Eliminating Disparate Impact in MCDM: The case of TOPSIS

Sandro Radovanović
University of Belgrade
Faculty of Organizational Sciences
Jove Ilića 154, Belgrade, Serbia
sandro.radovanovic@fon.bg.ac.rs

Andrija Petrović
Singidunum University
Danijelova 32, Beograd, Serbia
apetrovic@singidunum.ac.rs

Boris Delibašić, Milija Suknović
University of Belgrade
Faculty of Organizational Sciences
Jove Ilića 154, Belgrade, Serbia
{ boris.delibasic, milija.suknovic}@fon.bg.ac.rs

Abstract. *In today's business, decision-making is heavily dependent on algorithms. Algorithms may originate from operational research, machine learning, but also decision theory. Regardless of their origin, the decision-maker may create unwanted disparities regarding race, gender, or religion. These disparities may further lead to legal consequences. To mitigate unwanted consequences one must adjust either algorithms or decisions. In this paper, we adjust the popular decision-making method TOPSIS to produce utility scores without disparate impact. This is done is by introducing "fairness weight" that is used for the calculation of the utility function of TOPSIS method. Fairness weight should provide the smallest possible intervention needed for a decision without disparate impact. The effectiveness of the proposed solution is shown on the synthetic dataset, as well as on the exemplar dataset regarding criminal justice.*

Keywords. Algorithmic decision-making, TOPSIS, Disparate impact, Fairness in decision-making

1 Introduction

The impact of algorithms on business decision-making is immense in recent years. Managers tend to spend less time preparing the decision process, less effort in modeling decisions, but gaining more trust in algorithms and their decisions. As a result, the usage of algorithms in modern business is increasing and important decisions are being made automatically, without greater supervision of the decision-maker (Grgić-Hlača et al., 2018). One of the reasons for the minor involvement of the decision-maker is the fact that decision-making methods, especially ones that originate from the data mining and machine learning area do not have the power of interpretation and explanation of the decision (Dwork et al., 2020). These algorithms have shown experimentally that they superior to the human processing of data. More specifically, the cost of learning a decision model is lower than hiring an expert, the accuracy is greater than the one of an expert, and the decision-making process is faster (Corbett-Davies et al., 2017).

However, the rise of algorithmic decision-making raises a concern about the impact of algorithms on everyday lives. One can observe the injustices that algorithms have made in recent years. Algorithmic bias is present in many areas, not solely related to business applications, but social applications of algorithmic decision-making as well. One interesting application is Google Ads that is shown to promote higher-paying jobs (from STEM fields) to male individuals rather than female individuals (Lambrecht & Tucker, 2018). Although there is a rational explanation on why male individuals are being promoted (due to historical cultural factors male individuals are earning more compared to female individuals), one must ask a question whether historical injustices should be replicated in algorithmic decision-making? It is shown that not only do algorithms replicate historical bias, but also amplifies it (Barocas & Selbst, 2016). There are even examples of social unrest due to algorithmic decision-making. A crime risk assessment score named Post Conviction Risk Score (PCRA) is developed to help judges (as decision-makers) whether to convict a person or not. However, the benefits of such systems are neglected by an adverse effect. More specifically, African-American offenders have a 13.5 percent greater risk score compared to White-Caucasian offenders (Skeem & Lowenkamp, 2016). Due to many such (historical) biases that exist in the data, and decision-making as well, racial and social unrest are visible today (Szetela, 2020). Therefore, in recent years an increased effort is made to adjust algorithms to correct historical injustices and promote fairness and social welfare (Kasy & Abebe, 2021).

In this paper, we aim to adjust one of the most popular Multi-Criteria Decision-Making (MCDM) method called Technique for Order Preference by Similarity to Ideal Solution (TOPSIS). This technique is applied in many different industries and has shown that it can successfully solve the most complex decision-making problems (Behzadian et al., 2012). Compared to other MCDM algorithms it calculates the cardinal utility of alternatives: In other words, TOPSIS method provides a complete ranking of alternatives. One of the reasons why TOPSIS gained popularity is due to an intuitive explanation of the results. More specifically, the best alternative is the one that is at the same time the closest to the "ideal" solution, and farthest from the "worst" solution. However, even if the decision-making process is well defined, one can create a disparate impact regarding gender or race (Barocas & Selbst, 2016; Lepri et al., 2018).

To mitigate unwanted bias we introduce new variables called "fairness weights". Fairness weights aim to reduce the effect of criteria if that criteria produces unfairness in results. These weights are being applied during the calculation of the distance of alternatives to the ideal and the worst solution. The problems we faced during the creation of the model are the non-convexity of TOPSIS utility scores and the need to satisfy the decision-maker criteria weights. Non-convexity of utility scores is tackled by using logarithmic transformation. Because of that non-convex problem is transformed into a concave problem that can be solved using convex optimization techniques. The need to satisfy the decision-maker criteria weights is fulfilled by the definition of the optimization procedure that constrains fairness while minimizing the change in the decision-maker's original problem setup.

The remainder of the paper is structured as follows. In Section 2 we provide related work needed for the definition of fairness in the decision-making process. In Section 3 we explain the proposed method, and experimental setup as well. In Section 4 we present the results and the discussion of the results. Finally, we conclude the paper in Section 5.

2 Related Work

The related work section consists of two subsections, one explaining fairness in decision-making, and the other explaining the TOPSIS method.

2.1 Fairness in decision-making

In the majority of MCDM methods, the decision-maker aggregates criteria into composite criteria that are regarded as a utility of alternative (Zavadskas et al., 2014). However, the utility function can result in a ranking of alternatives that are deemed as unfair. On the other side, dealing with unfairness is not common in MCDM and motivation can be found in machine learning and economic theory (Hutchinson & Mitchell, 2019).

Many discussions from political philosophy regard fairness as systematic discrimination made by decision-maker based on a race, gender, or religion, or more generally on a personal attribute declared as a sensitive attribute (Dwork et al., 2012). A sensitive attribute is often declared with s and it presents affiliation of an individual with the group. More specifically, an individual can belong to a group or not, i.e. person is of the male gender, or not. Further, since unfair decisions result in different expected outcomes between groups of people one can simplify the groups into two groups, namely privileged group ($s = 0$) and discriminated group ($s = 1$). In other words, unfair decisions result in privileged group individuals get the higher expected utility of a decision-making method than discriminated group individuals.

The reason why tackling unfairness in algorithmic decision-making is the source of unfairness. More specifically, decision-making methods model the decision-maker's beliefs regarding the problem at hand, thus model biases that a decision-maker has. These biases might be intentional (i.e. decision-maker favors male individual for job place), or not. Similarly, data can inherit historical and cultural biases (i.e. females tend to have lower working experience due to maternal leave). Regardless of the source of unfairness, the responsibility of the decision is on the decision-maker. (Köchling & Wehner, 2020)

If we were to measure the level of unfairness, we would measure the disparate impact of the decision-making process. Mathematically, the disparate impact can be calculated using the following formula:

$$DI = \frac{E(u|s=1)}{E(u|s=0)} \tag{1}$$

Where E presents mathematical expectation of the utility score u that an individual obtains from the decision-making method. With an assumption that discriminated group individuals have a lower expected value of getting the desired outcome, the value of DI is bounded to the range [0, 1], where $DI = 0$ present the total unfair decision-making process (every individual from the discriminated group have $u = 0$) and $DI = 1$ present the fair decision-making process. It is worth noticing that the decision-making process can have some level of discrimination. However, this should be explainable either as a random effect (a very small value) or as a necessity of a decision-making process (Grgić-Hlača et al., 2020; Kasy & Abebe, 2021).

2.2 TOPSIS

Algorithmic decision-making is a subject of interest for many decades (Maček et al., 2020). In the decision theory, many methods are developed to model the utility (or preferences) of the decision-maker regarding alternatives. One of the most prominent method for MCDM is TOPSIS. (Abdel-Basset et al., 2020).

TOPSIS method finds the ranking of alternatives through the calculation of the positive ideal solution S_i^+ and negative ideal solution S_i^-, whereby the most suitable alternative is geometrically closest to the positive ideal and farthest from the negative ideal solution. TOPSIS can be performed using the following steps (Çelikbilek & Tüysüz, 2020):

1. Establishing a decision matrix

The decision matrix has the structure as presented in (2).

$$M = \begin{array}{c|cccc} & C_1 & C_2 & \dots & C_n \\ A_1 & x_{11} & x_{12} & \dots & x_{1m} \\ A_2 & x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & \vdots & \vdots & \ddots & \\ A_m & x_{n1} & x_{n2} & & x_{nm} \end{array} \quad (2)$$

where A_i represent an alternative i, C_j criteria j, and x_{ij} value of alternative i for criteria j. In total, the decision matrix consists of m alternatives and n criteria.

2. Normalization of the decision matrix using L_2 norm;

More specifically, each value x_{ij} in the decision matrix is normalized to range of values [0, 1] using formula (3).

$$x_{ij}^N = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{m} x_{ij}^2}} \quad (3)$$

3. Calculation of a weighted normalized decision matrix;

By multiplying each normalized value of the decision matrix with the appropriate weight of the criteria, as presented in (4), one obtains a weighted normalized decision matrix.

$$v_{ij} = \omega_j \times x_{ij}^N \quad (4)$$

4. Calculating the positive ideal and negative ideal solutions;

Once a weighted normalized decision matrix is calculated, one can obtain the best possible and the worst possible value that can be obtained in the decision matrix using (5), (6), (7), and (8). The ideal positive solution IPS is defined as:

$$IPS = \{v_1^+, v_2^+, \dots, v_n^+\} \quad (5)$$

where:

$$v_j^+ = \left(max \wedge min\ v_{ij} \leftrightarrow j = 1, \dots, \text{n}\right) \quad (6)$$

for j representing criteria. More specifically, if a criteria is of benefit type, one selects the largest possible value, while if a criteria is of cost type, one selects the worst possible value.

The ideal negative solution INS is a IPS counterpart defined as:

$$INS = \{v_1^-, v_2^-, \dots, v_n^-\} \quad (7)$$

where:

$$v_j^- = \left(min \wedge max\ v_{ij} \leftrightarrow j = 1, \dots, \text{n}\right) \quad (8)$$

5. Calculating distance of each alternative from ideal positive and ideal negative solutions;

The most suitable distance metric for the TOPSIS method (given the L_2 norm used in step 2) is Euclidean distance. Thus, the distance from the IPS and INS is calculated for each alternative using (9) and (10).

$$D_i^+ = \sqrt{\sum_{j=1}^{n} \left(v_{ij} - v_j^+\right)^2}, i = 1,2, \dots, m \quad (9)$$

$$D_i^- = \sqrt{\sum_{j=1}^{n} \left(v_{ij} - v_j^-\right)^2}, i = 1,2, \dots, m \quad (10)$$

where D_i^+, D_i^- are distances from the positive ideal and negative ideal solution, respectively.

6. Utility calculation for each alternative.

Finally, for each alternative one calculates the closeness coefficient CC that represents the utility score of an individual using (11).

$$CC_i = \frac{D_i^-}{D_i^- + D_i^+} \quad (11)$$

The best possible value of CC is one, that would be a dominant alternative (an alternative is the same as IPS). While the worst possible score is zero, that represents a completely dominated alternative (an alternative is the same as INS).

The goal of this paper is to integrate the disparate impact measure into the TOPSIS method in such a manner that a decision-maker does not change his or her opinion about criteria weights. As a suitable method for such integration, we propose a convex optimization procedure that optimizes for minimum change in distance measures with disparate impact constraint. More specifically, we introduce new variables that can be regarded as “fairness weights”. These weights control the disparate impact made by the decision-maker in the TOPSIS method.

3 Methodology

The methodology section consists of two parts. In the first part, we explain the proposed methodology. Next, we explain the experiments that are conducted.

3.1 Fair TOPSIS method

The inclusion of the disparate impact into decision-making methods requires the development of adequate mathematical modeling. Therefore, a convex mathematical model is proposed.

First, we explain the disparate impact integration. By observing the formula (1), one can notice that the measure is non-convex. Non-convexity poses a problem in optimization procedure since it requires global optimization techniques, such as meta-heuristics, that do not warranty the optimal solution (Gandomi et al., 2013). In that case, one would like to convert (i.e. approximate sufficiently well) the non-convex formulation to a convex one, or even better to a linear one. By using the trick presented in (Zafar et al., 2019; Radovanović et al., 2020) we convert disparate impact into a linear form. More specifically, the disparate impact can be calculated in the linear form using (12).

$$SP = E(u|s = 1) - E(u|s = 0) \tag{12}$$

The measure presented in (12) is known as statistical parity. The mathematical expectation of the utility score can be calculated as in (13).

$$SP = \frac{1}{\sum_{i=1}^{m} s_i} \sum_{i=1}^{m} s_i u_i - \frac{1}{\sum_{i=1}^{m} (1 - s_i)} \sum_{i=1}^{m} (1 - s_i) u_i \tag{13}$$

However, passing through two sums may be inefficient and can be reduced, without the loss of precision, into (14).

$$SP = \sum_{i=1}^{m} (s_i - \bar{s})\, u_i \tag{14}$$

where $\bar{s}$ present the ratio of discriminated individuals in the decision matrix. Since s can take values zero and one, $\bar{s}$ can be calculated as an average value of vector s. An explanation of the (14) is the following. If an alternative belongs to the discriminated group ($s_i = 1$), then it “positively discriminates” and increases the value (since $s_i > \bar{s}$) of statistical parity proportionally by u_i. However, if an alternative belongs to the privileged group ($s_i = 0$), then it “negatively discriminates” and decreases the value (since $s_i < \bar{s}$) of statistical parity proportionally by u_i.

Having in mind that one can efficiently measure the level of discrimination, we can formally define the way we control discrimination. One way we can control for discrimination is by introducing the “fairness weights” that can be used during the calculation of D_i^+ and D_i^-. More specifically, we can present Euclidean distance using linear algebra as presented for D_i^+ in (15).

$$D_i^+ = \sqrt{(v_i - v^+)^{\mathrm{T}} (v_i - v^+)} \tag{15}$$

Then, we can add fairness weights vector w that controls for fairness using (16).

$$D_i^+ = \sqrt{(v_i - v^+)^{\mathrm{T}} (wI) (v_i - v^+)} \tag{16}$$

where I presents a diagonal unit matrix. The same set of weights, as well as the same calculation, is used for D_i^-. More specifically, D_i^- is calculated as in (17).

$$D_i^- = \sqrt{(v_i - v^-)^{\mathrm{T}} (wI) (v_i - v^-)} \tag{17}$$

If we integrate statistical parity (14) and newly defined D_i^+ and D_i^- we obtain statistical parity as presented in (18).

$$SP = \sum_{i=1}^{m} (s_i - \bar{s})\, \frac{D_i^-}{D_i^- + D_i^+} \tag{18}$$

However, this is still non-convex. The first problem we face is that both D_i^+ and D_i^- square roots are making the variable we want to optimize non-linear. More importantly, the utility function that uses w would make the problem non-convex. If we observe formula (18), we can see that the weights vector exists in both numerator and denominator of the closeness coefficient ratio. To mitigate this problem, we can use the logarithm of the closeness coefficient, and the squared distances. Since these distances have non-negative values and it is not expected to have an alternative that is dominated by all others (i.e. an alternative is equal to INS) the logarithm will always be defined. The result is presented in formula (19).

$$SP = \sum_{i=1}^{m} (s_i - \bar{s}) (\log(D_i^{-2}) - \log(D_i^{-2} + D_i^{+2}) \tag{19}$$

Squared distances are used to cancel the square roots while calculating the distances. Since these transformations are monotonic, they will still represent utility scores of alternatives just in different measurement units. Further, the logarithmic transformation makes coefficients in terms of logarithms, more specifically in concave form.

Finally, we can set our mathematical model. Since fairness weights alter the initial weights that the decision-maker expressed, one would like to change them as little as possible to obtain a fair solution. Therefore, our optimization problem can be expressed as in (20).

$$\begin{aligned} &\min \sum_{j=1}^{n} (1_j - w_j)^2 \\ &s.t. \\ &\sum_{i=1}^{m} (s_i - \bar{s}) \left(\log(D_i^{-2}) - \log(D_i^{-2} + D_i^{+2})\right) - t \geq 0 \\ &w_j \geq 0,\ j = 1, \dots, n \\ &w_j \leq 1,\ j = 1, \dots, n \end{aligned} \tag{20}$$

where 1 presents a vector of ones with length n, and t the allowable discrimination. It can be observed that the fairness weights w are bounded between zero and one. Value 0 indicates that a criteria is completely unfair and should be discarded from the decision-making process, while value 1 indicates that a criteria does not make unwanted discrimination between alternatives. By setting the weight to a value lower than 1, we deviate from the beliefs of the decision-maker. Since decision-making should help the decision-maker express his/her beliefs, we propose minimization of expressed beliefs deviation subject to the fairness constraints. The fairness constraint allows for some level of discrimination through the parameter t. The interpretation of the parameter t is that the average expected logarithm of utility adjusted for the imbalance in sensitive attribute s should be at most t.

The decision-maker can, by using this mathematical model, promote positive discrimination in the decision-making process by setting parameter t to be a large positive value. In that case, the average expected utility score of the discriminated group would be higher than the expected utility score of the privileged group. However, one should be careful not to overcompensate expected utility when promoting fairness and positive discrimination. (Binns, 2018; Finocchiaro et al., 2021)

Finally, when fairness weights w are obtained they are applied in the TOPSIS method for the calculation of the D_i^+ and D_i^- as presented in (16) and (17), and further for calculation of the closeness coefficient as presented in (11).

3.2 Experimental Setup

To test the proposed method we experiment on synthetic data, as well as on the exemplar dataset regarding criminal justice.

A synthetic dataset has a small number of alternatives and is used to show how well the proposed methodology works. It consists of six criteria and eight alternatives. The data is presented in Table 1. Above the criteria name, one can notice the orientation of the criteria. Value max indicates benefit criteria, while value min indicates cost criteria. As a criteria weighting scheme, we selected uniform weights (all criteria have the same weight). During the creation of this dataset, criteria C1 and C3 are intentionally created to be unfair. This is done with the idea that proposed methodology identifies those criteria and assigns them a low score.

Table 1. Synthetic dataset

	s	max C1	max C2	min C3	max C4	max C5	min C6
A1	1	6	8	2	1	9	2
A2	1	7	2	7	4	1	3
A3	1	3	5	9	9	5	3
A4	1	1	5	9	1	9	7
A5	0	9	3	3	2	3	6
A6	0	6	7	2	4	2	3
A7	0	5	7	4	9	4	1
A8	0	2	6	6	3	7	3

Another dataset used regards criminal justice software called COMPAS (Dressel & Faried, 2018). This software is used in the US and caused a lot of discussion regarding racial unfairness in the decision-making process. Interested readers are referred to (Washington, 2018). The dataset consists of 361 individuals (more specifically, randomly selected 5% of all individuals) that committed a felony, and the software is used to predict whether an individual is likely to commit another one in the near future. If an individual is ranked higher, then an individual is sent to a jail sentence. It is assumed that software is more likely to rank African-Americans higher compared to White-Caucasians, thus more likely to be sentenced. This is due to historical biases and injustices that are inserted into the data collection process. For the sake of the paper, we selected six criteria. More specifically, age, number of juvenile felony crimes, number of juvenile misdemeanor crimes, number of other juvenile crimes, number of priors felony counts, and charge degree. All of the criteria are benefit criteria, and race on an individual presents a value of the sensitive attribute. In this case, the White-Caucasian race is the privileged group (where a lower utility score is expected), while the other races present the discriminated group (having a higher expected utility score). Weights are equal for all six criteria.

We measure and report average changes in utility scores for both the discriminated and the privileged groups after performing the proposed mathematical model. Also, as a measure of fairness, we report the disparate impact before and after performing the model optimization using the proposed approach. We also discuss the fairness weights obtained from the optimization procedure.

It is worth mentioning that the discrimination parameter t is going to be set on the value to ensure that $DI > 0.75$. This means that average expected logarithmic utility scores adjusted for the imbalance in the counts of the alternatives based on the sensitive attribute should be considered as a random effect (MacCarthy, 2017; Raub, 2018). This value is obtained using grid search of parameter t.

4 Results and Discussion

The results for the synthetic dataset are presented in Table 2. It can be observed that the proposed methodology has not resulted in an increase of utility scores of all alternatives in the discriminated group, nor a decrease of utility scores of all alternatives in the privileged group. Therefore, this cannot be deemed fully as affirmative action, but as promoting equity (Reich, 2021; Finocchiaro et al., 2021).

In both the discriminated and privileged groups, utility scores are increased for two alternatives and decreased for two alternatives. However, the average increase is higher for the discriminated group. More specifically, the average utility score increased from 0.4250 to 0.4497 for the discriminated group (an increase of 0.0247 in the utility score). On the other hand, the average utility score decreased from 0.5925 to 0.5238 for the privileged group (a decrease of 0.0688 in the utility score).

Fairness increased as well. The disparate impact is 0.7173 if the original TOPSIS method is used, while it is 0.8587 after using the proposed method. This indicates that the proposed methodology solves the problem of unwanted discrimination in the decision-making process.

Table 2. Results of the TOPSIS method and the proposed method on the synthetic dataset

	s	TOPSIS	Proposed method	*Δ* utility
A1	1	0.7219	0.6544	- 0.0675
A2	1	0.3836	0.2633	- 0.1204
A3	1	0.3600	0.5189	0.1590
A4	1	0.2347	0.3624	0.1278
A5	0	0.4376	0.1251	- 0.3125
A6	0	0.6229	0.4849	- 0.1380
A7	0	0.8367	0.8738	0.0371
A8	0	0.4731	0.6114	0.1383

By inspecting the fairness weights we can notice that one criteria is considered to be completely unfair. That is criteria C1. A very low score (0.1649) holds for criteria C3 too. These two criteria are intentionally defined to be discriminative. Therefore, the proposed method managed to find unfair criteria and remove their influence in the decision-making process. Other criteria are considered fair, where criteria C6 obtained weight 0.7683, C2 weight 0.9245, while C3 and C4 weight 1.

After testing the proposed method on the synthetic dataset, we experimented on the criminal recidivism dataset, namely the COMPAS dataset. In this dataset, the discriminated group receives a greater expected utility score since a greater utility score presents an undesired outcome. For convenience, we set that $s = 1$ present White-Caucasians, and $s = 0$ other races. Due to the presence of very high values for criteria (outliers) and high independence between criteria (correlations are low), utility scores are very low overall.

If we inspect the results from the original TOPSIS method, we can observe that there is a very high disparate impact. More specifically, other races are twice as likely to get undesired outcomes as the White-Caucasian race with $DI = 0.5398$. Utility scores are very low for both privileged and discriminated groups, and they are $E(u|s = 1) = 0.0149$ and $E(u|s = 0) = 0.0275$. After performing the proposed optimization for the fairness we improved to $DI = 0.7512$. This level of fairness is boundary fair (for some legal documents it would require an additional explanation). Regardless of that, fairness increased by over 20%, which means that intervention in the decision-making method works. Further, $E(u|s = 1) = 0.0160$ and $E(u|s = 0) = 0.0213$ which means that both groups tend to be more equal. White-Caucasian individuals have their utility score increased by 1% on average, while other races lowered their utility score by 6% on average. One interesting finding is that 36.56% of White-Caucasian individuals increased their utility score, while 63.44% reduced the score. That indicates that a minority of the privileged individuals have their score greatly increased (since on average utility score increased). On the other side, other race individuals reduced their utility score in 77.61% of the cases and increased utility score by 22.39%.

Inspection of fairness weights discovers that some criteria are considered unfair. Those are a number of juvenile misdemeanor crimes (fairness weight equal to zero), and a number of priors felony counts (fairness weight equal to 0.0423). These factors are suspected to increase the racial injustices in the US (Abrams et al., 2021; Beckman & Rodriguez, 2021). Being “young and black” makes one more likely to be suspected and reported for crime or felony (Leiber & Johnson, 2008). Due to historical biases that exist in culture, young African-Americans (and other races as well) are more times reported for felonies than White-Caucasians. It is suspected that White-Caucasians are not being reported, thus making these decisions based on these criteria is biased and increases the social injustices. Other criteria are fair in respect of decision-making with fairness weights one for age (the model is not making age discrimination), 0.9467 for a number of juvenile felony counts, 0.9191 for a number of juvenile other felonies, and 0.9609 for charge degree.

Based on two examples, one on a synthetic dataset and another on a real-world dataset we showed that it is possible to introduce fairness into MCDM models, more specifically the TOPSIS method. The model gets an additional level of interpretation with fairness weights. The decision-maker may inspect these weights and correct future decisions to promote equality and equity. Due to the hard fairness constraint, we can ensure that the decision made by the TOPSIS method is fair. Finally, the proposed approach is set as convex optimization, thus gradient-based optimization procedures can find an optimal or near-optimal solution.

However, setting fairness weights to zero might be inappropriate due to omitting criteria in decision-making. More specifically, if criteria has the fairness weight equal to zero then it is not influencing the utility score regardless of the initial weight that the decision-maker provided. During the calculation of the D_i^+ and D_i^-, and further, during the calculation of utility score, these criteria are multiplied by zero. To ensure that criteria influence decision-making, one would add a parameter that expresses the lower bound of the fairness weight. More specifically, one could set $w_j \geq l$, where l is the lowest acceptable fairness weight one criteria can obtain.

Our proposed approach has some limitations. One limitation is in the optimization procedure. Since a hard fairness constraint is used, the optimization procedure might result in an unfeasible solution. In that case, either a grid search of discrimination parameter t should be used, or a solution can be found or approximated using bargaining solutions (Haake & Trockel, 2020).

5 Conclusions

The issue of unfair decisions and injustices made by algorithms forced data scientists and algorithm designers to adjusted methodologies and include some notion of fairness during the decision model creation. This growing field is present in the field of automated decision-making. However, adaptations of the traditional decision-making algorithms are rarely to be found.

This paper introduces a fairness constraint into the TOPSIS method. To integrate this constraint, we transformed a non-linear constraint into a convex form using a linear approximation of the disparate impact measure, and a logarithmic transformation of the closeness coefficient (utility score) with the quadratic distance measures in the TOPSIS method. The fairness constraints introduce new variables called fairness weights that measure how fair a criteria is. Those weights are bounded between zero (criteria is unfair) and one (criteria is fair). Once these weights are found, the calculation of the utility score is adjusted by reducing the impact of unfair criteria. The usefulness of the proposed method is tested both on the synthetic data and (for decision-making methods large) data for criminal justice.

However, there are some unanswered questions in the methodology. Due to the transformation of the TOPSIS utility score, the interpretation of the discrimination parameter t is altered. In future work, we plan to provide a more detailed analysis (both theoretical and experimental) about how to select parameter t to ensure satisfactory or needed disparate impact. In addition, we plan to test the effect and stability of parameter t for various situations. For example, whether the disparate impact remains stable for similar problems (i.e. by performing bootstrap sampling of the same problem). Next, what would happen if alternatives are informed about the fairness adjustments and act adversarially? For that, we must design an experiment using the Stackelberg game (Hu et al., 2019; Tsirtsis, & Gomez-Rodriguez, 2020).

Another question that we plan to answer is the effect of multicollinearity between criteria on fairness. Some approaches mitigate the correlation of criteria in the TOPSIS method by using Mahalanobis distance (Vega et al., 2014). We plan to extend the proposed method with Mahalanobis distance, thus solving two problems at the same time.

Acknowledgments

This paper is the result of the project ONR - N62909-19-1-2008 supported by the Office of Naval Research, the United States: *Aggregating computational algorithms and human decision-making preferences in multi-agent settings*.

References

Abdel-Basset, M., Ding, W., Mohamed, R., & Metawa, N. (2020). An integrated plithogenic MCDM approach for financial performance evaluation of manufacturing industries. *Risk Management, 22*(3), 192-218.

Abrams, L. S., Mizel, M. L., & Barnert, E. S. (2021). The Criminalization of Young Children and Overrepresentation of Black Youth in the Juvenile Justice System. *Race and Social Problems, 13*(1), 73-84.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review, 104*, 671.

Beckman, L., & Rodriguez, N. (2021). Race, Ethnicity, and Official Perceptions in the Juvenile Justice System: Extending the Role of Negative Attributional Stereotypes. *Criminal Justice and Behavior*, 00938548211004672.

Behzadian, M., Otaghsara, S. K., Yazdani, M., & Ignatius, J. (2012). A state-of-the-art survey of TOPSIS applications. *Expert Systems with Applications, 39*(17), 13051-13069.

Binns, R. (2018, January). Fairness in machine learning: Lessons from political philosophy. In *Conference on Fairness, Accountability and Transparency* (pp. 149-159). PMLR.

Çelikbilek, Y., & Tüysüz, F. (2020). An in-depth review of theory of the TOPSIS method: An experimental analysis. *Journal of Management Analytics, 7*(2), 281-300.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017, August). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SigKDD International Conference on Knowledge Discovery and Data Mining* (pp. 797-806).

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science advances, 4*(1), eaao5580.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012, January). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (pp. 214-226).

Dwork, C., Ilvento, C., Rothblum, G. N., & Sur, P. (2020). Abstracting Fairness: Oracles, Metrics, and Interpretability. *arXiv preprint arXiv:2004.01840*.

Finocchiaro, J., Maio, R., Monachou, F., Patro, G. K., Raghavan, M., Stoica, A. A., & Tsirtsis, S. (2021, March). Bridging Machine Learning and Mechanism Design towards Algorithmic Fairness.

In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 489-503).

Gandomi, A. H., Yang, X. S., Talatahari, S., & Alavi, A. H. (2013). Metaheuristic algorithms in modeling and optimization. *Metaheuristic applications in structures and infrastructures*, 1-24.

Grgić-Hlača, N., Redmiles, E. M., Gummadi, K. P., & Weller, A. (2018, April). Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. In *Proceedings of the 2018 World Wide Web Conference* (pp. 903-912).

Haake, C. J., & Trockel, W. (2020). *Introduction to the Special Issue "Bargaining"*. Springer.

Hu, L., Immorlica, N., & Vaughan, J. W. (2019, January). The disparate effects of strategic manipulation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 259-268).

Hutchinson, B., & Mitchell, M. (2019, January). 50 years of test (un)fairness: Lessons for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 49-58).

Kasy, M., & Abebe, R. (2021, March). Fairness, equality, and power in algorithmic decision-making. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 576-586).

Köchling, A., & Wehner, M. C. (2020). Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. *Business Research*, 1-54.

Lambrecht, A., & Tucker, C. (2019). Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of stem career ads. *Management Science, 65*(7), 2966-2981.

Leiber, M. J., & Johnson, J. D. (2008). Being young and black: What are their effects on juvenile justice decision making?. *Crime & Delinquency, 54*(4), 560-581.

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology, 31*(4), 611-627.

MacCarthy, M. (2017). Standards of fairness for disparate impact assessment of big data algorithms. *Cumberland Law Review, 48,* 67.

Maček, D., Magdalenić, I., &Ređep, N. B. (2020). A Systematic Literature Review on the Application of Multicriteria Decision Making Methods for Information Security Risk Assessment. *International Journal of Safety and Security Engineering, 10*(2), 161-174.

Radovanović, S., Petrović, A., Delibašić, B., & Suknović, M. (2020, August). Enforcing fairness in logistic regression algorithm. In *2020 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)* (pp. 1-7). IEEE.

Raub, M. (2018). Bots, bias and big data: artificial intelligence, algorithmic bias and disparate impact liability in hiring practices. *Arkansas Law Review, 71*, 529.

Reich, C. L. (2021). Resolving the Disparate Impact of Uncertainty: Affirmative Action vs. Affirmative Information. *arXiv preprint arXiv:2102.10019*.

Skeem, J. L., & Lowenkamp, C. T. (2016). Risk, race, and recidivism: Predictive bias and disparate impact. *Criminology, 54*(4), 680-712.

Szetela, A. (2020). Black Lives Matter at five: limits and possibilities. *Ethnic and Racial Studies*, *43*(8), 1358-1383.

Tsirtsis, S., & Gomez-Rodriguez, M. (2020). Decisions, Counterfactual Explanations and Strategic Behavior. *arXiv preprint arXiv:2002.04333*.

Vega, A., Aguarón, J., García-Alcaraz, J., & Moreno-Jiménez, J. M. (2014). Notes on dependent attributes in TOPSIS. *Procedia Computer Science, 31*, 308-31

Washington, A. L. (2018). How to argue with an algorithm: Lessons from the COMPAS-ProPublica debate. *Colorado Technology Law Journal, 17,* 131.

Zafar, M. B., Valera, I., Gomez-Rodriguez, M., & Gummadi, K. P. (2019). Fairness Constraints: A Flexible Approach for Fair Classification. *Journal of Machine Learning Research, 20*(75), 1-42.

Zavadskas, E. K., Turskis, Z., & Kildienė, S. (2014). State of art surveys of overviews on MCDM/MADM methods. *Technological and economic development of economy, 20*(1), 165-179.