Selektywny naiwny algorytm Bayesa

Wstęp

Naiwny algorytm Bayesa jest jednym z najpowszechniejszych algorytmów eksploracji danych używanych w zadaniu klasyfikacji. Wnioskuje prawdopodobieństwo, że nowy przykład należy do pewnej klasy na podstawie założenia, że wszystkie atrybuty są od siebie niezależne w danej klasie. Założenie to jest motywowane potrzebą oszacowania wielowymiarowych prawdopodobieństw na podstawie danych uczących. W praktyce większość kombinacji wartości atrybutów albo nie występuje w danych uczących, albo nie występuje w wystarczającej liczbie. W konsekwencji bezpośrednie oszacowanie każdego istotnego prawdopodobieństwa wielowymiarowego nie będzie wiarygodne. Naiwny algorytm Bayesa omija tę sytuację dzięki założeniu warunkowej niezależności. Chociaż wykazał się już on niezwykłą dokładnością klasyfikacji, założenie o warunkowej niezależności rzadko jest prawdziwe w rzeczywistości. W rezultacie naturalne jest ulepszanie go poprzez rozluźnienie założenia o warunkowej niezależności. Co za tym idzie wybranie tylko niektórych atrybutów może spowodować lepsze wyniki klasyfikacji. Proponuję tutaj zastosowanie algorytmu zdolnego do selekcji atrybutów, zwanego selektywnym naiwnym algorytmem Bayesa.

Metodologia

Do opisu skuteczności algorytmów wykorzystam miarę dokładności (accuracy). Mówi ona o stosunku poprawnych klasyfikacji do wszystkich klasyfikacji.

Wyniki każdego algorytmu przedstawiam, prezentując wynik accuracy obydwu algorytmów dla każdego z rozważanych zbiorów danych ze współczynnikami split (parametr funkcji train_test_split()) w zakresach od 0.5 do 0.9.

Opis algorytmu

Zasadą działania selektywnego naiwnego algorytmu Bayesa jest wybranie odpowiedniego podzbioru cech i uruchomienie na tak zbudowanym zbiorze danych naiwnego algorytmu Bayesa, aby odczytać z niego wynik. Bardziej szczegółowy opis wygląda następująco:

1. Dla każdej cechy oblicz współczynnik Mutual Info Score (MIS) cechy oraz kolumny klasy.
2. Posortuj cechy rosnąco po współczynniku mutual info score;
3. Stwórz podzbiory cech według następującego wzoru:
    - Pierwszy zbiór to cecha o najmniejszym współczynniku MIS;
    - Każdy kolejny zbiór otrzymujemy przez dodanie kolejnej cechy z listy;
4. Uruchom Naiwny Algorytm Bayesa dla każdego podzbioru cech;
5. Wybierz wyniki tego podzbioru, które zwracają najmniejszy błąd średniokwadratowy;

Porównanie z naiwnym algorytmem Bayesa

Oba algorytmy porównałem na kilku zbiorach danych: Iris, Wine, Digits, Breast Cancer, Students i Boston Housing.

Dane Iris:

https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_iris.html#sklearn.data sets.load_iris

Dane zawierające długości i szerokości płatków irysa: 150 próbek, 3 klasy.

```
Split parameter: 0.5
Naive Bayes accuracy:           0.9466666666666667
Selective Naive Bayes accuracy: 0.9466666666666667
[1, 0, 3, 2]

Split parameter: 0.6
Naive Bayes accuracy:           0.9444444444444444
Selective Naive Bayes accuracy: 0.9555555555555556
[1, 0, 3]

Split parameter: 0.7
Naive Bayes accuracy:           0.9333333333333333
Selective Naive Bayes accuracy: 0.9333333333333333
[1, 0, 3]

Split parameter: 0.8
Naive Bayes accuracy:           0.9333333333333333
Selective Naive Bayes accuracy: 0.9333333333333333
[1, 0, 3, 2]

Split parameter: 0.9
Naive Bayes accuracy:           0.9481481481481482
Selective Naive Bayes accuracy: 0.9481481481481482
[1, 0, 3, 2]
```

Tutaj wyniki są bardzo zbliżone dla obu algorytmów. Jedynie przy podziale na dane treningowe i testowe w stosunku 60:40 nieco lepiej poradził sobie algorytm selektywny.

Dane Breast Cancer:

https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_breast_cancer.html#sklearn.datasets.load_breast_cancer

Zbiór zawierający dane o osobach chorych na raka i informacje czy jest złośliwy czy nie. Zawiera 569 próbek i 2 klasy.

```
Split parameter: 0.5
Naive Bayes accuracy:           0.9368421052631579
Selective Naive Bayes accuracy: 0.9578947368421052
[24, 8, 4, 18, 1, 9, 28, 17, 0, 21, 11, 29, 22, 20]

Split parameter: 0.6
Naive Bayes accuracy:           0.935672514619883
Selective Naive Bayes accuracy: 0.9444444444444444
[24, 8, 4, 18, 1, 9, 28, 17, 0, 21, 11, 29, 22, 20]

Split parameter: 0.7
Naive Bayes accuracy:           0.9373433583959899
Selective Naive Bayes accuracy: 0.9448621553884712
[24, 8, 4, 18, 1, 9, 28, 17, 0, 21, 11, 29, 22, 20]

Split parameter: 0.8
Naive Bayes accuracy:           0.9210526315789473
Selective Naive Bayes accuracy: 0.9429824561403509
[24, 8, 4, 18, 1, 9, 28, 17, 0, 21, 11, 29, 22, 20, 5, 27, 19, 25, 15, 13, 2, 16, 3]

Split parameter: 0.9
Naive Bayes accuracy:           0.9337231968810916
Selective Naive Bayes accuracy: 0.9571150097465887
[24, 8, 4, 18, 1, 9, 28, 17, 0, 21, 11, 29, 22, 20]
```

Tutaj wyniki są zdecydowanie na korzyść algorytmu selektywnego. Dla każdego podziału radzi sobie zdecydowanie lepiej niż klasyczny naiwny algorytm Bayesa.

Dane Digits:

https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_digits.html#sklearn.datasets.load_digits

Zbiór zawiera informacje o każdym pikselu grafiki 8x8 i informacje jaką liczbę przedstawia. Zawiera 1797 próbek i 10 klas.

```
Split parameter: 0.5
Naive Bayes accuracy:           0.8342602892102335
Selective Naive Bayes accuracy: 0.8342602892102335
[0, 32, 39, 56, 24, 16, 31, 48, 8, 40, 47, 23, 15, 7, 55, 49, 63, 57, 1, 11, 14, 4, 6, 17,

Split parameter: 0.6
Naive Bayes accuracy:           0.830398517145505
Selective Naive Bayes accuracy: 0.830398517145505
[0, 32, 39, 56, 24, 16, 31, 48, 8, 40, 47, 23, 15, 7, 55, 49, 63, 57, 1, 11, 14, 4, 6, 17,

Split parameter: 0.7
Naive Bayes accuracy:           0.8251192368839427
Selective Naive Bayes accuracy: 0.8251192368839427
[0, 32, 39, 56, 24, 16, 31, 48, 8, 40, 47, 23, 15, 7, 55, 49, 63, 57, 1, 11, 14, 4, 6, 17,

Split parameter: 0.8
Naive Bayes accuracy:           0.8164116828929068
Selective Naive Bayes accuracy: 0.8164116828929068
[0, 32, 39, 56, 24, 16, 31, 48, 8, 40, 47, 23, 15, 7, 55, 49, 63, 57, 1, 11, 14, 4, 6, 17,

Split parameter: 0.9
Naive Bayes accuracy:           0.8108776266996292
Selective Naive Bayes accuracy: 0.8108776266996292
[0, 32, 39, 56, 24, 16, 31, 48, 8, 40, 47, 23, 15, 7, 55, 49, 63, 57, 1, 11, 14, 4, 6, 17,
```

Tutaj z kolei oba algorytmy osiągają dokładnie takie same wyniki.

Dane Wine:

https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_wine.html#sklearn.datasets.load_wine

Zbiór zawiera dane o składzie chemicznym win i informacje o typie wina. Zawiera 178 próbek i 3 klasy.

```
Split parameter: 0.5
Naive Bayes accuracy:           0.9438202247191011
Selective Naive Bayes accuracy: 0.9775280898876404
[7, 4, 3, 2, 10, 8, 5, 1, 12, 0, 11, 9]

Split parameter: 0.6
Naive Bayes accuracy:           0.9345794392523364
Selective Naive Bayes accuracy: 0.9532710280373832
[7, 4, 3, 2, 10, 8, 5, 1, 12, 0, 11, 9]

Split parameter: 0.7
Naive Bayes accuracy:           0.968
Selective Naive Bayes accuracy: 0.968
[7, 4, 3, 2, 10, 8, 5, 1, 12, 0, 11]

Split parameter: 0.8
Naive Bayes accuracy:           0.965034965034965
Selective Naive Bayes accuracy: 0.9790209790209791
[7, 4, 3, 2, 10, 8, 5, 1, 12, 0, 11, 9]

Split parameter: 0.9
Naive Bayes accuracy:           0.906832298136646
Selective Naive Bayes accuracy: 0.9130434782608695
[7, 4, 3, 2, 10, 8, 5, 1, 12, 0, 11, 9]
```

Ponownie można zauważyć, że algorytm selektywny radzi sobie znacznie lepiej niemal dla każdego podziału.

Dane Students:

https://www.kaggle.com/csafrit2/higher-education-students-performance-evaluation

Dane o ocenach studentów na podstawie różnych parametrów: 145 próbek, na 30 parametrów. Sprawdzanym parametrem jest ocena końcowa (GRADE). Dość duża liczba parametrów może sprzyjać użyciu przez Selektywny algorytm mniejszego datasetu.

W tym przypadku algorytm selektywny osiąga gorsze wyniki.

```
Split parameter: 0.5
Naive Bayes accuracy:           0.1780821917808219
Selective Naive Bayes accuracy: 0.136986301369863
[19, 23]

Split parameter: 0.6
Naive Bayes accuracy:           0.21839080459770116
Selective Naive Bayes accuracy: 0.1839080459770115
[19]

Split parameter: 0.7
Naive Bayes accuracy:           0.13725490196078433
Selective Naive Bayes accuracy: 0.09803921568627451
[19, 23, 27, 5, 4, 6]

Split parameter: 0.8
Naive Bayes accuracy:           0.1724137931034483
Selective Naive Bayes accuracy: 0.1724137931034483
[19]

Split parameter: 0.9
Naive Bayes accuracy:           0.1984732824427481
Selective Naive Bayes accuracy: 0.16793893129770993
[19]
```

Dane Boston:

https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_boston.html

Dane o cenach domów w mieście Boston,w których klasyfikatorem jest cecha RAD. Dane wejściowe mają 506 wierszy i 13 parametrów. Dla tego Datasetu wyniki dla obu algorytmów wyszły następująco:

Tylko dla dużego Split Parameter (0.9) Selektywny algorytm wniósł lepsze wyniki.

```
Split parameter: 0.5
Naive Bayes accuracy:           0.6205533596837944
Selective Naive Bayes accuracy: 0.6205533596837944
[3, 1, 10, 9, 6, 8, 5, 11, 4, 2, 7, 0]

Split parameter: 0.6
Naive Bayes accuracy:           0.5888157894736842
Selective Naive Bayes accuracy: 0.5888157894736842
[3, 1, 10, 9, 6, 8, 5, 11, 4, 2, 7, 0]

Split parameter: 0.7
Naive Bayes accuracy:           0.5380281690140845
Selective Naive Bayes accuracy: 0.5380281690140845
[3, 1, 10, 9, 6, 8, 5, 11, 4, 2, 7, 0]

Split parameter: 0.8
Naive Bayes accuracy:           0.5185185185185185
Selective Naive Bayes accuracy: 0.5185185185185185
[3, 1, 10, 9, 6, 8, 5, 11, 4, 2, 7, 0]

Split parameter: 0.9
Naive Bayes accuracy:           0.484649122807017555
Selective Naive Bayes accuracy: 0.5
[3, 1, 10, 9, 6, 8]
```

Wnioski:

Jest pewien typ zbiorów danych, dla których selektywny algorytm potrafi dać lepsze rezultaty niż zwykły naiwny klasyfikator Bayesa. Niestety, na przykładowych danych, które zostały zbadane trudno ocenić, które parametry decydują o przewagach naiwnego algorytmu, ponieważ nie widać wielu wyraźnych zależności. W niektórych przypadkach oba algorytmy próbowały zwrócić dokładnie takie same wyniki.