Politechnika Wrocławska
Wydział Elektroniki
Rok akad. 2015/2016
Kierunek Informatyka

Komputerowe wspomaganie diagnozowania choroby niedokrwiennej u dzieci z wykorzystaniem algorytmów minimalno-odległościowych

Zastosowanie informatyki w medycynie: Projekt

Grupa projektowa:
Barłomiej Grzegorek, XXXXXX
Marcin Mantke, 200633

Prowadzący:
prof. dr hab. inż. Marek Kurzyński

Wrocław, 06.06.2016

Spis treści

Spis rysunków

Spis listingów

Rozdział 1

Charakterystyka analizowanego problemu

Rozdział 2

Opis stosowanych algorytmów

2.1 Metryka odległości

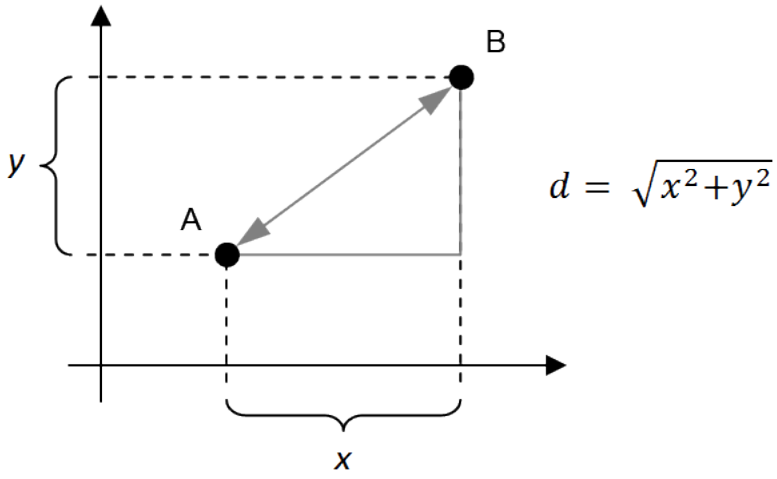
W algorytmach odległościowych istotną rolę odgrywa wybrana metryka wg której mierzone będzie odległość pomiędzy dwoma badanymi punktami. Spośród wielu wybrane zostały metryka euklidesowa i metryka Manhattan.

Metryka euklidesowa

Metryka euklidesowa, w której za odległość między dwoma punktami w przestrzeni przyjmuje się pierwiastek euklidesowego iloczynu skalarnego różnicy dwóch wektorów:

$$d_e(x, y) = \sqrt{(y - x, x - y)}$$

Przykład:



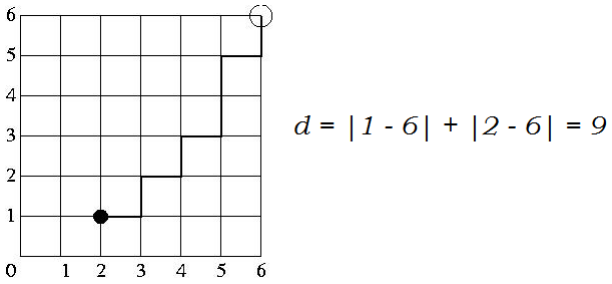
Rysunek 2.1: Przykład obliczania metryki eukliesowej.

Metryka Manhattan

Metryka Manhattan, w której odległość dwóch punktów to suma wartości bezwzględnych różnic ich współrzędnych, zgodnie z poniższym wzorem:

$$d_m(x, y) = \sum_{k=1}^{n} |x_k - y_k|$$

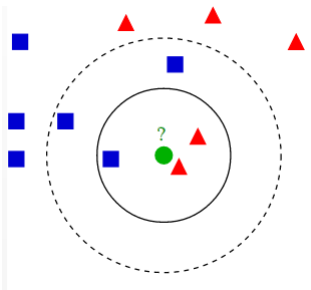
Przykład:



Rysunek 2.2: Przykład obliczania metryki manhattan.

2.2 Algorytm nearest neighbour

Algorytm k-NN pozwala wyszukać w zadanym zbiorze punktów, k punktów znajdujących się najbliżej nowego punktu, korzystając przy tym z wybranej miary odległości.
W zagadnieniu klasyfikacji, algorytm k-NN przyjmuje jako argument tzw. zbiór uczący składający się z wielowymiarowych wektorów cech oraz przypisanych do nich klas. Faza klasyfikacji nowego obiektu polega na przypisaniu odpowiedniej klasy nieznanemu wektorowi cech. Nowy obiekt przypisywany jest do klasy, która występuje najczęściej wśród k najbliżej znajdujących się obiektów ze zbioru uczącego, zgodnie z wybraną metryką. Specjalnym przypadkiem algorytmu jest sytuacja, w której $k = 1$. Nazywa się algorytmem najbliższego sąsiada, w którym klasa nowego obiektu ustalana jest na podstawie najbliżej leżącej próbki ze zbioru danych uczących. Na poniższym rysunku przestawiona została przykładowa przestrzeń zawierająca dane uczące (niebieskie kwadraty oraz czerwona trójkąty) oraz niezidentyfikowaną próbkę – zielone kółko):



Rysunek 2.3: Klasyfikacja w algorytmie k-nn.

Przyjmując liczbę najbliższych sąsiadów $k = 3$ (ciągła linia na rysunku), nowy obiekt zostanie przypisany do grupy czerwonych trójkątów, ponieważ jest ich więcej wśród 3 najbliższych próbek.

Przyjmując $k = 5$, nowy obiekt zostanie przypisany do klasy niebieskich kwadratów, ponieważ w najbliższym otoczeniu znajdują się 3 kwadraty i tylko 2 trójkąty.

2.3 Algorytm nearest mean

Algorytm nearest mean jest jednym z algorytmów minimalnoodległościowych. Jest on wykorzystywany w statystyce to prognozowania wartości pewnej zmiennej losowej bądź do zadania klasyfikacji.

Założenia:

- dany jest zbiór uczący, który zawiera obserwacje. Każda z obserwacji ma przypisany wektor zmiennych objaśniających $X_1 ... X_n$ oraz wartość zmiennej objaśnianej Y,
- dana jest obserwacja C z przypisanym wektorem zmiennch objaśniających $X_1 ... X_n$ dla której chcemy prognozować wartość zmiennej objaśnianej Y.

Przebieg algorytmu:

1. korzystając ze zbioru uczącego oblicz centroid dla każdej wartości zmiennej objaśnianej Y,
2. zaklasyfikuj obserwację C do jednej z klas zmiennej objaśnianej Y poprzez minimalizację odległości pomiędzy wektorem wybranych cech, a centroidem.

Rozdział 3

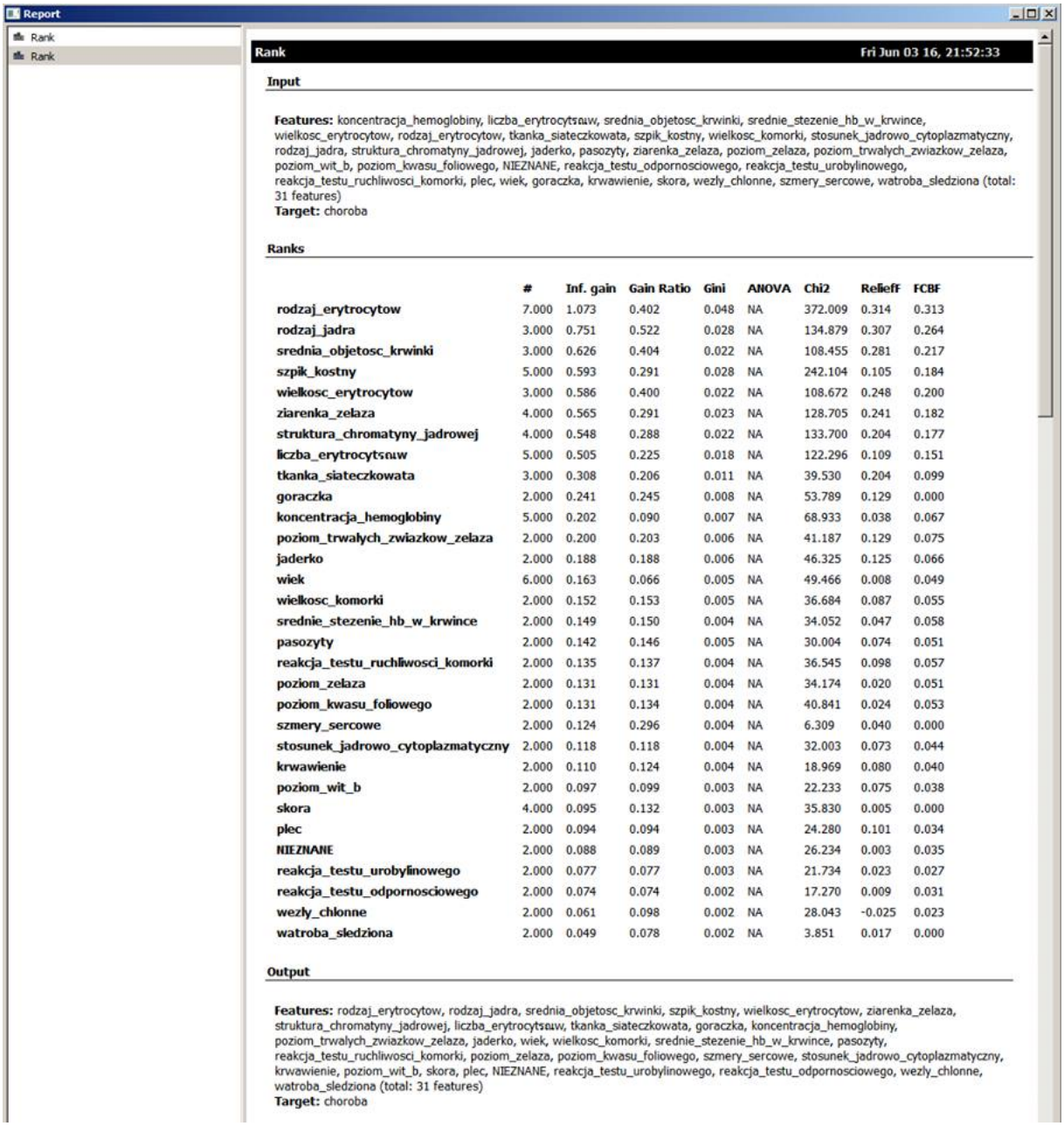
Implementacja

3.1 Informacja o środowisku implementacyjnym

Algorytmy zostały zaimplementowane w środowisku **MATLAB**. Ranking cech został uzyskany przy pomocy środowiska **Orange** (`http://orange.biolab.si/`). Jest to rozwiązanie open source, które umożliwia wizualizację oraz analizę danych.

3.2 Ranking cech

Korzystając z narzędzia **Orange** wygenerowaliśmy ranking cech przedstawiony na rysunku ??.



	#	Inf. gain	Gain Ratio	Gini	ANOVA	Chi2	ReliefF	FCBF
rodzaj_erytrocytow	7.000	1.073	0.402	0.048	NA	372.009	0.314	0.313
rodzaj_jadra	3.000	0.751	0.522	0.028	NA	134.879	0.307	0.264
srednia_objetosc_krwinki	3.000	0.626	0.404	0.022	NA	108.455	0.281	0.217
szpik_kostny	5.000	0.593	0.291	0.028	NA	242.104	0.105	0.184
wielkosc_erytrocytow	3.000	0.586	0.400	0.022	NA	108.672	0.248	0.200
ziarenka_zelaza	4.000	0.565	0.291	0.023	NA	128.705	0.241	0.182
struktura_chromatyny_jadrowej	4.000	0.548	0.288	0.022	NA	133.700	0.204	0.177
liczba_erytrocytsuw	5.000	0.505	0.225	0.018	NA	122.296	0.109	0.151
tkanka_siateczkowata	3.000	0.308	0.206	0.011	NA	39.530	0.204	0.099
goraczka	2.000	0.241	0.245	0.008	NA	53.789	0.129	0.000
koncentracja_hemoglobiny	5.000	0.202	0.090	0.007	NA	68.933	0.038	0.067
poziom_trwalych_zwiazkow_zelaza	2.000	0.200	0.203	0.006	NA	41.187	0.129	0.075
jaderko	2.000	0.188	0.188	0.006	NA	46.325	0.125	0.066
wiek	6.000	0.163	0.066	0.005	NA	49.466	0.008	0.049
wielkosc_komorki	2.000	0.152	0.153	0.005	NA	36.684	0.087	0.055
srednie_stezenie_hb_w_krwince	2.000	0.149	0.150	0.004	NA	34.052	0.047	0.058
pasozyty	2.000	0.142	0.146	0.005	NA	30.004	0.074	0.051
reakcja_testu_ruchliwosci_komorki	2.000	0.135	0.137	0.004	NA	36.545	0.098	0.057
poziom_zelaza	2.000	0.131	0.131	0.004	NA	34.174	0.020	0.051
poziom_kwasu_foliowego	2.000	0.131	0.134	0.004	NA	40.841	0.024	0.053
szmery_sercowe	2.000	0.124	0.296	0.004	NA	6.309	0.040	0.000
stosunek_jadrowo_cytoplazmatyczny	2.000	0.118	0.118	0.004	NA	32.003	0.073	0.044
krwawienie	2.000	0.110	0.124	0.004	NA	18.969	0.080	0.040
poziom_wit_b	2.000	0.097	0.099	0.003	NA	22.233	0.075	0.038
skora	4.000	0.095	0.132	0.003	NA	35.830	0.005	0.000
plec	2.000	0.094	0.094	0.003	NA	24.280	0.101	0.034
NIEZNANE	2.000	0.088	0.089	0.003	NA	26.234	0.003	0.035
reakcja_testu_urobylinowego	2.000	0.077	0.077	0.003	NA	21.734	0.023	0.027
reakcja_testu_odpornosciowego	2.000	0.074	0.074	0.002	NA	17.270	0.009	0.031
wezly_chlonne	2.000	0.061	0.098	0.002	NA	28.043	-0.025	0.023
watroba_sledziona	2.000	0.049	0.078	0.002	NA	3.851	0.017	0.000

Rysunek 3.1: Ranking cech.

3.3 Implementacja algorytmów

Nearest mean

Na poniższych listingach przedstawione są najważniejsze części programów.

Listing 3.1: Podział danych na zbiór uczący i testowy.

```
data = data(randperm(end), :);
train = data(1:floor(0.5*size(data, 1)), :);
test = data(floor(0.5*size(data, 1))+1:end, :);
```

Listing 3.2: Etap treningu - wyznaczanie centroidów.

```
centroid = [unique(data(:, 1)) zeros(size(unique(data(:, 1)), 1), size(data, 2)-1)
    ];

for label = unique(train(:, 1))'
    % zbierz wszystkie pr bki danej klasy
    train(train(:, 1) == label, 2:end)
    % oblicz centroid dla danej klasy
    centroid(centroid(:, 1) == label, 2:end) = mean(train(train(:, 1) == label, 2:
        end));
end
```

Listing 3.3: Wyznaczenie odległości próbki od centroida i przydzielenie do klasy.

```
pre_result = zeros(size(test, 1), 1);
for i = 1:size(test, 1)
    dist = pdist2(test(i, feature_rank(1:k)), centroid(:, feature_rank(1:k)));
    [~, templabel] = min(dist);
    pre_result(i) = centroid(templabel, 1);
end
```

Rozdział 4

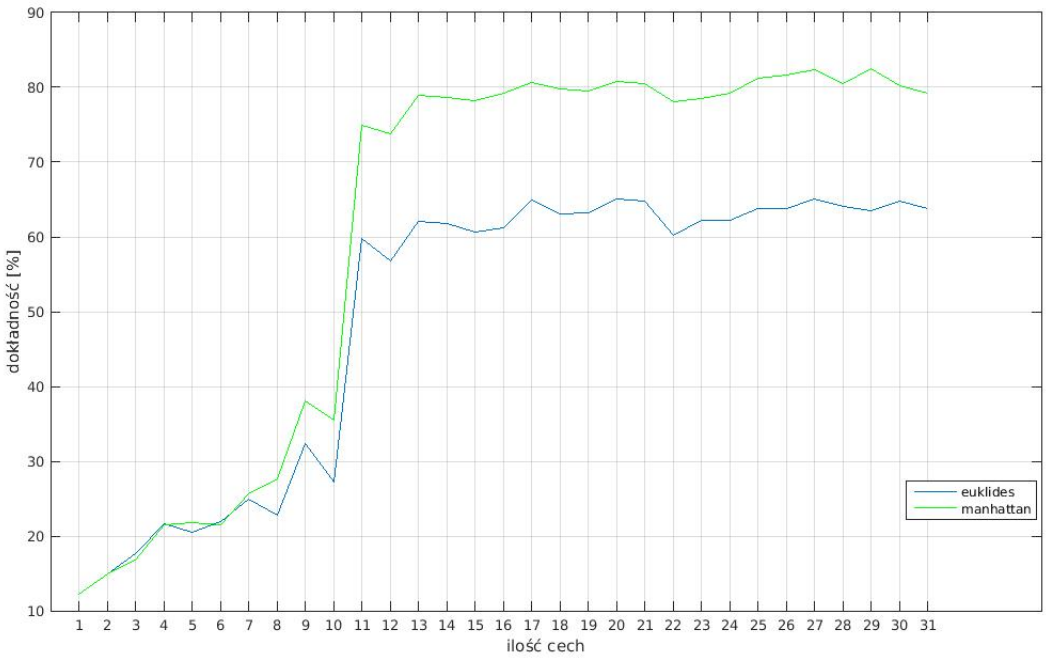
Opis badań eksperymentalnych

Badania eksperymentalne zostały przeprowadzone zgodnie z instrukcją. Zastosowano trenowanie i testowanie klasyfikatorów z wykorzystaniem 5 razy powtarzanej metody 2-krotnej walidacji krzyżowej. Zastosowane miary odległości to miara euklidesowa i Manhattan.

4.1 Wyniki badań

4.1.1 Algorytm nearest mean

Wyniki badań dla algorytmu nearest mean zostały przedstawione na rysunku **??**.



Rysunek 4.1: Wynik badań dla algorytmu nearest mean.

Jak widać na rysunku, dokładność klasyfikacji rośnie wraz ze zwiększaniem liczby cech biorącej udział w klasyfikacji. Największy skok dokładności występuje przy wykorzystaniu 11 cech (różnica

na poziomie 40%). Kolejne zwiększanie ilości cech utrzymuje dokładność klasyfikacji na względnie równym poziomie (nie ma dużych wzrostów ani spadków dokładności klasyfikacji).
Z rysunku można również odczytać, że wpływ na dokładność klasyfikacji ma wybrana metryka. Od momentu dołożenia 11 cechy, czyli największego wzrostu dokładności klasyfikacji, klasyfikacja dla metryki euklidesowej jest średnio 25-30% gorsza od metryki Manhattan.

Rozdział 5

Podsumowanie i wnioski

5.1 Wnioski płynące z analizy wyników

5.2 Ocena krytyczna i podsumowanie projektu

Bibliografia

[1] M.M. Sysło, N.Deo, J.S. Kowalik, *Algorytmy optymalizacji dyskretnej z programami w języku Pascal*, Wydawnictwo Naukowe PWN , 1999

[2] R. Neapolitan, K. Naimipour, *Podstawy Algorytmów z przykładami w C++*, Helion, 2004

[3] T. H. Cormen, C. E. Leiserson, R. L. Rivest, C. Stein, *Wprowadzenie do algorytmów*, Wydawnictwo Naukowe PWN, 2013