Komputerowe przetwarzanie wiedzy

Tomasz Kubik

5 grudnia 2005

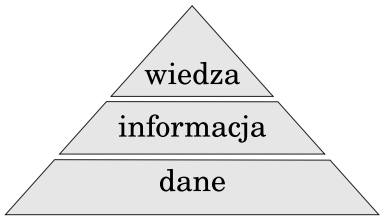
1 Wiedza

Najczęściej wiedza jest niepełna, nieprecyzyjna, niepewna ($3 \times N$):

- ograniczenia percepcji (nie da się obserwować wszystkiego jednocześnie, lub obserwacje są nieprecyzyjne - zasada Heisenberga?),
- ograniczenia reprezentacji otoczenia (gdy problem ma dużą złożoność zazwyczaj stosujemy uproszczenia)
- brak danych lub ich słabe oszacowanie.

1.1 Zarządzanie wiedzą

[1] Jednym z popularniejszych sposobów tłumaczenia, czym jest wiedza w systemach informatycznych, jest posłużenie się modelem piramidy, Rys.1. Dane, tworzące podstawę piramidy odpowied-

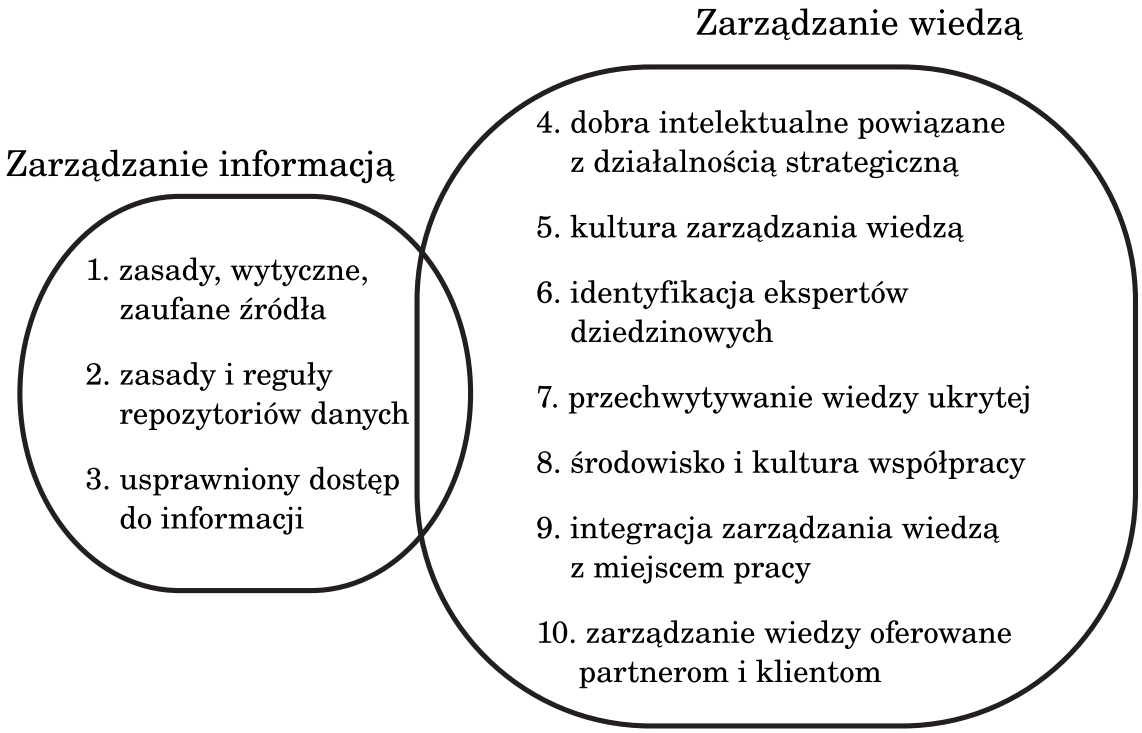


Rysunek 1: Piramida wiedzy

nio przetworzone (innymi słowy - umieszczone w odpowiednim kontekście), stają się informacją. Z kolei przetworzona (czytaj: umieszczona w kontekście) informacja staje się wiedzš. Przykładowo: z parametrycznego opisu danych `CENA_AKCJI(ACNE S.A.;10-12-2002;43,50)` można uzyskać informacje, że ćena zamknięcia akcji spółki ACNE S.A. wyniosła dziesiątego grudnia 43,40PLN, co daje 10 procentowy wzrost w stosunku do dnia poprzedniego i najwyższy wzrost na parkiecie tego dnia". Z tej informacji, poprzez wykorzystanie ludzkiego doświadczenia i intuicji (a więc ekspertyzy) możemy otrzymać wiedzę: „[...] co, analizując sytuację na rynku oraz przewartościowanie papierów każe spodziewać się szybkiej korekty kursu; REKOMENDACJA: sprzedaj". Na popularnych piramidach pojawia się jeszcze czwarty element, na samym szczycie - „mądrość". Tylko jak tu zdefiniować mądrość?

[1] Autorem niektórych fragmentów tekstu w tym podrozdziale jest Marek Kowalkiewicz

Zarządzanie wiedzą oraz zarządzanie informacjami pełnią niepoślednią rolę w nowoczesnych technikach zarządzania. Powiązanie tych dwóch obszarów ilustruje diagram Gartnera, Rys.2. Definicja Thomasa H. Davenporta, Larry'ego Prusaka (kluczowych postaci od zarządzania wie-



Rysunek 2: Zarządzanie wiedzą i informacjami

dzą), mówi, że wiedza jest to połączenie doświadczenia, ocen wartości, informacji o kontekście oraz analitycznego wglądu w zagadnienia, które zapewnia ramy dla oceny i wyłączania nowych doświadczeń i informacji; wiedza organizacji wywodzi się i jest charakterystyczna dla umysłów ludzi. Definicja ta dotyczy zagadnień, które w powszechnym rozumieniu nie mogą być przechowywane i przetwarzane na nośnikach elektronicznych (analityczny wgląd w zagadnienia, umysł ludzki).

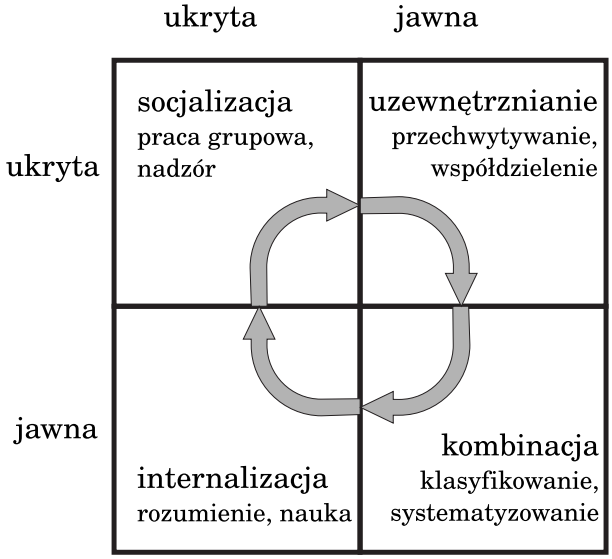
Pojęcia wiedzy i zarządzania wiedzą nie były i nie są ściśle powiązane z technologiami informacyjnymi. Natomiast technologie informacyjne z całą pewnością mogą posłużyć do wspomagania zarządzania wiedzą.

W kategoriach zarządzania wiedzą przyjęło się rozróżniać dwa typy wiedzy: wiedzę ukrytą (*Tacit knowledge*) oraz wiedzę jawną (*Explicit knowledge*). Wiedza ukryta jest bardzo trudna, lub niemożliwa, do skodyfikowania. Wiedza ukryta - „to ta, którą dysponuje garncarz i dzięki istnieniu której, nawet stos przeczytanych książek nie pozwoli zrobić informatykowi równie dobrych garnków". Nie obejdzie się bez kontaktu osobistego, dzięki któremu taka wiedza może zostać przekazana (proces ten nazywany jest również socjalizacją wiedzy). Wiedzę ukrytą, w przypadku systemów informacyjnych, możemy postrzegać jako naszą wewnętrzną wiedzę, kontekst, w jakim postrzegamy inne informacje. Dzięki temu, ten sam fakt, jest różnie interpretowany przez odbiorców. To dzięki wiedzy ukrytej ekspert jest ekspertem, to przez brak wiedzy ukrytej niektórzy nie radzą sobie z rozwiązywaniem problemów technicznych, które nie są wprost opisane w instrukcjach. Wiedza jawna z kolei, to wiedza, którą stosunkowo łatwo skodyfikować. Są to różnego rodzaju opisy procesów, sugestie określające sposoby wykonywania odpowiednich zadań, itd.

Nonaka i Takeuchi zapisali zależności pomiędzy wiedzą jawną a ukrytą w diagramie jak na

Rys. 3. Wiedza ukryta jest przekazywana innym w procesie socjalizacji, w trakcie pracy grupowej, oraz pod nadzorem „mistrzów". Gdy uda nam się przechwycić i współdzielić wiedzę ukrytą, następuje proces uzewnętrzniania (eksternalizacji). Tak uzewnętrzniona wiedza może być z kolei usystematyzowana i sklasyfikowana w procesie łączenia (kombinacji). Ostatecznie, taka wiedza może zostać przyswojona (internalizowana) w procesie rozumienia i nauki. W ten sposób zamyka się koło. Każde przejście cyklu zarządzania wiedzą powinno przenosić nas na wyższy poziom, poprzez zwiększenie jakości zarządzanej wiedzy. Tworzy to spiralę obiegu wiedzy.

Diagram Nonaki i Takeuchiego nie porusza problemu tzw. wiedzy nieznanej (hidden knowledge), tj. takiej, o której istnieniu uczestnik procesu nie zdaje sobie sprawy, albo nie przewidział jej istotności w kontekście procesu zarządzania wiedzą. Kolejny typ wiedzy, która nie jest tutaj zawarta, a o której warto wspomnieć, to tzw. wiedza potencjalna, a więc taka, której powstanie nie jest oczekiwanym efektem powyższego cyklu zarządzania wiedzą. Wiedza potencjalna może powstać w trakcie każdego z wymienionych etapów cyklu poprzez starcie kontekstu (a konkretniej wiedzy, doświadczenia oraz intuicji odbiorcy) z wiedzą przetwarzaną (uzewnętrznianą, łączoną, przyswajaną lub socjalizowaną).



Rysunek 3: Model zarządzania wiedzą

1.2 Przetwarzanie wiedzy

Przetwarzanie wiedzy podzielić można na dwa etapy:

- gromadzenie wiedzy, klasyfikacja, agregacja (uczenie)
- udostępnianie, wykorzystywanie (wnioskowanie)

1.3 Model Act*

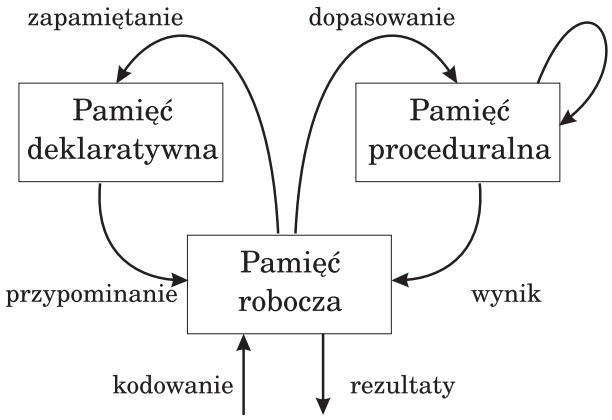
W modelu Act* (John Anderson, CMU) wyróżnia się 3 rodzaje pamięci:

- Pamięć deklaratywna (długotrwała)
  (czysta wiedza - co wiadomo o obiekcie, jak powiązany ?)

W pamięci tej przechowywane są abstrakcyjne pojęcia, reguły oraz łączące je relacje (sieć semantyczna + mechanizm asocjacji) np. krotka opisująca dane personalne osoby w bazie relacyjnej

- Pamięć proceduralna (długotrwała)
  (umiejętność - jak postępować ?)
  Wiedza zapisana niejawnie, trudna do przekazania i modyfikacji, może mieć postać reguł produkcji, z których każda zawiera zbiór warunków i akcji (taka wiedza może być np. algorytmem sortowania zapisany w postaci programu). Wiedza proceduralna może być rozszerzana przez wnioskowanie na bazie aktualnie posiadanej wiedzy. ACT* obsługuje trzy podstawowe typy uczenia: generalizację (w której zasięg zastosowań reguł produkcji poszerza się), dyskryminację (w której zasięg zastosowań zmniejsza się), wzmacnianie (w którym niektóre reguły produkcji stosowane są częściej niż inne). Nowe reguły produkcji tworzone są przez koniunkcje lub dysjunkcję istniejących reguł.

- Pamięć robocza (cache)
  pamięć robocza jest to ta część pamięci długotrwałej, której poziom aktywacji jest najwyższy.



Rysunek 4: Struktura modelu Act*

1.4 Reprezentacja

Jedna z najprostszych form reprezentacji wiedzy jest reprezentacja w przestrzeni cech/stanów, gdzie mamy:

- opis problemu (np. wektor parametrów opisujących stan obiektu)
- opis sposobów transformacji (operatory transformujące stan)

W opisie tym istotną rolę odgrywają parametry. Wartość parametru (atrybuty/cechy/parametru) pojedynczego obiektu może być:

- ciągła (np. liczba rzeczywista, wzrost, cena towaru),
- quazi-ciągła, tzn. ciągła z dokładnością do ziarna dyskretyzacji (np. liczba rzeczywista w komputerze, liczba całkowita, czas wyrażany w ilościach sekund/dni)

- dyskretna uporządkowana, jak w przypadku, gdy kolejne wartości atrybutu tworzą ciąg uporządkowany (według pewnego kryterium). Dla każdej pary wartości możemy określić ich relację. Np. postać leku: płynna < półpłynna-maść < sypka-proszek < stała-tabletka. Pora dnia: świt < rano < przedpołudnie < południe < popołudnie < ...

- dyskretna częściowo uporządkowana, gdy relacja uporządkowania jest określona tylko częściowo (tylko dla niektórych par wartości) np. marki samochodów: (Trabant, Maluch) < (Polonez, Łada) < (Opel, Volkswagen) < (Porsche, Jaguar)

- dyskretna nieuporządkowana, odzwierciedla fakt, iż atrybuty pozbawione jakichkolwiek relacji porządkujących rzadko występują w praktyce, ale często te relacje są bardzo słabe lub nieznane np. meble, gatunki filmów

Dodatkowe cechy atrybutów:

- informacja niepewna (prawdopodobieństwo, stochastyka, np. albo zdam egzamin albo nie zdam)

- informacja niepełna (NULL, np. znamy tylko fragment numeru rejestracyjnego, inicjały, itp)

- informacja nieprecyzyjna (niepewność pomiarowa, statystyka, np. średnica opony 45cm $\pm$5mm)

- informacja niewerbalna (trudna do wrażenia za pomocą liczb, niepewność lingwistyczna, np. interesujący, pompatyczny, skromny)

Rodzaje reprezentacji wiedzy (związane ze sposobem reprezentacji atrybutów):

- reprezentacja parametryczna, wiedza o przedmiotach i obiektach reprezentowana jest w przestrzeni cech (często jest to ciągła przestrzeń), np. rozmiar, waga, położenie, pojemność, kolor

- reprezentacja symboliczna, wiedza opisana jest identyfikatorami obiektów występujących we wzajemnych relacjach (zazwyczaj reprezentacja dyskretna), np. $X > Y$, gdzie X, Y - identyfikatory obiektów.

- reprezentacja temporalna i przestrzenna, wiedza reprezentowana jest przez własności obiektów zmieniające się „w czasie" i „w przestrzeni", np. równanie trajektorii ruchu, zmiany kursu walut, krzywa naładowania akumulatora w zależności od czasu,

Sposób reprezentacji wiedzy w dużej mierze zależy od jej natury. Do różnych problemów używa się więc różnych reprezentacji. Oto krótka ich lista:

- reprezentacja logiczna dwuwartościowa i wielowartościowa
  - logika boolowska,
  - logika pierwszego rzędu
  - logika rozmyta

- reprezentacja symboliczna (abstrakcyjna)
  - rachunek zdań
  - reprezentacja stanu otoczenia za pomocą predykatów:

```
{ on( floor, box_A ), on( box_A, box_B ), clear( box_B ) }
```

  - reprezentacja działań w postaci list warunków, skreśleń, dopisków:

```
put_on( obj_1, obj_2) {
    pre: clear( obj_1 ), clear(obj_2 );
    remove: clear( obj_1 );
    add: on( obj_1, obj_2 )
}
```

  - abstrakcyjna reprezentacja celu,

```
on( * , box_C )
```

- reprezentacja parametryczna
  - reprezentacja wektorowa:
    wady: duża złożoność w przypadku modelowania, kłopoty z modelowaniem niepewności, zwłaszcza dla obiektów „nieliniowych"
    zalety: szybkość przetwarzania, łatwość aktualizacji
  - reprezentacja rastrowa:
    wady: rozdzielczość, zajętość pamięci, złożoność obliczeniowa i kłopoty z aktualizacją
    zalety: łatwe modelowanie nieprecyzyjności lub niepewności
- reprezentacja za pomocą reguł generowania rozwiązań (systemy produkcyjne)

1.5 Planowanie działań

Algorytm rozwiązujący wyznacza, w ramach posiadanej wiedzy i przyjętej reprezentacji, rozwiązanie (plan działania) najlepiej spełniające zadany cel.

Planowanie działania polega na zdefiniowaniu relacji pomiędzy rejestrowanymi obserwacjami a wykonywanymi akcjami:

- w postaci dyskretnej plan działania przybiera formę tabeli przejść automatu skończonego (np. sterowniki prostych robotów)
- w postaci ciągłej jest to funkcja analityczna odwzorowująca przestrzeń percepcji w przestrzeń akcji (takie przyporządkowanie musi być jednoznaczne !)
- w postaci zbioru reguł (problem zupełności i spójności !)

Rodzaje algorytmów rozwiązujących:

- gdy przestrzeń jest ciągła:
  - analityczne wyznaczanie rozwiązań (teoria sterowania i optymalizacji)
  - wprowadzenie sztucznej dyskretyzacji przestrzeni, ziarno dyskretyzacji może być zmienne (metody numeryczne) lub stałe (gdy przestrzeń rozwiązań jest mała i przeliczalna)
- gdy przestrzeń jest dyskretna:
  - metody przeglądu grafu/drzewa reprezentującego taką przestrzeń:
    * poszukiwanie węzła
    * poszukiwanie ścieżki

    * poszukiwanie poddrzewa/podgrafu
    * poszukiwanie przekroju

- metody symulacyjne:
  - metody potencjałowe (dla przestrzeni ciągłej)
  - metody elementów skończonych, np. wave front
  - metody ewolucyjne

1.6 Perspektywa filozoficzna

[2] Problemy wiedzy i możliwości jej pozyskiwania zajmowały filozofów prawie od początku dziejów filozofii Zachodu. Poniżej wymieniam najbardziej znanych myślicieli, których myśl w jakiś sposób wydaje się z tymi problemami wiązać. Podsumowanie dorobku tych wybitnych postaci za pomocą jednego hasłowego zdania należy traktować ze sporym przymrużeniem oka (to jak kurs historii filozofii w 10 minut).

Sokrates: uczenie się utożsamione z przypominaniem sobie wiedzy wrodzonej (anamneza). Pewne pokrewieństwo z uczeniem się na podstawie wyjaśnień, które polega na przetwarzaniu wiedzy wrodzonej do dogodnej postaci pod wpływem obserwowanych przykładów.

Platon: realnie istniejące idee (pojęcia). Uczenie się nie polega na tworzeniu pojęć, lecz odkrywaniu (a właściwie przypominaniu sobie) istniejących odwiecznie pojęć.

Arystoteles: istnieją tylko rzeczy jednostkowe, pojęcia ogólne powstają na drodze abstrakcji. Nie są one jednak arbitralne i pozbawione znaczenia, gdyż to co ogólne (wspólne) w rzeczach zawiera ich forma. Rozróżnienie dedukcji i indukcji.

Augustyn: idee istnieją jako myśli Boga, który według nich stworzył rzeczy.

Abelard: rozstrzygnięcie sporu o uniwersalia według którego nie istnieją one realnie, ale mają podstawę we wspólnej formie rzeczy obejmowanych przez nie.

Tomasz z Akwinu: uniwersalia *ante rem*, *in re*, *post rem*, podział procesów umysłowych na trzy kategorie - tworzenie pojęć, wydawanie sądów oraz przeprowadzanie wnioskowania.
Powszechniki mogą mieć trojaką postać:

1) powszechnik może być zawarty w substancji jednostkowej, której istotę stanowi: to universale in re, zwane także przez Tomasza powszechnikiem bezpośrednim (universale directum)
2) powszechnik może być wyabstrahowany przez umysł: jest to universale post rem, które Tomasz nazwał refleksyjnym (reflexium). W tej postaci nie istnieje on w jednostkowych rzeczach; realnie (formaliter) istnieje tylko w umyśle, a jedynie podstawę ma (fundamentaliter) w rzeczach
3) poza tym należy jeszcze przyjąć powszechnik niezależny od rzeczy, universale ante rem: jest on ideą w umyśle Bożym, wzorem wedle którego Bóg stworzył świat realny.

Ockham: brzytwa Ockhama wzywająca do tłumaczenia świata za pomocą najprostszych hipotez. Nominalizm (pojęcia ogólne to tylko część języka).

[2] Wyciąg z wykładu Pawła Cichosza

Kartezjusz: idee wrodzone.

Hume: empiryzm, idee proste powstają na podstawie wrażeń, idee złożone są ich konstrukcjami. Idee ogólne powstają przez połączenie idei konkretnych. Krytyka zasady przyczynowości, która w konsekwencji podważa prawomocność indukcji (ale akceptacja dla posługiwania się nią w praktyce).

Kant: aprioryczne formy zmysłowości (czas, przestrzeń) i kategorie rozumu (m.in. substancja i przyczynowość) kształtujś poznanie. Z perspektywy indukcyjnego uczenia się maszyn można na to patrzeć jako na *bias*, czyli czynnik determinujący wybór przez ucznia określonej hipotezy na podstawie dostępnych danych.

Mill: kanony indukcji (zgodności, różnicy, reszt, zmian towarzyszących). Raczej chybione z punktu widzenia metodologii nauk przyrodniczych, lecz bliskie uczeniu się maszyn.

Avenarius: zasada ekonomii myślenia wzywająca do konstrukcji pojęć i praw, które opisują świat w najprostszy sposób.

James: pragmatyzm utożsamiający prawdę z długoterminową użytecznością (*truth = cash value*), co odpowiada w pewnym sensie założeniom uczenia się ze wzmocnieniem.

Koło wiedeńskie (m.im. Schlick, Carnap, Neurath) zdania protokolarne jako podstawa wiedzy, weryfikowalność jako kryterium sensowności.

Russell: podstawy współczesnej logiki matematycznej (także Frege, Whitehead). Odróżnienie formy gramatycznej zdań od ich formy logicznej.

Wittgenstein: zwrócenie uwagi na rolę języka, który wyznacza granice myśleniu (język reprezentacji hipotez określa, czego można się nauczyć). Język logiki jako język uniwersalny (w pierwszym okresie). Względność i równoprawność języków (w drugim okresie).

Carnap: problemy indukcji i prawdopodobieństwa wiedzy z niej pochodzącej.

Popper: antyindukcjonizm w metodologii nauk, poznanie przez proponowanie śmiałych falsyfikowalnych hipotez i ich testowanie.

1.7 Myślenie

[3] Matematyk Alan Turing (1912-1954) zaproponował następujący test, czy badany obiekt wykazuje się inteligencją: komunikujemy się z nim za pomocą neutralnego urządzenia (ekran tekstowy) i prowadzimy konwersację - jeżeli nie jesteśmy w stanie stwierdzić, czy po drugiej stronie jest człowiek, czy maszyna, obiekt badany posiada inteligencję nieodróżnialną od ludzkiej. Test ten nazywany jest testem Turinga. Test ten ma wyraźne zabarwienie behawiorystyczne - rozpoznajemy obiekt po jego zachowaniu, nie wdając się w analizę niedostępnego z zewnątrz jakiegoś życia wewnętrznego. Jeżeli chodzi jak kaczka, kwacze jak kaczka i wygląda jak kaczka, to prawdopodobnie jest to kaczka.

Podejmowano próby stworzenia programu, potrafiącego przejść test Turinga. Starym i najbardziej znanym przykładem, jest program ELIZA, prowadzący dialogi symulujące psychiatrę. Podaję link do oryginalnego artykułu Josepha Weizenbauma i do strony, gdzie można z tym programem porozmawiać. Każdy może ocenić, czy program zdał test.

[3] Autorem tekstu tego paragrafu jest Jan Śliwa

http://i5.nyu.edu/%7Emm64/x52.9265/january1966.htmli
http://www.nu-woman.com/eliza.htm

Z drugiej strony filozof John Searle (ur. 1932) sformułował tzw. argument chińskiego pokoju. Wyobraźmy sobie maszynę, która tak się zachowuje, jakby rozumiała chiński. Odbiera chińskie teksty i wykonywany przez nią program pozwala jej na wydawanie odpowiedzi po chińsku tak dobrze, że przechodzi test Turinga na osobnika rozumiejącego chiński. Ale czy naprawdę rozumie? Searle napisał: pójdźmy dalej - załóżmy, że to ja siedzę wewnątrz tej maszyny, wykonuję jej algorytm, przetwarzam wejściowe ciągi symboli na wyjściowe i zaliczam test Turinga. Ale ja WIEM, że chińskiego nie rozumiem, że manipuluję wyłącznie symbolami według zadanych przez kogoś reguł. Co dowodzi, że można zdać test, a nie rozumieć przetwarzanego tekstu. Argumentcja Searle'a sprowadza się w skrócie do dwóch punktów:

1. Symulacja procesów mentalnych, to nie proces mentalny (jak symulacja jedzenia hamburgerów, to nie jedzenie hamburgerów)

2. Przetwarzanie syntaktyczne, to nie przetwarzanie semantyczne

Wnioskowanie: przesłanki → konkluzja.

Schemat *modus ponens*: dla zdań p i q jeśli $p \rightarrow q$ jest prawdziwe i p jest prawdziwe, to q prawdziwe.

Schemat *modus tollens*: dla zdań p i q jeśli $p \rightarrow q$ jest prawdziwe i q jest fałszywe, to p jest fałszywe.

2 Prawdopodobieństwo

2.1 Definicja klasyczna (Laplace'a)

Autorem klasycznej definicji prawdopodobieństwa jest Pierre Simon de Laplace, który ogłosił ją w roku 1812.

> Prawdopodobieństwem zajścia zdarzenia A nazywamy iloraz liczby zdarzeń sprzyjających zdarzeniu A do liczby wszystkich możliwych przypadków, zakładając, że wszystkie przypadki wzajemnie się wykluczają i są jednakowo możliwe.

Definicję tą można zapisać również w bardziej formalny sposób. Oznaczmy zbiór wszystkich możliwych przypadków przez Ω. Elementami zbioru Ω są zdarzenia elementarne ω, zaś zbiór Ω to zbiór zdarzeń elementarnych. Zbiór zdarzeń sprzyjających A będzie w takim wypadku podzbiorem zbioru Ω: $A \subset \Omega$.

> Prawdopodobieństwo zajścia zdarzenia A możemy zapisać w postaci:
>
> $$P(A) = \frac{|A|}{|\Omega|}, \tag{1}$$
>
> gdzie $|A|$ oznacza liczbę elementów (moc) zbioru A, zaś $|\Omega|$ liczbę elementów (moc) zbioru Ω.

Definicja klasyczna pozwala obliczać prawdopodobieństwo w prostych przypadkach, jednak zawiera szereg wad: nie można jej stosować dla zbiorów nieskończonych, a przede wszystkim zawiera błąd logiczny. Zdarzenia elementarne muszą być jednakowo możliwe, co znaczy przecież to samo co jednakowo prawdopodobne.

2.2 Prawdopodobieństwo geometryczne

Definicja klasyczna nie pozwala obliczać prawdopodobieństwa w przypadku, gdy zbiory A i Ω są nieskończone, jednak jeśli zbiory te mają interpretację geometryczną, zamiast liczebności zbiorów można użyć miary geometrycznej (długość, pole powierzchni, objętość).

2.3 Definicja częstościowa

W 1931 roku Richard von Mises zaproponował, żeby zdefiniować prawdopodobieństwo jako granicę ciągu częstości:

$$P(A) = \lim_{n \to 0} \frac{k_n(A)}{n}, \tag{2}$$

gdzie $k_n(A)$ to liczba rezultatów sprzyjających zdarzeniu A po n próbach. Definicja ta nie mówi jednak nic o warunkach istnienia granicy i dlatego nie spełnia wymogów formalnych.

2.4 Aksjomaty Kołmogorowa

W 1933 Andriej Kołmogorow podał aksjomatyczną definicję prawdopodobieństwa.

> Zakładając, że Ω jest zbiorem zdarzeń elementarnych ω, zaś M jest σ-ciałem zbioru Ω, prawdopodobieństwem nazywa się funkcję $P : M \to R$ spełniająca następujące warunki:
>
> 1. $P(A) \geq 0$ dla każdego $A \in M$.
>
>   Oznacza to, że prawdopodobieństwo zajścia zdarzenia A wyraża się liczbą nieujemną.
> 2. $P(\Omega) = 1$.
>
>   Czyli prawdopodobieństwo, że wystąpi jakieś zdarzenie elementarne w przestrzeni Ω wynosi 1. Innymi słowy: nie ma zdarzeń elementarnych poza zbiorem Ω. Jeśli nie możemy określić zbioru Ω, nie jesteśmy też w stanie zdefiniować prawdopodobieństwa na tym zbiorze.
> 3. Jeśli (A_n) jest dowolnym ciągiem podzbiorów M parami rozłącznych, to:
> $P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n)$
>
>   To znaczy: prawdopodobieństwo zdarzenia, które jest sumą rozłącznych zdarzeń, obliczamy jako sumę prawdopodobieństw tych zdarzeń. Własność tą nazywamy *sigma-addytywnością*.

2.5 Ciało zbiorów

Ciało zbiorów to rodzina zbiorów A spełniająca następujące warunki:

1. zbiór pusty należy do A
2. dopełnienie zbioru należącego do A należy do A
3. suma dwóch (a więc i skończenie wielu) zbiorów należących do A należy do A.

Bardzo często rozważa się ciała zbiorów z dodatkowym warunkiem:

4. suma przeliczalnie wielu zbiorów z A należy do A.

Mówimy wtedy o σ-ciele (czytaj: sigma-ciele) .

Z warunków 2 i 3 wynika, że różnica dwóch zbiorów z A należy do A. Podobnie, część wspólna skończenie wielu zbiorów z A należy do A (w przypadku σ-ciała część wspólna przeliczalnie wielu zbiorów).

Czasem zamiast ciało (σ-ciało) zbiorów używa się terminu algebra zbiorów (odpowiednio σ-algebra).

Przykłady i własności:

1 Rodzina wszystkich podzbiorów dowolnego zbioru jest ciałem zbiorów (a nawet σ-ciałem).

2 Rodzina podzbiorów danego zbioru X złożona ze zbioru pustego i zbioru X jest ciałem zbiorów.

3 Dla danego podzbioru A zbioru X rodzina utworzona ze zbioru pustego, zbioru X, zbioru A i jego dopełnienia A' jest ciałem zbiorów.

4 Przekrój dowolnej rodziny ciał zbiorów (σ-ciał) jest znów ciałem zbiorów (σ-ciałem).

5 Dla dowolnej rodziny zbiorów A istnieje najmniejsze ciało (σ-ciało) zbiorów zawierające wszystkie zbiory tej rodziny. Nazywamy je ciałem (σ-ciałem) generowanym przez tę rodzinę.

2.6 Prawdopodobieństwo warunkowe

Prawdopodobieństwo B pod warunkiem A:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \tag{3}$$

stąd

$$P(A \cap B) = P(A)P(B|A) \tag{4}$$

Używając prawdopodobieństwa warunkowego można zdefiniować regułę całkowitego prawdopodobieństwa.

> Jeśli wynik doświadczenia jest jednym z n możliwych, wzajemnie wykluczających się zdarzeń A_i, (czyli $\Omega = A_1 + A_2 + \ldots + A_n$), to prawdopodobieństwo pojawienia się jakiegokolwiek zdarzenia o własności B wynosi
>
> $$P(B) = \sum_{i=1}^{n} P(A_i)P(B|A_i). \tag{5}$$

Dwa zdarzenia A i B są niezależne, jeśli fakt występowania A nie zmienia prawdopodobieństwa B i na odwrót, tzn. jeśli $P(B|A) = P(B)$ czyli $P(A \cap B) = P(A)P(B)$.

2.7 Zmienna losowa

Zmienna losowa X to funkcja mierzalna z przestrzeni probabilistycznej Ω do zbioru liczb rzeczywistych. Mierzalność rozumiemy względem σ-ciała zdarzeń w Ω i σ-ciała zbiorów borelowskich w R.

Odwzorowanie mierzalne z Ω w przestrzeń euklidesową R^n nazywa się wektorem losowym. Odwzorowanie takie ma postać $X(\omega) = (X_1(\omega), X_2(\omega), \ldots, X_n(\omega))$, gdzie $X_i(\omega)$ są zmiennymi losowymi.

Przykłady: Niech Ω będzie zbiorem wszystkich możliwych wyników dwukrotnego rzutu kostką - zbiór ten składa się z 36 par postaci (i, j), gdzie $1 \leq i, j \leq 6$. Następujące funkcje są oczywiście zmiennymi losowymi: „iloczyn liczby oczek wyrzuconej za pierwszym i drugim razem", „suma liczby oczek wyrzuconej za pierwszym i drugim razem", „liczba oczek wyrzuconych za pierwszym razem".

Rozróżniamy zmienne losowe typu skokowego (dyskretnego) i typu ciągłego.

2.8 Rozkład zmiennej losowej

Niech x będzie liczbą rzeczywistą o wartości zawartej pomiędzy $-\infty$ i ∞.

2.8.1 Dystrybuanta

Dystrybuantą zmiennej losowej X nazywamy funkcję:

$$F(x) = P(X < x), \tag{6}$$

określającą prawdopodobieństwo zajścia zdarzenia $X < x$. Pewne własności wynikające z tej definicji:

$\lim_{x\to\infty} F(x) = \lim_{x\to\infty} P(X < x) = P(\Omega) = 1.$

$P(X \geq x) = 1 - F(x) = 1 - P(X < x).$

$\lim_{x\to-\infty} F(x) = 0.$

2.8.2 Gęstość prawdopodobieństwa

Ciągłe dystrybuanty, które wszędzie mają pochodną dają się różniczkować. Funkcja gęstości prawdopodobieństwa zmiennej losowej X (lub w skrócie gęstość prawdopodobieństwa) wyraża się wzorem:

$$f(x) = \frac{dF(x)}{dx} = F'(x) \tag{7}$$

Pewne własności wynikające z tej definicji:

$$P(x < a) = F(a) = \int_{-\infty}^{a} f(x)dx$$

$$P(a \leq x < b) = F(b) - F(a) = \int_{a}^{b} f(x)dx$$

Istnieją funkcje gęstości prawdobodobieństwa, dla których nie istnieje analityczna postać odpowiadających im dystrybuant (jak ma to miejsca w przypadku funkcji gęstości prawdopodobieństwa rozkładu normalnego).

2.9 Funkcje zmiennej losowej, wartość oczekiwana, wariancja, momenty

2.9.1 Funkcja zmiennej losowej

Funkcja zmiennej losowej również jest zmienną losową

$$Y = H(X) \tag{8}$$

i posiada swoją dystrybuantę i gęstość prawdopodobieństwa.

2.9.2 Wartość oczekiwana

Wartość oczekiwana nazywana jest też wartością przeciętną, wartością średnią lub nadzieją matematyczną. Wartość oczekiwana nie jest ona zmienną losową.

W przypadku zmiennej losowej dyskretnej wartość oczekiwana jest sumą iloczynów wartości tej zmiennej losowej oraz prawdopodobieństw, z jakimi te wartości są przyjmowane:

$$E(X) = \bar{x} = \sum_{i=1}^{n} x_i P(X = x_i) \tag{9}$$

Wartość oczekiwana funkcji zmiennej losowej dyskretnej określa wzór:

$$E[H(X)] = \sum_{i=1}^{n} H(x_i) P(X = x_i) \tag{10}$$

W przypadku zmiennej losowej typu ciągłego (z różniczkowalną dystrybuantą) wartość oczekiwana dana jest wzorem:

$$E(X) = \bar{x} = \int_{-\infty}^{\infty} x f(x) dx \tag{11}$$

zaś wartość funkcji zmiennej losowej dana jest wzorem:

$$E[H(X)] = \int_{-\infty}^{\infty} H(x) f(x) dx \tag{12}$$

Wartość oczekiwana funkcji zmiennej losowej postaci

$$H(X) = (X - c)^k \tag{13}$$

nosi nazwę momentu rzędu k względem punktu c, μ_k:

$$\mu_k = E[(X - c)^k] \tag{14}$$

2.9.3 Momenty zmienej losowej

Moment zwykły rzędu k zmiennej losowej to wartość oczekiwana k-tej potęgi tej zmiennej dla $c = 0$:

$$m_k = E[(X)^k] \tag{15}$$

Wartość oczekiwana może być traktowana jako pierwszy moment zwykły m_1. Estymatorem wartości oczekiwanej rozkładu cechy w populacji jest średnia arytmetyczna.

Moment centralny rzędu k to moment wyznaczany względem wartości średniej:

$$\mu_k = E[(X - \bar{x})^k] \tag{16}$$

Momenty centralne mają szczególne znaczenie w statystyce. Wartości najniższych momentów centralnych wynoszą

$$\mu_0 = 1, \qquad \mu_1 = 0. \tag{17}$$

Wariancji zmiennej losowej to moment centralny rzędu 2

$$\mu_2 = \sigma^2(x) = \mathrm{var}(x) = E[(X - E(x))^2] = E[(X - \bar{x})^2]. \tag{18}$$

Zawiera on informację o średnim odchyleniu zmiennej losowej X od jej wartości średniej.

Odchylenie standardowe (dyspersja) zmiennej losowej X:

$$\sigma = \sqrt{\sigma^2(x)} \tag{19}$$

jest miarą średniego odchylenia wyników pomiarów zmiennej losowej X od jej wartości oczekiwanej. Często wartość ta utożsamiana jest z błędem pomiaru (jest tego samego wymiaru co X).

Współczynnik skośności (skośność) to moment centralny rzędu trzeciego, μ_3. Często zamiast niego korzysta się z bezwymiarowego parametru:

$$\gamma = \frac{\mu_3}{\sigma^3} \tag{20}$$

W przypadku gdy $H(X) = cX$ (c-stała), to:

$$E(cX) = cE(x) \tag{21}$$

$$\sigma^2(cX) = c^2\sigma^2(x) \tag{22}$$

Jeżeli istnieją $E(X)$ i $E(Y)$ to:

$$\forall_{a,b} \quad E(aX + bY) = aE(X) + bE(Y)(\text{liniowość}) \tag{23}$$

i stąd

$$\sigma^2(X) = E[(X - \bar{x})^2] = E(X^2 - 2X\bar{x} + \bar{x}^2) = E(X^2) - \bar{x}^2. \tag{24}$$

(Jeżeli zmienne X, Y są niezależne, to $E(XY) = E(X)E(Y)$).

Niech zmienna losowa U dana będzie wzorem

$$U = \frac{X - \bar{x}}{\sigma(X)} \tag{25}$$

Wtedy wartość oczekiwana

$$E(U) = \frac{1}{\sigma(X)} E(X - \bar{x}) = 0 \tag{26}$$

oraz wariancja

$$\sigma^2(U) = \frac{1}{\sigma^2(X)} E[(X - \bar{x})^2] = 1. \tag{27}$$

Zmienna standaryzowana (standardowa, normalizowana lub bezwymiarowa) to zmienna losowa o wartości średniej równej zeru i jednostkowej wariancji (jak powyższa zmienna U).

2.9.4 Wartość modalna

Wartość modalna x_m (wartość najbardziej prawdopodobna, moda) rozkładu to wartość zmiennej losowej, której odpowiada maksimum prawdopodobieństwa:

$$P(X = x_m) = \max. \tag{28}$$

Wartość modalna określona jest przez następujące warunki (dla rozkładów gęstości prawdopodobieństwa mających pierwszą i drugą pochodną):

$$\frac{d}{dx} f(x) = 0, \quad \frac{d^2}{dx^2} f(x) < 0. \tag{29}$$

Dany rozkład może bć jedno lub wielomodalny.

2.9.5 Mediana

Mediana $x_{0,5}$ rozkładu zdefiniowana jest jako wartość zmiennej losowej, dla której zachodzi

$$F(x_{0,5}) = P(X < x_{0,5}) = 0,5 \tag{30}$$

co w przypadku ciągłym daje się wyrazić jako:

$$\int_{-\infty}^{X_{0,5}} f(x)dx = 0,5. \tag{31}$$

2.9.6 Kwantyle

Kwantyle, x_q, dla $0 \leq q \leq 1$ definiuje się podobnie

$$F(x_q) = \int_{-\infty}^{X_q} f(x)dx = q. \tag{32}$$

Jeśli q jest wielokrotnością wartości 0,25 lub 0,1, to mówimy, odpowiednio, o kwadrylach lub decylach. Kwantyl $x_q(q)$ jest funkcją odwrotną do dystrybuanty.

2.10 Dwie zmienne losowe

2.10.1 Dystrybuanta

$$F(X,Y) = P(X < x, Y < y) \tag{33}$$

2.10.2 Łączna gęstość prawdopodobieństwa

$$f(x,y) = \frac{\partial}{\partial x}\frac{\partial}{\partial y}F(x,y) \tag{34}$$

Mając łączny rozkład prawdopodobieństwa można obliczyć prawdopodobieństwo zdarzenia

$$P(a \leq X < b, c \leq Y < d) = \int_a^b \left[\int_c^d f(x,y)dy\right] dx \tag{35}$$

Całkując po całym obszarze zmienności y otrzymujemy

$$P(a \leq X < b, c \leq Y < d) = \int_a^b \left[\int_{-\infty}^{\infty} f(x,y)dy\right] dx = \int_a^b g(x)dx \tag{36}$$

gdzie

$$g(x) = \int_{-\infty}^{\infty} f(x,y)dy \tag{37}$$

jest brzegową gęstością prawdopodobieństwa zmiennej X. Odpowiedni rozkład dla zmiennej Y dany jest przez:

$$h(y) = \int_{-\infty}^{\infty} f(x,y)dx \tag{38}$$

Posługując się funkcją łącznej gęstości można podać definicję:

zmienne X i Y są niezależne, jeśli

$$f(x,y) = g(x)h(y) \tag{39}$$

2.10.3 Wartość oczekiwana funkcji dwóch zmiennych losowych

$$E[H(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} H(x,y)f(x,y)dxdy \tag{40}$$

Wariancja

$$\sigma^2[H(X,Y)] = E[(H(X<Y) - E(H(X,Y)))^2] \tag{41}$$

W przypadku, gdy $H(X,Y) = aX + bY$

$$E(aX + bY) = aE(X) + bE(Y). \tag{42}$$

Jeśli $H(X,Y) = X^kY^l$, gdzie $k,l = 0,1,\ldots$ mamy momenty rzędu k i l względem X i Y:

$$\lambda_{kl} = E(X^kY^l). \tag{43}$$

W ogólniejszym przypadku, gdy $H(X,Y) = (X-a)^k(Y-b)^l$ mamy momenty rzędu k i l względem punktów a i b

$$\alpha_{kl} = E((X-a)^k(Y-b)^l). \tag{44}$$

Szczególne znaczenie mają momenty centralne (momenty względem $\lambda_{10}, \lambda_{01}$):

$$\mu_{kl} = E((X-\lambda_{10})^k(Y-\lambda_{01})^l). \tag{45}$$

Wartości momentów o szczególnym znaczeniu:

$$\begin{aligned}
&\mu_{00} = \lambda_{00} = 1,\\
&\mu_{10} = \mu_{01} = 0\\
\\
&\lambda_{10} = E(X) = \bar{x}\\
&\lambda_{01} = E(Y) = \bar{y}\\
\\
&\mu_{11} = \mathrm{cov}(X,Y)\\
&\mu_{20} = \sigma^2(X)\\
&\mu_{02} = \sigma^2(Y)
\end{aligned} \tag{46}$$

$$\begin{aligned}
\sigma^2(aX+bY) &= E(((aX+bY) - E(aX+bY))^2)\\
&= E((a(X-\bar{x}) + b(Y-\bar{y}))^2)\\
&= E(a^2(X-\bar{x})^2 + b^2(Y-\bar{y})^2 + 2ab(X-\bar{x})(Y-\bar{y}))
\end{aligned}$$

$$\sigma^2(aX+bY) = a^2\sigma^2(X) + b^2\sigma^2(Y) + 2ab\ \mathrm{cov}(X,Y) \tag{47}$$

3 Naiwny klasyfikator bayesowski

3.1 Teoria Bayesa

Teoria Bayes pokazuje, jak obliczyć prawdopodobieństwo *a posteriori*, $P(H|X)$. Prawdopodobieństwo to wyraża się równaniem:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \tag{48}$$

gdzie:

H - hipoteza (że X należy do klasy C);

X - przypadek (zestaw danych, fakty, obserwacje), które mogą wpłynąć na ocenę prawdopodobieństwa hipotezy;

$P(H)$ - prawdopodobieństwo *a priori*, że spełniona jest hipoteza H bez uwzględniania jakichkolwiek danych (np. $P(H)$ jest prawdopodobieństwem, że dowolny przypadek (zestaw danych) jest jabłkiem - nie zważając na kolor i kształt). $P(H)$, które nie zależy od X.

$P(H|X)$ - prawdopodobieństwo *a posteriori*, że hipoteza H jest prawdziwa dla zaobserwowanego przypadku X (np. prawdopodobieństwo, że owoc jest jabłkiem (H), jeśli znany jest warunek, że owoc ten jest okrągły i czerwony (X)).

$P(X|H)$ - wiarygodność hipotezy H ze względu na X (np. prawdopodobieństwo, że X jest czerwony i okrągły, gdy wiadomo, że prawdziwa jest hipoteza, że X jest jabłkiem).

$P(X)$ jest prawdopodobieństwem danych X (np. prawdopodobieństwo, że dany przypadek (zestaw danych) zarejestrowany wśród zbioru owoców jest czerwony i okrągły ang.*evidence*).

Jeśli X jest ciągłą zmienną losową, której dystrybucja zależy od H_i i jest wyrażona przez warunkową funkcję gęstości rozkładu prawdobodobieństwa $p(X|H_i)$, regułę Bayes'a można również przedstawić w postaci:

$$P(H_i|X) = \frac{p(X|H_i)P(H_i)}{p(X)} \tag{49}$$

gdzie $p(\cdot)$ to odpowiednie funkcje gęstości prawdopodobieństwa.

Dla dwóch hipotez dotyczących tego samego stanu rzeczy

$$p(x) = \sum_{i=1}^{2} p(X, H_i)P(H_i) \tag{50}$$

Jeśli dana jest obserwacja X oraz $P(H_1|X)$ jest większe od $P(H_2|X)$, to można byłoby wnioskować, że prawdziwą hipotezą jest H_1. Prawdopodobieństwo popełnienia błędu wyraża się wtedy przez:

$$P(error|x) = \begin{cases} P(H_1|X) & \text{jeśli zdecydowano, że prawdziwe jest} H_2 \\ P(H_2|X) & \text{jeśli zdecydowano, że prawdziwe jest} H_1. \end{cases} \tag{51}$$

Można dowieść, że reguła wyboru hipotezy:

$$\text{Wybierz } H_1 \text{ jeśli } P(H_1|X) > P(H_2|X);\ \text{w przeciwnym razie wybierz } H_2 \tag{52}$$

minimalizuje średnie prawdopodobieństwo popełnienia błędu $P(error)$

$$P(error) = \int_{\infty}^{\infty} P(error, X)dX = \int_{\infty}^{\infty} P(error|X)p(X)dX \tag{53}$$

Mianownik w równaniu 49, jak można zauważyć, jest współczynnikiem skalującym, zapewniającym $P(H_1|X) + P(H_2, X) = 1$. Można więc stosować alternatywną postać reguły decyzyjnej, w której ten współczynnik już nie występuje:

$$\text{Wybierz } H_1 \text{ jeśli } p(X|H_1) > P(X|H_2);\ \text{w przeciwnym razie wybierz } H_2 \tag{54}$$

3.2 Wielowymiarowe, ciagła przestrzeń cech

Niech $\omega_1, \ldots, \omega_k$ odpowiadają k stanom natury (hipotezom) oraz $\alpha_1, \ldots, \alpha_a$ będzie zbiorem możliwych akcji. W takim przypadku funkcja straty:

$$\lambda(\alpha_i|\omega_j) \tag{55}$$

opisuje stratę, jaką wywoła podjęcie akcji α_i, gdy stan natury jest ω_j.

Reguła Bayes'a dla przypadku wielowymiarowej, ciągłej przestrzeni cech i dla wielu stanów natury wyraża się wzorem:

$$P(\omega_j|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_j)P(\omega_j)}{p(\mathbf{x}} \tag{56}$$

gdzie

$$p(\mathbf{x}) = \sum_{j=1}^{k} p(\mathbf{x}|\omega_j)P(\omega_j). \tag{57}$$

Oczekiwana strata (ryzyko warunkowe) związane z podjęciem akcji α_i jest dana przez:

$$R(\alpha_i|\mathbf{x}) = \sum_{j=1}^{k} \lambda(\alpha_i|\omega_j)P(\omega_j|\mathbf{x}). \tag{58}$$

Oczekiwana strata związana z regułą decyzyjną jest dana wyraża się równaniem:

$$R = \int R(\alpha(\mathbf{x})|\mathbf{x})p(\mathbf{x})d\mathbf{x} \tag{59}$$

gdzie $\alpha(\mathbf{x})$ jest funkcją decyzyjną, która dla każdego $\mathbf{x}$ zakłada jedną z a wartości $\alpha_1, \ldots, \alpha_a$.

Reguła decyzyjna w takim przypadku polega na minimalizacji ryzyka całościowego, tj. na obliczeniu ryzyka warunkowego

$$R(\alpha_i|\mathbf{x}) = \sum_{j=1}^{k} \lambda(\alpha_i|\omega_j)P(\omega_j|\mathbf{x}). \tag{60}$$

dla $i = 1, \ldots, a$ i wybraniu akcji α_i, dla której $R(\alpha_i|\mathbf{x})$ osiąga minimum. Minimum takie nosi nazwę ryzyka Bayes'a. Oznacza się je przez R^* i jest to najlepszy wynik, jaki można osiągnąć.

3.3 Naiwny klasyfikator Bayesa

Naiwny klasyfikator Bayes'a jest metodą klasyfikacji powstała na bazie Teorii Bayesa. Zalicza się ją do grupy metod uczenia maszynowego. Określenie "naiwny" odnosi się do faktu, iż w zastosowanym modelu prawdopodobieństwa przyjęto pełną niezależność zmienny losowych, a to w rzeczywistości często mija się z prawdą.

W myśl tej teorii prawdopodobieństwo, że zaboserwowany przypadek (o cechach $X_1, \ldots, X_n$) należy do klasy C wyraża się wzorem:

$$P(C|X_1, \ldots, X_n) = \frac{P(C)P(X_1, \ldots, X_n|C)}{P(X_1, \ldots, X_n)} \tag{61}$$

Ponieważ mianownik w powyższym równaniu nie zależy od C, a ponadto dane są wartości cech X_i, przyjmuje się, że mianownik ten jest wartościa stałą. Licznik zaś przedstawić można jako:

$$\begin{aligned}
P(C, X_1, \ldots, X_n) &= P(C)P(X_1, \ldots, X_n|C) \\
&= P(C)P(X_1|C)P(X_2, \ldots, X_n|C, X_1) \\
&= P(C)P(X_1|C)P(X_2|C, X_1)P(X_3, \ldots, X_n|C, X_1, X_2) \\
&= P(C)P(X_1|C)P(X_2|C, X_1)P(X_3|C, X_1, X_2)P(X_4, \ldots, X_n|C, X_1, X_2, X_3) \\
&= \ldots
\end{aligned}$$

Z "naiwnego" założenia, że nie istnieje warunkowa zależność pomiędzy X_i oraz X_j $(i \neq j)$ mamy:

$$P(X_i|C, X_j) = P(X_i|C) \tag{62}$$

a więc

$$P(C, X_1, \ldots, X_n) = P(C)P(X_1|C)P(X_2|C)P(X_3|C)\ldots \tag{63}$$

Stąd

$$P(C|X_1 \ldots, X_n) = Z * P(C) \prod_{i=1}^{n} P(X_i|C) \tag{64}$$

gdzie Z - współczynnik skali zależny tylko od $X_1, \ldots, X_n$ (stała, jeżeli wartości cech są znane).

Naiwny klasyfikator Bayesa łączy przedstawiony model prawdopodobieństwa z regułą decyzyjną. Zazwyczaj reguła tą jest wybór tej hipotezy, która jest najbardziej prawdopodobna. Regułę tą nazywa się regułą MAP (*maximum a posteriori*). Odpowiadający jej klasyfikator jest funkcją $classify$ zdefiniowaną jak niżej:

$$classify(x_1, \ldots, x_n) = argmax_{c_k} P(C = c_k) \prod_{i=1}^{n} P(X_i = x_i|C = c_k) \tag{65}$$

gdzie $c_k, k = 1, \ldots, K$ to jedna z klas. W uczeniu z nauczycielem parametry modelu są estymowane. Z założenia o niezależności cech wystarczy estymować wartości prawdopodobieńst (class prior conditional feature model) używając metody *maximum likelihood, Bayesian inference* lub innych.

3.4 Przykład z klasyfikacją bayesowską *per pixel*

Problem polega tu na stwierdzeniu, do jakiej klasy należy punkt, gdy znane są odpowiadające mu poziomy szarości w kilku kanałach (tj. dany jest obraz wzorcowy z naniesionymi K klasami oraz obrazy odpowiadające wzorcowi w n kanałach spektralnych (256 poziomów szarości), szukany jest sposób klasyfikacji punktów innego obszaru, gdy znane są odpowiadające mu obrazy spektralne).

Na podstawie dostarczonych obrazów budowany jest naiwny klasyfikator Bayesa. Na początek obliczane jest prawdopodobieństwo $P(C = c_k)$. Aproksymowane jest ono licznością klasy k (ilością punktów, które na wzorcu zaklasyfikowano do tej klasy, wyrażoną przez N_k), podzieloną przez liczbę wszystkich punktów obrazu.

$$P(C = c_k) = \frac{N_k}{\sum_{k=1}^{n} N_k} \tag{66}$$

Następnie rozkładami normalnymi aproksymowane są prawdopodobieństwa warunkowe (prawdopodobieństwo, że poziom szarości punktu w kanale i wynosi X_i, gdy wiadomo, że punkt należy do klasy c_k):

$$P(X_i = x_i|C = c_k) \cong f(x_{k,i}) \tag{67}$$

Rozkłady normalne zdefiniowane są funkcją:

$$f(x_{k,i}) = \frac{1}{\sigma_{k,i}\sqrt{2\pi}} e^{\frac{-(x_{k,i}-\mu_{k,i})^2}{2\sigma_{k,i}}} \tag{68}$$

gdzie $x_{k,i}$ - poziom szarości punktu, $k = 1, \ldots, K$ - indeks odpowiadający klasie, $i = 1, \ldots, n$ - indeks odpowiadający kanałowi.

Odchylenie standardowe przypadku (będące aproksymacją $\sigma_{k,i}$) wynosi:

$$S_{N_k-1,i} = \sqrt{\frac{1}{N_k - 1}\sum_{j=1}^{N_k}(x_{k,i,j} - \bar{x}_{k,i})^2} \tag{69}$$

gdzie: $x_{k,i,j}$ - poziom szarości j - tego punktu z kanału i należącego (zgodnie z wzorcem) do klasy k.

Wartość średnia (będąca aproksymacją $\mu_{k,i}$) dana jest przez:

$$\bar{x}_{k,i} = \frac{1}{N_k}\sum_{j=1}^{N_k} x_{k,i,j} \tag{70}$$

4 Losowe pola Markowa

Klasyfikacja tekstur metodą losowych pól Markowa bazuje na modelu parametrycznym, w którym występują parametryzowane funkcje losowe. W ogólności klasyfikacja tekstur w tej metodzie polega na wyznaczeniu parametrów założonego modelu i dokonaniu klasyfikacji na ich podstawie.

4.1 Definicje podstawowe

Niech "pole widzenia" S będzie prostokątnym obszarem o rozmiarze $N_1 x N_2$ zdefiniowanym jak niżej:

$$S = \{(i,j) : 1 \leq i \leq N_1, 1 \leq j \leq N_2\} \tag{71}$$

Elementy należące $s \in S$ nazwiemy pikselami.
Niech "paleta" $\mathcal{L}$ będzie skończonym zbiorem, którego elementami są "kolory". W ogólności dla n kolorów można zapisać:

$$\mathcal{L} = \{1, \ldots, n\} \tag{72}$$

Funkcja $x : S \to X$ jest nazywana "obrazem". Wartość tej funkcji dla $s \in S$ oznaczymy przez x_s (pojedynczy indeks s jest uogólnieniem podwójnego indeksowania (i,j)).

Definicja 1 *System sąsiedztwa dla S zdefiniowany jest jako*

$$\sigma = \{\sigma_s \;\mid\; \forall s \in S\} \tag{73}$$

gdzie σ_s jest zbiorem pikseli sąsiadujących z pikselem s, zaś samo sąsiedztwo charakteryzuje się następującymi własnościami:

(1) piksel s nie sąsiaduje ze sobą samym: $s \notin \sigma_s$

(2) relacja sąsiedztwa jest wzajemna: $s \in \sigma_{s'} \Leftrightarrow s' \in \sigma_s$.

Piksele $s' \in \sigma_s$ nazywane są "sąsiadami" piksela s.

Dla regularnej kraty S zbiór sąsiadów piksela s zdefiniowany jest jako zbiór pikseli w promieniu r:

$$\sigma_s = \{s' \in S \;\mid\; [dist(s',s)]^2 \leq r, s' \neq s\} \tag{74}$$

gdzie $dist(s',s)$ oznacza odległość euklidesową pomiędzy pikselami s' i s. W przypadku dyskretnym (obrazów 2D) wyróżnia się otoczenia 1-go, 2-go, 3-go,... n-tego rzędu (zobacz Rys. 5).
Para (S,σ) tworzy graf, w którym węzłami są elementy S, zaś połączenia zdefiniowane są przez sąsiedztwo σ.

5	4	3	4	5
4	2	1	2	4
3	1		1	3
4	2	1	2	4
5	4	3	4	5

Rysunek 5: Otoczenia n-tego rzędu (pola oznaczone cyfrą 1 to otoczenie pierwszego rzędu, cyfrą 1 i 2 - to otoczenie drugiego rzędu; 1,2 i 3 - to otoczenie trzeciego rzędu, itd.).

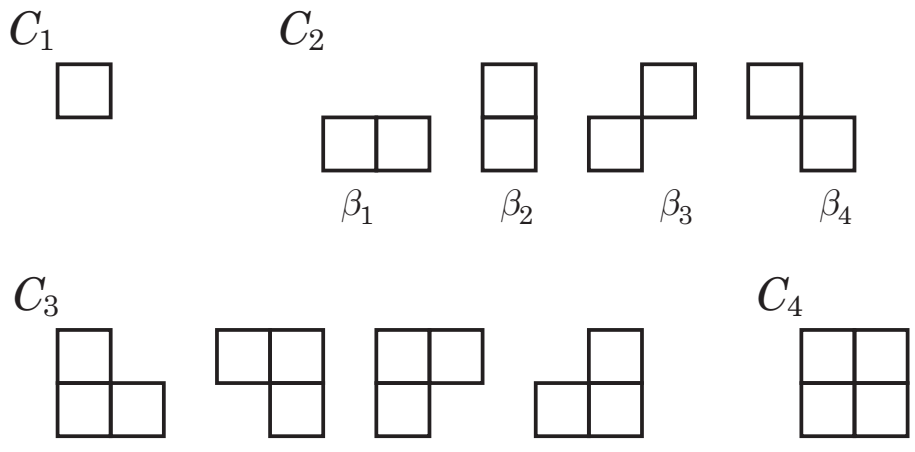
Definicja 2 *Klika c dla (S, σ) jest podzbiorem zbioru S. Klika zawierać może jeden piksel $c = \{s\}$, albo parę sąsiadujących pikseli $c = \{s, s'\}$, albo trójkę sąsiadujących pikseli $c = \{s, s', s''\}$, itd. Kolekcja klik zawierających pojedyncze piksele, pary oraz trójki oznaczona będzie, odpowiednio, przez*

$$C_1 = \{s \mid s \in S\} \tag{75}$$
$$C_2 = \{\{s, s'\} \mid s' \in \sigma_s, s \in S\} \tag{76}$$
$$C_3 = \{\{s, s', s'''\} \mid s, s', s'' \in S \text{ wszystkie są sąsiadami}\} \tag{77}$$

Kolekcja wszystkich klik oznaczana jest przez:

$$C = C_1 \bigcup C_2 \bigcup C_3 \cdots \tag{78}$$

Na Rys. 6 przedstawiono typy klik odpowiadających otoczeniu drugiego rzędu dla regularnej kraty.



Rysunek 6: Kliki dla otoczenia drugiego rzędu (z klikami należącymi do C_2 związano parametry β_k, $k = 1, 2, 3, 4$).

4.2 Pola losowe Markova i Gibbs'a

Niech $X = \{X_1, \ldots, X_m\}$ (w przypadku obrazu $m = N_1 \cdot N_2$) będzie rodziną zmiennych losowych zdefiniowanych na zbiorze S, w której każda zmienna X_s przyjmuje wartość $x_s \in \mathcal{L}$. Rodzina X nazywana jest polem losowym.
Niech zajście zdarzenia takiego, że X_s przyjęło wartość x_s oznaczone będzie przez $X_s = x_s$. Niech zajście łącznego zdarzenia oznaczone będzie przez $(X_1 = x_1, \ldots, X_m = x_m)$. Dla uproszczenia zdarzenie łączne oznaczać można przez $X = x$ gdzie $x = \{x_1, \ldots, x_m\}$ jest konfiguracją X odpowiadającą realizacji pola.
W przypadku dyskretnego zbioru X (ograniczona liczba kolorów) prawdopodobieństwo, że X_s przybierze wartość x_s oznaczone będzie przez $P(X_s = x_s)$ lub, w skrócie, przez $P(x_s)$. Łączne prawdopodobieństwo $P(X = x)$ oznaczone będzie przez $P(X_1 = x_1, \ldots, X_m = x_m)$ lub, w skrócie, przez $P(x)$ (w przypadku ciągłym $p(X_s = x_s)$ oraz $p(X = x)$ oznaczałyby odpowiednie funkcje gęstości prawdopodobieństwa).
Rozpatrując powyższe w terminach obróbki obrazów mamy:

> każdy obraz jest realizacją jakiegoś pola losowego. Każdy piksel z osobna jest realizacja zmiennej losowej, która przyjmuje wartości z przedziału zdefiniowanego paletą kolorów. $P(x_s)$ to prawdopodobieństwo wystąpienia piksela o kolorze x_s na pozycji s w obrazie.

Definicja 3 *X nazywane jest losowym polem Markowa na S względem systemu sąsiedztwa σ wtedy i tylko wtedy, gdy zachodzą dwa warunki:*

$$P(x) > 0, \forall x \in X \quad \textit{(positivity)} \tag{79}$$

$$P(x_s | x_{S-\{s\}}) = P(x_s \mid x_{\sigma_s}) \quad \textit{(Markovianity)} \tag{80}$$

gdzie $S - \{s\}$ jest to różnica zbiorów (wszystkie piksele poza pikselem s), $x_{S-\{s\}}$ to wartości (kolory) wszystkich pikseli poza pikselem s, oraz

$$x_{\sigma_s} = \{x_{s'} \mid s' \in \sigma_s\} \tag{81}$$

jest zbiorem zawierającym wartości (kolory) pikseli sąsiadujących z pikselem s.

Markovianity pokazuje lokalną charakterystykę pola X. Z własności tej wynika, że o kolorze danego piksela decyduje kolor jego sąsiadów. Z polem Markowa wiążą się jeszcze własności: homogeniczności i izotropiczności.

Definicja 4 *Zbiór zmiennych losowych X nazywany jest "polem losowym Gibbs'a" na S względem σ wtedy i tylko wtedy, gdy jego łączna dystrybucja ma postać:*

$$P(x) = \frac{1}{Z} e^{-\frac{1}{T}U(x)}, \tag{82}$$

gdzie

$$Z = \sum_{x \in X} e^{-\frac{1}{T}U(x)} \tag{83}$$

jest czynnikiem normalizującym (partition function), zapewniającym $\sum_{x \in X} p(x) = 1$; T jest stałą nazywaną "temperaturą" (można przyjąć $T = 1$);

$$U(x) = \sum_{c \in C} V_c(x) \tag{84}$$

jest "funkcją energii" będącą sumą potencjałów $V_c(x)$ wszystkich możliwch klik należących do C, przy czym wartość $V_c(x)$ zależy od lokalnej konfiguracji kliki c.

GRF jest polem homogenicznym jeśli $V_c(x)$ nie zależy od względnego położenia kliki c w S. Pole to jest izotropiczne, jeśli V_c nie zależy od orientacji kliki c.

Uwaga: MRF charakteryzowany jest przez swoją lokalną własność (the Markovianity) podczas gdy GRF charakteryzowany est przez swoją globalną własność (the Gibbs distribution). Teoria Hammersley-Clifford [Hammersley and Clifford 1971] ustanowiła równoważność pomiędzy tymi dwoma typami własności. Głosi ona, że X jest MRF na S względem σ wtedy i tylko wtedy, gdy X jest GRF na S względem σ.

Prawdopodobieństwo warunkowe $P(x_s \mid x_{\sigma_s})$ (wystąpienia piksela o kolorze x_s przy ustalonym jego otoczeniu) wyraża się wzorem:

$$P(x_s \mid x_{\sigma_s}) = \frac{e^{-\sum_{c\in\mathcal{A}} V_c(x_s)}}{\sum_{x'_s} e^{-\sum_{c\in\mathcal{A}} V_c(x'_s)}} \tag{85}$$

gdzie x'_s to wszystkie możliwe kombinacje jakie może przyjąć x_s, $\mathcal{A}$ to zbiór klik zawierających piksel s. Prawdopodobieństwo to zależy tylko od potencjału klik zawierających piksel s, a więc od kolorów pikseli z jego lokalnego otoczenia. Wynik ten dowodzi, że GRF jest MRF.

4.3 Model

Wybór postaci i parametrów funkcji potencjału jest podstawowym zadaniem w modelowaniu MRF. Formy funkcji potencjału jednoznacznie definiują postać dystrybucji Gibbs'a.

4.3.1 Auto-model

??????????????????????????????

$$U(x) = \sum_{s\in S} V_1(x_s) + \sum_{s\in S}\sum_{s'\in\sigma_s} V_2(x_s, x_{s'}) \tag{86}$$

auto-model

$$V_1(x_s) = x_s G_s(x_s), \quad V_2(x_s, x_{s'}) = \beta_{s,s'} x_s x_{s'} \tag{87}$$

$$U(x) = \sum_{\{s\}\in C_1} x_s G_s(x_s) + \sum_{\{s,s'\}\in C_2} \beta_{s,s'} x_s x_{s'} \tag{88}$$

gdzie $\beta_{s,s'}$ to stałe odzwierciedlające iterakcje między sąsiadami (punktami sąsiadującymi) s i s'

auto-logistic model Jeśli x_s przybiera wartości dyskretne ze zbioru $L = \{0, 1\}$ (lub $L = \{-1, +1\}$

$$U(x) = \sum_{\{s\}\in C_1} \alpha_s x_s + \sum_{\{s,s'\}\in C_2} \beta_{s,s'} x_s x_{s'} \tag{89}$$

gdzie $\beta_{s,s'}$ to współczynnik iterakcji.

Jeśli σ jest systemem najbliższego sąsiedztwa (4 najbliższych sąsiadów w 2D lub 2 sąsiadów w 1D) model ten redukuje się do **modelu Ising'a**. Prawdopodobieństwo warunkowe dla modelu auto-logistycznego z $L = \{0, 1\}$ dane jest przez:

$$P(x_s|x_{\sigma_s}) = \frac{e^{\alpha_s x_s} + \sum_{s'\in\sigma_s} \beta_{s,s'} x_s x_{s'}}{\sum_{x_s\in\{0,1\}} e^{\alpha_s x_s} + \sum_{s'\in\sigma_s} \beta_{s,s'} x_s x_{s'}} = \frac{e^{\alpha_s x_s} + \sum_{s'\in\sigma_s} \beta_{s,s'} x_s x_{s'}}{1 + e^{\alpha_s + \sum_{s'\in\sigma_s} \beta_{s,s'} x'_s}} \tag{90}$$

gdzie, dla dystrybucji homogenicznej, $\alpha_i = \alpha$, $\beta_{s,s'} = \beta$.

Auto-model nazywany jest modelem **auto-binomial** jeśli x_s przybiera wartości ze zbioru $\{0, 1, \ldots, M-1\}$ oraz każdy x_s ma conditional binomial distribution M prób z prawdopodobieństwem sukcesu q:

$$P(x_s|x_{\sigma_s}) = \left(\begin{array}{c} M-1 \\ x_s \end{array} \right) q^{x_s}(1-q)^{M-1-x_s} \tag{91}$$

gdzie

$$q = \frac{e^{\alpha_s + \sum_{s' \in \sigma_s} \beta_{s,s'} x_s x_{s'}}}{1 + e^{\alpha_s + \sum_{s' \in \sigma_s} \beta_{s,s'} x'_s}} \tag{92}$$

Odpowiednia funkcja energii ma postać

$$U(x) = -\sum_{\{s\} \in C_1} ln \left(\begin{array}{c} M-1 \\ x_s \end{array} \right) - \sum_{\{s\} \in C_1} \alpha_s x_s - \sum_{\{s,s'\} \in C_2} \beta_{s,s'} x_s x_{s'} \tag{93}$$

Dla $M = 1$ powyższy model redukuje się do modelu auto-logistycznego.

auto-normal model (Gaussian MRF) jeśli etykiety L mają ciągły charakter oraz dystrybucja łączna jest dystrybucją normalną wielu zmiennych. Wtedy funkcja rozkładu prawdopodobieństwa warunkowego ma postać:

$$p(x_s|x_{\sigma_s}) = \frac{1}{\sqrt{2\pi\underline{\sigma^2}}} e^{-\frac{1}{2\underline{\sigma^2}} \left(x_s - \mu_i - \sum_{s' \in \sigma_s} \beta_{s,s'} (x'_s = \mu_{s'}) \right)^2} \tag{94}$$

Wartość oczekiwana tej dystrybucji to

$$E(x_s|x_{\sigma_s}) = \mu_s - \sum_{s' \in \sigma_s} \beta_{s,s'} (x_{s'} - \mu_{s'}) \tag{95}$$

Wariancja zaś

$$\mathrm{var}(x_s|x_{\sigma_s}) = \underline{\sigma^2} \tag{96}$$

Wtedy łączne prawdopodobieństwo dystrybucji Gibbsa:

$$p(x) = \frac{\sqrt{det(B)}}{\sqrt{(2\pi\underline{\sigma^2})^m}} e^{\frac{(x-\mu)^T B (x-\mu)}{2\underline{\sigma^2}}} \tag{97}$$

gdzie x jest wektorem, μ jest wektorem $m \times 1$, $B = [b_{s,s'}]$ jest macierzą rozmiaru $m \times m$, której elementy są jednostkowe oraz elementy poza diagonalą o indeksie $(x_s, x_{s'})$ mają wartość $-\beta_{s,s'}$, tzn. $b_{s,s'} = \delta_{s,s'} - \beta s, s'$ z $\beta_{s,s'} = 0$. Stąd

$$V_1(x_s) = \frac{(x_s - \mu_s)^2}{2\underline{\sigma^2}} \tag{98}$$

$$V_2(x_s, x_{s'}) = \frac{\beta_{s,s'}(x_s = \mu_s)(x_{s'} - mu_{s'})}{2\underline{\sigma^2}} \tag{99}$$

Auto-regression model Danych jest m równań definiujących warunkowe gęstości prawdopodobieństwa.

$$x_s = x_s + \sum \beta_{s,s'} (f'_s - \mu_{s'}) + \varepsilon_s \tag{100}$$

gdzie $\varepsilon_s \sim N(0, \underline{\sigma^2})$

Łączna gęstość prawdopodobieństwa:

$$p(x) = \frac{\sqrt{det(B)}}{\sqrt{(2\pi\underline{\sigma^2})^m}} e^{\frac{(x-\mu)^T B^T B (x-\mu)}{2\underline{\sigma^2}}} \tag{101}$$

4.3.2 Multi-Level Logistic Model

W oryginalnym modelu potencjał kliki zawierającej przynajmniej dwa piksele wyraża się wzorem:

$$V_c(x) = \begin{cases} \zeta_c & \text{jeśli wszystkie piksele w klice } c \text{ mają ten sam kolor} \\ -\zeta_c & \text{w przeciwnym przypadku} \end{cases} \tag{102}$$

gdzie ζ_c jest potencjałem kliki typu c.

W naszym przypadku przyjęto założenie, że w modelu występują tylko potencjały klik o liczności 2 dla otoczenia rzędu drugiego, a potencjały innych klik równają się zeru. Niech więc z klikami należącymi do C_2 skojarzone będą parametry jak na Rys. 6. Potencjał kliki typu k należącej do C_2 określa teraz wzór:

$$V_2(x_s, x_{s'}) = \begin{cases} -\beta_k & \text{jeśli oba piksele w klice typu } k \text{ mają ten sam kolor} \\ \beta_k & \text{w przeciwnym przypadku} \end{cases} \tag{103}$$

gdzie β_k, $k = 1, 2, 3, 4$ to parametr związany z typem kliki.

Niech $\bar{\beta} = (\beta_1, \ldots, \beta_4)$ będzie wektorem parametrów. Potencjał kliki należącej do C_2 można zapisać jako $V_2(x_s, x_{s'}, \bar{\beta})$, podkreślając zależność od parametrów.

Ostatecznie prawdopodobieństwo warunkowe $P(x_s \mid x_{\sigma_s})$ dla tak przyjętego modelu wyraża się wzorem:

$$P(x_s \mid x_{\sigma_s}) = \frac{e^{-\sum_{s'} V_2(x_s, x_{s'})}}{\sum_{x'_s} e^{-\sum_{s'} V_2(x'_s, x_{s'})}} \tag{104}$$

gdzie $\sum_{s'}$ oznacza sumowanie po wszystkich sąsiadach piksela s tworzących z nim klikę należącą do C_2. Dla podwójnego indeksowania wzór ten można rozpisać następująco:

$$P(x_{i,j} \mid x_{\sigma_s}) = \frac{e^{-L(x_{i,j})}}{\sum_{x'_{i,j}} e^{-L(x'_{i,j})}} \tag{105}$$

gdzie

$$\begin{aligned} L(a_{i,j}) \quad = \quad & V_2(a_{i,j}, a_{i-1,j}) + V_2(a_{i,j}, a_{i+1,j}) + V_2(a_{i,j}, a_{i,j-1}) + V_2(a_{i,j}, a_{i,j+1}) + \\ & V_2(a_{i,j}, a_{i-1,j-1}) + V_2(a_{i,j}, a_{i+1,j+1}) + V_2(a_{i,j}, a_{i+1,j-1}) +_2 (a_{i,j}, a_{i-1,j+1}) \end{aligned} \tag{106}$$

Dla przykładu, jeśli we wszystkich klikach występują piksele jednakowego koloru, to L wyraża się przez:

$$L = 2\beta_1 + 2\beta_2 + 2\beta_3 + 2\beta_4 \tag{107}$$

jeśli zaś kolory pikseli w każdej klice są różne, to

$$L = -2\beta_1 - 2\beta_2 - 2\beta_3 - 2\beta_4 \tag{108}$$

4.4 Synteza tekstury

Generacja tekstury odpowiadającej realizacji dystrybucji Gibbsa $P(x) = \frac{1}{Z} e^{-U(x)}$ odbywa się przez próbkowanie dystrybucji. Często wykorzystywane w tym celu są dwa algorytmy: Metropolis sampler [Metropolis and et al 1953] and the Gibbs sampler [Geman and Geman 1984].

4.4.1 Metropolis sampler

Bazuje na metodzie opisanej w [Hammersley and Handscomb 1964].

Algorytm 1 Początek.
(1) Zainicjuj kolory pikseli danego obrazu x losowo wybranymi kolorami z palety $\mathcal{L}$;
(2) dla piksela $s \in S$ obrazu rób co następuje:
(2.1) niech y będzie kopią obrazu x;
(2.2) niech x_s będzie kolorem wylosowanym z palety $\mathcal{L}$;
(2.3) oblicz $p = min\ \{\ 1,\ P(y)/P(x)\ \}$, gdzie P jest daną dystrybucją Gibbsa (z przyjętymi dla syntezowanej tekstury parametrami);
(2.4) z prawdopodobieństwem p dokonaj przypisania $x = y$;
(3) powtórz (2) N razy.
Koniec.

Algorytm 1 jest algorytmem iteracyjnym, w którego iteracjach dla wybranego piksela losowany jest nowy kolor (z rozkładem równomiernym), przy czym akceptowany jest on z prawdopodobieństwem p.

4.4.2 Gibbs sampler

Bazuje na metodzie opisanej w [Geman and Geman 1984].

Algorytm 2 Początek.
(1) Zainicjuj kolory pikseli danego obrazu x losowo wybranymi kolorami z palety $\mathcal{L}$;
(2) dla piksela $s \in S$ obrazu rób co następuje:
(2.1) oblicz $p_l = P(x_s = l \mid x_{\sigma_s})$ dla wszystkich kolorów $l \in \mathcal{L}$ przy znanych kolorach sąsiadów x_{σ_s};
(2.2) z prawdopodobieństwem p_l dokonaj przypisania $x_s = l$
(3) powtórz (2) N razy.
Koniec.

Algorytm 2 jest algorytmem iteracyjnym, w którego iteracjach dla wybranego piksela proponowane są wszystkie możliwe kolory i akceptowane są one z prawdopodobieństwem obliczanym z dystrybucji Gibbsa w lokalnym otoczeniu.

4.5 Segmentacja tekstury

Podstawą do segmentacji były estymaty parametrów modelu obliczone dla kwadratowych obszarów. Do estymacji parametrów wykorzystano poniższe algorytmy [?]. W algorytmach tych korzysta się z pseudo-funkcji wiarygodności:

$$PL(x|\bar{\beta}) = \ln\left(\prod_{s\in S} P(x_s|x_{\sigma_s},\bar{\beta})\right) \tag{109}$$

Algorytm 3.2

(0) Inicjalizacja. Niech $t = 0$. Wybierz początkowy zestaw parametrów $\bar{\beta}_0$. Określ wartość η oraz maksymalną ilość iteracji N_1.

(1) Jeśli $t > N_1$ to koniec, w przeciwnym wypadku wykonuj co następuje:

(2) Wylosuj $\nu = \text{rand}(0,1)$.

Jeśli $\nu \leq \eta$, to

a) oblicz gradient funkcji wiarygodności dla $\bar{\beta} = \bar{\beta}_t$:

$$dPL = \left. \frac{\partial PL(x|\bar{\beta})}{\partial \bar{\beta}} \right|_{\bar{\beta}=\bar{\beta}_t} \tag{110}$$

b) wyznacz wartość $\bar{\beta}'$ na pozytywnym kierunku tego gradientu:

$$\bar{\beta}' = \bar{\beta}_t + \frac{dPL}{||dPL||} \cdot |\text{normal}(0,1)| \tag{111}$$

gdzie normal(μ, σ) to generator liczb losowych o rozkładzie normalnym ze średnią μ i wariancją σ;

c) oblicz $\alpha(\bar{\beta}', \bar{\beta}_t)$:

$$\alpha(\bar{\beta}', \bar{\beta}_t) = \min \left\{ 1, e^{PL(x|\bar{\beta}') - PL(x|\bar{\beta}_t)} \frac{1-\eta}{2\eta(\sqrt{2\pi}^3)} \right\} \tag{112}$$

w przeciwnym wypadku

a) niech $\bar{\beta}'$ będzie wartością otrzymaną z generatora liczb losowych o rozkładzie normalnym:

$$\bar{\beta}' = \text{normal}(\bar{\beta}_t, I_4) \tag{113}$$

gdzie I_4 to macierz jednostkowa.

b) oblicz $\alpha(\bar{\beta}', \bar{\beta}_t)$:

$$\alpha(\bar{\beta}', \bar{\beta}_t) = \min \left\{ 1, e^{PL(x|\bar{\beta}') - PL(x|\bar{\beta}_t)} \right\} \tag{114}$$

(3) Jeśli $\text{rand}(0,1) < \alpha(\bar{\beta}', \bar{\beta}_t)$, to $\bar{\beta}_{t+1} = \bar{\beta}'$; w przeciwnym $\bar{\beta}_{t+1} = \bar{\beta}_t$.

(4) $t = t + 1$, skocz do (1).

Algorytm 3.3

(0) Inicjalizacja. Niech $t = 0$. Wybierz początkowy zestaw parametrów $\bar{\beta}_0$. Określ temperaturę początkową T_0, maksymalną ilość iteracji N_2 oraz wyznacz wartość γ z równania $T_0 \gamma^{N_2} = 1$.

(1) Niech $\bar{\beta}'$ będzie wartością otrzymaną z generatora liczb losowych o rozkładzie normalnym:

$$\bar{\beta}' = \text{normal}(\bar{\beta}_t, I_4) \tag{115}$$

gdzie I_4 to macierz jednostkowa.

(2) Oblicz $\alpha(\bar{\beta}', \bar{\beta}_t)$:

$$\alpha(\bar{\beta}', \bar{\beta}_t) = \min \left\{ 1, e^{\frac{PL(x|\bar{\beta}') - PL(x|\bar{\beta}_t)}{T_t}} \right\} \tag{116}$$

(3) Jeśli $\text{rand}(0,1) < \alpha(\bar{\beta}', \bar{\beta}_t)$, to $\bar{\beta}_{t+1} = \bar{\beta}'$; w przeciwnym razie $\bar{\beta}_{t+1} = \bar{\beta}_t$.

(4) $T_{t+1} = \gamma\, T_t$, $t = t + 1$, skocz do (1).

Algorytm 3.4
(0) Inicjalizacja. Wybierz początkowy zestaw parametrów $\bar{\beta}_0$.
(1) Uruchom Algorytm 3.3 z wybranymi parametrami $\bar{\beta}_0$.
(2) Zapisz wynik z Algorytmu 3.3. jako początkowy zestaw parametrów $\bar{\beta}_0$ dla Algorytmu 3.2.
(3) Uruchom Algorytm 3.2.
(4) Powtórz M razy ($M \leq N_1$) kroki (0), (1), (2), (3) . Zapisz wyniki z kroku 3 jako $\tilde{\beta}_i$, $i = 1, \ldots, M$. Estymatę parametrów oblicz ze wzoru

$$\tilde{E}(\bar{\beta}|x) = \frac{1}{M} \sum_{i=1}^{M} \tilde{\beta}_i \tag{117}$$

4.6 Szczegóły implementacji

W naszej implementacji do generacji sztucznych tekstur wykorzystano Algorytm 2. Krok (2) zaimplementowano wyznaczając dystrubuantę rozkładu Gibbsa, obliczając następnie jej odwrotność dla wartości wylosowanej z przedziału $(0, 1)$ z równomiernym rozkładem prawdopodobieństwa. Algorytm uruchamiano 5 razy, przy czym każdy uruchomienie wiązało się z wyznaczeniem koloru dla wszystkich pikseli obrazu.

Krok (2) Algorytmu 2
(1) oblicz $Z = \sum_{l \in \mathcal{L}} U(l)$, gdzie $l = 1, \ldots, n$
(2) oblicz $Q_k = \sum_{l=1}^{k} U(l)/Z$, gdzie $k = 1, \ldots, n$
(3) wylosuj p = rand $(0, 1)$
(4) for($k = 1$; $k < n$; $k++$)
        if($p < Q_k$) break;
(5) nowy kolor równa się k

5 Metoda momentów chromatycznych

Klasyfikacja obszarów metodą obliczania momentów chromatycznych pierwotnie służyła do segmentacji kolorowych zdjęć dostarczonych w formacie RGB. W metodzie tej dla zadanego obszaru obliczane były momenty chromatyczne, które następnie interpretowane były jako wektor cech analizowanych w procesie klasyfikacji. Metoda ta została przystosowana do segmentacji obrazów satelitarnych.

Obrazy satelitarne rozpatrywane w opisywanym eksperymencie zawierały obszary o różnym typie pokrycia powierzchni. Chociaż obrazy niosły informacje w czterech różnych kanałach spektralnych, w opisywanym eksperymencie skorzystano z danych pochodzących tylko z trzech pierwszych kanałów. Kanały te zinterpretowano jako kolejne składowe RGB obrazu kolorowego.

5.1 Podstawy matematyczne

Niech $I(i, j)$ oznacza punkt (piksel) obrazu I o współrzędnych i, j, przy czym $0 \leq i \leq L_x - 1$ oraz $0 \leq j \leq L_y - 1$, gdzie L_x, L_y - szerokość i wysokość obrazu.

5.2 Konwersja RGB do CIE XYZ

Istnieje kilka definicji przestrzeni kolorów CIE XYZ. W eksperymencie przyjęto, że konwersji przestrzeni kolorów RGB do CIE XYZ dokonuje się wg. wzoru:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 0.607 & 0.174 & 0.200 \\ 0.299 & 0.587 & 0.114 \\ 0.000 & 0.066 & 1.111 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{118}$$

5.3 Chromatyczność

Chromatyczność, pomijając fizyczną interpretację, definiują następujące wzory:

$$x = \frac{X}{X+Y+X},\ y = \frac{Y}{X+Y+Z},\ z = \frac{Z}{X+Y+Z} \tag{119}$$

Mając x, y można obliczyć z jako uzupełnienie do 1: $z = 1 - x - y$. Tak więc z jest wartością nadmiarową. Co więcej, wszystkie współrzędne mieszczą się w przedziale od 0 do 1: $x, y, z \in [0, 1]$.

5.4 Dystrybucja chromatyczności oraz jej ślad

Dystrybucję chromatyczności definiuje poniższy wzór:

$$D(x, y) = K, \tag{120}$$

gdzie K to ilość pikseli, którym odpowiada chromatyczność (x,y). Inaczej zapisując:

$$D(x, y) = \text{card}\{(i, j) \mid I(i, j) \rightarrow (x, y)\} \tag{121}$$

Ślad dystrybucji jest rzutem $D(x, y)$ na płaszczyznę x, y. Wyraża się on przez:

$$T(x, y) = \begin{cases} 1, & \text{jeśli istnieje piksel, któremu odpowiada chromatyczność}(x, y) \\ 0, & \text{w przeciwnym wypadku.} \end{cases} \tag{122}$$

Inaczej zapisując:

$$T(x, y) = \begin{cases} 1, & \exists(i, j) : I(i, j) \rightarrow (x, y) \\ 0, & \neg\exists(i, j) : I(i, j) \rightarrow (x, y) \end{cases} \tag{123}$$

Obliczanie momentów chromatycznych (patrz następny podrozdział) wymaga zmiany skali i dyskretyzacji. Niech $X_s, Y_s \in \aleph$ będą współczynnikami skalującymi. Niech $x_s = Int(X_s x)$, $y_s = Int(Y_s y)$, gdzie $x_s, y_s \in \aleph$ oraz Int jest funkcją zwracającą część całkowitą swojego argumentu. Przeskalowana (dyskretna) dystrybucja i jej ślad mają teraz postać:

$$D_s(x_s, y_s) = \text{card}\{(i, j) | I(i, j) \underset{X_s, Y_s}{\rightarrow} (x_s, y_s)\} \tag{124}$$

$$T_s(x_s, y_s) = \begin{cases} 1, & \exists(i, j) : I(i, j) \underset{X_s, Y_s}{\rightarrow} (x_s, y_s) \\ 0, & \neg\exists(i, j) : I(i, j) \underset{X_s, Y_s}{\rightarrow} (x_s, y_s) \end{cases} \tag{125}$$

5.5 Momenty chromatyczne

Przy danej dystrybucji chromatyczności i danym śladzie dystrybucji można obliczyć momenty chromatyczne M_D (momenty dystrybucji) i M_T (momenty śladu dystrybucji). Momenty te definiują wzory:

$$M_D(k,l) = \sum_{x_s} \sum_{y_s} x_s^k y_s^l T_s(x_s, y_s) \tag{126}$$

$$M_T(k,l) = \sum_{x_s} \sum_{y_s} x_s^k y_s^l D_s(x_s, y_s) \tag{127}$$

6 Ocena wyników klasyfikacji

6.1 Pierwiastek błędu średniokwadratowego ($RMSE$)

Pierwiastek błędu średniokwadratowego ($RMSE$) [?]:

$$RMSE_k = \sqrt{\sum_{l=1}^{L} \left(\hat{N}_{kl} - N_{kl}\right)^2 \times \frac{m_l}{M}} \tag{128}$$

gdzie $\hat{N}_{kl}$ i N_{kl} to, odpowiednio, liczność punktów przypisanych do klasy k wg klasyfikatora oraz liczność punktów przypisanych do klasy k wg wzorca, przy czym klasyfikacja dotyczy obszaru l (obszar l może być zwartą powierzchnią na obrazku lub też zbiorem punktów wybranych losowo); m_l jest rozmiarem obszaru l wyrażającym się ilością zawartych w obszarze punktów; M jest całkowitym rozmiarem danych, $M = \sum_{l=1}^{L} m_l$ (gdzie L - liczba wyróżnionych obszarów).

Uwaga: miarę powyższą można użyć do oceny wyników klasyfikacji dla jednego obrazka (wtedy wybiera się w nim obszary do wyliczenia $RMSE$) lub też do oceny wyników klasyfikacji dla wielu obrazków (wtedy każdy z nich jest osobnym obszarem).

6.2 Progowany błędem procent poprawnej klasyfikacji (TCE)

Progowany błędem procent poprawnej klasyfikacji (TCE) [?]

$$TCE_\epsilon = \frac{1}{K} \sum_{k=1}^{K} \sum_{l=1}^{L} n_{kl} \times \frac{m_l}{M} \tag{129}$$

gdzie K to liczba klas, zaś dla danego błędu ϵ

$$n_{kl} = \begin{cases} 1 & \text{jeśli} \quad |\hat{N}_{kl} - N_{kl}| \leq \epsilon \\ 0 & \text{w przypadku przeciwnym} \end{cases} \tag{130}$$

6.3 Współczynniki obliczane na podstawie macierzy niezgodności

Niech A będzie macierzą zawierającą wyniki klasyfikacji:

$$A = [a_{ij}] \tag{131}$$

gdzie a_{ij} jest liczbą punktów należących do klasy j, które zostały zaklasyfikowane jako punkty należące do klasy i. W przypadku oceny wyników klasyfikacji dla jednej tylko klasy (punkty nie należące do klasy "tworzą" drugą klasę) tabela ta może mieć postać:

		wg wzorca należy do klasy	
		tak	nie
wg klasyfikatora należy do klasy	tak	a_{11}	a_{12}
	nie	a_{21}	a_{22}

W przypadku wielu klas tabela ta może mieć postać:

		wg wzorca należy do klasy			
		1	2	...	K
wg klasyfikatora należy do klasy	1	a_{11}	a_{12}	...	a_{1K}
	2	a_{21}	a_{22}	...	a_{2K}
	...	...	...	...	...
	K	a_{K1}	a_{K2}	...	a_{KK}

Na podstawie macierzy A obliczać można następujące współczynniki,[?, ?]:

- dokładność użytkownika klasy i:

$$u_i = a_{ii}/a_{ri}, \tag{132}$$

gdzie $a_{ri} = \sum_i a_{i.}$ (suma elementów w wierszu i);

– dokładność producenta klasy i:

$$p_i = a_{ii}/a_{ci}, \tag{133}$$

gdzie $a_{ci} = \sum_i a_{i.}$ (suma elementów w kolumnie is);

- całkowita dokładność klasyfikacji d:

$$d = \sum_i a_{ii}/a_t, \tag{134}$$

gdzie $a_t = \sum_i a_{ci} = \sum_i a_{ri}$ (liczba wszystkich punktów);

- prosty współczynnik Kappa:

$$\hat{\kappa} = \frac{P_o - P_e}{1 - P_e} \ , \ \text{gdzie} \ \ P_o = \frac{\sum_i a_{ii}}{a_t} \ \ \text{oraz} \ \ P_e = \frac{\sum_i a_{ri} a_{ci}}{a_t^2} \tag{135}$$

- ważony współczynnik Kappa:

$$\hat{\kappa}_w = \frac{P_{o(w)} - P_{e(w)}}{1 - P_{e(w)}} \ , \ \text{gdzie} \ \ P_{o(w)} = \frac{1}{a_t} \sum_i \sum_j w_{ij} a_{ij} \ \ \text{oraz} \ \ P_{e(w)} = \frac{1}{a_t^2} \sum_i \sum_j w_{ij} a_{ri} a_{cj} \tag{136}$$

Wagi w_{ij} spełniać powinny warunek: $0 <= w_{ij} < 1$. Zazwyczaj przyjmuje się, że $w_{ii} = 1$ (czyli najważniejsza jest zgodność dotycząca tych samych klas), natomiast dla $i \neq j$ przyjmuje się $w_{ij} = w_{ji} = 1 - \frac{|i-j|}{K-1}$. Takie postać współczynników wagowych jest słuszna jedynie wtedy, gdy klasy uporządkowane są w tabeli pod względem ważności. A ważność danej klasy można ocenić na podstawie jej liczności.

Ponieważ wartość ważonego współczynnika Kappa zależy od doboru wag, nie ma prostej interpretacji otrzymywanych wartości. Zgodnie z jedną z propozycji, [?]:

$\kappa > 0.75$	mocna zgodność
$0.4 < \kappa \leq 0.75$	dobra zgodność
$\kappa \leq 0.4$	słaba zgodność

Zgodnie z inna propozycją, [?]:

$$\begin{array}{ll} 0.81 \leq \kappa \leq 1 & \text{almost perfect} \\ 0.61 \leq \kappa \leq 0.80 & \text{substancial} \\ 0.41 \leq \kappa \leq 0.60 & \text{moderate} \\ 0.21 \leq \kappa \leq 0.40 & \text{fair} \\ 0.00 \leq \kappa \leq 0.20 & \text{slight} \\ \kappa < 0.0 & \text{poor} \end{array}$$

- (nie)dokładność Adasia, DA:

$$DA = \frac{a_{12} + a_{21}}{a_t} \tag{137}$$

Miara ta wyraża stosunek sumy źle zaklasyfikowanych punktów dla każdej z klas do całkowitej ich ilości. $0 \leq uDA \leq 1$. Można zauważyć, że DA to „przeciwieństwo" d dla macierzy o rozmiarze 2×2. Dla większej macierzy można obliczyć *uogolnioną (nie)dokładność Adasia* uDA:

$$uDA = \frac{\sum_i (a_{ri} - a_{ii})}{a_t} = 1 - d \tag{138}$$

- (nie)dokładość Witolda, DW:

$$\begin{aligned} err1 &= \begin{cases} \frac{a_{12}}{a_{11}+a_{12}} & \text{gdy } a_{11} + a_{12} > 0 \\ 0 & \text{w przypadku przeciwnym} \end{cases} \\ err2 &= \begin{cases} \frac{a_{21}}{a_{21}+a_{22}} & \text{gdy } a_{21} + a_{22} > 0 \\ 0 & \text{w przypadku przeciwnym} \end{cases} \\ DW &= \tfrac{err1+err2}{2} \end{aligned} \tag{139}$$

Można zauważyć, że DW to „przeciwieństwo" u_i dla macierzy o rozmiarze 2×2. Dla większej macierzy można obliczyć *uogolnioną (nie)dokładność Witolda* uDW:

$$\begin{aligned} err_{ij} &= \begin{cases} \frac{a_{ij}}{a_{ri}} & \text{gdy } a_{ri} > 0 \text{ oraz } i \neq j \\ 0 & \text{w przypadku przeciwnym} \end{cases} \\ err_i = \textstyle\sum_j err_{ij} &= \begin{cases} 1 - \frac{a_{ii}}{a_{ri}} = 1 - u_i & \text{gdy } a_{ri} > 0 \\ 0 & \text{w przypadku przeciwnym} \end{cases} \\ uDW &= \tfrac{1}{K} \textstyle\sum_i err_i \end{aligned} \tag{140}$$

6.4 Modele Markova

6.4.1 Model Markova k-tego rzędu

Załóżmy, że nastrój (humor) Kierownika można opisać jedną z trzech wartości: wesoły (☺), obojętny (😐) oraz smutny (☹) i że nie zmienia się on w ciągu dnia. Czy można przewidzieć humor kierownika w nadchodzącym dniu, jeśli znana jest historia jego nastrojów do dnia bieżącego?

Niech q_n wyraża humor w dniu n, $q_n \in \{☺, 😐, ☹\}$. Wtedy zadanie predykcji polega na określeniu prawdopodobieństwa

$$P(q_n | q_{n-1}, q_{n-2}, \ldots, q_1) \tag{141}$$

		q_{i-1}		
		☺	😐	☹
q_i	☺	0,7	0,14	0,16
	😐	0,2	0,5	0,3
	☹	0,1	0,36	0,54

Tablica 1: Model Markowa pierwszego rzędu

Prawdopodobieństwo to pozwala przewidzieć humor dla konkretnego, jednego dnia, przy znanych nastrojach z dni poprzednich.

Przykład: jeśli przez trzy ostatnie dni humor Kierownika tworzył sekwencję {☺,☺,😐} w chronologicznej kolejności, to prawdopodobieństwo, że Kierownik będzie smutny w dniu bieżącym wyraża się przez:

$$P(q_4 = ☹ | q_3 = 😐, q_2 = ☺, q_1 = ☺) \tag{142}$$

Prawdopodobieństwo to można byłoby wyznaczyć na drodze wyznaczenia częstości występowania sekwencji {☺,☺,😐,☹} w całej historii zaobserwowanych humorów Kierownika. W podejściu tym jednak tkwi pewien problem. Czy obserwacja sekwencji czterech dni jest wystarczająca? Jak długi powinien być okres obserwacji, tj. okres, w którym wyznacza się częstość wystąpień danej sekwencji?

Jeśli długość sekwencji nastrojów wynosi n, to liczba wszystkich możliwych sekwencji kończących się taką samą wartością wynosi 3^{n-1}. Stąd aby określić prawdobodobieństwo wystąpienia każdej z nich (tj. prawdopodobieństwo sekwencji kończącej się humorem ☺ albo 😐 albo ☹) należałoby zebrać 3^{n-1} statystyk.

Założenie Markova k-tego rzędu: prawdopodobieństwo q_n zależy tylko $q_{n-1}, \ldots, q_{n-k}$.

$$P(q_n | q_{n-1}, q_{n-2}, \ldots, q_1) = P(q_n | q_n - 1, \ldots, q_{n-k}) \tag{143}$$

W szczególności, dla $k = 1$ mamy mamy założenie Markova pierwszego rzędu. System, dla którego założenie piewszego rzędu jest prawdziwe nazywa się modelem Markova pierwszego rzędu, zaś sekwencje przez niego generowane $\{q_i\}$ nazywa się łańcuchem Markowa pierwszego rzędu.

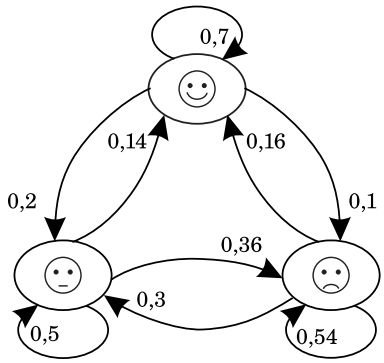
Prawdopodobieństwo wystąpienia sekwencji nastrojów $\{q_1, q_2, \ldots, q_n\}$ (czyli prawdopodobieństwo łączne wystąpienia jakichś humorów w dniu bieżącym i w dniach poprzednich), przy założeniu pierwszego rzędu, wyraża się przez:

$$P(q_1, \ldots, q_n) = \prod_{i=1}^{n} P(q_i | q_{i-1}) \tag{144}$$

Niech model Markowa pierwszego rzędu określa Tablica 1. Model ten można przedstawić również jako maszynę stanów skończonych (rysunek 7, w której wyróżnione są trzy stany, $S = \{☺, 😐, ☹\}$, zaś przejścia pomiędzy stanami określa tabela 1.

Można teraz zapytać się, jaki będzie prawdopodobieństwo, że po dniu bycia wesołym Kierownik będzie smutny przez dwa kolejne dni? Odpowiedź można obliczyć jak następuje:

$$\begin{aligned} P(q_3 = ☹, q_2 = ☹ | q_1 = ☺) &= P(q_3 = ☹ | q_2 = ☹, q_1 = ☺) \cdot P(q_2 = ☹ | q_1 = ☺) = \\ &= P(q_3 = ☹ | q_2 = ☹) \cdot P(q_2 = ☹ | q_1 = ☺) = \\ &= 0,54 \cdot 0,1 = 0,054 \end{aligned} \tag{145}$$



Rysunek 7: Maszyna stanów

humor Kierownika	prawdopodobieństwo 📢
☺	0,1
😐	0,2
☹	0,7

Tablica 2: Prawdopodobienstwo, że kierownik głośno mówi w zależności od jego nastroju

6.4.2 Ukryte modele Markova (*Hidden Markov Models - HMMs*)

Dotychczas humor Kierownika można było rozpoznać po jego twarzy. Przypuśćmy jednak, że kierownik w zależności od swojego humoru mówi głośniej (📢) lub ciszej (🔇). Jeśli słychać krzyki, to najprawdopodobniej jest w złym humorze, jeśli krzyków nie słychać - pewnie jest w dobrym humorze. Prawdopodobienstwo, że kierownik głośno mówi w zależności od jego nastroju przedstawia tabela 2.

Prawdopodobieństwo humoru można jedynie określić na podstawie parametru x_i, który przyjmuje wartości $\{📢, 🔇\}$.

$$P(q_i|x_i) = \frac{P(x_i|q_i)P(q_i)}{P(x_i)} \tag{146}$$

lub dla n dni, gdy sekwencja nastrojów $Q = \{q_1, \ldots, q_n\}$ i sekwencji głośności $X = \{x_1, \ldots, x_n\}$

$$P(q_1, \ldots, q_n|x_1, \ldots, x_n) = \frac{P(x_1, \ldots, x_n|q_1, \ldots, q_n)P(q_1, \ldots, q_n)}{P(x_1, \ldots, x_n)} \tag{147}$$

W powyższym równaniu $P(q_1, \ldots, q_n)$ to prawdopodobieństwo wystąpienia sekwencji nastrojów, $P(x_1, \ldots, x_n)$ to prawdopodobieństwo zaobserwowania sekwencji głośności.

?????????????? Przy założeniu, że x_i są niezależne od x_j oraz q_j dla wszystkich $j \neq i$:

$$P(x_1, \ldots, x_n|q_1, \ldots, q_n) = \prod_{i=1}^{n} P(x_i|q_i) \tag{148}$$

Naszym celem jest określenie nastroju Kierownika na podstawie zaobserwowanej jego głośności. Ponieważ prawdopodobieństwo $P(x_1, \ldots, x_n)$ jest niezależne od humorów (obserwacje głośności są ponadto znane!), można to prawdopodobieństwo pominąć. Miara prawdopodobieństwa, jaką otrzymamy, nazywa się wiarygodnością, L:

$$P(q_1, \ldots, q_n|x_1, \ldots, x_n) \propto L(q_1, \ldots, q_n|x_1, \ldots, x_n) = P(q_1, \ldots, q_n|x_1, \ldots, x_n)P(q_1, \ldots, q_n) \tag{149}$$

Z założenia pierwszego rzędu modelu Markova wynika:

$$P(q_1,\ldots,q_n|x_1,\ldots,x_n) \propto L(q_1,\ldots,q_n|x_1,\ldots,x_n) = \prod_{i=1}^{n} P(x_i|q_i) \prod_{i=1}^{n} P(q_i|q_n) \tag{150}$$

Przykład: Załóżmy, że pewnego dnia Kierownik był wesoły i zamknął się w gabinecie. W dniu następny słychać było zza drzwi jego podniesiony głos. Czy można zgadnąć, w jakim był wtedy humorze? Zobaczmy więc, jak przedstawia się funkcja wiarygodności, że Kierownik był w dobrym humorze drugiego dnia:

$$L(q_2 = \text{☺}|q_1 = \text{☺}, x_2 = \text{🔊}) = P(x_2 = \text{🔊}|q_2 = \text{☺})P(q_2 = \text{☺}|q_1 = \text{☺}) = ??????? \tag{151}$$

Funkcja wiarygodności, że Kierownik był w złym humorze w tym przypadku wyraża się przez:

$$L(q_2 = \text{☹}|q_1 = \text{☺}, x_2 = \text{🔊}) = P(x_2 = \text{🔊}|q_2 = \text{☺})P(q_2 = \text{☺}|q_1 = \text{☺}) = ??????? \tag{152}$$

zaświarygodność, że Kierownik był obojętny:

$$L(q_2 = \text{😐}|q_1 = happy, x_2 = \text{🔊}) = P(x_2 = \text{🔊}|q_2 = \text{☺})P(q_2 = \text{☺}|q_1 = \text{☺}) = ??????? \tag{153}$$

Przykład: Przypuśćmy, że przez trzy dni nie widziano Kierownika, choć było jednak słychać zza jego drzwi w tym czasie tylko przytłumione dźwięki. Spróbujmy określić wiarygodność faktu, że przez te trzy dni sekwencja humorów Kierownika była następująca: $q_1 = \text{☺}, q_2 = \text{😐}, q_3 = \text{☺}$. Ponieważ nic nie wiemy o humorze Kierownika w pierwszym dniu, traktujemy, że każda z trzech możliwości jest jednakowo prawdopodobna: *prior probability* $P(q_1 = \text{☺}) = p(q_1 = \text{☺}) = \frac{1}{3}$. Stąd:

$$\begin{aligned} &L(q_1 = \text{☺}, q_2 = \text{😐}, q_3 = \text{☺}|x_1 = \text{⊗}, x_2 = \text{⊗}, x_3 = \text{⊗}) = \\ &\quad P(x_1 = \text{⊗}|q_1 = \text{☺})P(x_2 = \text{⊗}|q_2 = \text{😐})P(x_3 = \text{⊗}|q_3 = \text{☺}) \\ &\qquad P(q_1 = \text{☺})P(q_2 = \text{😐}|q_1 = \text{☺})P(q_3 = \text{☺}|q_2 = \text{😐}) = \\ &= ???????????????????????????????? \end{aligned} \tag{154}$$

6.4.3 Terminologia

HMM jest określony przez:

- zbiór stanów $S = \{s_1, s_2, \ldots, s_{N_s}\}$,

oraz zbiór parametrów $B = \{\Theta, A, B\}$, gdzie:

- π jest wektorem, którego elementami są prawdopodobieństwa „a priori", $\pi_i = P(q_1 = s_i)$, określające prawdopodobieństwa pierwszych stanów s_i w sekwencji stanów. W przypadku braku wiedzy o początkowych stanach często nadaje się im równe wartości, tzn. $\pi_i = \frac{1}{N_s}$.
- A jest macierzą o elementach $a_{i,j} = P(q_{n+1} = s_j|q_n = s_i)$. Elementy te to prawdopodobieństwa przejść ze stanu i do stanu j.
- B określa prawdopodobieństwa emisji (*the emission probabilities*) mówiące o wiarygodność pewnej obserwacji x, jeśli model jest w stanie s_i. W zależności od rodzaju obserwacji, mamy:
  - dla przypadku dyskretnych obserwacji, tzn. gdy $x_n \in \{v_1, \ldots, v_K\}$, B jest macierzą, której elementami są prawdopodobieństwa zaobserwowania v_k, jeśli stanem bieżącym jest $q_n = s_i$. Wartość tych prawdopodobieństw wyraża się przez $b_{i,k} = P(x_n = v_k|q_n = s_i)$.

– dla przypadku ciągłego, gdy $x_n \in R^D$, $B(x)$ jest wektorem, którego elementami są funkcje gęstości prawdopodobieństwa nad przestrzenią obserwacji dla systemu będącego w stanie s_i, $b_i(x_n)$. Funkcje te wyraża się przez $b_i(x_n) = p(x_n|q_n = si)$. Są one często parametryzowane (jak *funkcje Gaussa*).

Działanie HMM charakteryzuje:

- sekwencja stanów ukrytych $Q = \{q_1, q_2, \ldots, q_N\}$, $q_n \in S$,
- sekwencja obserwacji $X = \{x_1, x_2, \ldots, x_N\}$.

HMM pozwalający na przejście z dowolnego emitującego stanu do każdego innego emitującego stanu nazywany jest *ergodic HMM*. W skrajnym przypadku, gdy w HMM istnieją tylko pojedyncze przejścia ze stanu do stanu (w tym przejścia do tego samego stanu) taki model nazywa się it left-right HMM.

6.4.4 Użyteczne formuły

- prawdopodobieństwo sekwencji stanów $Q = \{q_1, q_2, \ldots, q_N\}$ dla HMM z parametrami Θ odpowiada iloczynowi prawdopodobieństw przejść pomiędzy kolejnymi stanami:

$$P(Q|\Theta) = \pi_{q_1} \cdot \prod_{n=1}^{N-1} a_{q_n,q_{n+1}} = \pi_{q_1} \cdot a_{q_1,q_2} \cdot a_{q_2,q_3} \cdot \ldots \cdot a_{q_{N-1},q_N} \tag{155}$$

- wiarygodność sekwencji obserwacji $X = \{x_1, x_2, \ldots, x_N\}$ dla danej sekwencji stanów $Q = \{q_1, q_2, \ldots, q_N\}$ o tej samej długości (wiarygodność sekwencji obserwacji X wzdłuż pojedynczej ścieżki Q) dla HMM z parametrami B dana jest przez

$$P(X|Q, \Theta) = \prod_{n=1}^{N} P(x_n|q_n, \Theta) = b_{q_1,x_1} \cdot b_{q_2,x_2} \cdot \ldots \cdot b_{q_N,x_N} \tag{156}$$

  tzn. jest iloczynem prawdopodobieństw emisji obliczonych wzdłuż rozważanej ścieżki.

- Łączna wiarygodność sekwencji obserwacji X i ścieżki Q, dana jest przez regułę Bayesa:

$$P(X, Q|\Theta) = P(X|Q, \Theta) \cdot P(Q|\Theta) \tag{157}$$

- Wiarygodność sekwencji obserwacji $X = \{x_1, x_2, \ldots, x_N\}$ względem HMM z parametrami Θ dana jest przez:

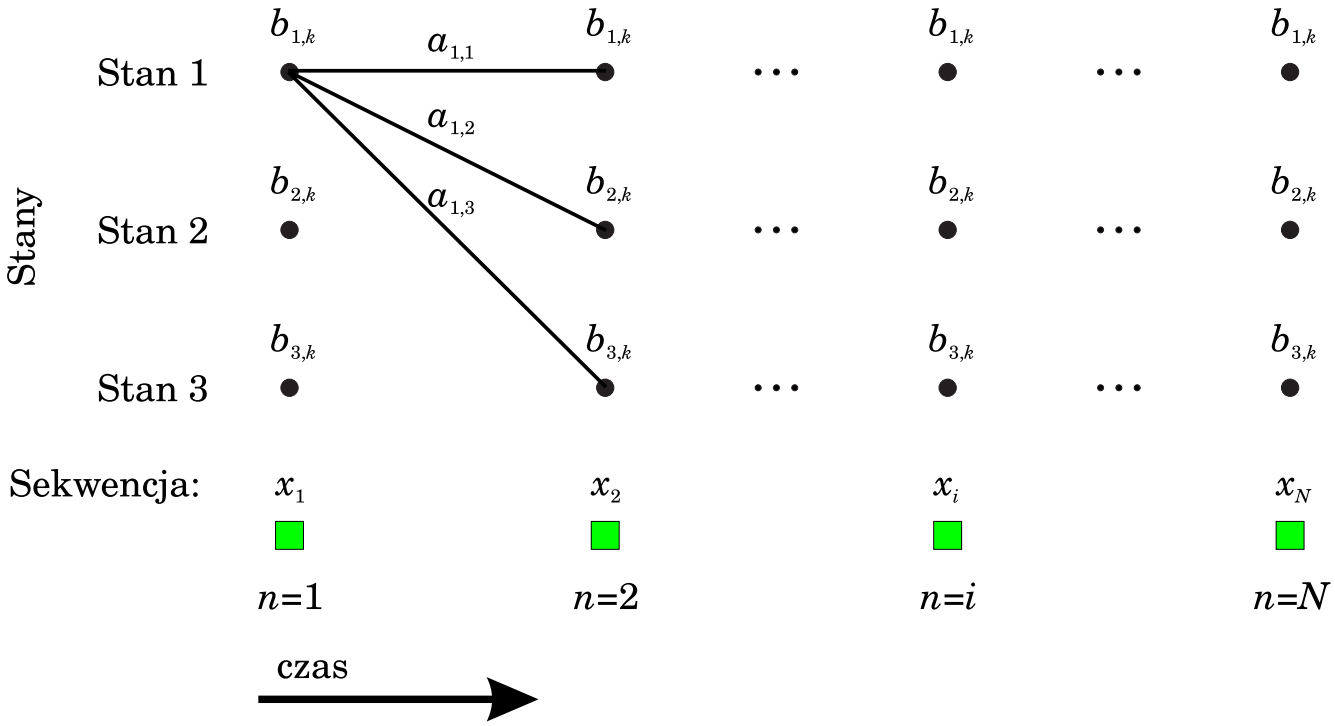
$$P(X|\Theta) = \sum_{allQ} P(X, Q|\Theta) \tag{158}$$

  tzn. jest sumą przez wszystkie możliwe sekwencje występujące w modelu.

6.4.5 Diagram kratowy (*trellis diagram*)

Diagram kratowy służy do wizualizacji obliczanych wiarygodności dla HMM. Na rysunku 8 pokazany jest diagram dla HMM z trzema stanami.

Każda kolumna w diagramie pokazuje możliwe stany w pewnej chwili czasu n. Każdy stan w kolumnie jest połączony ze stanem w dołączonej kolumnie przejściem o wiarygodności danej przez $a_{i,j}$ macierzy przejść A (jak Tablica 1). Na dole diagramu umieszczona jest sekwencja obserwacji $X = \{x_1, \ldots, x_N\}$. Wartości $b_{i,k}$ to wiarygodności obserwacji $x_n = v_k$ w stanie $q_n = s_i$ w chwili n.



Rysunek 8: Trellis diagram

6.4.6 Viterbi algorithm

7 Neuronowa implementacja metody pól sztucznego potencjału

Planowanie bezkolizyjnej ścieżki dla punktu p w trójwymiarowej, statycznej (znanej) przestrzeni roboczej E_0 polega na znalezieniu najkrótszego, bezkolizyjnego przejścia pomiędzy położeniem początkowym tego punktu, p_S, a położeniem docelowym, p_F. Dzięki wprowadzeniu pól sztucznego potencjału, zadanie to sprowadzić można do jednego z dwóch następujących zadań minimalizacji: a) zadania znalezienia ścieżki leżącej wzdłuż największego spadku energii pola; b) zadanie znalezienia ścieżki, minimalizującej energię związaną z jej długością i karą za zbliżanie się do przeszkód. Standardowo, pole sztucznego potencjału w zadaniu a) definiuje się jako funkcję $U : E_0 \rightarrow \mathbb{R}^+$, która każdemu punktowi z przestrzeni E_0 przypisuje wartość energii potencjalnej, odpowiadającej jego położeniu. Funkcja ta stanowi sumę dwóch funkcji podstawowych, mianowicie potencjału przyciągającego, U_a, i potencjału odpychającego, U_r,

$$U(p) = U_a(p) + U_r(p). \tag{159}$$

Potencjał przyciągający definiuje się tak, aby w miejscu punktu decelowego utworzyła się „energetyczna dolina". Definicja natomiast potencjału odpychającego zapewnić ma utworzenie się w obszarze zajmowanym przez przeszkody „energetycznych wzgórz". Posługując się taką fizyczną interpretacją, poszukiwana ścieżka będzie drogą, jaką zakreśli tocząca się wśród tych wzgórz bezwymiarowa kulka o masie jednostkowej. Drogę kulki wyznacza się przez rozwiązanie równania ruchu (poprzez np. całkowanie metodą Runge-Kutta):

$$\ddot{p} = F = F_a + F_r, \tag{160}$$

gdzie $F_a = -\frac{\partial U_a}{\partial p}$, $F_r = -\frac{\partial U_r}{\partial p}$ - odpowiednio siła przyciągająca i siła odpychająca; lub przez zastosowanie algorytmu gradientowego:

$$\dot{p} = -\alpha \frac{\partial U}{\partial p}. \tag{161}$$

Zazwyczaj potencjał przyciągający ma postać kwadratową:

$$U_a(p) = \frac{1}{2}\beta d_f^2(p) \tag{162}$$

lub jest złożeniem funkcji kwadratowej i liniowej:

$$U_a(p) = \begin{cases} \frac{1}{2}\beta d_f^2(p), & \text{gdy} \quad d_f < s \\ 2\beta s d_f(p) - \beta s^2, & \text{gdy} \quad d_f \leq s \end{cases}, \tag{163}$$

gdzie $\beta > 0$ - waga, s - parametr przełączania, $d_f^2(p) = (p - p_f)^T(p - p_f)$, $d_f(p) = (p - p_f)$.

Większą różnorodnością definicji charakteryzuje się potencjał odpychający. Dla prostej sceny wyraża się go jako sumę następujących potencjałów, [**?**, **?**]:

$$U_{ri}(p) = \begin{cases} \frac{1}{2}\beta(\frac{1}{d_{oi}(p)} - \frac{1}{\hat{d}_o}), & \text{gdy} \quad d_{oi}(p) < \hat{d}_o \\ 0, & \text{gdy} \quad d_{oi}(p) < \hat{d}_o \end{cases}, \tag{164}$$

gdzie $d_{oi}(p)$ - odległość punktu p od i-tej przeszkody, $\hat{d}_o$ - margines bezpieczeństwa. Wraz ze wzrostem ilości przeszkód w scenie rośnie niebezpieczeństwo pojawienia się w polu potencjałów minimów lokalnych. Aby zaradzić tej sytuacji wprowadzony został następujący potencjał odpychający, [**?**]:

$$U_{ri}(K(p)) = \begin{cases} \frac{A}{K_i(p)} \exp(-\alpha K_i(p)), & \text{dla} \quad K_i(p) > 1 \\ A \exp(-\alpha K_i^{1+\frac{1}{\alpha}}(p)), & \text{dla} \quad 0 \leq K_i(p) < 1 \end{cases}, \tag{165}$$

gdzie α - parametr odpowiedzialny za szybkość wzrostu potencjału, A - parametr skalujący, $K(p)$ - pseudoodległość punktu p od przeszkody i. Inną próbę skonstruowania pola potencjałów bez minimów lokalnych podjął Sato, [**?**], proponując do tego celu użycie pól potencjałów Laplace'a.

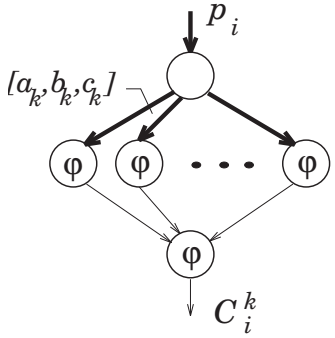
Jak można zauważyć, najbardziej kosztowną operacją w planowaniu bezkolizyjnej ścieżki za pomocą równań (160) i (161) jest, występujące przy obliczaniu gradientu potencjału odpychającego, obliczanie odległości. Aby obniżyć koszty planowania, wystarczy w miejscu tym zastosować neuronowo implementowalne algorytmy z podrozdziału **??** (zysk osiąga się przez wprowadzenie równoległych obliczeń). Efektywniejszym jednak sposobem wykorzystania wyników z podrozdziału **??** jest ich zastosowanie do rozwiązania zadania b).

Dla przypomnienia, zadanie b) polega na znalezieniu ścieżki pomiędzy punktem startowym a punktem docelowym, o minimalnej energii, związanej z jej długością i karą za zbliżanie się do przeszkód (statycznych), [**?**]. Zadaniu temu można nadać następującą fizyczną interpretację: szukana bezkolizyjna ścieżka jest kształtem, jaki przyjmie rozciągnięta w polu potencjałów elastyczna nitka, uwiązana końcami w punkcie startowym i docelowym.

Energię elastycznej nitki zdefiniować można w następujący sposób. Jeśli na nitce wyróżnionych zostanie $\mathcal{L}+1$ kolejnych punktów p_i (p_0 i $p_\mathcal{L}$ to odpowiednio, punkt startowy i punkt docelowy), wtedy energia sprężystości nitki [4] wyrażać się będzie równaniem:

$$E_{sp} = \frac{1}{2}\sum_{i=1}^{\mathcal{L}} L_i^2 = \frac{1}{2}\sum_{i=1}^{\mathcal{L}} (p_i - p_{i-1})^T(p_i - p_{i-1}). \tag{166}$$

[4] W rzeczywistości jest to energia sprężystości linii łamanej, składającej się z odcinków o długości L_i.



Rysunek 9: Sieć neuronowa obliczająca C_i^k.

Energia związana z karą za zbliżanie się do przeszkód dana zaś będzie wzorem:

$$E_{ka} = \sum_{i=1}^{\mathcal{L}} \sum_{k=1}^{\mathcal{K}} C_i^k, \tag{167}$$

gdzie $\mathcal{K}$ - liczba przeszkód, C_i^k - kara za zbliżanie się punktu p_i (należącego do nitki) do przeszkody k. Natomiast energia całkowita nitki, E, będzie następującą ważoną sumą energii sprężystości i energii związanej z karą za zbliżanie się do przeszkód:

$$E = w_{sp} E_{sp} + w_{ka} E_{ka}, \tag{168}$$

gdzie w_{sp}, w_{ka} są współczynnikami wagowymi.

Występującą we wzorze 167 karę C_i^k obliczyć można, jak w zadaniu a), zgodnie z formułami na potencjał odpychający, (164), (165). Aby jednak ominąć problem obliczania odległości, karę C_i^k otrzymać można bezpośrednio z przedstawionej na rysunku 9 sieci neuronowej. Sieć ta różni się od sieci przedstawionej na rysunku **??** tym, że: 1) neurony warstwy ukrytej i wyjściowej są tego samego typu, o sigmoidalnej funkcji aktywacji $\varphi(\cdot)$; 2) wszystkie wagi i wartości progowe (*bias'y*) warstwy ukrytej mają zmieniony znak (równania ścian są tak określone, że zwracają wartości dodatnie dla punktów należących do wnętrza danego obiektu); 3) wartość progowa (*bias*) neuronu warstwy wyjściowej równa jest liczbie neuronów warstwy ukrytej (liczbie ścian przeszkody k), M^k, pomniejszonej o 1/2. Analityczną postać funkcji kary przedstawia wzór:

$$C_i^k = \varphi(\sum_{j=1}^{M^k} \varphi(r_j^k) + M^k - 1/2), \tag{169}$$

gdzie $r_j^k(p_i) = A_j^k p_{ix} + B_j^k p_{iy} + C_j^k p_{iz} + D_j^k$ otrzymuje się z równania j-tej ściany k-tej przeszkody.

Mając zdefiniowaną energię całkowitą, zadanie znalezienia bezkolizyjnej ścieżki można utożsamić z zadaniem znalezienia takiego położenia punktów p_i, które tą energię minimalizuje. Do tego celu posłużyć się można, podobnie jak miało to miejsce przy liczeniu odwrotnej kinematyki, następującym algorytmem gradientowym:

$$\dot{p}_i = -\eta \left[\frac{\partial E}{\partial p_i} \right] \quad , \quad \eta > 0, i = 1, \ldots, \mathcal{L} - 1 \tag{170}$$

z warunkiem stopu:

$$|\dot{p}_i| < \epsilon \quad , \quad \epsilon \ll 1. \tag{171}$$

Zbieżność tego algorytmu wykazać można, sprawdzając znak $\dot{E}$ wzdłuż trajektorii (170). Ponieważ punkty p_0 i p_N są punktami statycznymi, pochodna energii w czasie ma postać:

$$\dot{E} = \sum_{i=1}^{\mathcal{L}-1} \left[\frac{\partial E}{\partial p_i}\right]^T \dot{p}_i = \sum_{i=1}^{\mathcal{L}-1} [w_{sp}\frac{\partial E_{sp}}{\partial p_i} + w_{ka}\frac{\partial E_{ka}}{\partial p_i}]^T \dot{p}_i, \tag{172}$$

gdzie

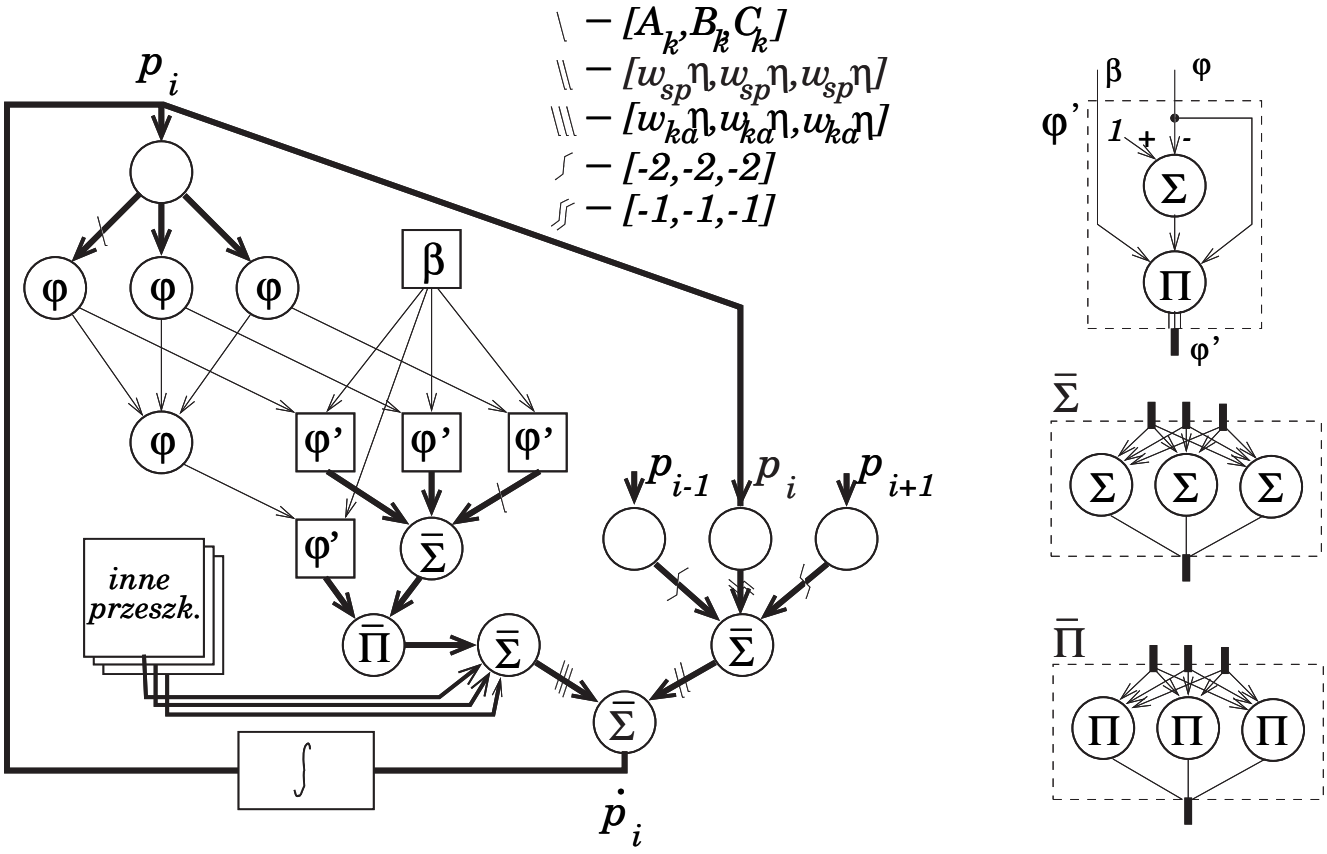
$$\begin{aligned}
\frac{\partial E_{sp}}{\partial p_i} &= 2p_i - p_{i-1} - p_{i+1}, \\
\frac{\partial E_{ka}}{\partial p_i} &= \sum_{k=1}^{\mathcal{K}} \frac{\partial C_i^k}{\partial p_i}, \\
\frac{\partial C_i^k}{\partial p_i} &= \varphi'(\sum_{j=1}^{M^k} \varphi(r_j^k(p_i)) + M^k - 1/2) \cdot (\sum_{j=1}^{M^k} \varphi'(r_j^k(p_i))) \cdot \frac{\partial r_j^k}{\partial p_i}, \\
\frac{\partial r_j^k}{\partial p_i} &= (A_j^k, B_j^k, C_j^k)^T, \\
\varphi'(\cdot) &= -\beta\varphi(\cdot)(1 - \varphi(\cdot)).
\end{aligned}$$

Jeśli $\dot{p}_i$ wybrane zostanie zgodnie z równaniem (170), to pochodna energii w czasie wzdłuż trajektorii tego systemu będzie ujemnie określona

$$\dot{E} = -\eta \sum_{i=1}^{\mathcal{L}-1} \dot{p}_i^T \dot{p}_i \leq 0 \tag{173}$$

wszędzie, z wyjątkiem punktów równowagi, gdzie $\dot{E} = 0$ ($\dot{E} = 0 \Leftrightarrow \dot{p}_i = 0, \forall i$). Fakt ten dowodzi, że zastosowanie algorytmu (170) zapewnia minimalizację energii E.

Aby ominąć problem minimów lokalnych (tzn. aby z algorytmu (170) uzyskać rzeczywiście najkrótszą, bezkolizyjną ścieżkę), obliczanie kolejnych położeń punktów p_i odbywać się musi równolegle z procesem wyżarzania (*annealing process*), [**?**]. Wyżarzanie to polega na zwiększaniu w miarę upływu czasu parametru β, występującego w wyrażeniu na funkcję sigmoidalną. Przy zmianie β (np. zgodnie z zależnością: $\beta = \beta_0 \log(1+t)$ lub $\beta = \beta_0(1+t)$, gdzie: β_0 - wartość początkowa, t - czas) zmienia się zasięg oddziaływania funkcji kary C_i^k. Im większe jest β, tym mniejszy jest

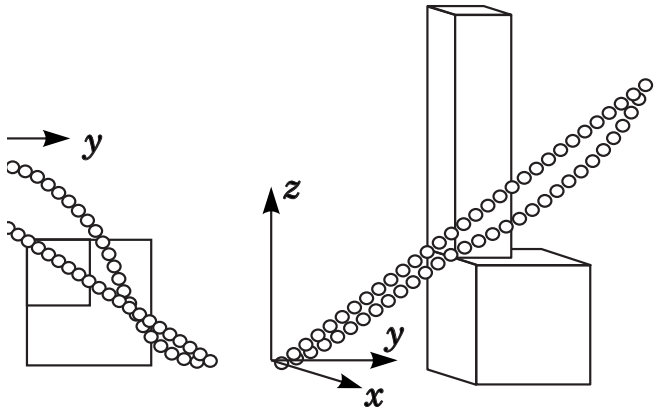


Rysunek 10: Neuronowa implementacja algorytmu $\dot{p}_i = -\eta \left[\frac{\partial E}{\partial p_i}\right]$.

zasięg oddziaływania. Małe β w chwili początkowej powoduje, że wygładzone kary C_i^k nakładają się na siebie, co w efekcie przynosi likwidację lokalnych minimów E. Wraz ze wzrostem β kary „wyostrzają się", co sprawia, że „nitka zaczyna rozciągać się w kierunku minimum globalnego". Na rysunku 10 przedstawiono schemat sieci neuronowej, implementującej algorytm (170).

Przykład: „Neuronowość" procesu planowania ścieżki wynika z faktu przyjęcia neuronowej reprezentacji sceny. Model ten różni się od modelu przyjętego przy obliczaniu odległości odwrotnym skierowanie wektorów normalnych ścian oraz nieco zmienionymi parametrami sieci (inne typy neuronów i inne ich wartości progowe (*bias'y*)). Natomiast sama zasada funkcjonowania użytego w procesie planowania ścieżki algorytmu jest podobna do zasady funkcjonowania algorytmu obliczania odległości. Rozwiązanie bowiem otrzymuje się poprzez minimalizację zdefiniowanej wcześniej funkcji kryterialnej. W przypadku planowania ścieżki funkcją kryterialną jest funkcja definiująca energię „elastycznej nitki" (zobacz równanie (168)), zaś minimalizujący ją algorytm[5] dany jest równaniem (170).

[5] Jest to algorytm gradientowy.



Rysunek 11: Planowania bezkolizyjnej ścieżki metodą „elastycznej nitki".

Na rysunku 11 pokazany jest wynik poszukiwań bezkolizyjnej ścieżki w trójwymiarowej przestrzeni. Widoczne na nim punkty stanowią ślad, jaki zostawiła rozciągana w polu potencjałów nitka. W położeniu początkowym nitka ma kształt prostoliniowego odcinka, przecinającego obecną na scenie przeszkodę. Po uruchomieniu algorytmu (z parametrem β zmienianym zgodnie z zależnością $\beta = \beta_0(1 + t_0)$) nitka zaczęła rozciągać się, wysuwając się poza obszar zajmowany przez przeszkodę. Ostateczny kształt, jaki przyjęła nitka jest kształtem poszukiwanej ścieżki.