Autoreferat

przedstawiający opis dorobku i osiągnięć naukowych

dr inż. Małgorzata Krętowska

Wydział Informatyki, Politechnika Białostocka
Białystok, 2019

1. **Imię i Nazwisko**

   Małgorzata Krętowska

2. **Posiadane dyplomy, stopnie naukowe – z podaniem nazwy, miejsca i roku ich uzyskania.**

1. Magister inżynier informatyki, Politechnika Białostocka, Instytut Informatyki, 1996, tytuł pracy: „Algorytmy wymiany rozwiązań bazowych i ich zastosowanie w syntezie klasyfikatorów", promotor: prof. dr hab. Leon Bobrowski.

2. Doktor nauk technicznych w dyscyplinie biocybernetyka i inżynieria biomedyczna (z wyróżnieniem), Instytut Biocybernetyki i Inżynierii Biomedycznej PAN, 2002, tytuł rozprawy: „Sieci neuropodobne jako narzędzie analizy historii zdarzeń", promotor: prof. dr hab. Leon Bobrowski; recenzenci: prof. dr hab. inż. Juliusz Lech Kulikowski, prof. dr hab. inż. Leszek Rutkowski.

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

   1. Politechnika Białostocka, Wydział Informatyki, asystent, 1996-2002
   2. Politechnika Białostocka, Wydział Informatyki, adiunkt, od 2002 roku

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. nr 65, poz. 595 z późn. zm.):**

   Tytuł osiągnięcia naukowego:
   Drzewiaste modele predykcyjne w analizie danych przeżycia

   W formie cyklu sześciu powiązanych tematycznie artykułów naukowych:

   [A1] Krętowska M. (2006) Random forest of dipolar trees for survival prediction, Rutkowski L. et al. (Eds.): ICAISC 2006, *Lecture Notes in Artificial Intelligence* 4029: 909-918, Springer

   [A2] Krętowska M. (2010) The Influence of censoring for the performance of survival tree ensemble, Rutkowski L. et al. (Eds.): ICAISC 2010, *Lecture Notes in Artificial Intelligence* 6114: 524-531, Springer

   [A3] Kretowska M. (2014) Comparison of tree-based ensembles in application to censored data, Rutkowski L. et al. (Eds.): ICAISC 2014, *Lecture Notes in Artificial Intelligence* 8467: 551-560, Springer

   [A4] Kretowska M. (2017) Piecewise-linear criterion functions in oblique survival tree induction, *Artificial Intelligence in Medicine* 75: 32-39 [IF2017=2,88; IF5y=2,86]

   [A5] Kretowska M. (2018) Tree-based models for survival data with competing risks, *Computer Methods and Programs in Biomedicine*, 159: 185-198 [IF2017=2,67; IF5y=2,84]

   [A6] Kretowska M. (2019) Oblique survival trees in discrete event time analysis, *IEEE Journal of Biomedical and Health Informatics*, DOI 10.1109/JBHI.2019.2908773 [IF2017=3,85; IF5y=3,94]

Publikacje [A1] – [A6] przedstawiają możliwości wykorzystania drzewiastych modeli predykcyjnych, w tym skośnych drzew przeżycia oraz rodzin tych drzew, do analizy trzech typów danych przeżycia: danych z ciągłym czasem przeżycia, z dyskretnym czasem przeżycia oraz z konkurencyjnym ryzykiem. W dalszych podrozdziałach niniejszego dokumentu przedstawiony zostanie szczegółowy opis przeprowadzonych badań naukowych.

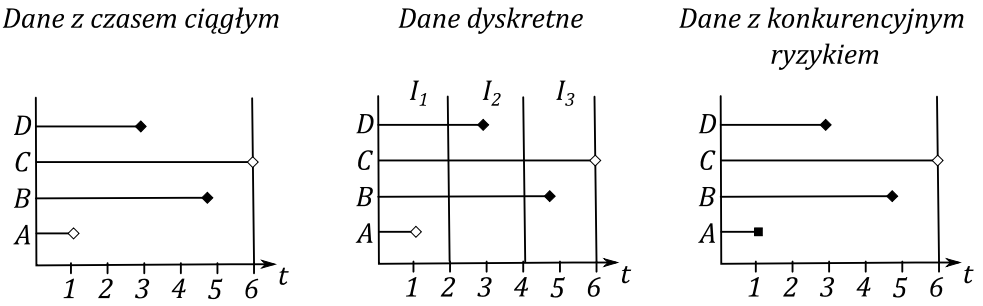
4.1 Wprowadzenie do obszaru badań

Analiza przeżycia [17] jest zbiorem metod mających na celu przewidywanie czasu wystąpienia określonego zdarzenia, zwanego porażką. Czas ten mierzony jest od zdarzenia początkowego, które w praktyce medycznej może oznaczać wystąpienie pierwszych symptomów danej choroby, operację czy też początek leczenia. Mianem porażki określa się z reguły nawrót choroby lub śmierć pacjenta.

Cechą charakterystyczną danych przeżycia jest występowanie tzw. obserwacji obciętych (ang. *censored cases*), dla których dokładny czas interesującego nas zdarzenia nie jest znany. Wiemy jedynie, że zdarzenie to nie wystąpiło przed zarejestrowanym w dokumentach czasem obserwacji pacjenta. Przyczyną obcięcia może być zakończenie badania klinicznego, czy też utrata kontaktu z pacjentem, związana na przykład ze zmianą jego miejsca zamieszkania. Czas przeżycia jest z reguły reprezentowany jako zmienna ciągła. Istnieje jednak szereg badań, w których można określić jedynie przedział czasu, w którym nastąpiło interesujące nas zdarzenie. Ma to miejsce wówczas, gdy pacjent wykonuje badania kontrolne co pewien, określony przez lekarza, czas, np. co kwartał, pół roku lub rok. Takie dane noszą nazwę dyskretnych danych przeżycia [26].

Naturalnym rozszerzeniem danych przeżycia są dane z konkurencyjnym ryzykiem, w przypadku których obserwacja pacjenta kończy się zarówno wystąpieniem interesującego nas zdarzeniem, jak też innych, konkurencyjnych w stosunku do niego, zdarzeń [22].

Różnice pomiędzy wymienionymi powyżej trzema typami danych, jak też ich reprezentacją w zbiorze uczącym przedstawia Rys. 1. Możemy zauważyć, że w przypadku danych ciągłych oraz dyskretnych wskaźnik przeżycia δ_i może przyjmować tylko dwie wartości informujące o wystąpieniu porażki ($\delta_i = 1$) lub jej braku ($\delta_i = 0$); w przypadku danych z konkurencyjnym ryzykiem jego wartość wskazuje również na typ zdarzenia kończącego obserwację ($\delta_i \in (0, p)$, gdzie p jest liczbą zdarzeń konkurencyjnych).



(x_i, t_i, δ_i)	(x_i, k_i, δ_i)	(x_i, t_i, δ_i)
$(x_A, 1, 0)$	$(x_A, 1, 0)$	$(x_A, 1, 2)$
$(x_B, 5, 1)$	$(x_B, 3, 1)$	$(x_B, 5, 1)$
$(x_C, 6, 0)$	$(x_C, 3, 0)$	$(x_C, 6, 0)$
$(x_D, 3, 1)$	$(x_D, 2, 1)$	$(x_D, 3, 1)$

Rys. 1 Trzy typy danych przeżycia wraz z ich reprezentacją w zbiorze uczącym

Istnieje szereg metod statystycznych pozwalających na przewidywanie ryzyka porażki. Najbardziej znanym i najczęściej wykorzystywanym modelem jest model proporcjonalnych hazardów Cox'a [5], który pozwala na ocenę wpływu poszczególnych cech na ryzyko porażki przy dosyć

rygorystycznych założeniach dotyczących rozkładu danych. Zastosowanie metod parametrycznych [15] wymagających określenia postaci funkcji rozkładu czasu przeżycia jest raczej marginalne.

Ze względu na dużą liczbę, często trudnych do spełnienia, założeń, którymi obwarowane są modele statystyczne, od wielu lat zaobserwować można rozwój alternatywnych sposobów analizy danych przeżycia. Bazują one często na sztucznej inteligencji, czy też uczeniu maszynowym i ich niekwestionowanym atutem jest brak założeń dotyczących rozkładu danych. Wśród nich istotną rolę pełnią modele drzewiaste. Wynik działania takich modeli w przypadku analizy przeżycia nie zawęża się do jednej wartości, lecz przyjmuje z reguły postać funkcji opisującej prawdopodobieństwo porażki w czasie.

Funkcje rozkładu czas przeżycia

Rozkład zmiennej losowej T reprezentującej czas przeżycia może być opisany przy użyciu kilku różnych funkcji. Najczęściej wykorzystywana jest funkcja przeżycia $S(t)=P(T>t)$ interpretowana jako prawdopodobieństwo wystąpienia porażki w czasie późniejszym niż t oraz funkcja hazardu $\lambda(t)$ rozumiana jako ryzyko porażki w czasie t. W przypadku danych dyskretnych wartości tych funkcji wyliczamy dla poszczególnych przedziałów czasowych: $S(I_k)=P(T>t_k)$, gdzie $I_k=[t_{k-1}, t_k)$, oraz $\lambda(I_k)=P(T\epsilon I_k|T>t_{k-1})$ interpretowana jako warunkowe prawdopodobieństwo porażki w przedziale I_k.

W przypadku danych z konkurencyjnym ryzykiem, pacjent obserwowany jest do momentu wystąpienia jednego z p ($p>1$) zdarzeń końcowych. Do opisu rozkładu czasu zajścia zdarzenia i, najczęściej wykorzystywana jest dystrybuanta (CIF, ang. *cumulative incidence function*) zdefiniowana jako prawdopodobieństwo zajścia tego zdarzenia do czasu t: $F_i(t)=P(T\leq t,\delta=i)$.

4.2 Stan wiedzy przed rozpoczęciem badań habilitacyjnych

Przegląd proponowanych w literaturze drzewiastych modeli predykcyjnych można podzielić na trzy odrębne nurty, rozwijane niezależnie dla każdego z trzech przedstawionych we wcześniejszym podrozdziale typów danych przeżycia.

Dane z ciągłym czasem przeżycia

Pierwsze doniesienia dotyczące zastosowania drzew do analizy danych przeżycia pojawiły się w latach osiemdziesiątych. Proponowane algorytmy można podzielić na dwie grupy. Pierwsza z nich bazowała na metodologii CART [3] i jako kryterium podziału wykorzystywała miary zróżnicowania wewnątrz węzłów, jak np. funkcję wiarogodności [18], czy też reszty martyngałowe [25]. Grupa druga, kryterium podziału budowała w oparciu o miary zróżnicowania pomiędzy węzłami, najczęściej bazujące na statystykach Taron-Ware'a, w tym teście log-rank [24, 19].

Pojedyncze drzewa przeżycia były też z powodzeniem łączone w rodziny modeli predykcyjnych, które charakteryzowały się lepszą stabilnością rozwiązań [10, 11, 13]. Proponowane algorytmy wykorzystywały dwie techniki generowania funkcji wynikowych. Pierwsza z nich, dla każdego czasu t wylicza średnią arytmetyczną funkcji uzyskiwanych z pojedynczych drzew wchodzących w skład rodziny [13]. Druga – na podstawie zawartości wybranych węzłów terminalnych pojedynczych drzew, tworzyła zagregowany zbiór danych uczących, na podstawie którego wyznaczany był estymator funkcji przeżycia dla nowego wektora cech [11].

Dane z dyskretnym czasem przeżycia

Ze względu na liczbę wyodrębnionych przedziałów czasowych wyróżniamy tutaj modele jedno- i wielopunktowe.

W podejściu jednopunktowym interesuje nas przeżycie pacjenta do pewnego ustalonego punktu czasu t_D (np. t_D = 5 lat). Jest to zadanie binarne, w którym celem jest stwierdzenie, czy do czasu t_D nastąpi porażka (wartość docelowa 1), czy też nie (wartość docelowa 0). Problem stanowią tu pacjenci, dla których obcięcie nastąpiło przed t_D, ponieważ ich stan w czasie t_D nie jest znany. Pacjenci Ci są z reguły wykluczani ze zbioru uczącego [9] lub ich wartości docelowe są przybliżane [8].

Modele wielopunktowe są wykorzystywane w przypadku podziału czasu przeżycia na więcej niż dwa przedziały czasowe. Celem analizy jest tutaj przewidywanie wartości funkcji przeżycia lub hazardu w poszczególnych przedziałach czasu. Proponowane modele są zawężone do drzew jednowymiarowych.

Jako kryterium podziału wykorzystuje się w nich zarówno miary bezpośrednio związane z analizą przeżycia, na przykład dyskretny model proporcjonalnych hazardów [2], jak też miary, wykorzystywane w drzewach decyzyjnych, np. indeks Gini [23].

Dane z konkurencyjnym ryzykiem

Rozwiązania dotyczące danych z konkurencyjnym ryzykiem różnicuje sposób postrzegania zdarzeń końcowych. Drzewa przeżycia budowane były w oparciu o pojedyncze wyróżnione zdarzenie, łącząc pozostałe w jedną grupę, lub uwzględniały wszystkie zdarzenia jednocześnie [12,4]. W pierwszym przypadku, jako wynik działania modeli drzewiastych, otrzymywano dystrybuantę dla pojedynczego zdarzenia, w drugim - zbiór funkcji CIF wyliczonych dla wszystkich zdarzeń końcowych. Podobny podział widoczny jest w pracy [14], w której do analizy danych z konkurencyjnym ryzykiem proponowane są lasy losowe. Nieco inne podejście zaproponowane zostało w pracy [21], gdzie wskaźnik obcięcia dla danych niepełnych został zastąpiony pseudo-wartościami. Proces ten spowodował zamianę danych obciętych na pełne, co uprościło algorytm indukcji lasów losowych.

Opisane powyżej modele drzew przeżycia oraz ich rodzin bazowały na strukturach wykorzystujących testy jednowymiarowych, w których każdy węzeł wewnętrzny weryfikował wartość tylko jednej cechy. Takie podejście, pomimo swojej prostoty i łatwości interpretacji, nie zawsze pozwala na uzyskanie dobrych właściwości uogólniających. Wykorzystanie drzew skośnych, w których podziały przestrzeni cech realizowane są za pomocą dowolnych hiperpłaszczyzn, może przyczynić się do poprawy jakości predykcji.

4.3 Cel i zakres badań

Przegląd istniejących narzędzi analizy danych przeżycia, w tym modeli drzewiastych, oraz umiejętność wykorzystania potencjału drzemiącego w odcinkowo-liniowych wypukłych funkcjach kryterialnych, skłoniło mnie do próby wykorzystania tego typu funkcji w procesie indukcji drzewiastych modeli prognostycznych. Zakres badań obejmuje następujące elementy:

(1) Wykorzystanie kryterium dipolowego w indukcji skośnych drzew przeżycia oraz ich rodzin [A1, A3, A4]

Istniejące drzewiaste modele predykcyjne, wykorzystywane do analizy danych z czasem ciągłym, były zawężone do drzew jednowymiarowych. Moim celem było opracowanie metody indukcji drzew skośnych z zastosowaniem wypukłych, odcinkowo-liniowych funkcji kryterialnych, tzw. kryterium dipolowego. Jest ono sumą funkcji kary powiązanych z elementami zbioru uczącego według określonych zasad – tzw. reguł tworzenia dipoli. Moim zadaniem było określenie reguł tworzenia dipoli, sposobu przycinania drzewa oraz porównanie otrzymanego modelu z istniejącymi metodami analizy danych przeżycia dla danych z czasem ciągłym. Pojedyncze skośne drzewa przeżycia będą stanowiły podstawę budowy modeli złożonych – rodzin drzew.

(2) Ocena wpływu liczby obserwacji obciętych na jakość działania narzędzi prognostycznych [A2]

Cechą charakterystyczną danych przeżycia są obserwacje obcięte, które zawierają niepełną informację o czasie porażki. Podstawą modeli analizy danych przeżycia jest między innymi umiejętność korzystania z tej niepełnej informacji. Moim celem była ocena wpływu liczby obserwacji obciętych w zbiorze uczącym na jakość predykcji.

(3) Indukcja drzewiastych modeli prognostycznych na potrzeby analizy danych z konkurencyjnym ryzykiem [A5]

Dane z konkurencyjnym ryzykiem stanowią naturalne rozszerzenie danych z czasem ciągłym. Istniejące rozwiązania również tutaj były zawężone do drzew jednowymiarowych. Mając do dyspozycji zbiór danych z wieloma zdarzeniami końcowymi oraz kryterium dipolowe, celem moim było

zaproponowanie algorytmu indukcji skośnych drzew przeżycia, jak również ich rodzin, do przewidywania ryzyka różnych typów porażki.

(4) Opracowanie modeli skośnych drzew przeżycia na potrzeby danych z czasem dyskretnym [A6]

Podział czasu przeżycia na rozłączne przedziały powoduje zmianę reprezentacji danych. Z konkretnym interwałem czasu wiążemy tutaj wielu pacjentów. Możemy więc rozpatrywać przedziały czasowe jako oddzielne klasy, wiedząc ponadto, że są one uszeregowane. Taka definicja problemu różni się od podejść proponowanych dla danych z ciągłym czasem przeżycia, stąd potrzeba stworzenia oddzielnych algorytmów, które uwzględniałyby specyfikę danych z czasem dyskretnym. Moim celem było zdefiniowane struktur modeli drzewiastych oraz opracowanie metod ich indukcji z zastosowaniem kryterium dipolowego do predykcji prawdopodobieństwa przeżycia oraz ryzyka porażki w poszczególnych interwałach czasu.

Na bazie powyższych rozważań i celów szczegółowych, można sformułować ogólny cel moich badań:

Opracowanie metod indukcji drzewiastych modeli predykcyjnych z wykorzystaniem wypukłych, odcinkowo-liniowych funkcji kryterialnych na potrzeby analizy przeżycia.

4.4 Osiągnięte wyniki

U podstaw proponowanych drzewiastych modeli prognostycznych znajdują się drzewa dipolowe. Proces indukcji tych drzew bazuje na minimalizacji wypukłych, odcinkowo-liniowych funkcji kryterialnych, tzw. kryterium dipolowego. Otrzymane drzewa są następnie przycinanie w celu uzyskania dobrych właściwości uogólniających. Opis wyników zacznę więc od przedstawienia podstaw tworzenia kryterium dipolowego oraz metod przycinania drzew. W kolejnych punktach zawarty będzie szczegółowy opis proponowanych rozwiązań.

4.4.1 Drzewa dipolowe

Drzewo binarne jest strukturą, w skład której wchodzą węzły wewnętrzne oraz liście. Każdy węzeł wewnętrzny, począwszy od korzenia, ma dwa węzły potomne, prawy i lewy. Poszczególne gałęzie drzewa zakończone są liśćmi, które nie mają potomków.

Tworzenie drzewa rozpoczynamy od jego korzenia. Na podstawie danych uczących i wybranej funkcji kryterialnej wyznaczana jest hiperpłaszczyzna dzieląca przestrzeń cech na dwie podprzestrzenie. Wektory cech ze zbioru uczącego znajdujące się po jej dodatniej stronie będą wchodziły w skład lewego potomka, natomiast wektory cech usytuowane po jej ujemnej stronie - w skład prawego. Taka procedura powtarzana jest dla kolejnych potomków do momentu, aż dany węzeł uznany zostanie za liść. Najczęściej ma to miejsce wówczas, gdy liczba wektorów cech w węźle jest odpowiednio mała lub wartość funkcji nie pozwala na dalszy ich podział. Liście, w odróżnieniu od węzłów wewnętrznych zawierających poszczególne podziały (testy), reprezentują otrzymane obszary przestrzeni cech.

Testy mogą przyjmować różne formy. Drzewa jednowymiarowe realizują podziały w oparciu o wartości tylko jednej zmiennej, sprawdzając zazwyczaj czy jest ona mniejsza czy większa od zadanego w węźle progu, lub, w przypadku dyskretnym, czy jest mu równa. Drzewo skośne wykorzystuje bardziej skomplikowane reguły podziału. W drzewach dipolowych podział jest realizowany przy użyciu hiperpłaszczyzn $H(w,\theta)=\{x: w^T x= \theta\}$ i ich indukcja wykorzystuje wypukłe, odcinkowo-liniowe funkcje kryterialne, tak zwane kryterium dipolowe [1].

4.4.1.1 Kryterium dipolowe

Dla dowolnego rozszerzonego wektora cech $z_j=[1, x_1, x_2, \ldots, x_N]^T$, $j=1,2,..,M$ ze zbioru uczącego L, zdefiniujmy dwa typy wypukłych, odcinkowo-liniowych funkcji kary:

$$\varphi_j^+(v) = \begin{cases} \gamma_j - v^T z_j & dla \quad v^T z_j \le \gamma_j \\ 0 & dla \quad v^T z_j > \gamma_j \end{cases} \qquad (1)$$

oraz

$$\varphi_j^-(v) = \begin{cases} \gamma_j + v^T z_j & dla \quad v^T z_j \geq -\gamma_j \\ 0 & dla \quad v^T z_j < -\gamma_j \end{cases} \qquad (2)$$

gdzie $v=[-\theta, w_1, w_2, \ldots, w_N]^T$ jest rozszerzonym wektorem wag, $\gamma_j \geq 0$ jest marginesem zazwyczaj równym 1.

Dipolowa funkcja kryterialna jest sumą funkcji kary (1) oraz (2) związanych z dipolami – parami wektorów cech (z_i, $z_{j'}$), $j \neq j'$, $j=1,2,\ldots,M$ oraz $j'=1,2,\ldots,M$, które podzielić możemy na dipole mieszane i czyste. Decyzja czy dana para wektorów tworzy dipol mieszany, czysty czy nie tworzy dipola podejmowana jest na etapie budowy funkcji. Jeżeli dwa wektory cech są jednorodne z punku widzenia analizowanego problemu i nie chcielibyśmy ich rozdzielać, wówczas tworzymy dipol czysty. Dipol mieszamy powstaje z dwóch wektorów, które powinny być od siebie odseparowane. W sytuacji, gdy nie jesteśmy w stanie stwierdzić, czy na podstawie posiadanych informacji dwa wektory są jednorodne czy nie, wówczas nie tworzą one dipola.

Funkcja związana z dipolem mieszanym $\varphi_{jj'}^m$, jest sumą dwóch różnych funkcji kary ($\varphi_{jj'}^m(v) = \varphi_j^+(v) + \varphi_{j'}^-(v)$ lub $\varphi_{jj'}^m(v) = \varphi_j^-(v) + \varphi_{j'}^+(v)$). Jej minimalizacja prowadzi do wyznaczenia hiperpłaszczyzny $H(v)$, która odseparowuje od siebie wektory z_j i $z_{j'}$. Z dipolem czystym łączymy funkcję $\varphi_{jj'}^p$, będącą sumą jednakowych funkcji kary ($\varphi_{jj'}^p(v) = \varphi_j^+(v) + \varphi_{j'}^+(v)$ lub $\varphi_{jj'}^p(v) = \varphi_j^-(v) + \varphi_{j'}^-(v)$). Wyliczona na jej podstawie hiperpłaszczyzna nie rozdziela wektorów cech z_j i $z_{j'}$. Kryterium dipolowe to suma funkcji skojarzonych ze wszystkimi dipolami utworzonymi dla danego zbioru danych:

$$\Psi_d(v) = \sum_{(j,j')\in I^p} \alpha_{jj'} \varphi_{jj'}^p(v) + \sum_{(j,j')\in I^m} \alpha_{jj'} \varphi_{jj'}^m(v) \qquad (3)$$

gdzie $\alpha_{jj'}$ jest ceną dipola (z_i, $z_{j'}$), I^p i I^m są odpowiednio zbiorami dipoli czystych i mieszanych.

Kryterium dipolowe wyliczane jest w każdym wewnętrznym węźle drzewa na podstawie wektorów cech zbioru uczącego, które w tym węźle się znalazły. Proces indukcji drzewa kończy się wówczas, gdy wszystkie dipole mieszane zostaną przecięte lub liczba obserwacji pełnych w węźle jest mniejsza niż 10.

4.4.1.2 Przycinanie drzew

Opisane powyżej kryterium stopu prowadzi przeważnie do uzyskania drzew z dużą liczbą węzłów, mających słabe właściwości uogólniające. Poprawę tych własności można uzyskać przycinając nadmiarowe gałęzie drzewa i zastępując je liśćmi. W przypadku drzew przeżycia algorytmy przycinania często bazują na metodzie split-complexity [19].

Dla danego drzewa T, algorytm przycinania minimalizuje miarę, zdefiniowaną jako:

$$G_\alpha = G(T) - \alpha|T| \ , \qquad (4)$$

gdzie α jest nieujemnym parametrem złożoności, $|T|$ jest liczbą węzłów wewnętrznych, a $G(T)$ jest sumą pewnych statystyk $G(w)$ po wszystkich węzłach wewnętrznych:

$$G(T) = \sum_{w\in In(T)} G(w) \ , \qquad (5)$$

gdzie $In(T)$ jest zbiorem węzłów wewnętrznych drzewa T. Dobór statystyki $G(w)$ zależy od rozwiązywanego problemu.

Wybór drzewa maksymalizującego miarę (4) jest ściśle związany z wartością α. Pomimo tego, że parametr α może przyjmować dowolne nieujemne wartości rzeczywiste, liczba poddrzew T jest skończona. Stąd, dane poddrzewo jest najlepszym rozwiązaniem (ma największą wartość G_α) dla całego przedziału wartości α. Pierwszym krokiem algorytmu, jest więc stworzenie sekwencji poddrzew skojarzonych z rosnącymi wartościami α. Para (T_i, α_i) będzie oznaczać, że T_i jest najlepszym poddrzewem dla $\alpha \in [\alpha_i, \alpha_{i+1})$. W wyniku działania algorytmu otrzymujemy sekwencję:

$$\begin{array}{ccccccc} T_0 & > & T_1 & > & \cdots & > & T_m \\ 0 & < & \alpha_1 & < & \cdots & < & \infty \end{array} , \qquad (6)$$

gdzie T_0 jest drzewem nieprzyciętym, T_m jest drzewem zbudowanym jedynie z korzenia, $T_i > T_{i+1}$ oznacza, że T_{i+1} jest poddrzewem drzewa T_i. Wybór najlepszego drzewa jest teraz zawężony do zbioru: $T_0, T_1, \ldots, T_m$. Przeprowadzamy go przy użyciu 10-cio punktowej walidacji krzyżowej (ang. *cross-validation*).

Algorytm split-complexity wywodzi się z metody cost-complexity [3], która maksymalizuje koszt:

$$C_\alpha = R(T) + \alpha|\tilde{T}| \ , \qquad (7)$$

gdzie α jest nieujemnym parametrem złożoności, $|\tilde{T}|$ jest liczbą liści, a $R(T)$ jest błędem wyliczanym na podstawie zbioru uczącego. Wybór najlepszego drzewa odbywa się podobnie jak w opisanym wyżej algorytmie split-complexity.

4.4.1.3 Rodziny drzew przeżycia

Jeżeli celem analizy jest przewidywanie wartości funkcji opisujących czas przeżycia, a nie podział przestrzeni cech realizowany przez pojedyncze drzewa, wówczas możemy wykorzystać tzw. rodziny drzew przeżycia. Dają one stabilniejsze wyniki, ale bez możliwości wglądu w reguły ich tworzenia.

Rodzina drzew przeżycia jest to zbiór K pojedynczych drzew indukowanych na bazie oddzielnych zbiorów uczących, których elementy są losowane (losowanie ze zwracaniem) ze zbioru bazowego L. Drzewa wchodzące w skład rodziny nie muszą mieć dobrych własności uogólniających, stąd też w procesie ich indukcji można pominąć etap przycinania.

W algorytmie tworzenia rodziny drzew przeżycia oraz generowania odpowiedzi dla danego na wejście nowego wektora cech x_{new} możemy wyróżnić następujące kroki [A6]:

1. Z bazowego zbioru uczącego L, wylosuj elementy K zbiorów treningowych ($L_1, L_2, .., L_K$) o liczności M (losowanie ze zwracaniem)
2. Utwórz K dipolowych drzew przeżycia T_k na bazie zbiorów L_k, $k=1,2,\ldots,K$,
3. Dla każdego drzewa T_k, $k=1,2,..,K$, wyodrębnij zbiór obserwacji $L_k(x_{new})$, które należą do tego samego liścia co nowa obserwacja x_{new},
4. Utwórz zagregowany zbiór $L_A(x_{new}) = [L_1(x_{new}), \ldots, L_K(x_{new})]$
5. Na bazie zbioru $L_A(x_{new})$ utwórz zagregowaną funkcję Kaplana-Meier'a [16] dla obserwacji x_{new}: $\hat{S}^A = (t|x_{new})$.
6. Na bazie zbioru $L_A(x_{new})$ wylicz wartości innych funkcji niezbędnych do analizy rozpatrywanego problemu.

Istnienie oraz treść ostatniego punktu zależy od typu zagadnienia, dla którego tworzymy rodzinę drzew przeżycia.

W celu dopasowania przedstawionego powyżej ogólnego algorytmu tworzenia dipolowego modelu prognostycznego do konkretnych problemów, należy skoncentrować się na modyfikacji następujących elementów:

1. Kryterium dipolowe: reguły tworzenia dipoli czystych i mieszanych oraz sposób reprezentacji otrzymanych obszarów przestrzeni cech (liści).
2. Przycinanie drzewa: metoda przycinania oraz dobór statystyki $G(w)$ lub $R(T)$.
3. Rodziny drzew: treść szóstego punktu algorytmu tworzenia rodziny drzew.

Znajdujący się w poniższych rozdziałach opis budowy drzew przeżycia do analizy danych z ciągłym czasem przeżycia, danych z konkurencyjnym ryzykiem oraz danych dyskretnych będzie uwzględniał właśnie te trzy elementy.

4.4.2 Modele predykcyjne dla danych z ciągłym czasem przeżycia

Istotnym elementem danych przeżycia jest zjawisko obcięcia, w wyniku którego część obserwacji uczących nie posiada dokładnej informacji o czasie zajścia porażki. Narzędzia prognostyczne dedykowane danym przeżycia już na etapie ich tworzenia powinny uwzględniać niepełną informację zawartą w danych obciętych [A4].

Kryterium dipolowe. Rozważając czas przeżycia jako zmienną ciągłą, naszym celem jest oddzielenie od siebie tych wektorów cech, dla których różnica pomiędzy ich czasami porażki jest duża. Z drugiej strony,

wektory, dla których interesujące nas zdarzenie miało miejsce w podobnym czasie, nie powinny być od siebie odseparowane. Aby określić, które pary wektorów powinny utworzyć dipole mieszane oraz czyste, wprowadźmy zbiór różnic czasów porażki D danego zbioru uczącego $L=(x_i, t_i, \delta_i)$, $i=1,\ldots,M$, gdzie x_i jest wektorem cech, t_i – czasem przeżycia, δ_i – wskaźnikiem porażki. Zakładając, że $D=\emptyset$, $i=1,2,\ldots,M$ oraz $j=i+1,\ldots,M$, tworzenie zbioru D odbywa się w następujący sposób:

1. Jeżeli $\delta_i=\delta_j=1$, to $D = D \cup \{|t_i - t_j|\}$.
2. Jeżeli $\delta_i=0$, $\delta_j=1$ i $t_i>t_j$, to $D = D \cup \{t_i - t_j\}$.
3. Jeżeli $\delta_i=1$, $\delta_j=0$ i $t_i<t_j$, to $D = D \cup \{t_j - t_i\}$.

Elementy zbioru $D=\{d_a\}$, $a=1,2\ldots, A$ są następnie sortowane rosnąco, tworząc sekwencję wartości:

$$d_{(1)} \leq d_{(2)} \leq \cdots \leq d_{(A)} \quad . \tag{8}$$

Stanowią one punkt wyjścia dla następujących reguł tworzenia dipoli:

1. Para wektorów cech (x_j, x_i) tworzy dipol czysty, jeżeli
   a. $\delta_j=\delta_i=1$ i $|t_j - t_i| < d_{(\lfloor \eta \cdot A \rfloor)}$
2. Para wektorów cech (x_j, x_i) tworzy dipol mieszany, jeżeli
   a. $\delta_j=\delta_i=1$ i $|t_j - t_i| \geq d_{(\lfloor \xi \cdot A \rfloor)}$
   b. $\delta_j=0$, $\delta_i=1$, $t_j>t_i$ i $t_j - t_i \geq d_{(\lfloor \xi \cdot A \rfloor)}$
   c. $\delta_j=1$, $\delta_i=0$, $t_j<t_i$ i $t_i - t_j \geq d_{(\lfloor \xi \cdot A \rfloor)}$

gdzie η i ζ to parametry mogące przyjmować wartości z zakresu od 0 do 1, $\lfloor x \rfloor$ to funkcja podłoga zwracającą największą liczbę całkowitą nie większą od x.

Analizując powyższe reguły możemy stwierdzić, że dipol czysty można utworzyć jedynie z obserwacji pełnych, gdyż tylko wówczas możemy być pewni, że różnica czasów porażki jest odpowiednio mała. W przypadku dipoli mieszanych jest możliwe wykorzystanie obserwacji obciętych wtedy, gdy druga obserwacja jest pełna oraz obcięcie nastąpiło dużo później niż porażka.

Każdy liść opisany jest za pomocą estymatora funkcji przeżycia Kaplana-Meier'a [16], wyliczonego na podstawie obserwacji ze zbioru uczącego znajdujących się w danym liściu.

Przycinanie drzewa. W algorytmie split-complexity, wykorzystywanym do przycinania otrzymanego drzewa przeżycia, funkcja $G(w)$ jest wartością statystyki log-rank mającej rozkład χ^2_1, weryfikującej hipotezę o jednakowych funkcjach przeżycia w prawym i lewym potomku węzła w.

Rodziny drzew. Wynikiem działania rodziny drzew jest tutaj zagregowana funkcja przeżycia Kaplana-Meier'a zdefiniowana w punkcie piątym procesu tworzenia rodziny drzew przeżycia; punkt szósty jest pomijany [A1]. Proponowane rozwiązania zostały porównane z innymi proponowanymi w literaturze złożonymi modelami predykcyjnymi [A3]. Na ich podstawie podjęte zostały również próby oceny wpływu frakcji obserwacji obciętych w zbiorze uczącym na jakość działania modeli prognostycznych [A2]. Uzyskane wyniki pokazują, że obserwacje obcięte występujące w danych już na poziomie 30-40 procent znacznie osłabiają możliwości predykcyjne uzyskanych struktur.

4.4.3 Modele predykcyjne dla danych z konkurencyjnym ryzykiem

W przypadku danych z konkurencyjnym ryzykiem możemy rozważać dwa podejścia. Pierwsze skupia się na przewidywaniu zajścia tylko jednego, wybranego zdarzenia, łącząc pozostałe typy porażek w jedną grupę, drugie – pozwala na analizę wszystkich rozpatrywanych typów zdarzeń równolegle [A5].

4.4.3.1 Model drzewa dla jednego zdarzenia

Punktem wyjścia do indukcji drzew dla jednego zdarzenia jest założenie Fine'a i Grey'a [6], które mówi, że obserwacje zakończone zdarzeniem konkurencyjnym ($\delta_i > 1$) narażone są na wystąpienie interesującego nas zdarzenia w każdym czasie t. Prowadzi to do stwierdzenia, że różnica czasów porażki pomiędzy interesującym nas zdarzeniem, a zdarzeniem konkurencyjnym jest zawsze na tyle duża, że umożliwia utworzenie dipola mieszanego.

Budowa zbioru różnic czasów porażki oraz reguł tworzenia dipoli obu typów jest taka sama jak w przypadku danych obciętych, za wyjątkiem dodatkowej reguły dotyczącej tworzenia dipoli mieszanych,

która uwzględnia zdarzenia konkurencyjne: para wektorów cech (x_j, x_i) tworzy dipol mieszany, jeżeli ($\delta_j=1$ i $\delta_i=z$) lub ($\delta_j=z$ i $\delta_i=1$), $z=2,3,\ldots,p$.

Rozpatrując tworzenie zbioru różnic czasów przeżycia D oraz dipoli w kontekście danych z konkurencyjnym ryzykiem, można stwierdzić, że dipole czyste mogą być budowana jedynie dla obserwacji zakończonych interesującym nas zdarzeniem, o ile różnica ich czasów porażki jest odpowiednio mała, natomiast obserwacje obcięte mogą być użyte do budowy dipoli mieszanych jedynie w połączeniu z obserwacjami, dla których $\delta_j=1$. Nie tworzą one dipoli mieszanych z obserwacjami zakończonymi zdarzeniami konkurencyjnymi.

Każdy liść drzewa opisany jest za pomocą estymatora funkcji CIF wyliczonego dla wyodrębnionego zdarzenia na podstawie obserwacji ze zbioru uczącego znajdujących się w danym liściu.

Przycinanie drzewa. W procesie przycinania, jako funkcję $G(w)$, wykorzystano statystykę Grey'a [7] weryfikującą hipotezę o równości funkcji CIF w prawym i lewym potomku węzła w.

Rodzina drzew. Punkt szósty algorytmu tworzenia rodziny drzew przeżycia ma tutaj postać: Na bazie zbioru $L_A(x_{new})$ wyznacz zagregowaną funkcję CIF dla interesującego nas zdarzenia dla obserwacji x_{new}: $\hat{F}_1^A(t|x_{new})$.

4.4.3.2 Model drzewa dla wielu zdarzeń

Model drzewa dla wielu zdarzeń jest wykorzystywany, gdy analizujemy wpływ wielu typów porażki równolegle. W tym celu, wektory cech są podzielone na grupy względem typu porażki. Przynależność do jednej grupy nie gwarantuje utworzenia pomiędzy parą wektorów dipola czystego. Typ dipola warunkuje również różnica czasów porażki.

W przypadku tego modelu dla każdego typu porażki tworzony jest oddzielny zbiór różnic czasów przeżycia D_z. Zakładając, że $D_1=D_2=\ldots D_p=\emptyset$, $z=1,2,\ldots,p$, $i=1,2,\ldots,M$ oraz $j=i+1,\ldots,M$, otrzymujemy:

1. Jeżeli $\delta_i=\delta_j=z$, to $D_z = D_z \cup \{|t_i - t_j|\}$.
2. Jeżeli $\delta_i=0$, $\delta_j=z$ i $t_i>t_j$, to $D_z = D_z \cup \{t_i - t_j\}$.
3. Jeżeli $\delta_i=z$, $\delta_j=0$ i $t_i<t_j$, to $D_z = D_z \cup \{t_j - t_i\}$.

Utworzone zbiory $D_z = \{d_{z_k}\}$, $z=1,2,\ldots,p$, $k=1,2,\ldots, z_A$, są następnie sortowane rosnąco tworząc p sekwencji:

$$d_{(z_1)} \leq d_{(z_2)} \leq \cdots \leq d_{(z_A)} \ . \qquad (9)$$

Reguły tworzenia dipoli są następujące:

1. Para wektorów cech (x_j, x_i) tworzy dipol czysty, jeżeli
   a. $\delta_j=\delta_i=z$ i $|t_j - t_i| < d_{(\lfloor \eta \cdot z_k \rfloor)}$
2. Para wektorów cech (x_j, x_i) tworzy dipol mieszany, jeżeli
   a. $\delta_j=\delta_i=z$ i $|t_j - t_i| \geq d_{(\lfloor \xi \cdot z_k \rfloor)}$
   b. $\delta_j=0$, $\delta_i=z$, $t_j>t_i$ i $t_j - t_i \geq d_{(\lfloor \xi \cdot z_k \rfloor)}$
   c. $\delta_j=z$, $\delta_i=0$, $t_j<t_i$ i $t_i - t_j \geq d_{(\lfloor \xi \cdot z_k \rfloor)}$
   d. ($\delta_j=s$, $\delta_i=z$) lub ($\delta_j=z$, $\delta_i=s$), $z \neq s$, $s=1,2,\ldots,p$, $z=1,2,\ldots,p$

gdzie η i ζ to parametry mogące przyjmować wartości z zakresu od 0 do 1, $\lfloor x \rfloor$ to funkcja podłoga zwracającą największą liczbę całkowitą nie większą od x.

Widzimy, że dipol mieszany jest tworzony pomiędzy wszystkimi parami obserwacji zakończonymi różnymi typami zdarzeń. W sytuacji, gdy oba wektory cech mają przypisany ten sam typ zdarzenia, lub gdy jeden z nich jest obcięty, wówczas dodatkowym elementem określającym typ dipola jest różnica czasów przeżycia.

Liście drzewa opisane są za pomocą estymatorów funkcji CIF wyliczonych dla wszystkich typów porażek na podstawie obserwacji ze zbioru uczącego wchodzących w skład liścia.

Przycinanie drzewa. W procesie przycinania wykorzystywany jest algorytm split-comlexity z $G(w)$ wyliczaną na podstawie wielowymiarowej statystyki proponowanej w pracy [20].

Rodzina drzew. Punkt szósty algorytmu tworzenia rodziny drzew przeżycia ma tutaj postać: Na bazie zbioru $L_A(x_{new})$ wyznacz zagregowane funkcje CIF dla wszystkich typów porażki z ($z=1,2, \ldots,p$) dla obserwacji x_{new}: $\hat{F}_z^A(t|x_{new})$.

4.4.4 Modele predykcyjne dla danych z dyskretnym czasem przeżycia

Do analizy dyskretnych danych przeżycia zaproponowane zostały trzy typy drzew: dyskretne drzewo przeżycia, drzewo klasyfikacyjne oraz modularne [A6]. Każda z zaproponowanych metod w inny sposób wykorzystuje informacje zawarte w dyskretnych danych przeżycia. Jest to możliwe dzięki odpowiedniemu przygotowaniu danych do analizy oraz wykorzystaniu różnych struktur drzewiastych.

4.4.4.1 Dyskretne drzewo przeżycia

Dyskretne drzewo przeżycia ma na celu podział przestrzeni cech na regiony reprezentujące poszczególne przedziały czasu. Nie chcemy więc separować obserwacji, dla których porażka miała miejsce w tym samym przedziale czasu. Natomiast obserwacje pochodzące z różnych przedziałów powinny być rozdzielane. Prowadzi to do następujących reguł formowania dipoli:

1. Para wektorów cech (x_j, x_i) tworzy dipol czysty, jeżeli
   a. $\delta_j = \delta_i = 1$ i $k_i = k_j$
2. Para wektorów cech (x_j, x_i) tworzy dipol mieszany, jeżeli
   a. $\delta_j = \delta_i = 1$ i $k_i \neq k_j$
   b. $\delta_j = 0$, $\delta_i = 1$ i $k_i \geq k_j$

Obserwacje obcięte biorą udział w tworzeniu dipoli mieszanych tylko z obserwacjami pełnymi, dla których numer przedziału czasowego jest nie większy niż czas obcięcia. Tylko wówczas mamy pewność, z ewentualna porażka obserwacji obciętej będzie miała miejsce w innym przedziale czasu.

Reprezentacja liści to wartości funkcji przeżycia oraz hazardu wyliczane dla wszystkich przedziałów czasu.

Przycinanie drzewa. Metoda split-complexity z funkcją $G(w)$ wyliczaną jako suma statystyk ilorazu wiarygodności.

4.4.4.2 Drzewo klasyfikacyjne

Drzewo klasyfikacyjne wymaga dodatkowego etapu przygotowania danych. Każda obserwacja ze zbioru uczącego jest powielana dla każdego przedziału czasowego, I_k, w którym była obserwowana. Nowy, powiększony zbiór uczący ma postać $L_{aug} = (x_i; k; d_{ik}) = (y_{ik}; d_{ik})$, gdzie $i = 1, \ldots, M$, $k = 1, \ldots, k_i$, d_{ik} to wskaźnik porażki, który jest równy 1 tylko dla ostatniego przedziału czasu obserwacji pełnej, tzn. przedziału, w którym miała miejsce porażka, natomiast $y_{ik} = [x_{ik}^T, k]^T$.

Dla każdej obserwacji z powiększonego zbioru danych znamy wartość d_{ik}, która informuje nas o stanie i-tego pacjenta w przedziale czasowym k. Wskaźnik ten jest również estymatorem funkcji hazardu λ_{ik}. Naszym celem jest zatem oddzielenie obserwacje z różnymi wartościami d_{ik} (wartość docelowa). Problem staje się zatem zadaniem klasyfikacji binarnej i drzewo jest indukowane w celu podzielenia przestrzeni wejściowej (powiększonej przestrzeni cech) na obszary o podobnej wartości wskaźników porażki, 1 lub 0. Uzyskuje się to przez minimalizację kryterium dipolowego otrzymanego na podstawie dipoli utworzonych zgodnie z poniższymi zasadami:

1. para wektorów wejściowych ($y_{ik}; y_{jk'}$) tworzy czysty dipol, jeżeli $d_{ik} = d_{jk'}$.
2. para wektorów wejściowych ($y_{ik}; y_{jk'}$) tworzy mieszaną dipol, jeśli $d_{ik} \neq d_{jk'}$.

Chociaż wartością docelową jest zmienna binarna, reprezentacją liścia jest wskaźnik ryzyka h_k wyliczony jako iloraz liczby porażek ($d_{ik} = 1$) do liczby wszystkich obserwacji w danym liściu.

Przycinanie drzewa. Metoda cost-complexity z funkcją $R(T)$ wyliczaną jako błąd średniokwadratowy.

4.4.4.3 Drzewo modularne

Drzewo modularne to zbiór K drzew T_k, $k = 1, \ldots, K$. Każde pojedyncze drzewo T_k ma na celu przewidywanie ryzyka porażki w k-tym przedziale czasowym. Drzewo $T_{k'}$ jest indukowane na podstawie obserwacji, dla których istnieje niezerowe prawdopodobieństwo porażki w okresie $I_{k'}$. Oznacza to, że wszystkie przypadki z powiększonego zbioru uczącego L_{aug} z $k = k'$, tworzą nowy zbiór uczący $L^{k'}_{aug} = (x_i,$

$k', d_{ik'}$), $i \in M_{k'}$, gdzie $M_{k'}$ oznacza zbiór indeksów tych wektorów cech, których obserwacja zakończyła się w przedziale I_k, gdzie $k \geq k'$.

Biorąc pod uwagę proces indukcji pojedynczego drzewa T_k jesteśmy zainteresowani takim podziałem przestrzeni cech, który odseparuje od siebie obserwacje z różnymi wartościami d_{ik}, tj. pacjentów, którzy doświadczyli porażki w przedziale k-tym od pozostałych pacjentów. Pojedyncze drzewo T_k indukujemy na podstawie zbioru uczącego L^k_{aug} zakładając, że:

1. para wektorów wejściowych (x_i; x_j) tworzy czysty dipol, jeżeli $d_{ik} = d_{jk}$.
2. para wektorów wejściowych (x_i; x_j) tworzy mieszany dipol, jeśli $d_{ik} \neq d_{jk}$.

Dla każdego liścia obliczamy warunkowe prawdopodobieństwo niepowodzenia w k-tym przedziale czasowym, h_k, wyliczony jako iloraz liczby porażek ($d_{ik} = 1$) do liczby wszystkich obserwacji w danym liściu.

Przycinanie drzewa. Do przycinania drzewa wykorzystana została metoda cost-complexity z funkcją *R(T)* wyliczaną jako błąd średniokwadratowy.

4.5 Podsumowanie

W ramach osiągnięcia naukowego, opisanego w pracach [A1-A6], zaproponowano szereg drzewiastych narzędzi prognostycznych pozwalających na analizę różnych typów danych przeżycia. Wspólną cechą proponowanych algorytmów jest sposób indukcji drzew wykorzystujący wypukłe, odcinkowo-liniowe funkcje kryterialne, tzw. kryterium dipolowe. Pozwala ono na tworzenie drzew skośnych, stanowiących nowość wśród istniejących już drzew przeżycia, zawężonych do struktur wykorzystujących testy jednowymiarowe.

Biorąc pod uwagę różne typy danych przeżycia, w ramach osiągnięcia naukowego zaproponowane zostały następujące elementy:

1. W zakresie analizy danych z ciągłym czasem przeżycia:
    a. Algorytm indukcji drzewa przeżycia
    b. Rodzina drzew przeżycia
    c. Ocena wpływu liczy obserwacji obciętych na jakość predykcji rodziny drzew przeżycia
2. W zakresie analizy danych z konkurencyjnym ryzykiem:
    a. Drzewo przeżycia dla jednego zdarzenia
    b. Drzewo przeżycia dla wielu zdarzeń
    c. Rodziny drzew przeżycia dla jednego oraz wielu zdarzeń
3. W zakresie analizy danych z dyskretnym czasem przeżycia zaproponowano:
    a. Dyskretne drzewo przeżycia
    b. Drzewo klasyfikacyjne
    c. Drzewo modularne

Wszystkie wymienione powyżej algorytmy zostały przeze mnie zaimplementowane oraz zweryfikowane pod kątem jakości predykcji. W tym celu, wyniki, uzyskane zarówno dla zbiorów syntetycznych jak też rzeczywistych, zostały porównane z rezultatami istniejących już metod analizy danych przeżycia. Jakość predykcji, w zależności od problemu mierzona była za pomocą indeksu C, współczynnika Brier'a (ang. *Brier score*) oraz średniego błędu względnego. Bazując na wynikach przeprowadzonych testów statystycznych, można powiedzieć, że jakość predykcji zaproponowanych drzewiastych modeli prognostycznych była lepsza lub nie różniła się istotnie od jakości predykcji istniejących rozwiązań.

Zaproponowane algorytmy zostały z powodzeniem wykorzystane do analizy rzeczywistych, dostępnych w literaturze zbiorów danych medycznych. Możliwości zastosowania drzewiastych modeli prognostycznych w praktyce są elementem składowym prac stanowiących osiągnięcie naukowe.

4.6 Literatura

[1] Bobrowski L, Kretowska M, Kretowski M.: Design of neural classifying networks by using dipolar criterions, Proceedings of the Third Conference on Neural Networks and Their Applications, Polish Society of Neural Networks, Polska, Kule, 1997, s. 689–94.

[2] Bou-Hamad I., Larocque D., Ben-Ameur H., Masse L., Vitaro F., Tremblay R., Discrete-time survival trees, *Canadian Journal of Statistics*, s. 17–32, 2009.
[3] Breiman L., Friedman J., Olshen R., Stone C.: Classification and regression trees, Belmont, CA: Wadsworth, 1984.
[4] Callaghan F.: Classification trees for survival data with competing risks, Praca doktorska, University of Pittsburgh, 2008.
[5] Cox D.: Regression models and life tables (with discussion), *Journal of Royal Statistical Society B*, Vol. 34, 1972, s. 187–220.
[6] Fine J., Gray R.: A proportional hazards model for the subdistribution of a competing risk, *Journal of the American Statistical Association*, Vol. 94, Nr 446, 1999, s. 496–509.
[7] Gray R.: A class of k-sample tests for comparing the cumulative incidence of a competing risk, *The Annals of Statistics*, Vol. 16, Nr 3, 1988, s. 1141–1154.
[8] Garcia-Laencina P.J., Abreu P.H., Abreu M.H., Afonoso N., Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values, *Computers in Biology and Medicine*, Vol. 59, 2015, s. 125–133.
[9] Gomez I., Ribelles N., Franco L., Alba E., Jerez J. M., Supervised discretization can discover risk groups in cancer survival analysis, *Computer Methods and Programs in Biomedicine*, Vol. 136, s. 11–19, 2016.
[10] Hothorn T., Buhlmann P., Dudoit S., Molinaro A. M., van der Laan, M. J.: Survival ensembles, *Biostatistics*, Vol. 7, 2006, s. 355–373.
[11] Hothorn T., Lausen B., Benner A., Radespiel-Troger M.: Bagging survival trees, *Statistics in Medicine*, Vol. 23, 2004, s. 77–91.
[12] Ibrahim N., Kudus A.: Decision tree for prognostic classification of multivariate survival data and competing risks, INTECH Open Access Publisher, 2009.
[13] Ishwaran H., Blackstone E. H., Pothier C. E., Lauer M. S.: Relative risk forest for exercise heart rate recovery as a predictor of mortality, *Journal of the American Statistical Association*, Vol. 99, 2004, s. 591–600.
[14] Ishwaran H., Kogalur U., Blackstone E., Lauer M., Random survival forests, *The Annals of Applied Statistics,* Vol. 2, Nr 3, 2008, s. 841–860.
[15] Kalbfleisch J., Prentice R.: The statistical analysis of failure time data, New York: John Wiley & Sons, 1980.
[16] Kaplan E., Meier P.: Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association*, Vol. 53, 1958, s. 457–81.
[17] Klein J., Moeschberger M., Survival Analysis, Techniques for Censored and Truncated Data. Springer, 1997.
[18] LeBlanc M., Crowley J.: Relative risk trees for censored survival data, *Biometrics*, Vol. 48, 1992, s. 411–25.
[19] LeBlanc M., Crowley J.: Survival trees by goodness of split, *Journal of American Statistical Association*, Vol. 88, Nr 422, 1993, s. 457–67.
[20] Luo X., Turnbull B.: Comparing two treatments with multiple competing risks endpoints, *Statistica Sinica*, Vol. 9, Nr 4, 1999, s. 985–987.
[21] Mogensen U., Gerds T., A random forest approach for competing risks based on pseudo-values, *Statistics in medicine*, Vol. 32, Nr 18, 2013, s. 3102–3114.
[22] Pintilie M.: Competing risks: A practical perspective, John Wiley & Sons, 2006.
[23] Schmid M., Kuchenhoff H., Hoerauf A., Tutz G., A survival tree method for the analysis of discrete event times in clinical and epidemiological studies, *Statistics in Medicine*, Vol. 35, Nr. 5, pp. 734–751, 2016.
[24] Segal M.: Regression trees for censored data, *Biometrics*, Vol. 44, 1988, s. 35–47.
[25] Therneau T., Grambsch P., Fleming T.: Martingale-based residuals for survival models, *Biometrika*, Vol. 77, Nr 1, 1990, s. 147–160.
[26] Tutz G., Schmid M., Modeling discrete time-to-event data, Springer Series in Statistics, Springer, 2016.

5 Omówienie pozostałych osiągnięć naukowo - badawczych

Po osiągnięciu stopnia doktora główny nurt moich badań dotyczył opracowania nowych metod analizy danych przeżycia, w tym metod z zastosowaniem kryterium dipolowego. Prowadzone w tym nurcie prace można podzielić na kilka grup:

- Badania dotyczące wykorzystania różnych typów sieci neuropodobnych do analizy danych przeżycia z czasem ciągłym, jako kontynuacja badań prowadzonych w ramach pracy doktorskiej. Zaproponowano rozwiązania z wykorzystania kryterium dipolowego w procesie syntezy sztucznych sieci neuronowych, w tym sieci rangowej oraz sieci modułowej. Dodatkowo zweryfikowano możliwości zastosowania uogólnionych sieci regresyjnych do analizy danych obciętych.

  1. Bobrowski L., Krętowska M., Ranked designing of prognostic neural models for survival data, *Biocybernetics and Biomedical Engineering*, 22(2-3): 223-239, 2002
  2. Krętowska M., Bobrowski L., Evaluation of dipolar neural networks in survival time prediction, In Proc. of the Sixth International Conf. on Neural Networks and Soft Computing, ed. by Rutkowski L., Kasprzyk J., *Advances in Soft Computing*, Physica-Verlag, Zakopane, 11-15 czerwca, 2002, s. 230-235, 2003
  3. Krętowska M., Bobrowski L., Artificial neural networks in identifying areas with homogeneous survival time, Rutkowski L. et al. (Eds.): ICAISC 2004, *Lecture Notes in Artificial Intelligence* 3070: 1008-1013, Springer-Verlag, Zakopane, Poland, 7-11 czerwca, 2004 [IF2004=0,25]
  4. Krętowska M., Uogólnione sieci regresyjne w przewidywaniu ryzyka porażki, XIV Krajowa Konferencja Naukowa : Biocybernetyka i Inżynieria Biomedyczna, s. 426-431, 2005
  5. Krętowska M., Ensemble of dipolar neural networks in application to survival data, L. Rutkowski *et al*. (Eds.): ICAISC 2008, Lecture Notes in Artificial Intelligence 5097, s. 78-88, 2008

- Badania dotyczące wykorzystania kryterium dipolowego do analizy danych z ciągłym czasem przeżycia, które przyczyniły się do powstania ostatecznej wersji algorytmu indukcji drzewa przeżycia, wchodzącego w skład osiągnięcia naukowego. Zaproponowane zostały podstawowe reguły tworzenia dipoli oraz zaimplementowane i zweryfikowane sposoby generowania i reprezentacji wyników działania rodzin drzew przeżycia

  1. Krętowska M., Induction of survival trees based on dipolar criterion, Fifth International Seminar Statistics and Clinical Practice: 68th Seminar of the International Centre of Biocybernetics, *Lecture Notes of the ICB Seminars*, ed. by Bobrowski L., Doroszewski J., Victor N., s. 175-179, 2002
  2. Krętowska M., Dipolowe drzewa regresyjne w analizie przeżyć, XIII Krajowa Konferencja Naukowa: Biocybernetyka i inżynieria biomedyczna, Vol.1, s.188-193, Gdańsk, 10-13 września 2003
  3. Krętowska M., Dipolar regression trees in survival analysis, *Biocybernetics and Biomedical Engineering*, 24 (3): 25-33, 2004
  4. Krętowska M., Ensembles of dipolar trees for survival prediction, Sixth International Seminar Statistics and Clinical Practice : 81th Seminar of the International Centre of Biocybernetics, *Lecture Notes of the ICB Seminars*, ed. by Bobrowski L., Doroszewski J., Kulikowski C., Victor N., s. 120-125, 2005
  5. Krętowska M., Lasy losowe – ocena jakości prognostycznej cech, *Zeszyty Naukowe Politechniki Białostockiej. Informatyka*, 2: 67-77, 2007
  6. Krętowska M., Ensembles of dipolar trees for prediction of survival time, *Biocybernetics and Biomedical Engineering*, 27 (3): 67-75, 2007
  7. Krętowska M., Prognostic abilities of dipoles based ensembles - comparative analysis, Zeszyty Naukowe Politechniki Białostockiej. Informatyka. 4: 73-83, 2009
  8. Kretowska M., Computational Intelligence in Survival Analysis, in Wang J. (Ed): Encyclopedia of Business Analytics and Optimization (5 Volumes). IGI Global, s. 491-501, 2014

- Prace prowadzone nad budową narzędzi prognostycznych na potrzeby danych z konkurencyjnym ryzykiem. Mechanizmy tworzenia rodzin drzew przeżycia dla danych obciętych zostały rozszerzone o możliwości wykorzystania informacji o dodatkowych zdarzeniach końcowych. Prowadzone badania dotyczyły możliwości wykorzystania kryterium dipolowego oraz sposobu generowania funkcji wynikowych.

1. Krętowska M., Competing risk analysis - graphical representation of variable influence, Zeszyty Naukowe Politechniki Białostockiej. Informatyka, 8: 19-30, 2011
2. Krętowska M., Competing risks and survival tree ensemble, L. Rutkowski et al. (Eds.): ICAISC 2012, Part I, Lecture Notes in Artificial Intelligence 7267, s. 387–393, 2012.
3. Kretowska M., Dipolar tree ensemble with and without adjustment to competing risks: Application to medical data, Studies in Logic, Grammar and Rhetoric, seria *Logical Statistical and Computer Methods in Medicine*, 35: 27-37, 2013
4. Kretowska M., Tree-based models in application to competing risks, L. Bobrowski et al. (Eds.): Lecture notes of the ICB seminar, Ninth International Seminar: Statistics and Clinical Practice, 97-101, 2014

Równolegle z głównym nurtem badań, w ramach współpracy z Uniwersytetem Medycznym w Białymstoku, prowadziłam prace w zakresie analizy danych medycznych. Badania te obejmowały swoim zasięgiem opracowanie sposobu analizy danych medycznych, jego realizację oraz interpretację uzyskanych wyników.

- W ramach współpracy z Kliniką Perinatologii i Położnictwa Uniwersytetu Medycznego w Białymstoku powstały następujące prace:

1. Laudańska H., Lemancewicz A., Krętowska M**.,** Reduta T., Laudański T., Permeability of human skin to selected anions and cations - in vitro studies, *Research Communications in Molecular Pathology and Pharmacology* 112(1-4):16-26, 2002 [IF2002=0,32]
2. Pierzyński P., Mazurek A., Kuć P., Krętowska M**.,** Mazurek E., Laudański T., Pretreatment levels of antigens CA 125 and CYFRA 21-1 in patients with endometrial cancer, *Polish Journal of Gynaecological Investigations* 4(3): 169-174, 2002
3. Urban R., Lemancewicz A., Urban J., Skotnicki M.Z., Krętowska M**.,** Misoprostol and dinoprostone therapy for labor induction: a Doppler comparison of uterine and fetal hemodynamic effects, *European Journal of Obstetrics and Gynecology and Reproductive Biology*, 106(1): 20-24, 2003 [IF2003=1,002]
4. Lemancewicz A., Urban R., Redźko S., Urban J., Krętowska M., Maternal serum concentration of sICAM-1 in women with a short cervix on transvaginal ultrasonography, Ginekologia Polska (75 supl): 185-188, II Interaktywna Konferencja Naukowa & Aktualne problemy w perinatologii i ginekologii, Bytom-Wisła, 2-4 grudnia 2004
5. Lemancewicz A.,Krętowska M.,Urban J.,Laudański P., Active matrix metalloproteinase-9 serum levels in women delivering preterm, *Polish Journal of Gynecological Investigation* 8(2):67-69, 2005
6. Urban R., Lemancewicz A., Przepieść J., Urban J., Krętowska M**.,** Antenatal corticosteroid therapy: a comparative study of dexamethasone and betamethasone effects effects on fetal Doppler flow velocity waveforms, *European Journal of Obstetrics and Gynecology and Reproductive Biology*, 120: 170-174-24, 2005 [IF2005=1,14]
7. Lemancewicz A., Laudański P., Kuć P., Pierzyński P., Krętowska M., Urban J., Matrix metalloproteinase-2 and tissue inhibitor of metalloproteinases-2 plasma levels in preterm labouring patients. *Archives of Perinatal Medicine*, 12(3):33-35, 2006
8. Lemancewicz A., Laudański P., Kuć P., Pierzyński P., Krętowska M., Urban J., Matrix metalloproteinase-2 and tissue inhibitor of metalloproteinase-2 urine levels in preterm labouring patients. *Archives of Perinatal Medicine* 13(2):50-52, 2007
9. Kuć P., Lemancewicz A., Laudański P., Krętowska M., Laudański T., Total matrix metalloproteinase-8 serum levels in patients labouring preterm and patients with threatened Praterm delivery, Folia Histochemica et Cytobiologia 48(3), s. 366-370, 2010 [IF2010=0,9; IF5y=1,36]
10. Laudanski P.,Lemancewicz A., Kuc P., Charkiewicz K., Ramotowska B., Kretowska M., Jasinska E., Raba G., Karwasik-Kajszczarek K., Kraczkowski J., Laudanski T., Chemokines Profiling of Patients with Preterm Birth, *Mediators of Inflammation*, vol. 2014, Article ID 185758, 7 pages, 2014. [IF2014=3,24; IF5y=3,52]

- Współpraca z Kliniką Okulistyki Dziecięcej Uniwersytetu Medycznego w Białymstoku zaowocowała zbiorem następujących publikacji:

1. Urban B., Bakunowicz-Łazarczyk A., Krętowska M., Śródbłonek rogówki u młodzieży z krótkowzrocznością, *Klinika Oczna*, 104(5-6): 381-383, 2002

2. Urban R., Lemancewicz A., Urban J., Krętowska M., Ocena stężeń cząsteczek adhezyjnych sICAM-1 i sVCAM-1 w surowicy krwi pacjentek z zagrażającym porodem przedwczesnym, *Ginekologia Polska*, 74(10): 1325-1328, 2003
3. Urban B., Ustymowicz A., Bakunowicz -Łazarczyk A., Krętowska M., Wpływ penoksyfiliny na dopplerowski parametry przypływu krwi w tętnicy środkowej siatkówki i w tętnicach rzęskowych tylnych krótkich u młodzieży z krótkowzrocznością postępującą, *Klinika Oczna* 3: 318-320, 2004
4. Urban B., Ustymowicz A., Bakunowicz-Łazarczyk A., Krętowska M., Ocena śródbłonka rogówki po operacjach usunięcia zaćmy u dzieci i młodzieży, *Klinika Oczna* 107(1-3): 43-45, 2005
5. Urban B., Peczyńska J., Urban M., Bakunowicz-Łazarczyk A.,Krętowska M., Śródbłonek rogówki u dzieci i młodzieży z cukrzycą insulinozależną, *Magazyn Okulistyczny* 2(4):262-267, 2005
6. Urban B., Peczyńska J., Głowińska-Olszewska B., Urban M., Bakunowicz-Łazarczyk A., Krętowska M., Analiza centralnej grubości rogówki u dzieci i młodzieży z cukrzycą insulinową, Klinika Oczna 10-12: 418-420, 2007
7. Urban B., Krętowska M., Szumiński M., Bakunowicz-Łazarczyk A., Ocena przedniego odcinka oczu normowzrocznych, nadwzrocznych i krótkowzrocznych u dzieci i młodzieży za pomocą optycznej koherentnej tomografii OCT Visante. Klinika Oczna 111(1):18-21, 2012
8. Urban B., Raczyńska D., Bakunowicz-Łazarczyk A., Raczyńska K., Kretowska M., Evaluation of Corneal Endothelium in Children and Adolescents with Type 1 Diabetes Mellitus, Mediators of Inflammation, 2013 [IF2013=2,42; IF5y=3,37]
9. Urban B., Bakunowicz-Łazarczyk A., Michalczuk M., Krętowska M., Evaluation of Corneal Endothelium in Adolescents with Juvenile Glaucoma, *Journal of Ophthalmology*, Volume 2015, Article ID 895428, 6 pages, 2015 [IF2015=1,46; IF5y=1,68]
10. Michalczuk M., Urban B., Chrzanowska-Grenda B., Oziębło-Kupczyk M., Bakunowicz-Łazarczyk A., Krętowska M., The assessment of multifocal ERG responses in school-age children with history of prematurity, *Documenta Opthalmologica* 132:47–55, 2016 [IF2016=1,92; IF5y=1,73]

- W ramach współpracy z Kliniką Endokrynologii, Diabetologii i Chorób Wewnętrznych Uniwersytetu Medycznego w Białymstoku powstała następująca publikacja:

1. Kretowski A., Wawrusiewicz N., Mironczuk K., Mysliwiec J., Kretowska M., Kinalska I., Intercellular adhesion molecule 1 gene polymorphisms in Graves' disease, *Journal of Clinical Endocrinology and Metabolism*, 88(10): 4945-4949, 2003 [IF2003=5,87]

02.04.2019 Małgorzata Krętowska

....

Data Podpis kandydata