Streszczenie rozprawy

Remark: The author of this abstract is NOT a native Polish speaker. For this reason, the Polish text may have some flaws.

I. Wprowadzenie i motywacja

Uczenie maszynowe - machine learning (**ML**) do początku XXI wieku przeżywa gwałtowny rozwój, stymulowany nie tylko zastosowaniami praktycznie w każdej dziedzinie naszego życia, ale także stale wzrastającą i taniejącą dostępną mocą obliczeniową (m.in., obliczenia na procesorach graficznych, obliczenia rozproszone w chmurze, superserwery z ogromną pamięcią RAM).

Typowy *przepływ zadań uczenia maszynowego* (przedstawiony na rys. 1) składa się z dwóch następujących części: (1) potoku danych (ang. data pipeline, data processing pipeline) i (2) budowania modelu (ang. model building). *Potok danych* obejmuje zadania znalezienia prawidłowej sekwencji transformacji wejściowego zbioru danych tak, aby wyjściowy zbiór danych nadawał się do przetworzenia algorytmem uczenia maszynowego. *Budowanie modelu* obejmuje zadania wyboru właściwego algorytmu uczenia maszynowego i jego hiper-parametrów tak, aby zbudowany model zapewniał dobrą generalizację w odniesieniu do danej miary wydajności.

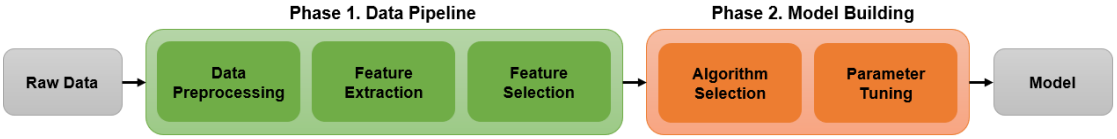


Figure 1: Typowy przepływ zadań uczenia maszynowego.

Ważnym trendem badawczym w zakresie uczenia maszynowego jest tzw. *kompleksowe uczenie maszynowe - end-to-end machine learning* (**E2EML**). Odnosi się ono do systemów, które są zdolne do budowania modeli z surowych danych bez ingerencji człowieka. Zwykle proces ten obejmuje czyszczenie i wstępne przetwarzanie danych, wybór odpowiedniego algorytmu i dostrojenie jego hiper-parametrów, podobnie jak przedstawiono na rysunku 1. Hiper-parametry algorytmu to parametry, które sterują procesem uczenia maszynowego, tj. wspomagają algorytm w znalezieniu dobrego modelu. Hyper-parametrów algorytm nie uczy się na podstawie danych ale są one dobierane ręcznie, są to np.: liczba drzew w lesie losowym (ang. random forrest), współczynnik uczenia (ang. learning rate) - w przypadku algorytmu gradientu prostego (ang. gradient descent) lub współczynnik mutacji (ang. mutation rate) - w przypadku algorytmu genetycznego (ang. genetic algorithm).

Zbudowanie wysokiej jakości modelu uczenia maszynowego dla zastosowań w przemyśle jest trudnym, czasochłonnym i złożonym obliczeniowo zadaniem, wymagającym wiedzy eksperckiej. Jest to główną przeszkodą dla powszechnego zastosowania rozwiązań ML i E2EML w firmach. Mimo, że większość firm gromadzi ogromne wolumeny danych w trakcie swojej działalności, możliwości analizy tych danych są ograniczone brakiem wystarczającej liczby pracowników posiadających odpowiednią wiedzę z zakresu przetwarzania danych i

uczenia maszynowego.

Z raportów ekspertów z dziedziny wynika, że od 50% do 80% czasu i zasobów w projektach data science jest poświęcane na budowanie przepływów zadań dla uczenia maszynowego i analizy danych [Chessell 2014, Pa1 2018, Cog 2019]. W tym procesie, równie ważnym jak budowanie samego potoku danych (por. rys. 1) jest wybór i strojenie algorytmu ML budującego model. Najnowsze rozwiązania częściowo umożliwiają budowanie wydajnych potoków danych i wybór dobrego modelu. Na etapie wyboru modelu, meta-optymalizatory są w stanie automatycznie wybrać algorytm i dostroić model bez pomocy użytkownika, kosztem dużego narzutu czasu obliczeniowego. Jednakże należy zauważyć, że rozwiązania te nie są szeroko stosowane ze względu na wymagania obliczeniowe i konieczną wiedzę ekspercką. W tym kontekście, dostrajanie hiper-parametrów często w ogóle nie jest realizowane [Couronné 2018]!

Jedną a technik ML najczęściej wykorzystywanych w praktyce jest klasyfikacja (and. classification). Polega ona na zbudowaniu modelu umożliwiającego przewidzenie na podstawie pewnych cech obiektu, do jakiej klasy należy ten obiekt. Wiele rzeczywistych scenariuszy może być modelowanych jako problem klasyfikacyjny. Przykładowo, techniki klasyfikacji są z powodzeniem stosowane w medycynie - do celów diagnostycznych, w bankowości - do oceny zdolności kredytowej, przemyśle spożywczym - do klasyfikowania produktów według jakości, w handlu - do budowania profili zakupowych klientów, czy w telekomunikacji - do podziału klientów ze względu na ich profile zachowania się.

Na drugim końcu spektrum zastosowań technik ML jest dziedzina prawna, która jest jedną z najsłabiej wspieranych przez te techniki. Dziedzina ta ma bardzo wysoki potencjał aplikacyjny technik ML. Pomoc systemowi prawnemu w podejmowaniu lepszych decyzji zmniejszyłaby koszty i stronniczość, tym samym dając każdemu obywatelowi szerszy dostęp do obiektywnego wymiaru sprawiedliwości. W ostatnich latach obserwuje się rosnące zainteresowanie firm informatycznych dostarczaniem nowoczesnych narzędzi dla dziedziny prawa. Firmy te, znane jako LegalTech, rosną w imponującym tempie (25% wzrostu rocznie na rynku szacowanym na ponad 1 miliard USD rocznie[1].

Prawo jest złożoną dziedziną (ang. messy concept) [Rissland 2006], która z natury rzeczy stwarza szereg trudności dla algorytmów uczenia maszynowego. Trudności te to przede wszystkim: możliwość wieloznacznej interpretacji, pojawiające się wyjątki od reguł, niestacjonarność obserwacji, rozumowanie dedukcyjne i indukcyjne, logika nieklasyczna. Co więcej, modele statystyczne często działają na zasadzie czarnej skrzynki, co znacznie ogranicza ich praktyczne zastosowania. Podobnie, uczenie maszynowe oparte o sztuczne sieci neuronowe nie jest możliwe do zastosowania w dziedzinie prawnej, ze względu na trudność związaną z objaśnianiem zbudowanego modelu orzekania o winie, a tym samym trudność uzasadniania wyroków otrzymanych przez taki model. Innymi słowy, dziedzina prawna łączy niektóre z najtrudniejszych wyzwań dzisiejszego uczenia maszynowego.

Podsumowując, konieczne jest zatem opracowanie technik ograniczających wymaganą wiedzę ekspercką i udział człowieka w budowaniu kompleksowych przepływów danych w uczeniu maszynowym bez drastycznego wydłużania czasu obliczeniowego, we wszystkich dziedzinach zastosowania technik ML. W przeciwnym razie, koszt przyjęcia rozwiązań w zakresie ML pozostanie wyższy niż utrzymanie obecnie istniejących, mniej wydajnych pro-

[1] https://prismlegal.com/legal-tech-market-sizing-and-opportunities/

cesów, szczególnie w dziedzinach tradycyjnie dalekich od dziedzin technicznych, takich jak dziedzina prawna.

II. Cel i zakres rozprawy

W tym kontekście, głównym celem niniejszej rozprawy doktorskiej jest opracowanie w pełni zautomatyzowanego (kompleksowego) rozwiązania wspierającego budowanie *modelu klasyfikacji* bez udziału człowieka. Rozwiązanie to ma pracować z dowolnymi typami danych, co implikuje konieczność budowania modeli dla przestrzeni bez żadnych metryk (ang. non-metric space). Jako dziedzinę aplikacyjną wybrano przewidywanie decyzji sądowych ze względu na wyzwania, jakie stwarza ta dziedzina oraz ze względu na to jak niewiele rozwiązań zastało zaproponowanych do tej pory w literaturze naukowej i dostępnym oprogramowaniu komercyjnym i niekomercyjnym.

Niniejsza rozprawa doktorska stara się odpowiedzieć na trzy następujące *pytania*:

- Czy algorytm klasyfikacji może nauczyć się modelu w przestrzeni bez żadnych metryk?
- W jakim stopniu przygotowanie danych wpływa na jakość modelu predykcji, tj. czy ważniejsze jest przygotowanie danych dla algorytmu, czy strojenie tego algorytmu?
- Jak skutecznie zautomatyzować fazę przygotowania danych (potok danych)?

W niniejszej rozprawie stawiamy dwie *hipotezy*.

- Po pierwsze, algorytm uczenia maszynowego może nauczyć się metryki na samych danych w oparciu o informacje zwrotne dostarczone przez zbiory uczące (ang. learning set).
- Po drugie, aby zbudować model, który będzie poprawnie działał także na nowych (nieznanych) danych, do budowy tego modelu należy dostarczyć danych wysokiej jakości, tj. jakość danych jest ważniejsza niż sam algorytm. Mówiąc dokładniej, jeżeli dany jest algorytm, który zachowuje się jak uniwersalny aproksymator [Csáji 2001], czyli jest w stanie nauczyć się prawie każdej funkcji ciągłej na zwartym podzbiorze R^m, wtedy głównym praktycznym ograniczeniem dla procesu uczenia jest jakość danych. Dla określonego budżetu czasu zakładamy, że ważniejsze może być poświęcenie większej jego części na wstępne przetwarzanie danych, niż na wybór algorytmu i dokładne dostrojenie jego hiper-parametrów.

III. Aktualny stan wiedzy

Przedstawiona w tym punkcie analiza stanu wiedzy dotyczy dwóch dziedzin objętych zakresem niniejszej rozprawy, tj. uczenia maszynowego (w szczególności zautomatyzowanego) i zastosowania technik ML w domenie prawnej.

Zautomatyzowane uczenie maszynowe i kompleksowe uczenie maszynowe

Techniki tradycyjnie nazywane jako *zautomatyzowane uczenie maszynowe* (**AutoML**) lub wspomniane wcześniej kompleksowe uczenie maszynowe, koncentrują się w praktyce

na problemie łączenia algorytmów i optymalizacji hiper-parametrów - nazywanym dalej CASH (ang. combined algorithm selection and hyperparameter optimization - CASH) [Kotthoff 2017, Feurer 2015]. Podejście takie całkowicie pomija znaczenie potoków danych dla jakości modelu [Crone 2006], koncentrując się na wyborze algorytmu i dostrajaniu hiperparametrów. Metoda sekwencyjnej optymalizacji w oparciu o model [Hutter 2011] (ang. Sequential Model-Based Optimization) może być zrealizowana na różne sposoby, między innymi przy użyciu Lasu Losowego [Hutter 2011], tzw. Estymatora Tree-Parzen [Bergstra 2015], lub Regresji Gaussa [Martinez-Cantin 2014].

W przypadku potoku danych i wstępnego przetwarzania, większość rozwiązań wykorzystuje półautomatyczne narzędzia wspierające naukowców danych (ang. data scientists). W [Polyzotis 2017] stosuje się wytyczne do weryfikacji jakości wstępnie przetworzonych danych w ciągłym uczeniu maszynowym, tj. modelach uczenia maszynowego w produkcji i otrzymywaniu w sposób ciągły nowych danych treningowych. Ostatnio zaproponowano metodę wykorzystującą meta-atrybuty do oszacowania wpływu operatorów przetwarzania wstępnego na dokładność modelu [Bilalli 2017]. Podejście to tworzy ukrytą przestrzeń wykorzystując meta-atrybuty (np. liczbę klas lub atrybutów, entropię, stosunek sygnału do szumu), w których można przedstawić dowolny zbiór danych. Moduł zwany meta-learner jest uczony na kilku różnych zbiorach danych. Meta-model jest zatem w stanie przewidzieć wpływ zastosowania różnych technik transformacji danych w potoku danych na jakość budowanego modelu predykcji, bez konieczności uczenia modelu i jego oceny za pomocą np. walidacji krzyżowej. Wreszcie, w innym podejściu użytkownik przekazuje do systemu informację zwrotną na temat jakości danych w celu optymalizacji przepływów [**?**].

Uczenie metryki polega na wyborze właściwej metryki, która umożliwia prawidłowe porównanie lub klasyfikację danych [Bellet 2013, Wang 2015]. Wybór odpowiedniej metryki do pomiaru odległości między dwoma punktami jest kluczowy dla jakości algorytmów klasyfikacji [Davis 2007]. Uczenie metryki polega na znalezieniu rzutu f z przestrzeni początkowej na przestrzeń euklidesową, tak że dla dowolnych elementów $\mathbf{x}$ i $\mathbf{x}'$, $d(\mathbf{x},\mathbf{x}') = ||f(\mathbf{x}) - f(\mathbf{x}')||$. Metryka powinna odzwierciedlać różnicę semantyczną między obiektami. Zaskakująco, większość metrycznych metod uczenia zakłada, że dane są początkowo reprezentowane w przestrzeni wektorowej, co może nie być właściwe dla wielu problemów, w których mogą pojawiać się dane ustrukturalizowane, częściowo ustrukturalizowane, lub nieustrukturalizowane.

Domena prawna

Jako dziedzina aplikacyjna rozwiązań opracowanych w ramach niniejszej rozprawy został wybrany wymiar sprawiedliwości. Nieliczne opublikowane wcześniej badania naukowe w zakresie stosowania uczenia maszynowego do wspomagania decyzji sądowych pokazały, że domena prawna jest szczególnie interesująca i trudna dla algorytmów uczenia maszynowego. Po pierwsze, ze względu na wielość i złożoność reguł prawnych oraz złożoność semantyczną aktów prawnych. Po drugie, ze względu na brak jednolitego repozytorium aktów prawnych i orzeczeń sądowych. Po trzecie, decyzje sądowe zmieniają się w czasie dla podobnych przypadków (tj. nie występuje stacjonarność obserwacji) i obserwuje się wielość odstępstw od reguł w wydawaniu orzeczeń. Zatem opracowanie całościowego

(zautomatyzowanego) podejścia do budowania modeli klasyfikacji dla wspomagania decyzji sądowych ma ogromny potencjał praktyczny (wdrożeniowy).

Przewidywanie decyzji sądowych stanowi wyzwanie samo w sobie, nawet dla najlepszych ekspertów prawnych: w przypadku *Supreme Court of the United States* (SCOTUS) osiągnięto 58% dokładność [Ruger 2004]. Natomiast projekt Fantasy SCOUTS[2], w którym mamy odczynienia z ogromną grupą wolontariuszy przewidujących jak dany członek Sądu Najwyższego Stanów Zjednoczonych będzie orzekał w danej sprawie, osiągnął 84,85% poprawnych prognoz. Brak jest podobnych wyników dla orzecznictwa europejskiego, za wyjątkiem badań na małych zbiorach danych [Aletras 2016].

Dotychczas zaproponowane podejścia do przewidywania decyzji sądowych można podzielić na trzy grupy: (1) modele statystyczne, (2) wnioskowanie na podstawie przypadków (ang. Case Based Reasoning - CBR) i (3) abstrakcyjną argumentację (ang. Abstract Argumentation – AA).

Modele statystyczne wykorzystano do przewidywania werdyktów sądu amerykańskiego - *Supreme Court of the United States* [Katz 2017a, Martin 2004b, Guimerà 2011]. Zgodnie z naszą najlepszą wiedzą, w odniesieniu do *European Court of Human Rights* istnieje niewiele modeli predykcji [Aletras 2016, Medvedeva 2020, Chalkidis 2019]. Zbiór danych użyty w [Aletras 2016] obejmuje wyłącznie kilka artykułów prawnych, z których każdy zawiera od 80 do 254 przypadków. Wykorzystane w pracach [Aletras 2016, Medvedeva 2020] modele predykcyjne wykorzystują liniowy klasyfikator SVM osiągając od 75% do 79% dokładności predykcji (accuracy). W [Chalkidis 2019] wykorzystano sztuczne sieci neuronowe, uzyskując wartość miary F1 maksymalnie 82%

Podejście CBR wykorzystuje podobieństwa pomiędzy cechami i rozwiązaniami poprzednich obserwacji w celu zbudowania nowego rozwiązania dla nowego przypadku (w kontekście niniejszej rozprawy - nowej sprawy sądowej). Metody CBR nie uwzględniają czynników pozaprawnych, a zatem nie są w stanie poradzić sobie z problemem prognozowania. Metody te dostarczają natomiast uzasadnienia dla swoich decyzji [Aleven 1997].

Podejście AA polega na modelowaniu informacji jako graf argumentów i wyciąganiu wniosków poprzez rozwiązywanie konfliktów za pomocą logiki lub ważenia argumentów. Mimo, że metody statystyczne dostarczają interesujących wyników dla problemu prognozowania [Guimerà 2011, Martin 2004a, Ruger 2004, Katz 2017b], nie są one w stanie dostarczyć prawnego uzasadnienia swoich prognoz. W AA pojawiły się dwa rodzaje przeciwstawnych podejść: pozytywne, które mają na celu modelowanie rzeczywistych procesów decyzyjnych [Baroni 2015] i normatywne, które próbują opracować metody wyboru spośród najlepszych alternatyw i argumentów [Dung 2006]. Pierwsze podejście może dobrze wspierać rozwiązanie problemu prognozowania, a drugie - problemu uzasadnienia. Oba podejścia w dużej mierze polegają na wiedzy eksperckiej koniecznej do konstruowania tzw. argumentów, co ogranicza zastosowanie AA.

[2] https://fantasyscotus.lexpredict.com/

IV. Kontrybucja rozprawy

W niniejszej rozprawie proponujemy alternatywne podejście do konstruowania potoku danych z uczeniem maszynowym, które zaprezentowano na rys. 2. Proponowany potok danych zakłada, że typy i formaty danych przetwarzanych przez potok danych nie są z góry znane i mogą ewoluować podczas przetwarzania danych. Stanowi to problem, ponieważ nie wszystkie algorytmy uczenia maszynowego mogą obsługiwać dowolny typ danych. W szczególności niektóre algorytmy działają tylko z danymi liczbowymi lub wartościami ciągłymi, niektóre nie mogą działać, gdy pojawiają się wartości puste, lub są wrażliwe na wartości odstające.

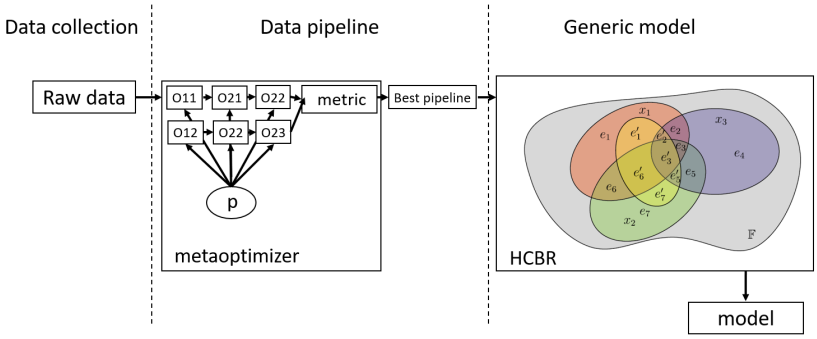


Figure 2: Zmodyfikowany przepływ zadań uczenia maszynowego zaproponowany jako rozwiązanie problemu kompleksowej klasyfikacji.

W szczególności, podejście zaproponowane w niniejszej rozprawie doktorskiej bazuje na trzech następujących rozwiązaniach.

- Po pierwsze, proponujemy **ogólny sposób automatycznego budowania i konfigurowania potoku** danych w celu przygotowania danych dla dowolnego algorytmu uczenia maszynowego. Konstrukcja potoku danych może być sformułowana jako problem optymalizacji, można go zatem rozwiązać automatycznie, w oparciu o istniejące meta-optymalizatory, przy wykorzystaniu minimalnej wiedzy specjalistycznej. Według naszej najlepszej wiedzy, dotychczas nie zaproponowano podobnego rozwiązania.
- Po drugie, proponujemy **zastosowanie metody Hypergraph Case-Based Reasoning** (HCBR), wykorzystującej zalety metod statystycznych, CBR i systemu argumentacji, jednocześnie unikając ich wad. W HCBR proponujemy zastosowanie generycznego algorytmu, który może przetwarzać dane dowolnego typu i uczyć się złożonych modeli, wykorzystujący przy tym niewiele hiper-parametrów lub nie wykorzystujący ich w ogóle. Dzięki temu, można zredukować czas potrzebny na budowanie modelu, bez konieczności udziału użytkownika.
- Po trzecie, opracowaliśmy **otwarte repozytorium danych prawnych**, zawierające sprawy sądowe i orzeczenia z Europejskiego Trybunału Praw Człowieka. Dane w repozytorium zostały wcześniej oczyszczone, uspójnione i przetransformowane (przez

zadania potoku danych) do postaci wymaganej przez algorytmy klasyfikacji. Repozytorium zostało upublicznione w postaci portalu (https://echr-opendata.eu/). Dzięki temu naukowcy z całego świata mogą korzystać ze zgromadzonych w nim danych i uruchamiać algorytmy uczenia maszynowego na danych przygotowanych do tego typu przetwarzania. Repozytorium stanowi tym samym benchmark dla algorytmów ML działających w domenie prawnej.

Budowa i optymalizacja zautomatyzowanego potoku danych

W prezentowanej rozprawie doktorskiej proponujemy, zgodnie z naszą najlepszą wiedzą, pierwszy ogólny sposób automatycznego budowania i konfigurowania potoku danych w celu przygotowania danych dla dowolnego algorytmu uczenia maszynowego. Zaproponowaliśmy zmodyfikowany przepływ zadań [Quemy 2020a, Quemy 2019b], zaprezentowany na rys. 3.

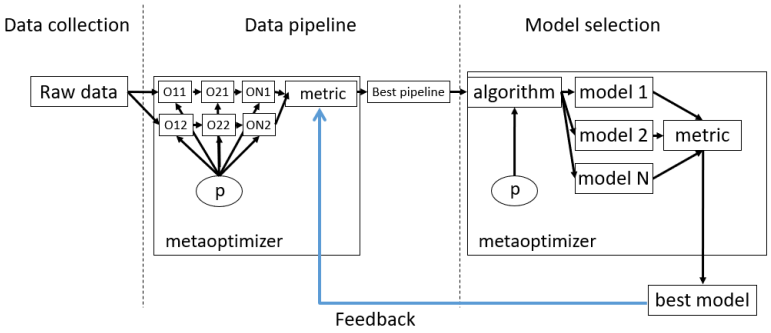
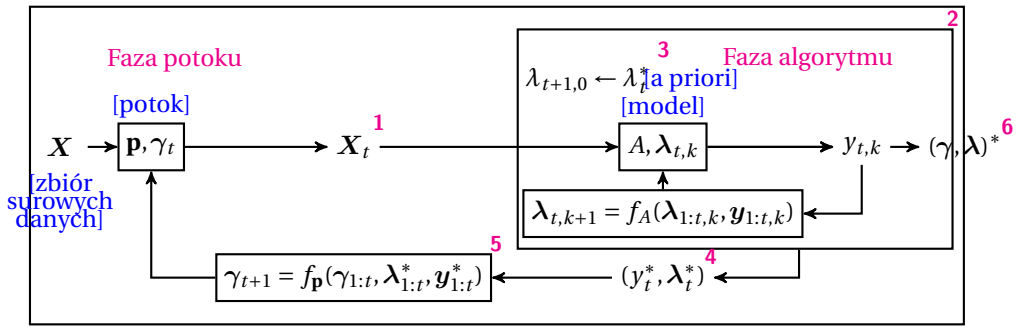


Figure 3: Przepływ zadań dla automatyzacji konstrukcji potoku danych. Główną ideą jest ponowne wykorzystanie meta-optymalizatora na podstawie danych zwrotnych dostarczonych przez działający model.

W celu wykazania potencjału takiego podejścia, zdefiniowaliśmy gramatykę, która umożliwia definiowanie potoków na wyższym poziomie abstrakcji, reprezentowanych jako grafy. Taką reprezentację nazwaliśmy prototypem potoku. Każdy węzeł może być utworzony za pomocą kilku operatorów (np. PCA), a każdy z nich ma swój własny zestaw parametrów (np. liczbę składników w PCA). Użytkownik końcowy nie musi posiadać żadnej wiedzy na temat tych operatorów.

W [Quemy 2020a, Quemy 2019b] zaproponowaliśmy dwu-etapowy proces optymalizacji budowania potoków dla uczenia maszynowego. Proces ten zilustrowano na rys. 4. W szczególności, zdefiniowaliśmy polityki alokacji czasu pomiędzy potok danych a algorytm budowania modelu. Pokazaliśmy, że często korzystniejsze jest przeznaczenie większej części czasu na konstruowanie potoku, niż na sam algorytm, oraz że polityki adaptacyjne podziału czasu pomiędzy potok danych a algorytm budowania modelu są lepsze niż polityki statycznego podziału.

Zaproponowane podejście zostało ocenione eksperymentalnie na wielu zbiorach danych i wielu potoków. Średnio, dla wszystkich testowanych zbiorów danych i metod, dla 20 potoków (0,42% przestrzeni przeszukiwania), zautomatyzowany proces był w stanie zmniejszyć błąd o 58,16% w porównaniu z podejściem, w którym cały dostępny czas został przeznaczony wyłącznie na strojenie hiper-parametrów.



[1] Pojedynczy potok przekształca cały zbiór danych podczas każdej iteracji.
[2] Wyjście $y_{t,k}$ pętli wewnętrznej jest miarą poprawności (np. walidacja krzyżowa).
[3] Pętla wewnętrzna jest inicjowana z poprzednią najlepszą konfiguracją (a priori).
[4] W t iteracji, wewnętrzna pętla zwraca najlepszą predykcję i konfigurację.
[5] f_M zwraca najbardziej korzystną konfigurację w odniesieniu do najlepszej osiągalnej metryki.
[6] Cały proces zwraca najlepszą konfigurację do wykorzystania w praktyce.

Figure 4: Dwu-etapowy proces optymalizacji budowania potoków uczenia maszynowego.

Przykładowo, wyniki dla lasu losowego na zbiorze danych Breast[3] pokazano na rys. 5. Widoczny z lewej strony żółty rozkład konfiguracji badanych przez algorytm jest przekrzywiony w kierunku większej dokładności, wskazując, że statystycznie nasze podejście tworzy dobre potoki. Wykres po prawej pokazuje, że algorytm jest efektywny w znajdowaniu potoku, który działa prawie najlepiej w przestrzeni przeszukiwania.

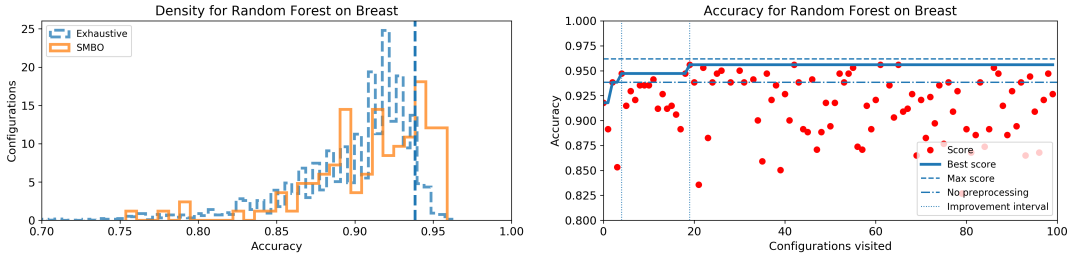


Figure 5: Przykład wyników dla lasu losowego na zbiorze danych Breast. Meta-optymalizator z większym prawdopodobieństwem próbkuje konfiguracje o wyższej dokładności.

Podsumowując, rozwiązując problem budowy i optymalizacji zautomatyzowanego potoku danych:

1. Wykazaliśmy, że wpływ konfiguracji potoku danych na dokładność klasyfikacji jest ogromny w porównaniu z wpływem wyboru hiper-parametrów i modelu.
2. Wykazaliśmy, że potoki danych można budować i konfigurować automatycznie przy użyciu istniejących meta-optymalizatorów, nawet przy ograniczonym budżecie obliczeniowym lub czasowym.

[3] https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original)

Hypergraph Case-Based Reasoning

Jako drugą kontrybucję w procesie budowy przepływu zadań (rys. 3) proponujemy zastosowanie generycznego algorytmu klasyfikacji - Hypergraph Case-Based Reasoning (HCBR) [Quemy 2019a, Quemy 2018b]. Może on przetwarzać dane dowolnego typu i uczyć się złożonych modeli. HCBR wymaga do pracy niewiele hiper-parametrów lub nie korzysta z nich w ogóle. Jak sugeruje nazwa, w HCBR, zbiór treningowy jest reprezentowany jako hipergraf. HCBR do oszacowania stopnia przypisania danego podzbioru atrybutów do klasy jest wykorzystywana partycja utworzoną przez tzw. sub-hiper-grafy.

Zaproponowana w rozprawie kontrybucja posiada kilka interesujących własności, użytecznych nie tylko w zastosowaniu w dziedzinie prawnej. W szczególności, przestrzeń modelu i reprezentacja danych jako hiper-graf zapewnia wygodny sposób wyjaśnienia każdej decyzji osobno, w oparciu o interakcje z decyzjami z przeszłości (np. postrzeganymi jako „kontrprzykłady" lub „analogie" w przypadku procesu sądowego, podobnie jak dla wnioskowania opartego na przypadkach). Ponadto, wrażliwość HCBR na hiper-parametry jest znikoma, dzięki czemu czasochłonne dostrajanie nie jest wymagane dla użytkownika końcowego. Hiper-parametry można natomiast wykorzystać do kontroli ryzyka związanego z prognozą, lepiej dostosowaną do potrzeb reprezentowanych przez konkretną dziedzinę (np. sędzia woli podejmować mniejsze ryzyko decyzji fałszywie pozytywnych, tj. wysyłania niewinnych do więzienia, podczas gdy lekarz woli podejmować mniejsze ryzyko fałszywie negatywnych, tj. niewykrycia nowotworu). Wreszcie, HCBR nie zakłada żadnej metryki w przestrzeni atrybutów (ang. feature space). Na działanie algorytmu nie wpływa reprezentacja atrybutu i może on pracować z niekompletnymi lub nieustrukturyzowanymi zbiorami danych.

HCBR został zaimplementowany w różnych wariantach (użyto do tego C++) [4] i oceniony eksperymentalnie. Eksperymenty pokazały że:

- zaproponowany przepływ zadań działa tak samo dobrze jak zbudowany w oparciu o standardowe metody, dla kilku referencyjnych zbiorów nieustrukturyzowanych danych;
- zaproponowane rozwiązanie sprawdza się lepiej niż rozwiązania konkurencyjne [Aletras 2016] w zakresie przewidywania decyzji Europejskiego Trybunału Praw Człowieka;
- HCBR średnio osiąga lepsze wyniki przy braku wiedzy specjalistycznej w porównaniu z 9 innymi uznanymi metodami: AdaBoost, k-Nearest Neighbors, Linear SVM, Radius-Based Function (RBF) SVM, Decision Tree, Random Forest, Neural Network i Quadratic Discriminant Analysis (QDA).

Otwarte repozytorium Europejskiego Trybunału Praw Człowieka

Jak wspomniano, w ramach rozprawy opracowano kompleksowe podejście do budowania przepływów danych, dla zastosowań w dziedzinie prawnej. W celu oceny jego działania niezbędnym było zbudowanie repozytorium danych sądowniczych. Repozytorium to integruje dane z Europejskiego Trybunału Praw Człowieka. Trybunał publikuje dokumenty

[4] https://github.com/aquemy/HCBR

związane ze sprawami sądowymi w języku naturalnym. Aktualnie dostępnych jest ponad 50 000 decyzji, gromadzonych od czasu utworzenia Trybunału. Oryginalne dane są dostępne w kilku formatach, min., tabelarycznym, JSON bez elementów zagnieżdżonych, CSV. W ramach projektu, z dostępnych dokumentów wyroków wyodrębniliśmy i zunifikowaliśmy standardowe atrybuty opisowe (ang. descriptive features), tworząc: (1) relacyjną bazę danych zawierającą sprawy sądowe i meta-dane o tych sprawach i (2) złożoną reprezentację *bag of words* z wyroków sądowych (uporządkowaną według paragrafów). Wstępne przetworzenie oryginalnych dokumentów (potok przygotowania danych) zostało przeprowadzone za pomocą algorytmu *entity matching* dostępnego w IBM Watson Services.

W celu zapewnienia powtarzalności zaprojektowanego przepływu danych i umożliwienia oceny jakości powstałych danych:

- każda wersja zbiorów danych jest wersjowana i publicznie dostępna, w tym także pliki pośrednie w celu zapewnienia tzw. data lineage;
- integralność procesu i wytworzonych danych jest dokładnie dokumentowana;
- skrypty do pobierania nieprzetworzonych dokumentów i tworzenia zbiorów danych są wersjowane i ogólnie dostępne;
- żadne dane nie są przetwarzane ręcznie na żadnym etapie konstruowania przepływu danych.

W celu przetestowania mocy predykcyjnej zbudowanego repozytorium, przeprowadziliśmy wiele eksperymentów, m.in., porównując 13 standardowych algorytmów uczenia maszynowego do klasyfikacji pod względem kilku wskaźników wydajności. Otrzymane wyniki dla zbiorów danych binarnych cechują się **dokładnością** (ang. accuracy) w zakresie od 75,86% do 98,32% i średnią 96,45%. Ponadto, eksperymenty pokazały, że niektóre atrybuty nadają się lepiej do predykcji decyzji sądowniczych niż inne. W szczególności stwierdziliśmy, że atrybuty tekstowe (ang. textual features) dobrze nadają się do przewidywania (binarnego) wyniku. Jednak po raz pierwszy pokazaliśmy, że nie są one tak dobre jak atrybuty czysto opisowe (ang. descriptive features) do określenia jakiego artykułu dotyczy dany przypadek sądowy.

V. Podsumowanie

Głównym celem niniejszej rozprawy było zapewnienie nowego, wydajnego podejścia do procesu budowy kompleksowego przepływu zadań dla problemu klasyfikacji oraz weryfikacja opracowanego podejścia w zastosowaniach klasyfikacji dokumentów prawnych. Główne kontrybucje rozprawy obejmują:

- ogólne podejście do automatycznej budowy i optymalizacji potoków danych przy użyciu standardowych technik meta-optymalizacji [Quemy 2020a, Quemy 2019b];
- nowy ogólny model matematyczny do klasyfikacji w przestrzeni nieustrukturyzowanych, zwany Hypergraph Case-Based Reasoning [Quemy 2019a, Quemy 2018b];

- zbudowanie wysokiej jakości repozytorium danych prawnych, dostępnego dla społeczności zajmującej się analizą danych i uczeniem maszynowym w dziedzinie prawnej [Quemy 2020c, Quemy 2020b].