Chapter XVII
Knowledge-Based Induction of Clinical Prediction Rules

Mila Kwiatkowska
Thompson Rivers University, Canada

M. Stella Atkins
Simon Fraser University, Canada

Les Matthews
Thompson Rivers University, Canada

Najib T. Ayas
University of British Columbia, Canada

C. Frank Ryan
University of British Columbia, Canada

ABSTRACT

This chapter describes how to integrate medical knowledge with purely inductive (data-driven) methods for the creation of clinical prediction rules. It addresses three issues: representation of medical knowledge, secondary analysis of medical data, and evaluation of automatically induced predictive models in the context of existing knowledge. To address the complexity of the domain knowledge, the authors have introduced a semio-fuzzy framework, which has its theoretical foundations in semiotics and fuzzy logic. This integrative framework has been applied to the creation of clinical prediction rules for the diagnosis of obstructive sleep apnea, a serious and under-diagnosed respiratory disorder. The authors use a semio-fuzzy approach (1) to construct a knowledge base for the definition of diagnostic criteria, predictors, and existing prediction rules; (2) to describe and analyze data sets used in the data mining process; and (3) to interpret the induced models in terms of confirmation, contradiction, and contribution to existing knowledge.

INTRODUCTION

The ever-increasing number of electronic patients' records, specialized medical databases, and various computer-stored clinical files provides an unprecedented opportunity for automated and semi-automated discovery of patterns, trends, and associations in medical data. Data mining (DM) techniques combined with the fast and relatively easy access to large databases of patients' records can support clinical research and clinical care. One of the promising applications of DM techniques and secondary data analysis is the creation of clinical prediction rules (CPRs). The major functions of CPRs are to simplify the assessment process, to expedite diagnosis and treatment for serious cases, and to reduce the number of unnecessary tests for low-probability cases. However, before the rules can be used as formal guidelines in diagnosis, prognosis, and treatment, they must be validated on large and diversified populations and evaluated in clinical settings. This lengthy and costly process can be mitigated by automated rule induction from the existing data sources.

The secondary use of previously collected data for supporting the creation and evaluation of CPRs has several advantages: a significant reduction of data collection time and cost, availability of data from wide-ranging populations, access to large data sets, and access to rare cases. Moreover, DM techniques provide flexible and adaptable methods for the *exploratory* as well as the *confirmatory* data analyses. In exploratory analysis, DM techniques can be used to identify potential predictors and generate hypothetical rules. In confirmatory analysis, they can be used to evaluate a hypothesis by confirming, contradicting, or refining.

However, DM techniques alone are not sufficient to address problems concerning secondary analysis of medical data and the complexity of medical reasoning. Each step of the DM process requires integration of medical knowledge –– from problem and data understanding, through to model induction, and finally to predictive model evaluation. To achieve this integration, the DM process needs an explicit knowledge representation which should be readable and verifiable by medical experts. Moreover, the models induced by DM techniques should be comprehensible, interpretable, and practical for clinical usage.

In this chapter, we concentrate on two major issues related to the applications of DM techniques in the creation of CPRs: (1) problems associated with the secondary use of medical data, which means that the data were originally collected for other purposes, so they may be only partially suitable for the new DM task; and (2) problems associated with the interpretation of the generated models in the context of existing knowledge. In order to address these two challenges, we have created a new knowledge representation framework, which combines a semiotic approach and a fuzzy logic approach, called by us, a *semio-fuzzy approach*. We have used this new framework to support medical knowledge representation in the diagnosis of obstructive sleep apnea (OSA) (Kwiatkowska & Atkins, 2004).

Our experience with medical DM is based on real clinical records. In our studies, we have used two types of data: patients' records from a specialized clinic and data collected for medical research. In the application of our framework, we focus on two DM phases: pre-processing and post-processing. First, we have constructed a knowledge base (KB) for OSA consisting of diagnostic criteria, known predictors, and existing CPRs. Then, in the pre-processing phase, we use the KB in the analysis of missing values, analysis of outliers, and analysis of the strength of predictors. Finally, in the post-processing phase, we utilize the KB to evaluate the induced models using three criteria: confirmation, contradiction, and contribution.