Mining and Model Understanding on Medical Data

Myra Spiliopoulou
Faculty of Computer Science
Otto-von-Guericke-University Magdeburg
Magdeburg, Germany
myra@ovgu.de

Panagiotis Papapetrou
Dept. of Computer and Systems Sciences
Stockholm University
Stockholm, Sweden
panagiotis@dsv.su.se

ABSTRACT

What are the basic forms of healthcare data? How are Electronic Health Records and Cohorts structured? How can we identify the key variables in such data and how important are temporal abstractions? What are the main challenges in knowledge extraction from medical data sources? What are the key machine algorithms used for this purpose? What are the main questions that clinicians and medical experts pose to machine learning researchers?

In this tutorial, we provide answers to these questions by presenting state-of-the-art methods, workflows, and tools for mining and understanding medical data. Particular emphasis is given on temporal abstractions, knowledge extraction from cohorts, machine learning model interpretability, and mHealth.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Applied computing** → **Health care information systems**.

KEYWORDS

medical mining, electronic health records, cohorts, deep learning, interpretability

ACM Reference Format:
Myra Spiliopoulou and Panagiotis Papapetrou. 2019. Mining and Model Understanding on Medical Data. In *The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19), August 4–8, 2019, Anchorage, AK, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3292500.3332274

1 TUTORIAL PERSPECTIVE

Medical research and patient care-taking are increasingly benefiting from advances in machine learning. The penetration of smart technologies and the Internet of Things give a further boost to initiatives for patient self-management and empowerment: new forms of health-relevant data become available and require new data acquisition and analytics workflows. As data complexity and model sophistication increase, model interpretability becomes mission-critical. But what constitutes model interpretation in the context of medical machine learning: what are the questions for which knowledge discovery from data should provide interpretable answers?

In this tutorial, we discuss basic forms of health-related data, including Electronic Health Records, cohort data from population-based studies and clinical studies, mHealth recordings, and data from internet-based studies. We elaborate on the questions that medical researchers and clinicians pose on those data, and on the instruments they use - giving some emphasis to the instruments "population-based study" and "Randomized Clinical Trial". We elaborate on what questions are asked with those instruments, on what questions can be answered from those data, on ML advances and achievements on such data, and on ways of responding to the medical experts questions about the derived models.

2 TARGET AUDIENCE

This tutorial is targeted to all KDD participants interested in the topic of learning and model understanding on medical data. Our particular focus group consists of junior researchers interested in knowledge discovery from health-related data and on how to convey the extracted models to medical experts.

The main prerequisites for the participants concerns basic knowledge within the areas of data mining, machine learning, and databases. The audience is expected to be familiar with standard concepts and methods of machine learning. Such knowledge can be expected from KDD participants, including students.

3 TUTORIAL OUTLINE

The tutorial is structured in six parts:

- **PART 1: Introduction**
  (1) The scope of the tutorial: tutorialists, structure, main topics
  (2) Introductory terms: What are patient data? Electronic Health Records (EHRs), social data, data collected in cohort studies
- **PART 2: EHRs and temporal abstractions**
  (1) Definition and examples of EHRs and EHR systems
  (2) Overview of the usage of EHRs globally
  (3) Predictive models on EHR data
  (4) Dealing with missing values in EHR variables
  (5) Survival analysis in EHRs
- **PART 3: Learning from cohorts**
  (1) Definition and examples of cohorts
  (2) Cohorts for clinical and population-based studies
  (3) Randomized clinical trials (RCTs)
  (4) Expert driven cohort refinement on EHR data
  (5) Cohort alignment for model validation
  (6) Expert inputs and what-if questions to models on cohorts
- **PART 4: Deep learning and interpretability**
  (1) Deep learning architectures for EHRs

KDD '19, August 4–8, 2019, Anchorage, AK, USA

ACM ISBN 978-1-4503-6201-6/19/08.
https://doi.org/10.1145/3292500.3332274

(2) Recurrent Neural Networks for diagnosis prediction
(3) Deep learning with attention mechanisms
(4) Interpretable model-specific methods for EHRs
(5) Interpretable model-agnostic methods for EHRs

- **PART 5: Learning from eHealth and mHealth data**

(1) Using the internet for therapy, the example of eCBT
(2) Potential challenges and pitfalls in mHealth
(3) Momentary assessments and the promise of smart devices
(4) Learning from the data of mobile devices
(5) Monitoring the momentary assessments of patients

- **PART 6: Conclusions**

(1) Summary and challenges in learning
(2) Challenges of small data
(3) Challenges on reliability
(4) Challenges in involving the expert
(5) Challenges in model explainability

4 TUTORIALISTS

Myra Spiliopoulou is Professor of Business Information Systems at the Otto-von-Guericke-University Magdeburg. Her research is on mining dynamic complex data, with focus on healthcare and social data. She is action editor for DAMI. In 2019 she is PC Chair of the IEEE Symposium of Computer Based Medical Systems 2019. In 2018 she was PC Chair of the Applied Data Science Track of KDD 2018. In the recent past, she was one of the four Journal Track Chairs for ECML PKDD 2017. She has held tutorials on topics of data mining at KDD 2009, 2015 and 2018, PAKDD 2013 and 2016, ICDM 2017, and in many ECML PKDD conferences. More details about the tutor can be found at http://www.kmd.ovgu.de/Team/Academic+Staff/Myra+Spiliopoulou.html.

Panagiotis Papapetrou is Professor at the Department of Computer and Systems Sciences at Stockholm University and Adjunct Professor at the Computer Science Department at Aalto University. His area of expertise is algorithmic data mining with particular focus on mining and indexing temporal data and healthcare data. Panagiotis received his PhD in Computer Science at Boston University in 2009, was a post-doctoral researcher at Aalto University during 2009-2013, and lecturer at the University of London during 2012-2013. He has participated in several national and international research projects. He is action editor of DAMI and board member of the Swedish AI Society. He has held tutorials on topics in data mining and healthcare at ECML/PKDD 2016, ICDM 2017, and KDD 2018. More details about the tutor can be found at https://papapetrou.blogs.dsv.su.se.

5 ACKNOWLEDGEMENTS

This material of this tutorial and the preparation was partly supported by the VR-2016-03372 Swedish Research Council Starting Grant, the EXTREME project funded by ICT-TNG, and by grants provided by Stockholm University and Stockholm County Council (SU-SLL).