Preface

The last decade has witnessed a revolution in interdisciplinary research where the boundaries of different areas have overlapped or even disappeared. New fields of research emerge each day where two or more fields have integrated to form a new identity. Examples of these emerging areas include bioinformatics (synthesizing biology with computer and information systems), data mining (combining statistics, optimization, machine learning, artificial intelligence, and databases), and modern heuristics (integrating ideas from tens of fields such as biology, forest, immunology, statistical mechanics, and physics to inspire search techniques). These integrations have proved useful in substantiating problem-solving approaches with reliable and robust techniques to handle the increasing demand from practitioners to solve real-life problems. With the revolution in genetics, databases, automation, and robotics, problems are no longer those that can be solved analytically in a feasible time. Complexity arises because of new discoveries about the genome, path planning, changing environments, chaotic systems, and many others, and has contributed to the increased demand to find search techniques that are capable of getting a good enough solution in a reasonable time. This has directed research into heuristics.

During the same period of time, databases have grown exponentially in large stores and companies. In the old days, system analysts faced many difficulties in finding enough data to feed into their models. The picture has changed and now the reverse picture is a daily problem–how to understand the large amount of data we have accumulated over the years. Simultaneously, investors have realized that data is a hidden treasure in their companies. With data, one can analyze the behavior of competitors, understand the system better, and diagnose the faults in strategies and systems. Research into statistics, machine learning, and data analysis has been resurrected. Unfortunately, with the amount of data and the complexity of the underlying models, traditional approaches in statistics, machine learning, and traditional data analysis fail to cope with this level of complexity. The need therefore arises for better approaches that are able to handle complex models in a reasonable amount of time. These approaches have been named data mining (sometimes data farming) to distinguish them from traditional statistics, machine learning, and other data analysis techniques. In addition, decision makers were not interested in techniques that rely too much on the underlying assumptions in statistical models. The challenge is to not have any assumptions about the model and try to come up with something new, something that is not obvious or predictable (at least from the decision makers' point of view). Some unobvious thing may have significant values to the decision maker. Identifying a hidden trend in the data or a buried fault in the system is by all accounts a treasure for the investor who knows that avoiding loss results in profit and that knowledge in a complex market is a key criterion for success and continuity. Notwithstanding, models that are free from assumptions–or at least have minimum assumptions–are expensive to use. The dramatic search space cannot be navigated using traditional search techniques. This has highlighted a natural demand for the use of heuristic search methods in data mining.

This book is a repository of research papers describing the applications of modern

heuristics to data mining. This is a unique–and as far as we know, the first–book that provides up-to-date research in coupling these two topics of modern heuristics and data mining. Although it is by all means an incomplete coverage, it does provide some leading research in this area.

This book contains open-solicited and invited chapters written by leading researchers in the field. All chapters were peer reviewed by at least two recognized researchers in the field in addition to one of the editors. Contributors come from almost all the continents and therefore, the book presents a global approach to the discipline. The book contains 13 chapters divided into five parts as follows:

- Part 1: General Heuristics
- Part 2: Evolutionary Algorithms
- Part 3: Genetic Programming
- Part 4: Ant Colony Optimization and Immune Systems
- Part 5: Parallel Data Mining

Part 1 gives an introduction to modern heuristics as presented in the first chapter. The chapter serves as a textbook-like introduction for readers without a background in heuristics or those who would like to refresh their knowledge.

Chapter 2 is an excellent example of the use of hill climbing for clustering. In this chapter, Vladimir Estivill-Castro and Michael E. Houle from the University of Newcastle and the University of Sydney, respectively, provide a methodical overview of clustering and hill climbing methods to clustering. They detail the use of proximity information to assess the scalability and robustness of clustering.

Part 2 covers the well-known evolutionary algorithms. After almost three decades of continuous research in this area, the vast amount of papers in the literature is beyond a single survey paper. However, in Chapter 3, Erick Cantú-Paz and Chandrika Kamath from Lawrence Livermore National Laboratory, USA, provide a brave and very successful attempt to survey the literature describing the use of evolutionary algorithms in data mining. With over 75 references, they scrutinize the data mining process and the role of evolutionary algorithms in each stage of the process.

In Chapter 4, Beatriz de la Iglesia and Victor J. Rayward-Smith, from the University of East Anglia, UK, provide a superb paper on the application of Simulated Annealing, Tabu Search, and Genetic Algorithms (GA) to nugget discovery or classification where an important class is under-represented in the database. They summarize in their chapter different measures of performance for the classification problem in general and compare their results against 12 classification algorithms.

Iñaki Inza, Pedro Larrañaga, and Basilio Sierra from the University of the Basque Country, Spain, follow, in Chapter 5, with an outstanding piece of work on feature subset selection using a different type of evolutionary algorithms, the Estimation of Distribution Algorithms (EDA). In EDA, a probability distribution of the best individuals in the population is maintained to sample the individuals in subsequent generations. Traditional crossover and mutation operators are replaced by the re-sampling process. They applied EDA to the Feature Subset Selection problem and showed that it significantly improves the prediction accuracy.

In Chapter 6, Jorge Muruzábal from the University of Rey Juan Carlos, Spain, presents the brilliant idea of evolving teams of local Bayesian learners. Bayes theorem was resurrected as a result of the revolution in computer science. Nevertheless, Bayesian approaches, such as

Bayesian Networks, require large amounts of computational effort, and the search algorithm can easily become stuck in a local minimum. Dr. Muruzábal combined the power of the Bayesian approach with the ability of Evolutionary Algorithms and Learning Classifier Systems for the classification process.

Neil Dunstan from the University of New England, and Michael de Raadt from the University of Southern Queensland, Australia, provide an interesting application of the use of evolutionary algorithms for the classification and detection of Unexploded Ordnance present on military sites in Chapter 7.

Part 3 covers the area of Genetic Programming (GP). GP is very similar to the traditional GA in its use of selection and recombination as the means of evolution. Different from GA, GP represents the solution as a tree, and therefore the crossover and mutation operators are adopted to handle tree structures. This part starts with Chapter 8 by Peter W.H. Smith from City University, UK, who provides an interesting introduction to the use of GP for data mining and the problems facing GP in this domain. Before discarding GP as a useful tool for data mining, A.P. Engelbrecht and L Schoeman from the University of Pretoria, South Africa along with Sonja Rouwhorst from the University of Vrije, The Netherlands, provide a building block approach to genetic programming for rule discovery in Chapter 9. They show that their proposed GP methodology is comparable to the famous C4.5 decision tree classifier–a famous decision tree classifier.

Part 4 covers the increasingly growing areas of Ant Colony Optimization and Immune Systems. Rafael S. Parpinelli and Heitor S. Lopes from Centro Federal de Educacao Tecnologica do Parana, and Alex A. Freitas from Pontificia Universidade Catolica do Parana, Brazil, present a pioneer attempt, in Chapter 10, to apply ant colony optimization to rule discovery. Their results are very promising and through an extremely interesting approach, they present their techniques.

Jon Timmis and Thomas Knight, from the University of Kent at Canterbury, UK, introduce Artificial Immune Systems (AIS) in Chapter 11. In a notable presentation, they present the AIS domain and how can it be used for data mining. Leandro Nunes de Castro and Fernando J. Von Zuben, from the State University of Campinas, Brazil, follow in Chapter 12 with the use of AIS for clustering. The chapter presents a remarkable metaphor for the use of AIS with an outstanding potential for the proposed algorithm.

In general, the data mining task is very expensive, whether we are using heuristics or any other technique. It was therefore impossible not to present this book without discussing parallel data mining. This is the task carried out by David Taniar from Monash University and J. Wenny Rahayu from La Trobe University, Australia, in Part 5, Chapter 13. They both have written a self-contained and detailed chapter in an exhilarating style, thereby bringing the book to a close.

It is hoped that this book will trigger great interest into data mining and heuristics, leading to many more articles and books!