POLITECHNIKA WROCŁAWSKA
WYDZIAŁ INFORMATYKI I ZARZĄDZANIA

KIERUNEK: INFORMATYKA

Systemy uczące się
Laboratorium

Algorytm k-najbliższych sąsiadów

AUTOR:
Bartosz Kardas

PROWADZĄCY:
dr inż. Paweł Myszkowski

OCENA PRACY:

WROCŁAW, 2017

Spis treści

1. Opis algorytmu k najbliższych sąsiadów 4
1.1. Klasyfikacja 4
1.2. Głosowanie 4
1.3. Metryki 5

2. Dane testowe - charakterystyka 6
2.1. Seeds 6
2.1.1. Opis 6
2.1.2. Dystrybucja klas 6
2.1.3. Atrybuty 6
2.2. Ecoli 7
2.2.1. Opis 7
2.2.2. Dystrybucja klas 7
2.2.3. Atrybuty 7
2.3. User Knwoledge Modelling 8
2.3.1. Klasy 8
2.3.2. Dystrybucja klas 8
2.3.3. Atrybuty 8

3. Wyniki eksperymentów 9
3.1. Porównanie normalizacji danych 9
3.2. Porównanie ilości najbliższych sąsiadów 10
3.3. Porównanie metod głosowania 12
3.4. Porównanie metryk 12
3.5. Porównanie z innymi klasyfikatorami 13

4. Wnioski 15
4.1. Normalizacja danych 15
4.2. Ilość najbliższych sąsiadów 15
4.3. Głosowanie 15
4.4. Metryki 15
4.5. Porównanie klasyfikatorów 16

Spis rysunków

3.1. Wyniki klasyfikatora dla różnych metod preprocessingu. 9
3.2. Wyniki klasyfikatora dla różnej ilości najbliższych sąsiadów 11
3.3. Wyniki klasyfikatora dla różnych metod głosowania. 12
3.4. Wyniki klasyfikatora dla różnych metryk. 13
3.5. Analiza porównawcza klasyfikatorów. 14

Spis tabel

2.1. Dystrybucja klas w zbiorze seeds. 6
2.2. Opis atrybutów w zbiorze seeds. 6
2.3. Dystrybucja klas w zbiorze ecoli. 7
2.4. Opis atrybutów w zbiorze seeds. 7
2.5. Dystrybucja klas w zbiorze seeds. 8
2.6. Opis atrybutów w zbiorze ukm. 8

3.1. Tabela normalizacja. 9
3.2. Tabela ilość najbliższych sąsiadów. 10
3.3. Tabela metody głosowania. 12
3.4. Tabela metryk. 13
3.5. Tabela porównawcza. 14

Rozdział 1

Opis algorytmu k najbliższych sąsiadów

1.1. Klasyfikacja

Klasyfikator k-najbliższych sąsiadów jest jednym z najprostszych klasyfikatorów. Jego działanie opiera się na zasadzie wyszukiwania wśród danych uczących najbliższych elementów dla elementu poddawanego predykcji. Innymi słowy dla elementu, którego klasę należy wyznaczyć, oblicza się odległość w danej metryce od pozostałych. Następnie wybieranych jest k najbliższych sąsiadów i dokonuje się głosowania na daną klasę.

1.2. Głosowanie

W niniejszej pracy zaimplementowano kilka różnych metod głosowania. Poniżej krótko przedstawiono zasadę każdej z nich.

- prosta - inaczej większościowa, wybrana zostaje klasa o największej ilości głosów.
- ważona - oddane głosy posiadają wagę względem odległości. (Bliżsi sąsiedzi mają większą wagę, bardziej znaczący głos.)
- rankingowa - każdy z sąsiadów oddaje swoje punkty na daną klasę. Punkty przydzielane są według rankingu (1 punkt - najdalszy sąsiad, k punktów - najbliższy sąsiad)
- rankingowo-ważona - połącznie metody ważonej i rankingowej

Warto zauważyć, że w każdej z metod istnieje szansa na remis, w tym przypadku wybierana jest klasa o większej liczności bądź losowa.

1.3. Metryki

Metryka jest to funkcja która określa odległość pomiędzy dwoma elementami w danej przestrzeni. Dotychczas opracowano wiele różnych metryk, do najpopularniejszych należą:

- metryka manhattan
- metryka euklidesowa
- metryka chebysheva
- metryka minkowskiego

Metryka Minkowskiego jest uogólnieniem pozostałych metryk, poniżej przedstawiono jej wzór.

$$minkowski(p, x, y) = (\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}}$$

Warto zauważyć, że dla $p = 1$ metryka minkowskiego jest metryką manhattan, dla $p = 2$ metryką euklidesową, a przy $p \to \infty$ jest metryką chebysheva.

Rozdział 2

Dane testowe - charakterystyka

2.1. Seeds

2.1.1. Opis

Zbiór zawiera dane dotyczące charakterystki ziaren. Zawiera 3 klasy ziaren: Kama, Rosa and Canadian, po 70 rekordów każdy. Każde z ziaren opisane jest przez 7 atrybutów rzeczywistych. Łącznie 210 rekordów. Każde z ziaren jest opisane, nie posiada brakujących wartości. Zbiór może być wykorzystywany do metod klasyfikacji oraz klasteryzacji.

2.1.2. Dystrybucja klas

Liczności poszczególnych klas zostały przedstawione w tabeli 2.1

Tab. 2.1: Dystrybucja klas w zbiorze seeds.

NAZWA	LICZNOŚĆ
Kama	70
Rosa	70
Canadian	70

2.1.3. Atrybuty

Poniżej 2.2 przedstawiono opis poszczególnych atrybutów.

Tab. 2.2: Opis atrybutów w zbiorze seeds.

NAZWA	RODZAJ	OPIS
Wielkość	REAL	area A
Perymetr	REAL	perimeter P
Kompaktowość	REAL	compactness C = $4 * pi * A/P^2$
Długość	REAL	length of kernel
Szerokość	REAL	width of kernel
Współczynnik asymetrii	REAL	asymmetry coefficient
Długość rowka	REAL	length of kernel groove

2.2. Ecoli

2.2.1. Opis

Zbiór zawiera dane dotyczące miejsca lokalizacji białek. Zawiera 8 klas lokalizacji przedstawione w tabeli 2.3. Każda z lokalizacji opisana jest przez 5 atrybutów rzeczywistych, 2 binarne, 1 kategoryczny, z czego ostatni jest nazwą konkretnej sekwencji (unikalna dla każdego rekordu). Łącznie 336 rekordów. Każdy z nich jest w pełni opisany, nie posiada brakujących wartości. Zbiór może być wykorzystywany do metod klasyfikacji.

2.2.2. Dystrybucja klas

Liczności poszczególnych klas zostały przedstawione w tabeli 2.3

Tab. 2.3: Dystrybucja klas w zbiorze ecoli.

NAZWA	LICZNOŚĆ	OPIS
cp	143	cytoplasm
im	77	inner membrane without signal sequence
pp	52	perisplasm
imU	35	inner membrane, uncleavable signal sequence
om	20	outer membrane
omL	5	outer membrane lipoprotein
imL	2	inner membrane lipoprotein
imS	2	inner membrane, cleavable signal sequence

2.2.3. Atrybuty

Poniżej 2.4 przedstawiono opis poszczególnych atrybutów.

Tab. 2.4: Opis atrybutów w zbiorze seeds.

NAZWA	RODZAJ
name	UNIQUE
mcg	REAL
gvh	REAL
lip	BINARY
chg	BINARY
aac	REAL
alm2	REAL
alm1	REAL

2.3. User Knwoledge Modelling

2.3.1. Klasy

Zbiór zawiera dane dotyczące koorelacji między wynikami testów badanych osób, a czasem spędzonym na naukę. Zawiera 4 klasy oznaczające wynik egzaminu, przedstawione w tabeli 2.5. Każda z lokalizacji opisana jest przez 5 atrybutów rzeczywistych. Łącznie 403 rekordów. Każdy z nich jest w pełni opisany, nie posiada brakujących wartości. Zbiór może być wykorzystywany do metod klasyfikacji oraz klasteryzacji.

2.3.2. Dystrybucja klas

Liczności poszczególnych klas zostały przedstawione w tabeli 2.5

Tab. 2.5: Dystrybucja klas w zbiorze seeds.

NAZWA	LICZNOŚĆ
very-low	50
low	129
middle	122
high	130

2.3.3. Atrybuty

Poniżej 2.6 przedstawiono opis poszczególnych atrybutów.

Tab. 2.6: Opis atrybutów w zbiorze ukm.

NAZWA	RODZAJ	OPIS
STG	REAL	Stopień poświęcenia czasu uczenia na główny cel
SCG	REAL	Stopień powtarzania informacji o głównym celu
STR	REAL	Stopień poświęcenia czasu uczenia na elementy powiązane z głównym celem
LPR	REAL	Wynik egzaminu powiązanego z głównym celem
PEG	REAL	Wynik egzaminu z głównym celem

Rozdział 3

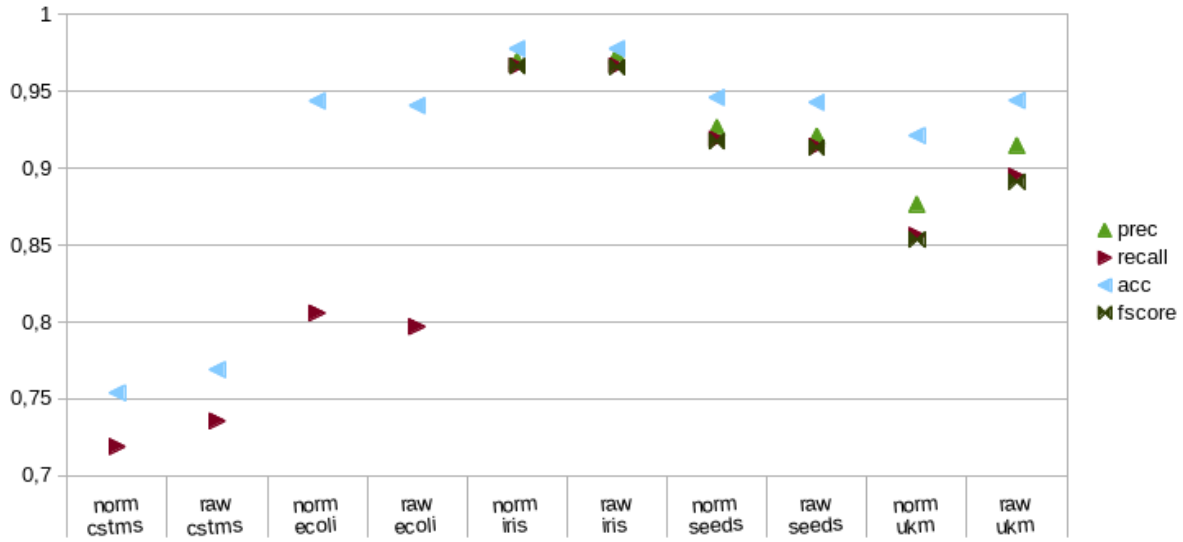
Wyniki eksperymentów

3.1. Porównanie normalizacji danych

Poniżej przedstawiono wyniki klasyfikatorów dla danych znormalizowanych i nieznormalizowanych.

Tab. 3.1: Tabela normalizacja.

name	normalized	prec	recall	acc	fscore
cstms	norm	nan	0,719	0,754	nan
cstms	raw	nan	0,736	0,769	nan
ecoli	norm	nan	0,806	0,944	nan
ecoli	raw	nan	0,797	0,941	nan
iris	norm	0,970	0,967	0,978	0,966
iris	raw	0,974	0,967	0,978	0,966
seeds	norm	0,927	0,919	0,946	0,918
seeds	raw	0,921	0,914	0,943	0,914
ukm	norm	0,877	0,856	0,921	0,854
ukm	raw	0,915	0,895	0,944	0,891

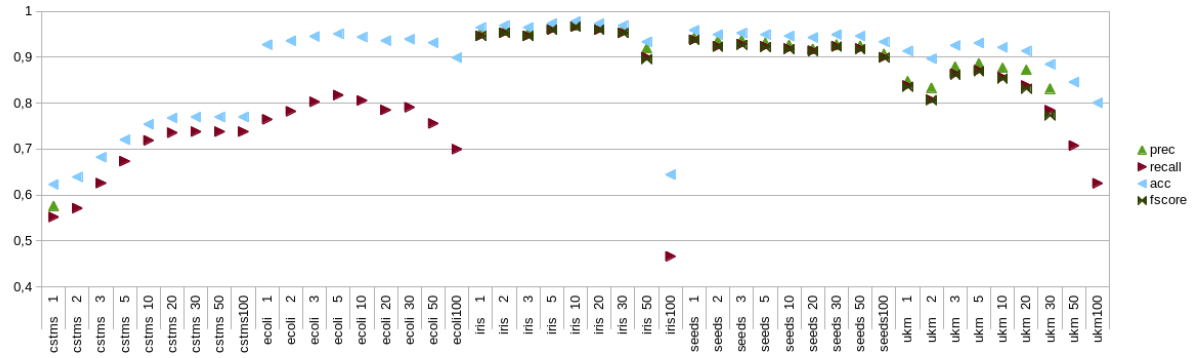


Rys. 3.1: Wyniki klasyfikatora dla różnych metod preprocessingu.

3.2. Porównanie ilości najbliższych sąsiadów

Tab. 3.2: Tabela ilość najbliższych sąsiadów.

cstms	1	0,576	0,552	0,623	nan
cstms	2	nan	0,571	0,639	nan
cstms	3	nan	0,626	0,682	nan
cstms	5	nan	0,674	0,720	nan
cstms	10	nan	0,719	0,754	nan
cstms	20	nan	0,736	0,768	nan
cstms	30	nan	0,738	0,770	nan
cstms	50	nan	0,738	0,770	nan
cstms	100	nan	0,738	0,770	nan
ecoli	1	nan	0,765	0,927	nan
ecoli	2	nan	0,782	0,936	nan
ecoli	3	nan	0,803	0,945	nan
ecoli	5	nan	0,818	0,951	nan
ecoli	10	nan	0,806	0,944	nan
ecoli	20	nan	0,785	0,936	nan
ecoli	30	nan	0,791	0,939	nan
ecoli	50	nan	0,756	0,931	nan
ecoli	100	nan	0,700	0,899	nan
iris	1	0,953	0,947	0,964	0,946
iris	2	0,959	0,953	0,969	0,953
iris	3	0,953	0,947	0,964	0,946
iris	5	0,964	0,960	0,973	0,960
iris	10	0,970	0,967	0,978	0,966
iris	20	0,964	0,960	0,973	0,960
iris	30	0,960	0,953	0,969	0,953
iris	50	0,919	0,900	0,933	0,896
iris	100	nan	0,467	0,644	nan
seeds	1	0,942	0,938	0,959	0,938
seeds	2	0,936	0,924	0,949	0,923
seeds	3	0,937	0,929	0,952	0,927
seeds	5	0,931	0,924	0,949	0,922
seeds	10	0,927	0,919	0,946	0,918
seeds	20	0,918	0,914	0,943	0,913
seeds	30	0,928	0,924	0,949	0,923
seeds	50	0,925	0,919	0,946	0,917
seeds	100	0,908	0,900	0,933	0,899
ukm	1	0,848	0,838	0,914	0,836
ukm	2	0,833	0,808	0,897	0,807
ukm	3	0,879	0,864	0,926	0,862
ukm	5	0,887	0,872	0,931	0,870
ukm	10	0,877	0,856	0,921	0,854
ukm	20	0,872	0,838	0,914	0,831
ukm	30	0,832	0,785	0,885	0,773
ukm	50	nan	0,708	0,846	nan
ukm	100	nan	0,626	0,801	nan



Rys. 3.2: Wyniki klasyfikatora dla różnej ilości najbliższych sąsiadów

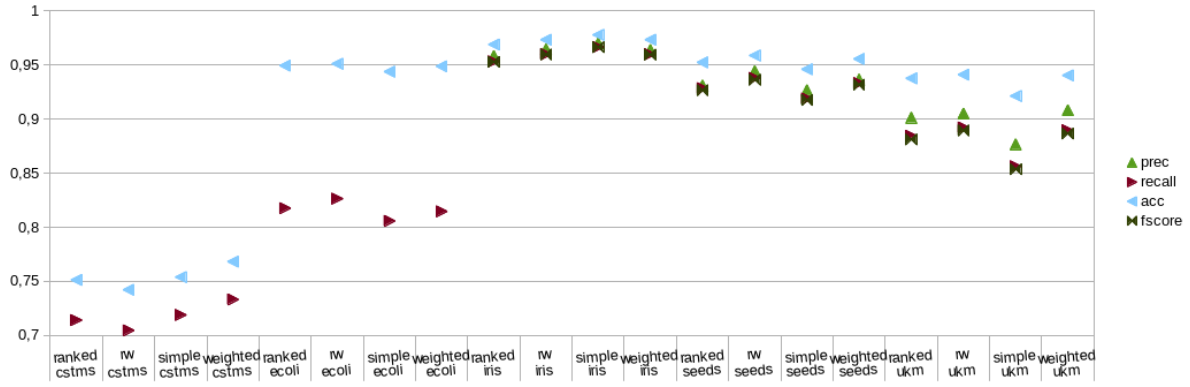
Powyżej przedstawiono wyniki klasyfikatorów dla różnej ilości najbliższych sąsiadów - 1, 2, 3, 5, 10, 20, 30, 50 oraz 100.

3.3. Porównanie metod głosowania

Poniżej przedstawiono wyniki klasyfikatorów dla różnych metod głosowania - proste, ważone, rankingowe, rankingowo-ważone.

Tab. 3.3: Tabela metody głosowania.

name	normalized	prec	recall	acc	fscore
cstms	ranked	nan	0,714	0,751	nan
cstms	rw	nan	0,705	0,742	nan
cstms	simple	nan	0,719	0,754	nan
cstms	weighted	nan	0,733	0,768	nan
ecoli	ranked	nan	0,818	0,949	nan
ecoli	rw	nan	0,826	0,951	nan
ecoli	simple	nan	0,806	0,944	nan
ecoli	weighted	nan	0,815	0,949	nan
iris	ranked	0,958	0,953	0,969	0,953
iris	rw	0,964	0,960	0,973	0,960
iris	simple	0,970	0,967	0,978	0,966
iris	weighted	0,964	0,960	0,973	0,960
seeds	ranked	0,931	0,929	0,952	0,927
seeds	rw	0,944	0,938	0,959	0,936
seeds	simple	0,927	0,919	0,946	0,918
seeds	weighted	0,937	0,933	0,956	0,932
ukm	ranked	0,901	0,885	0,938	0,881
ukm	rw	0,905	0,892	0,941	0,890
ukm	simple	0,877	0,856	0,921	0,854
ukm	weighted	0,908	0,890	0,940	0,887



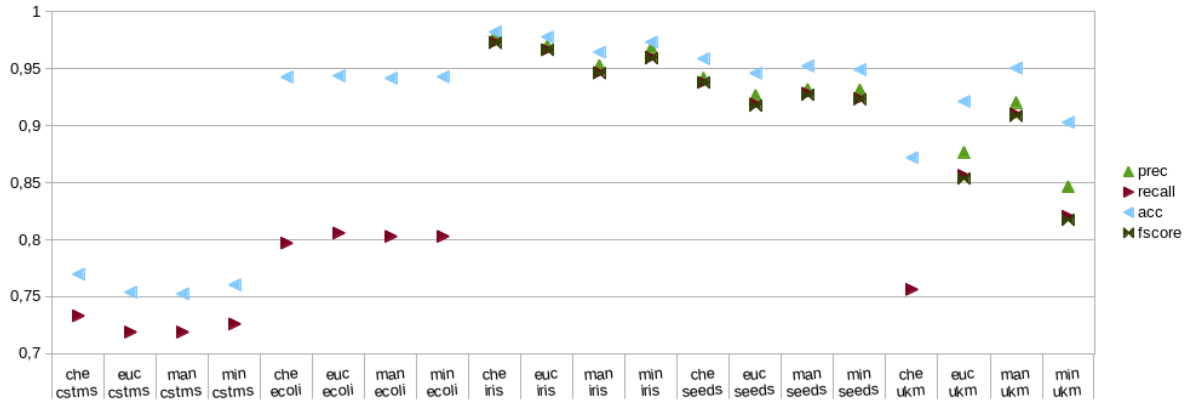
Rys. 3.3: Wyniki klasyfikatora dla różnych metod głosowania.

3.4. Porównanie metryk

Poniżej przedstawiono wyniki klasyfikatorów dla różnych metryk - manhattan, euklidesowej, chebysheva i minkowskiego o parametrze $p = 3$.

Tab. 3.4: Tabela metryk.

name	normalized	prec	recall	acc	fscore
cstms	che	nan	0,733	0,770	nan
cstms	euc	nan	0,719	0,754	nan
cstms	man	nan	0,719	0,753	nan
cstms	min	nan	0,726	0,760	nan
ecoli	che	nan	0,797	0,943	nan
ecoli	euc	nan	0,806	0,944	nan
ecoli	man	nan	0,803	0,942	nan
ecoli	min	nan	0,803	0,943	nan
iris	che	0,979	0,973	0,982	0,973
iris	euc	0,970	0,967	0,978	0,966
iris	man	0,953	0,947	0,964	0,946
iris	min	0,968	0,960	0,973	0,959
seeds	che	0,942	0,938	0,959	0,938
seeds	euc	0,927	0,919	0,946	0,918
seeds	man	0,932	0,929	0,952	0,927
seeds	min	0,931	0,924	0,949	0,923
ukm	che	nan	0,756	0,872	nan
ukm	euc	0,877	0,856	0,921	0,854
ukm	man	0,920	0,910	0,950	0,909
ukm	min	0,847	0,821	0,903	0,817



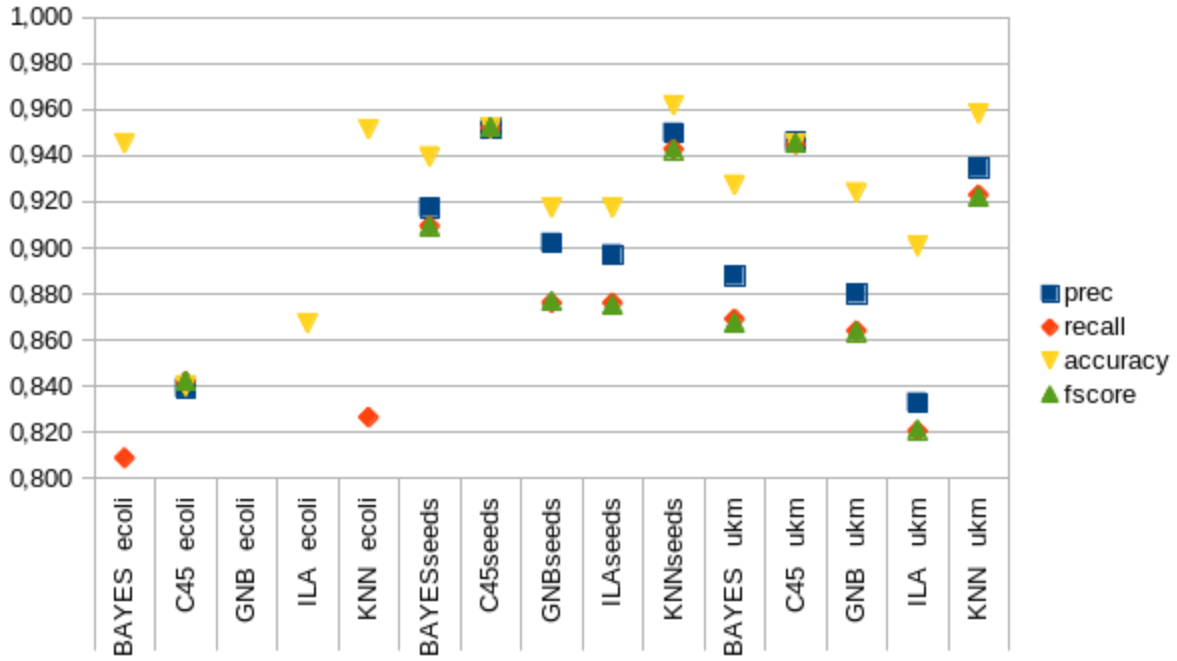
Rys. 3.4: Wyniki klasyfikatora dla różnych metryk.

3.5. Porównanie z innymi klasyfikatorami

Poniżej przedstawiono wyniki różnych klasyfikatorów. BAYES - standardowa implementacja naiwnego klasyfikatora bayesowskiego. GNB - naiwny klasyfikator bayesowski opierający się o wyliczanie prawdopodobieństwa z rozkładu normalnego. ILA - klasyfikator oparty na regułach, algorytm C45 - drzewo decyzyjne oraz KNN - k najbliższych sąsiadów. Dodatkowo wybrano jedynie najlepsze klasyfikatory spośród wszystkich wyników.

Tab. 3.5: Tabela porównawcza.

classifier	name	prec	recall	accuracy	fscore
BAYES	ecoli	nan	0,809	0,945	nan
C45	ecoli	0,839	0,842	0,840	0,842
GNB	ecoli	nan	0,412	0,680	nan
ILA	ecoli	nan	0,662	0,867	nan
KNN	ecoli	nan	0,826	0,951	nan
BAYES	seeds	0,917	0,910	0,940	0,909
C45	seeds	0,952	0,952	0,952	0,952
GNB	seeds	0,902	0,876	0,917	0,877
ILA	seeds	0,897	0,876	0,917	0,875
KNN	seeds	0,950	0,943	0,962	0,943
BAYES	ukm	0,888	0,869	0,927	0,867
C45	ukm	0,946	0,945	0,945	0,945
GNB	ukm	0,880	0,864	0,924	0,864
ILA	ukm	0,833	0,821	0,901	0,821
KNN	ukm	0,935	0,923	0,958	0,922



Rys. 3.5: Analiza porównawcza klasyfikatorów.

Rozdział 4

Wnioski

4.1. Normalizacja danych

Normalizacja danych nie miała większego wpływu dla zbioru iris oraz seeds. Dla obu podejść klasyfikatory uzyskały podobne właściwości. Dla zbirów ukm oraz customers dane surowe (bez normalizacji) okazały się lepszym wyborem dla klasyfikatora kNN. W przypadku zbioru ecoli dane znormalizowane uzyskały lepszy wynik niż dane surowe.

4.2. Ilość najbliższych sąsiadów

Dla wszystkich zbiorów można zauważyć wyraźną zależność pomiędzy ilością najbliższych sąsiadów a wynikami klasyfikatorów. Przeważnie ilość najbliższych sąsiadów nie powinna być za mała (1,2) ani za duża (50, 100). Optymalne wartości uzyskano dla około 10 - 20 najbliższych sąsiadów (zbiory customers, ecoli, iris). W przypadku zbioru ukm optymalną wartością było 5 sąsiadów, a dla zbioru seeds 1 najbliższy sąsiad.

4.3. Głosowanie

W przypadku głosowania nie można jednoznacznie stwierdzić, która z metod okazała się ogółem najlepsza. Dla różnych zbiorów, różne metody wykazały lepszą skuteczność. Zbiory ecoli, seeds oraz ukm uzyskały podobną charakterystykę pod tym względem: najsłabszą metodą głosowania okazała się metoda prosta (większościowa). Najlepszą zaś rankingowo-ważona. Dla zbioru iris odwrotnie - najlepszą metodą okazała się metoda większościowa. W przypadku zbioru customers metoda ważona uzyskała najlepsze wyniki.

4.4. Metryki

Podobnie jak dla głsowania nie można jednoznacznie stwierdzić, która z metod okazała się ogółem najlepsza. Dla różnych zbiorów, różne metody wykazały lepszą skuteczność. Różnice między poszczególnymi metrykami nie zmieniają się znacząco dla tych samych zbiorów oprócz zbioru ukm. W tym przypadku odnotowano wyraźne różnice. Najlepszą metryką w tym przypadku okazała się metryka manhattan, a najgorszą metryka chebysheva. Warto jednak zauważyć, że to właśnie metryka chebysheva uzyskała najlepsze wyniki dla zbiorów customers, iris i seeds.

4.5. Porównanie klasyfikatorów

Dla wszystkich zbiorów klasyfikator kNN wykazał się najlepszą skutecznością. Najlepsze z wyników jakie zostały przez niego osiągnięte wynoszą około 0.96, co jest bardzo dobrym wynikiem. Co w porównaniu do szybkości i prostoty implementacji stawia go na czele wśród badanych klasyfikatorów.