Preface

The basic notion of the book "*Data Mining and Medical Knowledge Management: Cases and Applications*" is knowledge. A number of definitions of this notion can be found in the literature:

- Knowledge is the sum of what is known: the body of truth, information, and principles acquired by mankind.
- Knowledge is human expertise stored in a person's mind, gained through experience, and interaction with the person's environment.
- Knowledge is information evaluated and organized by the human mind so that it can be used purposefully, e.g., conclusions or explanations.
- Knowledge is information about the world that allows an expert to make decisions.

There are also various classifications of knowledge. A key distinction made by the majority of knowledge management practitioners is Nonaka's reformulation of Polanyi's distinction between tacit and explicit knowledge. By definition, *tacit knowledge* is knowledge that people carry in their minds and is, therefore, difficult to access. Often, people are not aware of the knowledge they possess or how it can be valuable to others. Tacit knowledge is considered more valuable because it provides context for people, places, ideas, and experiences. Effective transfer of tacit knowledge generally requires extensive personal contact and trust. *Explicit knowledge* is knowledge that has been or can be articulated, codified, and stored in certain media. It can be readily transmitted to others. The most common forms of explicit knowledge are manuals, documents, and procedures. We can add a third type of knowledge to this list, the *implicit knowledge*. This knowledge is hidden in a large amount of data stored in various databases but can be made explicit using some algorithmic approach. Knowledge can be further classified into procedural knowledge and declarative knowledge. *Procedural knowledge* is often referred to as knowing how to do something. *Declarative knowledge* refers to knowing that something is true or false.

In this book we are interested in knowledge expressed in some language (formal, semi-formal) as a kind of model that can be used to support the decision making process. The book tackles the notion of knowledge (in the domain of medicine) from two different points of view: data mining and knowledge management.

Knowledge Management (KM) comprises a range of practices used by organizations to identify, create, represent, and distribute knowledge. Knowledge Management may be viewed from each of the following perspectives:

- **Techno-centric:** A focus on technology, ideally those that enhance knowledge sharing/growth.
- **Organizational:** How does the organization need to be designed to facilitate knowledge processes? Which organizations work best with what processes?

- **Ecological:** Seeing the interaction of people, identity, knowledge, and environmental factors as a complex adaptive system.

Keeping this in mind, the content of the book fits into the first, technological perspective. Historically, there have been a number of technologies "enabling" or facilitating knowledge management practices in the organization, including expert systems, knowledge bases, various types of Information Management, software help desk tools, document management systems, and other IT systems supporting organizational knowledge flows.

Knowledge Discovery or Data Mining is the partially automated process of extracting patterns from usually large databases. It has proven to be a promising approach for enhancing the intelligence of systems and services. Knowledge discovery in real-world databases requires a broad scope of techniques and forms of knowledge. Both the knowledge and the applied methods should fit the discovery tasks and should adapt to knowledge hidden in the data. Knowledge discovery has been successfully used in various application areas: business and finance, insurance, telecommunication, chemistry, sociology, or medicine. Data mining in biology and medicine is an important part of biomedical informatics, and one of the first intensive applications of computer science to this field, whether at the clinic, the laboratory, or the research center.

The healthcare industry produces a constantly growing amount of data. There is however a growing awareness of potential hidden in these data. It becomes widely accepted that health care organizations can benefit in various ways from deep analysis of data stored in their databases. It results into numerous applications of various data mining tools and techniques. The analyzed data are in different forms covering simple data matrices, complex relational databases, pictorial material, time series, and so forth. Efficient analysis requires knowledge not only of data analysis techniques but also involvement of medical knowledge and close cooperation between data analysis experts and physicians. The mined knowledge can be used in various areas of healthcare covering research, diagnosis, and treatment. It can be used both by physicians and as a part of AI-based devices, such as expert systems. Raw medical data are by nature heterogeneous. Medical data are collected in the form of images (e.g. X-ray), signals (e.g. EEG, ECG), laboratory data, structural data (e.g. molecules), and textual data (e.g. interviews with patients, physician's notes). Thus there is a need for efficient mining in images, graphs, and text, which is more difficult than mining in "classical" relational databases containing only numeric or categorical attributes. Another important issue in mining medical data is privacy and security; medical data are collected on patients, misuse of these data or abuse of patients must be prevented.

The goal of the book is to present a wide spectrum of applications of data mining and knowledge management in medical area.

The book is divided into 3 sections. The first section entitled "*Theoretical Aspects*" discusses some basic notions of data mining and knowledge management with respect to the medical area. This section presents a theoretical background for the rest of the book.

Chapter I introduces the basic concepts of medical informatics: data, information, and knowledge. It shows how these concepts are interrelated and how they can be used for decision support in medicine. All discussed approaches are illustrated on one simple medical example.

Chapter II introduces the basic notions about ontologies, presents a survey of their use in medicine and explores some related issues: knowledge bases, terminology, and information retrieval. It also addresses the issues of ontology design, ontology representation, and the possible interaction between data mining and ontologies.

Health managers and clinicians often need models that try to minimize several types of costs associated with healthcare, including attribute costs (e.g. the cost of a specific diagnostic test) and misclassification costs (e.g. the cost of a false negative test). Chapter III presents some concepts related to cost-sensitive learning and cost-sensitive classification in medicine and reviews research in this area.

There are a number of machine learning methods used in data mining. Among them, artificial neural networks gain a lot of popularity although the built models are not as understandable as, for example, decision trees. These networks are presented in two subsequent chapters. Chapter IV describes the theoretical background of artificial neural networks (architectures, methods of learning) and shows how these networks can be used in medical domain to solve various classification and regression problems. Chapter V introduces classification networks composed of preprocessing layers and classification networks and compares them with "classical" multilayer perceptions on three medical case studies.

The second section, "*General Applications*," presents work that is general in the sense of a variety of methods or variety of problems described in each of the chapters.

In chapter VI, biomedical image registration and fusion, which is an effective mechanism to assist medical knowledge discovery by integrating and simultaneously representing relevant information from diverse imaging resources, is introduced. This chapter covers fundamental knowledge and major methodologies of biomedical image registration, and major applications of image registration in biomedicine.

The next two chapters describe methods of biomedical signal processing. Chapter VII describes methods for preprocessing, analysis, feature extraction, visualization, and classification of electrocardiogram (ECG) signals. First, preprocessing methods mainly based on the discrete wavelet transform are introduced. Then classification methods such as fuzzy rule-based decision trees and neural networks are presented. Two examples, visualization and feature extraction from body surface potential mapping (BSPM) signals and classification of Holter ECGs, illustrate how these methods are used. Chapter VIII deals with the application of principal components analysis (PCA) to the field of data mining in electroencephalogram (EEG) processing. Possible applications of this approach include separation of different signal components for feature extraction in the field of EEG signal processing, adaptive segmentation, epileptic spike detection, and long-term EEG monitoring evaluation of patients in a coma.

In chapter IX, existing clinical risk prediction models are examined and matched to the patient data to which they may be applied, using classification and data mining techniques, such as neural Nets. Novel risk prediction models are derived using unsupervised cluster analysis algorithms. All existing and derived models are verified as to their usefulness in medical decision support on the basis of their effectiveness on patient data from two UK sites.

Chapter X deals with the problem of quality assessment of medical Web sites. The so called "quality labeling" process can benefit from employment of Web mining and information extraction techniques, in combination with flexible methods of Web-based information management developed within the Semantic Web initiative.

In medicine, doctors are often confronted with exceptions both in medical practice or in medical research; a proper method of how to deal with exceptions are case-based systems. Chapter XI presents two such systems. The first one is a knowledge-based system for therapy support. The second one is designed for medical studies or research. It helps to explain cases that contradict a theoretical hypothesis.

The third section, "*Specific Cases*," shows results of several case studies of (mostly) data mining, applied to various specific medical problems. The problems covered by this part range from discovery of biologically interpretable knowledge from gene expression data, over human embryo selection for the purpose of human in-vitro fertilization treatments, to diagnosis of various diseases based on machine learning techniques.

Discovery of biologically interpretable knowledge from gene expression data is a crucial issue. Current gene data analysis is often based on global approaches such as clustering. An alternative way is to utilize local pattern mining techniques for global modeling and knowledge discovery. The next two chapters deal with this problem from two points of view: using data only, and combining data with domain knowledge. Chapter XII proposes three data mining methods to deal with the use of local patterns, and chapter XIII points out the role of genomic background knowledge in gene expression data mining. Its application is demonstrated in several tasks such as relational descriptive analysis, constraint-based knowledge discovery, feature selection, and construction or quantitative association rule mining.

Chapter XIV describes the process used to mine a database containing data related to patient visits during Tinnitus Retraining Therapy.

Chapter XV describes a new multi-classification system using Gaussian networks to combine the outputs (probability distributions) of standard machine learning classification algorithms. This multi-classification technique has been applied to the selection of the most promising embryo-batch for human in-vitro fertilization treatments.

Chapter XVI reviews current policies of tuberculosis control programs for the diagnosis of tuberculosis. A data mining project that uses WHO's Direct Observation of Therapy data to analyze the relationship among different variables and the tuberculosis diagnostic category registered for each patient is then presented.

Chapter XVII describes how to integrate medical knowledge with purely inductive (data-driven) methods for the creation of clinical prediction rules. The described framework has been applied to the creation of clinical prediction rules for the diagnosis of obstructive sleep apnea.

Chapter XVIII describes goals, current results, and further plans of long time activity concerning application of data mining and machine learning methods to the complex medical data set. The analyzed data set concerns longitudinal study of atherosclerosis risk factors.

The book can be used as a textbook of advanced data mining applications in medicine. The book addresses not only researchers and students in the field of computer science or medicine but it will be of great interest also for physicians and managers of healthcare industry. It should help physicians and epidemiologists to add value to their collected data.

Petr Berka, Jan Rauch, and Djamel Abdelkader Zighed
Editors