What's inside the black-box?
A genetic programming method for interpreting complex machine learning models

Morgan Jones
mwj7@aber.ac.uk

Department of Computer Science,
Aberystwyth University,
Aberystwyth, Wales, UK

Introduction

- Best performing ML techniques are worst with respect to being interpretable/explainable
- Explainability key for adoption of AI in more areas
- Model extraction as addition to ML to generate understandable models
- Most natural example: decision trees
- Greedy tree-construction flawed
- Our approach: Multi-objective GP for model extraction

The New Method

We propose a novel model agnostic approach to XAI model extraction. We use NSGA-II paired with strongly typed GP (STGP) to evolve decision tree-like structures which simultaneously balance the complexity and accuracy of the trees. Complexity is minimised and accuracy maximised by our objective functions below.

$$maximise \frac{1}{k}\sum_{i=1}^{k} f1(predict(fold(i)), blackbox\_predict(fold(i)))$$

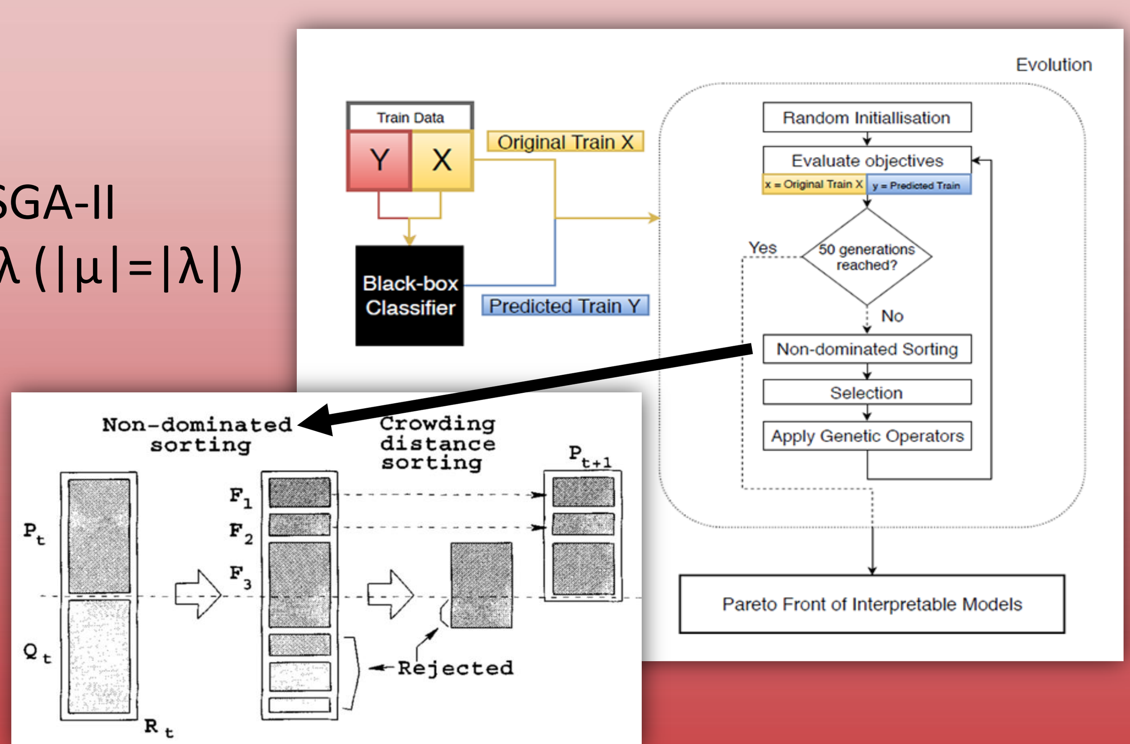
$$f1(predicted, real) = \left(\sum_{c \in C} |C| \times \frac{2 \times precision \times recall}{precision + recall}\right) / \sum_{c \in C} |C|$$

$$minimise \sum split\_points$$

We use subtrees to construct features as mathematical expressions, these implicit features allow our trees to learn simpler rules.

F1 metric is result of an internal 3 fold cross-validation (k=3) .

Elitist NSGA-II
with μ+ λ (|μ|=|λ|)



Evolutionary training process of our algorithm shown above alongside diagram of non-dominated sorting in NSGA-II.

Experiments & Results

4 Current Model Extraction methods
- Bayesian Rule Lists
- Logistic Regression
- Decision Tree
- Simplified Decision Tree

3 Black-Box Models
- Random Forests
- Gradient Boosting
- Deep Neural Network

The reconstruction ability was the f1 measure result of a 10 fold cross-validation averaged across all three black-box classifiers. For each model extraction method this was done for each dataset.

Used 30 datasets from the OpenML repository. These were restricted to <15000 instances, <5 classes, and no missing values.

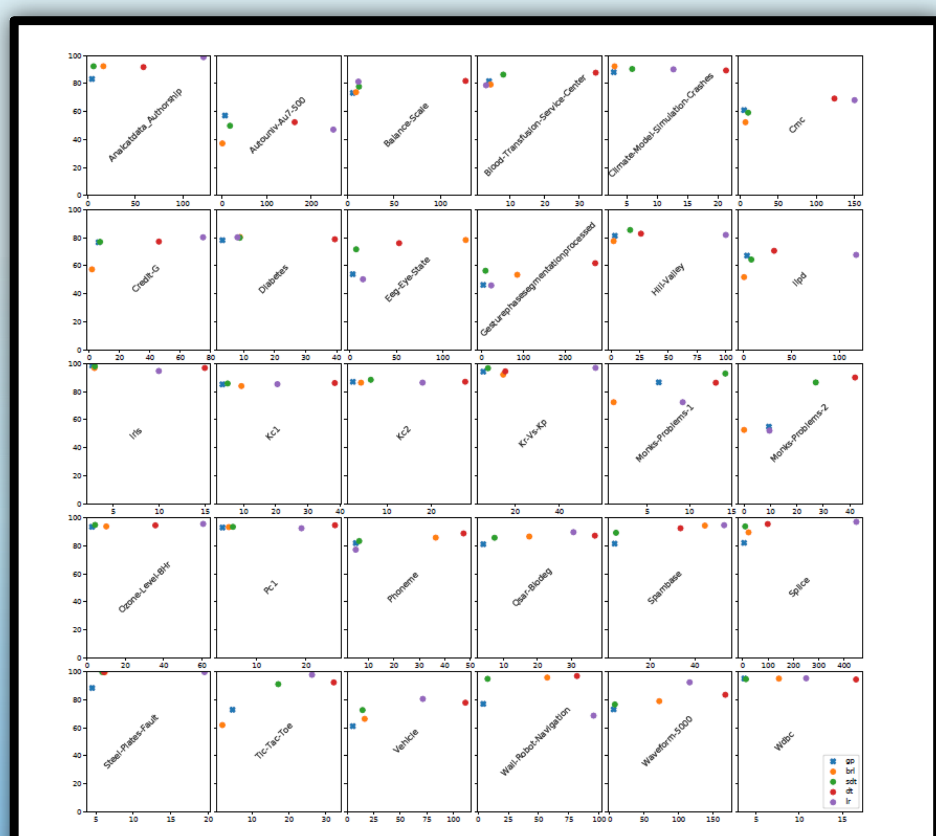


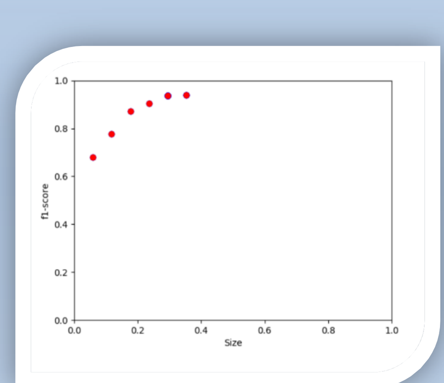
Figure 7: Testing recreation ability (y) vs Complexity (x). The ideal position is the top left corner of the graph.

Table 1: Summary of dataset characteristics

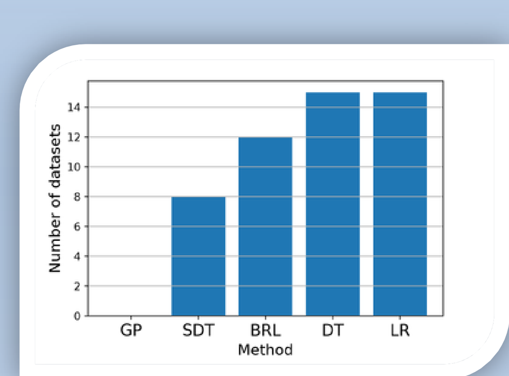
Dataset	Numeric Features	Categorical Features	Classes	Instances
analcatdata	70	0	4	841
autoUniv-au7-500	8	4	5	500
balance-scale	4	0	3	625
blood-transfusion	4	0	2	748
climate-model	20	0	2	540
cmc	2	7	3	1473
credit-g	7	13	2	1000
diabetes	8	0	2	768
eeg-eye-state	14	0	2	14980
GesturePhase	32	0	5	9873
hill-valley	100	0	2	1212
ilpd	9	1	2	583
iris	4	0	3	150
kc1	21	0	2	2109
kc2	21	0	2	522
kr-vs-kp	0	36	2	3196
monks-problems-1	0	6	2	556
monks-problems-2	0	6	2	601
ozone-level-8hr	72	0	2	2534
pc1	21	0	2	1109
phoneme	5	0	2	5404
qsar-biodeg	41	0	2	1055
spambase	57	0	2	4601
splice	0	60	3	3190
steel-plates-fault	33	0	2	1941
tic-tac-toe	0	9	2	958
vehicle	18	0	4	846
wall-robot-navigation	24	0	4	5456
waveform-5000	40	0	3	5000
wdbc	30	0	2	569

Table 2: Summary of the results. The average testing performance is presented.

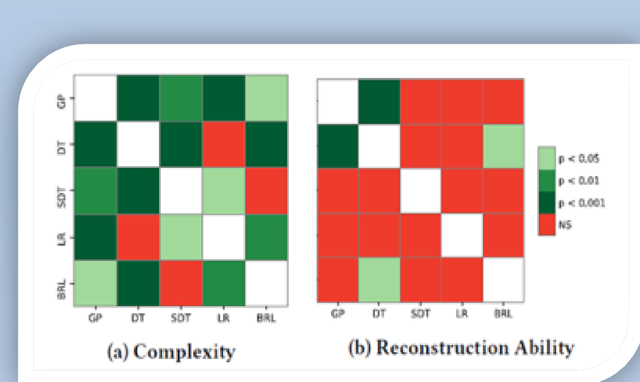
	Black-box Test Accuracy			Test Reconstruction Ability					Model Complexity				
	RF	GB	DL	GP	BRL	SDT	DT	LR	GP	BRL	SDT	DT	LR
Analcatdata_Authorship	99.40	98.70	99.80	83.21	92.33	92.3	91.71	98.81	5	17	6	59	122
Autouniv-Au7-500	47.50	44.52	38.30	55.69	36.83	49.49	50.64	46.57	7	1	18	163	250
Balance-Scale	82.70	86.40	96.90	72.9	73.06	77.6	81.44	80.68	6	9	13	127	12
Blood-Transfusion	66.70	73.70	70.50	81.65	79.24	86.34	87.71	78.73	4	5	8	35	4
Climate-Model	86.90	85.30	88.20	88.44	92.47	90.66	89.72	91.09	3	3	6	22	13
Cmc	54.30	54.20	45.90	60.58	51.84	58.6	68.62	67.81	6	7	11	123	150
Credit-G	73.60	75.70	72.40	76.61	57.3	76.92	77.12	80.19	7	2	8	46	76
Diabetes	74.20	73.40	71.00	77.5	79.56	78.72	78.15	79.26	4	9	9	39	8
Eeg-Eye-State	93.10	87.50	78.90	54.26	77.65	71.86	74.3	49.28	4	120	8	50	14
GesturePhase	67.00	62.80	60.00	47.51	53.16	55.94	61.35	47.14	5	81	10	270	23
Hill-Valley	35.70	52.70	64.10	80.57	76.68	84.96	82.26	81.4	4	3	18	27	100
Ilpd	65.40	66.60	70.00	67.2	51.29	63.98	70.13	67.19	4	1	8	32	118
Iris	93.30	94.64	98.30	98.93	97.37	98.46	96.89	95.4	3	3	3	15	10
Kc1	82.30	84.30	82.70	85.19	83.98	85.8	86.14	85.34	4	10	5	39	21
Kc2	77.80	80.06	86.10	86.79	86.38	87.33	85.96	85.7	3	5	6	27	18
Kr-Vs-Kp	98.80	99.40	99.00	94.22	92.25	96.62	94.48	96.8	6	15	8	16	57
Monks-Problems-1	99.80	98.90	99.10	86.63	72.44	92.9	86.34	72.44	7	2	15	14	10
Monks-Problems-2	92.30	97.10	99.80	55.11	52.85	86.3	89.93	52.42	10	0	28	43	10
Ozone-Level-8Hr	93.70	93.20	93.60	93.44	93.76	94.82	94.52	95.55	3	11	5	36	61
Pc1	91.90	92.80	91.50	92.95	93.26	93.45	94.6	92.46	4	5	6	26	20
Phoneme	91.10	88.50	90.71	81.77	85.02	82.17	88.36	76.96	6	35	6	47	5
Qsar-Biodeg	87.00	86.60	84.70	81.14	86.52	85.81	87.1	89.59	4	18	8	38	31
Spambase	95.20	95.40	93.90	81.59	94.27	89.27	92.42	94.65	4	45	5	33	54
Splice	97.30	96.20	95.10	82.09	89.66	93.62	95.47	96.99	6	24	10	100	449
Steel-Plates-Fault	99.70	94.50	99.80	88.21	99.67	99.78	99.8	99.74	5	7	6	7	20
Tic-Tac-Toe	98.80	97.40	97.60	73.07	61.3	91.1	92.25	97.83	6	3	18	33	27
Vehicle	75.20	77.30	84.30	61.14	66.36	72.65	77.9	80.59	6	17	15	112	72
Wall-Robot-Navigation	99.20	99.69	92.50	78.63	96.23	95.33	97.4	68.53	5	58	8	79	96
Waveform-5000	85.30	83.80	83.20	72.64	79.48	76.91	83.51	92.5	6	71	8	168	117
Wdbc	95.40	94.40	98.70	95.21	95.13	94.14	94.2	95.48	4	8	4	17	11



Pareto front of trees to choose from



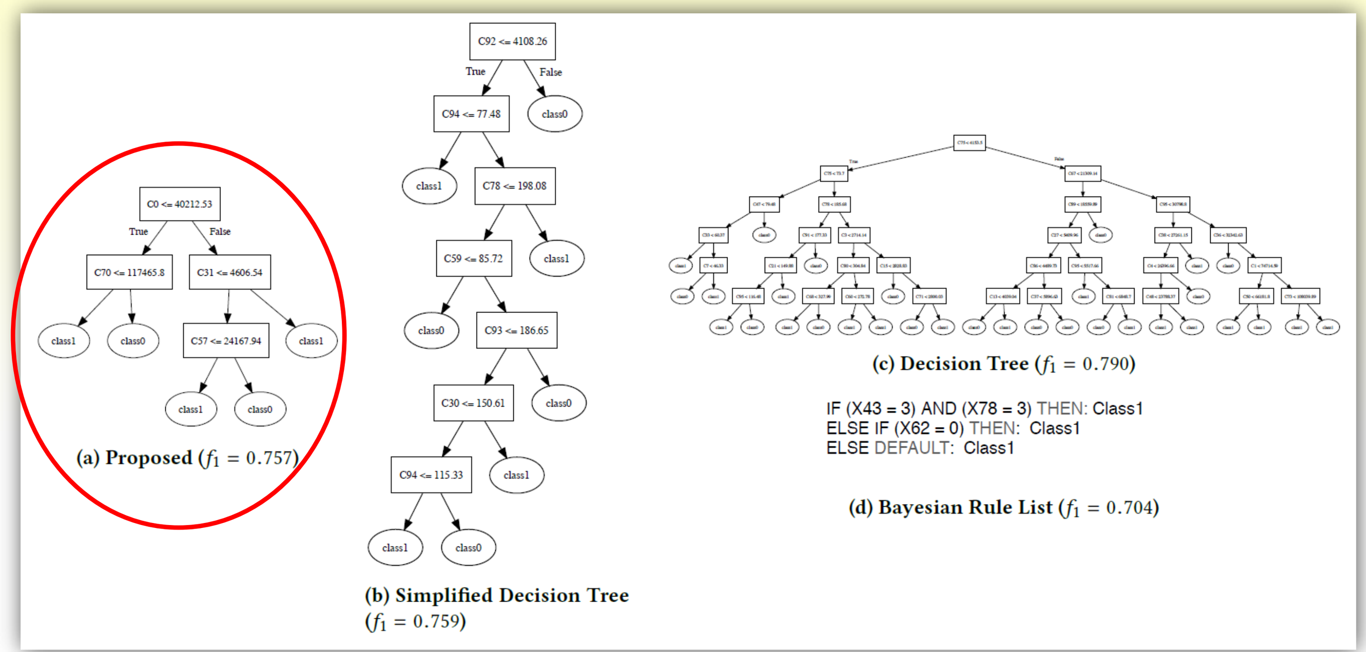
Method not dominated on any dataset



Significantly simpler interpretable models with equivalent accuracy

Further Analysis

Looking into our evolved tree we can see its splitting points make sense when considering the hill-valley dataset, which "when plotted in order the Y coordinate will create either a Hill or a Valley. We can see the tree is checking the first point, and comparing to the point at 30\%, or the point at 70\%, where the tree is trying to distinguish between classes by finding the common points for the hills/valleys and checking if these are high or low relative to the training data (e.g. a high point at the start, a low point at 30\%, then a high point at 57\% indicates a valley based on this tree).



Difficult Datasets		
Dataset	Reason	Evaluation
Autonuniv-Au7-500 & GesturePhase	5 classes	Relax push for simple trees on datasets with many classes
onks-Problems-2	entirely categorical features	Combining categorical features into a single branch
eeg-eye-state	data is sequential/time-series	The proposed method is not designed for such datasets

Conclusion

The new method was compared to existing approaches for model extraction, and was found to offer drastically simpler models, with statistically equivalent test accuracy. To our best knowledge, this is the first utilisation of multi-objective optimisation in explainable AI. We also believe this is the first application of GP for model extraction, and shows a promising direction for future developments.

Open Questions

- **Can recreation ability be improved without sacrificing simplicity?**
- **Can we find a more suitable measure of complexity to describe human interpretability?**
- **Is it possible to guide the evolution of the models based on human feedback?**