

CONTENTS

<b>Preface</b>	<b>xi</b>
James H. Harrison, Jr	
<b>Dedication</b>	<b>xv</b>
<b>Introduction to the Mining of Clinical Data</b>	<b>1</b>
James H. Harrison, Jr	

The increasing volume of medical data online, including laboratory data, represents a substantial resource that can provide a foundation for improved understanding of disease presentation, response to therapy, and health care delivery processes. Data mining supports these goals by providing a set of techniques designed to discover similarities and relationships between data elements in large data sets. Currently, medical data have several characteristics that increase the difficulty of applying these techniques, although there have been notable medical data mining successes. Future developments in integrated medical data repositories, standardized data representation, and guidelines for the appropriate research use of medical data will decrease the barriers to mining projects.

<b>Introduction to Data Mining for Medical Informatics</b>	<b>9</b>
Donald E. Brown	

Data mining consists of a series of techniques for the discovery of patterns in large databases. This article provides an introduction to common data mining techniques with a view toward their use. The article begins by describing methods for discovering and exploring associations in observations and variables. The discussion then turns to methods for prediction. These techniques discover relationships between sets of variables. The article concludes with a description of evaluative techniques that are useful for assessing the results from data mining.

## **Open-Source Tools for Data Mining**

37

Blaz Zupan and Janez Demsar

With a growing volume of biomedical databases and repositories, the need to develop a set of tools to address their analysis and support knowledge discovery is becoming acute. The data mining community has developed a substantial set of techniques for computational treatment of these data. In this article, we discuss the evolution of open-source toolboxes that data mining researchers and enthusiasts have developed over the span of a few decades and review several currently available open-source data mining suites. The approaches we review are diverse in data mining methods and user interfaces and also demonstrate that the field and its tools are ready to be fully exploited in biomedical research.

## **The Development of Health Care Data Warehouses to Support Data Mining**

55

Jason A. Lyman, Kenneth Scully, and James H. Harrison, Jr

Clinical data warehouses offer tremendous benefits as a foundation for data mining. By serving as a source for comprehensive clinical and demographic information on large patient populations, they streamline knowledge discovery efforts by providing standard and efficient mechanisms to replace time-consuming and expensive original data collection, organization, and processing. Building effective data warehouses requires knowledge of and attention to key issues in database design, data acquisition and processing, and data access and security. In this article, the authors provide an operational and technical definition of data warehouses, present examples of data mining projects enabled by existing data warehouses, and describe key issues and challenges related to warehouse development and implementation.

## **Multi-Database Mining**

73

Mir S. Siadaty and James H. Harrison, Jr

Biomedical data useful for data mining are often distributed across multiple databases. These databases may be aggregated using several techniques to create single data sets that may be mined using standard approaches; however, separate databases may, in their design or data representation, capture information that is analytically useful and that is lost on integration. Recent techniques for mining multiple databases simultaneously but separately may preserve and leverage the unique perspectives within each database. This article presents an example, "dual mining," in which concurrent analysis of a target database with a related knowledge base can improve the identification of association patterns in the target most likely to be of interest for further analysis.

## **Temporal Data Mining**

83

Andrew R. Post and James H. Harrison, Jr

Large-scale clinical databases provide a detailed perspective on patient phenotype in disease and the characteristics of health care processes. Important information is often contained in the relationships between the values and timestamps of sequences of clinical data. The analysis of clinical time sequence data across entire patient populations may reveal data patterns that enable a more precise understanding of disease presentation, progression, and response to therapy, and thus could be of great value for clinical and translational research. Recent work suggests that the combination of temporal data mining methods with techniques from artificial intelligence research on knowledge-based temporal abstraction may enable the mining of clinically relevant temporal features from these previously problematic general clinical data.

## **Regional and National Health Care Data Repositories**

101

James H. Harrison, Jr and Raymond D. Aller

Efforts are underway to define a national framework for secondary analysis of health-related data. In the meantime, regional health databases have been constructed using insurance claims data, clinical data from single large health care providers, clinical data from multiple collaborating health care providers, and public health data. Large-scale survey data also are available in government databases. Clinical laboratory results are an important component of all these databases because they can provide validation for manually assigned diagnostic and procedure codes and can support inference of key information not provided by coding, such as severity of disease and prevalence of risk factors.

## **Data Mining and Infection Control**

119

Stephen E. Brossette and Patrick A. Hymel, Jr

Patterns embedded in large volumes of clinical data may provide important insights into the characteristics of patients or care delivery processes, but may be difficult to identify by traditional means. Data mining offers methods that can recognize patterns in these large data sets and make them actionable. We present an example of this capability in which we successfully applied data mining to hospital infection control. The Data Mining Surveillance System (DMSS) uses data from the clinical laboratory and hospital information systems to create association rules linking patients, sample types, locations, organisms, and antibiotic susceptibilities. Changes in association strength over time signal epidemiologic patterns potentially appropriate for follow-up, and additional heuristic methods identify the most informative of these patterns for alerting.

## **Data Mining for Biomarker Development: A Review of Tissue Specificity Analysis**

127

Eric W. Klee

Novel biomarker development requires a significant resource commitment to translate candidate markers into clinical assays. Consequently, it is imperative high quality candidates are selected early in a biomarker development program. High throughput gene expression data are routinely used to identify transcripts differentially expressed in diseased versus normal samples. Data-mining Expressed Sequence Tag, Serial Analysis of Gene Expression, Massively Parallel Signature Sequencing, and microarray expression databases can provide additional information on the expression of candidate biomarkers across multiple tissues, organs, and disease states. From this information, quantitative measures of tissue-specific gene specificity are computed and used to guide candidate biomarker selection.

## **Data Mining in Genomics**

145

Jae K. Lee, Paul D. Williams, and Sooyoung Cheon

This article reviews important emerging statistical concepts, data mining techniques, and applications that have been recently developed and used for genomic data analysis. First, general background and some critical issues in genomic data mining are summarized. A novel concept of statistical significance is described, the so-called “false discovery rate”—the rate of false-positives among all positive findings—which has been suggested to control the error rate of numerous false-positives in large screening biological data analysis. Two recent statistical testing methods are then introduced: significance analysis of microarray and local pooled error tests. Statistical modeling in genomic data analysis is then presented, such as analysis of variance and heterogeneous error modeling approaches that have been suggested for analyzing microarray data obtained from multiple experimental or biological conditions. Two sections then describe data exploration and discovery tools largely termed as supervised learning and unsupervised learning. The former approaches include several multivariate statistical methods to investigate coexpression patterns of multiple genes, and the latter are the classification methods to discover genomic biomarker signatures for predicting important subclasses of human diseases. The last section briefly summarizes various genomic data mining approaches in biomedical pathway analysis and patient outcome or chemotherapeutic response prediction.

## **Index**

167