

Metody selekcji cech - 2. Selekcja genów lub białek o ekspresji różnicującej dwie grupy próbek przy użyciu metody Recursive Feature Elimination

1 OPIS PROBLEMU

Celem projektu była implementacja algorytmu SVM-RFE oraz sprawdzenie, jak ta metoda selekcji cech wpływa na skuteczność klasyfikacji binarnej.

1.1 SVM

SVM to algorytm uczenia maszynowego, tworzący specyficzny model sieci neuronowej, w której pod uwagę brane są jedynie najtrudniej separowalne punkty przestrzeni, zwane wektorami nośnymi. Jej uczenie polega na dobieraniu wag w celu maksymalizacji marginesu (separującego skrajne punkty), który definiuje w ten sposób różne klasy. Sieci te używane są głównie w zadaniach klasyfikacji, w których w sposób jednoznaczny rozdzielają jedną klasę od pozostałych, zwracając wynik z zakresu $\{-1, 1\}$.

1.2 RECURSIVE FEATURE ELIMINATION

Algorytm SVM-RFE został zaproponowany w 2002 roku[1]. Przy badaniach danych biologicznych obserwowany jest wysoki stopień korelacji między składowymi. Dodatkowo w danych zdarzają się cechy, które nie pomagają w poprawnej klasyfikacji i mogą być uznane za szum. Dlatego tak ważny jest wybór odpowiedniej liczby cech dla danego zadania. Jeśli wartościowych cech będzie zbyt mało, model nie będzie w stanie spełnić swojego zadania. Gdy cech będzie zbyt dużo, a większość z nich nie będzie wartościowa, model będzie trudniejszy w budowie i treningu. "Dobre" cechy ułatwiają modelowanie i pozwalają uzyskać lepsze wyniki, natomiast "złe" mogą sprawić, że do osiągnięcia tego samego poziomu skuteczności trzeba będzie zbudować bardziej skomplikowany model.

Zasada działania SVM-RFE[1]:

- wejście:
 - zbiór wektorów uczących: $\mathbf{X}_0 = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k, \dots, \mathbf{x}_l]^T$
 - zbiór etykiet klas: $y = [y_1, y_2, \dots, y_k, \dots, y_l]^T$
- inicjalizacja:
 - podzbiór pozostałych cech: $\mathbf{s} = [1, 2, \dots, n]$
 - lista z rankingiem cech: $r = []$

- powtarzaj aż $\mathbf{s} = []$:
 - ogranicz zbiór uczący tak, by zawierał tylko pozostałe cechy: $X = X_0(:, \mathbf{s})$
 - wytrenuj klasyfikator SVM
 - oblicz wektor wag \mathbf{w} na podstawie współczynników stworzonego SVM
 - oblicz ranking cech: $c_i = (w_i)^2$
 - znajdź cechę o najniższym ranking: $f = \operatorname{argmin}(\mathbf{c})$
 - zaktualizuj listę z rankingiem cech: $\mathbf{r} = [\mathbf{s}(f), r]$
 - usuń cechę o najniższym rankingu: $\mathbf{s} = \mathbf{s}(1 : f - 1, f + 1 : \operatorname{length}(\mathbf{s}))$
- wyjście: lista z rankingiem cech \mathbf{r} .

2 SZCZEGÓŁY IMPLEMENTACYJNE

Algorytm RFE został zaimplementowany samodzielnie w języku Python w wersji 3.6. Wykorzystano również gotowe biblioteki napisane w tym języku:

- *pandas* i *numpy* - do przygotowania zbiorów do formatu akceptowanego przez bibliotekę *scikit-learn*,
- *scikit-learn* - do budowy maszyny wektorów nośnych oraz do przeprowadzenia sprawdzianu krzyżowego zbudowanych modeli,
- *matplotlib* - do wizualizacji wyników.

Analiza została przeprowadzona w *Jupyter Notebooku*. W zeszycie **analiza_wstepna.ipynb** pokazano, jak wygląda zbiór danych i jakie transformacje należy na nim przeprowadzić, aby można było wykorzystać go do budowy SVM z *sklearna*. Pozostała część kodu i reszta analizy znajduje się w zeszycie **18z_mpb.ipynb** (zeszyt ten razem z przebiegiem eksperymentów dostępny jest pod adresem https://nbviewer.jupyter.org/github/Sowul/18z_mpb/blob/master/18z_mpb.ipynb lub https://18zmpb-anon20vjga.notebooks.azure.com/j/notebooks/18z_mpb.ipynb).

W związku z tym, że wszystkie zbiory danych są w tym samym formacie (tsv, czyli tab-separated values) i mają taką samą strukturę, proces ich przetwarzania można było zautomatyzować. Odpowiada za to funkcja **prepare_data**, która na wejściu przyjmuje nazwę pliku, a zwraca macierz ze zbiorem trenującym, listę etykiet klas dla poszczególnych próbek oraz słownik zawierający nazwy poszczególnych cech, który posłuży do ich identyfikacji w końcowym etapie analizy.

Następną sekcją zeszytu jest **Ładowanie danych**, gdzie przetwarzane są wszystkie pliki, a zbiory danych, etykiety i słowniki tarfiają do odpowiednich tablic, które ułatwią dalszą automatyzację testów.

Kolejne komórki to już **implementacja algorytmu RFE**. Nie udało się jej wykonać w "czystym" Pythonie, ponieważ skorzystano z jednej z możliwości jakie daje biblioteka *numpy*, a

mianowicie *boolean masking*¹. Implementacja nie była zbyt skomplikowana i została przeprowadzona według schematu pokazanego w sekcji o RFE (1.2). Skorzystano ze znanej funkcji **ordinal**, która posłużyła do formatowania numeru iteracji w pętli `while` w trybie `verbose`, który jest częścią funkcji **recursive_feature_elimination**. Funkcja ta przyjmuje zbiór uczący i listę etykiet, a zwraca ranking cech w formie listy tupli w formacie (numer cechy, ranking), słownik z rankingiem (kluczem jest numer cechy, a wartością jej ranking) oraz listę wyników sprawdzianu krzyżowego przeprowadzonego w trakcie procesu eliminacji cech, posortowaną zgodnie z rosnącą liczbą wykorzystanych w procesie uczenia cech. Algorytm jest następnie testowany na znanym zbiorze *digits*.

Kolejne komórki przedstawiają **funkcję pomocniczną do robienia wykresów** średniej wartości metryki *accuracy* dla 2-fold CV w funkcji liczby wybranych cech. Funkcja pozwala na wybranie "punktu odcięcia" wykresu, z uwagi na to, że skuteczność klasyfikacji dla dużej liczby cech nie zmienia się zbyt szybko lub utrzymuje się na stałym poziomie, przez co nie jest interesująca.

3 PRZEPROWADZONE TESTY

Komórka 10 to już właściwe testy algorytmu. Korzystając z przetworzonych wcześniej zbiorów danych, algorytm uruchamiany jest w pętli `for` dla każdego zbioru osobno, a wyniki zbierane są do słownika (na najwyższym poziomie kluczem jest nazwa przetwarzanego zbioru, na niższym klucze to *final_ranking*, *ranking_dict* i *cv_results*).

Algorytm sprawdzony został na zbiorach *Leukemia*, *LungCancer* i *Lymphoma* w wersji *_preprocessed* i 500. Zbiory *_preprocessed* to zbiory wstępnie przetworzone, zawierające mniej cech niż zbiory oryginalne, ale wciąż znacząco większe od zbiorów 500.

Zbiory 500 można traktować jak zbiór cech wybranych na podstawie metod statystycznych. Dlatego pierwsze pytanie, jakie można zadać brzmi: jak bardzo będą się różniły cechy ze zbiorów 500 od 500 cech wybranych przez RFE? Następnie zbadano, który ze zbiorów (już równolicznych) pozwoli na wytrenowanie lepszego modelu. Ostatnim etapem analizy było porównanie zbiorów najlepszych cech. Do tego celu stworzono 2 funkcje pomocnicze:

- *get_best_features* - zwraca nazwy *n* wybranych najlepszych cech,
- *compare_sets* - porównuje zbiory cech.

4 WYNIKI EKSPERYMENTÓW

4.1 WYBRANIE 500 CECH

Po wybraniu ze zbiorów *preprocessed* 500 cech sprawdzono stosunek liczby wspólnych cech między zbiorami *preprocessed* i 500 do liczności zbioru *preprocessed* oraz do sumy tych zbiorów. Wyniki przedstawiają Tabela 4.1 i 4.2.

¹<https://docs.scipy.org/doc/numpy-1.15.0/user/basics.indexing.html#boolean-or-mask-index-arrays>

Zbiory	Stosunek w %
Leukemia_preprocessed i Leukemia_500	48.2
LungCancer_preprocessed i LungCancer_500	10.8
Lymphoma_preprocessed i Lymphoma_500	44.83

Tabela 4.1: Stosunek liczby wspólnych cech zbiorów do liczności zbioru *preprocessed* dla 500 wybranych cech.

Zbiory	Stosunek w %
Leukemia_preprocessed i Leukemia_500	31.75
LungCancer_preprocessed i LungCancer_500	5.708
Lymphoma_preprocessed i Lymphoma_500	28.86

Tabela 4.2: Stosunek liczby wspólnych cech zbiorów do liczności sumy tych zbiorów dla 500 wybranych cech.

4.2 PORÓWNANIE SKUTECZNOŚCI KLASYFIKACJI

Oba zbiory *Lymphoma* klasyfikowały ze 100% skutecznością już dla pojedynczej cechy. Nie jest to interesujące z punktu widzenia działania RFE, dlatego zamieszczono tylko po jednym wykresie. Wszystkie wykresy umieszczono na końcu sprawozdania. Wyniki klasyfikacji przedstawia Tabela 4.3.

4.3 PORÓWNANIE NAJLEPSZYCH WYBRANYCH CECH

W Tabeli 4.4 zestawiono minimalną liczbę cech, dla której osiągnięto najlepszy średni wynik klasyfikacji. Kolejne wykresy również przedstawiają te zależności. Podobnie jak w Sekcji 4.1, sprawdzono, jak podobne są uzyskane zbiory. Wyniki przedstawione są w Tabelach 4.5, 4.6, 4.7 i 4.8.

4.3.1 LISTY CECH

Lista cech, dla której uzyskano największą wartość metryki *accuracy* na zbiorze *Leukemia_preprocessed*:

Zbiór	Accuracy
Leukemia_preprocessed	1.000
Leukemia_500	1.000
LungCancer_preprocessed	0.994
LungCancer_500	0.962
Lymphoma_preprocessed	1.000
Lymphoma_500	1.000

Tabela 4.3: Średnia wartość *accuracy* uzyskana w 2-fold CV w procesie uczenia na zbiorze złożonym z 500 cech.

Zbiór	Liczba cech
Leukemia_preprocessed	3
Leukemia_500	7
LungCancer_preprocessed	6
LungCancer_500	16
Lymphoma_preprocessed	1
Lymphoma_500	1

Tabela 4.4: Zestawienie liczby cech, dla której osiągnięto najlepszy wynik klasyfikacji dla danego zbioru.

Zbiory	Stosunek w %
Leukemia_preprocessed i Leukemia_500	65.0
LungCancer_preprocessed i LungCancer_500	15.0
Lymphoma_preprocessed i Lymphoma_500	63.16

Tabela 4.5: Stosunek liczby wspólnych cech zbiorów do liczności zbioru *preprocessed* dla 20 wybranych cech.

Zbiory	Stosunek w %
Leukemia_preprocessed i Leukemia_500	48.15
LungCancer_preprocessed i LungCancer_500	8.108
Lymphoma_preprocessed i Lymphoma_500	48.0

Tabela 4.6: Stosunek liczby wspólnych cech zbiorów do liczności sumy tych zbiorów dla 20 wybranych cech.

Zbiory	Stosunek w %
Leukemia_preprocessed i Leukemia_500	20.0
LungCancer_preprocessed i LungCancer_500	0.0
Lymphoma_preprocessed i Lymphoma_500	80.0

Tabela 4.7: Stosunek liczby wspólnych cech zbiorów do liczności zbioru *preprocessed* dla 5 wybranych cech.

Zbiory	Stosunek w %
Leukemia_preprocessed i Leukemia_500	11.11
LungCancer_preprocessed i LungCancer_500	0.0
Lymphoma_preprocessed i Lymphoma_500	66.67

Tabela 4.8: Stosunek liczby wspólnych cech zbiorów do liczności sumy tych zbiorów dla 5 wybranych cech.

1. M11722_at
2. X95735_at
3. U70063_at

Lista cech, dla której uzyskano największą wartość metryki accuracy na zbiorze *Leukemia_500*:

1. U70063_at
2. M31166_at
3. D88270_at
4. L09209_s_at
5. M96326_rna1_at
6. U72621_at
7. X06948_at

Lista cech, dla której uzyskano największą wartość metryki accuracy na zbiorze *LungCancer_preprocessed*:

1. 33109_f_at
2. 41449_at
3. 38202_at
4. 39240_at
5. 1005_at
6. 37332_r_at

Lista cech, dla której uzyskano największą wartość metryki accuracy na zbiorze *LungCancer_500*:

1. 34091_s_at
2. 35999_r_at
3. 40079_at
4. 36040_at
5. 37148_at
6. 37345_at

7. 35726_at
8. 33218_at
9. 36133_at
10. 41424_at
11. 32138_at
12. 33343_at
13. 41385_at
14. 41620_at
15. 41691_at
16. 34265_at

Lista cech, dla której uzyskano największą wartość metryki *accuracy* na zbiorze *Lymphoma_preprocessed*:

1. (Unknown; Clone=1352493)

Lista cech, dla której uzyskano największą wartość metryki *accuracy* na zbiorze *Lymphoma_500*:

1. (Unknown; Clone=1352493)

5 PODSUMOWANIE

Z uwagi na "dydaktyczny" charakter zbiorów danych, ciężko wyciągnąć wnioski na temat wpływu zastosowanego algorytmu RFE na skuteczność klasyfikacji. Można natomiast porównać jego działanie z innym sposobem wyboru cech, które znalazły się w zbiorach typu 500. O ile zbiory *Leukemia* i *Lymphoma* dzielą około 50% cech wśród 500 wstępnie wybranych, to zbiór *LungCancer_preprocessed* przetworzony algorytmem RFE znacząco różni się od zbioru *LungCancer_500* - dzielą one tylko 11% cech w odniesieniu do zbioru *preprocessed* oraz mają poniżej 6% cech wspólnych. Udział cech wspólnych w sumie obu zbiorów dla 20 wybranych cech poprawia się tylko nieznacznie (do 8%). Dla 5 najlepszych cech iloczyn tych zbiorów jest pusty.

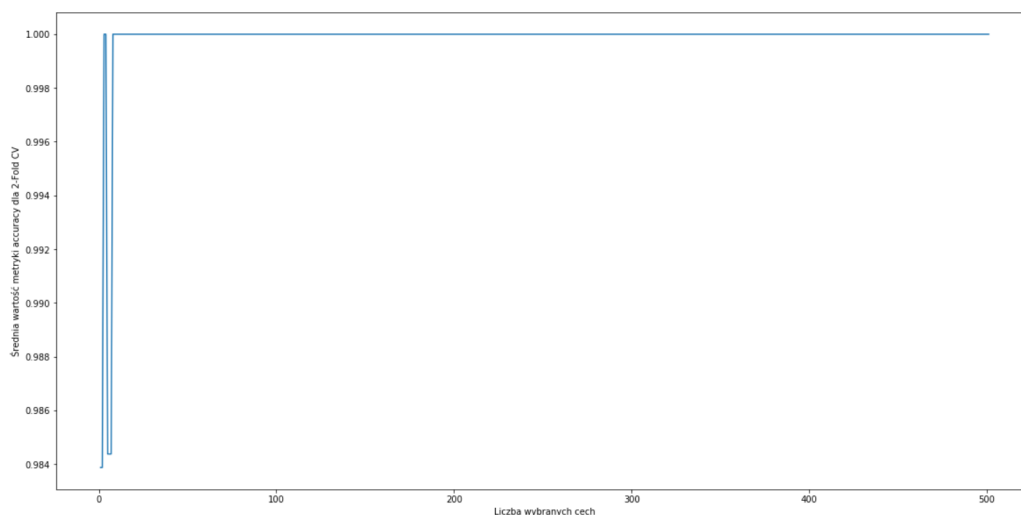
Przy dalszej selekcji cech, z 500 do 20, dla zbiorów *Leukemia* i *Lymphoma* można zauważyć podobną tendencję - wzrost udziału wybranych cech ze zbioru *preprocessed* w zbiorze 500, odpowiednio z 48% do 65% i z 45% do 63%. Podobnie jest z udziałem cech wspólnych w stosunku do sumy odpowiednich zbiorów *preprocessed* i 500 z 32% w przypadku *Leukemia* i z 28% *Lymphoma* do 48% dla 20 wybranych cech.

Zmniejszając badany zbiór do 5 najlepszych cech widać zupełnie odmienne zachowanie obu zbiorów. Tylko 1 cecha jest wspólna dla zbiorów *Leukemia_preprocessed* i *Leukemia_500*, a dla zbiorów *Lymphoma_preprocessed* i *Lymphoma_500* są to aż 4 z 5 cech.

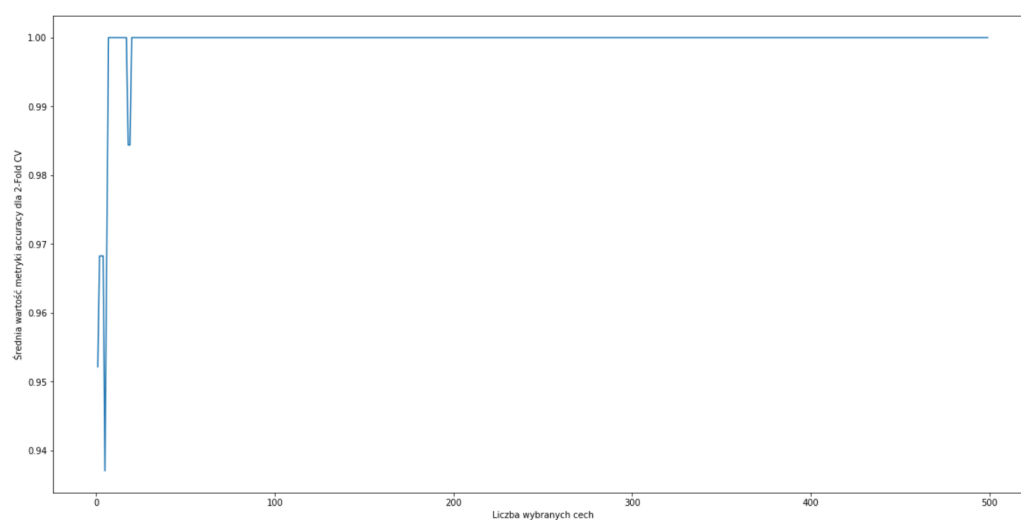
Algorytm SVM-RFE wydaje się być ciekawą, a jednocześnie prostą do zrozumienia i zaimplementowania, techniką selekcji cech, która może być stosowana zarówno do wstępnego wybrania cech, jak i do dalszego zawężania ich zbioru. Nie niesie on jednak informacji o poziomie przydatności danej cechy, więc warto stosować go w połączeniu na przykład z walidacją krzyżową lub innymi technikami oceny i selekcji cech.

BIBLIOGRAFIA

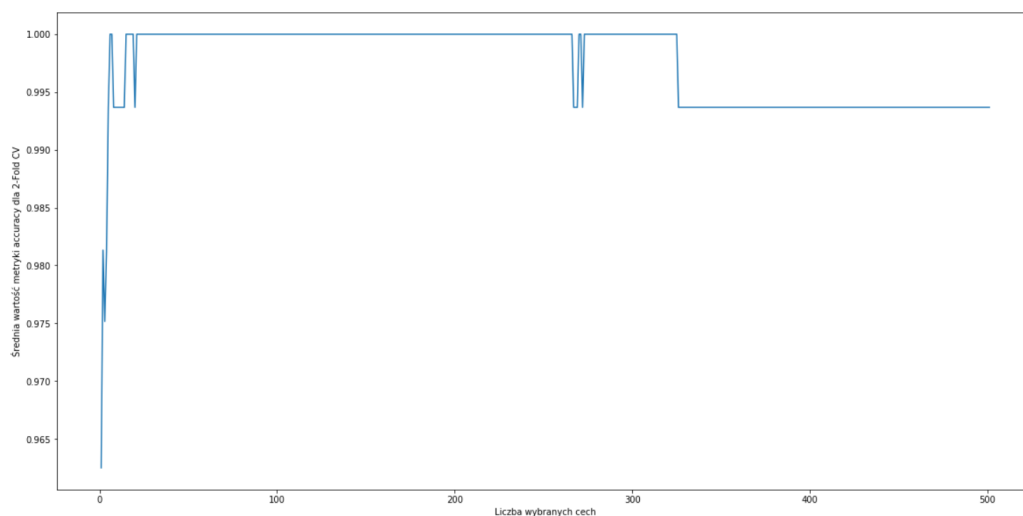
- [1] Guyon, I., Weston, J., Barnhill, S. et al., *Gene Selection for Cancer Classification using Support Vector Machines*, Machine Learning (2002) 46: 389., <https://doi.org/10.1023/A:1012487302797>.



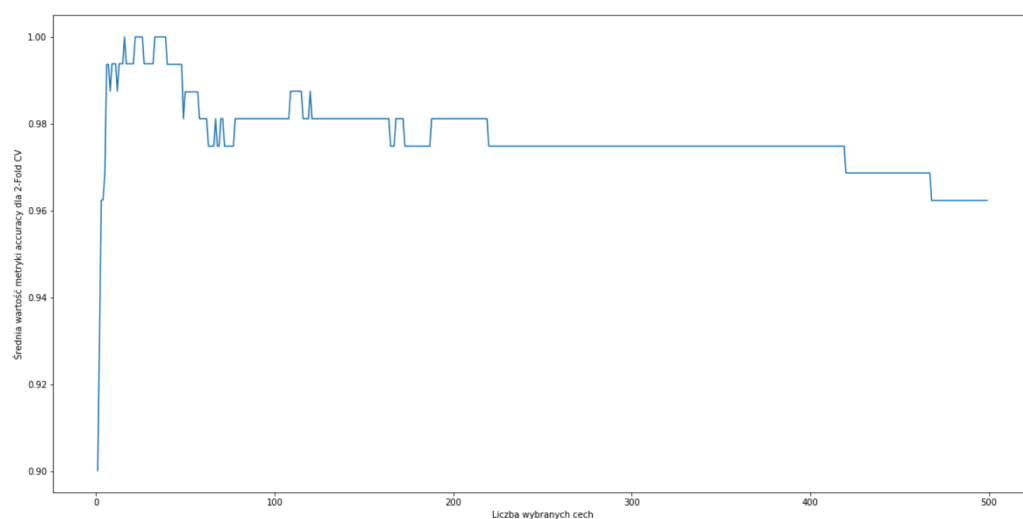
Rysunek 5.1: Wykres średniej wartości metryki accuracy uzyskanej w 2-fold CV w procesie uczenia na zbiorze Leukemia_preprocessed złożonym z 500 najlepszych cech.



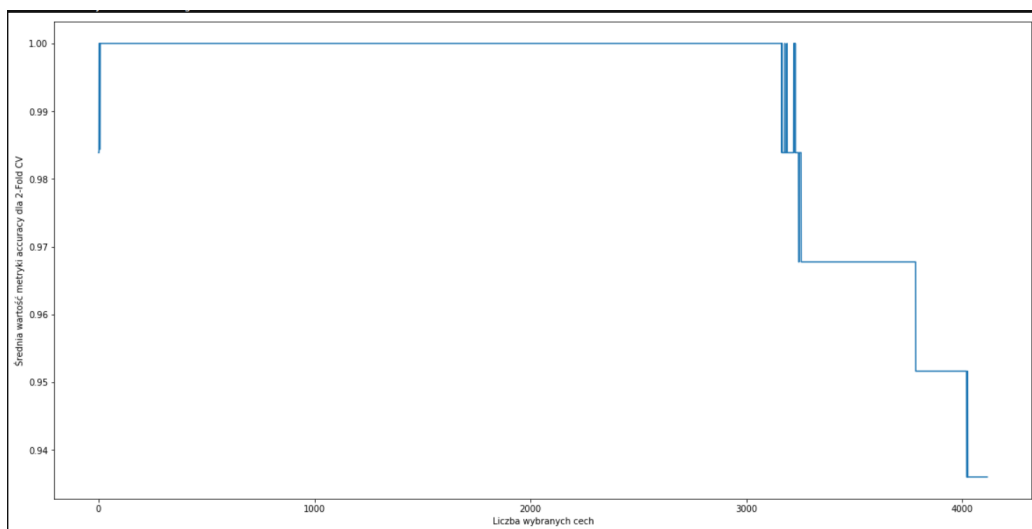
Rysunek 5.2: Wykres średniej wartości metryki accuracy uzyskanej w 2-fold CV w procesie uczenia na zbiorze Leukemia_500 złożonym z 500 najlepszych cech.



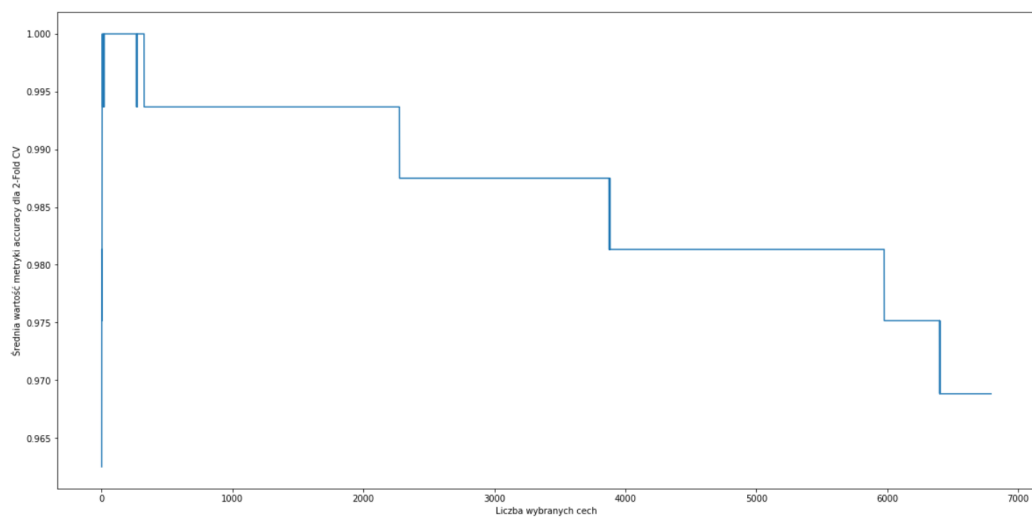
Rysunek 5.3: Wykres średniej wartości metryki accuracy uzyskanej w 2-fold CV w procesie uczenia na zbiorze LungCancer_preprocessed złożonym z 500 najlepszych cech.



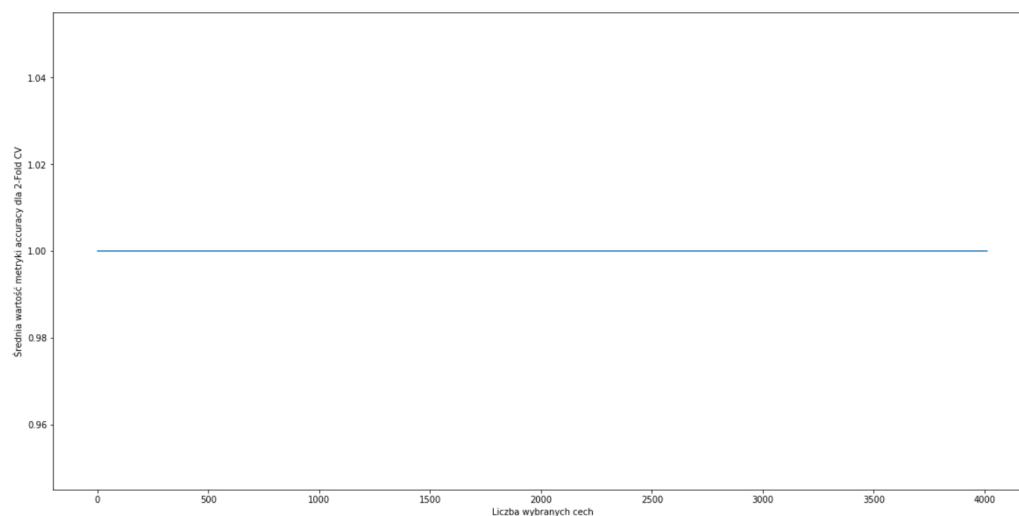
Rysunek 5.4: Wykres średniej wartości metryki accuracy uzyskanej w 2-fold CV w procesie uczenia na zbiorze LungCancer_500 złożonym z 500 najlepszych cech.



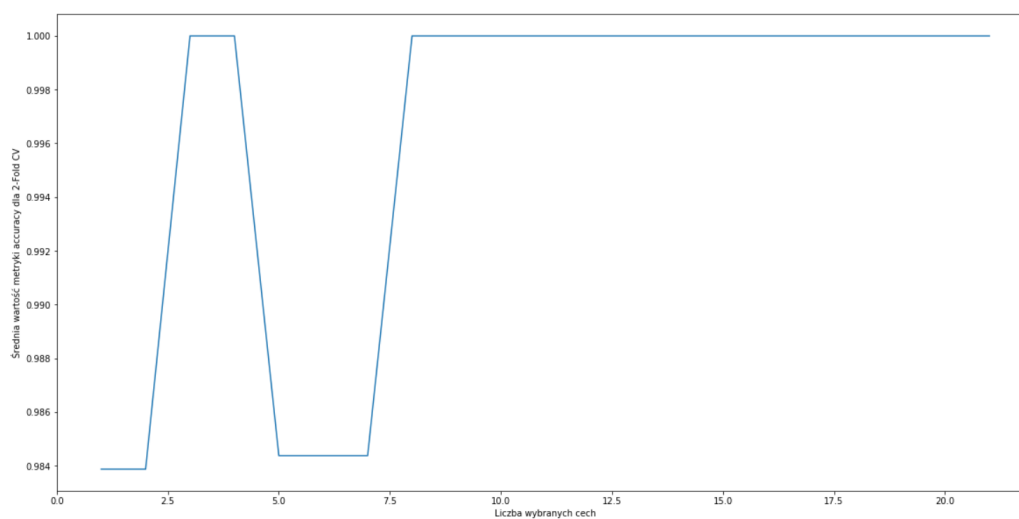
Rysunek 5.5: Wykres średniej wartości metryki accuracy uzyskanej w 2-fold CV w procesie uczenia na zbiorze Leukemia_preprocessed.



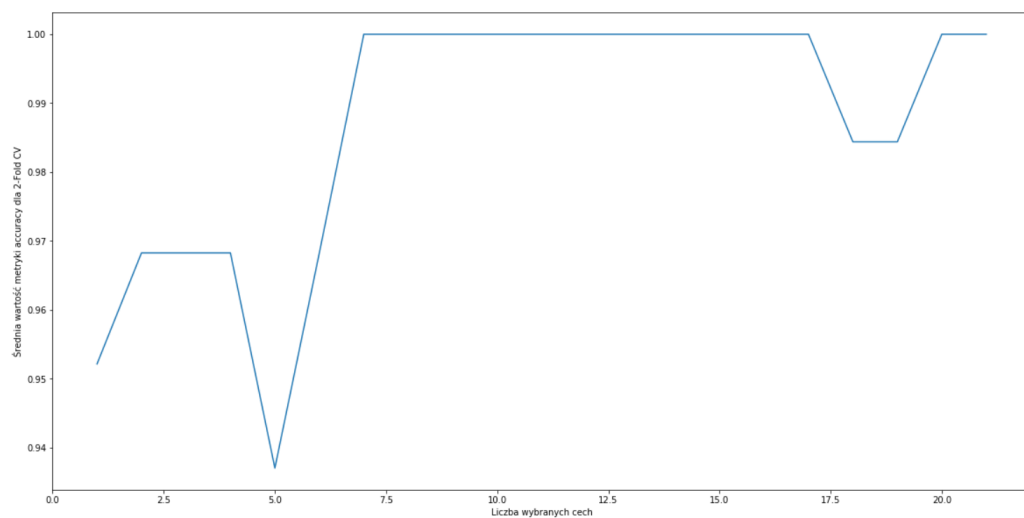
Rysunek 5.6: Wykres średniej wartości metryki accuracy uzyskanej w 2-fold CV w procesie uczenia na zbiorze LungCancer_preprocessed.



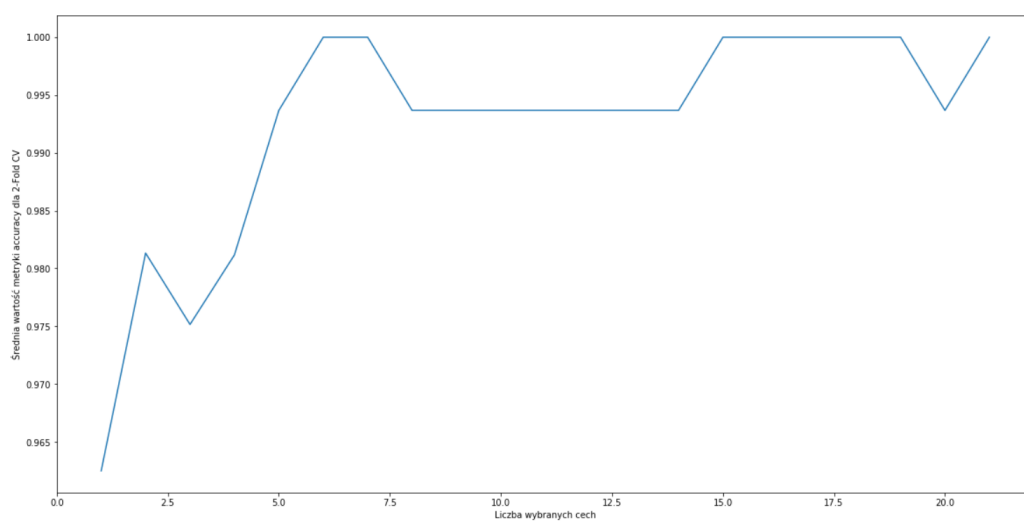
Rysunek 5.7: Wykres średniej wartości metryki accuracy uzyskanej w 2-fold CV w procesie uczenia na zbiorze Lymphoma_preprocessed.



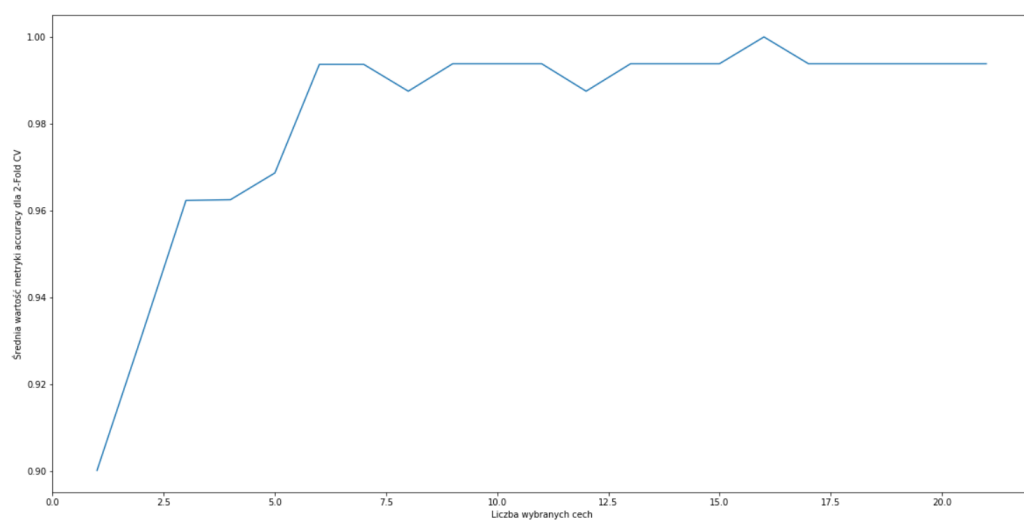
Rysunek 5.8: Wykres średniej wartości metryki accuracy uzyskanej w 2-fold CV w procesie uczenia na zbiorze Leukemia_preprocessed złożonym z 20 najlepszych cech.



Rysunek 5.9: Wykres średniej wartości metryki accuracy uzyskanej w 2-fold CV w procesie uczenia na zbiorze Leukemia_500 złożonym z 20 najlepszych cech.



Rysunek 5.10: Wykres średniej wartości metryki accuracy uzyskanej w 2-fold CV w procesie uczenia na zbiorze LungCancer_preprocessed złożonym z 20 najlepszych cech.



Rysunek 5.11: Wykres średniej wartości metryki accuracy uzyskanej w 2-fold CV w procesie uczenia na zbiorze LungCancer_500 złożonym z 20 najlepszych cech.