Obesity Risk Factors Ranking Using Multi-Task Learning

Lu Wang¹, Dongxiao Zhu¹, Elizabeth Towner², Ming Dong¹
1: Department of Computer Science, Wayne State University, Detroit, MI 48202
2: Department of Family Medicine & Public Health Sciences, Wayne State University, Detroit, MI 48202

Introduction
Obesity is one of the leading preventable causes of death in the United States (U.S.). Risk factor analysis is a process to identify and understand the risk factors contributing to a particular disease, and is an imperative component in the development of efficient and effective prevention and intervention efforts. Most existing methods usually aim to build a one-size-fits-all model to identify the risk factors at the population-level. However, this type of methods do not take into consideration of population heterogeneity.

Solution
To overcome this limitation, we employ the framework of multi-task learning (MTL) to identify a ranked list of obesity risk factors for each subpopulation/participant (task) simultaneously with utilizing appropriate shared information across tasks. By learning multiple related tasks at the same time, MTL provides a paradigm to rank risk factors at multi-levels.

Method
All tasks share a common set of features, $\ell_2,1$-norm regularized multi-task learning is able to capture the task relatedness from multiple related tasks:

$$\min_W \mathcal{L}(W) + \lambda \|W\|_{2,1}$$

- $W \in \mathbb{R}^{G \times T}$: weights.
- $T$: number of tasks, $t$: index of task, $g$: index of feature and $G$: number of continuous features.

$$\mathcal{L}(W) = \sum_{t=1}^{T} \|W^t X_t - Y_t\|_2^2$$

where $X_t$ is the input matrix of the $t^{th}$ task and $Y_t$ is the corresponding output variable.

$$\|W\|_{2,1} = \sum_{g=1}^{G} \|W_{g, 1}^t\|_1$$

the group sparse penalty, $\lambda$: tuning parameter.

Data

<table>
<thead>
<tr>
<th>Behavioral risk factor surveillance system (BRFSS)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Participants</strong></td>
</tr>
<tr>
<td><strong>Data collection</strong></td>
</tr>
<tr>
<td><strong>Multi-faced</strong></td>
</tr>
<tr>
<td><strong>Input variables</strong></td>
</tr>
<tr>
<td><strong>Output variable</strong></td>
</tr>
</tbody>
</table>

Table 1. Descriptions of data we use for experiments.

Conclusion

- MTL:
  - Capture relatedness of tasks;
  - Knowledge transfer among tasks ➔ Improve the data sufficiency;
  - Simultaneously learn all the related tasks;

- Multiple ranked list of obesity risk factors.

References


Contact
Dongxiao Zhu, Ph.D.
Associate Professor
http://www.cs.wayne.edu/dzhu/
dzhu@wayne.edu
Deep Reinforcement Learning for Optimizing Carpooling Policies

Ishan Jindal\textsuperscript{1}, Zhiwei Qin\textsuperscript{2}, Xuewen Chen\textsuperscript{1}, Matthew Nokleby\textsuperscript{1} and Jieping Ye\textsuperscript{2}  
\textsuperscript{1}Wayne State University, \textsuperscript{2}DiDi Research America

Analytic Approach: Learn Q-value function using the experiences generated by the Carpool Simulator.

Introduction

We developed a deep reinforcement learning (RL) based system to learn an efficient policy for rideshare carpooling that maximizes the driver’s revenue and minimizes traffic congestion using fewer cars to fulfill the same amount of demand. For this purpose, we developed a detailed carpooling simulator using the NYC taxi trip dataset. We observed that in the densely populated regions such as in Downtown Manhattan, it is always favorable to do a carpool. On the other hand, for the less populated areas such as Uptown Manhattan, RL learned policy significantly outperforms the fixed policy.

Challenges

- Development of carpooling simulator for generating training samples for Reinforcement Learning. Since, not availability of any carpooling datasets.
- End-to-End learning framework for a joint estimation of travel time and distance.
- Developed a system to predict efficient ridesharing policies for maximizing the drivers’ revenue.

Travel Time Estimation

Spatial-Temporal Neural Network (ST-NN) for travel time and distance estimation

The beauty of ST-NN is that it uses only the raw taxi trip data without requiring further feature engineering and provides a joint estimate of travel time and distance.

Policy Search (DQN)

Fixed Policy: Always favors Carpool.
DQN Policy: A more dynamic policy depends on the location and time of the driver generated using Deep Q networks.

MDP Formulation

State: (Lat, Lon, Time-of-the-day)  
Action: [Wait, Take 1, Carpool]  
Reward: Effective distance travelled: sum of the distances travelled by the driver on individual trips  
Episode: 0:00 AM $\rightarrow$ 23:59 PM  
State Transition: $T_0 \rightarrow T_1$

MDP Formulation

Policy Search (DQN)

Fixed Policy: Always favors Carpool.
DQN Policy: A more dynamic policy depends on the location and time of the driver generated using Deep Q networks.

Results and Observations

1) DQN Vs. Fixed Policy

Fixed Policy: Total Reward = 53.41  
DQN Policy: Total Reward = 73.01

2) Policy evaluation at different locations

Uptown Manhattan  
Fixed Policy: 340.06  
DQN: 339.42

Downtown Manhattan  
Fixed Policy: 41.5  
DQN: 46.07

More revenue at dense taxi call regions

Conclusion

- Learned RL policy always favors choosing an action which takes the taxi into such that the driver earns more revenue dense taxi call regions by carpooling which shows good promise towards a fully intelligent carpooling decision system.
- Comparing to baseline DQN policy always generate optimized policies to maximize driver’s reward.

Contact

Ishan Jindal, GTA (ECE)  
Information Processing Lab, Wayne State University  
E-Mail: ishan.jindal@wayne.edu
INTRODUCTION
Recently, data-driven methods have been increasingly applied to different domains ranging from business intelligence to medical diagnosis and manufacture design. Resistance Spot Welding (RSW) datasets can be used to build models, extract knowledge, and accelerate weld design process. Some welding quality prediction system [1] have been proposed by using shift-off Decision Tree and Random Forest algorithms in a single machine with small datasets. DATAVIEW [2], composing machine learning algorithms operators, such as Random Forest, Support Vector Machine, Artificial Neural Network, can predict weld quality in a high performance computing resources such as Amazon clouds, which makes the prediction process more efficient with large datasets.

Motivation
Resistance spot welding is a complex thermomechanical process with high veracity data and highly non-linear relation between process parameters. In Big Data scenarios, traditional implementation platforms like R are inefficient. DATAVIEW, a big data infrastructure, handling big data analyzing process with scaling the computing resource in both horizontal and vertical way, is more reliable and efficient.

OBJECTIVES
1- Design and development of implementation pipelines for predicting the weld quality (weld nugget width dimension) with R and DATAVIEW, respectively.
2- Performing statistically designed experiments to compare the efficiency of R and DATAVIEW in different data size scenarios.

METHODS
The experiment detail is shown below:
1) Design the two pipelines in R and DATAVIEW by using Random Forest algorithm to predict the nugget width of the joints.
2) Evaluate the implementation time (training and testing) in seconds.
3) Design of Experiment (DOE):
   • The experiment is a full factorial experiment.
   • There are two factors: Implementation Framework, Size of Dataset.
   • Execution Framework at two levels: R, DATAVIEW
   • Size of Dataset: at four levels 1 (514 kB), 2 (2515 kB), 3 (15,355 kB), 4 (50,000 kB).
R failed to analyze datasets 3 and 4, for this reason two approach adopted to perform DOE.
Scenario (1): Treating Failed results as NA and analyzing an unbalanced DOE.
Scenario (2): Assuming that R did not fail, implemented the process but at very high costs compared to DATAVIEW (10 times of costs of DATAVIEW). This is in fact penalizing R for not being able to implement the task efficiently. Note that the main purpose of this scenario is conducting statistical analysis on the results.

RESULTS
Scenario (1): By conducting unbalanced DOE the sum of Squares of the factors evaluated and partitioned.
• Both Execution Framework and Size of Dataset are statistically significant factors.
• Interaction between the factors is not statistically significant.

Scenario (2): By conducting this scenario, a balanced DOE will be returned and the Analysis of Variance (ANOVA) can be implemented.

CONCLUSIONS
Implementation platforms like R are not suitable for Big Data scenarios and computationally high-capacity frameworks are required. Results of this research show that at Execution Framework and Size of Dataset their interactions are statistically significant factors. Especially, interaction between execution platform and size of dataset is more important at bigger datasets. DATAVIEW outperforms R from the perspective of mean implementation time.

REFERENCES

ACKNOWLEDGEMENTS
• Digital Manufacturing and Design Innovation Institute (DMDII), Ford Motor Company, WSU

CONTACT
Changxin Bai, PhD Candidate Department of Computer Science 1; Big Data Lab Department of Industrial Engineering 2
changxin@wayne.edu
Anomaly detection; smart meter data; occupancy monitoring

**ABSTRACT**

Abnormal inactivities including immobilizing medical conditions or sudden deaths within a single-occupied household are usually not easy to be detected. The delayed discovery of such emergencies poses realistic concerns worldwide. In this paper, a novel algorithm is proposed to detect the inactivities based on smart meter readings. The pattern of power consumption variations is explored by distilling features of the data with the state-of-the-art probabilistic methods. The algorithm is able to cold start from limited historical data and perform well without extended parameter tuning, and therefore has potential to be implemented in a distributed embedded system.

**METHODOLOGY**

Home energy usage pattern

We assume that lighting and cooking constitute the main source of electricity consumption. By subsetting the Smart* Home Dataset [1], Figure 1 shows the electricity consumption in Home A over a 24-hour period. The occupant’s normal activity can be indicated by the intermittent usage of lighting and cooking appliances. Our study focuses on the main circuit readings.

**Qualitative analysis**

Figure 2 shows the normal power consumption curve (black) and three abnormal power consumption curves.

![Figure 2](image-url)

The grey area depicts the 95% confidence interval for the same-time meter readings in the past D=30 days, which is estimated by the following method. The expected consumption level

\[
\mu_{X_t,D} = \frac{1}{D} \sum_{d=1}^{D} x_{t-d}
\]

The standard deviation

\[
\sigma_{X_t,D} = \left( \frac{1}{D-1} \sum_{d=1}^{D} (x_{t-d} - \mu_{X_t,D})^2 \right)^{1/2}
\]

Changes in the variance provide a timely indication of an anomaly. Power level itself can also be used as a reliable signal.

We constitute a new series from the power consumption readings.

\[
\sigma_{T,J} = \left( \frac{1}{T} \sum_{t=1}^{T} (x_{t,J} - \mu_{T,J})^2 \right)^{1/2}
\]

\(\sigma_{T,J}\) is a stochastic process with noise smoothed out. We assume it follows the normal distribution. The expected value, and variance over D days for each t are estimated as follows.

\[
\mu_{\sigma_{T,J},D} = \frac{1}{D} \sum_{d=1}^{D} \sigma_{T,J-d}
\]

\[
\sigma_{\sigma_{T,J},D} = \left( \frac{1}{D-1} \sum_{d=1}^{D} (\sigma_{T,J-d} - \mu_{\sigma_{T,J},D})^2 \right)^{1/2}
\]

**EXPERIMENTS**

Home A’s meter data in 2014 is adopted as the normal case. Some abnormal situations are simulated by setting the readings after anomaly occurrence at t the same as the value at t.

**CONCLUSIONS**

The algorithm, without any parameter tuning, has achieved remarkable efficacy on randomly generated scenarios on a real data set, which demonstrated its cold-start capability. Furthermore, the detection operation has limited requirement on data storage, which shows potential for an embedded deployment.

**REFERENCES**


**CONTACT**

† Y. Liu is with the Department of Industrial and Systems Engineering, WSU
‡ T. Hu and C. Wang are with the Department of Electrical and Computer Engineering, WSU
‡ R code for this project is at https://github.com/profyliu/smpesgm2017
A classification tree (or decision tree classifier) is a predictive model represented in a tree-like structure. Without making excessive assumptions about the data distribution, a classification tree partitions the input space into rectilinear regions and ultimately gives a set of If...Then... rules to classify outputs or make predictions.

Prevailing tree-building methods, such as ctree and rpart available in R, utilize only a single variable in each split, which drastically limits the expressiveness and in some cases the predictive accuracy of the tree model.

This project aims to address the gap by providing an optimization-based framework to exploit multivariate splits and enable more expressive tree models.

**OBJECTIVES**

- Build an open-source, extensible framework for optimal decision-tree learning for academic and research users
- Provide a highly usable machine learning package for industrial and business users
- Utilize high-performance computing to build optimal tree models on big data
- Explore decomposition techniques such as column generation to solve large instances

In bsnsing, a mixed integer program (MIP) is solved at each node to identify the optimal combination of features used to split the node.

Currently, support CPLEX (commercial) and lpSolve (free). More solvers will be added in a future release.

**INTRODUCTION**

**METHODS**

Multi-variable rectilinear splits are more efficient than single-variable splits in carving out nonlinear features, and preserve interpretability better than linear combination splits.

<table>
<thead>
<tr>
<th>Illustration</th>
<th>Characteristics</th>
<th>Software Codes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**DEMONSTRATION**

**How it works**

How to use it (in R)

**PERFORMANCE**

bsnsing can discern nonlinear patterns better than alternative packages, rpart and ctree.

**CONTACT**

Yanchao Liu, Ph.D.
Assistant Professor
Industrial and Systems Engineering
Wayne State University
4185 Fourth Street, Rm 2169
Detroit, MI 48202
Phone: (313) 577-3301
Email: yanchaoliu@wayne.edu
INTRODUCTION

Streptococcus pyogenes, also known as group A strep, is a form of acute pharyngitis which often includes sore throat, fever, headache, abdominal pain, nausea, vomiting, and makes it painful to swallow. S. pyogenes occurs in 20%-30% of cases of pharyngitis in children from ages 5 to 15, the age group for which the infection is most common [1]. Current diagnosis test include rapid antigen detection test (RADT) and throat culture, which is the gold standard [1]. The drawback to RADT is that although it has high specificity, it has varying sensitivities, while the drawback to throat culture is that it takes 24-48 hours to complete [1].

Raman spectroscopy uses a wavelength of light, usually laser, to excite the molecules in a sample, 1 in 10^14 molecules will vibrate at a different frequency (inelastic scattering) and can be detected with Raman [2]. Raman spectroscopy includes information about nucleic acids, proteins, lipids, and carbohydrates. Raman can be used as a rapid point of car diagnostic for S. Pyogenes.

Previous work in our lab demonstrated that bacteria was easily differentiated from virus with Raman. Here we wanted to determine if Raman could distinguish one bacteria from others in the event a person has more than one infection.

OBJECTIVES

Classify S. pyogenes from filtered tap water and five different bacteria: Escherichia coli (E. coli), methicillin-resistant Staphylococcus aureus (MRSA), methicillin-sensitive Staphylococcus aureus (MSSA), Psuedomonas aeruginosa (P. aeruginosa), and Legionella using linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), support vector machine (SVM), and random forest (RF) to determine which classifier yields the best sensitivity and specificity.

MATERIALS & METHODS

The bacteria specimens were cultured in the lab, then washed to remove the media. The final rinse was done with water. The Raman spectroscopy reading was done with a 785 nm laser.

Prior to creating the model, the data was preprocessed using third order morphology, then baseline subtraction with cosmic ray removal for spectra from 400-2480 cm^-1. Next the data was min/max normalized to get the values between zero and one. The data was split into training data (80%) and test data (20%). For cross validation (CV), the training data was split into four parts. In this study, one method used was principle component analysis (PCA) to contend with the curse of dimensionality that comes with Raman spectroscopy to reduce the number of dimensions. The first 5 PCs explain over 95% of the variance, so that is the maximum number of PCs passed to LDA, QDA, and SVM.

RESULTS

According to the hierarchal clustering dendrogram, S. pyogenes is most like Legionella and least like filtered tap water. MRSA and MSSA are the most similar bacteria, which is expected because they are both strains of Staphylococcus aureus.

The RF classifier performed better than LDA, QDA, and SVM. RF 4 and RF 7 had the best sensitivity, 93% on CV and best specificity, 97% on CV.

One of the drawbacks of RF is that since it randomly selects features, it is not possible to tell which of the 1,368 features is most important.

CONCLUSIONS

The RF classifier can be used with Raman spectroscopy data to identify S. pyogenes from filtered tap water, E. coli, MRSA, MSSA. P. aeruginosa, and Legionella in order to get a real time diagnosis of S. pyogenes.

ACKNOWLEDGEMENTS

Thank you to Zach Auner & Kayla Trexler for culturing the bacteria.

CONTACT

Brandy Broadbent
Research Assistant/PhD Candidate - BME/SSIM
313.577.1306
brandybroadbent@wayne.edu
INTRODUCTION

Economic model predictive control (EMPC) is an optimization-based control design that seeks to:
• Enhance process economic performance
• Develop empirical process models

EMPC can be formulated to generate non-Guaranteed recursive feasibility, closed-loop stability/safety

Critical to its design are:
• A sufficiently accurate process model
• An economics-based objective function
• Safety-based/practical constraints

Methods for obtaining these functions from data are needed to enable practical design of the controller

OBJECTIVES:
• Use EMPC to manipulate the process state for obtaining data for developing/verifying empirical models
• Develop empirical process models, seeking a relationship to process physics
• Take advantage of such models in EMPC design for poorly-understood/simple chemical processes

CLASS OF SYSTEMS:
• Consider nonlinear process systems with the form:
  \[ \dot{x} = f(x, u, w) \]
  \[ x \in \mathbb{R}^n: \text{State vector} \]
  \[ u \in \mathbb{R}^m: \text{Input vector} \]
  \[ w \in \mathbb{R}^l: \text{Disturbance vector} \]
• Consider nonlinear empirical models with the form:
  \[ \dot{x} = f_N(x, u) \]
• Assume the existence of a controller \( h_N(x') \) that renders the origin of the empirical system asymptotically stable in the sense that a Lyapunov function \( V \) exists for the closed-loop system

CONTROL DESIGN

\[
\begin{align*}
\min_{u(t) \in S(\Delta)} & \int_{t_k}^{t_{N+1}} \left[ L_E(t, u(t)) + \delta_1 \sum_{i=1}^{N_u} \alpha_{u_i} (u_i(t) - u_{d, i})^2 \\
& + \delta_2 \sum_{i=1}^{N_u} \alpha_{u_i} \left( x_i(t) - x_{d, i} \right)^2 \\
& - \delta_3 \sum_{q=1}^{N_{u+2}} \alpha_{p_{uq}} (p_u(t_q) - p_{u, d, q})^2 \right] \, dt
\end{align*}
\]

s.t. \[ \dot{x} = f_N(x(t), u(t)) \]
\[ x(t_k) = x_{d, k} \]
\[ u(t) \in U, \forall t \geq t_k \]
\[ p_u(t) \leq p_{u, d, q}, \forall q = 1, \ldots, N_{u+2} \]
\[ \frac{dV}{dt}(x(t), u(t)) \]
\[ \leq \frac{dV}{dt}(x(t), p_u(t), h(x(t))) \]
\[ \text{if } x(t_k) \in \Omega_{p_u} \]

PARAMETER DESCRIPTION

\( u(t) \in S(\Delta): \) Piecewise-constant input trajectory
\( u_{d, i}, \alpha_{u_i}: \) Desired values of \( u_i \) and states \( x_i \)
\( \delta_1: \) State prediction from empirical dynamic model \( f_N(x) \)
\( t_k: \) Sampling time
\( \delta_2 = 0 \land \delta_3 = 1, i = 1, 2, 3, 4, 5: \) Activation condition
\( \Delta: \) Activation function
\( N: \) Prediction horizon
\( \Delta: \) Sampling period
\( x(t_k): \) State measurement
\( X, U: \) State and input bounds
\( \delta_{u,i}: \) Level set of \( V \) with upper bound \( \beta_u \)
\( y: \) Positive terminal bounding difference between \( x_i(t) \) and the steady-state value of \( x_i \)
\( t': \) Time when constraint enforcement policy changes

CONTROLLER PROPERTIES:

• Guaranteed recursive feasibility, closed-loop stability, and robustness
  • When controller parameters \( (\Delta, \beta, \rho) \) and disturbances are sufficiently small

RESULTS: CONTINUOUS STIRRED TANK REACTOR EXAMPLE

Dynamic Model Equations:
\[ \frac{dC_A}{dt} = \frac{E}{V} (C_{A0} - C_A) - \frac{k_B}{p} C_A \]
\[ \frac{dT}{dt} = \frac{F}{V} (T_0 - T) - \frac{\Delta H}{p} C_A \]
\[ \frac{dC_B}{dt} = \frac{F}{V} C_B + \frac{k_B}{p} C_A \]

\( C_A: \) Concentration of reactant species \( A \)
\( C_B: \) Concentration of product species \( B \)
\( F: \) Volumetric flow rate of inlet and outlet streams
\( T: \) Reactor temperature
\( T_0: \) Initial temperature
\( C_{A0}: \) Concentration of \( A \) in feed stream
\( \Delta H: \) Enthalpy of reaction
\( k_B: \) Pre-exponential factor
\( q: \) Liquid density in reactor

EMPC is used to collect specific data \((C_A \text{ or } T \text{ held constant})\) which aids in selecting terms in the model for which to fit parameters via regression

EMPC is used to collect non-routine operating data to validate the model with parameters determined via regression

CONCLUSIONS

• EMPC can be formulated to generate non-routine operating data for short periods of time
• Attempts can be made to develop data that will aid in developing improved (e.g., more physically-based) process models
• Attempting to obtain physically-meaningful models can aid in the development of economic objective functions and safety-based constraints for EMPC

REFERENCES


CONTACT

Helen Durand
Assistant Professor - Department of Chemical Engineering and Materials Science
313.577.3475
E-Mail: helen.durand@wayne.edu

Analytic Approach: Control Design with Nonlinear Empirical Models
INTRODUCTION

Clostridium difficile infection (CDI) is due to the effects of toxins, toxin A (TcdA) and toxin B (TcdB) on the host. CDIs account for half a million new cases and more than 29000 deaths each year in USA (1). Severe CDI is associated with systemic signs of infection and multiple organ failure. Animal models of CDI demonstrate a strong correlation between systemic toxemia and the occurrence of severe disease (2,3). This gives us impetus to detect these toxins in serum. However, current technologies have low sensitivity to detect C. difficile toxemia in human subjects (4). Raman spectroscopy (RS) is an upcoming technology that is used to detect bacteria and their toxins (5,6). We speculate that supervised machine learning of RS data may help us identify a serum biomarker signature to detect clinically relevant concentrations of C. difficile toxins in serum.

EXPERIMENTAL DESIGN

Serum samples were spiked with varying concentrations (1 ng/ml, 1 pg/ml, 0.1 pg/ml) of TcdA, TcdB and both (TcdA&B). RS was performed on an air-dried serum drop that was placed on a mirror polished stainless-steel slide with an inVia Raman microscope (Renishaw, Gloucestershire, UK) equipped with a 785nm excitation laser.

Raw spectra were preprocessed by applying an in-house developed LabVIEW procedure for smoothing (IIR zero phase filter), background subtraction (low pass filter by Don Roth), Cosmic ray removal (cutting the peak with large second derivative value) and normalization (Unit Vector Normalization).

Pre-processed spectra were analyzed by Partial Least Square Discriminant Analysis (PLS-DA) and Support Vector Classification (C-SVC) for toxin specific serum biomarker signature. Model accuracy was measured by cross-validation and bootstrap methods.

MATERIALS & METHODS

Raman spectral region between 1000-1008 cm⁻¹ which corresponds to Phenylalanine and region between 2900-3000 cm⁻¹ which corresponds to CH stretching vibrations of proteins and lipids, contributed significantly to the classification model.

RESULTS

Variable Importance Profiles (VIP)

C-SVC Model Performance

Significant Spectral regions

Phenylalanine

CH stretching vibrations

Raman Shift/cm

Peak Intensity (a.u.)

TcdA

TcdB

TcdA&B

Control serum

Spiked serum (TcdA)

Spiked serum (TcdB)

Spiked serum (TcdA&B)

Toxin-spiked serum samples of various conc. (1 ng, 1 pg & 0.1 pg/ml) were distinguished from control serum 100% with cross-validation error rate ranging from 0-18% and bootstrap error rate ranging from 0-12%. Sensitivity ranged from 87-100% and specificity ranged from 77-100% for various conc. of toxin-spiked serum.

CONCLUSIONS

Raman spectroscopy has the potential to rapidly detect C. difficile toxins in serum at clinically relevant concentrations and may be useful as a diagnostic tool to modify therapy and predict outcome in critically ill CDI patients.

REFERENCES


CONTACT

S. Kiran Koya, Ph.D.
Post-doctoral Associate,
SSIM, M&MI Dept. of Surgery, WSU
313.577.1306
skikoya@wayne.edu
**INTRODUCTION**

Cultural Algorithms (CA) is a dual inheritance, socially motivated, computational model based on evolutionary principles. Even CA are based on evolutionary theory, however, the original idea was derived form Nature-Based algorithms of which different species are able to obtain and pass along heritable characteristics to their offspring. In CA case belief space stores the experience (heritable characteristics) in knowledge sources. These experiences are being collected via acceptance function and then pass along to individuals (offspring) within population space via influence function.

![CA Framework](image1)

**OBJECTIVES**

Influence function (also called distribution mechanism) can control the way of which the knowledge sources influence individuals within population space. Different distribution mechanisms have been suggested to optimize problems solving more efficiently. This study suggests a new Aution-Based distribution mechanism (labeled CAT4) of which knowledge sources (bidders) leverage shared information (common value) about individuals before auction starts. In CAT4 we focus on the ability to track the system tolerance toward the perturbations (Robustness), and the ability to maintain an acceptable level of an average performance despite problem/configurations complexities (Resilience).

![Two-dimensional Space of Cones world](image2)

**MATERIALS & METHODS**

A new version of culture algorithm toolkit (CAT) was redesigned to inject CAT4 changes and additions. Main additions to original toolkit were: 1. Adding the common value elements that knowledge sources can acquire before determining bidding value. 2. Building a bidding strategy for each knowledge source of which will be utilized to determine if a knowledge source will bid on the designated individuals in population space or not.

The bidding values are not only determined by individual’s average performance, but also with the individual’s common value elements: 1. history of previous influencers (KS) and 2. Individual’s location in the network. These two elements will boost the bidding value for a knowledge source. Once the bidding values are collected from different KSs, the auction system will be initialized with all configuration parameters to determine the winner. Once the winner KS is declared, the individuals would be assigned to the winner KS, and the winner KS will be charged for Average Performance of this individuals.

A robust problem generator (cones world) was used to compare CAT4 results with that of Aution-Based Distribution mechanism.

![CAT4 Regression line over 50 runs for complexity, A = 3.35](image3)

**RESULTS**

To verify the performance improvement we did conduct multiple comparisons between CAT4 and CAT2 across different complexities [A=1.01, 3.35, 3.99] and different network topologies [L-Best, Square, Hexagon, Octagon, Sixteengon, Global]. We conduct 50 runs for each tuple (Complexity/topology), a total of 300 for each A-Value.

![CAT2 Regression line over 50 runs for complexity, A = 3.35](image4)

**CONCLUSIONS**

Empirical results have shown there are multiple areas: complexity/topology where CAT4 was more tolerance toward the perturbations (Robustness), and maintained an acceptable level of an average performance despite problem/configurations complexities (Resilience). Even though, both systems have run same problem with identical configurations, the results emphasis the importance of common value elements in the algorithm. We noticed that the more connections the topology has (Octagon, Sixteengon, Global) the more efficient CAT4 becomes in compare with CAT2. Despite the fact even while all knowledge resources are using the same topology(homogenous approach) results are promising. However, in the future work we plan to enhance CAT4 to allow heterogenous topology selection for the system to best fit the problem complexity, in order to find the solution more efficiently.

**REFERENCES**


**CONTACT**

Anas Al-Tirawi , Ph.D. Candidate
aaltirawi@wayne.edu

Prof. Robert Reynolds
reynolds@wayne.edu

Dept. of Computer Science
Wayne State University
COMMUNITY DETECTION IN PRODUCT-PRODUCT GRAPHS FOR MINING FREQUENT-ITEM-SETS

INTRODUCTION

Market basket analysis is the science of discovering customer purchase behavior in order to design marketing strategies. There are dozens of enterprise solutions to market basket analysis, based on FP-Growth algorithms. However, this method lacks robustness and laboriously handles temporal data.

Our proposed solution is based on the graph of transactions, which updates dynamically in time. A real-time controller monitors critical measures of the system and reports any change in customer behavior.

The output can be fed directly to other subsystems such as assortment planning, inventory planning, pricing and promotions, etc. to take suitable actions.

OBJECTIVES

Develop a robust algorithm to identify frequent-item-sets.

This algorithm must be:
• Dynamic
• Fast
• Work with real time data
• Handle large scale data

and design a controller for real-time monitoring of the vitals of the system as they evolve in time.

The controller must:
• Detect changes as soon as possible
• Be trained to predict changes
• Identify and reject the noises
• Decide on what action to take

MATERIALS & METHODS

A summary of the steps of the method is shown below.

- Generating the graph of items
- Community detection using Louvain
- Real-time measurement of the vitals of the network
- Reporting vitals to the controller
- Mining vitals to extract abnormalities
- Deciding on set of actions
- Feedback to the system and repeating the process

Initializing the graph of items is the most important and innovative part of this research. Graph is formed by following these steps:

- The community detection, which is the most time taking step does not need to be performed at every heartbeat. Only when the controller decides that current communities are no longer accurate, it sends a recalculation command to community detection kernel.

Measurements are quick and easy. Their complexities grow linearly in problem size. Data mining stage compares newly arrived data with historical data in to match the trends with its knowledge base. Based on the results of matching several set of actions are possible.

To test the controller, simulator randomly removed SKUs from the patterns, and controller could identify the removed SKUs/patterns after 3 or less time steps. Below is an example of control charts when one cluster has removed SKU.

RESULTS

To evaluate the accuracy of the method, several simulated sets of transaction were made and fed into the model. Number of items varies from 50 to 1000, and number of transactions varies from 20,000 to 2,000,000. A number of known patterns were put into simulated data, and the algorithm extracted more than 95% of the patterns.

Above, is a sample result. Trimming the SKU-SKU graph by removing all edges with pairwise support of less than elbow value, gives us another graph whose patterns are easily extracted using Louvain’s algorithm.

CONCLUSIONS

An effective graph based automated algorithm to find communities of frequently bought together SKUs is introduced.

Algorithm uses parallel and cluster computing capabilities to reduce execution time. Exhibiting good performance on synthetic datasets.

Next Steps for MBA:
• Control number of clusters
• Adapt method to monitor substitutable and complementary products and sales events
• Further operationalize method and enhance robustness
• Investigate Customer-Product Graphs

SPONSOR

Elham Nosrat, Research Assistant/Industrial & Systems Engineering Department
734.353.2159
fos7573@wayne.edu
INTRODUCTION
Availability of a validated, realistic fuel cost model is a prerequisite to the development and validation of new optimization methods and control tools. This research uses an autoregressive integrated moving average (ARIMA) model with historical fuel cost data in development of a three-step-ahead fuel cost distribution prediction. **MOTIVATION** of this research is to develop a forecasting algorithm for more accurate FC characterization and distribution estimation instead of relying on the data with a three-month delay.

DATA FEATURES
The public fuel cost data extracted from Form EIA-923 are analyzed and utilized for achieving more accurate probability density estimation using ARIMA model forecasting.

A. Form EIA-923 Exploration
The Date (Year and Month), Plant ID, Plant State, Energy Source, Quantity, Average Heat Content and Fuel Cost are archived and shown in Table I

<table>
<thead>
<tr>
<th>Table I Raw Data in Form EIA-923</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date</td>
</tr>
<tr>
<td>201301</td>
</tr>
<tr>
<td>201301</td>
</tr>
<tr>
<td>201301</td>
</tr>
<tr>
<td>201301</td>
</tr>
<tr>
<td>201301</td>
</tr>
</tbody>
</table>

Fuel cost data are required to be grouped together by Date, Plant ID and Energy Source for obtaining an updated overall fuel cost for each Energy Source in a month. The updated fuel cost **FC** are calculated by (1), (2) and (3):

- **Total Cost** by (1):
  \[
  \text{Total Cost} = \sum Q \times \text{AHC} \times \text{FC} \tag{1}
  \]
- **Total Heat** by (2):
  \[
  \text{Total Heat} = \sum Q \times \text{AHC} \tag{2}
  \]
- **FC** by (3):
  \[
  \text{FC}_{\text{Plant,Date,Source}} = \frac{\text{Total Cost}_{\text{Plant,Date,Source}}}{\text{Total Heat}_{\text{Plant,Date,Source}}} \tag{3}
  \]

B. Objective Data Selection
Texas is the top state in the energy consumption sources and taken as an example for the study. The probability density prediction and estimation focus on NG instead of SUB and BIT, because the fuel cost distributions of coal are stable.

MODEL DEVELOPMENT
Given the differential fuel cost \(\Delta FC\) and \(\Delta FC_{\text{hub}}\), without any seasonal pattern, a standard ARIMA model is proposed for forecasting the NG fuel cost series of each plant to estimate the distribution for the next month.

A. ARIMA Model
The ARIMA model for predicting \(\Delta FC\) and \(\Delta FC_{\text{hub}}\) is a standard ARIMA\((p,d,q)\) model described as follow:

\[
\phi_p(B)\theta_q(B)\Delta FC_t = \mu + \epsilon_t \tag{6}
\]

B. Normal Distribution Fitting
The standard normal distribution is fitted to the data with maximum likelihood method to estimate the probability density function.

\[
F(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{7}
\]

C. KL Divergence of Normal Distributions
To evaluate the forecasting performance with fitted normal distribution, the symmetric KL divergence between two distributions:

\[
KL(p,q) = \frac{1}{2}(D_{KL}(p||q) + D_{KL}(q||p)) \tag{8}
\]

PREDICTION PERFORMANCE
The training data set, from Jan. 2013 to June 2016, including 42 months of fuel cost data and NG hub spot price data are used to obtain the parameters of the ARIMA model. Six months of NG fuel cost data from July 2016 to Dec. 2016 are employed for validating the model. Then the normal distribution is used to fit the data to obtain an estimated distribution of the fuel cost.

KL divergence is implemented to check the divergence for the distributions of three-month delay fuel costs, forecasting fuel cost and the actual fuel cost.

CONCLUSION
The results show the proposed forecasting algorithm has a superior performance over the method that uses the three-month delayed data. The results also show that the method can better handle fuel cost volatility in prediction.

Contact: Prof. Caisheng Wang
The ECE Department, WSU
E-Mail: cwang@wayne.edu
Currently, most diseases are diagnosed only after significant disease-associated changes have occurred. This is particularly true in the case of complex conditions such as cancer. Here, we propose an approach able to identify when systemic qualitative changes in biological systems happen, thus opening the possibility of therapeutic interventions before the occurrence of symptoms. We propose a qualitative change detection (QCD) approach. QCD takes as input sequential measurements as described by time series (or progressive disease stages) together with known interactions described by biological networks. On this input, QCD applies an impact analysis approach to identify the time interval in which the system transitions to a different qualitative state.

**RESULTS**

**Table 1. The results of the QCD on 8 case studies:**

<table>
<thead>
<tr>
<th>Case Study</th>
<th>Pathway</th>
<th>Change Interval</th>
<th>Detected Change Interval</th>
<th>Significant Change Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>E. coli</td>
<td>Flagella building</td>
<td>3.5 h - 7.5 h</td>
<td>3.5 h - 7.5 h</td>
<td>3.5 h - 7.5 h</td>
</tr>
<tr>
<td>H. sapiens</td>
<td>Phenotype expression</td>
<td>2.5 h - 7.5 h</td>
<td>2.5 h - 7.5 h</td>
<td>2.5 h - 7.5 h</td>
</tr>
</tbody>
</table>

**Figure 4** displays the input and results of the qualitative change detector (QCD) for the *E. coli* flagella building phenomenon. Panel A: The multi-output coherent type 1 feed-forward loop (C-FFL) network that describes the flagellum building, together with the activation thresholds (B on the edges for each group of genes (dark green boxes)). The flagellum building is depicted in the cartoons matching the activation of each group of genes. The black box denotes building the flagellum hook which is the point of no return and hence the real change interval. Panel B: The heatmap of the sampled data (input QCD), and the real change interval (black arc and black vertical line positioned in the center of the interval). The change interval detected by QCD is shown by the green arc and the green vertical line positioned in the center of the interval (very close to the black line showing the actual point of no return). The stages of the flagella building are presented as cartoons in chronological order on the top part of the figure.

**Figure 5** displays the input and results of QCD on human hepatitis C virus (HCV) to hepatocellular carcinoma (HCC) progression. The input is gene expression data from GEO GSE6764 (right) and the viral carcinogenesis pathway from KEGG (hsa02020) - left. The data captures the progression from human HCV to HCC, specifically the transition from control (healthy) through the progressive stages of liver damage up to very advanced HCC. The right panel shows the heatmap of the disease progression for the measured KEGG pathway genes (in red), with the change interval detected for the phenomenon (green arc and green line in the center of the interval (control - high-grade dysplastic nodules)). The dark green vertical line (very early HCC) marks the pre-disease state detected by the DNBM method.

**CONCLUSIONS**

We designed and implemented QCD, an analytical method capable to detect qualitative changes in the state of a biological system by monitoring its gene expression levels. We did this with no training on previous examples, with no expert supervision, and with thresholds set using sound statistical criteria. The only hypothesis used in this study is that a qualitative change will involve enough pathway components to perturb the pathway in a significant way.

Due to the large variety of data sets used to validate the proposed approach, this work is expected to be of interest to a large category of life scientists ranging from basic scientists working with gene networks in model organisms such as fruit fly, yeast and mouse, to neuroscientists using neuropeptides to model nervous system responses, to clinical researchers studying human diseases such as cancer.

**ACKNOWLEDGEMENTS:** This work has been partially supported by the following grants: NIH RO1 DK089167, NIH STTR R42GM087013, NSF DBI-0965741 (to SD), by the Robert J. Sokol M.D. Endowment in Systems Biology (to SR), and by the Thomas Rumble Fellowship (to CM). This research was also supported, in part, by the Perinatology Research Branch, Division of Intramural Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, U.S. Department of Health and Human Services (NICHD/NIH/DHHS), and, in part, with federal funds from the NICHD/NIH/DHHS under Contract No. HHSN275201300006C.

**CONTACT:** Cristina Mitrea. cristina@wayne.edu

INTRODUCTION

- Parameter design as an important methodology in manufacturing industry applied to improve the quality of products and processes through a cost-effective approach.
- Purpose of parameter design: set the levels of control variables that maximize the quality while minimizing the process variability imposed by noise variables.
- Response surface methodology (RSM) has been known as a well method for parameter design.
- The observational data gathered during the operational processes can be useful to extract and select the control variables by data mining and machine learning methods.

OBJECTIVES

- The main objective of this study is to rely on using observational data to achieve robust parameter design of manufacturing processes.
- Controlled experiments can be challenging in production environments and this paves for an effective alternative approach to attain robust process parameter conditions.
- The proposed framework relies on an integrated feature selection, response surface modeling, and optimization methodology.
- We present results from numerical experiments that compare performance against traditional sequential methods (where variable selection precedes parameter design).

METHODOLOGY

An integrated feature selection approach with robust parameter design in 3 consecutive steps:

Step-1: Determine set of potential variable to control
Step-2: Build appropriate response surface model
Step-3: Optimize the process parameters with constrained optimization.

- Constrained optimization

\[ \text{Min} \quad \text{Var}_{z}[y(x,z)] \]

Subject to \( E_{z}[y(x,z)] = T \)
\( x \in R^k \)

- Mean Squared Error (MSE) loss function:

\[ \text{Min} \quad \text{MSE} \]

Subject to \( x \in R^k \)
\( E_{z}[y(x,z)] - T)^2 = [E_{z}[y(x,z)] - T]^2 + \text{Var}_{z}[y(x,z)] \)

RESULTS

The result of proposed approach has been compared with the results of two other non-integrated robust optimization methods which are random forest and a filter method based on mutual information.

Numerical Experiments

Settings for experiments to investigate the proposed method:

- Results of proposed approach compared to other methods:
  - Filter method with mutual information-based criterion
  - Random Forests

To investigate the accuracy of response model:

- Compare parameters obtained of response model with each element of true model parameters. \( \theta = [\beta_1, \beta_2, \beta_3, \beta_4] \)
- Comparison benchmark: \( \hat{\theta}_1 \) of true model \( \hat{\theta}_2 \) of obtained model

Case Study: Tire Compound Production

- In this research, we implemented the proposed method to select the effective variables and finding the settings of the control factors that minimizes variation in rubber compound (minimum) viscosity around the target value due to uncontrollable factors.
- Data Overview:
  - 214 Observations with Output Viscosity Measurements
  - Set of potential variables to control: 10 control variables, 6 noise variables

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean Response</th>
<th>Variance of Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ideal Target</td>
<td>65.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Proposed Method</td>
<td>64.2</td>
<td>17.4</td>
</tr>
<tr>
<td>Filter Method with MI</td>
<td>75.0</td>
<td>31.74</td>
</tr>
<tr>
<td>Random Forest</td>
<td>62.7</td>
<td>27.6</td>
</tr>
</tbody>
</table>

CONCLUSIONS

- We proposed a flexible and integrated approach for variable selection and robust design of manufacturing processes using observational data.
- Results from Numerical Experiments and Industry Case Study are promising.

REFERENCES


CONTACT

Ratna Babu Chinnam, Ph.D.
Professor, Industrial & Systems Engineering
Co-Director - Big Data & Business Analytics Group
Tel: 313.577.4846
Ratna.Chinnam@wayne.edu
INTRODUCTION

Increasing volume of Electronic Health Records (EHR) in recent years provides huge opportunities for data scientists to collaborate in different aspects of healthcare by applying advanced analytics on these EHR clinical data. A key requirement however is obtaining meaningful insights from high dimensional, sparse and complex biological/ biomedical data. Data mining and machine learning approaches can address this challenge by performing feature representation (feature engineering or unsupervised learning) in order to build more reliable and informative features from personalized clinical data and subsequently training the supervised learning using this represented features. In this study, we propose a prediction framework based on deep learning for feature engineering in a healthcare setting. In our approach, we use three different deep autoencoder networks including Stacked Autoencoders, Variational Autoencoders, and Adversarial Autoencoders.

We perform our approach on two EHR datasets obtained from eICU collaborative research database to train a predictive model of length of stay (LOS) in intensive care units for neuro-patients and cardiac-patients

OBJECTIVES

Based on two large and high dimensional datasets from eICU collaborative research database, our objectives in this research are:

1) Providing an accurate prediction of patient length of stay in ICU unit for two groups of patients.
2) Providing a comparison study for performance of different autoencoder-based networks especially using medical health records.

MATERIALS & METHODS

In this study, we propose an integrated predictive framework using representation learning. Our methodology follows the workflow shown in below Figure that includes three important steps:

**Step-1: Preprocessing and Word Embedding**

In the first step, we use the preprocessing methods such as outlier detection and imputation for missing values existed in dataset. We also transform categorical variables to vectors using a well-known word embedding algorithm called Glove algorithm.

**Step-2: Feature Representation using Deep Learning**

In the second step, all features will be represented in higher-level abstraction by three different autoencoder deep networks separately:
1) Stacked Autoencoder (SAE),
2) Variational Autoencoder (VAE),
3) Adversarial Autoencoder (AAE)

**Step-3: Supervised Learning**

In this step, we apply supervised learning models on the top of represented dataset for all three feature extraction approaches. Once the features are extracted, these representations from main dataset are entered in a well-known linear and non-linear supervised regression models.

EXPERIMENTAL STUDY

Medical doctors predict intensive care units (ICUs) length of stay for planning ICU capacity as an expensive unit in the hospital and identifying unexpectedly long ICU length of stay in special cases to better monitoring.

The data in the eICU database includes patients who were admitted to intensive care units during 2014 and 2015. Among different patients, we choose cardiovascular and Neurological patients admitted in the Cardiac-eICU and Neuro-eICU respectively.

IMPLEMENTATION & RESULTS

In this healthcare informatics case study, we implemented Variational autoencoder (VAE) and Adversarial autoencoder (AAE) by using TensorFlow and Theano libraries in Python and executed Stacked autoencoder (SAE) with H2O package in R.

All autoencoders are applied with 5 hidden layers (two hidden layers of encoders and decoders and one middle layer). For each deep architecture we consider specific parameters based on autoencoder types and data characteristics. We performed different deep networks (models).

The following Tables demonstrate MSE and R-squared results for different types of autoencoders and supervised learners in Cardiac-ICU data and Neuro-ICU respectively:

**CONCLUSIONS**

In this study, we proposed a novel application of deep learning for accurate prediction of length of stay (LOS) in intensive care units (ICUs) using high-dimensional EHRs (eICU database).

Empirically, our results demonstrate that:

1) Medical feature representation can improve the performance of prediction significantly.
2) Choice of representation can lead to different performances. Based on our findings, the performance of VAE and AAE is much better than SAE, and VAE demonstrated rather strong performance for LOS prediction.

REFERENCES


CONTACT

Dongxiao Zhu, Ph.D., Professor, Computer Science
dzhu@wayne.edu
Problem

Estimating $\sigma$

Deciding $p$

$a = \frac{\lambda}{\sigma}$

Demand

$\chi^2$

$\delta$

$\beta$

$\alpha$

$\gamma$

$\epsilon$

$\nu$

$\Gamma$

$\Delta$

$\Sigma$

$\Omega$

$\Lambda$

$\theta$

$\phi$

$\kappa$

$\tau$

$\phi$

$\psi$

$\chi$

$\zeta$

$\eta$

$\xi$

$\upsilon$

$\rho$

$\sigma$

$\nu$

$\kappa$

$\lambda$

$\theta$

$\phi$

$\psi$

$\chi$

$\zeta$

$\eta$

$\xi$

$\upsilon$

$\rho$

$\sigma$

$\nu$

$\kappa$

$\lambda$

$\theta$

$\phi$

$\psi$

$\chi$

$\zeta$

$\eta$

$\xi$

$\upsilon$

$\rho$

$\sigma$

$\nu$

$\kappa$

$\lambda$

$\theta$

$\phi$

$\psi$

$\chi$

$\zeta$

$\eta$

$\xi$

$\upsilon$

$\rho$

$\sigma$

$\nu$

$\kappa$

$\lambda$

$\theta$

$\phi$

$\psi$

$\chi$

$\zeta$

$\eta$

$\xi$

$\upsilon$

$\rho$

$\sigma$

$\nu$

$\kappa$

$\lambda$

$\theta$

$\phi$

$\psi$

$\chi$

$\zeta$

$\eta$

$\xi$

$\upsilon$

$\rho$

$\sigma$

$\nu$

$\kappa$

$\lambda$

$\theta$

$\phi$

$\psi$

$\chi$

$\zeta$
Analytic Approach: Sample Average Approximation-Based Parallel Algorithm

**INTRODUCTION**

The service placement problem in Mobile Edge Computing (MEC) has to consider several issues that were not present in the data-center settings:

- After the initial service placement, mobile users may move to different locations which may increase the execution time or the cost of running applications.
- The resource availability of servers may change over time.

**Motivation**

Markov Decision Processes (MDP) have been used by several researchers to model application placement problems in MEC. Most variants of the MDP problems are known to be P-complete, that is, it is not possible to design highly efficient parallel algorithms to solve them.

**Contributions**

- Developed a multi-stage stochastic programming model for the problem of application placement in MEC, and
- Designed a placement algorithm based on the Sample Average Approximation (SAA) method.

**Problem Definition**

We consider an MEC system, where:

- There is a set of M servers and a set of N users.
- Users request to execute components of an application on the edge servers.
- Each request from user has a due date within the T time periods.
- Users' locations are known in the first time period and change stochastically during the following time periods.
- The location of a user is specified by its coordinates in a two-dimensional grid of cells.

**Objective**

minimizing the total cost consisting of:

1) communication cost between users and servers
2) execution cost at each server
3) relocation cost

**Stochastic Program:**

\[
\min_{X,Y} \sum_{i=1}^{M} \sum_{j=1}^{N} c_i(x_i^j, y_j) + E[\sum_{i=1}^{M} \sum_{j=1}^{N} c_i(x_i^j, y_j)] + E[\sum_{i=1}^{M} \sum_{j=1}^{N} c_i(x_i^j, y_j)] + \sum_{i=1}^{M} \sum_{j=1}^{N} c_i(x_i^j, y_j)
\]

subject to:

1. \( \sum_{i=1}^{M} x_i^j \leq 1 \) if \( j \neq 0 \)
2. \( \sum_{i=1}^{M} x_i^j \geq 1 \) if \( j \neq 0 \)
3. \( x_i^j \geq 0 \) for all \( i, j \)

Algorithm 1 PG-SAA: Parallel Greedy SAA-based application placement algorithm

1. Generate F independent scenario samples: \( G_1, \ldots, G_F \), each of size \( |V| \).
2. Use \( G \) to initialize a parallel solution.
3. For \( i = 1 \) to \( N \) do in parallel
4. Create graph \( G_i(V,E) \) for user \( i \).
5. Use Dijkstra's algorithm to find the shortest path between node \( D \) and node \( S_i \) in graph \( G_i(V,E) \).
6. Update \( c_i \) cost of the shortest path
7. End for
8. Call RM-GAP to obtain the placement.
9. Record the optimal solution in \( (X^*, Y^*) \).
10. End for
11. Generate a sufficiently large sample of size \( U, U \gg N \).
12. For \( i = 1 \) to \( F \) do in parallel
13. For \( j = 1 \) to \( N \) do
14. Create graph \( G_i(V,E) \) for user \( i \).
15. Use Dijkstra's algorithm to find the shortest path between the dummy vertex \( D \) and vertex \( S_i \).
16. End for
17. End for
18. Out of \( F \) candidate solutions, choose the one that has the smallest estimated objective value
19. Allocate users' requests to servers according to \( X^* \).

Algorithm 2 RM-GAP algorithm

1. Input: \( G \)
2. \( Q_j \) for \( j = 1, \ldots, N \)
3. \( R_j \) for \( j = 1, \ldots, N \)
4. While \( U \neq 0 \) do
5. Construct the feasibility set \( F = \{ j : Q_j > 0 \} \)
6. For \( i = 1 \) to \( N \) do
7. Find the server with the minimum cost: \( j_i = \arg \min_{j \in F} c_i(y_j) \)
8. Assign the user with the greatest value of \( c_i \) to server \( j_i \)
9. Update the cost of the server \( j_i \)
10. End for
11. End while

EXPERIMENTAL ANALYSIS Set-up:

- Users and servers are located within a two-dimensional grid of 100 x 100 cells.
- For the sizes of the user requests, we use the data from the dataset provided by the LiveLab project on real-world smartphone usage [1].
- Other parameters are generated using the uniform distribution.

Metrics:

- Objective ratio
- Speed-up

Results:

Graph \( G_i(V,E) \) used for computing the assignment of user \( U \).

REFERENCES


CONTACT

Hossein Badri, Ph.D. Candidate
Department of Industrial & Systems Eng.
313.401.1993
Hossein.badri@wayne.edu
INTRODUCTION

**Importance of Driver Age Estimation**

- Feature extraction: geometry features, engineered features.
- Estimators: classification (SVM), regression (SVR), ranking.
- Deep learning: multi-class CNN, MR-CNN, DEX.

**RANKING-CNN**

- Existing Techniques in Human Facial Age Estimation:
  - Feature extraction: geometry features, engineered features.
  - Estimators: classification (SVM), regression (SVR), ranking.
  - Deep learning: multi-class CNN, MR-CNN, DEX.

Framework

- Contains a series of basic CNNs.
- Initialized with a pre-trained base CNN, fine-tuned with ordinal age labels.
- The binary outputs are aggregated to make the final age prediction.

THEORETICAL ANALYSIS

**A New Error Bound for Ranking**

**Theorem**  For any observation \((x,y)\), in which \(y > 0\) is the actual label (integer), then the following inequality holds:

\[ |r(x) - y| \leq \max_k e_k(x) \]

EXPERIMENTS

**Experiment Setup**

- Pre-train with 26,580 images from the unfiltered faces dataset.
- Fine-tune on the age estimation benchmark MORPH dataset.
- Randomly select 54,362 samples in the age range of 16 to 66.

**Comparison of MAE among different features and estimators.**

<table>
<thead>
<tr>
<th>Classification</th>
<th>SVM-Multi-CNN</th>
<th>Ranking-CNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>FBP+OLPP</td>
<td>3.59</td>
<td>3.45</td>
</tr>
<tr>
<td>ST</td>
<td>3.15</td>
<td>3.65</td>
</tr>
<tr>
<td>CNN Feature</td>
<td>4.99</td>
<td>5.10</td>
</tr>
<tr>
<td>Ranking-CNN Feature</td>
<td>3.99</td>
<td>3.94</td>
</tr>
</tbody>
</table>

Comparison with state-of-the-art models: MR-CNN, OR-CNN and DEX.

- Ranking-CNN: 2.96
- MR-CNN: 3.27
- OR-CNN: 3.34
- DEX: 3.25

For More Details


ACKNOWLEDGEMENTS

- NSF: CNS 1637312
- Ford URP: 2015-9186R

CONTACT

- Machine Vision and Pattern Recognition Lab: shixing.chen@wayne.edu
- Professor Ming Dong: mdong@wayne.edu
Analytic Approach: A Comparison of Common Data Types with Automated Vehicle Location Data

Introduction

Performance measures are essential for managing transportation systems, including signalized corridors. Coordination is an essential element of signal timing, enabling reliable progression of traffic along corridors. Improved progression leads to less user delay, which leads to user cost savings and lower vehicle emissions. This paper presents a comparative study of signal coordination assessment using four different technologies. These technologies include detector-based high-resolution controller data, Bluetooth/Wi-Fi sensor, segment-based probe vehicle data, and automated vehicle location data consisting of GPS-based vehicle trajectories, representing the data anticipated from emerging connected vehicle technologies. The data were compiled for a 4.2-mile corridor in Holland, Michigan. The results show that all of the data sources were able to identify, at some level where coordination issues existed. Detector-based controller data and GPS-based vehicle trajectory data were capable of showing greater detail, and could be used to make offset adjustments. The paper concludes by demonstrating the identification of signal coordination issues with the use of visual performance metrics incorporating automated vehicle location (AVL) trajectory data.

Methodology

US-31 Holland Corridor

Bluetooth/Wi-Fi Matching

Automated Traffic Signal Performance Measures

Coordination Issue Legend

- Random Arrivals at SB Quincy
- Poor PM Offset at SB James X-Over
- Poor AM/PM Offset at NB Quincy
- Poor PM Offset at NB Felch
- Poor AM/PM Offset at NB Felch X-Over
- Poor AM Offset at NB James
- Random Arrivals at NB James X-Over
Evaluating the Performance of Coordinated Signal Timing

Stephen Remias¹, Christopher Day², Jonathan Waddell¹, Jenna Kirsch¹, Ted Trepanier³
1: Wayne State University; 2: Iowa State University; 3: INRIX

AVL trajectory has a promising future with potential applications in planning, operations, maintenance,

Conclusions

- Bluetooth/WIFI, segment-based probe data, ATSPMs, and AVL trajectory data can all be used to locate potential coordination issues, however only ATSPMs and AVL trajectory data can be used to adjust offsets.
- Although, detector-based SPMs are still the gold standard for the evaluation of coordinated corridors, the scalability of AVL trajectory data can provide a systematic and lower cost approach to identify issues.
- AVL trajectory has a promising future with potential applications in planning, operations, maintenance, and model validation.
Deep neural networks (DNNs) make impressive progress in predictive modeling. DNNs’ merit of learning non-additive information make themselves promising in clinical predictive modeling. Successful DNNs require abundant data. In clinical research, we only have limited amount of data due to its expensive and time-consuming labeling process. We propose auxiliary-task-augmented network (ATAN) for predictive modeling, leveraging clinical measure as auxiliary tasks. ATAN falls into the framework of multi-task learning as a regularization method, alleviating the overfitting problem.

**INTRODUCTION**

- Feedforward DNNs are used to learn high-level feature representations for the primary target, auxiliary target.
- A shared feature representation is also learning to conceptually capture the clinical relevance between the primary and auxiliary targets.
- Feature representations are combined to make predictions.

**APPLICATION**

- We apply ATAN in a clinical hypertension dataset collected by Detroit Receiving Hospital.
- LVMI is the primary target, other CMR results are by-products of the labeling process, which are clinically related with LVMI.

**RESULTS**

- Mean square error (MSE), explained variance score (EVS) and median absolute error (MAE) are sued to evaluate the performance of ATAN.

**REFERENCES**


**ACKNOWLEDGEMENTS**

- Research in this research was supported by NSF/CCF 1451316 & 1637312.
- Cohort collection and related study were supported under NIMHD (1R01 MD005849), NHLBI (1R01 HL127215) and PCORI (FC14-1409-21656).

*Email: dzhu@wayne.edu*
Analytic Approach: A New Tool For Real Time Energy Emission Observation, Wayne State University Campus

INTRODUCTION

Dashboards are providing powerful means to monitor conditions at a glance. The main objective of this project was providing a visualized action tool for the energy used and emissions released due to electricity consumption at the Wayne State University campus area by providing a single page dashboard which is developed to show a graphical presentation of historical and real-time data. The real-time data is coming from an API containing a group of sensors in different locations named “CPN” which stands for Commercial Pricing Node, the CPN locations are shown below.

Effect of Emissions on Health

A) CO₂: Another effect is alterations in bone metabolism and related blood calcium concentrations.

B) NOx: Long term exposure can decrease lung function, asthma and increases the response to allergens which contributes to the formation of fine particles (PM) and ground level ozone, which are associated with adverse health effects.

C) SO₂: Emissions that lead to high concentrations of SO₂ in the air generally also lead to the formation of other sulfur oxides (SOx > PM). Gaseous SOx can harm trees and plants by damaging foliage and decreasing growth. SO₂ and other sulfur oxides can contribute to acid rain which can harm sensitive ecosystems.

D) Hg:

Mercury is toxic to the nervous system. Adults who have been exposed to too much methylmercury might begin to experience trembling hands and numbness or tingling in their lips, tongues, fingers or toes. At higher exposures walking could be affected, as well as vision, speech and hearing. Damage occurring before birth or in infancy can cause a child to be late in beginning to walk and talk and may cause lifelong learning problems.

E) Pb: Lead distributes throughout the body in the blood and is accumulated in the bones. The lead effects most commonly encountered in current populations are neurological effects in children and cardiovascular effects in adults. Infants and young children are especially sensitive to even low levels of lead, which may contribute to behavioral problems, learning deficits and lowered IQ.

MATERIALS & METHODS

To evaluate the sustainability of systems that draw power from electrical grids there is a need to rapidly and accurately quantify pollutant emissions associated with power generation. Air emissions resulting from electricity generation vary widely among power plants based on the types of fuel consumed, the efficiency of the plant, and the type of pollution control systems in service. There are 39 types of fuel, 17 types of prime movers and 116 combinations of prime mover and fuel type which demonstrates the LEEM approach complexity. The main purpose of LEEM is to estimate the LMP and emissions and shift the energy use to hours with lower electricity demand so the demand is not high enough to use a coal powerplant and therefore the cost and emission is going to decrease and renewable energy such as solar energy would be the priority. The LMP is indicating “Locational Marginal Price” which is calculated based on the real time energy use and power generation in order to estimate the price of electricity per kilowatt hour($/MWh). LMP is a function of time and location, in other words it varies dramatically with respect to time and amount of electricity demand in a region. Basically we have two types of real-time data for emissions in our dashboard regarding different calculation procedures by the LMP Emission Estimation Method (LEEM), first one is Highest probability which presents the most probable amount of emission($/MWh) released each 15 minutes and the other one is Weighted probability which is calculated by considering varied factors to provide another estimation of emission release to the air at each location.

RESULTS

Our real time data dashboard display:

CONCLUSIONS

The data is parsed from our source API which collects the data and applies the LEMP approach calculation behind the seen and sends information as a JSON file updating each five minutes. As the dashboard part, the data is stored at a database in a classified format, then a Java script reads the information from the source API each one minute and calls the information related to our specific CPN, then a PHP script answers different requests for our data and After all, there is a HTML with the role of projecting data set in desired diagrams and charts on the dashboard. The Gauge charts are presenting real-time data for the most recent data point which is reported by the API for the weighted probability, other graphs are presenting Weighted and highest probabilities estimated for each emission at the closest CPN to the Wayne State University campus area since past 21 hours to the present time. The Minimum, Average and Maximum range showed on the gauge charts as green, yellow and red are specified for each chart based on analyzing the historical data since last year. The colored zones are defined based on the lowest, average and highest ranges as the following table.

Here is the link of our dashboard for any one interested to check the emission release and electricity price for our Campus area: http://178.216.144.169/

REFERENCES

Database of the National Library of Medicine’s TOXNET system http://toxnet.nlm.nih.gov
LEEM Interface website http://leeminterface.org/
EIA Website https://www.eia.gov
LEEM Today http://leem.today/

CONTACT

Amir Shahin Kamjou, Graduate Student
Civil and Environmental Engineering Department
FX3555@wayne.edu
Carol J. Miller, PhD, PE . Wayne State University
ab1421@wayne.edu

Visualization Tools for Electricity Emission Intensity Data
Amir Shahin Kamjou, Carol J. Miller
Healthy Urban Waters, Wayne State University, Detroit, MI 48202
**INTRODUCTION**

Recently, data-driven methods have been increasingly applied to different domains ranging from business intelligence, medical diagnosis, to manufacturing design. Resistance Spot Welding (RSW) datasets can be used to build models, extract knowledge, and accelerate welding design process. Some welding quality prediction systems \[1\] have been proposed by using shift-off Decision Tree and Random Forest algorithms in a single machine with small datasets. DATAVIEW \[2\], with easy-to-use machine learning operators, including Random Forest, Support Vector Machine, Artificial Neural Network, can predict welding quality using high performance computer resources such as Amazon clouds, thus making the prediction process more efficient with large datasets.

**Motivation**

Resistance spot welding is a complex thermomechanical process with high vexity data and highly non-linear relation between process parameters. In big data scenarios, traditional implementation platforms like R are inefficient. DATAVIEW, a big data infrastructure, handling big data analytics process using scalable computing resources in both horizontal and vertical way, is more reliable and efficient.

**OBJECTIVES**

1. Design and development of implementation pipelines for predicting the welding quality (weld nugget width dimension) with R and DATAVIEW, respectively.
2. Performing statistically designed experiments to compare the efficiency of R and DATAVIEW in different data size scenarios.

**METHODS**

The experiment detail is shown below:

1. Design two pipelines in R and DATAVIEW by using the Random Forest algorithm to predict the nugget width of the joints.
2. Evaluate the implementation time (training and testing) in seconds.
3. Design of Experiment (DOE): The experiment is a full factorial experiment.
   - There are two factors: Implementation Framework, Size of Dataset.
   - Execution Framework at two levels: R, DATAVIEW.
   - Size of Dataset: at four levels 1 (514 kB), 2 (2515 kB), 3 (15,355 kB), 4 (50,000 kB).

R failed to analyze datasets 3 and 4, for this reason two approaches adopted to perform DOE.

Scenario (1): Treating Failed results as NA and analyzing an unbalanced DOE.

Scenario (2): Assuming that R did not fail, implemented the process but at very high costs compared to DATAVIEW (10 times of costs of DATAVIEW). This is in fact penalizing R for not being able to implement the task efficiently. Note that the main purpose of this scenario is conducting statistical analysis on the results.

**RESULTS**

Scenario (1): By conducting unbalanced DOE the Sum of Squares of the factors evaluated and partitioned.

Scenario (2): By conducting this scenario, a balanced DOE will be returned and the Analysis of Variance (ANOVA) can be implemented.

**ACKNOWLEDGEMENTS**

Digital Manufacturing and Design Innovation Institute (DMDII), Ford Motor Company, WSU

**CONTACT**

Changxin Bai, PhD Candidate Department of Computer Science 1; Big Data Lab Department of Industrial Engineering 2 changxin@wayne.edu

**REFERENCES**


In 2016, the Healthy Urban Waters program at Wayne State University (WSU) was asked to host and improve the data network which had been established through federal funding in 2007 to address the challenges of environmental degradation, creation of a sustainable urban environment, and increased public engagement and awareness. The data platform described here is a mass-oriented, user-friendly, and cloud-based system, and has been deployed to provide integrative water quality data in one of the most critical urban corridors of the Great Lake system. In addition, this poster provides several example applications of platform use for temporal and spatial characterization of intake water source quality and urban beach health, through consideration of Escherichia coli, Dissolved Oxygen, pH, and blue-green algae detections along the Huron-to-Erie corridor.

**OBJECTIVES**

The initial objective of the project was to provide early warning of chemical spills that may impact the source of drinking water along the Huron-to-Erie corridor and to communicate both concerns and assurances to the public through this platform. As organizations became aware of this platform, additional environmental data (beyond the source water data) was provided for hosting at this platform. We now have multiple regional projects integrated into the database.

**RESULTS**

- E. coli distribution.
- Significant positive correlation is observed between E. coli and the one-day antecedent precipitation.

**CONCLUSIONS**

Huron-to-Erie water quality data platform is a mass-oriented, user-friendly, and cloud-based system. This platform significantly facilitates the access of data across data providers and agencies for the Southeast Michigan region. Data archived in this platform continues to grow as does its value for environmental studies within the Huron-to-Erie corridor. Moreover, the availability of the data repository and examples of the types of applications that can be completed using the data and the tools associated with the platform are introduced.

**FUTURE WORK**

- Interesting points suggested by example applications deserve further study.
- More in-depth and complex data analytics will be provided in further papers.
- Seamlessly linked with others at the upstream and downstream boundary to expand the system utility.

**ACKNOWLEDGEMENTS**

This project was made possible through a program grant to Healthy Urban Waters from the ERB Family Foundation. We acknowledge the strong support from our collaborators at SEMCOG and the State of Michigan Office of the Great Lakes. In addition, Jamie Olson, must be recognized for his steadfast work to initiate this project.
Traffic crashes have a significant impact on the economy both in the form of property damage and also in the form of lost time. The congestion likely to happen in busy areas will cause waste of gas and air pollution. The worst are fatalities which no one wants to happen. Identifying the crash-prone locations will help traffic safety, transportation planning, and law enforcement to prioritize their efforts and resources to minimize the risk of accidents.

**INTRODUCTION**

There are many factors that can lead to crashes. Making use of data can guide the decision makers in the right direction.

- Location, time, speed, DUI are some of the factors that were analyzed.
- Mapping pedestrian and bike crashes on the same map can provide valuable insights.

**CONCLUSIONS**

- There are many factors that can lead to crashes. Making use of data can guide the decision makers in the right direction.
- Location, time, speed, DUI are some of the factors that were analyzed.
- Mapping pedestrian and bike crashes on the same map can provide valuable insights.

Data source:
https://www.michigantrafficcrashfacts.org/

**CONTACT**

Ilyas Ustun, Ph.D.
Research Fellow - Civil Engineering
Transportation Research Group
ilyasustun@wayne.edu
Automotive Dealership Management: Deriving Tailored Recommendations using Big Data

Haidar Almohri, PhD Student, Ratna Babu Chinnam, Ph.D. 
Industrial & Systems Engineering, Wayne State University, Detroit, MI 48202
Mark Colosimo, David Creech, Daniel Mathias - Urban Science

MOTIVATION
- Research drivers of total automotive dealership’s performance
- Define total dealership as comprised of three departments: New, Used, Aftersales (Primarily Service and Parts)
- Performance of new vehicle department, can be and likely is, influenced by performance of other departments
- First research key drivers within the departments themselves
  - Must be done before determining which relationships between departments matter

OBJECTIVE
- Understand performance dynamics at play within individual departments utilizing data from financial reports
  - Identify important Key Performance Indicators (KPIs) that have impact on performance of department

METHODOLOGY
- Step I: Select potentially important and “actionable” KPIs
- Step II: Define departmental response/performance variable(s)
- Step III: Modeling
  - Divide dealers by region and registration intensity groups (Low, Medium, and High Registrations)
  - Build models for each group within the region
- Step IV: Derive Recommendations

Parts and Service
- Department Performance Measures:
  - F-OP: Fixed Operating Profit % Sales
  - Service Retention: CP ROs Per UIO

Modeling F-OP & Service Retention
- LM using variables selected from Random Forest

F-OP
- Med Registration Group
  - Adjusted Fixed Expense Service % Sales
  - Service ROs Per UIO
  - Parts Revenue Return on Investment
  - Used Parts Mgr. AUC Per TDU
  - Sales Person Per Retail Sales
  - Internal Service % of Total Service
  - Total Parts Sales Per Revenue
  - Counter Retail Parts Per VTO

Service Retention
- Med Registration Group
  - Total CP ROs Per New and Used vehicle sales

Deriving Recommendations through Optimization
- Optimal Pareto Behavior

Comparison of KPIs

Dealer Case Study
Maximizing Dealer’s Profitability
- Estimated performance if recommendations are followed:

Used Department
- Department Performance Measures:
  - U-OP: Used Operating Profit % Sales
  - U-SE: Used “Sales Effectiveness”

Modeling U-OP & U-SE
- LM using variables selected from Random Forest

CONCLUSIONS
- Extensive analysis for automotive dealerships
  - Deriving recommendations through regression and optimization
  - Increasing the profitability of dealers by following recommendations

ACKNOWLEDGEMENTS
This work is supported by Urban Science.

CONTACT
Ratna Babu Chinnam, Ph.D.
Professor, Industrial & Systems Engineering
Co-Director - Big Data & Business Analytics Group
Tel: 313.577.4846
Ratna.Chinnam@wayne.edu
INTRODUCTION

- We study multi-modal retrieval, where each query and collection item can have both a textual fragment and an image and should be considered as an atomic unit. Multi-modal retrieval has been significantly less studied than cross-modal retrieval problem.
- We propose a gated neural architecture to project image and keyword queries as well as multi-modal retrieval units into the same low-dimensional embedding space and perform semantic matching in this space.
- The proposed architecture is trained to minimize structured hinge loss and can be applied to both cross-modal and multi-modal retrieval.

Projection of Textual and Visual Components

- Word2vec embeddings were used to represent the concepts in textual modality of queries that exist in the adopted controlled vocabulary.
- Embeddings of concepts in the visual modality were obtained by adopting the neural architecture that combines a deep CNN for image feature extraction and LSTM for caption generation, considering only the words in the adopted controlled vocabulary as candidate visual concepts.

Relevance matching layers

\[ p(d|q) = \sum_{i,j} d_i^j q_i^j + \rho(Q_d^i, D_d^i) p(D_d^i|Q_d^i) + \rho(Q_c^i, D_c^i) p(D_c^i|Q_c^i) + \rho(Q_t^i, D_t^i) p(D_t^i|Q_t^i) \]

\[ p(d|q) : \text{the probability of a collection item } d \text{ to be relevant to query } q \text{ that we decompose it into concept relevance matching and topic relevance matching scores as:} \]

\[ \rho((\cdot, \cdot) : \text{the prior probabilities for the concept and topical relevance,} \]

\[ \rho(D_d^i|Q_d^i) \rho(D_c^i|Q_c^i) \rho(D_t^i|Q_t^i) \]

Gating Network

\[ \rho(\phi, \psi) \approx 1 - \frac{\sum_{j \in \text{embeddings in } \phi} \min \{ d_j - \psi_j \}}{4k} \text{ for all embedding vectors in } \Psi \text{ and their most similar embeddings in } \Phi. \]

- The second term accounts for the cases when \(|\phi| \neq |\psi|.
- The gate function computes similarity between the most similar pairs in two sets of embedding vectors \(\Psi\) and \(\Phi\).

Proposed neural network architecture for T→TI task

\( S^T = \{ Q_t^1, Q_t^2, \ldots, D_t^{i_1}, D_t^{i_2}, \ldots, D_t^{i_k}, D_t^{\alpha_1}, D_t^{\alpha_2}, \ldots \} \) consists of topic embedding matrices that are used for computing the matching scores at the topic level.

\( S^C = \{ Q_c^i, D_c^i \} \) consists of concept embedding matrices that are used for computing the matching scores at the concept level.

We obtain the clusters of embedding vectors of all concepts in a given modality of a query or collection item

QTTi, CTTi and CITi are low-dimensional representations of the i-th topic in the query’s text, collection item’s text and collection item’s image, respectively.

QTC, CTC and CIC are low-dimensional representations of concepts in query’s text, collection item’s text and collection item’s image.

Results

<table>
<thead>
<tr>
<th>Task</th>
<th>CCQ</th>
<th>DVSH</th>
<th>THB</th>
<th>MCNN</th>
<th>JEMR</th>
</tr>
</thead>
<tbody>
<tr>
<td>I→T</td>
<td>0.4877</td>
<td>0.7230</td>
<td>0.7268</td>
<td>0.7381</td>
<td>0.7538</td>
</tr>
<tr>
<td>T→T</td>
<td>0.6281</td>
<td>0.7109</td>
<td>0.7321</td>
<td>0.7402</td>
<td>0.7835</td>
</tr>
<tr>
<td>I→I</td>
<td>0.5113</td>
<td>0.6083</td>
<td>0.6341</td>
<td>0.6278</td>
<td>0.6984</td>
</tr>
<tr>
<td>T→I</td>
<td>0.6317</td>
<td>0.7552</td>
<td>0.7649</td>
<td>0.7703</td>
<td>0.8092</td>
</tr>
<tr>
<td>T→I</td>
<td>0.4652</td>
<td>0.6745</td>
<td>0.7903</td>
<td>0.7113</td>
<td>0.7563</td>
</tr>
<tr>
<td>T→I</td>
<td>0.5163</td>
<td>0.7468</td>
<td>0.7572</td>
<td>0.7705</td>
<td>0.7825</td>
</tr>
</tbody>
</table>

We hypothesize that the proposed architecture can also be successfully applied to multi-modal e-commerce search and leave validation of this hypothesis to future work.

Conclusion

This paper presents a novel neural architecture for multi-modal retrieval when the query has a single modality and collection items can have multiple modalities.

We hypothesize that the proposed architecture utilizes a hybrid LSTM-CNN network to project the visual modalities, the skip-gram model to project the textual modalities into a common subspace, and a gating network to regulate the information flow by accounting for concept level and topic level matching scores.

The experiments on heterogeneous datasets indicate that the proposed method outperforms state-of-the-art baselines.

We hypothesize that the proposed architecture can also be successfully applied to multi-modal e-commerce search and leave validation of this hypothesis to future work.

CONTACT

Web: http://www.balaneshin.com/
Email: saeid@wayne.edu
Phone: +1 (313) 316 6723