Preface

A huge amount of data is collected every day in the form of sequences. These sequential data are valuable sources of information not only to search for a particular value or event at a specific time, but also to analyze the frequency of certain events or sets of events related by particular temporal/sequential relationship. For example, DNA sequences encode the genetic makeup of humans and all other species, and protein sequences describe the amino acid composition of proteins and encode the structure and function of proteins. Moreover, sequences can be used to capture how individual humans behave through various temporal activity histories such as weblog histories and customer purchase patterns. In general there are various methods to extract information and patterns from databases, such as time series approaches, association rule mining, and data mining techniques.

The objective of this book is to provide a concise state-of-the-art in the field of sequence data mining along with applications. The book consists of 14 chapters divided into 3 sections. The first section provides review of state-of-art in the field of sequence data mining. Section 2 presents relatively new techniques for sequence data mining. Finally, in section 3, various application areas of sequence data mining have been explored.

Chapter 1, Approaches for Pattern Discovery Using Sequential Data Mining, by Manish Gupta and Jiawei Han of University of Illinois at Urbana-Champaign, IL, USA, discusses different approaches for mining of patterns from sequence data. Apriori based methods and the pattern growth methods are the earliest and the most influential methods for sequential pattern mining. There is also a vertical format based method which works on a dual representation of the sequence database. Work has also been done for mining patterns with constraints, mining closed patterns, mining patterns from multi-dimensional databases, mining closed repetitive gapped subsequences, and other forms of sequential pattern mining. Some works also focus on mining incremental patterns and mining from stream data. In this chapter, the authors have presented at least one method of each of these types and discussed advantages and disadvantages.

Chapter 2, A Review of Kernel Methods Based Approaches to Classification and Clustering of Sequential Patterns, Part I: Sequences of Continuous Feature Vectors, was authored by Dileep A. D., Veena T., and C. Chandra Sekhar of Department of Computer Science and Engineering, Indian Institute of Technology Madras, India. They present a brief description of kernel methods for pattern classification and clustering. They also describe dynamic kernels for sequences of continuous feature vectors. The chapter also presents a review of approaches to sequential pattern classification and clustering using dynamic kernels.
Chapter 3 is *A Review of Kernel Methods Based Approaches to Classification and Clustering of Sequential Patterns, Part II: Sequences of Discrete Symbols* by Veena T., Dileep A. D., and C. Chandra Sekhar of Department of Computer Science and Engineering, Indian Institute of Technology Madras, India. The authors review methods to design dynamic kernels for sequences of discrete symbols. In their chapter, they have also presented a review of approaches to classification and clustering of sequences of discrete symbols using the dynamic kernel based methods.

Chapter 4 is titled, *Mining Statistically Significant Substrings Based on the Chi-Square Measure*, contributed by Sourav Dutta of IBM Research India along with Arnab Bhattacharya Dept. of Computer Science and Engineering, Indian Institute of Technology, Kanpur, India. This chapter highlights the challenge of efficient mining of large string databases in the domains of intrusion detection systems, player statistics, texts, proteins, et cetera, and how these issues have emerged as challenges of practical nature. Searching for an unusual pattern within long strings of data is one of the foremost requirements for many diverse applications. The authors first present the current state-of-art in this area and then analyze the different statistical measures available to meet this end. Next, they argue that the most appropriate metric is the chi-square measure. Finally, they discuss different approaches and algorithms proposed for retrieving the top-k substrings with the largest chi-square measure. The local-maxima based algorithms maintain high quality while outperforming others with respect to the running time.

Chapter 5 is *Unbalanced Sequential Data Classification Using Extreme Outlier Elimination and Sampling Techniques*, by T. Maruthi Padmaja along with Raju S. Bapi from University of Hyderabad, Hyderabad, India and P. Radha Krishna, Infosys Lab, Infosys Technologies Ltd, Hyderabad, India. This chapter focuses on the problem of predicting minority class sequence patterns from the noisy and unbalanced sequential datasets. To solve this problem, the authors proposed a new approach called extreme outlier elimination and hybrid sampling technique.

Chapter 6 is *Quantization Based Sequence Generation and Subsequence Pruning for Data Mining Applications* by T. Ravindra Babu and S. V. Subrahmanya of E-Comm. Research Lab, Education and Research, Infosys Technologies Limited, Bangalore, India, along with M. Narasimha Murty, Dept. of Computer Science and Automation, Indian Institute of Science, Bangalore, India. This chapter has highlighted the problem of combining data mining algorithms with data compaction used for data compression. Such combined techniques lead to superior performance. Approaches to deal with large data include working with a representative sample instead of the entire data. The representatives should preferably be generated with minimal data scans, methods like random projection, et cetera.

Chapter 7 is *Classification of Biological Sequences* by Pratibha Rani and Vikram Pudi of International Institute of Information Technology, Hyderabad, India, and it discusses the problem of classifying a newly discovered sequence like a protein or DNA sequence based on their important features and functions, using the collection of available sequences. In this chapter, the authors study this problem and present two techniques Bayesian classifiers: RBNBC and REBMEC. The algorithms used in these classifiers incorporate repeated occurrences of subsequences within each sequence. Specifically, RBNBC (Repeat Based Naive Bayes Classifier) uses a novel formulation of Naive Bayes, and the second classifier, REBMEC (Repeat Based Maximum Entropy Classifier) uses a novel framework based on the classical Generalized Iterative Scaling (GIS) algorithm.

Chapter 8, *Applications of Pattern Discovery Using Sequential Data Mining*, by Manish Gupta and Jiawei Han of University of Illinois at Urbana-Champaign, IL, USA, presents a comprehensive review of applications of sequence data mining algorithms in a variety of domains like healthcare, education, Web usage mining, text mining, bioinformatics, telecommunications, intrusion detection, et cetera.
Chapter 9, Analysis of Kinase Inhibitors and Druggability of Kinase-Targets Using Machine Learning Techniques, by S. Prashanthi, S. Durga Bhavani, T. Sobha Rani, and Raju S. Bapi of Department of Computer & Information Sciences, University of Hyderabad, Hyderabad, India, focuses on human kinase drug target sequences since kinases are known to be potential drug targets. The authors have also presented a preliminary analysis of kinase inhibitors in order to study the problem in the protein-ligand space in future. The identification of druggable kinases is treated as a classification problem in which druggable kinases are taken as positive data set and non-druggable kinases are chosen as negative data set.

Chapter 10, Identification of Genomic Islands by Pattern Discovery, by Nita Parekh of International Institute of Information Technology, Hyderabad, India addresses a pattern recognition problem at the genomic level involving identifying horizontally transferred regions, called genomic islands. A horizontally transferred event is defined as the movement of genetic material between phylogenetically unrelated organisms by mechanisms other than parent to progeny inheritance. Increasing evidence suggests the importance of horizontal transfer events in the evolution of bacteria, influencing traits such as antibiotic resistance, symbiosis and fitness, virulence, and adaptation in general. Considerable effort is being made in their identification and analysis, and in this chapter, a brief summary of various approaches used in the identification and validation of horizontally acquired regions is discussed.

Chapter 11, Video Stream Mining for On-Road Traffic Density Analytics, by Rudra Narayan Hota of Frankfurt Institute for Advanced Studies, Frankfurt, Germany along with Kishore Jonna and P. Radha Krishna, Infosys Lab, Infosys Technologies Limited, India, addresses the problem of estimating computer vision based traffic density using video stream mining. The authors present an efficient approach for traffic density estimation using texture analysis along with Support Vector Machine (SVM) classifier, and describe analyzing traffic density for on-road traffic congestion control with better flow management.

Chapter 12, Discovering Patterns in Order to Detect Weak Signals and Define New Strategies, by Anass El Haddadi of Université de Toulouse, IRIT UMR France Bernard Doussset, Ilham Berrada of Ensias, AL BIRONI team, Mohamed V University – Souissi, Rabat, Morocco presents four methods for discovering patterns in the competitive intelligence process: “correspondence analysis,” “multiple correspondence analysis,” “evolutionary graph,” and “multi-term method.” Competitive intelligence activities rely on collecting and analyzing data in order to discover patterns from data using sequence data mining. The discovered patterns are used to help decision-makers considering innovation and defining business strategy.

Chapter 13, Discovering Patterns for Architecture Simulation by Using Sequence Mining, by Pınar Senkul (Middle East Technical University, Computer Engineering Dept., Ankara, Turkey) along with Nilufer Onder (Michigan Technological University, Computer Science Dept., Michigan, USA), Soner Onder (Michigan Technological University, Computer Science Dept., Michigan, USA), Engin Maden (Middle East Technical University, Computer Engineering Dept., Ankara, Turkey) and Hui Meen Nyew (Michigan Technological University, Computer Science Dept., Michigan, USA), discusses the problem of designing and building high performance systems that make effective use of resources such as space and power. The design process typically involves a detailed simulation of the proposed architecture followed by corrections and improvements based on the simulation results. Both simulator development and result analysis are very challenging tasks due to the inherent complexity of the underlying systems. They present a tool called Episode Mining Tool (EMT), which includes three temporal sequence mining algorithms, a preprocessor, and a visual analyzer.

Chapter 14 is called Sequence Pattern Mining for Web Logs by Pradeep Kumar, Indian Institute of Management, Lucknow, India, Raju S. Bapi, University of Hyderabad, India and P. Radha Krishna,
Infosys Lab, Infosys Technologies Limited, India. In their work, the authors utilize a variation to the AprioriALL Algorithm, which is commonly used for the sequence pattern mining. The proposed variation adds up the measure Interest during every step of candidate generation to reduce the number of candidates thus resulting in reduced time and space cost.

This book can be useful to academic researchers and graduate students interested in data mining in general and in sequence data mining in particular, and to scientists and engineers working in fields where sequence data mining is involved, such as bioinformatics, genomics, Web services, security, and financial data analysis.

Sequence data mining is still a fairly young research field. Much more remains to be discovered in this exciting research domain in the aspects related to general concepts, techniques, and applications. Our fond wish is that this collection sparks fervent activity in sequence data mining, and we hope this is not the last word!

Pradeep Kumar
Indian Institute of Management Lucknow, India

P. Radha Krishna
Infosys Lab, Infosys Limited, India

S. Bapi Raju
University of Hyderabad, India