Preface

Rapid recent advances in automated data collection routines in clinical sciences have led to a tsunami of patient-oriented data stored in distributed, heterogeneous, and large databases and datamarts. The lack of existing computing tools to enable connectivity and interoperability between these fragmented sources and scattered locations (issues concerning accessibility), and to perform machine learning on heterogeneous and highly dimensional data sources (issues concerning complexity), is an overbearing impediment, not only to healthcare sciences, but also to computational research. Moreover, the rapid deployment of high speed networks coupled with developments in knowledge discovery, bolstered by mobile technologies has amplified the emphatic demand for a unifying, coherent computing resources designed to accommodate, enhance, and empower multidisciplinary, and multi-institutional healthcare informatics research.

Healthcare data is complex, highly context-dependent, inherently heterogeneous, and high dimensional—generating an amalgamation of computing research challenges that renders the extraction of insightful knowledge through interpretation of raw data a challenging computational task. These data resources encompass a spectrum of data types ranging from free-text notes to complex image types such as position emission tomography scans. As clinical data collection technologies continue to grow and storage costs continue to fall, more complex data types such as hyperspectral images are becoming available in abundance. These diverse and prolific data sources provide an outstanding research test bed for development of the novel machine learning algorithms that are at the heart of the current data-rich but information-poor paradigm, saddling many disciplines outside of just health care. It is evident that an integrated, panoramic view of data will provide an opportunity for previously impossible clinical insights and discoveries.

The book provides a unique compendium of current and emerging machine learning paradigms for healthcare informatics. Chapters provided by established scientists in the area with the wealth of experience in the area, and have been carefully selected to reflect the diversity, complexity, and the depth and breath of this multidisciplinary area. Machine learning paradigms in healthcare informatics such as the ones presented in the chapters offer the promise of precise, objective, and accurate in-silico analysis of this emerging area using information learning routines that reveal embedded patterns, trends, and anomalies in order to create models for faster and more accurate physiological and healthcare discovery.

Chapter 1 provides an introduction to machine learning in healthcare informatics. The chapter provides an overview of the data and knowledge discovery challenges associated in the field of healthcare informatics. It introduces the challenges of machine learning in the area and the relevant areas of investigation in the area. The chapter explains the taxonomy of the healthcare informatics area and the current and provides an overview of the current efforts and emerging challenges of the Electronic Health Records (EHR) systems.

Chapter 2 discusses a machine learning approach to screen arrhythmia from normal sinus rhythm from the ECG. The methodology consists of R-point detection using the Pan-Tompkins algorithm, discrete wavelet transform (DWT) decomposition, subband principal component analysis (PCA), statistical validation of features, and subsequent pattern classification. Different classifiers used were Gaussian mixture model (GMM), error back propagation neural network (EBPNN), and support vector machine (SVM). Results indicate that the Symlet-2 wavelet basis function provided the highest accuracy in classification. Among the classifiers, SVM yields the highest classification accuracy, whereas EBPNN yields a higher accuracy than GMM.

Uncontrolled diabetes may lead to many serious complications. The result may be ketosis, which is normally due to an increase of acetone (a toxic acid product) and may lead to a situation such as diabetic coma. A fuzzy logic control system for the regulation of glucose level for diabetic patients was proposed in Chap. 3. A mathematical model describing the relationship between the human glucose level, insulin, and food was first presented. Then, a generalized fuzzy logic controller, including a set of fuzzy logic rules, is introduced to regulate glucose levels for diabetic patients. Following the fuzzy logic controller, simulation is presented. The results show that the fuzzy logic control is effective for handling the glucose level based on feedback scheme.

An integrated methodology for electrocardiogram (ECG)-based differentiation of arrhythmia and normal sinus rhythm using genetic algorithm optimized *k*-means clustering was discussed in Chap. 4. Open source databases consisting of the MIT BIH arrhythmia and MIT BIH normal sinus rhythm data were used. The methodology consists of QRS-complex detection using the Pan-Tompkins algorithm, principal component analysis (PCA), and subsequent pattern classification using the *k*-means classifier, error back propagation neural network (EBPNN) classifier, and genetic algorithm optimized *k*-means clustering. The *k*-means classifier provided an average accuracy of 91.21 % over all folds, whereas EBPNN provided a greater average accuracy of 95.79 %. In the proposed method, the *k*-means classifier is optimized using the genetic algorithm (GA), and the accuracy of this classifier is 95.79 %, which is equal to that of EBPNN.

Pixel/voxel-based machine learning (PML) is a powerful tool in computer-aided diagnosis (CAD) schemes for detection of lesions in medical images. Massive-training ANNs (MTANNs) were used for improving the performance (i.e., both sensitivity and specificity) of CAD schemes for detection of lung nodules in computer tomography (CT) and the detection of polyps in CT colonography in Chap. 5. The MTANN supervised filter is effective for enhancement

of lesions including lung nodules and colorectal polyps and suppression of non-lesions in medical images, which contributed to the improvement of the sensitivity as well as specificity in the initial lesion detection stage in CAD schemes, whereas the classification MTANNs contributed to the improvement of specificity in the false positive (FP) reduction stage in CAD schemes.

Understanding the biomechanics of the human foot during each stage of walking is important for the objective evaluation of movement dysfunction, accuracy of diagnosis, and prediction of foot impairment. In Chap. 6 Bayesian Network (BN) was used to extract the probabilistic causal information of foot function data, such as muscle activities, plantar pressures, and toe trajectories, from different types of data on human walking phases. The graphical networks extracted from the three stages of the stance phase of gait measurement data were useful for understanding the foot function of the normal walking and simulated hemiplegic walking. Thus, understanding the foot function during walking is important for further analysis of diagnostic, therapy, and training programs for foot impairment.

Successful application of machine learning in health care requires accuracy, transparency, acceptability, ability to deal with complex data, ability to deal with background knowledge, efficiency, and exportability. Rule learning is known to satisfy the above criteria. Chapter 7 introduces rule learning in health care, presents very expressive attributional rules, briefly describes the AQ21 rule learning system, and discusses three application areas in healthcare and health services research.

In the past two decades, machine learning techniques have been extensively applied for the detection of neurologic or neuropsychiatric disorders, especially Alzheimer's disease (AD) and its prodrome, mild cognitive impairment (MCI). Chapter 8 presents some of the latest developments in the application of machine learning techniques to AD and MCI diagnosis and prognosis. Discussion on how various biomarkers as well as connectivity networks can be extracted from the various modalities, such as structural T1-weighted imaging, diffusion-tensor imaging (DTI), and resting-state functional magnetic resonance imaging (fMRI), for effective diagnosis and prognosis was provided in detail.

Chapter 9 discusses several examples of how machine learning algorithms can be used to guide clinical decision making, and to generate scientific insights about these decisions. The focus of the chapter has been on rehabilitation in home care. In clinical applications, it was shown that machine learning algorithms can produce better decisions than standard clinical protocols. A "simple" algorithm such as KNN may work just as well as a more complex one such as the SVM. More importantly, it was shown that machine learning algorithms can do much more than make "black-box" predictions; they can generate important new clinical and scientific insights. This can be used to make better decisions about treatment plans for patients and about resource allocation for healthcare services, resulting in better outcomes for patients, and in a more efficient and effective healthcare system.

The widespread adoption of electronic health records in large health systems, combined with recent advances in data mining and machine methods, creates opportunities for the rapid acquisition and translation of knowledge for use in clinical practice. One area of great potential is in risk prediction of chronic progressive diseases from longitudinal medical records. Chapter 10 illustrates this potential of using a case study involving prediction of heart failure. Throughout, we discuss challenges and areas in need of further development.

Chapter 11 provides a framework to improve the physicians' diagnostic accuracy with the aid of machine learning algorithm. The resulting system is effective in predicting patient survival, and rehab/home outcome. A method has been introduced that creates a variety of reliable rules that make sense to physicians by combining CART and C4.5 and using only significant variables extracted via logistic regression. A novel method for assessment of Traumatic Brain Injury (TBI) has also been presented. The ability of such a system to assess levels of Intracranial Pressure (ICP) as well as predict survival outcomes and days in ICU, together encompasses a wholesome diagnostic tool, which can help improve patient care as well as save time and reduce cost.

One of the most crucial problems facing the U.S. government is fraud in healthcare system. Due to a large amount of data, it is impossible to manually audit for fraud. Hence, many statistical approaches have been proposed to overcome this problem. As fraud can be committed in complex and numerous ways, fraud detection is challenging, and there is a greater need for working models for fraud detection, including types of fraud that are not yet in use, as these models will not be outdated quickly. To establish a well-functioning healthcare system, it is important to have a good fraud detection system that can fight fraud that already exists and fraud that may emerge in future. In Chap. 12 an attempt has been made to classify fraud in the healthcare system, identify data sources, characterize data, and explain the supervised machine learning fraud detection models.

A migraine is a neurological disorder that can be caused by many factors, including genetic mutations, lifestyle, cardiac defects, endocrine pathologies, and neurovascular impairments. In addition to these health problems, an association between some types of migraines and increased cardiovascular risk has emerged in the past 10 years. Moreover, researchers have demonstrated an association between migraines and impaired cerebrovascular reactivity. It is possible to observe carbon dioxide dysregulation in some migraineurs, while others show a markedly decreased vasomotor reactivity to external stimuli. Therefore, the assessment of the cerebrovascular pattern of migraineurs is important both for the onset of a personalized therapy and for follow-up care. Chapter 13 discusses the analysis of hemodynamic changes during external stimulation using near-infrared spectroscopy (NIRS) signals.

The segmentation of the carotid artery wall is an important aid to sonographers when measuring intima-media thickness (IMT). Automated and completely user-independent segmentation techniques are gaining increasing importance, because they avoid the bias coming from human interactions. Chapter 14 discusses the calculation of the large and overabundant number of parameters extracted from

ultrasound carotid images and then selects a smaller subset to classify the pixels into three classes (lumen, intima-media complex, and adventitia). The selection was obtained through a feature selection method based on rough set theory. In particular, the use of QuickReduct Algorithm (QRA), the Entropy-Based Algorithm (EBR), and the Improved QuickReduct Algorithm (IQRA) was discussed.

Many authors have contributed to this book with their tremendous hard work and valuable time. We deeply thank them for their great contributions. In no particular order, they are: Roshan Joy Martis, Chandan Chakraborty, Ajoy Kumar Ray, K. Y. Zhu, W. D. Liu, Y. Xiao, Teik-Cheng Lim, Hari Prasad, Kenji Suzuki, Myagmarbayar Nergui, Jun Inoue, Murai Chieko, Wenwei Yu, Janusz Wojtusiak, Chong-Yaw Wee, Daoqiang Zhang, Luping Zhou, Pew-Thian Yap, Dinggang Shen, Mu Zhu, Lu Cheng, Joshua J. Armstrong, Jeff W. Poss, John P. Hirdes, Paul Stolee, Walter F. Stewart, Jason Roy, Jimeng Sun, Shahram Ebadollahi, Ashwin Belle, Soo-Yeon Ji, Wenan Chen, Toan Huynh, and Kayvan Najarian, Sonali Bais, Samanta Rosati, Gabriella Balestra, Filippo Molinari, Samanta Rosati, Gabriella Balestra, and Jasjit S. Suri.

Sumeet Dua
U. Rajendra Acharya
Prerna Dua

http://www.springer.com/978-3-642-40016-2

Machine Learning in Healthcare Informatics

Dua, S.; Acharya, U.R.; Dua, P. (Eds.)
2014, XII, 332 p. 119 illus., 50 illus. in color., Hardcover
ISBN: 978-3-642-40016-2