Załącznik 2a: Autoreferat

dr inż. Dariusz Brzeziński

1. Imię i nazwisko

Dariusz Brzeziński

2. Posiadane stopnie i tytuły naukowe

1. Stopień doktora nauk technicznych – Politechnika Poznańska, Wydział Informatyki; dyscyplina: informatyka, 22 września 2015 r., „Block-based and Online Ensembles for Concept-drifting Data Streams", promotor: dr hab. inż. Jerzy Stefanowski, prof. nadzw.:
    - doktorat wyróżniony przez Radę Wydziału Informatyki,
    - osiągnięcia będące podstawą nadania stopnia doktora nagrodzone Nagrodą Ministerstwa Nauki i Szkolnictwa Wyższego za osiągnięcia naukowe II stopnia,
    - nagroda Polskiego Stowarzyszenia Sztucznej Inteligencji za najlepszą polską pracę doktorską ze sztucznej inteligencji.
2. Tytuł zawodowy magistra inżyniera – Politechnika Poznańska, Wydział Informatyki; kierunek: informatyka; specjalizacja: technologie przetwarzania danych, 2 września 2010 r., „Mining Data Streams with Concept Drift", promotor: dr hab. inż. Jerzy Stefanowski, prof. nadzw.:
    - medal „summa cum laude",
    - II nagroda w XXVII Ogólnopolskim Konkursie Polskiego Towarzystwa Informatycznego na najlepsze prace magisterskie z informatyki.
3. Tytuł zawodowy inżyniera – Politechnika Poznańska, Wydział Informatyki; kierunek: informatyka, 16 lutego 2009 r., „System wspomagania zarządzania zakładem opieki zdrowotnej z wykorzystaniem hurtowni danych", promotor: dr inż. Jacek Kobusiński.

3. Dotychczasowe zatrudnienie w jednostkach naukowych

1. Instytut Informatyki, Wydział Informatyki, Politechnika Poznańska,
   adiunkt, 1 etat, 1.10.2015 – obecnie.
2. Instytut Chemii Bioorganicznej Polskiej Akademii Nauk, Poznań,
   adiunkt, 0,3 etatu, 1.12.2018 – obecnie.
3. Instytut Informatyki, Wydział Informatyki, Politechnika Poznańska,
   asystent, 1 etat, 1.10.2010 – 30.09.2015.
4. Instytut Informatyki, Wydział Informatyki, Politechnika Poznańska,
   programista w projekcie Eskulap, umowa zlecenie, 1.07.2008 – 31.12.2015.

4. Wskazanie osiągnięcia naukowego wynikającego z ustawy o stopniach naukowych i tytule naukowym oraz o stopniach i tytule w zakresie sztuki

4.1. Tytuł osiągnięcia naukowego

Analiza miar oceny klasyfikatorów dla danych strumieniowych i niezbalansowanych

4.2. Lista prac wchodzących w skład osiągnięcia naukowego

[A1] Brzeziński, D., Stefanowski, J., Susmaga, R., Szczęch, I., 2019, *On the Dynamics of Classifier Performance Measures for Imbalanced and Streaming Data*. IEEE Transactions on Neural Networks and Learning Systems, DOI: 10.1109/TNNLS.2019.2899061. [**MNiSW 2016: 45 pkt.**, **IF 2017: 7,982**]

[A2] Brzeziński, D., Stefanowski, J., Susmaga, R., Szczęch, I., 2018, *Visual-based analysis of classification measures and their properties for class imbalanced problems*. Information Sciences, 462, 242-261. [**MNiSW 2016: 45 pkt.**, **IF 2017: 4,305**]

[A3] Brzeziński D., Stefanowski J., 2018, *Ensemble Classifiers for Imbalanced and Evolving Data Streams*. Data Mining in Time Series and Streaming Databases, World Scientific, pp. 44-68.

[A4] Brzeziński D., Stefanowski J., Susmaga R., Szczęch I., 2017, *Tetrahedron: Barycentric Measure Visualizer*. Machine Learning and Knowledge Discovery in Databases, Proceedings of ECML PKDD 2017, Part III, LNCS 10536, pp.419-422. [**Praca indeksowana w Web of Science**]

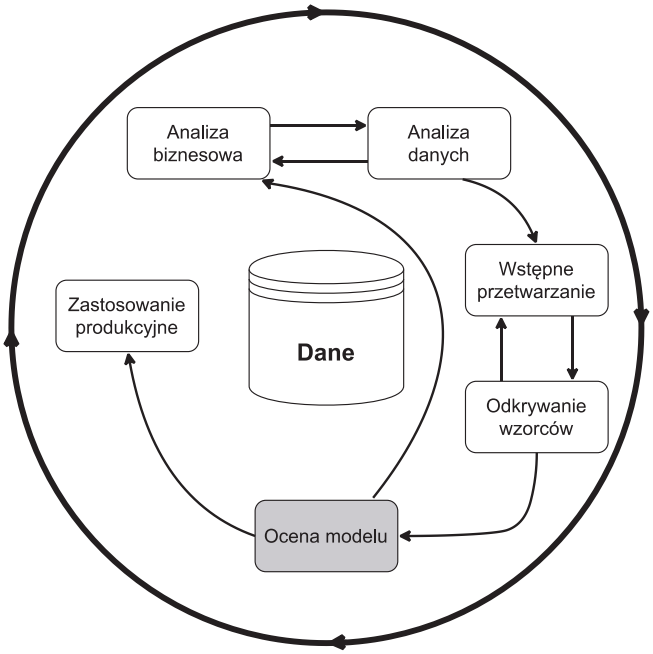
[A5] Brzeziński D., Stefanowski J., 2017, *Prequential AUC: Properties of the Area Under the ROC Curve for Data Streams with Concept Drift*. Knowledge and Information Systems, 52(2), 531-562. [**MNiSW 2016: 35 pkt.**, **IF 2016: 2,247**]

[A6] Lango M., Brzeziński D., Firlik S., Stefanowski J., 2017, *Discovering Minority Sub-clusters and Local Difficulty Factors from Imbalanced Data*. Discovery Science, LNCS 10558, pp. 324-339. [**Praca indeksowana w Web of Science**]

[A7] Brzeziński D., Stefanowski J., 2016, *Ensemble Diversity in Evolving Data Streams*. Discovery Science, LNCS 9956, pp. 229-244. [**Praca indeksowana w Web of Science**]

4.3. Omówienie celu naukowego ww. prac oraz osiągniętych wyników wraz z omówieniem ich ewentualnego wykorzystania

Cykl publikacji składających się na osiągnięcie naukowe dotyczy nurtu informatyki zawiązanego z eksploracją danych i uczeniem maszynowym. *Eksploracja danych* jest tu rozumiana jako iteracyjny, wieloetapowy proces, którego celem jest odkrywanie nowych, potencjalnie użytecznych, wzorców z danych [Fay99, Cha00]. Podstawowymi etapami tego procesu są: określenie celu eksploracji, selekcja danych, wstępne przetwarzanie danych, odkrywanie wzorców i ich ocena [Mar10] (Rysunek 1). Na etapie odkrywania wzorców wykorzystywane są często algorytmy *uczenia maszynowego*, działu informatyki zajmującego się tworzeniem algorytmów, które uczą się wnioskować o nowych danych na podstawie pewnego zbioru danych historycznych [Dom12]. Jednak kluczowym etapem eksploracji danych jest ocena odkrytych wzorców, gdyż wynik tego etapu steruje całym procesem – decyduje o tym czy cel eksploracji został osiągnięty, pokazuje które metody wstępnego przetwarzania najlepiej się sprawdzają, pozwala poprawnie sparametryzować i określić przydatność algorytmów uczenia maszynowego. Przedstawiany cykl prac koncentruje się właśnie na etapie oceny, w szczególności algorytmów uczenia maszynowego, w procesie eksploracji danych.



Rysunek 1. Proces eksploracji danych zgodnie ze standardem CRISP-DM [Cha00].

Algorytmy uczenia maszynowego mogą odkrywać wzorce reprezentowane w różnej postaci, na przykład w formie asocjacji, modeli klasyfikacyjnych, skupień czy regresji wielowymiarowej [Pia91]. Przedstawiane prace koncentrują się na ocenie *klasyfikatorów*, czyli algorytmów, które na podstawie zbioru danych uczących tworzą ogólną funkcję (model klasyfikacyjny), która potrafi przypisać nowy obiekt do jednej z wielu predefiniowanych klas. Do porównania klasyfikatorów można stosować takie kryteria oceny jak efektywność, odporność modelu, skalowalność, zróżnicowanie czy interpretowalność [Tan06]. Jednak obowiązkowym elementem oceny każdego klasyfikatora jest zbadanie jego zdolności predykcyjnych. W literaturze [Jap11] zostało zaproponowanych wiele miar, które starają się oszacować trafność przewidywania klas obiektów testowych, niedostępnych dla klasyfikatora na etapie tworzenia modelu. Mnogość miar zaproponowanych do tego zadania sprawia, że proces wyboru miary staje się nietrywialny, a co za tym idzie rodzi pokusę upraszczania wyboru przez ograniczanie się do kryterium popularności miary. Głównym celem niniejszego cyklu prac jest analiza i konstrukcja własności pozwalających podjąć świadomą decyzję o zastosowaniu wybranej miary oceny do konkretnego problemu klasyfikacji.

W prezentowanym cyklu analiza miar została zawężona do problemów, w których ocena klasyfikatora jest dodatkowo skomplikowana przez trudności występujące w danych. Dwie takie trudności, na których skupia się niniejszy cykl to niezbalansowanie klas i uczenie się ze zmiennych strumieni danych.

Niezbalansowanie klas to sytuacja, w której w zbiorze uczącym jedna z klas (zwana klasą mniejszościową) posiada znacząco mniej przykładów uczących niż inna klasa (zwana klasą większościową) [He09]. Wykorzystanie na niezbalansowanym zbiorze danych standardowej miary oceny jaką jest trafność klasyfikacji, powoduje promowanie trywialnych klasyfikatorów, które ignorują klasę mniejszościową. W tym kontekście szczególnie ważny jest dobór takiej miary oceny, która potrafi skutecznie promować nietrywialne klasyfikatory niezależnie od poziomu niezbalansowania klas. Ponadto wraz z niezbalansowaniem klas występują zwykle *lokalne czynniki trudności*, czyli cechy utrudniające klasyfikację, które charakteryzują tylko część przykładów uczących. W literaturze wyróżnia się takie lokalne czynniki trudności jak podział przykładów klasy mniejszościowej na wiele skupisk, nakładanie się klas czy występowanie wartości odstających [Nap16]. Określenie czy lokalne czynniki trudności występują w danych wymaga dedykowanych algorytmów i miar oceny.

Drugą z analizowanych trudności w uczeniu i ocenie klasyfikatorów jest zmienność definicji klas w strumieniach danych. *Strumień danych* to sekwencja przykładów uczących (np. pakietów sieciowych, wiadomości e-mail, odczytów z czujników), które napływają w sposób asynchroniczny ze zmienną intensywnością. Strumienie, podobnie jak inne duże wolumeny danych, mogą być przedmiotem eksploracji danych, a w szczególności ich klasyfikacji. Jednakże, ze względu na rozmiar i intensywność strumieni danych, ich klasyfikacja musi spełniać dodatkowe ograniczenia na czas przetwarzania oraz wykorzystywaną pamięć operacyjną. Co więcej, w przypadku klasyfikacji, strumienie danych podlegają zjawisku *dryftu* (ang. *concept drift*), czyli zmianom rozkładów prawdopodobieństwa definicji klas w źródle danych generującym nowe przykłady. Wspomniane ograniczenia na czas przetwarzania i dostępną pamięć sprawiają, że klasyfikatory dla strumieni danych są zwykle oceniane w sposób przyrostowy. Ponadto zjawisko dryftu powoduje konieczność wielokrotnej oceny klasyfikatora w celu wykrycia zmian zachodzących w strumieniu, a tym samym zmian zachodzących w zdolnościach predykcyjnych klasyfikatora. Dobór miary oceny klasyfikatora dla strumieni danych wymaga zatem wzięcia pod uwagę dodatkowych czynników w porównaniu do statycznego zbioru danych. Warto zauważyć, że wspomniane trudności nie wykluczają się i można analizować miary oceny dla strumieni, w których występuje niezbalansowanie danych. Takie połączenie wprowadza dodatkowe wyzwania dla klasyfikatorów, gdyż dane mogą zmieniać swój poziom niezbalansowania w czasie. Co więcej, metody próbujące wykrywać zmiany w strumieniu muszą być odporne na niezbalansowanie klas. Zatem połączenie niezbalansowania danych z przetwarzaniem strumieniowym rodzi nowe trudności w wyborze i analizie własności miar oceny klasyfikatorów.

Zgodnie z powyższą listą trudności, omówienie celów i wyników naukowych prac wchodzących w skład prezentowanego cyklu zostanie podzielone na dwie części obejmujące:

- analizę miar oceny klasyfikatorów oraz trudności dla danych niezbalansowanych [A1, A2, A4, A6];
- analizę i projektowanie miar oceny klasyfikatorów dla strumieni danych [A1, A3, A5, A7].

Analiza miar oceny klasyfikatorów oraz trudności dla danych niezbalansowanych

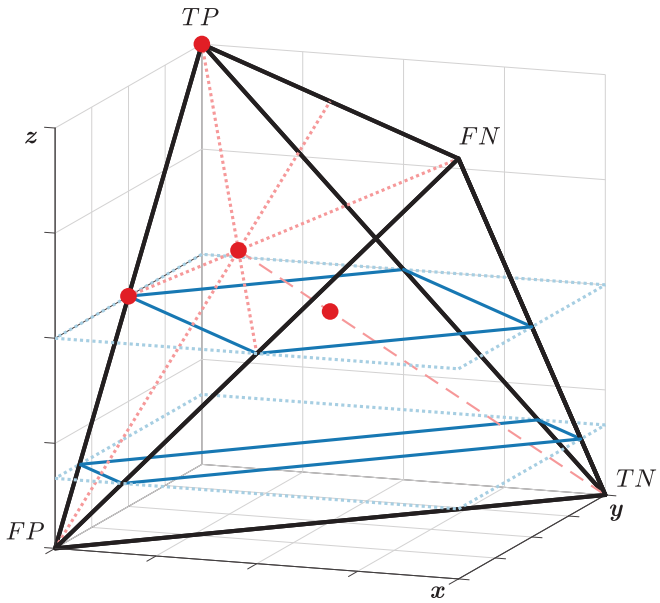
W większości problemów, w których występuje niezbalansowanie klas, miary oceny klasyfikatorów powinny skupiać się na klasie mniejszościowej [He13]. Takie miary są z reguły zdefiniowane jako funkcje binarnych macierzy pomyłek, gdzie klasa mniejszościowa nazywana jest *pozytywną* (*P*), a większościowa *negatywną* (*N*). W przypadku problemów klasyfikacji wieloklasowej, miary oceny są albo liczone dla różnych par klas osobno albo klasy nie będące klasą pozytywną są agregowane do jednej klasy negatywnej. Tabela 1 przedstawia binarną macierz pomyłek dla zbioru danych o rozmiarze *n*, której wpisy *TP*, *TN*, *FN*, *FP* oznaczają kolejno liczbę predykcji poprawnych pozytywnych, poprawnych negatywnych, niepoprawnie negatywnych i niepoprawnie pozytywnych. Dla tak zdefiniowanego problemu, *poziomem niezbalansowania* będziemy nazywać stosunek liczby przykładów pozytywnych do negatywnych (*P/N*).

Tabela 1. Macierz pomyłek dla problemu klasyfikacji binarnej.

Przewidziana / Rzeczywista	Pozytywna	Negatywna	łącznie
Pozytywna	*TP*	*FN*	*P*
Negatywna	*FP*	*TN*	*N*
łącznie	$\widehat{P}$	$\widehat{N}$	*n*

Ponieważ formalna analiza miar oceny klasyfikatorów jest pracą bardzo czasochłonną i wymagającą potencjalnie zaawansowanego aparatu matematycznego, w pracy [A2] zaproponowano oryginalną metodę wizualizacji miar oceny klasyfikatorów. Zaproponowana metoda wizualizacji jest zainspirowana podejściem znanym z analizy miar konfirmacji reguł [Sus15] i pozwala na łatwe badanie analizowanych miar w całych ich dziedzinach. Takie całościowe spojrzenie na wartości i dziedzinę miary jest bardzo istotne, bo pozwala na wnioskowanie o zachowaniu się miary w dowolnej sytuacji, w sposób niezależny od konkretnego zbioru danych. Jest to ważna cecha tego podejścia, gdyż tylko dostęp do informacji o zachowaniu się we wszystkich obszarach dziedziny pozwala na formułowanie ogólnych wniosków o miarach.

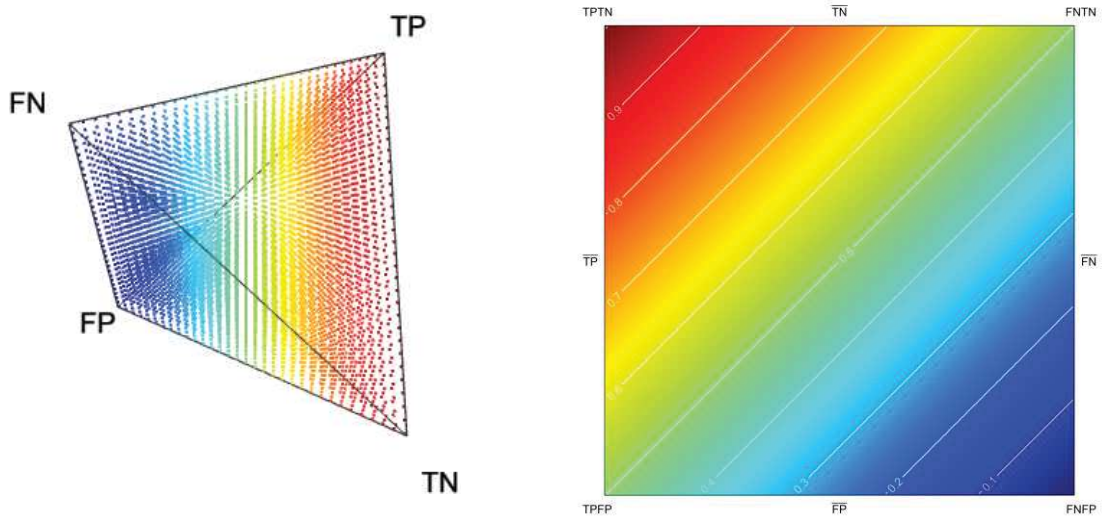
Proponowana metoda wizualizacji bazuje na tym, że większość miar oceny klasyfikatorów, w tym miar dla niezbalansowanych danych, zdefiniowana jest na wpisach *TP*, *TN*, *FN*, *FP* z binarnej macierzy pomyłek. Ponieważ suma tych czterech wartości jest stała dla danego zbioru danych o rozmiarze n, po odpowiednim przeskalowaniu możliwe jest reprezentowanie pierwotnie czterowymiarowej dziedziny miar oceny w postaci czworościanu w barycentrycznym układzie współrzędnych [Flo06]. Punkty wewnątrz wynikowego czworościanu (Rysunek 2) reprezentują macierze pomyłek o wartościach *TP*, *TN*, *FN*, *FP* proporcjonalnych do odległości między środkiem czworościanu a czterema wierzchołkami odpowiadającymi czterem wpisom w macierzy pomyłek. Przykładowo, punkt leżący na wierzchołku *TP* odpowiada macierzy pomyłek $\begin{bmatrix} n & 0 \\ 0 & 0 \end{bmatrix}$ a punkt leżący na środku ściany TP-FP-FN odpowiada macierzy pomyłek $\begin{bmatrix} n/3 & n/3 \\ n/3 & 0 \end{bmatrix}$. Co ważne, prostokątne przecięcia przedstawione na Rysunku 2 odpowiadają różnym poziomom niezbalansowania danych. Oznacza to, że zaproponowana w [A2] metoda wizualizacji nadaje się do analizy zachowania miar oceny przy różnych poziomach niezbalansowania danych.



Rysunek 2. Reprezentacja macierzy pomyłek w postaci czworościanu w barycentrycznym układzie współrzędnych. Czerwone punkty reprezentują cztery przykładowe macierze pomyłek. Niebieskie przecięcia odpowiadają zbiorom macierzy pomyłek dla dwóch poziomów niezbalansowania danych – 1:1 (górne przecięcie), 1:6 (dolne przecięcie).

Ponieważ miary oceny klasyfikatorów są zdefiniowane na podstawie macierzy pomyłek, można każdemu punktowi w czworościanie przyporządkować wartość miary reprezentowaną przez kolor. To z kolei pozwala analizować wartości miary oceny dla wszystkich możliwych wartości macierzy pomyłek, czyli dla całej dziedziny wartości. W efekcie metoda wykorzystywana w [A2] wizualizuje

miary w postaci kolorowych czworościanów, takich jak przedstawiony na Rysunku 3a. Wizualizując w podobny sposób przecięcie czworościanu (Rysunek 3b) można analizować wartości miary dla wybranego poziomu niezbalansowania danych. Tym samym proponowane przecięcia pozwalają na analizę podobną do tzw. ROC space [Fla03], z tą różnicą że dane w proponowanej wizualizacji nie są w żaden sposób przeskalowane (a tym samym zniekształcone) i proporcje długości boków proponowanego przecięcia ($\overline{TN}$:$\overline{TP}$) odpowiadają wprost proporcjom klas (P:N).



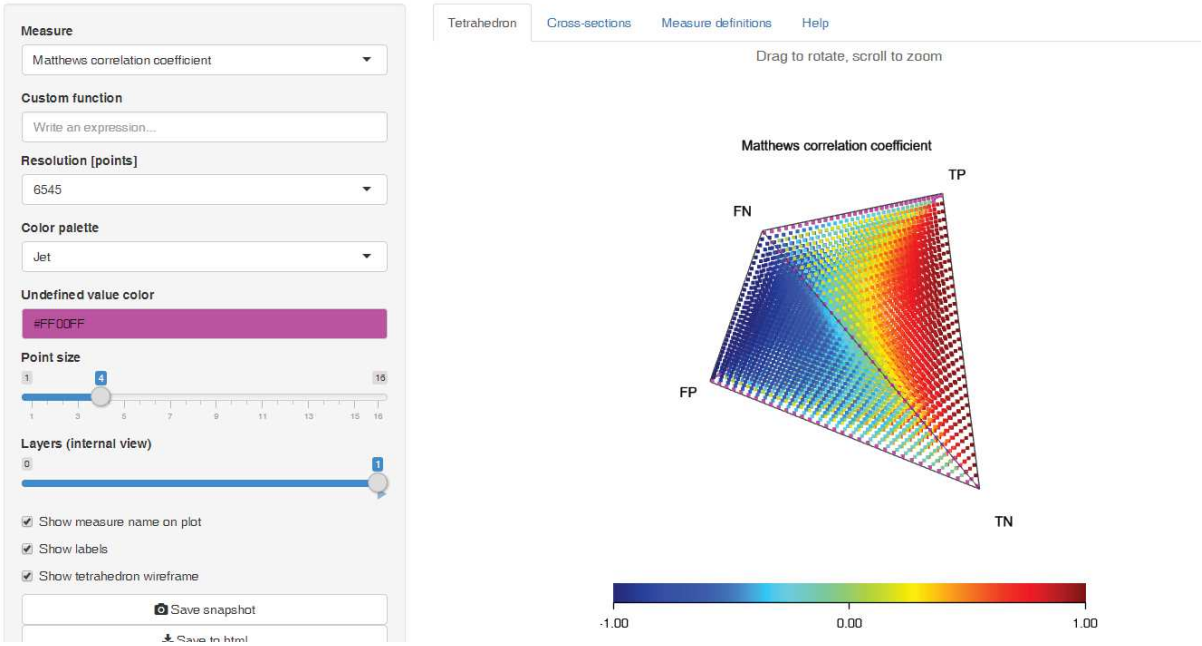
(a) Czworościan (b) Przekrój

Rysunek 3. Miara trafności klasyfikacji (ang. accuracy) reprezentowana jako (a) czworościan reprezentujący wszystkie możliwe macierze pomyłek, (b) prostokątne przecięcie przedstawiające wartości miary dla wybranego poziomu niezbalansowania danych (w tym przypadku zbioru o równolicznych klasach).

Korzystając z przedstawionej wizualizacji miar, w pracy [A2] zaproponowano dziesięć właściwości różnicujących miary oceny klasyfikatorów, w szczególności miary przeznaczone do danych niezbalansowanych. Proponowane właściwości obejmowały analizę wartości minimalnych, maksymalnych, monotoniczności, symetrii i występowania wartości nieokreślonych. Co ważne wszystkie właściwości można zweryfikować analizując czworościan wybranej miary lub kolejne jego przecięcia. Efektywność proponowanej metody analizy wizualnej zademonstrowano porównując 22 miary (m.in. accuracy, balanced accuracy, F1-score, G-mean, Kappa, precision, recall, specificity) za pomocą wspomnianych dziesięciu właściwości. Zauważono między innymi, że miara Optimized Precision [Ran06], która została zaprojektowana z myślą o danych niezbalansowanych, nie spełnia właściwości monotoniczności poprawnych predykcji. Oznacza to, że wartość miary Optimized Precision może maleć wraz ze wzrostem ułamka poprawnych predykcji. Pokazano również, że miary takie jak F1-score czy Mathews Correlation Coefficient mogą, w niektórych przypadkach, promować klasyfikatory, które ignorują klasę mniejszościową.

Oprócz typowych bezparametrowych miar oceny, w pracy [A2] przeanalizowano również miary, których działanie można dostroić za pomocą parametru. Przykładem takiej miary jest F_β, gdzie parametr β steruje wagami w średniej harmonicznej miar precision i recall. W ramach analizy takich miar w kontekście zaproponowanych dziesięciu właściwości, określono jakie wartości graniczne parametrów pozwalają (lub uniemożliwiają) spełnienie wybranych właściwości. Pokazano na przykład, że parametr β powinien być większy lub równy stosunkowi klasy negatywnej do pozytywnej (N/P), aby miara F_β nie faworyzowała klasy większościowej. Dwa podobne twierdzenia udowodniono dla miary IBA_α [Gar14].

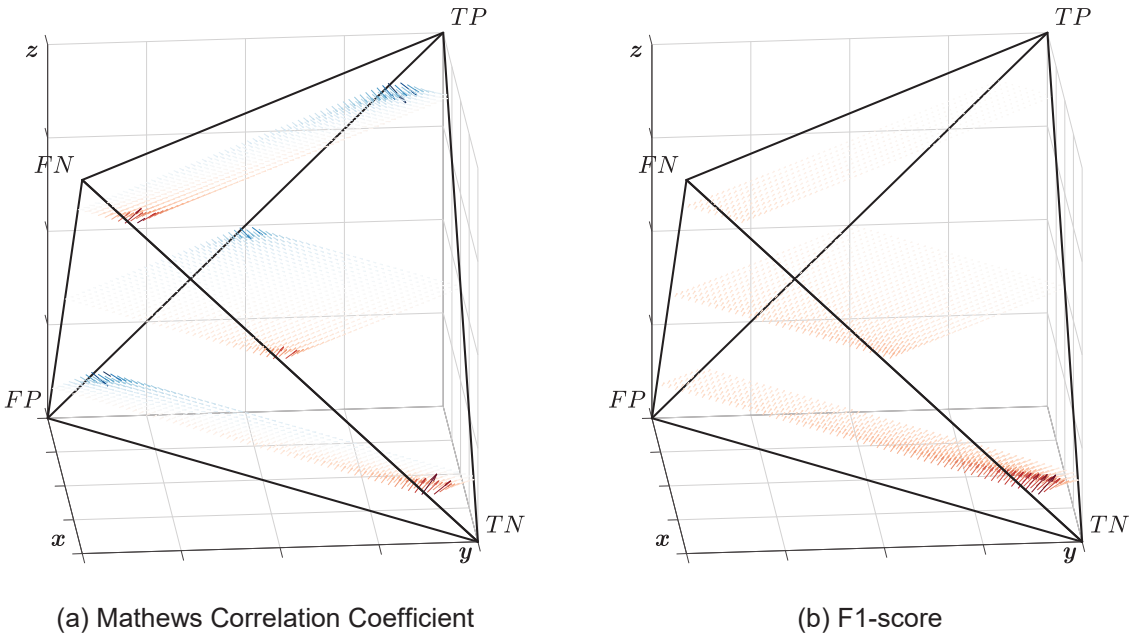
Aby umożliwić podobne analizy wszystkim zainteresowanym, oraz aby ułatwić projektowanie nowych miar oceny o wybranych właściwościach, w pracy [A4] zaproponowano narzędzie o nazwie Tetrahedron do wizualizacji miar w postaci czworościanu w układzie barycentrycznym. Tetrahedron umożliwia tworzenie interaktywnych trójwymiarowych wizualizacji miar w przeglądarce za pomocą standardu WebGL [Par12]. Narzędzie zawiera definicje 86 predefiniowanych miar, pozwala na wybór palety kolorów, umożliwia definiowania własnych bezparametrowych oraz parametrycznych miar i wizualizuje przekroje dla zdefiniowanych przez użytkownika poziomów niezbalansowania. Tetrahedron pozwala również animować wizualizacje aby ułatwić np. badanie zachowania się miary przy zmieniającym się poziomie niezbalansowania klas. Rysunek 4 przedstawia ekran główny aplikacji, której wersja demonstracyjna jest dostępna pod adresem: https://dabrze.shinyapps.io/Tetrahedron/.



Rysunek 4. Ekran główny aplikacji Tetrahedron pozwalającej wizualizować i projektować miary oceny za pomocą czworościanów w układzie barycentrycznym.

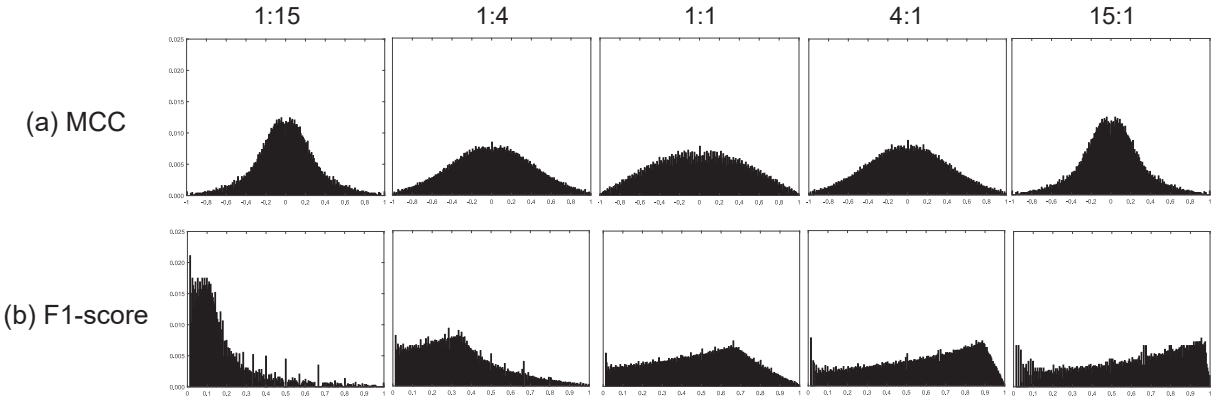
W pracy [A1] poszerzono badania własności miar o analizę ich gradientów. Gradient określa kierunek najszybszej zmiany wartości funkcji. W przypadku miary oceny gradient pokazuje zatem jakie zmiany w macierzy pomyłek powodują najszybszy wzrost wartości miary. W pracy [A1] zaproponowano wizualizację gradientów miar za pomocą strzałek w układzie barycentrycznym (Rysunek 5) wykorzystując kolor do określania kierunku zgodnego ze wzrostem bądź spadkiem liczby przykładów klasy mniejszościowej.

Zaproponowana analiza pozwala określić kierunek i prędkość zmian wartości różnych miar. Widać między innymi w jakim stopniu można osiągnąć wyższą wartość miary zmieniając proporcje klas w zbiorze danych. Taka informacja może zostać wykorzystana na przykład do generowania sztucznych przykładów uczących z klasy mniejszościowej lub do określania jak modyfikować etykiety przykładów na etapie uczenia z niezbalansowanych danych [Nap10]. Praca [A1] pokazała, które miary i w jakim stopniu byłyby podatne na takie zmiany. Przeprowadzona analiza miar uwidoczniła również jak miary się zachowują w przypadkach, gdy proporcje klas zmieniają się dynamicznie w czasie.



(a) Mathews Correlation Coefficient (b) F1-score

Rysunek 5. Gradient miary (a) Mathews Correlation Coefficient, (b) F1-score dla trzech różnych poziomów niezbalansowania danych. Czerwony kolor oznacza gradient w kierunku zgodnym ze wzrostem liczby przykładów pozytywnych. Kolor niebieski oznacza gradient zgodny ze wzrostem liczby przykładów negatywnych.



Rysunek 6. Funkcje masy prawdopodobieństwa miary (a) Mathews Correlation Coefficient, (b) F1-score dla pięciu różnych poziomów niezbalansowania danych. Poziom niezbalansowania od lewej do prawej (P:N): 1:15, 1:4, 1:1, 4:1, 15:1. Na osi x wartość miary [0-1]. Na osi y prawdopodobieństwo określonej wartości miary.

Różnice w zachowaniu miar przy zmieniającym się poziomie niezbalansowania danych wykazano również analizując ich funkcje masy prawdopodobieństwa. Funkcje masy prawdopodobieństwa miar zwizualizowane w postaci histogramów pokazują ich rozkład wartości. Można zatem za ich pomocą określić jak prawdopodobne jest uzyskanie konkretnej wartości miary przy wybranym poziomie niezbalansowania. Jak wykazała analiza przeprowadzona w pracy [A1] miary mogą znacząco zmieniać swoje funkcje masy prawdopodobieństwa wraz ze zmianą proporcji klas (Rysunek 6).

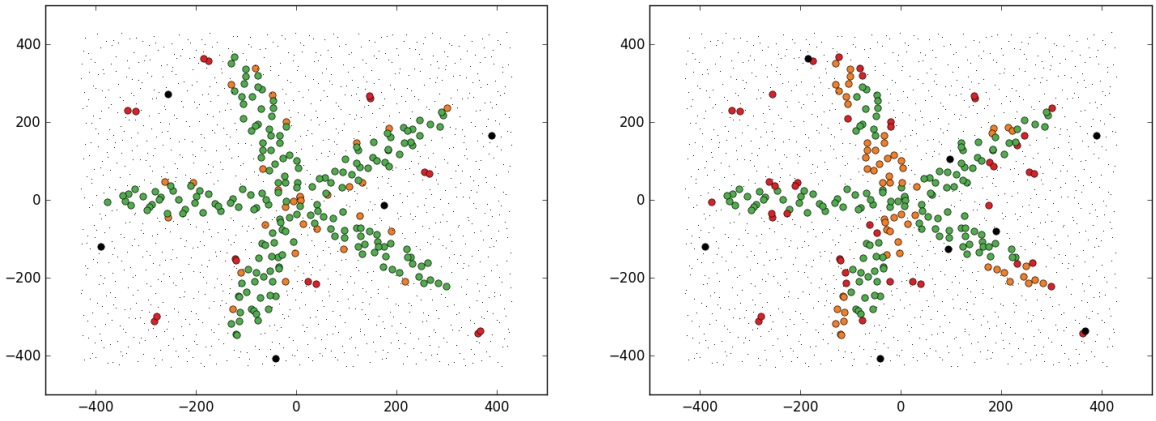
Zaobserwowane różnice w funkcjach masy prawdopodobieństwa sugerują, że wartości danej miary powinny być różnie interpretowane w zależności od poziomu niezbalansowania danych. O wiele bardziej prawdopodobne jest uzyskanie wartości F1-score≥0,5, gdy zbiór danych jest zbalansowany (1:1) niż gdy w zbiorze danych testowych jest mniej przykładów klasy pozytywnej (1:15) (Rysunek 6). Jednym z istotnych wyników pracy [A1] jest metoda normalizująca miarę w taki sposób, aby jej

interpretacja nie zależała od poziomu niezbalansowania zbioru danych. Dla wybranej miary *M* i zbioru danych o poziomie niezbalansowania *ir*, wartość miary *x* jest normalizowana za pomocą funkcji:

$$N(M, ir, x) = \sum_{t \leq x} pmf_{M,ir}(t)$$

gdzie $pmf_{M,ir}()$ to funkcja masy prawdopodobieństwa miary *M* przy poziomie niezbalansowania *ir*. Seria eksperymentów porównujących zachowanie różnych miar przed i po normalizacji na 12 zbiorach danych z repozytorium UCI [Lic13] wykazała, że normalizacja miary potrafi istotnie wpłynąć na ocenę działania klasyfikatora.

W pracy [A6] przedstawiono inne podejście do analizy problemu niezbalansowania – algorytm określający lokalne czynniki trudności w zbiorze danych. Zaproponowany algorytm, o nazwie ImGrid, automatycznie grupuje przykłady klasy mniejszościowej i wykrywa takie cechy w zbiorze danych jak: bezpieczne przykłady mniejszościowe, obserwacje na granicy dwóch klas, rzadkie skupienia mniejszościowe i wartości odstające [Nap16]. ImGrid oferuje tym samym kompromis między algorytmami analizy skupisk a podejściami opartymi o analizę sąsiedztwa przykładów. W efekcie proponowany algorytm potrafi dokładnie określić regiony w przestrzeni atrybutów o wybranej charakterystyce lokalnych czynników trudności (Rysunek 7). Co ważne w kontekście referowanego cyklu prac, wyniki zaprezentowane w [A6] otwierają drogę do tworzenia nowych miar kwantyfikujących lokalne czynniki trudności w niezbalansowanych zbiorach danych.



(a) Wygenerowane typy przykładów (b) Wynik działa algorytmu ImGrid

Rysunek 7. Przykład działania algorytmu ImGrid na zbiorze danych z 5 skupiskami bezpiecznych przykładów mniejszościowych (zielone punkty) wraz z przykładami granicznymi (pomarańczowe) oraz serią przykładów rzadkich (czerwone) i odstających (czarne). Przykłady klasy większościowej oznaczone małymi szarymi punktami.

Analiza i projektowanie miar oceny klasyfikatorów dla strumieni danych

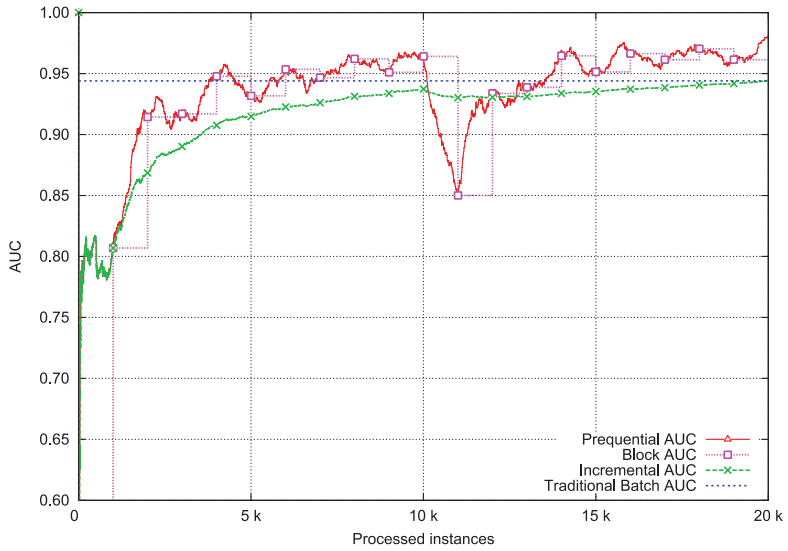
Oprócz typowej (jednokrotnej) oceny zdolności predykcyjnych klasyfikatorów na zbiorze testowym, istnieje szereg problemów, w których klasyfikator musi zostać oceniony wielokrotnie. Jednym z takich problemów jest przetwarzanie strumieni danych, gdzie nowe przykłady napływają w sposób ciągły i klasyfikator musi być oceniany na bieżąco. Z tego powodu w pracy [A7] postawiono sobie za cel analizę miar zróżnicowania klasyfikatorów dla strumieni danych.

Miary zróżnicowania klasyfikatorów to miary oceniające podobieństwo klasyfikatorów. Miary zróżnicowania są często stosowane do analizy składowych klasyfikatorów złożonych. *Klasyfikator złożony* (ang. c*lassifier ensemble)* to zespół klasyfikatorów (składowych), które wspólnie podejmują decyzję o ostatecznej predykcji. Przydatność klasyfikatorów złożonych zależy w dużej mierze od

zróżnicowania ich składowych – zespół wielu identycznych klasyfikatorów nie wnosiłby nic w stosunku do pojedynczego klasyfikatora. Choć istnieje szereg miar badających zróżnicowanie [Kun04] nie były one dotychczas analizowane w kontekście strumieni danych, gdzie ocena może zmieniać się w czasie.

Jednym z głównych oryginalnych wyników pracy [A7] jest opracowanie sposobu liczenia miar zróżnicowania w sposób przyrostowy z zapominaniem. Zaproponowano dwa sposoby liczenia sześciu różnych miar: Disagreement, Kohavi-Wolpert variance, Double fault, κ, Q i CFD [Kun04]. Pierwszy sposób wykorzystuje okno przesuwne, drugi działa w sposób w pełni przyrostowy i polega na ważeniu przykładów (mnożeniu każdego przykładu przez współczynnik) przy obliczaniu wartości miary. Wykorzystując zaproponowane przyrostowe wersje miar, przeprowadzono eksperymenty badające zróżnicowanie składowych klasyfikatorów blokowych (douczających składowe większymi partiami przykładów) i w pełni przyrostowych (douczających składowe przykład po przykładzie). Celem eksperymentów było określenie wpływu wybranej miary, typu klasyfikatora złożonego, liczby składowych i typu zmian zachodzących w strumieniu na ocenę zróżnicowania. Uzyskane wyniki pokazały, że zróżnicowanie składowych w klasyfikatorach strumieniowych jest z reguły stosunkowo niskie, lecz podatne na zmiany w czasie. Ważnym wynikiem tej analizy było również spostrzeżenie, że miara Disagreement może zostać wykorzystana do wykrywania zmian w strumieniu w sposób nienadzorowany.

Oprócz miar zróżnicowania, ocena zdolności predykcyjnych klasyfikatorów dla strumieni danych również wykonywana jest w sposób ciągły. W tym kontekście jednym z wyników prezentowanego cyklu prac jest propozycja i analiza właściwości miary Prequential AUC, czyli pola pod krzywą ROC liczonego w sposób przyrostowy z zapominaniem [A5]. Pole pod krzywą ROC (ang. *area under the Receiver Operator Characteristic curve*, w skrócie AUC) jest jedną z najpopularniejszych miar oceny dla danych niezbalansowanych. Jednakże ze względu na kosztowną procedurę obliczeniową, wykorzystanie tej miary było ograniczone do małych strumieni danych, a sama ocena nie odbywała się przyrostowo. Sposób przyrostowego obliczania AUC z zapominaniem (Prequential AUC) był pierwotnie rozważany przed doktoratem habilitanta [Brz15], ale dopiero w pracy [A5] algorytm ten został udoskonalony o obsługę predykcji o tym samym prawdopodobieństwie (tzw. remisów) i przeprowadzona została analiza właściwości proponowanej miary.



Rysunek 8. Porównanie oceny klasyfikatora w czasie za pomocą miary AUC liczonej: na całym zbiorze danych (Traditional Batch AUC), przyrostowo (Incremental AUC), blokami (Block AUC), przyrostowo z zapominaniem (Prequential AUC). Nagła zmiana w połowie strumienia pokazuje jak istotny jest element zapominania przy diagnozowaniu i wykrywaniu dryftu. Na osi x numer przykładu, na osi y wartość miary AUC.

W pracy [A5] porównano Prequential AUC z miarą AUC liczoną: na całym strumieniu, przyrostowo bez zapominania oraz blokami przykładów. Spośród wszystkich sposobów liczenia pola pod krzywą ROC, Prequential AUC jest miarą, która najlepiej nadaje się do wykrywania zmian w strumieniach danych. Wykazano również, że w przypadku stacjonarnych strumieni danych, Prequential AUC jest lepszym przybliżeniem tradycyjnej miary AUC liczonej na całym strumieniu niż AUC liczone blokami. Aby to określić, zastosowano miary zgodności (ang. *consistency*) i rozróżnialności (ang. *discriminacy*) [Hua05], które pozwoliły porównać Prequential AUC i AUC liczone blokami. W symulacji porównującej te dwie wersje miary, Prequential AUC okazało się bardziej zgodne z tradycyjnym AUC i zdolne do rozróżnienia większej liczby par klasyfikatorów.

Prequential AUC posłużyła również jako jedna z miar porównujących klasyfikatory dla niezbalansowanych danych. W pracy [A3] dokonano przeglądu klasyfikatorów złożonych zaprojektowanych do blokowego i przyrostowego uczenia się z niezbalansowanych strumieni danych. Dokonano również eksperymentalnego porównania wybranych algorytmów. W ramach eksperymentu przeanalizowano działanie miar: G-mean, recall, accuracy, Kappa, Kappa M i Prequential AUC.

Artykuły [A3, A5] są też jednymi z pierwszych prac, w których problem lokalnych czynników trudności (przykłady graniczne, rzadkie, odstające) został podniesiony w kontekście przetwarzania strumieniowego. W efekcie przeanalizowano nowe typy zmian w strumieniach danych takie jak: nagłe rozdzielenie się klasy mniejszościowej na wiele skupisk, pojawianie się małych skupisk klasy mniejszościowej, stopniowa zmiana poziomu niezbalansowania strumienia, nagła zmiana niezbalansowania strumienia. Przeprowadzenie analizy eksperymentalnej proponowanych typów zmian wymagało zaimplementowania nowych generatorów niezbalansowanych strumieni danych.

Z tematyką niezbalansowanych strumieni danych powiązana jest również praca [A1], gdzie analiza funkcji masy prawdopodobieństwa miar została odniesiona do zmian występujących w strumieniach danych. W [A1] pokazano między innymi, że zachowanie detektorów dryftu opartych o miary oceny [Gam13] zależy od niezbalansowania klas i jest to cecha wynikająca wprost z własności miar. W przyszłości ta wiedza może zostać wykorzystana do projektowania nowych miar przystosowanych do detektorów dryftu działających na strumieniach danych o dynamicznie zmieniającym się poziomie niezbalansowania klas.

Praca [A1] pokazała również, że gradienty miar mają znaczenie dla analizy niezbalansowanych strumieni danych. Wykazano, że dryft niezbalansowania strumienia ma bezpośredni wpływ na zmianę oceny klasyfikatora, a zmiana ta jest zgodna z gradientem miary oceny. To spostrzeżenie może być punktem odniesienia do projektowania nowych miar przystosowanych do analizy niezbalansowanych strumieni danych zmiennych w czasie.

Podsumowanie

Badania przedstawione w ramach cyklu publikacji [A1]-[A7] dotyczyły metodyki oceny klasyfikatorów. Dokonano w nich wnikliwej analizy istniejących miar oceny klasyfikatorów, opracowano nowe miary, a także zaproponowano szereg własności oraz metod wizualizacji przydatnych podczas wyboru miar. Wszystkie analizy brały pod uwagę sytuacje, w których ocena klasyfikatorów jest skomplikowana poprzez dodatkowe trudności w danych, takie jak niezbalansowanie klas czy zmienność charakterystyki danych w czasie.

Podstawowe elementy oryginalnego wkładu naukowego zawarte w siedmiu publikacjach wchodzących w skład przedstawionego cyklu są następujące:

- zaproponowano metodę wizualizacji miar oceny klasyfikatorów w postaci czworościanów w barycentrycznym układzie współrzędnych, umożliwiając tym samym pozyskiwanie informacji o zachowaniu się miary we wszystkich obszarach jej dziedziny [A2];
- przedstawiono sposoby wykorzystania zaproponowanej wizualizacji do badania cech charakterystycznych (m.in., ekstrema, wartości nieokreślone, monotoniczność, gradienty) miar [A1, A2];
- zaproponowano 10 właściwości miar oceny klasyfikatorów niezbalansowanych w oparciu o zaproponowaną metodę wizualizacji i dokonano szczegółowej analizy 22 miar według tych właściwości [A2];
- udowodniono trzy cechy miar parametryzowanych F_β i IBA_α [A2];
- zaimplementowano interaktywne narzędzie do wizualizacji i projektowania miar oceny w postaci czworościanu w układzie barycentrycznym [A4];
- zaproponowano analizę funkcji masy prawdopodobieństwa miar oceny dla różnych poziomów niezbalansowania danych [A1];
- pokazano zależność między poziomem niezbalansowania strumienia danych a zachowaniem się detektorów dryftu opartych o miary oceny klasyfikatora [A1];
- zaproponowano i przebadano metodę normalizacji miar oceny biorącą pod uwagę poziom niezbalansowania danych [A1];
- zaproponowano i eksperymentalnie zweryfikowano algorytm ImGrid do wykrywania skupisk klasy mniejszościowej i określania lokalnych czynników trudności przykładów uczących [A6];
- zaproponowano dwa sposoby liczenia miar zróżnicowania klasyfikatorów przyrostowo z zapominaniem [A7];
- przebadano zachowanie miar zróżnicowania klasyfikatorów w zmiennych strumieniach danych [A7];
- udoskonalono i przebadano własności miary Prequential AUC liczącej pole pod krzywą ROC w sposób przyrostowy z zapominaniem [A5];
- dokonano przeglądu klasyfikatorów i eksperymentalnego porównania miar ich oceny dla niezbalansowanych strumieni danych rozważając przy tym nowe typy zmian w strumieniu [A3];
- powiązano zmiany obserwowane w ocenie klasyfikatorów podczas dryftu niezbalansowania strumienia danych z gradientami miar oceny [A1].

5. Omówienie pozostałych osiągnięć naukowo-badawczych

5.1. Lista prac – pozostałe osiągnięcia

[O1] Kowiel, M., Brzeziński, D., Porebski, P., Shabalin, I., Jaskolski, M., Minor, W., 2019, *Automatic Recognition of Ligands in Electron Density by Machine Learning*. Bioinformatics, 35(3), 452-461. [**MNiSW 2016: 45 pkt.**, **IF 2017: 5,481**]

[O2] Gilski, M., Zhao, J., Kowiel, M., Brzeziński, D., Turner, D., Jaskolski, M., 2019, *Accurate geometrical restraints for WC base pairs*. Acta Crystallographica Section B, w druku. **[MNiSW 2016: 30 pkt., IF 2017: 6,467]**

[O3] Stefanowski J., Brzeziński D., 2017, *Stream Classification*. Encyclopedia of Machine Learning and Data Mining, Springer, pp. 1191-1199.

[O4] Piernik M., Brzeziński D., Morzy M., Morzy T., 2017, *Using Network Analysis to Improve Nearest Neighbor Classification of Non-Network Data*. Foundations of Intelligent Systems, LNCS 10352, pp. 105-115. [**Praca indeksowana w Web of Science**]

[O5] Brzeziński D., Grudziński Z., Szczęch I., 2017, *Bayesian Confirmation Measures in Rule-based Classification*. Proceedings of the 5th ECML PKDD Workshop on New Frontiers in Mining Complex Patterns, LNCS 10312, pp. 39-53.

[O6] Kowiel M., Brzeziński D., Jaskolski M., 2016, *Conformation-dependent restraints for polynucleotides: I. Clustering of the geometry of the phosphodiester group*. Nucleic Acids Research, 44, 8479-8489. [**MNiSW 2016: 40 pkt.**, **IF 2016: 10,162**]

[O7] Piernik M., Brzeziński D., Morzy T., 2016, *Clustering XML Documents by Patterns*. Knowledge and Information Systems, 46(1), 185-212. [**MNiSW 2016: 30 pkt.**, **IF 2016: 2,004**]

[O8] Lango, M.; Brzeziński, D., Stefanowski, J., 2016, *PUT at SemEval-2016 Task 4: The ABC of Twitter Sentiment Analysis*. Proceedings of the 10th International Workshop on Semantic Evaluation, 126-132.

[O9] Brzeziński D., Piernik M., 2015, *Structural XML Classification in Concept Drifting Data Streams*, New Generation Computing, 33(4), 345-366. [**MNiSW 2015: 25 pkt.**, **IF 2015: 0,533**]

po uzyskaniu stopnia doktora ↑

przed uzyskaniem stopnia doktora ↓

[O10] Brzeziński D., Stefanowski J., 2015, *Prequential AUC for Classifier Evaluation and Drift Detection in Evolving Data Streams*, New Frontiers in Mining Complex Patterns, LNCS 8983, pp. 87-101.

[O11] Piernik M., Brzeziński D., Morzy T., Leśniewska A., 2015, *XML Clustering: A Review of Structural Approaches*, Knowledge Enineering Review, 30(3):297-323. [**MNiSW 2015: 20 pkt.**, **IF 2015: 1,039**]

[O12] Brzeziński D., Stefanowski J., 2014, *Reacting to Different Types of Concept Drift: The Accuracy Updated Ensemble Algorithm*. IEEE Transactions on Neural Networks and Learning System, 25(1):81-94. [**MNiSW 2014: 45 pkt.**, **IF 2014: 4,291**]

[O13] Brzeziński D., Stefanowski J., 2014, *Combining Block-based and Online Methods in Learning Ensembles from Concept Drifting Data Streams*. Information Sciences, 265:50-67. [**MNiSW 2014: 45 pkt.**, **IF 2014: 4,038**]

[O14] Krempl G., Zliobaite I., Brzeziński D., Hüllermeier E., Last M., Lemaire V., Noack T., Shaker A., Sievi S., Spiliopoulou M. and Stefanowski J., 2014, *Open Challenges for Data Stream Mining Research*. SIGKDD Explorations, 16(1):1-10.

[O15] Brzeziński D., Piernik M., 2014, *Adaptive XML Stream Classification using Partial Tree-edit Distance*. Foundations of Intelligent Systems – 21st International Symposium, LNCS 8502, pp. 10-19.

[O16] Brzeziński D., Stefanowski J., 2013, *Classifiers for Concept-drifting Data Streams: Evaluating Things That Really Matter.* Proceedings of 1st ECML PKDD Workshop on Real-World Challenges for Data Stream Mining, pp. 10-14.

[O17] Brzeziński D., Stefanowski J., 2011, *Accuracy updated ensemble for data streams with concept drift*. Proceedings of 6th HAIS International Conference on Hybrid Artificial Intelligence Systems, Part II, LNCS 6679, pp. 155-163.

[O18] Brzeziński D., Leśniewska A., Morzy T., Piernik M., 2010, *Clustering XML Documents by Patterns*. Proceedings of 3rd Polish National Conference on Data Processing Technologies, pp. 297-308.

[O19] Brzeziński D., Leśniewska A., Morzy T., Piernik M., 2010, *XCleaner: A New Method for Clustering XML Documents by Structure*. Control and Cybernetics, 40 (3), 877-891. [**MNiSW 2010: 20 pkt.**, **IF 2010: 0,300**]

5.2. Opis pozostałych osiągnięć

Prace których nie ujęto w cyklu stanowiącym osiągnięcie habilitacyjne oraz te które zostały opublikowane przed uzyskaniem stopnia doktora dotyczą przede wszystkim teoretycznych oraz praktycznych aspektów: uczenia klasyfikatorów strumieniowych, grupowania danych semi-strukturalnych oraz zastosowania metod uczenia maszynowego do problemów krystalograficznych biologii strukturalnej.

Chemicy i biolodzy wykorzystują krystalografię jako jedną z najdokładniejszych metod określania struktury i budowy cząsteczek. Naświetlane promieniami rentgenowskimi kryształy odbijają uderzające w nie promienie, które są nagrywane przez detektory. Obrazy z detektorów (zdjęcia dyfrakcyjne), po wzięciu pod uwagę interferencji i innych zjawisk zachodzących podczas przechodzenia promieni przez kryształy, stanowią podstawę do stworzenia map gęstości elektronowej, czyli obrazów 3D określających prawdopodobieństwo występowania elektronów w przestrzeni [Rup09]. Takie mapy są następnie analizowane przez chemików, fizyków i biologów, którzy określają typy pierwiastków i połączenia między nimi.

Praca [O1] opisuje podejście do rozpoznawania tzw. ligandów, czyli małych cząsteczek chemicznych (np. leków) towarzyszących makromolekułom biologicznym (np. enzymom) w ich strukturach krystalicznych. Zaproponowany system uczenia maszynowego, o nazwie CheckMyBlob, poprawnie rozpoznaje ligand w 57-73% przypadków (w zależności od zbioru testowego), podczas gdy najlepsze z istniejących podejść [Ter07, Car14] osiągają trafność na poziomie 32-48% i wymagają średnio ponad dwa razy więcej czasu na dokonanie predykcji. Praktycznym rezultatem tych badań jest, obecnie rozwijany we współpracy z University of Virginia, serwer do rozpoznawania ligandów w mapach gęstości elektronowej: http://checkmyblob.bioreproducibility.org.

Z krystalografią są również związane prace [O2] i [O6], które dotyczą nowej parametryzacji więzów stereochemicznych dla zasad oraz reszt fosforanowych występujących w kwasach nukleinowych. Prace te wykorzystują metody wykrywania wartości odstających oraz analizy skupień do określenia grup typowych długości wiązań oraz kątów w kwasach nukleinowych występujących w Cambridge

Structural Database [All02]. Odkryte grupy parametrów, udostępnione w ramach usługi http://achesym.ibch.poznan.pl/restraintlib/, mogą zostać wykorzystane do poprawienia długości wiązań podczas modelowania map gęstości elektronowej kwasów nukleinowych.

Pozostałe prace spoza cyklu są bliższe podstawowej tematyce badawczej habilitanta prezentowanej w ramach głównego osiągnięcia. W pracy [O5] przeprowadzono weryfikację użyteczności miar konfirmacji do ograniczania liczby reguł asocjacyjnych wykorzystywanych do klasyfikacji danych zbalansowanych i niezbalansowanych. Na potrzeby tego zadania opracowano algorytm CM-CAR, który potwierdził użyteczność wybranych miar konfirmacji do generowania zawężonego zbioru reguł łączącego aspekty deskrypcyjne i predykcyjne. Z kolei w pracy [O4] zbadano przydatność miar centralności sieci do poprawiania zdolności klasyfikacyjnych algorytmu k-najbliższych sąsiadów. W tym celu podobieństwa między przykładami uczącymi zostały zamienione na graf zależności, który następnie posłużył do wyliczenia (za pomocą miar centralności) wag przykładów. Ponadto w pracy [O16] zaproponowano ważenie przykładów uczących po zajściu zmian w strumieniu jako sposób różnicowania oceny klasyfikatora przed i po dryfcie.

Tematyka będąca przedmiotem pracy doktorskiej habilitanta koncentrowała się wokół klasyfikatorów dla zmiennych strumieni danych. W ramach [O12, O17] opracowano algorytm AUE, który łączy elementy uczenia blokowego oraz przyrostowego. W pracy [O13] zaproponowano rozszerzenie tego podejścia dla strumieni gdzie możliwe jest uczenie klasyfikatora przykład po przykładzie. Z kolei w pracy [O10] zaproponowano pierwszą wersję algorytmu liczącego AUC w sposób przyrostowy z zapominaniem, ale nie wzięto pod uwagę możliwości remisowych predykcji, nie zbadano własności tej miary oraz nie porównano jej z innymi sposobami liczenia pola pod krzywą ROC. Wszystkie algorytmy zaproponowane w ramach pracy doktorskiej zostały dołączone do oficjalnej dystrybucji środowiska do testowania algorytmów strumieniowych o nazwie Massive Online Analysis (MOA) [Bif10].

Wieloletnia praca nad strumieniami danych zaowocowała współpracą nad artykułem przedstawiającym stanowisko badaczy tej tematyki dotyczące przyszłości badań nad uczeniem klasyfikatorów dla zmiennych strumieni danych [O14]. Innym potwierdzeniem rozpoznawalności w tym środowisku (już po doktoracie) było zaproszenie do sporządzenia wpisu encyklopedycznego o klasyfikacji w strumieniach danych [O3] w Encyclopedia of Machine Learning and Data Mining [Sam17].

Habilitant był również współautorem szeregu prac dotyczących przetwarzania danych tekstowych. Artykuły [O7, O18, O19] prezentują nowe podejścia do grupowania dokumentów XML, oparte o wykrywanie wzorców częstych. Praca [O11] to z kolei przegląd algorytmów do grupowania danych semi-strukturalnych. Dokumenty XML były również tematem prac [O9, O15], gdzie zaproponowano wykorzystanie wzorców częstych do klasyfikacji strumieni dokumentów XML. W pracy [O8] przedstawiono algorytm do klasyfikacji opinii, który wziął udział w konkursie SemEval-2016 [Nak16].

Literatura

[All02] Allen F., The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallographica Section B*, 58:308–388, 2002.

[Bif10] Bifet A., Holmes G., Kirkby R., Pfahringer B., MOA: Massive Online Analysis. *Journal of Machine Learning Research*, 11:1601–1604, 2010.

[Brz15] Brzeziński D., Stefanowski J., Prequential AUC for Classifier Evaluation and Drift Detection in Evolving Data Streams. *New Frontiers in Mining Complex Patterns*, LNCS 8983, 87–101, 2015.

[Car14] Carolan C., Lamzin V., Automated identification of crystallographic ligands using sparse-density representations. *Acta Crystallographica Section D*, 70:1844–1853, 2014.

[Cha00] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R., *CRISP-DM 1.0*. CRISP-DM Consortium, 2000.

[Dom12] Domingos P. M., A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87, 2012.

[Fay99] Fayyad U.M., Piatetsky-Shapiro G., Smyth P., Uthurusamy R., *Advances in knowledge discovery and data mining*, AAAI/MIT Press, 1999.

[Fla03] Flach P., The Geometry of ROC Space: Understanding Machine Learning Metrics through ROC Isometrics. In: *Proceedings of the 20th International Conference on Machine Learning*, 194–201, 2003.

[Flo06] Floater M., Hormann K., Kos G., A general construction of barycentric coordinates over convex polygons. *Advances in Computational Mathematics*, 24(1–4): 311–331, 2006.

[Gam13] Gama J., Sebastiao R., Rodrigues P., On evaluating stream learning algorithms. *Machine Learning*, 90(3):317–346, 2013.

[Gar14] Garcia V., Mollineda R., Sanchez J., A bias correction function for classification performance assessment in two-class imbalanced problems. *Knowledge-Based Systems*, 59:66–74, 2014.

[He09] He, H.,Garcia, E. A., Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21:1263–1284, 2009

[He13] He H., Ma Y. *Imbalanced Learning: Foundations, Algorithms and Applications*, Wiley, 2013.

[Hua05] Huang J., Ling C., Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 17(3):299–310, 2005.

[Jap11] Japkowicz N., Shah M., *Evaluating Learning Algorithms: A Classification Perspective*, Cambridge University Press, ISBN 9780521196000, 2011.

[Kun04] Kuncheva L., *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience, 2004.

[Lic13] Lichman M., UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, 2013.

[Mar10] Mariscal G., Marban O., Fernandez C., A survey of data mining and knowledge discovery process models and methodologies. *Knowledge Engineering Review*, 25(2):137–166, 2010.

[Nak16] Nakov P., Ritter A., Rosenthal S., Stoyanov V., Sebastiani F., SemEval-2016 task 4: Sentiment analysis in Twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, 2016.

[Nap10] Napierala K., Stefanowski J., Wilk S., Learning from imbalanced data in presence of noisy and borderline examples. In: *Proceedings of the 7th International Conference on Rough Sets and Current Trends Computing,* 158–167, 2010.

[Nap16] Napierala, K., Stefanowski, J., Types of minority class examples and their influence on learning classifiers from imbalanced data. *Journal of Intelligent Information Systems*, 46(3): 563–597, 2016.

[Par12] Parisi T., *WebGL: Up and Running*. O'Reilly Media, 2012.

[Pia91] Piatetsky-Shapiro G., Matheus C., *Knowledge discovery in databases*, AAAI/MIT Press, 1991.

[Ran06] Ranawana R., Palade V., Optimized Precision - A New Measure for Classifier Performance Evaluation. In: *Proceedings of the IEEE Congress on Evolutionary Computation*, 16–21, 2006.

[Rup09] Rupp, B., *Biomolecular Crystallography: Principles, Practice, and Application to Structural Biology*, CRC Press, 2009.

[Sam17] Sammut, C., Webb, G., *Encyclopedia of Machine Learning and Data Mining*, Springer, 2017.

[Sus15] Susmaga R., Szczęch I., Can Interestingness Measures Be Usefully Visualized?, I*nternational Journal of Applied Mathematics and Computer Science*, 25(2): 323–336., 2015.

[Tan06] Tan P-N., Steinbach M., Kumar V., Introduction to data mining, Addison Wesley, 2006.

[Ter07] Terwilliger T., Adams P. Moriarty N., Cohn J., Ligand identification using electron-density map correlations. *Acta Crystallographica Section D*, 63, 101–107, 2007.

D. Brzeziński

Dariusz Brzeziński