A Genetic Algorithm and an Exact Algorithm for Classifying the Items of a Questionnaire Into Different Competences

José Luis Galán-García[1], Salvador Merino[1]
Javier Martínez[1], Miguel de Aguilera[2]

[1]Department of Applied Mathematics.
[2]Department of Audiovisual Communication and Advertising
University of Málaga - Spain

4th European Seminar on Computing
ESCO 2014
June 15-20. Pilsen, Czech Republic

Contents

1. Likert Scale
   - Rensis Likert
   - Likert Scale
   - Likert Items
   - Analyzing data of a Likert Scale
2. Description of the problem
   - Description of the problem
   - Notation
   - Data
3. Solution
   - First approach
   - Numerical approximation using a GA
   - GaMMA Method: exact solution using a CAS
4. Conclusions and future work

Rensis Likert

- Born: August 5, 1903. Cheyenne, Wyoming.
- Received his B.A. in sociology from the University of Michigan in 1926.
- Ph.D. in psychology from Columbia University in 1932. In his thesis, he devised a survey scale (**Likert Scales**) for measuring attitudes and showed that it captured more information than competing methods. The 1-5 Likert Scales would eventually become Likert's best-known work.
- Died: September 3, 1981 (aged 78). Ann Arbor, Michigan.

Likert Scale

- A **Likert Scales** is a psychometric scale commonly involved in research that employs questionnaires.
- It is the most widely used approach to scaling responses in survey research.
- Respondents specify their level of agreement or disagreement on a symmetric agree-disagree scale for a series of statements.
- The **Likert Scales** is the sum of responses on several **Likert items**.

Likert Items

- A **Likert item** is simply a statement which the respondent is asked to evaluate according to any kind of subjective or objective criteria.
- It is considered symmetric or "balanced" because there are equal amounts of positive and negative positions.
- Often five ordered response levels are used.
- The format of a typical five-level Likert item could be:
  1. Strongly disagree
  2. Disagree
  3. Neither agree nor disagree
  4. Agree
  5. Strongly agree
- Many questionnaires use four-level Likert items in order to force the respondents not to answer the neutral position.

Analyzing data of a Likert Scale

The following procedure is used to analyze data from Likert scales:

- First, weights are assigned to the responses options, e.g. Strongly desagree=1, Desagree=2, etc.
- Then negatively-worded statements are reverse-coded (or reverse scored). E.g. a score of 2 for a negatively-worded statement with a 5-point response options is equivalent to a score of 4 on an equivalent positive statement.
- Next, scores are summed across statements to arrive at a total (or summated) score.
- Each respondent's score can then be compared with the mean score or the scores of other respondents to determine his level of attitude, loyalty, or other construct that is being measured.

Description of the problem

- Given a huge questionnaire with 170 four-level Likert items $(I_1, \ldots, I_{170})$.
- The questionnaire wants to evaluate the respondents' capabilities and skills to 23 different competences $(C_1, \ldots, C_{23})$.
- We have 173 responses to the questionnaire with the corresponding results for each competence.
- We know the number of items n_k for each competence C_k but not the items themselves.
- Each item I_j belongs to only one competence C_k. That is, $C_{k_1} \cap C_{k_2} = \emptyset$ for every $k_1 \neq k_2$.
- The problem consists on identifying the items of each competence.

Notation

- I_j is the jth item of the questionnaire ($j = 1, 2, \ldots, 170$).
- C_k is the kth competence ($k = 1, 2, \ldots, 23$).
- n_k is the number of items belonging to competence k. That is, $C_k = \{\pm I_{j1}, \ldots, \pm I_{jn_k}\}$ where sign $+$ or $-$ indicates if the corresponding item is a positive or negative worded statement.
- D_{ij} is the data matrix. That is, a matrix of size 173×170 where d_{ij} is the punctuation given by respondent i to item I_j (d_{ij} should be 1,2,3 or 4). Despite of this, some $d_{ij} = 0$ since the application which get the data, allows leaving some items unanswered.
- R_{ik} is the result matrix. That is, a matrix of size 173×23 where r_{ik} is the punctuation given to respondent i for competence C_k.

Data

- I_j are known.
- n_k are known.
- $C_{22} = \{16, 20, 41\}$.
- $C_{23} = \{61, -97, -149\}$.
- Matrices D_{ij} and R_{ik} are known.
- The goal is to distribute the remaining 164 items within the sets $C_1, C_2, \ldots, C_{21}$.

First approach

- The minimum value of n_k was 7.
- We tried to find all different combinations of 7 elements from the 164 remaining items but . . .
- $\begin{pmatrix} 164 \\ \\ 7 \end{pmatrix} =$ 556.052.759.712 different combinations.
- Computationally impossible to solve.

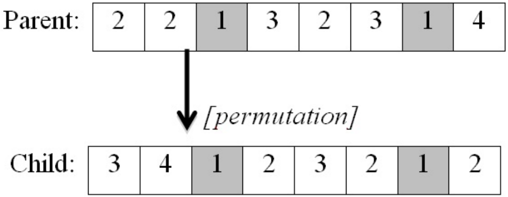
Numerical approximation using a GA

- In the computer science field of artificial intelligence, a genetic algorithm (GA) is a search heuristic that mimics the process of natural evolution.
- This heuristic is routinely used to generate useful solutions to optimization and search problems.
- Genetic algorithms belong to the larger class of evolutionary algorithms (EA), which generate solutions to optimization problems using techniques inspired by natural evolution, such as selection, genetic engineering, crossover, mutation and clonation.

Selection

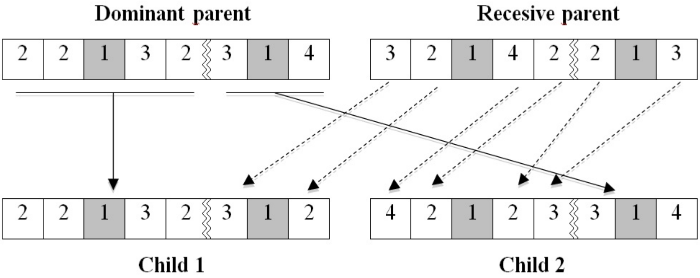
- The start point is the base solution, called “Adam”, obtained by analyzing the items and assigning them to the different competences.
- A set of clones that form the original generation of solutions (the first group of ”parents”) is generated. This part of the process is called “Selection”.
- Whenever you create new “children”, it is checked if they are better than their “parents”, and if so children replace their parents, improving the “species”.
- New children arising from their parents should keep those genes already fixed (we will note this fact with a shaded cell).

Genetic engineering

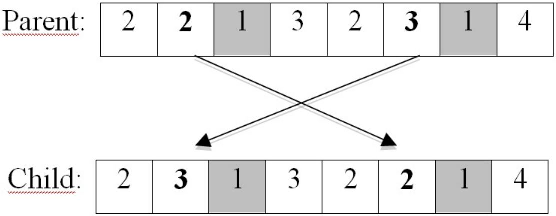


Parent:	2	2	1	3	2	3	1	4
Child:	3	4	1	2	3	2	1	2

Crossover

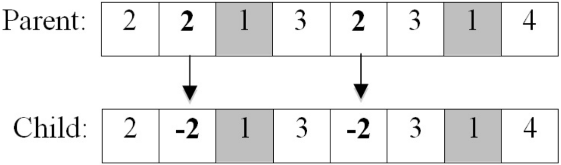

Dominant parent
2 2 1 3 2 3 1 4
Recesive parent
3 2 1 4 2 2 1 3
2 2 1 3 2 3 1 2
4 2 1 2 3 3 1 4
Child 1
Child 2

Mutation



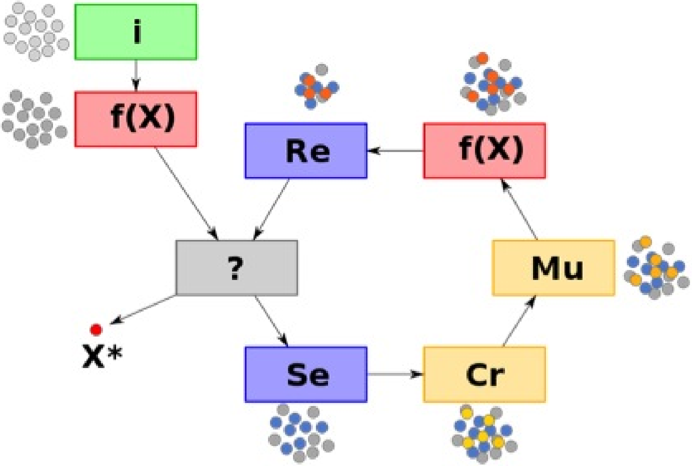
Parent:	2	**2**	1	3	2	**3**	1	4
Child:	2	**3**	1	3	2	**2**	1	4

Clonation



Parent:	2	**2**	1	3	**2**	3	1	4
Child:	2	**-2**	1	3	**-2**	3	1	4

Flux diagram



Solution

After more than 50 hours of continuously execution of the GA, we obtained the following solution:

- Exact solution for 15/21 competences (error zero).
- Very good approximation for the remaining 6 competences.
- After analyzing the data matrix D_{ij}, each of these 6 competences had items with punctuation zero ($d_{ij} = 0$).

GaMMA Method

- **GaMMA** (**Ga**lán **M**erino **M**artínez **A**guilera) Method will provide the exact solution using a CAS.
- It will start by building a quadratic system to solve the problem.
- The quadratic system will be converted to a linear system which will provide easily and quickly the solution.

Scoring

- We start with the data matrix D_{ij} and the result matrix R_{ik}.
- Remember that d_{ij} is the punctuation given from respondent i to item j. This punctuation will score for a competence k as follows:

$$\begin{array}{rll} 0 & \text{if} & I_j \notin C_k \\ d_{ij} & \text{if} & I_j \in C_k \text{ and } I_j \text{ is a positive-worded statement.} \\ 5 - d_{ij} & \text{if} & I_j \in C_k \text{ and } I_j \text{ is a negative-worded statement.} \end{array}$$

Establishing the unknowns

We now set the unknowns x_{jk} in order to solve the problem with the following meaning:

$$\begin{aligned} x_{jk} &= 0 \quad \text{if} \quad I_j \notin C_k \\ x_{jk} &= 1 \quad \text{if} \quad I_j \in C_k \text{ and } I_j \text{ is a positive-worded statement.} \\ x_{jk} &= -1 \quad \text{if} \quad I_j \in C_k \text{ and } I_j \text{ is a negative-worded statement.} \end{aligned}$$

We find now a function $f(x_{jk})$ such as $f(0) = 0$; $f(1) = d_{ij}$ and $f(-1) = 5 - d_{ij}$. This can be easily done by the quadratic function:

$$\frac{5}{2}x_{jk}^2 + \left(d_{ij} - \frac{5}{2}\right) x_{jk}$$

Quadratic system of unknowns

Therefore, the system to solve for a competence k is:

$$\frac{5}{2}x_{1k}^2+\left(d_{11}-\frac{5}{2}\right)x_{1k}+\cdots+\frac{5}{2}x_{(170)k}^2+\left(d_{1(170)}-\frac{5}{2}\right)x_{(170)k}=r_{1k}$$

$$\frac{5}{2}x_{1k}^2+\left(d_{21}-\frac{5}{2}\right)x_{1k}+\cdots+\frac{5}{2}x_{(170)k}^2+\left(d_{2(170)}-\frac{5}{2}\right)x_{(170)k}=r_{2k}$$

$$\vdots \qquad\qquad \vdots \qquad\qquad \vdots$$

$$\frac{5}{2}x_{1k}^2+\left(d_{(173)1}-\frac{5}{2}\right)x_{1k}+\cdots+\frac{5}{2}x_{(170)k}^2+\left(d_{(173)(170)}-\frac{5}{2}\right)x_{(170)k}=r_{(173)k}$$

The problem is that it is a quadratic system. But, fixing the first equation and subtracting it to the rest, we get:

Linear system of unknowns

$$\frac{5}{2}x_{1k}^2+\left(d_{11}-\frac{5}{2}\right)x_{1k}+\cdots+\frac{5}{2}x_{(170)k}^2+\left(d_{1(170)}-\frac{5}{2}\right)x_{(170)k}=r_{1k}$$

$$(d_{21}-d_{11})\,x_{1k}+\cdots+\left(d_{2(170)}-d_{1(170)}\right)x_{(170)k}=r_{2k}-r_{1k}$$

$$(d_{31}-d_{11})\,x_{1k}+\cdots+\left(d_{3(170)}-d_{1(170)}\right)x_{(170)k}=r_{3k}-r_{1k}$$

$$\vdots \qquad \vdots \qquad \vdots$$

$$\left(d_{(173)1}-d_{11}\right)x_{1k}+\cdots+\left(d_{(173)(170)}-d_{1(170)}\right)x_{(170)k}=r_{(173)k}-r_{1k}$$

The last 172 equations form a linear systems of 172 equations with 170 unknowns.

Matrix representation of the system

- Let A_{ij} be the 172×170 matrix where $a_{ij} = d_{(i+1)j} - d_{1j}$
- Let X_{jk} be the 170×21 matrix of unknowns x_{jk} (which value will be 0, 1 or -1).
- Let B_{ik} be the 172×21 matrix where $b_{ik} = r_{(i+1)k} - r_{1k}$

Therefore, the system to solve, in matrix form is:

$$A \cdot X = B$$

Solution

- rank$(A) = 170$ (computed with a CAS, specifically DERIVE).
- We found a sub-matrix $A_{\_}$ of A of size 170×170 such as rank$(A_{_}) = 170$ (specifically, removing files 96 and 172 from A).
- Let $B_{\_}$ be matrix obtained removing files 96 and 172 from B.

With this, the system $A_{\_} \cdot X = B_{\_}$ is consistent and the solution for the unknowns is given by:

$$X = {A_{\_}}^{-1} \cdot B_{\_}$$

which could be computed with DERIVE obtaining a matrix of 0, 1 and -1 values which provided us the solution of the problem.

Advantages and disadvantages of GA method

- The GA method does not require to have more equations than items.
- It is quite more slower than GaMMA method and may provide an approximation and not the exact solution.
- It needs to know the number n_k of items of each competence C_k (although the algorithm could be adapted but it would require quite more time for getting the solution).
- It only works when items belong to just one competence.

Advantages and disadvantages of GaMMA method

- GaMMA method is quite more faster and provide the exact solution.
- It does not require to know the number n_k of items of each competence C_k.
- It works although items belong to more than one competence. In fact, it has been already used to solve another similar problem with more items and items belonging to more than one competence.
- It requires to have more equations than items.
- The rank of the coefficient matrix A should be equal to the number of items (although it can be adapted to solve this situation).

Future work

- Extend the GA when n_k are not known.
- Extend the GA when items belong to more than one competence.
- Extend GaMMA Method when the rank of the coefficient matrix is less than the number of items.
- Extend both methods for other different type of questionnaires to measure attitudes (not only for Likert scale but for Category scale, Semantic differential scale, Stapel scale, Constant sum scale or Graphic scale).
- Encode the scoring of questionnaire to prevent solving this kind of problems. In this point we thanks the Spanish company QUORUM SELECTION which has partially support this research and it is interested in building a secure questionnaire with our future encode technique.

A Genetic Algorithm and an Exact Algorithm for Classifying the Items of a Questionnaire Into Different Competences

José Luis Galán-García[1], Salvador Merino[1]
Javier Martínez[1], Miguel de Aguilera[2]

[1] Department of Applied Mathematics.
[2] Department of Audiovisual Communication and Advertising
University of Málaga - Spain

4th European Seminar on Computing
ESCO 2014
June 15-20. Pilsen, Czech Republic