Genetic Algorithm with Logistic Regression for Alzheimer's Diagnosis and Prognosis

Piers Johnson

This project aims to develop a Genetic Algorithm (GA) in combination with Logistic Regression (LR) for Alzheimers' Disease (AD) progression prediction, which has not been reported in the literature. In this study, GA was used for finding one or more set(s) of neuropsychological tests which can predict the best of AD progression. LR was used for producing the fitness functions which is the evaluation of the predictive effect on AD progression with certain variables selected by the GA. The best solutions are defined as the set of variable which best classifying the conversions from HC to MCI/ AD or conversions from MCI to AD in 36 months. The classification models are evaluated with ROC. The models produce the highest ROC are considered as the best models.

A battery of neuropsychological tests, including depression and anxiety measures, from the Australian Imaging, Biomarker & Lifestyle (AIBL) study with 36 months follow up data was used for this study. Data from 31 healthy controls (HC) who converted to either mild cognitive impairment (MCI) or AD, and 604 who remained healthy were investigated for building the models for prediction of HC conversion. MCI cases which included 47 converters at 36 months and 30 non-converters were used to build the MCI conversion models. The neuropsychological test scores used for this study are normed and age adjusted [Ellis, 2009]. The whole set of neuropsychological variables used for GA to select the best subsets for prediction is listed in Table 1.

Table 1: Whole set of neuropsychological variables used for this study

1	MMSE	20	BNT...No.Cue..Australian.Z.Score.
2	Age.Scaled.Score..digit.span.	21	BNT...No.Cue..US.Z.score.
3	Age.Scaled.Score..digit.symbol.coding.	22	Clock.score.
4	Pass.Fail	23	Score..out.of.50.
5	List.A.1.5.T.score	24	UK.Pred.FSIQ
6	List.A.T6.Retention..Z.score.	25	US.Pred.FSIQ
7	List.A.Delayed.Recall..Z.score.	26	Dots.time..Z.score.
8	List.A.Recognition..Z.score.	27	Dots.errs
9	List.A.False.Positives..Z.score.	28	Words.time..Z.score.
10	Total.Recog.Discrim.d...Z.score.	29	Words.errs
11	RCFT.Copy..Z.score.	30	Colours.time..Z.score.
12	RCFT.Copy.time..Z.score.	31	Colours.errs
13	RCFT.3.min.delay..Z.score.	32	C.D.Stroop..Z.score.
14	RCFT.30.min.delay..Z.score.	33	CDR.Sum.of.Boxes
15	RCFT.Recog..Z.score.	34	HADS..D.
16	FAS..Age.Scaled.Score.	35	HADS..A.
17	Animals...Names..Age.scaled.score.	36	Recall.RAW..LM1.
18	Fruit.furniture.Total...Age.scaled.score.	37	Recall.RAW..LMII.
19	Fruit.furniture.Switching..Age.Scaled.Score.	38	Raw.score..digit.symbol.coding

Genetic Algorithms (GAs) are stochastic search mechanisms based on natural selection concepts. Potential solutions to a problem compete and mate with each other in order to produce increasingly stronger individuals. Each individual (called genome or chromosome in GA) in the population represents a potential solution to the problem that is to be solved, i.e. the optimization of some generally very complex function.

A binary genetic algorithm was used for this study. Each variable in the GA was represented as a bit in the individual genome. 2 point crossover was chosen for reproduction of the next generation, as it was noted as an optimal form of crossover for binary GAs in (Fvan Rooij, Jain et al. 1996). The type of mutation chosen was single bit flip mutation, this was chosen to minimise the changes to the binary genome as too much change would be more likely to cause a deleterious genome to be formed.

Each model from the GA was analysed further using a variation on Monte Carlo (MC) cross validation (CV). For each iteration within the MC the data was randomly split into 80% for training and 20% for validation. This was repeated for 1000 trials, and the areas under the ROC were averaged at the end to give a final number. The results are shown in Table 1 and Table 2 for prediction of conversion from HC to MCI/AD and MCI to AD respectively.

Table 1: GA results, HC conversion to MCI/AD over 36 months

GA_AUC	MC_AUC	#	Variables
0.894	0.866	9	1,5,8,10,15,17,19,21,38
0.916	0.877	10	1,5,7,8,15,17,18,30,33,38
0.926	0.854	10	5,6,8,10,15,18,22,27,33,38
0.941	0.864	10	1,2,4,5,6,15,17,18,20,28
0.907	0.892	11	5,7,8,13,14,17,18,20,32,33,38
0.905	0.873	13	1,2,5,9,12,13,15,17,19,20,21,32,38
0.912	0.812	14	6,7,8,9,14,15,16,18,10,23,24,27,28,38
0.913	0.869	14	1,2,5,7,8,12,15,17,18,19,30,33,36,38
0.913	0.890	15	1,5,7,8,9,16,17,18,20,22,23,24,25,27,38
0.918	0.830	15	1,2,5,6,8,9,16,17,19,22,26,30,32,34,36

Table 2: GA results, MCI conversion to AD over 36 months

GA_AUC	MC_AUC	#	Variables
0.903	0.852	6	1,6,13,15,19,25
0.907	0.849	6	1,15,19,25,27,31
0.900	0.838	8	1,8,15,16,19,23,24,31
0.922	0.846	8	1,5,7,15,19,24,27,31
0.931	0.870	9	1,5,9,15,19,21,23,31,34
0.923	0.844	10	1,9,13,14,15,19,21,22,24,33

In an effort to see how well the GA performs compared to more traditional statistical optimization techniques a stepwise(SW) algorithm was used in each of the cases. Whilst performing the stepwise optimization it was seen it was tending to produce models that were poor at converging if converging at all. GA was better or at least the same as any step model as far as the MC ROC value was concerned, the GA models were however as noted selecting much smaller variable sets than the stepwise ones. Some results are shown in Table 3.

Table 3: Comparison between GA and Stepwise results

Case	GA_ROC	SW_ROC	GA_Size	SW_SIZE
HC to AD/MCI	0.89	0.88	3	18
MCI to AD	0.87	0.87	5	14