Wrocław, 02 października 2014

prof. dr hab. Jacek Cichoń
Katedra Informatyki
Wydział Podstawowych Problemów Techniki
Politechnika Wrocławska
email: jacek.cichon@pwr.edu.pl

Recenzja rozprawy doktorskiej mgr Marka Śmieji

Mgr Marek Śmieja przedstawił rozprawę doktorskę pt. „Entropy i coding and data clustering". Jego promotorem jest dr hab. Jacek Tabor z Wydziału Matematyki i Informatyki Uniwersytetu Jagielońskiego. Rozprawę tą otrzymałem do recenzji w lipcu 20014 r. Na przedstawioną rozprawę składa się pięć prac:

1. Marek Śmieja, Jacek Tabor, *Entropy of mixture of sources and entropy dimension*, IEEE Trans. on Information Theory, 2012

2. Marek Śmieja, *Weighted approach to general entropy function*, IMA Journal of Math. Control and Information, 2014

3. Marek Śmieja, Jacek Tabor, *Renyi entropy dimension of the mixture of measures*, praca konferencyjna (SAI 2014)

4. Marek Śmieja, Dawid Warzycki, Jacek Tabor, Andrzej Bojarski, *Asymetric Clustering Index in the case study of* $5-\mathrm{HT}_{1A}$ *receptor ligands*, PLoS ONE (jest to recenzowane internetowe czasopismo naukowe wydawane przez Public Library of Science na zasadach otwartego dostępu)

5. Marek Śmieja, Jacek Tabor, *Image segmentation with use of cross-entropy clustering*, praca konferencyjna (CORES 2013)

Oprócz zestawienia tych prac autor przedstawił (według mnie zupełnie niepotrzebnie) dokument o nazwie „Przewodnik do rozprawy doktorskiej Entropia w kodowaniu i klasterowaniu danych".

1 Prace teoretyczne

Prace (1), (2) i (3) mają charakter teoretyczny. Prace (1) oraz (2) zostały opublikowane w recenzowanych czasopismach. Trzecia praca jest pracą konferencyjną (Science and Information Conference (SAI'2014), Londyn, sierpień 2014).

1.1 Praca „Entropy of mixture of sources and entropy dimension"

W pracy (1) autorzy definiują pojęcie entropii względem dowolnego pokrycia $\mathcal{Q}$ przestrzeni probabilistycznej. Określają ją w następujący sposób:

$$H(\mu;\mathcal{Q}) = \inf\{h(\mu,\mathcal{P}) : \mathcal{P} \in \mathbb{P} \wedge \mathcal{P} \prec \mathcal{Q}\}$$

gdzie $\mathbb{P}$ jest rodziną wszystkich przeliczalnych rozbić (na zbiory mierzalne) rozważanej przestrzeni probabilistycznej, symbol $\prec$ oznacza rozdrobnienie (czyli $\forall P \in \mathcal{P})(\exists Q \in \mathcal{Q})(P \subseteq Q)$) zaś $h(\mu,\mathcal{P})$ oznacza klasyczną entropię miary probabilistycznej μ względem rozbicia $\mathcal{P}$, czyli liczbę $\sum\{-\mu(P)\log_2(\mu(P)) : P \in \mathcal{P}\}$.

Głównym wynikiem tej pracy jest następująca nierówność (Theorem III.1):

$$H(\alpha\mu_1 + \beta\mu_2;\mathcal{Q}) \leqslant \alpha H(\mu_1;\mathcal{Q}) + \beta H(\mu_2;\mathcal{Q}) + sh(\alpha) + sh(\beta)\ ,$$

gdzie μ_1, μ_2 są dwoma miarami probabilistycznymi na ustalonej przestrzeni probabilistycznej, $\alpha,\beta \geqslant 0$, $\alpha+\beta = 1$ oraz $sh(x) = -x\log_2(x)$. Głównym pomysłem w dowodzie tego wyniku jest rozważanie rodziny miar związanych w rozważanym rozbiciem $\mathcal{Q}$:

$$W(\mu;\mathcal{Q}) = \{m \in M(X,\Sigma)^{\mathcal{Q}} : (\forall Q \in \mathcal{Q})(m_Q(X \setminus Q) = 0) \wedge \sum_{Q\in\mathcal{Q}} m_Q = \mu\}\ ,$$

gdzie $M(X,\Sigma)$ oznacza rodzinę miar probabilistycznych na σ-ciele Σ. Definiują następnie pojęcie entropii ważonej dla $m \in W(\mu;\mathcal{Q})$ (Definicja II.4): $h_W(\mu;m) = \sum\{sh(m_Q) : Q \in \mathcal{Q}\}$, pojęcie ważonej μ entropii dla pokrycia $\mathcal{Q}$: $H_W(\mu;\mathcal{Q}) = \inf\{h_W(\mu;m) : m \in W(\mu;\mathcal{Q})\}$ oraz pokazują, że $H_W(\mu;\mathcal{Q}) = H(\mu;\mathcal{Q})$. Ostatnia równość pozwala na zastąpienie rachunków związanych z wyznaczeniem klasycznej entropii $H(\mu;\mathcal{Q})$ pokrycia $\mathcal{Q}$ rachunkami związanych z wyznaczeniem $H_W(\mu;\mathcal{Q})$, czyli z obliczeniami związanymi z wyznaczeniem wartości $h_W(\mu;m)$ dla ciągu miar $m = (m_Q)_{Q\in\mathcal{Q}}$.

Rozważania prowadzone w tej pracy są elementarne: to, że udało się tak elementarnymi środkami udowodnić interesujący wynik świadczy o tym, że pomysł zastąpienia rachunków z danym pokryciem $\mathcal{Q}$ analizą ciągów z rodziny $M(X,\Sigma)$ jest bardzo efektywny. W dowodzie jednego faktu (Proposition II.2) autorzy wykorzystali uogólnienie klasycznej nierówności Hardy-Polya-Littlewooda (HPL) na przypadek nieskończony.

1.2 Praca „Weighted approach to general entropy function"

Praca (2) jest jedyną samodzielną pracą autora sposród prac wchodzących składających się na przedstawioną rozprawę. Autor uogólnia w niej część wyników z pracy (1). Zauważa mianowicie, że nierówność HPL stosowaną w poprzedniej pracy do sum postaci $\sum_{p\in\mathcal{P}} \mu(P)\log_2(\mu(P))$ można zastosować również do wyrażeń postaci $f(\sum_{p\in\mathcal{P}} g(\mu(P)))$ jeśli f jest malejąca zaś g jest super-addytywna i wypukła. Głównym wynikiem tej pracy jest Theorem 3.1, pokazujące, że ważona entropia jest równa klasycznej entropii również i w tym przypadku (jest to odpowiednik Theorem II.1 z pracy (1)). Jest to o tyle ciekawe, że w klasie funkcji postaci $f(\sum_{p\in\mathcal{P}} g(\mu(P)))$ zawiera się nie tylko klasyczne entropia Shannona, lecz również entropie Renyi oraz Tsallis'a.

1.3 Praca „Renyi entropy dimension of the mixture of measures"

W pracy (3) autorzy również zajmują się się "'mieszankami miar"'. Zajmują się, między innymi, pojęciem Renyi wymiaru entropijnego miary ($\overline{dim}_\alpha(\mu)$). Pokazują, dla dla dowolnych dwóch miar μ_1, μ_2 oraz liczb $b, b \geqslant 0$ takich, że $a+b=1$ zachodzą następujące nierówności (Theorem III.1):

$$\overline{dim}_\alpha(a_1\mu_1 + a_a\mu_2) \leqslant \max\{\overline{dim}_\alpha(\mu_1), \overline{dim}_\alpha(\mu_2)\}$$

dla $\alpha \in (0,1)$ oraz $\overline{dim}_\alpha(a_1\mu_1 + a_a\mu_2) \leqslant \min\{\overline{dim}_\alpha(\mu_1), \overline{dim}_\alpha(\mu_2)\}$ dla $\alpha \in (1,\infty)$. Podobny wynik otrzymują dla dualnego pojęcia $\underline{dim}_\alpha(\mu)$, a w przypadku, gdy obie miary mają wymiar entropijny Renyi zachodzą następujące dwie eleganckie równości (Corollary III.1):

$$dim_\alpha(a_1\mu_1 + a_a\mu_2) = \max\{dim_\alpha(\mu_1), dim_\alpha(\mu_2)\}$$

dla $\alpha \in (0,1)$ oraz

$$dim_\alpha(a_1\mu_1 + a_a\mu_2) = \min\{dim_\alpha(\mu_1), dim_\alpha(\mu_2)\}$$

dla $\alpha \in (1,\infty)$.

2 Prace aplikacyjne

Prace (4) oraz (5) mają charakter aplikacyjny. Nie znalazłem w niech żadnych wyników teoretycznych. Traktuję je jako raporty z przeprowadzonych obliczeń numerycznych. Co więcej, z oświadczenia promotora wynika, że wkład mgra M. Śmieji w te prace polegał głównie na pracy informatycznej oraz redakcyjnej.

2.1 Praca „Asymetric Clustering Index in the case study of $5-\mathrm{HT}_{1A}$ receptor ligand"

W pracy (4) autorzy analizują skuteczność pewnej klasy algorytmów automatycznej klasteryzacji. Do porównania ich skuteczności wybrano tzw. Asymetric Clustering Index (ACI) zdefiniowany jako

$$ACI_R(C) = \frac{MI(R,C)}{SE(R)},$$

gdzie $MI(R,C)$ oznacza "„mutual information" zaś $SE(R)$ oznacza entropię bazowej partycji. Za bazową partycję przyjmuje się partycje opracowaną przez eksperta. Dla konkretnych partycji $R = (R_i)_{i\in I}$ oraz $C = (C_j)_{j\in J}$ wyznaczanie współczynnika ACI sprowadza się do wyznaczenia wartości

$$\sum_{(i,j)\in I\times J} \Pr[R_i \cap C_j] \log_2(\Pr[R_i \cap C_j]) / \sum_{i\in I} \Pr[R_i] \log_2(\Pr[R_i])$$

Po przeprowadzeniu szeregu obliczeń numerycznych autorzy wskazali jeden zestaw znanych metod (complete linkage function, odcisk palca Klekoty-Roth'a i metryka podobieństwa Busera) jako najlepszą do klasteryzacji analizowanego materiału biologicznego.

2.2 Praca „Image segmentation with use of cross-entropy clustering"

W pracy tej mgr M. Śmieja wraz z promotorem badają eksperymentalnie skuteczność pewnej metody klasteryzacji obrazów bazującej na „cross-entropy clustering". Badania swoje testują na dwóch obrazach (na jednym z nich jest niedźwiedź a na drugim jest kobieta). Badana metoda jest ciekawa, gdyż jest mało wrażliwa na transformacje afiniczne obrazu.

Podsumowanie

Wyniki teoretyczne zawarte w pracach (1), (2) i (3) uważam za ciekawe. Zaproponowana jest w nich nowa metoda „entropii ważonej", która okazała się przydatna do otrzymania ciekawych i eleganckich wyników. Uważam, że prace (4) i (5) nie mają większej wartości merytorycznej - oceniam je jako typowe, przeciętne informatyczne prace konferencyjne. Swoją opinię o sprawności technicznej mgra M. Śmieji wyrobiłem sobie na podstawie pierwszych trzech prac. Mimo iż stosowane są w nich elementarne środki dowodowe, to rozumowania są pomysłowe i świadczą o dobrym opanowaniu przez autorów podstawowego warsztatu badawczego. Na uwagę zasługuje eleganckie uogólnienie nierówności HPL na ciągi nieskończone. Gdyby te trzy prace mgr M. Śmieja napisał samodzielnie, to rozprawę tą uznałbym za bardzo dobrą (i być może, w tej sytuacji, wnioskowałbym o uznanie jej za wyróżniającą).

Z informacji umieszczonych na stronach internetowych mgra M. Śmieji (http://ww2.ii.uj.edu.pl/ smieja/cv) dowiedziałem się, że w roku 2010 rozpoczął on studia doktoranckie. Znając realia obecnych studiów doktoranckich wiem, że przygotowanie ostatecznych wersji rozprawy doktorskich odbywa się często pod presją czasu. Z pewnością też negatywny wpływ na wielkość jego dorobku naukowego miało to, że w roku 2011 (a więc w trakcie trwania studiów doktoranckich) ukończył on drugi kierunek studiów (BSc in Computer Science) oraz że w trakcie trwania studiów doktoranckich angażował się we współpracę z firmami komercyjnymi (Comarch, Samsung).

W dołączonym oświadczeniu o współautorstwie, jego promotor, dr hab. Jacek Tabor, stwierdza, że wkład mgra M. Smieji do wspólnych prac (1) i (3) wynosi 50%. W związku z tym uważam, że jest on, w przybliżeniu, autorem dwóch $(1 + 0.5 \cdot 2 = 2)$ dobrych prac naukowych (jak już zaznaczyłem, nie biorę pod uwagę prac (4) i (5)). Biorąc pod uwagę obecne wymagania stawiane rozprawom doktorskim oświadczam, że rozprawa doktorska mgra M. Śmieji spełnia **minimalne** ustawowe jak i zwyczajowe wymagania. W związku z czym wnioskuję o dopuszczenie go do dalszych etapów przewodu doktorskiego.

Jacek Cichoń