

Using Genetic Algorithms and Boosting for Data Preprocessing

by

Celestino Lei

Master of Science in Software Engineering

2002



**Faculty of Science and Technology
University of Macau**

Contents

1	Introduction	1
I	Theoretical Background	5
2	Machine Learning Methods	7
2.1	Supervised learning	9
2.2	Connectionist Methods	11
2.2.1	Backpropagation Networks	12
2.3	Decision Tree	15
2.3.1	Description of ID3	18
2.3.2	Improvements in C4.5	20
2.4	Nearest Neighbor	21
2.5	Occam's Razor Principle	23
2.5.1	Theoretical framework	23
2.6	Classifier Properties	26
3	Classifier Committees	29
3.1	Bagging	30
3.2	Boosting	32
3.2.1	AdaBoost	36
3.2.2	Arc-x4	37
3.2.3	Further improvements	38
4	Genetic Algorithms	41
5	Data Preprocessing	45
5.1	Feature Selection	46
5.2	Features relevance	47
5.3	Feature Selection Approaches	48
5.3.1	Filter Approach	49
5.3.2	Wrapper Approach	54
5.4	Training Examples Selection	56
5.5	Genetic Algorithms in Data Preprocessing	58

5.5.1	GA applied to training examples selection	60
5.5.2	GA applied to feature selection	64
5.6	Boosting for Training Examples Selection	65
II	Practical Applications and Experiments	67
6	Applications	69
6.1	Boreholes and Piles	70
6.1.1	Application of Machine Learning Methods	72
6.1.2	Application of data preprocessing techniques	74
6.2	Automatic Questionnaire Input System	75
6.2.1	Digitization of the forms	77
6.2.2	Isolation of input boxes	77
6.2.3	Recognition of each box's input	80
6.2.4	Verification of suspicious recognition	81
6.2.5	Storage of the data	81
6.2.6	Feature extraction	81
6.2.7	The use of connectionist methods	82
6.2.8	The use of C4.5	86
6.2.9	The use of nearest neighbor methods	86
6.2.10	The use of classifier committees	88
6.2.11	The use of Genetic Algorithms	91
7	Experiments	95
7.1	Boreholes experiments	95
7.2	AutoQuIS experiments	100
7.2.1	Backpropagation classifier	101
7.2.2	C4.5 classifier	105
7.2.3	Nearest neighbor classifier	106
7.2.4	Application of Boosting to the C4.5 classifier	107
7.2.5	Application of data preprocessing techniques to the nearest neighbor classifier	111
7.2.6	Comparison between data preprocessing methods	116
8	Analysis	123
9	Conclusion and future work	127
A	Program class definitions	135
A.1	Backpropagation network	135
A.2	Arc-x4	137
A.3	Nearest neighbor	139
A.4	Genetic Algorithms for training examples selection	140