Proportions analogiques entre probabilités

Henri Prade[1] Gilles Richard[1],

[1] IRIT, CNRS & Université Toulouse III - Paul Sabatier, 118 route de Narbonne, 31062 Toulouse cedex 9, France

{henri.prade, gilles.richard}@irit.fr

Résumé

Les proportions analogiques sont des relations qui lient 4 items a, b, c et d et qui s'énoncent "a est à b comme c est à d". Ces 4 items sont souvent décrits par des vecteurs de valeurs booléennes, nominales, ou numériques. Les proportions analogiques peuvent aussi mettre en relation des formules logiques. L'article propose une première étude des proportions analogiques entre probabilités, qu'elles soient simplement entre des valeurs, ou entre des distributions (ce qui exige la préservation de leur normalisation). Les propriétés de définitions à base de proportion arithmétique, ou combinant cette dernière avec la proportion géométrique sont étudiées, et des utilisations potentielles décrites. Une annexe propose une preuve du théorème de Pythagore en termes de proportions géométriques.

Mots-clés

proportion analogique, probabilité, proportion numérique

Abstract

Analogical proportions are relations that link 4 items a, b, c and d and that are expressed as "a is to b as c is to d". These 4 items are often described by vectors of Boolean, nominal, or numerical values. Analogous proportions can also relate logical formulas. The article proposes a first study of analogous proportions between probabilities, whether they are simply between values, or between distributions (which requires the preservation of their normalization). The properties of definitions based on arithmetic proportion, or combining the latter with geometric proportion are studied, and potential uses are described. An appendix proposes a proof of the Pythagorean theorem in terms of geometric proportions.

Keywords

analogical proportion, probability, numerical proportion.

1 Introduction

Les analogies et les probabilités ne sont pas souvent considérées conjointement, bien qu'elles soient toutes deux liées à l'induction [15]. De fait déjà dans les années 1930, Janina Hosiasson-Lindenbaum [18], une philosophe, spécialiste du raisonnement probabiliste et logicienne, considérait le raisonnement analogique suivant : si deux conjectures f_1 et f_2 sont conséquences d'une même hypothèse h, l'observation de f_1 (qui devient donc un fait) induit h et augmente la croyance en la conjecture f_2. Elle cherchait des hypothèses additionnelles lui permettant une justification probabiliste de cette augmentation de croyance. Dans le même esprit, plus tard, Polya [11] considérera que f_1 et f_2 sont analogues s'ils sont impliqués par une hypothèse commune.

Dans cet article on s'intéresse à des analogies exprimées sous forme de proportions analogiques[1] c'est-à-dire des énoncés de la forme "a est à b comme c est à d". Tout comme l'inférence probabiliste bayésienne, l'inférence basée sur les proportions analogiques permet de faire de la classification avec succès [4, 2]. En classification, a, b, c, d sont des vecteurs de valeurs d'attributs décrivant les items à classer. Les attributs peuvent être booléens, nominaux, ou numériques. Cependant, les proportions analogiques peuvent être étendues à des cadres de représentation généraux comme la logique propositionnnelle [16]. Dans cet article, nous considérons le cadre probabiliste.

Cet article comporte quatre sections principales et une annexe. La section 2 présente les rappels nécessaires sur d'une part les proportions analogiques booléennes et d'autre part les proportions analogiques numériques. Ces dernières peuvent être envisagées soit sur le modèle des proportions arithmétiques, soit sur le modèle des proportions géométriques. La section 3 s'intéresse aux proportions analogiques entre des valeurs de probabilité : ces valeurs représentent les probabilités qu'un item ait des valeurs particulières d'attributs. La section 4 considère la possibilité de définir des proportions analogiques entre quatre distributions de probabilité. Pour ce faire deux définitions sont plus particulièrement considérées, l'une à base de proportion arithmétique, l'autre requiert à la fois la satisfaction de proportions arithmétiques et la satisfaction de proportions géométriques. Cette seconde définition est compatible avec une conservation de la divergence de Kullback-Leibler entre les paires de distributions. La section 5 discute des applications potentielles de proportions analogiques entre probabilités. Une annexe fournit une illustration du pouvoir des

1. Ces deux vues de l'analogie sont assez différentes. Cependant, si a et b d'une part, et c et d d'autre part sont "analogues", on peut arguer qu'une proportion analogique "a est à b comme c est à d" doit tenir. C'est ce qui est montré dans [12] où "a analogue à b" est modélisé, de façon un peu différente de Polya, par les relations de conséquence non monotone "si a généralement b" et "si b généralement a".

proportions géométriques en proposant une preuve élégante du théorème de Pythagore.

2 Rappels

Une proportion analogique "a est à b comme c est à d", notée $a : b :: c : d$, est censée satisfaire trois postulats :

— *réflexivité* : $a : b :: a : b$;
— *symétrie* : $a : b :: c : d \Rightarrow c : d :: a : b$;
— stabilité par *permutation centrale* : $a : b :: c : d \Rightarrow a : c :: b : d$.

Quand les items sont représentés de manière vectorielle, c'est-à-dire que $a = (a_1, \cdots, a_n)$, $b = (b_1, \cdots, b_n)$, $c = (c_1, \cdots, c_n)$, et $d = (d_1, \cdots, d_n)$, la proportion analogique $a : b :: c : d$ est définie composante par composante :

Définition 1 $a : b :: c : d$ *ssi* $\forall i = 1, \cdots, n,\ a_i : b_i :: c_i : d_i$.

Nous considérons successivement les cas où les composantes i correspondent à des variables booléennes, nominales, puis numériques. Dans ces sous-sections, on omettra les indices i pour alléger la notation, étant entendu qu'on ne s'intéresse qu'à une seule composante.

2.1 Proportion analogique booléenne

Le modèle booléen minimal des 3 postulats précédents est donné dans la table 1, ou l'on présente les valuations de a, b, c, d que doit contenir tout modèle des 3 postulats. Dans

a	b	c	d
0	0	0	0
1	1	1	1
0	0	1	1
1	1	0	0
0	1	0	1
1	0	1	0

TABLE 1 – Table de vérité de $a : b :: c : d$

ce modèle minimal, les 10 autres valuations ne sont pas valides. Une expression logique qui n'est vraie que pour les quadruplets de la table ci-dessus est donnée par la formule [9] :

$$a : b :: c : d = [(a \wedge \neg b) \equiv (c \wedge \neg d)] \wedge [(\neg a \wedge b) \equiv (\neg c \wedge d)] \quad (1)$$

Cette formule exprime précisément que "a diffère de b comme c diffère de d et b diffère de a comme d diffère de c" (et "lorsque a et b ne diffèrent pas, c et d ne diffèrent pas").

L'inférence analogique est basée sur la recherche de d connaissant a, b, et c et s'appuie sur le modèle minimal. L'examen de la table montre que s'il existe, d est unique, mais qu'il n'existe pas toujours de x tel que $a : b :: c : x$ soit vrai. Par exemple, il n'y a pas de solution dans $\{0, 1\}$ pour $1 : 0 :: 0 : x$ et $0 : 1 :: 1 : x$.

La formule (1) s'applique aussi à des variables booléennes représentant des formules propositionnelles quelconques. Elle indique que pour que la proportion $a : b :: c : d$ soit satisfaite il faut que $(a \wedge \neg b) \equiv (c \wedge \neg d)$ et que $(\neg a \wedge b) \equiv (\neg c \wedge d)$. Prenons l'exemple d'une proportion analogique qui apparaît dans [13] et qui est également valable pour toute structure de treillis [1] :

$$p \vee q : p :: q : p \wedge q$$

On peut facilement vérifier que $(p \vee q) \wedge \neg p \equiv q \wedge \neg(p \wedge q) \equiv \neg p \wedge q$ et $p \wedge \neg(p \vee q) \equiv (p \wedge q) \wedge \neg q \equiv \bot$.

Une formule propositionnelle à n variables peut être décrite sur les 2^n interprétations du langage considéré. Par exemple, si nous avons deux variables booléennes p et q, nous avons les interprétations pq, $p\neg q$, $\neg pq$, $\neg p \neg q$. En les prenant dans cet ordre, p est codé par 1100, $p \wedge q$ par 1000, $p \vee q$ par 1110, q par 1010, etc.

On peut montrer que si ces 4 formules logiques a, b, c, d, mettent en jeu un ensemble de n variables, et qu'elles sont donc représentables par des chaînes de bits de taille 2^n, cela revient à avoir une proportion analogique $a_i : b_i :: c_i : d_i$ sur chaque composante i des chaînes de bits qui représentent les formules, c'est-à-dire pour chaque interprétation possible [16] (notons que i fait ici référence à une interprétation, et non à une variable booléenne comme dans la définition 1).

Effectivement dans l'exemple ci-dessus on a bien

$$1110 : 1100 :: 1010 : 1000$$

composante par composante.

Etant donné trois formules propositionnelles, on peut trouver, si elle existe, la formule propositionnelle formant une proportion analogique avec ces trois formules. On peut aussi vérifier si quatre formules propositionnelles forment ou non un proportion analogique [16].

2.2 Proportions analogiques nominales

La description des items peut impliquer des attributs nominaux, c'est-à-dire des attributs avec un domaine fini dont la cardinalité peut être supérieure à 2.

Observons que la Table 1 présente trois types de motifs, à savoir $ssss$, $sstt$ et $stst$ (où $s, t \in \{0, 1\}$) ; les deux derniers motifs sont échangés par permutation centrale.

L'analogie booléenne s'étend alors facilement à un attribut nominal $\mathcal{A}$ prennant ses valeurs dans un domaine fini $\mathcal{D}_{\mathcal{A}}$. $a : b :: c : d$ est vrai pour une variable nominale si et seulement si (comme suggéré pour la première fois dans [10]) :

$$(a, b, c, d) \in \{(s, s, s, s), (s, t, s, t), (s, s, t, t) | s, t \in \mathcal{D}_{\mathcal{A}}\} \quad (2)$$

où s, t sont des valeurs distinctes quelconques du domaine de l'attribut $\mathcal{A}$. La condition (2) généralise clairement le cas booléen. Elle garantit la satisfaction des postulats des proportions analogiques.

2.3 Proportions analogiques numériques

Les proportions analogiques ont été imaginées par Aristote sur le modèle des proportions numériques étudiées dans la génération d'avant par Archytas de Tarente et Eudoxe de Cnide.
Parmi les proportions numériques, deux proportions qui mettent en jeu les quatre opérations élémentaires ont une place importante (a, b, c, d sont ici des nombres réels) :

- la proportion arithmétique $a - b = c - d$
- la proportion géométrique $\frac{a}{b} = \frac{c}{d}$

On prendra la convention $\frac{0}{0} = 1$ considérant que $\lim_{x \to 0} \frac{x}{x} = 1$.
Il est facile de vérifier que ces deux proportions sont bien des proportions analogiques puisqu'elles satisfont les trois postulats rappelés au début de cette section. Comme dans le cas booléen, elles expriment bien l'identité des résultats de comparaisons de a et b, et de c et d soit en termes de différence soit en termes de ratios. On peut aussi vérifier l'accord avec la table de verité (Table 1) : pour les 6 lignes de la table, les égalités des proportions numériques sont satisfaites (on prend $\frac{1}{0} = +\infty$) et il n'y a pas d'égalités pour les autres quadruplets possibles. A la différence du cas booléen, le cadre numérique offre des solutions pour les proportions analogiques *continues* qui sont de la forme $a : b :: b : c$. En effet, elles conduisent à la définition de la moyenne arithmétique ($b = \frac{a+c}{2}$) et de la moyenne géométrique ($b = \sqrt{ac}$).
Ces proportions peuvent aussi s'écrire en combinant les extrêmes (a et d) et les moyens (b et c) :

- $a + d = b + c$ (proportion arithmétique)
- $a \cdot d = b \cdot c$ (proportion géométrique)

Cependant pour la proportion géométrique, cela permet d'avoir $0 \cdot 1 = 0 \cdot 0$ alors que $0 : 0 :: 0 : 1$ n'est certainement pas une proportion analogique. On voit donc que l'équivalence entre les formes "quotient" et "produit" exclut 0.
Ces 2 proportions s'échangent par transformation *logarithmique / exponentielle* :
si $\frac{a}{b} = \frac{c}{d}$ alors nous avons $\ln(a) - \ln(b) = \ln(c) - \ln(d)$,
si $a - b = c - d$ alors nous avons $\frac{e^a}{e^b} = \frac{e^c}{e^d}$.
Rappelons enfin que la notation $a : b :: c : d$ a été utilisée (au moins) jusqu'à Gaspard Monge [14] pour signifier $\frac{a}{b} = \frac{c}{d}$ (la notation ' : :' a été introduite par le mathématicien anglais William Oughtred au début du XVII^{e}). On donne en Annexe une preuve du théorème de Pythagore en termes de proportions géométriques pour illustrer leur pouvoir expressif.

Il existe d'une part des extensions multivaluées de la proportion analogique booléenne [6] qui permettent d'évaluer sur $[0, 1]$ à quel point une proportion analogique entre quatre nombres de $[0, 1]$ tient approximativement. Des cadres généraux [8], [17] ont été récemment proposés pour définir si ou ou non une proportion analogique tient entre quatre nombres de $[0, 1]$. Ces cadres inclut les proportions arithmétiques et géométriques comme cas particuliers importants. C'est pourquoi dans la suite de cet article on ne considérera que ces deux proportions.

3 Proportion analogique entre valeurs de probabilité

Dans cette section, et la suivante, on s'intéresse à des proportions analogiques entre des nombres réels qui sont des probabilités, qui ont donc des valeurs entre 0 et 1.
On considère d'abord le cas de quatre probablités qui réfèrent à quatre populations différentes et qui concernent une certaine valeur d'un même attribut pour les élements de ces quatre ensemble. a, b, c, d sont donc les probabilités qu'un élément d'un ensemble A, respectivement B, C, D prenne la valeur v_i pour l'atttibut i. On peut bien sûr s'intéresser à différents attributs simultanément. C'est par exemple le cas si on dispose d'une collection d'exemples et qu'on peut faire des statistiques sur la valeur d'attributs pour 4 groupes de données différents.
Dans la suite on utilise les notations suivantes $::_{ari}$ et $::_{geo}$ pour distinguer les proportions analogiques arithmétiques er géométriques.
La propriété suivante, facile à vérifier, garantit que si la proportion analogique tient entre quatre probabilités, elle tient aussi si on s'intéresse à l'évènement contraire :

Propriété 1
Si $a : b ::_{ari} c : d$ *alors* $1 - a : 1 - b ::_{ari} 1 - c : 1 - d$.

Cette propriété désirable n'est pas vérifiée en général pour $::_{geo}$. Elle tient cependant si les probabilités sont aussi en proportion arithmétique. Cette propriété peut être considérée comme le pendant de l'implication suivante valide en logique Booléenne :

$$a : b :: c : d \Rightarrow \neg a : \neg b :: \neg c : \neg d$$

où a, b, c, d représentent des valeurs booléennes. Cela exprime l'indépendance au codage de la proportion booléenne, vérifiable sur la Table 1.

Propriété 2
Si $a : b ::_{ari} c : d$ *et si* $a : b ::_{geo} c : d$ *alors on a* $1 - a : 1 - b ::_{geo} 1 - c : 1 - d$.

Preuve On vérifie que $(1 - a)(1 - d) = (1 - b)(1 - c)$ si $ad = bc$ et $a + d = b + c$. $\square$

Comme dans les cas booléen et nominal, il n'existe pas toujours de x tel que $a : b ::_{ari} c : x$ tienne. En effet

$$0 \le a \le 1, 0 \le b \le 1, 0 \le c \le 1 \not\Rightarrow 0 \le x = b+c-a \le 1$$

De la même façon, il n'existe pas toujours de x tel que $a : b ::_{geo} c : x$ tienne, puisque

$$0 \le a \le 1, 0 \le b \le 1, 0 \le c \le 1 \not\Rightarrow 0 \le x = \frac{bc}{a} \le 1$$

Dans les deux cas, si b et c sont grands alors on doit avoir a grand, et si a est petit, on doit avoir b ou c petits pour qu'il y ait une solution.

4 Proportion analogique entre distributions de probabilité

Au lieu des vecteurs comme dans la définition 1, nous considérons maintenant des distributions de probabilité normalisées. a, b, c, d désignent maintenant quatre distributions de probabilité sur un domaine fini. On a donc $\Sigma_{i=1,n}a_i = 1$, $\Sigma_{i=1,n}b_i = 1$, $\Sigma_{i=1,n}c_i = 1$, $\Sigma_{i=1,n}d_i = 1$. On va examiner plusieurs définitions possibles de proportions analogiques entre des distributions de probabilité et étudier leurs propriétés.

4.1 Définition arithmétique

Considérons tout d'abord une définition basée sur la proportion arithmétique.

Définition 2 *Soient a, b, c, d quatre distributions de probabilité sur un domaine fini $X = \{x_1, \cdots, x_n\}$. Alors ces distributions sont en proportion analogique* arithmétique, *notée $a : b ::_{ari} c : d$ si pour tout i on a $a_i - b_i = c_i - d_i$, où $a_i = a(x_i)$, $b_i = b(x_i)$, $c_i = c(x_i)$, $d_i = d(x_i)$.*

Il est clair que $::_{ari}$ satisfait les 3 postulats des proportions analogiques. Cette définition préserve la normalisation des probabilités.

Proposition 1 *Soient a, b, c trois distributions de probabilité normalisées, si $a : b ::_{ari} c : d$ et si $0 \leq c_i + b_i - a_i \leq 1$, $\forall i$, alors d est une distribution de probabilité normalisée.*

Preuve
La condition $0 \leq c_i + b_i - a_i \leq 1$ assure que $0 \leq d_i \leq 1$. Comme $\forall i, a_i - b_i = c_i - d_i$, l'addition membre à membre de ces égalités montre que $\Sigma_{i=1,n}d_i = 1$ puisque $\Sigma_{i=1,n}a_i = 1; \Sigma_{i=1,n}b_i = 1$ et $\Sigma_{i=1,n}c_i = 1$. □

La condition $0 \leq c_i + b_i - a_i \leq 1$ est nécessaire pour que d soit une distribution de probabilité, comme le montre le contre-exemple suivant.

Contre-exemple 1 *Prenons $n = 2$. $a_1 = 0{,}7, a_2 = 0{,}3$; $b_1 = 0{,}3, b_2 = 0{,}7$; $c_1 = 0{,}2, c_2 = 0{,}8$.*
Ce qui donne $d_1 = -0{,}2, d_2 = 1{,}2$. □

La définition 2 couvre le cas déterministe :

Observation 1 *Dans le cas déterministe, les distributions de probabilités sont telles que $\exists i, e_i = 1$ and $\forall j \neq i, e_j = 0$, pour $e \in \{a, b, c, d\}$. Il y a alors trois possibilités pour que $a : b ::_{ari} c : d$ tienne :*
- *$\exists i, a_i = b_i = c_i = d_i = 1$;*
- *$\exists i, a_i = b_i = 1$ et $\exists j, c_j = d_j = 1$;*
- *$\exists i, a_i = c_i = 1$ et $\exists j, b_j = d_j = 1$.* □

Donnons maintenant quelques exemples, non extrêmes comme les précédents, de distributions pour lesquelles on obtiendra $a : b ::_{ari} c : d$.

Exemple 1

1. *En utilisant la Propriété 1, on voit que dès que $\exists i, \exists j$, tel que $a_i : b_i ::_{ari} c_i : d_i$, avec $a_j = 1 - a_i, b_j = 1 - b_i, c_j = 1 - c_i, d_j = 1 - d_i$ et $a_k = b_k = c_k = d_k = 0, \forall k \neq i, j$, on a $a : b ::_{ari} c : d$.*
2. *En fait si $a_i : b_i ::_{ari} c_i : d_i$, pour tout réel λ on a $\lambda - a_i : \lambda - b_i ::_{ari} \lambda - c_i : \lambda - d_i$. Ainsi en prenant des λ positifs dont la somme n'excède pas 1, on peut construire des paires $((a_i, b_i, c_i, d_i), (a_j, b_j, c_j, d_j))$ formant deux proportions arithmétiques telles $a_i + a_j = b_i + b_j = c_i + c_j = d_i + d_j = \lambda$, comme dans l'exemple suivant où on a des allocations partielles de probabilité $\lambda = 0{,}4$ et $\lambda' = 0{,}5$ complétées par un quadruplet (a_k, b_k, c_k, d_k) de valeurs égales à 0,1 :*
   $a_1 = 0{,}1; a_2 = 0{,}3; a_3 = 0{,}1; a_4 = 0{,}2; a_5 = 0{,}3.$
   $b_1 = 0{,}3; b_2 = 0{,}1; b_3 = 0{,}1; b_4 = 0{,}3; b_5 = 0{,}2.$
   $c_1 = 0{,}2; c_2 = 0{,}2; c_3 = 0{,}1; c_4 = 0{,}4; c_5 = 0{,}1.$
   $d_1 = 0{,}4; d_2 = 0; d_3 = 0{,}1; d_4 = 0{,}5; d_5 = 0.$
   On vérifie bien que $a : b ::_{ari} c : d$ tient entre des distributions normalisées, toutes différentes.
3. *Pour que $a : b ::_{ari} c : d$ soit satisfait, il n'est pas nécessaire de procéder comme précédemment, c'est-à-dire de construire des paires de quadruplets dont la somme terme à terme est constante. C'est ce que montrent les deux exemples suivants pour $n = 3$ et $n = 4$:*
   $a_1 = 0{,}3; a_2 = 0{,}2; a_3 = 0{,}5.$
   $b_1 = 0{,}5; b_2 = 0{,}1; b_3 = 0{,}4.$
   $c_1 = 0{,}4; c_2 = 0{,}2; c_3 = 0{,}4.$
   $d_1 = 0{,}6; d_2 = 0{,}1; d_3 = 0{,}3.$
4. $a_1 = 0{,}1; a_2 = 0{,}2; a_3 = 0{,}4; a_4 = 0{,}3.$
   $b_1 = 0{,}3; b_2 = 0{,}3; b_3 = 0{,}2; b_4 = 0{,}2.$
   $c_1 = 0{,}2; c_2 = 0{,}2; c_3 = 0{,}2; c_4 = 0{,}4.$
   $d_1 = 0{,}4; d_2 = 0{,}3; \quad d_3 = 0; d_4 = 0{,}3.$ □

Ces exemples montrent qu'ils existent des distributions non triviales satisfaisant la définition 2. Bien entendu, étant donné trois distributions a, b, c, il n'existe pas toujours de distribution d telle que $a : b ::_{ari} c : d$ tienne, puisque on doit avoir $\forall i, 0 \leq c_i + b_i - a_i \leq 1$. Notons cependant que étant donné a, b, il est toujours possible de trouver c tel que ces inégalités soient satisfaites pour chaque i. on peut alors trouver une distribution d en proportion anlogique arithmétique avec a, b, c.

Remarque 1 *Il est naturel de se demander ce qu'il en serait d'une définition analogue à la définition 2 en termes de proportion géométrique, c'est-à-dire considerer la définition suivante :*

Définition 3 *Soient a, b, c, d quatre distributions de probabilité sur un domaine fini $X = \{x_1, \cdots, x_n\}$. Alors ces distributions sont en proportion analogique* géométrique, *notée $a : b ::_{geo} c : d$ si pour tout i on a $\frac{a_i}{b_i} = \frac{c_i}{d_i}$,*
où $a_i = a(x_i)$, $b_i = b(x_i)$, $c_i = c(x_i)$, $d_i = d(x_i)$.

Malheureusement, comme le montre le contre-exemple suivant, la contrepartie de la proposition 1 pour la proportion géométrique est fausse. La normalisation de a, b, c ne suffit pas pour garantir celle de d quand $a : b ::_{geo} c : d$ tient.

Contre-exemple 2 *Prenons $n = 3$. On a $a_1 + a_2 + a_3 = 1$, $b_1 + b_2 + b_3 = 1$, $c_1 + c_2 + c_3 = 1$, et $a_1 \cdot d_1 = b_1 \cdot c_1$, $a_2 \cdot d_2 = b_2 \cdot c_2$, $a_3 \cdot d_3 = b_3 \cdot c_3$, avec la condition $b_i \cdot c_i \leq a_i$.*

- *$b_1 = 0{,}2; b_2 = 0{,}3; b_3 = 0{,}5$, $c_1 = 0{,}4; c_2 = 0{,}3; c_3 = 0{,}3$ donc $b_1 \cdot c_1 = 0{,}08; b_2 \cdot c_2 = 0{,}09; b_3 \cdot c_3 = 0{,}15$. On prend $a_1 = 0{,}3; a_2 = 0{,}4; a_3 = 0{,}3$. Ce qui donne $d_1 = \frac{4}{15}; d_2 = \frac{9}{40},; d_3 = \frac{1}{2}$, et finalement $\Sigma_i d_i = 0{,}991666 < 1$*
- *$b_1 = 0{,}1; b_2 = 0{,}5; b_3 = 0{,}4$, $c_1 = 0{,}2; c_2 = 0{,}2; c_3 = 0{,}6$. Donc $b_1 \cdot c_1 = 0{,}02; b_2 \cdot c_2 = 0{,}10; b_3 \cdot c_3 = 0{,}24$. En conservant le même a qu'au dessus, on obtient $d_1 = \frac{1}{15}; d_2 = \frac{1}{4}; d_3 = \frac{4}{5}$ et donc $\Sigma_i d_i = 1{,}116666 > 1$.*

Remarque 2 *Dans le même esprit, on pourrait se poser la question de définir une proportion analogique entre des distributions de* possibilité *[20, 5] a, b, c, d, où la normalisation est exprimée par $\max_i a_i = 1, \max_i b_i = 1, \max_i c_i = 1, \max_i d_i = 1$, les a_i, b_i, c_i, d_i étant compris entre 0 et 1. Comme une distribution de possibilité e peut être vue en termes de ses α-coupes $e_\alpha = \{i \mid e_i \geq \alpha\}$ qui sont des sous-ensembles emboités $e_\alpha \subseteq e_\beta$ si $\alpha \geq \beta$, on peut se ramener à l'approche pour le cas booléen pour chaque α-coupe [16].*

4.2 Définition combinant les deux types de proportions numériques

Dans cette sous-section, on étudie une définition de la proportion analogique entre distributions de probabilité, plus exigeante que la définition 2 car combinant proportion arithmétique et proportion géométrique.

Définition 4 *Soient a, b, c, d quatre distributions de probabilité sur un domaine fini $X = \{x_1, \cdots, x_n\}$. Alors ces distributions sont en proportion analogique* arithmético-géométrique, *notée $a : b ::_{arge} c : d$ si pour tout i on a $a_i - b_i = c_i - d_i$ et $\frac{a_i}{b_i} = \frac{c_i}{d_i}$, où $a_i = a(x_i)$, $b_i = b(x_i)$, $c_i = c(x_i)$, $d_i = d(x_i)$.*

Il est clair que $::_{arge}$ satisfait les 3 postulats des proportions analogiques. Malgré le caractère contraignant de cette définition, il existe des distributions de probabilité non triviales qui la satisfont.

Proposition 2 *Il existe des distributions de probabilité a, b, c, d telles que $a : b ::_{arge} c : d$ tient. Elles sont telles que pour chaque composante i,*

- *cas 1 : soit $a_i = b_i$ et $c_i = d_i$,*
- *cas 2 : soit $a_i = c_i$ et $b_i = d_i$.*

et ce sont les seules solutions.

Preuve
En posant $a_i + d_i = s_i$ et $a_i d_i = p_i$, on obtient $d_i = s_i - a_i$, puis $a_i(s_i - a_i) = p_i$, c'est-à-dire $a_i^2 - s_i \cdot a_i + p_i = 0$. Ainsi, a_i doit être une solution de l'équation du second degré [17] :

$$x^2 - s_i x + p_i = 0$$

Son discriminant est égal à $\Delta = s_i^2 - 4p_i = (b_i + c_i)^2 - 4b_i \cdot c_i = (b_i - c_i)^2$. Les deux solutions sont donc $a_i = \frac{s_i \pm \sqrt{\Delta}}{2} = \frac{b_i + c_i \pm (b_i - c_i)}{2}$. Ce qui donne $a_i = b_i$ ou $a_i = c_i$. Ce qui conduit à $(a_i, d_i) = (b_i, c_i)$ ou $(d_i, a_i) = (b_i, c_i)$. Donc $a : b ::_{arge} c : d$ si et seulement si pour chaque composante i, on a
- $a_i = b_i$ et $c_i = d_i$, ou
- $a_i = c_i$ et $b_i = d_i$ □

Voici un exemple d'une proportion analogique entre des distributions de probabilité a, b, c, d normalisées, toutes différentes, obéissant à la définition 4 :

Exemple 2
$a_1 = 0.1, a_2 = 0.3, a_3 = 0.2, a_4 = 0.3, a_5 = 0.1$
$b_1 = 0.1, b_2 = 0.4, b_3 = 0.2, b_4 = 0.2, b_5 = 0.1$
$c_1 = 0.1, c_2 = 0.3, c_3 = 0.3, c_4 = 0.3, c_5 = 0$
$d_1 = 0.1, d_2 = 0.4, d_3 = 0.3, d_4 = 0.2, d_5 = 0$

On peut vérifier que les quatre distributions sont bien normalisées, et que pour chaque i, on a à la fois $a_i - b_i = c_i - d_i$ et $\frac{a_i}{b_i} = \frac{c_i}{d_i}$ (rappelons qu'on a pris la convention $\frac{0}{0} = 1$). Observons que pour les i on a trois sortes de motifs possibles pour les valeurs de probabilités : (s, s, s, s), (s, t, s, t) et (s, s, t, t), ce sont précisément ceux déjà rencontrés pour les valeurs nominales en sous-section 2.2.2. Notons que les deux derniers motifs doivent être obligatoirement présents si on veut avoir des distributions distinctes, sinon on a $a = b$ (et donc $c = d$), ou $a = c$ (et donc $b = d$). Le motif (s, s, s, s) peut être absent. Une situation semblable avait déjà été observée dans le cas booléen [3].

Il est facile de trouver des distributions de probabilité a, b, c, d telles que $a : b ::_{arge} c : d$ tient. En effet, on a le résultat suivant.

Proposition 3 *Etant donné deux distributions de probabilité a et b différentes, il existe toujours deux distributions de probabilité c et d telles que $a : b ::_{arge} c : d$ tient. S'il existe au moins un i tel que $a_i = b_i$, les quatre distributions peuvent être différentes.*

Preuve
En effet, si $a_i \neq b_i$, alors $c_i = a_i$ et $d_i = b_i$. Remarquons, puisque $\Sigma_{i \mid a_i = b_i} a_i = \Sigma_{i \mid a_i = b_i} b_i \triangleq \rho$, que $\Sigma_{i \mid a_i \neq b_i} a_i = \Sigma_{i \mid a_i \neq b_i} b_i = 1 - \rho$. Donc on a aussi $\Sigma_{i \mid c_i \neq d_i} c_i = \Sigma_{i \mid c_i \neq d_i} d_i = 1 - \rho$. Pour les i tels que $a_i = b_i$, on affecte à $c_i = d_i$ des parts de la masse restante ρ de façon à ce que $\Sigma_{i=1,n} c_i = \Sigma_{i=1,n} d_i = 1$. □

Ces probabilités a, b, c, d satisfaisant $a : b ::_{arge} c : d$ ont, comme on vient le voir, une forme particulière. Elles satisfont une propriété remarquable en termes de divergence de Kullback-Leibler (notée KL), qui rappelons-le, évalue le changement entre deux distributions a et b par l'expression $KL(a,b) = \Sigma_{i=1,n} a_i \log \frac{a_i}{b_i}$ (dans le cas discret). En général, $KL(a,b) \neq KL(b,a)$. Cependant, on peut énoncer le résultat suivant :

Proposition 4 *Soient a, b, c, d quatre distributions de probabilité formant une proportion analogique arithmético-géométrique $a : b ::_{arge} c : d$. Alors on a :*

$$KL(a,b) = KL(c,d)$$

Preuve
Puisque $KL(a,b) = \Sigma_{i=1,n} a_i \log \frac{a_i}{b_i}$ et $KL(c,d) = \Sigma_{i=1,n} c_i \log \frac{c_i}{d_i}$, examinons les différents termes i : $a_i \log \frac{a_i}{b_i}$ et $c_i \log \frac{c_i}{d_i}$;
On sait par la Proposition 2, qu'il y deux cas :
- cas (1) $a_i = b_i$: les deux termes sont nuls.
- cas (2) : $a_i = c_i$: On a donc $a_i \log \frac{a_i}{b_i} = c_i \log \frac{c_i}{d_i}$ par simple remplacement puisque $\frac{a_i}{b_i} = \frac{c_i}{d_i}$.
D'où $KL(a,b) = KL(c,d)$. □

Puisque $::_{arge}$ satisfait les postulats des proportions analogiques, on a aussi

Propriété 3 *Si $a : b ::_{arge} c : d$, alors*
— *i) $KL(a,c) = KL(b,d)$;*
— *ii) $KL(b,a) = KL(d,c)$;*
— *iii) $KL(d,b) = KL(c,a)$.*

Preuve i) étant donné la proposition 4 et la permutation centrale ; ii) puisqu'on obtient $a : b ::_{arge} c : d \Rightarrow b : a ::_{arge} d : c$, par applications successives de la permutation centrale, de la symétrie, et de la permutation centrale ; iii) la dernière égalité, qui correspond à la permutation des extrêmes, peut s'obtenir par application de la seconde à la première. □

Un exemple de proportion analogique entre des distributions continues, définies à l'aide de fonctions linéaires par morceaux, est donné dans l'Annexe 2.

5 Perspectives

Dans cette section, on suggère quelques pistes d'applications des proportions analogiques entre probabilités. Puisque les résultats de la section 4 s'appliquent aussi à toute collection de nombres positifs ou nuls dont la somme est 1, on termine par une application aux sommes pondérées en aggrégation multi-critères (puisque comme dans une distribution de probabilité, la somme des coefficients y est égale à 1).
Commençons par rappeler le principle de la classification bayésienne. On a par le théorème de Bayes :

$$P(c|x)P(x) = P(x|c)P(c)$$

où x est un vecteur de valeurs d'attributs décrivant un item, c est une classe, et on attribue à x, parmi les classes possibles, la classe c qui maximise $P(c|x)$. On peut remarquer que cette égalité est une proportion géométrique :

$$P(x) : P(x|c) :: P(c) : P(c|x)$$

On peut au moins penser à trois pistes d'utilisation des proportions analogiques sur des probabilités :

1. **Parallèle analogie-probabilités**

   Tout comme l'approche bayésienne, les proportions analogiques offrent une application à la classification [4, 2]. Ceci suggère de faire un parallèle entre les deux. Considérons quatre items a, b, c, d dont les descriptions vectorielles forment une proportion analogique au sens de la définition 1. De plus on suppose que ces quatre items ont été classés, d'une part par la méthode bayésienne, et d'autre part par la méthode des proportions analogiques. Si les classes fournies pour a, b, c, d par le classifieur analogique sont elles-mêmes en proportion analogique, on peut bien sûr se demander si les deux classifieurs sont en accord, mais aussi plus généralement, si les distributions de probabilité fournies par la règle de Bayes forment une proportion analogique (au moins approximativement au sens de l'extension graduelle numérique des proportions analogiques entre valeurs nominales, appelées extension conservative dans [6]).

2. **Valeurs d'attributs incertaines**

   Dans un problème de classification, certains attributs peuvent présenter une incertitude quant à leur valeur. Considérons le cas d'un seul attribut incertain, qu'il soit booléen ou nominal. Pour un item donné, chaque valeur possible de cet attribut est associée à une probabilité. Supposons que quatre items correspondant à la valeur la plus probable de l'attribut forment une proportion analogique.

   Si les probabilités associées respectent également cette proportion (au moins approximativement comme plus haut) et que l'équation analogique sur les classes permette d'inférer une solution pour le quatrième item, alors on peut accepter cette solution avec la probabilité correspondante.

3. **Proportions analogiques entre distributions**

   Rappelons que l'apprentissage par transfert peut être vu comme une sorte de raisonnement analogique effectué au niveau méta (voir par exemple [19]), puiqu'il s'agit de tirer parti de ce qui a été appris sur un domaine source afin d'améliorer le processus d'apprentissage dans un domaine cible lié au domaine source. Comme le suggère le vocabulaire utilisé, la démarche est assez semblable à celle qui préside au raisonnement à partir de cas [7].

   On peut se demander si des proportions analogiques entre distributions de probabilités, en particulier au sens de la définition 4, pourrait permettre des transferts plus sophistiqués et plus contrôlés, en utilisant les Propositions 2 et 3.

   Plus simplement, on pourrait chercher à savoir pour deux attributs d'intérêt (l'âge et le sexe dans l'exemple ci-après), si les probabilités d'être dans une classe c

— pour un homme de moins de 40 ans,
— pour une femme de moins de 40 ans,
— pour un homme de plus de 40 ans,
— pour une femme de plus de 40 ans,

forment une proportion analogique pour les distributions de probabilité sur les classes de chacune des quatre populations.

Agrégation multi-critères

Des valeurs positives dont la somme fait 1 ne sont pas nécessairement des probabilités. Cela peut être aussi les coefficients d'une somme pondérée. Ainsi on pourra ajuster des coefficients d par rapport à d'autres jeux de coefficients a, b, c en résolvant une équation analogique $a : b :: c : x$ entre des distributions de coefficients, comme dans l'exemple suivant :

— a : coefficients des matières dans la filière scientifique de l'établissement 1,
— b : coefficients des matières dans la filière littéraire de l'établissement 1,
— c : coefficients des matières dans la filière scientifique de l'établissement 2,
— d : coefficients des matières dans la filière littéraire de l'établissement 2,

6 Conclusion

Cet article est vraisemblablement le premier à s'intéresser à des proportions analogiques sur des distributions de probabilité. On a notamment montré qu'il était possible d'avoir une définition combinant les proportions arithmétiques et géométriques. Cette définition garantit l'invariance, dans le cas discret, de la divergence de Kullback-Leibler entre les les distributions formant les deux paires de la proportion analogique. L'extension au cas continu nécessite une étude minutieuse, pour mieux comprendre quelles sont les situations possibles pour avoir une proportion analogique entre des distributions (au delà du premier exemple donné dans l'Annexe 2), et savoir quand la conservation de la divergence de Kullback-Leibler associée aux paires est préservée. Cela constitue donc une piste de recherche à poursuivre. Plus généralement, les applications concrètes restent à développer.

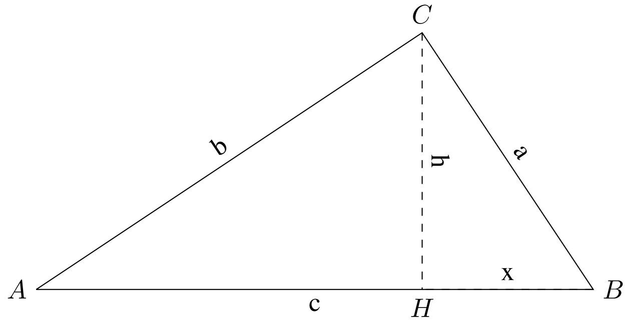
Remerciements

Les auteurs remercient Giuseppe Sanfilippo pour avoir posé la question “Do analogical proportions apply to probabilities ?” lors de la présentation de leur article [17]. Le présent article espère montrer que la question était féconde.

Cette recherche a été soutenue par le projet ANR « Analogies : from Theory to Tools and Applications » (AT2TA), ANR-22-CE23-0023.

Annexe 1 : Preuve du théorème de Pythagore par proportions géométriques

Considérons un triangle ABC rectangle en C, et la hauteur issue de C, comme sur la figure ci-après.
Les angles $\widehat{HAC}$ et $\widehat{HCB}$ sont égaux comme ayant le même complément $\widehat{ACH}$. Les triangles ABC, ACH et CBH sont donc semblables.



Dans cette figure, deux côtés d'un des trois triangles rectangles sont donc proportionnels aux deux côtés correspondants d'un autre triangle rectangle, et forment donc une proportion géométrique, en termes de longueurs.
Par exemple, on a
$AB : CB :: AC : HC$ c'est-à-dire $c : a :: b : h$
et donc $ab = ch$.
Si $a = 3$, $b = 4$, $c = 5$ on obtient $h = 2.4$.
On a aussi
$CH : AH :: BH : CH$ c'est-à-dire $h : c - x :: x : h$.
Ainsi
$h^2 = cx - x^2$, i.e., $h^2 + x^2 = cx$
Mais on a aussi
$AB : CB :: CB : BH$ c'est-à-dire $c : a :: a : x$
et donc $cx = a^2$.
On obtient $h^2 + x^2 = a^2$.   QED !

Pour $a = 3$, $c = 5$, on obtient $x = 1.8$.

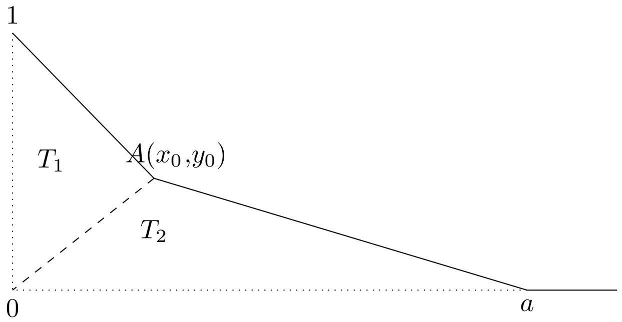
On peut aussi écrire
$CA : AB :: AH : CA$ c'est-à-dire $b : c :: c - x : b$
ce qui donne $b^2 = c^2 - cx$.
Ce qui conduit à
$c^2 = a^2 + b^2$.

NB : Notons l'usage d'une proportion continue
($c : a :: a : x$).

Annexe 2 : Un exemple de proportion analogique entre des distributions continues

Un exemple, non trivial, simple, peut être construit de la façon suivante, on prend
$a(x) = p(x)$ sur $(-\infty, +\infty)$
$b(x) = p(x)$ sur $(-\infty, 0]$; $b(x) = q(x)$ sur $[0, +\infty)$
$c(x) = q(x)$ sur $(-\infty, 0]$; $c(x) = p(x)$ sur $[0, +\infty)$
$d(x) = q(x)$ sur $(-\infty, +\infty)$

où p et q sont des fonctions linéaires par morceaux définies à partir d'une fonction paramétrée $S_{x_0,a}$, définie sur $[0, +\infty)$, représentée en trait plein sur la figure ci-après.



Le point A a pour coordonnées (x_0, y_0). On prend ensuite :
$p(x) = S_{x_0,a}(x)$ sur $[0, +\infty)$, $S_{x_0,a}(-x)$ sur $(-\infty, 0]$.
$q(x)$ est définie de manière similaire avec des paramètres x'_0 et a'.
Notons qu'on a $p(0) = q(0) = 1$, et que a, b, c, d sont donc continues.
L'aire en dessous de $S_{x_0,a}$ est égale à $aire(T_1) + aire(T_2) = x_0/2 + ay_0/2$.
Pour assurer la normalisation des distributions a, b, c, d, cette aire doit être égale à $1/2$, on doit donc imposer $(x_0 + ay_0) = 1$ et $(x'_0 + a'y'_0) = 1$. Dans ces conditions, on peut vérifier facilement que $a : b ::_{arge} c : d$ et que $KL(a, b) = KL(c, d)$.

En partant de l'exemple ci-dessus, et en utilisant des fonctions paramétriques $S_{h,x_0,a}$ et $S_{h,x'_0,a'}$ similaires aux précédentes, mais telles que $S_{h,x_0,a}(0) = S_{h,x'_0,a'}(0) = h$, on peut construire des exemples plus complexes :
Soient $p_1 < p_2$. Soient $\lambda < 1/2$, $\lambda' < 1/2$.
On prend alors
$a(x) = p(x)$ sur $(-\infty, p_1]$; $a(x) = r(x)$ sur $[p_1, p_2]$; $a(x) = p(x)$ sur $[p_2, +\infty)$;
$b(x) = p(x)$ sur $(-\infty, p_1]$; $b(x) = r(x)$ sur $[p_1, p_2]$; $b(x) = q(x)$ sur $[p_{2,} + \infty)$;
$c(x) = q(x)$ sur $(-\infty, p_1]$; $c(x) = r(x)$ sur $[p_1, p_2]$; $c(x) = p(x)$ sur $[p_2, +\infty)$;
$d(x) = q(x)$ sur $(-\infty, p_1]$; $d(x) = r(x)$ sur $[p_1, p_2]$; $d(x) = q(x)$ sur $[p_2, +\infty)$;
où

— $p(x) = S_{h,x_0,a}(x - p_1)$ sur $[p_1, +\infty)$,
  $p(x) = S_{h,x_0,a}(p_1 - x)$ sur $(-\infty, p_1]$, et
— $q(x) = S_{h,x'_0,a'}(x - p_2)$ sur $[p_2, +\infty)$,
  $q(x) = S_{h,x'_0,a'}(p_2 - x)$ sur $(-\infty, p_2]$,
— $r(x)$ est tel que $r(p_1) = r(p_2) = h$ (assurant la continuité des 4 distributions) et $\int_{p_1}^{p_2} r(x)dx = \lambda + \lambda'$ (r peut éventuellement être pris comme uniforme),

avec les contraintes
$(hx_0 + ay_0) = 1 - 2\lambda$,
$(hx'_0 + a'y'_0) = 1 - 2\lambda'$,
et $h(p_2 - p_1) \leq \lambda + \lambda'$ (il y a égalité si r est uniforme).
assurant ainsi la normalisation des 4 distributions.

Références

[1] Nelly Barbot, Laurent Miclet, and Henri Prade. Analogy between concepts. *Artif. Intel.*, 275 :487–539, 2019.

[2] Myriam Bounhas and Henri Prade. Analogy-based classifiers : An improved algorithm exploiting competent data pairs. *Int. J. Approx. Reason.*, 158 :108923, 2023.

[3] Myriam Bounhas and Henri Prade. Revisiting analogical proportions and analogical inference. *Int. J. Approx. Reason.*, 171 :109202, 2024.

[4] Myriam Bounhas, Henri Prade, and Gilles Richard. Analogy-based classifiers for nominal or numerical data. *Int. J. Approx. Reason.*, 91 :36–55, 2017.

[5] Didier Dubois and Henri Prade. *Possibility Theory : An Approach to Computerized Processing of Uncertainty*. Plenum Press, 1988.

[6] Didier Dubois, Henri Prade, and Gilles Richard. Multiple-valued extensions of analogical proportions. *Fuzzy Sets Syst.*, 292 :193–202, 2016.

[7] Béatrice Fuchs, Jean Lieber, Laurent Miclet, Alain Mille, Amedeo Napoli, Henri Prade, and Gilles Richard. Case-based reasoning, analogy, and interpolation. In Pierre Marquis, Odile Papini, and Henri Prade, editors, *A Guided Tour of Artificial Intelligence Research : Volume I : Knowledge Representation, Reasoning and Learning*, pages 307–339. Springer, 2020.

[8] Yves Lepage and Miguel Couceiro. Analogie et moyenne généralisée. In *Actes des 18èmes Journées d'Intelligence Artificielle Fondamentale, La Rochelle, 1-3 Juil.*, pages 114–124, 2024.

[9] Laurent Miclet and Henri Prade. Handling analogical proportions in classical logic and fuzzy logics settings. In *Proc. 10th Eur. Conf. on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU'09),Verona*, pages 638–650. Springer, LNCS 5590, 2009.

[10] Vito Pirrelli and François Yvon. Analogy in the lexicon : a probe into analogy-based machine learning of language. In *Proc. 6th Int. Symp. on Human Communication, Santiago de Cuba*, page 6 p., 1999.

[11] George Polya. *Mathematics and Plausible Reasoning-Vol.1 : Induction and analogy in Mathematics, Vol.2 : Patterns of Plausible Inference*. Princeton Univ. Press, 2nd ed. 1968, 1954.

[12] Henri Prade and Gilles Richard. Cataloguing / analogizing : A nonmonotonic view. *Int. J. Intell. Syst.*, 26(12) :1176–1195, 2011.

[13] Henri Prade and Gilles Richard. From analogical proportion to logical proportions. *Logica Universalis*, 7(4) :441–505, 2013.

[14] Henri Prade and Gilles Richard. Multiple analogical proportions. *AI Commun.*, 34(3) :211–228, 2021.

[15] Henri Prade and Gilles Richard. Analogical proportion- based induction : From classification to creativity. *J. of Applied Logics - IfCoLog J. of Logics and their Applicatiions*, 11 :51–87, 2024.

[16] Henri Prade and Gilles Richard. Diagrammatic analogical reasoning. In J. Lemanski, M. W. Johansen, E. Manalo, P. Viana, R. Bhattacharjee, and R. Burns, editors, *Proc. 14th Int. Conf. on Diagrammatic Representation and Inference (Diagrams'24), Münster, Sept. 27 - Oct. 1*, volume 14981 of *LNCS*, pages 485–489. Springer, 2024.

[17] Henri Prade and Gilles Richard. Frank's triangular norms in Piaget's logical proportions. In S. Destercke, M. V. Martinez, and G. Sanfilippo, editors, *Proc. 16th Int. Conf. on Scalable Uncertainty Management (SUM'24), Palermo, Nov. 27-29*, volume 15350 of *LNCS*, pages 369–377. Springer, 2024.

[18] Marta Sznajder. Janina Hosiasson-Lindenbaum on analogical reasoning : New sources. *Erkenntnis*, 89 :1349–1365, 2024.

[19] Hua-Yan Wang and Qiang Yang. Transfer learning by structural analogy. In Wolfram Burgard and Dan Roth, editors, *Proc. 25th AAAI Conf. on Artificial Intelligence, (AAAI'11), San Francisco, Aug.7-11*, pages 513–518. AAAI Press, 2011.

[20] Lotfi A. Zadeh. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Syst.*, 1(1) :3–28, 1978.