Andreas Holzinger

Biomedical Informatics

Discovering Knowledge in Big Data

Springer

Andreas Holzinger

Biomedical Informatics

Discovering Knowledge in Big Data

Springer

Andreas Holzinger
Medical University Graz and
Graz University of Technology
Graz, Austria

Additional material to this book can be downloaded from http://extra.springer.com.

ISBN 978-3-319-04527-6 ISBN 978-3-319-04528-3 (eBook)
DOI 10.1007/978-3-319-04528-3
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014932669

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To my family and friends

Preface

Amongst all the challenges which medicine and health care in all parts of our world are facing, including exploding costs, finite resources, aging population, etc., maybe one is *the* grand challenge: How to deal with the increasingly large, high-dimensional, weakly structured, and complex data sets and the growing masses of unstructured information produced by modern biomedical sciences. The future trend towards personalized medicine has resulted in an explosion in the amount of generated biomedical data sets from various sources, for example genomics, proteomics, metabolomics, lipidomics, transcriptomics, epigenetics, microbiomics, fluxomics, and phenomics.

This "Big Data" rapidly exceeds the capacity a single medical doctor may be able to handle manually. In the daily routine today, a medical doctor is no longer able to keep pace with these developments and to remember all the thousands of facts of the many diverse patients she or he has. Although human cognitive capacities are excellent in certain areas, in dealing with complex data sets the human learning is limited to a certain extent. This makes the application and further development of sophisticated machine learning approaches a pressing need.

For quite some time now, I have worked persistently on a synergistic combination of methodologies and approaches of two areas, which offer ideal conditions towards solving these challenges towards finding new, efficient, and user-centered algorithms and tools to deal with the "Big Data" challenge: Human–Computer Interaction (HCI) and Knowledge Discovery and Data Mining (KDD), with the goal of supporting human learning with machine learning to interactively discover new, previously unknown *insights into the data.*

Imagine a future where medical doctors would be able to ask questions (e.g., "What are the similarities/differences between patients with symptom X and patients with symptom Y") to their patient data *and get relevant answers.* The advantageous possibilities are endless if the doctors have an integrated overview on all *relevant* data at their clinical workplace and could find and diagnose diseases well in advance before they might become symptomatically apparent.

This book constitutes the lecture notes to my one-semester course where I try—and one can never be perfect—to provide undergraduate students of biomedical engineering, software engineering, and computer science a broad overview on "Biomedical Informatics" with emphasis on the aforementioned challenges, consequently focusing on a blend of aspects on data, information, and knowledge. Academic freedom lets teach in an egocentric way; consequently, I bring in what I find most interesting and which is amongst my research interests, following a research-based teaching style: I am passionate on extending advanced methods including time (e.g., information entropy) and space (e.g., computational topology), along with user-centered software engineering methods to create interactive software for mobile applications and content analytics techniques, and I follow three promising research streams: Graph-Based Data Mining, Entropy-Based Data Mining, and Topological Data Mining. Naturally, in the limited time of this course I can only scratch the surface and it is impossible to touch each and every interesting aspect—but as a first taster the goal is to open the eyes of the students for the current problems and to encourage them to study further.

Such a foreword is always an opportunity to thank people. This is not an easy task as many people contribute to the development of such a book, either directly or indirectly, and because there are so many, it is always the risk of forgetting somebody, so I will use the plural: I thank my Institutes both at Graz University of Technology and the Medical University of Graz for the academic freedom, the intellectual environment, and the opportunity to teach these fascinating topics to the students. I thank all my colleagues in academia, the clinical domain, and industry for their ongoing and constant fruitful support. I cordially thank all my team and the members from my research group and all my students for their valuable criticism and feedback, and my family and friends for their nurturing encouragement. Last but not least I thank the Springer management and production team for their great and smooth support!

Graz, Austria Andreas Holzinger

Overview

The Journey Through This Course

The goal of this one-semester course is to provide the students with a broad overview on "Biomedical Informatics" with focus on data, information, and knowledge. We will follow the hypothesis that information can bridge the gap between scientific biomedical research and clinical health practice. The course consists of the following 12 lectures (see Slide 0-1):

Slide 0-1: Roadmap Through This Course

The course consists of the following 12 lectures:

1. **Introduction: Computer Science Meets Life Sciences**. We start with the basics of life sciences, including biochemical and genetic fundamentals, some cell-physiological basics, and a brief overview of the human body; we answer the question "what is biomedical informatics," and we conclude with an outlook into the future.
2. **Fundamentals of Data, Information, and Knowledge**. In the second lecture we start with a look on data sources, review some data structures, discuss standardization versus structurization, review the differences between data, information, and knowledge, and close with an overview of information entropy.
3. **Structured Data: Coding, Classification (ICD, SNOMED, MeSH, UMLS)**. In the third lecture we focus on standardization, ontologies, and classifications, in particular on the International Statistical Classification of Diseases, the Systematized Nomenclature of Medicine, Medical Subject Headings, and the Unified Medical Language.

(continued)

(continued)

4. **Biomedical Databases: Acquisition, Storage, Information Retrieval, and Use**. In the fourth lecture we get a first impression of a hospital information system, we discuss some basics of data warehouse systems and biomedical data banks, and we concentrate on information retrieval.
5. **Semi-structured, Weakly Structured, and Unstructured Data**. In the fifth lecture we review some basics of XML, before we concentrate on network theory and discuss transcriptional regulatory networks, protein–protein networks, and metabolic networks.
6. **Multimedia Data Mining and Knowledge Discovery**. In the sixth lecture we determine types of knowledge, focus on the basics of data mining, and close with text mining and semantic methods, such as Latent Semantic Analysis, Latent Dirichlet Allocation, and Principal Component Analysis.
7. **Knowledge and Decision: Cognitive Science and Human–Computer Interaction**. In the seventh lecture we review the fundamentals of perception, attention, and cognition; discuss the human decision-making process, reasoning, and problem solving; and learn some principles of differential diagnosis and a few basics on human error.
8. **Biomedical Decision Making: Reasoning and Decision Support**. In the eighth lecture we start with the question "Can computers help doctors to make better decisions?" and apply the basics from Lecture 7 to the principles of decision support systems and case-based reasoning systems.
9. **Interactive Information Visualization and Visual Analytics**. In the ninth lecture we start with the basics of visualization science; review some visualization methods, including Parallel Coordinates, Radial Coordinates, and Star Plots; and learn a few things about the design of interactive visualizations.
10. **Biomedical Information Systems and Medical Knowledge Management**. In the tenth lecture we discuss workflow modeling, some basics of business enterprise hospital information systems, Picture Archiving and Communication Systems, and some standards, including DICOM and HL-7.
11. **Biomedical Data: Privacy, Safety, and Security.** In the eleventh lecture we start with the famous IOM "Why do accidents happen?" report and its influence on safety engineering, and concentrate on aspects of data protection and privacy issues of medical data.
12. **Methodology for Information Systems: System Design, Usability, and Evaluation**. Finally in the twelfth lecture we slip into the developer perspective and have a look on design standards, usability engineering methods, and on how we evaluate such systems.

Contents

Index

A. Holzinger, *Biomedical Informatics*, DOI 10.1007/978-3-319-04528-3,