Diagnozowanie przypadków zapalenia wątroby

Tomasz Szewczuk, 226163
Wydział Elektroniki
Politechnika Wrocławska
Wrocław, Polska
226163@student.pwr.edu.pl

Adam Węglowski, 226175
Wydział Elektroniki
Politechnika Wrocławska
Wrocław, Polska
226175@student.pwr.edu.pl

Streszczenie - W pracy przedstawione zostało działanie sieci neuronowych, naiwnego klasyfikatora Bayesa oraz drzew decyzyjnych na podstawie poprawności diagnozowania chorób. W tym przypadku wykorzystana została baza danych opisująca przypadki zapalenia wątroby. Eksperyment został wykonany, a jego wyniki przedstawione i omówione w poniższej pracy.

Słowa kluczowe: sieci neuronowe, drzewa decyzyjne, naiwny klasyfikator bayesowski, zapalenie wątroby, diagnozowanie

1. CEL BADAŃ

Podczas teoretycznego opracowywania tematu wyłonione zostały następujące cele projektowe:

- wspomaganie diagnostyki zapalenia wątroby,
- wykonanie badań,
- porównanie wyników dla wykorzystanych metod tzn. sztucznych sieci neuronowych, drzew decyzyjnych oraz naiwnego klasyfikatora bayesowskiego,
- zapoznanie czytelników ze sztucznymi sieciami neuronowymi, drzewami decyzyjnymi oraz naiwnym klasyfikatorem bayesowskim.

2. OPIS TECHNOLOGII

2.1 Sieci Neuronowe

Sieci neuronowe zyskały swoją popularność już w poprzednich dekadach ubiegłego wieku, jednak ich prawdziwy rozwój miał miejsce dopiero w XXI wieku. Są one jednym z wielu zagadnień, szeroko pojętej, sztucznej inteligencji. Zyskują one coraz większe zainteresowanie wśród naukowców i obserwatorów interesujących się nowymi technologiami. [1]

Podstawowym elementem budowy sieci neuronowych jest pojedynczy neuron, który łączy się z innymi neuronami, tworząc poszczególne warstwy sieci. Budowa oraz sposób działania ma naśladować działanie ludzkiego mózgu. Sieć neuronowa składa się najczęściej z trzech połączonych ze sobą warstw. Pierwsza warstwa nazywana jest warstwą wejścia – gromadzi ona dane, a następnie przesyła dalej (każdy neuron pierwszej warstwy dostarcza dane, do każdego neuronu drugiej warstwy, czyli warstwy ukrytej). Druga warstwa to już wspomniana wyżej warstwa ukryta – to w niej ma miejsce proces uczenia. Trzecia warstwa to warstwa wyjściowa – dostarcza wyniki przeprowadzonego eksperymentu.

Na pierwszy rzut oka układ wydaje się prosty, jednak nie ujmuje to złożoności, ponieważ sieć może opierać się na nieskończonej liczbie warstw neuronowych.

2.2 Naiwny klasyfikator Bayesowski

Naiwny klasyfikator Bayesowski jest prostym probabilistycznym klasyfikatorem, bazującym na twierdzeniu Bayesa (twierdzenie teorii prawdopodobieństwa, wiążące prawdopodobieństwa warunkowe dwóch zdarzeń warunkujących się nawzajem), nadaje się szczególnie do problemów o bardzo wielu wymiarach na wejściu. Mimo prostoty metody, często działa ona lepiej od innych, bardzo skomplikowanych metod klasyfikujących. Naiwne klasyfikatory bayesowskie są oparte na założeniu o wzajemnej niezależności predyktorów (zmiennych niezależnych). Często nie mają one żadnego związku z rzeczywistością i właśnie z tego powodu nazywa się je naiwnymi.

2.3 Drzewa decyzyjne

Drzewa decyzyjne są graficzną metodą wspomagania procesu decyzyjnego. Jest to jedna z najczęściej wykorzystywanych technik analizy danych. Drzewa składają się z korzenia oraz gałęzi prowadzących z korzenia do kolejnych wierzchołków. Wierzchołki, z których wychodzi co najmniej jedna krawędź, są nazywane węzłami, a pozostałe wierzchołki – liśćmi. W każdy węźle sprawdzany jest pewien warunek dotyczący danej obserwacji i na jego podstawie wybierana jest jedna z gałęzi prowadząca do kolejnego wierzchołka. Klasyfikacja danej obserwacji polega na przejściu od korzenia do liścia i przypisaniu do tej obserwacji klasy zapisanej w danym liściu. [2]

Drzewa decyzyjne znajdują praktyczne zastosowanie w różnego rodzaju problemach decyzyjnych, szczególnie takich, gdzie występuje dużo rozgałęziających się wariantów, a także w warunkach ryzyka. Wykorzystuje się je również w problemach związanych z klasyfikacją i predykcją pojęć typu: diagnostyka medyczna, przewidywanie wydajności, akceptacja oraz udzielanie kredytów.

3. BAZA DANYCH ORAZ ŚRODOWISKO

Do przeprowadzenia badań została użyta baza danych o nazwie ‘Hepatitis’ pobrana ze strony datahub.io. Baza ta posada 20 atrybutów. Została stworzona w 1988 roku.

Analiza badawcza przeprowadzona została w środowisku Waikato Environment for Knowledge Analysis (WEKA)

[3]. Wyżej wspomniany system umożliwia używanie różnych metod badawczych od sieci neuronowych i bayesowskich aż do zadań opartych na eksploracji danych. Klasyfikatory, które zostały wybrane do badań w oprogramowaniu WEKA znajdują się kolejno:

- NaiveBayes w grupie *bayes*
- MultilayerPerceptron w grupie *functions*
- RandomForest w grupie *trees*

4. STUDIA LITERATUROWE

W pracy badawczej pod tytułem *Artificial Intelligence Based Expert System For Hepatitis B Diagnosis* został stworzony inteligentny system do diagnozowania zapalenia wątroby typu B. System ten składa się z uogólnionej sieci neuronowej z regresją, która daje wynik określający, czy u pacjenta występuje wirusowe zapalenie wątroby typu B, czy nie oraz stopień zaawansowania choroby pacjenta. Z wyników przeprowadzonych badań można wywnioskować, że jest to metoda bardzo skuteczna, która w przyszłości będzie coraz częściej wykorzystywana w diagnozowaniu zapalenia wątroby. [4]

Mehrbakhsh Nilashia, Hossein Ahmadib, Leila Shahmoradicd, Othman Ibrahima oraz Elnaz Akbari w pracy pod tytułem *A predictive method for hepatitis disease diagnosis using ensembles of neuro-fuzzy technique* badali zastosowanie zespołów Neuro-Fuzzy Inference System do przewidywania choroby zapalenia wątroby. Korzystali również z drzew decyzyjnych do wyboru najważniejszych funkcji eksperymentalnego zestawu danych. Wyniki analiz wykazały, że wydajność tej metody jest lepsza niż w sieciach neuronowych, ANFIS oraz KNN. [5]

W artykule *Artificial Neural Network Accurately Predicts Hepatitis B Surface Antigen Seroclearance* przedstawiono badania, które miały na celu opracowanie sztucznych sieci neuronowych. Mogły one wykryć obecność antygenu HBs na podstawie dostępnych zmiennych w surowicy. Przeanalizowano dane od 203 nieleczonych pacjentów. W przypadku HBs wartości AUROC dla sztucznych sieci neuronowych wyniosły odpowiednio 0,96, 0,93 i 0,95 dla grup treningu, testowania i genotypu B. Były one znacznie wyższe niż dla LRM, qHBs i HBV DNA. Sieci neuronowe identyfikują pojawienie się antygenu u pacjentów z większą dokładnością na podstawie łatwo dostępnych danych surowicy. [6]

W pracy doktorskiej Rafała Adamczaka pod tytułem *Zastosowanie sieci neuronowych do klasyfikacji danych doświadczalnych* zaprezentowana została sieć MLP2LN, której celem jest ekstrakcja reguł. Sieć MLP2LN została zastosowana do analizy danych medycznych. Cel został osiągnięty, przez autora, poprzez modyfikację funkcji błędu stosowanej w sieciach MLP oraz opracowanie algorytmu budowania sieci. Modyfikacja funkcji błędu pozwala uniknąć problemu przeuczenia. Sposób konstrukcji sieci i jej specyficzna architektura rozwiązują problem liczby węzłów ukrytych oraz umożliwiają bardzo szybkie uczenie. Sieć MLP2LN może być zastosowana zarówno do danych dyskretnych, jak również do danych ciągłych. [7]

5. PRZYGOTOWANIE DO BADAŃ

5.1 Preprocess / selekcja danych

Plan eksperymentu zakładał przeprowadzenie badań dla kolejno: wszystkich dostępnych cech (20), 11 oraz 6 najistotniejszych atrybutów.

Środowisko badawcze WEKA posiada zaimplementowane filtry dla wstępnego przetwarzania danych. Filtr, który został wybrany do pracy z danymi to *AttributeSelection*. Jest to nadzorowany filtr, którego można używać do wybierania atrybutów. Cechuje go elastyczność oraz umożliwia on łączenie różnych metod wyszukiwania i oceny.

Pierwszym wybranym sposobem oceny atrybutów była metoda *CfsSubsetEval*. Oceniała ona wartość podzbioru atrybutów, biorąc pod uwagę indywidualną zdolność predykcyjną każdej cechy wraz ze stopniem redundancji między nimi. Po przeprowadzeniu tej operacji uzyskano 11 istotnych dla wyników atrybutów.

Do dalszych prac z atrybutami (celem było wyselekcjonowanie 6 najlepszych atrybutów) zastosowano metodę *InfoGainAttributeEval*. Wykorzystuje ona entropię do oceny istotności atrybutów. Bada istotność atrybutów ze względu współczynnik *InfoGain*.

$$InfoGain(Class, Attribute) = H(Class) - H(Class|Attribute)$$

Metoda wykonuje ocenę każdego atrybutu ze względu na przyjęte kryterium niezależnie, dla każdego z atrybutów osobno.

5.2 Wybór parametrów

Zanim przystąpiono do właściwych badań dokładności klasyfikacji wybranych metod, przeprowadzono wybór najlepszych parametrów do badań. Wybór ten został zrealizowany dla sieci neuronowej (momentum, wskaźnik uczenia, liczba epok, liczba warstw ukrytych) oraz drzewa decyzyjnego (maksymalna głębokość).

Tabela 1. Najlepsze parametry do testów dla sieci neuronowej

Badany parametr	Wartość
Momentum	0,2
Wskaźnik uczenia	0,3
Liczba epok	300
Liczba warstw ukrytych	a

Gdzie dla liczby warstw ukrytych:

$$a = \frac{(liczba\ atrybutów + liczba\ klas)}{2}$$

Tabela 2. Najlepszy parametr do testów dla drzew decyzyjnych

Badany parametr	Wartość
Maksymalna głębokość	10

6. BADANIA

Badania zostały przeprowadzone dla walidacji krzyżowej kolejno: 5-krotnej, 10-krotnej, 15-krotnej oraz 20-krotnej. Miarą badawczą jest dokładność (liczona w %). Natomiast statystyczną miarą porównawczą jest średnia arytmetyczna.

$$M = \frac{a_1 + a_2 + \cdots + a_n}{N}$$

gdzie:
M – średnia arytmetyczna
a_i – poszczególne wyniki
N – liczba wyników

6.1 Naiwny klasyfikator Bayesa

Tabela 3. Wyniki dokładności klasyfikacji w % dla naiwnego klasyfikatora Bayesa

Walidacja / Ilość cech / [%]	6	11	20
5	85,16	88,39	82,58
10	85,81	87,74	84,52
15	85,81	87,10	84,52
20	85,81	87,10	83,23

6.2 Sieci neuronowe

Tabela 4. Wyniki dokładności klasyfikacji w % dla sieci neuronowych

Walidacja / Ilość cech / [%]	6	11	20
5	86,67	78,71	79,36
10	86,00	82,58	81,94
15	86,00	85,81	82,58
20	87,34	81,29	80,65

6.3 Drzewa decyzyjne

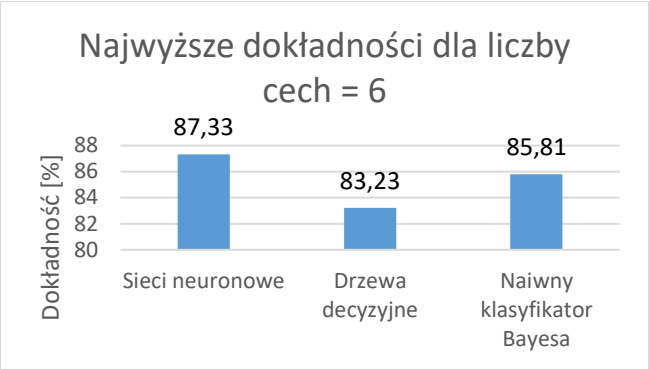
Tabela 5. Wyniki dokładności klasyfikacji w % dla drzew decyzyjnych

Walidacja / Ilość cech / [%]	6	11	20
5	83,23	84,52	85,81
10	82,58	84,52	85,81
15	83,23	81,94	83,87
20	83,23	84,52	85,16

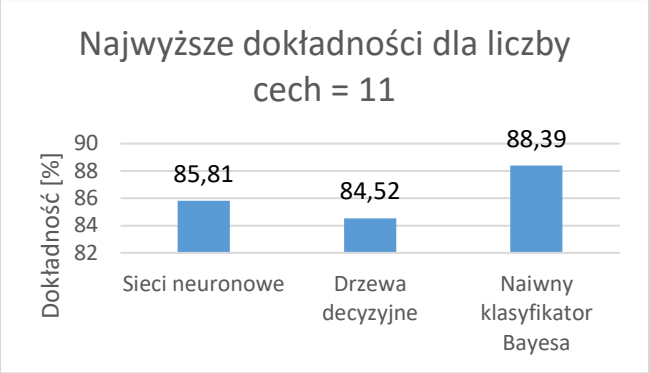
6.4 Porównanie wyników badań

Tabela 6. Porównanie czułość i swoistości dla wszystkich algorytmów

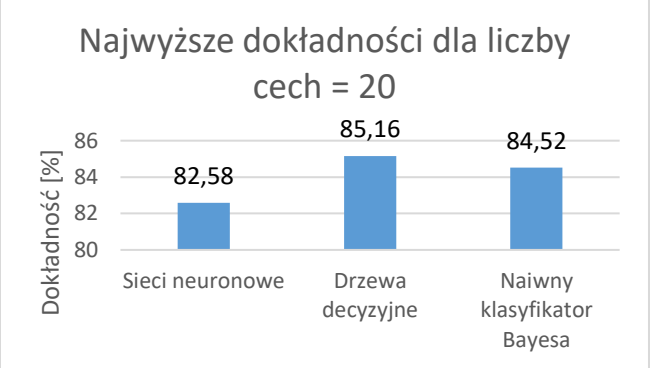
Algorytm	Czułość [%]	Swoistość [%]
Sieci neuronowe	88	56,7
Drzewa decyzyjne	88,1	76,2
Naiwny klasyfikator Bayesa	91,6	61,1



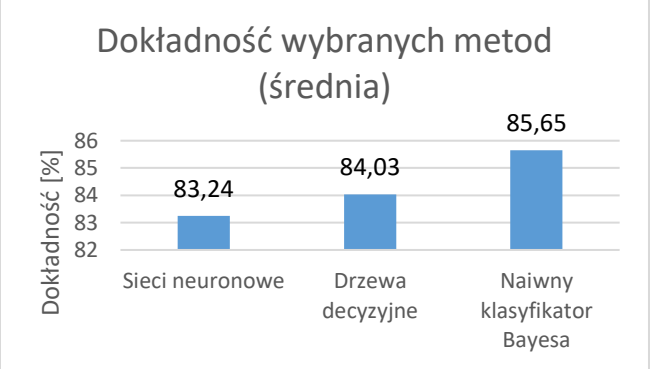
Rysunek 1. Najwyższe dokładności klasyfikacji w % dla danej liczby cech (6)



Rysunek 2. Najwyższe dokładności klasyfikacji w % dla danej liczby cech (11)



Rysunek 3. Najwyższe dokładności klasyfikacji w % dla danej liczby cech (20)



Rysunek 4. Dokładność klasyfikacji w % dla badanych metod (średnia z uzyskanych wyników)

Przeprowadzono analizę jakościową badanych systemów w celu lepszego poznania działania oraz zachowania się algorytmów. W wyniku zbadania swoistości oraz czułości można zauważyć, że niezależnie od algorytmu stosunek zbioru choroby błędnie sklasyfikowanej jest zdecydowanie większy dla danych przedstawiających die (‘0’) niż live (‘1’). Powodem takiego rezultatu może być baza danych, w której instancje klasy die stanowią tylko 20% całego zbioru. Wyniki eksperymentu przedstawione zostały w tabeli *Tabela 6*.

7. WNIOSKI I PODSUMOWANIE

Z przeprowadzonych badań wynika, że naiwny klasyfikator Bayesa jest najlepszym klasyfikatorem do rozpoznawania zapalenia wątroby. Naiwny klasyfikator Bayesa uzyskał najlepszy wynik pod względem dokładności oraz najlepszy średni wynik odnoszący się do dokładności. Spośród wybranych miękkich metod obliczeniowych najlepszy wynik uzyskały sieci neuronowe, jednakże najlepszy średni wynik uzyskały drzewa decyzyjne. Z wyników przeprowadzonego eksperymentu okazuje się, że sieci neuronowe najlepiej radzą sobie z mniejszą ilością cech, ale tracą na dokładności, kiedy cech zaczyna przybywać. Największą odporność na wpływ ilości cech na dokładność mają drzewa decyzyjne. Jednak zauważyć można, że ostatecznie wszystkie metody poradziły sobie dość dobrze z klasyfikacją choroby. Wszystkie wyniki na poziomie około 80-85%. Na uzyskany wynik mogła mieć jednak wpływ zastosowana baza „Hepatitis”. W powyższej bazie do dyspozycji oddano tylko 155 instancji problemu (123 z etykietą ‘live’ oraz 32 z etykietą ‘die’). Przy zastosowaniu innej bazy danych wynik mógłby być inny. Dalsze badania z różnymi bazami danych mogłyby ostatecznie ukazać, która metoda jest najlepsza.

Bibliografia

1) Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, “Deep learning”, Nature, vol. 521, no. 7553, p. 436, 2015
2) Przemysław Marynowski „Drzewa decyzyjne” Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie, 2013
3) datahub.io/machine-learning/hepatitis – baza danych dostęp 17.11.2019
4) Dakshata Panchal, Seema Shah „Artificial Intelligence Based Expert System For Hepatitis B Diagnosis”, 2011
5) Mehrbakhsh Nilashia, Hossein Ahmadib, Leila Shahmoradicd, Othman Ibrahima, Elnaz Akbari „A predictive method for hepatitis disease diagnosis using ensembles of neuro-fuzzy technique”, 2019
6) Ming-Hua Zheng, Wai-Kay Seto, Ke-Qing Shi, Danny Ka-Ho Wong, James Fung, Ivan Fan-Ngai Hung, Daniel Yee-Tak Fong, John Chi-Hang Yuen, Teresa Tong, Ching-Lung Lai, Man-Fung Yuen „Artificial Neural Network Accurately Predicts Hepatitis B Surface Antigen Seroclearance”, 2014
7) Rafał Adamczak „Zastosowanie sieci neuronowych do klasyfikacji danych doświadczalnych”, praca doktorska, 2001