Jakub Polak

Łukasz Główka

Grupowanie NBC ze względu na miary: NormManhattan_VDM i 1-Gower

Opis przydzielonego zadania projektowego

Naszym zadaniem była implementacja grupowania NBC ze względu na miary NormManhattan_VDM oraz 1-Gower.

Przyjęte założenia

Identyfikatory grup zaczynają się od 1.

W przypadku napotkania brakujących wartości atrybutów podczas obliczania miary NormManhattan_VDM odległość między tymi wartościami przyjmuje najwyższą możliwą wartość tj. 1 dla atrybutów numerycznych, lub 2 dla atrybutów symbolicznych.

Wartość R^+kNN jest wyznaczana podczas znajdowania k^+NN.

Wartość minimalna i maksymalna atrybutu numerycznego jest wyznaczana ze zbioru podczas jego załadowania.

Opis postaci danych wejściowych i wyjściowych

Dane wejściowe to wartości liczbowe lub napisy. Wartości brakujące powinny zostać zaznaczone jako NA, a identyfikator grupy powinien znaleźć się na ostatnim miejscu.

Struktura projektu

Projekt został napisany w języku C++. Pliki z rozszerzeniem .h znajdują się w folderze include/, a implementacyjne w folderze src/. Testy jednostkowe zostały zaimplementowane przy użyciu Google Test i znajdują się w folderze tests/.

Dane wczytywane są do klasy Dataset. Posiada ona klasy Datapoint oraz Parameter. Zawierają one informacje o wartości lub jej braku.

Przy wykonywaniu algorytmu NBC, tworzony zostaje GroupingSet. Znajduje on minimalne oraz maksymalne wartości danych numerycznych. Weryfikuje on także, czy dane typu int są numeryczne/prawda-fałsz. GroupingPoint przechowuje wszystkie potrzebne atrybuty do obsługi NBC, takie jak k^+NN, Rk^+NN oraz NBF.

Algorytm NBC, metryki, tworzenie plików wyjściowych oraz rand znajdują się w odosobnionych przestrzeniach nazw.

Podręcznik użytkownika (czyli jak korzystać z programu)

Aby wywołać program potrzebny jest nam plik z odpowiednio sformatowanymi danymi wejściowymi. Program odpalamy podając:

MED-project.out PLIK_WEJŚCIOWY.CSV WARTOŚĆ_K ALGORYTM(gower/m_vdm/m_vdm_0_5) FOLDER_WYJŚCIOWY TYPY_KOLUMN(string/int/bool/double)

np. MED-project.out ../data/test5.csv 3 m_vdm ./output_gower int,bool,string

Ilustracja działania zaimplementowanych algorytmów na przykładzie małego zbioru danych

x	y
4.2	4.0
5.9	3.9
2.8	3.5
12.0	1.3
10.0	1.3
1.1	3.0
0.0	2.4
2.4	2.0
11.5	1.8
11.0	1.0
0.9	0.0
1.0	1.5

Używamy tu przykład wykładowy.

Plik main.cpp przyjmuje oraz sprawdza argumenty programu. Wyznaczany jest plik wejściowy, wartości są ładowane do klasy Dataset, a następnie tworzona jest metryka.

Algorytm NBC tworzy GroupingSet oraz wyznacza k⁺NN oraz NDF dla punktów. Następnie tworzy on grupy oraz przypisuje klasy.

Wykorzystuje on do tego przekazaną metrykę, które obliczają odległość w zależności od typu danych przechowywanych w punkcie. Wykorzystywane są do tego funkcje pomocnicze oraz template.

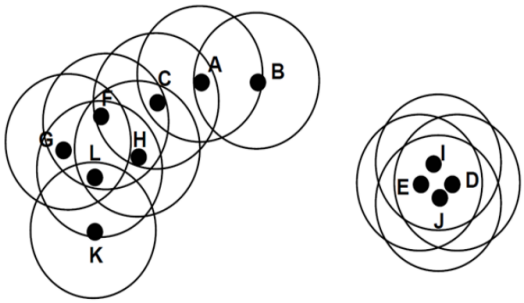
Po uruchomieniu algorytmu (metryka 1-gower), tworzą się następujące pliki:

-K+NN - dane z grupowania nbc, m.in. k⁺NN i odległości.

Id	Eps	maxEps	NDF	rk+nn	k+nn_value	k+nn
0	0.25	0.66	0.67	2	3	1 2 5
1	0.31	0.70	0.67	2	3	0 2 5
2	0.18	0.66	1.00	3	3	0 5 1
3	0.08	0.67	1.00	3	3	9 4 8
4	0.12	0.58	1.00	3	3	9 3 8
5	0.18	0.67	2.00	6	3	6 2 7
6	0.15	0.64	1.33	4	3	5 7 11
7	0.18	0.49	1.33	4	3	11 6 5
8	0.12	0.67	1.00	3	3	3 9 4
9	0.12	0.66	1.00	3	3	3 4 8
10	0.34	0.70	0.33	1	3	11 7 6
11	0.19	0.50	0.75	3	4	7 6 5 10

-OUT - zawiera klasę wyjściową. Możemy zobaczyć, że poprawnie wykryta została grupa 2 (EIJD/3489).

Id	Type	Group	Values
0	0	1	4.20, 4.00
1	0	1	5.90, 3.90
2	1	1	2.80, 3.50
3	1	2	12.00, 1.30
4	1	2	10.00, 1.30
5	1	1	1.10, 3.00
6	1	1	0.00, 2.40
7	1	1	2.40, 2.00
8	1	2	11.50, 1.80
9	1	2	11.00, 1.00
10	-1	-1	0.90, 0.00
11	0	1	1.00, 1.50

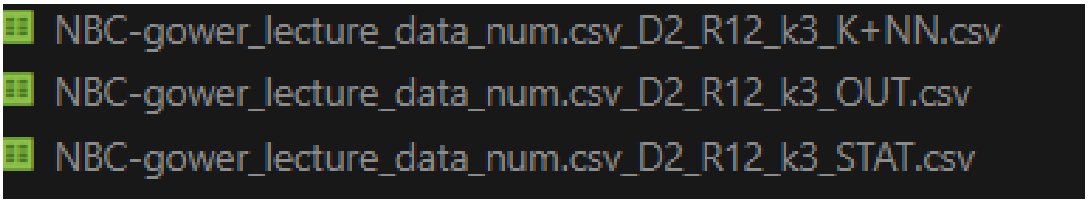


- STAT - statystyki oraz dane algorytmu.

```
Input filename: lecture_data_num.csv
Number of rows: 12
Number of attributes: 2
K: 3
File reading Time: 0 ms.
K+NN Time: 5 ms.
Grouping Time: 0 ms.
Sum: 5 ms.
Number of core points: 8
Number of boundary points: 3
Number of noise points: 1
TP: 21
TN: 33
Rand: 0.82
```

Plik K+NN zawiera dane odnośnie grupowania obiektów.

Powstałe nazwy plików:



Przykład z danymi mieszanymi

Id	Age	Car type	Risk
0	**23**	**Family**	**High**
1	**17**	**Sport**	**High**
2	43	Sport	Low
3	68	Family	Low
4	32	Truck	Low
5	20	Family	High

(metryka 1-gower)

```
Input filename: lecture_data_cars.csv
Number of rows: 6
Number of attributes: 2
K: 1
File reading Time: 0 ms.
K+NN Time: 1 ms.
Grouping Time: 0 ms.
Sum: 1 ms.
Number of core points: 4
Number of boundary points: 0
Number of noise points: 2
TP: 2
TN: 8
Rand: 0.67
```

Id	Type	Group	Values
0	1	1	23.00, Family
1	1	2	17.00, Sport
2	1	2	43.00, Sport
3	-1	-1	68.00, Family
4	-1	-1	32.00, Truck
5	1	1	20.00, Family

(metryka manhattan-vdm)

```
Input filename: lecture_data_cars.csv
Number of rows: 6
Number of attributes: 2
K: 1
File reading Time: 0 ms.
K+NN Time: 3 ms.
Grouping Time: 0 ms.
Sum: 3 ms.
Number of core points: 4
Number of boundary points: 0
Number of noise points: 2
TP: 2
TN: 9
Rand: 0.73
```

Id	Type	Group	Values
0	1	1	23.00, Family
1	1	2	17.00, Sport
2	1	3	43.00, Sport
3	-1	-1	68.00, Family
4	-1	-1	32.00, Truck
5	1	1	20.00, Family

W powyższym przypadku możemy zobaczyć działanie różnych metryk.

W przypadku 1-gower wykryte zostały 2 grupy, a przy metryce manhattan-vdm już 3.

Wyniki jakościowe i ilościowe dla 3 zbiorów danych

Za pomocą skryptu uruchamiany jest program dla 3 zbiorów danych:

Nazwa Zbioru	Typ danych	Liczba wierszy	Liczba atrybutów
German_Credit	mieszane	200	20
3D_Road_Network	numeryczne	300	3

Cervical_Cancer_Behavior_Risk	Przekonwertowane do nominalnych	72	19

1. German_Credit, zbiór mieszany

1-gower		
k=1	k=3	k=5
Input filename: german_credit.csv Number of rows: 100 Number of attributes: 20 K: 1 File reading Time: 10 ms. K+NN Time: 157793 ms. Grouping Time: 0 ms. Sum: 157803 ms. Number of core points: 53 Number of boundary points: 0 Number of noise points: 47 TP: 680 TN: 1447 Rand: 0.43	Input filename: german_credit.csv Number of rows: 100 Number of attributes: 20 K: 3 File reading Time: 5 ms. K+NN Time: 166651 ms. Grouping Time: 0 ms. Sum: 166656 ms. Number of core points: 50 Number of boundary points: 18 Number of noise points: 32 TP: 763 TN: 1426 Rand: 0.44	Input filename: german_credit.csv Number of rows: 100 Number of attributes: 20 K: 5 File reading Time: 5 ms. K+NN Time: 151063 ms. Grouping Time: 0 ms. Sum: 151068 ms. Number of core points: 48 Number of boundary points: 27 Number of noise points: 25 TP: 1484 TN: 836 Rand: 0.47
manhattan-vdm		
k=1	k=3	k=5

Input filename: german_credit.csv Number of rows: 100 Number of attributes: 20 K: 1 File reading Time: 6 ms. K+NN Time: 1030910 ms. Grouping Time: 0 ms. Sum: 1030916 ms. Number of core points: 62 Number of boundary points: 0 Number of noise points: 38 TP: 492 TN: 1622 Rand: 0.43	Input filename: german_credit.csv Number of rows: 100 Number of attributes: 20 K: 3 File reading Time: 6 ms. K+NN Time: 966183 ms. Grouping Time: 0 ms. Sum: 966189 ms. Number of core points: 52 Number of boundary points: 20 Number of noise points: 28 TP: 859 TN: 1444 Rand: 0.47	Input filename: german_credit.csv Number of rows: 100 Number of attributes: 20 K: 5 File reading Time: 6 ms. K+NN Time: 948619 ms. Grouping Time: 0 ms. Sum: 948625 ms. Number of core points: 53 Number of boundary points: 24 Number of noise points: 23 TP: 1558 TN: 1245 Rand: 0.57
Manhattan-vdm, categorical scaling=0.5		
k=1	k=3	k=5
Input filename: german_credit.csv Number of rows: 100 Number of attributes: 20 K: 1 File reading Time: 6 ms. K+NN Time: 1028432 ms. Grouping Time: 0 ms. Sum: 1028438 ms. Number of core points: 62 Number of boundary points: 0 Number of noise points: 38 TP: 465 TN: 1603 Rand: 0.42	Input filename: german_credit.csv Number of rows: 100 Number of attributes: 20 K: 3 File reading Time: 6 ms. K+NN Time: 942655 ms. Grouping Time: 0 ms. Sum: 942661 ms. Number of core points: 58 Number of boundary points: 19 Number of noise points: 23 TP: 850 TN: 1292 Rand: 0.43	Input filename: german_credit.csv Number of rows: 100 Number of attributes: 20 K: 5 File reading Time: 9 ms. K+NN Time: 939112 ms. Grouping Time: 0 ms. Sum: 939121 ms. Number of core points: 50 Number of boundary points: 28 Number of noise points: 22 TP: 1375 TN: 1272 Rand: 0.53

2. 3D_Road_Network, zbiór numeryczny

1-gower		
k=1	k=3	k=5

Input filename: 3d_road_network.c Number of rows: 300 Number of attributes: 3 K: 1 File reading Time: 5 ms. K+NN Time: 110405 ms. Grouping Time: 0 ms. Sum: 110410 ms. Number of core points: 205 Number of boundary points: 1 Number of noise points: 94 TP: 234 TN: 39185 Rand: 0.88	Input filename: 3d_road_network.csv Number of rows: 300 Number of attributes: 3 K: 3 File reading Time: 3 ms. K+NN Time: 121966 ms. Grouping Time: 0 ms. Sum: 121969 ms. Number of core points: 193 Number of boundary points: 47 Number of noise points: 60 TP: 514 TN: 41272 Rand: 0.93	Input filename: 3d_road_network.c Number of rows: 300 Number of attributes: 3 K: 5 File reading Time: 5 ms. K+NN Time: 121746 ms. Grouping Time: 0 ms. Sum: 121751 ms. Number of core points: 175 Number of boundary points: 82 Number of noise points: 43 TP: 666 TN: 41371 Rand: 0.94
Manhattan-vdm		
k=1	k=3	k=5
Input filename: 3d_road_network.csv Number of rows: 300 Number of attributes: 3 K: 1 File reading Time: 6 ms. K+NN Time: 175568 ms. Grouping Time: 0 ms. Sum: 175574 ms. Number of core points: 205 Number of boundary points: 1 Number of noise points: 94 TP: 234 TN: 39185 Rand: 0.88	Input filename: 3d_road_network.csv Number of rows: 300 Number of attributes: 3 K: 3 File reading Time: 4 ms. K+NN Time: 188652 ms. Grouping Time: 0 ms. Sum: 188656 ms. Number of core points: 193 Number of boundary points: 47 Number of noise points: 60 TP: 514 TN: 41272 Rand: 0.93	Input filename: 3d_road_network.csv Number of rows: 300 Number of attributes: 3 K: 5 File reading Time: 2 ms. K+NN Time: 189735 ms. Grouping Time: 0 ms. Sum: 189737 ms. Number of core points: 176 Number of boundary points: 81 Number of noise points: 43 TP: 666 TN: 41371 Rand: 0.94
Manhattan-vdm, categorical scaling=0.5		
k=1	k=3	k=5

Input filename: 3d_road_network.csv Number of rows: 300 Number of attributes: 3 K: 1 File reading Time: 2 ms. K+NN Time: 185368 ms. Grouping Time: 0 ms. Sum: 185370 ms. Number of core points: 205 Number of boundary points: 1 Number of noise points: 94 TP: 234 TN: 39185 Rand: 0.88	Input filename: 3d_road_network.csv Number of rows: 300 Number of attributes: 3 K: 3 File reading Time: 2 ms. K+NN Time: 187680 ms. Grouping Time: 0 ms. Sum: 187682 ms. Number of core points: 193 Number of boundary points: 47 Number of noise points: 60 TP: 514 TN: 41272 Rand: 0.93	Input filename: 3d_road_network.csv Number of rows: 300 Number of attributes: 3 K: 5 File reading Time: 2 ms. K+NN Time: 187161 ms. Grouping Time: 0 ms. Sum: 187163 ms. Number of core points: 176 Number of boundary points: 81 Number of noise points: 43 TP: 666 TN: 41371 Rand: 0.94

3. Cervical_Cancer_Behavior_Risk, zbiór nominalny

1-gower		
k=1	k=3	k=5
Input filename: cancer_risk.csv Number of rows: 72 Number of attributes: 19 K: 1 File reading Time: 5 ms. K+NN Time: 58781 ms. Grouping Time: 0 ms. Sum: 58786 ms. Number of core points: 41 Number of boundary points: 4 Number of noise points: 27 TP: 241 TN: 927 Rand: 0.46	Input filename: cancer_risk.csv Number of rows: 72 Number of attributes: 19 K: 3 File reading Time: 4 ms. K+NN Time: 61904 ms. Grouping Time: 0 ms. Sum: 61908 ms. Number of core points: 39 Number of boundary points: 9 Number of noise points: 24 TP: 380 TN: 778 Rand: 0.45	Input filename: cancer_risk.csv Number of rows: 72 Number of attributes: 19 K: 5 File reading Time: 6 ms. K+NN Time: 64109 ms. Grouping Time: 0 ms. Sum: 64115 ms. Number of core points: 38 Number of boundary points: 17 Number of noise points: 17 TP: 977 TN: 427 Rand: 0.55
Manhattan-vdm		
k=1	k=3	k=5

Input filename: cancer_risk.csv Number of rows: 72 Number of attributes: 19 K: 1 File reading Time: 6 ms. K+NN Time: 464895 ms. Grouping Time: 0 ms. Sum: 464901 ms. Number of core points: 39 Number of boundary points: 0 Number of noise points: 33 TP: 352 TN: 871 Rand: 0.48	Input filename: cancer_risk.csv Number of rows: 72 Number of attributes: 19 K: 3 File reading Time: 4 ms. K+NN Time: 464204 ms. Grouping Time: 0 ms. Sum: 464208 ms. Number of core points: 33 Number of boundary points: 11 Number of noise points: 28 TP: 402 TN: 924 Rand: 0.52	Input filename: cancer_risk.csv Number of rows: 72 Number of attributes: 19 K: 5 File reading Time: 4 ms. K+NN Time: 466861 ms. Grouping Time: 0 ms. Sum: 466865 ms. Number of core points: 33 Number of boundary points: 23 Number of noise points: 16 TP: 449 TN: 1032 Rand: 0.58
Manhattan-vdm, categorical scaling=0.5		
k=1	k=3	k=5
Input filename: cancer_risk.csv Number of rows: 72 Number of attributes: 19 K: 1 File reading Time: 4 ms. K+NN Time: 458142 ms. Grouping Time: 0 ms. Sum: 458146 ms. Number of core points: 39 Number of boundary points: 0 Number of noise points: 33 TP: 352 TN: 871 Rand: 0.48	Input filename: cancer_risk.csv Number of rows: 72 Number of attributes: 19 K: 3 File reading Time: 4 ms. K+NN Time: 466911 ms. Grouping Time: 0 ms. Sum: 466915 ms. Number of core points: 33 Number of boundary points: 11 Number of noise points: 28 TP: 402 TN: 924 Rand: 0.52	Input filename: cancer_risk.csv Number of rows: 72 Number of attributes: 19 K: 5 File reading Time: 4 ms. K+NN Time: 467619 ms. Grouping Time: 0 ms. Sum: 467623 ms. Number of core points: 33 Number of boundary points: 23 Number of noise points: 16 TP: 449 TN: 1032 Rand: 0.58

4. German_Credit, 10% wartości brakujących

1 - gower

```
Input filename: german_credit_missing.csv
Number of rows: 100
Number of attributes: 20
K: 3
File reading Time: 14 ms.
K+NN Time: 133822 ms.
Grouping Time: 0 ms.
Sum: 133836 ms.
Number of core points: 48
Number of boundary points: 15
Number of noise points: 37
TP: 813
TN: 1375
Rand: 0.44
```

Manhattan-vdm

```
Input filename: german_credit_missing.csv
Number of rows: 100
Number of attributes: 20
K: 3
File reading Time: 5 ms.
K+NN Time: 877045 ms.
Grouping Time: 0 ms.
Sum: 877050 ms.
Number of core points: 31
Number of boundary points: 8
Number of noise points: 61
TP: 1270
TN: 1063
Rand: 0.47
```

Manhattan-vdm, categorical scaling=0.5

```
Input filename: german_credit_missing.csv
Number of rows: 100
Number of attributes: 20
K: 3
File reading Time: 7 ms.
K+NN Time: 913839 ms.
Grouping Time: 0 ms.
Sum: 913846 ms.
Number of core points: 33
Number of boundary points: 6
Number of noise points: 61
TP: 1210
TN: 1032
Rand: 0.45
```

Wnioski z realizacji projektu

Użycie metryki Manhattan Vdm jest korzystniejsze przy ocenianiu zbioru, w których znajdują się kategorie. W zaimplementowanej przez nas wersji jest on jednak znacznie wolniejszy, w stosunku do użycia metryki 1-gower.

Parametr categoricalScaling nie wpłynął znacząco na indeks rand. Wyniki użytych przez nas metryk są porównywalne.

Bibliografia

German Credit https://github.com/deric/clustering-

benchmark/blob/master/src/main/resources/datasets/real-world/german.arff

3D Road Network
https://archive.ics.uci.edu/ml/datasets/3D+Road+Network+%28North+Jutland%2C+Denmark%29

Cervical Cancer Behavior Risk
https://archive.ics.uci.edu/ml/datasets/Cervical+Cancer+Behavior+Risk