Knowledge-Based Systems 220 (2021) 106916

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Classification by ordinal sums of conjunctive and disjunctive functions for explainable AI and interpretable machine learning solutions

Miroslav Hudec [a,c], Erika Mináriková [a], Radko Mesiar [b,e], Anna Saranti [c], Andreas Holzinger [c,d,*]

[a] Faculty of Economic Informatics, University of Economics in Bratislava, Slovakia
[b] Faculty of Civil Engineering, Slovak University of Technology, Slovakia
[c] Medical University Graz, Austria
[d] Alberta Machine Intelligence Institute, Edmonton, Canada
[e] Czech Academy of Sciences, Prague, Czech Republic

ARTICLE INFO

Article history:
Received 19 August 2020
Received in revised form 24 January 2021
Accepted 26 February 2021
Available online 2 March 2021

Keywords:
Explainable AI
Interpretable Machine Learning (ML)
Interactive ML
Aggregation functions
Ordinal sums
Glass-box
Transparency

ABSTRACT

We propose a novel classification according to aggregation functions of mixed behaviour by variability in ordinal sums of conjunctive and disjunctive functions. Consequently, domain experts are empowered to assign only the most important observations regarding the considered attributes. This has the advantage that the variability of the functions provides opportunities for machine learning to learn the best possible option from the data. Moreover, such a solution is comprehensible, reproducible and *explainable-per-design* to domain experts. In this paper, we discuss the proposed approach with examples and outline the research steps in interactive machine learning with a human-in-the-loop over aggregation functions. Although human experts are not always able to explain anything either, they are sometimes able to bring in experience, contextual understanding and implicit knowledge, which is desirable in certain machine learning tasks and can contribute to the robustness of algorithms. The obtained theoretical results in ordinal sums are discussed and illustrated on examples.

1. Introduction

We have plenty of evidence, that when people are reasoning, they do not always follow formal inference rules [1], e.g., modus ponens as it was used in traditional decision support systems [2]. Rather, they construct mental models [3] of the problem and interrogate these models to determine the best possible solution following simple expectation maximizing strategies [4]. However, when they do not find similar tasks or are under time pressure, humans often fail, biased by semantic explanations. It places the importance of semantic meaning ahead instead of logical reasoning (semantically independent). In many application areas this is a beneficial advantage against a formal system, such as understanding a sentence never heard before. Contrary, the illustrative examples are queries about the probability of the conjunction, where people tend to give higher probability to a conjunction of two predicates, rather than to one predicate [1,5–7]. Interestingly, this is fostering algorithms robustness and this is currently the hottest topic in the machine learning community [8,9], emphasized in the Posner-Lecture of Yoshua Bengio at NeurIPS in Vancouver in December 2019.

In many tasks ranging from everyday activities to medical diagnoses laypersons and domain experts classify entities into two classes, which we can mark as *yes–no*, *have illness–do not have illness*, *malign–benign*, and the like. However, for many entities this decision is not straightforward, so we need a class marked as *maybe*. In the three-valued logic *maybe* is expressed by 0.5. However, an entity might slightly or significantly incline to *yes* or *no*. Consequently, we need a many-valued logic. Classification by fuzzy sets and fuzzy logic supported by computing with words has shown its benefits in technical systems, and later in many other fields [10].

Neural networks have shown their efficiency in classification for some time, even beyond human-level performance [11–13]. For supervised learning, it holds true when well-designed (and of sufficient size) sets of input–output data are prepared for learning and validating. By well-designed sets we mean a sufficient amount of data of adequate quality, which covers the whole domain of input data. A neural network having an error equal to 0 might indicate that it works perfectly on a specific subset of input data, but when new data from other sub-domains are obtained, a neural network usually fails (catastrophic forgetting, [14]). The main problem is the lack of explainability: We do not know how the neural network has reached the solution and therefore we cannot explain the reasons why an entity is in,

* Corresponding author at: Medical University Graz, Austria.
E-mail address: andreas.holzinger@medunigraz.at (A. Holzinger).

https://doi.org/10.1016/j.knosys.2021.106916

let say, class *maybe* and moreover, whether it inclines to classes *yes* or *no* and *why*. Contrary, rule-based systems are explainable, but have problems with human interpretability due to the often high complexity [15]. Diverse indicators have been developed to measure interpretability as well as quality; these indicators are mentioned later on when an illustrative rule-based system has been introduced.

A possible solution is *classification by aggregation functions of mixed behaviour*, where parameters of functions and the key input parameters can be learnt from data. In this way, the domain experts can bring in their contextual implicit knowledge, which is described by the interactive machine learning approach [16] and has been proven as being useful within several scenarios [17–19].

Nevertheless, for domain experts one of the problems is the data distribution of the input attributes. Statistical interpretation is often used, but it is understandable only for people having a certain level of statistical literacy. Here, linguistic explanations by quantified summaries and summaries by modes of behaviour can be extremely useful [20]. Such summaries are also a valuable support for explaining solutions. On the entity level it is challenging to find the best explanation. However, on the global level, domain experts gets explanations about the distribution among classes for the whole data set, or for a particular time frame, or subset of considered attributes. This will be very important for future human-AI interfaces, supporting Question-Answering dialogues [21], which were proposed very early (e.g. the Advice Taker, [22]) and already in use in early medical decision support systems [23]. Nowadays, they are becoming important for future human-AI interfaces [24], in the context of explainable AI [25].

In this paper we describe the design and development of a framework for classification by ordinal sums of conjunctive and disjunctive functions and its perspective for so-called glass-box machine learning to support explainable AI [26].

In this framework we consider diverse conjunctive, disjunctive as well as averaging functions in order to propose a novel way for flexible classification. The remainder of this article is organized as follows. Section 2 introduces classical and fuzzy rule-based systems. Section 3 explains classification by aggregation functions. Section 4 is dedicated to ordinal sums in classification and illustrative examples, whereas Section 5 speculates applicability of aggregation functions in machine learning. Finally, Section 7 concludes the article.

2. Classification by the rule-based systems

The main goal of classification (usually by rule-based systems) is dividing entities into several classes. The binary classification divides entities into two distinct classes, usually *yes* and *no*. In the medical field, an illness exists or not, e.g. a melanoma is malign or not [13]. The obvious limitations have been reported in many related works, e.g., [27–29]. The extension is a classification into three classes *yes*, *maybe* and *no*. An example from business is: a promising customer — full discount (Y), a more or less promising one — medium (or average) discount (M), and a non-perspective one — no discount (N). A medical example is: patient has a diagnosis, presumably has (further evaluation is advisable), and does not have.

A rule base illustration of this classification is graphically shown in Fig. 1. Formally, this rule base is as follows:

IF $Atr_1 < a_1$ AND $Atr_2 < a_2$, THEN N;
IF $Atr_1 \geq a_1$ AND $Atr_2 < a_2$, THEN M;
IF $Atr1 < a_1$ AND $Atr_2 \geq a_2$, THEN M;
IF $Atr1 \geq a_1$ AND $Atr_2 \geq a_2$, THEN Y.

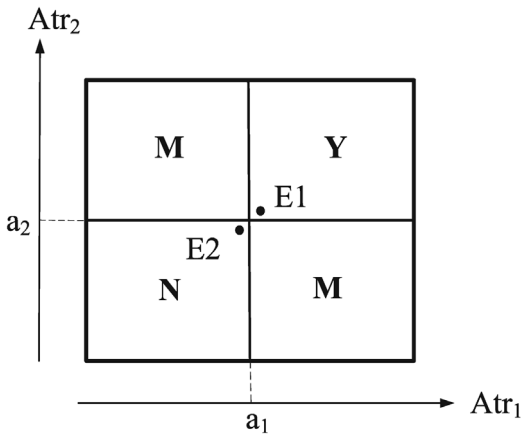

Fig. 1. Classical classification into three classes: *no*, *maybe*, *yes* expressed as, e.g., *no discount*, *average* and *yes (or full) discount* (bonus, belief, etc.).

where $a_1 \in X_1$ and $a_2 \in X_2$, (X_1 and X_2 are domains of attributes $Atr1$ and $Atr2$, respectively).

For illustrative purposes we have two input attributes. Generally, any rule base can be straightforwardly extended to n ($n > 2$) attributes.

Two clearly visible drawbacks are the following:

- The user should define crisp values to formalize crisp rules, which is not a usual human way of reasoning. If these parameters were learnt by machine learning approaches, the rationale for computing particular values and therefore the result of classification remain unexplained.
- Discontinuity, i.e., a small change in attributes' values might cause significant change in output (entities $E1$ and $E2$ in Fig. 1, for instance), which is also not an observable human evaluation.

Flexible (or fuzzy) classification is a way to mitigate the mentioned drawbacks. The modification to fuzzy classification space is formalized as (see Fig. 2)

IF $Atr1$ is *low* AND $Atr2$ is *low*, THEN N;
IF $Atr1$ is *high* AND $Atr2$ is *low*, THEN M;
IF $Atr1$ is *low* AND $Atr2$ is *high*, THEN M;
IF $Atr1$ is *high* AND $Atr2$ is *high*, THEN Y.

Entities $E1$ and $E2$ in Fig. 2 partially activate all rules and therefore belong to all classes, with $\sum_{i=1}^{4} \mu_{C_i}(e) = 1$, where μ_{C_i} is a membership degree to class C_i of entity e. Let in a business case, belonging to Y brings discount of 10, belonging to M brings 5 and belonging to N means no discount, then $E1$ gets discount slightly above 5, whereas $E2$ slightly below 5. In the classical case (Fig. 1), $E1$ gets 10, whereas $E2$ gets 0. Consequently, the resources assigned to motivation remains similar (not always the case, but the motivation is fairer). In a medical case, $\mu_Y(e) = 1$ means absolutely sure, $\mu_M(e) = 0.5$ means more or less sure and $\mu_N(e) = 0$ stands for no alarm. In fuzzy classification, we have medium belief that $E1$ and $E2$ incline to the illness, but $E1$ slightly more than $E2$.

Obviously, the problem of discontinuity is solved, but the task is more tedious for users, because they should assign a higher number of parameters to formalize terms such as *low* and *high*.

When interpretability and explainability are crucial factors, fuzzy-rule based systems are preferred. The reasons are [15]: integration, interaction, validation and trust. However, real-world inference systems consist of a higher number of input attributes and their granules and therefore a higher number of rules is

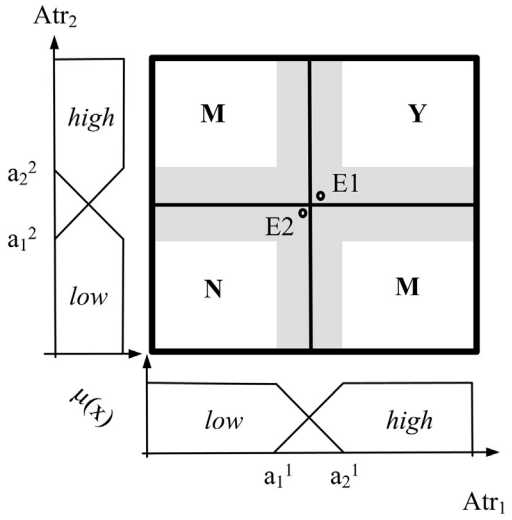

Fig. 2. Fuzzy classification into three classes: *no*, *maybe*, *yes* expressed as *no discount*, *average* and *full discount* (bonus, belief and the like).

constructed. It needs a set of quality indicators to manage consistency of a rule-based system. Some of criteria are widely accepted, whereas some other remain controversial. A deeper insight into interpretability of fuzzy rule-based systems is in, e.g., [30–32]. In order to manage interpretability of rule-based systems the suitable classification is due to [15]. On the fuzzy set level quality indicators include: normality, continuity and convexity. On the level of linguistic variables and fuzzy partitions indicators are: justifiable number of elements, coverage and relation preservation among others. On the fuzzy rules level indicators are description length and granular outputs. Finally, on the fuzzy rule bases level indicators are: compactness, average firing rules, completeness. The further indicators like dominance consistency [33] should not be neglected. When a rule-based system is growing, these indicators become more relevant. A rule-based classification system is a glass-box approach [34], but managing its quality might be very demanding. A complex rule-base can be simplified, e.g., by a graph theory approach suggested in [31], but the initial rule base should be of an acceptable quality. In a neuro-fuzzy system, we have a high demand for input–output data to learn higher number of parameters.

These observations motivated us to explore the possibilities for applying aggregation functions in classification, more precisely ordinal sums which belongs to the category of the mixed aggregation functions.

3. Prerequisites for applying aggregation functions in classification

Aggregation functions aggregate several input values into the most representative one, usually from the closed interval $[0, 1]$ to produce a real value in $[0, 1]$, i.e., $A : [0, 1]^n \to [0, 1]$ where A is an aggregation function which satisfies the following properties [35]:

$$A(1, 1, \ldots, 1) = 1 \qquad \text{boundary condition} \tag{1a}$$

$$A(0, 0, \ldots, 0) = 0 \qquad \text{boundary condition} \tag{1b}$$

$$x_i \le y_i,\ i = 1, \ldots, n \implies A(x_1, \ldots, x_n) \le A(y_1, \ldots, y_n) \quad \text{monotonicity} \tag{1c}$$

The main classification of aggregation functions is due to [36]: conjunctive $0 \le A(\mathbf{x}) \le \min(\mathbf{x})$, i.e., all atomic conditions should be at least partially met, or the property of simultaneity; averaging $\min(\mathbf{x}) \le A(\mathbf{x}) \le \max(\mathbf{x})$; disjunctive $\max(\mathbf{x}) \le A(\mathbf{x}) \le 1$, i.e., at least one condition should be satisfied, or the property of substitutability; and mixed ones, where $\mathbf{x}$ is a vector of degrees of satisfied predicates, $\mathbf{x} = (x_1, \ldots, x_n)$.

Remark. More generally, we can express conjunctive functions as $A(\mathbf{x}) \le x_i$ for each $i \in \{1, \ldots, n\}$, averaging functions as $xi \le A(\mathbf{x}) \le x_j$ for some $i, j \in \{1, \ldots, n\}$, disjunctive functions as $x_i \le A(\mathbf{x})$ for each $i \in \{1, \ldots, n\}$ and mixed as remaining aggregation functions.

Instead of constructing families of fuzzy sets (Fig. 2), the values of the input attributes in our proposal are transformed into the unit interval by the following rule: all values which are clear *low values*, or cause clear *no* assign value 0, whereas clear *high values* assign value 1. The other values should be transformed by a suitable function. This transformation is depicted in Fig. 3. In this way, we get the well-known structure of aggregation. Next, class N can be expressed by a conjunctive function, Y by a disjunctive function and M by an averaging function.

We denote by $\mathcal{C}$ the class of all conjunctive aggregation functions, analogously by $\mathcal{A}_{\mathcal{V}}$ the class of all averaging aggregation functions, by $\mathcal{D}$ the class of all disjunctive aggregation functions, and finally by $\mathcal{M}$ the class of all mixed aggregation functions.

Clearly, the class $\mathcal{C}$ is not suitable for the whole space. Observe the case $C(0, 1) = C(1, 0) = 0$, where $C \in \mathcal{C}$. In this way, rules IF *Atr*1 is *low* AND *Atr*2 is *high*, THEN M, and IF *Atr*1 is *high* AND *Atr*2 is *low*, THEN M are violated for some values in domains X_1 and X_2 (see, Figs. 2 and 3), even though the class $\mathcal{C}$ is suitable for the output class N. The dual observation holds for the class $\mathcal{D}$. Analogously, the class $\mathcal{A}_{\mathcal{V}}$ is not suitable due to compensation effect, which is not suitable for classes N and Y, but acceptable for class M.

Therefore, we need an aggregation function which emphasizes high values, attenuate low values and behave as an averaging function for the mixture of high and low values of input attributes. Thus, the possible solutions are gamma operators and ordinal sums of aggregation functions.

Gamma operators are an attempt to create aggregators compatible with human reasoning [37,38]. The conjunctive aggregation is performed by product t-norm (a strict t-norm having downward reinforcement property, i.e., $T(x, x) < x, x \in]0, 1[$). For the notation of real intervals, we use $[a, b]$ for closed intervals, $]a, b[$ for open intervals, and $[a, b[$ and $]a, b]$ for half-open intervals. The disjunctive aggregation is performed by the dual t-conorm, that is, the probabilistic sum having an upward reinforcement property ($S(x, x) > x, x \in]0, 1[$). Then, the gamma operator is a multiplicative combination (or weighted geometric mean) of product t-norm and its dual t-conorm [38]

$$\gamma_A(\mathbf{x}, \gamma) = \left(\prod_{i=1}^{n} x_i\right)^{(1-\gamma)} \cdot \left(1 - \prod_{i=1}^{n}(1 - x_i)\right)^{\gamma} \tag{2}$$

where $\gamma \in [0, 1]$ and n is the length of vector $\mathbf{x}$, or the following additive combination (or weighted arithmetic mean) of product t-norm and its dual t-conorm

$$\gamma_A(\mathbf{x}, \gamma) = (1 - \gamma)\prod_{i=1}^{n} x_i + \gamma(1 - \prod_{i=1}^{n}(1 - x_i)) \tag{3}$$

The parameter γ plays a role of a in Fig. 3. This operator is idempotent only for $\gamma = 0.5$ and $n = 2$. In our classification task, the multiplicative combination (2) is not suitable, because $\gamma_A((0, 1), \gamma) = 0$ for $\gamma \ne 1$. Next, for $\gamma = 0$, we get conjunction expressed by product t-norm, whereas for $\gamma = 1$ we get a disjunction expressed by the probabilistic sum t-conorm. On the

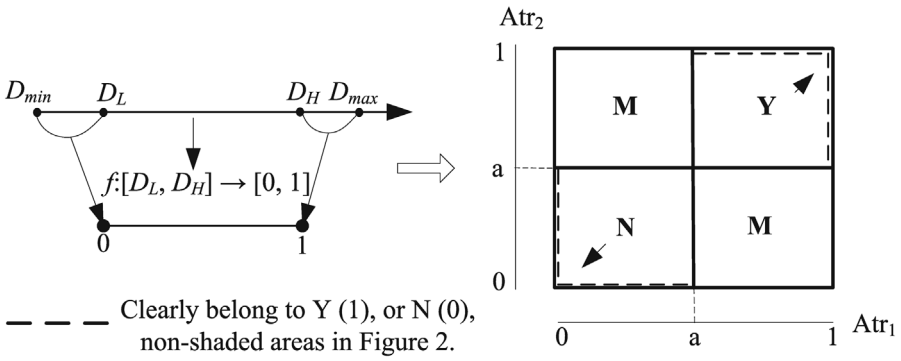

Fig. 3. A transformation from the classification space to the space of aggregation functions.

other hand, the additive combination (3) gives the right solution $\gamma_A((0, 1), \gamma) = 0.5$. However, $\gamma_A((0, x_2), \gamma) \neq 0$, when $x_2 < 0.5$ and $\gamma \in]0, 1]$.

In tasks of fitting functions to input–output data or in interactive machine learning, we adjust only the parameter γ. Conjunctive and disjunctive functions remain unchanged (product t-norm and probabilistic sum t-conorm, respectively). The possible problem is also the general non-idempotency of (2), (3). The question is, whether idempotency is a required property. The answer depends on the considered task. Several smaller values can cause a solution almost equal to 0, for some γ, which is not always desirable. Further, the solution for a vector of several 0.5 should be 0.5. Next, if we expect that value a (Fig. 3) is 0.5 for each attribute (if data distribution is normal, or value means a border between low and high influence), then it should be reflected in γ and therefore we lose the flexibility for adjusting.

4. Ordinal sums of conjunctive and disjunctive functions in classification

Ordinal sums in their origin were considered as extension methods for posets [39] or for semigroups [40]. Later, in the framework of fuzzy sets theory, they were considered to build new t-norms/t-conorms from the scaled versions of existing ones [41]. Let T_i, $i = 1, \ldots, k$ be a family of t-norms and $]a_i, b_i[$, $i = 1, \ldots, k$ be a family of non-empty pairwise disjoint open subintervals of the unit interval. Then function $T : [0, 1]^2 \to [0, 1]$ given by

$$T(x, y) = \begin{cases} a_i + (b_i - a_i) \cdot T_i(\frac{x-a_i}{b_i-a_i}, \frac{y-a_i}{b_i-a_i}) & (x, y) \in]a_i - b_i[^2 \\ \min(x, y) & \text{otherwise} \end{cases} \tag{4}$$

is a t-norm known as the ordinal sum of the summands $\langle a_i, b_i, T_i\rangle$, $i = 1, \ldots, k$. Analogously, we can create a t-conorm as the ordinal sum of the summands $\langle a_i, b_i, S_i\rangle$, $i = 1, \ldots, k$ where S_i is a disjunctive function expressed by t-conorm (then min in (4) is replaced by max).

The ordinal sum of conjunctive and disjunctive functions has been proposed by De Baets and Mesiar [42] as follows

For an n-ary aggregation function $B : [0, 1]^n \to [0, 1]$ and $[a, b] \subset \mathbb{R}$, denote $B_{[a,b]}(\mathbf{x}) = a + (b - a) \cdot B(\frac{\mathbf{x}-a}{b-a})$ Note that then $B_{[a,b]}$ is an n-ary aggregation function on $[a, b]$. Coming back to (4), we see that $(T_i)_{[a_i,b_i]}(x, y) = a_i + (b_i - a_i) \cdot T_i(\frac{x-a_i}{b_i-a_i}, \frac{y-a_i}{b_i-a_i})$.

For $B_1, \ldots, B_k : [0, 1]^n \to [0, 1], k \geq 2$, and $0 \leq a_0 < a_1 < \cdots < a_k = 1$ let $A_i : [a_{i-1}, a_i]^n \to [a_{i-1}, a_i]$ be given by $A_i = (B_i)_{[a_{i-1},a_i]}$. Then the ordinal sum $A : [0, 1]^n \to [0, 1], A = (\langle a_{i-1}, a_i, A_i\rangle)|i = 1, \ldots, k$ is given by

$$A(\mathbf{x}) = \sum_{i=1}^{k} (A_i(a_i \wedge (a_{i-1} \vee \mathbf{x})) - a_{i-1}) \tag{5}$$

is an aggregation function on $[0, 1]$. If all $B_1, \ldots, B_k$ are t-norms (t-conorms) then also A is a t-norm (t-conorm).

Note that, equivalently, $A(\mathbf{x}) = \sum_{i=1}^{k}(a_i - a_{i-1}) \cdot B_i(1 \wedge (0 \vee \frac{\mathbf{x}-a_{i-1}}{a_i-a_{i-1}}))$. For our purposes $n = k = 2$ is considered. Denoting $a_1 = a(a_0 = 0, a_2 = 1)$, we have two next forms of ordinal sums

(i) $B_1, B_2 : [0, 1]^2 \to [0, 1]$,

$$A(x, y) = a \cdot B_1(1 \wedge \frac{x}{a}, 1 \wedge \frac{y}{a}) + (1-a) \cdot B_2(0 \vee \frac{x-a}{1-a}, 0 \vee \frac{y-a}{1-a}) \tag{6}$$

(ii) $A_1 : [0, a]^2 \to [0, a], A_2 : [a, 1]^2 \to [a, 1]$,

$$A(x, y) = A_1(a \wedge x, a \wedge y) + A_2(a \vee x, a \vee y) - a \tag{7}$$

Then:

- if $(x, y) \in [0, a]^2$, $A(x, y) = a \cdot B_1(\frac{x}{a}, \frac{y}{a}) = A_1(x, y)$,
- if $(x, y) \in [a, 1]^2$, $A(x, y) = a + (1-a) \cdot B_2(\frac{x-a}{1-a}, \frac{y-a}{1-a}) = A_2(x, y)$,
- if $(x, y) \in [0, a] \times [a, 1]$, $A(x, y) = a \cdot B_1(\frac{x}{a}, 1) + (1 - a) \cdot B_2(0, \frac{y-a}{1-a}) = A_1(x, a) + A_2(a, y) - a$,
- if $(x, y) \in [a, 1] \times [0, a]$, $A(x, y) = a \cdot B_1(1, \frac{y}{a}) + (1 - a) \cdot B_2(\frac{x-a}{1-a}, 0) = A_1(0, y) + A_2(x, a) - a$.

Obviously, if B_1 is conjunctive and B_2 is a disjunctive aggregation function, then A is conjunctive on $[0, a]^2$ and disjunctive on $[a, 1]^2$. Moreover, if B_1 has a neutral element $e = 1$, i.e., B_1 is a semicopula [43], and B_2 has a neutral element $e = 0$, i.e., B_2 is a dual semicopula, then, for $(x, y) \in [0, 1]^2 \setminus ([0, a]^2 \cup [a, 1]^2)$ it holds $A(x, y) = x + y - a \in [\min(x, y), \max(x, y)]$, i.e., A is averaging on this domain.

Note that if B_1 is continuous but not a semicopula, i.e., $B_1(x_0, 1) < x_0$ or $B_1(1, x_0) < x_0$ for some $x_0 \in]0, 1[$, then A is not averaging on $[0, a] \times [a, 1] \cup [a, 1] \times [0, a]$. A similar clam holds for B_2. Hence, supporting the continuity of B_1 and B_2, A is continuous and reflects our demands depicted in Figs. 2 and 3. Thus, B_1 should be a semicopula and B_2 a dual semicopula.

4.1. Variations of conjunctive and disjunctive functions in ordinal sums

For simplicity, when $n = 2$, we denote the elements of vector $\mathbf{x}$ as x and y. To keep the requirement from Fig. 3, we need a conjunction for class N, a disjunction for class Y as well as an averaging function for class M.

4.1.1. Product t-norm and its dual probabilistic sum t-conorm

These functions (as representative of strict functions) are suitable to keep the solution equal to 0 for cases indicated by dashed line in class N, and an solution equal to 1 for cases indicated by dashed line in class Y (Fig. 3). The question is how to manage averaging behaviour. Just a reminder, product t-norm is expressed

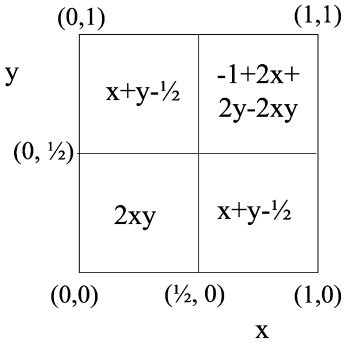

Fig. 4. The graphical interpretation of (8) for product t-norm, probabilistic sum t-conorm and arithmetic mean.

as $C_P(x, y) = x \cdot y$, whereas its dual t-conorm as $D_P(x, y) = x + y - x \cdot y$.

In order to keep the expected value on edges of subintervals $[0, a]^2$ and $[a, 1]^2$, when $a = 0.5$, the product t-norm is expressed as $C_P(x, y) = A_1(x, y) = 2x \cdot y$ (note that $2 \cdot 0.5 \cdot 0.5 = 0.5$). The dual observation holds for t-conorm: $D_P(x, y) = A_2(x, y) = -1 + 2x + 2y - 2x \cdot y$.

Next, the averaging function expressed by the arithmetic mean is as follows

$$AM(x, y) = A_1(x, \frac{1}{2}) + A_2(\frac{1}{2}, y) - \frac{1}{2} = x + y - \frac{1}{2} \tag{8}$$

The graphical interpretation can be seen in Fig. 4. The solution is shown in Table 1, column: solution for *AM*.

For simplicity, let values of both attributes be lower than or equal to 10 to indicate clear non-concern, and values higher or equal to 100 to indicate full concern. The transformation rule for the other values is linear, i.e.,

$$x = \begin{cases} 0 & \text{for } Atr1 \leq 10 \\ \dfrac{Atr1 - D_L}{D_H - D_L} & \text{for } Atr1 \in (D_L, D_H) \\ 1 & \text{for } Atr1 \geq 100 \end{cases}$$

Entities *E1*, *E2*, *E3* and *E4* are distinguishable (strict t-norm for *E3* and *E4* and averaging behaviour for *E1* and *E2*). Because, for *E2* both attributes have low values, therefore the intensity of the concern is decreased. For high value of one attribute and value 1, solution is 1. Clearly, low values have strict conjunctive behaviour, whereas high values have strict disjunctive behaviour and a mix of low and high has averaging behaviour, indicating that it is belonging to class M. In this case the averaging behaviour is managed by the arithmetic mean (one of its properties is the full compensation, an increased value of the first attribute by δ is compensated with a decreased value of the second attribute by δ).

Ordinal sum (5) introduced in [42] is based on the arithmetic mean *AM* as a solution of the equation

$$AM(A_1(a_1 \wedge \mathbf{x}), A_2(a_2 \wedge (a_1 \vee \mathbf{x})), \ldots, A_k(a_{k-1} \vee \mathbf{x})) = AM(a_1, \ldots, a_{k-1}, b) \tag{9}$$

for the variable b. Alternative ordinal sums (still covering both the ordinal sums of t-norms and ordinal sums of t-conorms) proposed in [42] are based on the quasi-arithmetic means. Recall that each quasi-arithmetic mean on $[0, 1]$ is generated by an additive generator $g : [0, 1] \to [-\infty, \infty]$ (g is continuous and strictly monotone), and then $QAM_g(x) = g^{-1}(\frac{1}{n}\sum_{i=1}^{n} g(x_i))$.

These ordinal sums can be seen as solutions of the equation $QAM_g(A_1(a_1 \wedge \mathbf{x}), A_2(a_2 \wedge (a_1 \vee \mathbf{x})), \ldots, A_k(a_{k-1} \vee \mathbf{x})) = QAM_g(a_1, \ldots, a_{k-1}, b_g)$ in the variable b_g. Thus, in our case when $n = k = 2$ and $a = a_1 = \frac{1}{2}$ we have for $Avg_P = A! = |_{[0,a]\times[a,1]}$ the next formula

$$Avg_P(x, y) = g^{-1}(g(A_1(x, \frac{1}{2})) + g(A_2(\frac{1}{2}, y)) - g(\frac{1}{2})) \tag{10}$$

When $g(t) = -\log t$ we get the geometric mean, and then G_P is given by (supporting the symmetry of A_1 and A_2)

$$G_P(x, y) = \frac{A_1(x, \frac{1}{2}) \cdot A_2(\frac{1}{2}, y)}{\frac{1}{2}} = 2x \cdot y \tag{11}$$

whereas for $g(t) = t^{-1}$ we get the harmonic mean, and then H_P is given by (again supporting the symmetry of A_1 and A_2)

$$H_P(x, y) = \frac{1}{\frac{1}{A_1(x, \frac{1}{2})} + \frac{1}{A_2(\frac{1}{2}, y)} - \frac{1}{\frac{1}{2}}} = \frac{x \cdot y}{x + y - 2x \cdot y} \tag{12}$$

The graphical interpretation for (11) is shown in Fig. 5. At first glance, averaging and conjunctive parts are managed by the same function. But, $2x \cdot y$ behaves in $[0, 0.5] \times [0, 0.5]$ like a conjunctive function, whereas in $[0, 0.5] \times [0.5, 1]$ and $[0.5, 1] \times [0, 0.5]$ like an averaging function. The solution is shown in Table 1, column: solution for *G*. Now, we have covered the averaging behaviour by the geometric mean. We can now observe that the solution for *E9* is 0, although we expected an averaging behaviour. In fact it is an averaging behaviour, because the value 0 is an annihilator for the geometric mean. By this case, we covered the behaviour of the gamma operator (2). In the case of a dual geometric mean, we get the annihilator equal to 1, which leads to the solution $(0, 1) = 1$. If the full influence of the extreme values is required in averaging part of the classification space, we are able to cover it. The *andness* measure of the geometric mean is higher than 0.5 [37] (namely it is $\frac{1}{3}$), i.e., the solution is lower than or equal to the arithmetic mean, for which the *andness* measure is 0.5. The opposite holds for the dual geometric mean.

In this way, we are able to cover diverse requirements for averaging behaviour when a conjunctive (resp. disjunctive) behaviour is managed by strict t-norm (resp. strict t-conorm).

Table 1
The classification solution when $a = 0.5$, A_1 is product, A_2 probabilistic sum and averaging behaviour is covered by arithmetic and geometric means.

Entity	Atr1	Atr2	**x**	**y**	Solution for *AM* (8)	Solution for *G* (11)
E1	28	77.5	0.2	0.75	0.45	0.3
E2	28	73	0.2	0.7	0.4	0.28
E3	28	28	0.2	0.2	0.08	0.08
E4	28	50.5	0.2	0.45	0.18	0.18
E5	82	91	0.8	0.9	0.96	0.96
E6	59.5	91	0.55	0.9	0.91	0.91
E7	132	73	1	0.7	1	1
E8	6	37	0	0.3	0	0
E9	146	4	1	0	0.5	0
E10	55	55	0.5	0.5	0.5	0.5
E11	55	28	0.5	0.2	0.2	0.2
E12	28	86.5	0.2	0.85	0.55	0.34
E13	37	73	0.3	0.7	0.5	0.42
E14	3	9	0	0	0	0
E15	102	117	1	1	1	1

4.1.2. Łukasiewicz t-norm and t-conorm

Let us now consider the Łukasiewicz t-norm ($C_L(x, y) = \max(0, x+y-1)$) and its dual t-conorm ($D_L(x, y) = \min(1, x+y)$). In this case (nilpotent t-norm and t-conorm), significantly low values of both attributes cause solution equal to 0, whereas two significantly high values cause a solution equal to 1. Thus, we are able to model the case when lower values indicate no concern and

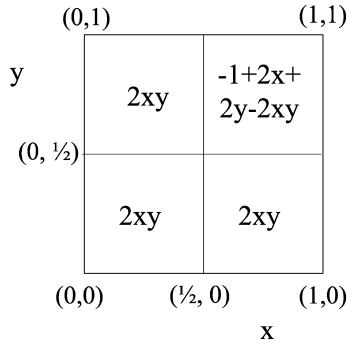

Fig. 5. The graphical interpretation of (11) for product t-norm, probabilistic sum t-conorm and geometric mean.

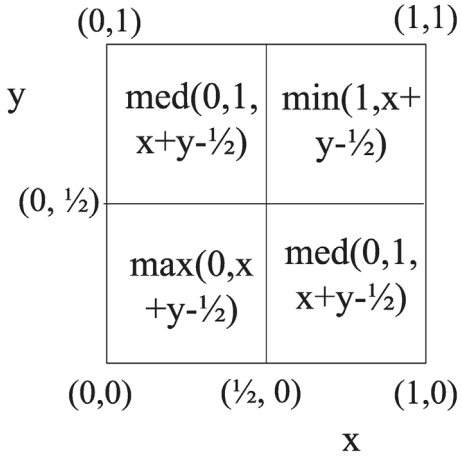

Fig. 6. The graphical interpretation of (8) for Łukasiewicz t-norm and t-conorm.

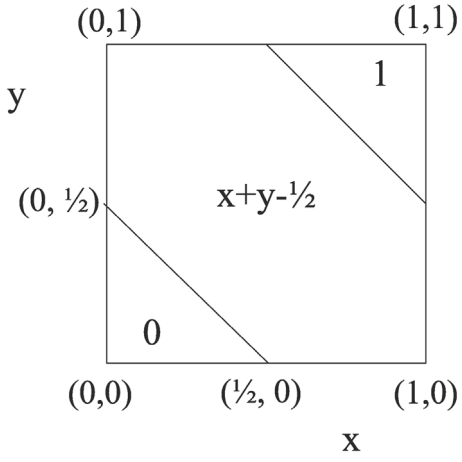

Fig. 7. The simplified graphical interpretation of (8) for Łukasiewicz t-norm and t-conorm.

higher values indicate full concern. Applying (8) and adjusting value 0.5 to edges of subintervals, we get functions shown in Fig. 6. The solution is shown in Table 2. We can observe that for two attributes ($n = 2$) this aggregation behaves as $AM_L(x, y) = \text{med}(0, 1, x + y - a)$.

Remark. Observe that the same aggregation (classification) space can be expressed by one function shown in Fig. 7.

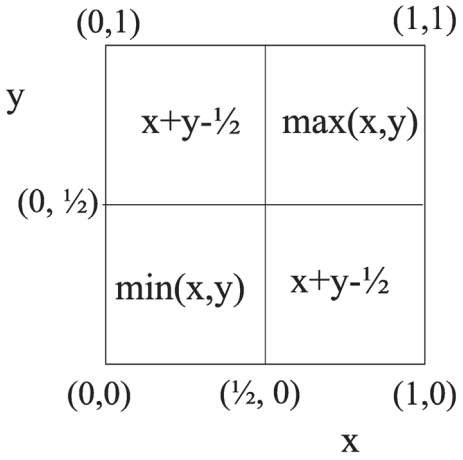

Fig. 8. The graphical interpretation of (8) for *MIN* and *MAX* functions.

Table 2
The classification solution when $a = 0.5$, A_1 is Łukasiewicz t-norm and A_2 is Łukasiewicz t-conorm.

Entity	x	y	Solution for *AM*
E1	0.2	0.75	0.45
E2	0.2	0.7	0.4
E3	0.2	0.2	0
E4	0.2	0.45	0.15
E5	0.8	0.9	1
E6	0.55	0.9	0.95
E7	1	0.7	1
E8	0	0.3	0
E9	1	0	0.5
E10	0.5	0.5	0.5
E11	0.5	0.2	0.2
E12	0.2	0.85	0.55
E13	0.3	0.7	0.5
E14	0	0	0
E15	1	1	1

Analogously as in Section 4.1.1, we can extent for other averaging functions.

4.1.3. Minimum t-norm and maximum t-conorm

Next, for idempotent conjunction (*MIN* function) and idempotent disjunction (*MAX* function) the graphical interpretation for arithmetic mean (8) in avg intervals is shown in Fig. 8, whereas the solution is in Table 3. Due to idempotency and a non-compensative effect, entities *E3* and *E4* are not distinguishable, but *E1* and *E2* are, because in this part of classification space we have averaging behaviour. This is the expected outcome as these two functions are limiting cases of conjunctive and averaging, and averaging and disjunctive functions. All subspaces in the classification space (Fig. 3) are idempotent.

Similarly as in Section 4.1.1, we can extent calculations for other averaging functions.

4.2. Further observations and variations of averaging functions

To summarize, with ordinal sums the classification is adjustable by assigning logic functions and an adjusting parameter a (when required) and therefore explainable. Strict, nilpotent or idempotent t-norms (resp. t-conorms) explain diverse classification requirements for class N (resp. Y). In addition, strict and nilpotent t-norms and t-conorms can be created by additive generators. It means, that we can further adjust functions to data by learning from data. In Section 5 we discuss this in more detail.

Table 3
The classification solution when $a = 0.5$, A_1 is *MIN* function and A_2 is *MAX* function.

Entity	**x**	**y**	Solution for *AM*
E1	0.2	0.75	0.45
E2	0.2	0.7	0.4
E3	0.2	0.2	0.2
E4	0.2	0.45	0.2
E5	0.8	0.9	0.9
E6	0.55	0.9	0.9
E7	1	0.7	1
E8	0	0.3	0
E9	1	0	0.5
E10	0.5	0.5	0.5
E11	0.5	0.2	0.2
E12	0.2	0.85	0.55
E13	0.3	0.7	0.5
E14	0	0	0
E15	1	1	1

The logical perspective of aggregation functions [37] considers global *andness* and *orness* where the arithmetic mean W is a logically neutral function, due to a full compensation effect (as arithmetic mean of conjunction and disjunction, where *andness* and *orness* values are equal to 0.5). The other averaging functions are either conjunctively or disjunctively polarized.

This implies that we are able to formalize diverse behaviours in class M (Fig. 3) and can logically explain it. We can observe that the same results are obtained when we apply the arithmetic mean regardless of different C and D in A_1 and A_2, respectively (Tables 1–3).

When using the geometric mean, we get an average behaviour with an absorbing element 0 (Table 1, column: solution for G). Let us check the geometric mean for the Łukasiewicz t-norm and t-conorm. Applying (11) for A_1 and A_2 shown in Fig. 6 we get

$$G_L(x,y) = \frac{\max(0, x+\frac{1}{2}-\frac{1}{2}) \cdot \min(1, y+\frac{1}{2}-\frac{1}{2})}{\frac{1}{2}}$$

and therefore $G_L(x,y) = 2x \cdot y$ which is an averaging function in the respective subintervals (see discussion in Section 4.1.1) and moreover value 0 is annihilator.

By this approach, the flexible classes *yes*, *no* and *maybe* are formalized by diverse conjunctive, disjunctive and averaging functions, respectively, and therefore can be explained by the logic properties of the chosen or learned functions.

4.3. A note to the internal and external validity (trustworthy) of the proposed aggregation

The internal validity shows that the results are less sensitive by the factors like a small imprecision in data. The similar entities are similarly treated, i.e., indicated by the intensities of belonging to the classes. For class M we see that small changes in data causes slight change in inclination towards Y or N. Next, the boundary conditions, a key requirement in aggregation is satisfied (see entities *E14* and *E15* in all tables). The monotonicity (if matching degree of one atomic predicate increases, whereas the other remains the same, the solution remains the same or increases) is a matter of direct verification.

The external validity generalizes to the other situations and data sets. It is a matter of various experiments on diverse larger data sets, but the initial observations reveal, whether we should do these experiments. When someone choose a large number of attributes, the limitation is the computational capacity. Mathematically, aggregation is not limited for $n \gg 2$ (we might create hierarchy of attributes). A higher number of entities is always good for achieving better learning results.

Furthermore, data incompleteness appears in the data sets. A topic for future research is adjusting the *missingness*-tolerant evaluation suggested for the logic scores of preferences [37]. Generally, we can assign values from 0 (full penalty for missingness) to 1 (full tolerance). This penalty presumably cannot be learnt from the known observations because it appears infrequently. Consequently, the human-in-the-loop (see next Section) is here very desirable. Considering other data types (e.g., images and short texts) the main problem is in the transformation into the unit interval to express relevance or severeness of findings in the attribute (e.g., intensity of colour in images or intensity of warnings in texts). These theoretical observations reveal as serious issues which should be in the focus of further research. In future work we will carry out extensive experiments on real-world data, see Section 6.

5. Aggregation functions in classification via interactive machine learning

AI became amazingly successful due to the huge success of statistical machine learning, particularly deep learning [44]. Current limitations are mainly due to lack of explainability [45] and the fact that these algorithms are extremely data hungry and the labelling of data is extremely costly in the medical domain. Consequently, the grand challenge of the future is in learning from little data. For example, with the best performing methods to date the classification of less-frequent illnesses (e.g. rare diseases), where we have a small amount of data sets cannot be properly learnt. One possibility to solve these problems is to follow a human-in-the-loop approach [16]. The human domain expert may be beneficial here in order to provide valuable information regarding the aggregation or classification into the learning process pipeline [46].

For example Convolutional Neural Networks (CNN) can be used as very effective classifiers, however, the classification layer is hidden and therefore the reasons for activating particular nodes and obtained solutions remain hidden to a human. The explainable AI community contributes with a variety of methods [47], one recent approach, Layer-wise Relevance Propagation [48] is applicable to various machine learning tasks providing a general framework for explaining predictions [49]. Classification by fuzzy IF–THEN rules is explainable by design, but users should provide a set of rules and assign parameters to each fuzzy set, which might be a tedious task. Hybrid systems also exist, but they require domain experts to intervene and a significant size of the input–output data make this also very cumbersome. In our work, we examine explainability by the properties of aggregation functions.

In the classification by aggregation functions, the classification space is visible to the user, who should assign values from the domain of considered attributes to values of 0 and 1 (see, Fig. 3). Domain experts are usually well aware of these values. In Business, for instance, this is often easy because even the workers recognize that the number of sold items lower or equal to D_L is without any doubt a weak performance, whereas the number of sold items above or equal to D_H is an excellent performance. The same might hold for the other attributes. In a medical case, for instance, the first attribute can be blood pressure, the second attribute the level of bad cholesterol, and so on. Having all attributes in the "red area" means full concern, whereas all values in a "green area" means no worry. Similarly, if one attribute is in the "red area" and the other is very close to this area, there is full concern. If all attributes are in the middle of the classification space, the concern is around 0.5, and so on.

When the set of input–output data is available, the learning is focused on adjusting data to the most suitable functions. In this section we consider functions from Section 4 covering all classes.

5.1. Adjusting functions

In the case of ordinal sums of conjunctive and disjunctive functions (6), we need to adjust the most suitable functions from $\mathcal{C}$ and $\mathcal{D}$. In this challenging task the most suitable value for parameter a, and the most suitable function can be learnt from the input–output data. Observe that the smallest t-norm is a drastic product and the largest is the min t-norm, i.e., $T_D(\mathbf{x}) \leq T(\mathbf{x}) \leq T_M(\mathbf{x})$. For the dual case (class Y) holds $S_M(\mathbf{x}) \leq S(\mathbf{x}) \leq S_D(\mathbf{x})$. Next, for the averaging part it holds $T_M(\mathbf{x}) = \min(\mathbf{x}) \leq Av(\mathbf{x}) \leq \max(\mathbf{x}) = S_M(\mathbf{x})$.

The theory offers several parametrized families of t-norms and t-conorms. More about these families can be found in e.g., [35,41]. These families usually cover basic t-norms as limiting cases. Several families do not cover both strict and nilpotent behaviour, or one of these behaviours is only for particular value of parameter. A suitable family for our purpose is from Schweizer & Sklar (1961) [50]. The family of t-norms is given as

$$T_\lambda^S(x,y) = \begin{cases} \min(x,y) & \text{for } \lambda = -\infty \\ T_P(x,y) & \text{for } \lambda = 0 \\ T_D(x,y) & \text{for } \lambda = \infty \\ (\max(x^\lambda + y^\lambda - 1, 0))^{\frac{1}{\lambda}} & \text{otherwise} \end{cases} \tag{13}$$

The limiting cases are: minimum t-norm for $\lambda = -\infty$, product t-norm for $\lambda = 0$, Łukasiewicz t-norm for $\lambda = 1$, Hamacher product ($\frac{xy}{x+y-xy}$) for $\lambda = -1$ and drastic product for $\lambda = \infty$. This family covers class N (see Fig. 3).

When we want to adapt (13) to the $[0, 0.5]$ interval we get $\min(x, y)$ for $\lambda = -\infty$, $2xy$ for $\lambda = 0$, $\min(x, y)$ if $\max(x, y) = 0.5$ and 0 otherwise for $\lambda = \infty$ and $(\max(x^\lambda + y^\lambda - 2^\lambda, 0))^{\frac{1}{\lambda}}$ for the remaining values of λ.

Analogously, the family of t-conorms is given as

$$S_\lambda^S(x,y) = \begin{cases} \max(x,y) & \text{for } \lambda = -\infty \\ S_P(x,y) & \text{for } \lambda = 0 \\ S_D(x,y) & \text{for } \lambda = \infty \\ 1 - (\max((1-x)^\lambda + (1-y)^\lambda - 1, 0))^{\frac{1}{\lambda}} & \text{otherwise} \end{cases} \tag{14}$$

Similarly, we can adapt this family to the $[0.5, 1]$ interval.

Thus, by learning λ from the input–output data sets, we are able to recall the nature of the classification (strict, nilpotent, *MIN*, non-continuous). Next, an ordinal sum A_1 and A_2 do not need necessarily to be dual. If for t-norm we get $\lambda > 0$, low values lead to the clear non-concern, whereas when we get $\lambda \leq 0$ for t-conorm, it means that high values do not lead to clear full-concern. Hence, this parameter reflects the behaviour and moreover explains it.

Following this approach we are able to reveal whether a limiting case of λ interprets the classification (as shown in Tables 1–3) or the best behaviour is between these limiting cases.

Functions belonging to the class $\mathcal{A}_\mathcal{V}$ can be expressed as power mean [35,37]:

$$A(x,y) = (0.5x^r + 0.5y^r)^{\frac{1}{r}}, \quad -\infty \leq r \leq \infty \tag{15}$$

on $[0, 0.5] \times [0.5, 1] \cup [0.5, 1] \times [0, 0.5]$ which leads to $(x^r + y^r - 0.5^r)^{\frac{1}{r}}$, $r \in]-\infty, 0[\cup]0, \infty[$.

For $r = -\infty$ we get *MIN*, whereas for $r = 0$ we get geometric mean, for $r = 1$ we get arithmetic mean. Finally, for $r = \infty$ we get *MAX*. The most suitable value of this parameter could be learnt from the input–output data sets.

This leads to the transparent and explainable classification into three classes marked as *yes*, *no*, *maybe* with flexible belonging, i.e., indicating the inclination to *yes* or *no*. For each new entity we can explain where it belongs, as well as how far it is from the clear *yes* or *no*.

5.2. Explainability by summarized sentences

Apart from the explainability discussed above, we can use linguistic summaries to explain the behaviour of the classified data and data distribution. Examples include, *most of the patients having higher values of Atr1 and Atr2 and Atr3 have a high possibility of illness I* and *most of the patients having very high values of Atr1 and Atr2 have for sure illness I*. These summaries can be an enormously useful feedback for medical doctors, or manager in business tasks, to learn from the solution or adjust values D_L and D_H and continue with examinations and/or experiments.

The next task should lead to learning the most suitable transformation $f : [D_L, D_F] \to [0, 1]$ for significantly non-uniform data distributions of input attributes, for other data types (e.g., images – where for instance intensity of a hue or colour indicates concern; text – where the number of worrying terms and their intensities (adjectives and adverbs) indicate concern; and imprecise data-values cannot be measured, but it is most likely *m*, not lower than *a* and not higher than *b*), and for the cases when low and high values indicate concern, but medium values do not. Consequently, a linear transformation has been examined in our work, and data distribution explained by linguistic summaries might be an input for the learning process. In any case this is valuable information for domain experts, but such summaries can also be learnt, as indicated in [51].

Sentences such as *the most of entities belong to class N* or *few entities belong to class N* might also express non-expected situations which should be examined to reveal, whether training data sets are not of a sufficient quality (e.g., do not cover the whole range of possible values) or problem is in transformation to $[0, 1]$ (Fig. 3).

5.3. Human perspectives

It is important to discriminate between explainability and causability. Explainability in a technical sense indicates decision-relevant parts of the used representations of an algorithm and of active parts in an algorithmic model, which contribute to the models accuracy on the training set, or to a specific prediction for one particular observation, and it does *not* refer to an explicit human model. On the other hand — Causability (the word comes from usability, because usability is an already well-known concept in software engineering [52]), is the extent to which an explanation of a statement to a human expert achieves a specified *level of causal understanding* [46]. This can be measured (similarly as usability can be measured) with effectiveness, efficiency and satisfaction in a specified context of use with the Systems Causability Scale [53]. In an ideal world human explanations and machine "explanations" are identical, and congruent with the ground truth, which is defined equally for both humans and machines. One important aspect for future human-AI interfaces is to map explainability with causability and to allow domain experts to interactively ask questions in order to get a deeper understanding why an AI came up with a result to gain insight into the underlying *independent* explanatory factors of a result [54]. In many domains, e.g. in the medical domain, various modalities contribute to a single result, which calls for multi-modal causability [55].

However, in the real world we face two problems: (1) ground truth, particularly in the medical domain is not well defined, especially when making a medical diagnosis; and (2) although human (scientific) models are often based on understanding causal mechanisms, today's successful machine learning is either model-free or the models or algorithms are typically based on correlation (and correlation is not causality in the sense of Judea Pearl [56], or related concepts of similarity and distance [46]). That means

given the data and specific requirements we can model a function that explains the output. Hence, the aspects of aggregation functions consisting of conjunctive and disjunctive ones can help with such problems. Consequently, our approach both expands and empowers the scope of classification into three overlapping classes by aggregation functions and discuss its applicability in machine learning for business and for the health domain.

We emphasize that this work is a theoretical one, therefore in our future work we plan to extent this work on experiments with real-world use-cases and data sets, refer to Section 6 Future Work for details.

The previously discussed classification tasks can be extended into various directions. The simplest one is adding input attributes. This approach works for any number of input attributes due to its associativity. However, it might become less human legible and manageable. Elementary attributes often belong to a group of similar ones. It leads to hierarchical aggregation of input attributes of the classification space. For instance, in the medical domain we have motor abilities and sensory feelings [37]. For the aggregation of elementary attributes into the category we have plenty of aggregation functions, which can be adjusted to the needs of the respective problem. Domain experts might pose mandatory, optional and sufficient requirements as well as preferred coalitions among the subsets of requirements, symptoms or observed values including:

IF $\bar{P}_1$ *is low* ***and*** *(*P_2 ***or else*** P_3 *is high)* ***and (most of*** *{*P_4*...* P_{10}*} is highly satisfied) THEN concern is high*

where P_1 is the first attribute, (P_2 or else P_3) is the second and the third one is a quantified aggregation, i.e., the classification space (Fig. 3) is three dimensional.

Furthermore, relative importance among predicates (or coalitions) might be considered, i.e., highly satisfied P_a *and* P_b is more serious than highly satisfied P_c and P_d (i.e., aggregation of elementary predicates by the Choquet integral, [57]) in $\bar{P}_1$. Next, in *OR ELSE* connective we are able to formalize the intensity of importance of optional requirement by aggregation functions [58].

As already mentioned, machine learning methods generally and deep learning methods particularly, are very data hungry [59,60] and the domain experts are usually not familiar with the mathematical formalization of their tasks. However, they are able to explain the expected aggregation linguistically [51]. Such domain expert linguistic explanations are required to recognize the most suitable subclass of aggregation functions, and most important, via machine learning the *most suitable function and its parameters* can be learnt. The answer could be provided by fitting empirical data to the recognizing subset of possible functions from user's linguistic explanations. This is formalized as [35]:

$$\begin{array}{c} \min \|\mathbf{r}\| \\ \text{subject to} \\ f \text{ satisfying } G_i,\ i = 1, \ldots, n \end{array} \tag{16}$$

where $\mathbf{r}$ is the norm of the vector of residuals (difference between the value calculated by function and expected output) and G_i is ith property of function f. Thus, a smaller amount of training input–output data for learning the most suitable functions and their respective parameters might suffice. This is a topic for our future research activities in augmenting the suggested classification space. In this way, we can also trace backward from output class to the elementary attributes and this would be a great contribution to the explainable AI community.

6. Future work

In this paper, we provide theoretical foundations in classification by ordinal sums of conjunctive and disjunctive functions for the future benefit of explainable AI. These results contribute to the ongoing debate between AI, ML, fuzzy rule-based systems, and aggregation functions communities to assess these concepts. This is highly relevant, because such systems should be optimized not only for pure accuracy [61] but also for other important – often ignored – criteria [62] including explainability which supports robustness, fairness, unbiasedness, trust, privacy and reliability.

Our future work will be focused on experiments on real-world data. Parameter learning in fuzzy systems can be achieved with evolutionary methods, as presented in [62]. Several works that are described need the computation of a multi-objective fitness function that strives to improve both performance and interpretability (by lowering the number and complexity of the fuzzy rules) at the same time. As stated in [62], to compute such a fitness function from data, is not a straightforward task. We will use a Reinforcement Learning solution with an informative reward strategy, so that this problem can be tackled — provided that the state and action space remain tractable. Deep Reinforcement Learning [63,64] can also be a viable technical solution, under the assumption that similar states are assumed to have the same expected reward (in the long term), so they do not need to be explicitly visited. Thereby, data can support the search of the most suitable parameters of t-norms and t-conorms and we can also interpret it whether it is strict or nilpotent. Furthermore, since Eq. (16) expresses the optimization of the proposed aggregation functions, it can be learned theoretically with the use of Reinforcement Learning. The rules of a learned fuzzy system are a list of interpretable IF-THEN rules. Current state-of-the-art explainable AI methods [65] focus on explanations that answer the question "Why not?" and could be expressed by statements "IF not, THEN ...". Here we envision counterfactuals of fuzzy rules as a promising approach when we apply a suitable negation (between the Gödel negation and its dual negation).

Some very remarkable approaches for the training of fuzzy rule based-systems and overcoming their weak points have been proposed in very recent work, demonstrating that this is a hot topic, [66–70]. They empower explainability and the ability to handle vagueness with accuracy and reliability, especially welcomed for the regulated applications (e.g. for our applications in the medical or life sciences domain). Here, type-2 fuzzy sets provide more freedom in creating rules [66], min–max probability Takagi–Sugeno–Kang fuzzy system determines reliability [70]. A next viable approach is the construction of the concise fuzzy systems of the reduced number of rules by the Lasso algorithm [68] and the integration of the various data and expert views (human-in-the-loop) on the problem [67] which improves legibility and explainability of a rule base.

On the other hand, it is not always suitable (or even possible) to simulate human inference mechanisms directly by rule base when a domain expert explains aggregations among input attributes linguistically and by conjunctive, disjunctive and mixed behaviour. Therefore, in our work we formalized the inference by the aggregation functions to match input data with three output classes (*yes*, *no*, *maybe*).

For more future work we envision to focus on extracting rules from the inference modelled by aggregation functions (i.e., relevant subareas could be converted into the fuzzy rules [69] of the Takagi–Sugeno–Kang fuzzy system due its intrinsic interpretability of the rule base and learning ability [68] and consequently learnt by the above discussed approaches. Next, the linguistic interpretation can be provided by several experts (or expert groups, a crowd of humans-in-the-loop), which leads to the multiple views where linguistic interpretation of aggregation and/or different input attributes are assigned by the human experts. Recent work [67] discusses the fusion of single views by a rule-based

system. The analogous work of fusion of multiple views computed by aggregation functions, which addresses an assumption of dissimilarity of views and may exploit consistency among different views, is also a topic for future work. Generally, the mutual benefit of classification by rule-based systems and by ordinal sums is a promising topic for future work. In this work, we formalized classification by ordinal sums as a novel approach. In future works, the synergy between them will be further evaluated.

7. Conclusion

In our work we have proposed classification into three classes *yes*, *no*, *maybe* by ordinal sums of conjunctive and disjunctive functions. Classification by machine learning is effective, but generally non-explainable. On the other hand, classification by rule-based systems is explainable in principle, but practically often lack human interpretability due to the high complexity. Consequently, our research proposes a framework for explainability of machine learning by ordinal sums.

In our work, we firstly converted classification into three classes from a rule-based system into the classification space managed by the ordinal sums of conjunctive and disjunctive functions. Secondly, we formalized diverse requirements for belonging to the respective classes by conjunctive, disjunctive and averaging functions. Finally, due to mixed behaviour of the ordinal sums we are able to classify into three classes *yes*, *no* and *maybe* when the averaging behaviour is activated.

Our research direction merges benefits of machine learning and explainability by aggregation functions, which will allow to make classification tasks more explainable for domain experts and at the same time less demanding for machine learning algorithms. Firstly, domain experts assign information for clear 0 and 1 (Fig. 3). Secondly, the learning process recognizes the most suitable disjunctive and conjunctive functions, which reveal the nature of classification into classes *yes* and *no*. Finally, we have considered the logical behaviour of averaging functions to reveal the nature of classification into the class *maybe*.

The next direction should lead to learning the most suitable transformation $f : [D_L, D_F] \to [0, 1]$ especially for significant non-uniform data distribution of input attributes and for other data types. In this direction, data distribution explained by linguistic summaries might be an enormously valuable input for learning and, at the same time, also very valuable information for the domain experts. Another direction is experimenting on real-world data. We have raised cases in business and health. A support for this direction is also observation of internal and external validity (trustworthy), which opens a lot of further research avenues.

Article summary

What is already known on the topic to the international research community?

(1) The commonalities and differences between fuzzy logic systems and neural networks in classification;
(2) The explainability and interpretability possibilities of fuzzy logic systems and neural networks in general, as well as their effects in the decision-making process of the domain experts;
(3) The need for combination of different methodologies that support and extend the functionality of neural networks towards more explainability;
(4) Ordinal sums are able to cover different behaviour of the input variables and reflect it in the output space.

What this paper contributes to the international research community?

(1) A novel approach how fuzzy logic and ordinal sums are coupled with the (usually) separable target classes useful in machine learning classification problems;
(2) The theoretical basis for practical applications of the proposed method that clarifies the difference in explainability between the invented rule-based system and neural networks;
(3) What are the advantages/disadvantages in explainability and interpretability for domain experts.
(4) Formalization of ordinal sums of conjunctive, disjunctive and averaging functions for the classification purposes.

CRediT authorship contribution statement

Miroslav Hudec: Conceptualization, Formal analysis, Methodology, Writing - original draft, Writing - review & editing. **Erika Mináriková:** Formal analysis, Methodology. **Radko Mesiar:** Formal analysis, Methodology. **Anna Saranti:** Investigation, Methodology. **Andreas Holzinger:** Conceptualization, Methodology, Writing - original draft, Writing - review & editing, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We are very grateful for the valuable comments of the anonymous reviewers and their encouragement for further work. Parts of this work have been funded by the Austrian Science Fund (FWF), Project: P-32554 explainable Artificial Intelligence. Moreover, the support of the projects VEGA 1/0466/19 and VEGA 1/0006/19 by the Ministry of Education, Science, Research and Sport of the Slovak Republic are kindly appreciated.

References

[1] M. Carter, Minds and Computers: An Introduction to the Philosophy of Artificial Intelligence, Edinburgh University Press, Edinburgh, UK, 2007.
[2] E.H. Shortliffe, B.G. Buchanan, A model of inexact reasoning in medicine, Math. Biosci. 23 (3–4) (1975) 351–379, http://dx.doi.org/10.1016/0025-5564(75)90047-4.
[3] A. Collins, D. Gentner, How people construct mental models, in: D. Holland, N. Quinn (Eds.), Cultural Models in Language and Thought, Cambridge University Press, Cambridge, 1987, pp. 243–265.
[4] J. Von Neumann, O. Morgenstern, Theory of Games and Economic Behavior, Princeton University Press, Princeton, 1944.
[5] A. Tversky, D. Kahneman, The framing of decisions and the psychology of choice, Science 211 (4481) (1981) 453–458, http://dx.doi.org/10.1126/science.7455683.
[6] A. Tversky, D. Kahneman, Judgment under uncertainty: Heuristics and biases, Science 185 (4157) (1974) 1124–1131, http://dx.doi.org/10.1126/science.185.4157.1124.
[7] D. Kahneman, Thinking, Fast and Slow, Macmillan, 2011.
[8] G. Marcus, The next decade in ai: four steps towards robust artificial intelligence, 2020, pp. 1–59, arXiv:2002.06177.
[9] J. Russin, R.C. O'Reilly, Y. Bengio, Deep learning needs a prefrontal cortex, in: Workshop Bridging Ai and Cognitive Science, ICLR, 2020.
[10] L.A. Zadeh, Fuzzy logic = computing with words, IEEE Trans. Fuzzy Syst. 4 (2) (1996) 103–111.
[11] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.
[12] T. Araújo, G. Aresta, E. Castro, J. Rouco, P. Aguiar, C. Eloy, A. Polónia, A. Campilho, Classification of breast cancer histology images using convolutional neural networks, PLoS One 12 (6) (2017) e0177544, http://dx.doi.org/10.1371/journal.pone.0177544.

[13] A. Esteva, B. Kuprel, R.A. Novoa, J. Ko, S.M. Swetter, H.M. Blau, S. Thrun, Dermatologist-level classification of skin cancer with deep neural networks, Nature 542 (7639) (2017) 115–118.
[14] R.M. French, Catastrophic forgetting in connectionist networks, Trends Cogn. Sci. 3 (4) (1999) 128–135, http://dx.doi.org/10.1016/s1364-6613(99)01294-2.
[15] J.M. Alonso, C. Castiello, C. Mencar, Interpretability of fuzzy systems: Current research trends and prospects, in: Springer Handbook of Computational Intelligence, Springer, 2015, pp. 219–237.
[16] A. Holzinger, Interactive machine learning for health informatics: When do we need the human-in-the-loop? Brain Inform. 3 (2) (2016) 119–131, http://dx.doi.org/10.1007/s40708-016-0042-6.
[17] D. Girardi, J. Küng, R. Kleiser, M. Sonnberger, D. Csillag, J. Trenkler, A. Holzinger, Interactive knowledge discovery with the doctor-in-the-loop: a practical example of cerebral aneurysms research, Brain Inform. 3 (3) (2016) 133–143, http://dx.doi.org/10.1007/s40708-016-0038-2.
[18] M. Hund, D. Boehm, W. Sturm, M. Sedlmair, T. Schreck, T. Ullrich, D.A. Keim, L. Majnaric, A. Holzinger, Visual analytics for concept exploration in subspaces of patient groups: Making sense of complex datasets with the doctor-in-the-loop, Brain Inform. 3 (4) (2016) 233–247, http://dx.doi.org/10.1007/s40708-016-0043-5.
[19] A. Holzinger, M. Plass, M. Kickmeier-Rust, K. Holzinger, G.C. Crişan, C.-M. Pintea, V. Palade, Interactive machine learning: experimental evidence for the human in the algorithmic loop, Appl. Intell. 49 (7) (2019) 2401–2414.
[20] M. Vučetić, M. Hudec, B. Božilović, Fuzzy functional dependencies and linguistic interpretations employed in knowledge discovery tasks from relational databases, Eng. Appl. Artif. Intell. 88 (2020) 103–118, http://dx.doi.org/10.1016/j.engappai.2019.103395.
[21] C.L. Paris, Generation and explanation: Building an explanation facility for the explainable expert systems framework, in: Natural Language Generation in Artificial Intelligence and Computational Linguistics, Springer, 1991, pp. 49–82.
[22] J. McCarthy, Programs with common sense, in: Teddington Conference on the Mechanization of Thought Processes, Her Majesty's Stationery Office London, 1959, pp. 75–91, http://www-formal.stanford.edu/jmc/mcc59/mcc59.html.
[23] E.H. Shortliffe, R. Davis, S.G. Axline, B.G. Buchanan, C.C. Green, S.N. Cohen, Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the mycin system, Comput. Biomed. Res. 8 (4) (1975) 303–320, http://dx.doi.org/10.1016/0010-4809(75)90009-9.
[24] E. Merdivan, D. Singh, S. Hanke, A. Holzinger, Dialogue systems for intelligent human computer interactions, Electron. Notes Theor. Comput. Sci. 343 (2019) 57–71, http://dx.doi.org/10.1016/j.entcs.2019.04.010.
[25] A. Holzinger, P. Kieseberg, E. Weippl, A.M. Tjoa, Current advances, trends and challenges of machine learning and knowledge extraction: From machine learning to explainable ai, in: Springer Lecture Notes in Computer Science LNCS 11015, Springer, Cham, 2018, pp. 1–8, http://dx.doi.org/10.1007/978-3-319-99740-7-1.
[26] A. Holzinger, From machine learning to explainable ai, in: 2018 World Symposium on Digital Intelligence for Systems and Machines (DISA), IEEE, 2018, pp. 55–66, http://dx.doi.org/10.1109/DISA.2018.8490530.
[27] M. Hudec, Fuzziness in Information Systems, Springer International Publishing.
[28] M. Hudec, M. Vujošević, Integration of data selection and classification by fuzzy logic, Expert Syst. Appl. 39 (10) (2012) 8817–8823, http://dx.doi.org/10.1016/j.eswa.2012.02.009.
[29] A. Meier, N. Werro, M. Albrecht, M. Sarakinos, Using a fuzzy classification query language for customer relationship management, in: Proceedings of the 31st International Conference on Very Large Data Bases, 2005, pp. 1089–1096.
[30] J. de Valente Oliveira, Semantic constraints for membrship function optimization, IEEE Trans. Syst. Man Cybern. 29 (1) (1999) 128–138.
[31] C. Fuchs, S. Spolaor, M. Nobile, U. Kaymak, A graph theory approach to fuzzy rule base simplification, in: Proceedings of the 18th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2020), 2020, pp. 387–401.
[32] C. Mencer, A. Fanelli, Interpretability constraints for fuzzy information granulation, Inform. Sci. 178 (24) (2008) 4585–4618.
[33] R. Słowiński, S. Greco, B. Matarazzo, Rough set methodology for decision aiding, in: Springer HandBook of Computational Intelligence, Springer, 2015, pp. 349–370.
[34] A. Holzinger, M. Plass, K. Holzinger, G.C. Crisan, C.-M. Pintea, V. Palade, A glass-box interactive machine learning approach for solving np-hard problems with the human-in-the-loop, arXiv:1708.01104.
[35] G. Beliakov, A. Pradera, T. Calvo, et al., Aggregation Functions: A Guide for Practitioners, Vol. 221, Springer, 2007.
[36] D. Dubois, H. Prade, On the use of aggregation operations in information fusion processes, Fuzzy Sets and Systems 142 (1) (2004) 143–161.
[37] J. Dujmović, Soft Computing Evaluation Logic: The LSP Decision Method and Its Applications, John Wiley & Sons, 2018.
[38] H.-J. Zimmermann, P. Zysno, Decisions and evaluations by hierarchical aggregation of information, Fuzzy Sets and Systems 10 (1–3) (1983) 243–260, http://dx.doi.org/10.1016/S0165-0114(83)80118-3.
[39] G. Birkhoff, Lattice Theory, third ed., Vol. XXV, AMS Colloqium publications, American Mathematical Society, Providence, 1967.
[40] A. Clifford, Naturally totally ordered commutative semigroups, Amer. J. Math. 76 (1954) 631–646, http://dx.doi.org/10.2307/2372706.
[41] E.P. Klement, R. Mesiar, E. Pap, Triangular Norms, Kluwer, Dordrecht, 2000.
[42] B. De Baets, R. Mesiar, Ordinal sums of aggregation operators, in: Technologies for Constructing Intelligent Systems, Vol. 2, Springer, 2002, pp. 137–147.
[43] F. Durante, C. Sempi, Semicopulae, Kybernetika 41 (3) (2005) 315–328.
[44] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444, http://dx.doi.org/10.1038/nature14539.
[45] A.B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, Inf. Fusion 58 (2020) 82–115, http://dx.doi.org/10.1016/j.inffus.2019.12.012.
[46] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, H. Müller, Causability and explainability of artificial intelligence in medicine, Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 9 (4) (2019) e1312.
[47] G. Montavon, W. Samek, K.-R. Müller, Methods for interpreting and understanding deep neural networks, Digit. Signal Process. 73 (2) (2018) 1–15, http://dx.doi.org/10.1016/j.dsp.2017.10.011.
[48] S. Lapuschkin, A. Binder, G. Montavon, K.-R. Müller, W. Samek, The lrp toolbox for artificial neural networks, J. Mach. Learn. Res. 17 (1) (2016) 3938–3942.
[49] W. Samek, K. Müller, Towards explainable artificial intelligence, in: W. Samek, G. Montavon, A. Vedaldi, L. Hansen, K. e. Müller (Eds.), Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Springer, Cham, 2019, pp. 5–22.
[50] B. Schweizer, A. Sklar, Associative functions and triangle inequalities, Publ. Math. Debrecen 8 (1961) 169–186.
[51] M. Hudec, E. Bednárová, A. Holzinger, Augmenting statistical data dissemination by short quantified sentences of natural language, J. Off. Stat. 34 (4) (2018) 981–1010.
[52] A. Holzinger, M. Errath, G. Searle, B. Thurnher, W. Slany, From extreme programming and usability engineering to extreme usability in software engineering education, in: 29th International Annual IEEE Computer Software and Applications Conference (IEEE COMPSAC 2005), IEEE, 2005, pp. 169–172, http://dx.doi.org/10.1109/COMPSAC.2005.80.
[53] A. Holzinger, A. Carrington, H. Müller, Measuring the quality of explanations: the system causability scale (scs), KI-Künstliche Intelligenz (2020) 1–6.
[54] A. Holzinger, Explainable ai and multi-modal causability in medicine, Wiley i-com J. Interact. Media 19 (3) (2020) 171–179, http://dx.doi.org/10.1515/icom-2020-0024.
[55] A. Holzinger, B. Malle, A. Saranti, B. Pfeifer, Towards multi-modal causability with graph neural networks enabling information fusion for explainable ai, Information Fusion 71 (7) (2021) 28–37, http://dx.doi.org/10.1016/j.inffus.2021.01.008.
[56] J. Pearl, Causality: Models, Reasoning, and Inference, second ed., Cambridge University Press, Cambridge, 2009.
[57] T. Murofushi, M. Sugeno, An interpretation of fuzzy measures and the choquet integral as an integral with respect to a fuzzy measure, Fuzzy Sets Syst. 29 (2) (1989) 201–227, http://dx.doi.org/10.1016/0165-0114(89)90194-2.
[58] M. Hudec, R. Mesiar, The axiomatization of asymmetric disjunction and conjunction, Inf. Fusion 53 (2020) 165–173.
[59] T. van der Ploeg, P.C. Austin, E.W. Steyerberg, Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints, BMC Med. Res. Methodol. 14 (1) (2014) 137.
[60] G. Marcus, Deep learning: A critical appraisal, arXiv:1801.00631.
[61] S.M. McNee, J. Riedl, J.A. Konstan, Being accurate is not enough: how accuracy metrics have hurt recommender systems, in: G. Olson, R. Jeffrey (Eds.), CHI'06 Extended Abstracts on Human Factors in Computing Systems, 2006, pp. 1097–1101, http://dx.doi.org/10.1145/1125451.1125659.
[62] A. Fernandez, F. Herrera, O. Cordon, M.J. del Jesus, F. Marcelloni, Evolutionary fuzzy systems for explainable artificial intelligence: Why, when, what for, and where to? IEEE Comput. Intell. Mag. 14 (1) (2019) 69–81, http://dx.doi.org/10.1109/MCI.2018.2881645.
[63] L. Graesser, W.L. Keng, Foundations of Deep Reinforcement Learning: Theory and Practice in Python, Addison-Wesley Professional, 2019.
[64] P. Winder, Reinforcement Learning, O'Reilly Media, 2020.
[65] P. Madumal, T. Miller, L. Sonenberg, F. Vetere, Explainable reinforcement learning through a causal lens, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 3, 2020, pp. 2493–2500.

[66] R. Chimatapu, H. Hagras, M. Kern, G. Owusu, Hybrid deep learning type-2 fuzzy logic systems for explainable ai, in: 2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), IEEE, 2020, pp. 1–6, http://dx.doi.org/10.1109/FUZZ48607.2020.9177817.

[67] T. Zhang, Z. Deng, D. Wu, S. Wang, Multiview fuzzy logic system with the cooperation between visible and hidden views, IEEE Trans. Fuzzy Syst. 27 (6) (2018) 1162–1173, http://dx.doi.org/10.1109/TFUZZ.2018.2871005.

[68] P. Xu, Z. Deng, C. Cui, T. Zhang, K.-S. Choi, S. Gu, J. Wang, S. Wang, Concise fuzzy system modeling integrating soft subspace clustering and sparse learning, IEEE Trans. Fuzzy Syst. 27 (11) (2019) 2176–2189, http://dx.doi.org/10.1109/TFUZZ.2019.2895572.

[69] J. Zhang, Z. Deng, K.-S. Choi, S. Wang, Data-driven elastic fuzzy logic system modeling: Constructing a concise system with human-like inference mechanism, IEEE Trans. Fuzzy Syst. 26 (4) (2018) 2160–2173, http://dx.doi.org/10.1109/TFUZZ.2017.2767025.

[70] Z. Deng, L. Cao, Y. Jiang, S. Wang, Minimax probability tsk fuzzy system classifier: A more transparent and highly interpretable classification model, IEEE Trans. Fuzzy Syst. 23 (4) (2014) 813–826, http://dx.doi.org/10.1109/TFUZZ.2014.2328014.