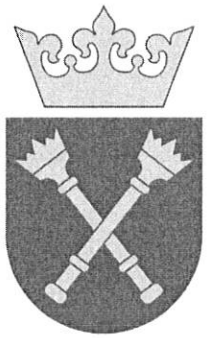
Uniwersytet Jagielloński

Wydział Matematyki i Informatyki



STRESZCZENIE ROZPRAWY DOKTORSKIEJ

Entropia w kodowaniu i klastrowaniu danych

mgr Marek Śmieja

Streszczenie rozprawy doktorskiej
napisanej pod kierunkiem dra hab. Jacka Tabora

Kraków, 2014

Niniejsza rozprawa doktorska ma na celu zaprezentowanie nowej teorii dotyczącej entropii, a także przedstawienie zastosowań entropii w problemach kodowania i klastrowania danych (analizy skupień). Na rozprawę składają się następujące artykuły naukowe:

[I] *Marek Śmieja, Jacek Tabor, „Entropy of the mixture of sources and entropy dimension", IEEE Transactions on Information Theory, 58/5, pp. 2719-2728, 2012 (impact factor: 2,621, punkty MNiSW: 45),*

[II] *Marek Śmieja, „Weighted approach to general entropy function", IMA Journal of Mathematical Control and Information, doi: 10.1093/imamci/dnt044, pp. 13, 2014 (impact factor: 0,741, punkty MNiSW: 20),*

[III] *Marek Śmieja, Jacek Tabor, „Rényi entropy dimension of the mixture of measures", przyjęte do Proceedings of Science and Information Conference SAI 2014, pp. 5, 2014 (praca konferencyjna),*

[IV] *Marek Śmieja, Dawid Warszycki, Jacek Tabor, Andrzej Bojarski, „Asymmetric Clustering Index in a case study of 5-HT_{1A} receptor ligands", przyjęte do PLoS ONE, pp. 13, 2014 (impact factor: 3,73, punkty MNiSW: 40),*

[V] *Marek Śmieja, Jacek Tabor, „Image segmentation with use of cross-entropy clustering", Proceedings of the 8th International Conference on Computer Recognition Systems CORES 2013, pp. 403-409, 2013 (praca konferencyjna).*

Pierwsze trzy prace stanowią teoretyczną część rozprawy, natomiast dwie ostatnie część poświęconą zastosowaniom. W każdej z tych dwóch części prace uporządkowane są względem ich ważności.

Część teoretyczna rozprawy dotyczy pojęcia entropii opisującego statystyczną ilość pamięci potrzebnej na zapis dowolnego elementu danych z zadanym maksymalnym błędem. Najważniejszym wynikiem tej części rozprawy jest opracowana nowa definicja entropii ważonej [I, Definition II.4], równoważna klasycznej [I, Theorem II.1], której zastosowanie pozwoliło na oszacowanie entropii i wymiaru entropijnego kombinacji miar [I, Theorem III.1, Theorem IV.1, Theorem IV.3], [II, Theorem 4.1], [III, Theorem III.1].

Poniżej postaram się omówić kolejne prace tej części rozprawy. Celem pracy [I] było wykazanie oszacowań na wymiar entropijny kombinacji miar. Narzędziem użytym do tego celu stała się opracowana definicja entropii ważonej, będąca pojęciem pokrewnym do ważonych miar Hausdorffa, które wykorzystał między innymi J. D. Howroyd do obliczania wymiaru Hausdorffa iloczynu kartezjańskiego:

$$\dim_H(X) + \dim_H(Y) \leq \dim_H(X \times Y).$$

Zachodzenie powyższej nierówności przez wiele lat było ważnym problemem otwartym.

Zanim przejdziemy do opisu najważniejszych rezultatów pracy [I] konieczne jest wprowadzenie podstawowych definicji oraz notacji. Podział $\mathcal{P}$ przestrzeni probabilistycznej (X, Σ, μ) na parami rozłączne zbiory zadaje stratne kodowanie deterministyczne: punkt $x \in X$ zapisujemy kodem stowarzyszonym z jedynym elementem $P \in \mathcal{P}$ spełniającym $x \in P$. Ustalając mierzalne pokrycie $\mathcal{Q}$ przestrzeni X rozważamy wszystkie możliwe podziały $\mathcal{P}$ wpisane w $\mathcal{Q}$, to znaczy spełniające warunek: dla każdego $P \in \mathcal{P}$ istnieje $Q \in \mathcal{Q}$ takie, że $P \subset Q$. Operując językiem kodowania, pokrycie $\mathcal{Q}$ definiuje maksymalny błąd jaki możemy popełnić w kodowaniu stratnym – podział $\mathcal{P}$ musi być akceptowalny przez to pokrycie (mówimy że $\mathcal{P}$ jest $\mathcal{Q}$-akceptowalny, co zapisujemy $\mathcal{P} \prec \mathcal{Q}$).

Powyższe rozumowanie prowadzi do dwóch podstawowych wersji definicji entropii Shannona: dla podziału $\mathcal{P}$ entropia Shannona jest zdefiniowana jako:

$$h(\mu; \mathcal{P}) = -\sum_{P \in \mathcal{P}} \mu(P) \log \mu(P),$$

natomiast dla pokrycia $\mathcal{Q}$ mamy:

$$H(\mu; \mathcal{Q}) := \inf\{h(\mu; \mathcal{P}) \, : \, \mathcal{P} \text{ jest przeliczalnym podziałem } X \text{ takim, że } \mathcal{P} \prec \mathcal{Q}\}.$$

Wielkość $H(\mu; \mathcal{Q})$ nawiązuje do entropii rozważanej przez A. Rényiego w definicji wymiaru entropijnego oraz ε-entropii opracowanej przez E. C. Posnera, z tym że jest określona na ogólnej przestrzeni probabilistycznej (niekoniecznie metrycznej).

Proponowane ważone podejście, zamiast podziału $\mathcal{P}$, dopasowuje odpowiedni ciąg miar $(\nu_Q)_{Q \in \mathcal{Q}}$ do ustalonego pokrycia $\mathcal{Q}$. W szczególności wymagane jest, aby z każdym elementem $Q \in \mathcal{Q}$ stowarzyszona była miara ν_Q spełniająca

$$\nu_Q(X \setminus Q) = 0. \tag{1}$$

Ponadto, suma tych miar powinna odzwierciedlać miarę wyjściową, to znaczy

$$\sum_{Q \in \mathcal{Q}} \nu_Q = \mu. \tag{2}$$

Nawiązując do poruszonego problemu kompresji, z każdym elementem $Q \in \mathcal{Q}$ możemy związać ustalony kod. Miara ν_Q określa prawdopodobieństwo z jakim element x zostanie zakodowany kodem stowarzyszonym z Q. W odróżnieniu od kodowania deterministycznego tym razem dopuszczamy losowość w sposobie kodowania – element może zostać zakodowany kodem zadanym przez dowolne Q, a miara ν_Q definiuje prawdopodobieństwo wystąpienia tego kodu.

Ważoną entropię Shannona dla ciągu miar $(\nu_Q)_{Q\in\mathcal{Q}}$ oraz dla pokrycia $\mathcal{Q}$ definiujemy jako:

$$h_W(\mu;(\nu_Q)_{Q\in\mathcal{Q}}) = -\sum_{Q\in\mathcal{Q}} \nu_Q(X)\log\nu_Q(X),$$

$$H_W(\mu;\mathcal{Q}) = \inf\{h(\mu;(\nu_Q)_{Q\in\mathcal{Q}}) : (\nu_Q)_{Q\in\mathcal{Q}} \text{ spełniający (1) i (2)}\}.$$

Najistotniejszy rezultat pracy [I] opisany jest w poniższym twierdzeniu:

RÓWNOWAŻNOŚĆ ENTROPII [I, THEOREM II.1]. *Entropia ważona jest równoważna klasycznej, to znaczy dla dowolnego mierzalnego pokrycia $\mathcal{Q}$ przestrzeni X zachodzi*

$$H(\mu;\mathcal{Q}) = H_W(\mu;\mathcal{Q}).$$

Wypowiadając to twierdzenie w języku kodowania możemy orzec, że dla każdego kodowania niedeterministycznego istnieje kodowanie deterministyczne o tej samej entropii i odwrotnie.

Podobnie jak J. D. Howroyd dzięki zastosowaniu ważonych miar Hausdorffa wykorzystał narzędzia analizy funkcjonalnej do wyznaczenia ograniczenia wymiaru Hausdorffa iloczynu kartezjańskiego przestrzeni metrycznych, entropia ważona umożliwiła wykorzystanie przestrzeni wektorowej funkcji do obliczania entropii (a w dalszej kolejności wymiaru entropijnego) mieszaniny źródeł. W szczególności wykazana została następująca nierówność [I, Theorem III.1]:

$$H(a_1\mu_1 + a_2\mu_2;\mathcal{Q}) \le a_1\, H(\mu_1;\mathcal{Q}) + a_2\, H(\mu_2;\mathcal{Q}) - a_1\log a_1 - a_2\log a_2, \tag{3}$$

gdzie $a_1, a_2 \in (0,1)$ spełniają $a_1 + a_2 = 1$, μ_1, μ_2 są miarami probabilistycznymi, a $\mathcal{Q}$ mierzalnym pokryciem. Powyższy rezultat oznacza w języku kodowania, że statystyczna pamięć potrzebna do zakodowania elementu wysłanego ze źródeł połączonych nie przekracza pamięci potrzebnej do zakodowania dowolnego elementu algorytmem stowarzyszonym ze źródłem składowym powiększonej o kod identyfikatora źródła (algorytmu kodującego). W pracy [I] podany został również zachłanny algorytm budujący kodowanie spełniające powyższe ograniczenie.

Podana następnie została formuła na wymiar entropijny wypukłej kombinacji miar [I, Corollary IV.1]:

$$\dim(a_1\mu_1 + a_2\mu_2) = a_1\dim(\mu_1) + a_2\dim(\mu_2),$$

co jest nawiązaniem do wspomnianych ważnych wyników uzyskanych przez J. D. Howroyda. Rozważona została także zależność pomiędzy górnym wymiarem entropijnym a górnym

wymiarem lokalnym miary (twierdzenie Younga). Uzyskany wynik [I, Theorem IV.3]:

$$\overline{\dim}(\mu) \leq \int_{\mathbb{R}^N} \overline{D}_\mu(x) d\mu(x),$$

stanowi poprawę oszacowania wykazanego przez A. Fan w 2002 roku.

W pracy [II] rozważony został problem równoważnej definicji ważonej wersji entropii dla szerokiej klasy funkcji. Jak wiadomo entropia Shannona określa w przybliżeniu statystyczną długość kodu w przypadku gdy koszt kodowania symbolu jest linową funkcją długości kodu. W przypadku wystąpienia innej zależności pomiędzy kosztem kodowania a długościami kodów inne funkcje entropii realizują statystyczny koszt kodowania. L. L. Campbell pokazał, że w przypadku gdy koszt kodowania jest wykładniczą funkcją długości, entropia Rényiego stanowi dolne ograniczenie kosztu kodowania. Zostały również wypowiedziane twierdzenia kodowe dla innych funkcji entropii.

W pracy [II] sformułowany został warunek ogólnej funkcji entropii, który pozawala określić, kiedy dana funkcja entropii daje się równoważnie wyrazić w formie ważonej:

RÓWNOWAŻNOŚĆ ENTROPII [II, THEOREM 3.1]. *Mając dane ciągłe funkcje rzeczywiste f, g, funkcja entropii postaci:*

$$f(\sum_{P\in\mathcal{P}} g(\mu(P))),$$

daje się równoważnie przedstawić w formie ważonej, jeśli zachodzi jeden z poniższych warunków:

— *f jest rosnąca, a g jest subaddytywna i wklęsła,*

— *f jest malejąca, a g jest superaddytywna i wypukła.*

Powyższy warunek jest spełniony przez najważniejsze funkcje entropii, takie jak entropia Shannona, Rényiego oraz Tsallisa. Jako przykład wykorzystania tego twierdzenia wyznaczone zostało oszacowanie na entropię Tsallisa H_α^T, $\alpha > 0, \alpha \neq 1$, źródeł połączonych μ_1, μ_2 [II, Theorem 4.1]:

$$H_\alpha^T(a_1\mu_1 + a_2\mu_2; \mathcal{Q}) \geq a_1 H_\alpha^T(\mu_1; \mathcal{Q}) + a_2 H_\alpha^T(\mu_2; \mathcal{Q}),$$

$$H_\alpha^T(a_1\mu_1 + a_2\mu_2; \mathcal{Q}) \leq a_1^\alpha H_\alpha^T(\mu_1; \mathcal{Q}) + a_2^\alpha H_\alpha^T(\mu_2; \mathcal{Q}) + \frac{a_1^\alpha + a_2^\alpha - 1}{1-\alpha},$$

gdzie a_1, a_2 są nieujemnymi liczbami spełniającymi $a_1 + a_2 = 1$, a $\mathcal{Q}$ mierzalnym pokryciem X. Pokazany został związek tego wyniku z oszacowaniem uzyskanym dla entropii Shannona analogicznej kombinacji [I, Theorem III.1].

Praca [III] koncentruje się oszacowaniu wymiaru entropijnego. Wymiar entropijny (skojarzony z odpowiednią funkcją entropii) opisuje asymptotyczne własności entropii w przestrzeni metrycznej. Dokładniej, określa on jak zachowuje się entropia pokrycia rodziną

kul $\mathcal{Q}_\delta$ o ustalonym promieniu $\delta > 0$ względem logarytmu tego promienia w przypadku granicznym (gdy promień zmierza do zera):

$$\dim(\mu) = \lim_{\delta \to 0} \frac{H(\mu; \mathcal{Q}_\delta)}{-\log(\delta)}.$$

Powyższa formułą jest dobrze zdefiniowana, jeśli badana granica istnieje. W przeciwnym przypadku mówimy jedynie o wymiarze dolnym i górnym.

I. Ciszár podał oszacowanie na entropię Rényiego kombinacji miar dla rzeczywistych przestrzeni probabilistycznych. Wprost z tego wynika formuła na wymiar entropijny Rényiego kombinacji. W pracy [III] rozważony jest analogiczny problem estymacji wymiaru, ale dla ogólnych przestrzeni metrycznych.

Przy pomocy ważonej wersji entropii, która to może zostać równoważnie zdefiniowana dla funkcji entropii Rényiego (jak wynika z [II]) uzyskujemy oszacowanie na entropię Rényiego kombinacji miar. Stosując podobne techniki jak w pracy [I] ten wynik wykorzystujemy do znalezienia formuły na wymiar entropijny Rényiego kombinacji:

WYMIAR ENTROPIJNY KOMBINACJI MIAR [III, COROLLARY III.1]. *Jeśli μ_1, μ_2 mają wymiary entropijne Rényiego, to kombinacja $a_1\mu_1 + a_2\mu_2$, gdzie $a_1, a_2 \in (0,1)$ oraz $a_1 + a_2 = 1$, też ma wymiar entropijny oraz zachodzi:*

$$\dim_\alpha(a_1\mu_1 + a_2\mu_2) = \begin{cases} \max\{\dim_\alpha(\mu_1), \dim_\alpha(\mu_2)\}, \text{dla } \alpha \in (0,1), \\ \min\{\dim_\alpha(\mu_1), \dim_\alpha(\mu_2)\}, \text{dla } \alpha \in (1,\infty). \end{cases} \tag{4}$$

Powyższa formuła jest identyczna z tą wyprowadzoną przez I. Ciszára, z tym że została ona formalnie dowiedziona dla ogólnych przestrzeni metrycznych.

Druga część rozprawy skupia się na zastosowaniach narzędzi entropijnych w klastrowaniu danych. Rozważone w tej części są dwa problemy – pierwszy dotyczy oceny zgodności uzyskanego podziału danych z zadanym podziałem referencyjnym, a drugi zastosowania klastrowania opartego o entropię krzyżową w segmentacji obrazów.

Praca [IV] rozważa problem podziału związków chemicznych aktywnych względem receptora 5-HT_{1A}. Jest to ośrodek odpowiedzialny za regulację niektórych funkcji układu nerwowego, a związki aktywne stanowią źródło leków na choroby tego układu. Jedną z podstawowych metod reprezentacji związków chemicznych są fingerprinty, czyli ciągi binarne. Kolejne współrzędne fingerprintu oznaczają obecność określonej cechy związku lub jej brak. W użyciu jest wiele reprezentacji fingerprintowych, a najbardziej złożone, takie jak fingerprint Klekota-Roth, zawierają nawet 4860 bitów. Celem pracy jest określenie, które metody klastrowania (wraz z reprezentacją fingerprintową danych) dokonują najwłaściwszego podziału

związków chemicznych pod względem ich cech strukturalnych. Jako podział referencyjny został przyjęty manualnie utworzony podział ekspercki przez D. Warszyckiego.

Do oceny wykorzystano wskaźnik ACI (Asymmetric Clustering Index), bazujący na informacji wzajemnej. Informacja wzajemna opisuje ilość informacji jaką jedno klastrowanie niesie o drugim. Aby utworzony wskaźnik przyjmował wartości z zakresu od zera do jeden, informacja wzajemna $\mathrm{I}(\mathcal{R}, \mathcal{P})$ została znormalizowana przez entropię podziału referencyjnego $h(\mathcal{R})$:

$$\mathrm{ACI}_{\mathcal{R}}(\mathcal{P}) = \frac{\mathrm{I}(\mathcal{R}, \mathcal{P})}{h(\mathcal{R})}.$$

Najważniejszą cechą wprowadzonego wskaźnika jest jego asymetria, co odróżnia go od innych powszechnie używanych wskaźników walidujących. Pozwala to wyróżnić podział referencyjny oraz określić na ile utworzony podział odzwierciedla referencję, a nie jak to się zwykle robi – na ile podziały są do siebie podobne.

Zbadano grupę hierarchicznych metod klastrowania z różnorodnym zestawem parametrów, takim jak metryka, funkcja łącząca oraz reprezentacja danych. W wyniku eksperymentu wyłoniono kombinację metryki Busera z funkcją łączącą "complete linkage function" oraz reprezentacją za pomocą fingerprintu Klekota-Roth, która pozwala na utworzenie podziału najlepiej opisującego charakter badanej przestrzeni. Badania dowiodły, że największy wpływ na dokonywany podział ma funkcja łącząca, w drugiej kolejności przyjęta reprezentacja danych, a ostatecznie metryka. Ten wynik pozwala mieć nadzieję, że wyznaczone parametry będą również właściwe do podziału danych w innych przestrzeniach chemicznych.

W pracy [V] zostało przedstawione zastosowanie metody klastrowania opartej o entropię krzyżową, CEC (cross-entropy clustering) do segmentacji obrazów. Algorytm CEC dzieli dane na grupy tak, aby minimalizować łączny koszt ich zapisu przyjmując dla każdej grupy pewien optymalny normalny rozkład prawdopodobieństwa. Dokładniej, mając dane parami rozłączne grupy $(U_i)_i$ przestrzeni danych $X \subset \mathbb{R}^N$ oraz rozważając klasę gaussowskich rozkładów prawdopodobieństwa minimalizowana jest funkcja uogólnionej entropii krzyżowej:

$$E(U_1, \ldots, U_k) = \sum_{i=1}^{k} p(U_i) \cdot [\frac{N}{2} \ln(2\pi e) - \ln(p(U_i)) + \frac{1}{2} \ln \det(\Sigma_{U_i})], \tag{5}$$

gdzie $p(U_i) = \frac{\mathrm{card}(U_i)}{\mathrm{card}(U)}$ oraz Σ_{U_i} oznacza macierz kowariancji U_i. Procedura minimalizacji jest przeprowadzana z użyciem powszechnie znanego podejścia Hartigana i znajduje ona pewne minimum lokalne funkcji (5). Z uwagi na obecność entropii Shannona w powyższym wzorze określającej koszt pamiętania identyfikatora grupy, wprowadzony jest dodatkowy koszt utrzymywania klastra, dlatego też algorytm jest w stanie redukować nadmiarowe grupy automatycznie.

Praca [V] skupia się na użyciu CEC w problemie segmentacji obrazów. Zaprezentowany jest proces przekształcenia danych obrazu do wejścia algorytmu klastrowania. Pojedynczy wektor stanowi zestawienie współrzędnych kolorów określonego otoczenia danego piksela (bloki wymiaru $b \times b$). Na takim zestawie przeprowadzamy procedurę PCA w celu redukcji nadmiarowych współrzędnych. Następnie każdy wektor wynikowy uzupełniamy współrzędnymi przestrzennymi badanego piksela. Zbiór takich wektorów stanowi wejście do algorytmu CEC.

Uzyskane wyniki pokazują przewagę przedstawionej metody nad segmentacją z użyciem k-means, jednej z najpopularniejszych metod klastrowania. Praktyczna przydatność zbadanej metody jest podyktowana jej dużą niezmienniczością na transformacje afiniczne takie jak skalowanie zdjęcia oraz automatycznym doborem wynikowej ilości grup.