k najbliższych sąsiadów

Indukcyjne metody analizy danych

Maksym Telepchuk

Maj 2020

1 Opis algorytmu

Klasyfikator k-NN tzw. k-najbliższych sąsiadów należy do grupy algorytmów opartych o analizę przypadku. Algorytm prezentuje swoją wiedzę o świecie w postaci zbioru przypadków lub doświadczeń. Idea klasyfikacji polega na metodach wyszukiwania tych zgromadzonych przypadków, które mogą być zastosowane do klasyfikacji nowych danych. Klasyfikacja nowych przypadków zgodnie z algorytmem k-NN jest realizowana na bieżąco, tzn. wtedy gdy pojawia się potrzeba klasyfikacji nowego przypadku. Algorytm k-NN nie buduje klasyfikatora.

Klasyfikacja nowego przypadku x jest realizowana w następujący sposób: Poszukujemy k najbliższych punktów w przestrzeni wzorców. Przypadek x jest klasyfikowany jako należący do klasy w sposób głosowania wyliczonych k najbliższych sąsiadów. „Głos" - klasa sąsiada. Waga głosu - na ile głos ma wpływ na końcowy wynik głosowania. „Głosują" tylko i wyłącznie k najbliższych sąsiadów.

2 Porównywane metody

Tutaj będą porównane następujące sposoby głosowania:

1. Większościowe. Zwracana klasa jest wynikiem głosowania większości głosów.
2. Ważone odległością. Waga głosu sąsiada jest odwrotnie proporcjonalna odległości od niego.
3. Ważone kwadratem odległości. Waga głosu sąsiada jest odwrotnie proporcjonalna kwadratu odległości od niego.

Najbliższych sąsiadów wybierano na podstawie miary odległości. W tym przypadku będą porównane trzy miary odległości:

1. Euclidean : $d(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$
2. Manhattan : $d(p,q) = \sum_{i=1}^{n}|q_i - p_i|$
3. Czebyszewa : $d(p,q) = \max_i |q_i - p_i|$

W tym badaniu również zostanie porównane jak algorytm radzi sobie z nieujednoliconymi danymi oraz z normalizowanymi i standaryzowanymi. Normalizacja oznacza przeskalowanie danych do przedziału (tutaj do przedziału $[0,1]$), a standaryzacja oznacza przeskalowanie danych tak, że $\mu = 1, \sigma = 1$.

Wydajność algorytmu będzie sprawdzona za pomocą kroswalidacji zwykłej oraz kroswalidacji stratyfikowanej.

2.1 Dane

Dla analizy zostały wybrane następujące zbiory danych :

- glass - 6 typów szkła, skład z 9 chemicznych elementów
- wine - 3 rodzaje wina, skład z 13 chemicznych elementów
- seeds - 3 typy nasienia, 7 pomiarów nasienia

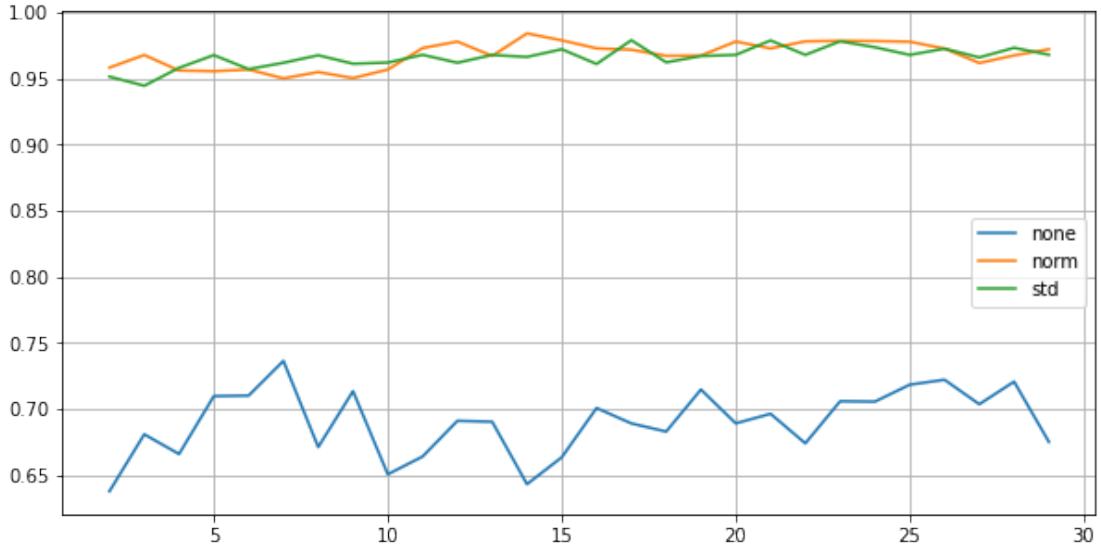
3 Wyniki badań

W tabelach oraz wykresach będą przestawione średnie wartości *f1score* przy 100 przeprowadzonych kroswalidacji stratyfikowanej dla 5 foldów. Za każdym razem zbiór jest mieszany.

3.1 Rola skalowania danych

Bardzo ważnym jest standaryzować lub normalizować dane, żeby wszystkie atrybuty miały taki sam wpływ na odległość pomiędzy próbkami.

Na zbiorze danych **wine** przeprowadzono skalowanie danych. Z rys. 1 i tab.1 widać, że standaryzacja i normalizacja polepszają wyniki klasyfikacji



Rysunek 1: Wpływ skalowania danych

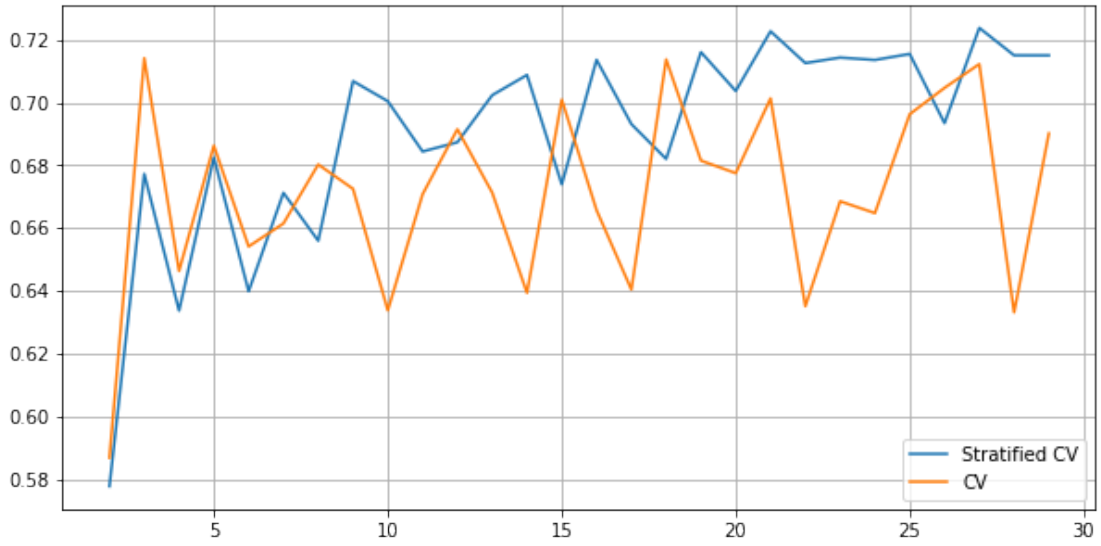
none	norm	std
0.61	0.96	0.94
0.68	0.96	0.96
0.64	0.96	0.95
0.69	0.96	0.97
0.66	0.97	0.96
0.65	0.96	0.98
0.67	0.96	0.97
0.7	0.96	0.97
0.68	0.96	0.98
0.67	0.97	0.95
0.68	0.97	0.96
0.69	0.97	0.97
0.66	0.97	0.96
0.69	0.98	0.96
0.74	0.97	0.96
0.68	0.98	0.97
0.71	0.98	0.97
0.69	0.98	0.96
0.71	0.98	0.97
0.68	0.97	0.98
0.71	0.98	0.97
0.72	0.97	0.96
0.66	0.97	0.97
0.67	0.97	0.97
0.7	0.97	0.98
0.69	0.97	0.97
0.69	0.97	0.97
0.73	0.97	0.97

Tablica 1: Wpływ skalowania danych

3.2 Kroswalidacja

Do ewaluacji modelu została użyta kroswalidacja stratyfikowana zamiast zwykłej. Jest potrzebnym zachowywać balans klas pomiędzy foldami, żeby klasyfikator nauczył się dobrze rozpoznawać wszystkie klasy.

Przebadano kroswalidację zwykłą oraz stratyfikowaną na zbiorze zdebalansowanym **wine**. Wyniki są na rys.2 i w tabeli 2



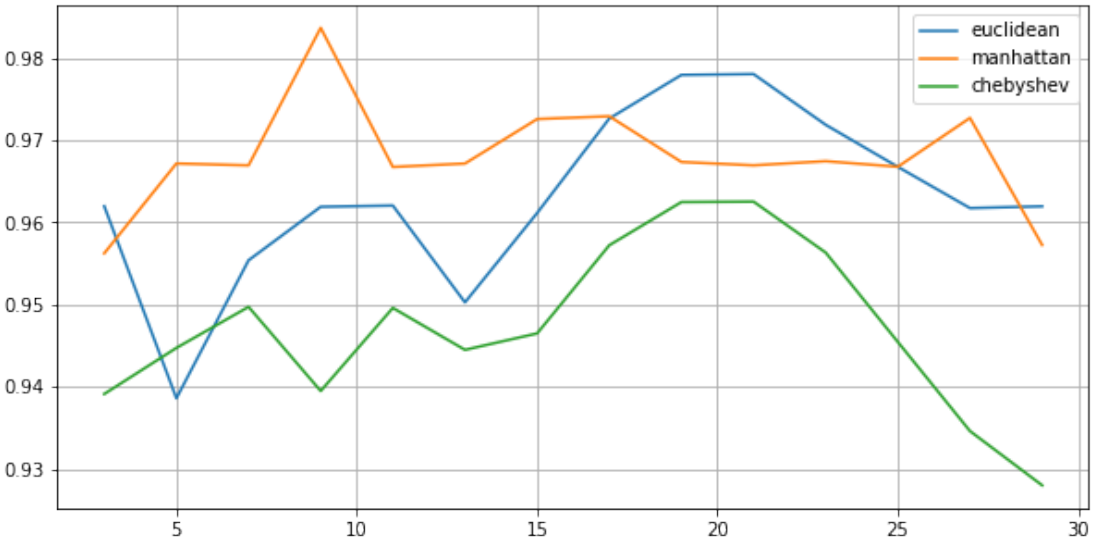
Rysunek 2: Wpływ typu kroswalidacji

stratified	norm
0.58	0.59
0.68	0.71
0.63	0.65
0.68	0.69
0.64	0.65
0.67	0.66
0.66	0.68
0.71	0.67
0.7	0.63
0.68	0.67
0.69	0.69
0.7	0.67
0.71	0.64
0.67	0.7
0.71	0.67
0.69	0.64
0.68	0.71
0.72	0.68
0.7	0.68
0.72	0.7
0.71	0.64
0.71	0.67
0.71	0.66
0.72	0.7
0.69	0.7
0.72	0.71
0.71	0.63
0.71	0.69

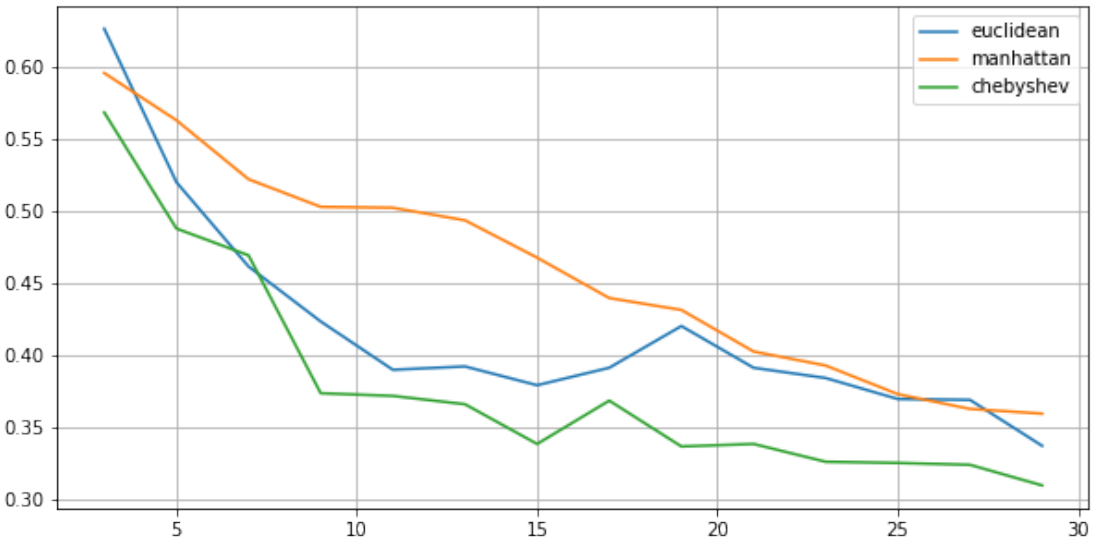
Tablica 2: Wpływ typu kroswalidacji

3.3 Metryki

Dla każdego ze zbioru została ustalona metoda głosowania równoważona, każdy zbiór został znormalizowany oraz przebadany pod kątem f1-score dla zakresu $k = [3, 30]$ oraz k - nieparzyste. Tutaj został wprowadzony warunek nieparzystości, bo przy takim głosowaniu może nastąpić 'remis'.



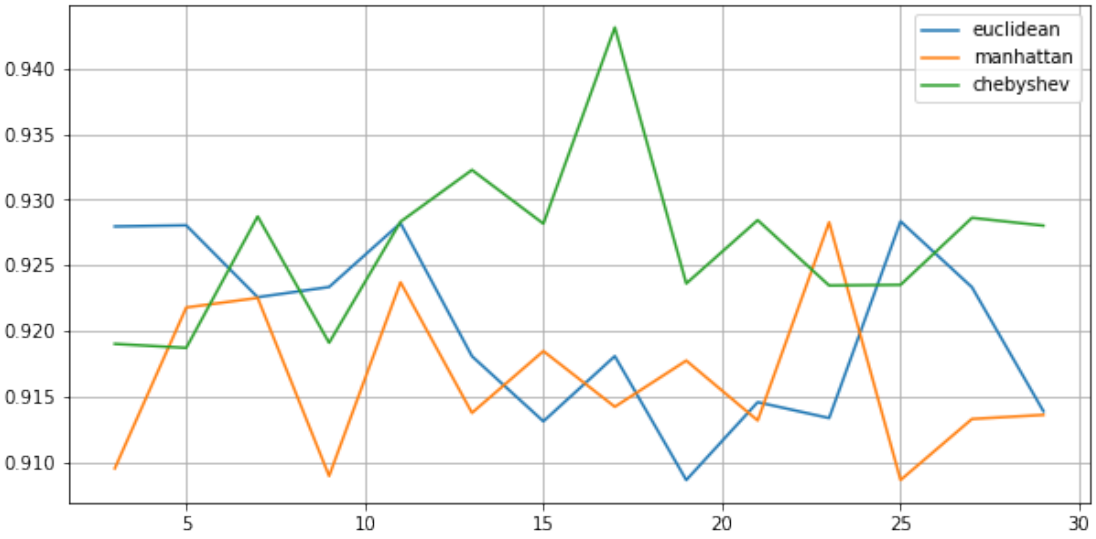
Rysunek 3: Porównanie metryk dla zbioru wine



Rysunek 4: Porównanie metryk dla zbioru glass

k	euclidean	manhattan	chebyshev
3.0	0.96	0.96	0.94
5.0	0.94	0.97	0.94
7.0	0.96	0.97	0.95
9.0	0.96	0.98	0.94
11.0	0.96	0.97	0.95
13.0	0.95	0.97	0.94
15.0	0.96	0.97	0.95
17.0	0.97	0.97	0.96
19.0	0.98	0.97	0.96
21.0	0.98	0.97	0.96
23.0	0.97	0.97	0.96
25.0	0.97	0.97	0.95
27.0	0.96	0.97	0.93
29.0	0.96	0.96	0.93

Tablica 3: Porównanie metryk dla zbioru wine



Rysunek 5: Porównanie metryk dla zbioru seeds

k	euclidean	manhattan	chebyshev
3.0	0.63	0.6	0.57
5.0	0.52	0.56	0.49
7.0	0.46	0.52	0.47
9.0	0.42	0.5	0.37
11.0	0.39	0.5	0.37
13.0	0.39	0.49	0.37
15.0	0.38	0.47	0.34
17.0	0.39	0.44	0.37
19.0	0.42	0.43	0.34
21.0	0.39	0.4	0.34
23.0	0.38	0.39	0.33
25.0	0.37	0.37	0.33
27.0	0.37	0.36	0.32
29.0	0.34	0.36	0.31

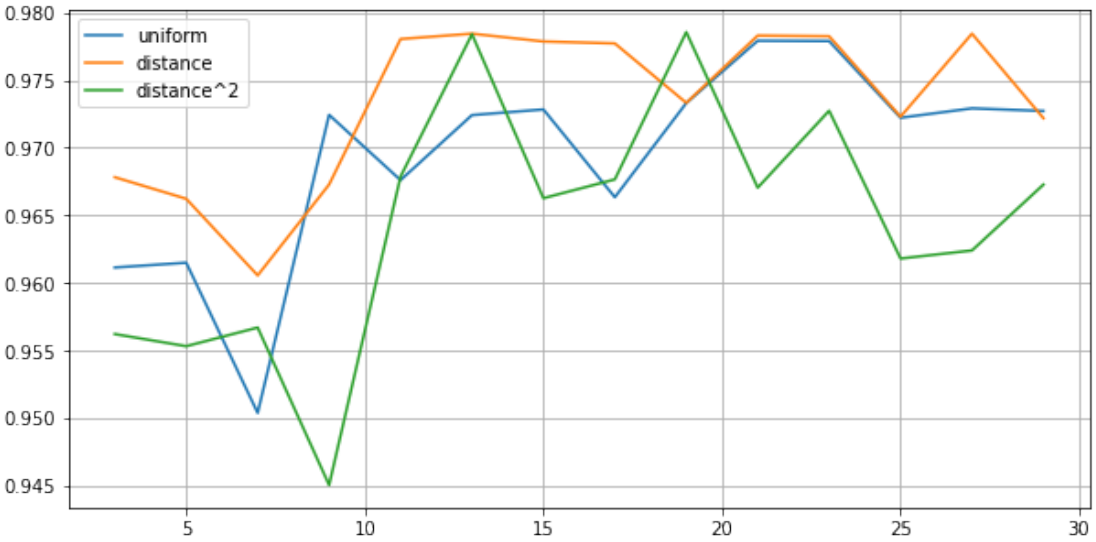
Tablica 4: Porównanie metryk dla zbioru glass

k	euclidean	manhattan	chebyshev
3.0	0.93	0.91	0.92
5.0	0.93	0.92	0.92
7.0	0.92	0.92	0.93
9.0	0.92	0.91	0.92
11.0	0.93	0.92	0.93
13.0	0.92	0.91	0.93
15.0	0.91	0.92	0.93
17.0	0.92	0.91	0.94
19.0	0.91	0.92	0.92
21.0	0.91	0.91	0.93
23.0	0.91	0.93	0.92
25.0	0.93	0.91	0.92
27.0	0.92	0.91	0.93
29.0	0.91	0.91	0.93

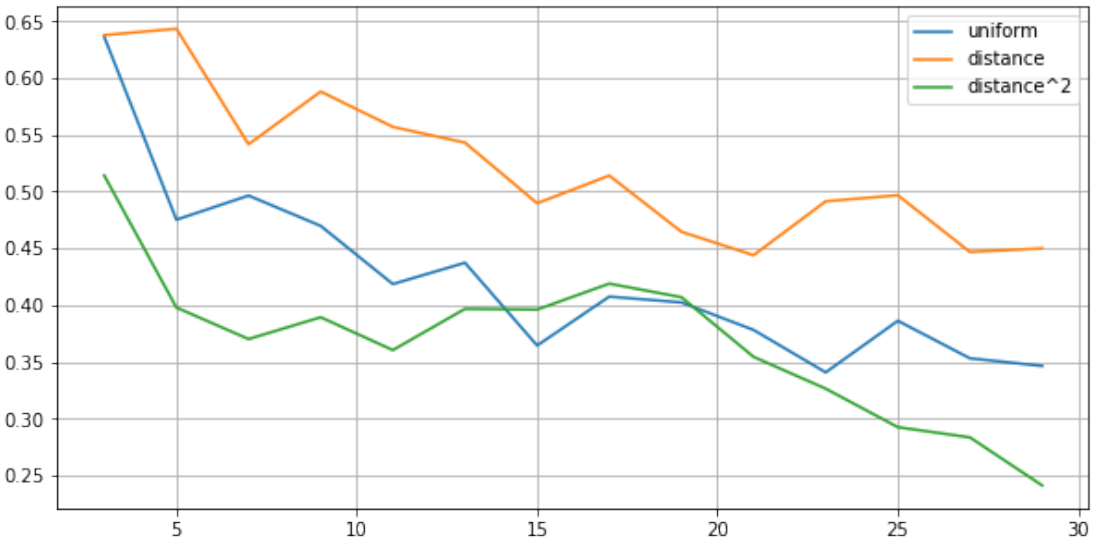
Tablica 5: Porównanie metryk dla zbioru seeds

3.4 Metody głosowania

Dla każdego ze zbioru została ustalona metryka euklidesowa, każdy zbiór został znormalizowany oraz przebadany pod kątem f1-score dla zakresu $k = [3, 30]$ oraz k - nieparzyste. Tutaj został wprowadzony warunek nieparzystości, bo przy takim głosowaniu może nastąpić 'remis'.



Rysunek 6: Porównanie metod głosowania dla zbioru wine



Rysunek 7: Porównanie metod głosowania dla zbioru glass

k	uniform	distance	distance2
3.0	0.96	0.97	0.96
5.0	0.96	0.97	0.96
7.0	0.95	0.96	0.96
9.0	0.97	0.97	0.95
11.0	0.97	0.98	0.97
13.0	0.97	0.98	0.98
15.0	0.97	0.98	0.97
17.0	0.97	0.98	0.97
19.0	0.97	0.97	0.98
21.0	0.98	0.98	0.97
23.0	0.98	0.98	0.97
25.0	0.97	0.97	0.96
27.0	0.97	0.98	0.96
29.0	0.97	0.97	0.97

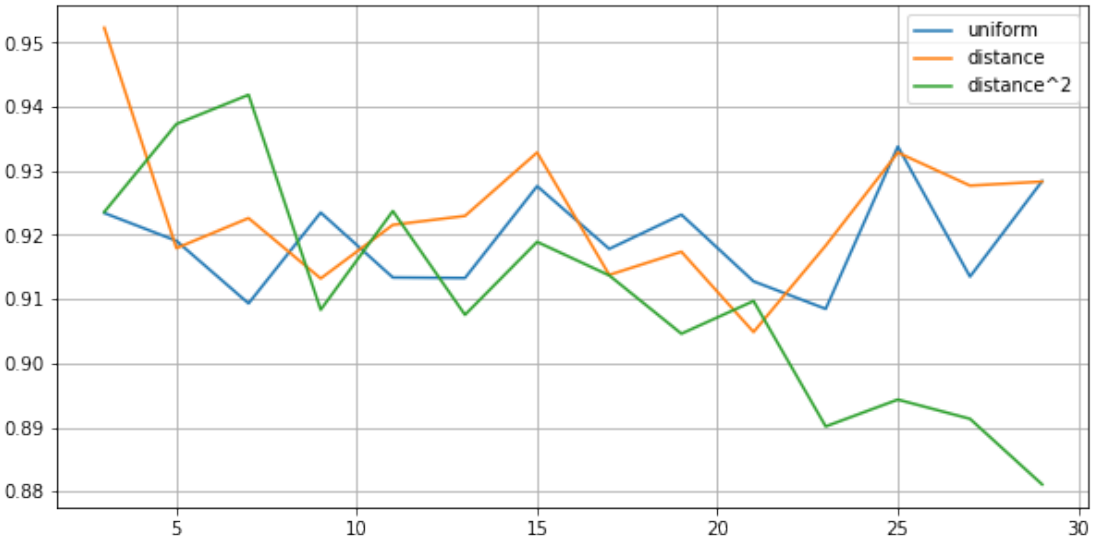
Tablica 6: Porównanie metod głosowania dla zbioru wine

k	uniform	distance	distance2
3.0	0.64	0.64	0.51
5.0	0.48	0.64	0.4
7.0	0.5	0.54	0.37
9.0	0.47	0.59	0.39
11.0	0.42	0.56	0.36
13.0	0.44	0.54	0.4
15.0	0.36	0.49	0.4
17.0	0.41	0.51	0.42
19.0	0.4	0.46	0.41
21.0	0.38	0.44	0.35
23.0	0.34	0.49	0.33
25.0	0.39	0.5	0.29
27.0	0.35	0.45	0.28
29.0	0.35	0.45	0.24

Tablica 7: Porównanie metod głosowania dla zbioru glass

k	uniform	distance	distance2
3.0	0.92	0.95	0.92
5.0	0.92	0.92	0.94
7.0	0.91	0.92	0.94
9.0	0.92	0.91	0.91
11.0	0.91	0.92	0.92
13.0	0.91	0.92	0.91
15.0	0.93	0.93	0.92
17.0	0.92	0.91	0.91
19.0	0.92	0.92	0.9
21.0	0.91	0.9	0.91
23.0	0.91	0.92	0.89
25.0	0.93	0.93	0.89
27.0	0.91	0.93	0.89
29.0	0.93	0.93	0.88

Tablica 8: Porównanie metod głosowania dla zbioru seeds



Rysunek 8: Porównanie metod głosowania dla zbioru seeds

	Naive Bayes	C5.0	k-NN
wine	0.98	0.93	0.98
glass	0.44	0.73	0.69
seeds	0.90	0.86	0.94

Tablica 9: F1Score dla różnych algorytmów

4 Wnioski

Warto wykonywać skalowanie danych (normalizacja lub standaryzacja) korzystają z algorytmu k-NN. Warto również używać kroswalidację stratyfikowaną zamiast zwykłej.

Dla zbioru wine najlepiej się sprawdziła metryka manhattan, oraz metoda głosowania distance.

Dla zbioru glass dobór matryki manhattan i wag distance również dają najlepsze wyniki z badanych metod. Klasyfikacja zbioru glass jest szczególnie skuteczna przy małych k.

Dla zbioru seeds najlepszą metryką też jest metryka Chebyszewa wydaje się sensowna. Metoda głosowania distance2 dała bardzo dobre wyniki dla małych k, ale dla większych wartości metoda distance ją przebiła.

Porównanie z algorytmem Naiwnego Bayesa są podane w tabeli 9. W tym przypadku dla k-NN są podane wyniki dla najlepszego k oraz połączenia najlepszych metod głosowania i metryk.

Można wywnioskować, że wybór stosowanej metryki i sposobu głosowania zależy od zbioru danych. Jak widać, dla wszystkich zbiorów danych, algorytm k-NN nie jest gorszym od naiwnego podejścia oraz w dwóch zbiorach jest lepszym od algorytmu C5.0.

Literatura

[1] http://wazniak.mimuw.edu.pl/index.php?title=ED-4.2-m09-1.0-toc.

[2] http://www.cs.ucc.ie/~dgb/courses/tai/notes/handout4.pdf.