Insight

SFI RESEARCH CENTRE FOR DATA ANALYTICS

A World Leading SFI Research Centre

Science Foundation Ireland sfi For what's next

Book of Abstracts

Insight Student Conference

6th edition

February 12, 2020
National University of Ireland, Galway

Sina Ahmadi, Bianca Pereira

Heike Vornhagen, Piyush Yadav, Alex Acquier, Mona Isazad

Rajdeep Sarkar, Koustava Goswami, Jefkine Kafunah, Bentolhoda Binaei

HOST INSTITUTIONS

DCU, NUI Galway OÉ Gaillimh, UCC University College Cork, Ireland Coláiste na hOllscoile Corcaigh, UCD Dublin

PARTNER INSTITUTIONS

Maynooth University National University of Ireland Maynooth, Tyndall National Institute, Trinity College Dublin Coláiste na Tríonóide, Baile Átha Cliath The University of Dublin, University of Limerick Ollscoil Luimnigh

FUNDED BY:

European Union European Regional Development Fund, Science Foundation Ireland sfi For what's next

Book of Abstracts
Insight Student Conference (ISC 2020)

National University of Ireland, Galway
Insight Centre for Data Analytics

Sina Ahmadi, **Bianca Pereira**
Scientific committe: Piyush Yadav, Rajdeep Sarkar, Koustava Goswami
Local committee: Heike Vornhagen, Alex Acquier, Mona Isazad, Jefkine Kafunah, Bentolhoda Binaei

February 12, 2020

Published by:
Insight Centre for Data Analytics
studentconference2020.insight-centre.org/

Credits:
Book and LaTeX editor: Sina Ahmadi
LaTeX templates for abstracts: Igor Brigadir, Emir Muñoz
using LaTeX's 'confproc' package, version 0.8 (by V. Verfaille)

The 6th Insight Student Conference

Sina Ahmadi, Bianca Pereira
Data Science Institute, Insight Centre for Data Analytics, NUI Galway
firstname.lastname@insight-centre.org

1. Motivation

The Insight Student Conference (ISC) is an annual event organised by students for students. It is a unique occasion for all members of the Insight SFI Research Centre for Data Analytics to network and get to know the research being conducted by students within various research teams. Moreover, it provides an opportunity for students to demonstrate and develop their academic writing, presentation and peer-reviewing skills.

The 6th edition of the conference brings together more than 130 Insight researchers, among students and staff, from six different institutions around Ireland (Figure 1). This Book of Abstracts condenses the rich research being conducted by our students.

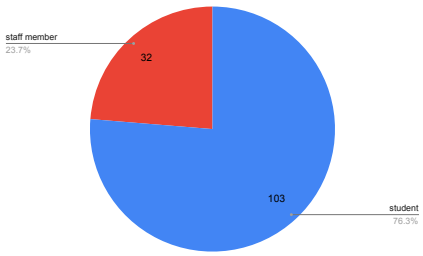


Figure 1: Distribution of the attendees

2. The Conference

The 6th Insight Student Conference carries the tradition of previous conferences, but also proposes innovation. First, our poster session has gone completely paper-free, by replacing posters by screens. We hope that this initiative, even if trivial, helps us to be more environmentally friendly and less wasteful. Second, we have replaced lecture-style presentations by networking and other fun activities such as treasure hunting and a business card sharing competition. Our aim is to create an entertaining environment that allows us to take the most advantage of our time together by increasing opportunities for engagement between students.

3. This Volume

This volume presents all abstracts submitted to the conference which have also been through internal peer-review. All students who expressed interest in submitting to the conference were invited to review at least one abstract. The number of authors and reviewers per Institution can be found in Figure 2. All abstracts within this volume were categorised according to key domain areas (Figure 3).

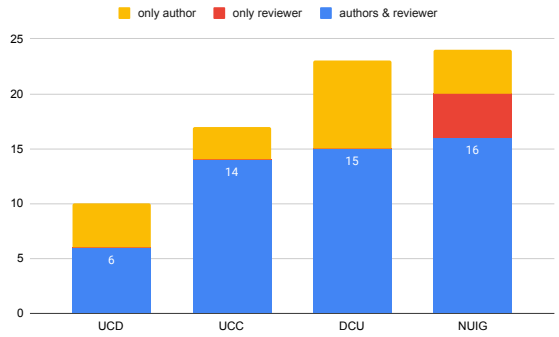


Figure 2: Number of first authors and reviewers per institution (70 first authors, 55 reviewers)

4. Conclusion and Future Work

We hope the discussions and innovations starting at the 6th Insight Student Conference provide a seed for the improvement and success of future conferences. Also, we hope more student-led activities and increased student participation in the future of Insight.

5. Acknowledgments

This conference would not have been possible without the support of everyone involved. We would like to thank the members of the Local Committee for all their logistic support, namely Heike Vornhagen, Alex Acquier, Mona Isazad, and Jefkine Kafunah; the members of the Scientific Committee for managing the submissions and the review process, namely Piyush Yadav, Koustava Goswami and Rajdeep Sarkar; all the administrative support provided by Hilda Fitzpatrick, Christiane Leahy-Coen and Claire Browne; and all the support and motivation provided by Prof. Noel O'Connor and Prof. Mathieu d'Aquin. Last, but not least, we would like to thank all authors, PC members, student supervisors and local volunteers.

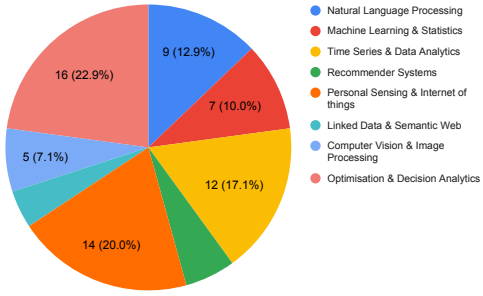


Figure 3: Track distribution of the submissions

Timetable

IS: Invited Speaker, KS: Keynote Speaker, PS: Presentation Session

Wednesday 12 February 2020

09:45–10:20	**Registration, bus arrival (breakfast served on arrival)**		
10:20–10:30	**Welcome remarks**		
10:30–11:00	IS	**Noel O'Connor** Bailey Allen Hall, NUIG	Opening Session
11:00–12:00	KS	**James Hayton** Bailey Allen Hall, NUIG	The basic principles every PhD student needs to know
12:00–13:30	**Lunch**		
13:30–14:30	PS	**First Presentation Session**	
14:30–14:45	**Coffee Break**		
14:45–15:30	PS	**Second Presentation Session**	
15:30–16:30	**World Café (Networking Session)**		
16:30–17:00	**Closing Session**		

Program Committee

(In alphabetical order by last name)

- Jaynal Abedin
- Alex Acquier
- Sina Ahmadi
- Elham Alghamdi
- Eric Arazo
- Felipe Arruda Pontes
- Fouad Bahrpeyma
- Camille Ballas
- Andrea Barraza
- Ardeshir Behrouzirad
- Eoin Brophy
- David Browne
- Diego Carraro
- Bharathi Raja Chakravarthi
- Charmaine Cruz
- James Davenport
- Eoin Delaney
- Oksana Dereza
- Sarah Dillon
- Jaime Fernandez
- Tadhg Fitzgerald
- Agustín García Pereira
- Koustava Goswami
- Venkatesh Gurram Munirathnam
- Agatha C Hennigen de Mattos
- Vitor Horta
- Hong Huang
- Changhong Jin
- Eoin Kenny
- Muhammad Imran Khan
- Liudmila Khokhlova
- Luis Lebron
- Clare Lillis
- Ajay Nair
- Mervyn O Luing
- Daniela Oliveira
- Bianca Pereira
- Sean Quinn
- Priya Rani
- Rajdeep Sarkar
- Sebastian Scheurer
- Yves Sohege
- Annanda Sousa
- Pheobe Wenyi Sun
- Shardul Suryawanshi
- Federico Toffano
- Rafael Torrecilla Rubio
- Mani Vegupatti
- Andrea Visentin
- Congcong Xing
- Piyush Yadav
- Andrea Yanez
- Tarek Zaarour
- Lili Zhang
- Mingchuan Zhao

LIST OF ABSTRACTS

Natural Language Processing

1 Discourse Analysis for Automatic Fake News Profiling
Lucas Azevedo

2 MWSAD-Monolingual Word Sense Alignment Datasets
Sina Ahmadi, John McCrae

3 Leveraging Orthographic Information to Improve Machine Translation of Under-resourced languages
Bharathi Raja Chakravarthi, Mihael Arcan, John P. McCrae

4 Novel Algorithm to build Under Resourced and High Resourced Code-Mix Annotated Corpora of Same Language Family
Koustava Goswami, John McCrae, Theodorus Fransen

5 Offensive content detection in Multimodality
Shardul Suryawanshi, Mihael Arcan, Paul Buitelaar

6 Chatbots in action: Towards the development of scalable domain specific chatbots
Rajdeep Sarkar, John McCrae, Mihael Arcan

7 Predicting Judicial Decisions: A Statistically Rigorous Approach and a New Ensemble Classifier
Andrea Visentin

8 Diachronic Word Embeddings for Historical Languages
Oksana Dereza

9 Challenges in Modelling Minority Languages on Social Media
Priya Rani, John McCrae, Theodorus Fransen

Time Series & Data Analytics

10 Moving Object Path Prediction for Advanced Driver Assistance Systems
Jaime B. Fernandez, Suzanne Little, Noel E. O'Connor

11 An Investigation of Quality Constituents in QoE Subjective Tests: A Case Study
Pheobe Sun

12 CubeMap Matching for On-Demand Query Processing
Suzanne McCarthy, Andrew McCarren, Mark Roantree

13 Using a Graph Model for Agri Data Integration
Congcong Xing

14 An approach to generate diverse time series
Fouad Bahrpeyma, Mark Roantree, Andrew McCarren

15 Time, Distance and Route Inference
Daniel A. Desmond

16 HoliFab: Precise Flow Control using Photo Actuated Hydrogel Valves and PI Controlled LED Actuation for Microfluidic MEMS.
Andrew Donohoe

17 Geospatial Analysis of Peatland Land use and Drainage in Ireland
Wahaj Habib, John Connolly, Kevin Mcguinness, Mathew Saunders, Shane Regan, Declan Delaney

18 Predicting Mastitis Using Machine Learning Applied To Milk Flow Profiles
Changhong Jin, Brian Mac Namee

19 On Privacy Comparison of SQL Queries
Muhammad Imran Khan

20 The Application of CNNs for Sustainable Agricultural Practices Classification
Agustin Garcia Pereira

21 Milk Supply Forecasting
Eoin Delaney

Recommender Systems

22 Debiased Offline Evaluation of Active Learning in Recommender Systems
Diego Carraro

23 Towards Sharing Recommender System Task Environments
Andrea Barraza-Urbina, Mathieu D'Aquin

24 Recommendation of Role Models for Personal Development Planning
Bianca Pereira

25 Pace my race: recommendations for marathon running
Jakim Berndsen

Personal Sensing & Internet of things

26 Person-Independent Multimodal Emotion Detection for Children with High-Functioning Autism
Annanda Sousa

27 Functional knee assessment via multi-sensor approach
Liudmila Khokhlova, Dimitrios-Sokratis Komaris, Salvatore Tedesco, Brendan O'Flynn

28 Deep Learning in Exercise-based CVD Rehabilitation
Ghanashyama Prabhu

29 Towards the Control of Epidemic Spread: Designing Reinforcement Learning Environments
Andrea Yanez

30 Subject-Dependent and -Independent Human Activity Recognition with Person-Specific and -Independent Models
Sebastian Scheurer

31 Evaluating the Impact of Data Loss on Orientation Estimation
Clare Lillis

32 Is Navicular Drop Associated with Running Related Injuries?
Sarah Dillon

33 Convolutional Neural Networks for Heart Rate Estimation and Human Activity Recognition in Wrist Worn Sensing Applications
Eoin Brophy, Tomas E. Ward

34 Developing approaches to biodiagnostics via wearable platforms
Melissa Finnegan, Aoife Morrin

35 SmartSense: Development of a machine learning algorithm for pump clogging prediction
Rafael Torrecilla Rubio

36 Asynchronous Distributed Clustering Algorithm for Wireless Sensor Networks
Cheng Qiao

37 Overview of the Habitat Mapping, Assessment and Monitoring with High-Resolution Imagery (iHabiMap) Project
Charmaine Cruz, John Connolly, Kevin McGuinness

38 Intra-Session Reliability & Discriminative Validity of IMUs as a Measure of the Forward Lunge
James Davenport

39 Bio-Impedance Measurement System for Biomedical Applications
Ardeshir Behrouzirad

Linked Data & Semantic Web

40 Annotating Library Data based on Existing Knowledge Graphs
Daniela Oliveira

41 Efficient Distributed Path Queries on Linked Data Using Partial Evaluation
Qaiser Mehmood, Mathieu d'Aquin

42 Model RDBMS to Knowledge Graph
Mani Vegupatti, Paul Buitelaar, Cécile Robin

Computer Vision & Image Processing

43 3D Object Detection for Instrumented Vehicles
Venkatesh Gurram Munirathnam

44 Mapping Informal Settlements with Machine Learning
Agatha Carolina Hennigen de Mattos, Gavin McArdle, Michela Bertolotto

45 INSIGHT@DCU TRECVID-VTT 2019 and beyond
Luis Lebron

46 POSTER News for Kids
Enric Moreu

47 Rule Based Approach for Facial Expression Recognition in the Wild
Alex Acquier

Optimisation & Decision Analytics

48 A Multi-objective Supplier Selection Framework-based on User-Preferences
Federico Toffano

49 Handling Noisy Constraints in Semi-supervised Overlapping Community Finding
Elham Alghamdi, Ellen Rushe, Mehran Hossein Zadeh Bazargani, Derek Greene, Brian Mac Namee

50 Impact of Segment Duration on DASH-based on-Demand Streaming in Wireless Environment
Abid Yaqoob, Gabriel-Miro Muntean

51 An Energy-Accuracy-Throughput Aware Scheduling for DNN-based Real-time Multimedia Event Processing Systems
Felipe Arruda Pontes

52 The Application of an Outcome-Representation Learning Model for Characterization of the Decision-Making Strategies Used by Younger and Older People
Lili Zhang, Tomas Ward

53 Real-time algorithm configuration through tournament ranking
Tadhg Fitzgerald

54 Blended Control and Deep Reinforcement Learning for Unknown Fault-Tolerant Control
Yves Sohege

55 A Grouping Genetic Algorithm for Joint Stratification and Sample Allocation Designs
Mervyn Oluing

56 Improved Drug Residue Withdrawal Method for Antibiotic usage in Cow's Milk
Cathal Ryan, Brendan Murphy

57 Feature set reduction for improved interpretability of machine learning in radiology
Jingwen Bian, Eric Wolsztynski

58 Optimisation and Acquiring the Solutions in Interactive Constraint Based System
Hong Huang

59 Replicating the VIX
Mingchuan Zhao

60 The Idea of Conceptual Consensus for Entity Management on the Blockchain
Atiya Usmani

61 The Good, the Bad, and the Unexpected: Factors Affecting the Valence of Unexpected Events
Molly Quinn, Mark Keane

62 Analysis of pathogenic and commensal bacterial volatile signatures using solid phase micro-extraction (SPME) coupled with Gas Chromatography – Mass Spectrometry
Shane Fitzgerald, Aoife Morrin, Emer Duffy, Linda Holland

63 Water Level Prediction using LSTMs
Asma Slaimi, Noel O'Connor, Fiona Regan, Susan Hegarty

Machine Learning & Statistics

64 Towards Architecture-Agnostic Neural Transfer
Sean Quinn

65 Pseudo-Labeling and Confirmation Bias in Deep Semi-Supervised Learning
Eric Arazo, Diego Ortego, Paul Albert, Noel O'Connor, Kevin Mcguinness

66 Towards Explaining Deep Neural Networks through Graph Analysis
Vitor Horta, Suzanne Little, Alessandra Mileo

67 Unsupervised PulseNet: K-Means guided Pruning of Convolutional Neural Networks
David Browne

68 Generating Counterfactual Explanations for Deep learning
Eoin Kenny, Mark Keane

69 Bivariate Gamma Mixture of Experts Models for Joint Insurance Claims Modelling
Sen Hu

70 Unstructured Pruning to Reduce the Cost of Training Deep Neural Network
Camille Ballas

73 **List of Keywords**

Discourse Analysis for Automatic Fake News Profiling

Lucas Azevedo
Insight Centre for Data Analytics
lucas.azevedo@insight-centre.org

1. Introduction

Defined as the intentional or unintentional spread of false information through context and/or content manipulation, fake news has become one of the most serious problems associated with online information. Consequently, it comes as no surprise that Fake News Detection has become one of the major foci of various fields of machine learning and while machine learning models have allowed individuals and companies to automate decision-based processes that were once thought to be only doable by humans, it is no secret that the real-life applications of such models are not viable without the existence of an adequate training dataset. In addition to the effort of leveraging such data, the author also provides a document classifier that uses a variety of semantic and syntactic aspects to support a language model that aims to solve the proposed task.

2. Veritas Dataset

There are a number of Fact Checking (FC) agencies that work on the hard tasks of: monitoring social media, identifying potential false claims and debunking or confirming them [1], while providing a narrative that includes sources related to that claim. In the vast majority of their articles, those agencies provide a claim and a veracity verdict - assigned by the FC journalist - for that claim in a structured way, but fail to do so for the web document that originated that claim, here referred as 'origin'. Also, despite the constant effort of the FC agencies, manual fact checking is an intellectually demanding and laborious process, thus the need for its automation.

There are indeed several collections of news that suit different approaches to the fake news detection task but all of those alternatives lack either a significantly large amount of annotated documents [2] or their original document. [3]

2.1. Our Dataset

In order to create the needed dataset, we first automatically crawled[1] the articles contained in these agencies, obtaining more than 11K claims and their veracity verdict.

Although the FC agencies provide claims and their veracity verdict in a structured manner, the articles that originally conveyed the claim is often contained in the narrative of the FC article in the form of a hyperlink and in a completely non-structured way. After obtaining poor results with three different automatic approaches that aimed to identify the origin, we have developed a web interface to accelerate a manual annotation process.

[1] A web crawler (also known as a web spider or web robot) is a program or automated script which browses the World Wide Web in a methodical, automated manner.

3. Proposed Model

The main contribution of this work is the investigation of the usage of linguistic aspects as discriminatory features in a text classifier that should determine whether a given document is more likely to convey false information or not.

Previous work [cite] has been done over the usage of linguistic aspects as features for similar tasks such as deception detection, document clustering, among others. Most of those projects use few (mainly one) linguistic aspects and the majority of them report an improvement of their results by doing so. Inspired by those results, we have defined different evaluation methods for multiple linguistic aspects and have also obtained an improvement in the Fake News classification task in comparison to using only the embedded representation from a language model, e.g., BERT, word2vec, etc.

Subjectivity, Specificity, Complexity, Uncertainty, Affect and Verbal Immediacy are the main classes of aspects measured by our model. We also make use of syntactic features as Diversity, Quantity, Pausality.

4. Conclusion

There is still some advancement to be done on the dataset creation process, although the data generated from the initial annotation sessions, which was used to train the classifier, have yielded very positive improvements to the general accuracy and F1-score of the model.

When analysing the classifier, it is safe to affirm that the usage of linguistic aspects to support the language model is very beneficial. We are aware of a imbued redundancy that our features might present, since the aspects analyzed by the different approaches, in some cases, overlap with each other, but expect that the eventual bias this redundancy might add to the model can be overcome by performing a PCA (principal component analysis) that would additionally provide us with the importance ranking for each of the features.

References

[1] M. Babakar and W. Moy. The state of automated factchecking. *Full Fact*, 2016.
[2] W. Ferreira and A. Vlachos. Emergent: a novel data-set for stance classification. 2016.
[3] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal. FEVER: a large-scale dataset for fact extraction and VERification. New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

MWSAD-Monolingual Word Sense Alignment Datasets

Sina Ahmadi, John P. McCrae
Data Science Institute, Insight Centre for Data Analytics, NUI Galway
sina.ahmadi@insight-centre.org

1 Introduction

Dictionaries form important foundations of numerous natural language processing (NLP) tasks, including word sense disambiguation, machine translation, question answering and automatic summarization. However, the task of combining dictionaries from different sources is difficult, especially for the case of mapping the senses of entries, which often differ significantly in granularity and coverage.

Different dictionaries and related resources such as wordnets and encyclopedia have significant differences in structure and heterogeneity in content, which makes aligning information across resources and languages a challenging task. Word sense alignment (WSA) is a more specific task of linking dictionary content at sense level which has been proved to be beneficial in various NLP tasks, such as word-sense disambiguation [2]. Moreover, combining LSRs can enhance domain coverage in terms of the number of lexical items and types of lexical-semantic information [1].

2 Objective

Given the current progress of artificial intelligence and the usage of data to train neural networks, annotated data with specific features play a crucial role to tackle data-driven challenges, particularly in NLP. In recent years, a few efforts have been made to create *gold-standard* dataset, i.e., a dataset of instances used for learning and fitting parameters, for aligning senses across monolingual resources including collaboratively-curated ones such as Wikipedia[1], and expert-made ones such as WordNet.

In this project, we present a set of datasets for the task of WSA containing manually-annotated monolingual resources in 15 languages. The annotation is carried out at sense level where four semantic relationships, namely, relatedness, equivalence, broadness, and narrowness, are selected for each pair of senses in the two resources by native lexicographers. Given the lexicographic context of this study, we have tried to provide lexicographic data from expert-made dictionaries. We believe that our datasets will pave the way for further developments in exploring statistical and neural methods, as well as for evaluation purposes.

3. Methodology

The main goal of the current study is to provide semantic relationships between two sets of senses for the same lemmas in two monolingual lexical sense resources (LSRs). The actual annotation was implemented by means of online spreadsheets that provide a simple but effective manner to complete the annotation. This also had the added advantage that the annotation task could be easily completed from any device. In order to collect the data that was required for the annotation, each of the participating institutes provided their data in some form providing the following data:

- An entry identifier, that locates the entry in the resource;
- A sense identifier marking the sense in the resource, for example the sense number;
- The lemma of the entry;
- The part-of-speech of the entry;
- The sense text, including the definition.

In order to facilitate the task of annotation, we convert the initial data into spreadsheets with dynamic drop-down lists.

4. Evaluation

We performed an intrinsic evaluation on our datasets by computing a number of resource statistics on the senses.

Danish	3595 (461802)	1667 (18599)
Dutch	176 (3576)	81 (1345)
English (KD)	125 (827)	102 (610)
English (NUIG)	2290 (19082)	1499 (22245)
Estonian	1143 (7178)	1144 (8636)
German	419 (2926)	247 (3013)
Hungarian	132 (1379)	103 (1150)
Irish	1003 (8642)	1324 (7072)
Italian	270 (1987)	208 (1401)
Serbian	386 (2798)	153 (1139)
Slovenian	100 (279)	48 (377)
Russian	100 (715)	147 (1438)

Table 1: Statistics of the datasets (number of tokens in parentheses)

References

[1] I. Gurevych, J. Eckle-Kohler, S. Hartmann, M. Matuschek, C. M. Meyer, and C. Wirth. Uby: A large-scale unified lexical-semantic resource based on LMF. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 580–590. Association for Computational Linguistics, 2012.

[2] R. Navigli and S. P. Ponzetto. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250, 2012.

[1] https://www.wikipedia.org

Leveraging Orthographic Information to Improve the Machine Translation of Under-resourced Languages

Bharathi Raja Chakravarthi, Mihael Arcan, John P.McCrae
Data Science Institute, National University of Ireland Galway
bharathi.raja@insight-centre.org

1. Abstract

The current study aims to examine the correlates of orthographic and cognate information in machine translation of closely related languages. Since closely related languages have a large overlap of vocabulary in-terms of a cognate, it is important to study how to utilize this extra information to improve the machine translation of under-resourced languages. This work investigates between English & Dravidian and English & Goidelic languages.

2. Introduction

Natural Language Processing (NLP) plays a significant role in keeping the languages alive and development of the languages in the digital devices era [4]. One of sub-part of NLP is Machine Translation (MT). MT has long been most promising application of Artificial Intelligence (AI). MT has shown to increase access of information through native language of the speakers in the many case, one such important case is spread of vital information during crisis or emergency in less common languages [5]. Machine translation for under-resourced languages suffers from the availability of data in the form of sentence aligned parallel corpora. However there are the languages which are closely related to the under-resourced language have abundant resources. The exploration of lexical, morphological, orthographic and syntactic similar languages balances the smaller volumes of available data.

We hypothesize that both orthography and cognate are linguistic determinants of mutual intelligibility which may facilitate interconnection. Thus, we need to analyze the impact of orthography and cognate information from closely related languages on the MT of under-resourced languages. We plan to utilize the multilingual neural machine translation system to analyze our hypothesis.

3. Multilingual Neural Machine Translation

Previous work by [3] and [2] extended the architecture of [1] to use a universal model to handle multiple source and target languages with a special tag in the encoder to determine which target language to translate. The idea is to use the unified vocabulary and training corpus without modification in the architecture to take advantage of the shared embedding. The goal of this approach is to improve the translation quality for individual languages pairs, for which parallel corpus data is scarce by letting the NMT to learn the common semantics across languages and reduce the number of translation systems needed.

4. Dravidian Language

Dravidian languages are a family of languages spoken primarily in the Southern part of India and spread over South Asia. Dravidian languages use abugida scripts consists of a consonant-vowel sequence with a consonant core (C^+) and dependent vowel.

5. Goidelic Language

The Goidelic languages form one of the two groups of Insular Celtic languages which are spoken in and around Ireland, Scotland and Isle of Man. The Goidelic branch includes Irish, Scottish Gaelic, and Manx Gaelic. Each is spoken by minority communities in Ireland, Scotland, and the Isle of Man, respectively. All the Goidelic language are considered as under-resourced languages compared to other European languages.

6. Evaluation

There are two major criteria for automatic MT evaluation: completeness and correctness, which are considered by BLEU an automatic evaluation technique which is a geometric mean of n-gram precision. BLEU score is language independent, fast, and has correlation with human judgment so we use it to evaluate our work.

References

[1] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[2] T. Ha, J. Niehues, and A. H. Waibel. Toward multilingual neural machine translation with universal encoder and decoder. In *Proceedings of the International Workshop on Spoken Language Translation*, 2016.

[3] M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado, M. Hughes, and J. Dean. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, Dec. 2017.

[4] A. Karakanta, J. Dehdari, and J. van Genabith. Neural machine translation for low-resource languages without parallel corpora. *Machine Translation*, 32(1):167–189, Jun 2018.

[5] G. Neubig and J. Hu. Rapid adaptation of neural machine translation to new languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 875–880, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics.

Novel Algorithm to Build Under Resourced and High Resourced Mix Annotated Corpora of Same Language Family

Koustava Goswami, Theodorus Fransen, John P McCrae
Data Science Institute, National University of Ireland

1. Abstract

Limited access to data sources is the biggest challenge of different under resourced language families. This problem leads to build less efficient machine learning models. We proposed a method to build code-mix annotated corpora of high resourced and low resourced languages which have similar written dialects. We are also publishing an under-resourced Celtic family language Manx and English code-mix corpora which is annotated and can be used further in any machine learning works and experiments.

2. Introduction

Data collection is one of the vital work that has to be carried out before building any machine learning or deep learning models. Under resourced languages are very hard to analyze and work on due to absence of proper data. Moreover sometimes language data are very ambiguous and very hard to combine and work with. Code-mix corpora of Under Resourced and High Resourced languages of same family is very hard to create. Also, the annotation of data is very hard to manage and work on. It is therefore necessary to make a novel way to build code-mix corpora of under resourced and high resourced languages with annotation at word label identifying individual words.

We have identified a very under resourced language which is Manx from the Celtic family. Manx is known as Manx-Gaelic. It was the first language of people on the Isle of Man. The last native speaker died in 1974[1]. The Manx language uses the same alphabets of English which makes language identification very hard in a English-Manx code-mix corpora on word label. As there is no specific code-mix corpora present for Manx-English, we have collected data from different sources written by native Manx speaker as well as from Wikipedia and created an algorithm to make an English-Manx mix annotated corpora.

3. Related Work

There are many initiatives have been taken by different people to revive the Manx language and make the data available to others work on. Website has been dedicated to work in this language[2].

To make a code-mix corpora from different language some research has been done. Research by (Zamin, 2012) shows how to align word based on the dictionary method. Most recent work has been done by (Chen, 2009) which shows how bootstraping annotation helps from one language to another. While this kind of work has been done, it does not work very well in under resourced languages as there are very few words and sometime directly bootstraping from high recourced languge to low resourced language will not work for annotation.

4. Methodology

To collect data from different online sources, a python pdf to document parser has been written which can access any Manx data source in the internet and parse the texts into a single document with the same name. One single code-mix English-Manx corpora has been built using the datasets.

The texts are tokenized into words to make a dictionary. It seems that the Manx language uses the same dialects of English but it has some different rules in compound word formation. With the help of only the NLTK word tokenizer it has been observed that the compound words are splited wrongly based on which an algorithm has been written to tokenize the word which involves a multistage process – 1. With the help of regular expressions and based on the same dataset 2. With the help of NLTK word tokenizers.

The Bootstrapping process has been followed to annotate the words in the sentences based in the dictionary. In this process the entire corpora have been tokenized at word level by four types of tokenizers – M(Manx), E(English), U(Unidentified), P(Punctuation).

5. Conclusion and Future Directions

The code-mix annotated corpora of Manx-English is one of the biggest datasets that are available for different kind of machine learning algorithm building which can successfully lead to make a model of machine translation and language Identification work. The evaluation has been done based on a source cleaned datasets which leads to quite accurate annotated sentences. In future the same handcrafted work will be built on by means of a deep learning algorithm which can lead to make novel unsupervised language annotation model for under resourced languages.

References

Zamin, N. a. (2012). A statistical dictionary-based word alignment algorithm: An unsupervised approach. 2012 International Conference on Computer \& Information Science (ICCIS) (pp. 396--402). IEEE.

Chen, Z. a. (2009). Can one language bootstrap the other: a case study on event extraction. Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing (pp. 66--74). Association for Computational Linguistics.

[1] https://en.wikipedia.org/wiki/Manx_language

2 https://www.learnmanx.com/

Multimodal Meme Dataset (MultiOFF): Identifying Offensive Content in Image and Text

Shardul Suryawanshi, Mihael Arcan,and Paul Buitelaar
Data Science Institute, National University of Ireland Galway
shardul.suryawanshi@insight-centre.org

1. Abstract

A meme is a form of media that spreads an idea or emotion across the internet. As the interaction over the web increased with multimodality, incidents of posting hateful memes and related events like trolling, cyberbullying have increased through memes. Offensive content detection has been extensively explored in single modality such as text or image. However, combining two modalities to identify offensive content is still a developing area.

2. Introduction

Identifying offensive content in memes is more challenging since it expresses humour and sarcasm in an implicit way. If we only consider the text or the image of a meme, it might not be offensive. Therefore, it is necessary to combine both modalities to identify whether a given meme is offensive or not. Since there was no publicly available dataset for offensive content in memes, we developed a new dataset called "MultiOFF " for offensive meme classification, using human annotators. This dataset is based on the Kaggle 2016 US presidential election meme dataset [1]. On the basis of this dataset we developed an early fusion approach to combine evidence from both the image and text modality. We compared our approach with a text-only and image-only baseline to investigate its performance. Results on combining modality for memes show improvement in terms of Precision, Recall, and F-Score.

3. Related work

Research by [2] proposes the deep neural network architecture to be followed to deal with multimodal data. Their dataset relies on the tag attached to the social media posts as a label while the dataset used in the proposed work has been annotated manually. In their research [2] emphasize more on emotion analysis, unlike our research which gives importance to the detection of offensive content. [1] proposes different types of architectural designs that can be used to classify or predict multimodal data. While their research delves into emotion classification based on multimodal data, it does not match with the objective of this research, i.e. binary classification of memes into offensive and non-offensive.

4. Methodology

Text baselines based on logistic regression (LR), naive bays (NB), stacked LSTM, Bidirectional LSTM (BiLSTM) and convolutional neural network (CNNText) has been used. On the other hand, VGG16 image classifier has been used as an image baseline. Stacked LSTM, BiLSTM and CNNText have been used in combination of VGG16 to build a multimodal classifier. The visual and textual features extracted from the meme has been concatenated in a vector representation to train the multimodal classifier. The multimodal classifier used in the experiments are Stacked LSTM + VGG16, BiLSTM + VGG16, CNNText + VGG16. Multimodal approach has been illustrated in figure 1.

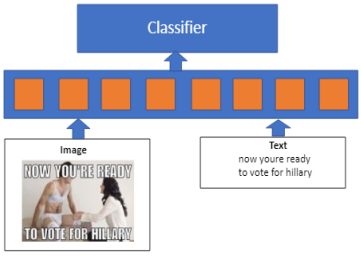


Figure 1: Multimodal classifier

5. Conclusion and Future Directions

From the experiments, it has been found out that the ability to retain most of the offensive content could be increased by a multimodal classifier. It is still debatable if the accuracy of such multimodal system is reliable. As a remedy, the manual evaluation from the administrator could be useful before blocking content identified as offensive. As a future direction, concatenating the image and text embeddings for representing meme could be improved upon by fusing embeddings.

References

[1] C. T. Duong, R. Lebret, and K. Aberer. Multimodal classification for analysing social media. *ArXiv*, abs/1708.02099, 2017.

[2] A. Hu and S. Flaxman. Multimodal sentiment analysis to explore the structure of emotions. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 350–358, London, UK, 2018. ACM.

[1] https://www.kaggle.com/SIZZLE/2016electionmemes

Chatbots in action: Towards the development of scalable domain specific chatbots

Rajdeep Sarkar, John McCrae, Mihael Arcan
Insight Centre for Data Analytics, NUI Galway
firstname.lastname@insight-centre.org

1. Abstract

The use of chatbots is gaining popularity with time in different segments of life. However, building high quality models which are scalable to multiple domains is time-consuming and expensive due to the lack of training data for each domain. In this work, I plan on using external knowledge in the form of knowledge graphs and recommendations to improve the performance of chatbots and extend then to multiple domains.

2. Introduction

Chatbots are becoming a ubiquitous part of customer service both for their cost savings and ease of use, however developing a chatbot for a new situation is still a time-consuming and expensive process. Recent works on chatbots show a shifting trend from template-based systems towards deep learning architectures [2] [5]. An encoder encodes the utterance text into a fixed size vector which is then fed into the decoder to generate a response text. Encoder-Decoder architectures use deep neural networks internally such as Bi-LSTMs, Bi-GRUs, Transformers etc. However, major drawbacks of such architectures are the scalability of the systems into multiple domains and lack of domain-specific training data. External knowledge in the form of knowledge graphs or recommendations (for e-commerce) contain rich knowledge about the user and the entities present in the chats [1]. Hence, incorporating external knowledge can help in enriching the knowledge learnt by the model. Recent works on learning embeddings for graphs such as rdf2vec[6], graph convolutional networks and graph attention networks enables us to incorporate the knowledge as fixed size vectors. However, there is still much to do in terms of intent detection, slot filling, natural language understanding, dialog management and response generation. In this work, I would be primarily working towards improving the state-of-the-art in the above tasks for domain specific chatbots.

3. Related Work

Recent trends show the use of deep learning architectures for end-to-end chatbots and dialog agents. [5] built an end-to-end model using fusion networks which tries to improve the performance of natural language understanding, dialog manager and response agent. [3] consider using future utterances to improce the performance of their end-to-end model. However, not much work has gone into studying the use of external knowledge in dialog agents. [1] show that including external knowledge in the form of recommendation can help in improving the performance of chatbots in e-commerce. However, external knowledge needs to be updated with time and [4] demonstrated that dialog agents can be used to learn factual knowledge in graphs such as DBpedia and ConceptNet.

4. Methodology

My major contribution would be the inclusion of external knowledge into dialog agents which would lead to improved performances for slot filling, intent detection and response generation. I would also work towards making the dialog-agent scalable to multiple domains without the need for high volume of domain specific data.

5. Conlusion

This work would lead to chatbots being scalable to multiple domains without the need of exorbitant amounts of data, by using domain specific external knowledge.

References

[1] Q. Chen, J. Lin, Y. Zhang, M. Ding, Y. Cen, H. Yang, and J. Tang. Towards knowledge-based recommender dialog system. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1803–1813, 2019.

[2] A. Gupta, J. Hewitt, and K. Kirchhoff. Simple, fast, accurate intent classification and slot labeling for goal-oriented dialogue systems. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 46–55, 2019.

[3] Z. Jiang, X.-L. Mao, Z. Huang, J. Ma, and S. Li. Towards end-to-end learning for efficient dialogue agent by modeling looking-ahead ability. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 133–142, 2019.

[4] S. Mazumder, B. Liu, S. Wang, and N. Ma. Lifelong and interactive learning of factual knowledge in dialogues. *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 21–31, 2019.

[5] S. Mehri, T. Srinivasan, and M. Eskenazi. Structured fusion networks for dialog. *arXiv preprint arXiv:1907.10016*, 2019.

[6] P. Ristoski and H. Paulheim. Rdf2vec: Rdf graph embeddings for data mining. In *The Semantic Web – ISWC 2016*, pages 498–514, 2016.

Predicting Judicial Decisions: A Statistically Rigorous Approach and a New Ensemble Classifier

Andrea Visentin
University College Cork
andrea.visentin@insight-centre.org

1. Introduction

Natural language processing and machine learning are gaining wide popularity in supporting judicial decision-making. Research in this area is particularly active. However, a methodological issue in the use of AI methods can lead to poor statistical soundness in the results. We consider and improve the work of Aletras et. al. [1] for predicting the outcome of cases at the European Court of Human Rights. We replicate their experiments using a more statistically reliable methodology and analyzed the results using state-of-the-art Bayesian techniques for classifier comparison. We also improved classification accuracy using an ensemble-based approach. These techniques will widely improve the statistical soundness of machine learning applications in law by providing robust baselines for comparison.

2. Research objectives

Two main technical issues are present in the work of [1] are a single run of the 10-cross validation and a non statistical sounding classifiers compariso Our method comprises the following:

- A Bayesian statistically sound method to compare classifier.
- A new way to make juridical decisions based on documents with different subsections, motivated by work from ensemble classifiers.
- An extension of the previous technique to multiple classifiers.

3. Datasets

The datasets are defined by a set of features extracted from cases from the European Court of Human Rights (ECtHR). The court has jurisdiction to rule on the possible violations of the European Convention of Human Rights (ECHR).
The datasets relate to cases associated with violations of Article 3, 6 and 8. Cases are divided into a strict number of sections and subsections: *Procedure*, *Circumstances*, *Relevant Law*, *Law*. The text features extracted from these cases are of two types.
N-gram features: Using the Bag-of-Words (BOW) model the different subsections are represented as numeric vectors.
Topics: N-grams that are semantically similar are clustered. These vectors represents the 30 most frequent topics in the court cases.
SVM is used as classifier.

4. Classifiers comparison

We decided to utilize the *Bayesian correlated t-test* presented in [2] to give statistical soundness to the experimental part of this work. It is a non-parametric test that allows us to evaluate the possibility of one classifier being better than the other or the probability of the two classifiers being *practically equivalent*.

5. Voting classifier

We introduced a **voting** classifier. Out of the classifiers presented herein, only those built from the subsections can be considered as fully independent. We combine these using a simple majority voting scheme. The performances of the voting classifier are better or equivalent than the the previous techniques on average and the difference is statistically significant in most of the comparison.

6. Ensemble

Different algorithms exploit the data in different ways, with the result that a given algorithm may significantly outperform another on one problem instance. Classifiers, other than than SVM, can outperform it on average or in particular subsection. We want to repeat the previous experiment with different classifiers, and then assign as a final decision the most common among those. It can be considered as an ensemble of ensembles. We selected two more classifiers: *gradient boosting machine (GBM)* and *K-nearest neighbour (KNN)*.
The improvement of the accuracy is minimal compared with the best of the individual classifier and it is not statistically significant. We plan to do more tests with bigger datasets and different classifiers.

References

[1] N. Aletras, D. Tsarapatsanis, D. Preoţiuc-Pietro, and V. Lampos. Predicting judicial decisions of the european court of human rights: A natural language processing perspective. *PeerJ Computer Science*, 2:e93, 2016.
[2] A. Benavoli, G. Corani, J. Demšar, and M. Zaffalon. Time for a change: a tutorial for comparing multiple classifiers through bayesian analysis. *The Journal of Machine Learning Research*, 18(1):2653–2688, 2017.

Diachronic Word Embeddings for Historical Languages

Oksana Dereza
Unit for Linguistic Data, Insight Centre for Data Analytics, NUIG
oksana.dereza@insight-centre.org

1. Abstract

The surge of interest to distributional semantics has lately reached historical linguistics. The recently emerged concept of diachronic, or dynamic embeddings transforms the task of language modelling into the task of modelling language change. This work is aimed at finding an optimal solution to this problem for historical languages, such as Old and Middle Irish, Gothic, Latin, Ancient Greek, Old Church Slavonic etc.

2. Related work

Most of the published work on diachronic word embeddings is focused on semantic change in modern languages and covers only a short (from the historical point of view) time span, not exceeding two centuries. For example, [6] explores semantic evolution in news articles in English over 27 years; [5] and [3] also focus on English, but take a broader time span of 110 and 160 years respectively using Google Books. The authors of [2] run their model on several languages, covering the period of 200 years for English, French and German, and 50 years for Chinese. In general, current research on the topic, extensively described in [4], is primarily algorithmic and centered around training and alignment of diachronic embeddings. However, the scope of languages that have been used in experiments so far is quite narrow, and none of the proposed methods has been tested in non-ideal conditions, with hindering factors such as high spelling variation, substantial grammatical changes or the lack of data.

3. Research questions

Given the open challenges outlined in the previous section, we would like to focus on historical languages as they allow us to address several aspects of diachronic language modelling that have not yet received proper attention. Firstly, working with a larger period of time makes it possible to track not only semantic shifts, but also developments in morphology and syntax, which evolve slower than lexicon. Secondly, historical language data tends to be both scarce and inconsistent, which provides a perfect test case to evaluate how robust existing algorithms are and to map out the ways of their improvement. This leads to the following research questions.

1. What will be captured by diachronic embeddings trained on historical language data?
    - Will they model both semantic and grammatical change?
    - Will they generalise spelling variation?
2. Is training across multiple time spans better than treating different historical stages of a language as separate languages and training models for them individually?
    - What is the best step size for splitting the data? How to deal with vague datings in the training corpus ("XII – XV centuries", "before 800")?
    - How can we overcome data scarcity for a particular period?
3. What is the most effective evaluation scenario for this task?

4. Approach

We follow [1] and [6] and take joint learning as a primary approach. Firstly, simultaneous training across different periods eliminates an additional alignment step by placing all embeddings in the same vector space from the beginning. Secondly, it allows using the data from well-resourced periods to improve embeddings for under-resourced periods. Jointly trained models will be evaluated against those trained separately for each period and aligned a posteriori. Currently, we are working on data collection, which includes both standardising existing historical corpora from UD, CLARIN and CLTK repositories, and web-scraping digital editions of medieval texts, published at UCC CELT, Rhyddiaith Gymraeg, IcePaHC, Project Madurai etc.

References

[1] R. Bamler and S. Mandt. Dynamic word embeddings. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 380–389. JMLR. org, 2017.

[2] W. L. Hamilton, J. Leskovec, and D. Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1489–1501, 2016.

[3] Y. Kim, Y.-I. Chiu, K. Hanaki, D. Hegde, and S. Petrov. Temporal analysis of language through neural language models. *Proceedings of the ACL Workshop on Language Technologies and Computational Social Science*, pages 61–65, 2014.

[4] A. Kutuzov, L. Øvrelid, T. Szymanski, and E. Velldal. Diachronic word embeddings and semantic shifts: a survey. *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, 2018.

[5] A. Rosenfeld and K. Erk. Deep neural models of semantic shift. In *NAACL 2018 Proceedings: Human Language Technologies, Volume 1*, pages 474–484, 2018.

[6] Z. Yao, Y. Sun, W. Ding, N. Rao, and H. Xiong. Dynamic word embeddings for evolving semantic discovery. In *Proceedings of the 11th International Conference on Web Search and Data Mining*, pages 673–681. ACM, 2018.

Challenges in Modelling Minority Languages on Social Media

Priya Rani, Theodorus Fransen, John P. McCrae
Data Science Institute, National University of Ireland Galway
Priya.Rani name@insight-centre.org, Theodorus.Fransen@insight-centre.org, John.Mccrae@insight-centre.org

1. Introduction

The languages which do not have the privilege to be called as a national or official language are known as minority language. These are languages that have no linguistic data for the implementation of language technologies. In recent decades the computer-mediated communication has been engaged and attracted millions of people to create social network around the world. Computer-mediated communication has become one of the biggest platforms to extract linguistic data for different natural language processing applications. As the evolution of social media has created an abundance of linguistic data for information access and language technology, it has posed new challenges for language technology.

Code mixing is the most frequent user-generated content on social media, primarily occurring in multilingual societies, where language change takes place over a very short geo-spatial distance. It is a phenomenon where the speaker switches from one language to another. In most of the cases of code-mixing mixing happens with the dominant or most influential language. The current paper presents an overview of challenges in modelling minority language on social media.

2. Literature review

Modelling of the minority languages is a research problem which has been studied form decades and researcher have succeeded to get valuable results from these research. The automatic query generation technique was used to built corpus of documents with frequency and have greater coverage of vocabulary [1]. Hybrid filtering takes all the the sentences except unknown vocabulary and apply minimal blocks filtering on the rejected sentences. In minimal block filtering it takes only those blocks which have at least 5 known vocabulary. These techniques had an excellent result on building the language model from web corpus.[2].

3. Challenges

The common and the significant challenges are the lack of interest and finance while dealing with minority languages, as these languages are excluded and do not have a large amount of available literature for language technology development. Recently due to the different state of art for modelling minority languages that have been framed and experimented with a change in the scenario of social media use and technicality, several new challenges have been developed. The previously used methods for modelling language such as automatic query generation and bootstrapping don't seem to be appropriate for the modelling of the languages with these new challenges.

3.1. Corpus Collection

To develop the corpora for minority language from social media several methods are used, such as language-specific keyword or region-specific URL. These methods have several other advantages and disadvantages. Though keyword search gives specific language data, still collecting and evaluating the keywords for an unknown language is challenging. Similarly, with specific region URLs, we may get the corpora of all the languages spoken in the area.

3.2. Language Identification

The second challenge that one faces after the collection of the corpus is language identification. As most of the corpus collected from social media are code mixed, It's tough to segment the specific language sentences. The widespread and challenging problem is to decide at what level we should identify the language. Is it at word level, phrase level or at sentence level?

3.3. Lexical Error

The lexical error includes typo error, loan words, abbreviations. When we talk about code-switching, we assume that the users know both the language, therefore the switching could occur between two closely related languages as well. The problem with the code-switching between closely related language is how to know whether the word is a cognate or just a typo error.

4. Conclusion and Future Work

In this paper, I tried to introduce a few challenges that one faces while modelling minority language on social media. The development of a good and functional language model would eliminate all the challenges and thus help in building language technology.

References

[1] R. Jones and R. Ghani. Automatically building a corpus for a minority language from the web. In *ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS*, volume 38, pages 29–36. UNKNOWN, 2000.

[2] V. B. Le, B. Bigi, L. Besacier, and E. Castelli. Using the web for fast language model construction in minority languages. In *Eighth European Conference on Speech Communication and Technology*, 2003.

Moving Object Path Prediction for Advanced Driver Assistance Systems

Jaime Fernandez
Insight Centre for Data Analytics, Dublin City University
jaime.fernandez@insight-centre.org

1. Introduction

Autonomous navigation requires appropriate models of the static and dynamic environment being explored. Significant advancement has been reached in moving-object-free scenarios such as robotics. In contrast, locations with moving objects, for instance, pedestrians and cars, still pose significant challenges [1], [2]. Vehicle and pedestrian detection [3], [4] have also witnessed significant progress over the years, and methods that allow for the detections of these are increasingly reliable. Knowing where an object is located is already useful, however, predicting its location in the future is of great importance for Advanced Driver Assistance Systems (ADAS) and for application in assistive technology such as navigation of blind people, where path prediction can be used to guide a person to avoid collision. In this research, the main focus is on egocentric cameras, as those mounted on a vehicle or a person. Another relevant point is the interest in working with a holistic approach in order to use more information in the prediction task.

2. Motivation

Automobiles equipped with ADAS and sensors such as cameras, radars and LIDARs are now common place. Many of the accidents on the road can be avoided or can be mitigated at least by acting seconds in advance. For this reason, safety on the road is one of the main objectives in the development of ADAS. Predicting where a pedestrian or a vehicle will be in the near future in a scene can provide useful information that allows an ADAS to react in those seconds. Another motivation is the project VI-DAS where the result of this research was applied for understanding the outside sensing.

3. Objective

Developing of a novel technique to predict the possible path of moving objects in a traffic scene such as pedestrians and vehicles based on egocentric cameras from a moving vehicle and intelligent map data.

4. Approaches

The approaches developed to address the problem of path prediction can be classified according to the information considered and the assumptions made when using them. From the simplest to the most complex, the following classification can be made: physical-based, manoeuvre-based and interaction-aware. Physical-based approaches only take into account basic information such as the previous location of the object. Manoeuvre-based approaches first classify the action that the object is likely to carry out, i.e.: stopping, accelerating. Given this, these approach assume that the next action will match the current manoeuvre. Finally, interaction-aware approaches do not only take into account the object as an isolated identity but also how its action and the action of other objects will affect its possible path. Currently, LSTMs are being explored and the combination with other approaches such as Mixture Density Models.

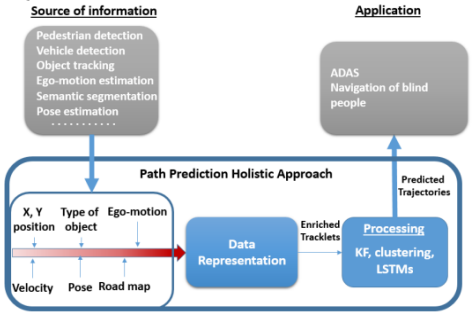


Figure 1. Holistic approach for path prediction.

5. Future

Experiments have shown that using more features that only *x,y* position improve accuracy. They have also evidenced that better metrics are need to really represent the similarity between two paths. The next steps will be to explore more features such as a semantic map of the road, ego-motion and interaction with other vehicles. New metrics will also be studied.

Acknowledgement

This research has been funded by EU H2020 Project VI-DAS under grant number 690772 and Insight Centre for Data Analytics funded by SFI, grant number SFI/12/RC/2289.

References

[1] Campbell, M., Egerstedt, M., How, J. P., & Murray, R. M. (2010). Autonomous driving in urban environments: approaches, lessons and challenges. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 368(1928), 4649-4672.
[2] Bengler, K., Dietmayer, K., Farber, B., Maurer, M., Stiller, C. and Winner, H., 2014. Three decades of driver assistance systems: Review and future perspectives. IEEE Intelligent Transportation Systems Magazine, 6(4), pp.6-22.
[3] Sivaraman, S. and Trivedi, M.M., 2013. Looking at vehicles on the road: A survey of vision-based vehicle detection, tracking, and behavior analysis. IEEE Transactions on Intelligent Transportation Systems, 14(4), pp.1773-1795.
[4] Dollar, P., Wojek, C., Schiele, B., & Perona, P. (2012). Pedestrian detection: An evaluation of the state of the art. IEEE transactions on pattern analysis and machine intelligence, 34(4), 743-761.

An Investigation of Quality Constituents in QoE Subjective Tests: A Case Study

Pheobe Wenyi Sun
Insight Centre for Data Analytics, UCD
wenyi.sun@insight-centre.org

1 Introduction

The objective quality of experience (QoE) prediction models are used to estimate human judgment of their satisfaction level of given stimuli. They are widely used to assess how human-centric a service design is. The predicted QoE scores can be used to either compare different algorithmic solutions or to be used as a guide to improve a service. In order to ensure the predicted QoE scores as close to human judgment as possible, subjective tests are used to generate the ground truth to build an objective model or to be used to check the reliability of an existing QoE prediction model under different conditions.

The most prevalent QoE models were developed and used in the telecommunication industry where the use scenarios were limited to voice transmission. Problems occur when researchers attempt to use similar human perception prediction methodologies, including conducting subjective assessment and building objective prediction models, in a broader scenario. For example, when comparing two algorithms that are used to combat the network jitter on the web, conflicting subjective responses were gathered from the subjective tests. Thereafter, the effect on the perceived quality effect was left unknown. To solve this problem, modification in the current subjective tests is needed.

2 Background

Jitter buffer is a technique used in the web-based real-time communications (WebRTC) when streaming media files to maintain the perceptual flow in users given fluctuating network conditions [1]. Different jitter buffer algorithms have different ramifications. The perceptual results are therefore different. Platform's choice to adopt a buffering algorithm depends on its impact on the actual human experience [4]. However, the comparison among different algorithms based on QoE tests is made difficult as conflicting results were found in both objective and subjective QoE tests.

3 Objectives

This research will investigate the questions currently asked in the most popular subjective tests, find the ambiguous concepts used in the questions and the various possible interpretations of each, and run a subjective experiment to test the aspects of concerns people have when assessing the overall quality in a particular user case – playing back audio over the web using WebRTC platforms.

In this scenario, people are facing the quality changes as a result of time scale modification. To see whether the subjective tests can still provide a robust subjective preference despite the interpersonal difference, we are going to answer the below questions first: 1) Do the traditionally designed QoE questions truly reflect what people experience in this scenario? 2) What aspects of speech features do people focus on when judging their QoE in this scenario?

4. Methodology

4.1 State-of-the-art Subjective Tests

Current subjective tests ask the subjects to rate their perceived 'quality' on a five-point category scale from 1 to 5 [3]. The tests draw conclusions on a service' perceptual quality based on the mean opinion score (MOS) [2].

4.2 Modified Subjective Tests

In order to investigate the many aspects people are concerned about when judging the 'quality' of the test stimuli in a listening test, we specify the question by asking the participants to rate different aspects of speech features that affect the overall QoE (e.g., intelligibility and naturalness) respectively. The correlation between the traditionally-measured QoE MOS score and the new MOS scores on many different aspects of speech features will be used to indicate the key factors forming people's QoE judgment in the case of WebRTC.

We will use the results from a subjective test comparing people's preferences over two different WebRTC platforms as the dataset for this experiment.

5. Current Progress

A trial experiment will be carried out following this research proposal. Depending on the result, subjective experiments on a larger scale will be carried out with the hope to inform novel QoE models applicable to a broader scope of scenarios.

References

[1] Y. Cinar, H. Melvin, and P. Pocta. A black-box analysis of the extent of time-scale modification introduced by webrtc adaptive jitter buffer and its impact on listening speech quality. *Communications - Scientific Letters of the University of Zilina*, 18(1):17–22, 2016.

[2] ITU-T P.800. Telephony transmission quality, Methods for subjective determination of transmission quality. *ITU-T Recommendation*, 800:29, 1996.

[3] ITU-T P.830. Subjective performance assessment of telephone-band and wideband digital codecs. 4, 1996.

[4] P. Počta, H. Melvin, and A. Hines. An analysis of the impact of playout delay adjustments introduced by VoIP jitter buffers on listening speech quality. *Acta Acustica united with Acustica*, 101(3):616–631, 2015.

CubeMap Matching for On-Demand Query Processing

Suzanne McCarthy
Insight Centre for Data Analytics
Suzanne.McCarthy@insight-centre.org

1. Introduction

The goal of this research is to develop a methodology that integrates new online source data with existing enterprise or web sources to automate the creation of new data marts. Data cubes are usually pre-computed to provide datasets for machine learning and decision support. Precomputation is too costly where a high volume of cubes are rarely used and is limited by its usage to in-house data only. Our approach uses on-the-fly cube construction, incorporating new data sources and reuses existing cube data where possible. In this work, we present CubeMaps which provide a fine-grained description of data cubes for easy matching with user queries represented using the same model.

2. Related Research

Data warehousing provides OLAP functionality and cubes as the richest representation of data to feed into analyses, decision support and machine learning. The issue with data warehouses is the expense in maintaining existing data cubes and building the gateways to introducing new data sources. Recent research[1,2,3] is now focused on finding a solution to a more on-demand style of cube construction where new data sources can be introduced as required.

The authors in [4] propose an On-Demand query fulfillment framework, very similar to ours. In this work, a dice is an abstraction of a set of facts (where, for the purpose of this particular work, facts refers to both measure data and dimensional data), where those facts may be contained in existing data cubes or may be missing; and may be required to fulfill a query or may be irrelevant to the query. In the case where the facts are contained within the existing cubes, but are irrelevant to the query, the facts will be dropped from the cube if space requires it. In the case where the facts are required for the query but missing from the cube, the facts will be extracted from their source. Therefore, the query is launched against a dice management system to determine whether to perform an ETL process, a dropping process or a filtering process in the case where the facts fetched from the source are a superset of those needed to fulfill the query. In our case, we drop data from cubes in a similar fashion but we do so as soon as it is no longer needed to fulfill the query. This leads to a more lightweight, dynamic, view-based query fulfillment rather than permitting data redundancy until the space is required. However, unlike the work in [4], we have the expectation that the data will be refreshed very frequently and that new data sources may be added and, when they do, these new sources will need to be integrated into the system seamlessly, without a large manual effort on the part of the developer.

3. Evaluation setup for Future Work

To setup our evaluation, we used our system [5,6] to import 61 data sources from 40 web sites and automatically create the 61 data cubes. As a consequence, we have 61 CubeMaps in the metadata repository. The evaluation and application of these CubeMaps for on-demand query fulfillment will now form the basis for future works.

4. References

[1] Yagiz Kargin, Holger Pirk, Milena Ivanova, Stefan Manegold, and Martin L. Kersten. 2012. Instant-On Scientific Data Warehouses - Lazy ETL for Data-Intensive Research. In Enabling Real-Time Business Intelligence - 6th International Workshop, BIRTE 2012, Held at the 38th International Conference on Very Large Databases, VLDB 2012, Istanbul, Turkey, August 27, 2012, Revised Selected Papers. 6075. https://doi.org/10.1007/978-3-642-39872-8_5

[2] Ying Yang. 2014. On-Demand Query Result Cleaning. In Proceedings of the VLDB 2014 PhD Workshop.

[3] Ying Yang, Niccol Meneghetti, Ronny Fehling, Zhen Hua Liu, and Oliver Kennedy. 2015. Lenses: An On-Demand Approach to ETL. PVLDB 8, 12 (2015), 15781589. https://doi.org/10.14778/2824032.2824055

[4] Lorenzo Baldacci, Matteo Golfarelli, Simone Graziani, and Stefano Rizzi. 2017. QETL: An approach to on-demand ETL from non-owned data sources. Data Knowl. Eng. 112 (2017), 1737. https://doi.org/10.1016/j.datak.2017.09.002

[5] Suzanne McCarthy, Andrew McCarren, and Mark Roantree. 2018. Combining Web and Enterprise Data for Lightweight Data Mart Construction. In Database and Expert Systems Applications - 29th International Conference, DEXA 2018, Regensburg, Germany, September 3-6, 2018, Proceedings, Part II. 138146. https://doi.org/10.1007/978-3-319-98812-2_10

[6] Suzanne McCarthy, Andrew McCarren, and Mark Roantree. 2019. A Method for Automated Transformation and Validation of Online Datasets. In Workshops Proceedings of the 23rd International IEEE Enterprise Distributed Object Computing Conference, ECOCW 2019, October 2019, Paris, France.

Using a Graph Model for Agri Data Integration

Congcong Xing
Dublin City University
congcong.xing2@mail.dcu.ie

1. Background and Motivation

Data Integration is a major process in data warehouse construction. It is essential for combining data from different data sources which contain heterogeneous data into a unified view of data to be used for downstream data analytics.

Agriculture (Agri) researchers tend to have large amounts of raw data that is disconnected and in different data types/formats. In terms of data analytics, a large amount of time will be consumed in data engineering work because of complicated data integration processes. In practice, manual intervention is difficult to avoid.

2. Proposed Solution

An automated data integration platform will be developed to speed up the process of integrating agri-data based on graph theory. In this Graph Based Integration Platform, data from heterogeneous sources will be divided into two categories: metadata and instance data, where metadata will be transformed to ontology description format (RDF/OWL) and feed into the ontology matching platform – YAM++ to be processed.

Upon completing the ontology matching process by consuming YAM++ APIs, an alignment file will be produced which will contain the matching results. Results from the alignment files can be used as a guide in the process of metadata integration. Once metadata integrated, the original raw data which is the instance data can be then easily integrated based on the integrated metadata.

3. Agri-Data Experiment

Agri-Data was extracted as **csv** format before the integration process. When put into Graph, metadata from **csv** files was extracted and processed into the Graph in a direct mapping method.

- **Direct Mapping Method of CSV/Table format:** Schema, Table and Column are the three main nodes in Metadata Graph. Table/csv spreadsheet names are extracted as Table nodes with the original name property. Column names are extracted as Column nodes with the original name property. The relationships between the three nodes are:

  Example 1. $(:Column) - [belong\_to] - > (:Table) - [:belong\_to] - > (:Schema)$

After being extracted and loaded as a meta-graph, metadata is exported as RDF (resource description framework) to be put into the ontology matching tool YAM++. YAM++ then processes the ontology using string matching, machine learning and information retrieval technology to get an ontology alignment result which is used as metadata integration instruction for downstream work. Finally, metadata is integrated with the instruction of YAM++ and domain experts' knowledge.

- **Nodes Integration in Graph:** After the alignment result is returned by YAM++, with domain knowledge of experts, Column nodes from different Tables which have suggested [:same_as] relationship are merged as one node.

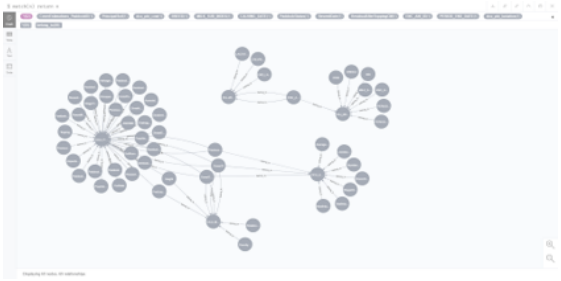


Figure 1: Agri-Data Integration Result

- There are 5 csv files loaded in the experiment:
  DCU_LIVESTOCKNUMBERS (Live stock intake)
  DCU_MANAGEMENDECISIONS (Pasture)
  DCU_PADDOCKESTIMATIONS (Pasture)
  dcu_pbi_cow (Cows information)
  dcu_pbi_lactation (Lactation records)

References

[1] Franck Michel. Integrating Heterogeneous Data Sources in the Web of Data. Databases [cs.DB]. Université Côte d'Azur, 2017. English. tel-01508602v2

[2] A. Petermann, M. Junghanns, R. Mu ller, and E. Rahm. BIIIG : Enbabling Business Intelligence with Integrated Instance Graphs. In Data Engineering Workshops (ICDEW), IEEE 30th Int. Conf. on, 2014.

[3] DuyHoa Ngo, Zohra Bellahsene. Overview of YAM++ - (not) Yet Another Matcher for ontology alignment task. Journal of Web Semantics (JWS), Volume 41, December 2016, Pages 30-49

An approach to generate diverse time series

Fouad Bahrpeyma, Mark Roantree, Andrew McCarren
The Insight Centre for Data Analytics, Dublin City University, School of Computing
Fouad.Bahrpeyma@insight-centre.org

1. Introduction

Time Series prediction has widespread applications in various domains such as finance, physics and agriculture. Practitioners use either traditional statistical approaches like ARIMA or machine learning approaches such as ANNs, RNNs and LSTM to implement prediction. Past studies have indicated that each time series is a unique problem and each prediction method has a unique set of abilities. In order to investigate such abilities, especially for new prediction methods, one should evaluate the algorithm against various types of time series. The available datasets mainly fail to provide sufficient diversity for providing an appropriate benchmark, in this regard. This study presents a new approach to generate time series of disparate characteristics to enable practitioners to evaluate prediction methods against a wide variety of time series.

2. Methodology

A time series is a sequence of values observed at equal intervals. To be able to create a time series, one should first identify time series constituent components. Generally, a time series is composed of four main components [1]: 1- *Trend*, which carries the magnitude of the signal, 2- *Seasonality*, a repeating pattern with a fixed length, 3- *Cyclicality*, (a set of) patterns repeating at variable intervals, and 4- the Irregular component which does not fall into the three mentioned categories (Note that each component is a time series) However, in this study, Trend and Cyclicality have been combined into a third level entity called the Trend-Cycle component in order to simplify their simulation.

Each time series component can be simulated in different ways which are described as follows:

- Trend-Cyclicality $\boldsymbol{T^C}$ was implemented in three forms: 1- a linear function, 2- a combination of several sinusoidal functions, and 3- a peace-wise linear function.
- Seasonality $\boldsymbol{S}$ was implemented via 1- a repeating sinusoidal function, 2- a smoothed impulsive function, 3- a frequently switching step function, and 4- a triangular function.
- The irregular component $\boldsymbol{I}$ was simulated using the fractional Gaussian noise (fGn), fractional Brownian motion (fBm), and multi-fractional Brownian motion (mBm). Brownian motion can be stationarized, which is a pre-requisite for prediction operations, via differencing steps [3].

In this study, time series components are incorporated using additive and multiplicative operations. With the three components ($\boldsymbol{T^C}$, $\boldsymbol{S}$ and $\boldsymbol{I}$) and two operations ("+" and "×"), there are 8 possible ways to generate time series, shown in Table 1.

Table 1: Ways to combine TS components

1. $y = T^C + S + I$	5. $y = T^C \times S + I$
2. $y = (T^C + S) \times I$	6. $y = T^C \times I + S$
3. $y = (T^C + I) \times S$	7. $y = S \times I + T^C$
4. $y = (S + I) \times T^C$	8. $y = T^C \times S \times I$

3. Evaluation

In this study, 5000 series were generated to be then evaluated in terms of disparity in this section. This study identifies several important time series characteristics, including: LRD, Hurst Exponent, Shannon Entropy, Dickey Fuller, Skewness, GOD, Fractal dimension and Fisher information [2].

Examining disparity can be realized via the histogram plot of time series characteristics. An example of the results (for Entropy) is shown in Figure 1.

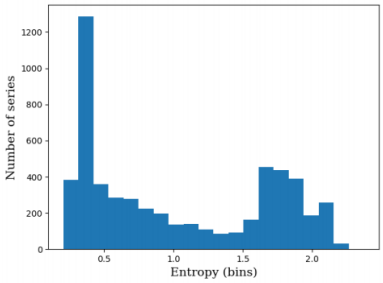


Figure 1: Results for Entropy

7. Conclusion

This study presented a method for generating time series of disparate characteristics, which is meaningfully important to evaluate prediction methods performances. We generated 5000 time series and evaluated them in terms of diversity for several characteristics. Further research might be undertaken to improve understanding of more aspects diversity in time series data.

8. References

[1] Brockwell, Peter J., and Richard A. Davis. Introduction to time series and forecasting. springer, 2016.

[2] Lütkepohl, Helmut, Markus Krätzig, and Peter CB Phillips, eds. Applied time series econometrics. Cambridge university press, 2004.

[3] Biagini, Francesca, et al. Stochastic calculus for fractional Brownian motion and applications. Springer Science & Business Media, 2008.

Time, Distance and Route Inference

Daniel A. Desmond
Insight Centre for Data Analytics, University College Cork
daniel.desmond@insight-centre.org

1. Introduction

Where we have historical trip data of New York taxi trips[1] consisting of the start and end points along with the distance travelled and time taken for the trip, we can use this information to create a more accurate map by inferring the delays due to traffic density and junctions. While more information is available, for example the taxi fare it is not being used as we wish to recreate the map using the minimum information unlike Zahn et al[2]

Our hypothesis is that in urban driving conditions a vehicle in isolation will travel at the posted speed limit, but due the effects of other vehicles and junctions the speed of travel will be less than the posted limit. Therefore the effects of other vehicles and junctions on the travel time can be simulated by setting the time to travel between junctions as a function of the posted limit while estimating the time to traverse a junction and the delay due to other vehicles and assign this time to the junctions.

2. Approach

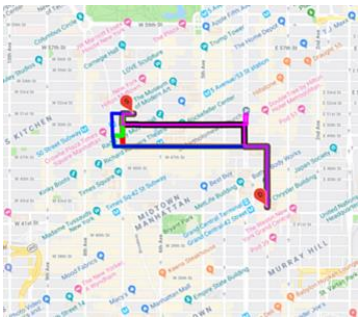


Figure 1: Set of possible routes between two points

The actual route taken from the start to the end of a trip is not known but a set of possible routes for the trip can be obtained as we know the recorded distance (d) of the trip. For each trip in our training set we create a set of feasible routes where the distance travelled (x) is

- $d \leq x \leq d + 0.11$.

The data consists of trips which began and ended on Manhattan Island in a one hour time slice.

Each possible trip is broken into two parts

- Travelling between junctions
- Waiting at and traversing junctions

The time to travel between junctions is set as the time to cover the distance at the posted speed limit.

When a vehicle arrives at a junction it may have a number of options depending on traffic restrictions.

These are

- To turn against the flow of traffic
- To turn with the flow of traffic
- To travel straight through the junction

These will be treated individually at each junction.

To infer the time required to traverse each junction we look to adjust the junction time iteratively based on the errors occurring to the possible routes and if the possible routes are feasible or not based on current junction estimates, and minimise the total absolute error on the possible routes that have the minimum absolute error for each trip.

A number of different initial values for each junction will be examined. These are

1. Setting each junction to zero
2. Average time to traverse the junctions for all feasible routes of each trip
3. Same as 2. But treating each junction type individually
4. Using the average traffic light timing and treating each junction type individually
5. Average time to traverse the junctions for the most feasible route of each trip and treating each junction type individually

3. Testing

The trips in the testing set were then processed using the calculated map searching for the quickest path. The inferred time was compared to the recorded values as shown in Table 1.

Initialization	Mean Trip Error (secs)	Mean Abs. Percentage Error (%)	RMSLE	Trips within 10% recorded time (%)	Coefficient of Determination (R^2)
Zeros	-38.49 (-483.10)	21.07 (64.91)	0.2832 (1.1624)	40.45 (0.46)	0.7605 (0.1580)
Average	-35.75 (-109.18)	21.07 (30.03)	0.2832 (0.4083)	40.82 (20.76)	0.7602 (0.6147)
Average with junction diff	-35.73 (-110.96)	21.07 (29.94)	0.2832 (0.4085)	40.81 (20.73)	0.7603 (0.6162)
Traffic lights with junction diff	-35.65 (-149.63)	21.08 (31.11)	0.2831 (0.4520)	40.70 (17.78)	0.7597 (0.6023)
Average closest with junction diff	-38.68 (-45.80)	21.06 (31.26)	0.2830 (0.3896)	40.73 (22.71)	0.7630 (0.6668)

Table 1: Errors between the recorded and calculated trip times

It should be noted that some of the trips take an excessive distance in relation to their start and end points. This can be accounted for in the training set but cannot be accounted for in testing.

Also that regardless of the initialization time for each junction the system in general will lead to similar results while individual results may be different

4. References

[1]http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml
[2] Xianyuan Zhan, Samiul Hasan, Satish V. Ukkusuri, Camille Kamga.: *Urban link travel time estimation using large-scale taxi data with partial information.* Transportation Research Part C 33 (2013) 37–49

HoliFab: Precise Flow Control using Photo Actuated Hydrogel Valves and PI Controlled LED Actuation for Microfluidic MEMS.

Andrew Donohoe
Insight Centre for Data Analytics, Dublin City University
andrew.donohoe@insight-centre.org

1. Abstract

Accurately controlling flow, using actuators that are fully integrated within a microfluidic device has long since been a challenge [1] with conventional methods requiring flow control, channel selection and sealing. Fluidic control is typically achieved using a series of external valves and pumps, increasing the overall scale of a microfluidic system. In recent years break throughs in material science have allowed for the use of photo responsive polymers imbedded within the fluidic device for fluid handling [2,3]. We present a system for control of flow within fluidic channels using photo responsive polymer valves. These valves can be polymerized in-situ. An LED platform and a PI algorithm has been used to achieved accurate flow control utilising these photo responsive polymers.

2. Experimental

The photo responsive valves were photo polymerised in situ using a monomeric cocktail containing 200 mg NIPAAm, 8.35 mg MBIS, 7.91 mg SPA-1, 7.42 mg PBPO and 6.05 µL Acrylic Acid (dissolved in 500 µL of the polymerisation solvent (2:1 v/v, THF:DI water) and an externally mounted LED.

The microfluidic chips that contain the photo responsive valve are fabricated using precision micro milling and thermal bonding [4]. The chip consists of a straight 1 mm wide channel and a valve chamber comprised of a 1mm diameter pillar centred within a circle of diameter 3mm as seen in figure 1.

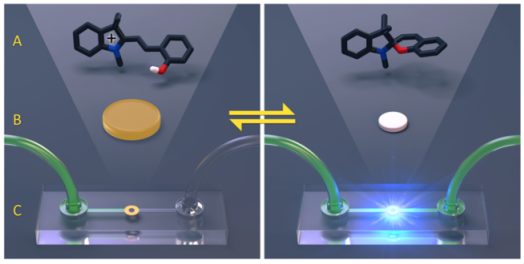


Figure 1: *Hydrogel valves within microfluidic fluidic channels under light irradiation (open) and in the fully swollen (closed) states.* [2]

Polymerisation and actuation was achieved using blue surface-mounted LEDs (Kingbright HB Blue 450nm, 600 mW/4.5 lm/1.3 cd, 3.5V) at 450 nm wavelength. The valves were polymerised around cylindrical pillars that provides an anchor point to prevent movement of the valve. A constant head of pressure of 3 mBar is maintained using two reservoirs of different volumes.

Precise programmable regulation of the relationship between flowrate and LED power has been achieved by using a flow meter to provide feedback to a PI controller. Through optimization of proportional (KP) and integral (KI) the system accurately delivers the programmed flow rates (figure 2) by detecting the disparity between the measured and set flow rates. The current design can also be used to maintain a constant flow rates for extended periods.

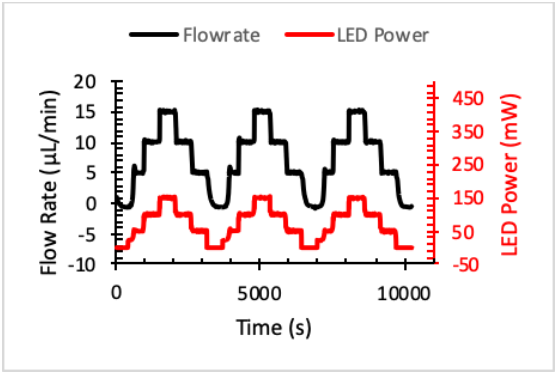


Figure 2: Flow rate control obtained using Photo responsive Polymer Valve and PI control. Cycles of (5.0, 10.0, 15.0, 10.0, 5.0 µL/min) shown in black (KP = 5 at 5 µL/min ,KP = 8.0 at 10.0 µL/min and 15.0 µL/min; KI = 0.1), overlaid with power supplied to the LED to achieve actuation in red.

3. Conclusion

We have demonstrated flow control through the incorporation of photo responsive materials within microfluidic channels with the use of a PI controller and an LED platform. Excellent stability at set flow rates of 5.0 µL/min, 10.0 µL/min and 15.0 µL/min is shown. The LED power data mirrors the flow rate with no indication of increasing demand over time.

8. References

[1] C. Delaney, P. McCluskey, S. Coleman, J. Whyte, N. Kent, D. Diamond, Precision control of flow rate in microfluidic channels using photoresponsive soft polymer actuators, Lab. Chip. 17 (2017) 2013–2021.

[2] J. ter Schiphorst, S. Coleman, J.E. Stumpel, A. Ben Azouz, D. Diamond, A.P.H.J. Schenning, Molecular Design of Light-Responsive Hydrogels, For in Situ Generation of Fast and Reversible Valves for Microfluidic Applications, Chem. Mater. 27 (2015) 5925–5931.

[3] S. Coleman, J. ter Schiphorst, A. Ben Azouz, S. Bakker, A.P.H.J. Schenning, D. Diamond, Tuning microfluidic flow by pulsed light oscillating spiropyran-based polymer hydrogel valves, Sens. Actuators B Chem. 245 (2017) 81–86.

[4] A. Donohoe, G. Lacour, D.J. Harrison, D. Diamond, M. McCaul, Fabrication of Rugged and Reliable Fluidic Chips for Autonomous Environmental Analyzers Using Combined Thermal and Pressure Bonding of Polymethyl Methacrylate Layers, ACS Omega. 4 (2019) 21131–21140.

Geospatial Analysis of Peatland Land use and Drainage in Ireland

[1,2]Wahaj Habib, [2]Kevin McGuiness & [1]John Connolly
[1]School of History and Geography, Dublin City University, [2]Insight Centre for Data Analytics, Dublin City University

1. Introduction

Peatlands are considered as an important ecosystem and there is increasing interest in their restoration. This is because Peatlands play an important role in land atmospheric exchange of C/GHGs (Carbon/Green House Gases), and a range of various other ecosystem services, such as climate regulation, water regulation, purification and treatment. Peatlands cover a small fraction (2-3 %) of the land and freshwater surface of the planet (Misney et al.,2019). Nevertheless, they account for approximately one third of global SOC (Soil Organic Carbon) stock. In Ireland (Fig. 1), peatlands cover up to 21% of the total land area and represent between 50-75% of the total SOC (Renou-Wilson et al.,2011). However, much of this area (~95%) has been degraded through anthropogenic activities for instance drainage for agriculture, forestry and peat extraction. Therefore, there is a need for the development of a system to identify management-related impacts on peatland function (ecology, hydrology and biogeochemistry). This will directly support rehabilitation and conservation activities and help to identify candidate sites for restoration. Moreover, it will also provide spatially explicit GHG emission factors associated with peatlands management. The main aim of this study is, to assess the impact of anthropogenic management of peatlands, using GIS (Geographic Information Systems), Earth Observation and Machine Learning (ML) techniques.

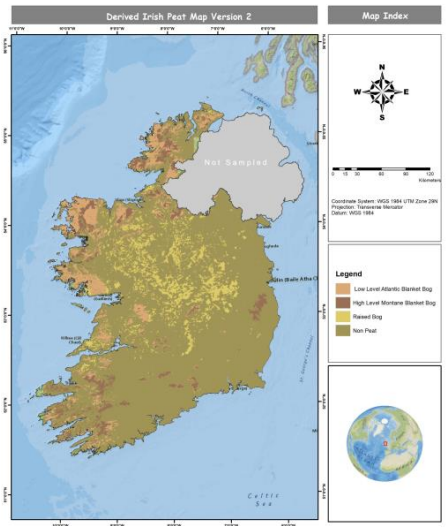


Figure 1: Derived Irish Peat Map version 2 (Connolly & Holden, 2009)

2. Main Objectives

The main objectives of the research are.

- Generate NDVI maps of the map using very high-resolution aerial photographs and high-resolution satellite imagery and correlate with data obtained from IoT sensors to estimate the GHG/C emission status of peatlands
- Development of National peatland LULC and drainage map of Ireland.

3. Methodology

Both high resolution satellite data and very high-resolution aerial photography will be used in this research. Peatland area will be defined by the Derived Irish Peat Map (DIPM2). Semi-automatic object-based image analysis, machine learning and automatic segmentation techniques will be used to map peatland drainage and habitat condition at national scale. Furthermore, a mapping approach from local to national scale will be implemented to generate NDVI (Normalized difference Vegetation Index) maps which will aid towards observing C/GHG emission status of peatlands. For this purpose, fine scale NDVI maps from in-situ sensors, and very high to high resolution NDVI maps from aerial photographs and Copernicus Sentinel-2 satellite imagery respectively, will be utilized.

3. Current status

The project is at initial startup phase. Currently literature review is being carried out with preliminary assessment of satellite images for the use of NDVI mapping.

Next steps include acquisition of very high-resolution aerial photos and initial assessment of test sites in the midland of Ireland.

5. Intended outcomes

Overall, the output (LULUF (Land Use Land Use Change Forestry), drainage and NDVI maps) generated from all these different datasets will be integrated together to determine the management impacts on peatlands function in Ireland.

6. References

Connolly, J. and Holden, N. M. (2009) Mapping peat soils in Ireland; updating the Derived Irish Peat Map. Irish Geography, 3, 343-352.

Minasny, Budiman, et al. (2009) Digital mapping of peatlands–A critical review. Earth Science Reviews, 196.

Renou-Wilson et al. (2011). BOGLAND Sustainable management of peatlands in Ireland. STRIVE report No 75 prepared for the Environmental Protection Agency (EPA), Johnstown Castle, Co Wexford, 157.

Predicting Mastitis Using Machine Learning Applied To Milk Flow Profiles

Changhong Jin
VistaMilk, Insight Centre for Data Analytics
changhong.jin@insight-centre.org

1. Introduction

Mastitis is the most costly disease in cows, leading to decreased milk yield and quality [2]. Mastitis can be characterised as clinical or sub-clinical. Clinical mastitis is defined as inflammation of a mammary gland with visible abnormalities in the milk which can be easily detected. Subclinical mastitis has no visible manifestations and is harder to detect.

The detection of mastitis in cows is essential to ensure that cows produce more milk and have a longer productive life. Furthermore, the earlier the detection of mastitis, the more it can limit the spread of the disease in the herd. Since the 1960s, the method of enumerating somatic cell count (SCC) has been widely used to detect mastitis in cows [2]. Somatic cells are mainly made up of white blood cells, or leukocytes, which increase in response to the bacteria that cause mastitis.

While the SCC of milk to be collected from farms is regularly measured many farms do not have access to cow level SCC data, or if they do, it will be sporadic in nature. It has also been shown that SCC values collected at one milking in isolation can be misleading when diagnosing cows with mastitis. However, modern milking machines are equipped with sensors that measure the rate of flow of milk from each cow at each milking (usually twice per day). The goal of our work is to build effective mastitis detection models by applying machine learning techniques to the readily available data source. This is challenging as it is likely that the relationship between milk flow data and mastitis is much weaker than the relationship between the SCC and mastitis.

2. Methodologies

Using milk flow data instead of somatic cell count to detect cows with mastitis can be seen as a time series classification problem. In this field, the k-nearest neighbors (kNN) technique coupled with dynamic time wraping (DTW) measure is widely accepted as the baseline [1], its working principle is shown in the Figure 1.

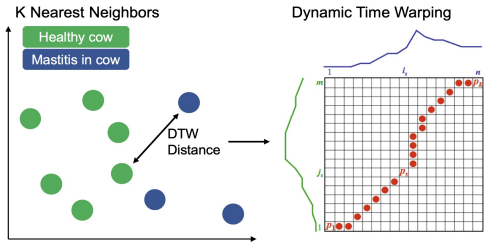


Figure 1: The principle of kNN with DTW

In addition, a feature-based classifier known as bag of patterns (BOP) [3] can also be used. Such method breaks down a time series into sub-sequences and represents them as discrete features, then builds a histogram of the pattern frequency for classification. Figure 2 illustrates details of this approach.

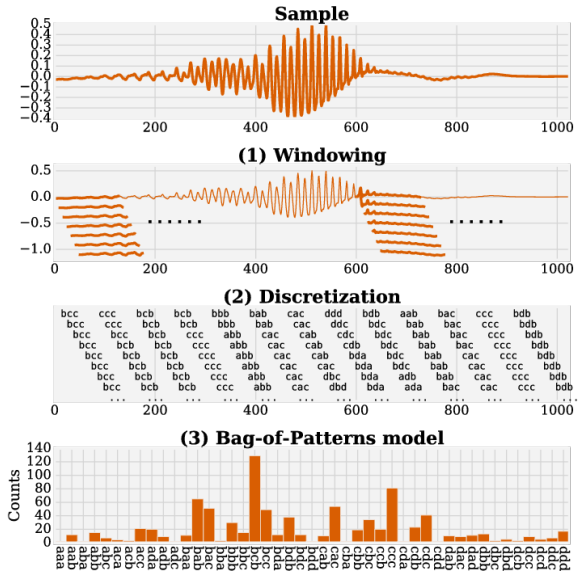


Figure 2: The details of BOP model (reproduced from [3])

Our work will explore the use of kNN technique with DTW measure and BOP model in the detection of mastitis based on milk flow data.

Acknowledgement

This work was performed in collaborate with Brian Mac Namee and John Upton in Teagasc, funded by Science Foundation Ireland under grant number 12/RC/2289_P2.

References

[1] A. Bagnall, J. Lines, A. Bostrom, J. Large, and E. Keogh. The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, 31(3):606–660, 2017.

[2] P. L. Ruegg. A 100-year review: Mastitis detection, management, and prevention. *Journal of dairy science*, 100(12):10381–10397, 2017.

[3] P. Schäfer and U. Leser. Fast and accurate time series classification with weasel. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 637–646. ACM, 2017.

On Privacy Comparison of SQL Queries

Muhammad Imran Khan
Insight Centre for Data Analytics, Department of Computer Science,
University College Cork, Cork, Ireland. imran.khan@insight-centre.org

1. Introduction

There exist scenarios where a comparison between two SQL queries is desired; for instance, in anomaly-based intrusion detection systems and log summarization for audit [1]. Query comparison techniques in literature measure how similar two SQL queries are based on data-centric and/or syntax-centric features. We are interested in comparing two SQL queries in terms of privacy whereby in terms of privacy we mean if one query is more (or less or equally) private than the other SQL query. For that reason, we measure the identification capability of SQL queries. The greater the identification capability of an SQL query the greater the risk it carries as compared to an SQL query with lesser identification capability. This provides the basis for a form of privacy-anomaly detection.

2. The Model

We use the query abstractions as depicted in Table 1 to compute the identification capability of an SQL queries. Given an SQL query Q_i, its abstraction is denoted as $A(Q_i)$, where the elements of $A(Q_i)$ are the attributes in the SQL query.

Query (Q_i)	SQL Query Abstraction A(Q_i)
`SELECT firstName, designation FROM companyRecord;`	{firstName, designation}

Table 1: SQL query abstraction deployed. The elements of $A(Q_i)$ are the attributes in the SQL query.

2.1 Discrimination Rate (DR)

The comparison is achieved by adopting a recently proposed privacy metric known as Discrimination Rate (DR) privacy metric (and Combined Discrimination Rate - CDR) [2] that measures the efficiency of an anonymity system based on information theory. The DR and CDR are given by $DR_X(Y) = 1 - \frac{H(X|Y)}{H(X)}$ and $CDR_X(Y_1, Y_2, \ldots Y_n) = 1 - \frac{H(X|Y_1,Y_2,\ldots Y_n)}{H(X)}$. Where X and Y, where X is the set of outcomes and Y is the attribute for which the measurement of the identification capacity is desired. $H()$ is represents the entropy.

2.2 Computing Identification Capability

Let a relation be denoted as $\mathcal{T}$ and let the attributes in $\mathcal{T}$ are denoted as $\{atr_1, atr_2, atr_3, \ldots, atr_n\}$. Given a set of SQL queries $\{Q_1, Q_2, Q_3, Q_4, \ldots, Q_m\}$ executed on relation $\mathcal{T}$. Let L be an audit log consisting of SQL queries executed on relation $\mathcal{T}$ that is $\{Q_1, Q_2, Q_3, Q_4, \ldots, Q_m\} \in L$, The abstraction of L is represented by $A(L)$ and the abstraction of an individual SQL query $Q_i \in L$ is represented as $A(Q_i)$. The identification capability of an SQL query Q_i is denoted as $DRSQL(A(Q_i))$ = $CDR_X(atr_1, atr_2, \ldots, atr_k)$ where $atr_1, atr_2, \ldots, atr_k \in A(Q_i)$. In a case where only a single attribute atr is queried, then the identification capability of that query Q_i is $DRSQL(A(Q_i))$= $CDR_X(atr)$ = $DR_X(atr)$.

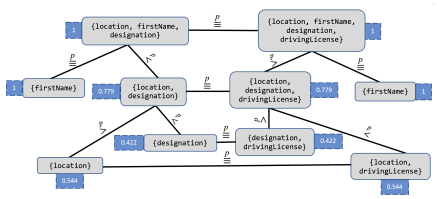


Figure 1: A fragment of privacy-aware attribute relationship diagram. The more-private and privacy equivalence relations are represented by $\overset{p}{>}$ and $\overset{p}{\equiv}$ respectively.

2.3 Comparing SQL Queries

To perform a privacy comparison between SQL queries, we define the following relations: *privacy-equivalence* relation, *less-private* relation and *more-private* relation. We denote the privacy equivalence relation as $\overset{p}{\equiv}$, and we generalize privacy equivalence relation for SQL query abstractions. We define a privacy equivalence relation on a set of SQL query abstractions $A(L)$ as a relation on $A(L)$. Given two SQL queries Q_i and Q_j and their abstractions $A(Q_i)$ and $A(Q_j)$, then the privacy equivalence between $A(Q_i)$ and $A(Q_j)$ is defined as $A(Q_i)$ being a subset of $A(Q_j)$ and for every $x \in z$ has a discrimination rate value of 0, where z = $A(Q_j) \cap A(Q_i)$ or an attribute or a set of attribute $\in A(Q_j) \cup A(Q_i)$ has $DR_X(atr)$ or $CDR_X(atr_1, \ldots, atr_k)$ = 1. Given two SQL queries Q_i and Q_j and their abstractions as $A(Q_i)$ and $A(Q_j)$, then the more-private relation is defined as $A(Q_i)$ being a subset of $A(Q_j)$ and the $DRSQL(A(Q_i))$ is less than the $DRSQL(A(Q_j))$ then we can say that Q_i is more-private than Q_j that is $A(Q_i) \overset{p}{>} A(Q_j)$ or Q_j is less-private than Q_i. In order to assist the privacy comparison and for a better explanation to compute the identification capability of the SQL queries, based on the discrimination rate and combined discrimination rate, one can then articulate a privacy-aware attribute relationship diagram for a relation $\mathcal{T}$. A fragment of the privacy-aware attribute relationship diagram is shown the Figure 1.

References

[1] G. Kul, D. T. Luong, T. Xie, V. Chandola, O. Kennedy, and S. Upadhyaya. Similarity measures for sql query clustering. *IEEE Transactions on Knowledge and Data Engineering*, Jul 2018.

[2] L. P. Sondeck, M. Laurent, and V. Frey. Discrimination rate: an attribute-centric metric to measure privacy. *Annals of Telecommunications*, 72(11):755–766, Dec 2017.

The Application of CNNs for Sustainable Agricultural Practices Classification

Agustin Garcia Pereira
Insight Centre for Data Analytics, NUI Galway
agustin.garciapereira@insight-centre.org

1. Introduction

The need for spatial information about agricultural practices is expected to grow rapidly due to environmental, agronomic, and economic reasons [1]. Modern agriculture is considered a significant source of environmental pollution and resources depletion. The simplification and intensification of agricultural systems where the same crop is grown time after time at the same place, incrementing the need for external chemical inputs, is threatening the worldwide sustainability of crop production [2]. This creates the need for promoting more sustainable agricultural practices, and thus, the need for tracing them. Remote sensing has shown to be an effective technology to monitor land use dynamics, and many algorithms have been developed and used to automate the land use classification task [3]. In the last time, the increasing availability of Earth Observation data, the growing collection of geospatial databases, and the advances in the field of Artificial Intelligence (AI) and computing power presents new opportunities to address such global problems related to sustainability and climate change. The intersection of geospatial science and AI is starting to be referred in the literature and in the industry as GeoAI. Even though a plethora of satellite images describe the same place on earth every day, very few AI solutions have harnessed the temporal dimension of remote sensed images for classifying land use dynamics. Moreover, only 9% of the total remote sensing and agriculture publications focus on cropping practices [3]. In this sense, there is a need for applying and assessing novel deep learning algorithms and its applicability in the domain of remote sensing to help address the important problems described above.

2. Research goals and related work

Our research is focused on 1) Assessing the performance of Convolutional Neural Networks (CNNs) on multivariate remotely sensed time series to classify land use and agricultural practices and compare it with state-of-the-art algorithms; 2) Experiment how different pre-processing tasks affect model's performance; 3) Evaluate classification accuracy at different early time stamps of the time-series; 4) Evaluate how spatial variabilities on the training/testing sets affect model's accuracy. Comparing our work with other related studies, ours makes use of publicly available satellite imagery, making a transfer learning approach that would make the process of fitting the models for other geographical locations, viable. We also utilize convolution layers to learn temporal patterns from land-use dynamics and we plan to compare the results with other state of the art algorithms, such as Long Short-Term Memory (LSTMs). While some studies only focused on the classification of a few agricultural types, we have trained a single model that is able to classify 20 agricultural classes with 89% accuracy. None of the studies analyzed have classified pure temporal dynamics, as we are doing with agricultural practices related to sustainability.

3. Milestones achieved

First, we conducted an extensive literature review about GeoAI applications following the Aristotle's Four Causes-derived framework. In order to advance our research, we listed and analyzed high quality ground truth data sources related with agriculture and openly available satellite imagery sources. Then, we built a pipeline to create high quality labeled time-series datasets that can be used to train CNNs. The pipeline consumes Landsat 7 and Landsat 8 satellite images, as well as a shapefile containing the ground truth information, and creates a tabulated file containing labeled, temporary sampled, missing data-free time series at the pixel level [4]. Using this pipeline, we have evaluated the use of CNNs with convolutions in the temporal dimension in a set of two distinct experiments, where we classified 20 agricultural land use classes and two agricultural practices related to sustainability, respectively. Our experiments showed promising results with 89% and 88% accuracy, respectively [5]. All the experiments are conducted using the Azure infrastructure grant obtained with the AI for Earth Microsoft's grants program.

4. Future work

Future work includes comparing the performance of CNNs with LSTMs. Evaluating the effect of averaging time series at the polygon level before training the models. Evaluate the performance of the models at early stages of the growing season (for land use classification). Evaluate the performance of the models using different spatial setting for the training/testing datasets.

5. References

[1] J. De Baerdemaeker, *Precision agriculture technology and robotics for good agricultural practices*, vol. 1, no. PART 1. IFAC, 2013.
[2] K. J. Noone and C. Folke, "A safe operating space for humanity," no. May 2014, 2013.
[3] A. Bégué *et al.*, "Agricultural Systems Studies using Remote Sensing," 2019.
[4] A. García Pereira, A. Ojo, E. Curry, and L. Porwol, "Data Acquisition and Processing for GeoAI Models to Support Sustainable Agricultural Practices," in *Proceedings of the 53rd Hawaii International Conference on System Sciences 2020 (HICSS 2020)*, 2020, pp. 922–931.
[5] A. G. Pereira, L. Porwol, A. Ojo, and E. Curry, "Towards a Temporal Deep Learning Model to Support Sustainable Agricultural Practices," *27th AIAI Irish Conf. Artif. Intell. Cogn. Sci.*, pp. 1–12, 2019.

Milk Supply Forecasting

Eoin Delaney, Derek Greene, Mark Keane
Insight University College Dublin, VistaMilk
Eoin.delaney@insight-centre.org

1. Introduction

The agri-food industry is one of the biggest drivers in the Irish economy, accounting for 10% of employment and 9.3% of exports [1]. The application of machine learning to make future predictions has exciting potential to optimize agricultural processes and reduce waste on farms. Our goal is to accurately predict the quantity of milk collected from a specified dairy herd at an industrial scale for short, medium and long term horizons by using features that are readily available at commercial dairy farms.

2. Previous Research

Previous forecasts on Irish farms have focused on data from individual cows in a research herd where the data collection process is consistent and reliable. Ideal conditions are maintained where stocking rates and calving seasons are kept constant and cows are free from disease [2]. We question whether these conditions are reflective of what happens on everyday farms. The most predictive features were found to be the number of cows milked (NCM) and the days in milk (DIM) [2]. The NARX model (non-linear autoregressive with exogenous inputs) performed best in 63.8% of prediction ranges, most notably in predicting peak milk production which is important for industry in determining processing capacity. In the model comparison study it was noted that different models outperform others at different stages of the year. The study mentions a need for models which can be applied at an industrial scale instead of at the individual cow level.

3. Methods

Five years of data from 2480 herds across fourteen counties was provided by an industry partner. Information regarding the milk collection quantity, date and time of collection, herd size and calving number are examined. From these values the cumulative calving number and the average herd DIM were determined. This data should be readily available at commercial dairy farms. As much of this data was collected manually, data preprocessing is an integral step of our analysis. One of the interesting aspects of our data is that the measurements are not taken at regular time intervals which is a property that many classical time series models such as ARIMA assume. As a consequence of this we also examine the cumulative supply quantity as a target variable.

3.1. Prophet Forecasting

Prophet is a generative additive forecasting model that was developed by Facebook and has three main components: trend, seasonality and holidays [3]. Some of the main benefits of this method are that the measurements do not need to be at regular time intervals and fitting is extremely fast. Using the last eight months of data from a herd as testing data for a univariate forecast we obtained a benchmark MAPE of 32.40%. Considering additional regressors such as cumulative calving number and average herd DIM reduced the MAPE to 11.53% (Fig 1). No hyperparameter tuning was implemented.

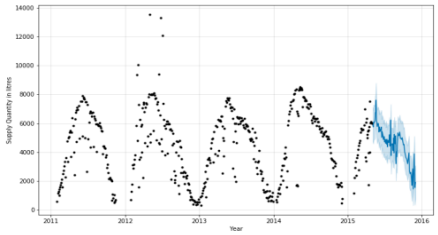


Figure 1: Prophet Forecast (Blue)

4. Future Work

We intend on extending our forecasting results by implementing different model set-ups. K nearest neighbors (k-NN) is one of the most intuitive and explainable models and has successfully been applied to predict grass growth in Irish farms. Recurrent neural networks are competitive forecasting models which have excelled in terms of prediction accuracy in forecasting competitions [4]. Incorporating weather and grass growth data into the models and examining ensembles may also help generate more accurate forecasts.

5. References

[1] Teagasc, *Agriculture in Ireland*. Accessed: 02/01/2020. [Online]. Available: https://www.teagasc.ie/rural-economy/rural-economy/agri-food-business/agriculture-in-ireland/

[2] Murphy, M.D., O'Mahony, M.J., Shalloo, L., French, P. & Upton, J., "Comparison of modelling techniques for milk-production forecasting'', *Journal of Dairy Science*, vol. 97, no. 6, pp. 3352-3363, 2014.

[3] S. J. Taylor, M. Park, U. States, B. Letham, M. Park, and U. States, "Forecasting at scale", *PeerJ Preprints*, pp. 1-25, 2017.

[4] H. Hewamalage, C. Bergmeir and K. Bandara, "Recurrent Neural Networks for Time Series Forecasting: Current Status and Future Directions", arXiv:1909.00590, 2019.

Debiased Offline Evaluation of Active Learning in Recommender Systems

Diego Carraro
Insight Centre for Data Analytics, University College Cork
diego.carraro@insight-centre.org

1. Introduction

Active Learning (AL) when applied to Recommender Systems (RSs) aims at proactively acquiring additional ratings data from the RS users in order to improve subsequent recommendation quality. AL strategies are typically evaluated offline first, but the classic AL offline evaluation methodology does not take into account the *bias problem* in RS offline evaluation. Such problem arises from the use of a *biased dataset* to perform the evaluation, which is biased due to many factors, known as confounders. For example, it has been recognized that features of the user interface play an important role: differences in the ways items are exposed to users (e.g. position on the screen) influence the likelihood of a user interacting with those items. The actions of the recommender itself set up a feedback loop: users are typically more likely to interact with the recommender's suggestions than with other items. We therefore argue that the bias problem also affects the evaluation of AL strategies as well as the RS one. For this reason, in this extended abstract, we propose a new AL offline evaluation methodology for RSs which mitigates the bias and thus facilitates a truer picture of the performances of the AL strategies under evaluation.

2 Our Approach to Unbiased AL Evaluation

In offline experiments, AL strategies are evaluated by randomly splitting the biased dataset into three sets: the known ratings (K), the hidden ratings (H) and the test ratings (T). Known ratings are the ones from which the initial recommender model is built, i.e. the ones we assume that the RS has at hand. The hidden ratings are the ones which the simulated users might reveal to the system if prompted to do so by the AL strategy. Subsequently, these elicited ratings can be added to the known ratings, and a new recommender model can be built. Test ratings are used to measure the performance of the RS both before applying the AL strategy and afterwards. Because the split is random, all three parts are biased. Thus, to study the impact of bias in the evaluation, in our work we propose to *debias* some of those sets, to resemble an unbiased evaluation. To debias, we use WTD [1], a weighted sampling approach which enjoys low overheads and high generality. WTD *intervenes* on biased data by sampling a subset of such data. The produced intervened sample is supposed to be less biased.

We design the following three evaluation methods to assess an AL strategy's performance in our experiments.

- *CLASSIC*: this method corresponds to the classic way of evaluating an AL strategy, where there is no attempt to mitigate the bias in the dataset, i.e. K, H, T are all biased.

	CLASSIC	*INT_T*	*INT_HT*
RND	+0.22%	+0.97%	-0.50%
POP	+9.67%	+3.01%	-0.27%
HP	+9.45%	+11.83%	+10.25%

Table 1: Results on ML dataset.

- *INT_T*: if we want an unbiased evaluation for an AL strategy, then we must, at the least, debias the test set. Therefore, this method debiases T and leaves K and H.
- *INT_HT*: in this method, both the test and the hidden sets are debiased. In fact, we argue that H has a big impact on the final performance of an AL strategy and we use this method to prove it. We leave K biased.

3. Experiments and Results

In the experiments that we report here on the MovieLens 1M dataset (ML), our goal is not to find the best AL strategy. Rather, our goal is simply to show that debiasing can affect the results, even to the extent of changing which strategy is the best one. We compare three well-known and easy-to-implement AL strategies from the literature. *RND* asks the users to rate random items. POP asks the users to rate popular items in the dataset and HP asks the users to rate items which the recommender thinks they will like.

Table 1 reports test results. For each AL strategy and for each evaluation method, we measure Recall@10 of recommendations provided to users, both before and after the AL step. The table shows the percentage change in the Recall@10 achieved after the AL step so that we can observe the impact of each AL strategy.

The results shows that the HP strategy is a good strategy according to all three methods. Also, while the performance of RND is poor for all three methods, the performance of the POP strategy is overestimated by *CLASSIC*. In fact, POP is no longer the 'same' as HP for *INT_T*, because it drops to second place in the ranking. Finally, according to *INT_HT*, POP drops even more and its performance is now similar to RND's.

In the light of our findings, this suggests the need to reconsider results presented in the literature and henceforth to use instead a debiased evaluation method. In our future work, we plan to use our debiased evaluation methods to assess the effectiveness of more AL strategies.

References

[1] D. Carraro and D. Bridge. Debiased offline evaluation of recommender systems: A weighted-sampling approach. In *Procs. of REVEAL 2019, 13th ACM Conference on Recommender Systems*, 2019.

Towards Sharing Recommender System Task Environments

Andrea Barraza-Urbina, Mathieu d'Aquin
Data Science Institute, NUI Galway
andrea.barraza@insight-centre.org, mathieu.daquin@insight-centre.org

1 Introduction

Recommender Systems (RS) help users discover interesting products by means of suggestions. Conventionally, the RS problem has been formalized as a Supervised Learning task (Batch Learning). However, recent works, as explained in [1], have proposed that Reinforcement Learning (RL) [4] can be a more appropriate paradigm (Online and Incremental Learning) to frame the "modern" and Interactive RS problem (view Figure 1). Under this paradigm, the RS is viewed as a sequential decision-making Agent focused on learning over time how to perform *actions* (e.g., offer item suggestions), in different *states* (e.g., to different users), using interactive feedback in the form of *reward* signals (e.g., user ratings).

Currently, the RS community has mostly focused on creating and sharing datasets and simulations to share experiments. We argue in [2] that this is not enough, and that Task Environments (as in RL) can encapsulate more assumptions and design decisions necessarily incurred in RS evaluation design.

2 RS ENVIRONMENTS IN A NUTSHELL

A *RS Task Environment* (or *RS Environment*) is the domain or application scenario where the RS Agent will be deployed, and embodies the most relevant characteristics of the considered problem setting. In its minimal form, the Environment is a function that for any action a_t performed by the Agent over the current Environment state s_t, will generate a reward r_{t+1} and the next Environment state s_{t+1}, i.e.: $(s_t, a_t) \mapsto (r_{t+1}, s_{t+1})$. Although we can describe Environments using Markov Decision Process (MDP) [4] constructs, in [2], we propose an alternative general framework of components that can help define a principled way to build Environments in the RS field. Figure 1, presents the main components a RS Task Environment should define, which are:

Simulator. Imitates the dynamics of the RS application domain (e.g., dynamics of users and items). Note that the Simulator only defines world dynamics and not the Agent's task/goals. To build a Simulator, it is common to use *Observed Data* (sampled from the real application) augmented with *Design Assumptions* about the world being modeled.

Reward Function. Uses the Simulator's output x_{t+1} to generate a bounded numeric reward r_{t+1}. Rewards define the RS goals and are used to motivate the RS Agent to learn.

State Feature Representation. This component takes as input the Simulator's output x_{t+1} to generate state s_{t+1}. A state encapsulates all the relevant information about the Environment made available to the Agent at a given time step.

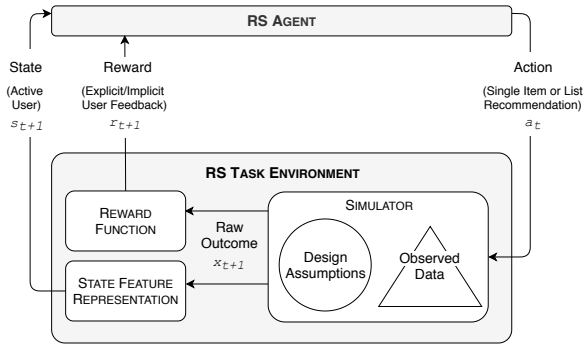


Figure 1: The Interactive RS problem as RL. More on [2, 1].

3 Conclusion

Describing Interactive RS evaluations in a way which is *complete* and *consistent* to be *efficiently reproduced* and *compared*, can quickly turn into an overwhelming task. In many ways, RS evaluation design is more an art than a science. *RS Task Environments* can be a way to encapsulate and share many of the complexities and design assumptions necessarily made when doing RS evaluations. Recent works such as [1, 3] are steps in this direction. We have presented a general and modular framework that identifies the main components of a RS Environment (more in [2]). There are multiple paths for *future research*, such as: (a) *A Framework to share RS Environments*: BEARS [1] presents a possible solution, (b) *Defining types of RS Environment:* Individual components can be built and combined to create multiple types of RS Environments, (c) *Assessing RS Environment Quality*: We cannot build perfect Environments, thus it is important to analyze Environments to understand in which ways they are imperfect and which problem settings they represent best. Overall, we hope that the introduced model helps structure a discussion around the meaning of Task Environments for RS, their role in evaluation design and encourages new research in this area.

Acknowledgement. This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2, cofunded by the European Regional Development Fund.

References

[1] Barraza-Urbina and et al. Bears: Towards an evaluation framework for bandit-based interactive recommender systems. In *REVEAL'18. Workshop of ACM RecSys*, 2018.

[2] A. Barraza-Urbina and M. d'Aquin. Towards sharing task environments to support reproducible evaluations of interactive recommender systems. *arXiv preprint arXiv:1909.06133*, 2019.

[3] D. Rohde and et al. Recogym: A reinforcement learning environment for the problem of product recommendation in online advertising. *arXiv preprint arXiv:1808.00720*, 2018.

[4] R. Sutton and B. A. *Reinforcement learning: An introduction*. MIT press, 2018.

Recommendation of Role Models for Personal Development Planning

Bianca Pereira
National University of Ireland, Galway
bianca.pereira@insight-centre.org

1. Introduction

Lifelong learning has challenges that differ from formal education. In formal education, a professional educator develops a course programme with learning goals to be achieved, as well as a series of activities in which learners need to engage to achieve these goals. On the other hand, in personal development planning, learners need to choose their own learning goals and activities. This process is call 'personal development planning'.

Personal development planning can benefit from observation of other people's learning journey as: (i) inspiration for the uptake of new goals, (ii) demonstration that certain goals are achievable, and (iii) demonstration on a possible path to achieve a given goal. In this PhD work, we aim at developing a system that recommends people (a.k.a. role models) who can serve as examples to support learners in developing their personal development plans.

2. Role Models

Role models are mental images developed by each person to represent their desired future selves. According to Gibson [3], role models are defined as "(...) active, cognitive constructions devised by individuals to construct their ideal, or 'possible' selves based on their own developing needs and goals".

When such cognitive constructs are, at least partially, represented by a person, that person is also called a 'role model'. According to the Motivational Theory of Role Models (MTRM) [4], role models can support the process of personal development planning by influencing the learner's achievements, motivations and goals through inspiration, demonstration of what is possible, and by acting as behavioural models.

3. Related Work

Role models are identified by four categories of attributes: learning goals, learning paths, social identities, and competency levels.

Role models are types of peers and the recommendation of peers is not a new task. However, in our preliminary literature review, there is no peer recommendation work that focuses on the four categories of attributes simultaneously. In Expert Search, the focus is only on peers who have a maximum level of competency on a given topic of interest (see [2] for an overview). Similarly, in the area of Technology-Enhanced Learning (TEL), [1] propose methods to recommend 'knowledgeable people' on a topic related a learner's current goal. Also in the TEL domain, [6] recommends peers based on shared social identities and similar learning paths; however, such recommendation is based on shared career milestones rather than learning goals.

4. Research Questions and Methodology

This research uses a Design Science Research methodology [5] focused on the following questions:

RQ1: What personal characteristics make someone a suitable role model?

RQ2: What information about a role model's learning journey support the design of personal development plans?

RQ3: What is the most suitable architecture for a role model recommendation system?

5. Approach and Ongoing Work

The proposed system is divided into three components: (i) goal modelling, where learner's goals are described; (ii) user profiling, where both learners and role models are described based on their learning goals, learning paths, social identities, and competency levels; and (iii) peer recommendation, composed of the recommendation engine responsible to recommend role models.

In this research, we have already conducted a preliminary literature review and initial user study to explore RQ1 and RQ2, with results to be submitted for publication in 2020.

6. Acknowledgements

I would like to acknowledge the support and guidance of Prof. Dr. Mathieu d'Aquin and Dr. Michael Hogan.

References

[1] G. Beham, B. Kump, T. Ley, and S. Lindstaedt. Recommending knowledgeable people in a work-integrated learning system. *Procedia Computer Science*, 1(2):2783–2792, 2010.

[2] G. Bordea. *Domain adaptive extraction of topical hierarchies for Expertise Mining*. PhD thesis, 2013.

[3] D. E. Gibson. Role models in career development: New directions for theory and research. *Journal of vocational behavior*, 65(1):134–156, 2004.

[4] T. Morgenroth, M. K. Ryan, and K. Peters. The motivational theory of role modeling: How role models influence role aspirants' goals. *Review of general psychology*, 19(4):465, 2015.

[5] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee. A design science research methodology for information systems research. *Journal of management information systems*, 24(3):45–77, 2007.

[6] N. Van Labeke, G. D. Magoulas, and A. Poulovassilis. Searching for "people like me" in a lifelong learning system. In *European Conference on Technology Enhanced Learning*, pages 106–111. Springer, 2009.

Pace my race: recommendations for marathon running

Jakim Berndsen
Insight Centre for Data Analytics @ University College Dublin
jakim.berndsen@insight-centre.org

1. Introduction

Due to the increasing ubiquity of smart devices amongst distance runners there now exist large quantities of data about the training programmes and race performances of athletes of varying ability levels. Recent work has involved the use of recommender systems in the domain of distance running to help runners improve their performance. Recommendations can be made to inform a runner's training [1], or to help them create a race plan of how they will divide their efforts of the course of the race [2]. While these recommender systems are undoubtedly useful their scope remains limited; both types of recommendation are done pre-race and cannot be adapted should race circumstances change. Recreational runners are well known for running too quickly in the opening kilometres of a marathon rendering even the best race strategy useless. This motivates the work presented in this abstract: devise an in-race system that can recommend pacing adjustments, if deemed necessary, to guide a runner safely to the end of a race without *hitting the wall*.

2. Recommendations

Our approach to making recommendations is two-fold. Firstly, we must identify runners that are at risk of hitting the wall later on in the marathon, and secondly, present a pacing adjustment that will see them finish the race with a minimal slowdown.

To accomplish the first of these tasks we generate features for each of the runners in our dataset ($n = 7931$). For each of these runners we have their pace (min/km), heart rate (bpm), and cadence (steps/min). We sample the average of these three features over 1km, 5km and race to point windows, at every 500 metre point of the marathon. We then train a XGBoost model for each segment to determine the *split* (second half time/first half time) of that particular marathon. Large positive splits are strongly associated with hitting the wall, and thus we determine that any runner with a large positive split is at risk of a detrimental slowdown and will present them with a recommendation.

To generate recommendations we leverage the fact that some experienced runners will make pacing adaptations when at risk of slowing down. These runners who about significant slowdown can be used as our exemplars. If our model predicts a runner is at risk of slowing down significantly we will suggest a pacing plan based on the strategies of similar, at-risk runners that managed to finish the race without a detrimental slowdown.

The recommended pacing plan for runner X at risk of slow down is generated with the following steps:

1. Find the most similar runners to X that do not slow down.
2. Calculate the average pacing profile these similar runners.
3. Normalise this average pacing profile. $\frac{P_i}{Mean(P)}$
4. Make a further finish time prediction based model trained on runners that do not slow down.
5. Calculate the required average pace to finish in that time.
6. Multiply the average pace over the normalised pacing profile to return a personalised pacing profile to the runner.

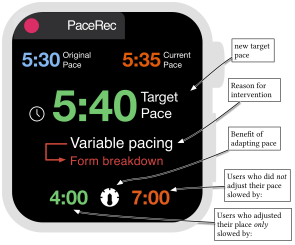


Figure 1: Example of recommendation on a smartwatch.

3. Explainability

As runners may not yet feel tired at the point a recommendation is made they may choose to simply ignore any advised pacing adaptations. To increase adherence to our adapted pacing plans we must therefore explain our recommendations. We use three methods to ensure runners will follow our new pacing strategy.

Firstly, we make only recommendations at key points of the marathons. These can be distance milestones, significantly landmarks, aid stations or any user defined interval. These milestones tend to be places that runners are making decisions about their future pacing and are thus more likely to regard our pacing adaptations. Additionally it ensures we make a limited number of recommendations rather than constantly adapting pace which may annoy the athlete.

Secondly, we give the reason a recommendation is being made. The features used to train the model correspond closely to terms runners use to describe their own performance, and thus the reason the decision is being made can be relaid to the runner in terms they understand. For example, a runner may be exhibiting variable cadence which could be indicative of a future form breakdown.

Lastly, we demonstrate the benefit of the recommendation. The time loss of following the recommendation is presented against the slowdown of similar runners that did not slow down. This demonstrates the net benefit of following the recommendation, and that the counter intuitive move of a slight slowdown early in the race can reap gains later on. Through these well timed explanations, that show both the reason and benefit behind the recommendations, users should trust and adhere to the pace adaptations presented resulting in more runners finishing the marathon enjoyably and safely.

References

[1] S. Mohan, A. Venkatakrishnan, M. Silva, and P. Pirolli. On Designing a Social Coach to Promote Regular Aerobic Exercise. *Proceedings of the 29th Conference on Innovative Applications of Artificial Intelligence (to appear)*, (February):4721–4727, 2017.

[2] B. Smyth and P. Cunningham. 'Running with Cases: A CBR Approach to Running Your Best Marathon'. *Proceedings of ICCBR 2017, Trondheim, Norway, June 2017*, 2017.

Person-Independent Multimodal Emotion Detection for Children with High-Functioning Autism

Annanda Sousa
Data Science Institute - NUI Galway
annanda.sousa@insight-centre.org

1. Background and Motivation

Automatic Emotion Detection (ED) aims to automatically identify people's cognitive states or emotions, e.g. happiness, anger, fear using different types of media inputs such as texts, video, audio and sensor signals. When combining more than one type of data, they are called *Multimodal* Emotion Detection systems and usually outperform unimodal systems.

Automatic ED is advancing to become an important component of Human-Computer Interaction (HCI) through affect-sensitive systems. An affect-sensitive system detects the user's emotions and automatically adapts its interaction with the human based on their emotions. Even with all the advancement of ED for users with typical neurological development, usually referred to as neurotypical, when applying those systems to children with autism they do not perform well, mainly because of this particular population's way to express emotions [2], motivating the need to develop ED systems specifically tailored to autistic children. Autism Spectrum Disorder (ASD) is a developmental disorder with spectrum manifestation of traits, characterised by impairments in social interaction, communication and repetitive patterns of behaviour and interests. High-Functioning Autism (HFASD) is defined as ASD without significant cognitive and language impairments [1].

Nowadays, the development of computer-based interventions tools for the treatment of children with autism has increased, and studies have shown evidence demonstrating the effectiveness of such tools to support ASD. Regardless, most technological tools that have been developed to support children on the autism spectrum do not use automatic ED which could be of great relevance to turn them into significant supplementary support to classic interventions that are usually expensive and very much dependent on human presence.

2. Related Work and Research Objective

Previous studies have developed ED systems tailored to children with autism [2]. Together these studies provide important evidences to show that it is viable to model and automatically identify emotions of children with ASD. However, such studies remain limited when considering two points: input multimodalities and generalisability of the model. To the best of our knowledge, none of their models used multimodal input data for emotion identification and most of the works created models that are individual-specific.

During this project, we aim to answer the following Research Question:

RQ1: How to create a multimodal Emotion Detection system which:
i) Is tailored to children with high-functioning autism;
ii) Is person-independent;
iii) Keeps the balance between performance and usability.

To be able to answer **RQ1**, we further need to explore answers to the following research question:

RQ2: How to build a ground truth dataset annotated with the emotional states we aim to identify?

Considering the problem stated above, the proposed model will be a multimodal ED system tailored to children with high-functioning autism. It will involve four input modalities: video, audio, text and physiological signals (i.e. heart rate measure). Based on those input, our model will use features extracted from: facial expressions, body movements, the words content of the speech, the tone of the voice and the heart rate values. All of them are broadly used in the ED field.

3. Research Approach and Methods

Our first challenge to address is to obtain the ground truth dataset. To achieve this, we are finishing the preparations to conduct a study with human participants to elicit, capture and tag spontaneous emotional zones expressions from children with high-functioning autism (**RQ2**). The subsequent challenge is to design and develop the Multimodal Emotion Detection system. To accomplish that, we will use the standard machine learning methodology for multimodal ED, which first includes the extraction of features from the dataset. Then, we will define/create the information fusion layer, the machine learning model, and to evaluate its accuracy. Finally, we will design computer experiments to analyse the relation between input modality data/features/data fusion and the accuracy of the Multimodal Emotion Detection model. The objective is to measure the impact of a given selection in the accuracy of the Multimodal ED model.

References

[1] V. L. Gaus. Cognitive behavioural therapy for adults with autism spectrum disorder. *Advances in Mental Health and Intellectual Disabilities*, 5(5):15–25, 2011.

[2] C. Liu, K. Conn, N. Sarkar, and W. Stone. Physiology-based affect recognition for computer-assisted intervention of children with Autism Spectrum Disorder. *International Journal of Human Computer Studies*, 2008.

Functional knee assessment via multi-sensor approach

Liudmila Khokhlova*, Dimitrios-Sokratis Komaris, Salvatore Tedesco, Brendan O'Flynn
Insight Centre for Data Analytics, Tyndall National Institute, University College Cork
liudmila.khokhlova@tyndall.ie

1. Introduction

The ability to move freely and without pain is an important quality of life's component. Among factors jeopardizing human mobility, knee injuries and disorders occupy a significant place. For instance, the anterior cruciate ligament (ACL) is the most commonly injured ligament of the knee: every year, there are around 30 cases of ACL injuries for every 100,000 people [1]. Degenerative disorders, such as knee osteoarthritis, also pose a serious problem; for example, the lifetime risk of symptomatic knee osteoarthritis is estimated to be equal to 44.7% [2] and more than 1,500 damaged knees are replaced in Ireland each year [3].

While modern medicine made great advancements in the surgical treatment of those conditions, rehabilitation is still an indispensable part of a successful clinical outcome. However, the effectiveness of rehabilitation can be influenced by several factors: the quality of progress monitoring, the correctness of the prescribed measures and the patient's motivation, especially during unsupervised at-home rehabilitation. Devices that can remotely gather biomechanical and muscle activity data can be a solution to this problem. By such means, recorded training sessions can be processed and used to assess the patient's current condition and rehabilitation progress. Further, clinicians can leverage this information to make personalized adjustments in a rehabilitation program, while reducing the number of face-to-face meetings and associated costs without sacrificing the quality of the provided care.

2. Related work

Currently, we are witnessing a growing interest in wearable devices for lower extremities rehabilitation. Most frequently, research investigations look into activity recognition and performance evaluation using inertial measurement units [4]. Only a few proposed solutions feature multiple sensors, capable of motion tracking, muscle electrical activity recording [5] or plantar pressure distribution [6].

Another promising method that can be employed for joints health and implant condition assessment is acoustic emission (AE) monitoring. AE is radiation of acoustic waves that occurs during deformation. AE monitoring is widely used in research to investigate signals from bone fractures. However, little work has been undertaken to explore the possible use of AE for detecting more subtle and time-varying conditions [7].

3. Current research

At present, we are conducting a study to determine physiological biomarkers (measurable indicators of biological condition) best suited for remote assessment of the knee's condition. We plan to recruit patients scheduled for ACL reconstruction and total knee arthroplasty (TKA) as well as age, height and weight-matched control subjects. To follow patients' progress, trials will be set pre-operatively, early postoperatively and after completion of the recommended rehabilitation regimen. Data will be captured using motion tracking, surface electromyography, AEs, and ground reaction forces. Machine learning approaches will be adopted on the collected dataset.

4. Conclusion and future work

Ultimately, we aim to use the identified biomarkers to develop a wearable device in order to facilitate remote at-home monitoring and provide a tool for objective assessment of the rehabilitation progress.

Moreover, utilizing new approaches such as AE monitoring along with current gold standard techniques, we aim to shed new light on the mechanisms of TKA failures or causes of patients' dissatisfaction. This multi-sensor approach in a longitudinal study, will also allow a broad analysis of compensating strategies which are usually developed by patients after ACL reconstruction. The obtained results can contribute to a better understanding of factors behind the increased risk of re-injury or osteoarthritis development.

5. References

[1] Knee surgery, anterior cruciate ligament. https://www.hse.ie/eng/health/az/k/knee-surgery,-anterior-cruciate-ligament/

[2] L. Murphy *et al.*, Lifetime risk of symptomatic knee osteoarthritis, *Arthritis Rheum.*, 59(9):1207–13, September 2008.

[3] Knee replacement. https://www.hse.ie/eng/health/az/k/knee-replacement/recovering-from-a-knee-replacement.html

[4] M. O'Reilly, B. Caulfield, T. Ward, W. Johnston, C. Doherty, Wearable Inertial Sensor Systems for Lower Limb Exercise Detection and Evaluation: A Systematic Review, *Sports Medicine*, 48(5):1221–1246, May 2018.

[5] R. D. Gurchiek *et al.*, Remote gait analysis using wearable sensors detects asymmetric gait patterns in patients recovering from ACL reconstruction, *IEEE International Conference on Body Sensor Networks (BSN)*, May 2019.

[6] M. Ianculescu, B. Andrei, A. Alexandru, A smart assistance solution for remotely monitoring the orthopaedic rehabilitation process using wearable technology: Re.flex system, *Stud. Informatics Control*, 28(3): 317–326, September 2019.

[7] C. N. Teague *et al.*, Novel methods for sensing acoustical emissions from the knee for wearable joint health assessment, *IEEE Trans. Biomed. Eng.*, 63(8): 1581–1590, August 2016.

Deep Learning in Exercise-based CVD Rehabilitation

Ghanashyama Prabhu, Prof. Noel E. O'Connor, Prof. Kieran Moran
Insight SFI Centre for Data Analytics, Dublin City University, Dublin, Ireland.
ghanashyama.prabhu@insight-centre.org

1. Introduction

Cardiovascular disease (CVD) is the leading cause of premature death and disability in Europe and worldwide. World Health Organization reports physical inactivity as one of the main triggering cause for deaths due to CVD [2]. The deaths due to CVD are expected to rise more than 23.6 million by 2030, and thus exercise-based cardiac rehabilitation gained significance in secondary prevention programs which intern are very effective in lowering the recurrence rate of CVD and reduce the need for medicines.

Community-based rehabilitation programs are one such variant, however, everyone may not want to join these classes due to personal or travel reasons. Home-based, self-monitoring exercising is another such rehabilitation program in which a personal instructor could visit the patient's home, provide a tailored programme, monitor and provide personalized, qualitative feedback on the completed exercise. Unfortunately, this is not generally feasible in practice for a variety of reasons, including financial.

A technological approach by integrating wearable sensor (Wrist-worn: for aesthetic reasons) for assessing movement into an appropriate smartphone application (i.e. eHealth and eRehabilitation). Here in this work, we are using an existing CNN architecture model, AlexNet architecture [1], to recognize an exercise from a set of local muscular endurance (LME) exercises and then the same architecture is used to count the repetition count for the data obtained from a single wrist-worn sensor.

2. Objective

This method in general addresses two key challenges: firstly, it is important to be able to track which exercise is being completed and secondly, it is useful to provide quantitative feedback on completed exercise in terms of volume (repetition count) in order to build the user's confidence.

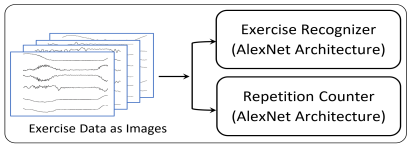


Figure 1: Pipeline for exercise-based CVD rehabilitation

3. Methodology

Pipeline for exercise recognition and repetition count is shown in Figure 1. AlexNet architecture is used for classification. Shimmer 3 (MSP430 CPU, 3D accelerometer [± 16g], 3D gyroscope [± 2000dps]), the robust, miniature wearable wireless sensor, calibrated and used to collect data from 76 participants (Males 46, Females 29; Age Group: 20 to 54) with a sampling rate of 512 Hz.

Data for a total of 10 LME exercises: six upper body LMEs (Bicep Curls, Frontal Raises, Lateral Raises, Triceps Extension of Right-arm, Pec Dec, and Trunk Twist) and four lower body LMEs (Squats, Lunges alternate side, Leg Lateral Raise - right, and Standing Bicycle Crunches) to be collected.

Participants performed each exercise for about 30 seconds and data from the wrist-worn sensor is Bluetooth streamed and stored using an exclusive MATLAB-GUI based module. Additionally, data collected for some commonly observed movements, like front bending, side bending/leaning, arm-stretching, upward leaning, and sit-to-stand to investigate how well the exercises could be distinguished by the deep learning models.

A 4-second window with 0.5second overlap is used in segmenting the sensor data. PNG images were generated with plots of each axis of accelerometer and gyroscope for each window. AlexNet architecture is used for 11-class exercise recognition. A second AlexNet architecture is used along with a counter for repetition recognition and counting from the labelled peak information of each image.

AlexNet models were trained for different combinations of optimizers(Adam, SGD, RMSprop), loss functions, learning rates and batch size to obtain the best model. Data from 46 distinct participants are used in model training. Data from 30 participants (15 each) used in validation and testing.

4. Results and Future Scope

Precision, recall, F1-score measures were computed on test-set on the best model possible with the validation set. The pipeline used in exercise-based CVD rehabilitation could recognize each exercise with overall 97.03% F1-score measure.

For repetition counting, in the case of upper-body exercises, zero error in counting was recorded with 98% for five exercises, however for Trunk-twist and other Lower body exercises except for lunges the zero error in counting was reported in 68%. It is observed that, in the case of lunges, repetition counting was very poor. Further, A single CNN model can be built which can perform both exercise recognition and repetition counting.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[2] W. H. Organization. `https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)`, 2017.

Optimal Intervention Strategies for Decision-Support in Epidemic Management

Andrea Yañez
Insight Centre for Data Analytics, National University of Ireland Galway, Ireland.
andrea.yanez@insight-centre.org

1. Introduction

Infectious disease spread remains as a threat to humankind. Emergence or re-emergence of pathogens can suddenly increment the number of ill individuals above the expected in an area (epidemic) or around the globe (pandemic). In the last decade, diseases such as Measles, AIDS, Malaria, Ebola, among others, still cause millions of deaths. The crucial question is not whether an epidemic will emerge, but when it does, *What are the most effective interventions (e.g., social distancing, school closure, vaccination) that can be applied to contain or reduce the spread?* This question poses a challenge to governments, public health officials and emergency response personnel; who must compare the potential impact of the vast number of possible decisions to select the most optimal interventions to implement. This task can easily become overwhelming. As a solution, decision makers generally use mathematical models or simulations to analyze and compare the performance of a limited number of *pre-selected* intervention strategies. Nonetheless, it is likely that this method could fail to consider the most optimal intervention strategies. My work aims to propose an optimization technique that can more effectively search the space of possible health policies and identify the most optimal interventions to control an epidemic. In this abstract, I present my proposed solution approach, its components and the challenges associated.

2 Problem Formulation

We frame the task of finding an optimal intervention strategy as a Reinforcement Learning (RL) problem in Figure 1. In RL, an agent within an environment executes an

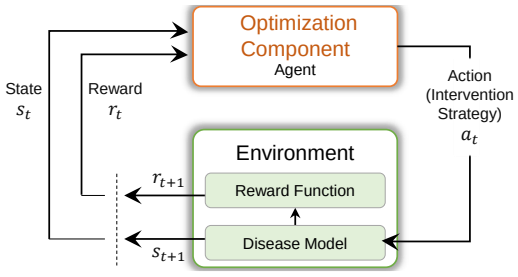


Figure 1: Identifying Optimal Intervention Strategies as a Reinforcement Learning Task.

action a_t over $t \in T$ discrete time steps or decision points. For each action, the agent receives a reward r_t. The goal is to find an optimal policy π^* that can indicate the best way to act for each environment state [1]. Similarly, we assume that public health officials employ mitigation strategies at discrete decision points during the evolution of an epidemic. The mapping of RL components to our problem setting can be summarised as follows (see Figure 1):

Agent/Optimisation Component: the Optimisation Component aims to explore the space of possible policies to learn an optimal policy. See subsequent sections for more details.

Environment: The environment is an abstraction of the problem that reduces it to signals that represent the option selected by the agent (interventions), the agent's new state and how the agent is performing (reward) [1]. Both the disease model and reward function will be further discussed in subsequent sections.

Reward r: The reward r_t is a numeric feedback signal given by the environment after an action a_t is executed. It is representative of the goals of the public health officials (e.g. reduce the number of infected individuals and costs).

State s: The state of the disease model has all the information that is necessary to represent any stage of the disease, e.g., number of infected individuals.

Action/Intervention Strategy a: In our proposed approach, an intervention strategy is the action executed by the optimisation component at a specific state s. An intervention x is a specific mitigation measure carried out to prevent or interrupt the spread of the disease. The set $\mathcal{A} = \mathcal{P}(\bigcup X_n)$ is the power set of all possible interventions, representing all combinations including the empty set indicating "no intervention".

Policy π: The policy is a function that indicates an action $a \in \mathcal{A}$ to perform for each state $s \in \mathcal{S}$, $\pi : \mathcal{S} \rightarrow \mathcal{A}$.

The main goal of this research is to design an optimisation component that can find an *optimal policy* which can be used by a public health official, to choose an action a in each disease state s to obtain the maximum expected reward after $t \in T$ decision points, i.e. $\pi^* = \arg\max_\pi E_\pi\left[\sum_{t=1}^{T} r_t\right]$.

3 Challenges

The main goal is to propose an optimization technique to find optimal policies for an existing disease model. My research faces a number of challenges which include: curse of dimensionality, defining reward functions and applying interventions whose impact is not immediately effective (e.g. vaccination) while maintaining Markov properties.

References

[1] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Subject-Dependent and -Independent Human Activity Recognition with Person-Specific and -Independent Models

Sebastian Scheurer
Insight Centre for Data Analytics, University College Cork
sebastian.scheurer@insight-centre.org

1. Introduction

Human Activity Recognition (HAR) systems are typically evaluated for their ability to generalise to either unknown or known users, which are referred to as, respectively, *subject-independent* and *subject-dependent* performance. If commissioning a HAR system entails obtaining examples of the activities from its end users, then the subject-dependent performance ought be optimised, suggesting that we train a personalised HAR model for each user. We refer to models obtained in this manner as *person-specific models* (PSMs), because they are tuned for a specific person. If the system is deployed without being trained on end users' data, then it must ship with a model pre-fitted to data from a (representative) sample of users. We refer to a model obtained in this manner as a *person-independent model* (PIM), because its performance is assumed to be independent of the person using it. Unfortunately, few papers assess both performance types for both PIM and PSMs, and none for all four combinations. This paper aims to close this gap with an empirical comparison of the subject-dependent and -independent performance achieved by combining PIM and PSM with four popular machine learning algorithms and applying each to eight HAR data-sets acquired from a body-worn inertial sensor.

2. Methods

We follow the standard approach to human activity recognition comprised of data pre-processing, segmentation into windows, feature extraction from those windows, and activity inference on them based on their features—with the inference step implemented with a machine learning algorithm [1]. We assess four popular machine learning algorithms—L_2 regularised logistic regression, kNN, SVM, and a gradient boosted ensemble of decision trees (GBT)—using a set of features extracted from eight publicly available data sets. We estimate the subject-independent performance via leave-one-subject-out cross-validation (CV), and the subject-independent performance via a separate k-fold CV for each user.

In addition to PIMs and PSMs, we also consider ensembles of PSMs (EPSMs). An EPSM maintains a PSM for each known user. When an instance for a known user needs to be classified, an EPSM simply applies that user's PSM, but when an instance originates with an unknown user, it applies each user's PSM to obtain confidence scores for each activity. Then the EPSM calculates each activity's mean score, and classifies the instance to the activity with the maximum mean score.

3. Results

Figure 1 illustrates the trade-off between the subject-dependent and -independent performance ($\kappa \times 100$). Each data point corresponds to a user, except for PIMs where it corresponds to the median. Each panel depicts the results for one data-set, and the symbol and colour indicate whether a PIM, PSM, or EPSM was used. A logistic mixed effects model analysis of these data show that GBT outperforms the other algorithms on both subject-dependent and -independent performance, PSMs outperform PIMs by 3.5% on known users, and PIMs outperform PSMs by 13.9% on unknown users. The analysis further shows that although PIMs outperform EPSMs on unknown users by 2.1%, that difference is not significant at the $\alpha = 0.05$ significance level. The estimated difference between subject-dependent and subject-independent performance ranges from 20.5% to 26.1% with PIMs, 27.5% to 33.3% with EPSMs, and from 42% to 48.5% with PSMs.

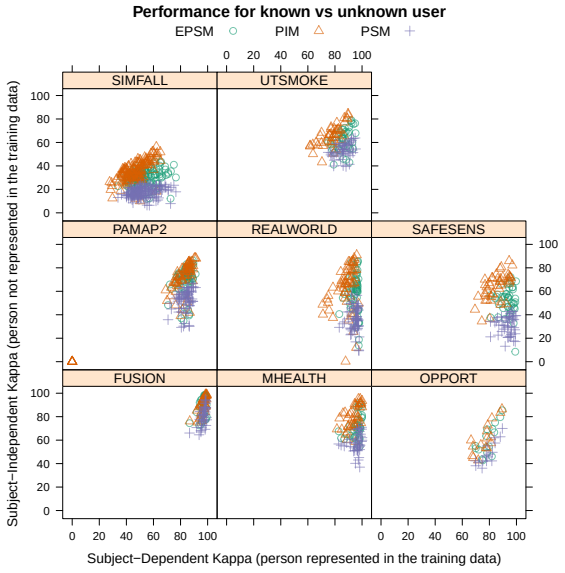


Figure 1: Subject-independent versus -dependent κ (%)

References

[1] A. Bulling, U. Blanke, and B. Schiele. "A Tutorial on Human Activity Recognition Using Body-worn Inertial Sensors". In: *ACM Computing Surveys* 46.3 (Jan. 2014). DOI: 10.1145/2499621.

Evaluating the Impact of Data Loss on Orientation Estimation

Clare Lillis
Dublin City University
Clare.lillis@insight-centre.org

1. Introduction

The advancement of micro-electro-mechanical systems has led to the use of inertial measurement units (IMUs) in areas related to Health, Sport and Game development. These IMUs are specifically suitable for use in human motion analysis, due to their small and non-intrusive size, portability and low-cost technologies. They usually contain a combination of sensors, for example, accelerometers, gyroscopes or magnetometers. The combination of sensors is used for more accurate sensing information [1]. IMUs have potential to perform at a sufficient level of accuracy when compared to laboratory-fixed motion analysis systems. IMUs have their own associated limitations including errors associated with the individual sensors within the sensing device. Inertial sensors can also suffer from occasional data loss.

2. Motivation

In many of our experiments data loss has been present in varying amounts. For some applications this loss of data may not be detrimental or noticeable in the signal. When capturing fast movements such as walking or running, depending on speed, a gait cycle may occur well within the number of samples sampled in 1 second. When information is lost it may be difficult to get an accurate analysis of spatiotemporal parameters and joint angles. Related studies have dealt with this by either disposing of windows of data which did not meet a set threshold for data loss, or simply recording multiple participant trials with the hope one would be useable. Our interest in this problem arose from calculating segment orientation from raw inertial sensor data using the Madgwick filter described in [1] for walking and running. When calculating orientation, the operations rely on the last calculated orientation to compute the new orientation using the sampled raw inertial data. We noticed how poorly the orientation algorithm dealt with missing data when the β parameter was not optimized for the specific task. Setting the parameter to the suggested value of 0.041 [1] gave us significant problems in the resulting orientation. It was observed that around points of data loss, the accuracy of the orientation calculation deteriorates as each calculation relies on the accuracy of the last. This is important to investigate as we were only able to optimize β for our experiments because we also had ground-truth data captured for comparison. Studies without reference data use the suggested value.

3. Methods

After ensuring the issues were not sensor specific, we investigated the impact of data loss on orientation accuracy. We did this on controlled loss-free inertial data and on (naturally occurring) loss present inertial data. Loss free test data was created from inertial running data by selecting a number of continuous gait cycles which were free from data loss. We experimentally removed large amounts of samples from the test data to recreate high data loss and then tested our gap filling methods (linear and cubic spline interpolation) before calculating orientation. This data was removed intentionally from two important regions in the gait pattern, peak and non-peak regions, to see how this would affect the algorithm. Filling large gaps of continuous missing data (i.e. >80 samples) had a negative effect on accuracy. Gaps not filled in the inertial data were filled in Euler data to keep file length uniform. We then investigated at what point the amount of missing data affected accuracy. We experimentally removed samples from the loss-free data to recreate random data loss, applied the filling methods and calculated orientation. In individual experiments we removed a number of consecutive samples <10, <20, <40, <80 and experimented with a combination of these data losses, applying gap filling up to a certain gap size threshold to see which produced the most accurate orientation. This was to find a gap filling threshold that up to which we fill consecutive gaps in the inertial data, and any larger gaps would be filled in the Euler data. We also ran these experiments on the data with natural data loss and computed the orientation. We compared the results of all experiments to both the ground-truth data and the orientation computed from the same inertial data where missing data was ignored and filled after calculating orientation.

4. Conclusion

The amount of data lost and the region of loss in the signal together impact how much the inertial signal is altered compared to no loss, and in turn adds to the errors when orientation is calculated. A lower gap filling threshold meant that the data filled was an accurate representation of the data lost. We determined a rule for filling missing data, any consecutive data loss of size >20 samples would not be filled in inertial data but filled after the orientation data was computed. We concluded that the best approach was to use was linear interpolation with <20 gapsize threshold. We found that depending of the amount of loss, gap filling the data before calculating orientation was either more accurate or had equal accuracy to ignoring the missing data.

References

[1] Madgwick, Sebastian OH, Andrew JL Harrison, and Ravi Vaidyanathan. "Estimation of IMU and MARG orientation using a gradient descent algorithm." 2011 IEEE international conference on rehabilitation robotics. IEEE, 2011

Is Navicular Drop Associated with Running Related Injuries?

Sarah Dillon[1,2], Dr Enda Whyte[2], Aoife Burke[1,2], Dr Siobhán O'Connor[2] ,Dr Shane Gore[1,2], Prof Kieran Moran[1,2]
[1]Insight Centre for Data Analytics, [2]Dublin City University
sarah.dillon@insight-centre.org

1. Introduction

Despite a high incidence of running-related injuries (RRIs) of between 7.7-17.3 per 100 hours of running[1], the etiology remains debated. One biomechanical factor which may be of importance in relation to the development of RRIs, is the magnitude of foot pronation. This is because increased pronation is associated with increases in loading[2], which may exceed the tissue's capabilities. As a commonly used clinical measure of subtalar pronation, navicular drop (ND) has been the heavily investigated[3]. However, research remains conflicting. One methodological consideration which may explain this conflicting evidence, is the comparison of injured subjects with control subjects with a history of injury. Using this methodological design, is unclear if an individual's current biomechanical state varies depending on the duration of time since injury. As such, a comparison of 'recently injured' (between 3 months and 2 years prior), 'injured > 2 years' and 'never injured' groups may reveal potentially important distinctions between groups. This research aims to investigate potential differences in ND between these groups.

2. Recruitment

Participants were recruited using posters, emails and social media. Inclusion criteria included; recreational runners, aged 18-65, no injury within the last three months, no lower limb surgery within the past six months and no involvement in contact sports. Recreational running was defined as a minimum of 10km per week, for at least six months prior.

3. Methods

To measure ND, a single tester palpated and marked the navicular[4]. Using a ruler, the distance (mm) from the navicular to the floor was measured in sitting and standing. The difference between measurements was calculated, with the average of three measurements recorded. Previous RRI history was obtained via online questionnaire. An RRI was defined as a pain in the lower limbs that required a restriction of running (distance, speed, duration or training) for at least 7 days, three consecutive training sessions or required consultation with a healthcare professional[5].

4. Data Analysis

Data were analyzed using SPSS (23; IBM, NY). A one-way between subjects ANOVA was conducted to compare the effect of navicular drop on injury status. Assumptions of the ANOVA were tested prior to running the analysis.The p value for determining significance was set at $p \leq 0.05$.

5. Results

Two hundred and eighty recreational runners participated in this study. Of these, 39 had never been injured (27 males, 12 females 41.9 ± 10.9 years), 51 had been injured more than 2 years prior (32 males, 19 females, 43 ± 7 years) and 190 had sustained an RRI between 3 months and 2 years prior to participating (116 males, 74 females, 43 ± 8 years). No significant differences were found between any of the demographic variables. No significant differences were found between groups for ND on dominant ($F(2,258)=.262$, $p=.770$) or non-dominant sides ($F(2,258)=1.703$, $p=.184$). Results are presented in table 1.

Table 1-Results of the 1- way ANOVA for ND. INJ- injured, SD- standard deviation. ND- navicular drop.

	INJ >2 years prior Mean ± SD	INJ 3 months-2 years prior Mean ± SD	Never INJ Mean ± SD	P value
ND Dominant (mm)	8.2 ± 3.3	8.0 ± 3.4	8.4 ± 3.2	.770
ND Non-Dominant (mm)	8.8 ± 3.5	8.6 ± 3.1	9.5 ± 3.5	.184

6. Discussion

No significant difference in ND was noted between the three groups. Though this contradicts the commonly held belief that excessive pronation may be associated with increased injury[2], this is in line with some of the research available[3]. The findings of this study suggest that there may be no link between pronation and RRI among runners. Alternatively, the lack of significance between previous RRI and ND may be because the measure itself may not accurately reflect the dynamic pronation that occurs during gait.

7. Limitations and future directions

The statistically non-significant finding in this study between retrospective RRI and ND, adds to the evidence refuting the association. Other measures of foot position that better capture pronation during movement may be more appropriate for use in a running population.

8. References

[1] Videbaek, S., Bueno, A. M., Rasmus, B., Nielsen, O., Rasmussen, S., Videbæk, S., Bueno, A. M., Nielsen, R. O., & Rasmussen, S. (2015). Incidence of Running-Related Injuries Per 1000 h of running in Different Types of Runners: A Systematic Review and Meta-Analysis. *Sports Med*, *45*(7), 1017–1026.

[2] Mei, Q. et al. (2019) 'Foot Pronation Contributes to Altered Lower Extremity Loading After Long Distance Running', Frontiers in Physiology. Frontiers, 10, p. 573.

[3] Buist, I. *et al.* (2010) 'Predictors of running-related injuries in novice runners enrolled in a systematic training program: A prospective cohort study', *American Journal of Sports Medicine*, 38(2), pp. 273–280.

[4] Eslami, M., Damavandi, M. and Ferber, R. (2013) Association of Navicular Drop and Selected Lower-Limb Biomechanical Measures During the Stance Phase of Running Journal of applied biomechanics 30(2).

[5] Yamato, T. P., Saragiotto, B. T. and Lopes, A. D. (2015) 'A Consensus Definition of Running-Related Injury in Recreational Runners: A Modified Delphi Approach', *Journal of Orthopaedic & Sports Physical Therapy*, 45(5), pp. 375–380.

Convolutional Neural Networks for Heart Rate Estimation and Human Activity Recognition in Wrist Worn Sensing Applications*

Eoin Brophy
Insight SFI Research Centre for Data Analytics, Dublin City University
eoin.brophy@insight-centre.org

1. Introduction

Wrist-worn smart devices are providing increasingly more useful insights into human health, behaviour and performance through ever more sophisticated analytics. However, battery life, device cost and sensor performance in the face of movement-related artefact all present challenges which must be further addressed to see more effective applications and wider adoption through commoditisation of the technology. We address these challenges by demonstrating that through using only a simple optical measurement, i.e. photoplethysmography (PPG), used conventionally for heart rate detection in wrist-worn sensors, we can provide improved heart rate and human activity recognition (HAR) simultaneously at low sample rates without the need for an inertial measurement unit. This simplifies hardware design, and reduces both bill of materials and power budgets, in turn facilitating the shift in health policy from reactive to proactive treatment-based models where the focus is increasingly on keeping people healthy.

2. Related Work

Biagetti *et al.* conducted a study on the same dataset used in this paper for activity recognition [1]. Using the PPG data only for HAR they achieved 44.7% classification accuracy using their feature extraction algorithm.

Reiss *et al.* sought to solve a regression problem by estimating heart rate from PPG and accelerometer data [3], using a standalone PPG we develop a convolutional neural network regression (CNNR) architecture for heart rate estimation on a single channel time series without any preprocessing.

3. Methodology

We implement two different deep learning pipelines, one for HAR and one for heart rate estimation. HAR is achieved by leveraging transfer learning to retrain a CNN that was pre-trained on ImageNet to distinguish between different patterns in the time domain characteristics of the PPG trace during different human activities.

For heart rate estimation we use a CNN adopted for regression which through training with PPG data maps noisy optical signals to heart rate estimates. In both cases, comparisons are made with leading conventional approaches. A publicly available PPG exercise dataset was used for the experiments in this paper [2].

*This work is funded by Science Foundation Ireland under grant numbers 17/RC-PhD/3482 and SFI/12/RC/2289

[1] HeartPy Project: `https://pypi.org/project/heartpy/`

4. Results

4.1. HAR

Our results demonstrate that a low sampling frequency can achieve good performance without too much degradation in accuracy. At 10Hz sampling frequency, we achieved 83.0% classification accuracy for HAR and the same frequency also yielded a robust heart rate estimation that performs similarly well to the heart rate error for 256Hz.

Table 1: F1 score of HAR Classifier on 10Hz sampling frequency

Exercise	High	Low	Run	Walk
F1-Score	0.851	0.745	0.837	0.865

4.2. Heart Rate Estimation

The average error across all exercises and sampling frequencies decreased using our method from 22.59% to 20.15%, an increase in over 2 percentage points relative to the performance of the open-source tool kit HeartPy[1] which served as a baseline here.

5. Conclusion

The results shown above demonstrate that more cost and power-efficient wearables are possible through the exploitation of secondary information available from a simple optical sensor. This suggests single-sensor based wearables can achieve much of the functionality and capabilities of more complex multi-modal wearables. An area of future work lies in the optimisation of the networks that will help our models rival the current commercial standards.

References

[1] G. Biagetti, P. Crippa, L. Falaschetti, S. Orcioni, and C. Turchetti. Human activity recognition using accelerometer and photoplethysmographic signals. In I. Czarnowski, R. J. Howlett, and L. C. Jain, editors, *Intelligent Decision Technologies 2017*, pages 53–62, Cham, 2018. Springer International Publishing.

[2] D. Jarchi and A. Casson. Description of a Database Containing Wrist PPG Signals Recorded during Physical Exercise with Both Accelerometer and Gyroscope Measures of Motion. *Data*, 2(1):1, 2016.

[3] A. Reiss, I. Indlekofer, P. Schmidt, and K. Van Laerhoven. Deep ppg: Large-scale heart rate estimation with convolutional neural networks. *Sensors*, 19(14):3079, Jul 2019.

Developing approaches to biodiagnostics via wearable platforms

Melissa Finnegan[1], Emer Duffy[1], Aoife Morrin[1]

National Centre for Sensor Research, School of Chemical Sciences, Dublin City University, Dublin 9[1]

melissa.finnegan@insight-centre.org

1. Introduction

The focus of this research is to develop and apply new and existing soft responsive materials capable of responding to target biomarkers on the body directly via the skin and also in the biofluid sweat. One of the approaches being investigated is the use of responsive colorimetric dyes to track the profile of skin volatiles that are being emitted from the skin. Specifically, the acidity of the volatile profile is being assessed using pH-responsive dyes. Early evidence from the group has shown that the fatty acid emission (C8-C16) correlates well with skin surface pH. This early study used a headspace-solid phase microextraction (HS-SPME) approach to sample the skin volatiles which were then analysed by GC-MS. This project hopes to exploit this correlation of acidic volatile emissions with pH to develop a non-contact, skin pH sensor.

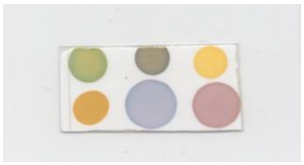


Figure 1: Colourimetric sensor

2. Experimental

A wide range of responsive colorimetric dyes are being tested in this work to understand their response to skin volatiles. pH, solvatochromic and metalloporphyrin dyes are being used in the sensor array and are immobilised to cellulose substrates using a sol-gel. The use of the sol-gel matrix allows soluble dyes to be converted into nanoporous pigments which are insoluble. The durability and stability of the sensor arrays is improved by using this method of immobilising the pH or porphyrin dye into the sol-gel matrix.[1] The response of each dye is recorded by imaging the sensor array before and after exposure to volatiles. ImageJ is used to calculate the red, green and blue value of each dye and the net colour response is quantified.

A host of different skin sampling methods have been examined including gauze sampling, swab sampling, tape stripping and also direct application on the skin incorporating a wire mesh spacer between the sensor array and the skin.

3. Results

The following results are from indirect experiments carried out. This data shows that for a higher pH value the pH indicator dyes will exhibit a high response while the porphyrin dyes will exhibit a low response.

For a lower pH value the pH indicator dyes (figure 2) will exhibit a lower response while the porphyrin dyes (figure 3) will exhibit a higher response.

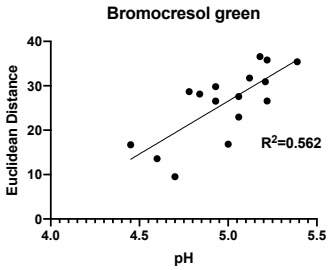


Figure 2: Bromocresol green results

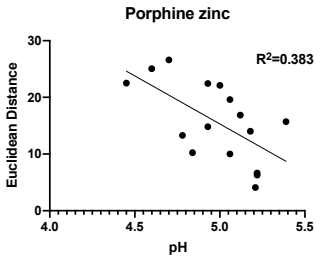


Figure 3: Porphine zinc results.

4. Conclusion

In conclusion, there are some promising results obtained from the indirect experiments carried out. For future work, more indirect sampling will be carried out to further increase the data set. Direct sampling will also be carried out in order to elicit a response from the sensors.

5. References

(1) Lim, S. H.; Musto, C. J.; Park, E.; Zhong, W.; Suslick, K. S. A Colorimetric Sensor Array for Detection and Identification of Sugars. Org. Lett. 2008, 10 (20), 4405–4408. https://doi.org/10.1021/ol801459k.

SmartSense:

Development of a machine learning algorithm for pump clogging prediction

Rafael Torrecilla Rubio
Insight – Dublin City University
rafael.torrecillarubio@insight-centre.org

1. Introduction

Current population expansion is leading to higher demand of more reliable wastewater systems [1]. Of particular concern is the increased risks posed by flexible thin film structures found in sewage water, such as rags or cloth-like objects. These are routinely found to cause pump clogging with significant and costly consequences on operation. Although anti-clogging solutions exist, they often compromise pump efficiency [2].

Even though research on anti-clogging solutions continues to play an important part in hydraulics design, there is currently no reliable solution [3].

Research aim is to allow a more predictive control to limit the impact of clogging by coupling a low-cost sensing solution with Machine Learning algorithms.

2. Methods

Literature review about applicable solutions is to be conducted for the selection of the appropriate low-cost technique.

Once that the sensing system has been identified, the proof of concept is validated in a water tunnel where rags are released and sensed with the prototype.

Satisfactory prototypes are then tested in a rig where a certain number of rags can previously be introduced in a tank from where they flow into the pump passing through the prototype.

Extracted data is analysed and a Machine Learning algorithm is developed and trained to upgrade the system. Coupled prototype with Machine Learning is tested again and reliability is measured.

3. Current results and future work

Available sensing solutions have been reviewed. The most appropriate techniques for real-time mapping of multiphase flows are Electrical Resistance Tomography and Electrical Impedance Tomography, both of which can easily be implemented as a ring in a pipe section [4] (see Figure 1).

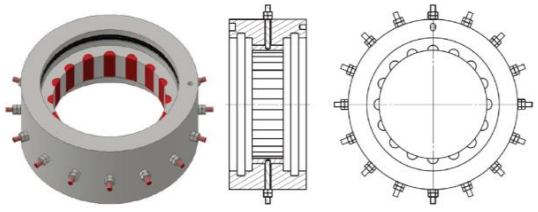


Figure 1 - Ring design for Electrical Resistance Tomography for installation in pipes [4].

Figure 2 explains the working principle. With a configuration of 16 electrodes, the current flows from the injection electrode 1 to the receiving electrode 2. Differences in voltage within the current field can be measured by the remaining pairs of electrodes. This process is repeated for all electrode combinations [5].

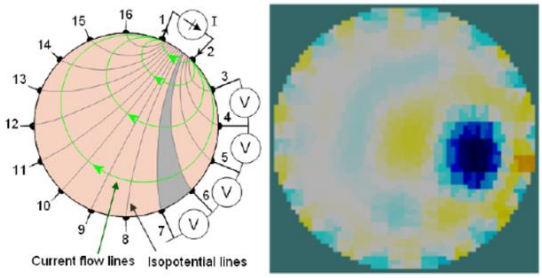


Figure 2 - Working principle for an Electrical Impedance Tomography system (left) [5] and illustration of image reconstruction result (right) [6].

As illustrated in Figure 2 (right), measured signals can be processed to reconstruct internal images of immersed structures by mapping the resistance or impedance between pairs of electrodes [6].

Current results show that thin rags can be detected. Although resulting signal is weak, signal treatment can be used to amplify the signal while filtering out noise.

Future work includes a prototype construction, which will be tested in a water tunnel to determine its capabilities and limitations and then, in a rig to extract necessary data to develop and train the Machine Learning algorithm to predict clogging and compensate poor resolution.

4. References

[1] R. Connolly, “An Experimental and Numerical Investigation into Flow Phenomena Leading to Wastewater Centrifugal Pump Blockage,” no. September, pp. 1–67, 2017.
[2] E. Akrami, M. Specklin, B. Breen, R. Connolly, A. Albadawi, and Y. Delauré, “Numerical characterisation of the dynamics of thin flexible structures in a sewage pump,” no. 037, pp. 1–10, 2019.
[3] A. Albadawi, M. Specklin, R. Connolly, and Y. Delauré, “A thin film fluid structure interaction model for the study of flexible structure dynamics in centrifugal pumps,” J. Fluids Eng. Trans. ASME, vol. 141, no. 6, 2019.
[4] P. Pavlíček, “Experimental equipment for an electrical resistance tomography of a gas lift flow,” AIP Conf. Proc., vol. 2047, no. November, 2018.
[5] J.Malmivuo, “Bioelectromagnetism,”no. November, 2017.
[6] "EIDORS3D," [Online].Available: http://eidors3d.sourceforge.net/tutorial/EIDORS_basics/tutorial110.shtml. [Accessed 11 December 2019].

Asynchronous Distributed Clustering Algorithm for Wireless Sensor Networks

Cheng Qiao
University College Cork
qiao.cheng@insight-centre.org

1. Introduction

In distributed clustering problems, nodes in a wireless sensor network must learn clusters from the data sensed across the network, without centralising the raw data. Each node acts as its own decision maker, but must communicate with other nodes to learn wider network patterns -e.g. temperature readings in indoor heating systems, or vehicle movements in traffic control. This method trades off potentially lower decision quality for reduced communication, reduced energy use and improved privacy.

Without any priori knowledge about the true number of clusters and how many different patterns across the whole network, we attempt to learn the global pattern across all the nodes in a wireless mesh network, while minimising inference time and communication cost, and respecting the privacy of the raw data.

2 Related work

The state-of-art technology is proposed by Datta, Fatta, Bénézit and Bendechache.

- Bendechache et al. [1] suggested that represent the cluster by boundary points in a tree-based network.
- Datta et al. [3] proposed a synchronous distributed k-means algorithm, exchanging centroids and counts each round.
- Bénézit, et al. [2] and Fatta et al. [4] offered a similar approach, but using gossip to exchange information.

However, the previous algorithms synchronise the behaviour of the sensors [2, 3, 4] and ignore the communication cost [1, 2, 3, 4].

3 Research questions and methodologies

Initially, we assume that the number of clusters K is known to agents in advance and all agents are receiving data from the same pattern, which is widely adopted by the state-of-art. Different in-network clustering approaches including k-means and Gaussian Mixture Models, and different methods of summarising clusters to exchanged between nodes are considered. In experiments on randomly generated network topologies, we demonstrate that methods which do more extensive clustering in each cycle, and which exchange descriptions of cluster shape and density instead of just centroids and data counts, achieve more consistent clustering, in significantly shorter elapsed time.

Then, we relax the assumption that K is fixed in advance. The basic idea is that each agent evaluates K by the silhouette method locally and centroids, size of clusters and shape of clusters are shared to its neighbours. Agents who received this summary description estimate the original information by regeneration. We compared our method to centralising methods, including centralising all raw data and centralising basic models to a central agent. Although the clustering accuracy achieved by proposed distributed methods dropped by around 9% and the communication cost increased by as much as 20%, the convergence time has been reduced, the problem of single point of failure has been avoided and data privacy is respected.

In some real applications, there might be sub-groups of agents that are receiving different patterns of data, and this must be identified. The agent that does the final clustering is responsible for determining what sub-patterns exist. Since K varies with agent and sub-pattern, Weighted Earth Mover's Distance (EMD) between centroids are used to measure the similarity between agents. Then a hierarchical dendrogram was used to evaluate the appropriate number of sub-patterns and spectral clustering is applied to find all sub-patterns. Experimental results showed that the proposed method could detect the right number of patterns, put agent into the right sub-pattern and preserves high clustering accuracy.

4 Future work

We will extend our algorithms and experiments to include networks in which nodes fail (e.g. because of limited batteries) and we will address application scenarios in which data distributions change over time. Finally, we will extend our methods to handle different inference problems, identifying which problems can or should be handled in the network, and which require transmission to a central server for more extensive analysis.

References

[1] M. Bendechache, N.-A. Le-Khac, and M.-T. Kechadi. Hierarchical aggregation approach for distributed clustering of spatial datasets. In *IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, pages 1098–1103. IEEE, 2016.

[2] F. Bénézit, V. Blondel, P. Thiran, J. Tsitsiklis, and M. Vetterli. Weighted gossip: Distributed averaging using non-doubly stochastic matrices. In *2010 IEEE International Symposium on Information Theory*, pages 1753–1757. IEEE, 2010.

[3] K. H. Datta S, Giannella C. Approximate distributed k-means clustering over a peer-to-peer network. *IEEE Transactions on Knowledge and Data Engineering*, 10(21):1372–1388, 2009.

[4] G. Di Fatta, F. Blasa, S. Cafiero, and G. Fortino. Fault tolerant decentralised k-means clustering for asynchronous large-scale networks. *Journal of Parallel and Distributed Computing*, 73(3):317–329, 2013.

Overview of the Habitat Mapping, Assessment and Monitoring with High-Resolution Imagery (iHabiMap) Project

Charmaine Cruz[1,2], Kevin McGuinness[1,3], John Connolly[1,2]
[1]*Insight Centre for Data Analytics, Dublin City University*
[2]*School of History and Geography, Dublin City University*
[3]*School of Electronic Engineering, Dublin City University*
charmaine.cruz@insight-centre.org

1. Introduction

Despite the ecological importance of natural habitats, they are facing threats of loss and degradation. The Habitats Directive [1] requires EU countries to accurately map and monitor the condition of these habitats. Ireland is committed to this Directive and must report, map, and monitor the conservation status of its habitats listed in Annex 1 of the Directive based on field-based ecological data. The ecologists evaluate Ireland's habitats every six years [2]. The field-based mapping and assessment methods, while still desirable, are time-consuming, difficult and expensive. Thus, another mapping approach should be considered as an important supplement to traditional methods. Mapping using remote sensing techniques could aid in the monitoring and evaluation of Irish habitats in a cost-effective and repeatable manner. The advent of Unmanned Aerial Vehicles (UAVs) delivers new developments in the field of remote sensing by providing multi-sensor images with centimeter-level resolution. In addition, UAVs offer flexible data acquisition suited for monitoring and change detection applications due to their independence from weather and cloud cover.

2. Project Aim

The "Habitat Mapping, Assessment and Monitoring using High-Resolution Imagery" Project or iHabiMap is a part of Ireland's initiative to produce a detailed assessment of its habitats using ultra-high resolution images acquired from UAVs. The project, led by Dublin City University, aims to develop analytical approaches by utilizing UAVs and machine learning algorithms to map, assess, and monitor three habitats - upland, grasslands, and coastal zones.

3. Methodology

Figure 1 shows the general methodology for this study. Multispectral data will be acquired and tested for each habitat assessment. The methodology will provide a reproducible automated technique to enable frequent habitat mapping in Ireland. Field surveys will be conducted alongside the UAV data acquisition at each study site. In-situ and UAV data will be acquired concurrently over a three or four-year period, which aims to capture both the intra- and inter-annual variability (Table 1).

Table 1. Field and UAV Surveys throughout the Project

Site Name	Y1	Y2	Y3	Y4
Slieve Mish	0	4	1	0*
Magharees	0	4	1	0*
Bull Island	1	1	0	0*
Glenasmole	0	4	1	0*
Liffey Head	1	1	0	0*
Total				**19**

4. Preliminary Activities and Future Work

UAV and ecological data were initially acquired for two sites (i.e., Bull Island and Liffey Head). These datasets will be processed using machine learning algorithms. The expected outputs of this study include (1) a hierarchical Random Forest algorithm developed to classify habitat data, and (2) accurate Annex 1 habitat maps for the study sites.

5. Reference

[1] Commission of the European Communities. (1992). Council Directive 92/43/EEC of 21 May 1992 on the conservation of natural habitats and of wild fauna and flora. In Official Journal of the European Union.
[2] NPWS (2013). The Status of Protected EU Habitats and Species in Ireland. Overview, Volume 1. Unpublished Report, National Parks & Wildlife Services.

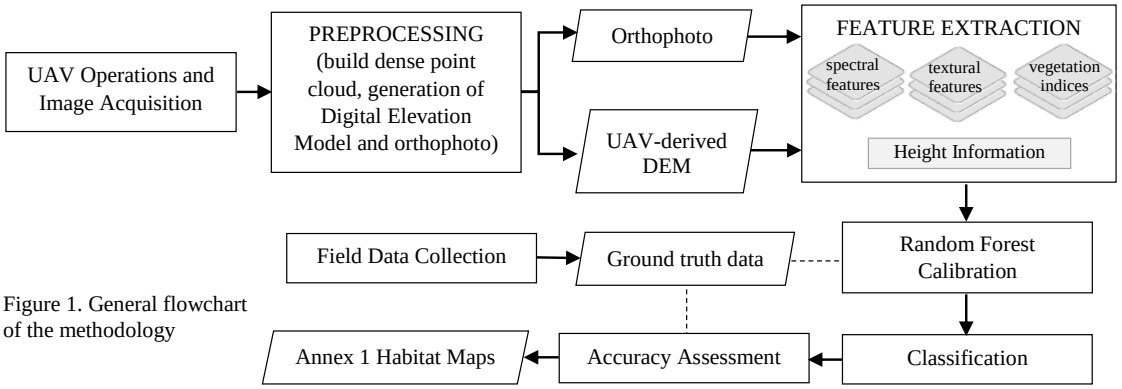


Figure 1. General flowchart of the methodology

Intra-Session Reliability & Discriminative Validity of IMUs as a Measure of the Forward Lunge

James Davenport
Insight Centre for Data Analytics, University College Dublin
james.davenport@insight-centre.org

1. Introduction

The Forward Lunge is a lower limb functional movement which incorporates components of strength, flexibility and balance [1]. It exaggerates the movements occurring at the lower limb during the gait cycle. Clinicians commonly use it in the identification of functional deficits helping steer them on the clinical decision-making pathway. Current measures of the forward lunge are restricted to either expensive laboratory based objective measures such as motion capture systems or subjective visual clinical interpretation, which demonstrates poor levels of reliability. Recent developments in the area of body worn inertial sensors has meant that the objective quantification of motor function tasks can be performed, without the barriers associated with other methods [2].

2. Aims

This research aims to establish the intra-session reliability and discriminative validity of kinetic measures derived from two individual shank based inertial sensors during the forward lunge.

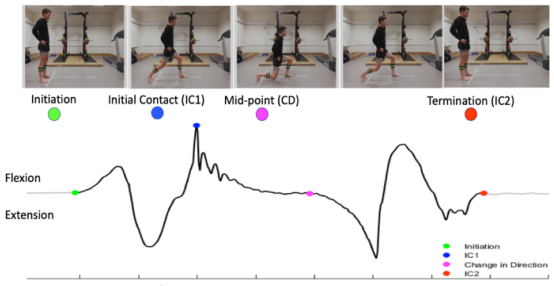


Figure 1: Gyro signal from lead shank IMU

3. Methods

Twenty-three healthy participants took part in the study (12 Male, 11 Female, 30.8 ± 8.6yrs, 1.7 ± 0.9m, 65.3 ± 10.8kg). Each participant performed 3 sets of 5 lunges bilaterally at 0-, 10-, and 20-minutes pre- and post-intervention (60-second modified Wingate test), aimed to introduce a short-term alteration in motor function through central and peripheral fatigue.

Lunge distance and stance was set as 100% of leg length and hip width respectively (± 5%). IMUs were worn on the lateral aspect of each shank. The lunge was segmented into initiation, initial contact, mid-point, and termination.

Peak & root mean squared (RMS) of total acceleration signals of the shank based IMU were taken for all lunges. Intraclass correlation coefficients (ICCs) were calculated based on a mean rating (k=3), absolute agreement two-way mixed effects model. Intra-session reliability was defined as poor (ICC<0.5), moderate (0.5-0.75), good (0.75-0.9) or excellent (>0.9). A series of one-way Analysis of Variance (ANOVA) with repeated measures were conducted to establish the discriminant validity of IMUs in capturing changes in features from altered motor control. The level of significance was set a priori at P<0.05.

4. Results

ICC values for peak acceleration ranged from 0.916 to 0.981 for left limb and 0.903 to 0.978 for right limb. ICCs for RMS of left and right limb ranged from 0.908 to 0.979 and 0.899 to 0.977 respectively.

The fatiguing exercise induced an immediate increase in peak acceleration; T_{0min} (P = 0.002) for the right limb. However this was not retained at T_{10min} (P = 0.111) and T_{20min} (P=0.254). The change in peak acceleration was not significant across any time points on the left limb T_{0min} (P<0.149), T_{10min} (P=0.808) and T_{20min} (P=0.869)

There was an immediate increase in RMS in the left limb T_{0min} (P<0.001) which was not retained at T_{10min} (P=0.186) and T_{20min} (P=0.409). Similarly, for the right limb, there was an immediate increase in RMS at T_{0min} (P<0.001). The increased RMS was retained at T_{10min} (P < 0.001) and T_{20min} (P = 0.019) for the right limb.

5. Discussion

The IMUs features showed good to excellent intra-session reliability across the pre-fatigue measures. Demonstrating the IMUs were capable of capturing repeated measures of the forward Lunge with excellent accuracy. The One-way repeated measures ANOVA were used to compare the final pre-fatigue to the post-fatigue measures to identify changes that had occurred and a pairwise comparison with Bonferroni post-hoc highlighted where these changes had occurred. It showed that changes differed between limb and over time points. This demonstrates potential clinical utility with scope for further investigation.

7. References

[1] Cook G, Burton L, Hoogenboom B. Pre-participation screening: the use of fundamental movements as an assessment of function - part 1. North American journal of sports physical therapy: NAJSPT. 2006;1(2):62-72.

[2] O'Reilly M, Caulfield B, Ward T, Johnston W, Doherty C. Wearable Inertial Sensor Systems for Lower Limb Exercise Detection and Evaluation: A Systematic Review. Sports medicine (Auckland, NZ). 2018;48(5):1221-46.

Bio-Impedance Measurement System for Biomedical Applications

Ardeshir Behrouzirad
PhD Student, Tyndall National Institute-University College Cork
ardeshir.behrouzirad@tyndall.ie

1. Introduction

Bioimpedance measurement has a wide range of applications including the estimation of body composition, early diagnosis of disease in the human body and cell health monitoring during the cell culturing process. Bioimpedance is a measure of the ability of biological tissue to impede electrical current. Bioimpedance is measured by detecting the response of an electric excitation applied to biological tissue[1]. The focus of this research work is to develop a wearable bioimpedance measurement system to perform early detection of disease. The bioimpedance system is in the early stages of development with the focus currently on the design of the electronics instrumentation. Bioimpedance instrumentation consists of an excitation circuit and a measurement circuit. As the first step towards this goal, the excitation circuit was designed, implemented and tested. The design objective was to produce a circuit with configurable excitation capability for amplitude and frequencies. This excitation circuit delivers an AC current to the body and is designed to meet the safety requirements defined by the IEC60601 standard. A modified Howland circuit has been implemented with commercially available components and measurements have been successfully made on an equivalent circuit representing the human body. The pros and cons of the different circuit implementations will be investigated and an optimized solution will be found. The outcome of this research will inform the specification and roadmap to achieve a complete system on a chip.

2. Howland Circuit as a Current Pump

A Howland circuit topology was used to implement the excitation circuit. The Howland circuit is a voltage-controlled current pump. It can both sink and source precise amounts of current, provide high output impedance and high frequency bandwidth[2]. The excitation circuit and the measurement circuit can affect the accuracy of the measured impedance. Therefore improving this circuit is important in increasing the accuracy of the system for measuring an unknown impedance.

Fig. 1 shows the Howland current pump that can excite an unknown impedance. As it can be seen, the circuit non-idealties can cause error in the output current at different frequencies. Two considerations are important here. Selecting proper R_a defines the current generated by V_{in}. Selecting suitable op amps with appropriate bandwidth insures constant current over the frequency range of interest.

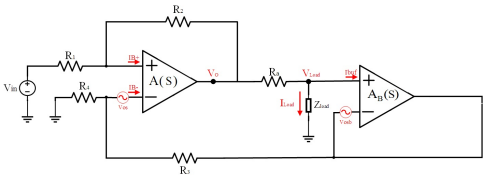


Figure 1: Howland circuit with modeled nonidealities

$$I_{load} \approx \frac{1}{\frac{2\times(Z_{load}+R_a)}{A(s)} + R_a} \times V_{in} \quad (1)$$

$$R_a \approx \frac{A(s) \times V_{in}}{(2 + A(s)) \times I_{load}} \quad (2)$$

(1) and (2) show the derived equations for selecting the proper R_a for a given current. This is important for applications that measure the impedance of the tissue in the body. The IEC60601 standard strictly limits the current that is allowed to be applied to the body at a certain frequency. Fig. 2 shows the prototype of the Howland circuit which excites the equivalent model of human body in order to measure the impedance. Tests were performed over impedance range of $1K\Omega$ to $100K\Omega$ and the circuit could generate $10\mu A_{rms}$ to $100\mu A_{rms}$ over the frequency range of 1KHz to 5MHz. The test results show that the circuit can produce the required current with less than 5% error with the required amplitude.

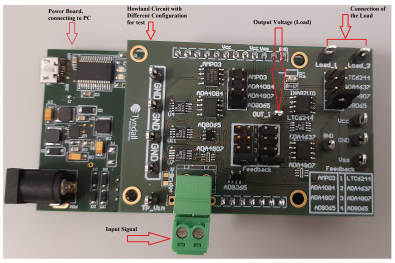


Figure 2: Manufactuerd PCB board of the Howland Circuit

References

[1] S. Khalil, M. Mohktar, and F. Ibrahim. The theory and fundamentals of bioimpedance analysis in clinical status monitoring and diagnosis of diseases. *Sensors*, 14(6):10895–10928, June 2014.

[2] A. Mahnam, H. Yazdanian, and M. M. Samani. Comprehensive study of howland circuit with non-ideal components to design high performance current pumps. *Measurement*, 82:94–104, March 2016.

Annotating Library Data based on Existing Knowledge Graphs

Daniela Oliveira
Data Science Institute, Insight Centre for Data Analytics, NUI Galway
daniela.oliveira@insight-centre.org

1. Introduction

Interoperability between library documents is achieved when the metadata is expressed by common vocabularies or equivalent conceptual models. Besides these well-known data models [1] several libraries develop their data model that can be based on several ontologies, leading to complex schemas that prove challenging to integrate.

Considering the case of a library that wants to provide integration with other resources, the integration task can prove to be challenging since there are many models and different ways to model the data in those models. We propose the develop a framework that, using existing Knowledge Graphs (KG) in the domain, facilitates the task of annotating structured and semi-structured data with a model that is interoperable with existing KGs.

2. Methods

The framework workflow has two main stages that are described in the following sections.

2.1 Build Stage

The workflow starts with the Build Stage, in which the KG is constructed from several existing KGs. This KG contains the data and implicitly contains the information about the ontologies used in each data model.

The triples in the data are parsed into documents that can be indexed by ElasticSearch[1] (ES) and the namespaces included in each KG are extracted.

These ontologies are loaded into a graph using graph-tool [3]. graph-tool is a Python module for the efficient manipulation of graphs. This graph is enriched with edges from direct mappings, obtained with ontology matching, and indirect ontology mappings, extracted from data relationships.

2.2 Annotation Stage

The main goal of the Annotation Stage is to take a dataset and integrate it with existing KGs in the same domain. These KGs were connected in the Build Stage to facilitate the annotation and interoperability between datasets.

The Annotation Stage starts with parsing the dataset from structured and semi-structured data into a common structure that is based on JSON documents. The entities and properties of these documents are matched with documents in the ES indices, and, finally, the candidates are ranked according to a scoring metric and a score confidence metric.

[1] https://www.elastic.co/what-is/elasticsearch

Library	Works	Work Types	Languages
Hardiman	1 561 167	11	532
British	8 125 515	4	216
French	133 172	1	216
German	19 710 630	11	355
Gutenberg	60 671	7	67
Portuguese	1 203 636	4	499
Spanish	4 591 742	15	387

Table 1: Characterisation of entities in the libraries.

2.3 Experimental Data

We chose six international libraries to serve as background knowledge: the British Library, the National Library of France, the German National Library, Project Gutenberg library, the National Library of Portugal, and the National Library of Spain.

The experimental data we used to test the annotation framework was provided by the Hardiman Library, in NUI Galway[2]. This data was provided in an RDF/XML that follows the MARC bibliographic conversion specifications since it was automatically generated by a marc2bibframe framework. Table 1 show the number of works of each library, the number of unique types of works, and the number of languages of each one of the chosen libraries.

2.4 Evaluation

The evaluation of the framework is two-fold: local evaluation and overall contribution of the framework. The local evaluation shows the performance of the overall annotations of instances, entities, and properties. The overall contribution showcases the impact of the methods proposed in improving the confidence of proposed annotations.

3 Conclusions and Future Work

At this moment, we still do not have the full results and evaluation to present but preliminary results of the graph building stage can be found in [2].

References

[1] BIBFRAME - Bibliographic Framework Initiative (Library of Congress).

[2] D. Oliveira, R. Sahay, and M. d'Aquin. Leveraging Ontologies for Knowledge Graph Schemas. In *Knowledge Graph Building Worshop at ESWC*, page 12, Portoroz, Slovenia, 2019.

[3] T. P. Peixoto. The graph-tool python library, 2017. type: dataset.

[2] https://library.nuigalway.ie

FedS: Distributed Path Query Over Linked Data

Qaiser Mehmood and Mathieu D'Aquin
Insight Centre for Data Analytics, NUIG
qaiser.mehmood@insight-centre.org

1. Introduction

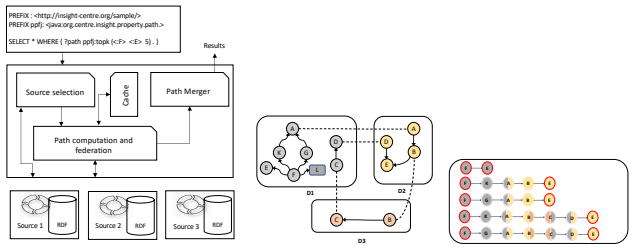
Massive amount of Linked Data is omnipresent in the form of Resource Description Format (RDF). An expressive way of querying this data is declarative queries, where a fundamental paradigm is called the *path query*. SPARQL1.1. provides a feature called Property Paths (PPs). Using PPs a user can check the existence of paths between two entities. However, PPs queries can only be executed against a single graph and has limitations such as; it can not enumerate the paths, no shortest pathfinding mechanism, and has worst query performance even for small amount of data. In our previous work [3], we addressed these limitations and proposed an extension, for PPs, that works for a single graph in a centralized way. However, as Linked Data by nature is distributed over the cloud. Thus, centralizing the data is not a feasible and prominent solution. Consequently, we extend our work and propose a distribute pathfinding approach. We evaluate our approach against state-of-the-art work.

2. Background and Motivation

It is very common in the biomedical domain that two biological entities (gene, protein, pathway, drug, etc.) are associated via several properties or paths. However, these entities may exist across different datasets. Centralizing this distributed data poses some challenges and drawbacks such as, (i) merging distributed data into a single graph is a tedious task, (ii) copied data may miss the opportunity to query the up-to-date data. While on the other-side, distributed query processing provides an opportunity to tackle these challenges. In the context of distributed path queries, some initiative [2, 4, 1] have been taken. However, some of these approaches (e.g., [2]) requires a pre-computed up-to-date index. Hence, path completeness is only assured if the index is up-to-date according to the current status of the underlying distributed RDF datasets. While the other [4] does not support SAPRQL1.1. Property Paths and requires sophisticated data fragmentation. Our objective in this work is to propose an index-free approach which guarantees the up-to-date path retrievals.

3. Approach

We implemented a *cache assisted* path query federation engine called FedS, and a shared algorithm that runs over the distributed datasources. FedS has three components i.e., (i)*source selection:*selects the relevant datasources, (ii) *path computation and federation:* assisted with *cache*, it computes the paths and delegates the path subqueries where required, (iii) *path merger:* evaluates the partial paths and merges those into complete paths. In summary, posing a path query, the FedS engine delegates the requests to the datasources. The shared algorithm, running on remote datasets, computes the paths (full or partial) against each query request and returns the answers back to engine. FedS engine, on receiving these answers, intelligently computes and generates the complete paths and results are presented to the user. Figure 1 depicts the architecture, a sample of distributed datasets and paths calculated by FedS engine.

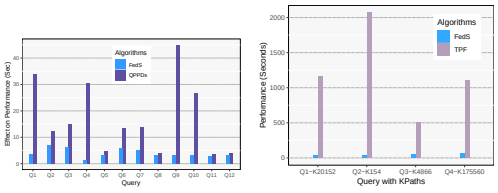


(a) Architecture (b) Datasets (c) Calculated paths

Figure 1: System architecture along sample distributed datasets and calculated paths

4. Results

We experimented with *real-world* and *synthatic* data. Figure 2 depicts the query performance of FedS against QPPDs and Triple Pattern Fragments (TPF). We can see that FedS outperformed the other systems.



(a) FedS vs QPPDs (b) FedS vs TPF

Figure 2: System architecture along sample distributed datasets and calculated paths

References

[1] L. De Vocht, R. Verborgh, and E. Mannens. Using triple pattern fragments to enable streaming of top-k shortest paths via the web. Springer.

[2] Q. Mehmood, M. Saleem, R. Sahay, A.-C. N. Ngomo, and M. D'Aquin. Qppds: Querying property paths over distributed rdf datasets. *IEEE Access*, 2019.

[3] V. Savenkov, Q. Mehmood, J. Umbrich, and A. Polleres. Counting to k or how sparql1. 1 property paths can be extended to top-k path queries. In *SEMANTiCS*. ACM, 2017.

[4] X. Wang, J. Wang, and X. Zhang. Efficient distributed regular path queries on rdf graphs using partial evaluation. In *25th ACM CIKM*. ACM, 2016.

Model RDBMS to Enterprise Knowledge Graph

Mani Vegupatti Selvanathan
Insight Centre for Data Analytics, National University of Ireland Galway
mani.vegupatti@insight-centre.org

1. Introduction

Enterprises generate large amounts of data from day-to-day operations. The competitiveness of enterprises is often determined by how quickly they can analyze the generated data to support their decision-making process by either a proactive or a reactive approach. In enterprises, the decision-making process is augmented by business intelligence solutions with capabilities of generic, predictive and prescriptive analytics.

Business intelligence solutions are normally based on predefined data models and structured reports. When a new type of analysis is to be carried out, the modeling and generation of reports can take several months and by the time reports are available, the information is not required anymore or has lost its business value.

A system, which can provide facts and details of the business entities or concepts without having to define structured reports would enable high competitive advantage to enterprises. To provide such a system, the semantics of underlying enterprise data should be well defined in the system by the enterprises.

The framework of Semantic Web as defined by the World Wide Web Consortium [1] (W3C) provides methodologies such as Linked Data, Vocabularies, and Query Language to define the semantics of the data. Using this framework, we can define a graph which consists of concepts, entities, and relations among them, which is commonly known as Knowledge Graph, a term coined by Google [2].

The overall research work is to provide a framework to define an Enterprise Knowledge Graph(EKG)-based Business Intelligence System. This consists of multiple tasks namely, 1) Extract, Transform and Load (ETL) data from a) structured data such as Relational Data Bases (RDB), b) unstructured data such as documents, 2) parse the information request by the user, 3) query the knowledge graph, and 4) visualize the query results. In this paper, we will discuss the ETL stage on RDB data, which represents the major portion of the data from enterprises to include in an EKG

2. Problem Statement

We have general domain systems like Google Search, which can provide details and facts on a real-world entity or concept based on Knowledge Graph [2]. However, the EKG is essentially different in some aspects:

1. The details and facts needed for an entity or concept is not a generic information retrieval problem, but rather involves domain-specific knowledge and complex calculations based on multiple relations
2. The majority of facts are temporal and continuously changing in nature. We need multiple copies of the same facts based on the timestamp and also the latest value of that instance of time.
3. The facts need to be analysed based on multidimensional aspects, to allow reasoning for aggregation or disaggregation using a hierarchy

[1] https://www.w3.org/

[2] https://developers.google.com/knowledge-graph

3. Previous Work

The framework of Linked Data with RDF for representing the knowledge as a triple of subject, predicate and object was released by W3C and below are some major work,

1. A direct mapping of relational data to RDF by W3C [1] which provides methods to map all the tables of a given database into RDF using the relational schema.
2. R2RML: RDB to RDF Mapping Language by W3C [3] which allows mappings to be customized between schema and RDF definition.
3. A framework to use the RDB data directly in Semantic Web using a wrapper approach rather than replication[4].

4. Future Work and Contributions

Multi-layer RDF(MRDF) can fulfill the limited capabilities of above methods and address the enterprise data needs such as multidimensional modelling, complex calculations and temporal aspects. Our contributions will be,

1. A framework of MRDF to represent EKG
2. A language for ETL of RDB to the above MRDF

References

[1] M. Arenas, A. Bertails, E. Prud'hommeaux, and J. Sequeda. A direct mapping of relational data to rdf. *W3C recommendation*, 27:1–11, 2012.

[2] L. Bellomarini, D. Fakhoury, G. Gottlob, and E. Sallinger. Knowledge graphs and enterprise ai: the promise of an enabling technology. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pages 26–37. IEEE, 2019.

[3] S. Das, S. Sundara, and R. Cyganiak. R2rml: Rdb to rdf mapping language, w3c recommendation 27 september 2012. *Cambridge, MA: World Wide Web Consortium (W3C)(www. w3. org/TR/r2rml)*, 2012.

[4] J. F. Sequeda. *Integrating relational databases with the Semantic Web*, volume 22. IOS Press, 2016.

3D Object Detection for Instrumented Vehicles

Venkatesh Gurram Munirathnam, Suzanne Little, Noel E. O'Connor
Insight Centre for Data Analytics,Dublin City University
venkatesh.gurummunirathnam@insight-centre.org

1. Introduction

The data acquisition process in the context of instrumented vehicles (i.e. autonomous driving) captures a huge amount of data due to the increasing availability of sensory technology. The purpose of collecting such huge data is to develop human-like perception capable to comprehend the environmental perception, precise positioning and path planning of the objects appearing in the complex mixed-traffic scenario while driving on the road. Object detection is an important part of the perception system of the safe driving vehicles and most of the 3D object detection approaches in the literature employ spatial features extracted from Image and LiDAR (Light Detection and Ranging) data. These features are either considered individually or fused together to predict the 3D position of the object in the scene. To understanding the complex and dynamic environment the perception system of instrumented vehicles, it is expected to have component which can transform the sensory data into semantic information and capable of incorporate the temporal cues extracted from multi-modal data while estimating the region of interest(ROI) containing the objects. The current research work focus on designing and developing 3d object detection for perception system in instrumented vehicles.

2. Related Work

Some of the recent approaches in the literature using LiDAR and Image sensory data are considered here. MV3D[1] uses bird-eye and front view projection of lidar point along with RGB data and all the 3D box proposals are generated based on birds-eye-view(b.e.v) feature only. In AVOD[3], both modalities are fused, further merged, and then passed to the region proposal. In F-PointNet[4], Object classification and bounding box regression are performed on the point cloud with 2D object detection positions on the image are extrapolated to a 3D space. In Yolo4D [2], the sequence of stacked 3D LiDAR point cloud data is used to learn temporal information via LSTM layer.

3. Proposed Methodology

The pipeline depicted in Figure 1 has two branch one for processing image data and other for LiDAR data and is inspired by the various work from the recent literature. Two similar deep convolution layers (like VGG16, ResNet50) one for each modalities (LiDAR,Image) are used for extracting feature maps. These extracted feature are used to learn the temporal cues and by the region proposal network module which extracts the objectness and box proposal in respective sensory data. objectness information is used for refining the temporal region learning. Fusion of the both the modalities are done twice in the proposed framework, first using only the box proposals and later using 3D object region proposal along with temporal cues extracted on LiDAR and Image data. The feature map obtained during the post fusion is passed to the 3D box regression module for the final prediction of the object class and 3D box parameter estimation.

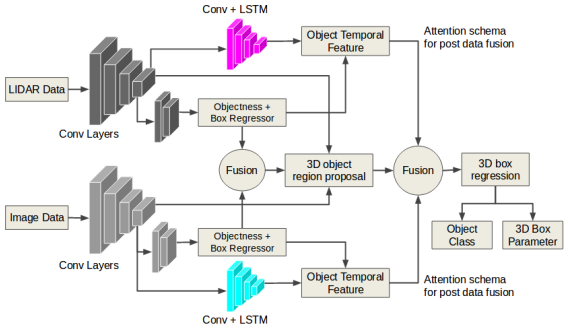


Figure 1: Proposed pipeline for 3D Object Detection

4. Ongoing and Future Work

Implementation of model using Only LiDAR data is in progress so no results are presented here. Next step in this research work will be to explore the strategies and efficacy of using the temporal information[2] of the objects in the mixed traffic environment from image and LiDAR data. Investigation of various schema[1] for fusing the temporal and multi-modal data during 3D box proposal. It would be interesting to investigate the possibility of exploiting the cross weights between representation of modalities and try to gradually learn interactions of the modalities in a deep network.

References

[1] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia. Multi-view 3d object detection network for autonomous driving. In *IEEE Conf. on CVPR*, pages 1907–1915, 2017.

[2] A. El Sallab, I. Sobh, M. Zidan, M. Zahran, and S. Abdelkarim. Yolo4d: A spatio-temporal approach for real-time multi-object detection and classification from lidar point clouds. 2018.

[3] J. Ku, M. Mozifian, J. Lee, A. Harakeh, and S. L. Waslander. Joint 3d proposal generation and object detection from view aggregation. In *2018 IEEE/RSJ Int. Conf. on IROS*, pages 1–8. IEEE, 2018.

[4] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas. Frustum pointnets for 3d object detection from rgb-d data. In *IEEE Conf. on CVPR*, pages 918–927, 2018.

Mapping Informal Settlements with Machine Learning

Agatha C. Hennigen de Mattos, Gavin McArdle and Michela Bertolotto
University College Dublin
agatha.hennigendemattos@ucdconnect.ie

1. Introduction

Nearly one-quarter of the world's urban population live in deprived areas under shocking and intolerable conditions. These areas, called informal settlements, are formally defined as households where the inhabitants suffer from one of the following deprivations: lack of water sources, sanitation, housing durability or security of tenure [1]. Slums have a similar definition, with additional lack of sufficient living area, and both terms will be used interchangeably in this article.

In this context, the United Nations ratified the Sustainable Development Goals (SDG) which ensure that slums are upgraded, and that adequate, safe and affordable housing and basic services are accessible for all by 2030. However, keeping track of the development and upgrade of slums through traditional data collection, such as census surveys, can be prohibitively expensive and institutionally difficult, as some governments may see little benefit in committing resources only to have their lacklustre performance officially documented. Figure 1 shows the latest information available at the SDG Tracker, which is dated from 2014 and aggregated at country level.

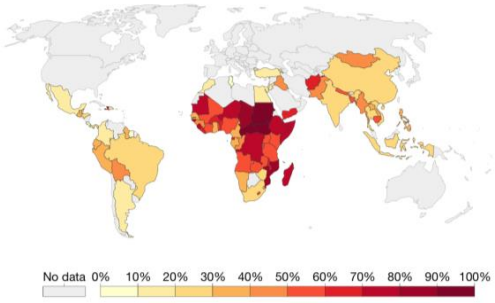


Figure 1: Share of urban population living in slums [2].

An alternative path to measuring the progress toward this goal is to use passively collected data, such as satellite imagery, to map these communities.

2. Remote Sensing Approaches to Mapping Informal Settlements

Recent reviews of remote sensing approaches using satellite imagery to map informal settlements reveal the different methods that have been used over the years, such as object-based image analysis and, more recently, machine learning [3], [4]. Yet, slums can be remarkably different from each other when seen from space, which causes algorithms trained in one location to be ill-suitable to identify settlements in other places. Additionally, some techniques require high resolution imagery which is costly to acquire and process. The absence of georeferenced and labelled data on the areas where slums are located is also an obstacle.

Despite these difficulties, progress has been made in the field and the latest work demonstrates machine learning techniques have the highest accuracies and could be promising in tackling this problem [3], [4].

3. The Aims of The Proposed Research

The proposed research seeks to advance the monitoring of the population living in informal settlements by achieving two objectives.

The first one is to develop an algorithm for mapping informal settlements using a georeferenced dataset of these communities made available by the Brazil's official statistics agency (IBGE). This dataset has a lot of important characteristics that make it well-suited for this research. First, the areas are already georeferenced and labelled. Second, to our best knowledge, it has not been used before. Thirdly, it has data on many cities, which would allow for the model to be trained in different urban contexts, something past reviewers have shown to be limited in current studies and that would, consequently, contribute to the advancement of the state-of-the-art on the topic. Additionally, the research will also test algorithms developed for other cities/countries, if they are accessible, to the above-mentioned dataset; and examine the differences between the official dataset and volunteered geographic information sources in terms of number of and total area measurement of informal settlements.

The second objective is to conduct a spatiotemporal analysis of the expansion of these areas and, with other auxiliary data, gain insight into the mechanisms that drive the development of these settlements.

4. References

[1] U. N. Habitat, 'Tracking progress towards inclusive, safe, resilient and sustainable cities and human settlements', Jul. 2018.

[2] 'Goal 11: Sustainable Cities and Communities - SDG Tracker', *Our World in Data*. [Online]. Available: https://sdg-tracker.org/cities. [Accessed: 03-Feb-2020].

[3] M. Kuffer, K. Pfeffer, and R. Sliuzas, 'Slums from Space—15 Years of Slum Mapping Using Remote Sensing', *Remote Sens.*, vol. 8, no. 6, p. 455, Jun. 2016, doi: 10.3390/rs8060455.

[4] R. Mahabir, A. Croitoru, A. T. Crooks, P. Agouris, and A. Stefanidis, 'A Critical Review of High and Very High-Resolution Remote Sensing Approaches for Detecting and Mapping Slums: Trends, Challenges and Emerging Opportunities', *Urban Sci.*, vol. 2, no. 1, p. 8, Mar. 2018, doi: 10.3390/urbansci2010008.

INSIGHT@DCU team in TRECVID-VTT 2019

Luis Lebron
Insight@DCU
luis.lebroncasas@insight-centre.org

1. Motivation

Video captioning, which consists in extracting a sentence or paragraph to describe a video, is a very challenging task in the computer vision field that has witnessed a renewed interest in the community since the arrival of deep learning. From this new set of techniques, it has become possible to generate more complex sentences. The application of these techniques is multiple from a textual description of movies for hearing-impaired to the automatic generation of guideline from example videos on the industry.

2 Problem Statement

There is a high level of complexity in capturing the fine details in a video. In this problem, we are looking for hours of videos where different actions are happening and to suppress the bias of the human we also need multiple annotations for the same clips or videos. Although this is a difficult constrain more and more dataset are being created. In terms of metric, most of the current metrics come from the text translation task and only evaluates the predicted caption versus the one provided by humans. These may be not the best approach as two fairly different captions could be describing the same scene from different points of views or levels of detail.

3 Related Work

Nowadays, deep learning has provided a new range of techniques to solve the video caption problem becoming the most used technique. Initial approaches to these techniques use the encoder-decoder structure to generate single sentences which describe the whole video. Datasets like the ones generated for TRECVID's video-to-text task [1] have helped to provide data to train these models.

With the emergence of more complex datasets like ActivityNet, these methods became the start of a new trend of algorithms that focus on not only generating a single sentence but on producing multiple sentences to explain multiple clips in the videos.

4. Proposed Solution

One our first look at the problem, we decided to focus on the data and how we can work with it in a simple scenario. To test our model and see its performance in a real case, we participated in the video to text task in TRECVID [1]. Our baseline model comes from [2] which uses LSTMs and an attention model. From there we explore the use of different techniques to improve the baseline results like replacing the LSTM for the Transforms. Our main idea was to include text extract from an image captioning module as input to provide a base description. However, only the use of more data and better embedding report better results in the metrics than these improvements.

Table 1: 3-top results and Insight@DCU submission in TRECVID-VTT 2019, reporting four of the commonly use automatic metric for this task.

	CIDEr	CIDEr-D	BLEU4	Meteor
RUC_AIM3	0.585	0.332	0.063	0.308
UTS_ISAUTS_ISA	0.496	0.243	0.043	0.286
FDU	0.428	0.180	0.027	0.243
Insight_DCU	0.035	0.015	0.004	0.141

5. Evaluation

For this first round of the experiment, we have work on the TRECVID-VTT task and got some qualitative and quantitative results. Although we didn't achieve one of the top scores (table 1), it helps us to identify the weakness of our models compare to the others. Our first mistake was to use BLEU as the references metric, other teams where using CIDEr for it and a reward base loss function base on CIDEr. Also, we only extend the number of samples with only one additional dataset while others use multiples of them.

6. Future work

On the future we plan to look on the automatic generation of captions, the use of more fine detailed information as a feature and the creation of a new pair of the loss function and metric to train deep learning models.

7 Acknowledgement

This research was supported by the Irish Research Council Enterprise Partnership Scheme together with United Technologies Research Center Ireland and Insight@DCU.

References

[1] G. Awad, A. Butt, K. Curtis, Y. Lee, J. Fiscus, A. Godil, A. Delgado, A. F. Smeaton, Y. Graham, W. Kraaij, and G. Quénot. Trecvid 2019: An evaluation campaign to benchmark video activity detection, video captioning and matching, and video search & retrieval. In *Proceedings of TRECVID 2019*. NIST, USA, 2019.

[2] Á. Peris, M. Bolaños, P. Radeva, and F. Casacuberta. Video description using bidirectional recurrent neural networks. In *International Conference on Artificial Neural Networks*, pages 3–11. Springer, 2016.

News for children

Enric Moreu
Insight Centre for Data Analytics, Dublin City University
enric.moreu@insight-centre.org

1. Motivation

We live in the Internet era, the information is easily accessible by anyone with a computer, smartphone or tablet. Everybody can produce, consume and share contents on the Internet while children spend large hours exposed to the internet, playing games and chatting, but also interacting with the media contents.

In that context, fake news and misinformation emerged, reaching epidemic proportions worldwide. Those malicious news manipulate the public opinion, and specially kids, that are really vulnerable because of their poor Media Literacy [4].

2. Proposed Solution

The proposed solution is re-use the adult's news to generate an adapted "cartoonized" version for kids using AI. The AI-generated broadcast will introduce the kids to news, while improving their Media Literacy skills, hopefully helping them to identify fake news in the future. Contents will be more dynamic and stimulating than the adults news, plus they will be curated by professionals. Thanks to the new AI generative techniques the style of the cartoon can be adapted to the children age group by modifying between reality and cartoon style.

Two approaches are presented to solve the challenge of cartoon generation: style transfer and 3D animation.

2.1. Style Transfer

The Style Transfer [3] technique apply the style of an image (source) into another (target) by using a two-way loss that regulates the information about how the content and the style are transferred to the output image.

Figure 1: Video Style Transfer

It uses two variables to emphasize the content or the style:

$$Ł_{total}(\vec{p}, \vec{a}, \vec{x}) = \alpha Ł_{content}(\vec{p}, \vec{x}) + \beta Ł_{style}(\vec{a}, \vec{x})$$

When applying Style Transfer to a video optical flow can be used to decide if a part of an image should be updated as Manuel Ruder [5] described. This prevents a flickering effect on the background.

2.2. 3D animation

The second approach to create cartoons is animating a 3D character that mimics the anchor using Blender [1] and OpenPose [2].

Figure 2: Pose to 3D character

Blender is used to render a 3D model rigged with a structure of bones that follow the body keypoints extracted using OpenPose. Some small bones like the hands are affected by undesired rotations because the keypoints only indicate the location of each bone.

3. Future Work

While both approaches look promising there are many problems that need to be solved on the hand, face and background of the scene. The combination of Style Transfer plus 3D animation in the same pipeline can generate very promising results.

References

[1] Blender Online Community. Blender - a 3d modelling and rendering package,
[2] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In *arXiv preprint arXiv:1812.08008*, 2018.
[3] L. A. Gatys, A. S. Ecker, and M. Bethge. A neural algorithm of artistic style. *CoRR*, abs/1508.06576, 2015.
[4] S. Livingstone. Developing social media literacy: How children learn to interpret risky opportunities on social network sites. *Communications*, 39(3):283–303, 2014.
[5] M. Ruder, A. Dosovitskiy, and T. Brox. Artistic style transfer for videos. *CoRR*, abs/1604.08610, 2016.

Rule Based Approach for Facial Expression Recognition in the Wild

Alex Acquier
Insight Centre for Data Analytics, NUI Galway
alex.acquier@insight-centre.org

1. Abstract

This paper is looking into developing an Automatic Facial Expression Recognition (AFER) system from a multimodal (2D + 3D) video feed. This project is undertaken using the Fast Action Coding System (FACS) principles which allow to develop an open AFER system given that it is easily extensible with diverse facial expressions in the wild.

2. Introduction

Open AFER systems could help management to improve productivity by estimating the moral of staff, event manager to estimate the real time success of an event or even marketer will be able to adapt the advertising of a live billboard in function of the person standing in front of it. There are presently only few real-time open AFER systems and the one which are available use machine learning classifiers which require large datasets which often contain posed expressions to train.

The system proposed in this paper aims to use the FACS principles to develop a new real time open AFER system. FACS uses the contraction of the face's muscles to generate what is called action unit (AUs) which correspond to different changes in distinct areas of the face. Different combinations of those AUs reflect different facial expressions and the scoring can also weigh the intensity of those AUs which means that the detection of micro expressions may be possible. This approach also has the advantage of requiring little to no data for training and to be usable in the real life situations.

3. Related Work

Upon reviewing the state of the art for facial expression recognition, two conclusions can be drawn: a large annotated datasets is needed to train a machine learning classifiers and there is no annotated facial expressions database in RGB-D either posed or in the wild.

FACS taxonomizes the face muscles movement by encoding the slight differences in the facial appearance and has been used for decades by psychologists, animators but also some modified version has been used on primates. Using a combination of different AUs which are determined through the change of the face, the facial expressions as well as their intensities can be estimated. This method offers the advantage to not require any datasets for the classifiers' training, to be have the potential to work in the wild and can be used in open system.

The system presented in [1] uses 2D + 3D data, FACS, is real-time but is not an open system and recognizes only four facial expressions (disgust, happiness, sadness and surprise) while six facial expressions can be recognized (the one cited above plus fear and anger). This approach is going to be the base for the open system that the author wants to develop: it should be real time and be able to recognize the six basic emotions in a real life environment using both 2D and 3D data.

4. Project state of completion

4.1 Completed task

- Face detection in 2D feed and acquisition of data
- Design of a lightweight landmark scheme
- Design and application of a series of filter banks
- 2D face alignment
- Experiments on the state of the art with [2] and [3] using both posed and in the wild datasets

4.2 Ongoing work

- Implementation of the landmark scheme
- Alignment of the 3D data
- Experiments on the state of the art with [4]

4.3 Future Work

- Coding of the different AUs for AFER
- Generation of datasets for calibration
- Experiment in real life conditions

References

[1] F. Tsalakanidou and S. Malassiotis, "Real-time 2D + 3D facial action and expression recognition", Pattern Recognition, Volume 45, Issue 5, pp 1763 – 1775, 2010

[2] S. Alizadeh and A. Fazel, "Convolutional Neural Network for Facial Expresion Recognition", https://arxiv.org/pdf/1704.06756.pdf, 2017

[3] J. Cai, Z.Meng, A.S. Khan, Z. Li, J. O'Reilly, Y. Tong, "Island Loss Learning for Discriminative Features in Facial Expression Recognition", https://arxiv.org/abs/1710.03144, 2017

[4] D. Acharya, Z.Huang, D.P. Paudel and L. Van Gool, "Covariance Pooling for Facial Expression Recognition", Conference on Computer Vision and Pattern Recognition Workshops, 2018.

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289.

A Multi-objective Supplier Selection Framework based on User-Preferences

Federico Toffano
Insight Centre for Data Analytics, University College Cork
federico.toffano@insight-centre.org

1. Introduction

We propose a novel resolution approach to the problem of selecting a set of suppliers in order to satisfy an input expected demand for a specific set of components, where different evaluation criteria need to be considered. More precisely, a solution to our problem is defined as a mapping from each component i with expected demand d_i to a set of suppliers S_i such that $\sum_{\forall j \in S_i} q_{i,j} \geq d_i$, where $q_{i,j} \geq 0$ is the quantity of component i to be ordered from supplier j. Let $\mathcal{X}$ be the state space of the solutions. To evaluate a solution $x \in \mathcal{X}$ we consider four different criteria or evaluators $f_k : \mathcal{X} \rightarrow I\!R$, with $k \in [1, 4]$. The first criterion f_1 is *cost* that includes the cost of ordering the components required and the cost of establishing business relationships. The second and third criteria f_2 and f_3 are *delay* and *lead time* that are computed as the worst case of expected delay and lead time of the selected suppliers of the corresponding solution. The last criterion f_4 is the worst *supplier reputation* among the selected suppliers. Our purpose to minimize cost, delay and lead time, and maximize the reputation.

2. User preferences and MILP model

In multi-criteria decision problems, it is often the case that the Pareto frontier (i.e. the set of undominated solutions) is huge and practically unfeasible to be computed or to be evaluated by a human decision maker. To deal with this kind of problems, it is common practice to define a scalar utility function u such that given two solutions x_1 and x_2, $u(x_1) \geq u(x_2)$ if and only if $x_1 \succcurlyeq x_2$, where $\succcurlyeq$ reads "is at least as good as" [1].

Let $\mathcal{W}_0 = \{w \in I\!R^4 : \sum_{k=1}^{4} w_k = 1, w_k \geq 0, \forall k = 1, \ldots, 4\}$ be the set of weights vectors representing the user preferences. We suppose that a user has an associated unknown weight vector w. Intuitively, each weight w_k represents the importance that the user gives to the corresponding criteria f_k.

We suppose a form of *additive independence* between the criteria used to evaluate a solution x defining the scalar utility function for a user with preference w as $u_w(x) = \sum_{k=1}^{3} -w_k f_k(x) + w_4 f_4(x)$, where the evaluators f_k for all $k \in [1, 4]$ are linear functions w.r.t. some variables used to represent a solution. Given a fixed user preference w, such scalar utility function u_w will be used as objective function for a mixed integer linear programming (MILP) model where the set of constraints is defined according to the input data of the problem. The solution x^* that maximize u_w will then be an optimal solution w.r.t. the user preference w.

3. Preference elictiation

In general, asking the decision maker to define its preference w is liable to be a difficult and error-prone task. Hence, the framework uses an alternative approach based on learning information about the real user preference by interacting directly with the user.

An input user preference $x_1 \succcurlyeq x_2$ can be translated into a linear the constraint $u_w(x_1) - u_w(x_2) \geq 0$ that can be added to the user preference state space $\mathcal{W}_0$ reducing the set of possible weights vector. Let V_Λ be a convex polyhedron in $I\!R^4$ defined by a set of non-strict linear inequalities Λ associated to input user preferences; we define $\mathcal{W}_\Lambda$ as the convex and closed polytope $\mathcal{W}_\Lambda = \mathcal{W}_0 \cap V_\Lambda$. Our idea is to iteratively ask to the user her preference between two solutions in order to reduce the initial user preference state space $\mathcal{W}_0$ to a smaller user preference state space $\mathcal{W}_\Lambda$ until we can recommend a solution with worst case loss below a certain threshold. Briefly, at each iteration of our framework, we compute the set of solutions $\mathcal{Y}$ associated to the extreme points of the current polytope $\mathcal{W}_\Lambda$, and then we select the query from $\mathcal{Y}$ trying to maximize the information gain that we can get from the user answer.

4 Results and conclusions

We tested our method on randomly databases simulating a real-world scenario. From our experimental results, it looks like that the number of queries asked to the user scale very well w.r.t. the complexity of the problem converging to a solution with 9 queries in average. A drawback of our method is the high sensitivity to incorrect answers w.r.t. the real user preference. This issue could be addressed by a Bayesian representation of the user preferences at the expense of an increased complexity in terms of total number of queries and computational time. Regarding the query selection method, we proposed a new approach and we compared it with the state of the art getting better results in terms of query computation time and roughly the same results in terms of information gain.

5 Acknowledgements

This work has been done in collaboration with Michele Garaffa (UTRC, Cork) and Nic Wilson (Insight Centre for Data Analytics, University College Cork)

References

[1] R. L. Keeney and H. Raiffa. *Decisions with multiple objectives: preferences and value trade-offs*. Cambridge university press, 1993.

Handling Noisy Constraints in Semi-supervised Overlapping Community Finding

Elham Alghamdi, Ellen Rushe, Mehran H.Z. Bazargani, Brian Mac Namee, and Derek Greene
Insight Centre for Data Analytics, elham.alghamdi@insight-centre.org

1. Introduction

Community structure is an essential property that helps us to understand the nature of complex networks. Since algorithms for detecting communities are unsupervised in nature, they can fail to uncover useful groupings, particularly when the underlying communities in a network are highly overlapping [1]. Recent work has sought to address this via semi-supervised learning [2], using a human annotator or "oracle" to provide limited supervision. This knowledge is typically encoded in the form of must-link and cannot-link constraints, which indicate that a pair of nodes should always be or should never be assigned to the same community. In this way, we can uncover communities which are otherwise difficult to identify via unsupervised techniques.

However, in real semi-supervised learning applications, human supervision may be unreliable or "noisy", relying on subjective decision making [3]. Annotators can disagree with one another, they might only have limited knowledge of a domain, or they might simply complete a labeling task incorrectly due to the burden of annotation. Thus, we might reasonably expect that the pairwise constraints used in a real semi-supervised community detection task could be imperfect or conflicting. The aim of this study is to explore the effect of noisy, incorrectly-labeled constraints on the performance of semi-supervised community finding algorithms for overlapping networks. Furthermore, we propose an approach to mitigate such cases in real-world network analysis tasks. We treat noisy pairwise constraints as outliers, and use outlier detection approaches such as autoencoder, a commonly-used method in the domain of anomaly detection, to identify such constraints.

2. Methods and Experimental Design

The key aspect of our work is an iterative approach using autoencoder to remove noisy pairwise constraints selected by the AC-SLPA algorithm [2]. An *autoencoder* (AE) refers to a neural network architecture that attempts to reconstruct a given input in an effort to learn an informative latent feature representation. Formally, for an input vector x, we attempt to map x to a reconstruction of itself x'. By doing this, a latent representation of the data is created in the hidden layer(s) of the network [4]. These networks can utilize a "bottleneck" configuration where the hidden layer(s) of the network compress the data [4]. The network is trained by minimizing the mean squared error (MSE) between the reconstruction and input. In our work we employ the above neural network architecture to identify potentially noisy pairwise constraints selected by AC-SLPA before applying the community detection process, see figure 1 for details.

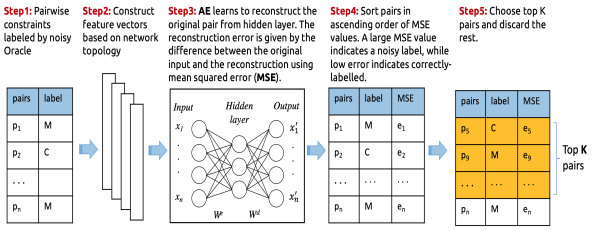


Figure 1: The process of detecting noisy constraints using autoencoder

3. Evaluation

The performance of both proposed methods is evaluated by running experiments on two groups of synthetic benchmark networks containing overlapping communities. The depth of the autoencoder is varied to assess its effect on performance. All models were trained with a learning rate of 103 for a maximum of 100 epochs and a batch size of 256. AUC score over the resulted MSE errors are calculated, which provides an estimate of the number of constraints that were successfully detected in the absence of a definitive threshold.

4. Conclusion

Based on extensive experiments, all AE models show promising results, high AUC scores with the lowest scores mostly around 70%. A second set of experiments are in progress, which include exploring other outlier detection approaches and clustering-based algorithms in this context. As a next step, we evaluate the performance of AC-SLPA when incorporating reliable constraints as selected by the best approach for detecting noisy constraints.

References

[1] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann. Link communities reveal multiscale complexity in networks. *Nature*, 466(7307):761–764, 2010.

[2] E. Alghamdi and D. Greene. Active semi-supervised overlapping community finding with pairwise constraints. *Applied Network Science*, 4, 2019.

[3] M. R. Amini and P. Gallinari. Semi-supervised learning with an imperfect supervisor. *Knowledge and Information Systems*, 8(4):385–413, 2005.

[4] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016.

Impact of Segment Duration on DASH-based on-Demand Streaming in Wireless Environment

Abid Yaqoob
Dublin City University, Ireland
abid.yaqoob@insight-centre.org

Gabriel-Miro Muntean
Dublin City University, Ireland
gabriel.muntean@dcu.ie

1. Abstract

The influence of segment duration on DASH-based on-demand streaming plays a critical role. The configuration of segment duration is important to achieve efficient adaptive playback experience. In this paper, we investigate the impact of segment duration on the DASH-based streaming system. The experimental evaluation reveals that the performance of DASH is optimum with (2-6) seconds segment duration.

2. Introduction

Multimedia streaming has become the most popular application due to the recent advancements in networking and computing technologies. Video delivery adaptation helps to improve video quality by dealing with different objectives which include quality, segment size, and load balancing, etc. on mobile, wireless and wired access networks [1]. Built on top of TCP and HTTP, DASH client uses the standard HTTP protocol to retrieve the video chunks, it maintains the playback session state to minimize the network load on the server-side. In this paper, we investigate the impact of segment duration on video bitrate adaptation algorithms in fluctuating wireless network environment.

3. Experimental Evaluation

NS-3.26[1] was used for simulation. The DASH server provides various encoded representations, i.e., 45, 363, 791, 1033, 1647, 2134, 2484, 3527, 3840, and 4220 Kbps, of the Big Buck Bunny[2] video sequence. The representations are segmented into different lengths, i.e., 1s, 2s, 4s, 6s, and 10s. The bandwidth link between the server and the wireless access point changes from 2 Mbps to 3 Mbps after the first 50s, then it switches repeatedly between 3 Mbps and 1 Mbps after every 50s until the end of simulation at 300 seconds. We evaluated our proposed Throughput and Buffer Occupancy-based Adaptation (TBOA) [1] algorithm against FDASH [2] and SFTM [3] rate adaptation algorithms. The performance is evaluated in terms of average video bitrate and the number of video bitrate fluctuations.

Fig. 1 illustrates the adaptive behavior of three clients under different segment lengths. TBOA provides high-quality video bitrates for all segment durations. TBOA and FDASH achieve nearly 1.67 Mbps and 1.48 Mbps for 4s segment duration, respectively. For all other segment lengths, TBOA achieves nearly the same video bitrate. FDASH achieves little improved video bitrate for longer segment duration and undergoes the lowest bitrate fluctuations for 10s segment duration. SFTM provides the highest video bitrate for the 2s segment duration. For longer segment duration, SFTM inaccurately estimates the available bandwidth and achieves low bitrates with a reduced number of bitrate fluctuations.

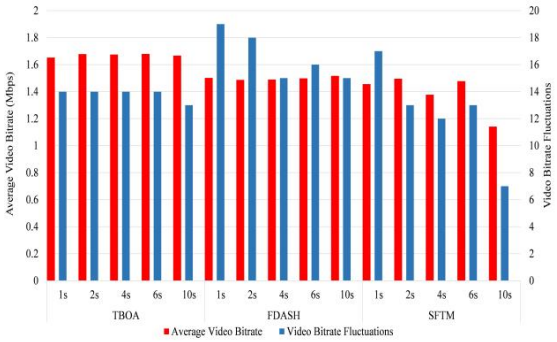


Figure 1: Adaptive behavior of the clients

4. Conclusion

This paper investigates the impact of segment duration on DASH-based on-demand streaming in a wireless network environment. From simulation results, we found that the longer segment duration does not significantly improve video quality. However, longer segment duration leads to the lower number of both the video bitrate fluctuations and the number of HTTP requests. On the other hand, a shorter segment duration reduces the data required to initiate the video playback.

References

[1] A. Yaqoob, T. Bi, and G. Muntean, "A dash-based efficient throughput and buffer occupancy-based adaptation algorithm for smooth multimedia streaming," in *2019 15th International Wireless Communications Mobile Computing Conference (IWCMC)*, June 2019, pp. 643–649.

[2] D. J. Vergados, A. Michalas, A. Sgora, D. D. Vergados, and P. Chatzimisios, "FDASH: A fuzzy-based MPEG/DASH adaptation algorithm," *IEEE Systems Journal*, vol. 10, no. 2, pp. 859–868, 2016.

[3] C. Liu, I. Bouazizi, M. M. Hannuksela, and M. Gabbouj, "Rate adaptation for dynamic adaptive streaming over HTTP in content distribution network," *Signal Processing: Image Communication*, vol. 27, no. 4, pp. 288–311, 2012

[1] https://www.nsnam.org/releases/ns-3-26/
[2] https://peach.blender.org/

An Energy-Accuracy-Throughput Aware Scheduling for DNN-based Real-time Multimedia Event Processing Systems

Felipe Arruda Pontes
Insight Centre for Data Analytics, National University of Ireland, Galway
felipe.arruda.pontes@insight-centre.org

1. Introduction

Because of recent advances on BigData we have seen an increasing amount of application from different fields being deployed in cloud environments, specially for programs that require a immediate response to events analysed from constant and large streams of video, usually through the use of Multimedia Event Processing systems . At the same time, the number of Multimedia Internet of Things (MIoT) devices, has increased dramatically [4]. Such devices are good candidates for inputting and processing large quantities of multimedia data in a stream-like fashion, and because of their mobility and portability they are often seen as good solutions in distributed scenarios and were there are continuously changing working spaces [2].

On another note, Deep Neural Network(DNN)-based solutions have been shown to generate state-of-the-art performance in many complex machine learning and computer vision problems, such Object Detection, Action Detection, and etc... Yet, with DNN-based solutions good results there comes also a great price in terms of computation resources, such as Memory, CPU, and GPU. This in turns becomes a limitation when working with a mixed Cloud-Edge infrastructure, were in some cases it might be OK to execute the stream analysis on a dedicated machine on the cloud, but not as much on a resource constrained device on the Edge [4].

This decision of were and when to run the data stream workloads is known as a scheduling problem. Yet, to the best of our knowledge, there are few approaches for scheduling workload of DNN-based real-time distributed event processing systems that takes into account the energy, accuracy and throughput requirements based on the users queries. This work will focus on the study, design and implementation of such scheduling plan, through the use of self-adaptative mechanisms, while aiming at maintaining good Quality of Service(QoS) metrics based on the user's queries requirements in terms of energy, accuracy and throughput.

2. Current Stage

In order to demonstrate and evaluate the scheduler solution, we had first to design and implement a fully working Distributed Multimedia Complex Event Processing framework as a basis. In it's current stage, the framework implements a event-based microservice architecture, as it is a distributed architecture that fits well with our problem [1], and it is running on top of containers engines.The framework basis as it is, can handle simple events queries, such as Object Detection, but still lacks the mechanisms for being self-adaptive. Figure 1 shows a representation of this framework.

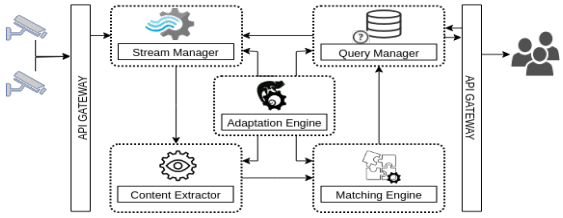


Figure 1: Current representation of our MCEP framework

3. Evaluation

As means of evaluation, we plan on executing benchmarks and comparison of the energy consumption, accuracy and throughput of the framework using the proposed scheduler on multiple Object Detection queries with different requirements, against a random scheduler, a statically defined one, as well comparing with results from related works. This way we will be able to show the results, and how much improvement the proposed solution may present.

4. Next Steps

For our next steps we plan on monitoring the energy consumption through the use of devices on for both cloud and edge scenarios, and implement a self-adaptive strategy, based on MAPE-k, but using immutable infrastructure and stateless services patterns in order to simplify the execution step of the adaptation in the framework. And finally we will identify and select the most promising scheduling solutions to archive the required QoS in our context, and perform the evaluations of the solutions experiments.

References

[1] P. Jamshidi, C. Pahl, N. C. Mendonça, J. Lewis, and S. Tilkov. Microservices: The journey so far and challenges ahead. *IEEE Software*, 35(3):24–35, 2018.

[2] J. Seo, S. Han, S. Lee, and H. Kim. Computer vision techniques for construction safety and health monitoring. *Advanced Engineering Informatics*, 29(2):239–251, 2015.

[3] H. Tann, S. Hashemi, and S. Reda. Flexible deep neural network processing. *arXiv preprint arXiv:1801.07353*, 2018.

[4] S. Teerapittayanon, B. McDanel, and H.-T. Kung. Distributed deep neural networks over the cloud, the edge and end devices. In *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*, pages 328–339. IEEE, 2017.

The Application of an Outcome-Representation Learning Model for Characterization of the Decision-Making Strategies Used by Younger and Older People

Lili Zhang, Tomas Ward
Dublin City University

1. Abstract

We use an existing IGT dataset which comprises 63 older people (65-88 years) and 90 younger people (18-34 years) who had completed a computerized version of the Iowa Gambling Task (IGT). We applied the Outcome Representation Learning (ORL) model in particular as this demonstrated the best fit to the collected data and whose parameters were estimated with the robust, contemporary approach of hierarchical Bayesian analysis. We first conducted a Bayesian behavioral analysis (repeated measures ANOVA) which demonstrated no significant differences in learning curves for this task. Secondly, after fitting the ORL model, the extracted parameters revealed a statistically significant reduced decay rate for the younger group compared to the older group. In addition, the difference between the positive and negative learning rates of the older group was larger than that of the younger group. We conclude that younger subjects exhibit increased dependency on memorizing decision history based on the reduced decay parameter observed. We also conclude that learning among older subjects is dominated by positive outcomes and is more tolerant of risk. Our findings both validate and expand upon previous work in such studies.

2. Introduction

Characterizing how decision-making changes with age is important for understanding the psychology of older people. The Iowa Gambling Task (IGT) [1] presents a well-established approach for collecting data relevant to this goal. This task requires participants to choose from decks of cards that provide losses and gains in money: two of the decks have low payouts but lower losses, and the remaining two decks have high payouts but even higher losses.

The outcome-Representation Learning (ORL) Model [2] is a novel reinforcement learning model which explicitly accounts for the effects of expected value, gain–loss frequency, choice perseveration, and reversal-learning in the IGT with only five free parameters.

3. Dataset

We used a subset of a public dataset [3] that integrated 10 studies assessing performance of healthy participants on the IGT. 153 participants were recruited in this data set, including 90 younger people and 63 older people. Younger adults ranged in age from 18 to 34 years old and older adults ranged in age from 65 to 88 years.

4 Results

We present a Bayesian approach to examine whether the older group and younger group differ in their IGT performance, encompassing both behavioral and model-based anslysis.

4.1 Bayesian Behavioral Data Analyses

The Bayesian data analysis was applied in the form of a $10(block) \times 2(age)$ repeated measures ANOVA. It is showed that the data are $4789000/785318.720 = 6.10$ times more likely under the "block model" that assumes an effect of block, but no effect of group than under the "Block + Group model" that assumes both group and block differences (i.e. the Bayes factor BF_{01} is 6.10 in favor of the model that includes no main effect of group). This result demonstrates there is no significant difference in learning between the two groups.

4.2 Cognitive Modeling Analysis

We fitted the ORL model in an R package called HBayesDM (hierarchical Bayesian modeling of Decision Making tasks) using Hierarchical Bayesian analysis.

Extracting the parameters from the ORL model demonstrated elevated reward sensitivity in older subjects relative to younger subjects (Bayesian Independent Samples T-test $BF_{10} = 3.257e + 6$). The difference between the positive and learning rates of the older group was significantly larger than that of the younger group ($BF_{10} = 841082.97$), from which we conclude that learning of the older subjects is dominated by positive results and they are more tolerant of risk. In addition, the decay rate for the younger was reduced significantly compared to the older group ($BF_{10} = 4.703e + 6$), which indicates the increasing dependency on memorizing decision history of younger group.

References

[1] A. Bechara, A. R. Damasio, H. Damasio, and S. W. Anderson. Insensitivity to future consequences following damage to human prefrontal cortex. 1995.

[2] N. Haines, J. Vassileva, and W.-Y. Ahn. The outcome-representation learning model: A novel reinforcement learning model of the iowa gambling task. *Cognitive science*, 42(8):2534–2561, 2018.

[3] H. Steingroever, D. J. Fridberg, A. Horstmann, K. L. Kjome, V. Kumari, S. D. Lane, T. V. Maia, J. L. McClelland, T. Pachur, P. Premkumar, et al. Data from 617 healthy participants performing the iowa gambling task: A" many labs" collaboration. *Journal of Open Psychology Data*, 3(1):340–353, 2015.

ReACTR: Real-time Algorithm Configuration through Tournament Ranking

Tadhg Fitzgerald
University College Cork
tadhg.fitzgerald@insight-centre.org

1. Algorithm Configuration

Automated algorithm configuration (AAC) has been shown time and again to improve the performance of SAT and other combinatorial optimisation solvers[3]. Identifying a good solver configuration not only allows these solvers to find solutions faster but also allows for the fairer comparison between approaches[4]. Without AAC it is possible that an inferior algorithm can outperform another as a result of superior parameter tuning. Finally, AAC automates the laborious, time-consuming process of identifying good parameters and allows the researcher to instead focus on important work.

2. Offline vs. Real-time AAC

Traditionally algorithm configuration techniques adopt an offline approach where a collection of sample problem instances are collected and an algorithm configurator trains on these for a period of time[3]. The best configuration discovered during this training phase is then used by the solver to solve all new instances encountered in the production phase. This methodology, though successful, makes a number of assumptions which may not necessarily hold true. Offline AAC requires a representative set of training instances, which may be difficult to collect. Additionally, a significant amount of time, in the order of hours or days, is required to train the configurator. Lastly, once a good configuration has been discovered it is used to solve all future problem instances. This assumes that the structure of incoming instances do not change over time.

When developing ReACT, our goal was to address the issues outlined above by both solving and performing algorithm configuration simultaneously[2, 1]. Solutions are returned as quickly or quicker than running the default unconfigured solver. ReACT works over a stream of instances so it is not necessary to collect a training set beforehand. This stream processing approach comes with the added advantage that the configuration is constantly improving so should the type of instances encountered change the configurator is able to adapt. The only requirement that ReACTR has is that the system supports parallel processing.

3. ReACTR

Parallel Racing Parallel racing forms the core of the ReACTR system. Incoming problem instances are solved by different solver configurations (selected from an internal pool) simultaneously. The configuration which solves the problem first is deemed the winner. All other solvers are terminated immediately so as not to incur any additional overhead. The race information is then used to update an internal leaderboard. ReACT uses the Bayesian ranking system TrueSkill to effectively rank multiple configurations after each race. This internal leaderboard is invaluable for the other core components of the ReACT system, namely selection, removal, and configuration generation.

Selection As a new instances arrive we wish to exploit the information we have on good configurations in order to solve them as quickly as possible, however, in order to find improving configurations we must explore other configurations. Luckily, this problem is known as the multi-armed bandit problem and is well studied in the literature. An effective approach is the ϵ-greedy strategy where the best option is chosen with probability $1-\epsilon$ and randomly otherwise. In ReACTR, $1-\epsilon$ of the simultaneous runs are solved using the best configurations (as determined by TrueSkill ranking) while the rest are sampled from the remaining configurations in the pool.

Pool Maintenance In order to drive improvement it is necessary to remove under performing configurations from the internal pool. TrueSkill's ranking and confidence metrics are invaluable here. Any configurations which the ranking system is confident are of low quality can safely be removed and replaced by a fresh configurations. As ReACT is constantly removing weak configurations the quality of the configurations rises over time. Replacing weak configurations with good quality configurations allows faster improvement. New configurations are generated using an evolutionary approach. Two highly ranked configurations are combined by crossover to produce a new configuration which is then added to the pool. To maintain diversity in the pool and exploration of the configuration space a certain amount of configurations are also generated at random.

Using a combination of the above techniques we have shown that ReACTR can match or even exceed the performance of SOTA offline configurators (see presentation or cited papers for exact details of this).

References

[1] T. Fitzgerald, Y. Malitsky, and B. O'Sullivan. Reactr: realtime algorithm configuration through tournament rankings. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[2] T. Fitzgerald, Y. Malitsky, B. O'Sullivan, and K. Tierney. React: Real-time algorithm configuration through tournaments. In *Seventh Annual Symposium on Combinatorial Search*, 2014.

[3] F. Hutter, H. H. Hoos, and K. Leyton-Brown. Sequential model-based optimization for general algorithm configuration. In *LION 5*, 2011.

[4] F. Hutter, M. Lindauer, A. Balint, S. Bayless, H. Hoos, and K. Leyton-Brown. The configurable sat solver challenge (cssc). *Artificial Intelligence*, 2017.

Blended Control and Deep Reinforcement Learning for Unknown Fault-Tolerant Control

Yves Sohège
University College Cork
yves.sohege@insight-centre.org

1. Introduction

Autonomous Vehicles such as self-driving cars and delivery quadcopters need to be able to make intelligent decisions in case of unknown operating scenarios. Most industrial control mechanisms require hand-tuning to be able to operate under known operating conditions. Since it is impossible to know all faults a system will experience at design time the control law must be able to reconfigure itself online. In this work we explore the application of *Deep Reinforcement Learning* on a blended control framework that enables the system to synthesise an improved controller for unknown faults. We show the proposed framework outperforms a traditional fault tolerant control (FTC) framework on a Quadcopter trajectory tracking task under unknown rotor loss of effectiveness.

2. Blended Control

A recent extension of the traditional switching control framework involves the weighted combination of all controllers. The **Blended Control** Framework utilizes a high-level controller to compute the optimal blending distribution for all pre-defined controllers. This approach has received limited success due to the difficulty in calculating the optimal blending distributions but can provide smoother transitions between controllers.

3. Methods

Our work aims to explore the use of DRL to **learn the optimal blending distributions** in a Blended Control Framework. An architecture diagram of the proposed approach can be seen in Figure 1. This framework can be broken down into 3 parts. (1) The low-level controllers are pre-defined and consisting of one nominal and $N-1$ fault controllers. (2) The high-level controller is a neural network that takes as observation the low-level controller output as well as the current system performance and outputs the blend weight distribution for the low-level controllers. (3) Lastly the blending function computes the weighted sum of the controller outputs. We add constraints to ensure that the blended control signal is bound *between* the low-level controller outputs.

For situations where no low-level controller performs well, such as unknown faults, the framework is capable of interpolating between the controllers to synthesise an improved control signal. Compared to standard approaches to learn a control policy of the system directly we apply learning on a high-level control space and rely on existing controller mechanisms.

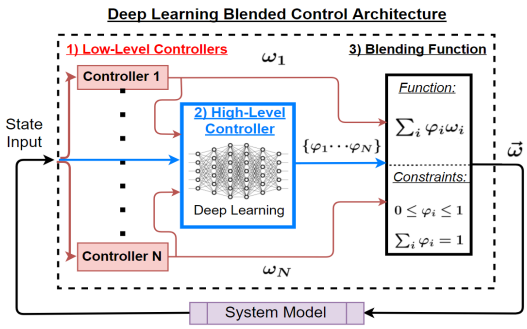


Figure 1: Deep Reinforcement Learning Blended Control Framework

4. Results

The presented approach was implemented on a Quadcopter trajectory tracking task and trained under unknown rotor faults using Deep Deterministic Policy Gradient Algorithm [1]. The full details can be found in [2]. In summary, we show that the framework has the ability to learn how to blend the low-level controllers optimally to synthesise a control signal that outperforms a traditional switched architecture on trajectory and attitude tracking accuracy under unknown rotor faults. We also remove the inherent delay in computing a fault diagnosis, which can be catastrophic for highly unstable systems, by using a neural network which has a continuous action space.

5. Future Work

Learning the blending distribution between low-level controllers has a major drawback which is that it fundamentally depends on the low-level controller tuning. We aim to address this enhancing one of the low-level controllers with learning so that for unknown faults the system is not limited to the static pre-tuned controllers.

References

[1] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra. Continuous control with deep reinforcement learning. 2015.

[2] Y. Sohège and G. Provan. Unknown fault tolerant control using deep reinforcement learning: A blended control approach. 2019.

A Grouping Genetic Algorithm for Joint Stratification and Sample Allocation Designs

Mervyn O'Luing

Insight Centre for Data Analytics, University College Cork
mervyn.oluing@insight-centre.org

1. Introduction

In this paper we propose an algorithm to partition (create subsets of) atomic strata into larger groupings or strata and search for the minimum sample size that meets accuracy requirements from all possible partitions. We build on the work of Ballin and Barcaroli (2013) who use a Genetic Algorithm (GA) to cut short the search of all possible partitions to find the minimum sample size. We propose an alternative GA that we claim is better suited to this application. We support our claim by comparing computational results on publicly-available test data.

2. Problem Description

Consider all possible partitions of the atomic strata and for each of these partitions: estimate the minimum sample size necessary to meet your accuracy requirements. Each partition is a candidate solution. The number of possible partitions is known as the search space, because, if it was feasible in terms of time and cost, the full list of candidate solutions would be evaluated for the optimum solution. However, given that the set of partitions grows exponentially with the set of atomic strata, search techniques such as genetic algorithms are used as they can find an optimum solution without evaluating all possible solutions.

3. Our Goal

To design a *grouping genetic algorithm* (GGA) (Falkenauer, 1998). GGAs have been shown to perform far better than standard GAs on grouping problems. Our GGA should a good quality or the best solution quicker than the GA. Smaller samples are cheaper to gather. It also means a saving in time taken to find the smaller sample size.

4. Why the GGA should outperform the GA

Strata should be independent (the same value cannot be in more than one stratum) but values in each stratum should be close in value (internally homogenous). There are two levels of strata in this problem. The basic level is the atomic strata. We intend to create higher level strata from subsets, partitions or groupings of atomic strata. If each atomic stratum contains values that are the same or close in value then smaller sample sizes are needed for a precise estimate of the mean or total for the target variables. It follows also that grouping homogeneous atomic strata into strata will also require a smaller sample size to meet precision constraints. This is what leads to a good candidate solution, i.e. the strata are internally homogeneous but independent and smaller samples sizes result.

The original GA typically prolongs the search for the the optimum solution, because good strata could be mixed with bad strata - which is a 'hit-or-miss' approach in practice and can push the sample size back up for the offspring. The GGA on the other hand preserves the information in good strata and this is more likely to create better quality offspring.

5. Comparing the Genetic Algorithms

The online crowdfunding platform kiva.org provides a dataset of loans issued to people living in poor and financially excluded circumstances around the world over a two period for a Kaggle Data Science for Good challenge. The dataset has 671,205 unique records.

The algorithms searched for the smallest sample size necessary to accurately describe *term in months, lender count and loan amount* after 100 iterations.

GA		**GGA**		**Reduction**	
Sample size	**Strata**	**Sample size**	**Strata**	**Sample size**	**Strata**
78018	43030	11963	1793	84.67%	95.83%

Table 1: Comparison of GA v GGA algorithms on Kiva Loans dataset

The above table shows an 84.67% in sample size and a 95.83% reduction in the number of strata after 100 iterations.

References

Ballin, M. and G. Barcaroli (2013). Joint determination of optimal stratification and sample allocation using genetic algorithm. *Survey Methodology 39*(2), 369–393.

Falkenauer, E. (1998). *Genetic algorithms and grouping problems*. John Wiley & Sons, Inc.

Improved Drug Residue Withdrawal Method for Antibiotic usage in Cow's Milk

Cathal Ryan
Insight Centre for Data Analytics, UCD
cathal.ryan@insight-centre.org

1. Introduction

The production of milk from cows both globally, and also within Ireland has been steadily growing over time. Thus the need and use of antibiotics to help keep the increased quantity of cows healthy has also been increasing. This has lead models to try to calculate when the potentially harmful antibiotics are discarded from the cow. Many worldwide organisations such as the European Medical Agency (EMA) [1] and the Food and Drug Association (FDA) [2] have published their own statistical methods for calculating the time at which a cows milk should be deemed healthy, both of which focus on creating a specific value called the time to safe concentration value (TTSC). While calculating the TTSC value both models compute what is called a tolerance upper limit which shows with a certain $\alpha\%$ confidence that $\delta\%$ proportion of the total population will be below the Maximum Residue Limit (MRL). There are many drawbacks of the two models discussed including the problem where the observations drop below the Limit of Quantification (LOQ) and thus the model should take this into account.

2. Alternative Model

Our alternative model tries to build upon the model created by the FDA which creates a single linear regression model for each cow and then calculates the tolerance upper limit from predictions on each seperate model.

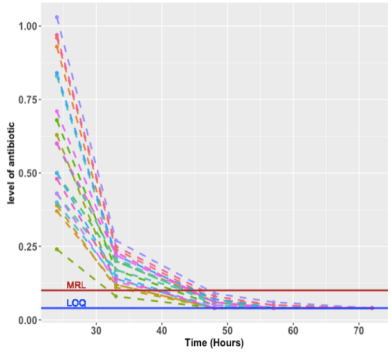


Figure 1: Antibiotic Residue over time

Yet when we look at Figure 1 we can see that many values are being recorded at the LOQ due to them being censored values thus our model takes this into account by using a seperate likelihood function for when the observation is censored. Thus leaving us with the likelihood function for a Tobit Regression model [3]:

$$\sum_{n=0}^{N}[I_n^{LOQ} log\phi(\frac{LOQ - x_n'\beta}{\sigma})+$$

$$(1 - I_n^{LOQ})(log\phi(\frac{y_n - x_n^{'\beta}}{\sigma}) - log\sigma)]$$

$$I_n^{LOQ} = 1, if y_n = LOQ, I_n^{LOQ} = 0, if y_n > LOQ$$

When we look at the data again we can also see another shortcoming of the FDA method, it assumes each cow is independent of each other. This is quite clearly false in our case as most cows reduce the antibiotic residue within their milk at roughly the same in this dataset. To take this into account we added a random intercept for each cow to allow for variation between cows which leads us to the final likelihood function:

$$L_n = \int_{-\infty}^{+\infty} \prod_{m=1}^{M} [\phi(\frac{a - x_{nm}'\beta - \mu_n}{\sigma_\epsilon}]^{I_{it}^{LOQ}} *$$

$$[\frac{1}{\sigma_\nu}\phi(\frac{y_{nm} - x_{nm}'\beta - \mu_n}{\sigma_\epsilon}]^{1-I_{nm}^{LOQ}}]\phi(\frac{\mu_n}{\sigma_\mu})d\mu_n$$

3. Results

The addition of using both a Tobit Regression model and also a random effect for each cow greatly increased the statistical power of our model compared to other existing models. This can be seen by calculating the Bayesian Inference Criteria (BIC) for each model. BIC returns a value of 65.9 for our Tobit mixed effect Regression compared to a value of 230.7483 for the Linear fixed effects model.
The occured with a value of 58.4 hours, while the lowest TTSC came about from our proposed Tobit mixed effects model with a value of 48.6 hours. Which would allow for a farmer using this particular antibiotic to keep the third milking if they used results computed from our model compared to the EMA and FDA method.

4. References

1. Ema.europa.eu. (2000). (online) Available at: https://tinyurl.com/s6yr9js [Accessed 11 Dec. 2019].

2. Fda.gov. (2019). [online] Available at: https://www.fda.gov/media/70028/download [Accessed 11 Dec. 2019]

3. Tobin, J. (1958). Estimation of Relationships for Limited Dependent Variables. Econometrica, 26(1), p.24.

Feature set reduction for improved interpretability of machine learning in radiology

Jingwen Bian, Eric Wolsztynski
Insight Centre for Data Analytics; School of Mathematical Science, University College Cork
jingwen.bian@insight-centre.org

1. Introduction

The application of machine learning (ML) to analyse high-throughput Positron Emission Tomography (PET) imaging datasets has become ubiquitous in Radiology in the last decade. This framework, called radiomics, involves a large number of image-derived agnostic features with limited biological interpretation. Clinical integration of artificial intelligence (AI) systems in this field will require interpretable predictive models for both diagnosis and therapeutic management. This challenging objective could be guided by the learning output from ML models.

2. Dataset

We consider radiological summaries derived for a set of primary non-small cell lung cancer PET studies imaged with fluorodeoxyglucose (FDG) acquired at Cork University Hospital over a three-year period starting in 2012 [1]. The final cohort comprised of 93 PET/CT patients after exclusion of unsuitable cases. A total of 133 variables were considered and may be identified in two frames: (i) routine clinical variables; (ii) image-derived variables, comprising of structural features with associated spatial uptake gradients as defined in [1] and a set of image summaries including morphologic and texture features, all derived by definitions provided by the Image Biomarker Standardization Initiative [2].

3. Methodology

3.1. Correlation and partial correlation analyses

Correlation analysis is used to identify associations and potential pathways for feature set simplification. Since correlation is largely induced by overlap in information with other variables within the dataset, Gaussian graphical models (GGMs) are also used to analyse partial correlation between variables, thus highlighting direct relationships between features that are conditionally dependent given all other variables.

3.2. Multilinear analysis of features

Direct association of features is also captured separately via multivariate modelling, using one feature as the dependent variable explained by subset of other features. Here linear modelling was considered and a final subset comprising of the 2 or 3 most important predictors in the model were selected for final prediction of each feature in the dataset. Lasso models and stepwise selection were both used to capture linear relationships. Repeated 5-fold cross-validation (CV), using 10 repetitions, was applied to a subset of 80 observations, fitting each model to each of the features successively. Final predictive models were fitted to the whole CV data, using only these most popular features determined by CV. Feature prediction potential was evaluated on the independent test set comprising of the remaining studies in the NSCLC dataset for prediction performance assessment.

3.3. Nonlinear analysis of features

In order to capture potential nonlinear relationships among the feature set, random forest and neural network were also fitted and tested in a similar CV framework as that used for the linear models.

4. Results and Future Work

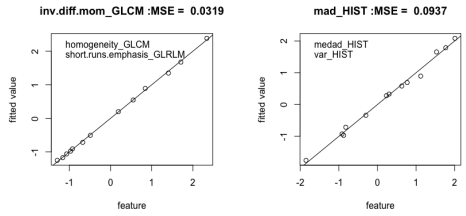


Figure 1: Scatter plots of feature values and predictions

This work aimed at identifying potential pathways for multivariate recombination of agnostic features, in view of simplifying radiomics datasets. For many of the agnostic image-derived features, predictors selected via cross-validated modelling could be used to describe the dependent features with high accuracy, thus allowing for significant reduction of the feature set. This in turn may facilitate clinical interpretation of the feature subsets. The next step of this work will focus on the question of integration of feature recombinations into predictive and prognostic models.

Acknowledgement

This work was supported by Science Foundation Ireland under Grant No. 12/RC/2289-P2.

References

[1] E. Wolsztynski, J. O'Sullivan, N. M. Hughes, T. Mou, P. Murphy, F. O'Sullivan, and K. O'Regan. Combining Structural and Textural Assessments of Volumetric FDG-PET Uptake in NSCLC. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 3(4):421–433, 2019.

[2] A. Zwanenburg, S. Leger, M. Vallières, and S. Löck. Image biomarker standardisation initiative. *arXiv preprint arXiv:1612.07003*, 2016.

Optimization and Acquiring the Solutions in Interactive Constraint Based System

Hong Huang
Insight Centre for Data Analytics, University College Cork, Ireland
hong.huang@insight-centre.org

Abstract

The decision optimization problem has been frequently raised in recent years. A decision optimization problem involves not just deliver a good enough solution to solve the problem, but also an efficient way to construct such a solution. In nowadays, technology allows human to express and respond very fast, i.e. 5G. The trend of the personalized and customized is rising and becoming more and more apparent. The novel technology allows the information system to exchange and process massive amounts of data. It shows the great challenge and opportunity for the decision optimization related field and research work. In past decades, the constraint programming (CP) has shown great potential in solving not just the mathematics problem but also the decision problem, etc.

In the real-life, the decision problem is solving by providing one or more optimal, or good enough in many circumstances, solutions through an agent. The searching and learning process for solutions are performed by the system. The system here could be interactive, where user can express their preference and react over one or more given solutions. However, in the real-life problem, searching for the optimal solution and mining the user's exact preferences are still popular topics. In this work, we try to address the problem of finding such an optimal solution and consider the less-cost way or tradeoff way to achieve such a goal.

1. Introduction

There are many recent works have shown that the Constraint Satisfy Problem (CSP) is extremely helpful formulating a lot of real-life problems. In some special CSP, like the Weighted Max Constraint Satisfy Problem (WMax-CSP), that allows the user to express the degree of satisfaction of constraints, rather than only satisfaction or violation of the whole constraint clauses. Such kind of real-life problem is like hotel searching, movie filtering and product selecting, etc. For these problems, much research has proven that constraint can be well applied to tackle the issue and find the solution. In some of those frameworks (e.g. [2, 1]), where the standard constraints are associated with a preference function which specifies the preference for each distance allowed by the constraint, it is allowed to express and modify the preference both over constraints and solutions.

The initial problem is if an interactive system that can acquire user's preference and take the known the preferences to filter the good solutions for the user, can the interacting and learning processes of it be optimized? How to achieve the balance between the cost of updating the system and acquiring the preference and learning good solutions to users?

Many techniques and extensions of classical CSP have been proposed, developed and well researched to make CSP more robust and reliable in solving the decision optimization problems such as acquiring the user's preferences and use the preferences to provide the best solution to the user. There are some well-studied example of the extension of classical CSP such as Fuzzy Semiring CSP (FSCSP) can be used in describing and solving the proposed problem.

2. Experiment

In each iteration, the system gives top 10 solutions to the user and asks for user's feedback. After the acquiring process, we synchronized the preference value of constraints for the rest of data. Then we re-rank all the solutions to find the new best 10 solutions to ask for the user's satisfaction. The acquiring process will stop when it finds all the optimal solution to the user or after 10 iterations.

Benchmark Details			Results					
#Total	#Opt	Sparse	Only Sols Preference			Both Sols & Single Art		
Items	Sols	Factor	#Iters	Time(s)	Eva	#Iters	Time(s)	Eva
1000	1	10	10	1.776	0.0432	**2.8**	1.829	**0.098**
1000	1	20	10	1.807	0.0184	**8.4**	1.793	**0.0268**
1000	15	10	**8.8**	1.947	**0.2426**	**4.6**	1.859	**0.3700**
1000	15	20	**4**	2.067	**0.2514**	**1**	1.843	**0.3059**
5000	1	10	10	8.914	0.0329	10	9.254	0.0264
5000	1	20	10	8.102	0.0264	**6.4**	7.898	**0.0086**
5000	15	10	10	8.846	0.2480	**8.2**	8.782	**0.3087**
5000	15	20	10	7.633	0.2736	**8.2**	6.622	**0.3431**
10000	1	10	10	22.402	0.0096	10	24.281	0.0097
10000	1	20	10	18.073	0.0149	10	17.128	0.0163
10000	15	10	**8.2**	22.023	**0.2672**	**8.2**	22.522	**0.3573**
10000	15	20	10	16.343	0.2347	**8.2**	16.917	**0.2706**

The datasets in our experiments are all randomly generated. The items are all containing 7 attributes and 4 constraints over the attributes. The deviation index of results are represented as Eva in Table 2. The results showed in Table 2 are the mathematical average of the results of 5 experiment runs. The calculation for the deviation index in our experiments is: $Eva = \frac{1}{n}\sum_{i=1}^{n}\frac{(O_i - i)}{m}$ (where n is the number of the Oracle optimal solutions, m is the number of total items, O_i is the index of solution i in the optimal order with respect to all the items). For the results, lower the deviation index value means better solutions it found.

References

[1] F. Rossi and A. Sperduti. Acquiring both constraint and solution preferences in interactive constraint systems. *Constraints*, 9(4):311–332, 2004.

[2] F. Rossi, K. B. Venable, and N. Yorke-Smith. Uncertainty in soft temporal constraint problems:a general framework and controllability algorithms forthe fuzzy case. *CoRR*, abs/1110.2212, 2011.

Replicating the VIX Index

Mingchuan Zhao
University of Limerick
mingchuan.zhao@insight-centre.org

1. Introduction

Cboe Global Markets, Incorporated (Cboe) introduced the famous Cboe Volatility Index, i.e. the VIX Index in 1993.[2] It is designed to measure the market's expectation of 30-day volatility implied by at-the-money S&P 100 Index option prices. Nowadays the new VIX Index is calculated based on S&P 500.

Meanwhile, a data frame of the monthly S&P 500 options data can be obtained from OptionMetrics. In order to examine whether the data frame can work properly in future researches, we can use the VIX Index to perform the check. The step-by-step calculation method of the VIX Index is explained in details in the Cboe VIX White Paper.[2] Thus, we can use our own data frame to calculate our homemade VIX Index.

2. Homemade VIX Index

We can mimic the calculation in the Cboe VIX White Paper to get a homemade VIX. Instead of using the daily data, we use the monthly S&P 500 option data from OptionMetrics database.

2.1 Selection of the Options

Firstly, for each date the options needs to be sifted for use in the VIX Index calculation. Only out-of-the-money calls and puts would be selected, since they do not carry intrinsic values. It is also necessary to decide the strike price range of options with non-zero bids. This range can reflect the market's expectation of future underlying prices in the following month. We select calls (puts) with strike prices higher (lower) than the forward price, and once two consecutive call (put) options are found to have zero bids, no calls (puts) with higher (lower) strikes are considered.

2.2 Calculation of Homemade VIX

According to the VIX White Paper, the Volatility implied by the option prices can be calculated based on the following formula:

$$\sigma^2 = \sqrt{\frac{2}{T}\sum_i \frac{\Delta K_i}{K_i^2} e^{RT} Q(K_i) - \frac{1}{T}\left[\frac{F}{K_0} - 1\right]^2}, \quad (1)$$

and the VIX Index is thus calculated by $VIX = 100 \times \sigma$. In formula (1), F is the forward price derived from the option prices by call-put parity. K_0 is the first strike below the forward price F. ΔK_i is the interval between strike prices. $Q(K_i)$ is the midpoint of the bid-ask spread for each option. All calculation steps are modified on the basis of the monthly data we have.

We can then compare our homemade VIX with the real-time VIX data from the Cboe database to see how much the two VIX indexes are similar and related to each other.

3. Conclusions and Future Work

The result shows that the homemade VIX is very closely related to the real one. The Pearson correlation coefficient is 0.989 and the linear regression model works very well between the two VIXs. All the coefficients are significant.

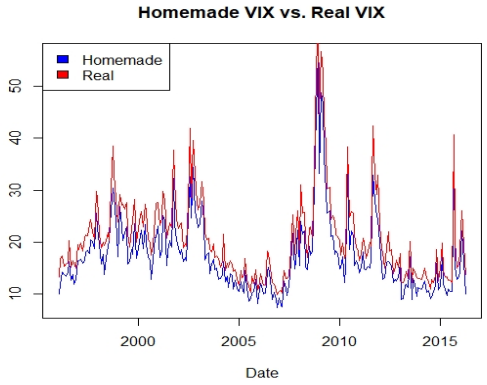


Figure 1: Homemade and Real VIX

This has proved that the S&P 500 data we have is working very well and the homemade VIX is a very good substitute of the real VIX. The next step for the research can be trying to find the replicating strategy of this homemade VIX Index by using the option data we have, or using the VIX Index to look into different financial derivatives, for example, variance swaps.[1]

The ultimate goal of the project is to utilize the data frame to help optimise option portfolios according to specific criterions.

References

[1] S. Bossu, E. Strasser, and R. Guichard. Just What You Need to Know about Variance Swaps. May 2005.
[2] Cboe Exchange Inc. Cboe Vix White Paper. pages 1–19, 2019.

The Idea of Conceptual Consensus for Entity Management on the Blockchain

Atiya Usmani, Dr. Edward Curry
Insight Centre for Data Analytics
{ atiya.usmani, edward.curry}@insight-centre.org

1. Motivation

Blockchain is becoming popular among data management applications, but they still remain inexplorable. It is a secure, tamper-proof, distributed ledger of linked encrypted blocks of data that establishes trust in a trustless system. Despite being transparent, it is very cumbersome to search for data and transactions relating to similar entities on the blockchain behind the cryptographic keys. Thus, there is a need to create a simple blockchain on which a decentralised registry can be created making the blockchain more queryable and explorable.

Most blockchains use transactional consensus like PoW, PoS, etc. to ensure that all nodes have the same copy of the chain. But these take time (non-finality problem) and computational power making them unfit for real-time and primitive IoT devices. Hence , we have abolished transactional consensus in order to make the application more simpler. A major issue then, is to acheive consensus in building the registry and resolving inconsistencies.

2. Problem Statement

When many different organisations are collaborating and exchanging data on the blockchain, how can we create a decentralised registry of different entities in the data ecosystem, such that consensus be achieved between two conficting entity descriptions. We coin the term - ”conceptual consensus” for it.

3. Related Work

Third and Domingue [1] have built a semantic index for Ethereum representing the data on distributed ledgers as linked data using Blondie ontology and Minimum Service Model Ontology (MSM) to map smart contracts in order to connect them to semantic web services. RDF triples are generated and are stored on in the RDF Store where they can be queried using SPARQL. Crowder [2] is an entity resolution system which uses t is a hybrid human-machine technique where machines do the initial coarse pass and humans verify them.

4. Proposed Solution and Discussion

The decentralised registry (see Figure 1) consists of four layers : i) participating data souces, ii) the blockchain upon which the entity registry is built- *entitychain*, iii) the querying layer and iv) the consensus layer. In our system, there is no mining of transactions. ”Conceptual Consensus” is achieved with the help of association graphs and user rankings.

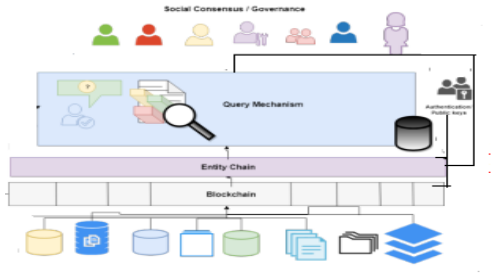


Figure 1: System in Layers.

First all nodes enter the metadata of their entities resulting in the initial population of the registry.Then initial associations are identified by machines using Wordnet and entity matchers to create association graphs. The fine tuning of these association graphs/trails are performed by levaraging user feedback by generating HITs or Human Intellegence Tasks wherein each user rates the similiraty between a set of attributes or entity pairs. The initial automatic matchers do help in generating meaningful comparison tasks for the user, thus reducing the human effort required and making it more scalable. Further the user is rewarded by providing various incentives. For attributes having inconsistent values in different datasets, both of them will be displayed along with the source IDs and the users can rank the results.The similarity measures are then calculated by combining the numerical values of the matchers and past user rankings. It also makes it easier to query and explore the blockchain and recieve more comprehensive results. When the user queries on one attribute from one database, the graph is traversed and the answer set also includes the synonymous attribute from the other dataset or the same dataset. The security of the system is maintained with the help of user authentication , and all user rankings are stored on the chain to avoid malicious behavior. The privacy of the user is maintained behind cryptographic hashes.

References

[1] Third A, Domingue J. Linked data indexing of distributed ledgers. InProceedings of the 26th International Conference on World Wide Web Companion 2017 Apr 3 (pp. 1431-1436). International World Wide Web Conferences Steering Committee.

[2] Wang J, Kraska T, Franklin MJ, Feng J.,” Crowder: Crowdsourcing entity resolution.” Proceedings of the VLDB Endowment. 2012 Jul 1;5(11):1483-94.

The Good, the Bad, and the Unexpected: Factors Affecting the Valence of Unexpected Events

Molly Quinn
University College Dublin
molly.quinn@insight-centre.org

1. Introduction

Though most of us probably worry about unexpectedly bad things happening to us, we sometimes daydream about unexpectedly good things happening, too. These opposing possibilities reflect the inherent ambiguity that arises when we consider what unexpected things may occur next. Researchers do not often ask what "the unexpected" truly means to their participants. In the present study, participants were presented with everyday scenarios and asked to think of unexpected next events.

2. Experiment 1

The first experiment revealed a potential countering effect in the valence of unexpected events produced for valenced scenarios. We first collected a total of 2,540 unexpected events (127 participants x 20 materials).

In a post-test, independent participants (N=31) rated the valence of the material scenarios on a 5-pt Likert-type scale from Very Negative to Very Positive. Materials were mainly positive, but some negative materials revealed a possible correlation, r = -.60, p = .005.

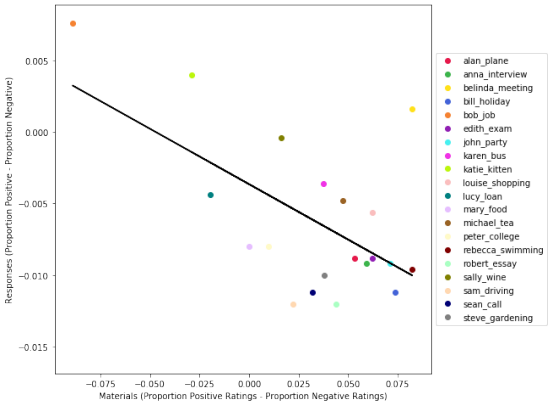


Figure 1: Potential relationship between valence of material and valence of response to UNEXP condition.

3. Experiment 2

The second experiment directly manipulated valence of the original scenarios to determine its effect on responses.

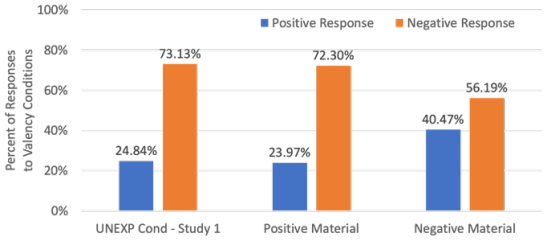
3.1. Methods

Participants viewed 10 edited scenarios and answered "Something unexpected happened, what do you think happened?" Participants saw only the positive or the negative version of a material. They saw five materials of each valence.

Sentence Type	***Sample Scenario – Louise Shopping***
Goal	Louise wants to shop at an expensive clothes store.
Positively-Valenced	The previous week, she received a pay rise because she won a big account for her company.
Negatively-Valenced	The previous week, she received a pay cut because she lost a big account for her company.
Further Action	Louise draws money from the ATM.

3.2. Results

We collected a total of 1,020 responses (102 Participants x 10 materials x 2 Valency Conditions) . There was still a majority of negative responses overall, however positively-valenced materials received more negative responses than negatively-valenced materials.



4. Discussion

The negativity bias has been well-established in many domains of psychological literature (Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001; Rozin & Royzman, 2001; Taylor, 1991). Results from the present studies suggest that while there may be a negativity bias in producing unexpected events, the extent of this bias is affected by the original state of the valence of events. Future work should examine how this bias might be reduced.

5. References

[1] I Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad Is Stronger Than Good. *Review of General Psychology*, *5*(4), 323–370

[2] Foster, M. I., & Keane, M. T. (2015). Why some surprises are more surprising than others: Surprise as a metacognitive sense *of* explanatory difficulty. *Cognitive psychology*, *81*, 74-116.

[3] Quinn, M. S., Campbell, K., & Keane, M. T. (2019). The Unexpected Unexpected and the Expected Unexpected: How People's Conception of the Unexpected is Not That Unexpected. *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci19).* Montreal, Canada (July).

[4] Taylor, S. E. (1991). Asymetrical effects of positive and negative events: The mobilization-minimization hypothesis. Psychological Bulletin, 110(1), 67–85.

Analysis of pathogenic and commensal bacterial volatile signatures using solid phase micro-extraction (SPME) coupled with Gas Chromatography – Mass Spectrometry

Shane Fitzgerald,
Dublin City University, Insight Centre for Data Analytics
Shane.fitzgerald28@mail.dcu.ie

1. Introduction

Wound infections play a major role in the incessant growth of chronic wounds. The management of chronic wounds has become a huge economic burden, highlighting the need for improved methods of wound infection monitoring. The detection of volatile organic compounds (VOCs) emitted by pathogenic bacteria has been proposed as a potential non-invasive approach for characterising wound infections. However, preliminary studies must be carried out to characterise microbial volatiles from both pathogens and skin commensals and to assess various factors that influence their VOC production. In this study, VOC profiles of prominent bacteria - present in wounds - were obtained using a simple solid phase micro-extraction (SPME) sampling step coupled with GC-MS analysis. The techniques used are quick, simple, and capable of recovering a sufficient amount of VOCs to discriminate prevalent pathogenic bacteria. As non-invasive sampling procedures are growing more desirable, these techniques may potentially be transferrable to clinical settings.

2. Clinical context

Bacteria emit VOCs as secondary metabolites in response to their environment (e.g. nutrients, substrate, and cell density), with each bacterial species emitting their own specific VOC signature. Detection of these VOC signatures could potentially be used to identify infection before it manifests. The potential clinical application of microbial VOC analysis has already been highlighted by studies carried out on the breath of cystic fibrosis patients[i]. Interest in the VOCs emitted from wounds began in 2010, when A.N. Thomas et al published a study that utilised the application of a skin sampling patch, which was placed over each participant's wound for 24 hours; after this period, the patches were removed and the absorbed VOCs were thermally desorbed and analysed using GC-MS [ii]. In 2018, a study was carried out on human biopsies to assess the microbial load using VOC emissions and confirmed that the use of VOCs has the potential for wound monitoring[iii]. Obtaining biopsies from patients is invasive and can be potentially painful so there is a desire for less-invasive sample collection procedures.

3. Proposal

The main objective of this study was to assess the capability of this method for the rapid *in vitro* discrimination of four prevalent microbial species present in infected wounds. We propose that in the future this analytical method will be potentially compatible with non-invasive sample collection methods such as swabbing, and yield important diagnostic information.

4. Results

Sampling was successfully carried out through a quick (20 minute), non-contact SPME sampling step that allowed a quick turn over of analysis results. The results of this study have shown that these bacteria can be differentiated based on their VOC signatures. PCA was performed on the bacterial VOC data to visualize discriminatory VOC signatures. The dendrogram in Figure 1 shows that the three pathogenic (harmful) bacteria *S.aureus, P.aeruginosa, and E.coli* were successfully clustered based on their VOC signatures.

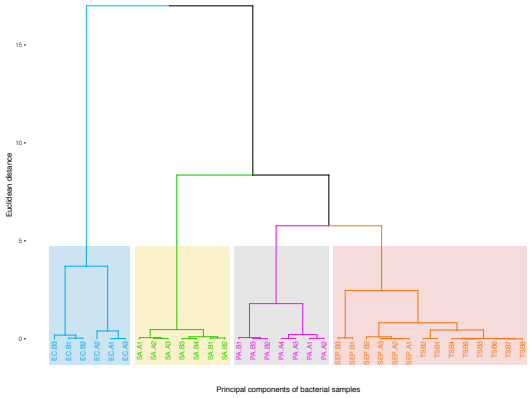


Figure 1: Dendrogram produced through hierarchical clustering analysis of the principal components of the bacterial samples (k = 4). E.coli (blue), EC.A: DSM30083 & EC.B: DSM103372; S.aureus (green), SA.A: DSM2569 & SA.B: DSM799; P.aeruginosa (pink), PA.A: DSM105372 & PA.B: DSM25642; S.epidermidis (orange), SEP.A: CSF41498 & SEP.B: RP62A; and TSB control media (orange).

8. References

[1] Bos, L., Meinardi, S., Blake, D. and Whiteson, K. (2016). Bacteria in the airways of patients with cystic fibrosis are genetically capable of producing VOCs in breath. *Journal of Breath Research*, 10(4), p.047103.

[2] Thomas, A., Riazanskaia, S., Cheung, W., Xu, Y., Goodacre, R., Thomas, C., Baguneid, M. and Bayat, A. (2010). Novel noninvasive identification of biomarkers by analytical profiling of chronic wounds using volatile organic compounds. *Wound Repair and Regeneration*, 18(4), pp.391-400.

[3] Ashrafi, M., Novak-Frazer, L., Bates, M., Baguneid, M., Alonso-Rasgado, T., Xia, G., Rautemaa-Richardson, R. and Bayat, A. (2018). Validation of biofilm formation on human skin wound models and demonstration of clinically translatable bacteria-specific volatile signatures. *Scientific Reports*, 8(1).

Water Level Prediction using LSTMs

Asma Slaimi, Noel O'Connor, Fiona Regan, Susan Hegarty
Dublin City University
asma.slaimi@insight-centre.org

1. Abstract

In recent years, more and more emphasis has been placed on the predicted rise in water level as becomes a more pressing issue as a result of global warming. Flooding is one of the most prevalent and most destructive natural disasters often resulting in loss of life. Flooding occurs when there is more water upstream than usual and as this flows downs stream to the low-lying areas, the excess water flow into the surrounding areas.

Application of machine learning (ML) techniques has received significant attention in recent years. One of the more popular applications of ML is prediction systems. Predicting water level changes is notoriously tricky, it depends on a complex mixture of data including the history of water level, precipitation among other data sources.

In this research, Artificial Neural Networks (ANNs) are used as a basis to create our predictive system. ANNs are a sub-field of machine learning where the algorithms are inspired by the structure of the human brain. It can be considered to be a "black box" winch takes one or multiple inputs, such as sensor data, processing them into one or multiple outputs. Our system uses data provided by sensors, to create a computer tool to predict river water level using ANNs in particular using Long Short-Term Memory models (LSTMs) [1].

2. Introduction

Developing accurate tools for analysing the hydrological cycle is an active research area. A key challenge is that development of such tools requires large amounts of data. Researchers often only have access to a small range of measurements due to the costly nature of data collection and a limited number of techniques used. Most research efforts in the literature focus on building specific/physical models. Building a generic model that can be applied to several different catchments, as targeted in this work, is extremely challenging.

3. Hydrological event prediction

Predictions based on machine learning could significantly contribute to water resource management (for both short and long term hydrological event prediction such as flood, rainfall or storms) providing better performance and cost-effective solutions. To date, physical models were used to predict hydrological events. These models require various type of hydro-geomorphological data sets and knowledge of hydrological parameters which can be highly challenging.

Recently developed prediction models are data driven. These models numerically model the inherent non-linearity, solely based on historical data without requiring knowledge about the underlying physical processes.

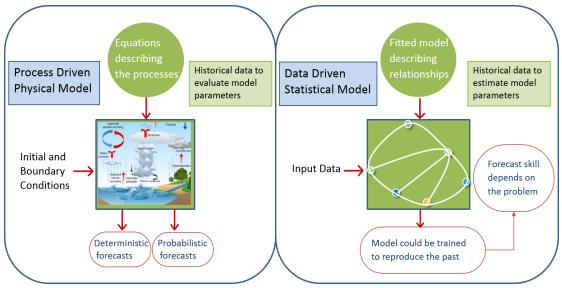


Figure 1: Data driven statistical models vs process driven physical models

Recurrent neural networks (RNNs) are a machine learning approach that can derive meaning from historical data (instead of catchment physical characterizations). Long Short-Term Memory models (LSTMs) have recently become popular choices for problems related to sequential time series data. They are very effective in handling long-term temporal dependencies.

4. Methods

We started by elaborating a literature review regarding the best practice in relation to how data is managed and applied to catchment scale water management (in order to understand the structure of current data storage and management platform). The initial idea was to introduce big data analysis methods to the catchment monitoring domain. So, we have conducted some initial experiments on local data-sets to determine the appropriateness of LSTMs Neural Networks for ahead-of-time water level prediction. This analysis was helpful in detecting water body response and recovery.

5. Conclusion

LSTMs networks can follow the general pattern of the water levels based on historical data. Further research can be carried out to investigate wither some other parameters can be incorporated to the model to enhance the prediction such as rainfall data.

References

[1] D. Zhang, B. Heery, M. O'Neil, S. Little, N. E. O'Connor, and F. Regan. A low cost smart sensor network for catchment monitoring. *Sensors*, 19(10), May 2019.

Towards Architecture-Agnostic Neural Transfer

Seán Quinn
Insight Centre for Data Analytics, Dublin City University
sean.quinn@insight-centre.org

1. Introduction

Modern Artificial Neural Networks can leverage large amounts of data to be trained to perform hard tasks such as recognising objects in an image or translating languages. The process they use is equivalent to a feature extraction with respect to the raw data and an optimisation goal. This process exposes the underlying compositional and hierarchical structure of the concepts contained within high dimensional data, but does not typically provide high level access to such structure or easily facilitate it's re-use in related tasks. Unlike Neural Networks, humans learn by building a conceptual model of the world, which relies on the persistence of known concepts across tasks, and the ability to carry a reservoir of background knowledge across domains. The authors of [2] argue that such a conceptual model is fundamentally incompatible with the purely connectionist learning of a neural network, as it may require the exploration of structural variations in the network's architecture, which goes beyond the capabilities of regular gradient-based learning in fixed weight space. One of the main challenges in making more human-like artificial intelligence is incorporating these properties of structured learning into the neural network learning paradigm. In addressing this challenge, we have been inspired by recent influential review articles within the Deep Learning community [2] which call for new approaches to enhance deep representations with prior knowledge. We consider this a significant stepping stone on the path to improving the ability of machines to learn new tasks faster and in a domain invariant way.

2. Related Work

The scope of this research is neural knowledge transfer methods which go beyond the basic direct transfer of layers of trained weights from one neural network to another, a method which enforces the maximum level of architectural constraint through making the transfer of knowledge dependant on a significant section of both networks being identical. In surveying the literature the main approaches relevant to this research include (i) Knowledge Distillation (ii) Maximum Mean Discrepancy (MMD) based methods (iii) Adversarial Domain Adaptation [5] (iv) methods based on aligning certain gramian matrices between networks [6] [3] (v) Kullback–Leibler divergence (KL-Divergence) based methods (vi) and Probabilistic Knowledge Transfer [4].

3. Scientific Approach

In addition to conducting a transparent comparison of the methods detailed in section 2 we aim to contribute towards the development of less architecturally constraining knowledge transfer methods. (a) For those methods which require the manual selection of layers for knowledge transfer in teacher and student networks, we aim to augment the method with a gated structure for automatically learning which layers to use, something akin to what is employed in LSTM models, allowing knowledge to flow from teacher to student across different layers depending on the stage of training. (b) Many of the current methods require pairing of layers across teacher and student such that each pair is the same size, the only current workaround to this constraint is to add a regressor on top of one layer so as to align their sizes, this has been shown to be less than optimal as information is lost in this process. We aim to explore the effectiveness of more sophisticated layer re-sizing methods such as [1] in removing this constraint in the knowledge transfer scenario. (c) Many of the methods we examine can be considered as a pre-training step conducted before regular neural training on a given task while others operate in conjunction with regular learning as a joint loss, we aim to investigate the suitability of various combinations of methods whereby one is a pre-training and the other is a joint loss objective. (d) Finally, it has been shown in the case of Neural Style Transfer, which is a method based on aligning Gramian matrices across networks, that when derived this method is equivalent to a form of MMD loss [3]. We aim to explore if another Gramian matrix based method, [6], can also be derived to a form of MMD and if so, to further investigate the relationship between the two methods.

References

[1] T. Chen, I. Goodfellow, and J. Shlens. Net2net: Accelerating learning via knowledge transfer. *arXiv preprint arXiv:1511.05641*, 2015.

[2] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, 2017.

[3] Y. Li, N. Wang, J. Liu, and X. Hou. Demystifying neural style transfer. *arXiv preprint arXiv:1701.01036*, 2017.

[4] N. Passalis and A. Tefas. Learning deep representations with probabilistic knowledge transfer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 268–284, 2018.

[5] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7167–7176, 2017.

[6] J. Yim, D. Joo, J. Bae, and J. Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4133–4141, 2017.

Pseudo-Labeling and Confirmation Bias in Deep Semi-Supervised Learning

Eric Arazo, Diego Ortego, Paul Albert, Noel E. O'Connor & Kevin McGuinness
Insight Centre for Data Analytics, Dublin City University (DCU)
Dublin, Ireland
eric.arazo@insight-centre.org

1. Introduction

Semi-supervised learning (SSL), i.e. jointly learning from labeled and unlabeled samples, is an active research topic due to its key role on relaxing human supervision. In the context of image classification, recent advances are mainly focused on consistency regularization methods that encourage invariant predictions for different perturbations of unlabeled samples [3]. We, conversely, propose to learn from unlabeled data by generating soft pseudo-labels using the network predictions. We show that a naive pseudo-labeling [2] overfits to incorrect pseudo-labels due to the so-called confirmation bias [3], which stems from using incorrect predictions on unlabeled data for training. This work shows that, contrary to previous attempts on pseudo-labeling [1], simple modifications to prevent confirmation bias lead to state-of-the-art performance without adding consistency regularization strategies.

2. Related work

Previous work on deep SSL differ in whether they use consistency regularization or pseudo-labeling to learn from the unlabeled set. The former imposes that the same sample under different perturbations must produce the same output [3] and the latter seeks the generation of labels or pseudo-labels for unlabeled samples to guide the learning process [1].

3. Pseudo-labeling

We train a CNN using categorical cross-entropy with the one-hot encoding label for the labeled samples and a pseudo-label for the unlabeled samples. We use the network output as soft pseudo-labels. In particular, we use the softmax predictions of the network in every mini-batch of every epoch as a soft pseudo-label. In the first epoch, however, we use the predictions from a model trained in a 10 epochs warm-up phase using the labeled data subset.

3.1 Confirmation bias

It is natural to think that reducing the confidence of the network on its predictions might alleviate confirmation bias and improve generalization. Recently, mixup data augmentation [4] introduced a strong regularization technique that combines data augmentation with label smoothing, which makes it potentially useful to deal with this bias. In the pseudo-labeling context, using soft-labels and mixup reduces overfitting to model predictions. In more extreme cases where few labeled samples per batch are shown, we find that setting a minimum number of labeled samples per mini-batch (as done in other works [3, 1]) provides a constant reinforcement with correct labels during training, reducing confirmation bias.

Table 1: Test error for the proposed approach using the 13-CNN network. Bold indicates lowest error.

	CIFAR-10	CIFAR-100	SVHN	Mini-Imagenet
Labeled images	500	4000	250	4000
Supervised (M)*	37.60 ± 0.65	52.70 ± 0.28	53.15 ± 6.54	72.03 ± 0.21
MT [3]	27.45 ± 2.64	45.36 ± 0.49	4.35 ± 0.50	72.51 ± 0.22
MT-LP [1]	24.02 ± 2.44	43.73 ± 0.20	-	72.78 ± 0.15
LP [1]	32.40 ± 1.80	46.20 ± 0.76	-	70.29 ± 0.81
Ours*	$\mathbf{8.80 \pm 0.45}$	$\mathbf{37.55 \pm 1.09}$	$\mathbf{3.66 \pm 0.12}$	$\mathbf{56.49 \pm 0.51}$

4. Experimental work

We use four image classification datasets, CIFAR-10/100, SVHN and Mini-ImageNet, to validate our approach. We use the "13-CNN" [3] architecture for CIFAR-10/100, and SVHN and ResNet-18 for Mini-ImageNet.

Table 1 shows that our approach outperforms other pseudo-labeling and consistency regularization methods from the state-of-the-art in the four datasets. Our approach still outperformed previous methods when trained with other levels of labeled data, proving its robustness to different scenarios.

5. Conclusions

This paper presented a semi-supervised learning approach for image classification based on pseudo-labeling. We proposed to directly use the network predictions as soft pseudo-labels for unlabeled data together with effective regularization techniques to alleviate confirmation bias. This conceptually simple approach outperforms related work in four datasets, demonstrating that pseudo-labeling is a suitable alternative to the dominant approach in recent literature: consistency-regularization.

References

[1] A. Iscen, G. Tolias, Y. Avrithis, and O. Chum. Label Propagation for Deep Semi-supervised Learning. In *CVPR*, 2019.
[2] A. Oliver, A. Odena, C. Raffel, E. Cubuk, and I. Goodfellow. Realistic Evaluation of Deep Semi-Supervised Learning Algorithms. In *NeurIPS*, 2018.
[3] A. Tarvainen and H. Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 2017.
[4] H. Zhang, M. Cisse, Y. Dauphin, and D. Lopez-Paz. mixup: Beyond Empirical Risk Minimization. In *ICLR*, 2018.

Towards Explaining Deep Neural Networks through Graph Analysis

Vitor A. C. Horta, Suzanne Little, Alessandra Mileo
Insight Centre for Data Analytics at Dublin City University
vitor.horta@insight-centre.org

1. Introduction

Deep learning is being used across many different areas due to its potential to solve complex tasks. However, the resulting complexity of neural networks makes it difficult to explain the whole decision process used by the model, which makes understanding deep learning models an active research topic.

Some of the challenges involved in this task are the different and complex architectures of deep networks and the difficulty in extracting knowledge from neurons in hidden layers. In this work, we address this issue by extracting the knowledge acquired by trained Deep Neural Networks (DNNs) and representing this knowledge in a graph. Our hypothesis is that knowledge contained in the proposed graph is compatible with knowledge acquired by the DNN and by using graph analysis tools we can gain insights on how the model works.

2. Method

In this work, a novel way to extract and represent knowledge from trained DNNs is proposed. The proposed method extracts knowledge from a DNN and represents it as a graph, which is called a co-activation graph. In the co-activation graph, nodes represent neurons in a DNN and weighted relationships indicate a statistical correlation between their activation values. Thus, it is a representation that connects neurons in any layer of the DNN, including hidden (convolutional and dense) and the output layer.

Given a trained DNN, we can generate a co-activation graph using three steps: *(i) Extract activation values; (ii) Define and calculate edge weights; (iii) build and analyse the co-activation graph.*. Once a co-activation graph is created, relationships between pairs of neurons in any depth of hidden layers and neurons in the output classes can be analysed using graph analysis methods. Our goal is to evaluate whether such analyses over this graph can lead to interesting insights on the inner workings of the deep learning model.

3. Experiments and Results

To evaluate whether the proposed method can help understanding how deep learning models work, three experiments were conducted on image classification. The used datasets were: *MNIST-handwritten [2], MNIST-fashion [4], CIFAR-10 [1].* For the first two datasets, we trained the respective convolutional models from scratch. For the third dataset we used a pre-trained state-of-the-art model, MobileNetV2 [3], containing 19 layers and separated convolutional layers with pointwise and depthwise convolutions. This model was chosen to check whether the consistency and flexibility of our approach when applied deeper models and different architectures. After building a co-activation graphs for each of the above deep learning models, we analysed these graphs using different methods to extract knowledge from their respective models. Three main results obtained from our analyses were:

Table 1: Classes and their communities in co-activation graphs.

	Fashion	Handwritten Digits
Community	Classes	Classes
C1	T-shirt/Top; Pullover; Coat; Shirt	Deer;Dog;Horse;
C2	Trouser; Dress;	Frog;Bird;Cat
C3	Sandal; Sneaker; Bag; Ankle Boot	Airplane; Ship; Truck; Automobile
Modularity	0.413	0.489

- Community detection methods over this graph can reveal which classes tare more similar from the DNN point of view. These communities also indicate classes with high semantic similarity, as shown in Table 1.
- Highly overlapping classes in the graph are responsible for most mistakes in the model.
- Central nodes in the graph represent the most important neurons in the respective deep neural network.

The achieved results point to the feasibility of our approach, indicating that, by using graph analysis over a co-activation graph, it is possible to gain relevant insights on how the respective deep learning model works.

4. Conclusion

In this work, we proposed a novel approach to analyse and explain the inner workings of deep learning models. For future work, we plan to explore how to include domain knowledge from external knowledge bases into our approach. In addition, we believe that our approach could be combined with other explainability methods, such as visualization techniques, in order to achieve more robust explanations for DNNs.

References

[1] A. Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.
[2] Y. LeCun and C. Cortes. MNIST handwritten digit database. 2010.
[3] M. Sandler, A. G. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, pages 4510–4520. IEEE Computer Society, 2018.
[4] H. Xiao, K. Rasul, and R. Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

UnSupervised PulseNet: K-Means guided Pruning of Convolutional Neural Networks

David Browne
University College Cork
david.browne@insight-centre.org

1. Introduction

The ability to run start-of-the-art deep Convolutional Neural Network (CNN) on restricted mobile devices, while maintaining their performance is desirable, but due to its size and computational expense not often possible. When a CNN is first fully trained, it contains a significant number of redundant filters and nodes within it. Pruning the network of these redundant elements helps to reduce space and complexity. Our method , unsupervised PulseNet, performs this iteratively, but unlike most approaches which need a pre-determined pruning parameter, we use unsupervised K-means to find a *good* number of clusters, and prune the redundant ones.

2. Related Work

Crowley et al [2] examine the effects of pruning a network followed by fine-tuning, the approach by our work, to pruning the network followed by retraining from scratch. Browne et al [1] and Zou et al [4] both use a simplistic pruning method that analyzes the networks feature maps. When the network is fully trained Browne et al used an L1 magnitude of each feature map to rank their corresponding filter, removing a predefined number of the lowest ranking ones. While Zou et al also used feature maps, their pruning metric was linear discriminant analysis. Both methods retrained the pruned networks to retrieve as much accuracy as possible. He et al [3] argued that instead of using smaller L1 norm as being the less important filters to prune, a better metric would be the geometric median. The geometric median is a well-known robust estimator of centrality in Euclidean space, and [3] uses it to get similar information within a layer.

3. Results

We show how PulseNet achieves SOTA results pruning CIFAR10, Table. 1.

Method	# Parameters	Energy (mJ)	ACC
CifarNet			
Original	797962 (100%)	0.06	85.66
PulseNet	148593 (18.62%)	0.04 (1.5)	83.20
AlexNet			
Original	46787978 (100%)	1.11	90.50
PulseNet	316571 (0.68%)	0.12 (9.25)	88.79
VGG16			
Original	33638218 (100%)	0.98	91.85
PulseNet	1590755 (4.73%)	0.12 (8.2)	89.32

Table 1: PulseNet results on CIFAR10 dataset.

4. Unsupervised PulseNet

Unsupervised PulseNet, Fig. 1, we believe is the first CNN pruning method that does not require user input. It iteratively prunes each layer of the network in a coarse, medium and fine manner. First it prunes all layers before extracting and fine-tuning the network. Then we prune the 2 sections individually (convolution and fully-connected layers). Finally, PulseNet prunes each layer of the network and performs finetuning right after it. If it is unable to recover from too harsh a pruning, it returns to its last best state and either prunes another part of the network, or stops pruning.

Algorithm 1 Pulse-Net

Initialize p_Layer_lst = [all,conv,fc,single]
Initialize λ, p_num = 2%, 0
Initialize p_Layer = $p\_Layer\_lst[p\_num]$
Train Network until validation loss convergence
Calculate validation acc and store as best acc
Repeat until Halt:
  For each i in p_Layer
    Cluster each $p\_Layer[i]$ using k-means
    Determine best k using elbow method
    Find Filter v nearest to each k center
    Remove all filters in $p\_Layer[i] != v$
  Extract weights, biases and batch-norm parameters
  Initialize new structure onto smaller network
  Fine-Tune Network till validation loss converges
  Calculate validation acc
  If validation acc - best acc $> \lambda$:
    best acc $\leftarrow$ validation acc
  Else:
    If p_Layer != single[last layer]
      $p\_num+ = 1$
    Else:
      Halt

References

[1] D. Browne, M. Giering, and S. Prestwich. Pulse-net: Dynamic compression of convolutional neural networks. In *IEEE 5th WF-IoT*, pages 346–351. IEEE, 2019.
[2] E. J. Crowley, A. Storkey, and M. O'Boyle. Pruning neural networks: is it time to nip it in the bud? *1810.04622*, 2018.
[3] Y. He, P. Liu, Z. Wang, Z. Hu, and Y. Yang. Filter pruning via geometric median for deep convolutional neural networks acceleration. In *Proc.IEEE CVPR*, pages 4340–4349, 2019.
[4] J. Zou, T. Rui, Y. Zhou, C. Yang, and S. Zhang. Convolutional neural network simplification via feature map pruning. *Computers & Electrical Engineering*, 70:950–958, 2018.

Generating Counterfactual Explanations in Deep Learning

Eoin M. Kenny and Mark T. Keane
University College Dublin
eoin.kenny@insight-centre.org

1. Introduction

The recent explosion in the use of AI systems has raised questions about the ability of these systems to explain their outputs. Counterfactual explanations are a popular method to help with this issue, but no convincing framework has been suggested to guide their usage. This research suggests that by utilizing the existing psychological literature, with generative adversarial networks (GANs), and case-based reasoning (CBR) methods, a framework can be designed to generate better explanations than the state-of-the-art methods.

2. Explanation Justification

Among other things, when generating counterfactuals, humans typically: (i) do so after a bad outcome, (ii) produce *additive* counterfactuals, and (iii) change *exceptional* events to be *non-exceptional* [1]. To introduce this in the current domain (MNIST dataset), counterfactual explanations are only generated: (i) when the model makes incorrect classifications, (ii) with the assistance of a GAN to "add" to the explanation, and (iii) by changing probabilistically low feature values to high.

3. Methodology and Experiment

Typically, CBR systems use the actual training data to produce explanations, but this approach is very limited. To overcome this, we use a GAN to produce a counterfactual **Explanation** image for the **Query** we wish to explain. To do this, it uses a **Target** image for the optimization process (see Fig. 1).

Formally, the GAN G has a latent input vector z and produces an image I, this is connected to the model m (a CNN) that we wish to explain, which produces a set of features x before the output classifier. The query we wish to explain (I_q) produces the features x_q, some of which *negatively contribute* to the classification of the *correct label* and are *probabilistically unlikely* to occur. Let this set of feature values (a.k.a., neuron activations) be $X_q^- = \{x_1^-, x_2^- \dots x_n^-\}$, we modify these to be values of higher probability (based on a KDE of the training data distribution for each neuron) by solving:

$$\arg\max_{x_i^-} \int_{x_i^- - e}^{x_i^- + e} f(x_i^-)dx$$

where e is normalized as one percent of the total range for the PDF f. Testing this at every value of a continuous distribution is intractable, so we discretize the search space to again be across one hundred steps. we find the closest real example (i.e., the *target*) to this updated vector of x_q using its L_2 distance to the training data's latent features. Subsequently, let the nearest neighbor's latent features be x_t, we can then use this for the optimization of z_q (i.e., the input to G which produces $\sim I_q$) towards an explanation image I_e by solving:

$$z_e = \arg\min_{z_q} \left\| m(G(z_q)) - x_t \right\|_2^2$$

where z_e is the latent representation needed to generate the *explanation* image (see Fig. 1).

4. Results

Fig. 1 shows an explanation for a *query* of a number 7 misclassified as 9. The *target* image features become visible in the *explanation* image as the optimization completes. Clearly, the *query* can "adopt" certain features of the *target* as it optimizes to the *explanation*.

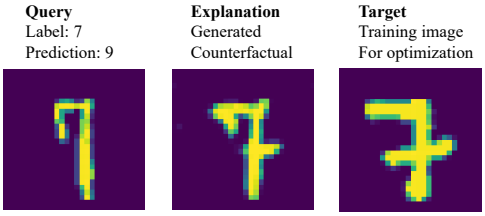


Figure 1: Here a *query* of a 7 is misclassified as a 9. The *explanation* says "If only the query looked more like this, I would have thought it was a 7 instead of a 9". The *target* is used for the optimization, and its features (such as the dash though the 7) are adopted by the *explanation*.

4. Future Work

Future work will test more complex datasets such as CelebA and ImageNet. Moreover, we would like to investigate the usage of Monte Carlo Dropout to gauge model uncertainty in classifications and guide the usage of similar explanations when the model is "lucky" to be correct, which is the other situation humans tend to create counterfactual explanations [1]. Finally, there has been no attention towards semi-factual explanations in deep learning, hence we would like to use this framework to investigate such explanations for good (e.g., the model is correct) and bad (e.g., the model is incorrect) outcomes.

8. References

[1] Byrne, R.M., 2019. Counterfactuals in explainable artificial intelligence (XAI): evidence from human reasoning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence.*

Bivariate Gamma Mixture of Experts Models for Joint Insurance Claims Modelling

Sen Hu, T. Brendan Murphy, Adrian O'Hagan
Insight Centre for Data Analytics, University College Dublin
sen.hu@insight-centre.org

1. Introduction

In general insurance, risks from different categories are often modelled independently and their sum is regarded as the total risk the insurer takes on in exchange for a premium. The dependence from multiple risks is generally neglected even when correlation could exist. It is desirable to take the covariance of different categories into consideration in modelling in order to better predict future claims and hence allow greater accuracy in ratemaking. In this work multivariate severity models are investigated using mixture of experts models with bivariate gamma distributions, where the dependence structure is modelled directly using a GLM framework, and covariates can be placed in both gating and expert networks. Furthermore, parsimonious parameterisations are considered, which leads to a family of bivariate gamma mixture of experts models. It can be viewed as a model-based clustering approach that clusters policyholders into sub-groups with different dependencies, and the parameters of the mixture models are dependent on the covariates. Clustering is shown to be important in separating the data into sub-groupings where strong dependence is often present, even if the overall data set exhibits only weak dependence. In doing so, the correlation within different components features prominently in the model.

2. Method

In this work we consider the bivariate and multivariate gamma definitions provided by Cheriyan (1941) [1]: let X_1, X_2, X_3 be independent gamma random variables, where $X_i \sim Gamma(\alpha_i, \beta)$ for $i = 1, 2, 3$, with different shape parameters $\alpha_i > 0$ and a common rate parameter $\beta > 0$. Then vector $\boldsymbol{Y}$ is defined as:

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} X_1 + X_3 \\ X_2 + X_3 \end{bmatrix} \sim \text{BG}(\alpha_1, \alpha_2, \alpha_3, \beta).$$

It follows that it has density

$$f(\boldsymbol{y}) = C \int_0^m e^{\beta x_3} x_3^{\alpha_3 - 1} (y_1 - x_3)^{\alpha_1 - 1} (y_2 - x_3)^{\alpha_2 - 1} dx_3.$$

where $m = \min(y_1, y_2), C = \frac{\beta^{\alpha_1+\alpha_2+\alpha_3} e^{-\beta(y_1+y_2)}}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)}$.

The set-up of the mixture of experts (MoE) model is that: let $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n$ be an i.i.d bivariate sample of outcomes from a population. Suppose the population consists of G components. Each component can be modelled by a bivariate gamma distribution $f(\boldsymbol{y}_i|\theta_g)$ with component-specific parameters $\theta_g = \{\alpha_{1g}, \alpha_{2g}, \alpha_{3g}, \beta_g\}$, for $g = 1, \ldots, G$ and $i = 1, \ldots, n$. There are also concomitant covariates $\boldsymbol{w}_i$ available which are used to predict future outcome variables. The observed density is

$$\begin{aligned} p(\boldsymbol{y}_i|\boldsymbol{w}_i) &= \sum_{g=1}^{G} \tau_g(\boldsymbol{w}_{0i}) p(\boldsymbol{y}_i|\theta_g(\boldsymbol{w}_i)) \\ \log(\alpha_{kig}) &= \boldsymbol{\gamma}_{kg}^\top \boldsymbol{w}_{ki}, \quad \text{for} \quad k = 1, 2, 3, \\ \log(\beta_{ig}) &= \boldsymbol{\gamma}_{4g}^\top \boldsymbol{w}_{4i}, \end{aligned}$$

and τ_g is the mixing proportion. Different concomitant covariates $\boldsymbol{w}_i$ can go to different parts of the regression models for different parameters. $\tau_g(\boldsymbol{w}_{0i})$ is called the gating network and $p(y_{1i}, y_{2i}|\theta_g(\boldsymbol{w}_i))$ is called the expert network. When the mixing proportion is regressed on covariates, it is typically modelled using a multinomial logistic regression. The expert network $p(y_{1i}, y_{2i}|\theta_g(\boldsymbol{w}_i))$ is typically modelled via a GLM framework with a log link function.

3. Results and conclusions

When applying the MoE model to an Irish insurer data set, investigating the dependency between accidental damage and third party property damage, the clustering result is shown in Figure 1. We conclude that when covariance is considered in modelling by employing such bivariate gamma distributions, claim heterogeneity can be identified and claim predictions could be improved.

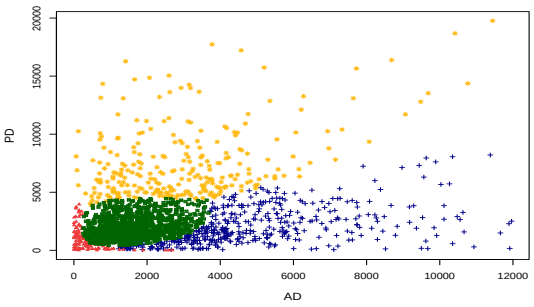


Figure 1: Classification plot when clustering accidental damage (AD) and property damage (PD) without covariates.

References

[1] K. Cheriyan. A bivariate correlated gamma-type distribution function. *Journal of the Indian Mathematical Society*, 5:133–144, 1941.

Unstructured Pruning to Reduce the Cost of Training Deep Neural Network

Camille Ballas
Dublin City University
camille.ballas@insight-centre.org

1. Introduction

Network pruning [3, 6, 7, 8, 10] aims at reducing the number of parameters of Deep Neural Networks while maintaining the original accuracy. By cutting down both its memory footprint and computational cost, punning enable to reduce the inference cost of such models and accelerate their deployment on limited computational resources devices, such as mobile phone or low power chips. The goal is to find which ones of those connections are important for the task, and thus should be kept, and which ones can be safely removed without hurting the performances of the network.

However, training deep neural network remains a heavy procedure, that is *time* and *computationally* expensive. My research is looking at applying pruning methods as early as possible during the training phase of deep networks to provide more sustainable AI with model that are *faster* and *lighter* to train.

2. Problem formulation

Unstructured pruning, as in [6], consists of cutting-off individual connections from the neural network by removing any parameters from the model, i.e. setting some weights to zero based on an oracle. Consider a Neural Network with parameters $\boldsymbol{\theta} \in \mathbb{R}^N$. In unstructured pruning, one want to sparsify the network, so that a fraction κ of elements in $\boldsymbol{\theta}$ are zero, while not changing too much the function computed by the network. This can be framed as the following optimization problem

$$\begin{aligned} \underset{\Delta\boldsymbol{\theta}}{\text{minimize}} \quad & D(f(\boldsymbol{\theta}+\Delta\boldsymbol{\theta}), f(\boldsymbol{\theta})) \\ \text{subject to} \quad & \frac{1}{N}\|\Delta\boldsymbol{\theta}\|_0 = \kappa \\ \text{and} \quad & \Delta\boldsymbol{\theta}_n \in \{-\boldsymbol{\theta}_n, 0\} \quad \forall n \in 1..N \end{aligned} \tag{1}$$

where $D(f(\cdot), f(\cdot)$ is a distance measure between two neural network, κ is the desired sparsity, ratio of parameters to be removes, and $\|\boldsymbol{x}\|_0$ is the L0 norm, i.e. the number of non-zero elements in $\boldsymbol{x}$.

One common way of solving pruning problem, introduced by [8], is to approximate D using a Taylor Series expansion, usually up to the second order, leading to the corresponding formulation:

$$\Delta\mathcal{L}(\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{\theta}+\Delta\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}) =$$

$$\underbrace{\frac{\partial\mathcal{L}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}^\top \Delta\boldsymbol{\theta}}_{A} + \underbrace{\frac{1}{2}\Delta\boldsymbol{\theta}^\top \mathbf{H}(\boldsymbol{\theta})\Delta\boldsymbol{\theta}}_{B} + \underbrace{\mathcal{O}(\|\Delta\boldsymbol{\theta}\|^3)}_{C} \tag{2}$$

The equation is composed of 3 terms: the first-order term (A), corresponding to *magnitude pruning* [6], the second order term (B), corresponding to pruning based on the Hessian or Fischer information, and the higher order terms (C) usually neglected.

3. Method and Challenges

Traditional pruning strategy consists of three stages: 1) train the full-sized model to convergence, 2) remove parameters from the model based on a certain criterion (*e.g.* the magnitude of the weights) and 3) fine-tune the pruned model for a few epochs to recover the original accuracy.

When pruning during the training phase, it is important to understand when it is reasonable to prune the network, as well as the quantity of parameters we should remove, and if we should remove a large quantity at once (*one-shot pruning*) or small quantities multiple times (*iterative pruning*).

To answer those questions, we are conducting an intensive empirical analysis using different pruning approaches over multiple convolutional network architectures and datasets with different complexity to understand the limitation of different unstructured pruning methods which often doesn't scale well.

4. Related Work

Deep learning models are by default over-parametrised. A common belief behind over-parametrisation is that starting with a large capacity model is important as it provide stronger learning representation and it is easier to optimise [11, 1, 12]. However, recent work have demonstrated that sparse models can learn perfectly well when train from scratch suggesting that large over-parametrisation is not necessity to train deep network [10, 4].

However, only a few work has been looking at reducing the size of the model while training [9] and recent study have shown that sparse network doesn't usually scale well [2, 5]. Applying unstructured pruning during the training of deep networks remain an open question.

5. Future Work

To prune the model during the training, a good understanding of both, the training and the pruning methodology is required. Over the past year, we have conducted an intensive study over different pruning criterion. Immediate future work will focus on understanding when to prune and how much to prune comparing one-shot pruning and iterative pruning strategy.

References

[1] M. A. Carreira-Perpinan and Y. Idelbayev. "Learning-Compression" Algorithms for Neural Net Pruning. 2018.

[2] U. Evci, F. Pedregosa, A. N. Gomez, and E. Elsen. The difficulty of training sparse neural networks. *CoRR*, abs/1906.10732, 2019.

[3] J. Frankle and M. Carbin. The lottery ticket hypothesis: Training pruned neural networks. *CoRR*, abs/1803.03635, 2018.

[4] J. Frankle and M. Carbin. The Lottery Ticket Hypothesis: Training Pruned Neural Networks. *arXiv*, 2018.

[5] T. Gale, E. Elsen, and S. Hooker. The state of sparsity in deep neural networks. *CoRR*, abs/1902.09574, 2019.

[6] S. Han, H. Mao, and W. Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. 10 2016.

[7] B. Hassibi, D. G.Stork, and G. Wolff. Optimal brain surgeon: Extensions and performance comparison. pages 263–270, 01 1993.

[8] Y. Lecun, J. Denker, and S. Solla. Optimal brain damage. volume 2, pages 598–605, 01 1989.

[9] N. Lee, T. Ajanthan, and P. H. S. Torr. SNIP: single-shot network pruning based on connection sensitivity. *CoRR*, abs/1810.02340, 2018.

[10] Z. Liu, J. Li, Z. Shen, G. Huang, S. Yan, and C. Zhang. Learning efficient convolutional networks through network slimming. *CoRR*, abs/1708.06519, 2017.

[11] J. H. Luo, J. Wu, and W. Lin. ThiNet: A Filter Level Pruning Method for Deep Neural Network Compression. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[12] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. 11 2016.

List of Keywords

3D, 46

A priori rate-making, 69
Absence of knowledge, 15
Active Learning, 22, 49
Active learning, 48
Adaptation approaches, 50
Aging, 52
Agri-Data, 13
AlexNet architecture, 28
Algorithm configuration, 53
Annex 1 habitats, 37
Anomaly Detection, 19
Architecture, 51
Artificial intelligence, 20
ASD, 26
Asynchronous, 36
Attention BLSTM, 45
Audio and Speech, 11
Autism, 26
Autonomous Vehicles, 10

Bacteria, 62
Bagging, 30
Bayesian Modelling, 52
Benchmark, 14
Bias in the data, 22
Bio-Impedance, 39
Biomechanics, 32
Biomedical Applications, 39
Bivariate gamma distributions, 69
Blended Control, 54
Blockchain, 60
Boosting, 30
Brownian motion, 14

C3D, 45
Caching, 41
Cartoon, 46
Chatbot, 6
Children, 26
Classifiers comparison, 7
Clogging, 35
Code mixing, 9
Cognitive science, 61
Colourimetric Sensors, 34
Community Finding, 49
Complex event processing, 51
Composite risk variables, 57
Computational linguistics, 2
Computer Vision, 43
Computer vision, 44, 47, 65
Conceptual Consensus, 60
Convolution Neural Networks, 67
Convolutional neural networks, 20

Data analytics, 19
Data cubes, 12
Data Integration, 13
Data linking, 2
Data Mining, 18
Data preprocessing, 12
Data Science, 19
Data warehousing, 12
Decision Making, 52
Decision optimisation, 58
Deep learning, 1, 4–6, 20, 28, 33, 43, 64–66, 68, 70
Deep learning theory, 70
Deep Reinforcement Learning, 54
Diachronic word embeddings, 8
Dialogue agents, 6
Discriminative Validity, 38
Disease model, 29
Distributed Clustering, 36
Distributed Graph Traversal, 41
Distributed Paths, 41
Drainage Mapping, 17
Dynamic time warping, 18
Dynamic word embeddings, 8

Earth Observation, 17
Electrical Impedance Tomography, 35
Electrical Resistance Tomography, 35
Electromyography, 27
Electronic Dataset Creation., 4
Electronic lexicography, 2
Embeddings, 8
Emotion detection, 26
Energy efficiency, 51
Ensemble classifiers, 7
Ensemble Methods, 30
Entity Management, 60
Environment, 29
ETL, 12
Evaluation, 23
Exercise Recognition, 28
Explainable AI, 66, 68
Explanation, 61

Facial expression recognition, 47

Fact Checking, 1
Fake News, 1
Financial mathematics, 59
Flow Control and Actuators, 16
Foot Position, 32
Forecasting, 21
Forward Lunge, 38
Fuzzy constraint, 58

GAN, 46
Gas Chromatography, 62
General insurance claim modelling, 69
Generalised linear model, 69
GeoAI, 20
Graph, 13
Graph analysis, 66
Grouping Genetic Algorithms, 55

Habitat conservation, 37
Habitats Directive, 37
Historical languages, 8
Human Activity Recognition, 30
Human Sensing, 30
Hydrological cycle, 63
Hyper-parameters, 53

Image Classification, 67
Image Processing, 43
Image processing, 47
Inertial sensing, 31
Inertial Sensors, 30
Informal Settlements, 44
Injury, 32
Interpretable AI, 68
Interpretable radiomics, 57
Intervention, 22
Intra-Session Reliability, 38
Intrusion Detection Systems, 19

Juridical decisions, 7

Knee condition assessment, 27
Knowledge Distillation, 64
Knowledge Graph, 42
Knowledge Graphs, 40
Knowledge graphs, 6
Knowledge Representation, 42
Knowledge Transfer, 64

Language change, 8
Language modelling, 1, 8, 9
Long Short-Term Memory models (LSTMs), 63
LSTMs, 10

Machine Learning, 4, 5, 17, 22, 30, 35, 43, 68
Machine learning, 7, 26, 37, 44, 53, 58
Machine Translation, 3
Marathon pacing, 25
Marathon running, 25
Mass spectrometry, 62
Mastitis in cows, 18
Microfluidics, 16
Microservices, 51
Mixed effects model, 56
Mixture of experts model, 69
Mobility Mining, 15
Model Efficiency, 67
Model-based clustering, 69
Motion analysis, 27
MPEG-DASH, 50
Multimedia, 51
Multimodal emotion detection, 26
Multimodality, 5

Natural Language Processing, 1
Natural language processing, 2, 7, 8
Navicular drop, 32
Network Compression, 67
Neural Network compression, 70
Neural Networks, 63, 64
Neural networks, 26
News, 46
Non invasive, 62

Object Detection and Tracking, 43
Offensive Language Detection, 5
Offline Evaluation, 22
Ontologies, 40
Optimisation, 53
Orthography, 3

Pairwise constraints, 49
Path Prediction, 10
Path Query, 41
Peatlands, 17
Peatlands Landuse, 17
Personal sensing, 31
pH, 34
Photoplethysmography, 33
Physiotherapy, 32
Prediction, 35
Prediction systems, 63
Predictive modelling, 57
Preference elicitation, 48
Privacy, 19
Pruning Networks, 67
Pseudo-labelling, 65
Psycholinguistics, 1
Pump, 35

Quality of Experience (QoE), 11
Query processing, 12

RDBMS, 42
RDF, 41
Real-time, 51
Recommendation Systems, 23
Recommendation systems, 24
Recommender systems, 22, 25
Recovery monitoring, 27
Regression, 56
Rehabilitation, 27
Reinforcement Learning, 23, 29
Remote sensing, 20, 37
Repetition Counting, 28
Replicating portfolio, 59
Representation learning, 65
Reproducibility, 23
Rule based approach, 47
Running, 31, 32

S&P 500, 59
Sampling, 55
Satellite images, 44
Segment duration, 50
Self-adaptability, 51
Semantic Web, 40
Semi supervised learning, 65
Semi-supervised learning, 49
Sentiment, 61
Skin, 34
Smart-Agriculture, 21
Social Media, 9
Socio-technical systems, 24
Solid phase micro extraction, 62
SPARQL 1.1 Property Path., 41
Sports analytics, 25
Sports Science, 32
Statistical learning, 57
Stimuli Responsive Materials, 16
Stratification, 55
Stylometry, 1
Subjective Experiment, 11
Suggestions, 22
Supplier selection, 48
Symbolic Representation, 18

Task Environment, 23
Thin film structure, 35
Time and Route Inference, 15
Time Series, 10
Time series characteristics, 14
Time series classification, 18
Time series construction, 14
Time series prediction, 14
Time-series, 21
Tobit Regression, 56
Tolerance interval, 56
Traffic Scenes, 10
Transfer Learning, 5, 64
Transfer learning, 33

Under Resource Languages, 4
Under-resourced languages, 3, 8, 9
Unknown Fault Tolerant Control, 54
Unmanned Aerial Vehicles, 37
User-driven design, 24

Valence, 61
Variance swap, 59
Video, 46
Video captioning, 45
VIX, 59
VOC, 62

Water Levels, 63
Wearables, 31, 39
Wireless Sensor Networks, 36