Kraków, 5 września, 2014

dr hab. Prof. AGH Andrzej Bielecki
Katedra Informatyki Stosowanej
Wydział Elektrotechniki, Automatyki, Informatyki i Inżynierii Biomedycznej
Akademia Górniczo-Hutnicza

Recenzja rozprawy doktorskiej:

Marek Śmieja
Entropia w kodowaniu i klastrowaniu danych

Niniejsza recenzja została sporządzona na zlecenie Dziekana Wydziału Matematyki i Informatyki Uniwersytetu Jagiellońskiego w związku ze staraniami Pana Marka Śmieji o uzyskanie stopnia doktora nauk matematycznych w dziedzinie informatyki.

Rozprawa doktorska Pana Marka Śmieji składa się z pięciu opublikowanych lub przyjętych do druku artykułów:

[1] Śmieja M., Tabor J. (2012),
Entropy of the mixture of sources and entropy dimension,
IEEE Transactions of Information Theory, vol.58, 2719-2728.

[2] Śmieja M. (2014),
Weighted approach to general entropy function,
IMA Journal of Mathematical Control and Information, przyjęte do druku.

[3] Śmieja M., Tabor J. (2014),
Renyi entropy dimension of the mixture of measures,
Proceedings of Science and Information Conference SAI, 2014, przyjęte do druku.

[4] Śmieja M., Warszycki D., Tabor J. Bojarski A. (2014),
Asymmetric clustering index in a case study of 5-HT_{1A} receptor ligands,
PLoS ONE, przyjęte do druku.

[5] Śmieja M., Tabor J. (2013),
Image segmentation with use of cross-entropy clustering,
Proceedings of the 8th International Conference on Computer Recognition Systems CORES, 403-409.

Prace [1], [2], [3] mają charakter teoretyczny, natomiast prace [4] i [5] – aplikacyjny. Nadmienić należy, że prace [1] i [4] ukazały się w bardzo prestiżowych czasopismach,

natomiast praca [2] w czasopiśmie prestiżowym. Z dołączonych oświadczeń wynika, że wkład doktoranta w publikacjach współautorskich wynosi:

- w pracy [1] 50%,
- w pracy [3] 50%,
- w pracy [4] 40% i jest większy od wkładu każdego spośród pozostałych współautorów,
- w pracy [5] 60%.

Ponadto wszyscy współautorzy wszystkich czterech publikacji współautorskich oświadczyli, że uważają Pana Marka Śmieję za głównego autora każdej z nich.

Zawartość pracy

W publikacji [1], po omówieniu entropii Shannona oraz problemu kodowania, wprowadzona jest definicja entropii ważonej (Def.II.4). Wprowadzona entropia jest obliczana jako infimum rodziny funkcji określonych na rodzinie mierzalnych podzbiorów ustalonej przestrzeni o wartościach w zbiorze miar tej przestrzeni – funkcje $h_W(\mu;m)$ (wzór (16)). Podejście takie pozwala na obliczanie entropii wypukłej kombinacji miar. Ma to ścisły związek z obliczaniem entropii mieszanego źródła sygnałów, składającego się ze skończonej liczby źródeł S_k wysyłających sygnały z prawdopodobieństwami a_k, przy czym $\Sigma a_k = 1$. Następnie wykazana jest równoważność entropii ważonej i klasycznej dla rodziny zbiorów mierzalnych (Th.II.1). Istotnymi wynikami publikacji [1] jest oszacowanie ważonej entropii wypukłej kombinacji miar (Th.III.1) oraz oszacowanie wymiaru entropijnego wypukłej kombinacji miar (Th.IV.1, Th.IV.2, Th.IV.3).

W publikacji [2] rozważana jest ogólna funkcja entropii. Sformułowany został warunek ogólnej funkcji entropii *(the condition of general entropy function)* oraz twierdzenie o równoważnej reprezentacji ogólnej funkcji entropii w postaci ważonej (Th.3.1). Powyższe twierdzenie jest zastosowane do oszacowania entropii Tsallisa mieszanego źródła sygnałów (Prop.4.1) oraz superpozycji miar (Th.4.1). Wykazano, że przy przyjętych założeniach oszacowania podane w twierdzeniu 4.1. są najlepsze.

W publikacji [3] podano oszacowanie na entropię Renyia dla kombinacji miar (Th.II.1). Bazując na powyższym wyniku podano oszacowanie na wymiar entropii Renyia dla kombinacji miar (Th.III.1).

W publikacji [4] poruszony jest problem klastrowania danych. Metoda entropijna oceny poprawności klasteryzacji została zastosowana do oceny różnych metod klastrowania pewnej klasy związków biochemicznych ze względu na ich własności. Podział referencyjny został dokonany przez eksperta. Autorzy wprowadzili asymetryczny,

znormalizowany wskaźnik pozwalający porównywać dwie klasteryzacje tego samego zbioru danych. Wprowadzony wskaźnik jest ilorazem miary wzajemnej informacji dokonanych klasteryzacji oraz entropii klasteryzacji referencyjnej i przybiera wartości z przedziału [0,1].

W publikacji [5] rozważany jest problem segmentacji obrazów. Użyty algorytm segmentacji wykorzystuje entropię krzyżową *(cross entropy),* której używa się w przypadku, gdy rozkład prawdopodobieństwa danych nie jest znany. Zdefiniowana jest funkcja energetyczna E (wzór (1)) oraz znajdowany jest taki podział, dla którego funkcja ta osiąga minimum. Wartość funkcji E jest miarą średniej długości kodowania elementu. W pracy podane są przykłady działania algorytmu dla dwóch obrazów przy różnych wartościach parametrów algorytmu. Wyniki są porównane z segmentacją opartą na algorytmie klasteryzacji metodą k-najbliższych sąsiadów.

Uwagi

Moim zdaniem dysertacja doktorska Pana Śmieji zawiera istotne wyniki naukowe. Za szczególne osiągnięcie uważam wprowadzenie w pracy [1] entropii ważonej wzorowanej na ważonych miarach Hausdorffa, co pozwala na szacowanie entropii kombinacji miar i, co za tym idzie, entropii mieszanych źródeł sygnałów. W pracach [1], [2], [3] zostały zbadane pewne własności entropii ważonej oraz związki między różnymi rodzajami entropii. Ponadto, wynik uzyskany w [3] jest istotnym uogólnieniem wyniku uzyskanego przez Csiszara (Csiszar I., *On the dimension and entropy of order α of the mixture of probabilisty distributions,* Acta Math. Hungarica, vol.13, 1962, 245-255). Podejście entropijne zastosowane w [4] do oceny porównawczej kilku zaproponowanych klasteryzacji tego samego zbioru danych w stosunku do klasteryzacji referencyjnej dokonanej przez eksperta pozwoliło na wyłonienie najskuteczniejszego algorytmu klasterowania danego zbioru. W przypadku związków biochemicznych ma to istotne znaczenie w kontekście badań nad nowymi lekami. Ze względu na czasochłonność oraz wysoką cenę tego typu badań wszelkie wspomaganie komputerowe w tym zakresie jest zarówno cenne jak też perspektywiczne. Podejście entropijne zastosowane w pracy [5] do segmentacji obrazów zawiera wyniki pozwalające domniemywać, że rozpatrywane metody entropijne wraz z opracowanym przez Autorów publikacji algorytmem segmentacji obrazu mogą być stosowane nie tylko do preprocesingu, co jest standardem, ale również do analizy syntaktycznej obrazu zwłaszcza w algorytmach hierarchicznej analizy obrazu. Jeśli to domniemanie by się potwierdziło, otworzyłoby to drogę do

rozwoju nowej klasy metod syntaktycznego rozpoznawania obrazów. Podsumowując uwagi o merytorycznej wartości przedstawionej dysertacji doktorskiej uważam, że nie tylko zawiera ona istotne wyniki naukowe, ale stanowi punkt wyjścia do dalszych, interesujących i perspektywicznych badań zarówno w aspekcie teoretycznym jak i aplikacyjnym. Potwierdzeniem tego jest fakt, że kolejne dwie publikacje będące kontynuacją badań prezentowanych w dysertacji doktorskiej znajdują się obecnie w recenzji (*Śmieja M., Tabor J., „Entropy estimation in lossy compression"* oraz *Śmieja M., Wiercioch M., „Constrained clustering based on inner partititions detection"*).

Przechodząc do uwag krytycznych, jedyne zauważone przeze mnie usterki dotyczą pracy [5] i mają wyłącznie charakter redakcyjny. Mianowicie, nie jest w pracy wyjaśnione co oznaczają zmienne ε oraz N we wzorze (1). Ponadto, w tym samym zbiorze, lewa strona jest zależna od obu wspomnianych parametrów a prawa nie, co niewątpliwie jest usterką formalną.

Konkluzja

Konkludując, uważam, że rozprawa doktorska Pana Marka Śmieji spełnia wymogi Ustawy o stopniach i tytułach naukowych i w związku z tym wnioskuję o dopuszczenie Doktoranta do dalszych etapów przewodu doktorskiego. Ponadto, ze względu na bardzo wysoki poziom merytoryczny i wartość uzyskanych wyników, zarówno w aspekcie teoretycznym jak i aplikacyjnym, co w szczególności jest poparte faktem, że wyniki te zostały opublikowane w bardzo prestiżowych czasopismach naukowych, oraz ze względu na staranną redakcję wnioskuję o wyróżnienie rozprawy.

Andrzej Bielecki