Wrocław, 2020

Komputerowe wspomaganie diagnozowania stanów ostrego brzucha z wykorzystaniem algorytmu k-NN

PROJEKT – Zastosowanie informatyki w medycynie

AUTORZY:

Karolina Filipiak, 226008

Karol Maśluch, 235044

Spis treści

[1. Opis problemu medycznego 3](#_Toc41404667)

[2. Ranking cech 4](#_Toc41404668)

[3. Informacje o środowisku implementacyjnym 7](#_Toc41404669)

[4. Opis badań eksperymentalnych 7](#_Toc41404670)

[a. Parametry klasyfikacji 7](#_Toc41404671)

[b. Badania 7](#_Toc41404672)

# Opis problemu medycznego

Naszym zadaniem było nabycie umiejętności zastosowania oraz implementacji algorytmu, który będzie wspomagał diagnostykę stanów ostrego brzucha. W tym celu, należało skorzystać z algorytmu minimalno-odległościowego k-NN [1], bazując na zbiorze danych dostarczonych przez prowadzącego.

Na początku przystąpiliśmy do zdefiniowania problemu medycznego. Na podstawie materiału empirycznego oraz implementacji odpowiedniego mechanizmu, należało dobrać taki zbiór cech, dzięki któremu z jak największym prawdopodobieństwem uda się wystawić prawidłową diagnozę dla pacjenta. Problemem było również określenie optymalnej liczby cech, dla których uda się wystawić trafną diagnozę pacjentowi.

Analizując dokument z danymi dotyczącymi stanów ostrego brzucha, wyodrębniliśmy:

* 476 pacjentów ze stwierdzonym stanem ostrego brzucha, gdzie każdy z pacjentów został opisany zestawem 31 cech, na podstawie których zdiagnozowano u nich jedną z 8 odmian powyższej przypadłości
* 31 cech – atrybuty/parametry uzyskane z wywiadu z pacjentem oraz badania wstępnego. Wszystkie poniżej opisane cechy mają charakter dyskretny (skokowy), ponieważ są opisane konkretnymi, dopuszczalnymi wartościami i jednocześnie nie mogą być wyrażone wartościami pośrednimi z danego przedziału (przykład: cecha „Apetyt” może przyjmować wartość 1, 2 lub 3 i nie może przyjmować żadnej innej wartości spoza tego przedziału - np. 1,5).

*Tabela 1. Opis cech.*

|  |  |  |  |
| --- | --- | --- | --- |
| Lp. | Nazwa cechy | Dopuszczalne wartości | Klasyfikacja cechy |
| 1 | Płeć | 1, 2 | Symptomy ogólne |
| 2 | Wiek | 1, 2, 3, 4, 5 |
| 3 | Lokalizacja bólu na początku zachorowania | 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 | Ból |
| 4 | Lokalizacja bólu obecnie | 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 |
| 5 | Intensywność bólu | 1, 2, 3 |
| 6 | Czynniki nasilające ból | 1, 2, 3, 4 |
| 7 | Czynniki przynoszące ulgę | 1, 2, 3 |
| 8 | Progresja bólu | 1, 2, 3 |
| 9 | Czas trwania bólu | 1, 2, 3, 4 |
| 10 | Charakter bólu na początku zachorowania | 1, 2, 3 |
| 11 | Charakter bólu obecnie | 1, 2, 3, 4 |
| 12 | Nudności i wymioty | 1, 2, 3 | Oddawanie moczu |
| 13 | Apetyt | 1, 2, 3 |
| 14 | Wypróżnienia | 1, 2, 3 |
| 15 | Oddawanie moczu | 1, 2 |
| 16 | Poprzednie niestrawności | 1, 2 | Historia |
| 17 | Żółtaczka w przeszłości | 1, 2 |
| 18 | Poprzednie operacje brzuszne | 1, 2 |
| 19 | Leki | 1, 2 |
| 20 | Stan psychiczny | 1, 2, 3 | Ogólne badanie |
| 21 | Skóra | 1, 2, 3 |
| 22 | Temperatura (pacha) | 1, 2, 3, 4, 5, 6 |
| 23 | Tętno | 1, 2, 3, 4, 5, 6, 7, 8, 9 |
| 24 | Ruchy oddechowe powłok brzusznych | 1, 2 | Oglądanie brzucha |
| 25 | Wzdęcia | 1, 2 |
| 26 | Umiejscowienie bolesności uciskowej | 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 | Badania palpacyjne |
| 27 | Objaw Blumberga | 1, 2 |
| 28 | Obrona mięśniowa | 1, 2 |
| 29 | Wzmożone napięcie powłok brzusznych | 1, 2 |
| 30 | Opory patologiczne | 1, 2 |
| 31 | Objaw Murphy’ego | 1, 2 |

* 8 klas – dana klasa oznacza diagnozę uzyskaną na podstawie kombinacji cech

*Tabela 2. Opis klas.*

|  |  |  |
| --- | --- | --- |
| Lp. | Klasa (diagnoza) | Ilość wystąpień |
| 1 | Ostre zapalenie wyrostka robaczkowego | 141 |
| 2 | Zapalenie uchyłków jelit | 17 |
| 3 | Niedrożnośc mechaniczna jelit | 29 |
| 4 | Perforowany wrzód trawienny | 28 |
| 5 | Zapalenie woreczka żółciowego | 55 |
| 6 | Ostre zapalenie trzustki | 32 |
| 7 | Niecharakterystyczny ból brzucha | 157 |
| 8 | Inne przyczyny ostrego bólu brzucha | 17 |

*Wykres 1. Rozkład procentowy częstości występowania danej przypadłości w stanach ostrego brzucha.*

# Ranking cech

Każdy odrębny przypadek stanu ostrego brzucha jest opisany stałą ilością cech i na ich podstawie jest formułowana diagnoza. Daną diagnozę opisuje pewien zbiór cech wraz z konkretnymi wartościami tych cech, które są charakterystyczne dla danej klasy. Pewne wartości mogą być określone jako te, które z większym prawdopodobieństwem będą wskazywać na daną diagnozę, podczas gdy niektóre wartości są traktowane tylko jako pomocnicze, które mogą jedynie sugerować daną diagnozę. W związku z tym stworzyliśmy ranking cech, który posortował dane atrybuty w kolejności od najbardziej znaczących do najmniej znaczących przy kwalifikacji do diagnozy. W tym celu skorzystaliśmy z metody rekurencyjnej eliminacji cech. Za pomocą zewnętrznego estymatora wagi są przepisywane do poszczególnych cech, a następnie cechy o najgorszym współczynniku są eliminowane. Ostatnia wyeliminowana cecha zajmuję najwyższą pozycję w rankingu, wyznacza najlepszą cechę pod względem otrzymanego współczynnika. Jako estymator zewnętrzny wykorzystaliśmy współczynnik modelu liniowego.

Dodatkowo w celu eliminacji mniej znaczących cechy otrzymane z rankingu cech skrzyżowaliśmy z dodatkową selekcją cech, którą była jednoczynnikową selekcją wyłaniającą   
k-najlepszych cech. Jako czynnik wybraliśmy metrykę X2 (chi kwadrat) . Jako k obraliśmy 25 cech. Tabela 4 zawiera ranking wybranych cech po skrzyżowaniu. Ostatecznie zostało wybrane 24 cechy z oryginalnych 31.

*Tabela 3. Ranking cech.*

|  |  |  |
| --- | --- | --- |
| Pozycja w rankingu: | Numer cechy: | Nazwa cechy: |
| 1 | 27 | Objaw Blaumberga |
| 2 | 24 | Ruchy oddechowe powłok brzusznych |
| 3 | 12 | Nudności i wymioty |
| 4 | 19 | Leki |
| 5 | 21 | Skóra |
| 6 | 9 | Czas trwania bólu |
| 7 | 31 | Objaw Murphy’ego |
| 8 | 7 | Czynniki przynoszące ulgę |
| 9 | 20 | Stan psychiczny |
| 10 | 14 | Wypróżnienia |
| 11 | 6 | Czynniki nasilające ból |
| 12 | 5 | Intensywność bólu |
| 13 | 11 | Charakter bólu obecnie |
| 14 | 8 | Progresja bólu |
| 15 | 2 | Wiek |
| 16 | 30 | Opory patologiczne |
| 17 | 17 | Żółtaczka w przeszłości |
| 18 | 23 | Tętno |
| 19 | 29 | Wzmożone napięcie powłok brzusznych |
| 20 | 18 | Poprzednie operacje brzuszne |
| 21 | 15 | Oddawanie moczu |
| 22 | 26 | Umiejscowienie bolesności uciskowej |
| 23 | 13 | Apetyt |
| 24 | 22 | Temperatura (pacha) |
| 25 | 10 | Charakter bólu na początku zachorowania |
| 26 | 25 | Wzdęcia |
| 27 | 4 | Lokalizacja bólu obecnie |
| 28 | 1 | Płeć |
| 29 | 16 | Poprzednie niestrawności |
| 30 | 3 | Lokalizacja bólu na początku zachorowania |
| 31 | 28 | Obrona mięśniowa |

*Tabela 4. Wartości statystyki X2 dla poszczególnych cech.*

|  |  |  |
| --- | --- | --- |
| Lp. | Nazwa cechy | Wartość statystyki X2 |
| 1 | Płeć | 6.19 |
| 2 | Wiek | 21.2 |
| 3 | Lokalizacja bólu na początku zachorowania | 31.4 |
| 4 | Lokalizacja bólu obecnie | 70.8 |
| 5 | Intensywność bólu | 18.4 |
| 6 | Czynniki nasilające ból | 59.3 |
| 7 | Czynniki przynoszące ulgę | 87.4 |
| 8 | Progresja bólu | 30.5 |
| 9 | Czas trwania bólu | 84.2 |
| 10 | Charakter bólu na początku zachorowania | 32.1 |
| 11 | Charakter bólu obecnie | 127.8 |
| 12 | Nudności i wymioty | 84.6 |
| 13 | Apetyt | 54.4 |
| 14 | Wypróżnienia | 42.1 |
| 15 | Oddawanie moczu | 12.3 |
| 16 | Poprzednie niestrawności | 4.4 |
| 17 | Żółtaczka w przeszłości | 0.5 |
| 18 | Poprzednie operacje brzuszne | 7.0 |
| 19 | Leki | 15.5 |
| 20 | Stan psychiczny | 15.9 |
| 21 | Skóra | 20.2 |
| 22 | Temperatura (pacha) | 17.7 |
| 23 | Tętno | 58.1 |
| 24 | Ruchy oddechowe powłok brzusznych | 31.1 |
| 25 | Wzdęcia | 8.7 |
| 26 | Umiejscowienie bolesności uciskowej | 77.4 |
| 27 | Objaw Blumberga | 21.3 |
| 28 | Obrona mięśniowa | 13.0 |
| 29 | Wzmożone napięcie powłok brzusznych | 11.4 |
| 30 | Opory patologiczne | 8.5 |
| 31 | Objaw Murphy’ego | 13.6 |

*Tabela 5. Ranking cech po skrzyżowaniu.*

|  |  |  |
| --- | --- | --- |
| Pozycja w rankingu: | Numer cechy: | Nazwa cechy: |
| 1 | 27 | Objaw Blaumberga |
| 2 | 24 | Ruchy oddechowe powłok brzusznych |
| 3 | 12 | Nudności i wymioty |
| 4 | 19 | Leki |
| 5 | 21 | Skóra |
| 6 | 9 | Czas trwania bólu |
| 7 | 31 | Objaw Murphy’ego |
| 8 | 7 | Czynniki przynoszące ulgę |
| 9 | 20 | Stan psychiczny |
| 10 | 14 | Wypróżnienia |
| 11 | 6 | Czynniki nasilające ból |
| 12 | 5 | Intensywność bólu |
| 13 | 11 | Charakter bólu obecnie |
| 14 | 8 | Progresja bólu |
| 15 | 2 | Wiek |
| 17 | 17 | Żółtaczka w przeszłości |
| 18 | 23 | Tętno |
| 19 | 29 | Wzmożone napięcie powłok brzusznych |
| 20 | 15 | Oddawanie moczu |
| 21 | 26 | Umiejscowienie bolesności uciskowej |
| 22 | 13 | Apetyt |
| 23 | 22 | Temperatura (pacha) |
| 24 | 10 | Charakter bólu na początku zachorowania |

# Informacje o środowisku implementacyjnym

Badania zostały przeprowadzone przy pomocy języka Python [2] w wersji 3.7.3. Do stworzenia rankingu cech zostały użyte klasy *SVR* oraz *CHI2* z modułu *sklearn.feature\_selection* [3]. Klasy te przyjmowały dwa wektory, pierwszym były wartości cech, a drugim przypisanie do poszczególnych diagnoz. Algorytm również został zaimplementowany z użyciem biblioteki *sklearn* wykorzystując metody *KNeighborsClassifier* zwracający klasyfikator k-najbliższych sąsiadów. Na jego podstawie trenowany był zbiór uczący przy pomocy metody *fit*, a następnie przy pomocy funkcji *predict* przyjmującej jako parametr zbiór testowy dokonywana była diagnoza. Otrzymane wartości klasyfikacji porównano z rzeczywistym stanem badanego obiektu wykorzystując opisane w kolejnej części metryki klasyfikacji.

# Opis badań eksperymentalnych

W badaniu zostały porównywane wyniki klasyfikacji algorytmu k-najbliższych sąsiadów. Jako parametry użyto: ilość cech, ilość sąsiadów oraz metrykę odległości. Algorytm na podstawie zbioru uczącego miał za zadanie przypisywać wartościom cech zestawu testującego odpowiedne klasy diagnozy. Wydajność algorytmu było reprezentowana jako ilość poprawnie przypisanych diagnoz na ilość przypisani. Celem badania było znalezienie jak najlepszych parametrów wejściowych dla posiadanych danych.

## Parametry klasyfikacji

W tabeli 6 zostały zawarte wszystkie parametry wejściowe użyte w procesie klasyfikacji. Pierwszym parametrem jest ilość sąsiadów dla algorytmu k-najbliższych sąsiadów, wartości te wynosiło odpowiednio: 1, 5, 10. Kolejnym parametrem była metryka odległości, wybrane zostały odpowiednio: metryka Euklidesowa oraz metryka Manhattan. Metryce euklidesowej odpowiada używanej najczęściej metodzie pomiaru, gdzie najkrótszy odcinek linii prostej wyznacza nam najkrótszą odległość, metryka ta posiada wartości ciągłe. W metryce Manhattan odległość może być liczona tylko po siatce składającej się z komórek o zdefiniowanej długości boku, metryka ta różni się tym od metryki euklidesowej, tym że jej wartości przyjmują tylko określone wartości dyskretne. W metryce tej nie możemy też poruszać się inaczej niż po krawędziach sześcianów tworzących siatkę. Ostatnim parametrem jest ilość cech użyta w procesie klasyfikacji, ilość ta mieści się w przedziale zamkniętym od 1 do 24 cech.

*Tabela 6. Parametry klasyfikacji.*

|  |  |
| --- | --- |
| Rodzaj parametru: | Testowane wartości: |
| Algorytm | 1-NN, 5-NN, 10-NN |
| Metryka odległości | Euklides [4], Manhattan |
| Liczba cech | <1,24> |

## Badania

Przeprowadzenie badań odbyło się z wykorzystaniem 5 razy powtarzanej metody 2-krotnej walidacji krzyżowej. Wszystkich pacjentów losowo podzielono na zbiór uczący i testujący w proporcjach 1:1. Następnie uruchomiono na testowanych obiektach poszczególne algorytmy zmieniając kolejno wyżej wymienione rodzaje parametrów. Po wykonaniu eksperymentu zbiór uczący był zamieniany ze zbiorem testującym i ponownie była wykonywana analiza. Losowanie i testowanie było powtarzane pięciokrotnie dla każdej permutacji parametrów klasyfikacji. Wynikiem poszczególnego eksperymentu był procentowy wynik poprawnie przewidzianych diagnoz. Wyniki z pięciu rund były uśrednione, a najlepszy wynik był zapamiętywany. Dodatkowo dla eksperymentu z najwyższym wynikiem zapamiętywana była jego macierz konfuzji.

# Bibliografia

|  |  |
| --- | --- |
| [1] | O. Harrison, "towardsdatascience.com," 10 9 2018. [Online]. Available: https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761. |
| [2] | "https://www.python.org," [Online]. Available: https://www.python.org. [Accessed 20 6 2020]. |
| [3] | "scikit-learn," [Online]. Available: https://scikit-learn.org/stable/modules/classes.html. [Accessed 20 5 2020]. |
| [4] | "Wikipedia," [Online]. Available: https://en.wikipedia.org/wiki/Euclidean\_distance. [Accessed 20 5 2020]. |