Dr inż. Maciej Zięba
Politechnika Wrocławska
Wydział Informatyki i Zarządzania

AUTOREFERAT

1. **Imię i Nazwisko:** Maciej Zięba
2. **Posiadane dyplomy, stopnie naukowe/ artystyczne – z podaniem nazwy, miejsca i roku ich uzyskania oraz tytułu rozprawy doktorskiej:**

   Uzyskany tytuł: *Doktor nauk technicznych w dyscyplinie informatyka*
   Rok uzyskania tytułu: *2014 (praca złożona do recenzji w czerwcu 2013)*
   Miejsce: *Instytut Informatyki, Politechnika Wrocławska.*
   Tytuł rozprawy: *Zespoły klasyfikatorów SVM dla danych niezbalansowanych.*
   Promotor: *prof. dr hab. inż. Jerzy Świątek.*
   Uwagi: *Rozprawa doktorska została wyróżniona.*

   Uzyskany tytuł: *Mgr w Naukach o Zarządzaniu*
   Rok uzyskania tytułu: *2011*
   Miejsce: *Instytut Organizacji i Zarządzania, Politechnika Wrocławska.*
   Tytuł Pracy: *Ensemble decision trees for customer classification in service oriented systems.*
   Promotor: *dr inż. Barbara Gładysz.*
   Uwagi: *Studia magisterskie ukończone z wynikiem bardzo dobrym.*
   Specjalność: *Technologie Informacyjne w Zarządzaniu*

   Uzyskany tytuł: *MSc in Computer Science*
   Rok uzyskania tytułu: *2010*
   Miejsce: *Blekinge Institute of Technology, Szwecja.*
   Tytuł pracy: *Multistage neural networks for pattern recognition.*
   Promotor: *Prof. Ludwik Kuźniarz.*
   Uwagi: *Praca napisana w ramach programu Double Diploma Programme.*
   Specjalność: *Computer Science.*

   Uzyskany tytuł: *Mgr inż. Informatyki*
   Rok uzyskania tytułu: *2009*
   Miejsce: *Instytut Informatyki, Politechnika Wrocławska.*
   Tytuł Pracy: *Multistage neural networks for pattern recognition.*
   Promotor: *prof. dr hab. inż. Jerzy Świątek.*
   Uwagi: *Studia magisterskie ukończone z wynikiem bardzo dobrym.*
   Specjalność: *Podstawowe Problemy Informatyki.*

3. **Informacje o dotychczasowym zatrudnieniu w jednostkach naukowych/ artystycznych.**

Okres: *11/2019 – obecnie*
Stanowisko: *Wykładowca - studia podyplomowe Biznes.AI*
Miejsce: *Akademia Leona Koźmińskiego, Warszawa.*

Okres: *9/2016 – 3/2017*
Stanowisko: *Post doc*
Miejsce: *University of Wollongong, Australia.*

Okres: *2016 – obecnie*
Stanowisko: *Adiunkt*
Miejsce: *Wydział Informatyki i Zarządzania, Politechnika Wrocławska.*

Okres: *2015 - 2016*
Stanowisko: *Wykładowca*
Miejsce: *Wyższa Szkoła Horyzont, Wrocław.*

Okres: *2013 – 2016*
Stanowisko: *Asystent naukowo-dydaktyczny*
Miejsce: *Wydział Informatyki i Zarządzania, Politechnika Wrocławska.*

Okres: *2012 – 2013*
Stanowisko: *Asystent naukowy*
Miejsce: *Wydział Informatyki i Zarządzania, Politechnika Wrocławska.*
Uwagi: *Praca w ramach projektu „Platforma optymalizacji procesów biznesowych w zintegrowanych systemach informacyjnych”.*

Okres: *2009 – 2012*
Stanowisko: *Samodzielny informatyk*
Miejsce: *Wydział Informatyki i Zarządzania, Politechnika Wrocławska.*
Uwagi: *Praca w ramach projektu „Nowe technologie innowacyjne dla elektronicznej gospodarki i społeczeństwa informacyjnego oparte na paradygmacie SOA”.*

4. **Wskazanie osiągnięcia wynikającego z art. 16 ust. 2 ustawy z dnia 14 marca 2003 r. o stopniach naukowych i tytule naukowym oraz o stopniach i tytule w zakresie sztuki (Dz. U. 2017 r. poz. 1789)**

A. Tytuł osiągnięcia naukowego

Metody i algorytmy rozwiązujące podstawowe problemy z jakością i strukturą danych w zadaniach uczenia maszynowego

B. Cykl publikacji powiązanych tematycznie

[P1] **Zięba, Maciej.** *"Service-oriented medical system for supporting decisions with missing and imbalanced data."* IEEE journal of biomedical and health informatics 18.5 (2014): 1533-1540.
Punktacja MNiSW (rok publikacji): 30.
IF (rok publikacji): 1.440.
Punktacja MNiSW (rok 2020): 100
Liczba cytowań (Google Scholar): 19

[P2] **Zieba, Maciej**, Piotr Klukowski, Adam Gonczarek, Yaroslav Nikolaev, and Michal J. Walczak *"Gaussian process regression for automated signal tracking in step-wise perturbed Nuclear Magnetic Resonance spectra."* Applied Soft Computing 68 (2018): 162-171.
Punktacja MNiSW (rok publikacji): 40.
IF (rok publikacji): 4.873
Punktacja MNiSW (rok 2020): 200
Liczba cytowań (Google Scholar): 3

Wkład Autorów w powstanie pracy (wkład z podziałem na zadania - *Załącznik 1*):

- Maciej Zięba (40%)
- Piotr Klukowski (30%)
- Adam Gonczarek (5%)
- Yaroslav Nikolaev (15%)
- Michał J. Walczak (10%)

[P3] **Zięba, Maciej**, Jakub M. Tomczak, and Adam Gonczarek. *"RBM-SMOTE: restricted Boltzmann machines for synthetic minority oversampling technique."* Asian Conference on Intelligent Information and Database Systems. Springer, Cham, 2015.
Punktacja MNiSW (rok 2020): 20
Liczba cytowań (Google Scholar): 11

Wkład Autorów w powstanie pracy (wkład z podziałem na zadania - Załącznik 2):

- Maciej Zięba (80%)
- Jakub M. Tomczak (10%)
- Adam Gonczarek (10%)

[P4] Tomczak, Jakub M., and **Maciej Zięba**. *"Probabilistic combination of classification rules and its application to medical diagnosis."* Machine Learning 101.1-3 (2015): 105-135.
Punktacja MNiSW (rok publikacji): 35.
IF (rok publikacji): 1.719
Punktacja MNiSW (rok 2020): 140
Liczba cytowań (Google Scholar): 19

Wkład Autorów w powstanie pracy (wkład z podziałem na zadania - Załącznik 3):

- Jakub M. Tomczak (70%)
- Maciej Zięba (30%)

[P5] **Zięba, Maciej**, Jakub M. Tomczak, and Jerzy Świątek. "*Self-paced learning for imbalanced data.*" Asian Conference on Intelligent Information and Database Systems. Springer, Berlin, Heidelberg, 2016.
Punktacja MNiSW (rok 2020): 20
Liczba cytowań (Google Scholar): 1

Wkład Autorów w powstanie pracy (wkład z podziałem na zadania - Załącznik 4):

- Maciej Zięba (80%)
- Jakub M. Tomczak (10%)
- Jerzy Świątek (10%)

[P6] **Zieba, Maciej**, and Wolfgang Karl Härdle. *"Beta-boosted ensemble for big credit scoring data."* Handbook of Big Data Analytics. Springer, Cham, 2018. 523-538.
Liczba cytowań (Google Scholar): 3

Wkład Autorów w powstanie pracy (wkład z podziałem na zadania - Załącznik 5):

- Maciej Zięba (85%)
- Wolfgang Karl Härdle (15%)

[P7] Tomczak, Jakub M., and **Maciej Zięba**. *"Classification restricted Boltzmann machine for comprehensible credit scoring model."* Expert Systems with Applications 42.4 (2015): 1789-1796.

Punktacja MNiSW (rok publikacji): 35
IF (rok publikacji): 2.981
Punktacja MNiSW (rok 2020): 140
Liczba cytowań (Google Scholar): 56

Wkład Autorów w powstanie pracy (wkład z podziałem na zadania - Załącznik 6):

- Jakub M. Tomczak (60%)
- Maciej Zięba (40%)

[P8] **Zięba, Maciej**, Sebastian K. Tomczak, and Jakub M. Tomczak. *"Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction."* Expert Systems with Applications 58 (2016): 93-101.
Punktacja MNiSW (rok publikacji): 35.
IF: 3.928
Punktacja MNiSW (rok 2020): 140
Liczba cytowań (Google Scholar): 151

Wkład Autorów w powstanie pracy (wkład z podziałem na zadania - Załącznik 7):

- Maciej Zięba (55%)
- Sebastian K. Tomczak (25%)
- Jakub M. Tomczak (20%)

[P9] **Zieba, Maciej**, Lei Wang. *"Training triplet networks with GAN."* International Conference on Learning Representations ICLR (workshop track). 2017.
Liczba cytowań (Google Scholar): 12

Wkład Autorów w powstanie pracy (wkład z podziałem na zadania - Załącznik 8):

- Maciej Zięba (85%)
- Lei Wang (15%)

[P10] **Zieba, Maciej**, Piotr Semberecki, Tarek El-Gaaly, and Tomasz Trzcinski. *"Bingan: Learning compact binary descriptors with a regularized GAN."* Advances in Neural Information Processing Systems. 2018.
Punktacja MNiSW (rok 2020): 200
CORE 2018 Rating: A*
Google Scholar H5-index: 169
Liczba cytowań (Google Scholar): 16

Wkład Autorów w powstanie pracy (wkład z podziałem na zadania - Załącznik 9):

- Maciej Zięba (75%)
- Piotr Semberecki (10%)
- Tarek El-Gaaly (5%)
- Tomasz Trzcinski (10 %)

[P11] Maciej Zamorski, **Maciej M. Zięba,** Piotr J. Klukowski, Rafał A. Nowak, Karol Kurach, Wojciech Stokowiec, Tomasz Trzciński "*Adversarial autoencoders for compact representations of 3D point clouds"*. Computer Vision and Image Understanding. 2020. vol. 193, art. nr 102921, s. 1-8.
IF: 2.645
Punktacja MNiSW (rok 2020): 100
Liczba cytowań (Google Scholar): 15

Wkład Autorów w powstanie pracy (wkład z podziałem na zadania - Załącznik 10):

- Maciej Zamorski (35%)
- Maciej Zięba (35%)
- Piotr Klukowski (5%)
- Rafał Nowak (5%)
- Karol Kurach (5%)
- Wojciech Stokowiec (5%)
- Tomasz Trzciński (10%)

Przedstawiony do oceny cykl publikacji obejmuje **11** pozycji, w tym **6** artykułów w czasopismach z **IF** (sumaryczny **IF** równy **17.586**), **3** referaty w materiałach konferencyjnych (wszystkie konferencje są punktowane na liście **MNiSW** z roku 2020**,** w tym jedna z prac została opublikowana na konferencji **NeurIPS,** punktowanej za **200 pkt.**), recenzowana publikacja na konferencji ICLR (workshop track) i jeden rozdział w książce. Suma punktów przedstawionych publikacji zgodna z listą **MNiSW** z roku 2020 wynosi **1060 pkt,** przy czym dwie publikacje osiągnęły maksymalną liczbę równą **200 pkt.** Łączna liczba cytowań dla wykazanych prac jest równa **306.**

C. Omówienie celu naukowego prac i osiągniętych wyników wraz z omówieniem ich ewentualnego wykorzystania

1. Wprowadzenie

We współczesnych czasach obserwujemy niepodważalny rozwój metod sztucznej inteligencji, głównie metod uczenia maszynowego. Rozwój ten możliwy jest dzięki coraz większej ilości danych zagregowanych w sposób cyfrowy, co w połączeniu z coraz większymi możliwościami obliczeniowymi umożliwia konstrukcję coraz bardziej zaawansowanych modeli uczenia maszynowego znajdujących zastosowanie w różnych

obszarach, takich jak diagnostyka medyczna, analiza ryzyka finansowego, rozpoznawanie, generowanie, segmentacja obrazów, czy też przetwarzanie języka naturalnego. Wzrost dostępności potencjalnych danych uczących z jednej strony niesie za sobą ogromne możliwości w konstrukcji coraz bardziej zaawansowanych rozwiązań AI, z drugiej jednak strony rzeczywiste dane mogą charakteryzować się szeregiem problemów, które uniemożliwiają bezpośrednie ich zastosowanie w procesie konstrukcji modelu uczenia maszynowego. Tego typu problemy mogą wynikać z niekompletności, błędnych wartości, niezbalansowania w ramach określonych klas, czy też samej struktury danych która utrudnia, bądź uniemożliwia bezpośrednie stosowanie znanych algorytmów sztucznej inteligencji. W swojej pracy badawczej koncentruję się na rozwiązywaniu problemów z danymi uczącymi na potrzeby budowy efektywniejszych modeli dla różnych obszarów uczenia maszynowego.

2. Cel i zakres prowadzonych badań

Celem prac było opracowanie szeregu metod rozwiązujących współczesne problemy z danymi w zadaniach uczenia maszynowego. W ramach swojej pracy badawczej wyodrębniłem, i podjąłem się rozwiązania następujących problemów z danymi:

- niekompletności danych uczących [P1, P2].
- niezbalansowania danych uczących [P3, P4, P5, P6].
- struktury danych uczących [P7, P8, P9, P10, P11]

Pierwszy z wymienionych problemów dotyczy sytuacji, w której dane zawarte w zbiorze uczącym zawierają brakujące wartości w obrębie pewnej grupy przykładów. W przypadku danych w postaci tabelarycznej pewne wartości atrybutów dla pewnych obiektów w zbiorze mogą być nieznane na poziomie zebranych danych treningowych. Brak wartości pewnych atrybutów może wynikać z ograniczonej możliwości wykonania pomiarów pewnych wartości cech obiektów, z pominięcia pewnych atrybutów w procesie cyfryzacji danych, rozszerzenia zakresu danych podczas agregacji, konieczności anonimizacji pewnych danych, oraz innych czynników. Problem brakujących danych może wystąpić również w przypadku przypadku obrazów, tekstu, dźwięku, oraz innego typu danych sekwencyjnych.

Drugi z wymienionych problemów odnosi się do zagadnień uczenia nadzorowanego, głównie zadań klasyfikacji, i charakteryzuje się nierównomiernym rozkładem klas w obrębie zbioru treningowego. Wyuczenie modelu predykcyjnego na tego typu danych uczących może skutkować tym, że model będzie obciążony w kierunku klas dominujących z zbiorze, czyli będzie miał tendencję do faworyzowania klas wiodących, a w ekstremalnych przypadkach może klasyfikować wszystkie obiekty testowe do najliczniejszych klas.

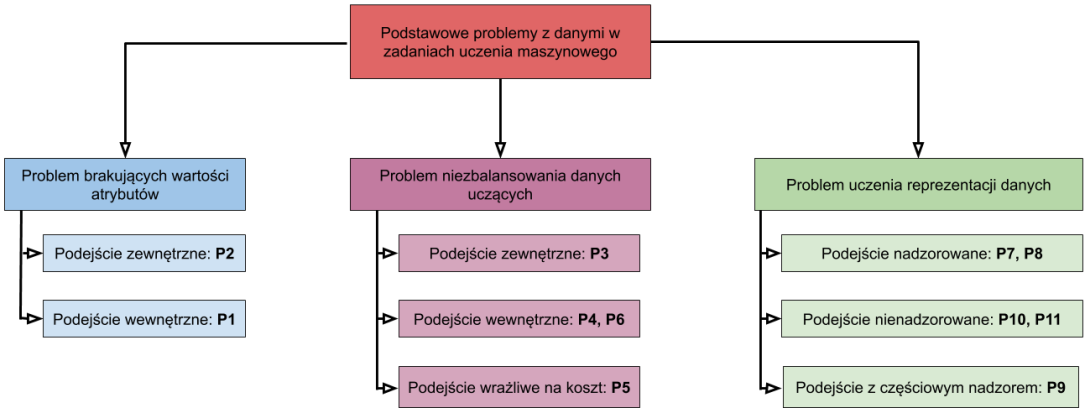
Trzeci problem odnosi się do przypadków, w których reprezentacja danych uczących nie pozwala bezpośrednio na efektywne stosowanie metod dla określonych zadań uczenia

maszynowego. W takich przypadkach konieczne jest opracowanie algorytmów, które pozwolą na przejście z przestrzeni danych na przestrzeń ukrytą (przestrzeń cech) która umożliwi efektywne wykonywanie zadań związanych z klasyfikacją, wyszukiwaniem informacji, czy też grupowaniem. Problem ten w literaturze, w zależności od typu danych i zastosowania, nazywamy ekstrakcją cech, uczeniem reprezentacji, bądź kodowaniem.

Zakres prowadzonych badań obejmuje algorytmy rozwiązujące przedstawione problemy z danymi:

- brakujących wartości atrybutów:
    - opracowanie autorskiego algorytmu klasyfikacyjnego dla danych niezbalansowanych wykorzystującego macierz koincydencji **[P1]**;
    - opracowanie modelu uzupełniania elementów w danych sekwencyjnych z wykorzystaniem procesów Gaussa **[P2]**;
- niezbalansowania danych uczących:
    - opracowanie modyfikacji algorytmu SMOTE na potrzeby generowania obserwacji syntetycznych celem balansowania zbioru uczącego i poprawiającego jakość generowanych próbek z wykorzystaniem ograniczonych maszyn Boltzmanna **[P3]**;
    - opracowanie mechanizmów estymacji rozkładu a priori na klasę pozwalającego na stosowanie reguł miękkich dla problemów o niezbalansowanych danych uczących **[P4]**;
    - opracowanie wariantu algorytmu uczenia we własnym tempie dla danych niezbalansowanych **[P5]**;
    - opracowanie autorskiej metody konstrukcji zespołu modeli klasyfikacyjnych dla danych niezbalansowanych **[P6]**;
- uczenia reprezentacji danych:
    - udział w opracowaniu interpretowanej reprezentacji w postaci tablicy scoringowej na podstawie ograniczonej maszyny Boltzmanna **[P7]**.
    - opracowanie ewolucyjnego podejścia do generowania syntetycznych cech **[P8]**.
    - opracowanie modelu uczenia się reprezentacji w trybie częściowego nadzoru z wykorzystaniem modeli generatywnych i sieci typu triplet **[P9]**;
    - opracowanie algorytmu uczenia reprezentacji danych w trybie nienadzorowanym z wykorzystaniem modeli generatywnych **[P10]**.
    - opracowanie algorytmu uczenia reprezentacji danych w trybie nienadzorowanym dla danych reprezentowanych przez chmury punktów **[P11]**;

Na Rysunku 1 przedstawione zostało powiązanie tematyczne cyklu publikacji w zależności od problemu, który podejmuje się rozwiązać w ramach danej pracy. W następnych rozdziałach pragnę opisać bardziej szczegółowo przedstawione rozwiązania.



Rysunek 1. Schemat powiązania tematycznego publikacji [**P1-P11**] w zależności od poruszanego problemu.

3. Problem brakujących wartości atrybutów

W literaturze istnieje wiele podejść które rozwiązują problem brakujących wartości danych [1]. Podejścia te można zaklasyfikować do dwóch grup: metod wewnętrznych, które dostosowują proces uczenia modelu bo niekompletnych danych, oraz algorytmów zewnętrznych które sprowadzają niekompletne dane do kompletnej postaci, stosując techniki uzupełniania brakujących atrybutów, bądź też eliminując atrybuty bądź rekordy z brakującymi wartościami. W ramach osiągnięcia badawczego opracowałem dwie metody eliminacji brakujących wartości atrybutów [**P1, P2]**. Pierwsza z opracowanych metod **[P1]** jest metodą wewnętrzną i zakłada konstrukcję zespołu modeli klasyfikacyjnych poprzez wykorzystanie pełnej informacji dostarczonej w ramach zbioru uczącego. Drugie z opracowanych rozwiązań **[P2]** stanowi podejście zewnętrzne, w ramach którego brakujące wartości atrybutów uzupełnianie są poprzez zastosowanie procesów Gaussa.

Istotą pierwszego z opracowanych algorytmów **[P1]** jest zastosowanie efektywnej metody podziału zbioru uczącego z brakującymi wartościami na kompletne podzbiory, które następnie były wykorzystane do wyuczenia zespołu klasyfikatorów bazowych. Podzbiory do wyuczenia poszczególnych modeli generowane są na podstawie autorskiego algorytmu wykorzystującego macierz koincydencji, który w swoim działaniu zakłada pewne zależności pomiędzy występowaniem brakujących wartości w ramach podgrup atrybutów. Zadaniem opracowanego podejścia była identyfikacja podgrup atrybutów celem dostarczenia kompletnych podzbiorów danych do wyuczenia modeli bazowych. Wyuczony zespół klasyfikatorów bazowych wykonywał finalną klasyfikację w ramach procedury ważonego głosowania. Cechą charakterystyczną opracowanego rozwiązania jest to, że działa ono niezależne od stosowanego modelu klasyfikatora, oraz stara się wydobyć jak najwięcej informacji z niekompletnych danych proponując

różne warianty podziału wejściowego zbioru uczącego. Opracowane rozwiązanie zostało wdrożone jako komponent zaproponowanego przeze mnie, działającego w oparciu o usługi webowe, systemu agregującego narzędzia uczenia maszynowego na potrzeby zadań analizy, przetwarzania i predykcji danych. Jakość opracowanej metody została przebadana na rzeczywistym zbiorze danych z obszaru diagnostyki medycznej.

Kolejnym rozwiązaniem elementem wchodzącym w skład osiągnięcia badawczego jest metoda uzupełniania brakujących wartości dla danych sekwencyjnych **[P2]**. W ramach tego podejścia zaproponowana została metoda przewidywania kolejnego elementu w sekwencji za pomocą procesu Gaussa. Dzięki zastosowaniu tego typu modelu możliwe było nie tylko określenie najbardziej prawdopodobnego położenia następnego elementu sekwencji, ale również określenie jego niepewności. Opracowane rozwiązanie zostało zastosowane do problemu śledzenia sygnałów NMR, gdzie konieczne było określenie położeń niewykrytych sygnałów w widmie, jednak może być z powodzeniem stosowana do innych problemów uzupełniania brakujących wartości dla atrybutów sekwencyjnych. Jakość opracowanego rozwiązania została poddana badaniom ilościowym i jakościowym, które w szegółach zostały opisane w pracy **[P2]**.

4. Problem niezbalansowanego zbioru uczącego

Problem zakłada dysproporcje między liczebnościami obiektów należących do różnych klas, który negatywnie wpływa na jakość procesu uczenia, skutkując modelem klasyfikacyjnym obciążonym decyzjami w kierunku klasy dominującej. W literaturze istnieje wiele podejść do rozwiązania problemu dysproporcji w danych uczących, które dzielimy na zewnętrzne (niezależne od modelu), oraz wewnętrzne, wprowadzające mechanizmy eliminujące problem niezbalansowania na poziomie modelu [2]. Typowym podejściem należącym do pierwszej grupy jest algorytm SMOTE (*ang. Synthetic Minority Over-sampling Technique*) [3], którego istota działania polega na generowaniu syntetycznych obserwacji celem zwiększenia liczebności obserwacji z nielicznie reprezentowanych klas. Syntetyczne próbki generowane są w losowym miejscu na odcinku łączącym dwie obserwacje z klasy zdominowanej. Zasadniczym ograniczeniem algorytmu SMOTE jest to, że wygenerowany obiekt znajdujący się na odcinku może nie reprezentować próbki z rozkładu danych. Przykładowo, zdjęcie będące wynikiem interpolacji dwóch innych zdjęć w przestrzeni danych nie gwarantuje, że tak zbudowany obraz przedstawia sensowny obiekt. W ramach osiągnięcia badawczego opracowałem modyfikację klasycznego algorytmu SMOTE, która eliminuje to ograniczenie **[P3]**. Istotą opracowanego rozwiązania było zastosowanie Ograniczonej Maszyny Boltzmanna (*ang. Restricted Boltzmann Machine, RBM*) [4] celem zakodowania przykładów z klasy zdominowanej do przestrzeni ukrytej w której następnie generowane były syntetyczne reprezentacje z wykorzystaniem algorytmu SMOTE. Wygenerowane syntetyczne reprezentacje były następnie transformowane do przestrzeni danych poprzez odwrotne zastosowanie modelu RBM. W konsekwencji, dzięki zastosowaniu algorytmu SMOTE w przestrzeni ukrytej, wygenerowane syntetyczne próbki reprezentowały rzeczywisty rozkład danych. Było to możliwe, gdyż że rozkład w przestrzeni ukrytej był dobrany w

ten sposób, aby w wyniku interpolacji otrzymane reprezentacje ukryte również pochodziły z założonego rozkładu prawdopodobieństwa. Opisane podejście zostało poddane analizie ilościowej i jakościowej na zbiorze danych z obszaru oceny ryzyka kredytowego **[P3]**. Warto nadmienić, iż opracowana metoda została zacytowana i opisana w pracy podsumowującej 15 lat rozwoju algorytmu SMOTE [5].

W ramach osiągnięcia badawczego opracowałem, bądź brałem udział w opracowaniu algorytmów rozwiązujących problem niezbalansowania danych na poziomie konstrukcji modelu predykcyjnego. W ramach pracy **[P4]** uczestniczyłem w opracowaniu algorytmu klasyfikacyjnego wykorzystującego *tzw. miękkie reguły decyzyjne*, które mogą być interpretowane jako probabilistyczne rozszerzenie klasycznych reguł klasyfikacyjnych. Istotą opracowanego rozwiązania jest wprowadzenie zmiennych ukrytych reprezentujących cechy koniunkcyjne, względem których obiekt i jego klasa stają się warunkowo niezależne. Dzięki wprowadzeniu takich mechanizmów możliwe było zastosowanie w modelu mechanizmów wnioskowania w oparciu o regułę Bayesa. Problem niezbalansowania danych został rozwiązany w metodzie poprzez zaproponowanie odpowiedniego estymatora prawdopodobieństwa klasy pod warunkiem cechy koniunkcyjnych uwzględniającego liczności obiektów należących do różnych klas. Jakość opracowanej metody została sprawdzona na zestawie 37 zbiorów referencyjnych pozyskanych z repozytorium UCI, oraz na 8 zbiorach danych z obszaru diagnostyki medycznej. Biorąc pod uwagę typowe wskaźniki od oceny metod klasyfikacji, takie jak średnia pod krzywą ROC (*AUC*), oraz średnią geometryczną pomiędzy czułością i specyficznością (*GMean*) opracowana metoda charakteryzuje się jakością działania na porównywalnym poziomie w stosunku do metod referencyjnych wykorzystujących w swoim działaniu zespoły klasyfikatorów, natomiast osiągnęła wyższą jakość niż referencyjne podejścia bazujące na regułach decyzyjnych.

Kolejnym algorytmem eliminującym problem niezbalansowania danych uczących opracowanym przeze mnie w ramach osiągnięcia badawczego był wariant metody uczenia się we własnym tempie (*ang. self-paced learning*) **[P5]**. Istotą uczenia we własnym tempie [6] jest wybór w pierwszej kolejności obserwacji prostych, sięgając po trudniejsze przykłady na dalszym etapie uczenia. Trudność obserwacji oceniana jest na bieżąco przez algorytm uczący, co odróżnia to podejście od uczenia tzw. programem nauczania (*ang. curriculum learning*) [7] w przypadku którego trudność obserwacji oceniana jest przed uruchomieniem procesu uczenia. W opracowanej przeze mnie metodzie trudność obserwacji jest oceniana na podstawie analizy wartości funkcji straty dla rozpatrywanego przykładu. Dodatkowo, celem balansowania danych uczących wprowadzono zmodyfikowane kryterium uczenia modelu we własnym tempie, uwzględniające niesymetryczną macierz kosztów. Dzięki zastosowaniu modyfikacji kryterium wrażliwego na koszt błędnej klasyfikacji możliwe było odpowiednie dostosowanie tempa uczenia obserwacji z klasy dominującej i zdominowanej. Jakość opracowanej metody została przebadana na zestawie 32 zbiorów referencyjnych, przy założeniu że bazowym modelem stosowanym w eksperymencie była sieć neuronowa,

uczona z wykorzystaniem kryterium w postaci entropii krzyżowej. W eksperymencie wykazano, iż zastosowanie wrażliwego na koszt podejścia do uczenia we własnym tempie w istotny sposób zwiększa jakość budowanego modelu w stosunku do modeli referencyjnych budowanych metodą ablacyjną (*ang. ablation studies*) biorąc pod uwagę AUC jako główne kryterium uczenia.

Następnym algorytmem eliminującym problem niezbalansowania danych na poziomie konstrukcji modelu i opracowanym przeze mnie w ramach osiągnięcia badawczego jest model *BetaBoost* **[P6]**. Jest to autorska metoda działająca w oparciu o zespoły klasyfikatorów wykorzystująca specyficzną metodę próbkowania danych bazującą na rozkładzie Beta-dwumianowym (*ang. Beta-binomial*). W każdej iteracji konstrukcji nowego modelu bazowego wchodzącego w skład zespołu klasyfikatorów obserwacje ze zbioru treningowego dla każdej z klas z osobna są sortowane względem rosnącego prawdopodobieństwa wystąpienia klasy pozytywnej. Następnie, dla każdej z rozpatrywanej klas próbkowania jest równa liczba obserwacji z rozkładu Beta-dwumianowego o zadanych (dla każdej klasy z osobna) wartościach parametrów, które są wykorzystywane do budowy nowego modelu bazowego. Parametry dla rozkładów dobierane są w zależności od przyjętej jednej z dwóch strategii. Pierwsza strategia zakłada silny stabilny model klasyfikacyjny i niewielkie zaufanie do jakości danych uczących. W przypadku drugiej strategii zakładamy słaby, niestabilny model bazowy (*ang. weak learner*) i duże zaufanie do jakości danych uczących. W ramach pierwszej strategii dobieramy tak parametry rozkładu aby z wyższym prawdopodobieństwem wybierać obserwacje z klasy pozytywnej znajdujące się wysoko w rankingu, natomiast, w przypadku obserwacji z klasy negatywnej wybierać z wyższym prawdopodobieństwem te obserwacje, które znajdują się nisko po posortowaniu względem prawdopodobieństwa wystąpienia klasy pozytywnej. W konsekwencji, do budowy kolejnych modeli będą wykorzystywane te obserwacje, które znajdują się wysoko (dla klasy pozytywnej), bądź nisko (dla klasy negatywnej) w konstruowanych rankingach. Strategia ta, zakładając brak zaufania do danych, mniejszą wagę przywiązuje do obserwacji ciężkich (*ang. hard examples*), których niekorzystna lokalizacja w rankingu może wynikać ze złej jakości próbek wynikających z obserwacji odstających, bądź przekłamanych wartości. W przypadku drugiej strategii z wyższym prawdopodobieństwem wybierane są przykłady znajdujące się w niekorzystnych lokalizacjach i w konsekwencji uruchamiana jest technika tzw. wiązania ciężkich przykładów (*ang. hard mining*), która polega na wybieraniu tych przykładów, z którymi model potencjalnie miał najwięcej problemów z poprawną klasyfikacją.

Pierwsza ze strategii została przetestowana z wykorzystaniem modelu bazowego w postaci regresji logistycznej, natomiast druga z wykorzystaniem prostego drzewa decyzyjnego. Biorąc pod uwagę dwa duże zbiory z obszaru oceny ryzyka kredytowego, oraz kryterium oceny jakości w postaci AUC, w badaniach ilościowych wykazano że pierwsza z wymienionych strategii istotnie poprawiła jakość działania zespołów klasyfikatorów z modelem bazowym w postaci regresji logistycznej, natomiast druga

strategia przyczyniła się do poprawy jakości, jeżeli modelem bazowym było proste drzewo decyzyjne. Jako modele referencyjne przyjęto w badaniu algorytm AdaBoost oraz zbalansowany model typu bagging. Na podstawie otrzymanych wyników stwierdzić można, że dla problemów o niezbalansowanym rozkładzie klas próbkowanie z wykorzystaniem rozkładów Beta-dwumianowych do budowy zbiorów uczących wykorzystywanych do konstrukcji kolejnych modeli bazowych prowadzi o istotnie lepszych wyników otrzymywanych przez model niż rozpatrywane modele referencyjne wykorzystujące typowe techniki próbkowania.

5. Uczenie się reprezentacji danych

Problem uczenia się reprezentacji jest zagadnieniem szeroko rozpatrywanym w literaturze i może odnosić się do interpretowalności reprezentacji, procedury ekstrakcji cech, czy uczenia się rozkładu danych na pewnej ukrytej rozmaitości. W ramach osiągnięcia badawczego współuczestniczyłem w konstrukcji interpretowanego modelu klasyfikacyjnego bazującego na Ograniczonych Maszynach Boltzmanna (*ang. Restricted Boltzmann Machines, RBM*) **[P7]**. Jedną z grup modeli interpretowanych stanowią tablice scoringowe, które charakteryzują się tym, że każdej wartości wejściowej określającej obiekt i poddanej wcześniejszej binaryzacji przypisana jest określona wartość punktowa. Finalna klasyfikacja odbywa się poprzez zsumowanie punktów dla cech określających obiekt i odniesienie jej do wartości punktu odcięcia (*ang. threshold value*), traktowanego jako hiperparametr metody. Klasyczne podejścia do budowy tablic scoringowych w swoim działaniu wykorzystują model regresji logistycznej. W ramach osiągnięcia badawczego wykazano, że interpretowalny model tablicy scoringowej może zostać zbudowany z wykorzystaniem klasyfikacyjnej wersji Ograniczonej Maszyny Boltzmanna [8]. W tym celu, w pierwszej kolejności konieczne jest wyuczenie klasyfikacyjnej wersji modelu RBM na danych treningowych. Następnie, w oparciu o wyuczony model dla każdej z rozpatrywanych klas binarnych wyznaczane jest prawdopodobieństwo pojawienia się danej cechy, przy założeniu, że pozostałe cechy nie są aktywowane, a rozpatrywany obiekt należy do klasy pozytywnej. Otrzymana w ten sposób wartość prawdopodobieństwa traktowana jest jako waga określająca istotność rozpatrywanej cechy w tablicy scoringowej. Punkt odcięcia traktowany jest jako hiperparametr i wyznaczany jest z wykorzystaniem wyodrębnionego zbioru walidacyjnego. Jakość opracowanej metody została przebadana na zestawie 4 zbiorów danych z obszaru oceny ryzyka kredytowego. Dla 3 z rozpatrywanych danych wykazano, że opracowany model tablicy scoringowej jest istotnie lepszy niż rozpatrywane podejścia referencyjne. W badaniach wykazano również, że opracowany model charakteryzuje się dużą odpornością na niezbalansowanie danych uczących.

W ramach osiągnięcia badawczego opracowałem algorytm do automatycznej ekstrakcji cech z wykorzystaniem podejścia ewolucyjnego **[P8]**. Istotą opracowanego rozwiązania jest zastosowanie iteracyjnej procedury selekcji najistotniejszych cech, które w dalszej kolejności są krzyżowane z wykorzystaniem zadanych operatorów reprezentujących

podstawowe działania matematyczne. Do wyboru istotnych cech zaproponowałem wykorzystanie wyuczonych gradientowo lasów losowych (XGBoost) [9]. Aby zwiększyć jakość predykcji, dla każdej iteracji w ramach której konstruowane były syntetyczne cechy budowany jest odrębny model w postaci lasu losowego. W fazie wnioskowania zespół zbudowanych modeli podejmuje finalną decyzję w oparciu o procedurę ważonego głosowania. Opracowana metoda została zastosowana do przewidywania upadłości przedsiębiorstw w oparciu o wskaźniki finansowe rozpatrywanych w eksperymencie organizacji. Ciekawym efektem uzyskanym w przypadku zastosowania metody do tak zdefiniowanego zadania klasyfikacji było uzyskanie cech, które można było interpretować jako syntetyczne wskaźniki finansowe. Opracowany model został poddany analizie ilościowej na rzeczywistym zbiorze danych zebranych przez dr inż. Sebastiana Tomczaka. Badania ilościowe, biorąc pod uwagę kryterium AUC, wykazały istotnie wyższą jakość modelu w stosunku do podejść referencyjnych.

Podczas dalszej pracy badawczej pracowałem nad rozwiązaniami mającymi na celu uczenie się reprezentacji z wykorzystaniem danych o częściowo, bądź całkowicie nieznanych etykietach. Efektem prac było zaproponowanie autorskiej metody uczenia się reprezentacji w trybie z częściowym nadzorem **[P9]**. Istotą opracowanego rozwiązania było zastosowanie modelu generatywnego typu GAN [10], który, poprzez zastosowanie odpowiedniej techniki dopasowywania cech (*ang. feature matching*) [11] w procedurze uczenia był w stanie wyuczyć się reprezentatywnych cech na warstwach pośrednich sieci dyskryminującej. Kryterium uczenia modelu zostało wzbogacone o trójkową funkcję strat (*ang. triplet loss*) [12], która umożliwiała wyuczenie się zadanych trybem nadzorowanym relacji pomiędzy porcją oznaczonych obiektów ze zbioru treningowego i pełniła rolę regularyzatora dla rozpatrywanej przestrzeni ukrytej sieci dyskryminującej. W badaniach przeprowadzonych na dwóch zbiorach referencyjnych wykazano, że jakość reprezentacji otrzymanej z sieci dyskryminującej, poprzez zastosowanie zagregowanego kryterium uczenia modelu typu GAN (dla danych zaetykietowanych) oraz trójkowej funkcji strat (dla oznaczonej części obserwacji) w istotny sposób zwiększyło jakość reprezentacji. Jakość reprezentacji była badana dla zadania klasyfikacji, poprzez zastosowanie algorytmu KNN, oraz dla zadania wyszukiwania obrazów (ang. image retrieval), z wykorzystaniem kryterium uśrednionej precyzji (ang. mean average precision, mAP). Jakość modelu badano w odniesieniu do modeli referencyjnych, takich jak zwykła sieć trójkowa, czy też model typu GAN bez regularyzacji sieci trójkowej.

Opracowana koncepcja polegająca na wykorzystaniu sieci dyskryminującej budowanej w ramach modelu GAN do reprezentowania danych została w dalszej kolejności rozszerzona o przypadek nienadzorowany, ze szczególnym naciskiem na konstrukcję kompaktowych reprezentacji binarnych **[P10]**. W tym celu, opracowałam dwie autorskie metody regularyzacji warstwy pośredniej sieci dyskryminującej. Pierwszy element regularyzacji wymuszał dopasowanie przeskalowanych odległości Hamminga pomiędzy reprezentacjami obiektów z wysoko wymiarowej przestrzeni (warstwy pośredniej sieci

dyskryminującej o dużej liczbie jednostek ukrytych) do przestrzeni o niskim wymiarze (warstwy pośredniej sieci dyskryminującej o niewielkiej liczbie jednostek ukrytych), docelowo wykorzystywanej do reprezentowania obiektów. Drugi z elementów regularyzacji miał na celu zwiększyć różnorodność binarnych kodowań poprzez maksymalizację entropię binarnych reprezentacji dla warstwy docelowej sieci dyskryminującej. Dzięki zastosowaniu wymienionych technik regularyzacji udało się uzyskać dobrej jakości kompaktową reprezentację binarną bezpośrednio z warstwy pośredniej dyskryminatora, który na ogół jest elementem niewykorzystywanym po wyuczeniu modelu typu GAN. Jakość opracowanego rozwiązania została przebadana dla dwóch zadań: dopasowywania oraz wyszukiwania obrazów. Otrzymane wyniki dla opracowanej metody były znacząco lepsze w odniesieniu do najlepszych rozwiązań referencyjnych rozpatrywanych w eksperymentach, które również były uczone w oparciu o dane bez nadzoru.

W ramach dalszych prac nad uczeniem się reprezentacji uczestniczyłem opracowaniu generatywnego modelu dla obiektów reprezentowanych poprzez chmury punktów 3D **[P11]**. Istotą opracowanego rozwiązania była architektura typu *adversarial autoencoder* [13] zaprojektowana dla danych reprezentowanych w postaci zbiorów punktów. W opracowanym modelu niewrażliwość na permutacje wynikająca z reprezentacji obiektu w postaci zbioru była zapewniona poprzez zastosowanie architektury typu *PointNet* [14]. Jako miary rekonstrukcji wykorzystywane w procesie uczenia przyjęto odległości Chamfera i Wassersteina [15]. Zadany rozkład na przestrzeni ukrytej wymuszany był poprzez zastosowanie sieci GAN operującej na przestrzeni ukrytej. Istotnym elementem wyróżniającym opracowany model było zaproponowanie odpowiednio sparametryzowanego rozkładu Beta dla przestrzeni ukrytej, dzięki czemu możliwe było wyuczenie się przez model ukrytej reprezentacji binarnej. Opracowany model został poddany analizie ilościowej na znanym zbiorze referencyjnym *ShapeNet* [16], jego właściwości generatywne zostały sprawdzone z wykorzystaniem standardowych miar wykorzystywanych do oceny jakości modeli generatywnych dla reprezentacji trójwymiarowych, takich jak pokrycie, czy minimalna odległość dopasowania. Ocena jakości reprezentacji binarnej odbyła się poprzez wyuczenie na reprezentacji liniowego klasyfikatora SVM, i zbadanie jego właściwości klasyfikacyjnych poprzez analizę wskaźnika poprawności klasyfikacji.

6. Podsumowanie osiągnięcia naukowego

W ramach prowadzonych badań zdiagnozowano szereg problemów z danymi uczącymi występującymi w uczeniu maszynowym i skoncentrowano się na trzech głównych: problemie niezbalasnowania danych uczących, problemie brakujących wartości atrybutów, oraz zagadnieniu uczenia się reprezentacji danych. Dla zidentyfikowanych problemów opracowano szereg algorytmów, które podejmują się ich rozwiązania zgodnie ze schematem przedstawionym na Rysunku 1. Jakość każdego z opracowanych rozwiązań została przebadana w sposób ilościowy, na rzeczywistych,

bądź referencyjnych danych, a otrzymane wyniki były zestawiane z metodami opisanymi w literaturze podejściami referencyjnymi podejmującymi rozpatrywane zagadnienie. Opracowane rozwiązania były prezentowane na renomowanych konferencjach, i publikowane w znanych czasopismach, gdzie były poddane rzetelnymi i szczegółowym recenzjom.

5. **Omówienie pozostałych osiągnięć naukowo - badawczych (artystycznych)**.

Najważniejsze osiągnięcia przed doktoratem

W ramach rozprawy doktorskiej opracowałem szereg rozwiązań bazujących na wzmacnianych modelach typu SVM dla danych niezbalansowanych. Istotą opracowanych rozwiązań było zaproponowanie autorskiej funkcji kosztu pozwalającej na iteracyjną konstrukcję wrażliwych na koszt modeli SVM jako klasyfikatorów bazowych. W procesie uczenia modelu polegającego na optymalizacji zaproponowanej funkcji celu konstruowany jest zespół modeli klasyfikacyjnych nieobciążony w kierunku klasy dominującej. Jakość opracowanego rozwiązania została przebadana na zestawie zbiorów referencyjnych dla danych niezbalansowanych, otrzymane wyniki zostały zestawione z wynikami metod referencyjnych. Ponadto, opracowany model klasyfikacyjny został zastosowany do wybranych problemów z obszaru oceny ryzyka kredytowego i diagnostyki medycznej z wykorzystaniem rzeczywistych zbiorów danych. Opracowana metoda została wzbogacona o mechanizmy aktywnego uczenia pozwalające na wybór jedynie istotnych próbek dla każdej z iteracji konstrukcji modelu bazowego. Do wyboru istotnych próbek wykorzystywany był poszerzony margines klasyfikacyjny ostatnio wyuczonego modelu SVM. Opracowana modyfikacja pozwoliła zwiększyć efektywność uczenia i jakość predykcyjną opracowanego algorytmu. Większość z wypracowanych wyników w ramach doktoratu została opublikowana w poniższych pracach:

Zięba Maciej, Tomczak Jakub *“Boosted SVM with active learning strategy for imbalanced data”*. Soft Computing. 2015, vol. 19, nr 12, s. 3357-3368.

Zięba Maciej, Tomczak Jakub, Lubicz Marek, Świątek Jerzy *“Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients”*. Applied Soft Computing. 2014, vol. 14, pt. A, s. 99-108.

Najważniejsze osiągnięcia po doktoracie

W ramach osiągnięć po doktoracie, brałem udział w opracowaniu jednego z wariantów algorytmów budowy modelu do detekcji sygnałów w widmach NMR. Opracowany wariant modelu zakładał, że dla wyznaczonych znaczących ekstremów lokalnych w widmie NMR, konstruowany był wycinek sygnału o zadanych wymiarach, dla którego

punkt ekstremum był punktem centralnym, i który następnie był dostarczany na wejście głębokiej sieci neuronowej. Zadaniem głębokiej sieci neuronowej była klasyfikacja, czy rozpatrywany punkt wyznaczonego ekstremum reprezentował znaczącą odpowiedź sygnału. Opracowana metoda została zwalidowana na rzeczywistych sygnałach NMR, a wyniki prac zostały opublikowane w renomowanym czasopiśmie Bioinformatics:

Klukowski Piotr, Augoff Michał, **Zięba Maciej**, Drwal Maciej, Gonczarek Adam, Walczak Michał. "*NMRNet: a deep learning approach to automated peak picking of protein NMR spectra*". Bioinformatics. 2018, vol. 34, nr 15, s. 2590–2597.

Kolejnymi zagadnieniami, w obrębie których prowadziłem prowadziłem prace badawcze po doktoracie były zagadnienia związane z modelami generatywnymi dla chmur punktów 3D. W ramach prowadzonych badań brałem udział w opracowaniu modelu generatywnego dla obiektów trójwymiarowych wykorzystującego dyskretne modele przepływowe [17]. Opracowany model zakładał, że chmura punktów była sprowadzana do niskowymiarowej przestrzeni z wykorzystaniem wielowymiarowego skalowania, następnie, poprzez zastosowanie dyskretnego modelu przepływowego transformowana do osadzenia z wymuszonym rozkładem a priori. Reprezentacja wektorowa chmury w otrzymanym osadzeniu stanowi czynnik warunkujący do drugiego modelu przepływowego przekształcającego trójwymiarowe punkty próbkowane z Gaussa na punkty reprezentujące trójwymiarowy kształt determinowany przez wektor warunkujący. Generowanie chmur punktów 3D z wykorzystaniem modelu odbywa się poprzez próbkowanie losowego wektora z zadanego rozkładu a priori, który wykorzystywany jest jako czynnik warunkujący w dyskretnym modelu przepływowym, który przekształca próbki z Gaussa na kształt 3D. Ze względu na przyjętą architekturę modelu możliwe jest generowanie kształtów 3D o dowolnej liczbie punktów. Możliwości generatywne opracowanej metody zostały przebadane na referencyjnym zbiorze danych i otrzymane wyniki zostały zaprezentowane na konferencji NeurIPS 2019:

Stypułkowski Michał, Zamorski Maciej, **Zięba Maciej**, Chorowski Jan *"Conditional Invertible Flow for Point Cloud Generation"* NeurIPS 2019, Sets & Partitions: NeurIPS 2019 Workshop.

W dalszej kolejności, przy współpracy z naukowcami z UJ, brałem udział w opracowaniu modelu *HyperCloud* - modelu generatywnego dla chmur punktów 3D wykorzystujący hipersieci [18]. Opracowany model zakładał istnienie sieci przekształcającej chmurę punktów do przestrzeni ukrytej regularyzowanej do zadanego rozkładu a priori. W ramach naszych prac rozważaliśmy dwa sposoby regularyzacji przestrzeni ukrytej: poprzez dywergencję Kullbacka-Leiblera, oraz poprzez zastosowanie dodatkowego, ciągłego modelu przepływowego na przestrzeni ukrytej [19]. Następnie, wektor reprezentujący chmurę punktów w przestrzeni ukrytej jest przekształcany z wykorzystaniem dekodującej sieci neuronowej. Dla przyjętego modelu sieć dekodująca reprezentowana jest przez tzw. hipersieć, czyli model przekształcający wektor ukryty do

przestrzeni parametrów innej sieci neuronowej, nazywanej siecią docelową. W przypadku opracowanego modelu sieć docelowa jest reprezentowana przez MLP i celem jej działania jest przekształcenie trójwymiarowych punktów z zadanego rozkładu a priori do przestrzeni reprezentującej kształt. W odróżnieniu od modelu wykorzystującego dyskretne modele przepływowe czynnikiem warunkującym dla modelu generującego punkty nie jest wektor w przestrzeni ukrytej, ale zestaw parametrów modelu przekształcającego punkty 3D. Praca dokumentująca opracowaną metodę została opublikowana na konferencji ICML 2020:

Przemysław Spurek, Sebastian Winczowski, Jacek Tabor, Maciej Zamorski, **Maciej Zięba**, Tomasz Trzciński, *Hypernetwork approach to generating point clouds,* ICML 2020.

Istotnym ograniczeniem opracowanego modelu był fakt, iż sieć docelowa reprezentowana była przez klasyczny model MLP, co uniemożliwiało dwukierunkowe transformacje punktów pomiędzy przestrzenią z założonym rozkładem a priori, a przestrzenią w której reprezentowana jest chmura punktów 3D. Ograniczenie to zostało wyeliminowane poprzez zastąpienie modelu MLP ciągłym modelem przepływowym. W konsekwencji, hipersieć działająca z architekturze jako dekoder przewiduje parametry funkcji modelującej ciągłe, dwukierunkowe przepływy pomiędzy reprezentacjami. Dodatkowo, jako rozkład z którego próbkowane są punkty przyjęto rozkład log-normalny zdefiniowany na powierzchni sfery jednostkowej, co umożliwiło efektywne generowanie obiektów 3D reprezentowanych za pomocą powłok. Wyniki ilościowe i jakościowe pokazują znaczną poprawę jakości zmodyfikowanego modelu w stosunku do oryginalnej architektury HyperCloud. Wypracowane wyniki zostały opublikowane w pracy:

Przemysław Spurek, **Maciej Zięba**, Jacek Tabor, Tomasz Trzciński, *HyperFlow: Representing 3D Objects as Surfaces*, https://arxiv.org/pdf/2006.08710.pdf.

Statystyki dorobku naukowego

- Sumaryczny IF: **26.557**
- Sumaryczny IF dla prac przed doktoratem: **4.44**
- Sumaryczny IF dla prac po doktoracie: **22.11**
- Suma cytowań (Google Scholar, 29.06.2020): **521**
- Suma cytowań (WoS, 29.06.2020): **210**
- Suma cytowań bez autocytowań (WoS, 29.06.2020): **202**
- Średnia liczba cytowań na pracę (WoS, 29.06.2020): **11,22**
- H-index (WoS, 29.06.2020): **7**
- H-index (Google Scholar, 29.06.2020): **11**

Potwierdzona wspólnymi publikacjami współpraca krajowa i międzynarodowa

- **Dr hab. Tomasz Trzciński (Politechnika Warszawska).** Współpraca w zakresie opracowywania metod do uczenia się niskowymiarowych reprezentacji binarnych i modeli generatywnych dla chmur punktów 3D.

  Wspólne publikacje:

  Maciej Zięba, Piotr Semberecki, Tarek El-Gaaly, and Tomasz Trzcinski. "Bingan: Learning compact binary descriptors with a regularized GAN." Advances in Neural Information Processing Systems. 2018.

  Maciej Zamorski, Maciej Zięba, Piotr J. Klukowski, Rafał A. Nowak, Karol Kurach, Wojciech Stokowiec, Tomasz Trzciński "Adversarial autoencoders for compact representations of 3D point clouds". Computer Vision and Image Understanding. 2020. vol. 193, art. nr 102921, s. 1-8.

  Przemysław Spurek, Sebastian Winczowski, Jacek Tabor, Maciej Zamorski, Maciej Zięba, Tomasz Trzciński, Hypernetwork approach to generating point clouds, ICML 2020.

  Przemysław Spurek, Maciej Zięba, Jacek Tabor, Tomasz Trzciński, *HyperFlow: Representing 3D Objects as Surfaces*, https://arxiv.org/pdf/2006.08710.pdf.

- **Dr Karol Kurach (Google Brain)**. Współpraca w zakresie opracowywania modeli generatywnych dla chmur punktów 3D.

  Wspólne publikacje:

  Maciej Zamorski, Maciej Zięba, Piotr J. Klukowski, Rafał A. Nowak, Karol Kurach, Wojciech Stokowiec, Tomasz Trzciński "Adversarial autoencoders for compact representations of 3D point clouds". Computer Vision and Image Understanding. 2020. vol. 193, art. nr 102921, s. 1-8.

- **Prof Lei Wang (University of Wollongong, Australia)** Współpraca w ramach której opracowano metodę uczenia się reprezentacji danych wizyjnych poprzez zastosowanie technik z częściowym nadzorem. Współpraca wynikająca z odbytego stażu podoktorskiego.

  Wspólne publikacje:

  Maciej Zięba, Lei Wang. "Training triplet networks with GAN." International Conference on Learning Representations ICLR (workshop track). 2017.

- **Prof Wolfgang Heardle (Humboldt-Universität zu Berlin, Niemcy)** Współpraca w zakresie modeli do oceny ryzyka kredytowego dla danych niezbalansowanych. Współpraca wynikająca z wizyty na **Humboldt-Universität** w roli wizytującego naukowca.
  Wspólne publikacje:

  Maciej Zięba, and Wolfgang Karl Härdle. "Beta-boosted ensemble for big credit scoring data." Handbook of Big Data Analytics. Springer, Cham, 2018. 523-538.

- **Dr Yaroslaw Nikolaev (ETH Zurich, Szwajcaria)**, Współpraca miała miejsce w okresie trwania projektu "Opracowanie narzędzi bioinformatycznych do poszukiwania leków (01.01.2016 - 28.02.2018)". W ramach współpracy opracowany został moduł do śledzenia sygnałów NMR. Głównym zadaniem po stronie ETH było opracowanie i przygotowanie danych na potrzeby walidacji opracowanego algorytmu.

  Wspólne publikacje:

  Maciej Zięba, Piotr Klukowski, Adam Gonczarek, Yaroslav Nikolaev, and Michał J. Walczak "Gaussian process regression for automated signal tracking in step-wise perturbed Nuclear Magnetic Resonance spectra." Applied Soft Computing 68 (2018): 162-171.

- **Dr Przemysław Spurek (Uniwersytet Jagielloński, Polska)**, oraz **prof. dr hab. Jacek Tabor (Uniwersytet Jagielloński, Polska)**, Współpraca w ramach opracowania nowego modelu generowania chmur punktów i siatek reprezentujących obiekty 3D z wykorzystaniem hypersieci.

  Wspólne publikacje:

  Przemysław Spurek, Sebastian Winczowski, Jacek Tabor, Maciej Zamorski, Maciej Zięba, Tomasz Trzciński, Hypernetwork approach to generating point clouds, ICML 2020

  Przemysław Spurek, Maciej Zięba, Jacek Tabor, Tomasz Trzciński, *HyperFlow: Representing 3D Objects as Surfaces*, https://arxiv.org/pdf/2006.08710.pdf.

- **Dr hab. Jan Chorowski (Uniwersytet Wrocławski, Polska)** Współpraca, której efektem była metoda generowania chmur punktów z wykorzystaniem dyskretnych modeli przepływowych.

Wspólne publikacje:

Michał Stypułkowski, Maciej Zamorski, Maciej Zięba, Jan Chorowski, Conditional Invertible Flow for Point Cloud Generation, Neural Information Processing Systems, NeurIPS 2019: Workshop on Sets & Partitions.

Literatura

[1] García-Laencina, Pedro J., José-Luis Sancho-Gómez, and Aníbal R. Figueiras-Vidal. *"Pattern classification with missing data: a review."* Neural Computing and Applications 19.2 (2010): 263-282.

[2] He, Haibo, and Edwardo A. Garcia. *"Learning from imbalanced data."* IEEE Transactions on knowledge and data engineering 21.9 (2009): 1263-1284.

[3] Chawla, Nitesh V., et al. *"SMOTE: synthetic minority over-sampling technique."* Journal of artificial intelligence research 16 (2002): 321-357.

[4] Hinton, Geoffrey E. *"A practical guide to training restricted Boltzmann machines."* Neural networks: Tricks of the trade. Springer, Berlin, Heidelberg, 2012. 599-619.

[5] Fernández, Alberto, et al. *"SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary."* Journal of artificial intelligence research 61 (2018): 863-905.

[6] Kumar, M. Pawan, Benjamin Packer, and Daphne Koller. *"Self-paced learning for latent variable models."* Advances in neural information processing systems. 2010.

[7] Bengio, Yoshua, et al. *"Curriculum learning."* Proceedings of the 26th annual international conference on machine learning. 2009.

[8] Larochelle, Hugo, and Yoshua Bengio. *"Classification using discriminative restricted Boltzmann machines."* Proceedings of the 25th international conference on Machine learning. 2008.

[9] Chen, Tianqi, and Tong He. *"Higgs boson discovery with boosted trees."* NIPS 2014 workshop on high-energy physics and machine learning. 2015.

[10] Goodfellow, Ian, et al. *"Generative adversarial nets."* Advances in neural information processing systems. 2014.

[11] Salimans, Tim, et al. *"Improved techniques for training gans."* Advances in neural information processing systems. 2016.

[12] Hoffer, Elad, and Nir Ailon. *"Deep metric learning using triplet network.*" International Workshop on Similarity-Based Pattern Recognition. Springer, Cham, 2015.

[13] Makhzani, Alireza, et al. *"Adversarial autoencoders."* arXiv preprint arXiv:1511.05644 (2015).

[14] Qi, Charles R., et al. *"Pointnet: Deep learning on point sets for 3d classification and segmentation."* Proceedings of the IEEE conference on computer vision and pattern recognition. 2017

[15] Achlioptas, Panos, et al. *"Learning representations and generative models for 3d point clouds."* International conference on machine learning. 2018.

[16] Chang, Angel X., et al. *"Shapenet: An information-rich 3d model repository."* arXiv preprint arXiv:1512.03012 (2015).

[17] Dinh, Laurent, Jascha Sohl-Dickstein, and Samy Bengio. *"Density estimation using real nvp."* arXiv preprint arXiv:1605.08803 (2016).

[18] Ha, David, Andrew Dai, and Quoc V. Le. *"Hypernetworks."* arXiv preprint arXiv:1609.09106 (2016).

[19] Yang, Guandao, et al. *"Pointflow: 3d point cloud generation with continuous normalizing flows."* Proceedings of the IEEE International Conference on Computer Vision. 2019.