2nd International Workshop on Mining Urban Data (Preface)

Ioannis Katakis KATAK@DI.UOA.GR
National and Kapodistrian University of Athens, Greece

François Schnitzler FRANCOIS@EE.TECHNION.AC.IL
Technion - Israel Institute of Technology, Haifa, Israel

Thomas Liebig THOMAS.LIEBIG@TU-DORTMUND.DE
University of Dortmund, 44221 Dortmund, Germany

Abstract

This paper presents an overview of the second International Workshop on Mining Urban Data (MUD2). The MUD2 workshop was held in conjunction with the 32nd International Conference on Machine Learning (ICML 2015) in Lille, France, July 11, 2015.

1. Introduction

We are gradually moving towards a smart city era. Many innovative applications arise daily utilizing massive urban data streams. Technologies that apply machine learning algorithms to urban data will have significant impact in many aspects of the citizens' everyday life. Examples of such applications include managing disastrous events, understanding the city's sentiment and opinion, tracking health issues, monitoring crucial environmental factors as well as improving energy efficiency and optimizing traffic. Unfortunately, urban data have some characteristics that hinder the state of the art in machine learning algorithms. Such are diversity, privacy, distributed and partitioned data, lack of labels, noise, complimentary of multiple sources and requirement for online learning. Many smart city applications require to tackle all these problems at once. This workshop aimed at discussing a set of new Machine Learning applications and paradigms emerging from the smart city environment. MUD2 will focus on presenting novel approaches that target some of the following applications: a) Traffic Management, b) Public Transport Adjustment, c) Accident Prevention, d) Resource Allocation, e) Energy Efficiency, f) Sentiment Analysis, g) Environment.

History A successful first edition of the MUD workshop was organized and co-located with EDBT/ICDT conference in March 24-28, 2014[1]. During the workshop 15 full and short papers were presented (50% full paper acceptance ratio). Approximately 30 people attended the workshop. As a follow-up, a special issue on Mining Urban Data has been organized (currently in the review phase) at the Journal of Information Systems[2]. 43 papers were submitted to the Special Issue. Workshops with a similar focus have been held recently: senseML, at ECML/PKDD 2014, and SenseMine, at ACM SenSys 2013. Finally, a workshop series dedicated to Computational Transportation Science is also organized.

2. Scope

The second version of MUD aimed at raising the awareness of the Machine Learning community on the challenges and opportunities of the Urban Data research arena. Mining Urban Data is a multidisciplinary field. The Data Management and Knowledge Discovery communities have already started working towards this direction (see first version of MUD at EDBT/ICDT conference[3]), however even though there are some first efforts from Machine Learning perspective, a greater ML involvement is required. ML will contribute to a better understand-

Proceedings of the 2nd International Workshop on Mining Urban Data, Lille, France, 2015.

[1] http://www.insight-ict.eu/mud/
[2] http://goo.gl/vUejYf
[3] http://www.insight-ict.eu/mud

ing and long term prediction of the urban sensor-citizen environment. The engagement of the community was achieved by inviting researchers and industries to present their work in the workshop. The workshop was open, but not limited to, the following topics:

Online learning - data generated from sensors can only partially/temporarily be stored. Thus, a major requirement is to process and analyse them as they arrive from the sources. Algorithms should be on-line and adaptive.

Large Scale Learning - the massive volume of data demands distributed / parallel processing technologies. Other issues include the complexity of the data coming from different sources with different spatial and temporal references or granularity.

Learning in Mobile Environment - special techniques are required for storing and learning in mobile environments.

Heterogeneous Data and Information Fusion - in many smart-city applications, different types of information (GPS, weather, Twitter, traffic data) should be analysed and combined in order to draw conclusions.

Learning with Social media - the main issue in mining micro-blogging data (e.g. Twitter) is that the text is very short, cursorily written and in different languages.

Event Detection - a very interesting research issue that arises from such data is the identification of real world events (e.g. "traffic jam", "accident", "flood").

Learning with Uncertain/Noisy Data - data generated by a smart city are typically very noisy. Uncertainty management procedures as well as crowdsourcing techniques might be required in order to aid the data models disambiguate the information.

Learning without Labels - with the size of the data sets and the associated area, labeling the full data set can be prohibitively expensive. Therefore, learning must typically be done with originally no or extremely few labels. Semi-supervised or active learning approaches could be very interesting for such applications.

Computer vision - CCTV cameras are a rich source of information. They can be used to count pedestrians, detect accidents, security etc.

3. Data Sets

We especially welcomed contributions based on data that can be reused by the community. Some published data sets are the following (the list is incomplete - please refer to the web site):

Dublin Bus GPS sample data from Dublin City Council. This data set comprises GPS traces of public buses within Dublin for a period of 1 month.
http://dublinked.ie/datastore/datasets/dataset-304.php

Traffic Volume Data for Dublin City. This data set originates from the traffic management system installed at Dublin city and contains performance values (traffic flow and saturation) at about 900 junctions for a period of 3 months.
http://dublinked.ie/datastore/datasets/dataset-305.php

Sales Prices in Helsinki. The dataset records advertised sales prices of 8337 residential apartments in Helsinki region along with features describing real estate characteristics, location characteristics, and accessibility in terms of travel distances and travel times. For more details see the paper Zliobaite et al (2015) published at the MUD2 workshop.
http://www.zliobaite.com/datahel.zip

An updated list of available data sets is available at the workshop website[4].

4. Invited Talks

This year, we welcome three invited speakers at the MUD workshop:

- Dr. Eleni Pratsini – Lab Director Smarter Cities Technology Center, IBM Research, Ireland, *"Using Big Mobile Data to Analyze Social Events in Cities"*
- Prof. Kristian Kersting – Fraunhofer IAIS and Technical University of Dortmund, Germany, *"Poisson Dependency Networks: Gradient Boosted Models for Multivariate Count Data"*
- Prof. Sharad Mehrotra – University of California, Irvine, USA *"Towards 'on the fly' data cleaning"*

5. Submissions and acceptance

We received 18 submissions from 11 countries. The number of authors per country is depicted in Figure 1. 15 submissions were accepted, 9 for regular talks and 6 for short talks. In the following, we present an overview of the submissions, grouped by topic.

5.1. Traffic

- Distributed Traffic Flow Prediction with Label Proportions: From in-Network towards High Per-

[4] http://www.insight-ict.eu/mud2/data.html

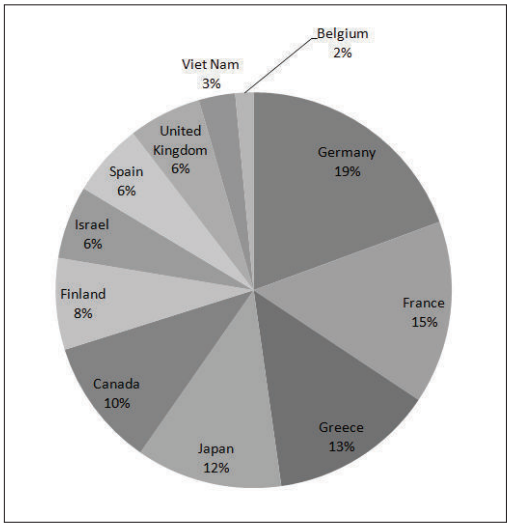

Figure 1. Distribution of MUD2 authors per country.

formance Computation with MPI, *Thomas Liebig, Marco Stolpe and Katharina Morik*

- Towards detection of faulty traffic sensors in real-time, *Nikolas Zygouras, Nikolaos Panagiotou, Nikos Zacheilas, Ioannis Boutsis, Vana Kalogeraki, Ioannis Katakis and Dimitrios Gunopulos*
- Car-traffic forecasting: A representation learning approach *Ali Ziat, Gabriella Contardo, Nicolas Baskiotis and Ludovic Denoyer*

5.2. Social Media

- Event-based Clustering for Reducing Labeling Costs of Incident-Related Microposts, *Axel Schulz, Petar Ristoski, Johannes Fürnkranz and Frederik Janssen*
- Modelling Time and Location in Topic Models, *Christian Pölitz*
- Stresscapes: Validating Linkages between Place and Stress Expression on Social Media, *Martin Sykora, Colin Robertson, Ketan Shankardass, Rob Feick, Krystelle Shaughnessy, Becca Coates, Haydn Lawrence and Thomas W. Jackson*

5.3. Trip Planning

- Automatic Extrapolation of Missing Road Network Data in OpenStreetMap, *Stefan Funke, Robin Schirrmeister and Sabine Storandt*
- Improved Trip Planning by Learning from Travelers' Choices, *Boris Chidlovskii*
- Report from Dagstuhl: SocioPaths - Multimodal Door-to-Door Route planning via Social Paths, *Thomas Liebig, Sabine Storandt, Peter Sanders, Walied Othman and Stefan Funke*

5.4. Bicycles and Public Transport

- Profiling users of the Velo'v bike sharing system, *Albrecht Zimmermann, Mehdi Kaytoue, Marc Plantevit, Céline Robardet and Jean-François Boulicaut*
- Accessibility by public transport predicts residential real estate prices: a case study in Helsinki region, *Indrė Žliobaitė, Michael Mathioudakis, Tuukka Lehtiniemi, Pekka Parviainen and Tomi Janhunen*
- On Predicting Traveling Times in Scheduled Transportation, *Avigdor Gal, Avishai Mandelbaum, François Schnitzler, Arik Senderovich and Matthias Weidlich*

5.5. Open, Mobile Data and Environment

- Analyzing Open Data from the City of Montreal, *Joelle Pineau and Pierre-Luc Bacon*
- Evaluating distance measures for trajectories in the mobile setting, *Nikolaos Larios, Christos Mitatakis, Vana Kalogeraki and Dimitrios Gunopulos*
- Airvlc: An application for real-time forecasting urban air pollution, *Lidia Contreras Ochando, Cristina I. Font Julián, Francisco Contreras Ochando and Cèsar Ferri*

6. Organization

Members of the Organizing Committee were: Ioannis Katakis (*National & Kapodistrian University of Athens*), François Schnitzler, (*Technion*), Thomas Liebig, (*TU Dortmund*), Gennady Andrienko, (*Fraunhofer IAIS and City University London*), Dimitrios Gunopulos, (*National & Kapodistrian University of Athens*), Katharina Morik, (*TU Dortmund*), Shie Mannor, (*Technion*). In addition, we would like to thank Marco Cuturi, workshop chair of ICML 2015 for the excellent collaboration.

Acknowledgemnts

The organizers received funding from the European Union's Seventh Framework Programme under grant agreement number FP7-318225, INSIGHT.