Andrzej Łachwa

Podobieństwo zbiorów

1. Wstęp

Niemal codziennie używamy określenia „podobieństwo" i wskazujemy rzeczy podobne do siebie. Na pierwszy rzut oka jest to pojęcie proste. Jednak informatyk patrzy na takie z pozoru proste pojęcia nieco inaczej. Musi nadać im tak precyzyjny sens, by nadawały się na elementy budowanego modelu rzeczywistości lub na operacje wchodzące w skład tworzonej metody obliczeniowej.

Pojęcie podobieństwa wykorzystujemy głównie do budowania zbiorów: zbiór składa się z elementów podobnych! I choć określenie „zbiór" zastępujemy często terminami „typ encji" czy „klasa obiektów", nie zmienia to istoty sprawy. Łącząc elementy w zbiór, podejmujemy decyzję, na czym ma polegać ich podobieństwo.

Jak już to wyjaśniałem w poprzednim tomie „Informatyki" (zob. [4]), zbiór może być pojmowany na wiele sposobów. Tutaj zajmę się pojęciem podobieństwa, w szczególności zaś podobieństwa zbiorów.

2. Podobieństwo obiektów

Równość w matematyce jest eksplikowana przez relację równoważności. Relacja ta to relacja jednocześnie zwrotna, symetryczna i przechodnia. Dzieli ona uniwersum, na którym jest określona, na klasy równoważności: klasy obiektów równoważnych. Podział taki jest rozłączny i zupełny.

Równość (jednakowość, identyczność) jest zatem przede wszystkim binarną relacją równoważności określoną na pewnym uniwersum. Ponadto jest wartością względną, zależy od sytuacji czy punktu widzenia obserwatora (na danym uniwersum można zwykle zdefiniować różne relacje równoważności). I wreszcie równość oznacza zastępowalność jednego obiektu drugim w określonej sytuacji (por. [5, s. 43–49]).

Podobieństwo oznacza tylko częściową zastępowalność – istnieje możliwość zastąpienia jednego obiektu drugim, ale z pewnym ryzykiem czy pewną stratą. Podobieństwo obiektów danego uniwersum w matematyce nazywa się tolerancją i jest ono relacją zwrotną i zarazem symetryczną. Przechodniość nie jest wymagana, a to dlatego, że obiekty podobne nie są identyczne, nieznacznie różnią się od siebie i te drobne różnice między kolejnymi podobnymi obiektami mogą doprowadzić do obiektów całkowicie różnych od tych początkowych (por. [5, s. 69 i n.]).

Przykładem tego jest znana zabawa ze słowami polegająca na przekształceniu słowa początkowego w słowo końcowe przez zmianę w każdym kolejnym słowie znajdującym się pomiędzy nimi tylko jednej litery, np. możemy w taki sposób przekształcić słowo „kot" w słowo „lew" ([5, s. 72]): *kot – kos – los – lis – lin – len – lew*.

Zbiór *U* z określoną na nim relacją tolerancji *T* nazywa się przestrzenią tolerancji. Struktura tej przestrzeni jest bardzo ciekawa. W szczególności okazuje się, że dowolną tolerancję można określić za pomocą zbioru cech elementów uniwersum *U* w taki sposób, że elementami podobnymi są te, które mają co najmniej jedną wspólną cechę (por. [5, s. 79 i n.]).

3. Podobieństwo zbiorów dystrybutywnych

Dwa zbiory dystrybutywne są równe, gdy mają te same elementy. Zbiory te są równoważne, gdy można wskazać pewne identyczne cechy tych zbiorów, np. równoliczność. Jeżeli osłabimy to wymaganie przez rezygnację z przechodniości, to otrzymamy relację tolerancji. A zatem, tak jak wyżej, jeżeli mamy dany zbiór cech, to możemy przyjąć, że zbiory są podobne, gdy mają co najmniej jedną wspólną cechę.

W matematyce podobieństwo zbiorów dystrybutywnych często definiuje się odmiennie, jako szczególną relację równoważności – *równokształtność*. Na przykład w geometrii dwa wielokąty uznaje się za podobne, gdy mają te same kąty i proporcje: mają taki sam kształt, ale mogą mieć różną wielkość. W algebrze dwa wyrażenia nazywa się podobnymi, gdy mają ten sam kształt z dokładnością do współczynników liczbowych. Przykłady takie można mnożyć.

4. Podobieństwo zbiorów rozmytych

Zbiorem rozmytym jest ogół tych elementów pewnego uniwersum, które można powiązać myślowo w całość na podstawie jakieś ich własności, zwykle nieostrej (por. [3, s. 12]).

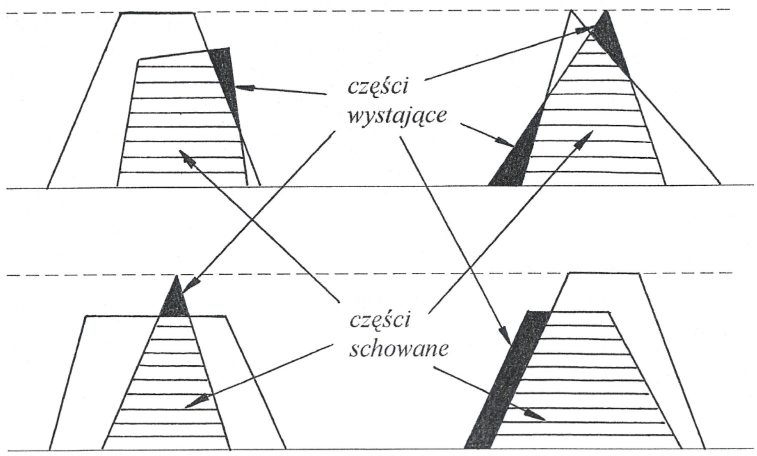
Dwa zbiory rozmyte określone na pewnym uniwersum są identyczne wtedy i tylko wtedy, gdy każdy element tego uniwersum należy do tych zbiorów w tym samym stopniu. Podobnie definiuje się inkluzję zbiorów rozmytych: zbiór rozmyty A jest zawarty w zbiorze rozmytym B (określonym na tym samym uniwersum, co A) wtedy i tylko wtedy, gdy każdy element uniwersum należy do zbioru A w stopniu nie większym, niż należy do zbioru B (por. [6]).

Tak rozumiane równość i inkluzja zbiorów rozmytych są wprawdzie formalnie poprawne i eleganckie, ale mają niewielkie znaczenie z punktu widzenia metod obliczeniowych. Po pierwsze, dlatego że w przypadku uniwersum nieskończonego (lub skończonego, ale bardzo dużego – por. [4, s. 37]) nie moglibyśmy sprawdzić, czy odpowiednia relacja zachodzi; po drugie zaś, dlatego że zdefiniowane wyżej dwie ostre własności zwykle nie mają zastosowania przy przetwarzaniu rozmytej informacji. Rozmytość to przecież niewyraźność i niepewność. Niewielkie różnice w sposobie rozumienia tej niepewności (niewielkie różnice w wartościach funkcji przynależności) nie powinny decydować o tak istotnych własnościach, jak równość i zawieranie. Zaproponowano zatem (por. [1, s. 43–46]) rozmycie własności identyczności i inkluzji zbiorów rozmytych. Jednak moim zdaniem, zamiast mówić o rozmytej równości zbiorów (identyczności zbiorów), lepiej stosować termin „podobieństwo". Podobieństwo rozumiane jest bowiem w języku naturalnym jako własność nieostra i niekoniecznie przechodnia, równość zaś (czy też identyczność) – jako własność ostra i przechodnia.

4.1. Podobieństwo oparte na rozmytej inkluzji

Rozmycie ostrego znaczenia inkluzji sprowadza się do tego, że własność zawierania się jednego zbioru w drugim będzie rozumiana jako stopniowalna: od stopnia 0 oznaczającego niezawieranie, poprzez przypadki zawierania częściowego, do stopnia 1 oznaczającego zawieranie całkowite. Dodatkowo oczekuje się, że taka rozmyta własność będzie zwrotna ($A \subset A$ w stopniu 1) i antysymetryczna (jeżeli $A \subset B$ w stopniu α i $B \subset A$ w stopniu β, to A jest podobne do B w stopniu $\alpha \wedge \beta$).

Oryginalny sposób obliczania rozmytej inkluzji wprowadziłem w pracy [3]. Zasadzał się on na spostrzeżeniu, że stopień częściowego zawierania się zbioru A w zbiorze B ma nas informować o tym, jak mają się do siebie części zbioru A, które „wystają" poza zbiór B, do części zbioru A, które „mieszczą się wewnątrz" zbioru B. Wzrokowo potrafimy to łatwo ocenić (por. rysunek 1) i wydaje się, że porównujemy przede wszystkim maksymalne odległości odpowiednich krzywych i powierzchnie odpowiednich figur (na rysunku 1 powierzchnie te zaczerniono i zakreskowano).



Rysunek 1. Części wystające i części schowane

Założyłem, że wzór na obliczanie stopnia zawierania się zbiorów powinien być na tyle prosty, by dało się go łatwo stosować w rozmaitych sytuacjach praktycznych. Nie możemy więc obliczać powierzchni, długości krzywych czy maksymalnych odległości między dowolnymi krzywymi.

Zaproponowany wzór dotyczy zbiorów rozmytych o funkcjach przynależności ciągłych i rzeczywistych, a ponadto odnosi się tylko do zbiorów, których nośniki są ograniczone. Niektóre z tych założeń można łatwo ominąć.

W omawianym wzorze wykorzystuję dwa łatwe do policzenia wskaźniki – wysokość oraz nośnik. Przypomnijmy:

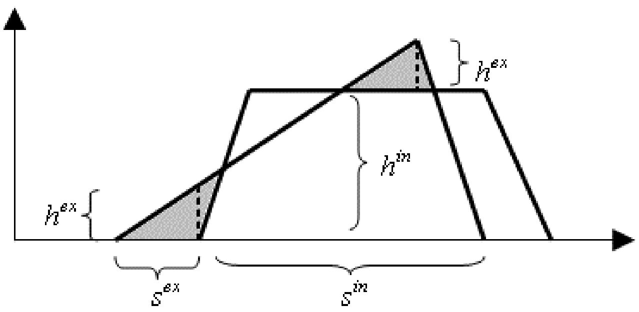
- wysokością zbioru rozmytego A nazywa się najwyższy stopień przynależności i oznacza się ją przez $h(A)$,
- nośnikiem zbioru rozmytego A nazywa się podzbiór tych elementów uniwersum, dla których stopień przynależności jest niezerowy, i oznacza przez $supp(A)$.

W omawianym wzorze wziąłem pod uwagę wysokość h^{ex} różnicy ograniczonej $A - B$, wysokość h^{in} iloczynu $A \cap B$, długość s^{in} nośnika części schowanej i długość s^{ex} nośnika części wystającej (por. rysunek 2).

Z pełnym zawieraniem A w B będziemy mieć do czynienia wtedy, gdy $h^{ex} = 0$ i $s^{ex} = 0$, czyli gdy A nigdzie nie wystaje poza B. Z brakiem zawierania mamy do czynienia w dwóch przypadkach: (1) wtedy gdy $h^{in} = 0$; (2) wtedy gdy $h^{ex} \geq h^{in}$ lub $s^{ex} \geq s^{in}$, przy czym ten drugi przypadek jest propozycją arbitralną i rozumiem, że może być ona dyskusyjna. Jeżeli jednak zgodzimy się z tym rozstrzygnięciem, to z częściowym zawieraniem będziemy mieć do czynienia w pozostałych przypadkach, tzn. wtedy gdy $0 \leq s^{ex} < s^{in}$ i zarazem $0 < h^{ex} < h^{in}$. Stopień taki może być liczony za pomocą każdego wzoru typu:

$$(1 - s^{ex} / s^{in})\ T\ (1 - h^{ex} / h^{in}),$$

gdzie T jest dowolną trójkątną operacją mnożenia (por. [3, s. 47–50), np. operacją minimum.



Rysunek 2. Wysokości i nośniki części schowanych i części wystających

Ostatecznie wzór na obliczanie stopnia zawierania się zbioru rozmytego A w zbiorze rozmytym B ma postać:

$$dg(A \subset B) = \begin{cases} 1 & \text{dla } h^{ex} = 0 \\ 0 & \text{dla } h^{ex} \geq h^{in} \text{ lub } s^{ex} \geq s^{in} \text{ lub } h^{in} = 0. \\ (1 - s^{ex} / s^{in})\ T\ (1 - h^{ex} / h^{in}) & \text{wpp} \end{cases}$$

Równość dwóch zbiorów ostrych zachodzi wtedy, gdy pierwszy z nich zawiera się w drugim, a drugi w pierwszym. Podobieństwo zbiorów rozmytych mogę więc zdefiniować jako mniejszy ze stopni tych dwóch rozmytych inkluzji:

$$dg(A \approx B) = dg(A \subset B) \wedge dg(B \subset A).$$

4.2. Podobieństwo oparte na metryce

Metryką na przestrzeni zbiorów rozmytych $F(X)$ na uniwersum X nazywamy funkcję m: $F(X) \times F(X) \rightarrow \mathbb{R}^+ \cup \{0\}$, taką, że:

- $m(A, B) \geq 0$
- $m(A, B) = m(B, A)$
- $A = B \Rightarrow m(A, B) = 0$
- $m(A, C) \leq m(A, B) + m(B, C)$.

Dla zbiorów rozmytych na uniwersum skończonym zaproponowano różne rodzaje odległości (por. [1, 2, 3]). Nas będą interesować tylko metryki znormalizowane, czyli przyjmujące wartości w przedziale [0, 1]. Przykładem takiej metryki jest tzw. odległość dwudzielna:

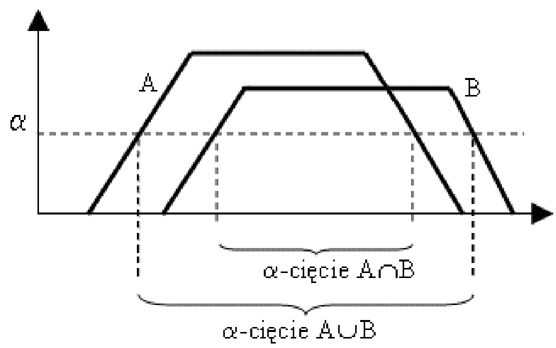
$$k(A, B) = 1 - |A \cap B| / |A \cup B|.$$

Wyznaczanie podobieństwa dwóch zbiorów rozmytych może polegać na obliczaniu dopełnienia ich znormalizowanej odległości. Stopień podobieństwa zbiorów rozmytych A i B to wówczas $E(A, B) = 1 - m(A, B)$, gdzie m jest dowolną metryką znormalizowaną. Na przykład dla metryki dwudzielnej mamy:

$$E(A, B) = |A \cap B| / |A \cup B|.$$

4.3. Podobieństwo oparte na przekrojach

Przyjmijmy, że X jest zbiorem liczb rzeczywistych, nośniki zbiorów rozmytych A i B są ograniczone, a każde α-przecięcie iloczynu i sumy tych zbiorów jest odcinkiem, skończoną sumą odcinków, zbiorem jednoelementowym (przecięcie na poziomie wartości szczytowej) albo zbiorem pustym (przecięcie powyżej wysokości). Przy takich założeniach możemy we wzorze $E(A, B) = |A \cap B| / |A \cup B|$ zastąpić iloczyn i sumę zbiorów rozmytych ich α-przekrojami, moce zaś tych zbiorów – ich długościami (długością zbioru ostrego będącego sumą odcinków jest suma długości tych odcinków). Stosunek długości α-cięcia iloczynu do długości α-cięcia sumy pokaże nam na każdym poziomie podobieństwo między zbiorami A i B (por. rysunek 3).

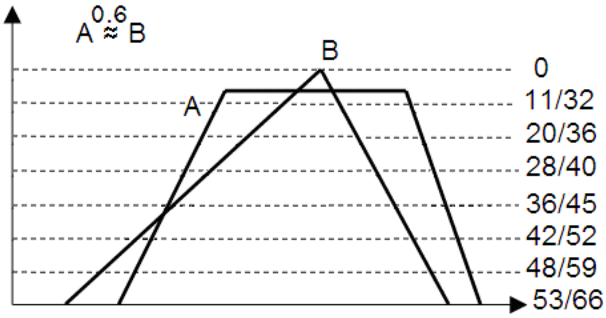


Rysunek 3. Przekrój sumy i iloczynu dwóch zbiorów na poziomie α

Jeśli policzymy średnią z tych liczb, to dla zbiorów identycznych uzyskamy wartość 1, dla zbiorów o pustym iloczynie – wartość 0, dla pozostałych zaś przypadków – stopnie dobrze oddające intuicyjnie pojmowane podobieństwo (por. rysunek 4).

W zastosowaniach praktycznych możemy ograniczyć się do przecięcia sumy i iloczynu zbiorów rozmytych A i B na skończonej, niewielkiej liczbie n poziomów:

$$E^n(A, B) = (1/n) \cdot \sum_{\alpha\ =\ 0,\ \lambda,\ 2\lambda,\ ...,\ n\lambda} |(A \cap B)_\alpha| \;/\; |(A \cup B)_\alpha|, \text{ gdzie } \lambda = [h(A) \vee h(B)] \;/\; n.$$



Rysunek 4. Stopień podobieństwa zbiorów dla $n = 8$

Podobieństwo zbiorów A i B z rysunku 4 policzono na dwa sposoby: dla $n = 8$ i dla $n = 80$. Otrzymane wyniki wynoszą 0,601 i 0,602, czyli są dostatecznie podobne, aby nie stosować tak wielu przecięć, jak w drugim z tych sposobów. W przypadku zbiorów łamanych (zbiorów rozmytych, których kształt przybliżono łamaną, por. [4, s. 45]) poziomy przecięć raczej nie powinny być rozłożone symetrycznie, tylko przechodzić przez wierzchołki łamanych.

5. Podsumowanie

Podobieństwo zbiorów dystrybutywnych jest często definiowane w matematyce klasycznej (np. w geometrii czy algebrze) niezbyt szczęśliwie, jako relacja równoważności. W matematyce zbiorów rozmytych eksplikacja pojęcia podobieństwa jest w pełni zgodna z jego znaczeniem w języku naturalnym i odpowiada relacji tolerancji.

Bibliografia

[1] Kacprzyk J., *Zbiory rozmyte w analizie systemowej*, Warszawa 1986.
[2] Lin C.-T., Lee C.S.G., *Neural Fuzzy Systems. A Neuro-Fuzzy Synergism to Intelligent Systems*, New York 1996.
[3] Łachwa A., *Rozmyty świat zbiorów, liczb, relacji, faktów, reguł i decyzji*, Warszawa 2001.

[4] Łachwa A., *Który zbiór wybrać?*, „Acta Academiae Modrevianae. Informatyka”, Kraków 2006, s. 35–49.
[5] Szrejder J.A., *Równość, podobieństwo, porządek*, Warszawa 1975.
[6] Zadeh L.A., *Fuzzy sets*, „Information and Control” 1965, vol. 8, s. 338–353.