Paper 3257-2015

A Genetic Algorithm for Data Reduction

Lisa Henley, University of Canterbury, New Zealand

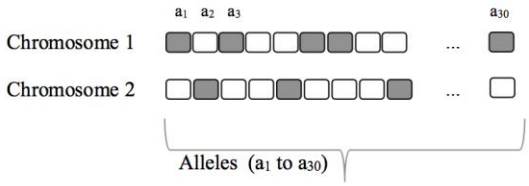
ABSTRACT

When large amounts of data are available, choosing the variables for inclusion in model building can be problematic. In this analysis, a subset of variables was required from a larger set. This subset was to be used in a later cluster analysis with the aim of extracting dimensions of human flourishing. A genetic algorithm (GA), written in SAS®, was used to select the subset of variables from a larger set in terms of their association with the dependent variable life satisfaction. Life satisfaction was selected as a proxy for an as yet undefined quantity, human flourishing. The data were divided into subject areas (for example health, environment). The GA was applied separately to each subject area to ensure adequate representation from each in the future analysis when defining the human flourishing dimensions.

GENETIC ALGORITHMS – A BRIEF INTRODUCTION

Genetic Algorithms are iterative, heuristic (experience based) search processes that can be used to find solutions to problems where an exhaustive search of all potential solutions would be impractical due to time or resource constraints. John Holland (1993) was credited with their invention in the early 1970s. They mimic natural evolution by using techniques such as natural selection, inheritance, mutation and crossover. They are used in fields and contexts where the optimal solution is required from within a large search space. This research attempts to find the best subset of variables to summarise a larger dataset.

Typically, a Genetic Algorithm (GA) starts with an initial population. This initial population consists of chromosomes that are traditionally a string of zeros and ones, and is usually a randomly generated sample of the solution search space. For example, for a variable selection / reduction exercise where there are 30 variables to choose from, each chromosome will be 30 bits long. Each of these bits is called an allele and each of these 30 alleles will consist of either a randomly generated 0 (indicating the variable is not to be selected) or 1 (indicating the variable is to be selected). Figure 1 is a visual example of two chromosomes, each 30 alleles long, from a population encoded to represent variable selection or non-selection from a total of 30 variables.



Variable (Allele) Name	Chromosome 1	Chromosome 2
Variable 1	Included	Not Included
Variable 2	Not Included	Included
Variable 3	Included	Not Included
Variable 4	Not Included	Not Included
Variable 5	Not Included	Included
Variable 6	Included	Not Included
Variable 7	Included	Not Included
Variable 8	Not Included	Not Included
Variable 9	Not Included	Included
…	…	…
Variable 30	Included	Not Included

Figure 1 Example chromosomes

The size of the initial population of chromosomes is only one of the control parameters that must be decided for a genetic algorithm. Recommendations in this area can suggest that the size of the initial population should represent the size of the problem space. There are suggested methods for finding optimal control parameters, for example Grefenstette (1986) suggests a two-stage process selecting firstly an appropriate GA to solve the problem and then secondly using further algorithms to find the optimal control parameters. Alternatively, as in Vinterbo (1999), a variety of parameters can be tried and the results evaluated in context. For example, increasing the population size may result in a better solution, but will also result in increased processing time.

The next step is to choose an appropriate fitness function to evaluate each of the chromosomes. The fitness function is important as it determines how the 'best' or maximum of the solution space is measured. For example, if the aims of the analysis were to find the best subset of potential independent variables in terms of explaining a dependent variable, as well as creating the simplest subset possible, then Akaike's Information Criterion (AIC) may be a suitable measure as AIC can meet both of these aims (Bozdogan, 1987). If however, simplicity was not paramount, and there was more interest in the explanatory power of the variable subset, then Root Mean Square Error (Beal, 2005) may be an appropriate fitness function.

The fitness function is evaluated for each member of the initial population. Using the example given, with 30 potential independent variables, and choosing the AIC as the fitness function, we would build a regression model for each member of the initial population. We would use the alleles with a value of 1 within each chromosome to decide which independent variables are in the model. Once the model has been run, the resultant AIC would be stored with the chromosome and this becomes the 'fitness' of that chromosome or solution.

Once the members of the initial population have had their fitness calculated, a selection method is chosen to decide which individual will carry on to the next step in the process. This can be done using a number of methods including, for example, 'roulette-wheel selection', where there is a chance for all to be selected, but a greater chance of selection for individuals of greater fitness, or 'tournament selection' where a number of individuals compete in a tournament, or competition, and the winner, the fittest individual, is selected to proceed to the next step. The total number of individuals chosen to move onto the next stage is another user-controlled parameter in the algorithm.

The winning chromosomes then undergo crossover and mutation. This is the method by which new potential solutions are generated. There are a number of crossover methods; the simplest is to randomly choose two parents from the winning group and swap their alleles from a random point on their chromosome to create two new offspring. The offspring may then undergo mutation that can also happen in a number of ways. The simplest way involves reversing the value of a random bit in a very small number of chromosomes, for example, changing a 0 to a 1 for 0.25% of chromosomes. Mutation helps to avoid local maxima (chromosomes with good, but not optimal fitness) and premature termination of the process, but the rate of mutation should be low. Figure 2 is a pictorial representation of the process of crossover and mutation. Two chromosomes shown one on top of the other swap a section of alleles (A and B) at the crossover point. Section A, (which was formerly attached to the upper chromosome) is now attached to the bottom chromosome and the final allele of this section is subsequently mutated.

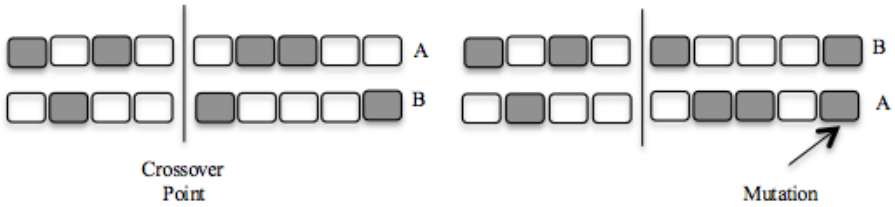


Figure 2 A pictorial representation of the crossover and mutation processes within the genetic algorithm

The newly formed offspring are then evaluated in terms of their fitness, combined with the parent population, and a new winner selection process occurs. This selection, crossover, mutation, evaluation loop continues until the 'best' solution is reached. The point of termination can be decided in a number of ways, for example, if the average fitness of the offspring does not change for 20 generations and the solution could therefore be judged ‘stable’.

In this next section a GA, written in SAS®, is used in conjunction with the REG procedure in order to undertake a data reduction exercise.

REDUCING THE DATASET

INTRODUCTION

Regression is a tool for analysing the relationship between a dependent variable, and one or more independent variables. It is often used for predictive modeling, but can also be used to determine which of the independent variables are most closely associated with the dependent variable. There are multiple forms of regression analysis, and the form chosen can have a substantial impact on the result. For example, using stepwise selection in a regression analysis may produce different results to an analysis on the same data using forward selection (Zhang & Xu, 2001). Therefore, even with classical regression, there is a large potential solution space to be explored - $2^n - 1$, where n is the number of variables to choose from.

In this work we are attempting to find a subset of variables from which to define as yet undefined dimensions of Human Flourishing, from a larger set of data. The larger set (290 variables) is stored at country level and comes from multiple open data sources, including the World Bank and International Labour Organization, each relating to a specific area defined in Stiglitz (2009) as being important in the development of future measures of human progress. These areas are:

1. Material living standards
2. Health
3. Education
4. Personal activities including work
5. Political voice and governance
6. Social connections and relationships
7. Environment
8. Insecurity

A GA written in BASE SAS®, is used to sample the solution space in order to find an optimal solution.

The dependent variable chosen is a measure of ‘life satisfaction’. This is a proxy for Human Flourishing. As life satisfaction is not expected to fully explain human flourishing, the GA is applied separately to the data associated with each of the areas in (Stiglitz, 2009). This is to ensure adequate representation from each of the areas when later calculating the Human Flourishing dimensions; as the data associated with each area has different levels of explanatory power in terms of life satisfaction. For example health related variables may be more closely associated with life satisfaction than variables relating to the environment, but variables related to the environment may be very important in the context of human flourishing.

The Genetic Algorithm attempts to find the subset of variables that, together, have the closest relationship with the dependent variable. This subset, with adequate representation from each of the Stiglitz (2009) areas will then become the 'best' summary of the dataset.

METHOD

A residual analysis was carried out on a test dataset to check the suitability of using regression.

Treating each area defined by Stiglitz (2009) separately, the human flourishing proxy (life satisfaction) was matched by country and year to the relevant records of each dataset. The human flourishing proxy, life satisfaction, was available for two years only, however, as mentioned, the aim of this analysis is to reduce the original number of variables from which to build the human flourishing dimensions at a later time, of which life satisfaction is but a part.

The GA written in BASE SAS® has user controlled parameters to allow multiple re-runs for each of the datasets. The parameters are: initial chromosome population size, chromosome length (number of independent variables), number of replicates being sampled from the population, number of replicates selected for crossover (the fittest *x*), number of elites (chromosomes retained from one generation to another without crossover), a 'stable' value (the number of iterations the average fitness function must remain the same before stability is deemed to be reached), maximum iterations (to prevent continuous running if no stable solution could be found), selection method (forward, backward or stepwise), significance level of (variable) entry and significance level of (variable to) stay. Although there are further, alternate forms of regression, for the sake of implementation simplicity, only the three mentioned are included in the analysis.

Firstly an initial randomly generated population of chromosomes is created, each chromosome consisting of 0/1 alleles. Each chromosome is of a length equal to the number of independent variables in the dataset to be reduced. Population size of either 100 or 200 is used here to check consistency of results. The code used to build the population is shown below:

```
data initpop (drop = i temp);
  *&outvars = number of variables in dataset to be reduced;
  array allele {&outvars} 3;
  *e.g. %let initpop=200;
  do until (_n_ = %eval(&initpop + 1));
    *number of variables in the dataset;
    do i = 1 to &outvars;
      temp=uniform(1);
     if temp >=0.5 then allele[i] = 1;
     else allele[i]=0;
   end;
  *make a chromosome of the alleles;
  chromosome = catt(of allele1-allele&outvars);
  output;
  _n_ + 1;
end;
run;
```

The population is randomly sampled without replacement in its entirety to create a number of replicates equal to half of the population. Then a regression equation, as follows, is built for each member of each replicate (binary tournament), using the variables selected by the alleles of the chromosome. Only main effects are included. That is:

$$Y = b_0 + b_1X_1 + b_2X_2 + \cdots + b_p X_p, where\ p \leq q$$

where *Y* is the response, or dependent, variable, and the proxy for human flourishing, life satisfaction; the *X*s represent the *p* explanatory variables selected by the GA using PROC REG, forward, backwards or stepwise regression; the *b*s are the regression coefficients, and *q* is the number of variables selected by the initial chromosome. The Root Mean Square Error (RMSE) was selected as the fitness function, as the aim is to find the subset of variables (within each dataset) that has the most explanatory power in regards to the dependent variable. This fitness function is another parameter that can be easily changed in the code. The algorithm checks which of the two tournament opponents has the lowest RMSE (fitness

function), and retains that chromosome as the fittest (winner of the tournament). It performs this operation for all of the pairs to obtain the 'best' of each replicate.

The fittest *x* replicates are selected for crossover as per the previously set parameter. A copy is taken of this group, as they become the 'parents' of the next generation. Then the average fitness of the parent group is retained in a macro variable. The 'stable' parameter value is set at 20, meaning the average fitness of the group needs to remain stable for 20 iterations before termination of the algorithm.

At this point, (for each iteration), the two fittest chromosomes are preserved as 'elites'. Then all of the remaining tournament winners, or fittest are re-sampled without replacement, again into replicates of sample size two. Crossover is performed at a random point along the chromosomes (the same point for each member of a pair and in 0.25% of cases mutation takes place.

A regression model is again built using PROC REG for each member of each tournament pair and the fitness calculated. The fittest from each pair is retained and this group is combined with the parents of that respective generation. Again, the fittest (parameterised) *x* of the group are selected for crossover and the process continues for either 500 generations or until the average fitness is stable.

As an example, shown in Figure 3, a population of 100 would divide into 50 replicates of sample size two, regression models would be built for each member of each group, and the fitness calculated. Forty (crossover parameter) of the fittest chromosomes would be selected for crossover. These would be copied (they are the parents), the average fitness of the whole group calculated and stored and the two fittest preserved as elites. The remaining 38 chromosomes would be re-sampled into 19 groups of two. A regression model built for each, the fitness calculated and the fittest 20 recombined with the 40 parents. The top 40 would be selected for crossover and the process begins again (copying the parents) until stability is reached.

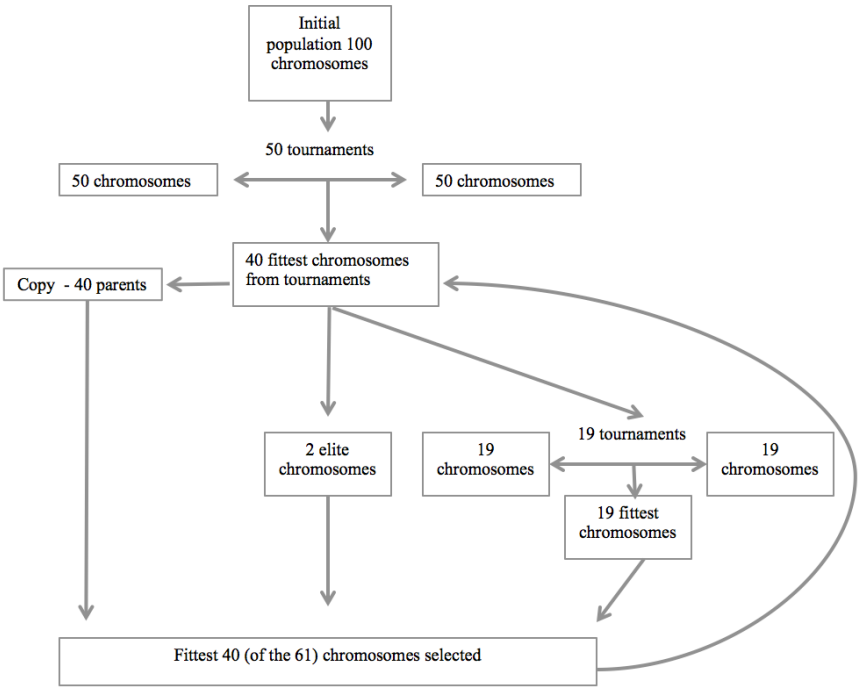


Figure 3 An example GA population evolution

The algorithm was run multiple times for each dataset (each Stiglitz (2009) subject area) and varying population size, crossover amount and the regression type. The significance level for entry (SLE) of a variable into the model was set to 0.20 and the significance level for a variable to stay in the model (SLS)

was set to 0.15. The aim of this analysis is to choose variables to be subsequently analysed, it is not a predictive analysis, so these levels were chosen to avoid excluding potentially valuable variables.

RESULTS

Tables similar to that shown in Table 1 An example summary table of GA runs were produced for each of the 15 datasets.

Trial	Population	Crossover Count	Elite Count	Time (minutes)	Generations	R-Squared	Variables Kept	Variable Description
Forward Selection								
1	100	40	2	3	55	0.707167	SH_IMM_IDPT	Immunization, DPT (% of children ages 12-23 months)
							SH_IMM_MEAS	Immunization, measles (% of children ages 12-23 months)
							SH_XPD_EXTR_ZS	External resources for health (% of total expenditure on health)
							SH_XPD_OOPC_TO_ZS	Out-of-pocket health expenditure (% of total expenditure on health)
							SH_XPD_PCAP_PP_KD	Health expenditure per capita, PPP (constant 2005 international $)
							SH_XPD_PUBL_GX_ZS	Health expenditure, public (% of government expenditure)
							SP_DYN_CBRT_IN	Birth rate, crude (per 1,000 people)
							SP_DYN_LE00_FE_IN	Life expectancy at birth, female (years)
							SP_DYN_LE00_MA_IN	Life expectancy at birth, male (years)
							SP_POP_0014_TO_ZS	Population ages 0-14 (% of total)
							SP_POP_65UP_TO_ZS	Population ages 65 and above (% of total)
2	100	40	2	3	55	0.707167	As above	
3	200	80	2	6	58	0.70559	As above	
4	200	80	2	7	58	0.70559	As above	

Table 1 An example summary table of GA runs

The R-squared values differed for each dataset. For example the R- squared of the World Bank Health analysis, shown in Table 1, were relatively good compared to the environment dataset, containing variables such as carbon dioxide emissions. The variables in that dataset had almost no explanatory power for life satisfaction.

This was to be expected and was the reason why each dataset was analysed separately. As mentioned, life satisfaction was only a proxy for human flourishing and it was important that variables from all of the areas outlined in 4 be included in the final dataset, for later calculation of the Human Flourishing dimensions.

It is important to note that the variables selected by each chromosome are not necessarily those that end up as 'selected' in the table. The chromosome defines the 'starting' dataset, the forward, backward or stepwise process then determines what remains at the end.

The variables selected within each of the subject areas were combined. In this way the original 290 variables were reduced to 107 variables retaining representation from each of the subject areas.

CONCLUSION

A GA can be successfully used to trim a large dataset; allowing a much more comprehensive search of the solution space (i.e. 2^n-1 combinations of variables) than could be done manually.

Using a GA allows the analyst to modify the search dependent on requirements e.g. parsimony vs. predictive ability.

Splitting the large dataset into subject areas, and applying the GA to these 'areas' individually ensures all areas are represented in the final dataset (if required).

REFERENCES

Beal, D. J. (2005). SAS Code to Select the Best Multiple Linear Regression Model for Multivariate Data Using Information Criteria (Vol. Paper SA01_05). Portsmouth: South East SAS User Group.

Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52*(3), 345:370.

Grefenstette, J. (1986). Optimisation of Control Parameters for Genetic Algorithms. *IEEE Transactions on Systems, Man and Cybernetics*, *SMC-16*(1).

Holland, J. (1993). *Adaptation in Natural And Artificial Systems* (2nd ed.). Massachusetts: Massachusetts Institute of Technology.

Stiglitz, J. (2009). *Report by the Commission on the Measurement of Economic Performance and Social Progress (Commission of the Government of France)*. Paris.

Vinterbo, S. (1999). A Genetic Algorithm to Select Variables in Logistic Regression. *Journal of the American Medical Informatics Association 1999 Symposium Supplement*, 984–988.

Zhang, W. J., & Xu, L. (2001). Comparison of Different Methods for Variable Selection. *Analytica Chimica Acta*, *446*(1-2), 475–481.

ACKNOWLEDGMENTS

This work would not have been possible without the never-ending support, advice and knowledge of my supervisors; in particular Professor Jennifer Brown.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Lisa Henley
lisa@statsgeeks.com