Explanation Sets: A general framework for Machine Learning Explainability

Rubén R. Fernández[a], Isaac Martín De Diego[a], Javier M. Moguerza[a], F. Herrera[b,c]

[a]*Data Science Laboratory (DSLAB) www.datasciencelab.es, Rey Juan Carlos University, C/ Tulipán, s/n, 28933, Móstoles, Spain*
[b]*Dept of Computer Science and Artificial Intelligence, Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), University of Granada, Granada 18071, Spain*
[c] *Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia*

Abstract

Explainable Machine Learning (*ML*) is an emerging field of Artificial Intelligence that has gained popularity in the last decade. It focuses on explaining *ML* models and their predictions, enabling people to understand the rationale behind them. Counterfactuals and semifactuals are two instances of Explainable *ML* techniques that explain model predictions using other observations. These techniques are based on the comparison between the observation to be explained and another one. In counterfactuals, their prediction is different, and in semifactuals, it is the same. Both techniques have been studied in the Social Sciences and Explainable *ML* communities, and they have different use cases and properties. In this paper, the Explanation Set framework, an approach that unifies counterfactuals and semifactuals, is introduced. Explanation Sets are example-based explanations defined in a neighborhood where most observations satisfy a grouping measure. The neighborhood allows defining and combining restrictions. The grouping measure determines if the explanations are counterfactuals (dissimilarity) or semifactuals (similarity). Besides providing a unified framework, the major strength of the proposal is to extend these explanations to other tasks such as regression by using an appropriate grouping measure. The proposal is validated in a regression and classification task using several neighborhoods and grouping measures.

Keywords: Explainable Machine Learning, Explanation Sets, Counterfactuals, Semifactuals, Example-based explanation

Email addresses: ruben.rodriguez@urjc.es (Rubén R. Fernández), isaac.martin@urjc.es (Isaac Martín De Diego), javier.moguerza@urjc.es (Javier M. Moguerza), herrera@decsai.ugr.es (F. Herrera)

1. Introduction

Machine Learning (*ML*) is becoming increasingly present in our day-to-day activities and decisions. This presence has brought innumerable benefits, such as product experience personalization and reducing the complexity of some tasks (e.g., route planning). However, this ubiquitous presence has downsides, and sometimes it is questionable whether this automatic decision-making is acting in our best interest. Instances of such downsides are biases towards minorities, objective mismatch, and mistrust due to the opacity in *ML* systems [1, 2]. Further, humans are sometimes reticent to adopt *ML* solutions if they are not deemed as interpretable and trustworthy [3, 4]. Another limitation of this opacity arises when these systems are used to model a process whose inner workings are of interest [2]. In some cases, it is possible to understand how a particular decision is made, but it becomes infeasible as the complexity of the systems increases.

Explainable *ML* is an area of *ML* whose aim is to deal with the opacity problem by providing explanations of how *ML* models and their predictions work. The benefits of this area are not limited to end-users, impacting all the users involved in the life cycle of these systems [5]. Data scientists can leverage these techniques to debug *ML* systems when the outcome is not the expected, they want to improve the model or make sure that the model is fair (e.g., not biased), or to enforce explainability constraints in the training step [6, 7]. Regulators can check if the system adheres to a given regulation by inspecting how the model works [8]. Domain experts and analysts can determine if a model is production-ready and look for domain-related problems (e.g., the model does not match the intuition of the domain expert) [7]. Finally, end-users can understand the rationale behind predictions and do not take them as given, in particular, when the consequences are negative [8].

There is no one-size-fits-all solution for explaining *ML* models [9]. Hence, several explanation methods emphasize different aspects of *ML* models. Following the taxonomy presented in [1], these techniques can be categorized based on their scope, origin, and applicability. The scope indicates if a technique explains the whole model (global) or a region of the input space (local). Regarding the origin, it can be intrinsic if the model can be understood directly (e.g., linear regression) or post-hoc when we use a technique to extract an explanation. Finally, the applicability defines if a technique can explain any model (model-agnostic) or a specific model (model-specific).

This paper focuses on a group of post-hoc techniques called example-based explanations. These techniques use observations, previously defined or synthetic, to explain *ML* models and/or their predictions. Counterfactuals and semifactuals are two example-based explanation techniques whose goal is to explain the outcome of an observation of interest (often referred to as factual sample). These explanation techniques are based on the comparison of the factual sample with another observation (or set of observations). In counterfactuals, the outcome of the *ML* model for the factual sample is different and in semifactuals is the same. The goal in both cases is to understand how changes to the factual sample affect the outcome of the model. Other instances of example-based

explanations are prototypes and criticisms [14] and influential observations [15].

Counterfactuals are one of the most widely used Explainable *ML* techniques because they resemble the human-thinking process [1, 10]. In contrast, semifactual-based techniques (although not often directly referred to like that in the *ML* literature) are gaining momentum [11, 12]. Techniques that combine both approaches also exist [11, 13].

Currently, most explainability approaches on counterfactual techniques present the following limitations: 1) they use only one observation, and 2) they are defined in a binary classification context and they do not consider other *ML* tasks like regression or outlier detection. Regarding semifactual approaches, they lack a general definition and they are also defined in a classification context. These limitations motivate the necessity of a general framework that formalizes counterfactuals and semifactuals to more than one observation and different *ML* tasks in terms of set theory and similarity measures.

This paper introduces an example-based explanation framework called Explanation Sets. It encompasses counterfactuals and semifactuals into a single framework. Explanation Sets are based on two restrictions over the observations: a restriction based on the observation attributes (neighborhood) and a restriction based on their prediction (grouping measure). The grouping measure determines if the explanations are semifactuals (similarity) or counterfactuals (dissimilarity). These two concepts enable to extend these explanations to other *ML* tasks and express new explanation preferences and/or restrictions without modifying the extraction procedure. Thus, Explanation Sets are agnostic of the *ML* task and model, although specific task and model aware instances can also be considered.

The proposal is evaluated in two use cases. The first use case is a regression task where several similarities and dissimilarities, and manifold-closeness constraints are considered. An explanation based on Decision Trees that uses both semifactual and counterfactual Explanation Sets is included. The second use case is a classification task. Different neighborhoods are used to enforce actionability and diversity. The explanations without neighborhood restrictions are compared to those with restrictions. In both cases, we study the effects of implementing this framework over previous extraction approaches and how these weaknesses can be overcome.

The remainder of the paper is organized as follows. Section 2 details the relevant previous works to understand the proposal. Section 3 introduces the proposal. Section 4 addresses the use cases to validate the proposal. Finally, Section 5 resumes the main conclusions of this paper and provides further research directions.

2. Related work

2.1. Explainable Machine Learning

Counterfactuals have been thoroughly studied in social sciences, whereas semifactuals have gotten less attention [16]. The representation of these two

explanation methods is similar, but their effect on our thinking is far from similar. Counterfactuals are represented by two statements, the conjecture "if only p, q" (e.g., study, pass the exam) and the presupposed fact "not-p, not-q" (e.g., not study, not pass the exam). In these conjectures, "p" is the predicate and "q" the consequent, which correspond to the observations and outcome, respectively, in the *ML* literature. In the case of semifactuals, the conjecture is "even if p, not-q" (e.g., study, not pass the exam) and the presupposed fact "not-p, not-q" (e.g., not study, not pass the exam).

Regarding their effect, McCloy [16] suggested that inferences about the relation between the antecedent and the consequent may trigger different emotional responses. Counterfactuals amplify this response and semifactuals reduce it. It has also been stated that a combination of counterfactuals (how to avoid an event) and semifactuals (hypothetical situations that would have led to the same event) help to correctly evaluate the causal structure of an event [16, 17]. Nevertheless, as previously mentioned, there is no silver bullet in explanation techniques, and the choice of counterfactuals, semifactuals, both, or other techniques depends on the problem and situation.

Similarly to the social sciences, counterfactuals have gotten more attention than semifactuals in the Explainable *ML* field. This is evidenced by the various terminologies that use semifactual-based techniques such as Anchors [12], factual rules [13], and pertinent positives [11]. Anchors are a rule-based representation that defines a sub-region of the feature space around the factual sample where the prediction does not change. A factual rule classifies the factual sample in a surrogate Decision Tree built in the neighborhood of the factual sample. Finally, pertinent positives state what is minimally sufficient to justify a prediction.

In contrast, most counterfactual based techniques in the *ML* literature use the same definition [18]:

> "Score p was returned because variables $\mathbf{v}$ had values $(v_1, v_2, \ldots)$ associated with them. If $\mathbf{v}$ instead had values $(v'_1, v'_2, \ldots)$, and all other variables had remained constant, score $p\prime$ would have been returned."

In this definition, $\mathbf{v}$, is the observation whose explanation is of interest, and p is the prediction of a *ML* model f, such that $f(\mathbf{v}) = p$. The outcome p is known as fact, which is the event that occurs in the reality of the model f under the parameters $\mathbf{v}$. The foil, $p\prime$, is the event that did not occur and was the expected outcome. The parameters $\mathbf{v}\prime$ represent one of the possible scenarios where the outcome $p\prime$ would have occurred. Notice that this definition could easily be adapted to semifactuals by having $p\prime = p$ and replacing "if" by "even if":

> "Score p was returned because variables $\mathbf{v}$ had values $(v_1, v_2, \ldots)$ associated with them. Even if $\mathbf{v}$ instead had values $(v'_1, v'_2, \ldots)$, and all other variables had remained constant, score p would have been returned."

There are several extensions to counterfactual explanations, but they share the same foundations. Counterfactual sets [19] is a counterfactual-based technique that uses several counterfactuals inside a given neighborhood to explain a prediction rather than a single counterfactual. Counterfactual rules [13] are similar to counterfactual sets, but the representation is constrained to be a rule of a surrogate Decision Tree built in the neighborhood of the factual sample. Other extensions include counterfactuals explained by regression coefficients [20] and minimal adversarial perturbation (the closest counterfactual) [21], among many others.

Another major area of research in counterfactual-based explanations is how to rank counterfactual explanations. The number of possible counterfactuals for a given observation in a *ML* model is usually infinite, but we do not have the capacity to consider all the alternatives [16]. Consequently, it is necessary to provide a measure to rank these explanations. This measure is used to enforce the desirable properties of a counterfactual explanation, such as sparsity, actionability, and data manifold-closeness [22].

The sparsity helps to understand counterfactuals as it promotes counterfactuals with few changes. Instances of distances used to enforce sparsity are $L0$, $L1$, and $L\infty$ Norms [23], the inverse of the median absolute deviation [24], and sGower [25]. Actionability establishes which features should not change (e.g., country of origin) [22] or how they can change (e.g., the age can only increase, not decrease) [23]. Data manifold closeness is a penalization to those counterfactuals that contain an unlikely combination of features. Examples of techniques to enforce this penalization are density estimation [26], autoencoders [11], and using prototypes as guides [27]. There are combinations of actionability and data manifold-closeness such as the penalization of those changes where the path between the factual sample and the counterfactual has a low density [26].

The desirable properties in the selection of counterfactuals have also been extensively studied in the social sciences. Thus, some factors are more mutable than others [16]. Instances of these preferences are voluntary changes over external changes, actions rather than inactions, or the last event in an independently related sequence. The main conclusion from these preferences is that the choice of a rank function is deeply connected with the domain and problem.

Some techniques like factual and counterfactual rules [13], and pertinent positives and negatives [11] combine semifactuals and counterfactuals. Pertinent positives and negatives explanations are stated as follows: "An input $\mathbf{x}$ is classified in class y because features $f_i, \ldots, f_k$ are present and because features $f_m, \ldots, f_p$ are absent.". The presence of the features $f_i, \ldots, f_k$ and the features $f_m, \ldots, f_p$ represent what should be minimally present and critically absent to justify the prediction, respectively. On the other hand, factual and counterfactual rules are presented as individual explanations.

Flip points [28] is a technique that falls between counterfactuals and semifactuals. They are defined as the points that lie on the decision surface (i.e., the positive and negative classes have the same score). The consideration of whether this technique generates counterfactuals or semifactuals is determined

by the implementation of the *ML* models.

2.2. Notation

The following notation will be used throughout this document. Let f be a *ML* model, $f : X \to Y$, that maps observations from the feature space, X, to the output space, Y. *ML* models output space is usually the real space, $Y = \mathbb{R}^p$, or a discrete space that contains the labels of the problem (e.g., $Y = \{-1, 1\}$ in a binary problem). Further, the output space usually contains a single output (i.e., $p = 1$). Regarding the input space, let X be a measurable set. Let μ be a measure of X. The output space is usually defined as the real space, thus $X = \mathbb{R}^n$, where n is the number of features. Let $\hat{\mathbf{x}} \in X$ be the sample whose explanation is of interest. Let $N(\mathbf{x}) \subseteq X$ be a neighborhood containing the instance $\mathbf{x}$. Finally, let $\hat{\mathbf{y}} = f(\hat{\mathbf{x}})$ be the output of the model f for the observation of interest, and let $\mathbf{y}' \in Y$ be an element of the output space.

3. Explanation sets

In this section, we propose the Explanation Sets framework, an approach that unifies counterfactuals and semifactuals. This framework provides the tools for expressing explanation preferences (neighborhood) and it enables the user to define when two outcomes are similar based on its preferences (grouping measure). In Section 3.1, the core concepts of the proposal are introduced. Section 3.2 and 3.3 focus on the grouping measures and neighborhood, respectively. Finally, Section 3.4 presents a taxonomy of the representations of Explanation Sets.

3.1. Concepts

Given two elements of the output space, we can define a surjective mapping:

$$m : Y \times Y \to \{0, 1\} \tag{1}$$

that indicates whether they should be grouped (1) or not (0). Without loss of generality, this mapping will be considered either a similarity or a dissimilarity measure. This mapping will be referred to as grouping measure. The similarity groups two elements if they are similar and the dissimilarity if they are not. On the basis of the definition of the grouping measure, there is a unique way to convert a similarity into a dissimilarity and vice versa using the following conversion (see uniqueness and validity proof in Appendix A):

$$m'(\hat{\mathbf{y}}, \mathbf{y}') = \begin{cases} 1 & if\ m(\hat{\mathbf{y}}, \mathbf{y}') = 0 \\ 0 & if\ m(\hat{\mathbf{y}}, \mathbf{y}') = 1 \end{cases} \tag{2}$$

Definition 1. *(Explanation Set). Given the observation to be explained $\hat{\mathbf{x}}$ and the neighborhood $N(\hat{\mathbf{x}})$, an Explanation Set, $ES_N(\hat{\mathbf{x}})$, is defined as a subset of the neighborhood $N(\hat{\mathbf{x}})$:*

$$ES_N(\hat{\mathbf{x}}) \subseteq N(\hat{\mathbf{x}})$$

Definition 2. *(Explanation Set fidelity). Given a* ML *model f, a grouping measure m, a set $S \subseteq X$, and a measure μ of S, the Explanation Set fidelity, $fidelity_{m,f,\mu}(S)$, is a measure of the proportion of observations that satisfy the grouping measure in the set S:*

$$fidelity_{m,f,\mu}(S) = \frac{\mu(\{\mathbf{x} \in S : m(f(\hat{\mathbf{x}}), f(\mathbf{x})) = 1\})}{\mu(S)}$$

The calculation of the fidelity is often intractable because it involves Explanation Sets with infinite elements (e.g., an Explanation Set represented by feature ranges with continuous variables). In such a case, it can be estimated by sampling observations from the Explanation Set and calculating the fidelity over the resulting set setting μ to be the cardinality of the set.

Definition 3. *(Hard Explanation Set). A Hard Explanation Set is a Explanation Set whose fidelity is 1.*

Explanation Sets are a technique for explaining the outcome of the model f for the observation $\hat{\mathbf{x}}$. They are defined as a subset of the neighborhood $N(\hat{\mathbf{x}})$ that contains observations whose comparison with the sample $\hat{\mathbf{x}}$ might be illustrative to explain the outcome $f(\hat{\mathbf{x}})$. The neighborhood enables to restrict the feature space to a set of observations of interest based on the user preferences. In addition, the observations in the Explanation Set should meet the groping measure m that restricts the observations from the neighborhood to those that are grouped with $\hat{\mathbf{x}}$ under the reality of the model f. The proportion of individuals meeting the restriction is the Explanation Set fidelity (see Definition 2). A higher Explanation Set fidelity should be preferred because it indicates that more observations meet the grouping function and the explanation is faithful to its purpose.

Hard Explanation Sets are a special case of Explanation Sets whose fidelity is 1. This implies that all observations in the Explanation Set are guaranteed to satisfy the grouping measure. A limitation of Hard Explanation sets is that providing this guarantee is complex and expensive when the Explanation Set is infinite depending on the underlying *ML* model. For instance, in Explanation Sets represented by feature intervals with a tree-based ensemble, it could be calculated by modifying the approaches [19] or [29]. However, this calculation might not be practical with models where the decision surface is not axis-parallel. Consequently, the goal is usually to find an Explanation Set (or several Explanation Sets) that satisfies a user-defined fidelity requirement while providing a broad and simple explanation (similarly to Anchor [12]). The task of generating Explanation Sets (also referred to as extraction procedure) consists of finding a subset of a given neighborhood that meets a series of desirable properties (e.g., high fidelity). An approach to generate Explanation Sets based on Anchor and Tree Parzen Estimators is described in Section 4.1.

3.2. Semifactuals and Counterfactuals based on similarity functions

Counterfactual and semifactuals are explanation methods that explain the outcome of a process by comparing a set of variables involved in the process with a hypothetical situation where some of these variables are different [16]. In counterfactual explanations, the outcome of the original scenario is different from the outcome of the hypothetical scenario, and in semifactual explanations, the outcome is the same. In Explainable *ML*, this set of variables represents an observation, and the process is a *ML* model. Regarding the comparison of the outcome, they are usually applied in classification problems. Thus, it is necessary to check if the outcome is different (in counterfactuals), or the same (in semifactuals). The differences and properties of counterfactual and semifactual explanations from previous studies have been explored, from a social sciences viewpoint, in Section 2.

The definition of counterfactuals and semifactuals in terms of the comparison of the classes poses two significant limitations. First, it limits the applicability of these explanation techniques to classification problems. Second, in classification problems, they implicitly use a simple matching similarity (i.e., 1 if the two outcomes are the same and 0 otherwise), but depending on the scenario, other similarities might be more insightful.

Alternative definition 1. *(Counterfactual). Given a similarity measure m and a* ML *model f, a counterfactual for the observation $\hat{\mathbf{x}}$ is an observation $\mathbf{x}$ such that $m(f(\hat{\mathbf{x}}), f(\mathbf{x})) = 0$.*

Proposition 1. *A counterfactual Explanation Set is an Explanation Set whose grouping measure is a dissimilarity.*

Alternative definition 2. *(Semifactual). Given a similarity measure m and a* ML *model f, a semifactual for the observation $\hat{\mathbf{x}}$ is an observation $\mathbf{x}$ such that $m(f(\hat{\mathbf{x}}), f(\mathbf{x})) = 1$.*

Proposition 2. *A semifactual Explanation Set is an Explanation Set whose grouping measure is a similarity.*

The proposed alternative definitions rely on a custom similarity measure instead of the simple matching similarity. In this way, the user can choose the similarity that best fits the problem and observation to be explained. In addition, the counterfactuals and semifactuals can be extended to other *ML* tasks by providing adequate similarity measures. Notice that these definitions use a similarity measure for convenience, but they can also be written in terms of a dissimilarity measure (see the conversion in Eq. 2).

Counterfactuals and semifactuals explanations are based on the comparison of the observation of interest with another observation that represents a hypothetical scenario. Counterfactual and semifactual Explanation Sets are not limited to one hypothetical scenario, and they can enrich the explanation with an arbitrary number of them. The properties of the Explanation Sets determine how the explanations are enriched, and it might provide additional explanation

properties. For instance, Explanation Sets whose elements are enumerated do not have additional explanation benefits other than having more instances to compare. Explanation Sets that can be represented using feature ranges can contain infinite elements while using a simple representation. This representation has been described in an earlier work where counterfactual sets were proposed [19] (a special case of counterfactual Explanation Sets for binary classification), and in Anchor explanations [12], which are a special case of semifactual Explanation Sets. The representation of Explanation Sets is further described in Section 3.4.

Examples. We provide two examples to motivate the benefits of the proposed alternative definitions and the different explainability properties of semifactuals and counterfactuals.

In the first example, we consider a *ML*-based system that estimates the price of a house. This system considers several attributes: location, size, and the number of bedrooms. In this context, the system gives an estimated price of p, and the user might wonder, for example, what would have to change to increase the price to $p + o$. Let $v = p$, then the dissimilarity, $m\_gt_{v,o}$, for this example could be defined as:

$$m\_gt_{v,o}(p, p') = \begin{cases} 0 & : otherwise \\ 1 & : min(p, p') <= v \wedge max(p, p') > v + o \end{cases} \tag{3}$$

where $o > 0$ and v is used to make the dissimilarity symmetric. Notice that if dissimilarity is not required to be symmetric (pseudo-dissimilarity), it could be simplified as follows:

$$m\_gt_{o}(p, p') = \begin{cases} 0 & : otherwise \\ 1 & : p + o < p' \end{cases} \tag{4}$$

As we are interested in changing the outcome, the explanations will be counterfactuals. The set of all counterfactuals will contain instances of houses whose price is greater than $p + k$ (if any). Therefore, these explanations will help to identify, if possible, changes that would increase the price of the house.

In the second example, the same *ML*-based system is considered. In this case, the user is interested in what features related to the house could change while keeping the price in the range $p - k \leq p \leq p + k$. Here, the similarity m_sr_k could be defined as:

$$m\_sr_{k}(p, p') = \begin{cases} 0 & : otherwise \\ 1 & : |p - p'| \leq k \end{cases} \tag{5}$$

As we are interested in keeping the outcome in a range, the explanations will be semifactuals. Thus, the explanations will enable to compare the house with other houses in the same price interval. In this scenario, the explanations might suggest the user to, for example, sell the furniture and appliances of the house because it does not affect the price of the house, and consequently, they will get more money.

Notice that in both cases, we could have used both explanation types because the set of all counterfactuals in the neighborhood characterizes the set of all semifactuals in the neighborhood and vice versa (their union is the neighborhood and the intersection is empty). However, counterfactual and semifactual Explanation Sets are better tailored for the first and second cases, respectively. For instance, in the first case, semifactual Explanation Sets discard the hypothesis instead of providing the actual explanations. While by discarding all the invalid hypotheses only the actual explanations will remain, the procedure is straightforward if we had used a counterfactual Explanation Set in the first case. In the second case, the reasoning is similar.

In the second example, we consider a ML-based system that estimates the probability of getting a certain disease based on the user activities and precautions. In this problem, the user might wonder why the system gives a certain risk estimation of r. The probability is a continuous measure and defining the similarity to be 1 when $r = r'$ where r' is the risk estimation of another observation will result in few similar observations, if any. Thus, we define the similarity, m_risk, to be equals when the two risk estimations are in the range $r - k \leq r' \leq r + k$ similarly to Eq. 5.

In this case, semifactual Explanation Sets will enable to devise activities that we are not currently doing, but we could without reducing or increasing the risk. In addition, they could be used to avoid taking precautions that are not currently affecting the risk estimation. On the other hand, counterfactual Explanation Sets provide with activities and precautions that would change the risk estimation so that we need to be careful about them. Therefore, the use of both counterfactual and semifactual Explanation Set allows providing better insights of the factors that affect the prediction and those which do not.

3.3. Explanation Set Neighborhood based on distances

Explanation Sets are defined in a given neighborhood rather than on the whole feature space. This allows restricting the explanations to a certain region of the feature space, if needed. Instances of such restrictions in the counterfactual literature are actionability, sparsity, data manifold-closeness, and diversity. In addition, restrictions can be used to give a local meaning to the explanations (i.e., consider only instances close under a given distance). These restrictions are mostly applied through ad-hoc approaches which limit and hinder their applicability to other works. We propose to unify these restrictions using neighborhoods facilitating their reusability, applicability, and composability.

In this work, neighborhoods are defined using a distance, d, and a radius, $r \in R$, whose center is the factual sample, $\hat{\mathbf{x}}$, such that $N_{d,r}(\hat{\mathbf{x}}) = \{\mathbf{x} \in X \mid d(\hat{\mathbf{x}}, \mathbf{x}) < r\}$. We propose a simple taxonomy of neighborhoods in Explanation Sets based on the type of distance: **user-defined** and **model-induced**. A user-defined distance can be any distance function. Model-induced metrics are based on the properties of the *ML* model. Examples of model-induced metrics are the kernel in Support Vector Machines (SVMs) and the proximity measure [30] in Random Forest. Model-induced distances have the additional property of

grouping individuals that are similar in the space where the *ML* model projects the observations.

We hypothesize that model-induced metrics are especially useful when debugging and developing *ML* models. In this scenario, Explanation Sets contain individuals similar under the model reality (model-induced metric) that should be grouped based on the user's criteria (mapping m). Ideally, the mapping m in this scenario should group similar outcomes using grouping measures like the similarity in Eq. 5 for continuous outcomes or simple matching for discrete outcomes. Thus, developers can determine if the way the model is grouping instances makes sense in the scenario defined by the mapping m. On the other hand, user-defined distances introduce a bias in the instance selection process. Consequently, user-defined distances might group instances that are not similar under the model perspective. This bias might provide better explanations from a user perspective, but it might hide relevant details in other scenarios like model debugging.

The definition of neighborhoods based on distances makes straightforward the composition of several restrictions by combining the distances of the individual restrictions. For instance, we can define a distance d_n by combining a base distance d_b (e.g., $L1$ norm to enforce sparsity), with terms that enforce plausibility d_p, data manifold-closeness d_c, and promote diversity d_d, as follows:

$$d\_n(\mathbf{x}, \mathbf{x}') = d\_b(\mathbf{x}, \mathbf{x}') + d\_c(\mathbf{x}, \mathbf{x}') + d\_p(\mathbf{x}, \mathbf{x}') + d\_d(\mathbf{x}, \mathbf{x}') \tag{6}$$

Actionability restricts a set of observations meeting some conditions, and consequently, it can be modeled as having the distance to those individuals that do not meet those properties higher than the radius of the neighborhood r (e.g., r plus a small positive value ϵ). For instance, we could define d_p to restrict those instances whose feature x_i differs by ξ as follows:

$$d\_p_i(\mathbf{x}, \mathbf{x}') = \begin{cases} 0 : |x_i - x'_i| < \xi \\ r + \epsilon : otherwise \end{cases} \tag{7}$$

In the case of diversity, it can be enforced by penalizing those observations that are close to specific regions of the input space (e.g., a region where an explanation was previously extracted). The diversity could be defined as follows:

$$d\_d(\mathbf{x}, \mathbf{x}') = \begin{cases} 0 : \mathbf{x} = \mathbf{x}' \\ \delta(\mathbf{x}) + \delta(\mathbf{x}') : otherwise \end{cases} \tag{8}$$

where δ is a function that gives the penalization of an observation. An example of a penalization function is the inverse of the distance between an observation and a penalization point plus one [31].

Finally, data manifold-closeness can be defined similarly to the diversity measure:

$$d\_c(\mathbf{x}, \mathbf{x}') = \begin{cases} 0 : \mathbf{x} = \mathbf{x}' \\ \rho(\mathbf{x}) + \rho(\mathbf{x}') : otherwise \end{cases} \tag{9}$$

where ρ is a function that penalizes observations that are not close to the observations in the training data. Examples of penalization functions in the literature are density estimation techniques [26] and autoencoders [11].

Notice that in the proposed definitions of diversity and data manifold-closeness, we also consider the observation to be explained ($\mathbf{x}$). This observation is considered so that the definition is symmetric and the resulting distance is valid. The resulting distance is not affected because those terms are constant. However, if the extraction algorithm does not require a valid distance (i.e., symmetry is not required), these terms can be dropped.

The presented combination procedure and the diversity, actionability, and data manifold-closeness terms are examples to illustrate the benefits of the methodology. Thus, more complex combination procedures (e.g., weighting the terms), new terms (e.g., transition penalization [26]), or other implementation for diversity, actionability, and data manifold-closeness terms could be used.

3.4. Taxonomy of explanation sets representations

The explanations from an Explanation Set can be represented in several ways. For instance, they can be represented by enumerating their elements. However, this representation becomes impractical as the number of elements grows and it can not be used when the number of elements is infinite. Thus, a more compact representation should be preferred when the number of elements in an Explanation Set is more than a handful. The choice of a representation depends on the problem, and the representation might require the Explanation Set to meet some properties.

We provide a simple taxonomy covering the representation of the current state-of-the-art observation-based explanation methods. These representations provide a compact representation of a set of observations, and consequently, can be used to represent an Explanation Set. The taxonomy is defined as follows:

- **Restrictive or non-restrictive**: Restrictive representations require the Explanation Set to fulfill some properties (e.g., simply connected or restricted to a region of the input space), whereas non-restrictive representations can apply to any Explanation Set.
- **Exact or approximate**: Exact representations contain all the information to generate the original observations of the ES, while approximate representations do not.

Exact representations are mostly restrictive because they usually require the Explanation Set to meet some properties. For instance, rule-based explanations require the explanations to be represented by feature ranges (see RF-OCSE [19] and Anchor [12]). Enumerating the elements of an Explanation Set is an exact

representation that might be considered either restrictive or non-restrictive, depending on whether we consider requiring finiteness a restriction. Notice that exact and restrictive representations can be converted into approximate and restrictive (or non-restrictive) representations by lifting some restrictions (e.g., in rule-based representations, discard those elements outside a rule, or use a rule that covers instances that were not initially defined in the Explanation Set).

Image anchors [12] are a representation for image explanations that requires all the elements within the explanation to have some attributes fixed (i.e., pixels that do not change), and the others can take any value. Hence, image anchors are an exact and restrictive representation. This representation is not limited to images and can be used in any set that meets the requirements. If the image anchors are invariant under translation (i.e., move those pixels along the image, keeping the spatial distribution), the representation could include all the valid translations of the fixed values.

Pertinent positives and Pertinent negatives [11] are a representation similar to image anchors, but in addition to the fixed attributes, it defines restrictions upon some unfixed attributes. The required factors to assert the outcome are pertinent positives and the factors whose absence is required are pertinent negatives. While this representation can be defined with either a counterfactual Explanation Set or a semifactual Explanation Set, it is best defined using a counterfactual Explanation Set for the pertinent positives and semifactual Explanation Set for the pertinent negatives. Thus, a single representation is used to combine the information of two Explanation Sets.

Prototypes and criticisms [14] are an example of non-restrictive and approximate representation. It allows explaining a set of observations by providing the most average instances (prototypes) and the instances that are not well represented by the prototypes (criticisms). In addition, techniques commonly used in the exploratory analysis to describe sets of data can be used (e.g., clustering and centrality measures).

4. Use cases for regression and classification

This section shows how the proposed methodology can enhance existing example-based explanation methods in the literature. For the sake of brevity, we focus only on the most common tasks in *ML*, namely, regression and binary classification, although extensions to other tasks such as anomaly detection and multi-class classification are straightforward. The framework used to extract the Explanation Sets is described in Section 4.1. The regression and classification use cases are described in Sections 4.2 and 4.3, respectively. In both use cases, a gradient boosting tree using Lightgbm [32] with default parameters is considered. The cases of study are carried out using 10-fold cross-validation, and the explanations are generated on the test partitions. Finally, Section 4.4 summarizes the main lessons learned in the use cases.

4.1. Framework

Semifactual-based explanations are extracted using an adaptation of the Anchor method to the proposed methodology [12, 33]. Specifically, the predict_fn parameter has been modified to incorporate the information regarding the grouping measure and neighborhood, and the initial dataset is also modified to include only the observations from the training set belonging to the neighborhood. Notice that this implementation can not generate counterfactual-based explanations because the factual observation is always contained in the explanation. Thus, Anchor can only use similarities as grouping measures. Regarding counterfactual Explanation Sets, they are extracted in two steps. In the first step, a counterfactual is generated using the Tree Parzen Estimators implementation from the Hyperopt library [34]. This optimization procedure is run for 50 iterations with default parameters. The search space is uniform for both categorical and numerical variables. Then, in the second step, the counterfactual Explanation Set is generated by extracting a semifactual Explanation Set for the counterfactual from the previous step. The source code and the data to reproduce the cases of study can be found in `https://github.com/URJCDSLab/explanation_sets_experiments`.

Explanation sets are evaluated in terms of coverage, fidelity (see Definition 2), and the number of conditions in the explanation. In this section, we will refer to the observation whose explanation is of interest as the factual sample and to the elements from the training set that belongs to a given neighborhood as training neighborhood. The coverage is the percentage of observations from the training neighborhood included in the Explanation Set. High coverages are desirable because it implies that the explanation is more generic. The fidelity is estimated as the proportion of samples in the Explanation Set from the training neighborhood that meets the grouping measure. Higher fidelities are also preferred because otherwise, the explanation would not be faithful. The number of conditions, which is the number of features with restrictions in the explanation, is a measure of the complexity of the explanation. A lower number of conditions is desirable because the explanation will be easier to understand. Finally, individual counterfactuals are evaluated in terms of distance to the factual sample and the number of changes (i.e., the number of features that differ between the factual sample and another observation). Small distances and a high sparsity are preferred because counterfactuals that are close to the factual sample and involve few changes are easier to understand.

4.2. Regression case of study

In this case of study, we consider the Concrete Compressive Strength dataset from the UCI repository [35]. This dataset consists of 1030 instances. The goal is to estimate the concrete compressive strength using 8 numeric features. These features measure the age of the concrete and the quantities of several elements of its elaboration. The output variable, the concrete compressive strength, ranges from 2.3 to 82.6. We will refer to the compressive strength of the factual sample as h and the compressive strength of any other observation as h'.

The main goal of this case of study is to illustrate counterfactuals and semifactuals based on the Definitions 1 and 2, respectively, and to study the effect of the grouping measure. These counterfactuals and semifactuals are generated using different grouping measures depending on the goal of the explanation. In addition, a neighborhood with manifold closeness constraints following Eq. 9 is considered. These constraints are implemented using the one-class SVM implementation from Scikit Learn [36] with default parameters. The observations classified as outliers are considered outside the neighborhood. An explanation that combines several counterfactual and semifactual Explanation Sets is also illustrated. Following the taxonomy in Section 3.4, the representations used for the Explanation Sets are rule-based (restrictive and approximate), a combination of rule-based explanations (restrictive and approximate), and finite enumeration (exact and non-restrictive).

First, we evaluate semifactual Explanation Sets explanations. These explanations are extracted using a similarity, $m\_sr_k(h, h')$, that considers two compressive strengths, h and h', to be equals when their difference is less than k (see Eq. 5).

The explanations are extracted using $k = 5$ and $k = 10$. Notice that in problems with a real-valued output (e.g., regression or probabilities), it does not make sense to group only elements whose prediction is the same (i.e., simple-matching similarity). Besides, there might be no difference between two predictions that are close from the user's perspective, or the *ML* model might not be sensitive enough to discriminate at such precision. For instance, in this use case, where the average gap between the sorted target output is 0.07, there is probably no noticeable difference between 50.01 and 50.005. The models obtained an average mean squared error over all folds of 18.44, and consequently, it is not sensible enough to discriminate at such precision or even a few units.

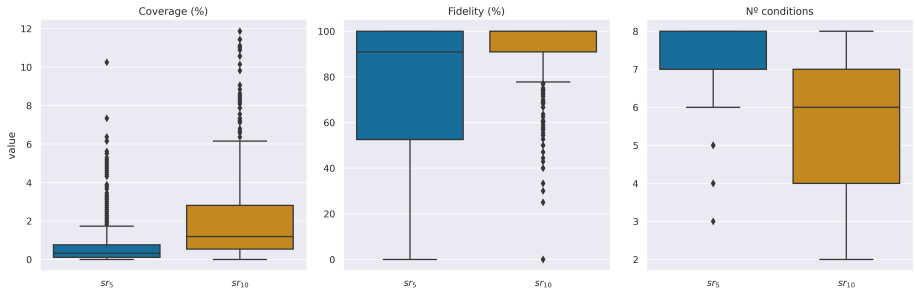

Figure 1: Semifactual Explanation Sets coverage, fidelity, and number of conditions for each similarity. Higher values in coverage and fidelity are preferred and lower values in the number of conditions.

A box plot for each metric and similarity is shown in Figure 1. It can be seen that the semifactual-based explanations using $k = 10$ are better than those obtained using $k = 5$, in all metrics. Notice that these results are not unexpected because the set of all semifactuals using m_sr_5 is a subset of m_sr_{10}.

Thus, the results using m_sr_{10} should be at least as good as m_sr_5. The average coverage is low, 2.02% and 0.67% for m_sr_{10} and m_sr_5, respectively. However, the average coverage in real-valued problems will mostly be far smaller than those of a binary classification problem. Consider the theoretical maximum coverage in this case of study, which is 32.34% and 15.96% for m_sr_{10} and m_sr_5, respectively. Conversely, in a balanced classification problem using a simple-matching similarity, the theoretical maximum coverage is 50.00%. This difference might be even bigger in practice if a lower radius is required or the label distribution is more spread.

The representation for the semifactual Explanation Sets is rule-based. As an example, we show an explanation for the observation with values: cement=540.00, blast furnace slag = 0.00, fly ash = 0.00, water = 162.00, superplasticizer = 2.50, coarse aggregate = 1040.00, fine aggregate = 676.00, and age = 28.00. The rule-based semifactual Explanation Set using the similarity m_sr_5 is the following: cement $>$ 350.00, water $\leq$ 164.90, age $>$ 7.00, fine aggregate $\leq$ 734.15 and 0.00 $<$ superplasticizer $\leq$ 6.50. Using the similarity m_sr_{10}, the explanation is: cement $>$ 349.00, water $\leq$ 165.60, age $>$ 14.00, fine aggregate $\leq$ 733.50, 0.00 $<$ superplasticizer $\leq$ 6.30, and fly ash $\leq$ 0.00.

The low maximum theoretical coverage also has implications on the high variance of the fidelity and the number of conditions. Given the factual sample, $\hat{\mathbf{x}}$, the grouping measure transforms the *ML* model in a binary classifier, where an observation $\mathbf{x}$ is grouped with $\hat{\mathbf{x}}$ if its prediction is 1 or not grouped if 0. Thus, a low maximum theoretical coverage implies that there will be much more 0s than $1s$. Consequently, the problem is imbalanced. The procedure that extracts Anchor explanations relies on sampling on its first stage that does not consider class imbalance. Thus, the observations from the non-grouping group will be over-sampled, decreasing the quality of the explanations. This problem could be partially addressed using an imbalance-aware sampling procedure or generating new observations from the minority class using synthetic methods.

Regarding the number of conditions, having a low coverage indicates that it is likely that the number of conditions is high or the conditions are very restrictive. This dependence arises from the fact that high coverages are only achievable when the rules are broad and cover several instances from the dataset.

Notice that obtaining better values in all metrics does not imply that these explanations are better from the user viewpoint because it depends on which is an acceptable k for the similarity. The choice of the k parameter depends both on the domain and the accuracy of the model.

Then, we evaluate counterfactual explanations without manifold restrictions. These explanations are represented using finite enumeration, and they are extracted with two dissimilarities. First, we consider the similarities m_sr_5 and m_sr_{10} converted into dissimilarities using Eq. 2. Then, we consider a dissimilarity, $d\_gt_{vo}(h, h')$, where a element h' is grouped with h if h' is greater than h plus a value o (see Eq. 3).

The dissimilarity m_gt_{vo} is considered with $o = 0$ and $o = 5$. The counterfactuals extracted using these dissimilarities are depicted in Figure 2. The method extracted valid counterfactuals (i.e., belong to the neighborhood and meet the

grouping measure) in all cases except 10 instances using the dissimilarity $m\_gt_{v,5}$ and one in $m\_gt_{v,0}$. The prediction of the factual samples that correspond to the invalid counterfactuals is contained in the range $[76.84, 80.52]$, which is close to the maximum value in the dataset 82.6. The reason because the method is not able to generate valid counterfactuals is that predicting values higher than the maximum value in the training set might be hard or even impossible (e.g., Decision Trees or Random Forest). Consequently, counterfactual explanations might not always be possible for a given observation and dissimilarity measure.

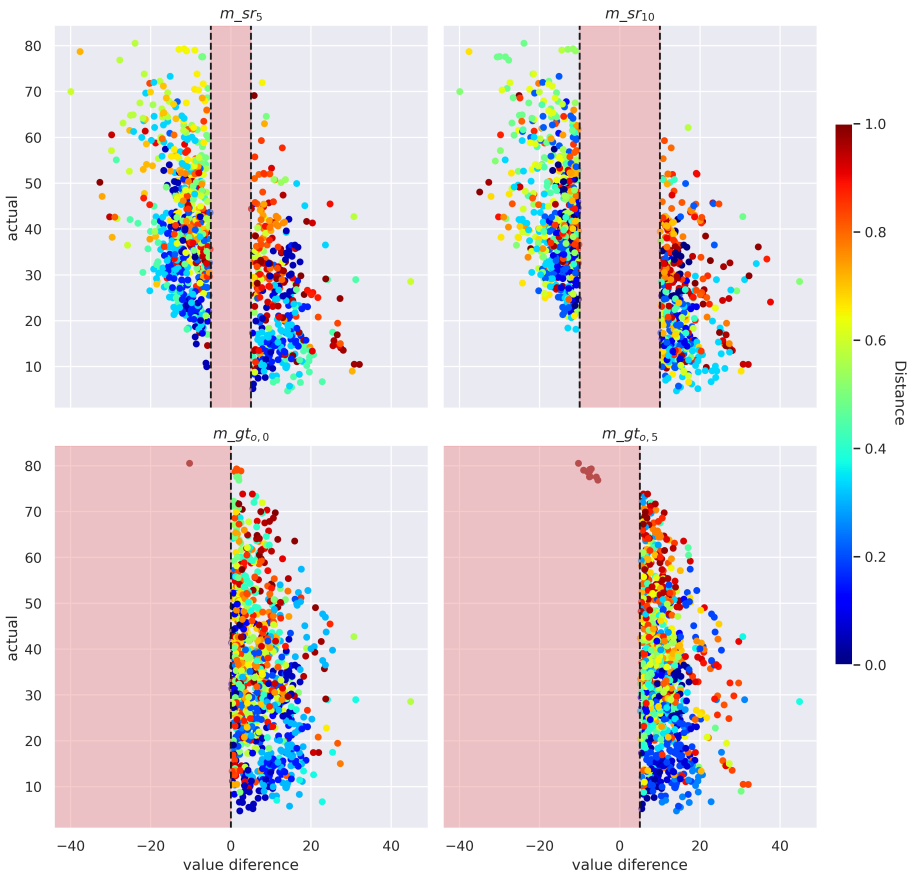

Figure 2: Counterfactuals extracted with several dissimilarities without manifold restrictions. The value difference is the difference between the factual sample prediction and the counterfactual prediction. The sweet pink regions denote the regions of the space where the grouping measure is met. The distance between the factual sample and the counterfactual is transformed using a quantile-based transformation to make it uniform in the interval [0,1].

In the counterfactuals extracted using the dissimilarities m_sr_5 and m_sr_{10}, there is a slight downwards trend. Thus, the counterfactuals extracted from observations with high values tend to have a lower value, and those with low values

a high value. This is explained by the fact that it is easier to find observations whose prediction is closer to the mean of the labels because its distribution is bell-shaped. A similar phenomena can be seen in the counterfactuals extracted using the dissimilarities $m\_gt_{v,0}$ and $m\_gt_{v,5}$. However, since the prediction of the counterfactuals should be higher than that of the factual sample, the observations with higher predictions cannot have counterfactuals with lower values. Therefore, the value difference is as close as possible to the valid frontier. Regarding the counterfactual distance to the factual sample, there does not seem to be a relation between the actual value, counterfactual value, and their distance.

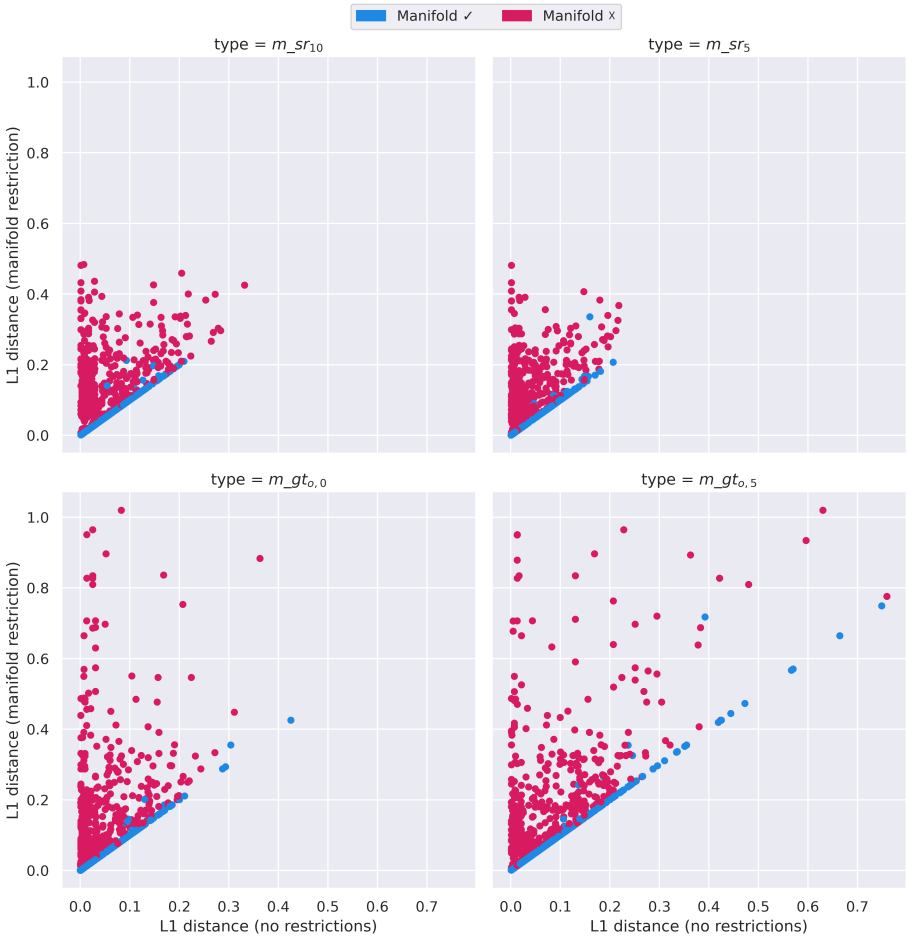

Figure 3: Comparison of the distance to the factual sample of the counterfactuals extracted with and without manifold closeness constraints for each dissimilarity. For a given point, the x-axis and y-axis represent the L1-distance from the counterfactual to the factual sample extracted with and without manifold constraints, respectively. The color indicates if the counterfactual extracted without restrictions meets the manifold restrictions (blue) or not (ruby).

Next, we compare the counterfactuals extracted with and without manifold closeness constraints using the previous dissimilarities. For the sake of clarity, the counterfactuals extracted without manifold closeness restrictions will be referred to as base counterfactuals. The comparison between the counterfactuals is depicted in Figure 3 for each dissimilarity. It can be seen that all observations lie above the diagonal which is expected because base counterfactuals are always closer to the factual sample. The 52.59% of the base counterfactuals meet the manifold closeness restrictions and they mostly lie in the diagonal. On the other hand, when the base counterfactuals do not meet the manifold restrictions the distance for the counterfactual extracted with restrictions increases. This arrangement implies that the extraction method using restrictions can find counterfactuals comparable to the base counterfactuals when possible. The main finding in this comparison is that the difference in distance between the base and restricted counterfactuals is very small in most cases, and consequently, a realistic counterfactual can be usually selected with a small distance penalization. In a small set of cases, especially using the dissimilarities $m\_gt_{v,0}$ and $m\_gt_{v,5}$, there is a notable difference between the distances. However, this is not a limitation of the extraction method and it might provide valuable information for understanding the *ML* method with the help of domain experts.

Finally, we show an explanation based on semifactual and counterfactual Explanation Sets for a random observation from the dataset. The observation has the following values: cement=332.5, blast furnace slag = 142.5, fly ash = 0.0, water = 228.0, superplasticizer = 0.0, coarse aggregate = 932.0, fine aggregate 594.0, and age 270.0. The age is measured in days and the other variables in kg within 1 m^3 mixture. The prediction for this observation is 41.87. The grouping measure used is $m\_gt_{v,0}$. To create this explanation, a new binary target using the grouping measure over all the samples from the training set is calculated. Then, a Decision Tree classifier with a maximum depth of 3 was trained on this data. The resulting Decision Tree is shown in Figure 4. From this tree, we can obtain 8 Explanation Sets, one for each leaf. Thus, the representation is a combination of several rule-based counterfactual and semifactual Explanation Sets. For instance, the leftmost leaf represents a counterfactual Explanation Set because 41.87 is not greater than itself, and it is defined as $age \leq 42$, $cement \leq 266.1$, $water \leq 156.2$. This counterfactual Explanation Set has a coverage of 2.46% and an Explanation Set fidelity of 76.47%. On the other hand, the fourth leaf represents a semifactual Explanation Set defined as $age \leq 42$, $cement > 266.1$, and $superplasticizer > 1.25$.

4.3. Classification case of study

In the second case of study, the proposed framework is evaluated in a classification task. For this purpose, the Adult dataset [35] has been selected. The goal of the dataset is to estimate if a given person earns less or more than $\$50k$. The dataset contains 29170 instances and 12 features. The features are numerical: age, capital gain, capital loss, and hours per week, and categorical: work class, education, marital status, occupation, relationship, race, sex, and country. The dataset is slightly imbalanced. Approximately 80% of the observations belong

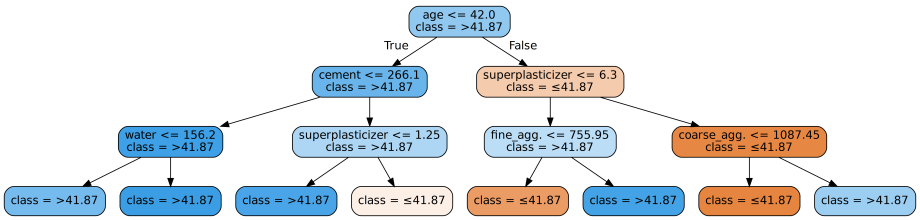

Figure 4: Semifactual and counterfactual Explanation Sets extracted for a random observation from the dataset. The prediction for this observation is 41.87. The conditions from the root to a leaf denote an Explanation Set. The leaves with value $\leq \$50k$ are semifactual Explanation Sets and the leaves with value $> \$50k$ are counterfactual Explanation Sets.

to the less than $\leq \$50k$ group, and the remaining 20% belong to the $> \$50k$ group. Regarding the preprocessing, it uses the same approach as in MACE [10].

In this case of study, the goal is to show explanations extracted from a neighborhood with several actionability restrictions (see Eq. 7). Thus, we compare the explanations extracted with and without these restrictions. Then, an example of a neighborhood with diversity constraints (see Eq. 8) is provided. Regarding the grouping measures, we consider the simple-matching similarity for semifactual-based explanations, and the negated simple-matching (see conversion in Eq. 2) for counterfactual-based explanations. Following the taxonomy in Section 3.4, the representations used for the Explanation Sets are rule-based (restrictive and approximate) and finite enumeration (exact and non-restrictive).

First, we evaluate the pattern of the changes of semifactual and counterfactual Explanation Sets, and the individual counterfactuals used to generate the counterfactual Explanation Sets. These explanations are extracted using two neighborhoods:

- Base explanations: The neighborhood considers the sGower distance [25] parametrized with the L1-Norm for numeric variables and simple-matching for categorical variables. The radius is a high value so that all observations belong to the neighborhood (i.e., no restrictions).

- Restricted explanations: In addition to the base sGower distance, this neighborhood includes other distances that restrict changes over some features (see Eq. 7). The radius is a high value r so that it does not exclude elements based on the base distance. These penalization distances return 0 when a change is allowed and $r+1$ when a change is not. The neighborhood distance is the sum of the base and penalization distances. In the individual counterfactual explanations, the features fixed are age, race, sex, marital status, and relationship status. In counterfactual and semifactual Explanation Sets, the values fix are race, sex, marital status, and relationship status. These values are fixed because they are not usually actuable.

The representation of the individual counterfactuals is finite enumeration (an Explanation Set with one element) and the counterfactual and semifactual Explanation Sets representation is rule-based. As an example with show the explanations for an observation with the following values: age=37, workclass=private, education=masters, marital status=married, occupation=white collar, relationship=wife, race=white, sex=female, capital gain=0, capital loss=0, hours per week=40, and country=United States. In the individual counterfactual, the changed values are: age=28 and education=associates. In the semifactual Explanation Set, the explanation is the following: education = masters, relationship = wife, occupation = white collar, sex = female, marital status = married, age > 30.00, race = white. Finally, the changes in the counterfactual Explanation Set are the following: age ≤ 28.00, hours per week ≤ 40.00.

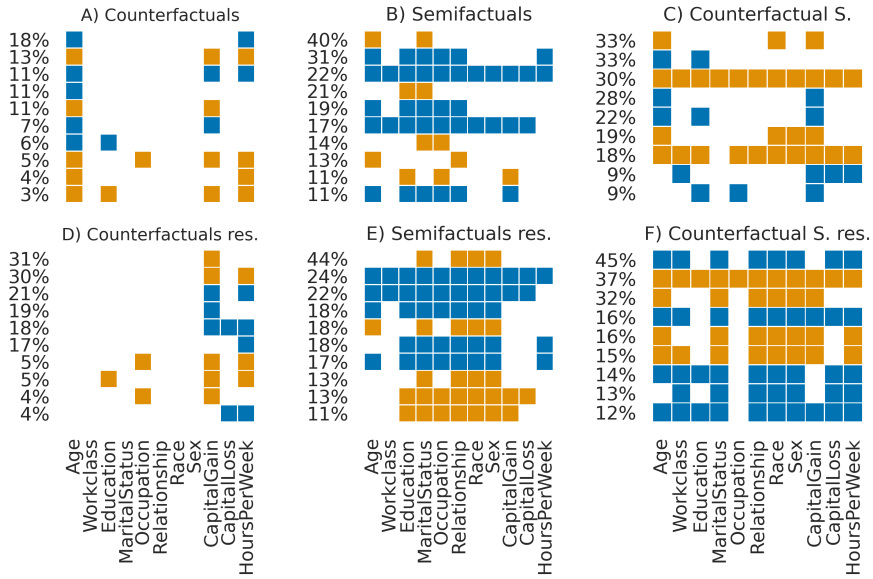

Figure 5: Counterfactual and semifactual Explanation Sets, and individual counterfactual change patterns. The patterns (rows) indicate the percentage of explanations sharing the same structure. The blue and orange squares indicate that there is a change in that feature in counterfactual-based explanations, and that there is a restriction over that value in semifactual-based explanations. Orange rows indicate that the pattern belongs to the $\leq \$50k$ class and the blue color is related to the $> \$50k$ class.

The patterns for each type of explanation are shown in Figure 5. As an example, we explain the most common patterns for counterfactuals (Figure 5 A) and semifactuals (Figure 5 B). In the counterfactuals, the most common pattern appears in the 18% of the explanations of the $> \$50k$ class. Notice that these counterfactuals have changes in age and hours per week. On the other hand, the most common pattern in the semifactual explanations indicates that in the 40% of cases, only the age and marital status have restrictions.

In the individual counterfactual explanations, there is not much difference between the number of feature changes for the two classes. Unlike Anchor explanations that are rule-based, counterfactuals are represented by an observation. Therefore, counterfactuals are not affected by too-specific regions (i.e., small simply connected area of the feature space that belongs to a given class) of the feature space that would result in low coverage anchor explanations. It can be seen that most counterfactuals involve changes over age which might suggest that this feature is very relevant in the classification. On the other hand, the age is fixed in the restricted neighborhood, and consequently, it does not change. The capital gain and hours per week are also relevant in the base neighborhood, and its presence gets magnified in the restricted neighborhood. Besides the age, the features: race, sex, and marital status are also fixed in the restricted neighborhood. In contrast with the feature age, these features only involve changes in a few patterns that are not common. This fact does not imply that these features are less relevant in the classification, and different neighborhoods might generate other patterns. While changes over these features might not be useful if the goal is to try to change the outcome, they are useful in other tasks like diagnosing possible biases or assessing their importance.

In the semifactual Explanation Sets (see Figure 5 B) and D)), there is a high difference between the number of conditions between the patterns of the group $\leq \$50k$ and the $> \$50k$ group. The reason behind this difference is that the classes are slightly imbalanced. This imbalance primarily affects the number of conditions and the coverage, as it promotes more specific explanations that result in more conditions and lower coverage. In addition, fixing the race, sex, marital status, and relationship status also decreases the average fidelity and coverage, and increases the number of conditions. Unlike the semifactual Explanation Sets, the number of conditions is lower in the $> \$50k$ group in counterfactual Explanation Sets. However, this does not contradict the previous findings since the counterfactuals for the $> \$50k$ group belong to the $\leq \$50k$ class and vice versa. Therefore, it supports the previous findings that Anchor obtains better explanations for the majority class in imbalanced problems.

A comparison of the quality metrics for both counterfactual and semifactual Explanation Sets is reported in Table 1. In the semifactual-based explanations, the average number of conditions increases a 35.36% when adding the neighborhood restrictions, and the coverage increases by 233.14%. The fidelity remains similar. Notice that the coverage is calculated over the training neighborhood. Thus, if the restrictions result in a more homogeneous neighborhood, it will increase the coverage. However, the standard deviation of the coverage also increased by 299.08% which suggests that there are several semifactual Explanation Sets with extreme coverage values (either very high or 0 coverage). As previously seen in Figure 5, there is a high difference in quality between the two classes.

Regarding counterfactual Explanation Sets, the number of conditions increased by 27.87% when adding the restrictions, and the fidelity decreased by 23.2%. On the other hand, the coverage increased by 1.88%. It can be seen that there is a high difference between the quality of semifactual Explanation

Sets and counterfactual Explanation Sets. This difference is primarily the effect of two factors. First, the counterfactual extraction procedure might generate counterfactuals in wiggly regions of the feature space (i.e., underrepresented for the counterfactual class) because it is optimizing the neighborhood distance, not for counterfactual Explanation Sets quality. Consequently, the counterfactual Explanation Sets generated there using Anchor will have low coverage or low fidelity. Second, in the same way, the neighborhood restrictions help to achieve a high coverage in the semifactuals Explanation Sets because of the homogeneity, it penalized the counterfactuals Explanation Sets.

Explanation type	Label	N. conditions	Coverage (%)	Fidelity (%)
S.F. Base	$\leq \$50k$	2.19 (0.64)	15.25 (8.7)	98.30 (1.69)
	$> \$50k$	7.50 (2.34)	0.40 (0.7)	93.33 (10.47)
S.F. Restricted	$\leq \$50k$	3.89 (1.47)	50.86 (37.73)	98.09 (3.37)
	$> \$50k$	6.64 (3.14)	1.15 (1.97)	92.85 (13.48)
C.S. Base	$\leq \$50k$	7.14 (2.36)	1.03 (1.64)	83.00 (12.57)
	$> \$50k$	3.05 (1.34)	22.33 (13.83)	95.73 (3.89)
C.S. Restricted	$\leq \$50k$	8.33 (2.33)	1.72 (2.72)	59.55 (32.35)
	$> \$50k$	6.95 (2.53)	20.2 (22.51)	83.45 (13.68)

Table 1: Explanation set quality metrics calculated for the semifactual and counterfactual sets explanations with the base neighborhood and the restricted neighborhood. The mean and standard deviation (in parenthesis) is calculated for each explanation type, neighborhood, and label.

Finally, we evaluate how the proposed methodology can enforce diversity over the counterfactuals extracted from the base neighborhood. The penalization considered is the inverse of the base distance plus 1 to the previous counterfactuals following Eq. 8. The average distance of the diverse counterfactuals increased by 62.14% (0.12 and 0.20 for the base and diverse counterfactuals, respectively). However, the average distance between the base and the diverse counterfactual was 0.42, which is bigger than the distance between the diverse counterfactual and the factual sample. Therefore, the method is capable of enforcing diversity. The representation for diverse counterfactual Explanation Sets is finite enumeration, in particular, a set with two elements: the counterfactual and the diverse counterfactual. Notice that more than one diverse counterfactual can be used to enrich the explanation.

4.4. Lessons learned

This section summarizes the main lessons learned from the proposal in the study cases. The first lesson is about how the grouping measure usually leads to imbalanced classification problems and how this affects the current extraction procedures. The next lesson shows that it is not always possible to generate an Explanation Set. Then, the last lesson is about how extracting counterfactual Explanation Sets by first obtaining a counterfactual and then extracting a semifactual Explanation Set might lead to poor quality explanations.

The proposal was evaluated on two use cases concerning classification and regression tasks. In the regression task, several dissimilarities and similarities were considered. In this case of study, we show how converting the regression problem into a binary classification problem using the grouping measure leads to an imbalanced classification problem. This fact is not a limitation of the proposal, but most counterfactual and semifactual extraction procedures perform worse in imbalanced problems. Further, a large difference in explanation set quality between the minority and majority classes should be expected because most observations of the training neighborhood belong to the majority class. This fact is also evidenced in the classification case of study, where the classes are slightly imbalanced. In particular, when the neighborhood restrictions result in a more homogeneous training neighborhood, this quality difference is magnified.

Another interesting finding is that depending on the *ML* model and task, it might not be possible to always extract a counterfactual explanation. This is because the output of some models such as Decision Trees or Random Forest is bounded, and they can not predict values lower or bigger than those of the training set. However, in such cases, the absence of counterfactuals also provides valuable information. It suggests that there does not exist a hypothetical scenario in which that dissimilarity will be fulfilled. Regarding semifactuals, there is always at least one semifactual because the observation to be explained is a semifactual itself (although not very useful). However, this scenario also provides valuable information, the constraints to extract the semifactual were very restrictive.

Finally, we show that extracting counterfactual sets by first extracting a counterfactual and then extracting a semifactual set might not be the best approach. This is because the counterfactual extraction procedure only optimizes for counterfactual quality, not counterfactual set quality. While it is possible to promote counterfactual Explanation Sets quality in the counterfactual extraction procedure, the approach will be similar to extract counterfactual Explanation Sets directly.

5. Conclusions and future work

In this work, a new explanation framework called Explanation Sets that unifies counterfactuals and semifactuals explanations has been presented. Explanations Sets are an example-based Explanaible *ML* technique to explain *ML* predictions. The key idea is simple yet powerful, explain *ML* predictions using observations from a sub-region of the feature space (neighborhood) and whose prediction compared with factual prediction satisfies a criterion based on the user preferences (grouping measure). The neighborhood restricts the feature space to an area of interest based on their distance to the factual sample and a radius (e.g., actionability and manifold-closeness). If the grouping measure is a dissimilarity, the Explanation Sets are counterfactual Explanation Sets, whereas if the grouping measure is a similarity, the Explanation Sets are semi-

factual Explanation Sets. Both counterfactuals and semifactuals have shown in the literature to be very powerful explanation techniques.

The benefits of the proposal are twofold. First, the neighborhood allows expressing preferences and restrictions to both semifactual and counterfactual-based explanations in a unified way. Second, the users can consider different grouping measures that best suit their understanding of being similar and dissimilar for a given problem and *ML* task. In short, the proposal introduces a general framework that formalizes counterfactual and semifactual explanations and gives the users a tool for expressing their preferences over the explanations in any *ML* task.

Regarding the downsides, the quality of the explanation drops with the current extraction procedure when the grouping measure is restrictive. The quality is worse in counterfactual Explanation Sets than in semifactual Explanation Sets because they include an initial stage where the Explanation Set quality is not optimized. Another downside arises in the definition of neighborhoods based on distances. If several continuous distances are used to determine if a given observation belongs or not to the neighborhood, the membership becomes difficult to understand from the user's perspective. This problem intensifies when these explanations are framed in a high-dimensionality setting (concentration of distances). However, this limitation can be mitigated by using mainly distances that only indicate if an observation belongs or not to the neighborhood.

Future work will focus on developing new methods to extract Explanation Sets from *ML* methods. While Explanation Sets is a generic framework, different extraction procedures might be required to generate different topologies of Explanations Sets (e.g., rule-based). Further, these new methods will focus on overcoming the limitation of the current approaches in imbalanced problems which might occur as a result of a restrictive grouping measure. End-to-end extraction approaches for counterfactuals Explanation Sets will likely improve the quality of the explanations because they could be optimized all along the process. Another future area of work will be to study how different distances affect the resulting explanations. Specifically, a comparison between user-defined distances and model-induced distances, as well as their possible use-cases. In this regard, the effect of high-dimensionality (concentration of distances) on the explanation quality will also be studied. Finally, continuous membership functions for the neighborhood and the grouping measure will be studied. Currently, these functions only indicate whether an observation meets the restrictions. However, by providing a continuous membership function extraction methods can get a sense of how far or close are from meeting the restrictions.

Acknowledgments

Rubén R. Fernández, Isaac Martín De Diego, and Javier M. Moguerza supported by grants from the Spanish Ministry of Economy and Competitiveness under the Retos-Colaboración program: SABERMED (Ref: RTC-2017-6253-1), Retos-Investigación program: MODAS-IN (Ref: RTI2018-094269-B-I00). F. Herrera is supported by the Andalusian Excellence project P18-FR-4961.

References

[1] C. Molnar, Interpretable machine learning, Lulu.com, 2018.

[2] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, arXiv preprint arXiv:1702.08608 (2017).

[3] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, F. Herrera, Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, Information Fusion 58 (2020) 82–115.

[4] J. Zhu, A. Liapis, S. Risi, R. Bidarra, G. M. Youngblood, Explainable ai for designers: A human-centered perspective on mixed-initiative co-creation, in: 2018 IEEE Conference on Computational Intelligence and Games (CIG), IEEE, 2018, pp. 1–8.

[5] V. Belle, I. Papantonis, Principles and practice of explainable machine learning, arXiv preprint arXiv:2009.11698 (2020).

[6] G. Plumb, M. Al-Shedivat, E. Xing, A. Talwalkar, Regularizing black-box models for improved interpretability (hill 2019 version), arXiv preprint arXiv:1906.01431 (2019).

[7] S. Krishnan, E. Wu, Palm: Machine learning explanations for iterative debugging, in: Proceedings of the 2Nd workshop on human-in-the-loop data analytics, 2017, pp. 1–6.

[8] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the gdpr, Harv. JL & Tech. 31 (2017) 841.

[9] V. Arya, R. K. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. V. Liao, R. Luss, A. Mojsilović, et al., One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques, arXiv preprint arXiv:1909.03012 (2019).

[10] A. Adhikari, D. M. Tax, R. Satta, M. Faeth, Leafage: Example-based and feature importance-based explanations for black-box ml models, in: 2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), IEEE, 2019, pp. 1–7.

[11] A. Dhurandhar, P.-Y. Chen, R. Luss, C.-C. Tu, P. Ting, K. Shanmugam, P. Das, Explanations based on the missing: Towards contrastive explanations with pertinent negatives, in: Advances in Neural Information Processing Systems, 2018, pp. 592–603.

[12] M. T. Ribeiro, S. Singh, C. Guestrin, Anchors: High-precision model-agnostic explanations., in: AAAI, Vol. 18, 2018, pp. 1527–1535.

[13] R. Guidotti, A. Monreale, F. Giannotti, D. Pedreschi, S. Ruggieri, F. Turini, Factual and counterfactual explanations for black box decision making, IEEE Intelligent Systems 34 (6) (2019) 14–23.

[14] B. Kim, R. Khanna, O. O. Koyejo, Examples are not enough, learn to criticize! criticism for interpretability, in: Advances in neural information processing systems, 2016, pp. 2280–2288.

[15] P. W. Koh, P. Liang, Understanding black-box predictions via influence functions, in: Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org, 2017, pp. 1885–1894.

[16] R. McCloy, R. M. Byrne, Semifactual "even if" thinking, Thinking & Reasoning 8 (1) (2002) 41–67.

[17] S. J. Sherman, A. R. McConnell, Dysfunctional implications of counterfactual thinking: When alternatives to reality fail us, What might have been: The social psychology of counterfactual thinking (1995) 199–231.

[18] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR, arXiv: 1711.00399 (Mar. 2018).

[19] R. R. Fernández, I. M. de Diego, V. Aceña, A. Fernández-Isabel, J. M. Moguerza, Random forest explainability using counterfactual sets, Information Fusion 63 (2020) 196–207.

[20] A. White, A. d. Garcez, Measurable counterfactual local explanations for any classifier, arXiv preprint arXiv:1908.03020 (2019).

[21] M. Chapman-Rounds, M.-A. Schulz, E. Pazos, K. Georgatzis, Emap: Explanation by minimal adversarial perturbation, arXiv preprint arXiv:1912.00872 (2019).

[22] S. Verma, J. Dickerson, K. Hines, Counterfactual explanations for machine learning: A review, arXiv preprint arXiv:2010.10596 (2020).

[23] A.-H. Karimi, G. Barthe, B. Balle, I. Valera, Model-agnostic counterfactual explanations for consequential decisions, in: International Conference on Artificial Intelligence and Statistics, 2020, pp. 895–905.

[24] T. Laugel, M.-J. Lesot, C. Marsala, X. Renard, M. Detyniecki, Inverse classification for comparison-based interpretability in machine learning, arXiv preprint arXiv:1712.08443 (2017).

[25] R. R. Fernández, I. M. d. Diego, V. Aceña, J. M. Moguerza, A. Fernández-Isabel, Relevance metric for counterfactuals selection in decision trees, in: International Conference on Intelligent Data Engineering and Automated Learning, Springer, 2019, pp. 85–93.

[26] R. Poyiadzi, K. Sokol, R. Santos-Rodriguez, T. De Bie, P. Flach, Face: feasible and actionable counterfactual explanations, in: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2020, pp. 344–350.

[27] A. Van Looveren, J. Klaise, Interpretable counterfactual explanations guided by prototypes, arXiv preprint arXiv:1907.02584 (2019).

[28] R. Yousefzadeh, D. P. O'Leary, Interpreting neural networks using flip points, arXiv preprint arXiv:1903.08789 (2019).

[29] P. Blanchart, An exact counterfactual-example-based approach to tree-ensemble models interpretability, arXiv preprint arXiv:2105.14820.

[30] L. Breiman, Manual on setting up, using, and understanding random forests v3. 1. 2002, URL: http://oz. berkeley. edu/users/breiman/Using_random_forests_V3 1.

[31] R. K. Mothilal, A. Sharma, C. Tan, Explaining machine learning classifiers through diverse counterfactual explanations, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 607–617.

[32] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, Lightgbm: A highly efficient gradient boosting decision tree, in: Advances in neural information processing systems, 2017, pp. 3146–3154.

[33] J. Klaise, A. Van Looveren, G. Vacanti, A. Coca, Alibi: Algorithms for monitoring and explaining machine learning models (2019).
URL `https://github.com/SeldonIO/alibi`

[34] J. Bergstra, D. Yamins, D. Cox, Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures, in: International conference on machine learning, PMLR, 2013, pp. 115–123.

[35] D. Dua, C. Graff, UCI machine learning repository (2017).
URL `http://archive.ics.uci.edu/ml`

[36] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

Appendix A. Conversion uniqueness of Equation 2

Proposition 3. *The conversion defined in Equation 2 is unique.*

Proof. There are two possible mappings between the similarities and dissimilarities defined following Equation 1: the identity transformation (i.e., no transformation) and the conversion in Equation 2. Let a similarity m_s, then by definition of similarity $m_s(\hat{\mathbf{y}}, \hat{\mathbf{y}}) = 1$. If m_s is converted into a dissimilarity m_d using the identity transformation, then $m_d(\hat{\mathbf{y}}, \hat{\mathbf{y}}) = 1$ which contradicts the definition of dissimilarity. The reasoning is similar starting from a dissimilarity, but it would require the similarity of an element to itself to be 0. Now, suppose m_s is converted into m_d using the conversion in the Definition 2, then $m_s(\hat{\mathbf{y}}, \hat{\mathbf{y}}) = 1$ and $m_d(\hat{\mathbf{y}}, \hat{\mathbf{y}}) = 0$. The reasoning is similar starting from a dissimilarity. Henceforth, the conversion between similarities and dissimilarities is unique. Finally, the symmetry property is preserved and the relations $m_d(\hat{\mathbf{y}}, \hat{\mathbf{y}}) \leq m_d(\hat{\mathbf{y}}, \hat{\mathbf{y}'})$ and $m_s(\hat{\mathbf{y}}, \hat{\mathbf{y}}) \geq m_d(\hat{\mathbf{y}}, \hat{\mathbf{y}'})$ hold for the converted similarity and dissimilarity, respectively. □