A Grouping Genetic Algorithm for Joint Stratification and Sample Allocation Designs

Mervyn O'Luing, Steven Prestwich, S. Armagan Tarim

Insight Centre for Data Analytics, University College Cork
mervyn.oluing@insight-centre.org
steven.prestwich@insight-centre.org
armagan.tarim@ucc.ie

1. Introduction

In this paper we propose an algorithm to partition (create subsets of) atomic strata into larger groupings or strata and search for the minimum sample size that meets accuracy requirements from all possible partitions. We build on the work of Ballin and Barcaroli (2013) who use a Genetic Algorithm (GA) to cut short the search of all possible partitions to find the minimum sample size. We propose an alternative GA that we claim is better suited to this application. We support our claim by comparing computational results on publicly-available test data.

2. Problem Description

Consider all possible partitions of the atomic strata and for each of these partitions: estimate the minimum sample size necessary to meet your accuracy requirements. Each partition is a candidate solution. The number of possible partitions is known as the search space, because, if it was feasible in terms of time and cost, the full list of candidate solutions would be evaluated for the optimum solution. However, given that the set of partitions grows exponentially with the set of atomic strata, search techniques such as genetic algorithms are used as they can find an optimum solution without evaluating all possible solutions.

3. Our Goal

To design a *grouping genetic algorithm* (GGA) (Falkenauer, 1998). GGAs have been shown to perform far better than standard GAs on grouping problems. Our GGA should normally provide a good quality or the best solution quicker than the GA. Good solutions mean smaller samples which are cheaper to gather. It also means a saving in time taken to find the smaller sample size.

4. Why the GGA should outperform the GA

Strata should be independent (the same value cannot be in more than one stratum) but values in each stratum should be close in value (internally homogenous). There are two levels of strata in this problem. The basic level is the atomic strata. We intend to create higher level strata from subsets, partitions or groupings of atomic strata. If each atomic stratum contains values that are the same or close in value then smaller sample sizes are needed for a precise estimate of the mean or total for the target variables. It follows also that grouping homogeneous atomic strata into strata will also require a smaller sample size to meet precision constraints. This is what leads to a good candidate solution, i.e. the strata are internally homogeneous but independent and smaller samples sizes result.

Genetic algorithms create new solutions by mixing current solutions together with an occasional adjustment to see if it makes a difference. The original GA typically prolongs the search for the optimum solution, because in the mixing process good strata could be split-up and their information lost - which is a 'hit-or-miss' approach in practice and can push the sample size back up for the generated solution (offspring). The GGA on the other hand preserves the information in good strata and this is more likely to create better quality offspring.

5. Comparing the Genetic Algorithms

The online crowdfunding platform kiva.org provides a dataset of loans issued to people living in poor and financially excluded circumstances around the world over a two period for a Kaggle Data Science for Good challenge. The dataset has 671,205 unique records.

The algorithms searched for the smallest sample size necessary to accurately describe *term in months, lender count and loan amount* after 100 iterations.

GA		GGA		Reduction	
Sample size	**Strata**	**Sample size**	**Strata**	**Sample size**	**Strata**
78018	43030	11963	1793	84.67%	95.83%

The above table shows an 84.67% reduction in sample size and a 95.83% reduction in the number of strata after 100 iterations. In this test we achieve our goal of developing a GGA to find a smaller size size in less time that the GA.

References

Ballin, M. and G. Barcaroli (2013). Joint determination of optimal stratification and sample allocation using genetic algorithm. *Survey Methodology 39*(2), 369–393.

Falkenauer, E. (1998). *Genetic algorithms and grouping problems*. John Wiley & Sons, Inc.