LNAI 3944

Joaquin Quiñonero-Candela   Ido Dagan
Bernardo Magnini   Florence d'Alché-Buc (Eds.)

Machine Learning Challenges

Evaluating Predictive Uncertainty
Visual Object Classification,
and Recognizing Textual Entailment

First PASCAL Machine Learning Challenges Workshop, MLCW 2005
Southampton, UK, April 2005
Revised Selected Papers

Springer

Joaquin Quiñonero-Candela   Ido Dagan
Bernardo Magnini   Florence d'Alché-Buc (Eds.)

Machine Learning Challenges

Evaluating Predictive Uncertainty Visual Object Classification and Recognizing Textual Entailment

First PASCAL Machine Learning Challenges Workshop, MLCW 2005
Southampton, UK, April 11-13, 2005
Revised Selected Papers

Springer

Volume Editors

Joaquin Quiñonero-Candela
Max Planck Institute for Biological Cybernetics
Spemannstr. 38, 72076 Tübingen, Germany
E-mail: joaquin@first.fraunhofer.de

Ido Dagan
Bar Ilan University
Ramat Gan 52900, Israel
E-mail: dagan@macs.biu.ac.il

Bernardo Magnini
ITC-irst, Centro per la Ricerca Scientifica e Tecnologica
Via Sommarive 14, 38050 Povo (Trento), Italy
E-mail: magnini@itc.it

Florence d'Alché-Buc
Université d'Evry-Val d'Essonne
IBISC CNRS FRE 2873 and GENOPOLE
523, Place des terrasses, 91000 Evry, France
E-mail: florence.dalche@ibisc.univ-evry.fr

Library of Congress Control Number: 2006924677

CR Subject Classification (1998): I.2.6-8, I.2.3, I.4-7, F.1, F.2, F.4

LNCS Sublibrary: SL 7 – Artificial Intelligence

ISSN 0302-9743
ISBN-10 3-540-33427-0 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-33427-9 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media

springer.com

Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11736790 06/3142 5 4 3 2 1 0

Lecture Notes in Artificial Intelligence 3944

Edited by J. G. Carbonell and J. Siekmann

Subseries of Lecture Notes in Computer Science

Lecture Notes in Artificial Intelligence (LNAI)

Vol. 3946: T.R. Roth-Berghofer, S. Schulz, D.B. Leake (Eds.), Modeling and Retrieval of Context. XI, 149 pages. 2006.

Vol. 3944: J. Quiñonero-Candela, I. Dagan, B. Magnini, F. d'Alché-Buc (Eds.), Machine Learning Challenges. XIII, 462 pages. 2006.

Vol. 3930: D.S. Yeung, Z.-Q. Liu, X.-Z. Wang, H. Yan (Eds.), Advances in Machine Learning and Cybernetics. XXI, 1110 pages. 2006.

Vol. 3918: W.K. Ng, M. Kitsuregawa, J. Li, K. Chang (Eds.), Advances in Knowledge Discovery and Data Mining. XXIV, 879 pages. 2006.

Vol. 3910: S.A. Brueckner, G.D.M. Serugendo, D. Hales, F. Zambonelli (Eds.), Engineering Self-Organising Systems. XII, 245 pages. 2006.

Vol. 3904: M. Baldoni, U. Endriss, A. Omicini, P. Torroni (Eds.), Declarative Agent Languages and Technologies III. XII, 245 pages. 2006.

Vol. 3900: F. Toni, P. Torroni (Eds.), Computational Logic in Multi-Agent Systems. XVII, 427 pages. 2006.

Vol. 3899: S. Frintrop, VOCUS: A Visual Attention System for Object Detection and Goal-Directed Search. XIV, 216 pages. 2006.

Vol. 3898: K. Tuyls, P.J. 't Hoen, K. Verbeeck, S. Sen (Eds.), Learning and Adaption in Multi-Agent Systems. X, 217 pages. 2006.

Vol. 3891: J.S. Sichman, L. Antunes (Eds.), Multi-Agent-Based Simulation VI. X, 191 pages. 2006.

Vol. 3890: S.G. Thompson, R. Ghanea-Hercock (Eds.), Defence Applications of Multi-Agent Systems. XII, 141 pages. 2006.

Vol. 3885: V. Torra, Y. Narukawa, A. Valls, J. Domingo-Ferrer (Eds.), Modeling Decisions for Artificial Intelligence. XII, 374 pages. 2006.

Vol. 3881: S. Gibet, N. Courty, J.-F. Kamp (Eds.), Gesture in Human-Computer Interaction and Simulation. XIII, 344 pages. 2006.

Vol. 3874: R. Missaoui, J. Schmidt (Eds.), Formal Concept Analysis. X, 309 pages. 2006.

Vol. 3873: L. Maicher, J. Park (Eds.), Charting the Topic Maps Research and Applications Landscape. VIII, 281 pages. 2006.

Vol. 3863: M. Kohlhase (Ed.), Mathematical Knowledge Management. XI, 405 pages. 2006.

Vol. 3862: R.H. Bordini, M. Dastani, J. Dix, A.E.F. Seghrouchni (Eds.), Programming Multi-Agent Systems. XIV, 267 pages. 2006.

Vol. 3849: I. Bloch, A. Petrosino, A.G.B. Tettamanzi (Eds.), Fuzzy Logic and Applications. XIV, 438 pages. 2006.

Vol. 3848: J.-F. Boulicaut, L. De Raedt, H. Mannila (Eds.), Constraint-Based Mining and Inductive Databases. X, 401 pages. 2006.

Vol. 3847: K.P. Jantke, A. Lunzer, N. Spyratos, Y. Tanaka (Eds.), Federation over the Web. X, 215 pages. 2006.

Vol. 3835: G. Sutcliffe, A. Voronkov (Eds.), Logic for Programming, Artificial Intelligence, and Reasoning. XIV, 744 pages. 2005.

Vol. 3830: D. Weyns, H. V.D. Parunak, F. Michel (Eds.), Environments for Multi-Agent Systems II. VIII, 291 pages. 2006.

Vol. 3817: M. Faundez-Zanuy, L. Janer, A. Esposito, A. Satue-Villar, J. Roure, V. Espinosa-Duro (Eds.), Nonlinear Analyses and Algorithms for Speech Processing. XII, 380 pages. 2006.

Vol. 3814: M. Maybury, O. Stock, W. Wahlster (Eds.), Intelligent Technologies for Interactive Entertainment. XV, 342 pages. 2005.

Vol. 3809: S. Zhang, R. Jarvis (Eds.), AI 2005: Advances in Artificial Intelligence. XXVII, 1344 pages. 2005.

Vol. 3808: C. Bento, A. Cardoso, G. Dias (Eds.), Progress in Artificial Intelligence. XVIII, 704 pages. 2005.

Vol. 3802: Y. Hao, J. Liu, Y.-P. Wang, Y.-m. Cheung, H. Yin, L. Jiao, J. Ma, Y.-C. Jiao (Eds.), Computational Intelligence and Security, Part II. XLII, 1166 pages. 2005.

Vol. 3801: Y. Hao, J. Liu, Y.-P. Wang, Y.-m. Cheung, H. Yin, L. Jiao, J. Ma, Y.-C. Jiao (Eds.), Computational Intelligence and Security, Part I. XLI, 1122 pages. 2005.

Vol. 3789: A. Gelbukh, Á. de Albornoz, H. Terashima-Marín (Eds.), MICAI 2005: Advances in Artificial Intelligence. XXVI, 1198 pages. 2005.

Vol. 3782: K.-D. Althoff, A. Dengel, R. Bergmann, M. Nick, T.R. Roth-Berghofer (Eds.), Professional Knowledge Management. XXIII, 739 pages. 2005.

Vol. 3763: H. Hong, D. Wang (Eds.), Automated Deduction in Geometry. X, 213 pages. 2006.

Vol. 3755: G.J. Williams, S.J. Simoff (Eds.), Data Mining. XI, 331 pages. 2006.

Vol. 3735: A. Hoffmann, H. Motoda, T. Scheffer (Eds.), Discovery Science. XVI, 400 pages. 2005.

Vol. 3734: S. Jain, H.U. Simon, E. Tomita (Eds.), Algorithmic Learning Theory. XII, 490 pages. 2005.

Vol. 3721: A.M. Jorge, L. Torgo, P.B. Brazdil, R. Camacho, J. Gama (Eds.), Knowledge Discovery in Databases: PKDD 2005. XXIII, 719 pages. 2005.

Vol. 3720: J. Gama, R. Camacho, P.B. Brazdil, A.M. Jorge, L. Torgo (Eds.), Machine Learning: ECML 2005. XXIII, 769 pages. 2005.

Vol. 3717: B. Gramlich (Ed.), Frontiers of Combining Systems. X, 321 pages. 2005.

Vol. 3702: B. Beckert (Ed.), Automated Reasoning with Analytic Tableaux and Related Methods. XIII, 343 pages. 2005.

Vol. 3698: U. Furbach (Ed.), KI 2005: Advances in Artificial Intelligence. XIII, 409 pages. 2005.

Vol. 3690: M. Pěchouček, P. Petta, L.Z. Varga (Eds.), Multi-Agent Systems and Applications IV. XVII, 667 pages. 2005.

Vol. 3684: R. Khosla, R.J. Howlett, L.C. Jain (Eds.), Knowledge-Based Intelligent Information and Engineering Systems, Part IV. LXXIX, 933 pages. 2005.

Vol. 3683: R. Khosla, R.J. Howlett, L.C. Jain (Eds.), Knowledge-Based Intelligent Information and Engineering Systems, Part III. LXXX, 1397 pages. 2005.

Vol. 3682: R. Khosla, R.J. Howlett, L.C. Jain (Eds.), Knowledge-Based Intelligent Information and Engineering Systems, Part II. LXXIX, 1371 pages. 2005.

Vol. 3681: R. Khosla, R.J. Howlett, L.C. Jain (Eds.), Knowledge-Based Intelligent Information and Engineering Systems, Part I. LXXX, 1319 pages. 2005.

Vol. 3673: S. Bandini, S. Manzoni (Eds.), AI*IA 2005: Advances in Artificial Intelligence. XIV, 614 pages. 2005.

Vol. 3662: C. Baral, G. Greco, N. Leone, G. Terracina (Eds.), Logic Programming and Nonmonotonic Reasoning. XIII, 454 pages. 2005.

Vol. 3661: T. Panayiotopoulos, J. Gratch, R. Aylett, D. Ballin, P. Olivier, T. Rist (Eds.), Intelligent Virtual Agents. XIII, 506 pages. 2005.

Vol. 3658: V. Matoušek, P. Mautner, T. Pavelka (Eds.), Text, Speech and Dialogue. XV, 460 pages. 2005.

Vol. 3651: R. Dale, K.-F. Wong, J. Su, O.Y. Kwong (Eds.), Natural Language Processing – IJCNLP 2005. XXI, 1031 pages. 2005.

Vol. 3642: D. Ślęzak, J. Yao, J.F. Peters, W. Ziarko, X. Hu (Eds.), Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, Part II. XXIII, 738 pages. 2005.

Vol. 3641: D. Ślęzak, G. Wang, M. Szczuka, I. Düntsch, Y. Yao (Eds.), Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, Part I. XXIV, 742 pages. 2005.

Vol. 3635: J.R. Winkler, M. Niranjan, N.D. Lawrence (Eds.), Deterministic and Statistical Methods in Machine Learning. VIII, 341 pages. 2005.

Vol. 3632: R. Nieuwenhuis (Ed.), Automated Deduction – CADE-20. XIII, 459 pages. 2005.

Vol. 3630: M.S. Capcarrère, A.A. Freitas, P.J. Bentley, C.G. Johnson, J. Timmis (Eds.), Advances in Artificial Life. XIX, 949 pages. 2005.

Vol. 3626: B. Ganter, G. Stumme, R. Wille (Eds.), Formal Concept Analysis. X, 349 pages. 2005.

Vol. 3625: S. Kramer, B. Pfahringer (Eds.), Inductive Logic Programming. XIII, 427 pages. 2005.

Vol. 3620: H. Muñoz-Ávila, F. Ricci (Eds.), Case-Based Reasoning Research and Development. XV, 654 pages. 2005.

Vol. 3614: L. Wang, Y. Jin (Eds.), Fuzzy Systems and Knowledge Discovery, Part II. XLI, 1314 pages. 2005.

Vol. 3613: L. Wang, Y. Jin (Eds.), Fuzzy Systems and Knowledge Discovery, Part I. XLI, 1334 pages. 2005.

Vol. 3607: J.-D. Zucker, L. Saitta (Eds.), Abstraction, Reformulation and Approximation. XII, 376 pages. 2005.

Vol. 3601: G. Moro, S. Bergamaschi, K. Aberer (Eds.), Agents and Peer-to-Peer Computing. XII, 245 pages. 2005.

Vol. 3600: F. Wiedijk (Ed.), The Seventeen Provers of the World. XVI, 159 pages. 2006.

Vol. 3596: F. Dau, M.-L. Mugnier, G. Stumme (Eds.), Conceptual Structures: Common Semantics for Sharing Knowledge. XI, 467 pages. 2005.

Vol. 3593: V. Mařík, R. W. Brennan, M. Pěchouček (Eds.), Holonic and Multi-Agent Systems for Manufacturing. XI, 269 pages. 2005.

Vol. 3587: P. Perner, A. Imiya (Eds.), Machine Learning and Data Mining in Pattern Recognition. XVII, 695 pages. 2005.

Vol. 3584: X. Li, S. Wang, Z.Y. Dong (Eds.), Advanced Data Mining and Applications. XIX, 835 pages. 2005.

Vol. 3581: S. Miksch, J. Hunter, E.T. Keravnou (Eds.), Artificial Intelligence in Medicine. XVII, 547 pages. 2005.

Vol. 3577: R. Falcone, S. Barber, J. Sabater-Mir, M.P. Singh (Eds.), Trusting Agents for Trusting Electronic Societies. VIII, 235 pages. 2005.

Vol. 3575: S. Wermter, G. Palm, M. Elshaw (Eds.), Biomimetic Neural Learning for Intelligent Robots. IX, 383 pages. 2005.

Vol. 3571: L. Godo (Ed.), Symbolic and Quantitative Approaches to Reasoning with Uncertainty. XVI, 1028 pages. 2005.

Vol. 3559: P. Auer, R. Meir (Eds.), Learning Theory. XI, 692 pages. 2005.

Vol. 3558: V. Torra, Y. Narukawa, S. Miyamoto (Eds.), Modeling Decisions for Artificial Intelligence. XII, 470 pages. 2005.

Vol. 3554: A.K. Dey, B. Kokinov, D.B. Leake, R. Turner (Eds.), Modeling and Using Context. XIV, 572 pages. 2005.

Vol. 3550: T. Eymann, F. Klügl, W. Lamersdorf, M. Klusch, M.N. Huhns (Eds.), Multiagent System Technologies. XI, 246 pages. 2005.

Vol. 3539: K. Morik, J.-F. Boulicaut, A. Siebes (Eds.), Local Pattern Detection. XI, 233 pages. 2005.

Vol. 3538: L. Ardissono, P. Brna, A. Mitrović (Eds.), User Modeling 2005. XVI, 533 pages. 2005.

Vol. 3533: M. Ali, F. Esposito (Eds.), Innovations in Applied Artificial Intelligence. XX, 858 pages. 2005.

Vol. 3528: P.S. Szczepaniak, J. Kacprzyk, A. Niewiadomski (Eds.), Advances in Web Intelligence. XVII, 513 pages. 2005.

Vol. 3518: T.-B. Ho, D. Cheung, H. Liu (Eds.), Advances in Knowledge Discovery and Data Mining. XXI, 864 pages. 2005.

Vol. 3508: P. Bresciani, P. Giorgini, B. Henderson-Sellers, G. Low, M. Winikoff (Eds.), Agent-Oriented Information Systems II. X, 227 pages. 2005.

Preface

The first **PASCAL** Machine Learning Challenges Workshop (MLCW 2005) (see, `www.pascal-network.org/Workshops/PC04/`) was held in Southampton, UK, during April 11-13, 2005. This conference was organized by the Challenges programme of the European Network of Excellence PASCAL (Pattern Analysis, Statistical modelling and ComputationAl Learning) in the framework of the IST Programme of the European Community. First annually and now quarterly, the PASCAL Challenges Programme plays the role of selecting and sponsoring challenging tasks, either practical or theoretical. The aim is to raise difficult machine learning questions and to motivate innovative research and development of new approaches. Financial support covers all the work concerning the cleaning and labelling of the data as well as the preparation of evaluation tools for ranking the results. For the first round of the programme, four challenges were selected according to their impact in the machine learning community, supported from summer 2004 to early spring 2005 by PASCAL and finally invited to participate in MLCW 2005 :

- The first challenge, called "Evaluating Predictive Uncertainty", dealt with the fundamental question of assigning a degree of confidence to the outputs of a classifier or a regressor.
- The goal of the second challenge, called "Visual Object Classes", was to recognise objects from a number of visual objects classes in realistic scenes.
- The third challenge task, called "Recognizing Textual Entailment", consisted in recognizing, given two texts fragments, whether the meaning of one text can be inferred (entailed) from the other.
- The fourth challenge was concerned with the assessment of "Machine Learning Methodologies to Extract Implicit Relations from Documents".

Each of these challenges raised noticeable attention in the research community, attracting numerous participants. The idea behind having a unique workshop was to make participants in different challenges exchange and benefit from the research experienced in other challenges. For the workshop, the session chairs made a first selection among submissions leading to 34 oral contributions. This book is concerned with selected proceedings of the first three challenges, providing a large panel of machine learning issues and solutions. A second round of selection was made to extract the 25 contributed chapters that make up this book, resulting in a selection rate of one half for the three considered challenges whose description follows.

Evaluating Predictive Uncertainty Challenge
When making decisions based on predictions, it is essential to have a measure of the uncertainty associated to them, or predictive uncertainty. Decisions are of course most often based on a loss function that is to be minimized in expectation. One common approach in machine learning is to assume knowledge of the loss

function, and then train an algorithm that outputs decisions that directly minimize the expected loss. In a realistic setting, however, the loss function might be unknown, or depend on additional factors only determined at a later stage. A system that predicts the presence of calcification from a mammography should also provide information about its uncertainty. Whether to operate or not will depend on the particular patient, as well as on the context in general. If the loss function is unknown, expressing uncertainties becomes crucial. Failing to do so implies throwing information away.

There does not seem to be a universal way of producing good estimates of predictive uncertainty in the machine learning community, nor a consensus on the ways of evaluating them. In part this is caused by deep fundamental differences in methodology (classical statistics, Bayesian inference, statistical learning theory). We decided to organize the Evaluating Predictive Uncertainty Challenge (`http://predict.kyb.tuebingen.mpg.de/`) to allow the different philosophies to compete directly on the empirical battleground. This required us to define losses for probabilistic predictions. Twenty groups of participants competed on two classification and three regression datasets before the submission deadline of December 11, 2004, and a few more after the deadline. We present six contributed chapters to this volume, by all the winners plus authors of other outstanding entries.

Visual Objects Classes
The PASCAL Visual Object Classes Challenge ran from February to March 2005 (`http://www.pascal-network.org/challenges/VOC/`). The goal of the challenge was to recognize objects from a number of visual object classes in realistic scenes (i.e., not pre-segmented objects). Although there already exist benchmarks such as the so-called 'Caltech 5' (faces, airplanes, motorbikes, cars rear, spotted cats) and UIUC car side images, largely used by the community of image recognition, it appears now that the developed methods are achieving such good performance that they have effectively saturated on these datasets, and thus the datasets are failing to challenge the next generation of algorithms. Such saturation can arise because the images used do not explore the full range of variability of the imaged visual class. Some dimensions of variability include: clean vs. cluttered background; stereotypical views vs. multiple views (e.g., side views of cars vs. cars from all angles); degree of scale change, amount of occlusion; the presence of multiple objects (of one or multiple classes) in the images.

Given this problem of saturation of performance, the Visual Object Classes Challenge was designed to be more demanding by enhancing some of the dimensions of variability listed above compared to the databases that had been available previously, so as to explore the failure modes of different algorithms. Four object classes were selected: motorbikes, bicycles, cars and people. Twelve teams entered the challenge. This book includes a contributed review chapter about the methods and the results achieved by the participants.

Recognizing Textual Entailment
Semantic analysis of language has been addressed traditionally through interpretation into explicitly stipulated meaning representations. Such semantic interpretation turned out to be a very difficult problem, which led researchers to approximate semantic processing at shallow lexical and lexical-syntactic levels. Usually, such approaches were developed in application-specific settings, without having an encompassing application-independent framework for developing and evaluating generic semantic approaches.

The Recognizing Textual Entailment (RTE) challenge was an attempt to form such a generic framework for applied semantic inference in text understanding. The task takes as input a pair of text snippets, called *text* (T) and *hypothesis* (H), and requires determining whether the meaning of T (most likely) entails that of H or not. The view underlying the RTE task is that different natural language processing applications, including question answering, information extraction, (multi-document) summarization, and machine translation, have to address the language variability problem and recognize that a particular target meaning can be inferred from different text variants. The RTE task abstracts this primary inference need, suggesting that many applications would benefit from generic models for textual entailment.

It is worth emphasizing some relevant features of the task, which contributed to its success:

- RTE is interdisciplinary: the task has been addressed with both machine learning and resource-based NLP techniques. It also succeeded to bridge, as a common benchmark, over different application-oriented communities.
- RTE was a really challenging task: RTE-1, in several respects, was a simplification of the complete task (e.g., we did not consider temporal entailment), but it proved to be at the state of the art of text understanding.
- The challenge attracted 17 participatants and made a strong impact in the research community, followed by a related ACL 2005 workshop and a dozen more conference publications later in 2005, which used the publicly available RTE-1 dataset as a standard benchmark.

February 2006

Joaquin Quiñonero-Candela
Ido Dagan
Bernardo Magnini
Florence d'Alché-Buc
MLCW 2005

Organization

The first (PASCAL) Machine Learning Challenges Workshop (MLCW 2005) was organized by the Challenges programme of the Network of Excellence PASCAL in Southampton, UK, April 11-13, 2005.

PASCAL Challenges Programme Committee

Programme Chairs	Florence d'Alché–Buc (Université d'Evry) Steve Gunn (University of Southampton) Michèle Sebag (Université de Paris XI)
Programme committee	Samy Bengio (IDIAP-Martigny) Alex Clark (Royal Holloway, University of London) Walter Daelemans (University of Antwerp) Cyril Goutte (Xerox Research Centre Europe) Steve Gunn (University of Southampton) Klaus-Robert Mueller (Fraunhofer FIRST) John Shawe-Taylor (University of Southampton) Bill Triggs (INRIA) Chris Watkins (Royal Holloway, University of London)

Programme Committee of (PASCAL) MLCW 2005

Conference Chair	Florence d'Alché–Buc (Université d'Evry) Steve Gunn (University of Southampton)
Local Organization	Eileen Simon (University of Southampton)

Session Chairs

Evaluating Predictive Uncertainty	Joaquin Quiñonero-Candela (Max Planck Institute for Biological Cybernetics, Fraunhofer FIRST and TU Berlin)
Visual Object Classes	Christopher Williams (University of Edinburgh) Andrew Zisserman (University of Oxford)

Recognizing Textual Entailment	Ido Dagan (Bar-Ilan University) Oren Glickman (Bar-Ilan University) Bernardo Magnini (ITC-irst, Istituto per la Ricerca Scientifica e Tecnologica)
Evaluating Machine Learning for Information Extraction	Mary Elaine Califf (Illinois State University) Fabio Ciravegna (University of Sheffield) Dayne Freitag (Fair Isaac Corporation) Nicholas Kushmerick (University College Dublin) Neil Ireson (University of Sheffield) Alberto Lavelli (ITC-irst, Istituto per la Ricerca Scientifica e Tecnologica, Povo)

Sponsoring Institution

PASCAL (Pattern Analysis, Statistical modelling and ComputationAl Learning): European Network of Excellence, IST Programme of the European Community, grant number IST-2002-506778.

Acknowledgements

FAB would like to thank cheerfully Steve Gunn and Michèle Sebag for co-managing the Challenges Programme, Eileen Simon for the nice local organization of the workshop at Southampton and John Shawe-Taylor for his support and his coordination of the European Network of Excellence PASCAL.

Table of Contents

Evaluating Predictive Uncertainty Challenge

Joaquin Quiñonero-Candela[1,2,3], Carl Edward Rasmussen[1], Fabian Sinz[1], Olivier Bousquet[1,4], and Bernhard Schölkopf[1]

[1] Max Planck Institute for Biological Cybernetics, Spemannstr. 38, D-72076 Tübingen, Germany
{carl, fabee, bernhard.schoelkopf}@tuebingen.mpg.de
[2] Fraunhofer FIRST.IDA, Kekuléstr. 7, D-12489 Berlin, Germany
joaquin@first.fraunhofer.de
[3] TU Berlin, SWT, Franklinstr. 28/29, D-10587 Berlin, Germany
[4] Pertinence, 32, rue des Jeûneurs, F-75002 Paris, France
olivier.bousquet@pertinence.com

Abstract. This Chapter presents the PASCAL[1] Evaluating Predictive Uncertainty Challenge, introduces the contributed Chapters by the participants who obtained outstanding results, and provides a discussion with some lessons to be learnt. The Challenge was set up to evaluate the ability of Machine Learning algorithms to provide good "probabilistic predictions", rather than just the usual "point predictions" with no measure of uncertainty, in regression and classification problems. Participants had to compete on a number of regression and classification tasks, and were evaluated by both traditional losses that only take into account point predictions and losses we proposed that evaluate the quality of the probabilistic predictions.

1 Motivation

Information about the uncertainty of predictions, or *predictive uncertainty*, is essential in decision making. Aware of the traumatic cost of an operation, a surgeon will only decide to operate if there is enough evidence of cancer in the diagnostic. A prediction of the kind "there is 99% probability of cancer" is fundamentally different from "there is 55% probability of cancer", although both could be summarized by the much less informative statement: "there is cancer". An investment bank trying to decide whether to invest or not in a given fund might react differently at the prediction that the fund value will increase by "10%± 1%" than at the prediction that it will increase by "10%± 20%", but it will in any case find any of the two previous predictions way more useful than the point prediction "the expected value increase is 10%". Predictive uncertainties are also used in active learning to select the next training example which will bring most information. Given the enormous cost of experiments

[1] Pattern Analysis, Statistical Modelling and Computational Learning (PASCAL) Network of Excellence, part of the IST Programme of the European Community, IST-2002-506778.

J. Quiñonero-Candela et al. (Eds.): MLCW 2005, LNAI 3944, pp. 1–27, 2006.

with protein binding chips, a drug making company will not bother making experiments whose outcome can be predicted with very low uncertainty.

Decisions are of course most often based on a loss function that is to be minimized in expectation. One common approach in Machine Learning is to assume knowledge of the loss function, and then train an algorithm that outputs decisions that directly minimize the expected loss. In a realistic setting however, the loss function might be unknown, or depend on additional factors only determined at a later stage. A system that predicts the presence of calcification from a mammography should also provide information about its uncertainty. Whether to operate or not will depend on the particular patient, as well as on the context in general. If the loss function is unknown, expressing uncertainties becomes crucial. Failing to do so implies throwing information away.

One particular approach to expressing uncertainty is to treat the unknown quantity of interest ("will it rain?") as a random variable, and make to predictions in the form of probability distributions, also known as *predictive distributions.* We will center our discussion around this specific representation of the uncertainty. But, how to produce reasonable predictive uncertainties? What is a reasonable predictive uncertainty in the first place?

Under the Bayesian paradigm, posterior distributions are obtained on the model parameters, that incorporate both the uncertainty caused by the noise, and by not knowing what the true model is. Integrating over this posterior allows to obtain the posterior distribution on the variables of interest; the predictive distribution arises naturally. Whether the resulting predictive distribution is meaningful depends of course on the necessary prior distribution, and one should be aware of the fact that inappropriate priors can give rise to arbitrarily bad predictive distributions. From a frequentist point of view, this will be the case if the prior is "wrong". From a Bayesian point of view, priors are neither wrong nor right, they express degrees of belief. Inappropriate priors that are too restrictive, in that they discard plausible hypotheses about the origin of the data, are sometimes still used for reasons of convenience, leading to unreasonable predictive uncertainties (Rasmussen and Quiñonero-Candela, 2005). If you believe your prior is reasonable, then the same should hold true for the predictive distribution. However, this distribution is only an updated belief — the extent to which it is in agreement with reality will depend on the extent to which the prior encompasses reality.

It is common in Machine Learning to not consider the full posterior distribution, but to rather concentrate on its mode, also called the Maximum a Posteriori (MAP) approach. The MAP approach being equivalent to maximum penalized likelihood, one could consider that any method based on minimizing a regularized risk functional falls under the MAP umbrella. The MAP approach produces predictions with no measure of the uncertainty associated to them, like "it will rain"; other methods for obtaining predictive uncertainties are then needed, such as Bagging for example (Breiman, 1996). More simplistic approaches would consist in always outputting the same predictive uncertainties, independently of the input, based on an estimate of the overall generalization error. This generalization error can in turn be estimated empirically by cross-validation, or theoretically by means Statistical Learning bounds on the generalization error. This simplistic approach should of course be regarded as a baseline, since any reasonable method that individually estimates predictive uncertainties depending on the input could in principle be superior.

It appears that there might not be an obvious way of producing good estimates of predictive uncertainty in the Machine Learning (or Statistical Learning) community. There is also an apparent lack of consensus on the ways of evaluating predictive uncertainties in the first place. Driven by the urgent feeling that it might be easier to validate the goodness of the different philosophies on the empirical battleground than on the theoretical, we decided to organize the Evaluating Predictive Uncertainty Challenge, with support from the European PASCAL Network of Excellence. The Challenge allowed different Machine Learning approaches to predictive uncertainty in regression and classification to be directly compared on identical datasets.

1.1 Organization of This Chapter

We begin by providing an overview and some facts about the Challenge in Sect. 2. We then move on to describing in detail the three main components of the Challenge: 1) in Sect. 3 we define what is meant by probabilistic predictions in regression and in classification, and explain the *format of the predictions* that was required for the Challenge, 2) in Sect. 4 we present the *loss functions* that we proposed for the Challenge, and 3) Section 5 details the five *datasets*, two for classification and three for regression, that we used for the Challenge. In Sect. 6 we present the results obtained by the participants, and in Sect. 7 we focus in more detail on the methods proposed by the six (groups of) participants who contributed a Chapter to this book. The methods presented in these six contributed chapters all achieved outstanding results, and all the dataset winners are represented. Finally, Sect. 8 offers a discussion of results, and some reflection on the many lessons learned from the Challenge.

2 An Overview of the Challenge

The Evaluating Predictive Uncertainty Challenge was organized around the following website: http://predict.kyb.tuebingen.mpg.de. The website remains open for reference, and submissions are still possible to allow researchers to evaluate their methods on some benchmark datasets.

The results of the Challenge were first presented at the NIPS 2004 Workshop on Calibration and Probabilistic Prediction in Machine Learning, organized by Greg Grudic and Rich Caruana, and held in Whistler, Canada, on Friday December 17, 2004. The Challenge was then presented in more depth, with contributed talks from some of the participants with best results at the PASCAL Challenges Workshop held in Southampton, UK, on April 11, 2005.