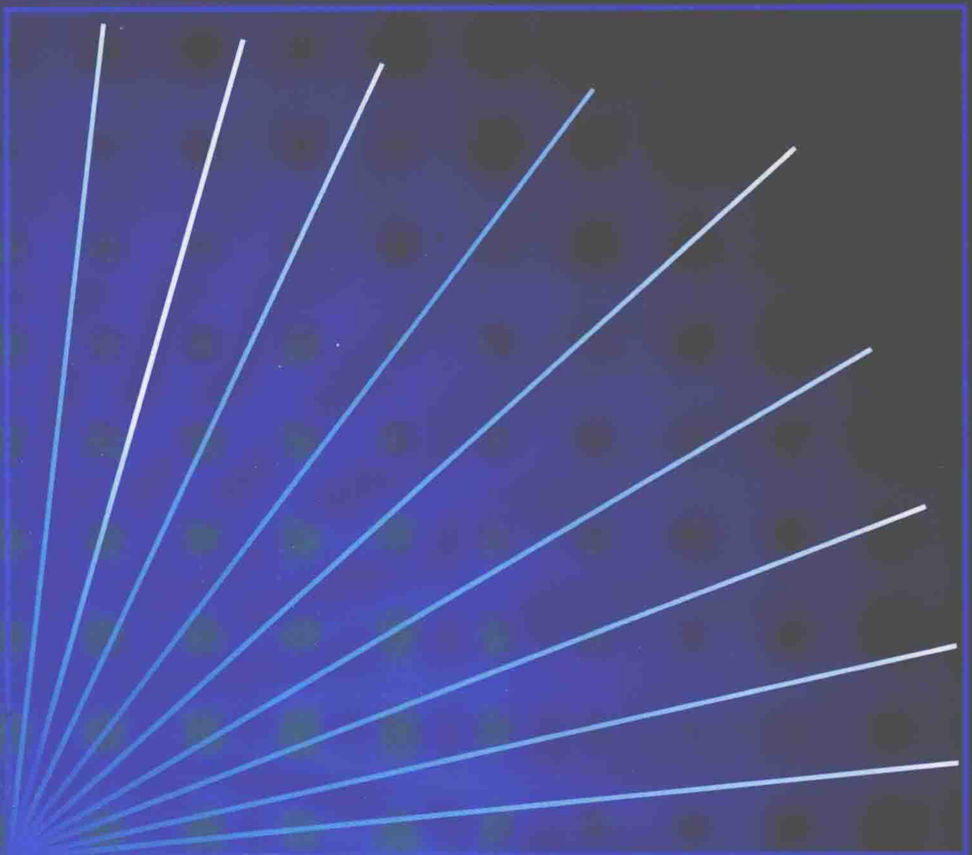
Evaluating Learning Algorithms

A Classification Perspective

Nathalie Japkowicz ▪ Mohak Shah



CAMBRIDGE

Evaluating Learning Algorithms

A Classification Perspective

NATHALIE JAPKOWICZ
University of Ottawa

MOHAK SHAH
McGill University

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore,
São Paulo, Delhi, Dubai, Tokyo, Mexico City

Cambridge University Press
32 Avenue of the Americas, New York, NY 10013-2473, USA

www.cambridge.org
Information on this title: www.cambridge.org/9780521196000

First published 2011

Printed in the United States of America

A catalog record for this publication is available from the British Library.

Library of Congress Cataloging in Publication data

Japkowicz, Nathalie.
Evaluating Learning Algorithms : A Classification Perspective / Nathalie Japkowicz, Mohak Shah.
p. cm.
Includes bibliographical references.
ISBN 978-0-521-19600-0
1. Machine learning. 2. Computer algorithms – Evaluation. I. Shah, Mohak. II. Title.
Q325.5.J37 2011
006.3′1–dc22 2010048733

ISBN 978-0-521-19600-0 Hardback

Cambridge University Press has no responsibility for the persistence or accuracy of URLs for external or third-party Internet Web sites referred to in this publication and does not guarantee that any content on such Web sites is, or will remain, accurate or appropriate.

Evaluating Learning Algorithms

The field of machine learning has matured to the point where many sophisticated learning approaches can be applied to practical applications. Thus it is of critical importance that researchers have the proper tools to evaluate learning approaches and understand the underlying issues.

This book examines various aspects of the evaluation process with an emphasis on classification algorithms. The authors describe several techniques for classifier performance assessment, error estimation and resampling, and obtaining statistical significance, as well as selecting appropriate domains for evaluation. They also present a unified evaluation framework and highlight how different components of evaluation are both significantly interrelated and interdependent. The techniques presented in the book are illustrated using R and WEKA, facilitating better practical insight as well as implementation.

Aimed at researchers in the theory and applications of machine learning, this book offers a solid basis for conducting performance evaluations of algorithms in practical settings.

Nathalie Japkowicz is a Professor of Computer Science at the School of Information Technology and Engineering of the University of Ottawa. She also taught machine learning and artificial intelligence at Dalhousie University and Ohio State University. Along with machine learning evaluation, her research interests include one-class learning, the class imbalance problem, and learning in the presence of concept drifts.

Mohak Shah is a Postdoctoral Fellow at McGill University. He earned a PhD in Computer Science from the University of Ottawa in 2006 and was a Postdoctoral Fellow at CHUL Genomics Research Center in Quebec prior to joining McGill. His research interests span machine learning and statistical learning theory as well as their application to various domains.

This book is dedicated to the memory of my father, Michel Japkowicz (1935–2008), who was my greatest supporter all throughout my studies and career, taking a great interest in any project of mine. He was aware of the fact that this book was being written, encouraged me to write it, and would be the proudest father on earth to see it in print today.

Nathalie

This book is dedicated to the loving memory of my father, Upendra Shah (1948–2006), who was my mentor in life. He taught me the importance of not falling for means but looking for meaning in life. He was also my greatest support through all times, good and bad. His memories are a constant source of inspiration and motivation. Here's to you Dad!

Mohak

Preface

This book was started at Monash University (Melbourne, Australia) and Laval University (Quebec City, Canada) with the subsequent writing taking place at the University of Ottawa (Ottawa, Canada) and McGill University (Montreal, Canada). The main idea stemmed from the observation that while machine learning as a field is maturing, the importance of evaluation has not received due appreciation from the developers of learning systems. Although almost all studies make a case for the evaluation of the algorithms they present, we find that many (in fact a majority) demonstrate a limited understanding of the issues involved in proper evaluation, despite the best intention of their authors. We concede that optimal choices cannot always be made due to limiting circumstances, and trade-offs are inevitable. However, the methods adopted in many cases do not reflect attention to the details warranted by a proper evaluation approach (of course there are exceptions and we do not mean to generalize this observation).

Our aim here is not to present the readers with yet another recipe for evaluation that can replace the current default approach. Rather, we try to develop an understanding of and appreciation for the different concerns of importance in the practical application and deployment of learning systems. Once these concerns are well understood, the other pieces of the puzzle fall quickly in place since the researcher is not left shooting in the dark. A proper evaluation procedure consists of many components that should all be considered simultaneously so as to correctly address their interdependence and relatedness. We feel that the best (read most easily understood) manner to bring this holistic view of evaluation to the fore is in the classification setting. Nonetheless, most of the observations that we make with regard to the various evaluation components extend just as well to other learning settings and paradigms since the underlying evaluation principles and objectives are essentially the same.

Altogether, this book should be viewed not only as a tool designed to increase our understanding of the evaluation process in a shared manner, but also as a first step in the direction of stimulating a community-wide debate on the relevance and importance of the evaluation of learning algorithms.

Incorporating concepts from both machine learning and statistics proved to be a bit more involved than we had first imagined. The main challenge was to integrate the ideas together and present them in a coherent manner. Indeed, sometimes the same terms are used in the two fields to mean different quantities while at other times, the same quantities are referred to by multiple names and notations. We have tried to put some aspects under a unified scheme (of both terminology and notation) but have left others to their more conventional usage, just to make sure that the reader can relate these to other texts. For instance, while we have used α for the confidence parameter in the statistical significance testing, we have also, in some places, used the common notion of p-value to relate to other discussions. Similarly, both P and Pr frequently appear in probabilistic contexts. We have used both these terms, keeping in mind their common use as well as a better readability of the text. To achieve this, we have used Pr when referring to events or probabilities for discrete variables. For other cases, e.g., distributions over continuous variables and priors, we use P or other symbols, as indicated in the text. However, with some exceptions, most notations are used locally and explained in their proper context to avoid confusion.

We have tried to illustrate the various methods and tests presented in the book with the use of the freely available R statistical package and WEKA machine learning toolkit. Our code, however, is in no sense optimal. Our main aim here was to illustrate the concepts in the simplest possible manner so that even the least experienced programmers could apply the code easily in order to immediately utilize the tools presented in the book. We hope to post better optimized code on the book Web page in the near future.

While our names figure on the cover, we cannot claim complete credit for the work presented in this book. This work was made possible thanks to the support of many people. The deficiencies or errors, however, are solely due to us. We would now like to take some space to thank them and acknowledge their support, advice, and understanding.

We would like to thank all our colleagues at the various institutions that hosted us while this book was in progress. They helped us form and develop our ideas on evaluation and stimulate our thoughts on various aspects of the problem, either directly or indirectly. These include: Peter Tischer, Ingrid Zuckerman, and Yuval Marom at Monash; Mario Marchand, Jacques Corbeil, and Francois Laviolette at Laval; Stan Matwin and Marcel Turcotte at the University of Ottawa; Chris Drummond and Peter Turney at the University of Ottawa and the National Research Council of Canada; Tal Arbel, D. Louis Collins, Doina Precup, and Douglas L. Arnold at McGill; the graduate students and postdoctoral Fellows William Klement, Guichong Li, Lisa Gaudette, Alex Kouznetsov, and Shiven Sharma at the University at Ottawa; Heidar Pirzadeh and Sara Shanian at Laval; and Dante De Nigris and Simon Francis at McGill. William, Alex, Guichong, and Shiven were also instrumental in running certain experiments, checking some of our formulas and code, and helping with the presentation, in various parts of the book. We also benefited greatly from discussions with Rocio Alaiz-Rodriguez during her visit to the University of Ottawa and, later, on-line. Conversations held about evaluation in the context of a collaboration with Health Canada were also quite enlightening and helped shape some of the ideas in this book. In particular, we would like to thank Kurt Ungar, Trevor Stocki, and Ian Hoffman for sharing their thoughts with us, as well as for providing us with data on Radioxenon Monitoring for the Detection of Nuclear Explosions.

Nathalie would like to thank, most particularly, James Malley of the National Institute of Health for helping her recognize the inadequacy of current evaluation practices in machine learning and the repercussions they may have in collaborative settings; and Chris Drummond with whom she had numerous discussions on evaluation, some of which have been ongoing for the past ten years.

Mohak would also like to extend a note of thanks to Ofer Dekel and Microsoft Research, Seattle, for hosting him there and the immensely productive discussions that helped invoke novel thoughts and ideas.

We would also like to acknowledge financial support from the Natural Science and Engineering Research Council of Canada.

Many thanks to our first editor at Cambridge University Press, Heather Bergman, whose confidence in our project was very uplifting. She made contract negotiations very easy, with her dynamism and encouragement. Lauren Cowles, who succeeded her as our editor, has been equally competent and helpful. Lauren indeed made the administrative process extremely easy and efficient, allowing us to devote more time to the ideas and contents of the book. Our copy editor Victoria Dahany deserves a special thank you for her meticulous work and the painstaking effort to refine our discussion without which this book would not have been in its present form. We would also like to thank Victoria for her encouraging notes during the copyediting phase that reinforced our belief in both the importance and pertinence of the subject matter. We would also like to thank David Jou, Marielle Poss, Katy Strong, and the Cambridge marketing team for their thorough professionalism and help with processing the book and disseminating the information as well as with design aspects of the marketing material. Also, the team at Aptara, especially Sweety Singh, Tilak Raj, and Pushpender Rathee, has been thoroughly professional in taking the book publication forward from copyediting to its final version.

Nathalie would also like to thank her husband, Norrin Ripsman, for sharing his experience with writing and publishing books. His advice on dealing with presses and preparing our material was particularly helpful. On a more personal note, she appreciated him for being there every step of the way, especially at times when the goal seemed so far away. Her daughter Shira also deserves great thanks for being the excellent girl that she is and bearing with her Mum's work all along. The baby-to-be, now lovely little Dafna, showed tremendous patience (in both her fetal and infant states), which made it possible for Nathalie to continue working on the project prior to and after her birth. Nathalie's father, Michel Japkowicz, and her mother, Suzanne Japkowicz, have also always been an unconditional source of loving support and understanding. Without their constant interest in her work, she would not be where she is today. Nathalie is also grateful to her in-laws, Toba and Michael Ripsman, for being every bit as supportive as her own parents during the project and beyond.

On the personal front, Mohak would like to acknowledge his mother Raxika Shah and his sister Tamanna Shah for their unconditional love, support, and encouragement. It is indeed the unsung support of family and friends that motivates you and keeps you going, especially in difficult times. Mohak considers himself exceptionally fortunate to have friends like Sushil Keswani and Ruma Paruthi in his life. He is also grateful to Rajeet Nair, Sumit Bakshi, Arvind Solanki, and Shweta (Dhamani) Keswani for their understanding, support, and trust.

Finally, we heartily apologize to friends and colleagues whose names may have been inadvertently missed in our acknowledgments.

Nathalie Japkowicz and Mohak Shah
Ottawa and Montreal
2010

Acronyms

2D	two-dimensional
3D	three-dimensional
ALL	acute lymphoblastic leukemia
AML	acute myloid leukemia
ANOVA	analysis of variance
ARI	adjusted Rand index
AUC	area under the (ROC) curve
Bin	Binomial (distribution)
BIR	Bayesian information reward
CD	critical difference
CDF	cumulative distribution function
CTBT	Comprehensive Nuclear Test Ban Treaty
CV	cross-validation
DEA	data envelopment analysis
DET	Detection Error Trade-Off
ERM	empirical risk minimization
exp	exponential
HSD	honestly significant difference
IBSR	Internet Brain Segmentation Repository
iff	if and only if
i.i.d.	independently and identically distributed
Inf	infimum
KDD	Knowledge Discovery in Databases (Archive)
KL	Kullback–Leibler
KS	Kolmogorov–Smirnov
LOO	leave-one-out
MAP	maximum a posteriori
MDS	multidimensional scaling
MRI	Magnetic Resonance Imaging
NEC	normalized expected cost
NHST	null hypothesis statistical testing
NPV	negative predictive value
PAC	probably approximately correct
PPV	positive predictive value
PR	precision-recall
RMSE	root-mean-square error
ROC	receiver operating characteristic (curve)
ROCCH	ROC convex hull
ROCR	ROC in R package
SAR	metric combining squared error (S), accuracy (A), and ROC area (R)
SAUC	scored AUC
SCM	set covering machine

SIM	simple and intuitive measure	SVM	support vector machine
		UCI	University of California, Irvine
SRM	structural risk minimization	VC	Vapnik–Chervonenkis
SS	sums of squares	w.r.t.	with regard to

Algorithms

1NN	1-nearest-neighbor	NN	nearest neighbor
ADA	AdaBoost using decision trees	RF	random forest
		RIP	Ripper
C45	decision tree (c4.5)	SCM	set covering machine
NB	naive Bayes	SVM	support vector machine

Algorithms are set in small caps to distinguish them from acronyms.

Acronyms used in tables and math

CI	confidence interval	LR	likelihood ratio
FN	false negative	Pr	probability
FP	false positive	TN	true negative
FPR	false-positive rate	TP	true positive
IR	information reward	TPR	true-positive rate

These are not acronyms, although sometimes TPR and FPR will appear as such. Authors' preferences were followed in this case.

Contents

1

Introduction

Technological advances in recent decades have made it possible to automate many tasks that previously required significant amounts of manual time, performing regular or repetitive activities. Certainly, computing machines have proven to be a great asset in improving human speed and efficiency as well as in reducing errors in these essentially mechanical tasks. More impressive, however, is the fact that the emergence of computing technologies has also enabled the automation of tasks that require significant understanding of intrinsically human domains that can in no way be qualified as merely mechanical. Although we humans have maintained an edge in performing some of these tasks, e.g., recognizing pictures or delineating boundaries in a given picture, we have been less successful at others, e.g., fraud or computer network attack detection, owing to the sheer volume of data involved and to the presence of nonlinear patterns to be discerned and analyzed simultaneously within these data. Machine learning and data mining, on the other hand, have heralded significant advances, both theoretical and applied, in this direction, thus getting us one step closer to realizing such goals.

Machine learning is embodied by different learning approaches, which are themselves implemented within various frameworks. Examples of some of the most prominent of these learning paradigms include supervised learning, in which the data labels are available and generally discrete; unsupervised learning, in which the data labels are unavailable; semisupervised learning, in which some, generally discrete, data labels are available, but not all; regression, in which the data labels are continuous; and reinforcement learning, in which learning is based on an agent policy optimization in a reward setting. The plethora of solutions that have been proposed within these different paradigms yielded a wide array of learning algorithms. As a result, the field is at an interesting crossroad. On the one hand, it has matured to the point where many impressive and pragmatic data analysis methods have emerged, of course, with their respective strengths and limitations.[1] On the other hand, it is now overflowing with hundreds of studies trying to improve the basic methods, but only marginally succeeding in doing so (Hand, 2006).[2] This is especially true on the applied front. Just as in any scientific field, the practical utility of any new advance can be accepted only if we can demonstrate beyond reasonable doubt the superiority of the proposed or novel methods over existing ones in the context in which it was designed.

This brings the issue of evaluating the proposed learning algorithms to the fore. Although considerable effort has been made by researchers in both developing novel learning methods and improving the existing models and approaches, these same researchers have not been completely successful at alleviating the users' scepticism with regard to the worth of these new developments. This is due, in big part, to the lack of both depth and focus in what has become a ritualized evaluation method used to compare different approaches. There are many issues involved in the question of designing an evaluation strategy for a learning machine. Furthermore, these issues cover a wide range of concerns pertaining to both the problem and the solution that we wish to consider. For instance, one may ask the following questions: What precise measure is best suited for a quantified assessments of different algorithms' property of interest in a given domain? How can these measures be efficiently computed? Do the data from the domain of interest affect the efficiency of this calculation? How can we be confident about whether the difference in measurement for two or more algorithms denotes a statistically significant difference in their performance? Is this statistical difference practically relevant as well? How can we best use the available data to discover whether such differences exist? And so on. We do not claim that all these issues can be answered in a definitive manner, but we do emphasize the need to understand the issues we are dealing with, along with the various approaches available to tackle them. In particular, we must understand the strengths and limitations of these approaches as well as the proper manner in which they should be applied. Moreover, we also need to understand what these methods offer and how to properly interpret the results of their application. This is very different from the way evaluation has been perceived to date in the machine learning community, where we have been using a routine, de facto, strategy, without much concern about its meaning.

In this book, we try to address these issues, more specifically with regard to the branch of machine learning pertaining to classification algorithms. In particular, we focus on evaluating the performance of classifiers generated by supervised learning algorithms, generally in a binary classification scenario. We wish to emphasize, however, that the overall message of the book and the

[1] These developments have resulted both from empirically studied behaviors and from exploiting the theoretical frameworks developed in other fields, especially mathematics.

[2] Although the worth of a study that results in marginal empirical improvements sometimes lies in the more significant theoretical insights obtained.

insights obtained should be considered in a more general sense toward the study of all learning paradigms and settings. Many of these approaches can indeed be readily exported (with a few suitable modifications) to other scenarios such as unsupervised learning, regression and so on. The issues we consider in the book deal not only with evaluation measures, but also with the related and important issues of obtaining (and understanding) the statistical significance of the observed differences, efficiently computing the evaluation measures in as unbiased a manner as possible, and dealing with the artifacts of the data that affect these quantities. Our aim is to raise an awareness of the proper way to conduct such evaluations and of how important they are to the practical utilization of the advances being made in the field. While developing an understanding of the relevant evaluation strategies, some that are widely used (although sometimes with little understanding) as well as some that are not currently too popular, we also try to address a number of practical criticisms and philosophical concerns that have been raised with regard to their usage and effectiveness and examine the solutions that have been proposed to deal with these concerns.

Our aim is not to suggest a recipe for evaluation to replace the previous de facto one, but to develop an understanding and appreciation of the evaluation strategies, of their strengths, and the underlying caveats. Before we go further and expand our discussion pertaining to the goals of this book by bringing forth the issues with our current practices, we discuss the de facto culture that has pervaded the machine learning community to date.

1.1 The De Facto Culture

For over two decades now, with Kibler and Langley (1988) suggesting the need for a greater emphasis on performance evaluation, the machine learning community has recognized the importance of proper evaluation. Research has been done to both come up with novel ways of evaluating classifiers and to use insights obtained from other fields in doing so. In particular, researchers have probed such fields as mathematics, psychology, and statistics among others. This has resulted in significant advances in our ability to track and compare the performance of different algorithms, although the results and the importance of such evaluation has remained underappreciated by the community as a whole because of one or more reasons that we will soon ponder. More important, however, is the effect of this underappreciation that has resulted in the entrenchment of a *de facto* culture of evaluation. Consider, for example, the following statement extracted from (Witten and Frank, 2005b, p. 144), one of the most widely used textbooks in machine learning and data mining:

> The question of predicting performance based on limited data is an interesting, and still controversial one. We will encounter many different