Komputerowe wspomaganie diagnozowania zawałów z wykorzystaniem algorytmu kNN

Karolina Działek[242040] i Damian Koper[241292]

Politechnika Wrocławska, Wydział Elektroniki,
wybrzeże Stanisława Wyspiańskiego 27, 50-370 Wrocław
{242040, 241292}@student.pwr.edu.pl

Streszczenie Algorytm K-najbliżych sąsiadów (kNN) stanowi jedną z metod klasyfikacji. Jest prosty w implementacji w swojej podstawowej formie oraz wykonuje dość złożone zadania klasyfikacyjne. Cel niniejszego projektu to stworzenie programu do komputerowego wspomagania diagnozowania zawałów z wykorzystaniem algorytmu kNN. Do realizacji zadania wykorzystano pięć plików tekstowych jako dane wejściowe. Każdy z nich odpowiada osobnej klasie.

Słowa kluczowe: knn, myocardial, infarction

1 Wprowadzenie

Cel projektu stanowi stworzenie programu do komputerowego wspomagania diagnozowania zawałów z wykorzystaniem algorytmu kNN.

1.1 Problem medyczny jako zadania klasyfikacji

Zadanie klasyfikacji w projekcie polega na tym, aby wspomóc rozpoznawanie stanów zwałowych wśród pacjentów na podstawie danych zgromadzonych podczas badań na ludziach, u których potwierdzono jedną z następujących diagnoz:

- ból nie pochodzący z serca,
- dusznica bolesna – dławica piersiowa,
- dusznica Prinzmetala – dławica naczynioskurczowa,
- pełnościenny zawał serca,
- podwsierdziowy zawał serca.

W zadaniu klasyfikacji wyróżnić można pewne pojęcia, w celu lepszego jego opisu:

- Klasa – pewna podprzestrzeń wartości zestawu danych, która w uczeniu nadzorowanym posiada swoją etykietę. Problem klasyfikacji jest odpowiedzią na pytania do jakiej klasy przyporządkować nowo napotkany zestaw wartości. Z punktu widzenia medycznego jest to zakwalifikowanie pacjenta jako zdrowego lub chorego z wyróżnieniem chorób na podstawie liczebności klas w danych uczących.

– Cecha – właściwość, która opisuje daną klasę. W medycynie jest to między innymi płeć, wiek, samopoczucie, czy też wynik badań.

Wynik zadania klasyfikacji to przyporządkowanie każdego z pacjentów do jednej z wymienionych klas. Jakość klasyfikacji za pomocą klasyfikatora k najbliższych sąsiadów została zbadana w zależności od liczby cech uwzględnionych podczas uczenia, a także zastosowanej metryki odległości.

1.2 Opis cech

Dane uczące to pięć plików tekstowych, przy czym każdy z nich odpowiada osobnej klasie i zawiera opis tego samego zestawu cech. Zbiór danych zawiera 5 klas, 59 cech oraz 901 rekordów. Opis poszczególnych cech z podziałem na ich charakter i możliwe do przyjęcia wartości zawiera tabela 1.

Tabela 1. Opis zbioru cech danych uczących i treningowych klasyfikatora.

L.p.	Cecha	Charakter	Wartości
	Ogólne		
1	wiek	dyskretny	liczby naturalne
2	płeć	dychotomiczny	0 - K, 1 - M
	Ból		
3	miejsce	kategoryczny	tabela 2
4	promieniowanie w klatce piersiowej	kategoryczny	tabela 3
5	charakter	kategoryczny	tabela 4
6	początek występowania	kategoryczny	tabela 5
7	liczba godzin od rozpoczęcia	dyskretny	liczby naturalne
8	długość trwania poprzedniego	kategoryczny	tabela 6
	Powiązane objawy		
9	nudności	dychotomiczny	0 - brak, 1 - obecny
10	potliwość	dychotomiczny	0 - brak, 1 - obecny
11	kołatanie serca	dychotomiczny	0 - brak, 1 - obecny
12	duszności	dychotomiczny	0 - brak, 1 - obecny
13	zawroty głowy/omdlenia	dychotomiczny	0 - brak, 1 - obecny
14	odbijanie	dychotomiczny	0 - brak, 1 - obecny
	Czynniki paliatywne		
15	czynniki paliatywne	kategoryczny	tabela 7

L.p.	Cecha	Charakter	Wartości
	Historia podobnego bólu		
16	wcześniejszy, tego samego rodzaju w klatce piersiowej	dychotomiczny	0 - brak, 1 - obecny
17	konsultacja lekarska przy wcześniejszym bólu	dychotomiczny	0 - brak, 1 - obecny
18	wcześniejszy, powiązany z sercem	dychotomiczny	0 - brak, 1 - obecny
19	wcześniejszy, spowodowany zawałem	dychotomiczny	0 - brak, 1 - obecny
20	wcześniejszy, spowodowany chorobą niedokrwienną serca	dychotomiczny	0 - brak, 1 - obecny
	Historia medyczna		
21	wcześniejszy zawał serca	dychotomiczny	0 - brak, 1 - obecny
22	wcześniejsza choroba niedokrwienna serca	dychotomiczny	0 - brak, 1 - obecny
23	wcześniejszy nietypowy ból w klatce piersiowej	dychotomiczny	0 - brak, 1 - obecny
24	niewydolność serca	dychotomiczny	0 - brak, 1 - obecny
25	choroba naczyń obwodowych	dychotomiczny	0 - brak, 1 - obecny
26	przepuklina rozwory przełykowego	dychotomiczny	0 - brak, 1 - obecny
27	nadciśnienie tętnicze	dychotomiczny	0 - brak, 1 - obecny
28	cukrzyca	dychotomiczny	0 - brak, 1 - obecny
29	palacz	dychotomiczny	0 - brak, 1 - obecny
	Obecne użycie leków		
30	diuretyki	dychotomiczny	0 - brak, 1 - obecny
31	azotany	dychotomiczny	0 - brak, 1 - obecny
32	beta-blokery	dychotomiczny	0 - brak, 1 - obecny
33	digoksyna	dychotomiczny	0 - brak, 1 - obecny
34	niesteroidowe leki przeciwzapalne	dychotomiczny	0 - brak, 1 - obecny
35	leki zobojętniające kwas żołądkowy, blokery H2	dychotomiczny	0 - brak, 1 - obecny

L.p.	Cecha	Charakter	Wartości
	Badanie fizyczne		
36	skurczowe ciśnienie tętnicze	dyskretny	liczby naturalne
37	rozkurczowe ciśnienie tętnicze	dyskretny	liczby naturalne
38	tętno	dyskretny	liczby naturalne
39	szybkość oddychania	dyskretny	liczby naturalne
40	rzężenia	dychotomiczny	0 - brak, 1 - obecny
41	sinica	dychotomiczny	0 - brak, 1 - obecny
42	bladość	dychotomiczny	0 - brak, 1 - obecny
43	szmery skurczowe	dychotomiczny	0 - brak, 1 - obecny
44	szmery rozkurczowe	dychotomiczny	0 - brak, 1 - obecny
45	obrzęk	dychotomiczny	0 - brak, 1 - obecny
46	trzeci ton serca	dychotomiczny	0 - brak, 1 - obecny
47	czwarty ton serca	dychotomiczny	0 - brak, 1 - obecny
48	tkliwość ściany klatki piersiowej	dychotomiczny	0 - brak, 1 - obecny
49	potliwość	0 - brak, 1 - obecny	
	Badanie EKG		
50	nowy załamek Q	dychotomiczny	0 - brak, 1 - obecny
51	jakikolwiek załamek Q	dychotomiczny	0 - brak, 1 - obecny
52	nowe uniesienie odcinka ST	dychotomiczny	0 - brak, 1 - obecny
53	jakiekolwiek uniesienie odcinka ST	dychotomiczny	0 - brak, 1 - obecny
54	nowe obniżenie odcinka ST	dychotomiczny	0 - brak, 1 - obecny
55	jakiekolwiek obniżenie odcinka ST	dychotomiczny	0 - brak, 1 - obecny
56	nowy odwrócony załamek T	dychotomiczny	0 - brak, 1 - obecny
57	jakikolwiek odwrócony załamek T	dychotomiczny	0 - brak, 1 - obecny
58	nowe zaburzenie przewodnictwa śródkomorowego	dychotomiczny	0 - brak, 1 - obecny
59	jakiekolwiek zaburzenie przewodnictwa śródkomorowego	dychotomiczny	0 - brak, 1 - obecny

Tabela 2. Opis wartości cechy *miejsce bólu*.

Wartość	Znaczenie
1	zamostkowy
2	lewa strona, w okol. serca
3	prawa strona na wys. serca
4	lewy bok klatki piersiowej
5	prawy bok klatki piersiowej
6	brzuch
7	plecy
8	inne

Tabela 3. Opis wartości cechy *promieniowanie bólu w klatce piersiowej*.

Wartość	Znaczenie
1	szyja
2	szczęka
3	lewe ramię
4	lewa ręka
5	prawe ramię
6	plecy
7	brzuch
8	inne

Tabela 4. Opis wartości cechy *charakter bólu*.

Wartość	Znaczenie
1	szyja
2	szczęka
3	lewe ramię
4	lewa ręka
5	prawe ramię
6	plecy
7	brzuch
8	inne

Tabela 5. Opis wartości cechy *początek występowania bólu*.

Wartość	Znaczenie
1	podczas wysiłku
2	w spoczynku
3	podczas snu

Tabela 6. Opis wartości cechy *długość trwania ostatniego bólu*.

Wartość	Znaczenie
1	poniżej 5 min
2	5 - 30 min
3	30 - 60 min
4	1 - 6 godz.
5	6 - 12 godz.
6	powyżej 12 godz.

Tabela 7. Opis wartości cechy *czynniki paliatywne*.

Wartość	Znaczenie
1	brak
2	nitrogliceryna w ciągu 5 min
3	nitrogliceryna po upływie 5 min
4	leki zobojętniające kwas żołądkowy
5	znieczulenie poza morfiną
6	morfina

1.3 Algorytm selekcji cech według rankingu

Za pomocą rankingu cech można wyróżnić najważniejsze cechy ze zbioru wszystkich cech. Jest to konieczne do przeprowadzenia procesu klasyfikacji. Aby wyznaczyć wspomniany ranking wykorzystano metodę ANOVA[1] - analizę wariancji. ANOVA stanowi popularną oraz często stosowaną analizę statystyczną służącą do porównania wpływu zmiennej na wartość analizowanej funkcji. Można podzielić analizę wariancji na 3 grupy:

- jednoczynnikowa - wpływ każdego czynnika analizowany jest oddzielnie,
- wieloczynnikowa - wpływy czynników analizowane są razem,
- analiza wariancji dla czynników wewnątrzgrupowych.

Zdarza się także, że łączy się różne rodzaje: międzygrupową (jedno lub wieloczynnikową) z wewnątrzgrupową, co nazywa się mianem analizy wariancji w schemacie mieszanym. Ideę analizy wariancji stanowi sprawdzenie, czy pewne zmienne niezależne wpływają na poziom zmiennej zależnej (testowanej).

W celu porównania wariancji dla analizowanych cech wykorzystano jednoczynnikową analizę wariancji. Porównanie wszystkich cech wraz z wartością statystyki F dla metody ANOVA widoczne jest na wykresie na rysunku 1. Działania matematyczne niezbędne do obliczenia statystyki F przedstawia równanie 1.

$$
\begin{aligned}
F &= \frac{MSTR}{MSE} \\
MSTR &= \frac{1}{r-1}\sum_{i=1}^{r} n_i(\overline{x_i} - \overline{x})^2 \\
MSE &= \frac{1}{n-r}\sum_{i=1}^{r}\sum_{j=1}^{n_i} n_i(x_{ij} - \overline{x_i})^2
\end{aligned}
\tag{1}
$$

gdzie:

n_i – liczba pomiarów i-klasy

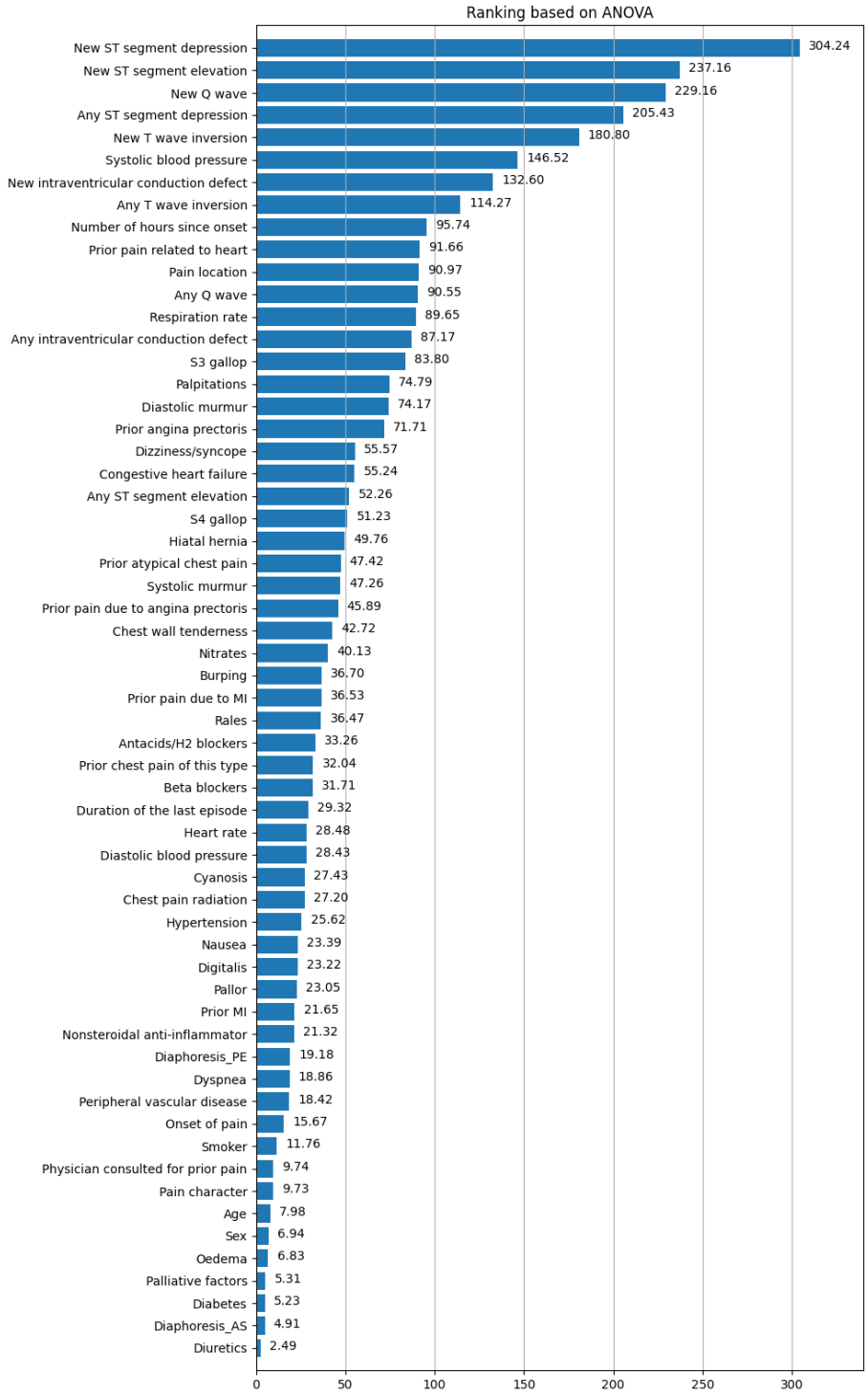
$\overline{x_i}$ – średnia arytmetyczna wartości pomiarów i-klasy

$\overline{x}$ – średnia arytmetyczna wartości pomiarów wszystkich klas

x_{ij} – wartości pomiaru j klasy i

r – liczba klas

W dalszej analizie brano pod uwagę k cech, dla których metoda analizy wariancji zwróciła najwyższy wynik, gdzie parametr k był analizowany pod kątem liczby cech dających najlepsze rezultaty.



Rysunek 1. Ranking cech na podstawie metody ANOVA

Literatura

1. Wahid, Z.: Application of one-way anova in completely randomized experiments. Journal of Physics: Conference Series (949) (2017)