Challenges of Automated Machine Learning on Causal Impact Analytics for Policy Evaluation

Prof. (Dr.) Yuh-Jong Hu and Shu-Wei Huang
hu@cs.nccu.edu.tw, wei.90211@gmail.com

Emerging Network Technology (ENT) Lab.
Department of Computer Science
National Chengchi University, Taipei, Taiwan

IEEE 2nd Int. Conf. TEL-NET-2017, Noida, India
Outline

1. INTRODUCTION

2. RESEARCH GOALS AND RESULTS

3. AUTOML FOR CLASSIFICATION AND REGRESSION

4. AUTOML FOR CAUSAL IMPACT ANALYTICS

5. A SIMPLIFIED CAUSAL INFERENCE MODEL

6. CONCLUSION AND FUTURE WORK

7. REFERENCES
Motivations

1. The main goals of big data analytics are to determine *correlation*, *prediction*, and *cause-effect* among high-dimensional data features.

2. Automated machine learning (AutoML) refers to the full aspects of automated machine learning without human in the analytics loop.

3. Until now, AutoML systems were primarily proposed for classification and regression, but lacked causal impact analytics.

4. We address the possible challenges of extending AutoML on causal impact analytics for a policy evaluation.
Intensive computation and storage requirements to satisfy the resources’ need of ETL processing and model construction.

Balancing the bias-variance while searching for an optimal model from an enormous amount of features combinations.

Need to address the problems of data heterogeneity, noise accumulation, spurious correlations, and incidental endogeneity.

Three levels of data analytics goals: correlation, prediction, and causal inference. However, correlation does not imply causation.

AutoML provides services to achieve three levels of big data analytics goals without human intervention in the processing loop.
Outline

1. Introduction
2. Research Goals and Results
3. AutoML for Classification and Regression
4. AutoML for Causal Impact Analytics
5. A Simplified Causal Inference Model
6. Conclusion and Future Work
7. References
Aiming at the following research goals:

1. How do we construct an AutoML system through optimization of a machine learning algorithm selection with its hyperparameter?
2. How can we apply causal inference on high-dimensional datasets?
3. What are the major challenges to extend current AutoML systems for causal model discovery and causal inference?
4. How can a real policy evaluation scenario be implemented to justify the feasibility of using a (simplified) causal impact analytics technique?
Contributions

1. We have established an extended AutoML system with causal impact analytics services in the AWS/EC2 cloud computing environment.

2. We have investigated the major challenges and obstacles on establishing a full set of AutoML systems for causal inference.

3. We have implemented a real policy evaluation scenario on stock market, i.e., S&P, impacts analytics by using the GDELT world’s collective news media datasets.
Related Work

- Scikit-learn[8]
- Auto-WEKA 2.0[9]
- BSTS[6][21]
- Spark ML Pipeline
Outline

1 Introduction

2 Research Goals and Results

3 AutoML for Classification and Regression

4 AutoML for Causal Impact Analytics

5 A Simplified Causal Inference Model

6 Conclusion and Future Work

7 References
Why is an AutoML System so Hard to Build?

- Need to address automatic ETL data preprocessing while facing data heterogeneity, sparsity, missing values, and noise accumulation.
- Numerous machine learning algorithms are available for selection on an optimal model construction. But, which one is the best?
- Once a specific machine learning algorithm is selected, the next challenge will be a search of optimal hyperparameter.
- We should provide an automated machine learning (AutoML) system with least human in the analytics processing loop.
AutoML can be formalized as a *Combined Algorithm Selection and Hyperparameter optimization (CASH)* problem.

Find the joint machine learning algorithm and hyperparameter setting that minimizes the loss function over the training and validation datasets.

Four major steps for AutoML systems:

1. ETL data preprocessing
2. Meta-learning for a model selection
3. Optimal hyperparameter and parameter set tuning
4. Optimal model testing and selection
Outline

1. Introduction
2. Research Goals and Results
3. AutoML for Classification and Regression
4. AutoML for Causal Impact Analytics
5. A Simplified Causal Inference Model
6. Conclusion and Future Work
7. References
Three approaches have been proposed for structural causal models:

1. Potential outcomes for counterfactuals analysis [19].
2. Structured equation models [6].
3. Causal graph for probability reasoning and causal analysis [3].

In this study, a type of simplified structured equation model, Bayesian Structural Time Series (BSTS), is used for causal impact analytics [6].

Google’s CausallImpact R package provides BSTS computation capability.
Challenges of Causal Impact Analytics for AutoML

- Only preprocessing observational datasets through ETL might not be enough for a causal model search and inference.
- How to smoothly integrate three structural causal model techniques into current AutoML systems is still unknown.
- The performance of automated causal model discovery algorithms is difficult to evaluate.
- We must apply several assumptions when computing the causal impact probability density.
- When the vertexes number of possible causal graphs increases, the number of directed acyclic graphs (DAGs) increase exponentially.
Outline

1. Introduction
2. Research Goals and Results
3. AutoML for Classification and Regression
4. AutoML for Causal Impact Analytics
5. A Simplified Causal Inference Model
6. Conclusion and Future Work
7. References
In the pre-intervention period, *train* a treatment group’s time series pattern from a control group viewpoint.

Then, in the *validation* phase, a cross-validation techniques is applied on a treatment group time series pattern, given a set of observation data only.

In the intervention and post-intervention periods, we predict what would happen for a treatment group’s unobserved *counterfactuals*.

Compare a treatment group’s real observed data to unobserved predicted counterfactuals with a control group’s learning model for a specific intervention factor, i.e., a policy evaluation.

BSTS technique is applied for a policy evaluation.
A Simplified Causal Inference Model

Bayesian Structural Time Series (BSTS)

- AutoML system is initiated at the pre-intervention period through rolling window Random Forest (RF) regression algorithm on the Spark ML pipeline system.
- In the intervention and post-intervention phases, the BSTS technique is manually applied for a policy evaluation to discover the counterfactuals and average treatment effects.
- BSTS uses a Markov Chain Monte Carol (MCMC) algorithm for posterior inference of simulated regression outcomes with its hyperparameter.
- We have built a simplified causal inference model for Spark ML pipeline on the AWS/EC2 cloud platform.
Spark ML Pipeline on the AWS/EC2 Platform
AutoML Pipeline for a Causal Inference Model

Auto ML Pipeline

Preprocessing
- Category Encoding
- Long to Wide Converting
- Standardization
- Features Selection

Modeling
- Decision Tree Regression
- Gradient Boosting Regression
- Random Forest Regression

Hyperparameter Tuning
- Param, Grid Tuning
- Cross Validation
- 1-day ahead Prediction

Causal Impact Analyzing
- BSTS Model
- Rolling RF Model
- Treatment Group
- Control Group

Merge while stock market opened
A Detailed Machine Learning (ML) Pipeline Stages

ML Pipeline

- Labeled Training set
  - pipeline.fit
  - fit & transform

- Vector Assembler
  - Feature processing
  - transform

- Vector Indexer
  - fit & transform

- RF Regressor
  - Model Training & Hyperparameter Tuning
  - predict

- Predictive Model
  - Class labels

- Testing set
  - pipeline.predict
A Policy Evaluation Scenario for Causal Impact Analytics

- Global Database of Events, Language, and Tone (GDELT) datasets, the world’s collective news media, are used for predicting S&P 500 stock market index variation with a specific intervention factor.
- GDELT datasets are public available via Google’s BigQuery; they are created through TABARI system based on taxonomy of CAMEO event data types.
- A policy evaluation refers to the concepts of news media report about the Occupy Wall Street (OWS) events’ influences.
- The US government economic stimulating policies were evaluated to confirm that the OWS events have positive influences on S&P 500 stock market indexes.
**Figure:** A machine learning pipeline with 1-day ahead prediction of S & P stock market indexes
Figure: The Occupy Wall Street (OWS) events on the influences of S & P stock market indexes
**Figure:** The counterfactuals of Random Forest (RF) regression algorithm with tree sizes 2000 and 200
Figure: The prediction errors of the rolling-Random Forest (RF) regression model by using 15-minute GDELT streaming datasets.
Conclusions

● Preliminary Results:
  1. Present the possible research challenges on tackling the big data analytics problem.
  2. Show the emerging techniques to achieve the vision of AutoML system without human in the analytics processing loop.
  3. Address the potential research challenges of empowering causal impact analytics in the AutoML system.
  4. A real policy evaluation scenario has been implemented on the AWS/EC2 cloud platform by using GDELT datasets with positive impacts of stock market prediction.

● Future Work:
  1. Exploiting on the seamless integration of automatic causal structure discovery and inference into the fully loaded AutoML systems for big data analytics on the public cloud platform.
Conclusions

Preliminary Results:

1. Present the possible research challenges on tackling the big data analytics problem.
2. Show the emerging techniques to achieve the vision of AutoML system without human in the analytics processing loop.
3. Address the potential research challenges of empowering causal impact analytics in the AutoML system.
4. A real policy evaluation scenario has been implemented on the AWS/EC2 cloud platform by using GDELT datasets with positive impacts of stock market prediction.

Future Work:

1. Exploiting on the seamless integration of automatic causal structure discovery and inference into the fully loaded AutoML systems for big data analytics on the public cloud platform.
Conclusions

- **Preliminary Results:**
  1. Present the possible research challenges on tackling the big data analytics problem.
  2. Show the emerging techniques to achieve the vision of AutoML system without human in the analytics processing loop.
  3. Address the potential research challenges of empowering causal impact analytics in the AutoML system.
  4. A real policy evaluation scenario has been implemented on the AWS/EC2 cloud platform by using GDELT datasets with positive impacts of stock market prediction.

- **Future Work:**
  1. Exploiting on the seamless integration of automatic causal structure discovery and inference into the fully loaded AutoML systems for big data analytics on the public cloud platform.
Conclusions

Preliminary Results:

1. Present the possible research challenges on tackling the big data analytics problem.
2. Show the emerging techniques to achieve the vision of AutoML system without human in the analytics processing loop.
3. Address the potential research challenges of empowering causal impact analytics in the AutoML system.
4. A real policy evaluation scenario has been implemented on the AWS/EC2 cloud platform by using GDELT datasets with positive impacts of stock market prediction.

Future Work:

1. Exploiting on the seamless integration of automatic causal structure discovery and inference into the fully loaded AutoML systems for big data analytics on the public cloud platform.
Conclusions

Preliminary Results:

1. Present the possible research challenges on tackling the big data analytics problem.
2. Show the emerging techniques to achieve the vision of AutoML system without human in the analytics processing loop.
3. Address the potential research challenges of empowering causal impact analytics in the AutoML system.
4. A real policy evaluation scenario has been implemented on the AWS/EC2 cloud platform by using GDELT datasets with positive impacts of stock market prediction.

Future Work:

1. Exploiting on the seamless integration of automatic causal structure discovery and inference into the fully loaded AutoML systems for big data analytics on the public cloud platform.
Outline

1 Introduction
2 Research Goals and Results
3 AutoML for Classification and Regression
4 AutoML for Causal Impact Analytics
5 A Simplified Causal Inference Model
6 Conclusion and Future Work
7 References


