University of Exeter
Department of Computer Science

Multi-Objective ROC learning for classification

Andrew Robert James Clark

December 2011

Submitted by Andrew Robert James Clark, to the University of Exeter as a thesis for the degree of Doctor of Philosophy in Computer Science , December 2011.

I certify that all material in this thesis which is not my own work has been identified and that no material has previously been submitted and approved for the award of a degree by this or any other University.

(signature) .....

Abstract

Receiver operating characteristic (ROC) curves are widely used for evaluating classifier performance, having been applied to e.g. signal detection, medical diagnostics and safety critical systems. They allow examination of the trade-offs between true and false positive rates as misclassification costs are varied. Examination of the resulting graphs and calculation of the area under the ROC curve (AUC) allows assessment of how well a classifier is able to separate two classes and allows selection of an operating point with full knowledge of the available trade-offs.

In this thesis a multi-objective evolutionary algorithm (MOEA) is used to find classifiers whose ROC graph locations are Pareto optimal. The Relevance Vector Machine (RVM) is a state-of-the-art classifier that produces sparse Bayesian models, but is unfortunately prone to overfitting. Using the MOEA, hyper-parameters for RVM classifiers are set, optimising them not only in terms of true and false positive rates but also a novel measure of RVM complexity, thus encouraging sparseness, and producing approximations to the Pareto front. Several methods for regularising the RVM during the MOEA training process are examined and their performance evaluated on a number of benchmark datasets demonstrating they possess the capability to avoid overfitting whilst producing performance equivalent to that of the maximum likelihood trained RVM.

A common task in bioinformatics is to identify genes associated with various genetic conditions by finding those genes useful for classifying a condition against a baseline. Typically, datasets contain large numbers of gene expressions measured in relatively few subjects. As a result of the high dimensionality and sparsity of examples, it can be very easy to find classifiers with near perfect training accuracies but which have poor generalisation capability. Additionally, depending on the condition and treatment involved, evaluation over a range of costs will often be desirable. An MOEA is used to identify genes for classification by simultaneously maximising the area under the ROC curve whilst minimising model complexity. This method is illustrated on a number of well-studied datasets and applied to a recent bioinformatics database resulting from the current InChianti population study.

Many classifiers produce "hard", non-probabilistic classifications and are trained to find a single set of parameters, whose values are inevitably uncertain due to limited available training data. In a Bayesian framework it is possible to ameliorate the effects of this parameter uncertainty by averaging over classifiers weighted by their posterior probability. Unfortunately, the required posterior probability is not readily computed for hard classifiers. In this thesis an Approximate Bayesian Computation Markov Chain Monte Carlo algorithm is used to sample model parameters for a hard classifier using the AUC as a measure of performance. The ability to produce ROC curves close to the Bayes optimal ROC curve is demonstrated on a synthetic dataset. Due to the large numbers of sampled parametrisations, averaging over them when rapid classification is needed may be impractical and thus methods for producing sparse weightings are investigated.

Contents

1 Introduction **9**

1.1 Structure of the thesis . . . 13

1.2 Principal Contributions . . . 14

2 Background **15**

2.1 Supervised Learning . . . 17

2.1.1 Maximum a posteriori estimation . . . 22

2.1.2 Sampling approaches . . . 22

2.1.3 Summary . . . 25

2.2 Measuring classifier performance . . . 25

2.2.1 Confusion Matrix . . . 26

2.2.2 Receiver Operating Characteristic graphs . . . 30

2.3 Sparse Kernel Models . . . 36

2.3.1 Relevance Vector Machines . . . 39

2.3.2 Summary . . . 47

2.4 Multi-Objective Optimisation . . . 48

2.4.1 Dominance and Pareto Optimality . . . 49

2.4.2 Multi-objective Evolutionary Algorithms . . . 50

2.4.3 Multi-objective Optimisation of ROC curves . . . 54

2.4.4 Summary . . . 55

3 Controlling Complexity in RVMs **57**

3.1 Introduction . . . 57

3.2 The Relevance Vector Machine . . . 58

3.3 Measuring complexity of RVM classifiers . . . 60

3.4 Evolving Sparse RVMs . . . 62

3.4.1 Multi-objective optimisation of RVM . . . 62

3.4.2 Illustration . . . 64

3.5 Cross Validation Methods for EAs . . . 66

3.5.1 Two-archive methods . . . 67

3.5.2 K-fold Cross Validation . . . 70

3.6 Benchmark Dataset Results . . . 71

3.7 Locating fRVM equivalent solutions . . . 76

3.8 Conclusion . . . 78

3.9 Future Work . . . 79

4 Gene Selection from Classification **81**

4.1 Introduction . . . 81
4.2 Background . . . 83
4.3 Multi-Objective Evolutionary Algorithm . . . 85
4.4 Splitting Procedure . . . 87
4.5 Experimental Results . . . 88
4.5.1 Leukaemia Dataset . . . 89
4.5.2 Colon Cancer . . . 90
4.5.3 Hereditary Breast Cancer . . . 93
4.6 Analysis of gene selection counts . . . 96
4.6.1 Feature selection counts . . . 96
4.6.2 Summary . . . 100
4.7 InChianti Dataset . . . 101
4.8 Further Analysis of Extreme Cognitive Divergence Results . . . 109
4.9 Potential models for Extreme Cognitive Divergence classification . . . 112
4.9.1 Negative Correlation Investigation . . . 116
4.9.2 Summary . . . 119
4.10 Chapter Summary and Conclusion . . . 119
4.11 Future Work . . . 120
4.12 Acknowledgements . . . 121

5 Soft classifiers from hard: Using Approximate Bayesian Computation to average hard classifiers 122
5.1 Introduction . . . 122
5.2 Approximate Bayesian Computation . . . 124
5.3 AUC error as a distance function . . . 127
5.4 Illustration . . . 129
5.4.1 MLP . . . 129
5.4.2 Priors . . . 130
5.4.3 AUC error ABC MCMC Algorithm Implementation . . . 131
5.4.4 Results . . . 132
5.5 Sparse Ensembles . . . 134
5.5.1 Performance-based selection . . . 136
5.5.2 fRVM based weighting . . . 138
5.5.3 Summary . . . 140
5.6 Illustration: Waveform Data . . . 141
5.7 Conclusion . . . 143
5.8 Future Work . . . 144

6 Conclusions 146
6.1 Multi-objective optimisation of RVM classifiers . . . 146
6.2 Gene Selection . . . 147
6.3 Approximate Bayesian Computation MCMC . . . 149
6.4 Overview . . . 150

A Additional Plots for Chapter 3 151

B InChianti Dataset Cofactors **152**

C Extreme Cognitive Divergence Results analysis **153**

Bibliography **154**