MACHINE LEARNING IN BIOINFORMATICS

Edited by
Yan-Qing Zhang
Jagath C. Rajapakse

A JOHN WILEY & SONS, INC., PUBLICATION

MACHINE LEARNING IN BIOINFORMATICS

Wiley Series on
Bioinformatics: Computational Techniques and Engineering

Bioinformatics and computational biology involve the comprehensive application of mathematics, statistics, science, and computer science to the understanding of living systems. Research and development in these areas require cooperation among specialists from the fields of biology, computer science, mathematics, statistics, physics, and related sciences. The objective of this book series is to provide timely treatments of the different aspects of bioinformatics spanning theory, new and established techniques, technologies and tools, and application domains. This series emphasizes algorithmic, mathematical, statistical, and computational methods that are central in bioinformatics and computational biology.

Series Editors: **Professor Yi Pan** and **Professor Albert Y. Zomaya**
pan@cs.gsu.edu zomaya@it.usyd.edu.au

Knowledge Discovery in Bioinformatics: Techniques, Methods, and Applications
Xiaohua Hu and Yi Pan

Grid Computing for Bioinformatics and Computational Biology
Edited by El-Ghazali Talbi and Albert Y. Zomaya

Bioinformatics Algorithms: Techniques and Applications
Ion Mandiou and Alexander Zelikovsky

Analysis of Biological Networks
Edited by Björn H. Junker and Falk Schreiber

Machine Learning in Bioinformatics
Edited by Yan-Qing Zhang and Jagath C. Rajapakse

MACHINE LEARNING IN BIOINFORMATICS

Edited by
Yan-Qing Zhang
Jagath C. Rajapakse

A JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey
Published simultaneously in Canada

Library of Congress Cataloging-in-Publication Data:
Zhang, Yan-Qing.
Machine learning in bioinformatics / edited by Yan-Qing Zhang, Jagath C. Rajapakse.
p. cm.
Includes bibliographical references.
ISBN 978-0-470-11662-3 (cloth)
1. Bioinformatics. 2. Machine learning. I. Rajapakse, Jagath Chandana. II. Title.
QH324.2.Z46 2008
572.80285'61–dc22
2008017095

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

CONTENTS

FOREWORD

Machine learning is a subfield of artificial intelligence and is concerned with the development of algorithms and techniques that allow computers to learn. It has a wide spectrum of applications such as natural language processing, search engines, medical diagnosis, bioinformatics and cheminformatics, stock market analysis, computer vision, and game playing. Recently, the amount of biological data requiring analysis has exploded and many machine learning methods have been developed to deal with this explosion of data. Hence, machine learning in bioinformatics has become an important research area for both computer scientists and biologists.

The aim of this book is to provide applications of machine learning to problems in the biological sciences, with particular emphasis on problems in bioinformatics. The book consists of a number of stand-alone chapters that explain and apply machine learning methods to central bioinformatics problems such as feature selection, sequence-based prediction of residue-level properties, promoter recognition, protein structure prediction, gene selection, and SNPS selection, classification, and data mining. This book represents the unification of two important fields in sciences—biology and computer science—with machine learning as a common theme. The chapters are written by well-known researchers in these interdisciplinary areas, and applications of various machine learning methods to different bioinformatics problems are presented. Students and scientists in biology and computer science will find this book valuable and accessible.

Several books in the similar areas have been published. However, this book is unique in that it presents cutting-edge research topics and methodologies in the area of machine learning methods when applied to bioinformatics. Many results presented in this book have never been published in the literature and represent the most advanced technologies in this exciting area. It also provides a comprehensive and balanced blend of topics, implementations, and case studies. I firmly believe that this book will further facilitate collaboration between machine learning researchers and bioinformaticians.

Both editors, Dr. Yan-Qing Zhang and Dr. Jagath C. Rajapakse, are rising stars in the areas of machine learning and bioinformatics. They have achieved a lot of research results in these areas. Their vision of creating such a book in a timely manner deserves our loud applause. This book is ideally suited both as a reference and as a text for a graduate course on machine learning or bioinformatics. This book can also serve as a

repository of significant reference materials because the references cited in each chapter serve as useful sources for further study in this area.

I highly recommend this timely and valuable book. I believe that it will benefit many readers and contribute to the further development of machine learning in bioinformatics.

Atlanta, Georgia
August 2008

DR. YI PAN
Chair and Professor
Georgia State University

PREFACE

In recent decades, machine learning techniques have been widely applied to bioinformatics. Many positive results have indicated that machine learning methods are useful for solving complex biomedical problems too difficult to solve by experts. Traditionally, researchers do biomedical research by using their knowledge and intelligence, performing experiments by hands and eyes, and processing data by basic statistical and mathematical tools. Due to huge amounts of biological data and a very large number of possible combinations and permutations of various biological sequences, the conventional human intelligence-based methods cannot work effectively and efficiently. So artificial intelligence techniques such as machine learning can play a critical role in complex biomedical applications.

Experts from different domains have contributed chapters to this book, which feature novel machine learning methods and their applications in bioinformatics. Relevant machine learning methods include support vector machines, kernel machines, feature selection, neural networks, evolutionary computation, statistical learning, fuzzy logic, supervised learning, clustering, ensemble learning, Bayesian networks, linear regression, principal components analysis, hidden Markov models, entropy-based information methods, and many others. The 20 chapters of the book are organized in a convenient order, based on their contents, so as to enable the readers to easily gather information in a progressive manner. A concise summary of each chapter follows.

In Chapter 1, Kung and Mak present feature selection methods such as the support vector machine recursive feature elimination (SVM-RFE), filter methods, and wrapper methods, in application to microarray data. Filter methods are based on input and output correlation statistics between input and predictions, or signal-to-noise (SNR) statistics, independent of the classifier or predictor. The development of microarray technology has brought with it problems that are interesting, both from statistical and biological perspectives. One important problem is to identify important genes that are relevant to distinguish cancerous samples from benign samples, or different cancer types. In the SVM-RFE, the magnitude of the weight connected to a particular feature is used as the ranking criteria for selection. The methods are illustrated in selection of important genes and in prediction of protein subcellular localization. In the protein subcellular localization, whether a protein lies in the cytoplasm, nuclear, extracellular, mitochondrial, or nuclear location is predicted from its amino acid sequence.

In Chapter 2, Menjoge and Welsch give a new feature selection method using 1-norm SVM and 2-norm SVM techniques, where the weights are used as regularization terms of 1-norm and 2-norm forms. Results show that these methods perform well as compared with other methods. The elastic net, in particular, demonstrates excellent classification accuracy. However, none of the methods dominate the other methods in both selecting a small number of variables and classifying data sets.

In Chapter 3, Kim and Park discuss adaptive supervised machine learning algorithms since the adaptive classifiers avoid expensive recomputation of the solution from scratch. Both an adaptive KDA/RMSE (aKDA/RMSE) based on updating the QR decomposition and an adaptive KDA/MSE based on updating the UTV decomposition KDA/MSE-UTV is proposed. These new kernel classifiers can be applied to compute leave-one-out cross-validation efficiently for bioinformatics applications.

In Chapter 4, Pang, Havukkala, Hu, and Kasabov propose a new gene selection method with better bootstrapping consistency for reliable microarray data analysis. The method ensures the reliability and generalizability of microarray data analysis, which thereby leads to an improvement of disease classification performance. Compared with the traditional gene selection methods without using consistency measurement, bootstrapping consistency method provides more accurate classification results. More importantly, results demonstrate that gene selection with the consistency measurement is able to enhance the reproducibility and consistency in microarray data analysis and proteomics-based diagnostics systems.

In Chapter 5, Wang and Palade introduce a series of fuzzy-based techniques, including the fuzzy gene selection method, the fuzzy C-mean clustering-based enhanced gene selection method, and the neuro-fuzzy ensemble approach for building a microarray cancer classification system. Three benchmark microarray cancer data sets, namely, the leukemia cancer data set, colon cancer data set, and lymphoma cancer data set, are used for simulations. The experimental results show that fuzzy-based systems can be efficient tools for microarray data analysis.

In Chapter 6, Li and Yang provide an ensemble learning method with feature selection to improve generalization performance of single classifiers from three aspects. Experiments on benchmark data show that genetic algorithm-based multitask learning (GA-MTL) is more effective than the earlier heuristic algorithms. The algorithms are demonstrated on a brain glioma data set to show the use of the algorithm as an alternative tool for bioinformatics applications.

In Chapter 7, Ahmad, Singh, Araúzo-Bravo, and Sarai study machine learning methods such as neural networks and support vector machines to predict one-dimensional features of protein structures, such as secondary structure, solvent accessibility, and coordination number, and more recently one-dimensional functional properties such as binding sites. The prediction techniques have been shown to have good performance even in the absence of known homology to other proteins. The computational similarities of the methods are highlighted. Common standards for making such sequence-based predictions are also developed.

In Chapter 8, Bu, Li, Gao, Yu, Xu, and Li give a new protein structure prediction method. Despite significant progresses made recently, every protein structure prediction method still possesses limitations. To overcome such shortcomings, a natural idea

is integrating the strengths of different methods to obtain more accurate structures by boosting some weaker predictors into a stronger one. As suggested by recent CASP competitions, the consensus-based prediction strategies usually outperform others by generating better results.

In Chapter 9, Gubbi, Shilton, and Palaniswami investigate different kernel machines in relation to protein structure prediction. Amino acids arrange themselves in 3D space in stable thermodynamic conformations, referred to as native conformation, and the protein becomes active in this state. Thermodynamic interactions include formation of hydrogen bonding, hydrophobic interactions, electrostatic interactions, and complex formation between metal ions. Protein molecules are quite complex in nature and often made up of repetitive subunits.

In Chapter 10, Jin and Zhang give a new method to predict protein subcellular locations based on SVM with evolutionary granular kernel trees (EGKT) and the one-versus-one voting approach. The new method can effectively incorporate amino acid composition information and combine binary SVM models for protein subcellular location prediction.

In Chapter 11, Liao discusses three applications, where the long-range correlations are believed to be essential, by using specific classification and prediction schemes: hidden Markov models for transmembrane protein topology, stochastic context-free grammars for RNA folding, and global structural profiling for antisense oligonucleotide efficacy. By first examining the limitations of present models, some expansions to capture and incorporate long-range features from the aspects of model architecture, learning algorithms, hybrid models, and model equivalence are made. The performance has been improved consequently.

In Chapter 12, Reddy, Weng, and Chiang give a novel optimization framework that searches the neighborhood regions of the initial alignment in a systematic manner to explore the multiple local optimal solutions. This effective search is achieved by transforming the original optimization problem into its corresponding dynamical system and estimating the practical stability boundary of the local maximum. Results show that the popularly used EM algorithm often converges to suboptimal solutions, which can be significantly improved by the proposed neighborhood profile search.

In Chapter 13, Rajapakse and Ho give a novel approach to encode inputs to neural networks for the recognition of transcription start sites in RNA polymerase II promoter regions. The Markovian parameters are used as inputs to three neural networks, which learn potential distant relationships between the nucleotides at promoter regions. Such an approach allows for incorporating biological contextual information at the promoter sites into neural networks and in general implementing higher-order Markov models of the promoters. Experiments on a human promoter data set show an increased correlation coefficient rate of 0.69 on average, which is better than the earlier reported by the NNPP 2.1 method.

In Chapter 14, Xie, Wu, and Yan propose three eukaryotic promoter prediction algorithms, PromoterExplorer I, II, and III. PromoterExplorer I is developed based on relative entropy and information content. PromoterExplorer II takes different kinds of features as the input and adopts a cascade AdaBoost-based learning procedure to select features and perform classification. The outputs of these two methods are combined to

build a more reliable system, PromoterExplorer III. Consistent and promising results have been obtained, indicating the robustness of the method. The new promoter prediction technique compares favorably with the existing ones, including Promoter-Inspector, Dragon Promoter Finder (DPF), and First Exon Finder (FirstEF).

In Chapter 15, Yang, King, and Elnitski introduce a bidirectional promoter—a region along a strand of DNA that regulates the expression of genes that flank the region on either side. An algorithm is developed for the purpose of finding uncharacterized bidirectional promoters. Results of the analysis have identified thousands of new candidate head-to-head gene pairs, corroborated the 5′ ends of many known human genes, revealed new 5′ exons of previously characterized genes, and in some cases identified novel genes. More effective machine learning approaches to classifying these features will be useful for future computational analyses of promoter sequences.

In Chapter 16, Zhang and Nam review computational methods used for miRNA research with a special emphasis on machine learning algorithms. In particular, detailed descriptions of the case studies based on the kernel methods (support vector machines), probabilistic graphical models (Bayesian networks and hidden Markov models), and evolutionary algorithms (genetic programming) are given. The effectiveness of these methods was validated by various approaches including wet experiments and their contributions were successful in the domain of miRNA. A well-defined generative model, such as Bayesian networks or hidden Markov models, constructed from a known data set in the prediction of miRNAs, can be used for the rational design of artificial pre-/shRNAs.

In Chapter 17, Lee and Shatkay present several works on tag SNP selection and mapping disease locus based on association study using SNPs. Tag SNP selection uses redundancy in the genotype/haplotype data to select the most informative SNPs that predict the remaining markers as accurately as possible. In general, machine learning methods tend to do better than purely combinatorial methods and also are applicable to bigger data sets with hundreds of SNPs. Identifying SNPs in disease association study is more difficult, largely depends on the population under study, and often faces the problem of replication.

In Chapter 18, Chanda, Zhang, and Ramanathan elaborate the application of some well-known machine learning techniques such as support vector machines, neural networks, linear regression, principal components analysis, hidden Markov models, and entropy-based information theoretic methods to locate genetic factors for complex diseases such as cystic fibrosis and multiple myeloma. They focus on two aspects, namely, tag SNP selection or selectively choosing some SNPs from a given set of possibly thousands of markers as representatives of the remaining markers (that are not chosen) and machine learning models for detecting markers that have potential high association with given disease phenotypes.

In Chapter 19, Winters-Hilt presents a new channel current-based nanopore cheminformatics to provide an incredibly versatile method for transducing single molecule events into discernable channel current blockade levels. The DNA–DNA, DNA–protein, and protein–protein binding experiments that were described were novel in that they made critical use of indirect sensing, where one of the molecules in

the binding experiment is either a natural channel blockade modulator or is attached to a blockade modulator.

In Chapter 20, Ganta, Narasimhamurthy, Kasturi, and Acharya propose an information fusion model-based analytical and exploratory framework for biomedical informatics. The framework presents a suite of tools and a workflow-based approach to analyze and explore multiple biomedical information sources through information fusion. The goal is to discover hidden trends and patterns that could lead to better disease diagnosis, prognosis, treatment, and drug discovery. However, there is a limit to the extent of knowledge that can be extracted from individual data sets. Recent focus on techniques analyzing genomic data sources in an integrated manner through information fusion could alleviate problems with individual techniques or data sets.

We sincerely thank all the authors for their important contributions and timely cooperation for publication of this book. We also thank Jung-Hsien Chiang, Arpad Kelemen, Rui Kuang, Ying Liu, Xinghua Lu, Lakshmi K. Matukumalli, Tuan D. Pham, and Changhui C. Yan for their valuable comments. We thank editors Paul Petralia and Anastasia Wasko from Wiley and Sanchari Sil of Thomson Digital for their guidance and help. We would like to thank Nguyen N. Minh for formatting the book. Finally, we would like to thank Dr. Yi Pan for his constant guidance.

Atlanta, Georgia
Nanyang, Singapore
August 2008

YAN-QING ZHANG
JAGATH C. RAJAPAKSE

CONTRIBUTORS

Raj Acharya, Pennsylvania State University, University Park, Pennsylvania.

Shandar Ahmad, Kyushu Institute of Technology, Kyushu, Japan, and Jamia Millia Islamia, New Delhi, India.

Marcos J. Araúzo-Bravo, Kyushu Institute of Technology, Kyushu, Japan.

Dongbo Bu, University of Waterloo, Waterloo, Ontario, Canada, and Institute of Computing Technology, China

Pritam Chanda, The State University of New York, Buffalo, New York.

Hsiao-Dong Chiang, Cornell University, Ithaca, New York.

Laura L. Elnitski, National Institutes of Health, Bethesda, Maryland.

Xin Gao, University of Waterloo, Waterloo, Ontario, Canada.

Srivatsava R. Ganta, Pennsylvania State University, University Park, Pennsylvania.

Jayavardhana Gubbi, The University of Melbourne, Melbourne, Australia.

Ilkka Havukkala, Auckland University of Technology, Auckland, New Zealand.

Sy Loi Ho, Nanyang Technological University, Nanyang, Singapore.

Yingjie Hu, Auckland University of Technology, Auckland, New Zealand.

Bo Jin, Georgia State University, Atlanta, Georgia.

Nikola Kasabov, Auckland University of Technology, Auckland, New Zealand.

Jyotsna Kasturi, Pennsylvania State University, University Park, Pennsylvania.

Hyunsoo Kim, Georgia Institute of Technology, Atlanta, Georgia.

David C. King, Pennsylvania State University, University Park, Pennsylvania.

Sun-Yuan Kung, Princeton University, Princeton, New Jersey.

Phil H. Lee, Queen's University, Kingston, Ontario, Canada.

Guo-Zheng Li, Shanghai University, Shanghai, China.

Ming Li, University of Waterloo, Waterloo, Ontario, Canada.

ShuaiCheng Li, University of Waterloo, Waterloo, Ontario, Canada.

Li Liao, University of Delaware, Newark, Delaware.

Man-Wai Mak, The Hong Kong Polytechnic University, Hong Kong, China.

Rajiv S. Menjoge, Massachusetts Institute of Technology, Cambridge, Massachusetts.

Jin-Wu Nam, Seoul National University, Seoul, Korea.

Anand Narasimhamurthy, Pennsylvania State University, University Park, Pennsylvania.

Vasile Palade, Oxford University, Oxford, United Kingdom.

Marimuthu Palaniswami, University of Melbourne, Melbourne, Victoria, Australia.

Shaoning Pang, Auckland University of Technology, Auckland, New Zealand.

Haesun Park, Georgia Institute of Technology, Atlanta, Georgia.

Jagath C. Rajapakse, School of Computer Engineering, and The Bioinformatics Research Center, Nanyang Technological University, Nanyang, Singapore.

Murali Ramanathan, The State University of New York, Buffalo, New York.

Chandan K. Reddy, Wayne State University, Detroit, Michigan.

Akinori Sarai, Kyushu Institute of Technology, Kyushu, Japan.

Hagit Shatkay, Queen's University, Kingston, Ontario, Canada.

Alistair Shilton, University of Melbourne, Melbourne, Australia.

Yemlembam Hemjit Singh, Jamia Millia Islamia, New Delhi, India

Zhenyu Wang, Oxford University, Oxford, United Kingdom

Roy E. Welsch, Massachusetts Institute of Technology, Cambridge, Massachusetts.

Yao-Chung Weng, Cornell University, Ithaca, New York.

Stephen Winters-Hilt, University of New Orleans, New Orleans, Louisiana.

Shuanhu Wu, City University of Hong Kong, Hong Kong, China.

Xudong Xie, City University of Hong Kong, Hong Kong, China.

Jinbo Xu, Toyota Technological Institute, Chicago, Illinois.

Hong Yan, City University of Hong Kong, Hong Kong, China.

Jack Y. Yang, Harvard University, Cambridge, Massachusetts.

Mary Q. Yang, National Institutes of Health, Bethesda, Maryland.

Libo Yu, University of Waterloo, Waterloo, Ontario, Canada.

Aidong Zhang, The State University of New York, Buffalo, New York.

Byoung-Tak Zhang, Seoul National University, Seoul, Korea.

Yan-Qing Zhang, Georgia State University, Atlanta, Georgia.

1

FEATURE SELECTION FOR GENOMIC AND PROTEOMIC DATA MINING

Sun-Yuan Kung and Man-Wai Mak

1.1 INTRODUCTION

The extreme dimensionality (also known as the curse of dimensionality) in genomic data has been traditionally a serious concern in many applications. This has motivated a lot of research in feature representation and selection, both aiming at reducing dimensionality of features to facilitate training and prediction of genomic data.

In this chapter, N denotes the number of training data samples, M the original feature dimension, and the full feature is expressed as an M-dimensional vector process

$$\mathbf{x}(t) = [x_1(t), x_2(t), \ldots, x_M(t)]^T, \quad t = 1, \ldots, N.$$

The subset of features is denoted as an m-dimensional vector process

$$\mathbf{y}(t) = [y_1(t), y_2(t), \cdots, y_m(t)]^T \tag{1.1}$$

$$= [x_{s_1}(t), x_{s_2}(t), \cdots, x_{s_m}(t)]^T, \tag{1.2}$$

where $m \leq M$ and s_i stands for index of a selected feature.

From the machine learning's perspective, one metric of special interest is the sample–feature ratio N/M. For many multimedia applications, the sample–feature

Machine Learning in Bioinformatics. Edited by Yan-Qing Zhang and Jagath C. Rajapakse

ratios lie in a desirable range. For example, for speech data, the ratio can be as high as 100 : 1 or 1000 : 1 in favor of training data size. For machine learning, such a favorable ratio plays a vital role in ensuring the statistical significance of training and validation.

Unfortunately, for genomic data, this is often not the case. It is common that the number of samples is barely compatible with, and sometimes severely outnumbered by, the dimension of features. In such situation, it becomes imperative to remove the less relevant features, that is, features with low signal-to-noise ratio (SNR) [1].

It is commonly acknowledged that more features means more information available for disposal, that is,

$$I[A] \leq I[A \cup B)] \leq \cdots, \tag{1.3}$$

where A and B represent two features, say x_i and x_j, respectively, and $I(X)$ denotes information of X. However, the redundant and noisy nature of genomic data makes it not always advantageous but sometimes imperative to work with properly selected features.

1.1.1 Reduction of Dimensionality (Biological Perspectives)

In genomic applications, each gene (or protein sequence) corresponds to a feature in gene profiling (or protein sequencing) applications. Feature selection/representation has its own special appeal from the genomic data mining perspective. For example, it is a vital preprocessing stage critical for processing microarray data. For gene expression profiles, the following factors necessitate an efficient gene selection strategy.

1. *Unproportionate Feature Dimension w.r.t. Number of Training Samples.* For most genomic applications, the feature dimension is excessively higher than the size of the training data set. Some examples of the sample–feature ratios N/M are

protein sequences	$\rightarrow$	1 : 1
microarray data	$\rightarrow$	1 : 10 or 1 : 100

   *S*uch an extremely high dimensionality has a serious and adverse effect on the performance. First, high dimensionality in feature spaces increases the computational cost in both (1) the learning phase and (2) the prediction phase. In the prediction phase, the more the features used, the more the computation required and the lower the retrieval speed. Fortunately, the prediction time is often linearly proportional to the number of features selected. Unfortunately, in the learning phase, the computational demand may grow exponentially with the number of features. To effectively hold down the cost of computing, the features are usually quantified on either *individual* or *pairwise* basis. Nevertheless, the quantification cost is in the order of $O(M)$ and $O(M^2)$ for individual and pairwise quantification, respectively (see Section 1.2).

3. *Plenty of Irrelevant Genes.* From the biological viewpoint, only a small portion of genes are strongly indicative of a targeted disease. The remaining "housekeeping" genes would not contribute relevant information. Moreover, their participation in the training and prediction phases could adversely affect the classification performance.
4. *Presence of Coexpressed Genes.* The presence of coexpressed genes implies that there exists abundant redundancy among the genes. Such redundancy plays a vital role and has a great influence on how to select features as well as how many to select.
5. *Insight into Biological Networks.* A good feature selection is also essential for us to study the underlying biological process that lead to the type of genomic phenomenon observed. Feature selection can be instrumental for interpretation/tracking as well as visualization of a selective few of most critical genes for *in vitro* and *in vivo* gene profiling experiments. The selective genes closely relevant to a targeted disease are called biomarkers. Concentrating on such a compact subset of biomarkers would facilitate a better interpretation and understanding of the role of the relevant genes. For example, for *in vivo* microarray data, the size of the subset must be carefully controlled in order to facilitate an effective tracking/interpretation of the underlying regulation behavior and intergene networking.

1.1.2 Reduction of Dimensionality (Computational Perspectives)

High dimensionality in feature spaces also increases uncertainty in classification. An excessive dimensionality could severely jeopardize the generalization capability due to overfitting and unpredictability of the numerical behavior. Thus, feature selection must consider a joint optimization and sometimes a delicate trade-off of the computational cost and prediction performance. Its success lies in a systematic approach to an effective dimension reduction while conceding minimum sacrifice of accuracy.

Recall from Equation 1.3 that the more the features the higher the achievable performance. This results in a monotonically increasing property: the more the features selected, the more the information is made available, as shown in the lower curve in Fig. 1.1a.

However, there are a lot of not-so-informative genomic features that are noisy and unreliable. Their inclusion is actually much more detrimental (than beneficial), especially in terms of numeric computation. Two major and serious adverse effects are elaborated below:

- *Data Overfitting.* Note that overoptimizing the training accuracy as the exclusive performance measure often results in overfitting the data set, which in turn degrades generalization and prediction ability.

  It is well known that data overfitting may happen in two situations: one is when the feature dimension is reasonable but too few training data are available; the other is when the feature dimension is too high even though there is a

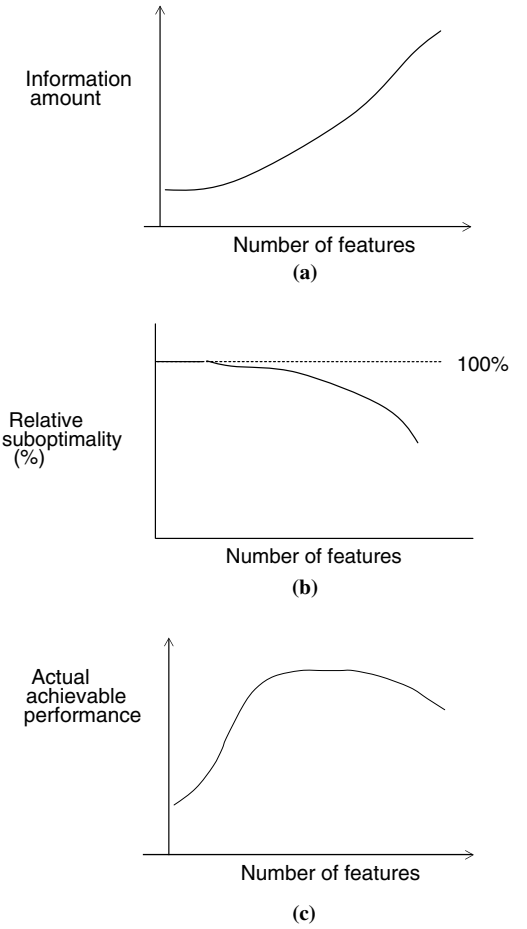


Figure 1.1 (a) Monotonic increasing property of the total information available. (b) Relative performance versus the feature size taking into consideration data overfitting and limited computational resources. (c) Nonmonotonic increasing property of the actual classification performance achievable. The best performance is often achieved by selecting an optimal size instead of the full set of available features.

reasonable amount of training data. What matters most is the ratio between the feature dimension and the size of the training data set. In short, classification/generalization depends on the sample–feature ratio.

Unfortunately, for many genomic applications, the feature dimension can be as high or much higher than the size of the training data set. For these applications, overtraining could significantly harm generalization and feature reduction is an effective way to alleviate the overtraining problem.

- *Suboptimal Search.* Practically, the computational resources available for most researchers are deemed to be inadequate, given the astronomical amounts of genomic data to be processed. High dimensionality in feature spaces increases

uncertainty in the numerical behaviors. As a result, a computational process often converges to a solution far inferior to the true optimum, which may compromise the prediction accuracy.

In conclusion, when the feature size is too large, the degree of suboptimality must reflect the performance degradation caused by data overfitting and limiting computational resource (see Fig. 1.1b). This implies a nonmonotonic property on achievable performance w.r.t. feature size, as shown in Fig. 1.1c. Accordingly, but not surprisingly, the best performance is often achieved by selecting an optimal subset of features. The use of any oversized feature subsets will be harmful to the performance. Such a nonmonotonic performance curve, together with the concern on the processing speed and cost, prompts the search for an optimal feature selection and dimension reduction.

Before we proceed, let us use a subcellular localization example to highlight the importance of feature selection.

Example 1 (Subcellular localization). *Profile alignment support vector machines (SVMs) [2] are applied to predict the subcellular location of proteins in an eukaryotic protein data set provided by Reinhardt and Hubbard [3]. The data set comprises 2427 annotated sequences extracted from SWISSPROT 33.0, which amounts to 684 cytoplasm, 325 extracellular, 321 mitochondrial, and 1097 nuclear proteins. Fivefold cross-validation was used to obtain the prediction accuracy. The accuracy and testing time for different number of features selected by a Fisher-based method [4] are shown in Fig. 1.2. This example offers an evidence of the nonmonotonic performance property based on real genomic data.* □

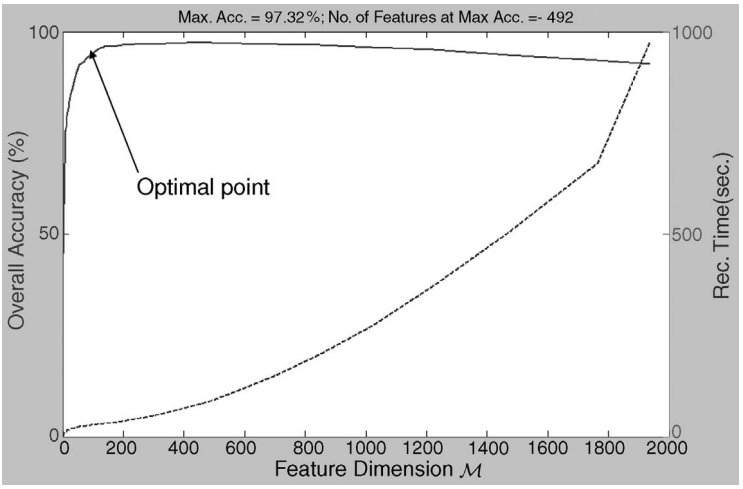


Figure 1.2 Real data supporting the monotonic increasing property. *Upper curve*: performance reaches a peak by selecting an optimal size instead of the full set of the features available. *Lower curve*: the computational time goes up (more than linear rate) as the number of features increases.

1.1.3 How Many Features to Select or Eliminate?

The question now is how many features should be retained, or equivalently how many should be eliminated? There are two ways to determine this number.

1. *Predetermined Feature Size.* A common practice is to have a user-defined threshold, but it is hard to determine the most appropriate threshold. For some applications, we occasionally may have a good empirical knowledge of the desirable size of the subset. For example, how many genes should be selected from, say, the 7129 genes in the leukemia data set [5]? Some plausible feature dimensions are as follows:
   (a) From classification/generalization performance perspective, a sufficient sample–feature ratio would be very desirable. For this case, empirically, an order of 100 genes seems to be a good compromise.
   (b) If the study concerns a regulation network, then a few extremely selective genes would allow the tracking and interpretation of cause–effect between them. For such an application, 10 genes would be the right order of magnitude.
   (c) For visualization, two to three genes are often selected for simultaneous display.
2. *Prespecified Performance Threshold.* For most applications, one usually does not know *a priori* the right size of the subset. Thus, it is useful to have a preliminary indication (formulated in a simple and closed-form mathematical criterion) on the final performance corresponding to a given size. Thereafter, it takes a straightforward practice to select/eliminate the features whose corresponding criterion functions are above/below a predefined threshold.

1.1.4 Unsupervised and Supervised Selection Criteria

The features selected serve very different objectives for unsupervised versus supervised learning scenarios (see Fig 1.3). Therefore, each scenario induces its own type of criterion functions.

1.1.4.1 Feature Selection Criteria for Unsupervised Cases In terms of unsupervised cases, there are two very different ways of designing the selection criteria. They depend closely on the performance metric, which can be either fidelity-driven or classification-driven.

1. *Fidelity-Driven Criterion.* The fidelity-driven criterion is motivated by how much of the original information is retained (or lost) when the feature dimension is reduced. The extent of the pattern spread associated with that feature is evidently reflected in the second-order moment for each feature x_i, $i = 1, \cdots, M$. The larger the second-order moment, the wider the spread, thus the more likely the feature x_i contains useful information.

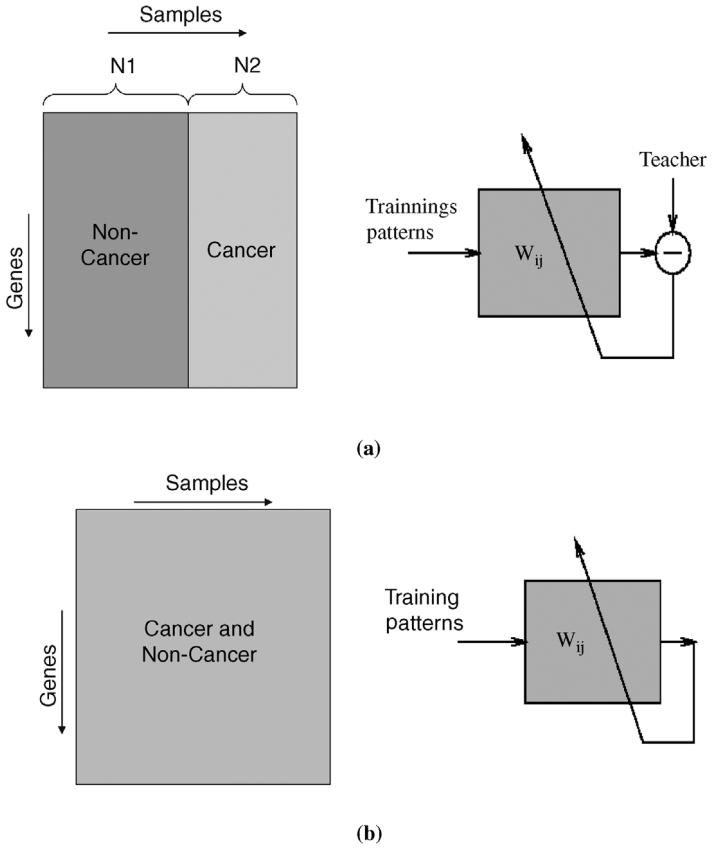


Figure 1.3 Difference between (a) supervised and (b) unsupervised feature selection.

There are two major types of fidelity-driven metrics:

- A performance metric could be based on the so-called mutual information: $I(\mathbf{x}|\mathbf{y})$.
- An alternative measure could be one which minimizes the reconstruction error:

$$\epsilon(\mathbf{x}|\mathbf{y}) \equiv \min_{\mathbf{y}\in\Re^m} ||\mathbf{x} - \hat{\mathbf{x}}_\mathbf{y}||,$$

where $\hat{\mathbf{x}}_\mathbf{y}$ denotes the estimate of $\mathbf{x}$ based on $\mathbf{y}$.

2. *Classification-Driven Criterion.* From the classification perspective, separability of data subclusters plays an important role. Thus, the corresponding criterion depends on how well can the selected features reveal the subcluster structure. The higher-order statistics, known as independent component analysis (ICA), has been adopted as a popular metric. For more discussion on this subject, see Ref. [6].

1.1.4.2 Feature Selection Criteria for Supervised Cases The ultimate objective for supervised cases lies in a high classification/predition accuracy. Ideally speaking, if the classification information is known, denoted by C, the simplest criterion will be $I(C|\mathbf{y})$. However, the comparison between $I(C|\mathbf{x})$ and $I(C|\mathbf{y})$ often provides a more useful metric. For example, it is desirable to have

$$I(C|\mathbf{y}) \rightarrow I(C|\mathbf{x}),$$

while keeping the feature dimension m as small as possible. However, the above formulation is numerically difficult to achieve. The only practical solution known to exist is the one making the full use of the feedback from the actual classification result, which is computationally very demanding. (The feedback-based method is related to the wrapper approach to be discussed in Section 1.4.5.)

To overcome this problem, an SNR-type criterion based on the Fisher discriminant analysis is very appealing. (Note that the Fisher discriminant offers a convenient metric to measure the interclass separability embedded in each feature.) Such a feature selection approach entails computing Fisher's discriminant denoted as FD_i, $i = 1, \ldots, M$, which represents the ratio of intercluster distance to intracluster variance for each individual feature. (This related to the filter approach to be discussed in Section 1.4.1.)

1.1.5 Chapter Organization

The organization of the chapter is as follows. Section 1.2 provides a systematic way to quantify the information/redundancy of/among features, which is followed by discussions on the approaches to ranking the relevant features and eliminating the irrelevant ones in Section 1.3. Then, in Section 1.4, two supervised feature selection methods, namely filter and warper, are introduced. For the former, the features are selected without explicit information on classifiers nor classification results, whereas for the latter, the select requires such information explicitly. Section 1.5 introduces a new scenario called self-supervised learning in which prior known group labels are assigned to the features, instead of the vectors. A novel SVM-based feature selection method called Vector-Index-Adaptive SVM, or simply VIA-SVM, is proposed for this new scenario. The chapter finishes with experimental procedures showing how self-supervised learning and VIA-SVM can be applied to (protein-sequence-based) subcellular localization analysis.

1.2 QUANTIFYING INFORMATION/REDUNDANCY OF/AMONG FEATURES

Quantification of information and redundancy depends on how the information is represented. A representative feature is the one that can represent a group of similar features. Denote S as a feature subset, that is, $S \equiv \{y_i\}$, $i = 1, \ldots, m$. In addition to the general case, what of most interest is either a single individual feature $m = 1$ or a pair of features $m = 2$. A generic term $I(S)$ will be used temporarily to denote the information pertaining to S, as the exact form of it has to depend on the application scenarios.

Recall that there are often a large number of features in genomic data sets. To effectively hold down the cost of computing, we have to limit the number of features simultaneously considered in dealing with the interfeature relationship. More exactly, such computational consideration restricts us to three types of quantitative measurements of the feature information:

1. *Individual Information*: The quantification cost is in the order of $O(M)$.
2. *Pairwise Information*: The quantification cost becomes now $O(M^2)$.
3. *Groupwise Information:* (with three or more features).

The details can be found in the following text.

1.2.1 Individual Feature Information

Given a single feature x_i, its information is denoted as $I(x_i)$. Such a measure is often the most effective when the features are statistically independent. This leads to the individual ranking scheme in which only the information and/or discriminative ability of individual features are considered. This scheme is the most straightforward, since each individual feature is independently (and simultaneously) evaluated. Let us use a hypothetical example to illustrate the individual ranking scheme.

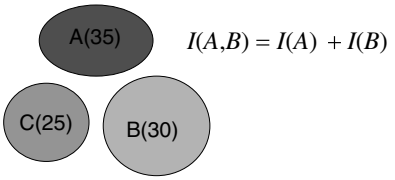
Example 2 (Three-party problem—without interfeature redundancy). *The individual ranking method works the best when the redundancy plays no or minimal role in affecting the final ranking. In this example, each area in Fig. 1.4 represents one feature. The size of the area indicates the information or discriminativeness pertaining to a feature. In the figure, no "overlapping" between elements symbolizes the fact that there exists no mutual redundancy between the features. In this case, the combined information of any two features is simply the sum of two individual amounts. For example,* $I(A \cup B) = I(A) + I(B) = 35 + 30 = 65$.

When all the features are statistically independent, it corresponds to the fact that there is no overlap pictorially. All methods lead to the same and correct result. It is, however, a totally different story with the statistically dependent cases. □

Unfortunately, the downside of considering the feature individually is that it does not fully account for the redundancy among the features. For example, it is very possible that two highest-rank individual features share a great degree of similarity. As a result, the inclusion of both features would amount to a waste of resource. In fact, one needs to take the interfeature relationship (such as mutual similarity/redundancy) into account. This problem can be alleviated by adopting either pairwise or groupwise information to be discussed next.

1.2.2 Pairwise Feature Information

Given a pair of features x_i and x_j, its information is denoted as $I(x_i \cup x_j)$. The main advantage of studying the pairwise relationship is to provide a means to identify the *similariy/redundancy* of the pair. A fundamental and popular criterion is based on



(a) Three-party problem

(b) Forward selection

(c) Backward elimination

Selection Method	Select 1 Feature	Select 2 out of 3 Features
Independent ranking	A	AB
Forward selection	A	AB
Backward elimination	A	AB
Correct solution	A	AB

(d)

Figure 1.4 (a) Three-party problem without redundancy. No "overlapping" between elements symbolizes the fact that no mutual redundancy exists between the features. (b) Consecutive result of the step-by-step forward selection. (c) Consecutive result of the step-by-step backward elimination. (d) Table illustrating search results of different strategies.

correlation. For example, the Pearson correlation coefficient is defined as

$$r_{x_i x_j} = \frac{E[x_i x_j]}{\text{var}(x_i)\text{var}(x_j)} = E[x_i x_j].$$

Without loss of generality, here we shall simply assume that both the features x_i and x_j are zero mean with unit variance $\text{var}(x_i)\text{var}(x_j) = 1$.

From the practical decision perspective, there are again two pairwise criteria: (1) mutual predictability and (2) mutual information.

1. *Mutual Predictability.* The mutual predictability represents the ability of estimating one feature from another feature. Such a metric is also closely tied with the (Pearson) correlation coefficients, that is,

$$\hat{x}_j = E[x_j | x_i] = r_{x_i x_j} x_i. \tag{1.4}$$

   When there is no correlation, that is, $r_{x_i x_j} = 0$, then $\hat{x}_j = 0$ regardless of whatever the value of x_i is. In other words, the information of x_i offers no information about x_j. In general, the predictability is a function of $r_{x_i x_j}$; the higher the correlation, the more predictable is x_j given x_i.
2. *Mutual Information.* Suppose that there exists pairwise redundancy, then $I(x_i \cup x_j) \leq I(x_i) + I(x_j)$. The mutual information $I(x_i, x_j)$ is also a function of $r_{x_i x_j}$, the higher the correlation, the greater the mutual information.