A class of monotone kernelized classifiers on the basis of the Choquet integral

Ali Fallah Tehrani1 | Marc Strickert2 | Diane Ahrens1

1Technology Campus Grafenau, Deggendorf Institute of Technology, Deggendorf, Germany
2Justus Liebig University Giessen, Giessen, Germany

Correspondence
Ali Fallah Tehrani, Technology Campus Grafenau, Technische Hochschule Deggendorf, Deggendorf, Germany. Email: ali.fallahtehrani@th-deg.de

Abstract
The key property of monotone classifiers is that increasing (decreasing) input values lead to increasing (decreasing) the output value. Preserving monotonicity for a classifier typically requires many constraints to be respected by modelling approaches such as artificial intelligence techniques. The type of constraints strongly depends on the modelling assumptions. Of course, for sophisticated models, such conditions might be very complex. In this study, we present a new family of kernels that we call Choquet kernels. Henceforth, it allows for employing popular kernel-based methods, such as support vector machines. Instead of a naïve approach with exponential computational complexity, we propose an equivalent formulation with quadratic time in the number of attributes. Furthermore, because coefficients derived from kernel solutions are not necessarily monotone in the dual form, different approaches are proposed to monotonize coefficients. Finally, experiments illustrate beneficial properties of the Choquet kernels.

KEYWORDS
Choquet integral, Choquet kernels, isotonic regression, kernel machines, monotone classification

1 | INTRODUCTION

Classification and regression are essential tasks in machine learning and data analytics. Although very powerful artificial intelligence tools meet these demands, the integration of prior knowledge into such approaches requires specific solutions (Ajraoui, Parra-Robles, & Wild, 2013; d. Bezenac, Pajot, & Gallinari, 2018; Du, Feldman, Li, & Jin, 2007; Duivesteijn & Feelders, 2008; Finsy, de Groen, Deriemaeker, Gelad, & Joosten, 1992; Moreno, Regueiro, Iglesias, & Barro, 2004; Muralidhar, Islam, Marwah, Karpatne, & Ramakrishnan, 2018; Thompson & Kramer, 1994; Yu, Jan, Simoff, & Debenham, 2007). Here, we want to establish a formal model to ensure a monotonic dependency between input vectors and their associated output states (J.-R. Cano, Gutiérrez, Krawczyk, Woźniak, & García, 2018). In this regard, a common question is addressed to bias the observations in a proper way to ensure desirable properties such as monotonicity. Taking into account the so-called prior knowledge, underlying model assumption can lead to improving model quality. The prior knowledge can be thought of as a prime structure between input space and output space. For instance, a waiting time for a bus has the Poisson distribution. In this case, establishing a learner preserving the Poisson distribution indeed might improve quality of model. To incorporate prior knowledge into a learning process, two major steps can be considered:

- Identifying the type of prior knowledge.
- Designing/choosing a model which is in agreement with the prior knowledge.
In this paper, we focus specifically on monotonicity, which is an important property in many applications. In medicine, the relationship between smoking and lung cancer is recognized as a monotone relationship; the higher the consumption of tobacco, the higher the risk of getting cancer. In particular, the probability of lung cancer is affected by the expression magnitudes of groups of genes. Also, biological cell size and heat dissipation share a monotone relationship. In modern cameras, the monotonicity can be identified as the inverse relationship between the intensity of the light and reaction of light sensors, to name a few. Including monotonicity in a hypothesis space filters existing hypothesis to a subhypothesis space, called the monotone space. Note that, given an arbitrary hypothesis space, the monotone space might be empty, due to the fact that there is no possibility of conforming to monotonicity. So the goal of monotonization is to enforce that a learned model that complies with the prior knowledge, typically by choosing a hypothesis within the monotone hypothesis space, if possible. In this regard, we advocate the usefulness of a new concept that we call the Choquet integral. Its capability to capture higher order dependencies between sets of data attributes, and thus, its complexity, can be controlled by an additivity parameter. This gives rise to a whole family of models for which we introduce several methods to guarantee monotonicity.

This paper can be regarded as an extension of Fallah Tehrani (2014), which extends the concept of the Choquet kernel. In Fallah Tehrani (2014), we have presented the compact form of the full Choquet kernel; however, the present version proposes the compact form of the Choquet kernel representation for arbitrary degrees of k-additivity. This degree varies between 1 and the number of attributes.

The rest of the paper is organized as follows: In the next section, we give an overview of related works to monotone classifiers. Section 3 is dedicated to introduce the Choquet integral. In Section 4, the Choquet integral is embedded into support vector machine framework. Section 5 presents a new family of kernels called the Choquet kernels. In Section 6, we deal with the concept of monotonization. Ultimately, the results are presented in Section 7.

2 RELATED WORK

As already mentioned, the problem of monotone learning has received a notable attention in AI community recently (Ben-David, 1994; Potharst, Ben-David, & van Wezel, 2009), despite having been presented in the literature much earlier (Ben-David, Sterling, & Pao, 1989; Giannopoulos, Moulianitis, & Nearhou, 2012; Sanchez, Couso, & Blanco, 2017). Commonly, for ensuring monotonicity, two different approaches are pursued:

- Approaches that ensure a monotonic relationship (Pareto dominance) among instances; to name a few: nearest neighbor classification (J. Cano, Aljohani, Ayaz Abbasi, Alowidbi, & Garcia, 2017; Duivesteijn & Feelders, 2008), neural networks (Sill, 1998), decision tree learning (Ben-David, 1995), rule induction (Dembczyński, Kotłowski, & Słowiński, 2009), isotonic separation Chandrasekaran, Ryu, Jacob, and Hong (2005), and piecewise linear models (Dembczyński, Kotłowski, & Słowiński, 2006).

- Approaches that modify training data to ensure a monotonic dependency: Towards this end, data preprocessing methods such as relabeling techniques have been developed. Such methods seek to repair inconsistencies in the training data, so that standard classifiers trained on that data will tend to be monotone. Often, they still do not guarantee this property though (Feelders, 2010; Kotłowski, Dembczyński, Greco, & Słowiński, 2008).

In this study, we follow in essence the first approach, which the core idea is to embed the Choquet integral in kernel-based support vector machine to utilize the advantages of quadratic programming. In the continuation of this section, approaches regarding the Choquet integral specifically for binary class classification are reviewed.

The problem of extracting a Choquet integral in a data-driven way has been addressed in the literature. This is a constraint optimization problem, which can be solved for example using the sum of squared errors as an objective function (Grabisch, 2003; Torra & Narukawa, 2007). To this end, Mori and Murofushi (1989) proposed an approach based on the use of quadratic forms, whereas an alternative, gradient-based method called heuristic least mean squares was introduced in Grabisch, 1995. Particularly, methods for binary class classification based on the Choquet integral were developed in Grabisch and Nicolas (1994) and Yan, Wang, and Chen (2007). In Grabisch and Nicolas (1994), the Choquet integral as a fusion operator has been applied in this context. More concretely, for an instance \( x = (x_1, \ldots, x_m) \), let \( \phi^{(i)}(x) \) express a measure of confidence (provided by feature \( c_i \)) that \( x \) belongs to class \( j \in \{0,1\} \). The global confidence for class \( j \) as an aggregation of these confidence degrees is formalized as follows:

\[
\phi^{(i)}(x) \equiv C_{\mu}(\chi^{(i)}, \ldots, \chi^{(m)}),
\]

where \( C_{\mu} \) denotes the discrete Choquet integral with respect to the fuzzy measure \( \mu \). In order to determine the fuzzy measure, the following objective function is minimized:

\[
J = \sum_{x \in T_0} \left( \phi^{(i)}(x) - \phi^{(i)}(x) - 1 \right)^2 + \sum_{x \in T_1} \left( \phi^{(i)}(x) - \phi^{(i)}(x) - 1 \right)^2,
\]

where \( T_0 \) and \( T_1 \) denote the data sets of non-events and events, respectively.
that is, the sum of squared differences between predicted and given output values. Similarity, Yan, Wang, and Chen (Yan et al., 2007) tackle this problem, albeit using another optimization criterion (which can be seen as a kind of relaxed class separability criterion). Besides, the authors define the Choquet integral based on a so-called signed nonadditive measure Murofushi, Sugeno, and Machida (1994). A signed nonadditive measure $\mu$ defined on $C = \{c_1, ..., c_m\}$ is a function $\mu: \mathcal{P}(C) \to (-\infty, \infty)$ satisfying $\mu(\emptyset) = 0$. Based on signed nonadditive measure, the Choquet integral can be reformulated as follows:

$$(c) \int f d\mu = \sum_{j=1}^{2^m-1} z_j \mu_j,$$

where

$$z_j = \begin{cases} \min(x_{i1}) - \max(x_{ij}) & \text{if } j > 0 \text{ or } j = 2^m - 1 \\ 0 & \text{otherwise} \end{cases},$$

with $a_j = \text{frc}(j/2) \in [0.5, 1)$ and $b_j = \text{frc}(j/2) \in [0, 0.5)$. Here, $\text{frc}(j/2)$ is the fractional part of $j/2$, and the maximum operation on the empty set is zero. For binary class classification, the following setting is proposed:

$$\max_{\beta} \sum_{i=1}^{n} \beta_i \text{ such that }$$

$$(c) \int f d\mu - b \leq \beta \text{ if the case belongs to the first class,}$$

$$(c) \int f d\mu - b \geq -\beta \text{ if the case belongs to the second class,}$$

$\beta_i \geq 0,$

where $\beta = \left(\begin{array}{c} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{array}\right)$, and there are in total $n$ records. Fallah Tehrani, Cheng, Dembczyński, and Hüllermeier (2011), Fallah Tehrani, Cheng, Dembczyński, and Hüllermeier (2012), and Fallah Tehrani, Strickert, and Hüllermeier (2014) generalize the common logistic regression by employing the Choquet integral, which is called choquistic regression. This setting allows using formal statistical techniques for adapting Choquet integral such as maximum likelihood. Apart from this, the choquistic regression is capable to handle a large number of data points, specifically under two-additive scenario (Hüllermeier & Fallah Tehrani, 2013). In addition, Lucca et al. (2016) introduce preaggregation functions, that is, the functions require obviously less conditions to ensuring monotonicity. In fact, instead of satisfying monotonicity constraints, it is enough to ensure monotonicity along some directions. Then, this approach has been used for fuzzy rule-based classification.

### 3 | THE DISCRETE CHOQUET INTEGRAL

In this section, we give the core motivation with respect to the feature mapping $\phi(\cdot)$ presented in the previous section. To this end, we require to introduce preliminaries to delve deeper into the Choquet integral and its characteristics.

Based on Rota’s definition (Rota, 1964), let $(X, \leq)$ be a partially ordered set, which is finite and bounded. For any function $f(\cdot)$ on $(X, \leq)$, the Möbius transformation of $f(\cdot)$ is the function $m : L \to \mathbb{R}$, which satisfies the following equality:

$$f(b) = \sum_{a \leq b} m(a).$$

Now, given a function $f(\cdot)$ on $X$, the solution of the above equation, namely $m(\cdot)$, is obtained as follows:

$$m(b) = \sum_{a \leq b} \mu(a, b)f(a),$$
where \( \nu(\cdot, \cdot) \) is termed as follows:

\[
\nu(p, q) = \begin{cases} 
1 & \text{if } p = q, \\
- \sum_{t \not\in q} \nu(p, t) & \text{if } p \prec q, \\
0 & \text{otherwise}.
\end{cases}
\]

Particularly, any measure \( \mu(\cdot) \) on \( X \) can be substituted by \( f(\cdot) \), meaning that for any measure \( \mu(\cdot) \) on \( X \), there exists a unique solution in terms of its Möbius transformation (Grabisch, 2009). We continue this topic with a brief overview of additive measures; an additive measure \( \mu: 2^X \to [0,1] \) has three identifications:

- \( \mu(A) \geq 0, \forall A \subseteq X \)
- \( \mu(\emptyset) = 0 \) and \( \mu(X) = 1 \)
- For two disjoint sets \( A \) and \( B \),

\[
\mu(A \cup B) = \mu(A) + \mu(B).
\]

The conventional aggregation operators, such as weighted mean, consider that the measure \( \mu(\cdot) \) is additive, though, unfortunately such measures do not represent real world interaction among the attributes very well. On contrary, nonadditive measures do not hold the third property and hence can exhibit more flexibility and therefore model a larger class of measures. In the following, the formal definition of such measures are presented.

### 3.1 Nonadditive measures and discrete Choquet integral

Let \( C = [c_1, \ldots, c_m] \) be a finite set and \( \mu(\cdot) \) a measure \( 2^C \to [0,1] \). For each \( A \subseteq C \), we interpret \( \mu(A) \) as the weight or, say, the importance of the set of elements \( A \). As an illustration, one may think of \( C \) as a set of criteria (binary features) relevant for a job, like “speaking French” and “programming Java”, and of \( \mu(\cdot) \) as the evaluation of a candidate satisfying Criteria \( A \) (and not satisfying \( C \setminus A \)). Formally, nonadditive measures, also called capacities or fuzzy measures, are defined as follows (Sugeno, 1974):

\[
\mu(\emptyset) = 0, \quad \mu(\{c\}) = 1, \quad \text{and} \quad \mu(A) \leq \mu(B) \quad \text{for all } A \subseteq B \subseteq C.
\]

A useful representation of nonadditive measure, that we shall explore reformulating Choquet integrals, is the Möbius transform: The Möbius transform \( m_\mu \) of a fuzzy measure \( \mu \) is defined as follows:

\[
m_\mu(A) = \sum_{B \subseteq A} (-1)^{|A| - |B|} \mu(B),
\]

for all \( A \subseteq C \). So far, the Criteria \( c_i \) were simply considered as binary features. Mathematically, \( \mu(A) \) can thus also be seen as an integral of the indicator function of \( A \), namely, the function \( f_A \) given by \( f_A(c) = 1 \) if \( c \in A \) and 0 otherwise. Now, suppose that \( f: C \to \mathbb{R}_+ \) is any non-negative function that assigns a value to each Criterion \( c_i \), mathematically, the overall evaluation can be considered as an integral \( C_\mu(f) \) of the function \( f(\cdot) \) w.r.t. the measure \( \mu \). The Choquet integral given a fuzzy measure \( \mu \) is formally defined as follows:

\[
C_\mu(f) = \sum_{i=1}^{m} \left( f(c_{(i)}) - f(c_{(i-1)}) \right) \mu(A_{(i)}),
\]

where the subscript notation \( (\cdot) \) refers to a permutation of \( \{1, \ldots, m\} \) such that \( 0 \leq f(c_{(1)}) \leq f(c_{(2)}) \leq \ldots \leq f(c_{(m)}) \) (and \( f(c_{(0)}) = 0 \) by definition), and \( A_{(0)} = [c_{(0)}, \ldots, c_{(0)}] \). In terms of the Möbius transform of \( \mu \), the Choquet integral can also be termed as follows:

\[
C_\mu(f) = \sum_{i \in C} m(T) \cdot \min(c_i).
\]

The term in (3) implies that the Choquet integral underlying its Möbius representation is equal to
where $m^* = \{m[c_1], \ldots, m[c_m], m[c_1, c_2], \ldots, m[c_1, \ldots, c_m]\}$. In fact, the ordering in $m^*$ is the same in $\varphi(\cdot)$. Whereas (2) requires more computational efforts, the term in (4) in essence takes advantage of its compact representation. In particular, one can easily identify that estimating a proper Möbius transformation in a data-driven way is carried out by enforcing monotonicity constraints to that setting, namely, considering the following constraint in the optimization setting:

$$\sum_{B \subseteq A \cup \{c_i\}} m(B) \geq 0, \quad \forall A \subseteq \mathcal{C}, \forall C \in \mathcal{C}.$$  \hspace{1cm} (4)

### 4 | BINARY CLASSIFICATION USING THE CHOQUET INTEGRAL

As stated in the Introduction, the goal of this research is to establish monotone classifiers in an efficient way. Because the Choquet integral intrinsically is a monotone function, it is a suitable candidate for these kinds of classifiers. In essence, the ultimate goal is to estimate a proper Choquet integral from data; estimating here can be understood as data-driven induction of the Möbius transformation (or equivalently, a proper fuzzy measure). Once the parameters of the Möbius transformation have been identified, the model underlying the Choquet integral can be applied for the classification task.

#### 4.1 | Primal form

Typically, estimating proper parameters given data is making use of the induction principle. Here, we are interested in the structural risk minimization, for which it is possible to apply efficient linear programming. To this end, recall again the notation in (3). Basically, this representation can be seen as an inner product between Möbius transformation vector and the basis functions. The set of basis functions is defined by $[\min_{i \in T}[x] \mid T \subseteq \{1, \ldots, m\}]$. Let us define the mapping $\varphi : \mathbb{R}^m \rightarrow \mathbb{R}^{2^m-1}$ as follows:

$$\varphi(x) = \varphi((x_1, \ldots, x_m)) = (x_1, \ldots, x_m, \min(x_1, x_2), \ldots, \min(x_{m-1}, x_m), \min(x_1, x_2, x_3), \ldots, \min(x_1, \ldots, x_m)).$$  \hspace{1cm} (5)

Given the Möbius transformation $m_\varphi$, the discrete Choquet integral is given by

$$\mathcal{C}_{m_\varphi}(x) = \langle m_\varphi, \varphi(x) \rangle,$$

where $m_\varphi$ denotes the Möbius transformation vector in the following ordering:

$$m_\varphi = \langle m(\{c_1\}), \ldots, m(\{c_m\}), m(\{c_1, c_2\}), \ldots, m(\{c_{m-1}, c_m\}), \ldots, m(\{c_1, \ldots, c_m\}) \rangle.$$  

Note that the vector $m_\varphi$ has the same ordering like $\varphi(x)$. Here, the plain definition of the discrete Choquet integral is considered, namely, instead of evaluation function $f(\cdot)$, we proceed with an instance $x$. Thus far, we assumed that the Möbius vector is given. Now, the objective is to estimate a proper Möbius vector from a given data. To this end, assume some instances labelled by two different classes, to be i.i.d.:

$$\{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^m \times \{-1, +1\}.$$  

The inner product representation allows for using $\varphi(x)$ as a feature mapping in support vector machine. To guarantee monotonicity property, we take into account monotonicity constraints as well. The Möbius transformation parameters then are estimated by soft margin optimization:

$$\min_{m \in \mathbb{R}^m} \left\{ \sum_{i=1}^n \left[ 1 - y_i (\omega \cdot \varphi(x_i) + b) \right] + \frac{1}{n} \sum_{i=1}^n \xi_i \right\}$$

subject to

$$y_i (\langle \omega, \varphi(x_i) \rangle + b) \geq 1 - \xi_i, \quad \forall i \in \{1, \ldots, n\}$$

$$\xi_i \geq 0, \quad \forall i \in \{1, \ldots, n\}$$

$$\sum_{B \subseteq A \cup \{c_i\}} m(B) \geq 0, \quad \forall A \subseteq \mathcal{C}, \forall C \in \mathcal{C},$$

where the vector $\omega$ is equal to $m_\varphi$. Figure 1 shows the flexibility of the Choquet integral facing different binary classification problems.
The expression in (3) can be reformulated as an inner product, namely, the complexity of formalism that we call it Choquet kernel. Essentially, a kernel equivalent to the Choquet integral is formulated that reduces the straightforward complexity of $O(2^m)$ to $O(m^2 \log m)$. Estimating proper parameters through kernel setting is accomplished in the dual form:

$$\min_{\alpha} \left\{ \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j a_i a_j K_{q}^i \phi_j(x_i, x_j) - \sum_{i=1}^{n} \alpha_i \right\}$$

s.t.

$$\sum_{i=1}^{n} y_i a_i = 0$$

$$0 \leq a_i \leq \nu \quad \forall i \in \{1, \ldots, n\}.$$  

Here, $\nu$ is a typical support vector machine trade-off parameter, and $K_{q}^{i} \phi(x, \cdot)$ is referred to the Choquet kernel introduced below that accounts for $q$-additive. In the Section 5, we present an efficient way to calculate the Choquet kernel. The kernel-based support vector machine optimization does not ensure monotonicity constraints automatically. In the sequel, it is possible to solve the soft margin optimization setting without considering any monotonicity constraint (relaxed version). Once the parameters have been determined, the learner tries to fix violated monotonicity conditions.

### 4.2 | Dual form

Now, we address an important question, namely, how to embed the Choquet integral in a kernel-based framework. In the following, we propose a formalism that we call it Choquet kernel. Essentially, a kernel equivalent to the Choquet integral is formulated that reduces the straightforward complexity of $O(2^m)$ to $O(m^2 \log m)$. Estimating proper parameters through kernel setting is accomplished in the dual form:

$$\min_{\alpha} \left\{ \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j a_i a_j K_{q}^i \phi_j(x_i, x_j) - \sum_{i=1}^{n} \alpha_i \right\}$$

s.t.

$$\sum_{i=1}^{n} y_i a_i = 0$$

$$0 \leq a_i \leq \nu \quad \forall i \in \{1, \ldots, n\}.$$  

### 5 | THE FAMILY OF THE CHOQUET KERNELS

The expression in (3) can be reformulated as an inner product, namely,

$$\langle m_3, \phi(f(x)) \rangle,$$

with the mapping $\phi : \mathbb{R}^m \rightarrow \mathbb{R}^{2^n-1}$ given by

$$\phi(x) = \phi(x_1, \ldots, x_m) = (x_1, \ldots, x_m, \min \{x_1, x_2\}, \ldots, \min \{x_{m-1}, x_m\}, \min \{x_1, x_2, x_3\}, \ldots, \min \{x_1, \ldots, x_m\}).$$

In addition, $m_3$ denotes the vector

$$\langle m(\{c_1\}), \ldots, m(\{c_n\}) \rangle, m(\{c_1, c_2\}), \ldots, m(\{c_1, \ldots, c_n\}) \rangle.$$  

Assuming $x, x' \in \mathbb{R}^m$, this notation allows for the definition of an inner product between $\phi(x)$ and $\phi(x')$ as follows:

$$\langle \phi(x), \phi(x') \rangle = x_1 x'_1 + \ldots + x_n x'_n + \min \{x_1, x_2\} \min \{x'_1, x'_2\} + \ldots + \min \{x_1, x_2, \ldots, x_m\} \min \{x'_1, x'_2, \ldots, x'_m\}.$$  

Following conventional kernel approaches, $K_{q}^{i} \phi(x, x') = \langle \phi(x), \phi(x') \rangle$ is a kernel (Choquet kernel), because it is defined on a Hilbert space equipped with the feature mapping $\phi(.)$.

**Theorem 1** For given $x, x' \in \mathbb{R}^m$, the $q$-additive Choquet kernel can be reformulated as follows:

$$K_{q}^{i} \phi(x, x') = \langle x, x' \rangle + \sum_{i=1}^{n} x_i \left\{ \sum_{k=1}^{n-1} \left( 1 + \sum_{j=1}^{n-2} \binom{m-k-i}{j} \right) x'_j + \sum_{j=1}^{n-1} \binom{m-p_i-l+1}{j} x'_j \right\}.$$  

**FIGURE 1** Exemplary decision boundary for the Choquet integral in case of $(x, y) \in [0, 1]^2$ with $y = 1 - \min \{y_1, y_2\} > .5$, $(x, y) \in [0, 1]^2$ with $y > .5$ and $(x, y) \in [0, 1]^2$ with $\min \{x, y\} > .5$, respectively.
where \( x_{q}^* \) is the k-th ordered (increasing) value among \( \{ x_1, \ldots, x_n \} \), and \( p_i \) is the position where \( x_{p_i}^* = x_{q}^* \); this refers to the ordered position of \( x_{p_i}^* \) in \( \{ x_1, \ldots, x_n \} \). Moreover, \( \sigma \) is a permutation that sorts the elements of \( x \) in an increasing order.

**Proof of Theorem 1.**

**Proof.** We start first with \( (\varphi(x), \varphi(x')) \). Assuming \( \sigma \) is a permutation that sorts the element \( x \) in increasing order

\[
\begin{align*}
& x_{p_1} \leq \ldots \leq x_{p_m} \\
& \downarrow \quad \ldots \quad \downarrow \\
& x_{q_1} \ldots \ldots \ x_{q_m}.
\end{align*}
\]

The inner product \( (\varphi(x), \varphi(x')) \) for \( q \)-additive can be termed as follows:

\[
(\varphi(x), \varphi(x')) = (\varphi(\sigma(x)), \varphi(\sigma(x'))) = x_{p_1}x_{q_1}^* + \ldots + x_{p_q}x_{q_{q-1}}^* + \min\{x_{p_q}, x_{q_{q-1}}\} \min\{x_{q_1}, x_{q_{q-2}}\} + \ldots + \min\{x_{p_1}, x_{q_1}\} + \min\{x_{q_2}, x_{q_{q-1}}\} + \ldots + \min\{x_{p_q}, x_{q_1}\}
\]

\[
= (x, x') + \sum_{i=1}^{m-1} x_{p_i} \left\{ \sum_{s=1}^{\min\{q-1, m-i\}} \varphi(i,s) \right\},
\]

where additionally,

\[
\varphi(i,s) = \sum_{l<s<k<\ldots<p \leq m} \min\{x_{p_l}, x_{p_{l}}, x_{p_{l}}^*, \ldots, x_{p_{q-1}}^*\} = \sum_{j=r+1}^{m} \sum_{k=r+1}^{m} \ldots \sum_{p=q+1}^{m} \min\{x_{p_l}, x_{p_{l}}, x_{p_{l}}^*, \ldots, x_{p_{q-1}}^*\}. 
\]

The proof is carried out through a separation on \( x_{q}^* \):

1. \( x_{q}^* \) is not the smallest element among the values in a subset.
2. \( x_{q}^* \) is the smallest element among the values in a subset.

1. For this case, \( x_{q}^* \) does not appear (because of min term). In this case, given the element \( x_{q}^* \), there are \( m-i \) elements in \( S = \{ x_{q}^*, \ldots, x_{q_{q-1}}^* \} \). Bascially, each subset of \( S \) containing at most \( q-1 \) elements contributes to the inner product. In this regard, the question is for \( s \)-way subsets of \( S \), how often do the elements smaller than \( x_{q}^* \) appear? This can be summarized as follows: for \( l = 1 \), there is a factor of \( m-i \) \( s \)-1 for \( x_{q}^* \), because \( x_{q}^* \) is the smallest element in \( S \). For \( l = 2 \) there is a factor of \( m-i \) \( s \)-2 for \( x_{q}^* \), because the \( x_{q}^* \) is the second element in \( S \) and so on. In the case of \( (k = q) \)-additivity, the following compact form can be developed:

\[
\begin{align*}
& \sum_{k=1}^{n-1} (m-i-k) x_{p_k}^* \quad \text{(2-way subsets)}, \\
& \sum_{k=1}^{n-1} (m-i-k) x_{q_k}^* \quad \text{(3-way subsets)}, \\
& \sum_{k=1}^{n-1} (m-i-k) x_{r_k}^* \quad \text{(4-way subsets)}, \\
& \sum_{k=1}^{n-1} (m-i-k) x_{s_k}^* \quad \text{(q-way subsets)}.
\end{align*}
\]

Summing up the above terms is equal to

\[
\sum_{k=1}^{n-1} \left( \sum_{j=0}^{q-2} \binom{m-k-1}{j} \right) x_{q_k}^*.
\]

2. On the other hand, for the values greater than \( x_{q}^* \) in the vector \( \mathbf{x} \), we are dealing with the following set:

\[
\left\{ \{ x_{q_{q-1}}, x_{q_{q-2}}, \ldots, x_{q_{q-1}} \} \mid S_k \geq x_{q_{q-1}}, \forall k, 1 \leq k \leq q-1 \right\}.
\]
Because the operator is min in these cases, only $x'_i$ is extracted by a factor. This factor is just the cardinality of the above set, namely,

$$\sum_{j=1}^{q-1} \binom{m-p_i-j+1}{j}.$$  

Considering the two complementing terms above, the resulting expression for $x_{\alpha i}$ is

$$x_{\alpha i} = \left\{ \sum_{k=1}^{q-1} \left( 1 + \sum_{j=1}^{q-2} \binom{m-k-1}{j} x'_{\alpha k} \right) + \sum_{j=1}^{q-1} \binom{m-p_i-j+1}{j} x'_i \right\}.$$  

So far, we only looked at the $i$-th permutation, but finally, the whole index range must be considered. Thus, the nonlinear part of the $q$-additive Choquet kernel can be written down as follows:

$$\sum_{i=1}^{m-1} x_{\alpha i} \left\{ \sum_{k=1}^{q-1} \left( 1 + \sum_{j=1}^{q-2} \binom{m-k-1}{j} x'_{\alpha k} \right) + \sum_{j=1}^{q-1} \binom{m-p_i-j+1}{j} x'_i \right\}.$$  

It is worth mentioning that the two binomial terms act as a switch to select the smallest elements and thus emulate the behaviour of the essential min-operator in the Choquet integral. The linear part $(x, x')$ is still missing in the above term and needs to be added. Adding the inner product, namely $(x, x')$, to the above term, the $q$-additive Choquet integral is revealed, and the proof is completed.

By way of example, given $x=(1,2,3,4)$ and $x'=(9,6,10,8)$, the expanded vectors through the feature mapping $\varphi(\cdot)$ are equal to

$$\varphi(x) = (1,1,1,2,2,3,1,1,1,2,1),$$  
$$\varphi(x') = (9,6,10,8,6,9,8,6,6,8,6,6,8,6,6,6,6).$$  

By extending the previous terms, the full Choquet kernel is calculated as follows:

$$1 \cdot (6+9+8+6+6+9)=1+2 \cdot (6+6+6) + 3 \cdot (9)$$  
$$= 1 \cdot (2^2 + 2^1 + 2^0) + 2 \cdot \left( \sum_{j=1}^{3} \binom{2}{j} \right) + 3 \cdot 2^0.$$  

Example: Given $x=(1,2,3,4,5)$ and $x'=(8,9,10,6,7)$, the different degrees of the Choquet kernels are depicted in Table 1.

It is easy to investigate that the Choquet kernel has a computational complexity of $O(m^2 \log m)$. First of all, the sorting operator $\sigma[\cdot]$ has a computational complexity of $O(\log m)$. In addition, the inner summands in (6), namely,

$$s_i \left\{ \sum_{k=1}^{q-1} \left( 1 + \sum_{j=1}^{q-2} \binom{m-k-1}{j} x'_{\alpha k} \right) x'_{\alpha i} + \sum_{j=1}^{q-1} \binom{m-p_i-j+1}{j} x'_i \right\},$$  

possesses the computational complexity of $O(m^2 \log m)$ (the term also contains sorting operator). The second term in (6) can be termed as an inner product between sorted values and binomial summands so that it contains a linear computational complexity. Finally, the whole term can be expressed as

**Table 1**  Example calculation of (1–5) additive Choquet kernels in an efficient way

| Choquet kernel | $(x, x')$ | 1       | 2       | 3       | 4       | 5       | $\sum$ |
|----------------|-----------|---------|---------|---------|---------|---------|-------|
| $K_1^{(1)}(x,x')$ | 115       | 0       | 0       | 0       | 0       | 0       | 115   |
| $K_2^{(1)}(x,x')$ | 115       | 6 + 7 + 2 * 8 | 6 + 7 + 9 | 6 + 7 | 6 | 0 | 251   |
| $K_2^{(2)}(x,x')$ | 115       | 4 * 6 + 3 * 7 + 3 * 8 | 3 * 6 + 2 * 7 + 1 * 9 | 2 * 6 + 1 * 7 | 6 | 0 | 347   |
| $K_2^{(3)}(x,x')$ | 115       | 7 * 6 + 4 * 7 + 3 * 8 | 4 * 6 + 2 * 7 + 1 * 9 | 2 * 6 + 1 * 7 | 6 | 0 | 384   |
| $K_2^{(4)}(x,x')$ | 115       | 8 * 6 + 4 * 7 + 3 * 8 | 4 * 6 + 2 * 7 + 1 * 9 | 2 * 6 + 1 * 7 | 6 | 0 | 390   |
\[ K_{\xi}^2(\mathbf{x}, \mathbf{x}^*) = \langle \mathbf{x}, \mathbf{x}^* \rangle + \langle \mathbf{x}_s, \mathbf{s} \rangle, \]

where \( \mathbf{x}_s = (x_1, ..., x_{m-1}) \) and respectively \( \mathbf{s} = (s_1, ..., s_{m-1}) \). All these together, the computation complexity of (6) is equal to \( O(m^7 \log m) \).

### 6 | MEASURE CORRECTION: FROM NON-MONOTONE MEASURE TO MONOTONE MEASURE

Thus far, the algorithm for learning the Choquet integral from data in the primal form setting considers enforcing monotonicity with auxiliary constraints, whereas for the dual form, up to now, we have neglected this issue. In the case of monotone learning, we assume that the data have almost a monotone structure; however, this does not imply that the Pareto dominance meets all datapoints. Before continuing on the topic, recall that given optimal parameters \( \{ \alpha_i^* \}_{i=1}^n \) derived from dual form and the feature mapping in (5), the optimal but not necessarily monotone weights can be expressed as follows:

\[ \omega = \sum_{i=1}^n \alpha_i^* y_i \phi(x_i). \]

The core idea of "relaxation" is to learn optimal parameters without enforcing any monotonicity constraint and ultimately monotonizing them to induce a fuzzy measure. In this regard, a nontrivial question is that how the learned weights can be monotonized? Basically, the structure of a fuzzy measure can be seen as a directed acyclic graph (DAG) structure (\( \Theta \)), where the vertices \( V = \{ V_1, ..., V_p \} \) correspond to the elements of \( \mathcal{P} \{ \varepsilon_1, ..., \varepsilon_m \} \) (\( \mathcal{P} \) is referred to the powerset), and \( E = \{ (V_i, V_j) | V_i \subset V_j \} \) is the set of directed edges from \( V_i \) to \( V_j \). We use the notation \( V_i \subset V_j \) if \( (V_i, V_j) \in E \). Next, associated to each vertex \( V \), the optimal learned parameter \( \mu(V) \) (which does not necessarily obey monotonicity constraints) is assigned. The main objective here is to identify a fuzzy measure \( \mu^{**}(\cdot) \) so that it is closest fuzzy measure to the original measure in terms of a given distance. More precisely, our objective here is to determine the following measure:

\[ \mu^{**} \leftarrow \arg\min_{\mu} \{ d(\mu^*, \mu) \ | \mu \text{ is a fuzzy measure on } \mathcal{C} \}, \]

underlying a distance function \( d(\cdot, \cdot) \). Roughly speaking, here, the objective is to determine a set of values that are in accordance with the structure of a DAG and also has a minimal distance to the original values. Burdakov et al. proposed (Burdakov, Grimwall, & Hussain, 2004) an algorithm based on the pool-adjacent-violator (PAV) algorithm. They proposed a generalization of PAV called the GPAV for the purpose of multidimensional isotonic regressions. The main idea is to collect vertices in clusters and to refresh clusters through appending the vertex which violates the monotonicity constraints. They proved that the GPAV algorithm has a computational complexity of \( O(p^2) \), where \( p \) is the number of vertices. Gebhardt, 1970 generalizes PAV for multidimensional structures. Pardalos and Xue (Pardalos & Xue, 1999) presented the IRT-BIN algorithm to tackle the isotonic regression problem for a DAG structure. The solution of measure correction can be studied also under minimal reassignment under \( L_1 \) norm. The minimal reassignment problem is derived from the ordinal classification in which some instances and their labels are given, and the task is to monotonize the data through minimal changing. In this regard, Dembczyński (Dembczynski, Salvatore, Kotlowski, & Slowinski, 2007) proposed an approach to tackle this problem. To this end, the following theorem is presented:

**Theorem 2** Assume that \( \mu \) is an arbitrary measure on set \( \mathcal{C} \). In addition, suppose

\[ \Gamma = \{ \mu^* | \mu^* \text{ is closest monotone measure to } \mu \text{ in terms of the } L_1 \text{ metric} \}. \]

Then \( \mu^{**} \in \Gamma \) exists, which holds the following property:

\[ \text{Im}(\mu^{**}(\cdot)) \subseteq \text{Im}(\mu(\cdot)). \]

This implies that by a proper rearranging of original responses, monotonization can be accomplished.

**Proof.** The proof is deferred to Appendix.
7 | EXPERIMENTS

This part presents some applications of our proposal from a real world perspective. In the following, a case study underlying LETOR dataset will be presented. After that, we introduce several aspects of the proposal in terms of gain and also runtime and computational complexity. For the accuracy comparison, we report the accuracy of the Choquet kernel versus accuracy of state-of-the-art binary classifiers. Moreover, to illustrate its efficiency, we compare runtime of the Choquet kernel versus primal form, namely, its feature mapping.

7.1 | Case study with LETOR 3.0

To further investigate the usefulness of our choquistic approach, we tested it on the OHSUMED corpus, which is a part of LETOR 3.0 collection. The OHSUMED corpus is a well-studied learning to rank dataset in information retrieval. It contains 16,140 instances, and every instance represents a query-document pair with a relevance judgement from 0 to 2, indicating irrelevant, partially relevant, and definitely relevant. There are in total 45 information retrieval features in the OHSUMED corpus. Each of them indicates how relevant or important the document is with respect to a query. For a compact representation and the computational concern, we have selected six representative features among those 45:

- \( \sum_{e \in C} \log (\frac{df}{C^j}) \) in "title" (QT).
- \( \sum_{e \in C} \log (\frac{df}{C^j}) \) in "abstract" (QA).
- \( \sum_{e \in C} \log (\frac{df}{C^j}) \) in "title + abstract" (QTA).
- LMIR with JM smoothing in "title" (LT).
- LMIR with JM smoothing in "abstract" (LA).
- LMIR with JM smoothing in "title + abstract" (LTA).

That is, we have chosen the features derived from the scores of query frequency and LMIR, which are both widely applied numeric measures of the query-document relation. Here, \( |C| \) and df means total number of documents and document frequency, respectively. A detailed description of these features and their generation can be found at Qin, Liu, Xu, and Li (2010), which are beyond the scope of this paper. A very good property of the LETOR features is they are monotone by design. For example, in our case, a higher LMIR score corresponds to a document that is more likely to be relevant to the given query. One can hence expect advantages in terms of predictive performance by using monotone learners, such as the approach we proposed here. In order to follow the same line of evaluation as we did for the benchmark datasets, we have binarized the relevance judgments, so that a binary classifier can be directly applied. Specifically, we associate instances with ratings 2 to the first class and instances with ratings 1 and 0 to the second class. Our evaluation is based on the randomized 50% split with 50 repetitions, where the number of two classes is balanced. The results in 0/1 loss are shown in Table 2. Similar to the results in the benchmark datasets, CK shows superior performance comparing with other logistic regression baselines. The differences, however, are not as large as the ones we have in the benchmark datasets. It is not beyond our expectation. Previous researches have already shown the collections in LETOR are difficult to learn, and it is very challenging to improve upon the baselines (Qin et al., 2010). Recent developments in learning to rank also confirm that the existing solutions to ranking problem are quite mature and the gap between different approaches are very small (Chapelle & Chang, 2011). For example, in the well-known Netflix competition, the difference between the first and the 10th team in terms of RMSE is around 0.005.

| Methods          | 0/1 loss  |
|------------------|-----------|
| CK + IRT-BIN     | .3088 ± .0045 |
| LR               | .3107 ± .0039 |
| KLR-ply, d = 2   | .3115 ± .0036 |
| KLR-rbf          | .3101 ± .0040 |
| MORE             | .3098 ± .0031 |
| LMT              | .3093 ± .0035 |

**TABLE 2** Classification performance on OHSUMED in terms of the mean and standard deviation of 0/1 loss

Abbreviations: CK, Choquet kernel; KLR-rbf, kernelized logistic regression for rbf-kernel; KLR-ply, kernelized logistic regression for polynomial kernel; LMT, logistic model tree; LR, linear logistic regression; MORE, learning rule ensembles under monotonicity constraints.
7.2 | Datasets

We conducted experiments on nine datasets, notably from the UCI repository\(^5\) and the WEKA machine learning framework (Hall et al., 2009; see Table 3 for a summary. Moreover, all input attributes have been normalized in unit interval.

### Table 3 Datasets and their properties

| Dataset               | Number of instances | Number of attributes | Source               |
|-----------------------|---------------------|----------------------|----------------------|
| DenBosch              | 120                 | 8                    | Daniels and Kamp (1999) |
| CPU                   | 209                 | 6                    | UCI                  |
| Breast Cancer (BCC)   | 286                 | 9                    | UCI                  |
| Car-MPG               | 392                 | 7                    | UCI                  |
| Employee Selection (ESL) | 488           | 4                    | WEKA                 |
| Mammographic (Mammo)  | 961                 | 6                    | UCI                  |
| Employee Rejection/Acceptance (ERA) | 1,000   | 4                    | WEKA                 |
| Lecturers Evaluation (LEV) | 1,000        | 4                    | WEKA                 |
| Car Evaluation (CEV)  | 1,728               | 6                    | UCI                  |
| OHSUMED               | 16,410              | 6                    | Qin et al. (2010)    |

7.3 | Methods

Methods that have been applied for the experimental part can be categorized into three categories: linear and monotone, nonlinear and non-monotone, and finally nonlinear and monotone. Nonlinear and non-monotone models are accomplished through kernelized logistic regression: two types of kernels are considered: (a) a degree 2 polynomial kernel and (b) radial basis function kernel; the first kernel is able to model low-level dependencies, whereas the radial basis function kernel is more powerful for modelling dependencies and can capture higher degree of dependencies. For experimental part, the width parameter of the Gaussian kernel was identified through \([10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, \text{ and } 10^0]\) in the most favourable way. For monotone and nonlinear models, we used two methods that are both monotone and flexible, namely, the MORE algorithm for learning rule ensembles under monotonicity constraints (Dembczynski et al., 2009) and the logistic model tree algorithm for logistic model tree.
induction (Landwehr, Hall, & Frank, 2003). Following the idea of forward stagewise additive modelling (Tibshirani, Hastie, & Friedman, 2001), the MORE algorithm treats a single rule as a subsidiary base classifier in the ensemble. Each rule is fitted by concentrating on the examples that are most challenging to classify correctly by rules. The logistic model tree algorithm constructs tree-structured models that contain logistic regression functions at the leaves by incrementally refining those constructed at higher levels in the tree structure.

### TABLE 4  The comparison results for diverse methods in terms of 0/1 loss

| dataset  | CK     | CK + IRT- BIN | LR     | KLR-ply | KLR-rbf | MORE   | LMT     |
|----------|--------|---------------|--------|---------|---------|--------|---------|
| ESL      | .051 ± .0191(1) | .051 ± .0191(1) | .066 ± .0203(3) | .092 ± .0279(6) | .065 ± .0229(2) | .066 ± .0219(4) | .069 ± .0228(5) |
| ERA      | .297 ± .0237(5) | .290 ± .0292(2) | .284 ± .0302(1) | .291 ± .0290(4) | .290 ± .0312(2) | .298 ± .0276(6) | .291 ± .0290(3) |
| LEV      | .141 ± .0188(2) | .142 ± .0191(3) | .162 ± .0249(7) | .147 ± .0231(4) | .149 ± .0233(6) | .139 ± .0214(1) | .147 ± .0232(5) |
| MMG      | .164 ± .0316(3) | .164 ± .0279(3) | .165 ± .0232(4) | .174 ± .0246(6) | .169 ± .0271(5) | .164 ± .0235(2) | .159 ± .0283(1) |
| CPU      | .060 ± .0326(5) | .059 ± .0575(4) | .064 ± .0335(6) | .075 ± .0372(7) | .040 ± .0284(2) | .041 ± .0299(3) | .033 ± .0352(1) |
| CEV      | .037 ± .0076(5) | .041 ± .0073(6) | .132 ± .0173(7) | .028 ± .0075(4) | .023 ± .0066(3) | .019 ± .0070(2) | .008 ± .0047(1) |
| BCC      | .270 ± .0404(5) | .249 ± .0478(1) | .277 ± .0548(7) | .256 ± .0506(2) | .259 ± .0529(4) | .257 ± .0463(3) | .270 ± .0554(6) |
| DBS      | .125 ± .0589(2) | .127 ± .0730(4) | .161 ± .0742(7) | .126 ± .0663(3) | .134 ± .0672(5) | .124 ± .0609(1) | .143 ± .0667(6) |
| MPG      | .059 ± .0319(1) | .061 ± .0320(4) | .061 ± .0263(2) | .072 ± .0268(5) | .074 ± .0284(7) | .073 ± .0269(6) | .061 ± .0251(3) |
| avg. Rank | 3.2   | 3.1            | 4.8   | 4.5    | 4       | 3.1    | 3.4     |

Abbreviations: CK, Choquet kernel; CK + IRT-BIN, monotonized Choquet kernel; KLR-ply, kernelized logistic regression for polynomial kernel; KLR-rbf, kernelized logistic regression for rbf-kernel; LMT, logistic model tree; LR, linear logistic regression; MORE, learning rule ensembles under monotonicity constraints.

### FIGURE 4  Illustration of runtime with respect to primal and dual setting for different datasets. The dark parts weighted average fraction of runtime for primal setting to the runtime of primal setting plus runtime of dual setting with our proposed setting for each dataset individually. The dashed line cuts the rectangles in the middle, which means if the dark part ends on the dash line, both methods would have the same runtime. BCC, breast cancer; CEV, car evaluation; ERA, employee rejection/acceptance; ESL, employee selection; LEV, lecturer’s evaluation; Mammo, mammographic.
For the dual form, the correction module underlying IRT-BIN method (Pardalos & Xue, 1999) is presumed. Please note that the learned weights of the Choquet kernel from dual form do not obey monotonicity necessarily. The correction module, in fact, monotonizes the learned weights. As discussed, the IRT-BIN method gives the optimal solution for the measure correction under $L_2$ norm. For the accuracy comparison, we applied the following setting: the dataset was randomly split into two parts; 80% for training and 20% for testing. This procedure is repeated 50 times, and the results are averaged.

7.4 | Results

We performed two types of experiments, namely, experiments in terms of runtime and also experiments in terms of performance. For the first part, we measured the runtime and compared it for two cases: (a) primal form and (b) dual form. An overview of proposed results is depicted in Figure 4 for the case of the full Choquet kernel. In fact, the runtime involves measure modification part to ensure monotonicity. The results indicate that if the number of attributes is few and the number of training examples is considerable, the primal setting may have an advantage, whereas having a large number of attributes and a few number of training examples our proposed kernels in their dual formulation is obviously desirable. In addition, we performed experiments to evaluate the performance of our proposal. Table 4 shows performance comparisons in terms of conventional 0/1 loss. In terms of performance, in the case of the Choquet kernel, there is no significant gain for IRT-BIN though, which it means that in the light of almost monotone dataset, there is no need to perform any correction procedure. Last but not least, a comparison for original measure derived from kernel setting and the modified measure (fuzzy measure) is depicted in Figure 5. To illustrate this, we sorted the values of fuzzy measure increasingly, and we took the same order for the original measure. Note that such a sorting based serialization of DAG edge weights leads to a shuffled view of the attribute subset measure; the topology of DAG siblings, ancestors, and children cannot be validly represented in such a figure, yet, an intuitive comparison of measures is possible. An interesting observation is that, while for some datasets, the original learned measure is itself monotone, for others, the measure must be monotonized.

To verify the feasibility of the Choquet kernel in terms of runtime, we have performed a runtime comparison between the full Choquet kernel and its feature mapping representation. The results are reported in milliseconds, and for the comparison, 10, 20, and 50% of data have been considered. The results are depicted in Figure 6, with the runtime of the Choquet kernel plotted by dashed lines. As demonstrated, for the feature mapping case, the runtime grows exponentially as the number of datapoints is increased. Meanwhile, it is worth mentioning that although the time complexity is one of the facets of computational complexity, another critical issue is storage demand. Because the feature mapping has a
exponential number of components, the storage of the components is not always tractable. In point of fact, for the number of attributes larger than 14, our machine was not capable of computing components.

8  |  CONCLUDING REMARKS

In this paper, we proposed a novel kernel underlying the Choquet integral, called the Choquet kernel. From the point of view of computational complexity, the Choquet kernel has lower complexity than the integral; however, the weights extracted from it do not necessarily obey monotonicity. To monotonize the weights, we conveyed a survey on several approaches, mainly based on isotonic regression. The monotonized weights then can be used for identifying the Choquet integral. The presented framework allows for theoretically founded modelling of higher order interactions between subsets of attributes in monotone setups. Programs are online available at https://mloss.org/software/view/537/.

ORCID
Ali Fallah Tehrani https://orcid.org/0000-0002-7979-774X

ENDNOTES

1 m in this paper overall is referred to the number of attributes.

2 If (V, V) ∈ E, then V (V) ≤ V (V), where (V) is a measure.

3 http://research.microsoft.com/en-us/um/beijing/projects/letor/

4 http://archive.ics.uci.edu/ml/

5 The experiments have been conducted by Intel(R) Core(TM) i5-4300U CPU @ 1.90 GHz 2.50 Ghz equipped with 8 GB RAM.

6 The experiments have been conducted by Intel(R) Core(TM) i5-4300U CPU @ 1.90 GHz 2.50 Ghz equipped with 8 GB RAM.

REFERENCES

Ajraoui, S., Parra-Robles, J., & Wild, J. M. (2013). Incorporation of prior knowledge in compressed sensing for faster acquisition of hyperpolarized gas images. Magnetic Resonance in Medicine, 69(2), 360–369. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/mrm.24252

Ben-David, A. (1994). A parallel monotonic classification machine. Engineering Applications of Artificial Intelligence, 7(1), 77–82. Retrieved from http://www.sciencedirect.com/science/article/pii/0952178694900450

Ben-David, A. (1995). Monotonicity maintenance in information-theoretic machine learning algorithms. Machine Learning, 19, 29–43.

Ben-David, A., Sterling, L., & Pao, Y. H. (1989). Learning and classification of monotonic ordinal concepts. Computational Intelligence, 5(1), 45–49.

Bezenac, E., Pajot, A., & Gallinari, P. (2018). Deep learning for physical processes: Incorporating prior scientific knowledge. In International conference on learning representations. Retrieved from https://openreview.net/forum?id=By4HsfWAZ

Burdakov, O., Grimmel, A., & Hussain, M. (2004). A generalised pav algorithm for monotonic regression in several variables. In Compstat.

Cano, J., Alijohani, N. R., Ayaz Abbasi, R., Alowidji, J. S., & Garcia, S. (2017). Prototype selection to improve monotonic nearest neighbor. Engineering Applications of Artificial Intelligence, 60(Supplement C), 128–135. Retrieved from http://www.sciencedirect.com/science/article/pii/S0952178617300295

Cano, J.-R., Gutierrez, P. A., Krawczyk, B., Wozniak, M., & Garcia, S. (2018). Monotonic classification: An overview on algorithms, performance measures and data sets.

Chandrasekaran, R., Ryu, Y., Jacob, V., & Hong, S. (2005). Isotonic separation. INFORMS Journal on Computing, 17, 462–474.

Chapelle, O., & Chang, Y. (2011). Yahoo! learning to rank challenge overview. In JMLR: Workshop and conference proceedings (yahoo! learning to rank challenge) (Vol. 14, pp. 1–24).

Daniels, H., & Kamp, B. (1999). Applications of MLP networks to bond rating and house pricing. Neural Computation and Applications, 8, 226–234.

Dembczyński, K., Kotłowski, W., & Słowiński, R. (2006). Additive preference model with piecewise linear components resulting from dominance-based rough set approximations. In International conference on artificial intelligence and soft computing 2006 (Vol. 4029, pp. 499–508).

Dembczyński, K., Kotłowski, W., & Słowiński, R. (2009). Learning rule ensembles for ordinal classification with monotonicity constraints. Fundamenta Informaticae, 94(2), 163–178.

Dembczynski, K., Salvatore, G., Kotlowski, W., & Slowiński, R. (2007). Optimized generalized decision in dominance-based rough set approach. In J. Yao, P. Lingras, W.-Z. Wu, M. Szcuka, N. J. Cercone, & D. Slezak (Eds.), Rough sets and knowledge technology (Vol. 4481, p. 118-125). Springer Berlin Heidelberg. Optimized Generalized Decision in Dominance-Based Rough Set Approach.

Du, H., Feldman, M. W., Li, S., & Jin, X. (2007). An algorithm for detecting community structure of social networks based on prior knowledge and modularity. Complexity, 12(3), 53–60. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/cplx.20166 doi: 10.1002/cplx.20166

Duivesteijn, W., & Feelders, A. (2008). Nearest neighbour classification with monotonicity constraints. In Machine learning and knowledge discovery in databases (Vol. 5211, pp. 301–316). Springer.

Fallah Tehrani, A. (2014). Learning nonlinear monotone classifiers using the choquet integral. Doctoral dissertation, Philipps-University Marburg. https://doi.org/10.17192/z2014.0358

Fallah Tehrani, A., Cheng, W., Dembczyński, K., & Hüllermeier, E. (2011). Learning monotone nonlinear models using the choquet integral. In Proceedings of the european conference on machine learning and principles and practice of knowledge discovery in databases. Springer.
Fallah Tehrani, A., Cheng, W., Dembczyński, K., & Hüllermeier, E. (2012). Learning monotone nonlinear models using the Choquet integral. Machine Learning, 89(1-2), 183–211.

Fallah Tehrani, A., Strickert, M., & Hüllermeier, E. (2014). The choquet kernel for monotone data. In 22th european symposium on artificial neural networks, esann 2014, Bruges, Belgium, april 23–25. Retrieved from http://www.elen.ucl.ac.be/Proceedings/esann/esannpdf/es2014-156.pdf

Feelders, A. (2010). Monotone relabeling in ordinal classification. In Proceedings of the 10th ieee international conference on data mining (pp. 803–808). IEEE Computer Society.

Finsy, R., de Groen, P., Deriemaeker, L., Geladé, E., & Joosten, J. (1992). Data analysis of multi-angle photon correlation measurements without and with prior knowledge. Particle & Particle Systems Characterization, 9(1–4), 237–251. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/ppsc.19920090133 doi: 10.1002/ppsc.19920090133

Gebhardt, F. (1970). An algorithm for monotone regression with one or more independent variables. Biometrika, 57, 263–271.

Gianopoulos, N., Moulianitis, V. C., & Nearchou, A. C. (2012). Multi-objective optimization with fuzzy measures and its application to flow-shop scheduling. Engineering Applications of Artificial Intelligence, 25(7), 1381–1394. Retrieved from http://www.sciencedirect.com/science/article/pii/S0952179612001546

Grabisch, M. (1995). A new algorithm for identifying fuzzy measures and its application to pattern recognition. In Proceedings of the 5th fuzzy system symposium (Vol. 1, pp. 145–150). IEEE.

Grabisch, M. (2003). Modelling data by the Choquet integral. In Information fusion in data mining (pp. 135–148). Springer.

Grabisch, M. (2009). Belief functions on lattices. International Journal of Intelligent Systems, 24(1), 76–95. Retrieved from https://doi.org/10.1002/int.20321 doi: 10.1002/int.20321

Grabisch, M., Murofushi, T., & Sugeno, M. (Eds.). (2000). Fuzzy measures and integrals: Theory and applications. Physica.

Grabisch, M., & Nicolas, J. M. (1994). Classification by fuzzy integral: performance and tests. Fuzzy Sets and Systems, 65(2–3), 255–271.

Grabisch, M., & Roubens, M. (2000). Application of the Choquet integral in multicriteria decision making. In M. Grabisch, T. Murofushi, & M. Sugeno (Eds.), Fuzzy measures and integrals: Theory and applications (pp. 348–374). Physica.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. ACM SIGKDD Explorations Newsletter, 11(1), 10–18.

Hüllermeier, E., & Fallah Tehrani, A. (2013). Efficient learning of classifiers based on the 2-additive choquet integral. In C. Moews & A. Nürnberg (Eds.), Computational intelligence in intelligent data analysis (Vol. 445, p. 17-29). Springer Berlin Heidelberg. Retrieved from https://doi.org/10.1007/978-3-642-32378-2_2.

Kotłowski, W., Dembczyński, K., Greco, S., & Slowiński, R. (2008). Stochastic dominance-based rough set model for ordinal classification. Information Sciences, 178(21), 3989–4204.

Landwehr, N., Hall, M., & Frank, E. (2003). Logistic model trees. In Proceedings of the 14th european conference on machine learning (pp. 241–252). Springer.

Lucca, G., Sanz, J. A., Dimuro, G. P., Bedregal, B., Mesiar, R., Kolesarova, A., & Bustince, H. (2016, April). Preaggregation functions: Construction and an application. IEEE Transactions on Fuzzy Systems, 24(2), 260–272. https://doi.org/10.1109/TFUZZ.2015.2453020

Moreno, D. L., Regueiro, C. V., Iglesias, R., & Barro, S. (2004). Using prior knowledge to improve reinforcement learning in mobile robotics. In Proc. towards autonomous robotics systems. univ. of.

Mori, T., & Murofushi, T. (1989). An analysis of evaluation model using fuzzy measure and the Choquet integral. In Proceedings of the 5th fuzzy system symposium (pp. 207–212). Japan Society for Fuzzy Sets and Systems.

Muralidhar, N., Islam, M., Marwah, M., Karpatne, A., & Ramakrishnan, N. (2018, 12). Incorporating prior domain knowledge into deep neural networks. In (p. 36–45). https://doi.org/10.1145/2018.8621955

Murofushi, T., Sugeno, M., & Machida, M. (1994). Non-monotonic fuzzy measures and the Choquet integral. Fuzzy Sets and Systems, 64(1), 73–86.

Pardalos, P. M., & Xue, G. (1999). Algorithms for a class of isotonic regression problems. Algorithmica, 23(3), 211–222. Retrieved from https://doi.org/10.1007/PL00009258 doi: 10.1007/PL00009258

Pothisart, R., Ben-David, A., & van Wezel, M. (2009). Two algorithms for generating structured and unstructured monotone ordinal data sets. Engineering Applications of Artificial Intelligence, 22(4), 491–496. Retrieved from http://www.sciencedirect.com/science/article/pii/S0952179609000529 https://doi.org/10.1016/j.engappai.2009.02.004

Pothisart, R., & Feelders, A. (2002). Classification trees for problems with monotonicity constraints. ACM SIGKDD Explorations Newsletter, 4(1), 1–10.

Qin, T., Liu, T.-Y., Xu, J., & Li, H. (2010). LETOR: A benchmark collection for research on learning to rank for information retrieval. Information Retrieval, 13 (4), 346–374.

Rota, G. C. (1964). On the foundations of combinatorial theory i. theory of möbius functions. Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete, 2, 340–368.

Sanchez, L., Cousso, I., & Blanco, C. (2017). A class of monotone fuzzy rule-based wiener systems with an application to li-ion battery modelling. Engineering Applications of Artificial Intelligence, 64(Supplement C), 367–377. Retrieved from http://www.sciencedirect.com/science/article/pii/S0952179617301483 https://doi.org/10.1016/j.engappai.2017.06.029

Sill, J. (1998). Monotonic networks. In Advances in neural information processing systems (pp. 661–667). Denver, USA: The MIT Press.

Sugeno, M. (1974). Theory of fuzzy integrals and its application (Unpublished doctoral dissertation). Tokyo Institute of Technology.

Thompson, M. L., & Kramer, M. A. (1994). Modeling chemical processes using prior knowledge and neural networks. AIChE Journal, 40(8), 1328–1340. Retrieved from: doi:10.1002/aic.690400806. https://aiche.onlinelibrary.wiley.com/doi/abs/10.1002/aic.690400806

Tibshirani, R. J., Hastie, T. J., & Friedman, J. (2001). The elements of statistical learning: Data mining, inference, and prediction. Springer.

Torra, V., & Narukawa, Y. (2007). Modeling decisions: Information fusion and aggregation operators. Springer.

Yan, N., Wang, Z., & Chen, Z. (2007). Classification with Choquet integral with respect to signed non-additive measure. Seventh IEEE International Conference on Data Mining Workshops, ICDMW, 283–288.

Yu, T., Jan, T., Simoff, S., & Debenham, J. (2007, 01). Incorporating prior domain knowledge into inductive machine learning.
AUTHOR BIOGRAPHIES

Ali Fallah Tehrani received the B.Sc. degree in pure mathematics from the University of Tehran, Tehran, Iran, the M.Sc. degree in mathematical logic and theoretical informatics from the Tarbiat Modares University, Tehran, Iran, and his PhD degree in Artificial Intelligence from the Philipps University of Marburg, Marburg, Germany. He is senior machine learning researcher at the Deggendorf Institute of Technology, Deggendorf, Germany. His research interests include machine learning, statistical learning theory, data mining, pattern recognition, data analysis, efficient algorithms, mathematical models, monotone learning, preference learning, and fuzzy integrals.

Marc Strickert obtained his PhD in Computer Science in 2005 at the University of Osnabruck in the research group ‘Learning with Neural Methods on Structured Data’. Later he was the group leader of data inspection in the Pattern Recognition Group at the Leibniz-Institute of Plant Genetics and Crop Plant Research Gatersleben (IPK), Germany. 2012 he joined SYNMIKRO-group at the Philipps university Marburg. He is now working with Justus-Liebig-University Giesen, Giesen, Germany. His research interests include self-organizing neural learning architectures for high-dimensional data analysis, time series processing, machine learning, bioinformatics, pattern recognition, unsupervised learning, and similarity-based clustering.

Diane Ahrens received the German diploma degree (adequate to MSc) and PhD degree in business administration and economics at the Passau University, Passau, Germany. She is an expert in global sourcing as well as in supply chain management, combined with international expertise she gained during her industrial career with Siemens AG and her academic career in China, India, Australia and among other countries. In 2003 she was appointed full professor at the Hof University of Applied Sciences, and 2009 at the Deggendorf Institute of Technology, Deggendorf, Germany. Currently, she is also director of the supply chain management, procurement, and big data analytics Research Institute at the Deggendorf Institute of Technology, Germany.

How to cite this article: Fallah Tehrani A, Strickert M, Ahrens D. A class of monotone kernelized classifiers on the basis of the Choquet integral. Expert Systems. 2020;37:e12506. https://doi.org/10.1111/exsy.12506

APPENDIX SECTION A.

Proof. (proof by negation) We suppose that there does not exist such measure. Let us assume that \( A \subset C \) is the smallest subset (minimal, note that the subset \( A \) is not unique) in terms of cardinality, so that \( \forall E \subset C \mu'(A) \neq \mu(E) \). It leads to the following inequality: \( \max_{K \subset A} \mu'(K) < \mu'(A) \leq \min_{L \subset A} \mu'(L) \). The proof proceeds in the following steps:

- If \( \mu(A) \leq \max_{K \subset A} \mu'(K) \), then by defining

\[
\mu_\triangledown(S) = \begin{cases} 
\mu^*(S) & S \neq A \\
\max_{K \subset A} \mu^*(K) & S = A
\end{cases}
\]

the updated measure \( \mu_\triangledown \) is closer to \( \mu \), which is an obvious contradiction. The reason is \( |\mu '_{(A)}-\mu (A)| < |\mu ^*(A)-\mu (A)| \). Note that in this case, \( \mu(A) \leq \max_{K \subset A} \mu'(K) < \mu^*(A) \).

- If \( \max_{K \subset A} \mu'(K) < \mu(A) < \min_{L \subset A} \mu'(L) \), then by defining

\[
\mu_\triangledown(S) = \begin{cases} 
\mu^*(S) & S \neq A \\
\mu(A) & S = A
\end{cases}
\]

the updated measure \( \mu_\triangledown \) is closer to \( \mu \), which surely contradicts with assumption.
• If $\max_{K \subset A} \mu^*(K) \leq \min_{L \subset L} \mu^*(L) < \mu(A)$, then by defining

$$\mu^\Diamond(S) = \begin{cases} 
\mu^*(S) & S \neq A \\
\min_{L \subset L} \mu^*(L) & S = A 
\end{cases}$$

the measure $\mu^\Diamond$ is closer to $\mu$. It is obviously a contradiction.

• If $\max_{K \subset A} \mu^*(K) = \min_{L \subset L} \mu^*(L) < \mu(A)$, then there exists $A', A \subset A$ t.s. $\mu^*(A) = \mu^*(A')$. Assume $A_U$ is the largest (maximal) subset in terms of cardinality (note that the subset $A_U$ in general is not unique), which has the above property. In other words,

$$\forall F \ s.t. \ A_U \subset F \Rightarrow \mu^*(A_U) < \mu^*(F).$$

Moreover, let $\mathcal{I} = \{T \subset C | A \subset T \subset A_U\}$. It is simple to verify that it is not possible that $\forall B \in \mathcal{I}$

$$\mu(B) < \mu^*(B).$$

It is simple to check that it is not possible that $\forall B \in \mathcal{I}$

$$\mu^*(B) < \mu(B).$$

(Otherwise, by defining $\mu^*(A) = \max_{B \in \mathcal{T}} \mu(B)$ and $\mu^*(A) = \min_{B \in \mathcal{T}} \mu(B)$, respectively, is in contradiction to our assumption.) In addition,

$$\sum_{B \in \mathcal{I}} |\mu(B) - \mu^*(B)| = \sum_{B \in \mathcal{I}} |\mu(B) - C|.$$

Suppose $B^* = \arg\min_{B \in \mathcal{T}} |\mu(B) - C|$. By defining $\mu^\Diamond(\cdot)$ as follows:

$$\mu^\Diamond(S) = \begin{cases} 
\mu^*(S) & S \in C \setminus \mathcal{I} \\
\mu(B^*) & S \in \mathcal{I} 
\end{cases}$$

the measure $\mu^\Diamond(\cdot)$ contains the values from the original measure and has the minimal distance to the original measure, which means it is in agreement with measure $\mu^\Diamond$. 

FALLAH TEHRANI ET AL. 17 of 17

Expert Systems

WILEY