Chapter 13
Final Remarks and Further Work for the Hybrid-AutoML System

ABSTRACT

This chapter addresses that the various use cases have proved that the aims and contributions of this research to conceptualise, design, and develop a scalable and flexible toolkit for automatic big data ML mode and model selection, on single or multi-varying datasets has been achieved. A major benefit of the hybrid-autoML toolkit is that it reduces the time data scientists and researchers in the field spend, searching through the algorithm selections and hyper parameter space. This advantage was discussed in Section 5.2 where the authors compared the hybrid-autoML tool with autoWeka on about 35 datasets using measures such as accuracy, mean absolute error (MAE), and time.

1. CONCLUSION

In this research, we have presented a toolkit for automatic machine learning (ML) mode and model selection on single or multi-varying datasets.

First, the basic concept of big data ML, ML tools, the algorithm selection problem, the meta-learning (learning-to-learn) paradigm and automated machine learning (autoML) was discussed. We discussed that although some hybrid autoML systems exists, e.g. autoWeka and auto-Sklearn, they do not consider knowledge known about mode selection but focus mainly on the supervised learning space for model selection. Some on one hand do not determine the importance and influence that knowledge of data sets meta features have over the choice of selecting the best ML mode and model automatically. Lastly, none of the known autoML system allows for automatic mode and model selection on multi-varying datasets at the same time. However, the hybrid-autoML system and functions designed in this research eliminates all that by taking them into consideration appropriately.

DOI: 10.4018/978-1-7998-7316-7.ch013

Second, provides more details and discussions from the literatures, that show the link between big data classification or clustering, the Meta learning paradigm, and how generic knowledge obtained about a dataset or about supervised and unsupervised learning, can be used to design a set of functions for automatic ML mode selection and model building on single or multi-varying datasets.

Third, we show and discuss some preliminary experimentations carried out in this research, using Weka (a well known data mining tool in the research community). The purpose of the pre experiments carried out, was to prove, properly identify and define the problems identified from previous discussions of literatures reviewed in chapter 2. The knowledge gained from this pre experiments helped define the rules for the hybrid-autoML system's model and design. The rule based functions modelled, takes into account the execution semantics for automatic ML mode and model selection.

Fouth, reported on the implementation details of the hybrid-autoML, visualisations, simulations and analysis. More specifically, we discussed and showed the design architecture (design consisting of three layers), components, testing strategy and materials of hybrid-autoML, and provided the relevant algorithms.

The toolkit named hybrid-autoML is an open source project that can be retrieved from github and easily used or extended. Hybrid-autoML provides a simple graphical user interface that facilitates automated ML mode and models selection, visualisation or evaluation and prediction capabilities.

Fifth, we addressed the unfolding of hybrid-autoML by evaluating its performance using 5 practical use cases and well known statistical and non-statistical measures. Based on the performance results of the experiments, a variety of observations are made. For example, use case 1 in section 5.1.1 shows an unsupervised mode and a simple and lightweight autoProb clustering function desgined in this research is chosen authomatically, for building a model on a small unlabelled dataset. While use case 2 in section 5.1.2 shows an unsupervised ML mode with a readily available EM clustering algorithm selected automatically for building a model on a larger unlabeled dataset. Use cases 3 and 5 from sections 5.1.3 and 5.1.5, proves that the hybrid-autoML tool knows when to automatically use a supervised ML mode to build an appropriate model on multi-varying datasets in the shortest time possible as compared to conventional autoWeka.

In conclusion, the various use cases have proved that the aims and contributions of this research to conceptualise, design, and develop a scalable and flexible toolkit for automatic big data ML mode and model selection, on single or multi-varying datasets has been achieved. A major benefit of the hybrid-autoML toolkit is that it reduces the time data scientists and researchers in the field spend, searching through the algorithm selections and hyper parameter space. This advantage was discussed in section 5.2 where we compared the hybrid-autoML tool with autoWeka on about 35 datasets using measures such as: accuracy, mean absolute error (MAE) and time.

2. FUTURE WORK

- Expanding the rule based function for model selection to accommodate more practical use case scenarios and algorithms, to further improve the automatic decision learning process.
- Expand the rules to accomodate better automatic data cleansing strategies before the automatic mode and model selection is performed.
- Considering the challenges of big data, incorporate some big data processing methods such as parallel processing to further optimize the process.