Automated Machine Learning for Internet of Things

Che-Min Chung, Cai-Cing Chen, Wei-Ping Shih, Ting-En Lin, Rui-Jun Yeh, and Iru Wang, *MoBagel Inc.*

Abstract – ***This paper presents Decanter AI, a new approach to machine learning that uses automated machine learning techniques to resolve the massive data problem in the rapidly industry of the Internet of Things (IoT). This solution is specialized in IoT data and applied to a real-world example of a smart building with over 100 connected sensors and its performance is compared to industry benchmarks.***

I. INTRODUCTION

By 2020, the world will have over 50 billion IoT devices connecting to each other and streaming over 60ZB data. Sufficient background in data science is necessary to uncover insights from this colossal amount of IoT data due to its complex natures of data variety, veracity, volume, and velocity. However, according to McKinsey's report, by 2018, the United States alone could face a shortage of 190,000 data scientists, meaning that a lot of IoT decision makers will be sitting on gold mines without the necessary tools to explore it.

Automated machine learning is a direct solution to the shortage of data scientists as it can drastically increase the performance and productivity of data scientists by speeding up work cycles, improving model accuracy, and ultimately, even possibly replace the need for data scientists.

II. THE ALGORITHMS

In this paper, we propose Decanter AI: a novel approach to automatically analyze IoT data using semi-supervised machine learning techniques. Decanter AI can precisely select the most optimal algorithms that can be used to build a predictive model. Decanter AI automates the data analytics workflow from feature engineering, algorithms selection, model building, parameters tuning to evaluating results based on our experience from analyzing data from various industries.

Decanter AI has more than 95 built-in algorithms and integrates with the latest open source machine learning algorithms including regression algorithms, instance-based algorithms, regularization algorithms, tree based algorithms, bayesian algorithms, clustering algorithms, dimensionality reduction algorithms, and ensemble algorithms. When a user defines a problem in Decanter AI, Decanter AI automatically builds a predictive model based on the user's problem. When IoT devices stream data into Decanter AI, Decanter AI would efficiently search for thousands of possible combinations of algorithms based on the statistical characteristics of the data in order to find the best algorithm for the user's data set and forecasting goals. For IoT data that does not have labels, Decanter AI can automatically assign classes for data with similar behaviors using clustering.

Decanter AI generates associated features by decomposing and restructuring the data and uses feature selection to find an optimal feature set by calculating the influence of features. Decanter AI defines lots of states according to recognized patterns and statistical characteristics of user's data. It can automatically take appropriate actions to train predictive models for each data state, tune parameters using grid search, evaluate models by cross validation referring to multiple statistical indicators, and finally rank all the performances. Using reinforcement learning, Decanter AI keeps all the training history to improve the policies to each data state of auto-modeling system.

Decanter AI uses the best model set to build an ensemble predictive model that can dynamically refine itself to make accurate real-time predictions. It changes the composition of the model set according to the training data due to the fact that the characteristics of IoT data might change as devices upload more data. This incremental approach of IoT data analysis is capable of capturing the behavior of large numbers of connected IoT devices based on time and location. The incremental analysis significantly reduces the real-time prediction error of current analysis methods.

III. EXPERIMENTATION

For the experiment, we use a smart building with over 100 connected motion sensors and collect real-time data that indicates the occupancy of every desk, space, and designated areas within the building. Data is collected for 6 months, and is used to predict the occupancy status of every hour in the next month. Accurate prediction of the occupancy at any given time in different areas of the smart building enables more efficient management of air conditioning, electricity, maintenance, thus leading to lower energy bills and repair costs.

For the prediction model, we used Decanter AI to automatically build and train the best model and compared the results to industry benchmarks of Amazon Machine Learning and Google Prediction on metrics of Adjusted R-squared, RMSE, and training time. This experiment is conducted using a 10-fold cross validation, where the original sample is randomly partitioned into 10 equal size subsamples. Each subsample is then trained using Decanter AI, Amazon Machine Learning, and Google Prediction and the results are shown below, which includes the minimum, maximum, and average of the training results.

978-1-5090-4017-9/17/$31.00 ©2017 IEEE

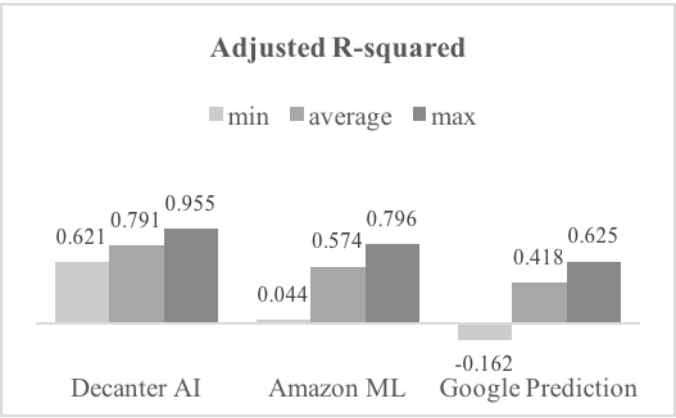

Fig. 1. Comparison of adjusted R-squared

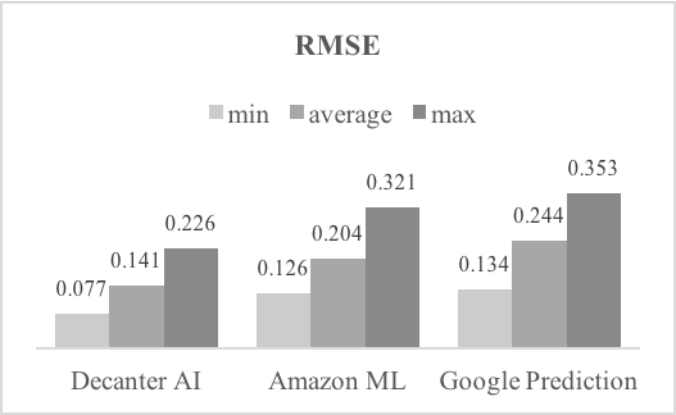

Fig. 2. Comparison of RMSE

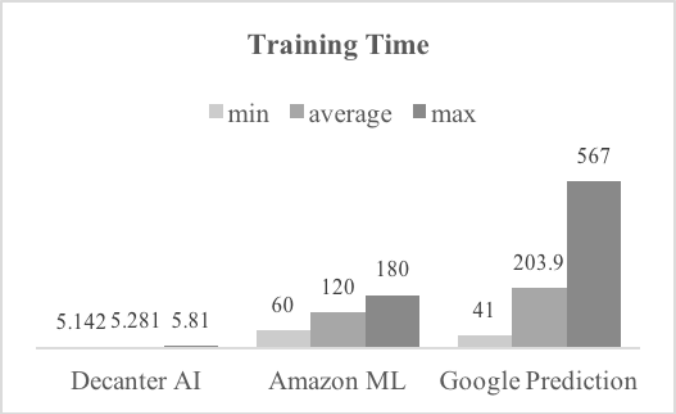

Fig. 3. Comparison of training time (in seconds)

TABLE I
Average Benchmark Results for IoT

Metrics	Decanter AI	Amazon ML	Google Prediction
Adjusted R-squared	0.791	0.574	0.418
RMSE	0.141	0.204	0.244
Training Time (second)	5.281	120	203.9

Based on the benchmark results, Decanter AI performed better in all three metrics of Adjusted R-squared, RMSE and training time compared to Amazon Machine Learning and Google Prediction. In addition, we discovered that Decanter AI's training time is much shorter than the other two services; however, since we do not know their precise computing specifications, this difference in training time is for reference only.

IV. CONCLUSION

According to the experiment result, the performance of our system, Decanter AI, is better than other popular semi-automatic data analytics platform. Decanter AI could reduce huge amount of time and increase the efficiency of data scientists due to the automated analytic process and pre-model building. Our system now could be the basis for data analytic beginner, but a domain expert is still needed for defining the question itself. Thus, the ultimate goal for us is to create a system that could help data scientists from all industries to build a basic model based on machine learning quickly. Moreover, the accuracy of prediction would be kept tuning by accumulating more domain knowledge and enriching in feature engineering procedure. In that sense, we would be able to recognize more patterns from IoT data, to find the best analytic process for each kind of data, and to really achieve the goal of fully automated machine learning. Also, we want to provide some useful guidelines for improving the quality of pattern recognition in IoT data analytics, and to suggest related open issues to foster research in this area.

V. REFERENCES

[1] B. Biggio G. Fumera F. Roli "Pattern recognition systems under attack: Design issues and research challenges" International Journal of Pattern Recognition and Artificial Intelligence vol. 28 no. 07 2014.

[2] M. Bailey J. Oberheide J. Andersen Z. M. Mao F. Jahanian and J. Nazario "Automated classification and analysis of internet malware " in Proceedings of the 10th international conference on Recent advances in intrusion detection Gold Goast Australia 2007 pp. 178-197.

[3] K. Rieck T. Holz C. Willems P. Düssel and P. Laskov "Learning and Classification of Malware Behavior " in Detection of Intrusions and Malware and Vulnerability Assessment vol. 5137 Springer Berlin /Heidelberg 2008 pp. 108-125.

[4] K. Rieck P. Trinius C. Willems and T. Holz "Automatic analysis of malware behavior using machine learning " Journal of Computer Security vol. 19 no. 4 pp. 639-668 Jan. 2011.

[5] Ronghua Tian L. Batten R. Islam and S. Versteeg "An automated classification system based on the strings of trojan and virus families " in 2009 4th International Conference on Malicious and Unwanted Software (MALWARE) 2009 pp. 23-30.

[6] A. Zarras A. Papadogiannakis R. Gawlik T. Holz "Automated Generation of Models for Fast and Precise Detection of HTTP-Based Malware" Annual Conference on Privacy Security and Trust (PST) 2014.

[7] S. Zander T. Nguyen G. Armitage "Automated traffic classification and application identification using machine learning" IEEE 30th Conference on Local Computer Networks (LCN 2005) 2005-November.

[8] J. Manyika, M. Chiu, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. Byers "Big data: The next frontier for innovation, competition, and productivity" McKinsey Global Institute May. 2011

[9] Ryan Elwell, Robi Polikar, "Incremental Learning of Concept Drift in Nonstationary Environments" IEEE 2011