dr hab. inż Jan Chorowski
Instytut Informatyki Uniwersytetu Wrocławskiego
ul. F. Joliot-Curie 15
50-383 Wrocław

Wrocław, 29.12.2021

Recenzja wniosku habilitacyjnego dr Marka Śmieji

Z przyjemnością zapoznałem się z pracami wykonanymi przez dra Śmieję przez ostatnie 6 lat po uzyskaniu stopnia doktora. Przedstawione przez dra Śmieję prace rysują spójną wizję możliwości przetwarzania danych przy brakujących informacjach dotyczących zarówno etykiet (zmiennych zależnych) jak również zmiennych niezależnych. Spośród załączonych przez dra Śmieję prac tylko jedną znałem przed przygotowaniem recenzji (A8, dotycząca przetwarzania niepełnych danych za pomocą sieci neuronowych), z pozostałymi pracami zapoznałem się na potrzeby przygotowania niniejszej recenzji. Z przyjemnością stwierdzam, że przedstawiony wybór prac jednoznacznie wskazuje na spełnienie przez dra Śmieję kryteriów dotyczących osiągnięcia naukowego dla stopnia dra habilitowanego.

1 Tematyka osiągnięcia

Przedstawione prace przedstawiają kompleksowo możliwości wykorzystania niepełnych danych podczas uczenia maszynowego. W szczególności przedstawiają one jak można wykorzystać:

- częściowe informacje dotyczące etykiet (zmiennych zależnych), podane w formie więzów (prace A1, A2), zgrubnej kategoryzacji (A4), zaszumionej i zgrubnej kategoryzacji (A5), lub częściowym wyspecyfikowaniem wielu kategoryzacji (A6);
- niepełne informacje o zmiennych niezależnych przez przybliżone uśrednienie odpowiedzi modelu (A7, A8), budowę modeli ignorujących brakujące dane (A9) próbkowanie uzupełnianie brakujących informacji (A10).

Przedstawię teraz bardziej szczegółowo treść wskazanych w osiągnięciu prac

- **A1: Semi-supervised discriminative clustering with graph regularization.** Knowledge-Based Systems 2018
  W pracy rozważane jest zadanie klasteringu (grupowania) danych przy niepełnej informacji dotyczącej grup - wybrane pary są oznaczone jako należące do tej samej gruby bądź jako należące do innych grup. Zaproponowano model dyskryminatywnie przypisujący każdą próbkę danych do

jednego z K klastrów. Następnie skonstruowano funkcje kosztu wykorzystującą podane więzy. Przedstawiono rozszerzenie wykorzystujące informację o podobieństwie danych. Osiągnięto dobre wyniki a danych testowych.

- **A2: A Classification-Based Approach to Semi-Supervised Clustering with Pairwise Constraints.** Neural Networks, 2020
  W pracy rozszerzono model model przypisujący dane do klastrów z pracy A1, zastępują regresję softmax głęboką siecią neuronową. Dzięki temu umożliwienie stworzenie bardziej skomplikowanych klastrów. Ponadto zaproponowano dwuetapowy proces uczenia w którym najpierw sieć w konfiguracji bliźniaczej (siamese network) uczona jest aby kategoryzować pary danych, a następnie wykorzystując koszt z pracy A1 przypisywać dane do klastrów. Wiele z pomysłów przedstawionych w pracy A2 przypomina stosowane obecnie procedury wstępnego samo-uczenia sieci neuronowych (contrastive self-training).

- **A3: Constrained clustering with a complex cluster structure.** Adv. in Data Analysis and Classification, 2017
  W pracy przedstawiono inny sposób na utworzenie klastrów o skomplikowanej budowie. Na początku więzy na parach danych użyte są do podziału danych na rozłączne podzbiory. Następnie każdy z podzbiorów jest niezależnie klasteryzowany. Końcowy klastering łączy niektóre z klastrów odkrytych w poprzednim kroku, pozwalając na zapis pojedynczego klastra za pomocą mikstury rozkładów.

- **A4: Semi-supervised cross-entropy clustering with information bottleneck constraint.** Information Sciences, 2017
  W pracy przedstawiono jak stworzyć klastering zgody z zadanym zgrubnym podziałem części danych. Założono klastry będące miksturami rozkładów normalnych i zaproponowano funkcję straty łączącą jakość klasteringu z miarą zgodności klastrów z dostarczonym podziałem. W pracy zastosowano przybliżenie log-prawdopodobieństwa mieszaniny rozkładów normalnych przez maksimum z prawdopodobieństwa pojedynczego elementu mikstury (metoda CEC). Dzięki temu uzyskano praktyczny w implementacji algorytm optymalizacji funkcji kosztu.

- **A5: Semi-supervised model-based clustering with controlled clusters leakage.** Expert Systems with Applications 2017
  Praca rozważa zadanie klasteryzacji przy dodatkowo dostarczonej hiperpłaszczyźnie w przybliżeniu rozdzielającą dane. Zaproponowano dopasowanie mieszaniny rozkładów normalnych starających się zachować zadany przez hiperpłaszczyznę podział. Metoda może być stosowana do poprawy podziału danych wyznaczonego przez klasyfikator uczony na bardzo małej ilości etykiet.

- **A6: SeGMA: Semi-Supervised Gaussian Mixture Auto-Encoder.** IEEE Transactions on Neural Networks and Learning Systems, 2020
  W pracy rozszerzono auto-enkodery Wassersteina do pracy z miksturami rozkładów normalnych w przestrzeni ukrytej. Ponadto pokazano jak

można wykorzystać dodatkowe informacje o danych (np. częściowo etykietowane dane można przypisać do składników mikstury). Uzyskano dobre wyniki na zdjęciach twarzy.

- **A7: Generalized RBF kernel for incomplete data.** Adv. in Data Analysis and CKnowledge-Based Systems, 2019
  W pracy przedstawiono sposób na uczenie i ewaluację modeli wykorzystujące funkcje jądrowe (kernel functions) przy niepełnych danych wejściowych. Pokazano że możliwe jest modelowanie gęstości prawdopodobieństwa brakujących danych oraz obliczanie wartości oczekiwanej funkcji jądrowych przy brakujących danych. Wykazano, że używanie wartości oczekiwanej funkcji jądrowej poprawia jakość klasyfikatorów względem bezpośredniego zastępowania brakujących danych.

- **A8: Processing of missing data by neural networks.** NeurIPS 2018
  W pracy dostosowany pomysł przedstawiony w A7 dla głębokich sieci neuronowych, pokazując jak obliczać wartość oczekiwaną neuronu z aktywacją ReLU przy niepełnych danych wejściowych. Uzyskano praktyczny model w którym estymacja rozkładów prawdopodobieństwa brakujących danych uzyskiwana jest podczas uczenia modelu. Zaproponowana metoda osiągała bardzi dobre wyniki w przeprowadzonych eksperymentach.

- **A9: Processing of missing data by (Graph) Convolutional Neural Networks.** ICONIP 2020
  W pracy zaproponowane zamianę warstw splotowych sieci neuronowych na warstwy grafowe. W efekcie stworzono sieci pomijające nieznane fragmenty danych wejściowych. Jest to alternatywny sposób przetwarzania danych brakujących - zamiast zastępowania ich innymi wartościami, stworzono model który po prostu ignoruje brakujące wartości.

- **A10: Iterative Imputation of Missing Data using Auto-encoder Dynamics.** ICONIP 2020
  W pracy zaproponowano metodę uzupełniania brakujących danych za pomocą sieci neuronowych. W tym celu wytrenowano auto-enkodery na pełnych danych. Następnie, w celu uzupełnienia niepełnych danych podążano w kieruku wskazywanym przez rekonstrukcję autoenkodera, który odpowiada pochodnej z log-prawdopodobieństwa danych w rozkładzie implikowanycm przez auto-enkoder.

Opisane w pracach A1-A10 metody są praktyczne, łatwe w implementacji i wykorzystują sprawdzone techniki (np. optymalizację gradientową). Eksperymenty sprawdzające ich jakość są wykonane poprawnie, przeprowadzono analizy statystyczne uzyskiwanych wyników. W pracach przedstawiono dowody wyprowadzeń i wprowadzonych własności modeli. Cechy te wskazują na biegłość habilitanta w rozwoju i analizie metod uczenia maszynowego.

1.1 Znaczenie osiągnięcia

Zaprezentowane prace kompleksowo traktują zagadnienie przetwarzania niepełnych danych. Najbardziej cytowaną pracą dr Śmieji jest praca A8. Niewątpliwie wynika to z tematyki pracy - wprowadza ona praktyczną metodą radzenia sobie

z brakującymi danymi za pomocą sieci ReLU - jednej z najpopularniejszych architektur głębokich sieci neuronowych. Cytowania pozostałych prac wskazują na umiarkowane zainteresowanie świata naukowego (mają po parenaście zewnętrznych cytowań w Google Scholar). Prawdopodobnie wynika to z ich specjalizacji – rozwiązują bardzo specyficzne zadania.

Obecnie popularne jest wykorzystanie danych nieopisanych do samo-uczenia sieci (self-training). Jest to wariant uczenia sieci rozwiązywania zadań dla których etykietowanie danych może być zautomatyzowane (np. łączenie dwóch przekształceń tego samego zdjęcia i odróżnianie ich od przekształcenia innego zdjęcia). Podejście to przypomina uczenie z więzami rozważane w pracach A1 i A2, ciekawym rozszerzeniem tych brac byłoby połączenie ich z samo-uczeniem.

2 Nawiązane współprace

Przedstawione prace były realizowane we współpracy z Uniwersytetm Lizbońskim, Politechniką Graz, Polską Akademią Nauk i Politechniką Wrocławską. Wykazują to zdolność habilitanta do nawiązywania współpracy i zespołowej realizacji prac badawczych. Są to niezbędne umiejętności, nie jest możliwe samodzielne prowadzenie zaawansowanych badań.

3 Podsumowanie

Przestawiony zbiór prac, stanowiący tylko fragment osiągnięć dra Śmieji, jednoznacznie wskazuje na jego umiejętność prowadzenia prac badawczych. W mojej opinii dorobek doktora Śmieji spełnia kryteria dotyczące osiągnięcia naukowego i pozostałego dorobku dla stopnia dr habilitowanego w dyscyplinie informatyka w dziedzinie nauk ścisłych i przyrodniczych.