Scalable Pattern Recognition Algorithms

Pradipta Maji · Sushmita Paul

Scalable Pattern Recognition Algorithms

Applications in Computational Biology and Bioinformatics

Springer

Pradipta Maji
Indian Statistical Institute
Kolkata, West Bengal
India

Sushmita Paul
Indian Statistical Institute
Kolkata, West Bengal
India

ISBN 978-3-319-05629-6 ISBN 978-3-319-05630-2 (eBook)
DOI 10.1007/978-3-319-05630-2
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014933668

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To my daughter

Pradipta Maji

To my parents

Sushmita Paul

Foreword

It is my great pleasure to welcome a new book on "Scalable Pattern Recognition Algorithms: Applications in Computational Biology and Bioinformatics" by Prof. Pradipta Maji and Dr. Sushmita Paul.

This book is unique in its character. Most of the methods presented in it are based on profound research results obtained by the authors. These results are closely related to the main research directions in bioinformatics. The existing conventional/traditional approaches and techniques are also presented, wherever necessary. The effectiveness of algorithms that are proposed by the authors is thoroughly discussed along with both quantitative and qualitative comparisons with other existing methods in this area. These results are derived through experiments on real-life data sets. In general, the presented algorithms display excellent performance. One of the important aspects of the methods proposed by the authors is their ability to scale well with the inflow data. It shall be mentioned that the authors provide in each chapter the directions for future research in the corresponding area.

The main aim of bioinformatics is the development and application of computational methods in pursuit of biological discoveries. Among the hot topics in this field are: sequence alignment and analysis, gene finding, genome annotation, protein structure alignment and prediction, classification of proteins, clustering and dimensionality reduction of gene expression data, protein–protein docking or interactions, and modeling of evolution. From a more general view, the aim is to discover unifying principles of biology using tools of automated knowledge discovery. Hence, knowledge discovery methods that rely on pattern recognition, machine learning, and data mining are widely used for analysis of biological data, in particular for classification, clustering, and feature selection.

The book is structured according to the major phases of a pattern recognition process (clustering, classification, and feature selection) with a balanced mixture of theory, algorithms, and applications. Special emphasis is given to applications in computational biology and bioinformatics.

The reader will find in the book a unified framework describing applications of soft computing, statistical, and machine learning techniques in construction of efficient data models. Soft computing methods allow us to achieve high quality

solutions for many real-life applications. The characteristic features of these methods are tractability, robustness, low-cost solution, and close resemblance with humanlike decision making. They make it possible to use imprecision, uncertainty, approximate reasoning, and partial truth in searching for solutions. The main research directions in soft computing are related to fuzzy sets, neurocomputing, genetic algorithms, probabilistic reasoning, and rough sets. By integration or combination of the different soft computing methods, one may improve the performance of these methods.

The authors of the book present several newly developed methods and algorithms that combine statistical and soft computing approaches, including: (i) neural network tree (NNTree) used for identification of splice-junction and protein coding region in DNA sequences; (ii) a new approach for selecting miRNAs from microarray expression data integrating the merit of rough set-based feature selection algorithm and theory of *B*. 632+ bootstrap error rate; (iii) a robust thresholding technique for segmentation of brain MR images based on the fuzzy thresholding technique; (iv) an efficient method for selecting set of bio-basis strings for the new kernel function, integrating the Fisher ratio and a novel concept of degree of resemblance; (v) a rough set-based feature selection algorithm for selecting sets of effective molecular descriptors from a given quantitative structure activity relationship (QSAR) data set.

Clustering is one of the important analytic tools in bioinformatics. There are several new clustering methods presented in the book. They achieve very good results on various biomedical data sets. That includes, in particular: (i) a method based on Pearson's correlation coefficient that selects initial cluster centers, thus enabling the algorithm to converge to optimal or nearly optimal solution and helping to discover co-expressed gene clusters; (ii) a method based on Dunn's cluster validity index that identifies optimal parameter values during initialization and execution of the clustering algorithm; (iii) a supervised gene clustering algorithm based on the similarity between genes measured with use of the new quantitative measure, whereby redundancy among the attributes is eliminated; (iv) a novel possibilistic biclustering algorithm for finding highly overlapping biclusters having larger volume and mean squared residue lower than a predefined threshold.

The reader will also find several other interesting methods that may be applied in bioinformatics, such as: (i) a computational method for identification of disease-related genes, judiciously integrating the information of gene expression profiles, and the shortest path analysis of protein–protein interaction networks; (ii) a method based on *f*-information measures used in evaluation criteria for gene selection problem.

This book will be useful for graduate students, researchers, and practitioners in computer science, electrical engineering, system science, medical science, bioinformatics, and information technology. In particular, researchers and practitioners in industry and R&D laboratories working in the fields of system design, pattern

recognition, machine learning, computational biology and bioinformatics, data mining, soft computing, computational intelligence, and image analysis may benefit from it.

The authors and editors deserve the highest appreciation for their outstanding work.

Warsaw, Poland, December 2013 Andrzej Skowron

Preface

Recent advancement and wide use of high-throughput technologies for biological research are producing enormous size of biological data distributed worldwide. With the rapid increase in size of biological data banks, understanding the biological data has become critical. Such an understanding could lead us to the elucidation of the secrets of life or ways to prevent certain currently non-curable diseases. Although laboratory experiment is the most effective method for investigating the biological data, it is financially expensive and labor intensive. A deluge of such information coming in the form of genomes, protein sequences, and microarray expression data has led to the absolute need for effective and efficient computational tools to store, analyze, and interpret these multifaceted data.

Bioinformatics is the conceptualizing biology in terms of molecules and applying informatics techniques to understand and organize the information associated with the molecules, on a large scale. It involves the development and advancement of algorithms using techniques including pattern recognition, machine learning, applied mathematics, statistics, informatics, and biology to solve biological problems usually on the molecular level. Major research efforts in this field include sequence alignment and analysis, gene finding, genome annotation, protein structure alignment and prediction, classification of proteins, clustering and dimensionality reduction of microarray expression data, protein–protein docking or interactions, modeling of evolution, and so forth. In other words, bioinformatics can be described as the development and application of computational methods to make biological discoveries. The ultimate attempt of this field is to develop new insights into the science of life as well as creating a global perspective, from which the unifying principles of biology can be derived. As classification, clustering, and feature selection are needed in this field, pattern recognition tools and machine learning techniques have been widely used for analysis of biological data as they provide useful tools for knowledge discovery in this field.

Pattern recognition is the scientific discipline whose goal is the classification of objects into a number of categories or classes. It is the subject of researching object description and classification method. It is also a collection of mathematical, statistical, heuristic, and inductive techniques of the fundamental role in executing the tasks like human beings on computers. In a general setting, the process of pattern recognition is visualized as a sequence of a few steps: data acquisition; data preprocessing; feature selection; and classification or clustering. In the first step,

data are gathered via a set of sensors depending on the environment within which the objects are to be classified. After data acquisition phase, some preprocessing tasks such as noise reduction, filtering, encoding, and enhancement are applied on the collected data for extracting pattern vectors. Afterward, a feature space is constituted to reduce the space dimensionality. However, in a broader perspective this stage significantly influences the entire recognition process. Finally, the classifier is constructed, or in other words, a transformation relationship is established between features and classes.

Pattern recognition, by its nature, admits many approaches, sometimes complementary, sometimes competing, to provide the appropriate solution for a given problem. For any pattern recognition system, one needs to achieve robustness with respect to random noise and failure of components and to obtain output in real time. It is also desirable for the system to be adaptive to the changes in the environment. Moreover, a system can be made artificially intelligent if it is able to emulate some aspects of the human reasoning system. Soft computing and machine learning approaches to pattern recognition are attempts to achieve these goals. Artificial neural network, genetic algorithms, fuzzy sets, and rough sets are used as the tools in these approaches. The challenge is, therefore, to devise powerful pattern recognition methodologies by symbiotically combining these tools for analyzing biological data in more efficient ways. The systems should have the capability of flexible information processing to deal with real-life ambiguous situations and to achieve tractability, robustness, and low-cost solutions.

Various scalable pattern recognition algorithms using soft computing and machine learning approaches, and their real-life applications, including those in computational biology and bioinformatics, have been reported during the last 5–7 years. These are available in different journals, conference proceedings, and edited volumes. This scattered information causes inconvenience to readers, students, and researchers. The current volume is aimed at providing a treatise in a unified framework describing how soft computing and machine learning techniques can be judiciously formulated and used in building efficient pattern recognition models. Based on the existing as well as new results, the book is structured according to the major phases of a pattern recognition system (classification, feature selection, and clustering) with a balanced mixture of theory, algorithm, and applications. Special emphasis is given to applications in computational biology and bioinformatics.

The book consists of 11 chapters. Chapter 1 provides an introduction to pattern recognition and bioinformatics, along with different research issues and challenges related to high-dimensional real-life biological data sets. The significance of pattern recognition and machine learning techniques in computational biology and bioinformatics is also presented in Chap. 1. Chapter 2 presents the design of a hybrid learning model, termed as neural network tree (NNTree), for identification of splice-junction and protein coding region in DNA sequences. It incorporates the advantages of both decision tree and neural network. An NNTree is a decision tree, where each non-terminal node contains a neural network. The versatility of this method is illustrated through its application in splice-junction and gene

identification problems. Extensive experimental results establish that the NNTree produces more accurate classifier than that previously obtained for a range of different sequence lengths, thereby indicating a cost-effective alternative in splice-junction and protein coding region identification problem.

The prediction of protein functional sites is an important issue in protein function studies and drug design. In order to apply the powerful kernel-based pattern recognition algorithms such as support vector machine to predict functional sites in proteins, amino acids need encoding prior to input. In this regard, a new string kernel function, termed as the modified bio-basis function, is presented in Chap. 3. It maps a nonnumerical sequence space to a numerical feature space using a bio-basis string as its support. The concept of zone of influence of bio-basis string is introduced in the new kernel function to take into account the influence of each bio-basis string in nonnumerical sequence space. An efficient method is described to select a set of bio-basis strings for the new kernel function, integrating the Fisher ratio and the concept of degree of resemblance. The integration enables the method to select a reduced set of relevant and nonredundant bio-basis strings. Some quantitative indices are described for evaluating the quality of selected bio-basis strings. The effectiveness of the new string kernel function and bio-basis string selection method, along with a comparison with existing bio-basis function and related bio-basis string selection methods, is demonstrated on different protein data sets using the new quantitative indices and support vector machine.

Quantitative structure activity relationship (QSAR) is one of the important disciplines of computer-aided drug design that deals with the predictive modeling of properties of a molecule. In general, each QSAR data set is small in size with a large number of features or descriptors. Among the large amount of descriptors present in the QSAR data set, only a small fraction of them is effective for performing the predictive modeling task. Chapter 4 presents a rough set-based feature selection algorithm to select a set of effective molecular descriptors from a given QSAR data set. The new algorithm selects the set of molecular descriptors by maximizing both relevance and significance of the descriptors. The performance of the new algorithm is studied using the R^2 statistic of support vector regression method. The effectiveness of the new algorithm, along with a comparison with existing algorithms, is demonstrated on several QSAR data sets.

Microarray technology is one of the important biotechnological means that allows to record the expression levels of thousands of genes simultaneously within a number of different samples. An important application of microarray gene expression data in functional genomics is to classify samples according to their gene expression profiles. Among the large amount of genes present in microarray gene expression data, only a small fraction of them is effective for performing a certain diagnostic test. In this regard, mutual information has been shown to be successful for selecting a set of relevant and nonredundant genes from microarray data. However, information theory offers many more measures such as the f-information measures that may be suitable for selection of genes from microarray gene expression data.

Chapter 5 presents different *f*-information measures as the evaluation criteria for gene selection problem. The performance of different *f*-information measures is compared with that of mutual information based on the predictive accuracy of naive Bayes classifier, *k*-nearest neighbor rule, and support vector machine. An important finding is that some *f*-information measures are shown to be effective for selecting relevant and nonredundant genes from microarray data. The effectiveness of different *f*-information measures, along with a comparison with mutual information, is demonstrated on several cancer data sets.

One of the most important and challenging problems in functional genomics is how to select the disease genes. In Chap. 6, a computational method is reported to identify disease genes, judiciously integrating the information of gene expression profiles and shortest path analysis of protein–protein interaction networks. While the gene expression profiles have been used to select differentially expressed genes as disease genes using mutual information-based maximum relevance-maximum significance framework, the functional protein association network has been used to study the mechanism of diseases. Extensive experimental study on colorectal cancer establishes the fact that the genes identified by the integrated method have more colorectal cancer genes than the genes identified from the gene expression profiles alone. All these results indicate that the integrated method is quite promising and may become a useful tool for identifying disease genes.

The microRNAs or miRNAs regulate expression of a gene or protein. It has been observed that they play an important role in various cellular processes and thus help in carrying out normal functioning of a cell. However, dysregulation of miRNAs is found to be a major cause of a disease. Various studies have also shown the role of miRNAs in cancer and utility of miRNAs for the diagnosis of cancer. In this regard, Chap. 7 presents a new approach for selecting miRNAs from microarray expression data. It integrates the merit of rough set-based feature selection algorithm reported in Chap. 4 and theory of B. 632+ bootstrap error rate. The effectiveness of the new approach, along with a comparison with other algorithms, is demonstrated on several miRNA data sets.

Clustering is one of the important analyses in functional genomics that discovers groups of co-expressed genes from microarray data. In Chap. 8, different partitive clustering algorithms such as hard *c*-means, fuzzy *c*-means, rough-fuzzy *c*-means, and self-organizing maps are presented to discover co-expressed gene clusters. One of the major issues of the partitive clustering-based microarray data analysis is how to select initial prototypes of different clusters. To overcome this limitation, a method is reported based on Pearson's correlation coefficient to select initial cluster centers. It enables the algorithm to converge to an optimum or near optimum solutions and helps to discover co-expressed gene clusters. In addition, a method is described to identify optimum values of different parameters of the initialization method and the clustering algorithm. The effectiveness of different algorithms is demonstrated on several yeast gene expression time-series data sets using different cluster validity indices and gene ontology-based analysis.

In functional genomics, an important application of microarray data is to classify samples according to their gene expression profiles such as to classify

cancer versus normal samples or to classify different types or subtypes of cancer. Hence, one of the major tasks with the gene expression data is to find groups of co-regulated genes whose collective expression is strongly associated with the sample categories or response variables. In this regard, a supervised gene clustering algorithm is presented in Chap. 9 to find groups of genes. It directly incorporates the information about sample categories into the gene clustering process. A new quantitative measure, based on mutual information, is reported that incorporates the information about sample categories to measure the similarity between attributes. The supervised gene clustering algorithm is based on measuring the similarity between genes using the new quantitative measure. The performance of the new algorithm is compared with that of existing supervised and unsupervised gene clustering and gene selection algorithms based on the class separability index and the predictive accuracy of naive Bayes classifier, *k*-nearest neighbor rule, and support vector machine on several cancer and arthritis microarray data sets. The biological significance of the generated clusters is interpreted using the gene ontology.

The biclustering method is another important tool for analyzing gene expression data. It focuses on finding a subset of genes and a subset of experimental conditions that together exhibit coherent behavior. However, most of the existing biclustering algorithms find exclusive biclusters, which is inappropriate in the context of biology. Since biological processes are not independent of each other, many genes may participate in multiple different processes. Hence, nonexclusive biclustering algorithms are required for finding overlapping biclusters. In Chap. 10, a novel possibilistic biclustering algorithm is presented to find highly overlapping biclusters of larger volume with mean squared residue lower than a predefined threshold. It judiciously incorporates the concept of possibilistic clustering algorithm into biclustering framework. The integration enables efficient selection of highly overlapping coherent biclusters with mean squared residue lower than a given threshold. The detailed formulation of the new possibilistic biclustering algorithm, along with a mathematical analysis on the convergence property, is presented. Some quantitative indices are reported for evaluating the quality of generated biclusters. The effectiveness of the algorithm, along with a comparison with other algorithms, is demonstrated on yeast gene expression data set.

Finally, Chap. 11 reports a robust thresholding technique for segmentation of brain MR images. It is based on the fuzzy thresholding techniques. Its aim is to threshold the gray level histogram of brain MR images by splitting the image histogram into multiple crisp subsets. The histogram of the given image is thresholded according to the similarity between gray levels. The similarity is assessed through a second-order fuzzy measure such as fuzzy correlation, fuzzy entropy, and index of fuzziness. To calculate the second-order fuzzy measure, a weighted co-occurrence matrix is presented, which extracts the local information more accurately. Two quantitative indices are reported to determine the multiple thresholds of the given histogram. The effectiveness of the algorithm, along with a comparison with standard thresholding techniques, is demonstrated on a set of brain MR images.

The relevant existing conventional/traditional approaches or techniques are also included wherever necessary. Directions for future research in the concerned topic are provided in each chapter. Most of the materials presented in the book are from our published works. For the convenience of readers, a comprehensive bibliography on the subject is also appended in each chapter. It might have happened that some works in the related areas have been omitted due to oversight or ignorance.

The book, which is unique in its character, will be useful to graduate students and researchers in computer science, electrical engineering, system science, medical science, bioinformatics, and information technology both as a textbook and as a reference book for some parts of the curriculum. The researchers and practitioners in industry and R&D laboratories working in the fields of system design, pattern recognition, machine learning, computational biology and bioinformatics, data mining, soft computing, computational intelligence, and image analysis will also be benefited.

Finally, the authors take this opportunity to thank Mr. Wayne Wheeler and Mr. Simon Rees of Springer-Verlag, London, for their initiative and encouragement. The authors also gratefully acknowledge the support provided by Dr. Chandra Das of Netaji Subhash Engineering College, Kolkata, India and the members of Biomedical Imaging and Bioinformatics Lab, Indian Statistical Institute, Kolkata, India for preparation of a few chapters of the manuscript. The book has been written when one of the authors, Dr. S. Paul, held a CSIR Fellowship of the Government of India. This work is partially supported by the Indian National Science Academy, New Delhi (grant no. SP/YSP/68/2012).

Kolkata, India, January 2014

Pradipta Maji
Sushmita Paul

Contents

Part II Feature Selection