AVATAR - Machine Learning Pipeline Evaluation Using Surrogate Model

Tien-Dung Nguyen[1(✉)], Tomasz Maszczyk[1], Katarzyna Musial[1], Marc-André Zöller[2], and Bogdan Gabrys[1]

[1] University of Technology Sydney, Sydney, Australia
TienDung.Nguyen-2@student.uts.edu.au,
{Tomasz.Maszczyk,Katarzyna.Musial-Gabrys,Bogdan.Gabrys}@uts.edu.au
[2] USU Software AG, Karlsruhe, Germany
m.zoeller@usu.de

Abstract. The evaluation of machine learning (ML) pipelines is essential during automatic ML pipeline composition and optimisation. The previous 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 requires 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, and it is unnecessary to execute them to find out whether they are good pipelines. To address this issue, we propose a novel method to evaluate the validity of ML pipelines using a surrogate model (AVATAR). The AVATAR enables to accelerate automatic ML pipeline composition and optimisation by quickly ignoring invalid pipelines. Our experiments show that the AVATAR is more efficient in evaluating complex pipelines in comparison with the traditional evaluation approaches requiring their execution.

1 Introduction

Automatic machine learning (AutoML) has been studied to automate the process of data analytics to collect and integrate data, compose and optimise ML pipelines, and deploy and maintain predictive models [1–3]. Although many existing studies proposed methods to tackle the problem of pipeline composition and optimisation [2,4–9], these methods have two main drawbacks. Firstly, the pipelines' structures, which define the executed order of the pipeline components, use fixed templates [2,5]. 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 a variety of structures. Secondly, while evolutionary algorithms based methods [4] enable the randomness of the pipelines' structure using the concept of evolution, this randomness tends to construct more invalid pipelines than valid ones.

M. R. Berthold et al. (Eds.): IDA 2020, LNCS 12080, pp. 352–365, 2020.
https://doi.org/10.1007/978-3-030-44584-3_28

Besides, the search spaces of the pipelines' structures and hyperparameters of the pipelines' components expand significantly. Therefore, the existing approaches tend to be inefficient as they often attempt to evaluate invalid pipelines. There are several attempts to reduce the randomness of pipeline construction by using context-free grammars [8,9] or AI planning to guide the construction of pipelines [6,7]. Nevertheless, all of these methods evaluate the validity of a pipeline by executing them (T-method). After executing a pipeline, if the result is a predictive model, the T-method evaluates the pipeline to be valid; otherwise it is invalid. If a pipeline is complex, the complexity of preprocessing/predictor components within the pipeline is high, or the size of the dataset is large, the evaluation of the pipeline is expensive. Consequently, the optimisation will require a significant time budget to find well-performing pipelines.

To address this issue, 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. We use the business process model and notation (BPMN) [10] to represent ML pipelines. BPMN was invented for the purposes of a graphical representation of business processes, as well as a description of resources for process execution. In addition, BPMN simplifies the understanding of business activities and interpretation of behaviours of ML pipelines. The ML pipelines' components use the Weka libraries[1] for ML algorithms. The evaluation of the surrogate models requires a knowledge base which is generated from many synthetic datasets. To this end, this paper has two main contributions:

- We conduct experiments on current state-of-the-art AutoML tools to show that the construction of invalid pipelines during the pipeline composition and optimisation may lead to bad performance.
- We propose the AVATAR to accelerate the automatic pipeline composition and optimisation by evaluating pipelines using a surrogate model.

This paper is divided into five sections. After the Introduction, Sect. 2 reviews previous approaches to representing and evaluating ML pipelines in the context of AutoML. Section 3 presents the AVATAR to evaluate ML pipelines. Section 4 presents experiments to motivate our research and prove the efficiency of the proposed method. Finally, Sect. 5 concludes this study.

2 Related Work

Salvador et al. [2] proposed 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). This proposed method is implemented in the tool AutoWeka4MCPS [2] developed on top of Auto-Weka 0.5 [11]. The pipelines, which are generated by

[1] https://www.cs.waikato.ac.nz/ml/weka/.

AutoWeka4MCPS, are represented using Petri nets [12]. A Petri net is a mathematical modelling language used to represent pipelines [2] as well as data service compositions [13]. The main idea of Petri nets is to represent transitions of states of a system. Although it is not clearly mentioned in these previous works [4–7], directed acyclic graph (DAG) is often used to model sequential pipelines in the methods/tools such as AutoWeka4MCPS [14], ML-Plan [6], P4ML [7], TPOT [4] and Auto-sklearn [5]. DAG is a type of graph that has connected vertexes, and the connections of vertexes have only one direction [15]. In addition, a DAG does not allow any directed loop. It means that it is a topological ordering. ML-Plan generates sequential workflows consisting of ML components. Thus, the workflows are a type of DAG. The final output of P4ML is a pipeline which is constructed by making an ensemble of other pipelines. Auto-sklearn generates fixed-length sequential pipelines consisting of scikit-learn components. TPOT construct pipelines consisting of multiple preprocessing sub-pipelines. The authors claim that the representation of the pipelines is a tree-based structure. However, a tree-based structure always starts with a root node and ends with many leaf nodes, but the output of a TPOT's pipeline is a single predictive model. Therefore, the representation of TPOT pipeline is more like a DAG. P4ML uses a tree-based structure to make a multi-layer ensemble. This tree-based structure can be specialised into a DAG. The reason is that the execution of these pipelines will start from leaf nodes and end at root nodes where the construction of the ensembles are completed. It means that the control flows of these pipelines have one direction, or they are topologically ordered. Using a DAG to model an ML pipeline makes it easy to understand by humans as DAGs facilitate visualisation and interpretation of the control flow. However, DAGs do not model inputs/outputs (i.e. possibly datasets, output predictive models, parameters and hyperparameters of components) between vertexes. Therefore, the existing studies use ad-hoc approaches and make assumptions about data inputs/outputs of the pipelines' components.

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 uses a fixed pipeline template including preprocessing, predictor and ensemble components. AutoWeka4MCPS also uses a fixed pipeline template consisting of six components. TPOT, ML-Plan and P4ML use grammars/primitive catalogues, 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 wasted time used to evaluate the invalid pipelines is significant. Moreover, using fixed templates, grammars and primitive catalogues reduce search spaces of potential pipelines, which is a drawback during pipeline composition and optimisation.

3 Evaluation of ML Pipelines Using Surrogate Models

Because the evaluation of ML pipelines is expensive in certain cases (i.e., complex pipelines, high complexity pipeline's components and large datasets) in the

context of AutoML, we propose the AVATAR[2] to speed up the process by evaluating their surrogate pipelines. The main idea of the AVATAR is to expand the purpose and representation of MCPS introduced in [12]. 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 transformed-features, and the ML algorithms by transition functions to calculate the output from the input tokens (i.e., the matrices of transformed-features). 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.

3.1 The AVATAR Knowledge Base

We define transformed-features as the features, which represent dataset's characteristics. These characteristics can be changed because of the transformations of this dataset by ML algorithms. Table 1 describes the transformed-features used

Table 1. Descriptions of the transformed-features of a dataset.

Transformed-feature	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

[2] https://github.com/UTS-AAi/AVATAR.

for the knowledge base. We select these transformed-features because the capabilities of a ML algorithm to work with a dataset depend on these transformed-features. These transformed-features are extended from the capabilities of Weka algorithms[3].

The purpose of the AVATAR knowledge base is for describing the logic of transition functions of the surrogate pipelines. The logic includes the capabilities and effects of ML algorithms (i.e., pipeline components).

The capabilities are used to verify whether an algorithm is compatible to work with a dataset or not. For example, whether the linear regression algorithm can work with missing value and numeric attributes or not? The capabilities have a list of transformed-features. The value of each capability-related transformed-feature is either 0 (i.e., the algorithm can not work with the dataset which has this transformed-feature) or 1 (i.e., the algorithm can work with the dataset which has this transformed-feature). Based on the capabilities, we can determine which components of a pipeline (i.e., ML algorithms) are not able to process specific transformed-features of a dataset.

The effects describe data transformations. Similar to the capabilities, the effects have a list of transformed-features. Each effect-related transformed-feature can have three values, 0 (i.e., do not transform this transformed-feature), 1 (i.e., transform one or more attributes/classes to this transformed-feature), or -1 (i.e., disable the effect of this transformed-feature on one or more attributes/classes).

To generate the AVATAR knowledge base[4], we have to use synthetic datasets[5] to minimise the number of active transformed-features in each dataset to evaluate which and how transformed-features impact on the capabilities and effects of ML algorithms[6]. Real-world datasets usually have many active transformed-features that make them not suitable for our purpose. We minimise the number of available transformed-features in each synthetic dataset so that the knowledge base can be applicable in a variety of pipelines and datasets. Figure 1 presents the algorithm to generate the AVATAR knowledge base. This algorithm has four main stages:

1. Initialisation: The first stage initialises all transformed-features in the capabilities and effects to 0.
2. Execution: Run ML algorithms with every synthetic dataset and get outputs (i.e., output datasets or predictive models).
3. Find capabilities: If the execution is successful, we set the active transformed-features of the input dataset for the ones in the capabilities.
4. Find effects: If an algorithm is a predictor/transformed-predictor, we set *PREDICTIVE_MODEL* for its effects. If the algorithm is a filter and its

[3] http://weka.sourceforge.net/doc.dev/weka/core/Capabilities.html.
[4] https://github.com/UTS-AAi/AVATAR/blob/master/avatar-knowledge-base/avatar_knowledge_base.json.
[5] https://github.com/UTS-AAi/AVATAR/tree/master/synthetic-datasets.
[6] https://github.com/UTS-AAi/AVATAR/blob/master/supplementary-documents/avatar_algorithms.txt.

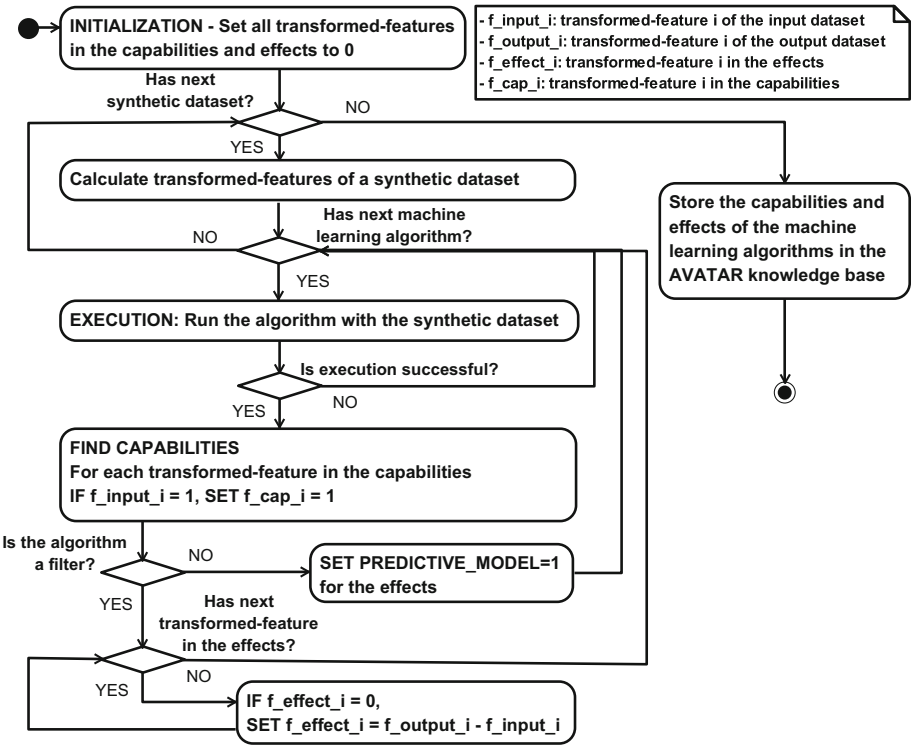

Fig. 1. Algorithm to generate the knowledge base for evaluating surrogate pipelines.

current value is a default value, we set this effect-related transformed-feature equal the difference of the values of this transformed-feature of the output and input dataset.

3.2 Evaluation of ML Pipelines

The AVATAR evaluates a ML pipeline by mapping it to its surrogate pipeline and evaluating this surrogate pipeline. BPMN is the most promising method to represent an ML pipeline. The reasons are that a BPMN-based ML pipeline can be executable, has a better interpretation of the pipeline in terms of control, data flows and resources for execution, as well as integrates into existing business processes as a subprocess. Moreover, 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 simplified ML pipeline.

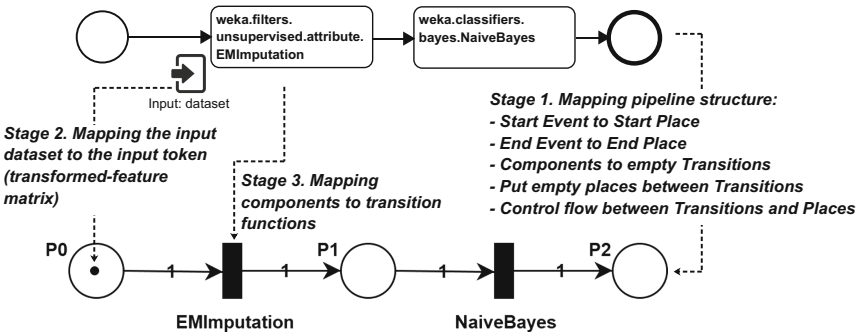

Fig. 2. Mapping a ML pipeline to its surrogate model.

Mapping a ML Pipeline to Its Surrogate Model. The AVATAR maps a BPMN pipeline to a Petri net pipeline via three stages (Fig. 2).

1. The structure of the BPMN-based 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 are put between all transitions. Finally, all flows are mapped to arcs.
2. The values of transformed-features are calculated from the input dataset to form a transformed-feature matrix which is the input token in the start place of the surrogate pipeline.
3. The transition functions are mapped from the components. In this stage, only the corresponding algorithm information is mapped to the transition function.

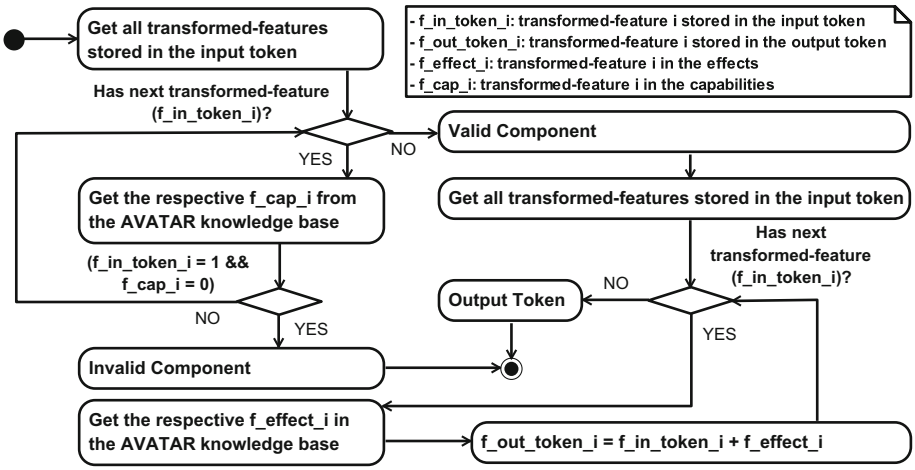

Fig. 3. Algorithm for firing a transition of the surrogate model.

Evaluating a Surrogate Model. The evaluation of a surrogate model will execute a Petri net pipeline. This execution starts by firing each transition of the Petri net pipeline and transforming the input token. As shown in Fig. 3, 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 transformed-feature stored in the input token ($f\_in\_token\_i$) is 1 and the corresponding transformed-feature in the component's capabilities (f_cap_i) is 0, this component is invalid. Otherwise, this component is always valid. If a component is invalid, the surrogate pipeline is evaluated as invalid. The second task calculates each transformed-feature stored in the output token ($f\_out\_token\_i$) in the next place from the input token by adding the value of a transformed-feature stored in the input token ($f\_in\_token\_i$) and the respective transformed-feature in the component's effects (f_effect_i).

4 Experiments

To investigate the impact of invalid pipelines on ML pipeline composition and optimisation, we have first conducted a series of experiments with current state-of-the-art AutoML tools. After that, we have conducted the experiments to compare the performance of the AVATAR and the existing methods.

4.1 Experimental Settings

Table 2 summarises characteristics of datasets[7] used for experiments. We use these datasets because they were used in previous studies [2,4,5]. The AutoML tools used for the experiments are AutoWeka4MCPS [2] and Auto-sklearn [5]. These tools are selected because their abilities to construct and optimise hyper-parameters of complex ML pipelines have been empirically proven to be effective in a number of previous studies [2,5,16]. However, these previous experiments

Table 2. Summary of datasets' characteristics: the number of numeric attributes, nominal attributes, distinct classes, instances in training and testing sets.

Dataset	Numeric	Nominal	No. of distinct classes	Training	Testing
abalone	7	1	26	2,924	1,253
car	0	6	4	1,210	518
convex	784	0	2	8,000	50,000
gcredit	7	13	2	700	300
wineqw	11	0	7	3,429	1,469

[7] https://archive.ics.uci.edu.

had not investigated the negative impact of the evaluation of invalid pipelines on the quality of the pipeline composition and optimisation yet. This is the goal of our first set of experiments. In the second set of experiments, we show that the AVATAR can significantly reduce the evaluation time of ML pipelines.

4.2 Experiments to Investigate the Impact of Invalid Pipelines

To investigate the impact of invalid pipelines, we use five iterations (Iter) for the first set of experiments. We run these experiments on AWS EC2 *t3a.small* virtual machines which have 2 vCPU and 2 GB memory. Each iteration uses a different seed number. We set the time budget to 1 h and the memory to 1 GB. We evaluate the pipelines produced by the AutoML tools using three criteria: (1) the number of invalid/valid pipelines, (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 runtime of the pipeline composition and optimisation. The wasted evaluation time represents the degree of negative impacts of invalid pipelines.

Tables 3 and 4 present negative impacts of invalid pipelines in ML pipeline composition and optimisation of AutoWeka4MCPS and Auto-sklearn using the above criteria. These tables show that not all of constructed pipelines are valid. Because AutoWeka4MCPS can compose pipelines which have up to six components, it is more likely to generate invalid pipelines and the evaluation time

Table 3. Negative impacts of invalid pipelines in pipeline composition and optimisation of 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)	16/26	90/79	69/88	34/29	53/80
	(2)	3607.7/1322.5	2007.1/1236.4	4512.9/2172.3	3615.4/277.6	23.2/3509.0
	(3)	73.18	61.88	67.51	92.87	0.66
car	(1)	205/152	108/70	197/313	139/156	85/64
	(2)	3818.1/291.8	3498.5/113.0	4523.6/532.6	5232.2/251.3	4365.1/90.1
	(3)	92.90	96.87	89.47	95.42	97.98
convex	(1)	18/20	2/0	17/11	crashed	crashed
	(2)	76.3/3588.1	3475.2/0.0	1324.7/2331.8		
	(3)	2.08	100.00	36.23		
gcredit	(1)	112/195	229/364	208/166	12/54	30/54
	(2)	2821.0/2260.1	3829.8/285.6	3933.8/184.0	3667.6/34.1	3634.8/64.7
	(3)	55.52	93.06	95.53	99.08	98.25
wineqw	(1)	203/213	121/139	crashed	201/302	36/54
	(2)	4880.6/1052.9	4183.4/1078.6		2418.5/1132.2	1639.2/862.2
	(3)	82.26	79.50		68.11	65.53

Table 4. Negative impacts of invalid pipelines in pipeline composition and optimisation of 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
car		crashed	crashed	crashed	crashed	crashed
convex	(1)	2/13	2/6	2/8	2/6	2/8
	(2)	560.8/2981.8	537.7/629.2	584.1/1537.5	558.1/977.1	560.0/1655.9
	(3)	15.76	15.07	16.39	15.66	15.72
gcredit		crashed	crashed	crashed	crashed	crashed
wineqw	(1)	0/42	0/22	0/42	0/32	0/32
	(2)	0.0/3523.4	0.0/909.7	0.0/3197.4	0.0/3054.0	0.0/3163.5
	(3)	0.00	0.00	0.00	0.00	0.00

of these invalid pipelines are significant. For example, the wasted evaluation time is 97.98% in the case of using the dataset car and Iter 5. We can see that changing the different random iterations 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 0.66% and 92.87%. The reason is that Weka libraries them-self 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, this impact is negligible in the case of Auto-sklearn. The reason is that Auto-sklearn uses meta-learning to initialise with promising ML pipelines. The experiments with the datasets abalone, car and gcredit show that Auto-sklearn limits the generation of invalid pipelines by making assumption about cleaned input datasets, because the experiments crash if the input datasets have multiple attribute types. It means that Auto-sklearn can not handle invalid pipelines effectively.

4.3 Experiments to Compare the Performance of AVATAR and the Existing Methods

In order to demonstrate the efficiency of the AVATAR, we have conducted a second set of experiments. We run these experiments on a machine with an Intel core i7-8650U CPU and 16 GB memory. We compare the performance of the AVATAR and the 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 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

Table 5. Comparison of the performance of the AVATAR and T-method

Dataset		abalone	car	convex	gcredit	winequality
T-method	Invalid/valid pipelines	683/ 1,097	4,387/ 6,817	252/ 428	4,557/ 7,208	1,276/ 1,951
	Total evaluation time of invalid/valid pipelines (s)	27,711.9/ 15,484.1	18,627.9/ 24,459.4	5,818.3/ 37,765.1	19,597.9/ 23,452.5	10,830.1/ 32,326.9
AVATAR	Invalid/valid pipelines	663/ 1,117	4,387/ 6,817	250/ 430	4,552/ 7,213	1,262/ 1,965
	Total evaluation time of invalid/valid pipelines (s)	3.5/4.9	43.1/64.8	19.6/131.1	57.0/89.2	17.1/25.4
Pipelines have different/similar evaluated results		20/1,760	0/11,204	2/678	5/11,760	14/3,213
The percentage of pipelines that the AVATAR can validate accurately (%)		98.88	100.00	99.71	99.96	99.57

of the pipelines because a valid pipeline always has a predictor at the end. Each pipeline is evaluated by the AVATAR and the T-method. We set the time budget to 12 h per dataset. We use the following criteria to compare the performance: the number of invalid/valid pipelines, the total evaluation time of invalid/valid pipelines (seconds), the number of pipelines that have the same evaluated results between the AVATAR and the T-method, and the percentage of the pipelines that the AVATAR can validate accurately (%) in comparison to the T-method.

Table 5 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. While the evaluation time of pipelines of the AVATAR is quite stable, the evaluation time of pipelines of the T-method is much higher and depends on the size of the datasets. It means that the AVATAR is faster than the T-method in evaluating both invalid and valid pipelines regardless of the size of datasets. Moreover, we can see that the accuracy of the AVATAR is approximately 99% in comparison with the T-method. We have carefully reviewed the pipelines which have different evaluated results between the AVATAR and the T-method. Interestingly, the AVATAR evaluates all of these pipelines to be valid and vice versa in the case of the T-method. 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. 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 large, we may continue to run into the problem of out of memory. We cannot conclude that if we allocate more memory, whether the executions of these pipelines would be successful or not. It proves 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 should be addressed in the future.

Table 6. Five invalid pipelines with the longest evaluation time using the T-method on the gcredit dataset.

Pipeline	#1	#2	#3	#4	#5
T-method (s)	11.092	11.068	11.067	11.067	11.066
AVATAR (s)	0.014	0.012	0.011	0.011	0.011

Finally, we take a detailed look at the invalid pipelines with the longest evaluation time using the T-method on the gcredit dataset, as shown in Table 6. Pipeline #1 (11.092 s) has the structure *ReplaceMissingValues* → *PeriodicSampling* → *NumericToNominal* → *PrincipalComponents* → *SMOreg*. This pipeline is invalid because *SMOreg* does not work with nominal classes, and there is no component transforming the nominal to numeric data. We can see that the AVATAR is able to evaluate the validity of this pipeline without executing it in just 0.014 s.

5 Conclusion

We empirically demonstrate the problem of generation of invalid pipelines during pipeline composition and optimisation. 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. In future, we will improve the AVATAR to evaluate pipelines' quality besides their validity. Moreover, we will investigate how to employ the AVATAR to reduce search spaces dynamically.

Acknowledgment. This research is sponsored by AAi, University of Technology Sydney (UTS).

References

1. Kadlec, P., Gabrys, B.: Architecture for development of adaptive on-line prediction models. Memetic Computing **1** (2009). https://doi.org/10.1007/s12293-009-0017-8. Article number. 241

2. Salvador, M.M., Budka, M., Gabrys, B.: Automatic composition and optimization of multicomponent predictive systems with an extended auto-WEKA. IEEE Trans. Autom. Sci. Eng. **16**(2), 946–959 (2019)
3. Zöller, M.A., Huber, M.F.: Survey on automated machine learning. arXiv preprint arXiv:1904.12054 (2019)
4. Olson, R.S., Moore, J.H.: TPOT: a tree-based pipeline optimization tool for automating machine learning. In: Workshop on Automatic Machine Learning, pp. 66–74 (2016)
5. Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M., Hutter, F.: Efficient and robust automated machine learning. In: Advances in Neural Information Processing Systems, pp. 2962–2970 (2015)
6. Mohr, F., Wever, M., Hüllermeier, E.: ML-Plan: automated machine learning via hierarchical planning. Mach. Learn. **107**, 1495–1515 (2018). https://doi.org/10.1007/s10994-018-5735-z
7. Gil, Y., et al.: P4ML: a phased performance-based pipeline planner for automated machine learning. In: AutoML Workshop at ICML (2018)
8. de Sá, A.G.C., Pinto, W.J.G.S., Oliveira, L.O.V.B., Pappa, G.L.: RECIPE: a grammar-based framework for automatically evolving classification pipelines. In: McDermott, J., Castelli, M., Sekanina, L., Haasdijk, E., García-Sánchez, P. (eds.) EuroGP 2017. LNCS, vol. 10196, pp. 246–261. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-55696-3_16
9. Tsakonas, A., Gabrys, B.: GRADIENT: grammar-driven genetic programming framework for building multi-component, hierarchical predictive systems. Expert Syst. Appl. **39**, 13253–13266 (2012)
10. Chinosi, M., Trombetta, A.: Modeling and validating BPMN diagrams. In: 2009 IEEE Conference on Commerce and Enterprise Computing, pp. 353–360. IEEE (2009)
11. Thornton, C., Hutter, F., Hoos, H.H., Leyton-Brown, K.: 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, pp. 847–855. ACM (2013)
12. Salvador, M.M., Budka, M., Gabrys, B.: Modelling multi-component predictive systems as Petri nets (2017)
13. Tan, W., Fan, Y., Zhou, M., Tian, Z.: Data-driven service composition in enterprise SOA solutions: a Petri net approach. IEEE Trans. Autom. Sci. Eng. **7**, 686–694 (2010)
14. Martin Salvador, M., Budka, M., Gabrys, B.: Towards automatic composition of multicomponent predictive systems. In: Martínez-Álvarez, F., Troncoso, A., Quintián, H., Corchado, E. (eds.) HAIS 2016. LNCS (LNAI), vol. 9648, pp. 27–39. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-32034-2_3
15. Barker, A., van Hemert, J.: Scientific workflow: a survey and research directions. In: Wyrzykowski, R., Dongarra, J., Karczewski, K., Wasniewski, J. (eds.) PPAM 2007. LNCS, vol. 4967, pp. 746–753. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-68111-3_78
16. Balaji, A., Allen, A.: Benchmarking automatic machine learning frameworks. arXiv preprint arXiv:1808.06492 (2018)