Efficient Boosted Ensemble-based Machine Learning in the Context of Cascaded Frameworks

A thesis presented in partial
fulfilment of the requirements

for the degree of

Doctor of Philosophy in Computer Science

at Massey University,
Auckland, New Zealand

Teo Sušnjak

2012

Abstract

The ability to both efficiently train robust classifiers and to design them for fast detection is an important goal of machine learning. With an ever increasing amount of available data being generated, the task of expeditiously producing real-time capable classifiers is becoming more challenging. In the context of the increasing complexity of the task, ensemble-based learning methods have proven themselves to be effective approaches for satisfying these requirements.

Ensemble methods produce a number of weak models that are strategically combined into a single classifier. They have been particularly effective when combined with boosting algorithms and strategies that structure the ensembles into cascades. The strength of cascaded-ensembles lies in the separate-and-conquer approach they employ during the training of each layer. Class decision-boundaries for trivial cases are learned in the early rounds, while more difficult decision boundaries are refined with each succeeding layer. In a two-class problem domain, non-target instances learned in initial layers are removed and replaced by more complex samples, frequently referred to as bootstrapping. With this procedure, efficient coarse-to-fine learning is accomplished.

The contribution of this thesis lies in three main areas that centre around the concept of improving the efficiency in the training and execution process. The first explored ways in which the conventional ensemble-cascades could be combined with an even more aggressive separate-and-conquer strategy that further partitions the ensemble inside each layer. The focus was on the two-class learning problem and used face detection as the medium to observe the trade-offs involved concerning both the accuracy and the efficiency of the resulting classifiers. The algorithm was further developed in a way that enabled the bootstrapping of positive samples within a cascade, alongside the conventional approach that bootstraps only the negative samples. Secondly, the negative effect of dynamic environments on static classifiers on binary class problems was considered. A method was developed which enabled the cascaded classifiers to efficiently adapt to the changing environment on domains with high volume streaming data. This environment was simulated using face detection as well. Lastly, the open problem of creating integrated multiclass cascades was researched and an algorithm was devised.

Overall, the findings have shown that invariably a trade-off is incurred between reduced training runtimes resulting from aggressive separate-and-conquer strategies and the accuracy of the final classifiers. Using the CMU MIT test dataset, the experiments showed that though the proposed positive sample bootstrapping component succeeded in significantly reducing the training runtimes without compromising the accuracy, the general decomposition strategy did lower the accuracy when compared to the benchmark Viola-Jones classifiers. The proposed adaptive cascade learning algorithm for drifting concepts was also evaluated on a face detection problem set. The results demonstrated its ability to effectively adapt to dynamic environments in high speed data streams without requiring explicit re-training of the individual classifiers. The multiclass cascaded algorithm was compared to three existing algorithms on 18 UCI datasets. It was found to be, on average, several times faster to train and to execute, while generating comparable accuracy rates. The algorithm exhibited scalability to large datasets but was found to be susceptible to producing overly complex classifiers on datasets with a large number of class labels.

Acknowledgements

My thanks to the Institute of Information and Mathematical Sciences at Massey University for enabling me to grow over the years in my academic pursuits and for fostering an environment which allowed this to take place. Thanks in particular to the faculty at the Department of Computer Science. You have inspired, encouraged and taught me not just knowledge, but the value of it. My gratitude goes to all the administration staff at the institute for their help and assistance over the years, as well as to all my colleagues from whom I have learned much.

I am grateful to the Tertiary Education Commission for providing me with a generous doctorate scholarship. I express my thanks also to the Ministry of Science and Innovation for their internship grant that enabled me to implement research and expertise arising from this research into the industrial setting. A special thanks to Compac Sorting Ltd. for their support in the final year and for providing me with an opportunity to apply my research to solving real-world problems.

I would like to extend my thanks to the members of the doctoral defence committee, Prof. Pitoyo Hartono, Prof. Alvis Fong and Assoc. Prof. Chris Scogings for their time as well as valuable input and engagement with me in a fruitful discussion. Also a special thanks to Prof. Anne de Bruin for her role as convenor of the examination.

Thanks to my friends and family for your support over the years and a special thanks to a few of you who have reviewed the initial manuscript. I look forward to now seeing and spending more time with all of you. Christoph Schumacher, without your encouragement I do not think I would have embarked on this path. Thank you for your friendship and continuous support.

Brian and Beverley for never doubting, but always believing. Expressing my gratitude to you both seems so inadequate. To my mum and dad who have unceasingly encouraged and urged me on from distant shores, thank you; ovo je i vaš uspjeh, hvala vam za sve.

Ken Hawick, my heartfelt gratitude to you for being my co-supervisor. You have been a source of wisdom, advice and encouragement during many phases of this research. To my supervisor Andre Barczak, I am so indebted to you. You have been my teacher, my mentor and have also become a close friend. I look forward to our conversations over coffee continuing for many years to come.

To my beautiful wife Sarah; I love you. You have been my rock every step of the way.

Teo Sušnjak
Auckland, New Zealand
October, 2012

Contents

List of Figures

List of Tables

List of Algorithms

List of Acronyms

BPSL	Bootstrapped PSL
BPSL.r	Bootstrapped PSL with re-sampling
CMU MIT	Carnegie Mellon University - Massachusetts Institute of Technology (Face detection dataset)
CTF	Coarse-To-Fine learning
DP	Domain Partitioning (Multiclass weak learner)
ECC	AdaBoost.ECC using Error Correcting Codes
ECOC	Error Correcting Output Codes
M2	Adaboost.M2
MCS	Multiple Classifier Systems
OAA	One-Against-All
OAO	One-Against-One
OC	AdaBoost.OC using Output Coding
PSL	Parallel Strong classifier within the same Layer algorithm
RIPPER	Repeated Incremental Pruning to Produce Error Reduction algorithm
ROC	Receiver Operating Curve
UCI	University of California Irvine (Machine learning dataset repository)
VJ	Viola-Jones (Ensemble-cascade algorithm)