



Universiteit  
Leiden  
The Netherlands

## Massively collaborative machine learning

Rijn, J.N. van

### Citation

Rijn, J. N. van. (2016, December 19). *Massively collaborative machine learning*. *IPA Dissertation Series*. Retrieved from <https://hdl.handle.net/1887/44814>

Version: Not Applicable (or Unknown)

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/44814>

**Note:** To cite this publication please use the final published version (if applicable).

Cover Page



Universiteit Leiden



The handle <http://hdl.handle.net/1887/44814> holds various files of this Leiden University dissertation

**Author:** Rijn, Jan van

**Title:** Massively collaborative machine learning

**Issue Date:** 2016-12-19

# **Massively Collaborative Machine Learning**

Jan N. van Rijn



Universiteit  
Leiden

The author of this PhD thesis was employed at Leiden University, and also used facilities of the University of Waikato.



The work in this thesis has been carried out under the auspices of the research school IPA (Institute for Programming research and Algorithmics).



The author was funded by the Netherlands Organization for Scientific Research (NWO) as part of the project 'Massively Collaborative Data Mining' (number 612.001.206).

Copyright 2016 by Jan N. van Rijn

Open-access: <https://openaccess.leidenuniv.nl>

Typeset using  $\text{\LaTeX}$ , diagrams generated using GGPLOT and GNUPLOT

Cover image by Olivier H. Beauchesne and SCImago Lab (used with permission)

Printed by Ridderprint B.V.

ISBN 978-94-6299-506-2

# Massively Collaborative Machine Learning

Proefschrift

ter verkrijging van  
de graad van Doctor aan de Universiteit Leiden,  
op gezag van Rector Magnificus prof. mr. C.J.J.M. Stolker,  
volgens besluit van het College voor Promoties  
te verdedigen op maandag 19 december 2016  
klokke 12:30 uur

door

Jan Nicolaas van Rijn  
geboren te Katwijk  
in 1987

## Promotiecommissie

Promotor: prof. dr. J. N. Kok  
Copromotores: dr. A. J. Knobbe  
dr. J. Vanschoren (Technische Universiteit Eindhoven)  
Promotiecommissie: prof. dr. T. H. W. Bäck (secretaris)  
prof. dr. E. Marchiori (Radboud Universiteit)  
prof. dr. B. Pfahringer (University of Waikato, Nieuw Zeeland)  
prof. dr. A. Plaat (voorzitter)

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Introduction . . . . .	1
1.2	Publications . . . . .	3
<b>2</b>	<b>Machine Learning</b>	<b>5</b>
2.1	Introduction . . . . .	5
2.2	Data . . . . .	6
2.2.1	Iris . . . . .	6
2.2.2	Mushroom . . . . .	8
2.3	Tasks . . . . .	10
2.4	Models . . . . .	11
2.4.1	Decision rules . . . . .	11
2.4.2	Decision trees . . . . .	12
2.4.3	Probabilistic reasoning . . . . .	14
2.4.4	Nearest Neighbour models . . . . .	15
2.4.5	Logistic Regression . . . . .	16
2.4.6	Support Vector Machines . . . . .	18
2.4.7	Neural Networks . . . . .	18
2.5	Evaluation . . . . .	21
2.6	Discussion . . . . .	23
<b>3</b>	<b>Meta-Learning</b>	<b>25</b>
3.1	Introduction . . . . .	25

---

3.2	Learning Approach . . . . .	27
3.2.1	Feature space . . . . .	28
3.2.2	Performance space . . . . .	29
3.3	Search Approach . . . . .	30
3.3.1	Combining Search and Learning . . . . .	32
3.4	Ensembles . . . . .	33
3.5	Conservation for Generalization Performance . . . . .	35
3.6	Model Characteristics . . . . .	36
3.7	Bias Variance Profile . . . . .	37
3.8	Discussion . . . . .	38
<b>4</b>	<b>Experiment Databases</b>	<b>41</b>
4.1	Introduction . . . . .	41
4.2	Networked science . . . . .	42
4.2.1	Designing networked science . . . . .	43
4.3	Machine learning . . . . .	44
4.3.1	Reusability and reproducibility . . . . .	45
4.3.2	Prior work . . . . .	45
4.4	OpenML . . . . .	46
4.4.1	Datasets . . . . .	46
4.4.2	Task types . . . . .	48
4.4.3	Tasks . . . . .	49
4.4.4	Flows . . . . .	49
4.4.5	Setups . . . . .	51
4.4.6	Runs . . . . .	51
4.4.7	Studies . . . . .	52
4.4.8	Plug-ins . . . . .	52
4.5	Learning from the past . . . . .	55
4.5.1	Model-level analysis . . . . .	56
4.5.2	Data-level analysis . . . . .	63
4.5.3	Method-level analysis . . . . .	68
4.6	Conclusions . . . . .	69
<b>5</b>	<b>Data Streams</b>	<b>71</b>
5.1	Introduction . . . . .	72
5.2	Related Work . . . . .	73
5.3	Methods . . . . .	76
5.3.1	Online Performance Estimation . . . . .	76
5.3.2	Ensemble Composition . . . . .	79
5.3.3	BLAST . . . . .	80



---

5.3.4	Meta-Feature Ensemble . . . . .	83
5.4	Experimental Setup . . . . .	84
5.4.1	Data Streams . . . . .	84
5.4.2	Parameter Settings . . . . .	87
5.4.3	Baselines . . . . .	87
5.5	Results . . . . .	88
5.5.1	Ensemble Performance . . . . .	88
5.5.2	Effect of Parameters . . . . .	93
5.6	Designed Serendipity . . . . .	98
5.7	Conclusions . . . . .	100
<b>6</b>	<b>Combining Accuracy and Run Time</b>	<b>103</b>
6.1	Introduction . . . . .	103
6.2	Related Work . . . . .	105
6.3	Methods . . . . .	107
6.3.1	Pairwise Curve Comparison . . . . .	107
6.3.2	Active Testing . . . . .	111
6.3.3	Combining Accuracy and Run Time . . . . .	113
6.4	Experiments . . . . .	114
6.4.1	Predicting the Best Classifier . . . . .	116
6.4.2	Ranking of Classifiers . . . . .	118
6.4.3	Loss Time Space . . . . .	121
6.4.4	Optimizing on Accuracy and Run Time . . . . .	124
6.5	Conclusion . . . . .	125
<b>7</b>	<b>Conclusions</b>	<b>129</b>
7.1	Open Machine Learning . . . . .	129
7.2	Massively Collaborative Machine Learning . . . . .	130
7.3	Community Adoption . . . . .	131
7.4	Future Work . . . . .	131
	<b>Bibliography</b>	<b>133</b>
	<b>Dutch Summary</b>	<b>147</b>
	<b>English Summary</b>	<b>149</b>
	<b>Curriculum Vitae</b>	<b>151</b>
	<b>Acknowledgements</b>	<b>153</b>

<b>Publication List</b>	<b>155</b>
<b>Titles in the IPA Dissertation Series since 2013</b>	<b>159</b>

## List of Figures

2.1	Scatter plot of the 'iris' dataset . . . . .	7
2.2	Decision rule model of the 'mushroom' dataset . . . . .	11
2.3	Decision tree model of the 'mushroom' dataset . . . . .	13
2.4	Logistic Regression model of the 'iris' dataset . . . . .	17
2.5	Support Vector Machine model built upon the 'iris' dataset . . . . .	19
2.6	Example of a neural network . . . . .	20
2.7	ROC curves of three classifiers on the 'German credit' dataset . . . . .	22
3.1	The Algorithm Selection Framework . . . . .	27
3.2	Example of Bayesian Optimization . . . . .	31
3.3	Bias and Variance . . . . .	38
4.1	Example of an OpenML task description . . . . .	49
4.2	Example of an OpenML flow . . . . .	50
4.3	Example of an OpenML run . . . . .	53
4.4	WEKA integration of OpenML . . . . .	54
4.5	Example of a RapidMiner workflow . . . . .	55
4.6	R integration of OpenML . . . . .	56
4.7	Various algorithms on the 'letter' dataset . . . . .	57
4.8	Effect of optimizing parameters . . . . .	59
4.9	Effect of gamma parameter for SVM's . . . . .	60
4.10	Ranking of algorithms over all datasets . . . . .	61
4.11	Results of Nemenyi test on classifiers in OpenML . . . . .	62

4.12	Optimal values of parameters . . . . .	64
4.13	The effect of feature selection . . . . .	66
4.14	Hierarchical clustering of stream classifiers . . . . .	69
5.1	Performance of four classifiers on intervals of the ‘electricity’ dataset . . . . .	73
5.2	Schematic view of Windowed Performance Estimation . . . . .	77
5.3	The effect of a prediction when using Fading Factors . . . . .	78
5.4	Online Performance Estimation . . . . .	79
5.5	Performance of 25 data stream classifiers based on 60 data streams. . . . .	81
5.6	Effect of the ensemble size parameter . . . . .	89
5.7	Performance of the various meta-learning techniques . . . . .	90
5.8	Accuracy per data stream . . . . .	92
5.9	Results of Nemenyi test . . . . .	93
5.10	Effect of the decay rate and window parameter . . . . .	94
5.11	Effect of the grace parameter on accuracy . . . . .	95
5.12	Performance for various values of $k$ . . . . .	97
5.13	Performance differences between Leveraging Bagging ensembles and single classifiers . . . . .	98
5.14	Performance differences between Online Bagging ensembles and single classifiers . . . . .	99
6.1	Learning curves on the ‘letter’ dataset . . . . .	108
6.2	Number of datasets with maximum number of learning curve samples . . . . .	114
6.3	Performance of meta-algorithm on predicting the best classifier . . . . .	117
6.4	Example Loss Curves . . . . .	119
6.5	Average Area Under the Loss Curves for various meta-algorithms . . . . .	120
6.6	Results of Nemenyi test on the Area Under the Loss Curve scores . . . . .	121
6.7	Example Loss Time Curves . . . . .	122
6.8	Average Area under the Loss Time Curve scores for the various meta-algorithms . . . . .	123
6.9	Results of Nemenyi test on the Area Under the Loss Time Curves scores . . . . .	123
6.10	Loss Time Curves . . . . .	126
6.11	Results of Nemenyi test on the Area Under the Loss Time Curve scores . . . . .	128