

# *Learning from Imbalanced Data Sets*

---

Papers from the AAI Workshop  
Technical Report WS-00-05



AAAI Press

American Association for Artificial Intelligence

# *Learning from Imbalanced Data Sets*

Papers from the AAAI Workshop

*Nathalie Japkowicz, Chair*

Technical Report WS-00-05

AAAI Press  
Menlo Park, California

Copyright © 2000, AAAI Press

The American Association for Artificial Intelligence  
445 Burgess Drive  
Menlo Park, California 94025 USA

AAAI maintains compilation copyright for this technical report and retains the right of first refusal to any publication arising from this AAAI event. Please do not make any inquiries or arrangements for hardcopy or electronic publication of all or part of the papers contained in these working notes without first exploring the options available through AAAI Press and AI Magazine. A signed release of this right by AAAI is required before publication by a third party.

ISBN 1-57735-120-7      WS-00-05

Manufactured in the United States of America

**AAAI Press**

445 Burgess Drive  
Menlo Park, California 94025

ISBN 1-57735-120-7      WS-00-05

ISBN 1-57735-120-7



9 781577 351207

# Organizing Committee

Robert C. Holte, *University of Ottawa*  
Nathalie Japkowicz (Chair), *Dalhousie University*  
Charles X. Ling, *University of Western Ontario*  
Stan Matwin, *University of Ottawa*

This AAAI Workshop was held July 31, 2000 in Austin, Texas

# Contents

- Machine Learning from Imbalanced Data Sets 101 / 1  
*Foster Provost*
- Open Mind Animals: Insuring the Quality of Data Openly Contributed over the World Wide Web / 4  
*David G. Stork and Chuck P. Lam*
- Learning from Imbalanced Data Sets: A Comparison of Various Strategies / 10  
*Nathalie Japkowicz*
- Correlates of State Failure / 16  
*Pamela Surko and Alan N. Unger*
- Measuring Performance when Positives are Rare / 18  
*S.H. Muggleton, C.H. Bryant, and A. Srinivasan*
- Feature Scaling in Support Vector Data Descriptions / 25  
*David M.J. Tax and Robert P.W. Duin*
- Using Autoencoding Networks for Tramp Metal Detection / 31  
*V. Bulitko, R. Greiner, R. Kube, and W. Zhou*
- A Recognition-Based Alternative to Discrimination-Based Multi-Layer Perceptrons / 32  
*Todd Eavis and Nathalie Japkowicz*
- Exploiting the Cost (In)sensitivity of Decision Tree Splitting Criteria / 39  
*Chris Drummond and Robert C. Holte*
- When Does Imbalanced Data Require more than Cost-Sensitive Learning? / 47  
*Dragos Margineantu*
- Learning from Imbalanced Data: Rank Metrics and Extra Tasks / 51  
*Rich Caruana*
- Handling Imbalanced Data Sets in Insurance Risk Modeling / 58  
*Edwin P. D. Pednault, Barry K. Rosen, and Chidanand Apte*
- Learning to Predict Extremely Rare Events / 64  
*Gary M. Weiss and Haym Hirsh*
- An Approach to Imbalanced Data Sets Based on Changing Rule Strength / 69  
*Jerzy W. Grzymala-Busse, Linda K. Goodwin, Witold J. Grzymala-Busse, and Xinqun Zheng*