Maximum Diversity Problem. A Multi-Objective Approach

Katherine Vera[1], Fabio Lopez-Pires[2,1], Benjamin Baran[1] and Fernando Sandoya[3]
katherinevera94@gmail.com, fabio.lopez@pti.org.py, bbaran@pol.una.py, fsandoya@espol.edu.ec
[1]National University of Asuncion, San Lorenzo, Paraguay
[2]Itaipu Technological Park, Hernandarias, Paraguay
[3]ESPOL Polytechnic University, Guayaquil, Ecuador

Abstract

The Maximum Diversity (MD) problem is the process of selecting a subset of elements where the diversity among selected elements is maximized. Several diversity measures were already studied in the literature, optimizing the problem considered in a pure mono-objective approach. This work presents for the first time multi-objective approaches for the MD problem, considering the simultaneous optimization of the following five diversity measures: (i) Max-Sum, (ii) Max-Min, (iii) Max-MinSum, (iv) Min-Diff and (v) Min-P-center. Two different optimization models are proposed: (i) Multi-Objective Maximum Diversity (MMD) model, where the number of elements to be selected is defined a-priori, and (ii) Multi-Objective Maximum Average Diversity (MMAD) model, where the number of elements to be selected is also a decision variable. To solve the formulated problems, a Multi-Objective Evolutionary Algorithm (MOEA) is presented. Experimental results demonstrate that the proposed MOEA found good quality solutions, i.e. between 98.85% and 100% of the optimal Pareto front when considering the hypervolume for comparison purposes.

Keywords: Multi-Objective Maximum Diversity, Multi-Objective Maximum Average Diversity, Multi-Objective Evolutionary Algorithm.

1 Introduction

In multiple contexts, the process of selecting objects, ideas, people, projects or resources is an activity frequently performed by individuals, companies or governments and in many cases it is required that the selected elements have different characteristics, so they can represent diversity. The quality of the selection is based on one or more criteria: economical, spatial, affective, political, among others. For example, people must select the methods of transportation and the routes they will use to reach their destination according to price, travel time and comfort. At the end of the month, a worker will have to select which goods or services are the right ones to spend his salary on, in order to satisfy his needs. Another case is vacation planning; a person will have to choose between a given number of destinations to visit according to price and the pleasure they will provide (1). An investor will seek to distribute his or her capital among different investments in such a way as to diversify the risk but to maximize the expected return, and therefore will try not to have much money invested in similar activities that may be affected by an economical crisis that impacts operations across a sector (2).

The growing interest in dealing with diversity in recent years has motivated efforts to study the management of equity, i.e., organizations are becoming increasingly interested in ensuring an equitable treatment of individuals or institutions (3). In fact the Operational Research area has studied several cases, which are summarized as follows:

In logistics, a key problem is locating logistic units that are mutually competitive, such as warehouses, plants or cargo transportation centers, so that their activities are not redundant between logistic units (4). An additional problem is the selection of sites in a number of different locations for facilities that represent some level of danger to the population, such as weapons and explosives stores, prisons or garbage dumps, while also providing an optimal location in terms of location (5).

 https://doi.org/10.19153/cleiej.21.2.2

Another case could be the problem of jury panel composition, which is the list of people from which jury members can be selected for a particular trial (6). Efficiently searching information on the Internet could be another relevant problem, so that the result should be a group of sites with diverse information (7).

In the design of new drugs, through the process known as *High Throughput Screening*, companies have a large library of molecules (in the millions) that they use repeatedly for each new project in order to identify successes, while the selection of a diverse subset would ensure that the cost of this process does not become excessive. In this context, Meinl proposed in (8) some diversity measures for this case.

The problem of selecting a work team to undertake a project, could also be considered as a diversity problem, where it is usually desirable to choose members according to their aptitudes, abilities to work in teams, ages and other factors that seem relevant. In the analysis of a social group, one might be interested in studying its diversity by means of a detailed study of some representative individuals; these selected individuals should have diverse characteristics which must be observed across the whole group, while also being the best for the objective of the team.

Recent studies, such as that of Castillo et al. (9), have determined that diversity in a group of people increases the ability of these groups to solve problems, and therefore contributes to obtaining more efficient working groups in companies, schools, government, etc. These authors provide a theoretical justification for this empirically known phenomenon. Pioneering researchers in this area such as Page (10) state that: *Various perspectives and tools allow groups of people to find more and better solutions and contribute to total productivity.* As a result, the problem of identifying diverse groups of people becomes a key point in large companies and other institutions.

In all the previously mentioned cases, it is a question of choosing the best subset of elements from a large set of possibilities. In many cases it is desirable to establish that the selected elements are not alike, but rather retain sufficiently different characteristics to represent the diversity existing in the original set.

Maximizing the diversity selection, i.e. finding the most diverse subset is known to be an NP-hard decision problem (11). This problem is commonly known as a *Maximum Diversity* (MD) problem (12). It is important to consider that simple procedures that seemingly offer efficient solutions lead to poor decisions; only efficient mathematical models can guarantee efficient solutions.

The MD problem presents two main research challenges:

- first, there are many existing distance measures to evaluate different elements;
- second, there are many existing diversity measures to evaluate different subsets of elements.

This work focuses on the second presented challenge studying the most relevant diversity measures in the literature.

The most studied measure is the Max-Sum Diversity Measure (13), in which the sum of distances between the selected elements is maximized. Heuristics (14) and meta-heuristics (15) have been proposed considering this diversity measure.

The Max-Min Diversity Measure, in which the minimum distance between the selected elements is maximized, has been well documented in recent studies (16). The specialized literature also includes the Max-MinSum Diversity Measure (17), in which the smallest total distance associated with each selected element is maximized. In the Minimum Differential (Min-Diff) Diversity Measure (17), the difference between the maximum sum and the minimum sum of the distances to the other selected elements is minimized. The Min-P-center Diversity Measure (2), in which the greatest of the minimum distances is minimized was also presented.

Prokopyev et al. introduced in (17) the Max-Mean Diversity Measure that consists in maximizing the average distance between the selected elements. Martí et. al. (18) proposed a GRASP algorithm with a Path Relinking in which the local search is based on the *Variable Neighborhood* methodology for this diversity measure.

If only one diversity measure is optimized, a classical mono-objective approach is considered. However, different diversity measures may be studied as conflicting objective functions; consequently, it could be desirable to study the MD problem considering a multi-objective optimization approach, simultaneously optimizing the following diversity measures: (i) Max-Sum, (ii) Max-Min, (iii) Max-MinSum, (iv) Min-Diff and (v) Min P-Center Diversity Measure.

The main contributions of this work are two multi-objective optimization models:

1. *Multi-Objective Maximum Diversity (MMD)* model, where the number of elements to be selected is defined a-priori.
2. *Multi-Objective Maximum Average Diversity (MMAD)* model, where the number of elements to be selected is a decision variable and the target is to optimize the average of each diversity model above presented.

 https://doi.org/10.19153/cleiej.21.2.2

The remainder of this work is structured as follows: Section II describes preliminary concepts while Section III presents multi-objective optimization problems. Section IV describes the proposed multi-objective problem formulation while Section V presents the proposed algorithm. Section VI summarizes experimental results while conclusions and future works are left to Section VII.

2 Preliminary Concepts

The following sub-sections present relevant concepts related to the main contributions of this work.

- The set of elements from which a subset should be chosen is denoted as $N = \{n_i\}$ and its cardinality is denoted as $|N|$.
- A selected subset is denoted as $M = \{m_i\}$, and its cardinality is denoted as $|M|$.
- The distance between each element n_i to every element n_j is denoted as $d_{i,j}$.

In the Operations Research literature, the classical mono-objective MD problem can be formulated as a maximization problem:

$$Max\ f(M) = div(M) \tag{1}$$

where: $div(M)$ is a chosen diversity measure that could be any of the five alternative measures above mentioned, to be presented in the following sub-sections. These diversity measures require different ways of measuring distance between elements, according to the specific context as described next.

2.1 Distance Measure 1: Euclidean Distance

To measure the diversity of a subset M given by a diversity measure $div(M)$, it is required to first have a clear definition of the connection, distance, or dissimilarity between each pair $(m_i, m_j) \in M$. The estimation of this distance depends on the particular problem that is being studied. In complex systems like social groups, a fundamental operation is the assessment of the dissimilarity between each individual pair. In the majority of applications, each element is supposed to be representable by a collection of attributes (1).

Let $x_{i,k}$ be the k^{th} attribute of element m_i, where $k = 1, ..., K$. Then the distance between elements m_i and m_j may be defined as:

$$d_{i,j} = \sqrt{\sum_{k=1}^{K}(x_{i,k} - x_{j,k})^2} \tag{2}$$

In this case $d_{i,j}$ is the Euclidean distance (3) between m_i and m_j, in which larger values indicate greater dissimilarity, but other distance measures can also be considered.

2.2 Distance Measure 2: Cosine Distance

Another common distance measure in the context of dissimilarity is the cosine distance, formulated as:

$$d_{i,j} = \frac{\sum_{k=1}^{K} x_{i,k} \times x_{j,k}}{\sqrt{\sum_{k=1}^{K} x_{i,k}^2} \times \sqrt{\sum_{k=1}^{K} x_{j,k}^2}} \tag{3}$$

The cosine distance between two elements can be viewed as the angle between their attribute vectors, where a large angle between elements indicates a large degree of dissimilarity. It takes values in [-1, 1] reflecting the affinity between the individuals, as presented in (3).

Several distance measures were already studied in the literature. Sung-Hyuk Cha presents in (19) 45 distance measures definitions. However, this work will not consider the origin of the considered distance, it will simply take for granted that a distance measure between each pair of elements exist.

Once the problem of estimating the distance between pairs of elements is determined, the following step is establishing the problem of measuring diversity.

 https://doi.org/10.19153/cleiej.21.2.2

2.3 Diversity Measure 1: Max-Sum

The diversity of a given subset is represented as the maximum sum of the distances between the selected elements (18). The Max-Sum diversity measure can be formulated as:

$$Max \;\; div_1(M) = \sum_{i<j, m_i, m_j \in M} d_{i,j} \tag{4}$$

This measure is used in the problem of locating logistics units that are mutually competitive, such as warehouses, plants or cargo transportation centers, in such a way that their activities are not redundant between different logistic units (4). It is also used in the problem of jury panel composition, which is the list of people from which jury members can be selected for a particular trial (6).

2.4 Diversity Measure 2: Max-Min

The diversity of a given subset is represented as the minimum distance between the selected elements, that have to be maximized (18). The Max-Min diversity measure can be formulated as:

$$Max \;\; div_2(M) = \min_{i<j, m_i, m_j \in M} d_{i,j} \tag{5}$$

This measure is used for example in the location of dangerous, contaminating, undesirable, or strategic installations, where it is important to avoid the vulnerability, hazards or accidents, caused by their proximity to each other. In those examples, the diversity measure that is defined is that of the shortest distance that exists between the selected elements (5).

2.5 Diversity Measure 3: Max-MinSum

It consists in selecting a set $M \subset N$ of $|M|$ elements such that the smallest total distance associated with each selected element m_i is maximized (3). The Max-MinSum diversity measure can be formulated as:

$$Max \;\; div_3(M) = \min_{m_i \in M} \sum_{m_j \in M, j \neq i} d_{i,j} \tag{6}$$

This measure is used in equity problems, for example in the context of facility location problems, where the fairness among candidate facility locations is as relevant as the diversity of the selected locations. These kind of problems also have applications in the context of urban public facility location or the selection of homogeneous groups (20).

2.6 Diversity Measure 4: Minimum Differential (Min-Diff)

It consists in minimizing the difference between the maximum sum and the minimum sum of the distances to the other selected elements. It is known to be strongly NP-hard, and it remains NP-hard even if sign restrictions for distances between elements are imposed (18). The Min-Diff diversity measure can be formulated as:

$$Min \;\; div_4(M) = \max_{m_i \in M} \sum_{m_j \in M, j \neq i} d_{i,j} - \min_{m_i \in M} \sum_{m_j \in M, j \neq i} d_{i,j} \tag{7}$$

This measure like the Max-MinSum diversity measure is used in equity problems and it has the same applications (20).

2.7 Diversity Measure 5: Min-P-Center

In contrast to the measurements mentioned above, the Min-P-center measure is not calculated with the selected items, but requires the whole original set. Note that in this case it is important to minimize the largest of these minimum distances (8). The Min-P-center diversity measure is formulated as:

$$Min \;\; div_5(M) = \max_{n_i \in N} \left\{ \min_{m_j \in M, j \neq i} d_{i,j} \right\} \tag{8}$$

The evaluation of Equation (8) requires $O(m \times n)$ operations, the assessment of possible diversity is more costly than in previous cases. This measure is useful, for example, to (8):

- locate a school so that the most distant student, walks as little as possible;
- locate ambulances so that the most distant patient is treated as soon as possible.

 https://doi.org/10.19153/cleiej.21.2.2

2.8 Proposed Diversity Model 1: Multi-Objective Maximum Diversity (MMD) Model

The proposed MMD model consists in simultaneously optimizing the previously mentioned diversity measures: (i) Max-Sum, (ii) Max-Min, (iii) Max-MinSum, (iv) Min-Diff and (v) Min-P-center, taking for granted that a distance measure between each pair of element is given and the number of elements to be selected is defined a-priori. This proposed MMD model is later formally defined in Equation (12).

2.9 Proposed Diversity Model 2: Multi-Objective Maximum Average Diversity (MMAD) Model

In the case that the number of selected elements $|M|$ is unknown, all considered diversity measures mentioned before can be also used to find convenient numbers of selected elements $|M|$, by obtaining the average of each diversity measure. The proposed MMAD model can be formulated as:

$$Max \;\; \overline{div_q(M)} = \frac{div_q(M)}{|M|} \tag{9}$$

where: q may identify any of the above mentioned diversity measures, any combination of them or even all of them.

In the literature, the only diversity measure that also optimizes $|M|$ is the Max-Sum ($div_1(M)$), which is also known as Max-Mean diversity measure (18). This work also optimizes the number $|M|$ of selected elements like the Max-Mean for the other diversity measures (Max-Min, Max-MinSum, Min-Diff and Min-P-center). To the best of the authors' knowledge, this has not previously been proposed in the literature. This proposed diversity model is formally defined in Equation (13).

3 Multi-Objective Optimization Problems

A general Multi-Objective Optimization Problem (MOP) includes a set of p decision variables, r objective functions, and s restrictions. Objective functions and restrictions are functions of decision variables. This can be expressed as (21):

$$\begin{aligned} &Optimize: \\ &\qquad y = f(x) = (f_1(x), f_2(x), ..., f_r(x)) \\ &Subject\ to: \\ &\qquad e(x) = (e_1(x), e_2(x), ..., e_s(x)) \geq 0, \\ &where: \\ &\qquad x = (x_1, ..., x_p) \in X \\ &\qquad y = (y_1, ..., y_r) \in Y \end{aligned} \tag{10}$$

X denotes the decision space while the objective space is denoted by Y. Depending on the problem, optimize could mean minimize or maximize. The set of restrictions $e(x) \geq 0$ determines the set of feasible solutions X_f and its corresponding set of objective vectors Y_f. The problem consists in finding x that optimizes $f(x)$. In general, there is no unique best solution but a set of solutions, none of which can be considered better than the others when all objectives are considered at the same time. This derives from the fact that there can be conflicting objectives. Thus, a new concept of optimality should be established for MOPs.

Given two decision vectors u, $v \in X_f$ in a pure minimization context:

$$\begin{aligned} f(u) = f(v) &\implies \forall l \in 1, 2, ..., r : f_l(u) = f_l(v) \\ f(u) \leq f(v) &\implies \forall l \in 1, 2, ..., r : f_l(u) \leq f_l(v) \\ f(u) < f(v) &\implies f(u) \leq f(v) \wedge f(u) \neq f(v) \end{aligned} \tag{11}$$

Thus, the following relations are possible between u and v:

- $u \succ v \longrightarrow (u \text{ dominates } v) \iff f(u) \leq f(v)$
- $v \succ u \longrightarrow (v \text{ dominates } u) \iff f(v) \leq f(u)$
- $u \sim v \longrightarrow u$ and v are not comparable. Neither one dominates the other.

A set of all non-dominated solutions is considered as the optimal Pareto set and it is denoted as: $P_{true} = \{x \in X_f \mid x \text{ is not dominated by other } x \in X_f\}$. The corresponding image of the optimal Pareto set is known as the optimal Pareto front and it is denoted as: $F_{true} = \{y = f(x) \mid x \in P_{true}\}$.

 https://doi.org/10.19153/cleiej.21.2.2

4 Proposed Multi-Objective Maximum Diversity Problem Formulation

This section firstly presents the proposed formulation for the *Multi-Objective Maximum Diversity* (MMD) model. This model considers the simultaneous optimization of the following five objective functions: (i) Max-Sum, (ii) Max-Min, (iii) Max-MinSum, (iv) Min-Diff and (v) Min P-Center diversity measures, that can be generalized as:

$$Optimize\ f(M) = \begin{bmatrix} div_1(M) \\ div_2(M) \\ div_3(M) \\ div_4(M) \\ div_5(M) \end{bmatrix} \tag{12}$$

where: $div_1(M)$, $div_2(M)$ and $div_3(M)$ are maximized while $div_4(M)$ and $div_5(M)$ are minimized.

In Equation (12), each objective function $div_q(M)$, where $q = 1, ..., 5$ represents the diversity measure corresponding to the definition, previously studied in Section II.

In addition, the proposed formulation for the *Multi-Objective Maximum Average Diversity* (MMAD) model considers the average optimization of the five objective functions presented in the MMD model, that can be generalized as:

$$Optimize\ f(M) = \begin{bmatrix} \overline{div_1(M)} \\ \overline{div_2(M)} \\ \overline{div_3(M)} \\ \overline{div_4(M)} \\ \overline{div_5(M)} \end{bmatrix} \tag{13}$$

where: again $\overline{div_1(M)}$, $\overline{div_2(M)}$ and $\overline{div_3(M)}$ are maximized while $\overline{div_4(M)}$ and $\overline{div_5(M)}$ are minimized.

In Equation (13), it is seen that the number $|M|$ of selected elements is also a decision variable that is optimized. In both MMD and MMAD models, several other diversity measures can be easily included, but in this work, only the five presented diversity measures are considered as objective functions.

Example 1:

The following example is presented to better understand the proposed problem formulation. A small problem instance is studied, considering $|N| = 5$, $|M| = 3$ and the distance matrix shown in Table 1.

For this example, there are $\binom{5}{3} = 10$ possible solutions, but for simplicity the five considered diversity measures are only calculated for the following two solutions:

- M_1: [1, 2, 3]
- M_2: [1, 2, 4]

Diversity Measure 1: Max-Sum

- $div_1(M_1) = div_1([1, 2, 3]) = d_{1,2} + d_{1,3} + d_{2,3} = 12.19$
- $div_1(M_2) = div_1([1, 2, 4]) = d_{1,2} + d_{1,4} + d_{2,4} = 12.28$

Table 1: Distance Matrix $d_{i,j}$ between elements (n_i, n_j) in N for Example 1.

	$\mathbf{n_1}$	$\mathbf{n_2}$	$\mathbf{n_3}$	$\mathbf{n_4}$	$\mathbf{n_5}$
$\mathbf{n_1}$	0	4.12	5.83	5.0	7.07
$\mathbf{n_2}$	4.12	0	2.24	3.16	7.81
$\mathbf{n_3}$	5.83	2.24	0	2.24	7.21
$\mathbf{n_4}$	5.0	3.16	2.24	0	5.0
$\mathbf{n_5}$	7.07	7.81	7.21	5.0	0

 https://doi.org/10.19153/cleiej.21.2.2

Considering the above mentioned possible solutions M_1 and M_2, the best solution taking into account the Max-Sum diversity measure is M_2, composed by [1, 2, 4], with 12.28 of diversity.

Diversity Measure 2: Max-Min

- $div_2(M_1) = div_2([1,2,3]) = \min\{d_{1,2}, d_{1,3}, d_{2,3}\} = 2.23$
- $div_2(M_2) = div_2([1,2,4]) = \min\{d_{1,2}, d_{1,3}, d_{2,3}\} = 3.16$

Considering the above mentioned possible solutions M_1 and M_2, the best solution taking into account the Max-Min would be the maximum, which is again M_2, composed by [1, 2, 4], with 3.16 of diversity.

Diversity Measure 3: Max-MinSum

- $div_3(M_1) = div_3([1,2,3]) = \min\{s_1, s_2, s_3\} = 6.36$
  where:
  - $s_1 = d_{1,2} + d_{1,3} = 9.95$
  - $s_2 = d_{1,2} + d_{2,3} = 6.36$
  - $s_3 = d_{1,3} + d_{2,3} = 8.07$
- $div_3(M_2) = div_3([1,2,4]) = \min\{s_1, s_2, s_4\} = 7.28$
  where:
  - $s_1 = d_{1,2} + d_{1,4} = 9.12$
  - $s_2 = d_{1,2} + d_{2,4} = 7.28$
  - $s_4 = d_{1,4} + d_{2,4} = 8.16$

Considering the above mentioned possible solutions M_1 and M_2, the best solution taking into account the Max-MinSum diversity measure is again M_2, composed by [1, 2, 4], with 7.28 of diversity.

Diversity Measure 4: Min-Diff

- $div_4(M_1) = div_4([1,2,3]) = \max\{s_1, s_2, s_3\} - \min\{s_1, s_2, s_3\} = 9.95 - 6.36 = 3.59$
  where:
  - $s_1 = d_{1,2} + d_{1,3} = 9.95$
  - $s_2 = d_{1,2} + d_{2,3} = 6.36$
  - $s_3 = d_{1,3} + d_{2,3} = 8.07$
- $div_4(M_2) = div_4([1,2,4]) = \max\{s_1, s_2, s_4\} - \min\{s_1, s_2, s_4\} = 9.12 - 7.28 = 1.84$
  where:
  - $s_1 = d_{1,2} + d_{1,4} = 9.12$
  - $s_2 = d_{1,2} + d_{2,4} = 7.28$
  - $s_4 = d_{1,4} + d_{2,4} = 8.16$

Considering the above mentioned possible solutions M_1 and M_2, the best solution taking into account the Min-Diff diversity measure is again M_2, composed by [1, 2, 4], with 1.84 of diversity.

Diversity Measure 5: Min-P-Center

- $div_5(M_1) = div_5([1,2,3]) = \max\{min_1, min_2, min_3, min_4, min_5\} = 7.07$
  where:
  - $min_1 = \min\{d_{1,2}, d_{1,3}\} = 4.12$
  - $min_2 = \min\{d_{2,1}, d_{2,3}\} = 2.24$
  - $min_3 = \min\{d_{3,1}, d_{3,2}\} = 2.24$
  - $min_4 = \min\{d_{4,1}, d_{4,2}, d_{4,3}\} = 2.24$
  - $min_5 = \min\{d_{5,1}, d_{5,2}, d_{5,3}\} = 7.07$

 https://doi.org/10.19153/cleiej.21.2.2

- $div_5(M_2) = div_5([1,2,4]) = \max\{min_1, min_2, min_3, min_4, min_5\} = 5.0$
  where:
  - $min_1 = \min\{d_{1,2}, d_{1,4}\} = 4.12$
  - $min_2 = \min\{d_{2,1}, d_{2,4}\} = 3.16$
  - $min_3 = \min\{d_{3,1}, d_{3,2}, d_{3,4}\} = 2.24$
  - $min_4 = \min\{d_{4,1}, d_{4,2}\} = 3.16$
  - $min_5 = \min\{d_{5,1}, d_{5,2}, d_{5,4}\} = 5.0$

Considering the above mentioned possible solutions M_1 and M_2, the best solution taking into account the Min-P-Center diversity measure is again M_2, composed by [1, 2, 4] with 1.84 of diversity.

In this first example, clearly M_2 is better than M_1 and therefore, it can be seen that M_2 dominates M_1 ($M_2 \succ M_1$) or equivalently, it can be said that M_1 is dominated by M_2 (22).

Example 2:

To demonstrate that the considered diversity measures may result in conflicting objective functions when optimizing the MD problem, and consequently some solutions may be better when considering some diversity measures, but not for others. A second example is presented considering $|N| = 10$, $|M| = 3$ and the distance matrix shown in Table 2.

Table 2: Distance Matrix $d_{i,j}$ between elements (n_i, n_j) in N for Example 2.

	$\mathbf{n_1}$	$\mathbf{n_2}$	$\mathbf{n_3}$	$\mathbf{n_4}$	$\mathbf{n_5}$	$\mathbf{n_6}$	$\mathbf{n_7}$	$\mathbf{n_8}$	$\mathbf{n_9}$	$\mathbf{n_{10}}$
$\mathbf{n_1}$	0	77.83	73.83	73.67	21.76	78.19	49.90	65.00	28.67	66.65
$\mathbf{n_2}$	77.83	0	57.73	73.84	71.28	94.36	71.45	55.59	54.07	62.80
$\mathbf{n_3}$	73.83	57.73	0	63.13	54.40	96.60	46.42	19.87	70.02	53.60
$\mathbf{n_4}$	73.67	73.84	63.13	0	68.17	37.37	32.03	74.87	73.88	11.93
$\mathbf{n_5}$	21.76	71.28	54.40	68.17	0	82.85	39.27	45.34	35.53	59.84
$\mathbf{n_6}$	78.19	94.36	96.60	37.37	82.85	0	56.30	104.17	80.86	44.33
$\mathbf{n_7}$	49.90	71.45	46.42	32.03	39.27	56.30	0	52.94	57.33	26.60
$\mathbf{n_8}$	65.00	55.59	19.87	74.87	45.34	104.17	52.94	0	60.19	64.24
$\mathbf{n_9}$	28.67	54.07	70.02	73.88	35.53	80.86	57.33	60.19	0	64.41
$\mathbf{n_{10}}$	66.65	62.80	53.60	11.93	59.84	44.33	26.60	64.24	64.41	0

For this example, there are $\binom{10}{3} = 120$ possible solutions, but for simplicity the five considered diversity measures are only calculated for the following two solutions:

- M_1: [3, 8, 9]
- M_2: [8, 9, 10]

Calculated results for the two considered solutions taking into account the five diversity measures are summarized as follows:

- $\mathbf{div_1(M_1) = 180.0}$
- $div_1(M_2) = 170.48$
- $div_2(M_1) = 52.94$
- $div_2(M_2) = 52.94$
- $div_3(M_1) = 108.54$
- $\mathbf{div_3(M_2) = 110.28}$
- $div_4(M_1) = 18.52$
- $\mathbf{div_4(M_2) = 7.25}$

 https://doi.org/10.19153/cleiej.21.2.2

- $\mathbf{div_5(M_1) = 56.31}$
- $div_5(M_2) = 57.33$

In this particular case, both have the same result for div_2, however in two out of five considered diversity measures ($div_1(M_1)$ and $div_5(M_1)$, M_1 is better than M_2, while M_2 is better than M_1 in two other objective functions ($div_3(M_2)$ and $div_4(M_2)$. It can be seen in this example that M_1 and M_2 are not comparable ($M_1 \sim M_2$) or equivalently neither M_1 dominates M_2 nor M_2 dominates M_1 (22).

5 Proposed Multi-Objective Evolutionary Algorithm

It has been demonstrated that evolutionary algorithms in general are an efficient way of solving optimization problems. A suitable representation of solutions is needed, as well as an objective function, which is subsequently optimized. This algorithm encodes potential solutions as chromosomes and uses genetic operators such as crossover and mutation as well as fitness functions (for each single objective) (23). This work proposes an efficient *Multi-Objective Evolutionary Algorithm* (MOEA) inspired in the *Non-dominated Sorting Genetic Algorithm II* (NSGA-II), which according to the specialized literature, is currently a reference algorithm even used for comparison with new multi-objective methods (24).

5.1 NSGA-II-based Algorithm

The NSGA-II algorithm proposed by Deb et al. (25) is presented in Algorithm 1 and it has the characteristics described below. As input data, it receives a data file containing:

Algorithm 1: Considered MOEA for the proposed MMD and MMAD models.

Data: $|N|$, $|M|$, matrix $D = \{d_{i,j}\}$
Result: Pareto set approximation P_{known}
initialize set of solutions P_0
P'_0 = repair unfeasible solutions of P_0
update set of solutions P_{known} from P'_0
$t = 0; P_t = P'_0$
while *is not stopping criterion* **do**
  Q_t = selection of solutions from $P_t \cup P_{known}$
  Q'_t = crossover and mutation of solutions of Q_t
  Q''_t = repair unfeasible solutions of Q'_t
  update set of solutions P_{known} from Q''_t
  increment t
  P_t = non-dominated sorting from $P_t \cup Q''_t$
end
return Pareto set approximation P_{known}

- the values of $|N|$ and $|M|$ and
- the distances $d_{i,j}$ between each pair of elements represented by a matrix $D = \{d_{i,j}\}$.

In the algorithm, a solution for the MMD model is represented by an array of $|M|$ elements. However, for the MMAD model, in which the number $|M|$ has to be optimized, the array has size $ceil(|N|/2)$ and if the number of selected elements is less than the size of the array, the remaining elements are completed with the value zero. The algorithm defines some parameters such as: Population Size and Maximum Number of Evaluations for the Stopping Criterion.

In the NSGA-II-based algorithm, first a set of solutions is generated randomly for the population. Then, a solution repair is performed in case the generated solutions are unfeasible because a solution can not contain repeated elements. The repair of unfeasible solutions can be carried out in two steps: first, a feasibility verification process is performed, where the duplicate elements are detected. Then, the process of repairing unfeasible solutions is performed, where the solutions that do not follow the feasibility criteria are repaired, eliminating duplicates and generating random elements not repeated in the analyzed solutions.

As long as the stopping criterion is not reached, the following steps are performed: First, a selection operation is performed. The binary tournament selection is used, which involves "tournaments" among pairs of solutions chosen at random from the population. The winner of each tournament (the one with the best fitness) is selected for crossover (26). Then, a crossover operation is performed. The Simulated Binary

 https://doi.org/10.19153/cleiej.21.2.2

Crossover (SBX) is used, which also takes two solutions and performs a crossover, returning as a result the two obtained children (26). Next, a mutation operation is performed. The Polynomial Mutation is used, which is typically applied to single solutions, modifying them accordingly; this returns the mutated solution (26). All considered genetic operators and configuration parameters for the algorithm are stardard values of the JMETAL framework (27) and also correspond to the test problems considered in (25).

After the crossover and mutation operation, the repair solutions function is performed again to repair unfeasible solutions. Finally, a non-dominated sorting is performed on each front. These results in the first front being completely non-dominant set in the current population and the second front being dominated by the solutions in the first front only and so on. Each solution in each front is assigned rank (fitness) values. Solutions in the first front are given a fitness value of 1 and solutions in the second front are assigned a fitness value of 2 and so on. The last accepted front is sorted according to a crowded-comparison operator, which is calculated for each solution and guides the selection process at the various stages of the algorithm toward a uniformly spread-out Pareto Optimal Front and only the best solutions are selected. The proposed algorithm stops after a stopping criterion (as a maximum number of evaluations is performed).

6 Experimental Evaluation

The following sub-sections summarize the experimental environment as well as the main findings identified in the experiments performed to validate the optimization scheme for the MMD and MMAD models. With the presented experiments, the following issues are demonstrated:

- the conflicting correlation between objective functions;
- the quality of solutions obtained by the proposed algorithm; and
- the scalability and the evolutionary improvement of solutions obtained by the proposed algorithm.

In order to study these issues, an Exhaustive Search Algorithm (ESA) was implemented to compare the quality of solutions obtained by the NSGA-II-based Algorithm to an optimal solution set, considering the Hypervolume performance metric, which is the most accepted comparison metric in the literature (28).

6.1 Experimental Environment

Experiments were performed on an Intel(R) Pentium(R) CPU B960 @ 2.20GHz with 4GB of RAM and Windows 10 Pro 64 bits OS. The proposed MMD and MMAD models were implemented with the JMETAL framework (27), requiring Java 8. The Hypervolume metric was calculated with a Python Project (29).

Problem instances available in the website of the OPTSICOM project (30) were considered for the evaluations. Specifically, it has been used the instances in the folder GKD, which have distance matrices whose values were calculated as Euclidean distances from randomly generated points with coordinates in [0-10].

6.2 Experimental Results

The main goal of the presented experimental evaluation is to get a set of solutions that simultaneously optimize the five diversity measures studied: (i) Max-Sum, (ii) Max-Min, (iii) Max-MinSum, (iv) Min-Diff and (v) Min-P-Center. The presented NSGA-II-based algorithm was executed with the following empirically-chosen parameters:

- population size: 500;
- maximum number of evaluations: 250,000.

6.2.1 Experiment 1: Objective Function Correlation

Correlation between diversity measures can be seen in Table 3, obtained considering the result of the NSGA-II-based algorithm for a problem instance with $|N| = 50$ elements from which $|M| = 5$ had to be selected. This analysis coincides with the correlations presented by Sandoya in (3).

Correlation between Max-Sum and Max-Min is 0.64, which can be considered relatively low. Therefore, this indicates that it is not expected that good solutions for one diversity measure will be necessarily good solutions for the other. There is no significant relation between the Min-Diff and the Min-P-Center solutions and the rest of the diversity measures. Their correlations are very low in the best cases. So, these problems can be considered different enough to the rest of the diversity measures.

 https://doi.org/10.19153/cleiej.21.2.2

Table 3: Correlation coefficients among the results in a problem instance with $|N| = 50$ and $|M| = 5$.

	Max-Sum	**Max-Min**	**Max-MinSum**	**Min-Diff**	**Min-P-Center**
Max-Sum	1	0.64	0.95	-0.024	-0.85
Max-Min	0.64	1	0.65	0.13	-0.76
Max-MinSum	0.95	0.65	1	0,26	-0,89
Min-Diff	-0.024	0.13	0.26	1	-0.27
Min-P-Center	-0.85	-0.76	-0.89	-0.27	1

Table 4: Hypervolume and Elapsed Time for the ESA and NSGA-II for the MMD model.

|N|	**|M|**	**Exhaustive Search**			**NSGA-II**			***PercHypDif***
		Hyperv.	**Time**	***NofSol***	**Hyperv.**	**Time**	***NofSol***	
10	3	1.47×10^{15}	0.199 sec	26	1.47×10^{15}	34.537 sec	21/21	0.0 %
15	3	1.65×10^{15}	0.454 sec	57	1.65×10^{15}	37.164 sec	50/50	0.0 %
25	7	1.08×10^{16}	3.866 min	1706	1.08×10^{16}	2.157 min	285/294	0.128 %
30	6	8.15×10^{15}	4.752 min	1141	8.12×10^{15}	1.341 min	313/347	0.423 %
50	5	4.84×10^{15}	21.757 min	1065	4.81×10^{15}	3.093 min	207/340	0.590 %
100	10	-	-		1.23×10^{16}	30.42 min	458	N/A
125	12	-	-		1.06×10^{16}	17.02 min	375	N/A
150	15	-	-		7.62×10^{16}	45.37 min	458	N/A
500	50	-	-		2.90×10^{16}	28.66 hour	438	N/A

Table 5: Hypervolume and Elapsed Time for the ESA and NSGA-II for the MMAD model.

|N|	**Exhaustive Search**			**NSGA-II**			***PercHypDif***
	Hyperv.	**Time**	***NofSol***	**Hyperv.**	**Time**	***NofSol***	
10	1.26×10^{15}	0.301 sec	208	1.24×10^{15}	32.578 sec	63/63	1.152 %
15	1.47×10^{15}	1.033 sec	1033	1.45×10^{15}	43.268 sec	354/362	1.098 %
25	-	-	-	2.57×10^{15}	19.867 sec	359	N/A
30	-	-	-	2.89×10^{15}	1.429 min	452	N/A
50	-	-	-	3.33×10^{15}	6.869 min	428	N/A
100	-	-	-	3.61×10^{15}	33.95 min	298	N/A
125	-	-	-	2.93×10^{15}	5.412 min	100	N/A
150	-	-	-	9.18×10^{15}	9.754 min	148	N/A
500	-	-	-	2.68×10^{15}	5.06 hour	64	N/A

The largest correlation coefficient is obtained between Max-Sum and Max-MinSum. Although these correlation values seem large enough, empirically it can be found that one diversity measure does not necessarily perform well on another or viceversa. Considering these correlations, it is clear that the objective functions are in conflict, thus the proposed multi-objective approach is useful for getting trade-off solutions for the MD problem. It is important to mention that as the number of evaluations and the population size increase, better solutions are usually provided by the proposed evolutionary algorithm (see Section 6.2.3).

6.2.2 Experiment 2: Quality of Solutions

To check the quality of solutions found by the proposed NSGA-II-based algorithm, an *Exhaustive Search Algorithm* (ESA) was also executed on five different instances randomly chosen from the OPTSICOM (30) website. Experimental results are presented in Tables 4 and 5. For cases where it was not possible to get solutions with the ESA, the acronym N/A is used, meaning that evaluations were not available.

Both Tables 4 and 5 contain the following columns:

Hypervolume: This column shows the hypervolume value for the Exhaustive Search and the proposed NSGA-II-based algorithms respectively.

Time: This column shows the execution time of the algorithms for the evaluated instance as a simple reference. A more sophisticated exact algorithm should be considered to compare execution times of algorithms.

Number of found Solutions (*NofSol:*) In this column the number of solutions found by the ESA and the NSGA-II-based algorithm are shown. The number of solutions that are finally non-dominated is also

 https://doi.org/10.19153/cleiej.21.2.2

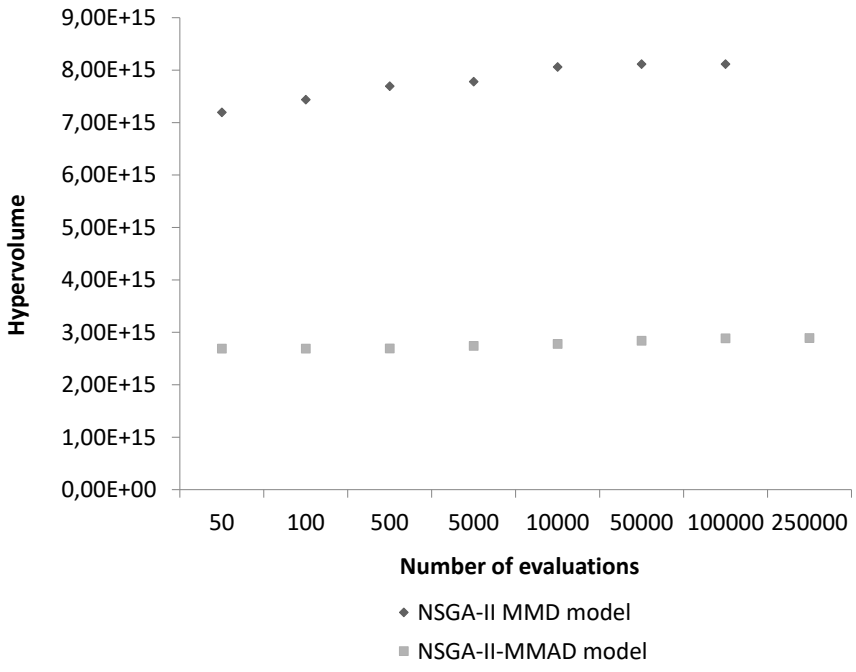

Figure 1: Hypervolume Evolution in each iteration for the MMD and MMAD model with $|N| = 30$ and $|M| = 6$.

specified; calculated by counting how many solutions are among the optimal solutions and how many ones are not. It can be seen that not all the solutions are non-dominated, this occurs because the NSGA-II-based shows all the non-dominated solutions found at the moment it stops.

Percentage of Hypervolume Difference (*PercHypDif*): This column is calculated using the *Exhaustive Search Hypervolume* (HV_{ES}) and the *NSGA-II-based Hypervolume* ($HV_{NSGA-II}$) as follows:

$$PercHypDif = \frac{|HV_{ES} - HV_{NSGA-II}|}{HV_{ES}} \times 100 \tag{14}$$

For the MMD model, the execution of the ESA was not possible for values of $|N| = 100$, 125, 150 and 500. For the other instances (i.e. $|N| = 10$ and 15), it can be seen that the ESA reaches all the solutions in a lower execution time than the proposed NSGA-II-based algorithm but in instances of $|N| = 25$, 30 or 50, it can be seen that the NSGA-II-based algorithm found optimal solutions in a considerable lower time. In both Tables 4 and 5, it can be seen that the proposed algorithm was able to find near-optimal solutions found by the ESA, as shown in the column (*PercHypDif*), where error values were between 0.0% and 1.152%.

6.2.3 Experiment 3: Scalability and Evolutionary Hypervolume Improvement

Considering that for the MMAD model, the ESA no longer ended for values of $|N|$ greater than 15, the scalability of the NSGA-II-based algorithm is demonstrated, taking into account that it found solutions for all the experimental instances and it can be easily verified that as the number of elements increases, the NSGA-II-based algorithm gets solutions faster than the ESA (see Tables 4 and 5).

Additionally to the good quality of solutions obtained by the proposed NSGA-II-based algorithm and its scalability for solving large problem instances, it can be observed that the quality of solutions is improved on each generation (or iteration). As shown in Figures 1 and 2, the hypervolume improves (increases) as the number of generations (or iteration) increases. It also can be seen in Figures 3 and Figure 4 how the hypervolume improves towards optimal values. Consequently, the good quality of solutions for large problem instances may also be guaranteed using according number of generations.

In this work, tables with the obtained results considering all objectives and charts of the Pareto front with only two objectives are presented instead of charts or graphs with all objectives, given that the representation

 https://doi.org/10.19153/cleiej.21.2.2

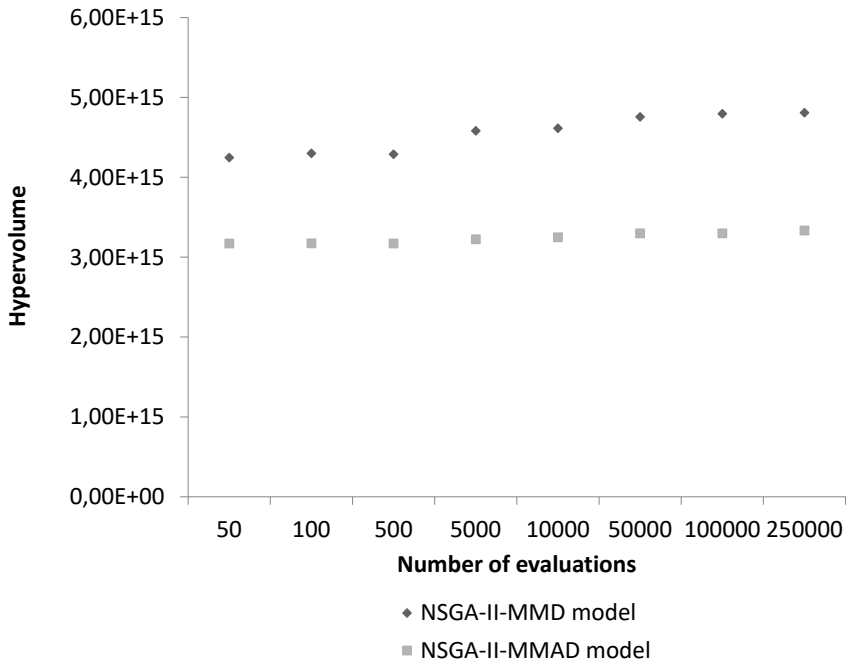

Figure 2: Hypervolume Evolution in each iteration for the MMD and MMAD model with $|N| = 50$ and $|M| = 5$.

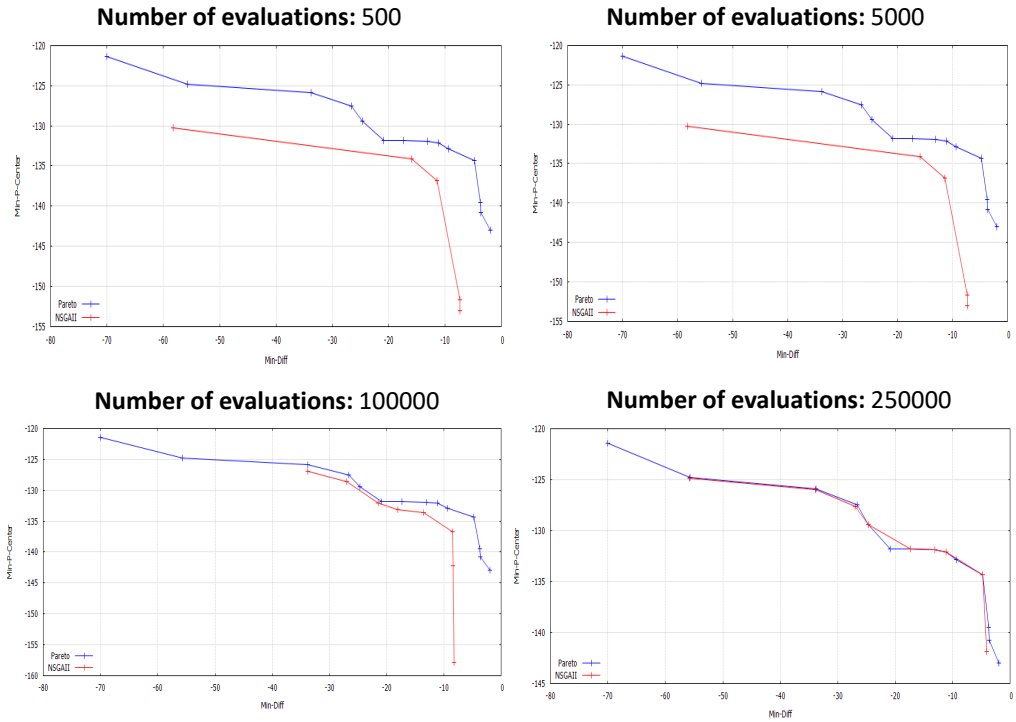

Figure 3: Pareto front vs NSGA-II front in each iteration considering Min-Diff and Min-P-Center with $|N| = 50$ and $|M| = 5$ for the MMD model.

of obtained Pareto fronts considering more than three objectives (five are considered in this work) in a single figure can be very confusing (22).

 https://doi.org/10.19153/cleiej.21.2.2

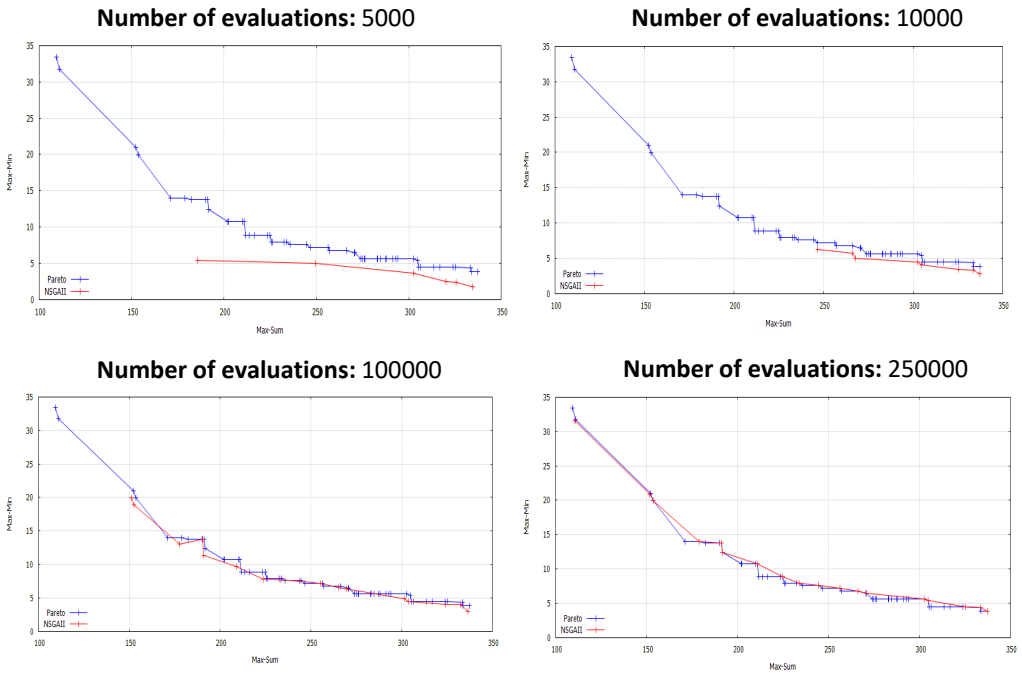

Figure 4: Pareto front vs NSGA-II front in each iteration considering Max-Sum and Max-Min with $|N| = 15$ for the MMAD model.

Figure 3 shows how the Pareto Front obtained by the NSGA-II-based algorithm is approximated at each iteration to the optimal Pareto Front found by the ESA for the Min-Diff and Min-P-Center Diversity Measures for the MMD model. In the same way for the MMAD model, Figure 4 shows how the Pareto Front obtained by the NSGA-II- based is approximated at each iteration to the Optimal Pareto front for the Max-Sum and Max-Min Diversity Measures.

7 Conclusions and Future Works

This work extended concepts of Multi-Objective Maximum Diversity (MMD) models previously presented in (31) and demonstrates its usefulness to obtain the most diverse subset of elements considering five existing diversity measures: (i) Max-Sum, (ii) Max-Min, (iii) Max-MinSum, (iv) Min-Diff and (v) Min-P-center.

Additionally, the Multi-Objective Maximum Average Diversity (MMAD) model was presented, where the optimization of the number $|M|$ of selected elements for each considered diversity measure was studied.

In the literature, only optimization of $|M|$ in the Max-Sum (also known as Max-Mean) is found, but in this work the concept was extended to all five diversity measures and a new multi-objective diversity problem was formulated and then solved.

Considering the experimental evaluations on several problem instances, it was shown that the proposed NSGA-II-based algorithm is able to find good quality solutions for the MMD and MMAD models as well as scaling for solving large problem instances, iteratively improving the hypervolume metric.

Focusing on the first MD problem challenge discussed in Section I, the proposed approach can be extended as a future work to consider several different distance measures between elements, as well as other diversity measures to be optimized.

References

[1] F. Sandoya and R. Aceves, *Grasp and path relinking to solve the problem of selecting efficient work teams.* INTECH Open Access Publisher, 2013, doi:https://doi.org/10.5772/53700.

[2] R. L. Hagin, *Investment management: Portfolio diversification, risk, and timing–fact and fiction.* John Wiley & Sons, 2004, vol. 235, doi:https://doi.org/10.2469/faj.v60.n6.1912.

 https://doi.org/10.19153/cleiej.21.2.2

[3] F. Sandoya, A. Martınez-Gavara, R. Aceves, A. Duarte, and R. Martı, "Diversity and equity models," in *Handbook of Heuristics*. Springer, 2015, pp. 1–20.

[4] F. Glover, C.-C. Kuo, and K. S. Dhir, "Heuristic algorithms for the maximum diversity problem," *Journal of information and Optimization Sciences*, vol. 19, no. 1, pp. 109–132, 1998, doi:https://doi.org/10.1080/02522667.1998.10699366.

[5] E. Erkut and S. Neuman, "Analytical models for locating undesirable facilities," *European Journal of Operational Research*, vol. 40, no. 3, pp. 275–291, 1989, doi:https://doi.org/10.1016/0377-2217(89)90420-7.

[6] M. Lozano, D. Molina, C. Garcı *et al.*, "Iterated greedy for the maximum diversity problem," *European Journal of Operational Research*, vol. 214, no. 1, pp. 31–38, 2011, doi:https://doi.org/10.1016/j.ejor.2011.04.018.

[7] A. Huang, "Similarity measures for text document clustering," in *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*, 2008, pp. 49–56.

[8] T. Meinl, "Maximum-score diversity selection," Ph.D. dissertation, Universitat Konstanz, 2010.

[9] O. Castillo, P. Melin, J. E. Gamez, V. Kreinovich, and O. Kosheleva, "Intelligence techniques are needed to further enhance the advantage of groups with diversity in problem solving," in *Hybrid Intelligent Models and Applications, 2009. HIMA'09. IEEE Workshop on*. IEEE, 2009, pp. 48–55, doi:https://doi.org/10.1109/hima.2009.4937825.

[10] S. E. Page, *The difference: How the power of diversity creates better groups, firms, schools, and societies*. Princeton University Press, 2008.

[11] C.-C. Kuo, F. Glover, and K. S. Dhir, "Analyzing and modeling the maximum diversity problem by zero-one programming," *Decision Sciences*, vol. 24, no. 6, pp. 1171–1185, 1993, doi:https://doi.org/10.1111/j.1540-5915.1993.tb00509.x.

[12] R. Martí, M. Gallego, A. Duarte, and E. G. Pardo, "Heuristics and metaheuristics for the maximum diversity problem," *Journal of Heuristics*, vol. 19, no. 4, pp. 591–615, 2013, doi:https://doi.org/10.1007/s10732-011-9172-4.

[13] J. B. Ghosh, "Computational aspects of the maximum diversity problem," *Operations research letters*, vol. 19, no. 4, pp. 175–181, 1996, doi:https://doi.org/10.1016/0167-6377(96)00025-9.

[14] F. Glover, K. Ching-Chung, and K. S. Dhir, "A discrete optimization model for preserving biological diversity," *Applied mathematical modelling*, vol. 19, no. 11, pp. 696–701, 1995, doi:https://doi.org/10.1016/0307-904x(95)00083-v.

[15] A. Duarte and R. Martí, "Tabu search and grasp for the maximum diversity problem," *European Journal of Operational Research*, vol. 178, no. 1, pp. 71–84, 2007, doi:https://doi.org/10.1007/s10288-007-0033-9.

[16] M. G. Resende, R. Martí, M. Gallego, and A. Duarte, "Grasp and path relinking for the max–min diversity problem," *Computers & Operations Research*, vol. 37, no. 3, pp. 498–508, 2010, doi:https://doi.org/10.1016/j.cor.2008.05.011.

[17] O. A. Prokopyev, N. Kong, and D. L. Martinez-Torres, "The equitable dispersion problem," *European Journal of Operational Research*, vol. 197, no. 1, pp. 59–67, 2009, doi:https://doi.org/10.1016/j.ejor.2008.06.005.

[18] R. Martí and F. Sandoya, "Grasp and path relinking for the equitable dispersion problem," *Computers & Operations Research*, vol. 40, no. 12, pp. 3091–3099, 2013, doi:https://doi.org/10.1016/j.cor.2012.04.005.

[19] S.-H. Cha, "Comprehensive survey on distance/similarity measures between probability density functions," *City*, vol. 1, no. 2, p. 1, 2007.

[20] A. Duarte, J. Sánchez-Oro, M. G. Resende, F. Glover, and R. MARTÍ, "Grasp with exterior path relinking for differential dispersion minimization," *Information Sciences, In press. TABLA V*, 2014, doi:https://doi.org/10.1016/j.cor.2012.04.005.

[21] J. Crichigno and B. Barán, "Multiobjective multicast routing algorithm for traffic engineering," in *Computer Communications and Networks, 2004. ICCCN 2004. Proceedings. 13th International Conference on.* IEEE, 2004, pp. 301–306, doi:https://doi.org/10.1109/icccn.2004.1401652.

[22] C. von Lücken, B. Barán, and C. Brizuela, "A survey on multi-objective evolutionary algorithms for many-objective problems," *Computational Optimization and Applications*, pp. 1–50, 2014, doi:https://doi.org/10.1007/s10589-014-9644-1.

[23] T. Meinl, C. Ostermann, and M. Berthold, "Maximum-score diversity selection for early drug discovery," *Journal of chemical information and modeling*, vol. 51, no. 2, pp. 237–247, 2011, doi:https://doi.org/10.1186/1758-2946-2-s1-p33.

[24] C. A. C. Coello, G. B. Lamont, D. A. Van Veldhuizen *et al.*, *Evolutionary algorithms for solving multi-objective problems.* Springer, 2007, vol. 5, doi:https://doi.org/10.1007/978-0-387-36797-2.

[25] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: Nsga-ii," *IEEE transactions on evolutionary computation*, vol. 6, no. 2, pp. 182–197, 2002, doi:https://doi.org/10.1109/4235.996017.

[26] A. J. Nebro and J. J. Durillo, "jmetal 4.3 user manual," *Available from Computer Science Department of the University of Malaga*, 2013.

[27] A. Nebro and J. Durillo. (2014) Github - jmetal source. [Online]. Available: https://github.com/jMetal/jMetal

[28] N. Riquelme, C. Von Lücken, and B. Baran, "Performance metrics in multi-objective optimization," in *Latin American Computing Conference (CLEI), 2015.* IEEE, 2015, pp. 1–11, doi:https://doi.org/10.1109/clei.2015.7360024.

[29] S. Wessing. (2016) Hypervolume implementation. [Online]. Available: https://ls11-www.cs.uni-dortmund.de/rudolph/hypervolume/start

[30] O. Project. (2012) Maximum diversity problem. [Online]. Available: http://www.optsicom.es/mdp/

[31] K. Vera, F. Lopez-Pires, B. Baran, and F. Sandoya, "Multi-objective maximum diversity problem," in *XLIII Computer Conference Latin American (CLEI), 2017.* IEEE, 2017, pp. 1–9.