

Data oddania: _____

Ocena: _____

Sebastian Kaźmierski 216795

Bartosz Paluszkiewicz 216856

Zadanie 1: Ekstrakcja cech, miary podobieństwa, klasyfikacja

1. Cel

Celem zadania jest stworzenie aplikacji która wykorzystując metodę Knn będzie wykonywała klasyfikację artykułów w dwóch kategoriach (Places, Topics). Aplikacja ma również implementować moduł ekstrakcji cech który będzie obliczał wartości cech klasyfikowanych artykułów. Po stworzeniu aplikacji przy jej wykorzystaniu zostaną przeprowadzone eksperymenty z różnymi wartościami parametrów metody Knn.

2. Wprowadzenie

2.1. Klasyfikator

Zaimplementowany klasyfikator zbioru tekstów wykorzystuje algorytm k najbliższych sąsiadów (k-nn, k nearest neighbours), który jest jednym z bezparametrowych metod klasyfikacji statystycznej.

W trakcie działania nie tworzy wewnętrznej reprezentacji danych uczących, lecz przechowuje wszystkie wzorce uczące i dopiero w momencie pojawienia się wzorca testującego szuka rozwiązania, wyznaczając odległość do wszystkich wzorców.

Bezparametrowość algorytmu objawia się brakiem założeń co do rozkładu podstawowych danych, np. rozkładu jednostajnego.

2.2. Wektor cech

Do ekstrakcji danych wykorzystaliśmy wektor następujących cech:

1. Liczba wszystkich słów kluczowych w pierwszych 10% procentach tekstu

$$\sum_{i=0}^{10} k_i \quad (1)$$

k_i - liczba słów kluczowych w i-tym procencie tekstu

2. Liczba wszystkich słów kluczowych w całym tekście

$$\sum_{i=0}^{100} k_i \quad (2)$$

k_i - liczba słów kluczowych w i-tym procencie tekstu

3. Liczba wszystkich nazw własnych w stosunku do liczby słów w tekście (po stop liście)

$$\frac{P_{STOP}}{W_{STOP}} \quad (3)$$

P_{STOP} - liczba nazw własnych* po stopliście

W_{STOP} - liczba wszystkich słów w tekście po stopliście

4. Średnia długość unikalnych nazw własnych*

$$\frac{\sum_{i=0}^n len(p_i)}{n} \quad (4)$$

len - funkcja długości słowa

p_i - i-ta unikalna nazwa własna*

n - liczba unikalnych nazw własnych*

5. Długość tekstu (liczba wszystkich słów po stop liście)

$$W_{STOP} \quad (5)$$

W_{STOP} - liczba wszystkich słów w tekście po stopliście

6. Liczba słów które występują więcej niż raz w stosunku do długości tekstu (po stop liście i po stemizacji)

$$\frac{M}{W_{STOP \text{ } STEM}} \quad (6)$$

M - liczba słów, które wystąpiły w tekście więcej niż raz

$W_{STOP \text{ } STEM}$ - liczba słów w tekście po stopliście i stemizacji

7. Średnia długość zdań (liczba słów) (po stop liście i po stemizacji)

$$\frac{\sum_{i=0}^S len(s_i)}{n} \quad (7)$$

len - funkcja długości zdania po stopliście i stemizacji

S - liczba wszystkich zdań w tekście po stopliście i stemizacji

8. Liczba unikalnych słów w stosunku do długości tekstu (po stop liście)

$$\frac{U}{W} \quad (8)$$

U - liczba unikalnych słów po stopliście

W - liczba wszystkich słów po stopliście

9. Liczba akapitów w stosunku do długości tekstu (po stop liście)

$$\frac{S}{W_{STOP}} \quad (9)$$

S - liczba akapitów w tekście

W_{STOP} - liczba słów w tekście po stopliście

10. Średnia długość akapitu (liczba słów) (przed stop listą)

$$\frac{\sum_{i=0}^S len(s_i)}{S} \quad (10)$$

len - funkcja długości akapitu

s_i - i-ty akapit

S - liczba wszystkich akapitów

11. Liczba wszystkich słów usuniętych przez stop listę w stosunku do długości tekstu po stop liście

$$\frac{R_{STOP}}{W_{STOP}} \quad (11)$$

R_{STOP} - liczba słów usuniętych przez stoplistę

W_{STOP} - liczba słów w tekście po stopliście

12. Liczba słów kluczowych dla WEST_GERMANY

$$W_{WG} \quad (12)$$

K_{WG} - liczba słów kluczowych dla WEST_GERMANY

13. Liczba słów kluczowych dla USA

$$K_{USA} \quad (13)$$

K_{USA} - liczba słów kluczowych dla USA

14. Liczba słów kluczowych dla FRANCE

$$K_{FRA} \quad (14)$$

K_{FRA} - liczba słów kluczowych dla FRANCE

15. Liczba słów kluczowych dla UK

$$K_{UK} \quad (15)$$

K_{UK} - liczba słów kluczowych dla UK
 16. Liczba słów kluczowych dla CANADA

$$K_{CAN} \quad (16)$$

K_{CAN} - liczba słów kluczowych dla CANADA
 17. Liczba słów kluczowych dla JAPAN

$$K_{JAP} \quad (17)$$

K_{JAP} - liczba słów kluczowych dla JAPAN

* nazwa własna – słowo pisane wielką literą które nie jest na początku zdania lub jest na początku zdania, ale przynajmniej raz występuje w środku

* długość tekstu – liczba wszystkich słów w tekście

2.3. Zastosowane metryki

Do obliczania odległości zastosowaliśmy trzy metryki (2.3.1 - 2.3.3), miarę własną (2.3.6) oraz dwie miary prawdopodobieństwa (2.3.4, 2.3.5). Ze względu na to, że miary podobieństwa służą określaniu "bliskości" wektorów, a nie odległości między nimi, do obliczeń musieliśmy wykorzystać odwrotności uzyskanych wartości.

2.3.1. Metryka Euklidesa

Odległość euklidesową określamy pierwiastek sumy kwadratów różnic wartości cech wektorów A, B i opisujemy wzorem:

$$d(A, B) = \sqrt{\sum_{i=1}^n ((x_{iA} - x_{iB})^2)} \quad (18)$$

2.3.2. Metryka Czebyszewa

Odległością Czebyszewa określamy największy moduł różnic między wartościami cechami wektorów A, B i określamy wzorem:

$$d(A, B) = \max_i |x_{iA} - x_{iB}| \quad (19)$$

2.3.3. Metryka Uliczna

Odległością w metryce ulicznej (odległością taksówkową, uliczną, Manhattan) nazywamy sumę modułów różnic wartości w każdym wymiarze wektorów A, B i określamy wzorem:

$$d(A, B) = \sum_{i=1}^n |x_{iA} - x_{iB}| \quad (20)$$

2.3.4. Minimum - maksimum

Odległość między wektorami A, B określamy jako odwrotność stosunku sumy minimów wartości cech do sumy maksimumów wartości cech, czyli stosunek maksimumów do minimów wartości cech wektorów.

$$d(A, B) = \left(\frac{\sum_{i=1}^n \min(x_{iA}, x_{iB})}{\sum_{i=1}^n \max(x_{iA}, x_{iB})} \right)^{-1} = \frac{\sum_{i=1}^n \max(x_{iA}, x_{iB})}{\sum_{i=1}^n \min(x_{iA}, x_{iB})} \quad (21)$$

2.3.5. Średnia arytmetyczna - minimum

Odległość między wektorami A, B określamy jako odwrotność stosunku sumy minimalnych wartości cech wektorów i średniej arytmetycznej wartości cech wektorów, czyli stosunek średniej arytmetycznej do sumy minimów wartości cech wektorów.

$$d(A, B) = \left(\frac{\sum_{i=1}^n \min(x_{iA}, x_{iB})}{\frac{1}{2} \sum_{i=1}^n (x_{iA} + x_{iB})} \right)^{-1} = \frac{\frac{1}{2} \sum_{i=1}^n (x_{iA} + x_{iB})}{\sum_{i=1}^n \min(x_{iA}, x_{iB})} \quad (22)$$

2.3.6. Miara własna

Odległość obliczoną przy użyciu autorskiej miary określamy jako sumę iloczynów ułożonych rosnąco różnic wartości między cechami wektora i współczynnika zależnego od pozycji tej odległości w posortowanym zbiorze, a także wielkości zbioru.

$$d(A, B) = \sum_{i=0}^{N-1} C_i \left(1 - \frac{i}{2N} \right) \quad (23)$$

$$C(A, B) = \left(\sum_{i=1}^N |x_A - x_{iB}| \right) \uparrow \quad (24)$$

2.4. Miary jakości

2.4.1. Accuracy - dokładność

Dokładnością nazywamy stosunek liczby poprawnych prognoz do wszystkich wykonanych prognoz,

$$accuracy = \frac{\text{correct predictions}}{\text{total predictions}} \quad (25)$$

W celu obliczenia kolejnych miar tworzymy macierz błędu, której uproszczoną formą jest poniższa tabela [2]:

	Positive prediction	Negative prediction
Positive class	True Positive (TP)	False Negative (FN)
Negative class	False Positive (FP)	True Negative (TN)

W powyższej tabeli wyróżniamy cztery rodzaje klasyfikacji:

1. True Positive - prawdziwie pozytywna - wartość, która powinna zostać sklasyfikowana jako *Positive*, została tak sklasyfikowana
2. True Negative - prawdziwie negatywna - wartość, która powinna zostać sklasyfikowana jako *Negative*, została tak sklasyfikowana
3. False Positive - fałszywie pozytywna - wartość, która **nie** powinna zostać sklasyfikowana jako *Positive*, została tak sklasyfikowana
4. False Negative - fałszywie negatywna - wartość, która **nie** powinna zostać sklasyfikowana jako *Negative*, została tak sklasyfikowana

W naszym przypadku musimy zakwalifikować 6, a nie 2 klasy, więc macierz błędu ma w rzeczywistości wymiary 6 x 6.

2.4.2. Precision - precyzja

Precyzja określa stosunek poprawnie zakwalifikowanych obiektów danej klasy do wszystkich prognoz, które wskazują na tę klasę (suma słusznie i niesłusznie zakwalifikowanych obiektów)

$$precision = \frac{\sum_c^C TruePositive_c}{\sum_c^C (TruePositive_c + FalsePositive_c)} \quad (26)$$

2.4.3. Recall - skuteczność

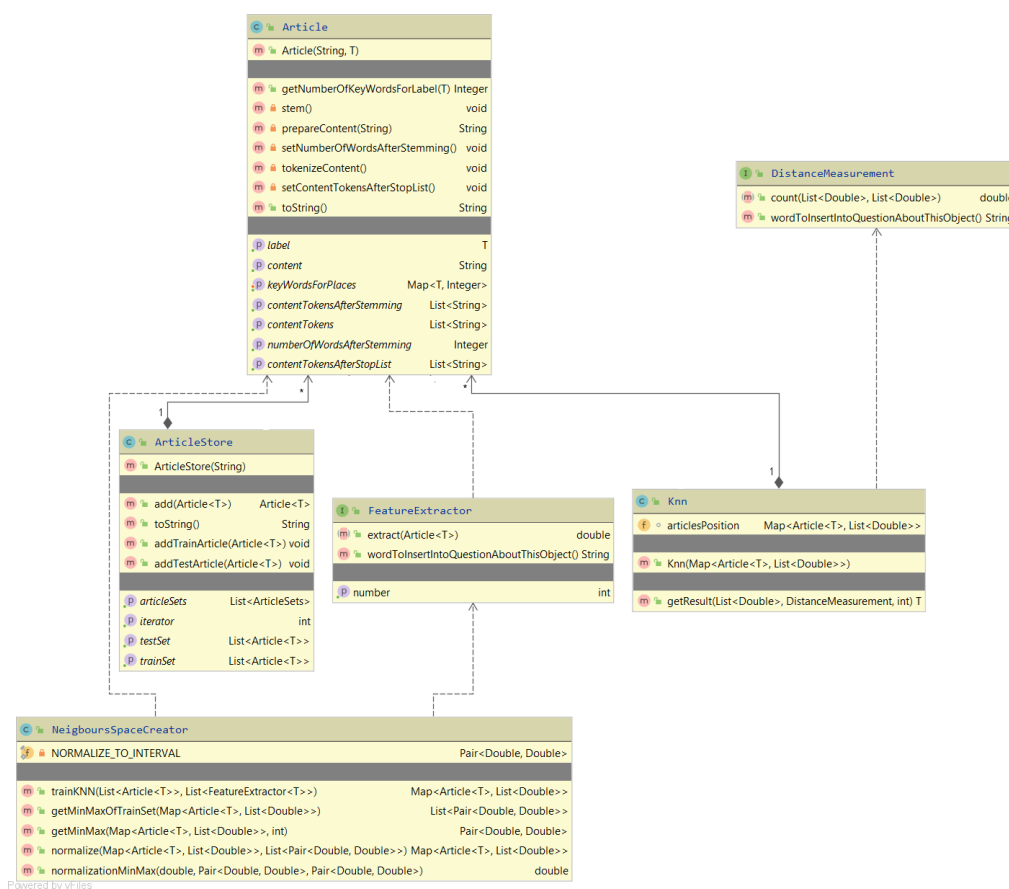
Skuteczność określa stosunek poprawnie zakwalifikowanych obiektów danej klasy do sumy poprawnie zakwalifikowanych i poprawnie odrzuconych obiektów

$$recall = \frac{\sum_c^C TruePositive_c}{\sum_c^C (TruePositive_c + FalseNegative_c)} \quad (27)$$

3. Opis implementacji

Podstawową klasą jest klasa „Article” która przechowuje w sobie treść artykułu oraz informację o etykiecie którą ten artykuł jest oznaczony. Wszystkie artykuły są przechowywane w obiekcie klasy „ArticleStore” która odpowiada za podział artykułów na zbiór uczący i testowy. Podczas tworzenia programu został wykorzystany wzorec projektowy strategii a świadczą o tym interfejsy „FeatureExtractors” implementowany przez klasy które są odpowiedzialne za obliczanie wartości cech tekstu oraz „DistanceMeasurement” implementowany przez klasy które reprezentują poszczególne miary/metryki. Klasa „NeighboursSpaceCreator” jest odpowiedzialna za przygotowanie zbioru uczącego czyli przechowywanie wartości cech dla artykułów ze zbioru uczącego. Klasa „Knn” na podstawie k najbliższych sąsiadów przydziela elementowi ze zbioru uczącego odpowiednią etykietę. Program został napisany w języku Java w wersji 13.

Rysunek 1. Diagram UML



4. Materiały i metody

Badania zostały przeprowadzone na zbiorze artykułów w języku angielskim które były oznaczone etykietami w dwóch kategoriach w kategorii „places” (west-germany, usa, france,uk ,canada, japan) oraz w kategorii „topics” (earn, acq)

4.1. Wpływ liczby sąsiadów na jakość klasyfikacji

Badanie zostało przeprowadzone z różną ilością sąsiadów 1,3,4,7,11,12,13,19,37,59
 Pozostałe parametry były stałe:
 Metryka: Metryka Euklidesa
 Podział danych: 70% - zbiór uczący (Places), 50% - zbiór uczący (Topics)
 Cechy: 1,2,3,4,5,6,7,8,9,10,11

4.2. Wpływ podziału danych na jakość klasyfikacji

Badanie zostało przeprowadzone z różną wielkość zbioru uczącego 30%,40%,50%,60%,70%
 Pozostałe parametry były stałe:
 Metryka: Metryka Euklidesa
 Liczba sąsiadów: 37
 Cechy: 1,2,3,4,5,6,7,8,9,10,11

4.3. Wpływ metryki/miary na jakość klasyfikacji

Badanie zostało przeprowadzone z różnymi metrykami/miarami: Średnia arytmetyczna minimum, Metryka Czebyszewa, Metryka Euklidesa, Minimum-maximum, Metryka Uliczna, Nasza Metryka. Pozostałe parametry były stałe:

Podział danych: 60% - zbiór uczący (Places), 50% - zbiór uczący (Topics)

Liczba sąsiadów: 37

Cechy: 1,2,3,4,5,6,7,8,9,10,11

4.4. Wpływ cech na jakość klasyfikacji

Badania zostały przeprowadzone w taki sposób, że od podstawowego zbioru cech czyli cech: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 odejmowaliśmy jedną cechę (lub dodawaliśmy w przypadku cech związanych z etykietami) i najbardziej interesujące przypadki zamieściliśmy w wynikach. Pozostałe parametry były stałe:

Metryka: Metryka Euklidesa (Topics), Metryka Uliczna (Places)

Liczba sąsiadów: 37

Podział danych: 60%

4.5. Prównanie naszej metryki i metryki Czebyszewa

Badanie zostało przeprowadzone z różną ilością sąsiadów 1,3,4,7,11,12,13,19,37,59 i przy użyci Naszej Metryki i Metryki Czebyszewa. Pozostałe parametry były stałe:

Podział danych: 60%

Cechy: 1,2,3,4,5,6,7,8,9,10,11

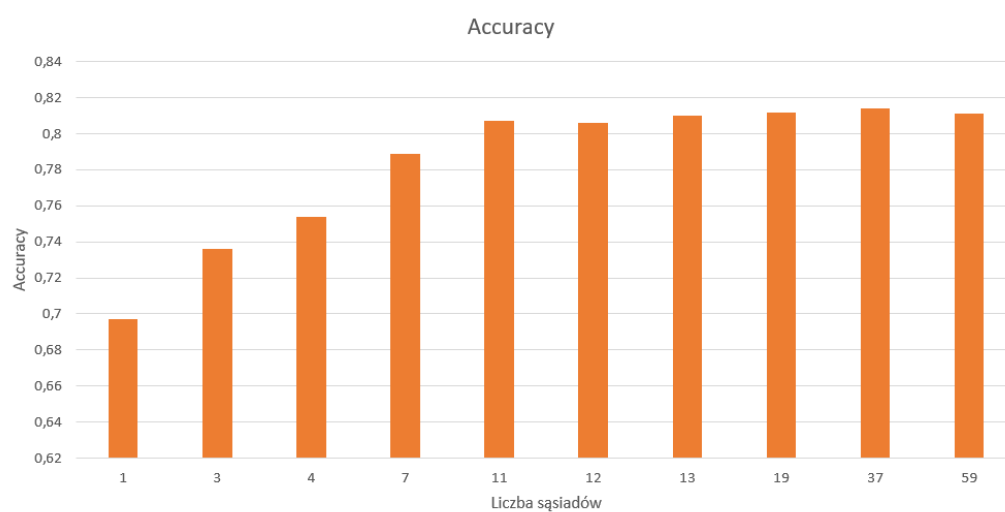
5. Wyniki

5.1. Wyniki klasyfikacji metodą k-NN dla 10 różnych wartości parametru k (Places)

Tabela 1. Wyniki klasyfikacji dla różnej liczby sąsiadów

Liczba sąsiadów	Accuracy	Średni Recall	Średni Precision
1	0,697	0,314	0,287
3	0,736	0,307	0,292
4	0,754	0,311	0,305
7	0,789	0,309	0,363
11	0,807	0,301	0,398
12	0,806	0,290	0,372
13	0,81	0,291	0,373
19	0,812	0,263	0,364
37	0,814	0,246	0,538
59	0,811	0,229	0,369

Rysunek 2. Accuracy w zależności od liczby sąsiadów



Rysunek 3. Recall i Precision w zależności od liczby sąsiadów

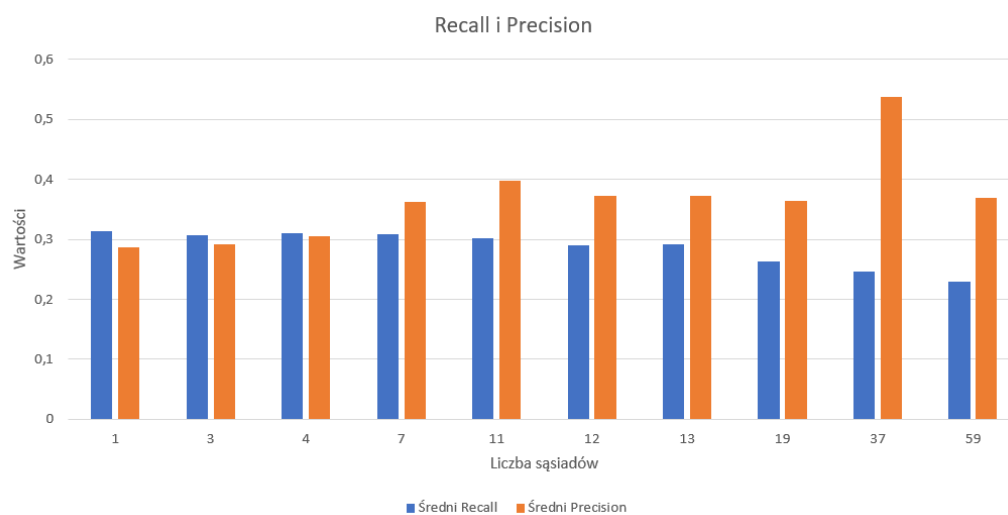


Tabela 2. Wyniki klasyfikacji dla różnej liczby sąsiadów z podziałem na etykiety

Liczba sąsiadów	Recall west-germany	Precision west-germany	Recall usa	Precision usa	Recall france	Precision france
1	0,16	0,136	0,813	0,876	0,085	0,054
3	0,113	0,16	0,864	0,875	0,017	0,048
4	0,075	0,154	0,886	0,875	0,017	0,034
7	0,132	0,326	0,936	0,862	0,034	0,111
11	0,075	0,333	0,962	0,854	0,017	0,111
12	0,094	0,333	0,964	0,852	0	0
13	0,094	0,345	0,969	0,853	0	0
19	0,038	0,25	0,981	0,841	0	0
37	0,028	0,429	0,991	0,828	0,017	1
59	0,009	0,143	0,993	0,823	0,017	0,25

Tabela 3. Wyniki klasyfikacji dla różnej liczby sąsiadów z podziałem na etykiety

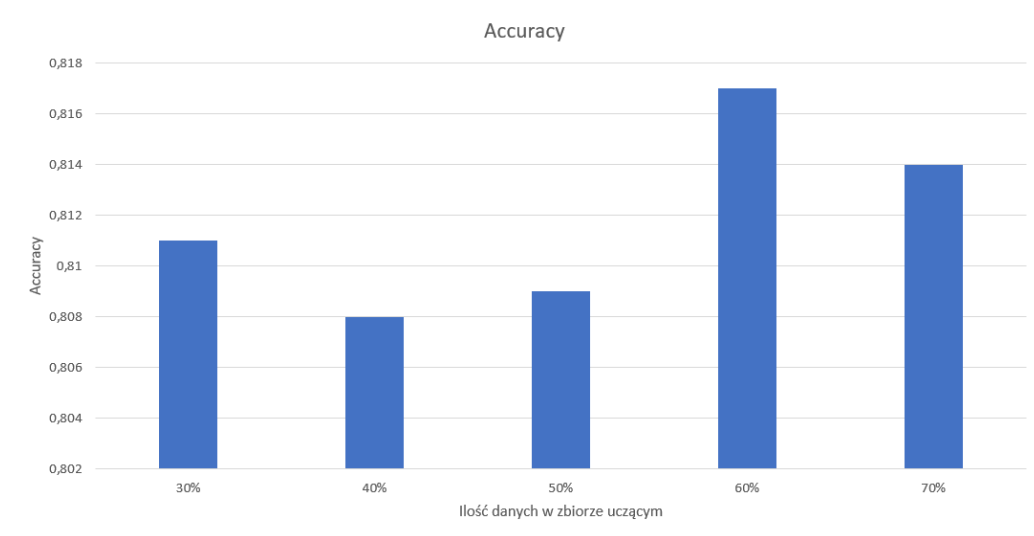
Liczba sąsiadów	Recall uk	Precision uk	Recall canada	Precision canada	Recall japan	Precision japan
1	0,332	0,298	0,189	0,138	0,303	0,219
3	0,378	0,303	0,119	0,184	0,352	0,184
4	0,401	0,35	0,139	0,179	0,345	0,24
7	0,385	0,409	0,057	0,187	0,31	0,283
11	0,378	0,478	0,041	0,25	0,331	0,361
12	0,363	0,468	0,037	0,231	0,283	0,35
13	0,37	0,469	0,025	0,188	0,29	0,382
19	0,332	0,503	0,012	0,214	0,214	0,378
37	0,282	0,607	0	0	0,159	0,365
59	0,252	0,595	0	0	0,103	0,405

5.2. Wyniki klasyfikacji metodą k-NN dla 5 różnych podziałów na zbiór uczący i testowy (Places)

Tabela 4. Wyniki klasyfikacji dla różnych podziałów na zbiór uczący i testowy

Ilość danych w zbiorze uczący	Accuracy	Średni Recall	Średni Precision
30%	0,811	0,229	0,387
40%	0,808	0,235	0,369
50%	0,809	0,240	0,334
60%	0,817	0,239	0,404
70%	0,814	0,246	0,538

Rysunek 4. Accuracy w zależności od podziału na zbiór uczący i testowy



Rysunek 5. Recall i Precision w zależności od podziału na zbiór uczący i testowy

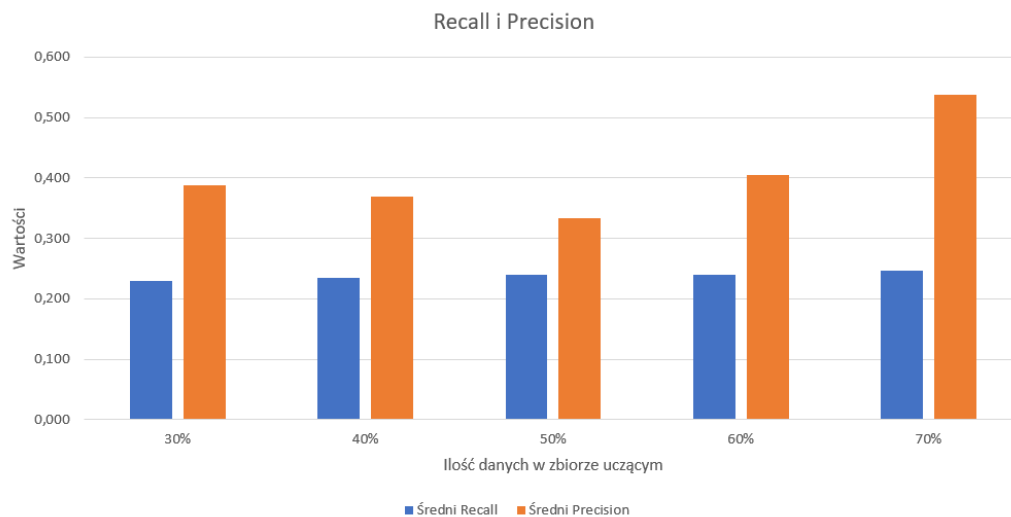


Tabela 5. Wyniki klasyfikacji dla różnych podziałów na zbiór uczący i testowy dla etykiet

Podział danych	Recall west-germany	Precision west-germany	Recall usa	Precision usa	Recall france	Precision france
30%	0,018	0,444	0,99	0,825	0,019	0,2
40%	0,031	0,333	0,99	0,823	0	0
50%	0,019	0,375	0,987	0,83	0	0
60%	0,033	0,444	0,989	0,833	0,011	0,25
70%	0,028	0,429	0,991	0,828	0,017	1

Tabela 6. Wyniki klasyfikacji dla różnych podziałów na zbiór uczący i testowy dla etykiet

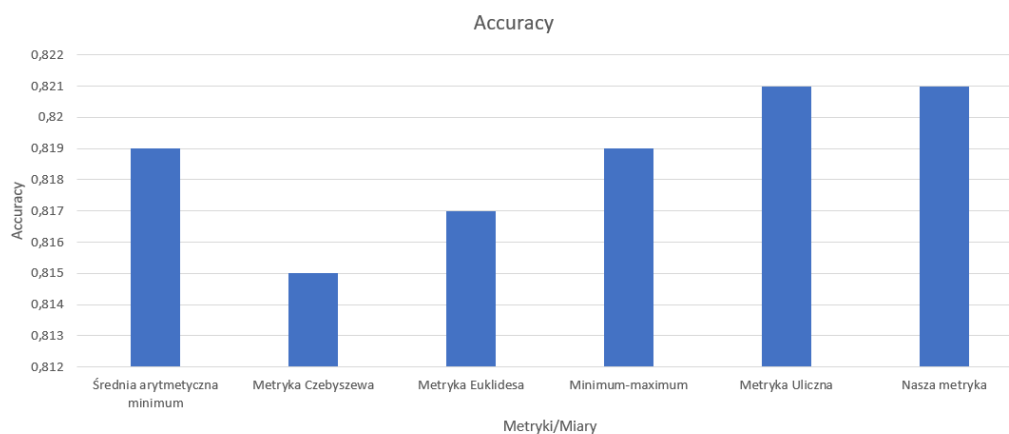
Podział danych	Recall uk	Precision uk	Recall canada	Precision canada	Recall japan	Precision japan
30%	0,272	0,561	0	0	0,075	0,293
40%	0,279	0,515	0	0	0,108	0,544
50%	0,329	0,536	0	0	0,102	0,26
60%	0,312	0,558	0	0	0,089	0,34
70%	0,282	0,607	0	0	0,159	0,365

5.3. Wyniki klasyfikacji metodą k-NN dla 6 różnych metryk/miar (Places)

Tabela 7. Wyniki klasyfikacji dla różnych metryk/miar

Metryka	Accuracy	Średni Recall	Średni Precision
Średnia arytmetyczna minimum	0,819	0,229	0,384
Metryka Czebyszewa	0,815	0,223	0,379
Metryka Euklidesa	0,817	0,238	0,399
Minimum-maximum	0,819	0,229	0,384
Metryka Uliczna	0,821	0,242	0,407
Nasza metryka	0,821	0,242	0,42

Rysunek 6. Accuracy w zależności od miary/metryki



Rysunek 7. Recall i Precision w zależności od miary/metryki

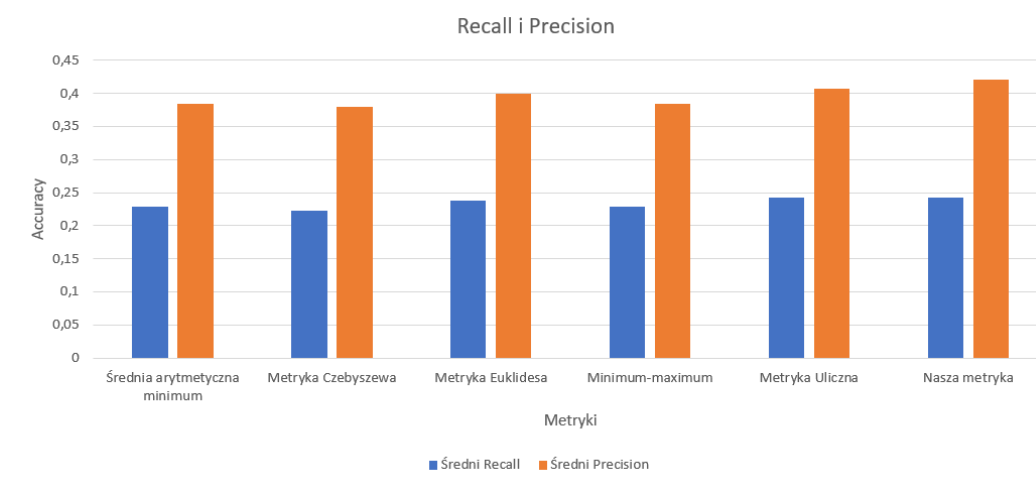


Tabela 8. Wyniki klasyfikacji dla różnych mair/metryk z podziałem na etykiety

Metryka	Recall west-germany	Precision west-germany	Recall usa	Precision usa	Recall france	Precision france
Średnia arytmetyczna minimum	0,008	0,333	0,995	0,827	0	0
Metryka Czebyszewa	0,016	0,4	0,994	0,825	0	0
Metryka Euklidesa	0,033	0,5	0,989	0,832	0,011	0,167
Minimum-maximum	0,008	0,333	0,995	0,827	0	0
Metryka Uliczna	0,016	0,5	0,994	0,832	0	0
Nasza metryka	0,008	0,333	0,994	0,83	0,011	0,167

Tabela 9. Wyniki klasyfikacji dla różnych mair/metryk z podziałem na etykiety

Metryka	Recall uk	Precision uk	Recall canada	Precision canada	Recall japan	Precision japan
Średnia arytmetyczna minimum	0,271	0,63	0	0	0,099	0,514
Metryka Czebyszewa	0,244	0,58	0	0	0,084	0,471
Metryka Euklidesa	0,312	0,555	0	0	0,084	0,34
Minimum -maximum	0,271	0,63	0	0	0,099	0,514
Metryka Uliczna	0,3	0,637	0	0	0,141	0,474
Nasza metryka	0,285	0,63	0	0	0,152	0,558

5.4. Wyniki klasyfikacji metodą k-NN dla różnych cech (Places)

Podstawowe cechy to 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11

Tabela 10. Wyniki klasyfikacji dla 4 podzbiorów cech

Różnice pomiędzy cechami podstawowymi	Accuracy	Średni Recall	Średni Precision
Cechy podstawowe	0,817	0,238	0,399
Dodatkowe cechy: 12 13 14 15 16 17	0,875	0,492	0,892
Brak cechy 8	0,805	0,179	0,25
Brak cechy 6	0,811	0,22	0,351
Brak cechy 9	0,819	0,241	0,417

Tabela 11. Wyniki klasyfikacji dla 4 podzbiorów cech z podziałem na etykiety

Różnice w cechach	Recall west-germany	Precision west-germany	Recall usa	Precision usa	Recall france	Precision france
Cechy podstawowe	0,033	0,5	0,989	0,832	0,011	0,167
Dodatkowo 12 13 14 15 16 17	0,325	0,952	0,995	0,873	0,389	0,833
Brak 8	0	0	0,998	0,806	0	0
Brak 6	0,024	0,429	0,99	0,825	0	0
Brak 9	0,033	0,5	0,991	0,833	0,011	0,2

Tabela 12. Wyniki klasyfikacji dla 4 podzbiorów cech z podziałem na etykiety

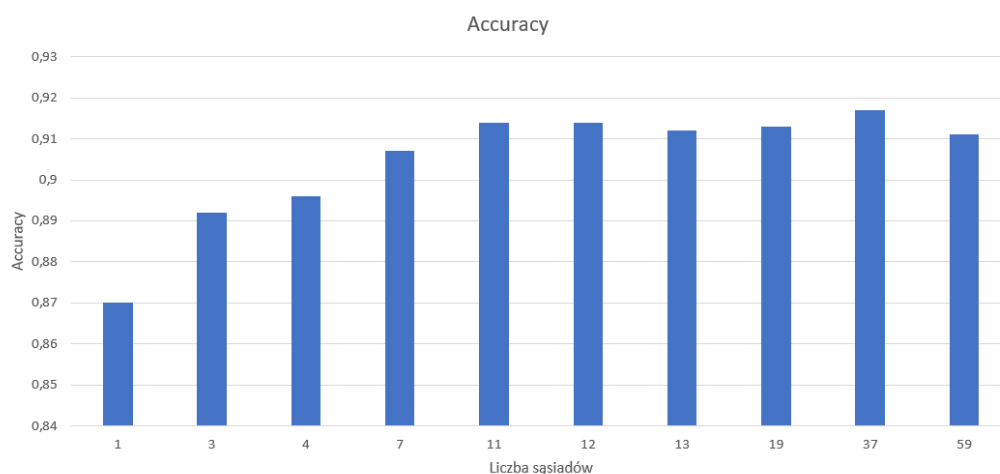
Różnice w cechach	Recall uk	Precision uk	Recall canada	Precision canada	Recall japan	Precision japan
Cechy podstawowe	0,312	0,555	0	0	0,084	0,34
Dodatkowo 12 13 14 15 16 17	0,468	0,874	0,254	0,86	0,524	0,962
Brak 8	0,074	0,694	0	0	0	0
Brak 6	0,212	0,514	0	0	0,094	0,34
Brak 9	0,321	0,565	0	0	0,089	0,405

5.5. Wyniki klasyfikacji metodą k-NN dla 10 różnych wartości parametru k

Tabela 13. Wyniki klasyfikacji dla różnej liczby sąsiadów (Topics)

Liczba sąsiadów	Accuracy	Średni Recall	Średni Precision
1	0,87	0,86	0,859
3	0,892	0,889	0,881
4	0,896	0,905	0,883
7	0,907	0,907	0,896
11	0,914	0,917	0,902
12	0,914	0,921	0,902
13	0,912	0,916	0,901
19	0,913	0,918	0,902
37	0,917	0,925	0,905
59	0,911	0,922	0,899

Rysunek 8. Accuracy w zależności od liczby sąsiadów



Rysunek 9. Recall i Precision w zależności od liczby sąsiadów

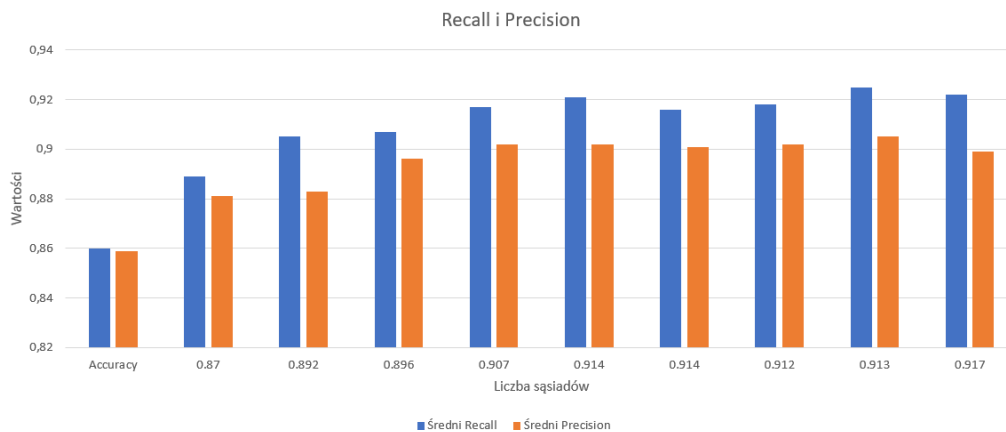


Tabela 14. Wyniki klasyfikacji dla różnej liczby sąsiadów z podziałem na etykiety

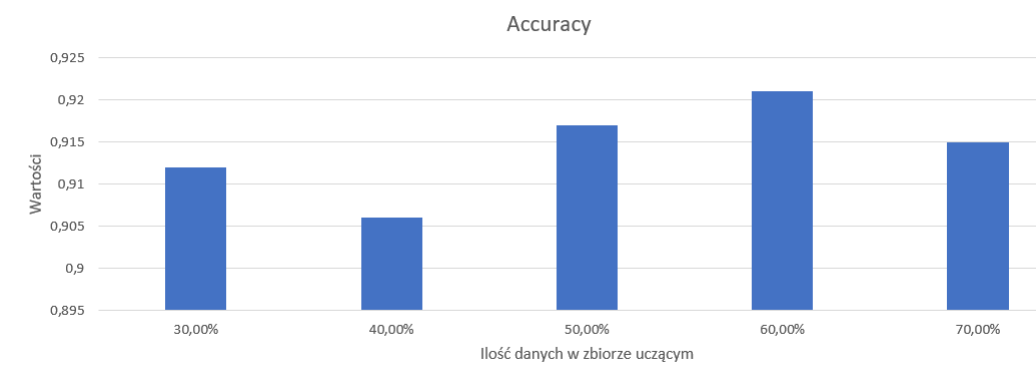
Liczba sąsiadów	Recall earn	Precision earn	Recall acq	Precision acq
1	0,897	0,899	0,822	0,819
3	0,901	0,928	0,877	0,834
4	0,873	0,96	0,936	0,807
7	0,906	0,946	0,908	0,845
11	0,906	0,957	0,928	0,847
12	0,894	0,968	0,948	0,835
13	0,902	0,958	0,93	0,843
19	0,901	0,961	0,935	0,843
37	0,898	0,97	0,951	0,841
59	0,884	0,974	0,959	0,824

5.6. Wyniki klasyfikacji metodą k-NN dla 5 różnych podziałów na zbiór uczący i testowy (Topics)

Tabela 15. Wyniki klasyfikacji dla różnych podziałów na zbiór uczący i testowy

Podział danych	Accuracy	Średni Recall	Średni Precision
30,00%	0,912	0,921	0,901
40,00%	0,906	0,921	0,893
50,00%	0,917	0,925	0,905
60,00%	0,921	0,926	0,909
70,00%	0,915	0,925	0,902

Rysunek 10. Accuracy w zależności od podziału na zbiór uczący i testowy



Rysunek 11. Recall i Precision w zależności od podziału na zbiór uczący i testowy

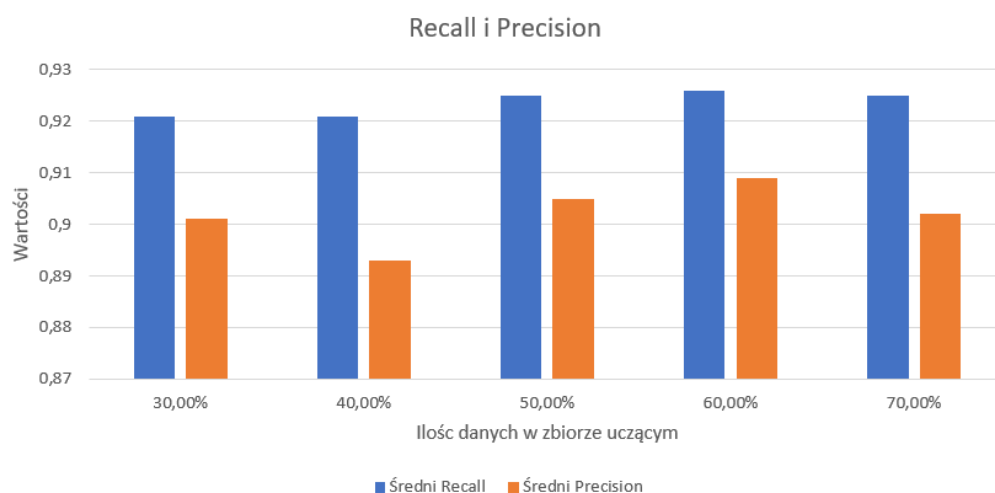


Tabela 16. Wyniki klasyfikacji dla różnych podziałów na zbiór uczący i testowy dla różnych etykiet

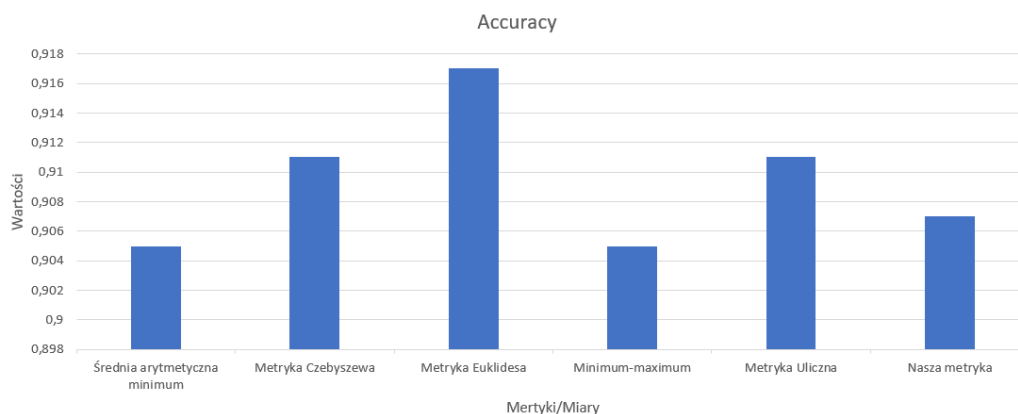
Podział danych	Recall west-germany	Precision west-germany	Recall usa	Precision usa
30,00%	0,018	0,444	0,99	0,825
40,00%	0,031	0,333	0,99	0,823
50,00%	0,019	0,375	0,987	0,83
60,00%	0,033	0,444	0,989	0,833
70,00%	0,028	0,429	0,991	0,828

5.7. Wyniki klasyfikacji metodą k-NN dla 6 różnych metryk/miar (Topics)

Tabela 17. Wyniki klasyfikacji dla różnych metryk/miar

Metryka	Accuracy	Średni Recall	Średni Precision
Średnia arytmetyczna minimum	0,905	0,918	0,893
Metryka Czebyszewa	0,911	0,919	0,899
Metryka Euklidesa	0,917	0,925	0,905
Minimum-maximum	0,905	0,918	0,893
Metryka Uliczna	0,911	0,92	0,899
Nasza metryka	0,907	0,916	0,894

Rysunek 12. Accuracy w zależności od miary/metryki



Rysunek 13. Recall i Precision w zależności od miary/metryki

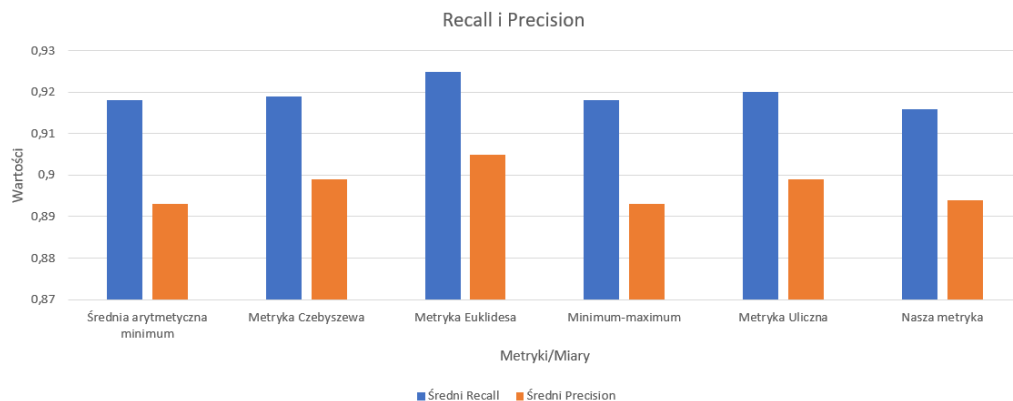


Tabela 18. Wyniki klasyfikacji dla różnych metryk z podziałem na etykiety

Metryka	Recall earn	Precision earn	Recall acq	Precision acq
Średnia arytmetyczna minimum	0,871	0,978	0,966	0,809
Metryka Czebyszewa	0,89	0,968	0,947	0,829
Metryka Euklidesa	0,898	0,97	0,951	0,841
Minimum -maximum	0,871	0,978	0,966	0,809
Metryka Uliczna	0,887	0,971	0,953	0,826
Nasza metryka	0,882	0,969	0,949	0,82

5.8. Wyniki klasyfikacji metodą k-NN dla różnych cech (Topics)

Podstawowe cechy to 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11

Tabela 19. Wyniki klasyfikacji dla 4 podzbiorów cech

Różnice pomiędzy cechami podstawowymi	Accuracy	Średni Recall	Średni Precision
Cechy podstawowe	0,921	0,926	0,909
Dodatkowe cechy: 12 13	0,938	0,944	0,926
Brak cechy 11	0,894	0,899	0,88
Brak cechy 3	0,871	0,888	0,861
Brak cechy 10	0,925	0,928	0,913

Tabela 20. Wyniki klasyfikacji dla 4 podzbiorów cech z podziałem na etykiety

Różnice w cechach	Recall earn	Precision earn	Recall acq	Precision acq
Podstawowe	0,909	0,966	0,943	0,852
Dodatkowo 12 13	0,921	0,981	0,967	0,872
Brak 11	0,879	0,951	0,919	0,809
Brak 3	0,829	0,966	0,948	0,755
Brak 10	0,917	0,964	0,938	0,863

5.9. Porównanie naszej metryki z metryką Czebyszewa (Places)

Tabela 21. Wyniki klasyfikacji dla różnej liczby sąsiadów

Liczba sąsiadów	Accuracy Nasza Metryka	Accuracy Metryka Czebyszewa
1	0,701	0,694
3	0,751	0,735
4	0,763	0,749
7	0,802	0,79
11	0,816	0,803
12	0,816	0,804
13	0,816	0,806
19	0,819	0,813
37	0,821	0,815
59	0,817	0,814

Tabela 22. Wyniki klasyfikacji dla różnej liczby sąsiadów z podziałem na etykiety
(Nasza metryka)

Liczba sąsiadów	Recall west-germany	Precision west-germany	Recall usa	Precision usa	Recall france	Precision france
1	0,171	0,149	0,808	0,883	0,111	0,075
3	0,081	0,204	0,882	0,885	0,156	0,083
4	0,098	0,231	0,897	0,879	0,078	0,062
7	0,057	0,179	0,952	0,866	0,1	0,155
11	0,049	0,261	0,973	0,856	0,089	0,216
12	0,033	0,2	0,974	0,854	0,089	0,235
13	0,049	0,353	0,975	0,851	0,056	0,185
19	0,024	0,273	0,985	0,843	0,044	0,222
37	0,008	0,333	0,994	0,83	0,011	0,167
59	0	0	0,998	0,821	0	0

Tabela 23. Wyniki klasyfikacji dla różnej liczby sąsiadów z podziałem na etykiety
(Nasza Metryka)

Liczba sąsiadów	Recall uk	Precision uk	Recall canada	Precision canada	Recall japan	Precision japan
1	0,382	0,297	0,187	0,125	0,33	0,267
3	0,335	0,376	0,165	0,141	0,236	0,241
4	0,368	0,379	0,152	0,155	0,204	0,235
7	0,374	0,439	0,076	0,224	0,215	0,293
11	0,365	0,506	0,057	0,273	0,194	0,349
12	0,359	0,508	0,06	0,311	0,194	0,352
13	0,365	0,512	0,044	0,311	0,178	0,327
19	0,324	0,545	0,029	0,375	0,178	0,405
37	0,285	0,63	0	0	0,152	0,558
59	0,232	0,681	0	0	0,058	0,611

Tabela 24. Wyniki klasyfikacji dla różnej liczby sąsiadów z podziałem na etykiety
(Metryka Czebyszewa)

Liczba sąsiadów	Recall west-germany	Precision west-germany	Recall usa	Precision usa	Recall france	Precision france
1	0,179	0,142	0,807	0,875	0,089	0,066
3	0,089	0,224	0,866	0,878	0,089	0,052
4	0,114	0,255	0,884	0,868	0,1	0,087
7	0,081	0,27	0,944	0,856	0,078	0,091
11	0,065	0,242	0,964	0,847	0,033	0,088
12	0,041	0,208	0,969	0,843	0,022	0,061
13	0,024	0,167	0,973	0,843	0	0
19	0,049	0,316	0,983	0,838	0,011	0,062
37	0,016	0,4	0,994	0,825	0	0
59	0	0	0,997	0,82	0	0

Tabela 25. Wyniki klasyfikacji dla różnej liczby sąsiadów z podziałem na etykiety (Metryka Czebyszewa)

Liczba sąsiadów	Recall uk	Precision uk	Recall canada	Precision canada	Recall japan	Precision japan
1	0,261	0,159	0,114	0,309	0,245	0,267
3	0,309	0,152	0,134	0,225	0,207	0,241
4	0,325	0,124	0,132	0,209	0,227	0,235
7	0,422	0,051	0,165	0,162	0,215	0,293
11	0,466	0,019	0,146	0,178	0,276	0,349
12	0,473	0,016	0,167	0,162	0,29	0,352
13	0,479	0,016	0,161	0,157	0,286	0,327
19	0,518	0	0	0,131	0,316	0,405
37	0,58	0	0	0,084	0,471	0,558
59	0,636	0	0	0,052	0,5	0,611

6. Dyskusja

Liczba sąsiadów

W poniższej dyskusji odwołujemy się do Rysunków 2 i 8 oraz Tabel 1 i 13

Po wyznaczeniu klasyfikacji metodą Knn dla różnej liczby sąsiadów otrzymaliśmy podobne wyniki w przypadku klasyfikacji w kategorii „places” (6 etykiet) i „topic” (2 etykiety). W obu przypadkach można zauważyć znaczny wzrost dokładności klasyfikacji przy zwiększaniu liczby sąsiadów od 1 do 11. Wynika z tego, że zbyt mała liczba sąsiadów znacznie obniża accuracy klasyfikacji. Powyżej 11 sąsiadów accuracy w przypadku kategorii „topic” się stabilizuje i waha wokół wartości osiągniętej w przypadku 11 sąsiadów. W przypadku kategorii „places” accuracy bardzo powoli ale jednak rośnie aż do 37 sąsiadów. Wynika z tego, że liczba sąsiadów przy której metoda Knn osiąga największe accuracy rośnie wraz z liczbą etykiet na które dzielimy dany zbiór. Wynika z tego również, że zbyt duża liczba sąsiadów zwiększa tylko czas obliczeń a nie poprawia otrzymanego wyniku.

Podział danych na zbiór uczący i testowy

Poniżej odwołujemy się do Rysunków 4 i 10 oraz Tabel 4 i 15

W oby kategoriach „places” i „topic” otrzymaliśmy podobne wyniki. Różnica pomiędzy najlepszym i najgorszym wynikiem nie przekraczała 2%. Najlepszy wynik otrzymaliśmy dla podziału 60% danych na zbiór uczący oraz 40% danych na zbiór testowy. Natomiast najgorszy wynik otrzymaliśmy w zestawie który jest lustrzanym odbiciem poprzedniego czyli 40% na zbiór uczący i 60% na zbiór testowy. Wynika z tego, że podział na zbiór uczący i testowy nie ma dużego wpływu na otrzymane wyniki. Wynika z tego również, że to który podział jest najlepszy a który najgorszy jest niezależne od kategorii i ilości etykiet według których chcemy klasyfikować nasze dane. Dodatkowo wynika z tego że podział 60% na zbiór uczący i 40% na zbiór testowy jest dużo lepszym podziałem niż podział 40% na zbiór uczący i 60% na zbiór

testowy.

Miara/Metryka

Poniżej odwołujemy się do Rysunków 6 i 12 oraz Tabel 7 i 17

W obu kategoriach „places” i „topic” otrzymaliśmy różne wyniki. Różnica pomiędzy najlepszym a najgorszym wynikiem znowu nie przekracza 2%. W przypadku kategorii „places” dwie najlepsze metryki to „Metryka Uliczna” i „Nasza Metryka” a dwie najgorsze to „Metryka Euklidesa” i „Metryka Czebyszewa”. W przypadku kategorii „topic” dwie najlepsze metryki to „Metryka Czebyszewa” i „Metryka Euklidesa” a dwie najgorsze to „Średnia arytmetyczna minimum” i „Minimum maximum”. Wynika z tego, że metryka nie ma dużego wpływu na wynik. Wynika z tego również, że to która metryka jest najlepsza zależy od kategorii oraz ilości etykiet na które dzielimy nasz zbiór.

Różne cechy.

Poniżej odwołujemy się do Tabel 10 i 19

W przypadku kategorii „places” znaczną przewagę nad innymi mają cechy z słowami kluczowymi dla poszczególnych etykiet (jednak jest ich 6 i muszą występować wszystkie razem). Wyniki dla pozostałych cech bardzo mało się różnią. Jednak za dwie najlepsze to cechy 8 i 6 a za najgorszą ponieważ jej obecność pogarsza wynik cechę 9. W przypadku kategorii „topic” cechy ze słowami kluczowymi dla poszczególnych etykiet nie mają już tak dużego wpływu jak poprzednio. Dwie najlepsze cechy w tym przypadku to 11 i 3 a najgorsza jest cecha 10.. Wynika z tego, że jakość cechy bardzo mocno zależy od kategorii oraz etykiet na które dzielimy nasz zbiór.

7. Wnioski

1. Znaczny wzrost accuracy występuje wraz ze wzrostem liczby sąsiadów ale tylko to pewnego momentu.
2. Zbyt duża liczba sąsiadów od pewnego momentu nie zwiększa już accuracy
3. Optymalna liczba sąsiadów jest niezależna od liczby etykiet na które dzielimy dany zbiór
4. Podział danych na zbiór uczący i testowy nie ma dużego wpływu na wyniki.
5. To który podział jest najlepszy a który najgorszy jest niezależne od kategorii i ilości etykiet na które dzielimy nasz zbiór
6. Najlepszy jest podział 60% na zbiór uczący i 40% na zbiór testowy a najgorszy 40% na zbiór testowy i 60% na zbiór uczący.
7. Metryka nie ma dużego wpływu na accuracy otrzymanego wyniku
8. To która metryka jest najlepsza silnie zależy od kategorii i tego na jakie etykiety dzielimy nasz zbiór.
9. Cechy mają duży wpływ na otrzymane wyniki.
10. To które cechy są najlepsze silnie zależy od kategorii i tego na jakie etykiety dzielimy nasz zbiór.

11. Metoda Knn daje dobre wyniki w przypadku gdy do każdej etykiet należy zbliżona liczba elementów
12. Metoda Knn nie sprawdza się dobrze w przypadku gdy w zbiorze elementów z jednej etykiet jest znacznie więcej niż elementów należących do innych etykiet.
13. Accuracy nie jest dobrą miarą skuteczności metody aby otrzymać pełne informacje o działaniu metody należy obliczyć recall i precision dla każdej z etykiet
14. W przypadku gdy elementy należące do jednej z etykiet są dużo bardziej liczne niż elementy należące do innych etykiet metoda knn zdecydowanie częściej przypisuje testowy element do tej etykiety niż do innych niezależnie od tego do jakiej etykiety powinien należeć element.

Literatura

- [1] Adam Niewiadomski "Materiały, przykłady i ćwiczenia do przedmiotu Komputerowe Systemy Rozpoznawania 1"
- [2] <https://machinelearningmastery.com/precision-recall-and-f-measure-for-imbalanced-classification/>
- [3] <https://machinelearningmastery.com/confusion-matrix-machine-learning/>
- [4] <http://home.agh.edu.pl/horzyk/lectures/miw/KNN.pdf>