Marcin Kurdziel
Instytut Informatyki
Wydział Informatyki, Elektroniki i Telekomunikacji
Akademia Górniczo-Hutnicza im. St. Staszica w Krakowie
al. Mickiewicza 30, 30-059 Kraków
kurdziel@agh.edu.pl

Recenzja w postępowaniu w sprawie nadania stopnia doktora habilitowanego doktorowi Markowi Śmieji

1 Osiągnięcie naukowe

Osiągnięciem naukowym przedstawionym przez doktora Marka Śmieję jest cykl powiązanych tematycznie artykułów naukowych, zatytułowany:

METODY UCZENIA MASZYNOWEGO DLA DANYCH NIEKOMPLETNYCH

Na cykl składa się 10 publikacji:

[A1] Marek Śmieja, Oleksandr Myronov, Jacek Tabor, *Semi-supervised discriminative clustering with graph regularization*, Konwledge-Based Systems, 151, 24–36, 2018.

[A2] Marek Śmieja, Łukasz Struski, Mario A. T. Figueiredo, *A Classification-Based Approach to Semi-Supervised Clustering with Pairwise Constraints*, Neural Networks, 127, 193–203, 2020.

[A3] Marek Śmieja, Magdalena Wiercioch, *Constrained clustering with a complex cluster structure*, Advances in Data Analysis and Classification, 11, 493–518, 2017.

[A4] Marek Śmieja, Bernhard C. Geiger, *Semi-supervised cross-entropy clustering with information bottleneck constraint*, Information Sciences, 421, 245–271, 2017.

[A5] Marek Śmieja, Łukasz Struski, Jacek Tabor, *Semi-supervised model-based clustering with controlled clusters leakage*, Expert Systems with Applications, 85, 146-157, 2017.

[A6] Marek Śmieja, Maciej Wołczyk, Jacek Tabor, Bernhard C. Geiger, *SeGMA: Semi-Supervised Gaussian Mixture Auto-Encoder*, IEEE Transactions on Neural Networks and Learning Systems, 32, 3930 - 3941, 2021[1].

[1] We wniosku wskazana jest publikacja w wersji *Early Access*. W wersji ostatecznej praca została opublikowana we wrześniu 2021 r.

[A7] Marek Śmieja, Łukasz Struski, Jacek Tabor, Mateusz Marzec, *Generalized RBF kernel for incomplete data*, Knowledge-Based Systems, 173, 150–162, 2019.

[A8] Marek Śmieja, Łukasz Struski, Jacek Tabor, Bartosz Zieliński, Przemysław Spurek, *Processing of missing data by neural networks*, Advances in Neural Information Processing Systems 31 (NeurIPS), 2719–2729, 2018.

[A9] Tomasz Danel, Marek Śmieja, Łukasz Struski, Przemysław Spurek, Łukasz Maziarka, *Processing of Incomplete Images by (Graph) Convolutional Neural Networks*, International Conference on Neural Information Processing (ICONIP), 512–523, 2020.

[A10] Marek Śmieja, Maciej Kołomycki, Łukasz Struski, Mateusz Juda, Mario A. T. Figueiredo, *Iterative Imputation of Missing Data using Auto-encoder Dynamics*, International Conference on Neural Information Processing (ICONIP), 258–269, 2020.

Przedstawione osiągnięcie naukowe obejmuje badania w dwóch powiązanych obszarach uczenia maszynowego: algorytmy uczenia pół-nadzorowanego oraz algorytmy uczenia maszynowego dla danych z brakującymi atrybutami.

Algorytmy uczenia pół-nadzorowanego

Tematem przewodnim prac [A1–A5] są algorytmy grupowania dla danych ze słabymi lub niepełnymi etykietami. W klasycznym (nienadzorowanym) zagadnieniu grupowania poszukujemy podziału zbioru obserwacji X (najczęściej punktów w $\mathbb{R}^d$) na rozłączne podzbiory (skupiska), $X_i \subset X$, który minimalizuje zadaną funkcję kosztu. Przykładowo, w algorytmie k-średnich minimalizujemy sumę kwadratów odległości obserwacji do geometrycznych środków skupisk. Alternatywnie, problem grupowania możemy sformułować jako estymację gęstości w modelu mieszanki rozkładów: $p(\mathbf{x} \mid \boldsymbol{\theta}) = \sum_{i=1}^{k} \phi_i p_i(\mathbf{x} \mid \boldsymbol{\theta}_i)$, z parametrem $\boldsymbol{\theta} = \{\phi_1, \boldsymbol{\theta}_1, \ldots \phi_k, \boldsymbol{\theta}_k\}$, $\sum_{i=1}^{k} \phi_i = 1$. Skupiska utożsamiamy wówczas z komponentami $p_i(\mathbf{x} \mid \boldsymbol{\theta}_i)$. W pracach [A1–A5] rozważane są oba warianty problemu grupowania. Zakłada się przy tym, że zbiór obserwacji wzbogacony jest o pewną dodatkową informację, np. zgrubny podział X.

W pracach [A1] i [A2] rozważany jest problem grupowania z dodatkową wiedzą wyrażoną w postaci zbioru więzów wskazujących, że pewne obserwacje należą do tego samego skupiska (relacja *must-link*) lub do różnych skupisk (relacja *cannot-link*). Grupowanie z takimi więzami jest sformułowane w postaci problemu klasyfikacji obserwacji do jednego z określonej liczby skupisk (tzw. grupowanie dyskryminatywne, *discriminative clustering*). W [A1] wykorzystano w tym celu model wieloklasowej regresji logistycznej z funkcją kosztu w postaci oczekiwanej liczby spełnionych więzów *must-link* i *cannot-link*. W przypadku gdy liczba więzów jest niewielka (dla większości obserwacji nie jest znana żadna relacja *must-link* lub *cannot-link*), w modelu wprowadzana jest funkcja wagowa zastępująca relacje binarne. Następnie tworzone są dodatkowe więzy opisujące ε-otoczenie każdej obserwacji. Model ten rozwinięto dalej w pracy [A2], proponując odmienny sposób wzbogacania zbioru więzów pomiędzy obserwacjami. W tym celu wykorzystano tzw. syjamską sieć neuronową[2] uczoną na dostępnych więzach *must link* i *cannot-link*. Brakujące więzy są uzupełnianie poprzez progowanie odpowiedzi takiej sieci. Sam model klasyfikacji został zmodyfikowany poprzez wprowadzenie w regresji logistycznej nieliniowego odwzorowania realizowanego perceptronem wielowarstwowym lub siecią konwolucyjną (w zależności od rodzaju grupowanych obserwacji). Obie prace prezentują eksperymentalną ocenę zaproponowanych modeli

[2]Dla pary obserwacji $(\mathbf{x}, \mathbf{y})$ syjamska sieć neuronowa $s(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta})$ zwraca odległość $\|f(\mathbf{x}; \boldsymbol{\theta}) - f(\mathbf{y}; \boldsymbol{\theta})\|$ pomiędzy ich osadzeniami. Funkcja obliczająca osadzenia $f(\cdot; \boldsymbol{\theta})$ to klasyczna skierowana sieć neuronowa. W procesie uczenia parametr $\boldsymbol{\theta}$ jest dobierany tak, by obserwacje podobne (np. dwa zdjęcia tej samej twarzy) miały bliskie osadzenia, zaś obserwacje niepodobne miały odległe osadzenia.

dla kilku wzorcowych zbiorów danych. Uzyskane wyniki pokazują poprawę jakości grupowania względem metod referencyjnych. W pracy [A2] pokazano również istotną statystycznie poprawę jakości grupowania względem modelu zaproponowanego w [A1].

Praca [A3] podejmuje zagadnienie grupowania z dodatkową wiedzą w sytuacji, gdy skupiska nie dają się opisać prostym rozkładem prawdopodobieństwa, np. wielowymiarowym rozkładem normalnym. Zaproponowane rozwiązanie polega na osobnym grupowaniu klas równoważności w X narzuconych przez dostępne więzy *must-link*. W kolejnym kroku budowane jest końcowe grupowanie X poprzez agregację skupisk zbudowanych dla klas równoważności. W pracy podano warunki konieczne i wystarczające, by agregacja dwóch skupisk nie naruszała więzów *cannot-link*. Sam algorytm przedstawiony jest jako procedura niezależna od sposobu wstępnego grupowania w obrębie klas równoważności. W eksperymentach do grupowania wykorzystano algorytm *Cross-Entropy Clustering*[3] (CEC). W mojej ocenie problem poruszany w [A3] warto byłoby rozważyć z perspektywy estymacji gęstości prawdopodobieństwa. Jeśli bowiem przyjąć, że rozkład obserwacji (w tym przypadku w obrębie klasy równoważności) jest nieznany, to pojawia się pytanie, dla jakich rozkładów nasz estymator (np. mieszanka Gaussowska) jest zgodny? W algorytmie CEC gęstość modelowana jest mieszanką co najwyżej k komponentów Gaussowskich (gdzie k to określona z góry stała, hiper-parametr modelu), przy czym liczba komponentów dobierana jest tak, by zminimalizować oczekiwaną długość kodu obserwacji. Gdyby skorzystać z odpowiedniej mieszanki o skończonej, lecz nieznanej (a priori) liczbie komponentów[4], otrzymamy model zgodny co do liczby komponentów (zakładając, że rozkład obserwacji faktycznie jest mieszanką skończoną i znamy postać rozkładu komponentu). Ciekawym problemem badawczym może być wówczas wplecenie więzów w taki model. O ile więzy *must-link*, a ściślej, indukowane przez nie grupy obserwacji, były wprowadzane w rozkład a priori (w celu zapewnienia identyfikowalności mieszanki[5]), to uwzględnienie więzów *cannot-link* wydaje się ciekawym problemem badawczym.

W pracach [A4] i [A5] rozważany jest problem grupowania, w którym dodatkowa informacja wyrażona jest w postaci zgrubnego podziału obserwacji na podzbiory. Celem algorytmu grupowania jest *rozdrobnienie* podziału zgrubnego. Zakłada się przy tym, że podział zgrubny może być niedokładny. W obu pracach proponowane są modyfikacje algorytmu CEC mające na celu zapewnienie zgodności z podziałem zgrubnym. W pracy [A4] funkcja kosztu w algorytmie CEC została rozbudowana o entropię warunkową $H(\mathcal{Z} \mid \mathcal{Y})$ podziału zgrubnego $\mathcal{Z}$ względem rozdrobnienia $\mathcal{Y}$. Wagę tego dodatkowego członu określa hiper-parametr wprowadzony do modelu. W pracy scharakteryzowano wpływ tego hiper-parametru na końcowy wynik grupowania. W pracy [A5] zgrubny podział wyrażony jest w postaci granicy decyzyjnej rozdzielającej dwie klasy obserwacji. Obserwacje w tej postaci objaśniane są mieszanką Gaussowską, w której zmienna rozdzielająca klasy jest niezależna od pozostałych zmiennych. Następnie formułowany jest warunek zapewniający, że nie więcej niż $\alpha < 0.5$ masy prawdopodobieństwa komponentu przypada na obszar po niewłaściwej stronie granicy decyzyjnej. W przypadku algorytmu CEC narzucone ograniczenie oraz przyjęta postać mieszanki Gaussowskiej prowadzą do dodatkowego członu regularyzującego w funkcji kosztu. Dla modelu ograniczonego do dwóch skupisk (po przeciwnych stronach granicy decyzyjnej) pokazano, że gdy $\alpha \to \infty$ reguła decyzyjna wyznaczona przez różnicę gęstości względem obu komponentów mieszanki zbiega do granicy decyzyjnej. Skuteczność zaproponowanych rozszerzeń algorytmu CEC oceniono w stosunkowo obszernym zestawie eksperymentów. Algorytm zaproponowany w pracy [A4] poprawiał skuteczność grupowania (względem metod referencyjnych) głównie w sytuacji, gdy podział zgrubny był niedokładny. Algorytm

[3] Jacek Tabor i Przemyslaw Spurek. W: *Pattern Recogniton* 47 (2014), s. 3046–3059.

[4] Sylvia. Richardson i Peter J. Green. W: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 59 (1997), s. 731–792.

[5] Hwan Chung, Eric Loken i Joseph L. Schafer. W: *The American Statistician* 2 (2004), s. 152–158.

zaproponowany w pracy [A5] w większości przypadków poprawiał wyniki względem rozwiązań alternatywnych.

W przedstawionym osiągnięciu tematykę uczenia pół-nadzorowanego poruszono również w kontekście tzw. modelowania generatywnego [A6]. Punktem wyjścia jest tu parametryczny model generatywny ze zmienną ukrytą: $\mathbf{z} \sim p(\mathbf{z})$, $\mathbf{x} \sim p(\mathbf{x} \mid \mathbf{z}; \boldsymbol{\theta})$. Jednym z bardziej popularnych modeli tego typu jest autokoder wariacyjny (*varational autoencoder*, VAE). W pracy rozważano model alternatywny, tzw. *Wasserstein Autoencoder*[6] (WAE). Zaproponowano modyfikację tego modelu mającą na celu uwzględnienie klas obserwacji w procesie generatywnym. Założono przy tym, że etykiety klas dostępne są jedynie dla części obserwacji w zbiorze uczącym. Zasadniczy pomysł w [A6] polega na zbudowaniu rozkładu a priori $p(\mathbf{z})$ w postaci mieszanki Gaussowskiej, której komponenty opisują zmienne ukryte dla poszczególnych klas. W funkcji kosztu wprowadzony jest następnie człon dyskryminatywny, w którym klasa obserwacji objaśniana jest przez kod ukryty. Ten człon uwzględniany jest jedynie dla obserwacji, dla których dostępne są etykiety. Pół-nadzorowany model generatywny był uprzednio rozważany dla klasycznych VAE[7]. W [A6] podkreślono jednak, że zaproponowany model operuje na ciągłej przestrzeni ukrytej, w przeciwieństwie do pół-nadzorowanego VAE, w którym etykieta klasy jest zmienną kategoryczną. Pozwala to wykorzystać zaproponowany model do generowania obserwacji będących pewnego rodzaju interpolacjami pomiędzy wyuczonymi klasami. Przedstawione wyniki eksperymentalne potwierdzają poprawę jakości tworzonych w ten sposób próbek w stosunku do modeli referencyjnych. W mojej ocenie uzyskane wyniki dodatkowo wzmocniłoby porównanie do pół-nadzorowanego modelu VAE, w którym klasa opisana jest ciągłą relaksacją zmiennej kategorycznej[8].

Algorytmy dla danych z brakującymi atrybutami

Prace [A7–A10] poświęcone są algorytmom uczenia maszynowego dla obserwacji, w których wartości pewnych atrybutów nie są dostępne. Popularny sposób wnioskowania z takich danych polega na oszacowaniu wartości brakujących atrybutów, biorąc pod uwagę atrybuty dostępne oraz inne obserwacje w zbiorze uczącym. W pracach [A7] i [A8] zaproponowano sposób uwzględnienia niepewności takiego oszacowania w metodach wykorzystujących funkcje jądrowe oraz w sieciach neuronowych. Przyjmijmy, że $[\mathbf{u}, \mathbf{v}]^{\mathsf{T}} \in \mathbb{R}^d$ to obserwacja, w której wartości atrybutów $\mathbf{u}$ są znane, zaś wartości atrybutów $\mathbf{v}$ są niedostępne. Jeśli $\mathbf{u}$, $\mathbf{v}$ mają łącznie wielowymiarowy rozkład normalny, to rozkład warunkowy $p(\mathbf{v} \mid \mathbf{u})$ również jest rozkładem normalnym. W [A7] fakt ten wykorzystano do uogólnienia funkcji jądrowej RBF do przypadku obserwacji z brakującymi atrybutami. Otrzymano funkcję jądrową (w postaci jawnej) będącą wartością oczekiwaną funkcji RBF dla rozkładów warunkowych. Funkcję tą wykorzystano w maszynie wektorów wspierających, uzyskując skuteczny algorytm klasyfikacji dla danych z brakującymi atrybutami. Badania te były kontynuowane w [A8], gdzie zaproponowano wykorzystanie rozkładów warunkowych $p(\mathbf{v} \mid \mathbf{u})$ (w tym przypadku dla mieszanki wielowymiarowych rozkładów Normalnych) do wyznaczenia wartości oczekiwanej funkcji aktywacji na zbiorze brakujących atrybutów. Uzyskano rozwiązania dla funkcji aktywacji ReLU i RBF. Co ważne, pokazano, że jeśli oczekiwane wartości aktywacji ReLU (lub RBF) dla miar probabilistycznych ν, μ są sobie równe (dla każdej wartości parametru funkcji aktywacji), to $\nu = \mu$. W tym sensie wnioskowanie z wartości oczekiwanych nie skutkuje niemożnością rozróżnienia obserwacji. Uzyskane wyniki eksperymentalne świadczą

[6] Ilya O. Tolstikhin i in. W: *6th International Conference on Learning Representations, (ICLR)*. 2018.

[7] Diederik P. Kingma i in. W: *Advances in Neural Information Processing Systems 27 (NIPS)*. 2014, s. 3581–3589.

[8] Na przykład rozkładem Gumbel-Softmax: Eric Jang, Shixiang Gu i Ben Poole. W: *5th International Conference on Learning Representations, ICLR*. arXiv preprint arXiv:1611.01144v5. 2017.

o skuteczności algorytmów zaproponowanych w [A7] i [A8].

W pracach [A9] i [A10] zajmowano się algorytmami uczenia maszynowego dla częściowo maskowanych obrazów. W [A9] wykorzystano przestrzenną grafową sieć konwolucyjną do klasyfikacji takich obrazów oraz rekonstrukcji maskowanego obszaru. Uzyskano poprawę wyników w stosunku do podstawowej grafowej sieci konwolucyjnej oraz kilku rozwiązań wykorzystujących klasyczne sieci konwolucyjne. W [A10] wykorzystano znane oszacowanie gradientu log-wiarygodności w modelu generatywnym *Denoising Autoencoder* do konstrukcji algorytmu proponującego wypełnienie maskowanej część obrazu. Algorytm konstruuje wypełnienie tak, by (lokalnie) zmaksymalizować log-wiarygodność. Lokalne maksimum jest wyznaczane metodą gradientu prostego. Uzyskano poprawę jakości wypełnień w stosunku do kilku metod referencyjnych.

Ocena osiągnięcia naukowego

W przedstawionym osiągnięciu naukowym zaproponowano szereg algorytmów uczenia maszynowego dla danych o niepełnych (lub słabych) etykietach oraz danych z brakującymi atrybutami. Za szczególnie cenny uważam algorytm przedstawiony w pracy [A8]. Zapewnia on uwzględnienie w odpowiedzi modelu nie tylko oszacowanej wartości brakującego atrybutu lecz również niepewności jego oszacowania. Choć część wyników przedstawiono dla wybranych funkcji aktywacji, to jest wśród niech funkcja ReLU – obecnie najpowszechniej używana funkcja aktywacji w sieciach neuronowych. Praca została opublikowana na konferencji NeurIPS, jednej z najbardziej prestiżowych konferencji poświęconych uczeniu maszynowemu (ranga A* według rankingu CORE). Cenny jest również analogiczny wynik dla algorytmów wykorzystujących funkcje jądrowe [A7]. Obie prace są bardzo dobrze cytowane[9]. Wysoko oceniam rezultaty uzyskane dla pół-nadzorowanych algorytmów grupowania. Cenne są w szczególności rozszerzenia algorytmu CEC na przypadek danych ze zgrubnym podziałem obserwacji ([A4] i [A5]), jak również algorytmy grupowania dyskryminatywnego dla danych z więzami ([A1], [A2]). We wszystkich pracach wymienionych w osiągnięciu Autor dokłada starań, by rezultaty teoretyczne poprzeć przekonującą oceną eksperymentalną na tle istniejących już rozwiązań. Wyniki eksperymentalne zaprezentowane są w sposób umożliwiający ich niezależną replikację.

Rezultaty przedstawione w osiągnięciu naukowym zostały opublikowane w cenionych czasopismach poświęconych tematyce uczenia maszynowego oraz dobrych lub bardzo dobrych konferencjach naukowych. W każdym przypadku Habilitant precyzyjne zakreślił swój wkład w osiągnięcie i był to najczęściej wkład wiodący.

Podsumowując, rezultaty uzyskane w osiągnięciu naukowym oceniam jako dobre lub bardzo dobre. W mojej opinii stanowią one znaczny wkład w rozwój dyscypliny naukowej Informatyka.

2 Pozostała aktywność naukowa

W latach 2018-2019 Habilitant był na trzymiesięcznym stażu podoktorskim w Uniwersytecie Lizbońskim oraz trzymiesięcznym stażu w Instytucie Podstaw Informatyki PAN. Efektem obu staży były publikacje naukowe. Dwie z tych publikacji, powstałe we współpracy z prof. M. Figueiredo, stanowią część zgłoszonego osiągnięcia naukowego (pozycje [A2] i [A10]). Habilitant spędził również kilka dni w TU Graz oraz Politechnice Wrocławskiej. Efektem pierwszej wizyty była publikacja [A6]. Poza pracami wymienionymi w osiągnięciu, Habilitant opublikował po dok-

[9]Liczba cytowań pracy [A8] jest obecnie kilkukrotnie większa od wartości podanej w dokumentacji wniosku.

toracie 18 prac naukowych w cenionych czasopismach i konferencjach naukowych. Publikacje te dotyczą uczenia głębokiego oraz algorytmów grupowania dla obserwacji o dużej liczbie wymiarów i obserwacji opisanych rzadkimi wektorami binarnymi. Na uwagę zasługują również publikacje dotyczące zastosowania metod uczenia maszynowego w chemoinformatyce. Wysoko oceniam aktywność Habilitanta w realizacji projektów naukowych – Habilitant kierował trzema projektami NCN (Preludiom, Sonata, Opus) i jednym projektem MNiSW. Aktywność naukową habilitanta potwierdza również udział w komitetach programowych wiodących konferencji z obszaru uczenia maszynowego (ICML, NeurIPS) i sztucznej inteligencji (AAAI). Na koniec warto wspomnieć, że Habilitant był promotorem 10 prac magisterskich. Spośród dyplomantów, którzy przygotowali prace magisterskie pod opieką Habilitanta troje kontynuuje pracę naukową w ramach studiów doktoranckich.

W podsumowaniu, pozostałą aktywność naukową Habilitanta oceniam wysoko. Uważam, że spełnia ona ustawowe warunki nadania stopnia doktora habilitowanego.

3 Konkluzja

W mojej ocenie osiągnięcie naukowe oraz pozostała aktywność naukowa Habilitanta spełniają warunki uzyskania stopnia doktora habilitowanego wymienione w Art. 219 Ustawy *Prawo o szkolnictwie wyższym i nauce*. Popieram nadanie doktorowi Markowi Śmieji stopnia doktora habilitowanego w dyscyplinie Informatyka.

Z poważaniem,
Marcin Kurdziel