

Aspects of Online Learning

Edward Francis Harrington

6 December 2004

A thesis submitted for the degree of
Doctor of Philosophy at
The Australian National University



Research School of Information Sciences and Engineering

Declaration

Except where otherwise indicated, this thesis is my own original work. Some of the material in thesis has already been published elsewhere in collaboration with others, the details of which are:

1. Harrington E., Kivinen J. and Williamson R. C., Channel Equalization and the Bayes Point Machine, International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 4, pp. 493-496, 2003.
2. Harrington, E., Kivinen, J., Williamson, R. C., Herbrich, R., & Platt J. Online Bayes Point Machine. Proceedings of PAKDD-03, 7th Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 241-252, Heidelberg: Springer Verlag, 2003.
3. Harrington E., Expert mixture methods for adaptive channel equalization, Kaynak, O. et al. (Eds.): International Conference of Artificial Neural Networks and Neural Information Processing (ICANN/ICONIP), LNCS 2714, pp. 538-545, Springer Verlag, 2003.
4. Harrington E. F., Online Ranking/Collaborative Filtering using the Perceptron Algorithm, Proceedings of the Twentieth International Conference of Machine Learning (ICML-2003), pp. 250-257, Morgan Kaufmann.

Edward Harrington

6 December 2004

Computer Sciences Laboratory
Research School of Information Sciences and Engineering
The Australian National University

Acknowledgements

My warmest thanks must go to my wife Annie, who provided me with great support and love, especially whilst writing this thesis.

I have been fortunate to have had the supervision of Bob Williamson, the many “fire side” like chats with him have kept my on track to complete this thesis. His guidance in machine learning and signal processing have proven to be invaluable. I would like to thank a number of people in machine learning community for their help over last three or so years: Peter Bartlett, Ralf Herbrich, Jyrki Kivinen, Shahar Mendelson, John Platt, Gunnar Rätsch, and Alex Smola. I am very grateful to Jyrki Kivinen for his supervision and his many thoughtful comments concerning this thesis which have lead to its improvement. I am also grateful to Ralf Herbrich for his guidance and encouragement throughout the last few years. I would like to thank the examiners for their constructive suggestions, as they resulted in a much improved thesis.

I am very grateful to my employer, the Defence Science and Technology Organisation, for their financial support in allowing me to undertake this PhD. Several people at DSTO have provide me with assistance during my PhD: Ian Doherty, Geoff Latham, Nigel McGinty and Garry Newsam. I am especially appreciative of the encouragement and the support given to me by Ian Doherty, and the late Geoff Latham, when applying for study leave to undertake a PhD. I would also like to thank Garry Newsam for his support during the later stages of my PhD.

Thanks to the head of Computer Sciences Laboratory, John Lloyd, for funding some of my travel during 2003. To Michelle Moravec many thanks go to her for her help with all my administration needs whilst at the Computer Sciences Laboratory.

The friendship of my fellow PhD students has helped me during the last three, or so years: Evan Greensmith, Omri Guttman, Kee Siong Ng and Cheng Soon Ong. I also acknowledge the proof reading efforts of Evan Greensmith.

Lastly, I would like to thank my parents and family for their love and support.

Abstract

Online learning algorithms have several key advantages compared to their batch learning algorithm counterparts: they are generally more memory efficient, and computationally more efficient; they are simpler to implement; and they are able to adapt to changes where the learning model is time varying. Online algorithms because of their simplicity are very appealing to practitioners. This thesis investigates several online learning algorithms and their application. The thesis has an underlying theme of the idea of combining several simple algorithms to give better performance. In this thesis we investigate: combining weights, combining hypothesis, and (sort of) hierarchical combining.

Firstly, we propose a new online variant of the *Bayes point machine* (BPM), called the *online Bayes point machine* (OBPM). We study the theoretical and empirical performance of the OBPM algorithm. We show that the empirical performance of the OBPM algorithm is comparable with other large margin classifier methods such as the *approximately large margin algorithm* (ALMA) and methods which maximise the margin explicitly, like the *support vector machine* (SVM). The OBPM algorithm when used with a parallel architecture offers potential computational savings compared to ALMA. We compare the test error performance of the OBPM algorithm with other online algorithms: the Perceptron, the voted-Perceptron, and Bagging. We demonstrate that the combination of the voted-Perceptron algorithm and the OBPM algorithm, called voted-OBPM algorithm has better test error performance than the voted-Perceptron and Bagging algorithms. We investigate the use of various online voting methods against the problem of ranking, and the problem of collaborative filtering of instances. We look at the application of online Bagging and OBPM algorithms to the telecommunications problem of channel equalization. We show that both online methods were successful at reducing the effect on the test error of label flipping and additive noise.

Secondly, we introduce a new mixture of experts algorithm, the *fixed-share hierarchy* (FSH) algorithm. The FSH algorithm is able to track the mixture of experts when the switching rate between the best experts may not be constant. We study the theoretical aspects of the FSH and the practical application of it to adaptive equalization. Using simulations we show that the FSH algorithm is able to track the best expert, or mixture of experts, in both the case where the switching rate is constant and the case where the switching rate is time varying.

Contents

Declaration	iii
Acknowledgements	v
Abstract	vii
List of Figures	xiii
List of Tables	xv
1 Introduction	1
1.1 Online versus batch learning	1
1.2 Motivations	2
1.3 Notational convention	2
1.4 Learning setting	3
1.5 Measures of performance	4
1.6 The Perceptron and large margin variants	6
1.7 The Bayes point machine	9
1.7.1 The Bayes point	9
1.7.2 Estimating the Bayes point	10
1.8 Aggregate, ensemble and voting methods	11
1.8.1 Bagging and the voted-Perceptron	11
1.8.2 Boosting	12
1.9 Tracking experts	13
1.10 Thesis outline	14
2 Online Bayes Point Machine	15
2.1 Introduction	15
2.2 The Algorithm	15
2.2.1 OBPM kernel implementation	17
2.3 Analysis of OBPM	18
2.3.1 Correlation between Perceptrons	19
2.3.2 Mistake bounds	20
2.4 Experiment results	22
2.4.1 Synthetic data set	22
2.4.2 Learning with large data sets	23

2.4.3	The kernel trick and UCI data set	25
2.4.4	USPS hand digit recognition	27
2.4.5	Drifting concept	28
2.5	Summary	30
3	Online Voting Methods	31
3.1	Introduction	31
3.2	Algorithms for online voting	32
3.3	Empirical results	34
3.3.1	Online voting on large data sets	36
3.3.2	Kernel and real world data sets	37
3.4	Summary	38
4	The Perceptron, Ranking and Collaborative Filtering	39
4.1	Introduction	39
4.2	PRank in a nutshell	41
4.3	Improved generalisation (large margin) versions	42
4.3.1	Bayes point approach	43
4.3.2	Online voting methods	45
4.4	Experiments and results	46
4.4.1	Ranking	47
4.4.2	Collaborative filtering	48
4.5	Summary	51
5	Adaptive Equalizer Aggregation	53
5.1	Introduction	53
5.2	System model	54
5.3	Stochastic gradient methods	55
5.4	Aggregation methods for equalization	56
5.4.1	Regression	57
5.4.2	Classification	58
5.5	Online implementations	58
5.5.1	Buffered approach	59
5.5.2	Subsampling approach	59
5.6	Experiments	60
5.6.1	Subsampled experiments	61
5.6.2	Buffered experiments	63
5.7	Summary	64
6	Tracking the Best Equalizer	65
6.1	Introduction	65
6.2	More adaptive equalization methods	66
6.3	Preliminaries: mixture of experts	67

6.4	The fixed-share hierarchy	71
6.5	Fixed-share hierarchy bound	73
6.6	Example scenarios	74
6.6.1	Model tracking (multipath)	74
6.6.2	Determining model order	75
6.6.3	Tracking LMS step-sizes	76
6.6.4	ANG prior parameters selection	76
6.6.5	Switching from blind to decision-directed equalization	77
6.7	Experiments	78
6.7.1	Model tracking (multipath)	78
6.7.2	Determining model order	81
6.7.3	Tracking LMS step-sizes	84
6.7.4	ANG prior parameters selection	85
6.7.5	Switching from blind to decision-directed equalization	86
6.8	Summary	87
7	Conclusions and Further Work	89
7.1	Main contributions	89
7.2	Conclusions and further work	90
	Thesis proofs	93
A.1	Lower bound proof of (2.1)	93
A.2	Proof of Theorem 2.3.1	94
A.3	Proof of Theorem 2.3.2	95
	Statistics	99
B.1	Student's t-distribution and associated statistics	99
B.2	Cochran test	99
	Several Blind Equalization Methods	101
C.1	Constant Modulus Algorithm	101
C.2	Decision Directed Decision Feedback Equalizer	101
	Glossary of Symbols	109

List of Figures

1.1	Binary classification example in two dimensions illustrating the hyper-plane with γ being the maximum margin separating the two classes (squares representing instances in $+1$ class and circles representing instances in -1 class).	3
1.2	Example scenario where classification is better than regression.	5
1.3	Illustration of the Bayes point for linear classifiers $\mathbf{w} \in \mathbb{R}^3$ trained on five examples (defining the planes which cut the surface of a sphere). We are looking at the surface squashed onto a two dimensional space.	11
2.1	Synthetic data set averages from 30 Monte Carlo simulations training OBPM with no label noise.	22
2.2	(a) Prediction errors for OBPMs $\tilde{\mathbf{w}}_t$ ($\tau=0.3$, $N=100$) and ALMA's weight solutions and (b) total number updates of both ALMA and OBPM for the Web data after training t examples.	24
2.3	Prediction errors from the drifting concept experiment with the MNIST OCR data set comparing LMS (step size 0.5), ALMA ($B=2$, $C=\sqrt{2}$), Perceptron ($\rho = 0$) and OBPM ($\tau=0.1$, $N=100$).	29
3.1	Histograms showing the frequency of occurrence of particular numbers of correctly classified training examples (v) made by each stored weight $\mathbf{w}_i$ for $i = 1, \dots, m$. Each algorithm (a) voted Perceptron and (b) OBPM was trained for a single epoch and trial of the Adult data set. The total number stored Perceptron weights was $m = 6945$ for the voted Perceptron and for OBPM $m = 5469$	34
4.1	Illustrative example of an update at iteration t of the PRank algorithm.	43
4.2	Illustration as to why OAP-BPM (Online Aggregate PRank-Bayes Point Machine) gives improved generalisation.	45
4.3	Experimental results for synthetic data set comparing the rank learning algorithms PRank, WH with step size $\eta = 0.1$, OAP-BPM with Bernoulli probability $\tau = 0.3$, OAP-BPM ensemble variants OAP-Bagging (Bagging) and OAP-Majority (Majority vote).	48
4.4	Collaborative filtering experimental results, comparing the averaged rank loss for the PRank, WH, OAP-BPM, OAP-Bagging, and OAP-Majority algorithms.	52

5.1	Communications system discrete time model with channel equalization applied at the receiver.	55
5.2	Architecture of the buffer method.	58
5.3	Architecture of the subsampling method.	60
5.4	Subsampling experiment comparing the test error performance on two different channels with AWGN.	61
5.5	Test error performance for channel B: (a) NLMS with various step sizes versus RLS with forgetting factor $\beta = 0.999$ in the presence of AWGN of 10dB (no label flipping), and (b) NLMS, OBPM and RLS with 10 percent label flipping and 20dB AWGN.	62
5.6	Test error (a), and diversity (b) results for Channel A with AWGN, comparing Bagging(LMS), Bagging(LMA) Bagging(NLMS), and BPM algorithms.	63
5.7	Label noise results for (a) Channel A and (b) Channel B using BPM equalizer, where only the training examples had label flipping, test examples had no label flipping.	63
6.1	Plot of the error term of bound versus the switching rate.	71
6.2	Model of the proposed system for switching from CMA to DD-DFE equalization.	77
6.3	Results for FSH algorithm tracking the equalizer model which matches the given (time-dependent) multipath.	79
6.4	Various weights generated by FSH algorithm.	80
6.5	Results of tracking the best model order of LSALE.	82
6.6	Tracking the step-size/learning rate which produces the smallest cumulative loss (and MSE) during the convergence of LMS.	83
6.7	Tracking the prior parameters of the ANG algorithm which produce the smallest cumulative loss (and MSE).	85
6.8	Results of using the fixed-share hierarchy algorithm to switch from blind (CMA) to decision directed feedback equalization (DD-DFE).	87

List of Tables

2.1	Table of test error percentages produced by averaging over 30 Monte-Carlo simulations of a single epoch of the parameter selection experiment, with 95 percent confidence interval for Student's t -distribution.	23
2.2	Real world data set test errors in percent run for 3 epochs. Result in bold indicates significantly different.	25
2.3	UCI data set test errors in percent produced by 100 random trials, for up to 10 epochs. Results in bold indicate that they are the smallest and statistically significant using the 95 percent confidence interval of the Student's t -distribution.	26
2.4	USPS data set test errors in percent run for 10 epochs for a RBF kernel with $\sigma = 3$	28
3.1	Test error comparing online voting methods for large data sets with linear classifiers.	35
3.2	Test error comparing online voting methods using kernels.	36
4.1	Synthetic data set experimental results produced by test sample (not used in training), showing the averaged rank loss.	47
4.2	Collaborative filtering experimental results produced by test sample (not used in training), showing the averaged rank loss.	49
B.1	Cochran test table	100

