Metody techniki systemów w medycynie 2

Stanisław Strauchold 259142
Rafał Kwiecień 259146

26 listopada 2023

Temat projektu: Kardiotokografia (CTG) z wykorzystaniem 10 klas

1 Analiza problemu oraz przegląd literatury

1.1 Analizowany problem

Analizowanym przez nas problemem jest zbiór danych z pomiaru kardiotokograficznego (CTG). Naszym zadaniem jest stworzenie modelu, który na podstawie danych z badania będzie w stanie przyporządkować wyniki do jednej z 10 klas.

Tabela 1: Charakterystyka analizowanego zbioru danych

Liczba instancji	2126
Liczba cech	35
Liczba klas	10

Tabela 2: Liczność klas

Numer klasy	Liczność	Udział w zbiorze [%]
0	197	9,27
1	384	18,06
2	579	27,23
3	53	2,49
4	81	3,81
5	72	3,39
6	332	15,62
7	252	11,85
8	107	5,03
9	69	3,25

Tabela 3: Charakterystyka funkcji w zbiorze

Cecha	Liczba brakujących	Wartość [%]
v1	382	17,97
v2	388	18,25
v3	380	17,87
v4	372	17,50
v5	429	20,18
v6	377	17,73
v7	398	18,72
v8	405	19,05
v9	374	17,59
v10	388	18,25
v11	370	17,40
v12	359	16,89
v13	385	18,11
v14	395	18,58
v15	418	19,66
v16	394	18,53
v17	404	19,00
v18	426	20,04
v19	358	16,84
v20	367	17,26
v21	379	17,83
v22	365	17,17
v23	387	18,20
v24	377	17,73
v25	373	17,54
v26	358	16,84
v27	381	17,92
v28	388	18,25
v29	362	17,03
v30	360	16,93
v31	387	18,20
v32	388	18,25
v33	405	19,05
v34	385	18,11
v35	401	18,86

Po analizie zbioru ilość brakujących cech wynosi 13645, co stanowi 18,10% całego zbioru

1.2 Problem danych niezbalansowanych

O problemie danych niezbalansowanych mówimy kiedy liczność poszczególnych klas różni się od siebie. Analizowany przez nas zbiór danych jest niezbalansowany, co widać w tabeli 2. Najmniej liczna klasa (klasa 3) występuje 53 razy, natomiast najliczniejsza (klasa 2) występuje 579 razy. W przypadku zastosowania mechanizmów uczenia na tym zbiorze może zajść zjawisko, które doprowadzi do częstszego przewidywania przez klasyfikator klas dominujących (bardziej licznych). Sposobem rozwiązania problemu danych niezbalansowanych jest preprocesing, polega na wyrównaniu liczności klas. [2]

Dwie najpopularniejsze metody preprocesingu danych:

- oversampling – powielanie instancji klasy o mniejszej liczności
- undersampling – zmniejszenie instancji klasy o większej liczności

1.3 Problem brakujących wartości

[3]W naszym zbiorze danych występuje problem braku niektórych wartości cech. Z problemem tym można sobie poradzić na różne sposoby:

- usunięcie wierszy z brakującymi danymi
- uzupełnienie brakujących wartości medianą dla każdej kolumny (cechy)
- stworzenie modelu przewidującego

1.4 Uczenie maszynowe

Uczenie maszynowe – gałąź informatyki, służy do analizy danych poprzez wykorzystanie automatycznych modeli analitycznych.

Jedną z kategorii uczenia maszynowego jest uczenie nadzorowane, które dzieli się na regresję oraz klasyfikację. Nasz problem należy do klasyfikacji, ponieważ zadaniem jest przewidzenie kategorii. Uczenie nadzorowane wykorzystuje zbiór trenujący, który składa się z danych wejściowych (cech) oraz przypisanych etykiet (klas). Po nauce na zbiorze trenującym system mając dane, ma przypisać im odpowiednią klasę.

1.5 Wstępne przetwarzanie danych

[4]Celem wstępnego przetwarzania danych jest odpowiednie przygotowanie zbioru, tak by algorytmy zbudowały możliwie najbardziej precyzyjny model.

Do wstępnego przetwarzania danych zaliczają się:

- selekcja danych

- czyszczenie danych
- redukcja liczby cech
- transformacja
- dyskretyzacja wartości

1.6 Redukcja wymiarowości

Wielowymiarowość jest jednym z podstawowych problemów w klasyfikacji. Ma ona negatywny wpływ na efektywność algorytmów. Dzięki redukcji wymiarowości obiektów uzyskujemy poprawę wyników predykcji, zmniejszamy złożoność obliczeniową oraz poprawiamy jakość danych. Wielowymiarowość może zostać rozwiązana z wykorzystaniem procesu selekcji cech, którego zadaniem jest wybranie cech istotnych i odrzucenie nieistotnych. [1]

1.7 Selekcja cech

Metody selekcji cech zawierają cztery etapy:

- generacja podzbioru cech
- ocena podzbioru cech
- kryterium stopu
- walidacja rezultatów

Wśród tych metod wyróżniamy dwa rodzaje procedur: filtry oraz wrappery. Filtry bazują na niezależnej ocenie cech wykorzystując ogólne charakterystyki danych. Wykorzystuje się do tego na przykład współczynniki korelacji między wartościami cech, a przynależnością do klasy. Zbiór cech poddawany jest filtracji w celu określenia najbardziej obiecującego podzbioru atrybutów. Wykonuje się to przed trenowaniem algorytmu.

Wrappery oceniają poszczególne podzbiory cech wykorzystując do tego algorytmy uczenia maszynowego. Algorytmy te zostaną potem wykorzystane w zadaniu klasyfikacji. Algorytm uczący zawarty jest w procedurze selekcji cech.

Proces generowania podzbioru atrybutów może się odbyć na różne sposoby: tworzenie indywidualnego rankingu, przeszukiwanie w przód, przeszukiwanie wstecz.

2 Plan eksperymentu

2.1 Cel eksperymentu

Celem eksperymentu jest zbadanie wpływu metod imputacji oraz ilości selekcjonowanych cel na jakość klasyfikatorów opartych o uczenie nadzorowane. Do tego zbadany zostanie również wpływ stosowanych metod selekcji cech. W eksperymencie zostaną użyte następujące klasyfikatory:

- CART - drzewo klasyfikacyjne i regresyjne
- GNB - naiwny klasyfikator bayesowski
- KNN - klasyfikator k-najbliższych sąsiadów

2.2 "Research questions"

- Które metody imputacji najlepiej wpływają na jakość poszczególnych klasyfikatorów?
- Jak metoda selekcji cech wpływa na jakość klasyfikacji?
- Jaki jest wpływ ilości selekcjonowanych cech na jakość klasyfikacji?

2.3 Podejścia

2.3.1 Podejście 1

- **sposób radzenia sobie z problemem brakujących wartości:** uzupełnienie brakujących wartości medianą wartości cech
- **selekcja cech:** metoda Select From Model z biblioteki scikit-learn
- **algorytm klasyfikacji:** CART

2.3.2 Podejście 2

- **sposób radzenia sobie z problemem brakujących wartości:** uzupełnienie brakujących wartości medianą wartości cech
- **selekcja cech:** rekurencyjna eliminacja cech
- **algorytm klasyfikacji:** GNB

2.3.3 Podejście 3

- **sposób radzenia sobie z problemem brakujących wartości:** uzupełnienie brakujących wartości medianą wartości cech
- **selekcja cech:** metoda współczynnika Pearsona
- **algorytm klasyfikacji:** KNN

2.4 Plan eksperymentu

1. Wczytania danych z pliku.
2. Wyodrębnienie cech oraz klas
3. Badanie wpływu metody imputacji na jakość klasyfikatorów
    - Uzupełnienie brakujących wartości jedną z trzech metod imputacji
    - selekcja stałej liczby cech wykorzystując metodę współczynnika Pearsona
    - Stworzenie modeli klasyfikacyjnych z wykorzystaniem walidacji krzyżowej oraz jednego z trzech klasyfikatorów
    - Zapisanie średniej wartości trafności
    - powtórzenie pętli dla każdej z par metoda imputacji - klasyfikator
4. Uzupełnienie brakujących wartości metodą imputacji, dla której uzyskano najlepsze wyniki w punkcie 3.
5. Badanie wpływu metod selekcji cech na jakość klasyfikatora
    - Wybór stałej liczby cech wykorzystując jedną z 3 metod selekcji
    - Stworzenie modeli klasyfikacji danych z wykorzystaniem walidacji krzyżowej oraz klasyfikatora CART
    - Zapisanie średniej wartości trafności
    - powtórzenie pętli dla każdej pary metoda selekcji cech - klasyfikator CART
6. Badanie wpływu ilości selekcjonowanych cen na jakość modelu
    - Wybór od 1 do 20 cech metodą rekurencyjnej eliminacji cech
    - Stworzenie modeli klasyfikacji danych z wykorzystaniem walidacji krzyżowej oraz klasyfikatora CART
    - Zapisanie średniej wartości trafności dla każdej pary

2.5 Opis środowiska eksperymentalnego

Eksperyment zostanie przeprowadzony z wykorzystaniem języka Python 3 w środowisku Visual Studio Code. W badaniu wykorzystamy biblioteki: scikit-learn, pandas oraz numpy.

3 Wyniki eksperymentów

3.1 Eksperyment 1

Trafność dla każdej pary metoda imputacji - klasyfikator GNB:
GNB - imputacja metodą mediany cech: 0.789 (0.03)
GNB - imputacja metodą średniej wartości cech: 0.842 (0.03)
GNB - imputacja metodą najbliższych sąsiadów: 0.753 (0.02)

Tabela 4: Tabela obserwacji końcowych dla 3 metod imputacji oraz klasyfikatora GNB

	GNB median	GNB mean	GNB nearest
GNB median	0	0	1
GNB mean	1	0	1
GNB nearest	0	0	0

Trafność dla każdej pary metoda imputacji - klasyfikator KNN:
KNN - imputacja metodą mediany cech: 0.333 (0.01)
KNN - imputacja metodą średniej wartości cech: 0.328 (0.01)
KNN - imputacja metodą najbliższych sąsiadów: 0.344 (0.02)

Tabela 5: Tabela obserwacji końcowych dla 3 metod imputacji oraz klasyfikatora KNN

	KNN median	KNN mean	KNN nearest
KNN median	0	0	0
KNN mean	0	0	0
KNN nearest	0	0	0

Trafność dla każdej pary metoda imputacji - klasyfikator CART:
CART - imputacja metodą mediany cech: 0.827 (0.02)
CART - imputacja metodą średniej wartości cech: 0.844 (0.02)
CART - imputacja metodą najbliższych sąsiadów: 0.794 (0.02)

Tabela 6: Tabela obserwacji końcowych dla 3 metod imputacji oraz klasyfikatora CART

	CART median	CART mean	CART nearest
CART median	0	0	1
CART mean	1	0	1
CART nearest	0	0	0

Wnioski: Wyniki dla każdej z par metoda imputacji - klasyfikator KNN okazały się bardzo niesatysfakcjonujące, stąd można wysnuć przypuszczenie, iż klasyfikator ten źle sprawdza się w przypadku klasyfikowania zbiorów zawierających brakujące wartości. Dla klasyfikatorów GNB oraz CART najlepsze wyniki zostały uzyskane dla metody imputacji jaką jest metoda średniej wartości dla danej cechy, zaś najgorsze wyniki uzyskano dla metody najbliższych sąsiadów.

3.2 Eksperyment 2

Trafność dla każdej pary metoda selekcji cech - klasyfikator CART:
CART - selekcja cech metodą korelacji Pearsona (P): 0.849 (0.03)
CART - selekcja cech metodą SelectFromModel (SFM): 0.902 (0.02)
CART - selekcja cech metodą rekurencyjnej eliminacji cech (RFE): 0.950 (0.01)

Tabela 7: Tabela obserwacji końcowych dla 3 metod selekcji cech oraz klasyfikatora CART

	CART P	CART SFM	CART RFE
CART P	0	0	0
CART SFM	1	0	0
CART RFE	1	1	0

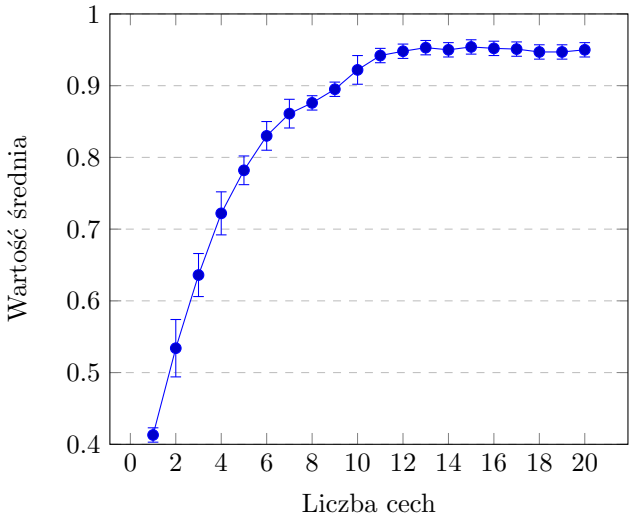
Wnioski: Spośród 3 metod selekcji cech najlepsze wyniki zostały uzyskane dla rekurencyjnej eliminacji cech, zaś najgorsze wyniki uzyskała metoda współczynnika korelacji Pearsona.

3.3 Eksperyment 3

Trafność dla każdej pary metoda ilość selekcjonowanych cech - klasyfikator CART:
CART - 1 cecha: 0.413 (0.01)
CART - 2 cech: 0.534 (0.04)
CART - 3 cech: 0.636 (0.03)
CART - 4 cech: 0.722 (0.03)
CART - 5 cech: 0.782 (0.02)
CART - 6 cech: 0.830 (0.02)
CART - 7 cech: 0.861 (0.02)
CART - 8 cech: 0.876 (0.01)
CART - 9 cech: 0.895 (0.01)
CART - 10 cech: 0.922 (0.02)
CART - 11 cech: 0.942 (0.01)
CART - 12 cech: 0.948 (0.01)

CART - 13 cech: 0.953 (0.01)
CART - 14 cech: 0.950 (0.01)
CART - 15 cech: 0.954 (0.01)
CART - 16 cech: 0.952 (0.01)
CART - 17 cech: 0.951 (0.01)
CART - 18 cech: 0.947 (0.01)
CART - 19 cech: 0.947 (0.01)
CART - 20 cech: 0.950 (0.01)

Wykres 1. Średnia wartość dla każdej pary: ilość selekcjonowanych cech - klasyfikator CART



Wnioski: Jakość klasyfikacji rosła logarytmicznie w zależności od ilości selekcjonowanych cech dla metody RFE. Od pewnego momentu różnice w jakości klasyfikacji względem ilości selekcjonowanych cech były coraz mniejsze, wręcz nieznaczące.

Literatura

[1] Paweł Zimeba, Redukcja wymiarowości i selekcja cech w zadaniach klasyfikacji i regresji z wykorzystaniem uczenia maszynowego. Zeszyty naukowe Uniwersytetu Szczecińskiego 2012

[2] Jason Brownlee, Random oversampling and undersampling for imbalanced classification. Styczeń 2020

[3] https://www.jmlr.org/papers/volume8/saar-tsechansky07a/saartsechansky07a.pdf

[4] Vahid Mirjalili Sebastian Raschka. Python. uczenie maszynowe. wydanie ii.