European Journal of Engineering Education

ISSN: 0304-3797 (Print) 1469-5898 (Online) Journal homepage: https://www.tandfonline.com/loi/ceee20

A new algorithm to create balanced teams promoting more diversity

Teresa Galvão Dias & José Borges

To cite this article: Teresa Galvão Dias & José Borges (2017) A new algorithm to create balanced teams promoting more diversity, European Journal of Engineering Education, 42:6, 1365-1377, DOI: 10.1080/03043797.2017.1296411

To link to this article: https://doi.org/10.1080/03043797.2017.1296411

Published online: 05 Mar 2017.

Submit your article to this journal

Article views: 221

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=ceee20

EUROPEAN JOURNAL OF ENGINEERING EDUCATION, 2017
VOL. 42, NO. 6, 1365–1377
https://doi.org/10.1080/03043797.2017.1296411

Taylor & Francis
Taylor & Francis Group

Check for updates

A new algorithm to create balanced teams promoting more diversity

Teresa Galvão Dias and José Borges

INESC TEC, Faculty of Engineering, University of Porto, Porto, Portugal

ABSTRACT

The problem of assigning students to teams can be described as maximising their profiles diversity within teams while minimising the differences among teams. This problem is commonly known as the maximally diverse grouping problem and it is usually formulated as maximising the sum of the pairwise distances among students within teams. We propose an alternative algorithm in which the within group heterogeneity is measured by the attributes' variance instead of by the sum of distances between group members. The proposed algorithm is evaluated by means of two real data sets and the results suggest that it induces better solutions according to two independent evaluation criteria, the Davies–Bouldin index and the number of dominated teams. In conclusion, the results show that it is more adequate to use the attributes' variance to measure the heterogeneity of profiles within the teams and the homogeneity among teams.

ARTICLE HISTORY
Received 13 January 2016
Accepted 29 January 2017

KEYWORDS
Team formation; team roles; student diversity

1. Introduction

The task of creating effective teams for a given assignment or job is a well know decision problem. Studies have suggested that diversity leads to better outcomes since a group of diverse members is able to bring up a variety of perspectives and problem solving skills that are essential to achieve better results (Page 2007).

An application example of such a problem could be an academic setting in which it may be desirable to create heterogeneous groups of students for class projects or to allocate students to classes in such a way that students are integrated in a diverse environment. Studies have suggested that heterogeneous groups promote positive interdependence, better group performance and interaction among students (Cohen 1994). The characteristics taken into account can range from race and gender to technical ability or psychological features.

The problem of grouping students into heterogeneous teams can be defined as follows. Having as input a set of students, each characterised by a set of attributes, the goal is to distribute the students among groups in such a way that the diversity of characteristics within groups is maximised and the groups are balanced. This problem is commonly known as the *maximally diverse grouping problem* (MDGP).

The traditional approach to the MDGP is to consider it as an extension of the *Maximum Diversity Problem* (MDP) formulation (Kuo, Glover, and Dhir 1993). While the MDP is defined as the task of selecting a subset of elements from a larger set of elements in such a way that the sum of the distances between the chosen elements is maximised (Martí and Duarte 2010), the MDGP refers to the problem of grouping a given set of elements into several mutually disjoint subsets to maximise

CONTACT José Borges jlborges@fe.up.pt
© 2017 SEFI

the overall diversity between elements (Fan et al. 2010). With these formulations a distance metric, such as the euclidian distance, is used to measure the similarity of profiles between pairs of students and the overall solution is measured by the sum of the distances between every pair of group members. Since the objective function takes into account just the distances between group members the homogeneity among groups is not explicitly taken into account by the method. We argue that by maximising the sum of the pairwise distances between elements the MDGP method may concentrate a big part of the diversity in a small number of the groups, thus obtaining unbalanced groups. We provide examples that illustrate this effect.

Herein we propose an alternative formulation for the MDGP in which the within group heterogeneity is measured by the attributes' variance instead of by the sum of distances between group members. Thus, the proposed formulation is focused in achieving solutions in which groups are composed of heterogeneous members while promoting equilibrium among groups.

We have been using this approach for several years with students from Information Systems courses in Engineering programmes, (Borges, Galvão Dias, and Cunha 2009). Students are assigned to teams and each team has to develop a class project during the semester, for which one of the main goals is the development of a range of skills such as organisation, chairing, team working and creative thinking.

In order to understand how a student naturally behaves when working in a team, we ask them to fill a questionnaire (Fraser and Neville 1993) that scores them against the following eight profiles: (1) president, (2) strategist, (3) intellectual, (4) monitor/evaluator, (5) operative, (6) team worker, (7) prospector, and (8) finisher/retoucher. For example, a student with leadership competencies will probably have a higher score in the president profile, while another student, more interested in conceptual models, may have a higher score in the intellectual profile. Some students may show a clear stronger profile while others may show similar scores for all profiles. Hence, in order to promote the development of the desired soft skills it is important to guarantee that the highest number of different profiles are present in each team. At the same time, it is also important to guarantee that diversity is balanced among the teams.

The students are grouped into teams by means of an efficient tabu search algorithm that was devised to solve the problem. The algorithm is given as well as the results of a set of experiments in which we compare the results of our formulation with those of the MDGP formulation for a set of real questionnaire results. In order to evaluate the solutions we make use of a modification of the Davies–Bouldin validity index (Davies and Bouldin 1979) which is commonly used to measure the similarity of clusters.

The rest of the paper is organised as follows, Section 2 gives two examples that illustrate the limitations of the MDGP formulation, Section 3 reviews related work, Section 4 presents our alternative formulation for the problem, Section 5 presents the experimental results and Section 6 our concluding remarks.

2. Limitations of the MDGP formulation

We will now give two examples to highlight the limitations of the MDGP formulation. The first example illustrates a case in which the MDGP formulation does not penalise teams composed by several identical members and the second is a graphical example that illustrates the concept of group dominance.

2.1. Example 1

Consider the task of grouping five elements, $(e_1, \ldots, e_5)$, into two teams. An element is characterised by a score in each of three attributes (a_1, a_2, a_3) in such a way that his scores add up to 30, Table 1(a). In addition, we define a matrix with the euclidian distance between each pair of elements, Table 1(b).

Table 1. An example with five elements evaluated by three attributes and the corresponding matrix of euclidian distances.

(a) The elements scores					(b) The distances matrix					
	a_1	a_2	a_3	Sum		e_1	e_2	e_3	e_4	e_5
e_1	15	10	5	30	e_1	0				
e_2	10	15	5	30	e_2	50	0			
e_3	10	10	10	30	e_3	50	50	0		
e_4	10	10	10	30	e_4	50	50	0	0	
e_5	10	10	10	30	e_5	50	50	0	0	0

We observe that in the example there are three identical elements and, intuitively, it would be preferable not to have a team composed only by elements with identical profiles.

There are $\binom{5}{2} = 10$ ways of grouping five elements into two teams, which are enumerated in Table 2. For each solution we give the sum of the Euclidean distances of the elements composing each team (dist), the variance of the team attributes (var), the overall sum of the distances for a solution (MDGP) and the global variance for the solution (VAR). For example, solution 1 has the first team $t_1 = \{e_1, e_2\}$ with $\text{dist} = d_{12} = 50$ and $\text{var} = 25$; and the second team $t_2 = \{e_3, e_4, e_5\}$ with $\text{dist} = 0$ and $\text{var} = 0$. Thus, this solution has an MDGP value of $d_{12} + d_{34} + d_{45} + d_{35} = 50 + 0 + 0 + 0 = 50$ and a VAR value of 25. In uppercase we highlight the elements with identical profiles E_3, E_4 and E_5.

The results given in Table 2 show that the MDGP formulation only penalises solution 1, that corresponds to having the three identical members in a group of three elements. All other solutions have identical value. In particular we note that the solutions that correspond to having a two member's team composed by two identical elements are not penalised by the MDGP formulation, solutions 8, 9 and 10. We argue that such solutions are not desirable since all diversity is concentrated in one team. In contrast, if the formulation that aims to maximise the within group variance is used all solutions inducing teams composed by a set of identical members are penalised.

2.2. Example 2

A second example is given in Figure 1 that provides a graphical illustration of a scenario in which the variance formulation is able to induce more heterogeneous teams than the MDGP formulation. In the example there are five elements, each having a score on two skills (in this case the sum of the scores is not constant across the students). The plots in the figure represent the optimal solution for each of the two formulations. We observe that in the solution induced by the MDGP formulation the team composed by two elements has much less variability of profiles than the other team, since in both axes the team skills are inside the other team skills boundaries.

We refer to this as a team dominating another team. For a given skill, we define *dominance* of a team over another as having the amplitude of the latter bounded by the amplitude of the former. When a team dominates another over n skills we say that there is a n-dominance. Following this

Table 2. The enumeration of all the admissible team configurations for the example given in Table 1.

Sol	Team 1		dist	var	Team 2			dist	var	MDGP	VAR
1	e_1	e_2	50	25	E_3	E_4	E_5	0	0	50	25
2	e_1	E_3	50	25	e_2	E_4	E_5	100	33.3	150	58.3
3	e_1	E_4	50	25	e_2	E_3	E_5	100	33.3	150	58.3
4	e_1	E_5	50	25	e_2	E_3	E_4	100	33.3	150	58.3
5	e_2	E_3	50	25	e_1	E_4	E_5	100	33.3	150	58.3
6	e_2	E_4	50	25	e_1	E_3	E_5	100	33.3	150	58.3
7	e_2	E_5	50	25	e_1	E_3	E_5	100	33.3	150	58.3
8	E_3	E_4	0	0	e_1	e_2	E_5	150	50	150	50
9	E_3	E_5	0	0	e_1	e_2	E_4	150	50	150	50
10	E_4	E_5	0	0	e_1	e_2	E_3	150	50	150	50

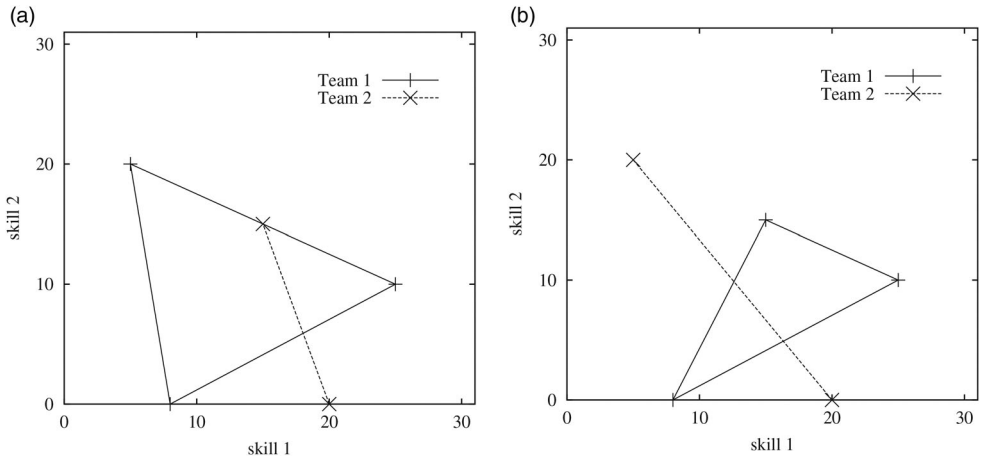

Figure 1. An example to illustrate the difference between the two formulations. (a) Distance-based grouping (MDGP) and (b) variance-based grouping (VAR).

definition, in the example given in Figure 1 the distance formulation induces a solution in which there is a 2-dominance of Team 1 over Team 2. The solution given by the variance formulation does not have dominated teams.

3. Related work

The problem of teaming students in a way that promotes a diverse environment has been extensively studied. Several alternative formulations have been proposed. For example, Mingers and O'Brien (1995) presents a method in which students are characterised by a vector of boolean variables and an information theory measure is used to measure the quality of the solution. An allocation is a good one if the size of the groups and the share of each attribute between them are equitable.

Graf and Bekele (2006) proposed an ant colony optimisation algorithm to maximise the heterogeneity of the formed groups. The students' heterogeneity is measured by the euclidian distance on the vectors representing the students' characteristics and a measure of goodness of heterogeneity based on minimising the absolute difference between the average of the highest and the lowest score and the scores of the remaining students. The authors claim that the approach is able to induce stable solutions. In Wang, Lin, and Sun (2007) it is proposed a computer-supported grouping system named Diana to create heterogeneous groups. Diana was created to induce groups that exhibit internal diversity and external balance with other groups. The similarity is assessed by measuring the distances among identified triangles. In Yeoh and Nor (2011) an approach whose main goal is the simplicity of the algorithm is proposed. The method pursues groups having an equal number of students of a particular race and gender (or some other similar criteria) and the differences in grade point average fall within a user specified tolerance.

In this work we propose an alternative formulation for the MDGP.

Maximum diversity models have been proposed in various contexts such as teams formation, curriculum design, exam scheduling or very-large-scale integration circuit design (Weitz and Lakshminarayanan 1997; Baker and Powell 2002). They involve the selection of elements from a population based on some measure of diversity. Each individual is characterised by a set of attributes that is dependant of the problem context. The attributes may reflect skills, background, experience or demographic information and can be described by numbers or by categories (Huxham and Land 2000; Belbin 2010). It is usual to collect this information through the use of questionnaires (Weitz and Lakshminarayanan 1997).

In the Operations Research literature we can find two different problems that fall in this category: the MDP and the MDGP, being the latter an extension of the former.

The MDP consists in identifying, in a population, a subset of elements that present the most diverse characteristics among themselves according to the attributes characterising them (Kuo, Glover, and Dhir 1993). An example application can be a college in which the admission policies has the objective to achieve a diverse student body in addition to take into account the students scores in the tests. The usual formulation is to measure students' similarity by means of a euclidian distance measure and the diversity of the selected set of elements as the sum of the euclidian distances.

For the context of students team formation in a classroom the usual approach is to formulate the problem as an MDGP. The MDGP can be seen as an extension of the MDP in which the goal is to group the students into a set of disjoint groups (Weitz and Lakshminarayanan 1997). An application is the formation of teams in a class in a way that creates a maximally diverse environment. This problem has been extensively studied and several efficient algorithms have been proposed for the MDGP formulation (Fan et al. 2011; Gallego et al. 2013; Rodriguez et al. 2013).

For both problems it is assumed that a distance function d_{ij}, measuring the inter-element distance, serves as a diversity measure. Different authors proposed alternative distance functions (Adil and Ghosh 2005), being the most common the Euclidean Distance. We let the distance between students i and j be:

$$d_{ij} = \sqrt{\sum_{p=1}^{P} (s_{ip} - s_{jp})^2},$$

where s_{ip} is the score of student i in the skill p with $p \in \{1, \ldots, P\}$.

The MDP can be formulated with an integer programming model as follows, where N is the total number of elements and m the subgroup size:

$$\begin{aligned} \text{Maximise} \quad & \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d_{ij} x_i x_j \\ \text{Subject to} \quad & \sum_{i=1}^{N} x_i = m \\ & x_i = \{0, 1\}, \quad 1 \le i \le N. \end{aligned} \tag{1}$$

The decision variables x_i are binary variables that take the value 1 if element i was selected and 0, otherwise.

The MDGP can be formulated as a quadratic integer programming as follows:

$$\begin{aligned} \text{Maximise} \quad & \sum_{g=1}^{G} \sum_{i=1}^{N-1} \sum_{j>i}^{N} d_{ij} x_{ig} x_{jg} \\ \text{Subject to} \quad & \sum_{g=1}^{G} x_{ig} = 1, \quad 1 \le i \le N \\ & \sum_{i=1}^{N} x_{ig} = G_g, \quad 1 \le g \le G \\ & x_{ig} = \{0, 1\}, \quad 1 \le i \le N, \ 1 \le g \le G, \end{aligned} \tag{2}$$

where N is the total number of students and G is the total number of groups. G_g is the number of students in group g. The decision variables x_{ig} are binary variables such that $x_{ig} = 1$ if student i is in group g and $x_{ig} = 0$, otherwise.

4. The alternative formulation for MDGP

Consider that N is the total number of students, P is the total number of profiles and G is the total number of groups. G_g is the number of students in group g. Consider that a particular assignment of students to groups was performed. The evaluation of a student i, assigned to group g, for each profile p is denoted by x_{pgi}.

We can measure the total variance of this particular solution using the ANOVA decomposition of variance as a sum of squares (SS). The total variance (SS_{total}) has two components, the first component (SS_{B}) measures the variance of the profiles between groups and the second component measures the variance of profiles within groups (SS_{W}), as presented in Equation (3).

$$SS_{\text{total}} = SS_{\text{B}} + SS_{\text{W}}. \tag{3}$$

Equation (3) can also be written as

$$\sum_{p=1}^{P}\sum_{g=1}^{G}\sum_{i=1}^{N}(x_{pgi} - \bar{x}_p)^2 = \sum_{p=1}^{P}\sum_{g=1}^{G} G_g.(\bar{x}_{pg} - \bar{x}_p)^2 + \sum_{p=1}^{P}\sum_{g=1}^{G}\sum_{i=1}^{N}(x_{pgi} - \bar{x}_{pg})^2, \tag{4}$$

where $\bar{x}_p$ is the mean for each profile p,

$$\bar{x}_p = \frac{\sum_{i=1}^{N} x_{ip}}{N}, \quad p = \{1, \ldots, P\} \tag{5}$$

and $\bar{x}_{pg}$ is the mean for each profile p for the students in group g

$$\bar{x}_{pg} = \frac{\sum_{i=1}^{N} x_{pgi}}{G_g}, \quad p = \{1, \ldots, P\}, \quad g = \{1, \ldots, G\}. \tag{6}$$

The total variance (SS_{total}) is constant, therefore, as it is shown by Equation (4) maximising the variance within groups (SS_{W}) is equivalent to minimising the variance between groups. We propose the following formulation:

$$\begin{aligned}
\text{Minimise} \quad & \sum_{p=1}^{P}\sum_{g=1}^{G} G_g * \left(\frac{\sum_{i=1}^{N} x_{pgi} \cdot y_{ig}}{G_g} - \bar{x}_p\right)^2 \\
\text{Subject to} \quad & \sum_{g=1}^{G} y_{ig} = 1, \quad 1 \le i \le N \\
& \sum_{i=1}^{N} y_{ig} = G_g, \quad 1 \le g \le G \\
& y_{ig} = \{0, 1\}, \quad 1 \le i \le N, \ 1 \le g \le G.
\end{aligned} \tag{7}$$

The decision variables y_{ig} are binary variables that take the value 1 if student i is in group g and the value 0, otherwise.

In order to solve this nonlinear problem, we implemented a Tabu Search algorithm.

Tabu Search is a metaheuristic search strategy, based on local search, proposed by Glover (1986). The procedure starts from an initial feasible solution and iteratively improves it by the application of small, local modifications (moves). The search space of a Tabu Search heuristic is composed by the set of all possible solutions that can be considered during the search. In order to avoid local optima, non-improving moves are allowed. The use of memory structures, called tabu lists, prevents cycling back to previous solutions. A move is declared tabu (is prohibited) if it leads to a previously visited solution within a number of iterations. Tabu moves are stored in a tabu list that is maintained throughout the execution of the algorithm. However, tabu lists may lead to stagnation of the search process (premature convergence to local optima) or forbid attractive moves even when there is no danger of cycling.

Aspiration criteria are mechanisms that, when satisfied, can revoke a tabu move. A simple aspiration criterion could be, for example, allowing a (tabu) move if the resulting solution has a better objective function value than that of the current best solution. We used a reactive tabu search procedure (Battiti and Tecchiolli 1994) since the tabu list size (tabu tenure) is not constant during the search. The reactive tabu search keeps a record of all the solutions visited during the search. If a solution is revisited within a specified number of iterations, then tenure is increased by a predetermined factor to diversify the search. Otherwise the tabu tenure is decreased to avoid excessive increase in tenure and to intensify the search. Finally, when the algorithm determines that all possible moves are tabu and none satisfy the aspiration criterion, then the tabu tenure is decreased, with the first solution on the elite list of solutions selected as the new incumbent solution and the tabu memory structure is reset. A simplified pseudo code for the tabu search algorithm is presented in Algorithm 1.

Algorithm 1 Tabu search pseudocode

```
build(S0) Build initial feasible solution;
S ← S0; f(S) ← f(S0); S is the current solution
S* ← S; f* ← f(S); S* is the best known solution
tabulist ← ø; initialise tabu list
while (stopping criterion) not met do
  for s ∈ Neighbourhood(S) do
    if s ∉ tabulist or aspiration(s) = TRUE then
      if f(s) < f(S*) then
        S* ← s
        f* ← f(s)
      end if
    end if
  end for
  update(tabulist, s);
end while
```

For the MDGP, each solution is coded by an array of size N (number of students). Each position of the array corresponds to a student and holds the group to which the student was assigned. For example, suppose that, for a class of 12 students, we want to form 3 groups with 4 students in each group. A possible solution could be, for example, the groups presented in Table 3.

This solution can be represented as array of size 12 (see Table 4), where the first position of the array means that student 1 is in group 2.

The algorithm starts with a random solution in which, each student i is randomly assigned to a group j, ensuring that the size of each group is respected. At each iteration, the neighbourhood of the current solution is defined as the set of solutions that can be obtained from S through the

Table 3. An example of a possible assignment of students to groups.

Groups	Students			
1	2	7	10	11
2	1	3	6	9
3	4	5	8	12

Table 4. A solution S for tabu search algorithm.

2	1	2	3	3	2	1	3	2	1	1	3

Table 5. A neighbour solution of *S*.

1	1	2	3	3	2	2	3	2	1	1	3

application of a small transformation. The neighbourhood of a given solution is composed by the set of solutions obtained by all possible single swaps between two students in different groups. For the example presented above, a feasible move could be exchanging the groups of students 7 and 1 (belonging to Groups 1 and 2 respectively). Only full swaps are allowed or, in other words, a student i belonging to group G_i can move to group G_j if and only if other student moves from group G_j to group G_i. In this way we guarantee that all moves generate feasible solutions, preserving the number of students allowed in each group (see Table 5). The value of the objective function (the average of the variance between groups) is then calculated for each solution in the neighbourhood of *S*.

5. Experimental evaluation

In order to highlight the differences between the variances formulation and the distances formulation for the MDGP we have conducted an extensive experimental evaluation.

5.1. Data sets and experiments setup

Two real data sets containing the students' answers to a questionnaire that is aimed at classifying their type of behaviour when working in teams are used. After filling the questionnaire a student is characterised by a score in each of eight profiles that correspond to their preferred behaviour when integrating a team. The goal is to distribute students in teams in a way that teams are composed by students with a diversity of profiles and that the diversity is evenly spread across the set of teams.

The *mec* data set contains the questionnaire results for 128 students of a mechanical engineering course at FEUP in the academic year of 2009/2010, and (ii) the *inf* data set contains the questionnaire results for 144 students of a Informatics and Computing Engineering course at FEUP in the academic year of 2009/2010.

For each data set the algorithm was run for several team's setups, as given in Table 6. Each row indicates the data set, the experiment id and the size of the teams created. The notation $\binom{N}{m}$ states that N teams of m students were created. For example, in the experiment mec04 the 128 students are organised into 8 teams of 4 students, 8 teams of 8 students and 2 teams of 16 students. As shown in the table the mec data set was evaluated with three setups in which all teams have identical size and two setups in which three different team sizes are considered.

Each setup was run 10 times, each with the two distinct objective functions: (i) maximising the within group variance; (ii) maximising the sum of the distances of group members. In order two compare and evaluate the two formulations we make use of the variance, the distance, the DBi index and the number of dominated solutions.

Table 6. The experiments configuration.

data set	experiment id	configuration
mec	mec01	$\binom{32}{4}$
mec	mec02	$\binom{16}{8}$
mec	mec03	$\binom{8}{16}$
mec	mec04	$\binom{8}{4}$ $\binom{8}{8}$ $\binom{2}{16}$
mec	mec05	$\binom{6}{4}$ $\binom{8}{6}$ $\binom{7}{8}$
inf	inf01	$\binom{8}{4}$ $\binom{8}{8}$ $\binom{3}{16}$
inf	inf02	$\binom{6}{4}$ $\binom{8}{6}$ $\binom{8}{8}$

The Davies–Bouldin index (DBi), (Davies and Bouldin 1979), is a metric commonly used for assessing the compactness of the clusters induced by a clustering algorithm. It is defined as the ratio of the sum of the within-cluster scatter to the between-cluster separation in such a way that the ratio is small if the clusters are compact and far from each other. In the context of a clustering algorithm the DBi index is useful for determining the number of clusters that is more adequate for a given data set.

Herein, the aim is to achieve a configuration that maximises the variability within clusters, thus, we utilise the DBi index in the opposite way. That is, a solution is perceived as better if it has a higher value for the DBi. The motivation for utilising the DBi was to have a metric to assess the quality of the solutions that is neutral relatively to the two formulations being discussed.

5.2. Assessing the groups' diversity

Table 7 gives the average result for 10 runs of each of the setups given in Table 6. Every solution was obtained by the Tabu-search algorithm while optimising the variance (VAR) or optimising the distance (DIST). The results show that the two formulations are searching for distinct solutions since the solutions reached by one formulation are different to the solutions found by the other formulation. That is observed for every individual run of the tests, for example for mex01 the optimum reached when using the VAR criterion corresponds to a global distance of 9116.70 and the optimum reached when using the DIST criterion corresponds to a global distance of 9202.70. That means that the solutions reached are distinct.

With respect to the DBi score analysis the variance formulation reaches solutions that have higher values than the corresponding distance formulation. This is an indication that the variance formulation induces teams composed of members with a higher variability of profiles.

Figure 2 shows the evolution of the DBi score throughout the optimisation process for the two formulations. Each line in the plot corresponds to a distinct run of the algorithm for corresponding setup. The analysis of the plot shows that maximising the variance is not equivalent to maximising the DBi score. In fact, although there is an increasing tendency for the DBi score there are intermediate solutions having values for the DBi that are higher than the final solution. Similar results were obtained for the other configurations. These results confirm that the DBi score is a metric independent of the two formulations.

In addition, in both plots the variance formulation has a DBi score that is superior to the corresponding distance formulation. We note that, while Figure 2(a) corresponds to a setup in which all teams are of identical size Figure 2(b) corresponds to a setup in which three different team sizes are taken into account.

5.3. Dominance analysis

In the context of student team formation one very important criteria is to obtain a set of balanced teams, that is, to avoid concentrating the variety of profiles in some teams at the cost of having other teams composed of similar members. We utilise the concept of dominance as a criterion to indicate the number of teams that present inferior variability relatively to other team. For a given skill, we

Table 7. Results obtained with each of the two formulations for the configurations given in Table 6.

		mec01	mec02	mec03	mec04	mec05	inf01	inf02
VAR	variance	119,173.61	119,401.12	119,439.85	119,361.79	119,351.27	119,370.19	119,325.75
	distance	9116.70	19,832.60	41,095.12	22,637.21	15,910.46	28,835.24	18,205.07
	DBi	20.02	42.33	84.83	33.69	33.60	42.54	32.25
DIST	variance	119,044.08	119,334.32	119,420.37	119,308.81	119,288.10	119,278.88	119,249.31
	distance	9202.70	20,002.46	41,377.65	25,098.27	16,632.70	31,990.66	19,200.63
	DBi	17.33	28.36	46.93	23.40	24.73	23.96	23.48

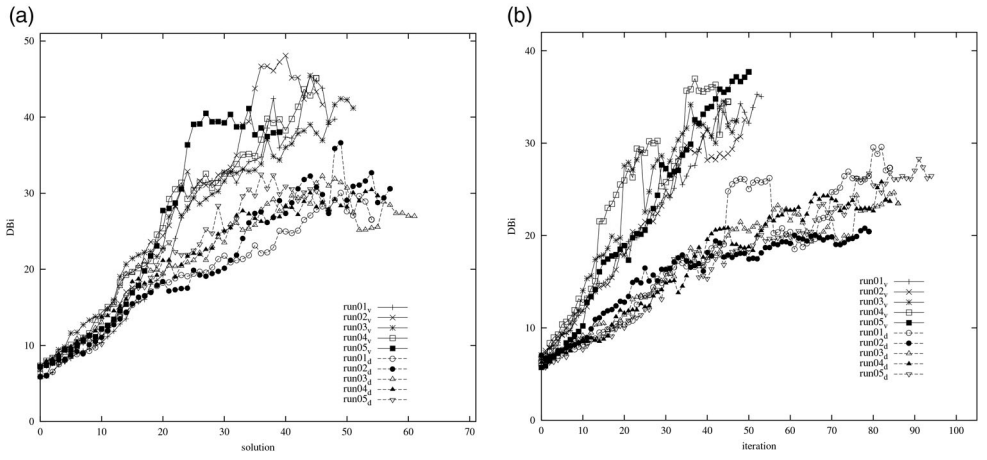

Figure 2. The evolution of the DBi score during the optimisation process for the variance formulation. (a) mec02 configuration and mec04 configuration.

define dominance of a team over another as having the amplitude of the latter bounded by the amplitude of the former. When a team dominates another over n skills we say that there is a n-dominance.

Table 8 gives the number of times a team is dominated by another team over the eight profiles for each setup. It is interesting to note that when the students are to be organised into teams of identical size neither formulation induces solutions in which there are teams dominated by other teams. When the aim is to obtain a set of teams with three different sizes the variance formulation does a better job with respect to the number of dominated teams. We stress that we are counting the number of times each team is dominated by another team. For example, for the run04 of the inf01 setup the aim is to create 19 teams having three different sizes. In this case, the solution obtained by the distance formulation is such that there are 10 dominance occurrences and therefore the solution is not fair according to such criterion. We believe this is an importance drawback of the distance formulation for the particular context of student team building.

5.4. Brief analysis of the TS algorithm

In order to assess if our tabu-search algorithm is able to get good solutions, we have performed and exhaustive search for the solution of a small toy example with 5 teams of 4 students each when using

Table 8. Number of dominated teams for the solutions induced by the two formulations.

Configuration	Formulation	run01	run02	run03	run04	run05	run06	run07	run08	run09	run10
mec01	var	–	–	–	–	–	–	–	–	–	–
	dist	–	–	–	–	–	–	–	–	–	–
mec02	var	–	–	–	–	–	–	–	–	–	–
	dist	–	–	–	–	–	–	–	–	–	–
mec03	var	–	–	–		–	–	–	–	–	–
	dist	–	–	–	–	–	–	–	–	–	–
mec04	var	–	–	1	–	–	–	–	–	–	–
	dist	2	–	2	2	3	5	–	–	–	–
mec05	var	–	–	2	–	–	–	–	–	1	–
	dist	6	4	3	–	–	5	7	–	4	–
inf01	var	–	1	–	–	–	–	–	–	–	–
	dist	–	–	4	10	–	–	–	8	–	–
inf02	var	1	–	–	3	–	–	–	–	–	1
	dist	–	–	–	3	8	–	–	–	15	9

Table 9. Per cent error of tabu search solutions.

Run	% Error
1	0.025
2	0.029
3	0.000
4	0.036
5	0.022
6	0.000
7	0.017
8	0.022
9	0.024
10	0.056

the VAR formulation. The exhaustive search for this solution took several hours and larger examples were unmanageable. Table 9 presents the per cent error of the solution obtained in each run. The tabu-search algorithm was able to obtain the optimal solution in two of the 10 runs (run03 and run06) and the maximum percentage error is 0.056%.

Figure 3 provides a comparison of the variance score and the DBi score for each of the 10 runs. In order to enable a comparison we standardised both scores by subtracting the average and dividing by the standard deviation. The results show some variability on the final scores and confirms the previous conjecture that maximising the variance within groups it is not equivalent to maximising the DBi score. In fact, run04 obtains a value for the DBi score that is much higher than the value obtained for the optimal solution.

6. Conclusions

We have proposed an alternative formulation for the MDGP. Our formulation measures the variance of the attributes characterising students. The goal is to maximise the variance within the groups while minimising the variance between the groups. The formulation of the new approach is given and the results of the experimental evaluation presented. In the experimental evaluation the proposed

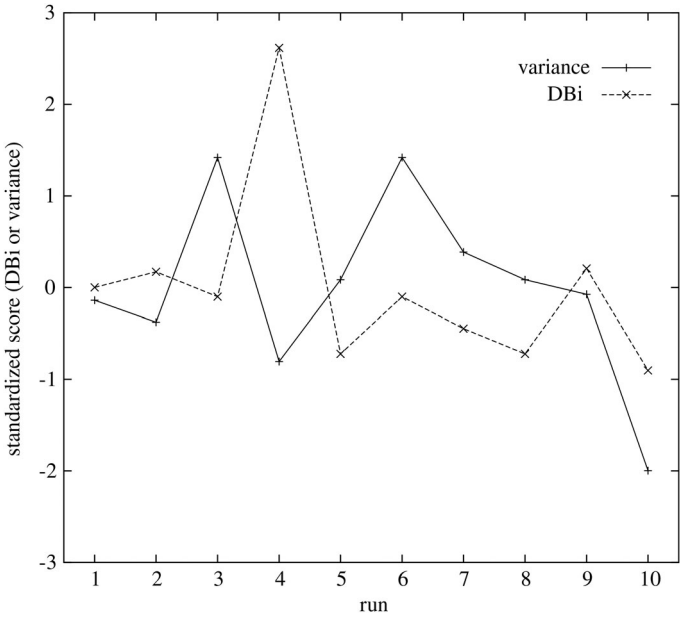

Figure 3. The results of the exhaustive search for a small example.

formulation is compared to the standard MDGP formulation. Two independent measures are used to assess the quality of the resulting teams, the DBi and the number of dominated solutions.

The results show that the formulation aiming at maximising the variance of the attributes within teams is able to achieve better solutions with respect to achieving heterogeneity of profiles within the teams and the homogeneity among teams.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

Teresa Galvão Dias has a background in Mathematics and holds a PhD in Sciences of Engineering (University of Porto, Portugal). She is Assistant Professor in the Faculty of Engineering of University of Porto. She is involved in lecturing Information Systems, Operations Research and Human Computer Interaction courses. She has actively participated in several large national and European R&D projects in areas such as decision support systems, transportation systems and mobility. Her main research interests are combinatorial optimization problems, metaheuristics, human-computer interaction and transportations systems and he regular publishes papers in *International Scientific Journals*.

José Borges received his PhD in Computer Science from University College of London, an MSc in Electrical Engineering and Computer Science and a first degree in Mechanical Engineering, both from University of Porto. He is currently an Assistant Professor and Researcher at the Department of Industrial Engineering and Management. He teaches courses in Statistics, Data Mining, Information Systems and Human Computer Interaction. His research interests include web data mining, data analysis and data science, information visualization and teamwork analysis. He has published 25+ papers in *International Journals*, ISI proceedings and book chapters.

References

Adil, G. K., and J. B. Ghosh. 2005. "Maximum Diversity/Similarity Models with Extension to Part Grouping." *International Transactions in Operational Research* 12 (3): 311–323.

Baker, K., and S. Powell. 2002. "Methods for Assigning Students to Groups: A Study of Alternative Objective Functions." *Journal of the Operational Research Society* 53 (4): 397–404.

Battiti, R., and G. Tecchiolli. 1994. "The Reactive Tabu Search." *ORSA Journal on Computing* 6 (2): 126–140.

Belbin, R. M. 2010. *Team Roles at Work*. 2nd ed. New York: Routledge.

Borges, J., T. Galvão Dias, and J. F. Cunha. 2009. "A New Group-formation Method for Student Projects." *European Journal of Engineering Education* 34 (6): 573–585.

Cohen, E. G. 1994. "Restructuring the Classroom: Conditions for Productive Small Groups." *Review of Educational Research* 64 (1): 1–35.

Davies, D. L., and D. W. Bouldin. 1979. "A Cluster Separation Measure." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1 (2): 224–227.

Fan, Z., Y. Chen, J. Ma, and S. Zeng. 2010. "A Hybrid Genetic Algorithmic Approach to the Maximally Diverse Grouping Problem." *Journal of the Operational Research Society* 62: 1423–1430.

Fan, Z. P., Y. Chen, J. Ma, and S. Zeng. 2011. "A Hybrid Genetic Algorithmic Approach to the Maximally Diverse Grouping Problem." *Journal of the Operational Research Society* 62 (1): 92–99.

Fraser, A., and S. Neville. 1993. *Teambuilding: A Practical Guide*. London: Industrial Society.

Gallego, M., M. Laguna, R. Marti, and A. Duarte. 2013. "Tabu Search with Strategic Oscillation for the Maximally Diverse Grouping Problem." *Journal of the Operational Research Society* 64 (5): 724–734.

Glover, F. 1986. "Future Paths for Integer Programming and Links to Artificial Intelligence." *Computers and Operations Research* 13: 533–549.

Graf, S., and R. Bekele. 2006. "Forming Heterogeneous Groups for Intelligent Collaborative Learning Systems with Ant Colony Optimization." In *Proceedings of the International Conference on Intelligent Tutoring Systems*, 217–226. LNCS 4053.

Huxham, M., and R. Land. 2000. "Assigning Students in Group Work Projects. Can We Do Better than Random?" *Innovations in Education and Training International* 37 (1): 17–22.

Kuo, C.-C., F. Glover, and K. S. Dhir. 1993. "Analyzing and Modeling the Maximum Diversity Problem by Zero-One Programming." *Decision Sciences* 24 (6): 1171–1185.

Martí, R., M. Gallego, and A. Duarte. 2010. "A Branch and Bound Algorithm for the Maximum Diversity Problem." *European Journal of Operational Research* 200 (1): 36–44.

Mingers, J., and F. O'Brien. 1995. "Creating Student Groups with Similar Characteristics: A Heuristic Approach." *Omega, International Journal of Management Science* 23 (3): 313–321.

Page, S. E. 2007. *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies*. New Jersey: Princeton University Press.

Rodriguez, F. J., M. Lozano, C. García-Martínez, and J. D. González-Barrera. 2013. "An Artificial Bee Colony Algorithm for the Maximally Diverse Grouping Problem." *Information Sciences: An International Journal* 230: 183–196.

Wang, D.-Y., S. S. J. Lin, and C.-T. Sun. 2007. "DIANA: A Computer-Supported Heterogeneous Grouping System for Teachers to Conduct Successful Small Learning Groups." *Computers in Human Behavior* 23 (4): 1997–2010.

Weitz, R. R., and S. Lakshminarayanan. 1997. "An Empirical Comparison of Heuristic and Graph Theoretic Methods for Creating Maximally Diverse Groups, VLSI Design, and Exam Scheduling." *Omega* 25 (4): 473–482.

Yeoh, H. K., and M. I. M. Nor. 2011. "An Algorithm to Form Balanced and Diverse Groups of Students." *Computer Applications in Engineering Education* 19 (3): 582–590.