26	38.58	100.00	0.97
krvskp	(1)	10/46	31/142	22/117	20/74	28/99
	(2)	3,614/1,223	4,292/1,506	8,788/766	5,609/224	3,792/1,243
	(3)	74.73	74.03	91.99	96.16	75.32
madelon	(1)	11/47	12/38	6/4	11/33	20/65
	(2)	6,474/3,034	7,170/3,257	5,108/514	3,705/2,469	6,022/3,151
	(3)	68.09	68.77	90.86	60.01	65.65
mnist	(1)	2/11	5/5	11/0	0/13	0/2
	(2)	29/7,171	10,641/3,660	5,780/0	0/7,228	0/7,214
	(3)	0.40	74.41	100.00	0.00	0.00
secom	(1)	19/89	2/14	17/97	17/118	3/31
	(2)	3,974/4,236	2,534/5,676	6,382/2,404	4,405/5,523	4/9,414
	(3)	48.40	30.86	72.64	44.37	0.04
semeion	(1)	27/84	5/0	22/23	20/51	14/23
	(2)	1,935/6,224	9,103/0	2,759/3,459	3,165/3,647	3,311/1,731
	(3)	23.71	100.00	44.37	46.46	65.67
shuttle	(1)	16/57	2/0	11/0	10/49	1/10
	(2)	3,727/1,620	4,873/0	7210/0	7,220/2,276	1/7,915
	(3)	69.71	100.00	100.00	76.03	0.01
waveform	(1)	18/42	5/5	11/0	25/84	20/33
	(2)	4,639/173	4,158/410	5,511/0	4,496/1,058	5,146/2,714
	(3)	96.41	91.02	100.00	80.95	65.47
wineqw	(1)	27/90	2/0	12/0	15/62	20.32
	(2)	6,082/725	4,444/0	8,376/0	4,573/306	8,526/1,412
	(3)	89.35	100.00	100.00	98.73	85.79
yeast	(1)	39/178	5/0	49/83	49/173	28/77
	(2)	5,828/325	7,033/0	8,821/224	10,630/1,172	7,501/660
	(3)	94.73	100.00	97.52	86.13	91.92

Table 5: Negative impacts of invalid pipelines using Auto-sklearn. (1): the number of invalid/ valid pipelines, (2): the total evaluation time of invalid/ valid pipelines (s), (3): the wasted evaluation time (%).

Dataset	Criteria	Iter 1	Iter 2	Iter 3	Iter 4	Iter 5
abalone		crashed	crashed	crashed	crashed	crashed
adult		crashed	crashed	crashed	crashed	crashed
amazon		crashed	crashed	crashed	crashed	crashed
car		crashed	crashed	crashed	crashed	crashed
cifar10small	(1)	4/4	7/1	6/4	6/7	6/2
	(2)	4579/2457	6819/263	5885/1194	3867/3253	6613/508
	(3)	65.08	96.29	83.14	54.31	92.86
convex	(1)	9/19	9/18	9/16	8/18	9/14
	(2)	2692/4475	2625/4539	2880/4291	2646/4524	3121/4052
	(3)	37.57	36.65	40.16	36.90	43.51
dexter	(1)	9/44	9/42	9/44	9/44	9/44
	(2)	4917/2177	3773/3333	4742/2366	4704/2415	4028/3079
	(3)	69.31	53.10	66.72	66.08	56.68
dorothea		crashed	crashed	crashed	crashed	crashed
gcredit		crashed	crashed	crashed	crashed	crashed
gisette	(1)	6/7	5/6	4/3	5/4	3/5
	(2)	5227/1920	4380/2723	6734/387	6125/1002	5156/1988
	(3)	73.13	61.67	94.56	85.94	72.17
kddcup		crashed	crashed	crashed	crashed	crashed
krvskp		crashed	crashed	crashed	crashed	crashed
madelon	(1)	9/50	11/47	9/49	9/49	9/49
	(2)	4215/2925	4763/2374	4914/2225	4774/2366	3718/3423
	(3)	59.03	66.74	68.84	66.86	52.07
mnist	(1)	3/8	6/10	6/10	7/5	5/9
	(2)	2802/4355	3379/3795	4013/3160	5837/1337	5516/1658
	(3)	39.15	47.10	55.94	81.37	76.89
secom		crashed	crashed	crashed	crashed	crashed
semeion	(1)	4/32	12/55	7/51	5/37	4/26
	(2)	6122/1042	5906/1228	5412/1731	5715/1444	5943/1230
	(3)	85.46	82.79	75.77	79.83	82.85
shuttle	(1)	3/10	3/10	2/18	2/15	3/19
	(2)	4329/2853	4312/2870	2087/5087	1897/5282	1957/5215
	(3)	60.27	60.03	29.09	26.42	27.28
waveform	(1)	3/21	4/27	3/27	3/30	3/22
	(2)	4081/3098	4295/2873	3679/3489	3890/3277	3876/3299
	(3)	56.84	59.92	51.32	54.28	54.02
wineqw	(1)	3/54	2/39	3/55	2/57	2/43
	(2)	1891/5244	2845/4309	1815/5322	9/7127	2140/5012
	(3)	26.50	39.77	25.44	0.12	29.92
yeast		crashed	crashed	crashed	crashed	crashed

Dataset	Number of invalid pipelines			Number of valid pipelines			Total evaluation time of invalid pipelines (s)			Total evaluation time of valid pipelines (s)			Wasted evaluation time (%)		
	Mean	StdDev	Max	Mean	StdDev	Max	Mean	StdDev	Max	Mean	StdDev	Max	Mean	StdDev	Max
abalone	16	10	29	57	27	97	5,531	3,328	8,639	2,063	2,738	7,284	70.18	38.04	99.10
adult	9	7	22	23	2	25	4,093	2,418	7,321	3,977	3,036	8,188	53.55	34.75	95.36
amazon	12	9	26	9	10	23	5,861	1,600	7,701	962	1,230	3,056	89.12	14.38	100.00
car	36	18	71	148	59	263	6,527	3,754	13,092	2,347	1,735	5,039	70.31	22.36	97.35
cifar10small	8	6	16	12	11	30	3,300	2,676	6,590	4,544	3,747	9,700	46.81	39.63	100.00
convex	5	3	10	19	13	40	4,448	1,540	6,622	3,711	956	4,968	53.75	13.62	66.31
dexter	5	5	14	18	20	55	3,969	2,879	7,210	3,397	3,242	8,796	55.44	39.07	100.00
dorothea	5	3	9	7	8	17	3,661	2,207	7,208	2,450	2,728	7,261	65.01	36.11	100.00
gcredit	75	38	139	354	191	706	2,458	1,672	5,233	3,400	1,968	5,908	43.09	28.11	84.77
gisette	5	3	11	15	9	28	3,092	1,920	5,239	5,379	1,768	7,853	35.74	22.18	59.80
kddcup	15	16	46	15	12	32	3,876	2,300	7,309	4,016	2,560	7,219	49.95	32.04	100.00
krvskp	22	7	31	96	33	142	5,219	1,916	8,788	992	452	1,506	82.45	9.59	96.16
madelon	12	5	20	37	20	65	5,696	1,200	7,170	2,485	1,022	3,257	70.68	10.55	90.86
mnist	4	4	11	6	5	13	3,290	4,302	10,641	5,055	2,876	7,228	34.96	43.42	100.00
secom	12	7	19	70	40	118	3,460	2,121	6,382	5,451	2,303	9,414	39.26	23.80	72.64
semeion	18	8	27	36	29	84	4,055	2,569	9,103	3,012	2,080	6,224	56.04	25.68	100.00
shuttle	8	6	16	23	25	57	4,606	2,671	7,220	2,362	2,917	7,915	69.15	36.69	100.00
waveform	16	7	25	33	30	84	4,790	481	5,511	871	989	2,714	86.77	12.44	100.00
wineqw	15	8	27	38	39	90	6,400	1,771	8,526	489	533	1,412	94.77	6.01	100.00
yeast	34	16	49	102	67	178	7,963	1,642	10,630	476	408	1,172	94.06	4.80	100.00

Table 6: Mean, standard deviation (StdDev), and Maximum (Max) of the number of invalid and valid pipelines, the total evaluation time of invalid and valid pipelines, and the wasted evaluation time using AutoWeka4MCPS.

Dataset	Number of invalid pipelines			Number of valid pipelines			Total evaluation time of invalid pipelines (s)			Total evaluation time of valid pipelines (s)			Wasted evaluation time (%)		
	Mean	StdDev	Max	Mean	StdDev	Max	Mean	StdDev	Max	Mean	StdDev	Max	Mean	StdDev	Max
abalone	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
adult	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
amazon	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
car	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
cifar10small	6	1	7	4	2	7	5,553	1,151	6,819	1,535	1,148	3,253	78.34	16.18	96.29
convex	9	0	9	17	2	19	2,793	187	3,121	4,376	185	4,539	38.96	2.59	43.51
dexter	9	0	9	44	1	44	4,433	448	4,917	2,674	449	3,333	62.38	6.31	69.31
dorothea	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
gcredit	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
gissette	5	1	6	5	1	7	5,524	819	6,734	1,604	818	2,723	77.49	11.49	94.56
kddcup	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
krvskp	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
madelon	9	1	11	49	1	50	4,477	448	4,914	2,663	449	3,423	62.71	6.29	68.84
mnist	5	1	7	8	2	10	4,309	1,184	5,837	2,861	1,180	4,355	60.09	16.49	81.37
secom	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
semeion	6	3	12	40	11	55	5,820	241	6,122	1,335	235	1,731	81.34	3.31	85.46
shuttle	3	0	3	14	4	19	2,916	1,148	4,329	4,261	1,145	5,282	40.62	15.97	60.27
waveform	3	0	4	25	3	30	3,964	209	4,295	3,207	208	3,489	55.28	2.91	59.92
wineqw	2	0	3	50	7	57	1,740	939	2,845	5,403	933	7,127	24.35	13.13	39.77
yeast	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Table 7: Mean, standard deviation (StdDev), and Maximum (Max) of the number of invalid and valid pipelines, the total evaluation time of invalid and valid pipelines, and the wasted evaluation time using Auto-sklearn. Symbol '-' presents crashed experiments.

Table 4 and 5 present the number of invalid pipelines and the wasted time used to evaluate these invalid pipelines in ML pipeline composition and optimisation of AutoWeka4MCPS and Auto-sklearn. Table 6 and Table 7 present mean, standard deviation and maximum of the number of invalid and valid pipelines, the evaluation time of the invalid and valid pipelines, and the wasted evaluation time. Firstly, these tables show that not all of the constructed pipelines are valid. We can see the presence of invalid pipelines for both AutoML tools across different datasets and iterations. For example, there are 9 invalid and 54 valid pipelines in the case of using AutoWeka4MCPS for the dataset *abalone* in *Iter 1*. Mean, standard deviation and maximum of the number of invalid pipelines in the case of using AutoWeka4MCPS for the dataset *abalone* are 16, 10 and 29 respectively. There are 9 invalid and 19 valid pipelines in the case of using Auto-sklearn, the dataset *convex* and *Iter 1*. Mean, standard deviation and maximum of the number of invalid pipelines in the case of using Auto-sklearn for the dataset *convex* is 9, 0 and 9 respectively. Secondly, the evaluation time of these invalid pipelines may be significant in several cases. For example, the wasted evaluation time is 75.48% in the case of using the dataset *car* and *Iter 5*. We can see that changing seed numbers has a strong impact on the wasted evaluation time in the case of AutoWeka4MCPS. For example, the experiments with the dataset *abalone* show that the wasted evaluation time is in the range between 0.21% and 99.10%. The reason is that Weka libraries themselves can evaluate the compatibility of a single component pipeline without execution. If the initialisation of the pipeline composition and optimisation with a specific seed number results in pipelines consisting of only one predictor, and these pipelines are well-performing, it tends to exploit similar ML pipelines. As a result, the wasted evaluation time is low. However, if the initialisation results in a complex pipeline which is invalid, there is no guidance for the next promising pipelines. Therefore, the next promising pipelines are randomly selected. As a result, the upcoming pipelines that will be evaluated may also be invalid. This impact is negligible in the case of Auto-sklearn. In other words, the impact of changing seed numbers on the variance of the wasted evaluation in the case of Auto-sklearn is less than in the case of AutoWeka4MCPS. For example, standard deviation of the number of invalid pipelines in the case of using Auto-sklearn and the dataset *cifar10small*, *convex* and *dexter* are 1, 0 and 0 respectively. However, standard deviation of the number of invalid pipelines in the case of using AutoWeka4MCPS and the dataset *cifar10small*, *convex* and *dexter* are 6, 3, 5 respectively. The reason is that Auto-sklearn uses meta-learning to initialise with promising ML pipelines. The experiments with the datasets *abalone*, *adult*, *amazon*, *car*, *dorothea*, *gcredit*, *kddcup*, *krvskp*, *secom*, and *yeast* show that Auto-sklearn limits the generation of invalid pipelines by making assumption about cleaned input datasets. The experiments crash if the input datasets have data quality issues (i.e. missing values) or not transformed into a specific, required format (i.e., all attributes must be in numeric format). Similar to AutoWeka4MCPS, the Auto-sklearn can not handle invalid pipelines effectively even with the initialisation using the meta-learning. In conclusion, we empirically prove the presence of invalid

pipelines and that the wasted time used to evaluate these invalid pipelines can be significant.

4.2. Experiments to compare the performance of the AVATAR and the traditional method that requires executions

4.2.1. Experiment settings

In order to demonstrate the efficiency of the AVATAR, we compare the performance of the AVATAR and the traditional method (which we refer to as T-method) that requires the executions of pipelines. The T-method is used to evaluate the validity of pipelines in the pipeline composition and optimisation of AutoWeka4MCPS and Auto-sklearn. We use the same datasets as in the first set of experiments. The goal of the second set of experiments is to show that the AVATAR can evaluate the validity of pipelines faster than the T-method and the evaluated results of the AVATAR are similar to the T-method.

We randomly generate ML pipelines which have up to six components (i.e. these component types are missing value handling, dimensionality reduction, outlier removal, data transformation, data sampling and predictor). The predictor is put at the end of the pipelines because a valid pipeline always has a predictor at the end. Each pipeline is evaluated using both the AVATAR and the T-method. We set the time budget to 10 hours per dataset instead of 2 hours in the first set of experiments. The reason is that we want to use the same total time budget of the first set of experiments. Each iteration of the first set of experiments has 2 hours optimisation time for each dataset. Therefore, five iterations have 10 hours of optimisation time. Please note that the generation of random pipelines in this set of experiments is not a part of the full ML pipeline composition and optimisation process because each random pipeline is evaluated by both the AVATAR and the T-method. We use the following criteria to compare the performance: (1) the number of invalid/ valid pipelines, (2) the total evaluation time of invalid/ valid pipelines (seconds), (3) the number of pipelines that have the same evaluated results between the AVATAR and the T-method, and (4) the percentage of the pipelines that the AVATAR can validate accurately (%) in comparison to the T-method.

4.2.2. Experiment results

Table 8 compares the performance of the AVATAR and the T-method using the above criteria. We can see that the total evaluation time of invalid/valid pipelines of the AVATAR is significantly lower than the T-method. For example, in the case of using the dataset *convex*, the number of invalid pipelines of the AVATAR and the T-method are approximately the same, 154 in comparison with 159. However, the T-method requires more time than the AVATAR to evaluate these invalid pipelines significantly, 9,655 seconds in comparison with 33 seconds. The number of valid pipelines of the AVATAR and the T-method are also approximately the same, 248 in comparison with 253. However, the T-method requires more time than the AVATAR to evaluate these valid pipelines significantly, 26,170 seconds in comparison with 219 seconds. While the evaluation time of pipelines of the AVATAR is quite stable,

Table 8: Comparison of the performance of the AVATAR and T-method

Dataset	T-method		AVATAR		Pipelines have different/similar evaluated results	The percentage of pipelines that the AVATAR can evaluate accurately (%)
	Invalid/valid pipelines	Total evaluation time of invalid/valid pipelines (s)	Invalid/valid pipelines	Total evaluation time of invalid/valid pipelines (s)		
abalone	553/780	18,609/17,336	537/796	23/27	16/1,317	98.80
adult	441/663	6,041/29,823	430/674	28/169	11/1,093	99.00
amazon	40/60	552/38,053	38/62	18/94	2/98	98.00
car	3,159/4,858	13,561/22,277	3,156/4,861	65/92	3/8,014	99.96
cifar10small	20/31	817/34,939	19/32	96/205	1/50	98.04
convex	159/248	9,655/26,170	154/253	33/219	5/402	98.70
dexter	274/412	2,137/33,528	272/414	54/290	2/684	99.71
dorothea	95/100	4,677/30,354	83/112	293/1,423	12/183	93.85
gcredit	3,616/5,373	16,033/19,755	3,613/5,376	82/125	3/8,986	99.97
gisette	83/122	2,418/33,179	79/126	134/386	4/201	98.05
kddcup	38/30	3,238/33,326	34/34	236/235	4/64	94.12
krvskp	3,106/4,858	14,084/21,703	3,096/4,868	74/132	10/7,954	99.87
madelon	1,032/1,633	5,846/29,893	1,028/1,637	53/209	4/2,661	99.85
mnist	70/135	1,014/36,940	70/135	48/195	0/205	100.00
secom	860/1,337	6,599/32,203	852/1,345	59/203	8/2,189	99.64
semeion	1,141/1,783	5,755/30,123	1,141/1,783	35/86	0/2,924	100.00
shuttle	553/829	9,237/27,299	540/842	19/64	13/1,369	99.06
waveform	2,096/3,201	12,185/23,604	2,031/3,266	62/147	65/5,232	98.77
wineqw	837/1,319	14,246/22,188	823/1,333	24/37	14/2,142	99.35
yeast	2,868/4,527	13,672/22,177	2,849/4,546	60/87	19/7,376	99.74

the evaluation time of pipelines of the T-method is much higher and depends on the size of the datasets. For example, the dataset *kddcup* is larger than *abalone* in terms of the number of instances and the number of attributes. The sum of the evaluation time of invalid and valid pipelines in case of using the T-method with the dataset *kddcup* and *abalone* are approximately the same, 36,564 seconds in comparison with 35,945 seconds. However, the sum of the number of invalid and valid pipelines in case of using the T-method with the dataset *kddcup* and *abalone* are 68 and 1,333 respectively.

It means that the AVATAR is faster than the T-method in evaluating both invalid and valid pipelines regardless of the size of datasets. The reason is that the AVATAR does not execute pipelines with a dataset. The AVATAR maps these pipelines to Petri net based pipelines. After that, the AVATAR finds the validity by firing the transitions of these surrogate pipelines with the input token mapped from the input dataset. These mappings and calculations are less time-consuming than executions of the original pipelines.

In order to prove that the AVATAR and the T-method deliver similar evaluated results, we compare the evaluated results of pipelines between the AVATAR and the T-method. We calculate the accuracy of the AVATAR by the percentage of the number of pipelines that have similar evaluated results over total number pipelines. Table 8 shows that the percentage of pipelines that the AVATAR can evaluate in the same way as T-method is more than 98% on all datasets excluding the cases of using the datasets *dorothea* and *kddcup* for which it is approximately 94%. We have carefully reviewed the pipelines which have different evaluation results between the AVATAR and the T-method. Interestingly, the AVATAR evaluates all of these pipelines to be valid, and the T-method evaluates these pipelines to be invalid. The reason is that executions of these pipelines cause the out of memory problem. In other words,

the AVATAR does not consider the allocated memory as an impact on the validity of a pipeline. To deal with the pipelines that were classified as invalid because of insufficient memory to validate them, a promising solution is to reduce the size of an input dataset by adding a sampling component with appropriate hyperparameters. If the sampling size is too small, we may miss important features. If the sampling size is too large, we may continue to run into the out of memory problem. We cannot conclude that if we allocate more memory, whether the executions of these pipelines would be successful or not. It shows that the validity of a pipeline also depends on its execution environment such as memory. These factors have not been considered yet in the AVATAR. This is an interesting research gap that will be addressed in the future.

Table 9: Five invalid pipelines with the longest evaluation time using the T-method on the dataset *mnist*

Pipeline	#1	#2	#3	#4	#5
T-method (s)	39.779	39.334	30.904	30.418	29.438
AVATAR (s)	0.375	0.016	0.062	0.063	0.032

Finally, we take a detailed look at the invalid pipelines with the longest evaluation time using the T-method on the dataset *mnist*, as shown in Table 9. Pipeline #1 (39.779s) has the structure *Discretize* → *RemovePercentage* → *RemoveUseless* → *RemoveOutliers* → *EMImputation* → *LogitBoost*. All of the attributes of the dataset *mnist* are numeric. The component *Discretize* transforms numeric to nominal attributes in the forms of bins. The components *RemovePercentage*, *RemoveUseless*, *RemoveOutliers* are compatible with nominal attributes, and do not have any effect on nominal attributes. The component *EMImputation* is not compatible with nominal attributes in the output dataset of the component *RemoveOutliers* that leads to an invalid pipeline.

This pipeline is invalid because *EMImputation* does not work with nominal attributes, and there is no component transforming the nominal to numeric attributes. We can see that the AVATAR is able to evaluate the validity of this pipeline without executing it in just 0.375 s.

4.3. Experiments to compare the performance of SMAC with and without the AVATAR

4.3.1. Experiment Settings

The AutoML tool used for the experiments is AutoWeka4MCPS⁸ with integrated AVATAR. We use the same experiment settings as for the first set of experiments. We use SMAC as a pipeline composition and optimisation method. We evaluate the experimental results of the five iterations with and without the AVATAR. We compare the performance of SMAC with and without the AVATAR using the following criteria:

- Mean, standard deviation and min of the error rate; and

⁸<https://github.com/UTS-AAI/autoweka>

- Mean, standard deviation and max of the number of the successfully evaluated pipelines.

4.3.2. Experiment Results

Table 10 presents mean, standard deviation and minimum of the error rate, and mean, standard deviation and maximum of the number of evaluated pipelines of 5 iterations with and without the AVATAR using the datasets presented in Table 3. We can see that using the AVATAR can accelerate the ML pipeline composition and optimisation. Additionally the results support the following observations:

- SMAC can evaluate more pipelines with the AVATAR than without it. The mean of the number of evaluated pipelines, when AVATAR is used, is higher than in the cases without the AVATAR for all datasets.
- SMAC finds better pipelines with the AVATAR than without it. Table 10 shows that mean of the error rate in experiments with the AVATAR is smaller than in the cases without the AVATAR in 15/20 datasets.
- The impact of SMAC’s initialisation on the convergence. Standard deviation of the error rate indicates the degree of the difference of the iterations’ convergences given different initialisation configurations of SMAC. This initialisation randomly selects first configurations based on seed numbers. The high value of this statistical metric indicates that the iterations are not able to converge given the time budget. Table 10 shows that the standard deviation of the error rate for two datasets in both cases with and without the AVATAR is greater than 5%. They are the iterations using the *amazon* and *cifar10small* datasets. It means that the convergence of the results for these datasets highly depends on the initialisation of SMAC because these datasets are large and the time budget for the optimisation is not enough. We have investigated the impact the initialisation of SMAC with multiple configurations in the fourth set of experiments.

Table 11 shows the best pipelines of each iteration with and without the AVATAR. This experiment results confirm that the most promising pipelines are often short pipelines (i.e. the pipelines only have one predictor/meta-predictor). There are two cases where the most promising pipelines include data preprocessing components. They are (kddcup, Iter 1) and (secom, Iter 2). We try to interpret the semantics of the first pipeline structure, *ClassBalancer* → *RemoveOutliers* → *InterquartileRange* → *AttributeSelection* → *Resample* → *PART*. The dataset *kddcup* is imbalanced, therefore, using the ML component *ClassBalancer* has a class balancing effect. Moreover, this dataset is high dimensional, therefore, using ML components such as *InterquartileRange* and *AttributeSelection* can reduce the number of dimensions. This is a large dataset; therefore, *ReSample* can reduce the size of the dataset to accelerate the training of predictive models. Figure 6 presents the histograms of three selected attributes of this

dataset. We see that there are potential outliers which can negatively affect the ML model training. Therefore, using the component *RemoveOutliers* can remove these outliers.

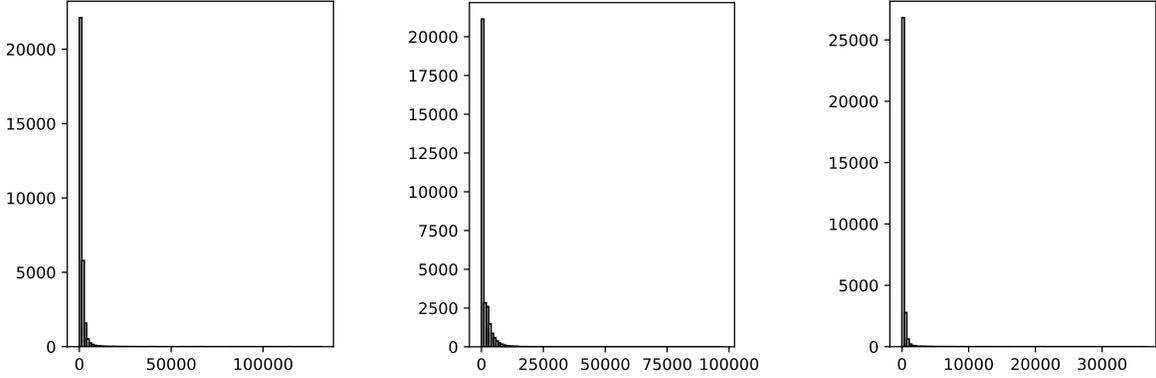


Figure 6: The histograms of three selected attributes of the dataset *kddcup*.

Table 12 shows the error rate and the number of evaluated pipelines of five iterations with and without the AVATAR. We group the experiment results into three main cases:

- *The error rate is lower with the AVATAR* (55/100 iterations) [Denoted by \uparrow]: In these cases, the AVATAR shows its effectiveness to quickly ignore invalid pipelines. Therefore, the number of evaluated pipelines are higher than without the AVATAR. As a result, the convergence of SMAC with AVATAR is faster. An example of these cases is the dataset *cifar10small* with Iter 5. This case clearly shows that the number of evaluated pipelines is higher with the AVATAR than without, 50 compared to 27. As a result, the error rate is lower with the AVATAR, 64.94 compared to 84.29. The results confirm that within the same time budget, the AVATAR allows an ML pipeline and optimisation method (i.e., SMAC) to explore the search space faster than without it.
- *The same error rate with and without the AVATAR* (39/100 iterations) [Do not have either \uparrow or \downarrow]:
 - *The cases with a similar number of evaluated pipelines.* (22/39 iterations): These are cases that have the difference between the number of pipelines less than or equal 10. An example of these cases is the dataset *dorothea* with Iter 5. The error rate of both cases is 9.32%. The number of evaluated pipelines with and without the AVATAR is 22 and 21 respectively. All of the datasets in this group are large and require long training time. Therefore, the number of evaluated pipelines are small. As a result, there is not a significant difference between the cases with and without the AVATAR.

- *The cases with significantly different number of evaluated pipelines. (7/39 iterations)*: These are cases that have the difference between the number of pipelines greater than or equal 100. An example from this group is the dataset *secom* with Iter 1. The error rate of both cases is 6.11%, and the number of evaluated pipelines with and without the AVATAR are 307 and 108 respectively. We have observed that in this group of cases, the convergence to very good solutions was quick with and without the AVATAR because it is quite easy to find well-performing pipelines for these datasets.
- *The error rate is higher with the AVATAR (6/100 iterations)* [Denoted by \downarrow]: These cases are (kddcup, Iter 2), (abalone, Iter 3), (abalone, Iter 4), (adult, Iter 4), (shuttle, Iter 4) and (madelon, Iter 5). What is common for these cases is that they have quite a small number of evaluated pipelines and a similar number of evaluated pipelines. The reason is that the training time of these pipelines is significant. Therefore, there is an element of luck related to the initialisation and the benefits of quick validity check that AVATAR provides cannot be realised.

Table 12 also shows the not completed iterations (i.e., represented by -) where no result is returned. The reason for that is that no valid, fully trained pipeline has been returned by SMAC. In these cases, we do not use these runs to calculate the error rate and relevant metrics. There are 21 and 40 incomplete iterations with and without the AVATAR respectively. The number of incomplete cases is almost double without using the AVATAR. Using the AVATAR has helped SMAC to overcome these situations in 19 out of 40 incomplete iterations when AVATAR was not used.

Figure 7 visualises the convergence of the error rate of the best pipelines found with and without the AVATAR. The lower and upper boundaries representing the min and the max error rates of the five iterations are shown in the form of the shaded area. The solid and the dashed lines represent the mean of the error rate of the best pipelines of the five iterations found at the time points with and without the AVATAR respectively. There are 5/20 cases (*adult*, *car*, *gisette*, *krvskp*, *secom*) for which SMAC has a similar convergence with and without the AVATAR. There are 15/20 cases for which SMAC converges faster by using the AVATAR and 2 hours of optimisation time. In addition, we can see that the shaded areas of SMAC with the AVATAR are narrower and lay lower than without the AVATAR in these 15 cases. It means that the stability of SMAC is better by using the AVATAR. For example, in the case of the dataset *madelon*, SMAC without the AVATAR converges to the error rate between 20% and 50% after 80 minutes, whereas SMAC with the AVATAR converges to the error rate less than 20% after 30 minutes.

Overall, we can observe that the use of AVATAR enables SMAC to find better pipelines faster. A disadvantage of SMAC is that SMAC does not work effectively with large datasets within a limited time budget. We can observe this for a number of cases which have a small number of evaluations (e.g.,

kddcup, Iter2). Although the AVATAR can quickly find valid pipelines in these cases, finding both valid and well-performing pipelines with limited time budget and resources (i.e., virtual machines) still remains a challenge for future studies.

Table 10: Mean, standard deviation (StdDev) and minimum (min) of the error rate, and mean, standard deviation and maximum (max) of the number of evaluated pipelines of AutoWeka4MCPS with (y) and without (n) the AVATAR using 2 hours time budget. The better results are in bold.

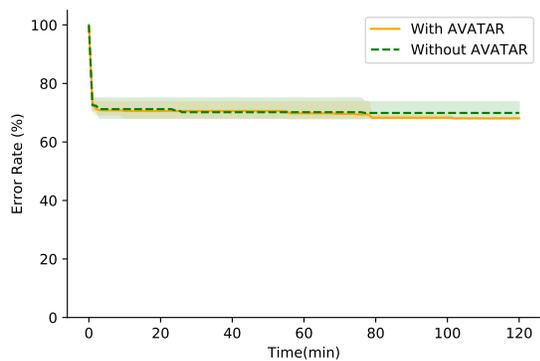
Dataset	% Error Rate						Number of Evaluated Pipelines					
	Mean		StdDev		Min		Mean		StdDev		Max	
	y	n	y	n	y	n	y	n	y	n	y	n
abalone	73.86	73.47	0.94	0.23	72.91	73.22	106	73	73	36	229	126
adult	14.60	14.38	0.60	0.70	13.41	13.41	60	32	28	7	93	45
amazon	46.03	58.57	10.68	6.38	38.48	52.19	44	22	38	14	103	36
car	0.58	2.73	0.30	0.95	0.33	1.07	1136	185	122	77	1269	334
cifar10small	71.94	75.44	8.76	8.86	64.94	66.58	35	20	24	13	67	41
convex	33.52	38.99	5.74	2.39	25.69	35.93	49	24	30	12	82	44
dexter	8.09	8.81	0.85	0.00	6.90	8.81	33	23	23	24	66	69
dorothea	8.02	9.32	1.31	0.00	6.71	9.32	17	12	14	9	42	22
gcredit	22.26	22.83	0.39	0.28	21.71	22.43	808	430	270	229	1298	845
gisette	2.58	2.93	0.19	0.54	2.39	2.39	27	20	20	12	54	39
kddcup	1.80	1.80	0.00	0.00	1.80	1.80	67	30	33	24	108	68
krvskp	0.44	0.67	0.04	0.12	0.40	0.49	898	118	128	40	1071	173
madelon	22.86	26.05	0.22	3.34	22.75	22.75	279	49	59	24	350	85
mnist	16.38	-	0.00	-	16.38	-	23	18	27	17	77	50
secom	6.11	6.11	0.00	0.00	6.11	6.11	277	81	160	47	490	135
semeion	4.95	8.37	0.41	1.98	4.66	5.64	338	54	118	36	515	111
shuttle	0.09	0.03	0.10	0.01	0.03	0.02	91	31	48	29	159	73
waveform	12.53	12.72	0.13	0.06	12.46	12.66	333	49	138	37	492	109
winequality	33.44	37.85	0.38	5.65	33.10	33.39	432	52	208	42	744	117
yeast	38.02	39.15	0.60	0.53	36.96	38.50	1116	136	349	80	1785	222

Table 11: The best pipelines of five iterations with and without the AVATAR.

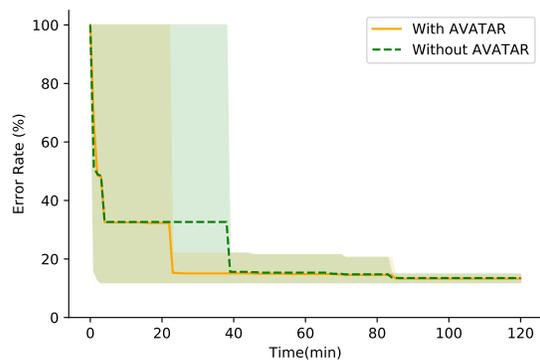
Dataset	Iter 1		Iter 2		Iter 3		Iter 4		Iter 5	
	With AVATAR	Without AVATAR	With AVATAR	Without AVATAR	With AVATAR	Without AVATAR	With AVATAR	Without AVATAR	With AVATAR	Without AVATAR
abalone	SMO	SimpleLogistic	DecisionTable	DecisionTable	NaiveBayes	Logistic	Kstar	SimpleLogistic	-	-
adult	PART	PART	Logistic	Logistic	PART	Logistic	NaiveBayes	DecisionTable	SGD	-
amazon	PART	Jrip	-	-	NaiveBayes Multinomial	-	NaiveBayes Multinomial	NaiveBayes	-	-
car	SMO	LMT	SMO	LMT	SMO	SMO	SMO	SMO	SMO	SMO
cifar10small	-	-	-	-	DecisionStump	-	-	-	RandomForest	DecisionStump
convex	RandomTree	VotedPerceptron	Jrip	Jrip	-	-	-	-	RandomForest	PART
dexter	SGD	NaiveBayes Multinomial	SimpleLogistic	-	-	-	-	-	NaiveBayes Multinomial	-
dorothea	DecisionStump	-	-	-	-	-	-	-	PART	PART
gcredit	RandomForest	NaiveBayes	SMO	NaiveBayes	SMO	NaiveBayes	SMO	SMO	RandomForest	SMO
gisette	VotedPerceptron	VotedPerceptron	-	-	-	-	-	-	VotedPerceptron	SGD
kddcup	ClassBalancer → RemoveOutliers → InterquartileRange → AttributeSelection → Resample → PART	-	-	DecisionStump	ZeroR	DecisionStump	DecisionStump	-	-	-
krvskp	J48	J48	J48	LMT	J48	LMT	J48	J48	J48	RandomForest
madelon	Jrip	REPTree	Jrip	REPTree	Jrip	-	Jrip	Jrip	Jrip	Jrip
mnist	-	-	-	-	PART	-	-	-	-	-
secom	SimpleLogistic	ClassBalancer → EMImputation → Normalize → PrincipalComponent → Kstar	DecisionStump	-	SMO	SMO	RandomForest	DecisionStump	REPTree	J48
semeion	SMO	SMO	SMO	-	SMO	SimpleLogistic	SMO	Multilayer Perceptron	SMO	LMT
shuttle	RandomForest	REPTree	RandomForest	-	DecisionTable	-	REPTree	RandomForest	-	-
waveform	LMT	Logistic	SimpleLogistic	-	SimpleLogistic	-	SimpleLogistic	LMT	SimpleLogistic	SimpleLogistic
winequality	RandomForest	RandomForest	RandomForest	-	RandomForest	-	Kstar	Logistic	Kstar	RandomForest
yeast	RandomForest	-	RandomForest	-	RandomForest	RandomForest	RandomForest	SMO	SMO	SMO

Table 12: Error rate (%) and the number of evaluated pipelines of five iterations with (y) and without (n) the AVATAR. The better results are in bold. \uparrow and \downarrow shows the lower and higher error rate by using the AVATAR respectively.

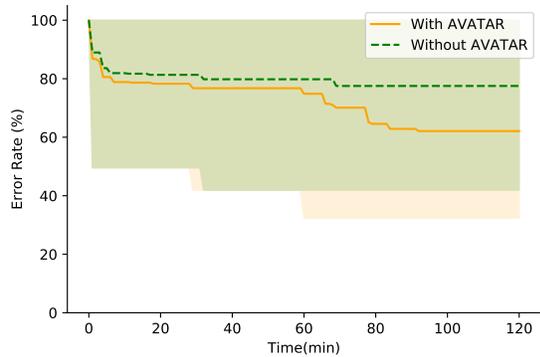
Dataset	Iter 1				Iter 2				Iter 3				Iter 4				Iter 5			
	% error		pipelines		% error		pipelines		% error		pipelines		% error		pipelines		% error		pipelines	
	y	n	y	n	y	n	y	n	y	n	y	n	y	n	y	n	y	n	y	n
abalone	\uparrow 72.91	73.26	229	63	73.60	73.60	132	91	\downarrow 75.41	73.78	48	69	\downarrow 73.50	73.22	99	126	-	-	21	16
adult	13.41	13.41	82	30	14.89	14.89	27	27	\uparrow 14.11	14.89	70	45	\downarrow 14.82	14.33	93	33	\uparrow 15.79	-	28	23
amazon	\uparrow 61.14	64.95	37	35	-	-	10	10	\uparrow 38.48	-	103	26	\uparrow 38.48	52.19	71	36	-	-	1	1
car	\uparrow 0.33	3.88	1209	116	\uparrow 0.41	3.31	1269	149	\uparrow 1.16	2.48	1123	164	\uparrow 0.59	1.07	912	334	\uparrow 0.41	2.89	1167	160
cifar10small	-	-	8	8	66.58	66.58	44	41	\uparrow 84.29	-	67	16	-	-	8	7	\uparrow 64.94	84.29	50	27
convex	\uparrow 35.59	35.93	71	29	39.28	39.28	66	44	-	-	14	7	-	-	10	17	\uparrow 25.69	41.76	82	25
dexter	\uparrow 6.90	8.81	66	69	\uparrow 8.57	-	52	21	-	-	19	4	-	-	4	4	\uparrow 8.81	-	24	19
dorothea	\uparrow 6.71	-	42	22	-	-	12	12	-	-	8	4	-	-	1	1	9.32	9.32	22	21
gcredit	\uparrow 22.86	23.14	503	389	\uparrow 21.71	23.00	861	261	\uparrow 22.00	23.00	686	464	\uparrow 22.29	22.57	691	189	22.43	22.43	1298	845
gisette	2.39	2.39	47	27	-	-	21	19	-	-	8	6	-	-	6	7	\uparrow 2.76	3.47	54	39
kddcup	\uparrow 1.80	-	108	26	\downarrow -	1.80	66	68	1.80	1.80	83	43	\uparrow 1.80	-	73	4	-	-	7	7
krvskp	\uparrow 0.40	0.72	761	56	\uparrow 0.40	0.58	997	173	\uparrow 0.49	0.80	1071	139	\uparrow 0.40	0.76	918	94	0.49	0.49	745	127
madelon	\uparrow 22.75	30.11	253	58	\uparrow 22.75	28.57	326	50	\uparrow 22.75	-	350	10	22.75	22.75	182	44	\downarrow 23.30	22.75	282	85
mnist	-	-	13	13	-	-	11	50	\uparrow 16.38	-	77	11	-	-	13	13	-	-	2	2
secom	6.11	6.11	307	108	\uparrow 6.11	-	37	16	6.11	6.11	490	114	6.11	6.11	164	135	6.11	6.11	385	34
semeion	\uparrow 4.93	5.64	339	111	\uparrow 4.75	-	341	5	\uparrow 5.74	10.03	515	45	\uparrow 4.66	7.34	353	71	\uparrow 4.66	10.48	143	37
shuttle	0.04	0.04	92	73	\uparrow 0.03	-	117	2	\uparrow 0.26	-	72	11	\downarrow 0.03	0.02	159	59	-	-	16	11
waveform	12.80	12.80	84	60	\uparrow 12.46	-	325	10	\uparrow 12.46	-	492	11	\uparrow 12.46	12.66	420	109	\uparrow 12.49	12.69	346	53
winequality	33.39	33.39	582	117	\uparrow 34.18	-	194	2	\uparrow 33.10	-	405	12	\uparrow 33.30	45.82	744	77	\uparrow 33.22	34.33	234	52
yeast	\uparrow 36.96	39.94	776	217	\uparrow 38.59	-	1092	5	\uparrow 37.83	38.50	954	132	\uparrow 38.11	39.27	1785	222	\uparrow 38.60	38.89	972	105



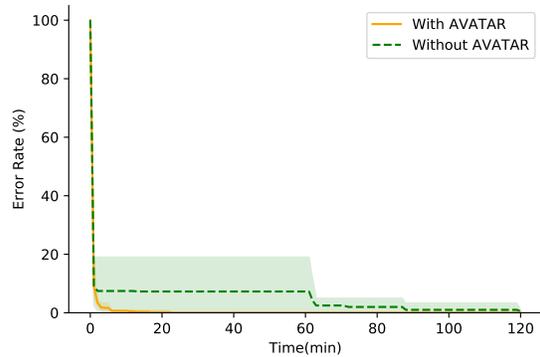
(a) abalone



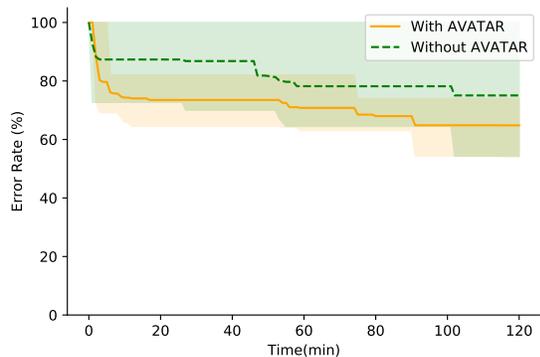
(b) adult



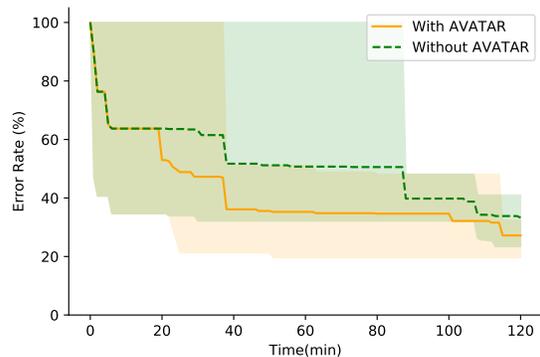
(c) amazon



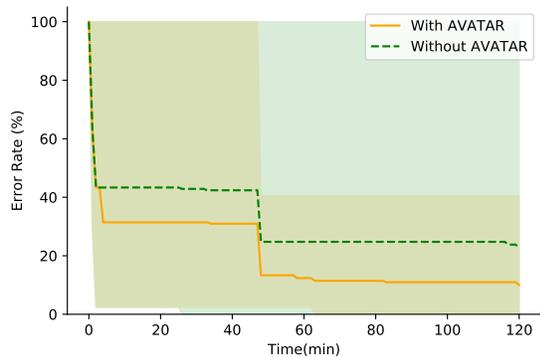
(d) car



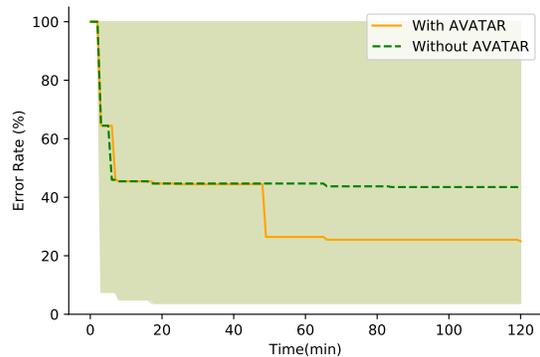
(e) cifar10small



(f) convex

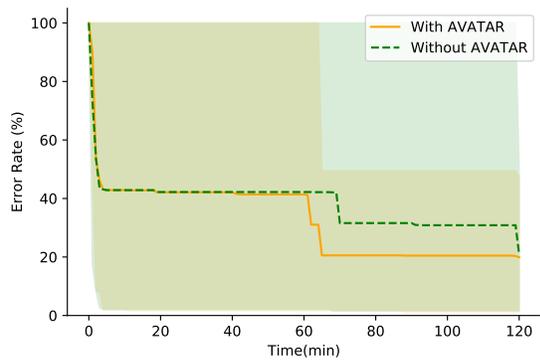


(g) dexter

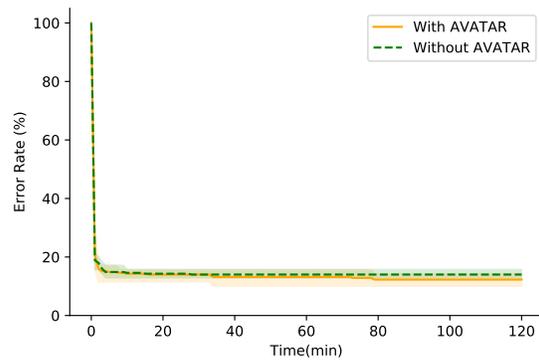


(h) dorothea

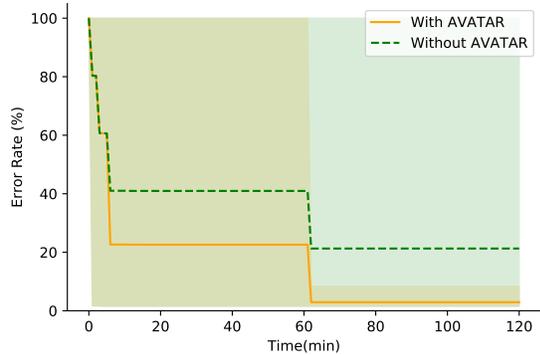
Figure 7: Compare the convergence of the error rate of the best pipelines found with and without the AVATAR.



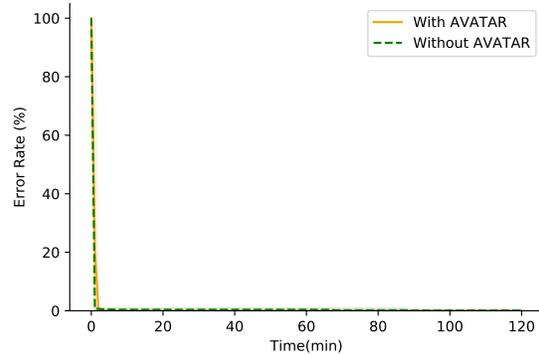
(i) gisette



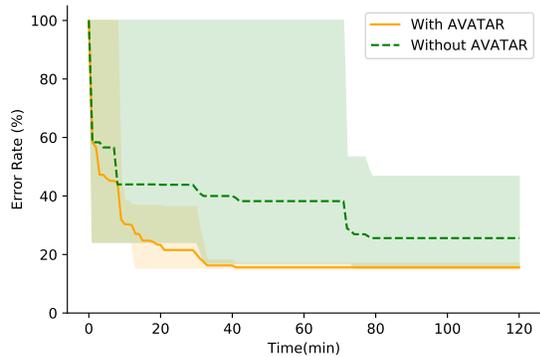
(j) gcredit



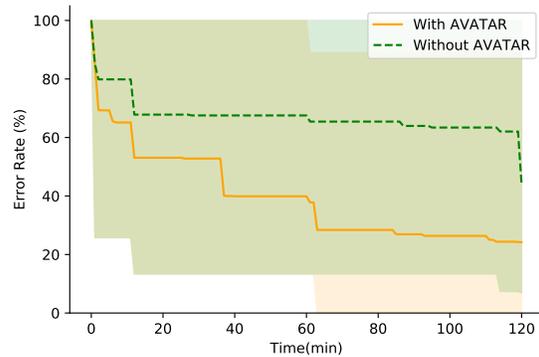
(k) kddcup



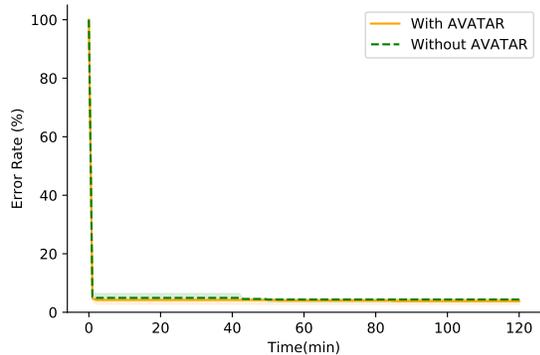
(l) krvskp



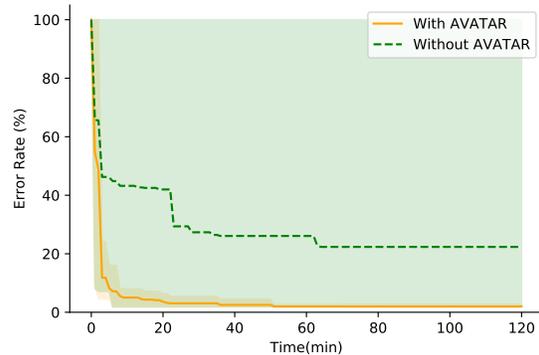
(m) madelon



(n) mnist

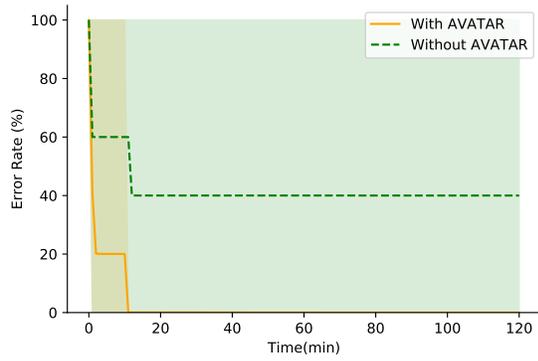


(o) secom

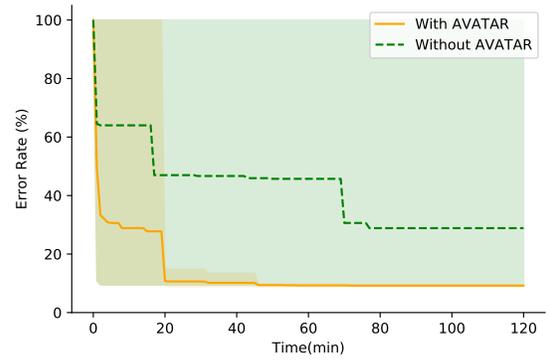


(p) semeion

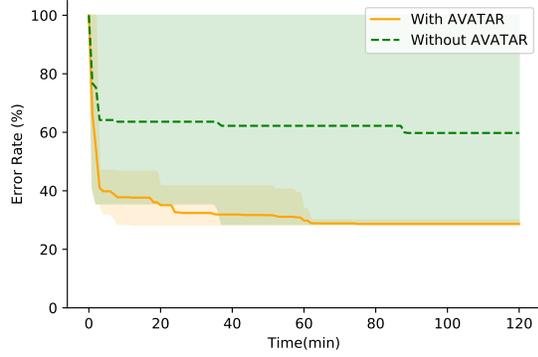
Figure 7: Comparison of the convergence of the error rate of the best pipelines found with and without the AVATAR.



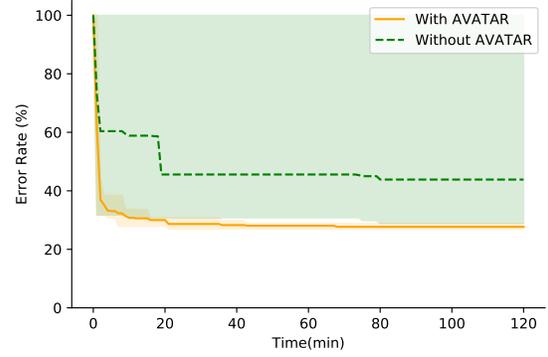
(q) shuttle



(r) waveform



(s) winequality



(t) yeast

Figure 7: Compare the convergence of the error rate of the best pipelines found with and without the AVATAR.

Table 13: Compare mean, minimum (min), difference of mean (dif_mean error), difference of minimum (dif_min error) and standard deviation (StdDev) of the error rate, and mean and standard deviation of the number of evaluated pipelines of SMAC integrated the AVATAR with 1 (SMAC1) and 5 (SMAC5) random initialising configurations. The better results are in bold.

Dataset	Mean error (%)		Min error (%)		Dif_mean error	Dif_min error	StdDev error		Mean pipelines		StdDev pipeline	
	SMAC1	SMAC5	SMAC1	SMAC5			SMAC1	SMAC5	SMAC1	SMAC5	SMAC1	SMAC5
abalone	73.86	73.44	72.91	73.26	0.42	0.35	0.94	0.16	106	161	73	93
adult	14.60	14.27	13.41	13.77	0.33	0.36	0.60	0.46	60	56	28	9
amazon	46.03	46.57	38.48	38.48	0.54	0	10.68	8.10	44	28	38	11
car	0.58	0.38	0.33	0.33	0.2	0	0.30	0.04	1136	1288	122	83
cifar10small	71.94	70.72	64.94	70.45	1.22	5.51	8.76	0.27	35	19	24	10
convex	33.52	33.39	25.69	26.78	0.13	1.09	5.74	4.56	49	44	30	15
dexter	8.09	10	6.9	7.86	1.91	0.96	0.85	3.34	33	75	23	26
dorothea	8.02	7.21	6.71	7.21	0.81	0.5	1.31	0.00	17	20	14	9
gcredit	22.26	22.37	21.71	22	0.11	0.29	0.39	0.23	808	948	270	331
gisette	2.56	3.66	2.39	2.04	1.1	0.35	0.19	2.05	27	35	20	18
kddcup	1.80	1.84	1.80	1.78	0.04	0.02	0.00	0.06	67	35	33	16
krvskp	0.44	0.49	0.40	0.40	0.05	0	0.04	0.08	898	798	128	87
madelon	22.86	23.88	22.75	22.64	1.02	0.11	0.22	1.24	279	200	59	64
mnist	16.38	9.17	16.38	9.17	7.21	7.21	0.00	0.00	23	15	27	10
secom	6.11	6.11	6.11	6.11	0	0	0.00	0.00	277	339	160	145
semeion	4.95	5.32	4.66	4.84	0.37	0.18	0.41	0.57	338	328	118	88
shuttle	0.09	0.02	0.03	0.02	0.07	0.005	0.10	0.00	91	140	48	21
waveform	12.53	12.49	12.46	12.43	0.04	0.03	0.13	0.07	333	357	138	88
winequality	33.44	33.61	33.10	33.04	0.17	0.06	0.38	0.89	432	428	208	148
yeast	38.02	38.09	36.96	37.25	0.07	0.29	0.60	0.56	1116	1171	349	103

4.4. Experiments to investigate the impact the initialisation of SMAC with multiple configurations

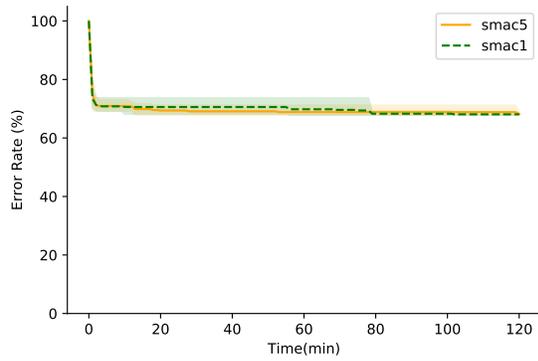
4.4.1. Experiment Settings

Because the AVATAR enables SMAC to evaluate more promising pipelines within the same optimisation time budget and we have observed for a number of datasets that the convergence can be relatively quick, we take this advantage of the AVATAR to conduct the fourth set of experiments to investigate the impact of using multiple initialisations within a single 2-hour optimisation slot. In this set of experiments, we customise SMAC to be initialised with five random configurations instead of one and record the best found pipeline from such five optimisation runs. We use the same datasets as in the previous set of experiments. As previously, we also use SMAC as a pipeline composition and optimisation method. We evaluate the experiment results of the five iterations with and without the AVATAR. As in the previous experiments, we set the total time budget to 2 hours and the memory to 1GB per iteration. We compare the performance of SMAC with the initialisation of one (SMAC1) and five configurations (SMAC5) using the following criteria:

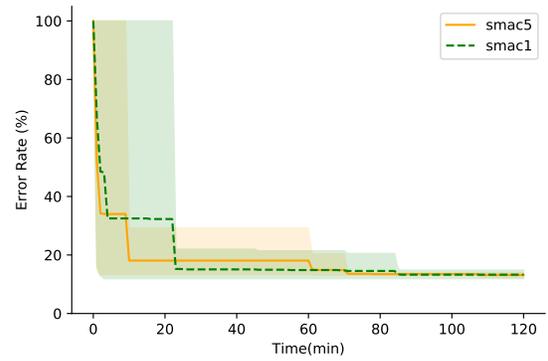
- Mean, standard deviation and min of the error rate;
- Mean, standard deviation and max number of the evaluated pipelines.

4.4.2. Experiment Results

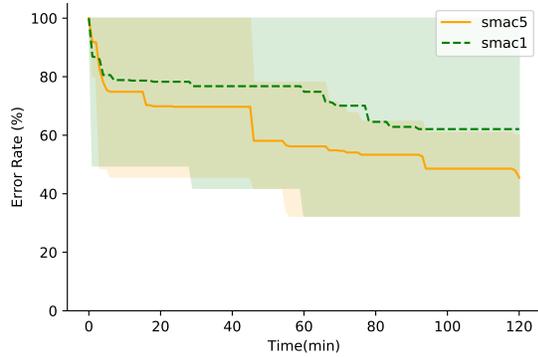
Table 13 compares the performance of SMAC1 and SMAC5. We can see that the number of evaluated pipelines of SMAC1 and SMAC5 is similar. In addition, the quality of the best pipelines found by SMAC1 and SMAC5 are also similar. Mean of the error rate of SMAC5 is higher than, equal and lower than SMAC1 in 11, 1 and 8 out of 20 datasets respectively. Minimum of the error rate of SMAC5 is higher than, equal and lower than SMAC1 in 8, 4 and 8 out of 20 datasets respectively. Therefore, we can conclude that the quality of the best pipelines found by SMAC1 and SMAC5 is similar based on the mean and min error rate of the five iterations. However, the standard deviation of the error rate for SMAC5 is lower than, equal and higher than SMAC1 in 12, 2 and 6 out of 20 datasets. It means that the convergence of SMAC5 is more stable than SMAC1. In other words, the initialisation of SMAC with multiple configurations within a single optimisation run is more likely to result in SMAC finding better ML pipelines in comparison to the single-configuration initialisation regardless of the input seed. Figure 8 visualises the convergence of the error rate of the best pipelines found by SMAC1 and SMAC5. The lower and upper boundary of the shaded area of the SMAC1 and SMAC5 are formed by the min and the max error rate of the five iterations respectively. The dashed and the solid lines represent the mean of the error rate of the best pipelines of the five iterations found at the time points for SMAC1 and SMAC5 respectively. These visualisations confirm that the stability of SMAC5 is better than SMAC1. A good example is the case of using the dataset *kddcup*. Although the SMAC5 and SMAC1 can find



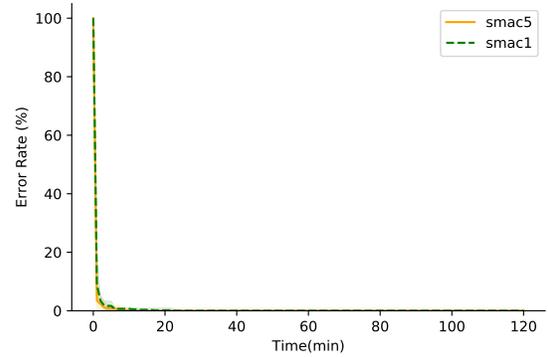
(a) abalone



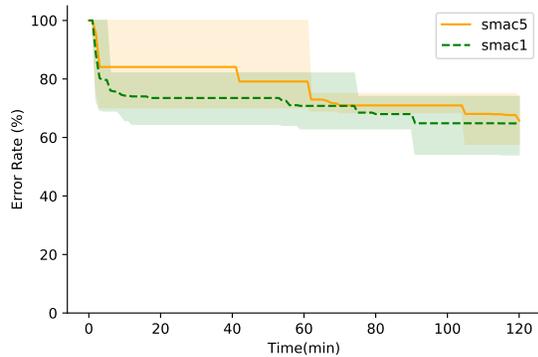
(b) adult



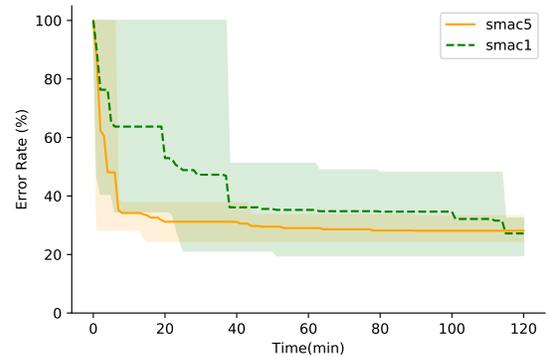
(c) amazon



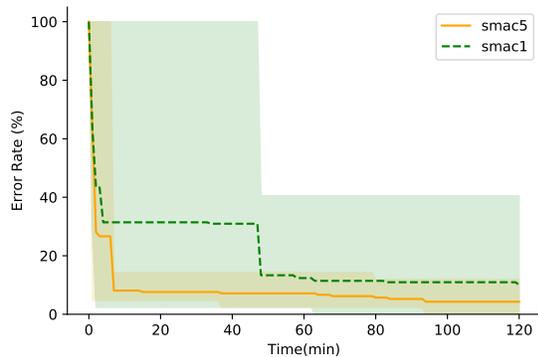
(d) car



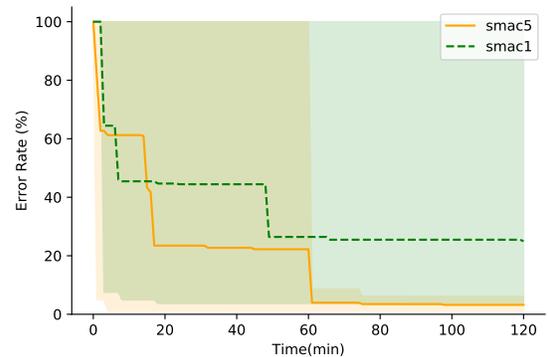
(e) cifar10small



(f) convex

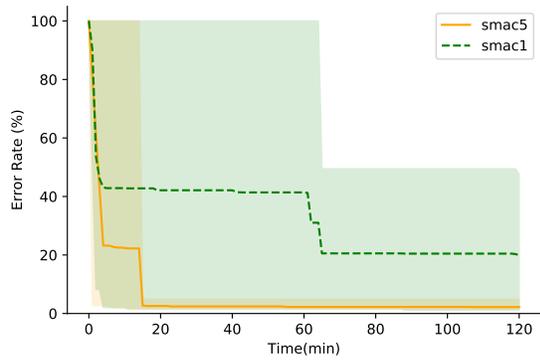


(g) dexter

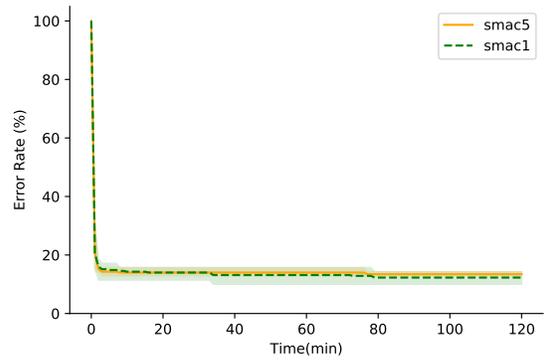


(h) dorothea

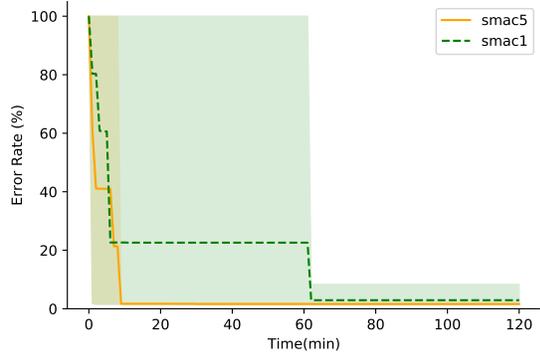
Figure 8: Compare the convergence of the error rate of the best pipelines found by SMAC1 and SMAC5.



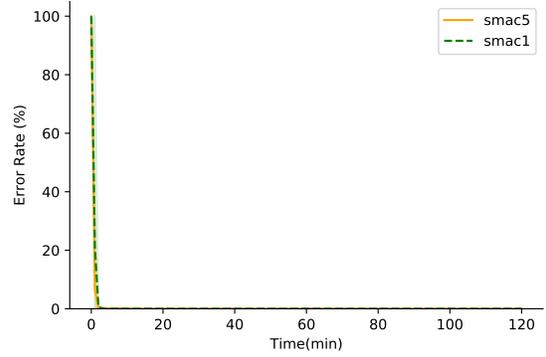
(i) gisette



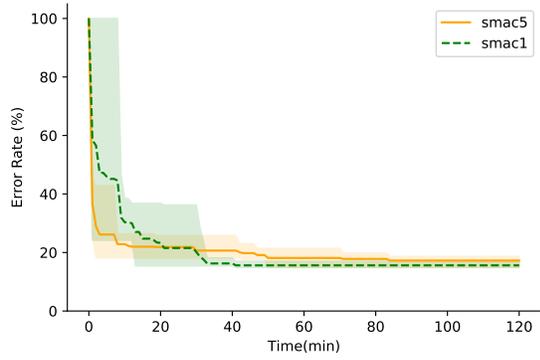
(j) gcredit



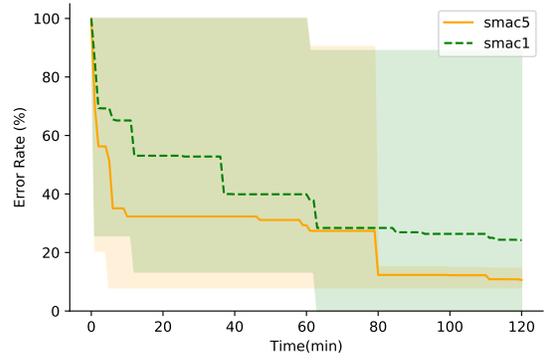
(k) kddcup



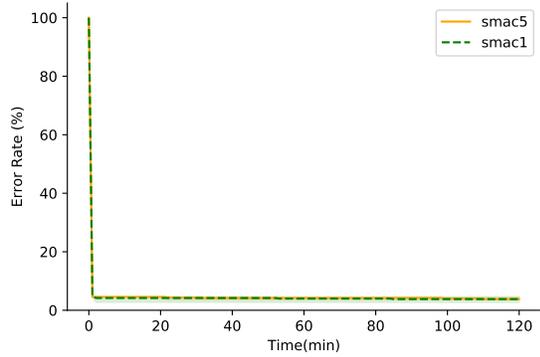
(l) krvskp



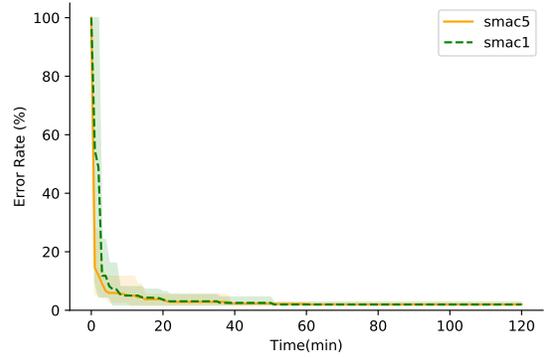
(m) madelon



(n) mnist

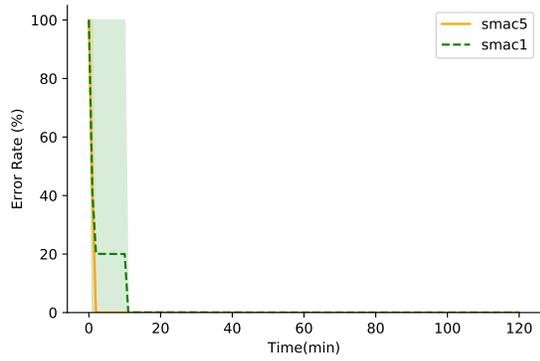


(o) secom

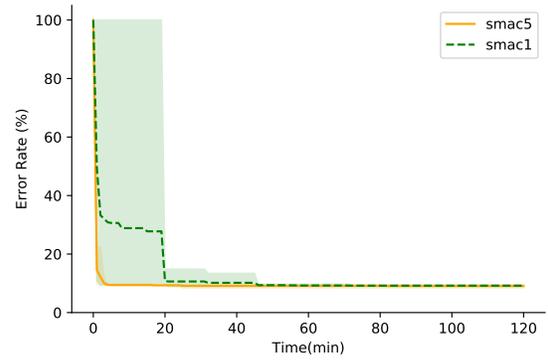


(p) semeion

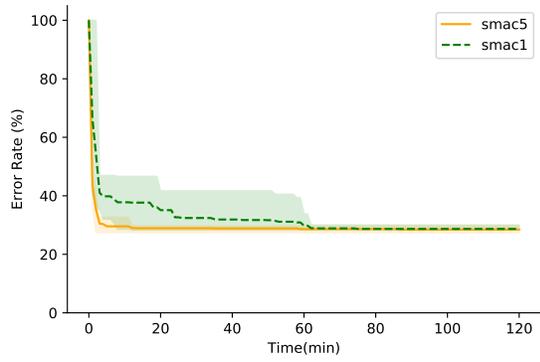
Figure 8: Compare the convergence of the error rate of the best pipelines found by SMAC1 and SMAC5.



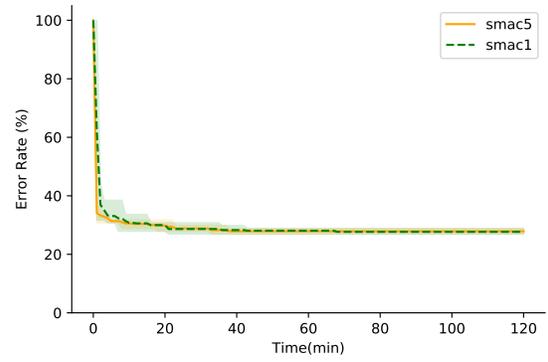
(q) shuttle



(r) waveform



(s) winequality



(t) yeast

Figure 8: Compare the convergence of the error rate of the best pipelines found by SMAC1 and SMAC5.

the pipelines that have similar error rate after 120 minutes, SMAC5 found these pipelines after around 10 minutes, but SMAC1 requires more than 60 minutes to find pipelines that have a similar error rate. Figure 8 also shows that the mean of the error rate of SMAC5 is lower than SMAC1 after the first 30 minutes of the optimisation time in 10 out of the 20 datasets (*abalone*, *amazon*, *convex*, *dexter*, *dorothea*, *gisette*, *kddcup*, *mnist*, *waveform* and *winequality*). For example, in case of the dataset *mnist*, we can see that the mean of the error rate of SMAC5 is lower than SMAC1 after approximately 5 minutes of the time budget. After that, SMAC5 continues to find better pipelines that have lower mean error rate than pipelines found by SMAC1 after 30 minutes of the optimisation time. The reason is that the multiple configurations based initialisation enables SMAC to reduce a chance of selecting bad-performing pipelines at the initialisation stage which then leads to faster convergence of SMAC. The mean of the error rate of SMAC5 is equal to SMAC1 after the first 30 minutes of the optimisation time in further 7 out of 20 datasets (*car*, *gcredit*, *krvskp*, *secom*, *semeion*, *shuttle* and *yeast*). For example, in case of the dataset *gcredit*, we can see that the means of the error rates of SMAC1 and SMAC5 are similar because this dataset is simple for SMAC to find well-performing pipelines regardless of the initialisation. The mean of the error rate of SMAC5 convergence is slower than SMAC1 after the first 30 minutes of the optimisation time in only 3 out of 20 datasets (*adult*, *cifar10small* and *madelon*). This illustrates that using multiple configurations for the initialisation of SMAC is not a good solution in these cases as they require longer continuous optimisation time to find well-performing pipelines. This in fact is pointing to a more fundamental problem of identifying data sets which require longer optimisation time due to their complexity and/or size and better identification of potentially bad initialisation as the key reason for longer convergence times or variability of the quality of the found ML pipelines. This topic remains as one of our future research priorities.

5. Conclusions

In this study, we empirically demonstrated the challenges encountered when generating invalid pipelines during pipeline composition and optimisation process within automated machine learning. These challenges are how to find valid and well-performing pipelines in large configuration spaces. The reason is that the configuration spaces of the pipelines' structures and hyperparameters of the pipelines' components expand significantly with the growing number of available methods and hyperparameters. For example, there are approximately eight billions unique pipelines if we limit the maximum length of a pipeline to eight components with the available methods in WEKA, one of the most popular ML libraries. We showed that the wasted evaluation time of invalid pipelines is significant in the current ML composition and optimisation process. We propose the AVATAR which is a pipeline evaluation method using a surrogate model. The AVATAR can be used to accelerate pipeline composition and optimisation methods by quickly ignoring invalid pipelines to improve the effectiveness of the AutoML

optimisation process. We demonstrate that SMAC can find better ML pipelines by using rather than not using the AVATAR within the same optimisation time. In addition, we leverage the advantage of the AVATAR’s ability to quickly evaluate the validity of promising pipelines to initialise SMAC with multiple configurations. The experiments show that this approach can stabilise the quality of the best pipelines found by SMAC regardless of the random effects of the initialisation. In future, we intend to improve the AVATAR to estimate pipelines’ quality besides their validity. Using various meta-learning approaches, as already illustrated for the purpose of warm start for AutoML in [Feurer et al. \(2015\)](#), will be further explored to achieve that goal.

References

- Balaji, A., Allen, A., 2018. Benchmarking automatic machine learning frameworks. arXiv preprint arXiv:1808.06492 .
- Deb, K., Pratap, A., Agarwal, S., Meyarivan, T., 2002. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE transactions on evolutionary computation* 6, 182–197.
- Feurer, M., Klein, A., Eggenberger, K., Springenberg, J., Blum, M., Hutter, F., 2015. Efficient and robust automated machine learning, in: *Advances in Neural Information Processing Systems*, pp. 2962–2970.
- Gil, Y., Yao, K.T., Ratnakar, V., Garijo, D., Ver Steeg, G., Szekely, P., Brekelmans, R., Kejriwal, M., Luo, F., Huang, I.H., 2018. P4ml: A phased performance-based pipeline planner for automated machine learning, in: *AutoML Workshop at ICML*.
- Kadlec, P., Gabrys, B., 2009. Architecture for development of adaptive on-line prediction models. *Memetic Computing* 1, 241.
- Mohr, F., Wever, M., Hüllermeier, E., 2018a. MI-plan: Automated machine learning via hierarchical planning. *Machine Learning* 107, 1495–1515.
- Mohr, F., Wever, M., Hüllermeier, E., 2018b. MI-plan: Automated machine learning via hierarchical planning. *Machine Learning* 107, 1495–1515.
- Nguyen, T.D., Maszczyk, T., Musial, K., Zöllner, M.A., Gabrys, B., 2020. Avatar-machine learning pipeline evaluation using surrogate model, in: *The Eighteenth International Symposium on Intelligent Data Analysis (IDA 2020)*, Springer.
- Olson, R.S., Moore, J.H., 2016. Tpot: A tree-based pipeline optimization tool for automating machine learning, in: *Workshop on Automatic Machine Learning*, pp. 66–74.

- de Sá, A.G., Pinto, W.J.G., Oliveira, L.O.V., Pappa, G.L., 2017. Recipe: a grammar-based framework for automatically evolving classification pipelines, in: European Conference on Genetic Programming, Springer. pp. 246–261.
- Salvador, M.M., Budka, M., Gabrys, B., 2016a. Adapting multicomponent predictive systems using hybrid adaptation strategies with auto-weka in process industry, in: Workshop on Automatic Machine Learning, pp. 48–57.
- Salvador, M.M., Budka, M., Gabrys, B., 2016b. Towards automatic composition of multicomponent predictive systems, in: International conference on hybrid artificial intelligence systems, Springer. pp. 27–39.
- Salvador, M.M., Budka, M., Gabrys, B., 2017. Modelling multi-component predictive systems as petri nets .
- Salvador, M.M., Budka, M., Gabrys, B., 2018. Automatic composition and optimization of multicomponent predictive systems with an extended auto-weka. *IEEE Transactions on Automation Science and Engineering* .
- Tan, W., Fan, Y., Zhou, M., Tian, Z., 2010. Data-driven service composition in enterprise soa solutions: A petri net approach. *IEEE Transactions on Automation Science and Engineering* 7, 686–694.
- Thornton, C., Hutter, F., Hoos, H.H., Leyton-Brown, K., 2013. Auto-weka: Combined selection and hyperparameter optimization of classification algorithms, in: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM. pp. 847–855.
- Tsakonas, A., Gabrys, B., 2012. Gradient: Grammar-driven genetic programming framework for building multi-component, hierarchical predictive systems. *Expert Systems with Applications* 39, 13253–13266.
- Zöller, M.A., Huber, M.F., 2019. Survey on automated machine learning. arXiv preprint arXiv:1904.12054 .