Coil 2000 Challenge Submission:
Genetic Algorithms and Hill-climbers

Dr Jonathan Carter
Huxley School for the Environment, Earth Sciences and Engineering
Imperial College of Science Technology and Medicine
South Kensington, London SW7 2BP, UK
tel: +44 (0)20 75947322; fax: +44 (0)20 75947444
email: j.n.carter@ic.ac.uk

METHOD

We have used a combination of Genetic Algorithms and Hill-climbers to generate if-then type rules that select sub-populations that have high probability of containing policy holders.

STAGE 1

We used a GA to develop multicomponent rules of the form:

if ((lower<=variable<=upper) and (lower<=variable<=upper) ...
    ...) then select individual

where "variable" is any of the 85 variables that categorise the population/ "lower" and "upper" are relevant bounds for that variable. No limit was place on the number of variables tested. The GA was asked to maximise the number of policy holders selected, whilst maximising the number of variables tested and minimizing the difference between lower and upper bounds.

STAGE 2

A selection of multi-component rules reduced to single component rules using a simple hill-climber. The object was to identify those simple rules that were most discriminating between the two sub-populations.

STAGE 3

We recombined the optimised single component rules to obtain multi-component rules that select small sub-populations. Finally we select the 800 solutions that are most likely to contain policy holders.

DESCRIPTION

The method identified the following simple rules:

Rule 1: variable 3 range 3-4.
Average size of house hold three or four people.
This is selecting for areas with small families, probabily two
adults and one/two childern.

Rule 2: variable 18 range 0-4.
Lower level education, 0%-49%.

This is selecting area in which people with lower education are poorly represented, ie wealthier areas.

Rule 3: variable 21 range 0-0.
Farmers, 0%.
This appears to select against country areas, ie those with farmers.

Rule 4: variable 29 range 0-1.
Social Class D, 0%-10%.
This is selecting for areas with few in social class D.

Rule 5: variable 34 range 0-2.
No car, 0%-23%.
This is choosing areas where most households have cars.

Rule 6: variable 37 range 0-2.
Income <30000, 0%-23%.
This is selecting against areas with low incomes, ie choosing wealthy areas.

Rule 7: variable 47 range 6-6.
Contribution to car policies, 1000-4999.
This is selecting areas with a particular attitude to buying car insurance.

Rule 8: variable 65 range 1-1.
Number of third party insurers, one company.
This selects areas where an individual buys third party insurance from one company only.

Rule 9: variable 9 range 1-2.
Number of car policies, 1-2.
This is selecting areas where most households have one or two separate car insurance policies.

PREDICTION AND SELECTION

Applying these rules to the two test populations gives the following statistics:

Training Population

Rule	Policy holders	Non-policy holders	Total selected	Probability
1	221	3118	3339	6.6%
2	219	2521	2740	8.0%
3	284	3892	4176	7.3%
4	286	3884	4170	6.9%
5	277	3574	3851	7.2%
6	243	2785	3028	8.0%
7	262	2057	2319	11.3%
8	201	2133	2334	8.6%
9	275	2475	2750	10.0%

Rule	Selected population	Expected policy holders
1	2306	152.6
2	1825	145.9
3	2809	205.0

4	2864	196.4
5	2562	184.3
6	2089	167.6
7	1591	179.8
8	1575	135.6
9	2006	200.6

Applying the rules in combination gave the following statistics:
142 individuals were selected by all 9 rules
312 individuals were selected by some combination of 8 rules
424 individuals were selected by some combination of 7 rules

giving a total of 878 individuals.

To get our selected population down to 800 we randomly deleted 78 individuals from the set of 424 who were selected by only 7 rules. These individuals were from a sub-population that were selected by rules 1, 3 and 5 (the least discriminatory of our rules).

The various rules were triggered by different numbers of individuals:

	Population Sizes	
Rule	878	800
1	736	658
2	693	647
3	778	700
4	839	773
5	818	740
6	792	728
7	721	683
8	570	526
9	795	741