Preface

A FEW WORDS ABOUT THE BOOK

This book represents my attempt to create a unique book on machine learning aspects of one of the important tasks in bioinformatics – microarray gene expression based cancer classification. That is, the input called a training set consists of expression level values measured for many (order of thousands or even tens of thousands) genes at once. Such measurements were however done only for a few patients, thus making the number of features (genes) much larger than the number of instances. Given such information, the task is to correctly assign class labels to the instances outside the training set. In this book, only binary or two-class classification problems are treated. Two classes of the data are 'tumor' ('cancer', 'diseased') and 'normal' ('healthy').

Although this book considers gene expression based cancer classification as an application, this does not imply that the methods described in this book cannot be used as they are for other bioinformatics tasks, where a data set is small but high dimensional; hence, the word 'bioinformatics' in the book title.

The uniqueness of this book stems from the combination of three topics:

- machine learning,
- bioinformatics,
- MATLAB®.

There are a plenty of books on machine learning (check, e.g., the web site of IGI Global http://www.idea-group.com/ and search there for the phrase "machine learning"). There are a few books covering both machine learning and bioinformatics. However, to my best knowledge this is one of very few books that cover all three topics above in one volume.

As follows from its topic, the subject of the book lies on a crossroad between computer science and biology. Hence, the main intention was to write a book that could be used either as a textbook or a reference book by researchers and students

from both fields. Also my purpose was to write a book that could be suitable both for novices and seasoned practitioners, for people from both academia and industry. So, it was a very challenging and ambitious undertaking, and you, the readers of this book, will decide whether it was successfully accomplished or not. In this book I often use 'we' when addressing to you, dear readers, since I always assume your invisible presence and do not want to turn the whole process exclusively into my monolog. Everything we do on this "journey", we do together.

To meet all diverse goals and the needs of broad audience, I decided that each chapter shall comprise an independent or almost independent containment of knowledge that can be read and understood independently of other chapters. Each chapter aims at explaining one machine learning method only. In addition, the organization of all chapters tries to follow the same structure. Namely, each chapter begins with the main idea and theory, which a given method is based on, followed by algorithm description in pseudo code, demonstrating how to implement the method step-by-step. After that, MATLAB® code is given together with detailed comments on it. MATLAB® was chosen because it is widely used by the research community; it includes many useful toolboxes such as Statistics Toolbox™ and Bioinformatics Toolbox™.

Thus, as a whole, each chapter describes the process of algorithm design from the very beginning to the very end as it was my humble hope to teach the readers the best practices that they could apply in their everyday research work. In other words, I tried to provide a kind of a standard of how algorithms shall be described (though I am far from imposing the rules I deem to be good on readers) in giving lectures for students and researchers and in research reports/theses prepared by both students and researchers.

Of course, many people may object me at this point. For example, some (but not all, I believe) adepts of extreme programming (for a quick introduction to this promising programming technique to write good software, read (chromatic, 2003); 'chromatic' is not a mistake but a nickname) may say that documentation is unnecessary for already heavily commented code. From my industrial and academic experience, I disagree with this statement as professionally composed documentation can save many days and weeks of work both for users/customers and for researchers that did not take part in the development of that piece of software. For example, imagine a situation when all people who developed software suddenly left a company (optimists may say that they got better job offers elsewhere, while pessimists (and many realists nowadays) may say that all of them were laid off because of a worsened financial situation) and a new staff needs to quickly advance further while utilizing code developed by their predecessors. How do you think this could happen if the previous staff members did not leave any traces of their thoughts or

their code comments are scarce and not very informative? This is, of course, not an impossible mission but anyway difficult.

I am not an inventor of the principles I advocate here. For instance, Prof. David Donoho from Stanford University (http://www-stat.stanford.edu/~donoho/index.html) some time ago suggested and actively promotes open code (in many cases it is MATLAB®) that can be used to reproduce research results described in scientific articles. I see it as a good practice to adhere. Besides, several MATLAB®-based books influenced me while I was writing my own book. Among them are (Nabney, 2002), (Martinez & Martinez, Computational statistics handbook with MATLAB®, 2002), (Stork & Yom-Tov, 2004), (van der Heijden, Duin, de Ridder, & Tax, 2004), and (Martinez & Martinez, Exploratory data analysis with MATLAB®, 2005). These books as well as MATLAB® code taught me best practices to utilize in my own book. As a quick reference guide to MATLAB®, I can recommend a little book by Davis and Sigmon (Davis & Sigmon, 2005).

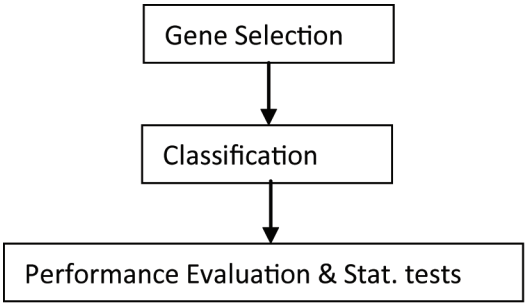
Other books that are relevant to the statistical aspects of bioinformatics and that target biologists include but not limited to (Zar, 1999), (van Emden, 2008), (Lee, 2010). The book of van Emden has a funny title "Statistics for the terrified biologists", lol. Does this mean that all or at least a majority of statisticians are cold-blooded (neither in the biological, nor in the criminal sense of these words) folks with unshaken resolution? Based on the laws of statistics, this is, of course, not true. But it would be interesting to see a book titled "Biology for the terrified statisticians". Hey, biologists out there, who of you is ready to write such a sweet revenge book?

Other books that may be complementary to the subject of my book are (Cohen, 2007), (Cristianini & Hann, 2007), (Alterovitz & Ramoni, 2007) (the last two books also concentrate on MATLAB® as the programming environment).

Although I mentioned earlier that each chapter can be read independently of other chapters, it does not mean that there are absolutely no connections between chapters. The book is structurally divided into five parts. Each part (except for the last one) begins with an introductory chapter providing a compressed summary of the topics and problems discussed in the chapters that follow.

The first group of chapters concerns the classification algorithms (or simply classifiers) most commonly employed in bioinformatics research working with gene expression data. The second group of chapters deals with feature or gene selection algorithms. Due to the huge number of algorithms, I tried to choose the algorithms built on as diverse principles as possible, though covering all existing types would certainly be unrealistic (e.g., searching with Google Scholar for the exact phrase "feature selection" returned 96,200 links while searching for "gene selection" resulted in 11,700 links (searches were performed in May of 2009)). The third group of chapters concentrates on classifier ensembles, i.e. several classifiers whose predictions are combined together in order to form the final vote.

Figure 1. General scheme involving a single classifier

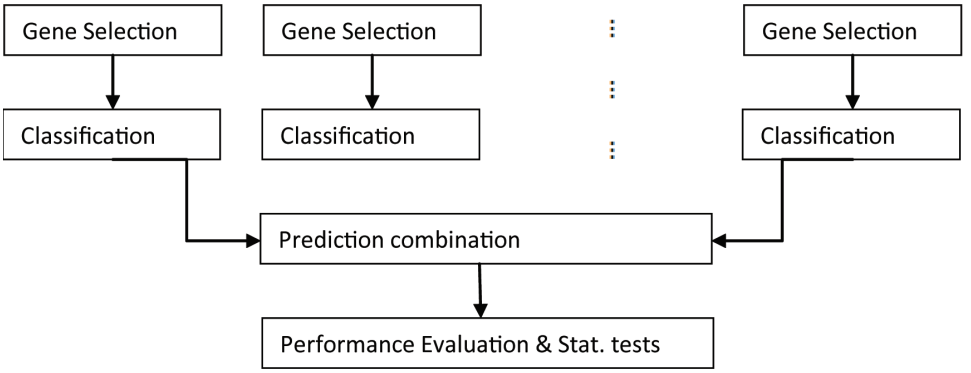


Compared to a single classifier, a classifier ensemble can deliver better and more stable performance. The fourth group of chapters describes advanced performance evaluation methods for single classifiers and classifier ensembles and statistical tests related to this evaluation. These methods, unlike many currently employed, are especially tuned to small-sample size problems such as classification of microarray gene expression data. Finally, the last part comprises a single chapter demonstrating how to utilize code spread across other four groups of chapters. It includes four examples putting feature selection, data classification, ensemble of classifiers, and performance evaluation techniques into one piece.

Such a book structure well matches to the general scheme used to solve problems like gene expression based cancer classification, where the number of features far exceeds the number of available instances (samples taken from patients). In the case when a single classifier is used, this scheme is shown in Figure 1, while in the case of a classifier ensemble, it is given in Figure 2.

Since there are too many genes compared to the number of instances, it is obvious that not all of them are related to cancer. In other words, many of them can be

Figure 2. General scheme involving an ensemble of classifiers



safely removed from the classification model so that they do not participate in the classification process. Their removal is also necessitated by the fact that such redundant genes, if left in a data set, will degrade the generalization ability of a classifier, i.e. the classifier will perfectly classify the training data but will suddenly degrade in performance on new, previously unseen (out-of-sample, test) instances.

As the task is not to perfectly classify the training data that are utilized for learning how to assign class labels but to correctly classify test data, generalization of the trained classifier is of great importance. With poor generalization, the classification mode is unusable as biologists and doctors cannot trust to the reliability of results. In other words, poor generalization is associated with 'no trustful outcome'. The removal of irrelevant and redundant genes out of a data set or alternatively the selection of highly relevant for disease prediction genes is called feature or gene selection. Usually, a small subset of the original genes remains after this procedure.

Once relevant genes have been selected, the next stage is the classification with selected genes, which typically consists of two steps: classifier training (optional if a classifier does not need it) on the training data and testing of the trained classifier on the test data. As the available data are scarce, the good solution is to artificially generate test instances based on the training instances. This is a rather new approach (see, e.g., (Braga-Neto & Dougherty, Bolstered error estimation, 2004), (Braga-Neto, Fads and fallacies in the name of small-sample microarray classification - A highlight of misunderstanding and erroneous usage in the applications of genomic signal processing, 2007), (Li, Fang, Lai, & Hu, 2009)), but I consider it very promising and appealing, compared to the other known techniques that reserve a part of the training instances for testing, thus reducing both training and test set sizes. When every extra instance is of importance, reducing either training or test data can easily bias classification results, i.e. to make them over-optimistic.

After a classifier or an ensemble of classifiers assigned class labels and/or class probability estimates to test instances, the final stage is the computation of the performance evaluation characteristics and running statistical tests in order to discover the statistically significant difference in classification performance (or the absence of such a difference) between several competing classification models.

Thus, the book covers the entire classifier design when using microarray gene expression data as input. What may look missing is probably the links to gene expression data sets to be used in experiments with the algorithms in this book. The paper by Statnikov et al. (Statnikov, Wang, & Aliferis, 2008) cites 22 data sets. Another paper (Yoon, Lee, Park, Bien, Chung, & Rha, 2008) refers to three prostate cancer data sets. Links to some data sets can be found at http://www.genecbr.org/links.htm. As active researchers, I suppose that you, dear readers, also know well where the microarray data are located on the web.

I hope readers will find this book useful for their work and education. Any questions, suggestions or comments as well as bug reports can be sent to me (olegokun@yahoo.com).

This book would not appear without several people whom I would like to express my gratitude. I would like to thank Dr. Mehdi Khosrow-Pour, President of IGI Global, for his kind invitation to write this book. I deeply appreciate patience and professional support of Julia Mosemann, Director of Book Publications who was always eager to help in difficult situations. Critical comments of two anonymous reviewers provided the valuable and objective opinion about my work and thus helped me to dramatically improve the final book draft before sending it to IGI. I am also indebted to my parents, Raisa and Gregory Okun, for their never-ending support, wise advice and love through my entire life.

This book is dedicated to my parents and my son Antoshka.

Oleg Okun
Malmo, Sweden, 9 May 2010

REFERENCES

Alterovitz, G., & Ramoni, M. F. (Eds.). (2007). *Systems bioinformatics: an engineering case-based approach*. Norwood, MA: Artech House.

Braga-Neto, U. M. (2007). Fads and fallacies in the name of small-sample microarray classification - A highlight of misunderstanding and erroneous usage in the applications of genomic signal processing. *IEEE Signal Processing Magazine*, *24*(1), 91–99. doi:10.1109/MSP.2007.273062

Braga-Neto, U. M., & Dougherty, E. R. (2004). Bolstered error estimation. *Pattern Recognition*, *36*(7), 1267–1281. doi:10.1016/j.patcog.2003.08.017

Chromatic. (2003). *Extreme programming pocket guide.* Sebastopol, CA: O'Reilly Media.

Cohen, W. W. (2007). *A computer scientist's guide to cell biology.* New York: Springer Science+Business Media.

Cristianini, N., & Hann, M. W. (2007). *Introduction to computational genomics: a case studies approach*. Cambridge, UK: Cambridge University Press.

Davis, T. A., & Sigmon, K. (2005). *MATLAB® Primer* (7th ed.). Boca Raton, FL: Chapman & Hall/CRC Press.

Lee, J. K. (Ed.). (2010). *Statistical bioinformatics: for biomedical and life science researchers*. Hoboken, NJ: Wiley-Blackwell.

Li, D.-C., Fang, Y.-H., Lai, Y.-Y., & Hu, S. C. (2009). Utilization of virtual samples to facilitate cancer identification for DNA microarray data in the early stages of an investigation. *Information Sciences*, *179*(16), 2740–2753. doi:10.1016/j.ins.2009.04.003

Martinez, W. L., & Martinez, A. R. (2002). *Computational statistics handbook with MATLAB*. Boca Raton, FL: Chapman & Hall/CRC Press.

Martinez, W. L., & Martinez, A. R. (2005). *Exploratory data analysis with MATLAB*. Boca Raton, FL: Chapman & Hall/CRC Press.

Nabney, I. T. (2002). *NETLAB: algorithms for pattern recognition*. London: Springer-Verlag.

Statnikov, A., Wang, L., & Aliferis, C. F. (2008). A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC Bioinformatics*, *9*(319).

Stork, D. G., & Yom-Tov, E. (2004). *Computer manual in MATLAB to accompany Pattern Classification* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

van der Heijden, F., Duin, R., de Ridder, D., & Tax, D. M. (2004). *Classification, parameter estimation and state estimation: an engineering approach using MATLAB*. Hoboken, NJ: John Wiley & Sons. doi:10.1002/0470090154

van Emden, H. (2008). *Statistics for the terrified biologists*. Hoboken, NJ: John Wiley & Sons.

Yoon, Y., Lee, J., Park, S., Bien, S., Chung, H. C., & Rha, S. Y. (2008). Direct integration of microarrays for selecting informative genes and phenotype classification. *Information Sciences*, *178*(1), 88–105. doi:10.1016/j.ins.2007.08.013

Zar, J. H. (1999). *Biostatistical analysis* (4th ed.). Upper Saddle River, NJ: Prentice Hall/Pearson Education International.