**Uczenie maszynowe w bioinformatyce**

**Sprawozdanie z projektu**

Sawilska Ada 300466

Król Maria 284178

1. **Temat projektu**

Tematem projektu jest selekcja genów lub białek różnicujących próbki z dwóch klas przy użyciu liniowego klasyfikatora SVM i metody Recursive Feature Elimination. Uzyskane wyniki zostały porównane z wynikiem selekcji w oparciu o miarę f-score.

1. **Wprowadzenie**
   1. **Selekcja cech**

Selekcja cech polega na wybraniu ze zbioru cech, które spełniają pewne kryteria. Najczęściej wybierane są cechy, których wartości w istotny sposób różnią się między klasami. W przypadku wykonania selekcji cech różnicujących na potrzeby klasyfikacji, celem jest uzyskanie jak najmniejszego zbioru cech, które wspólnie posłużą do zbudowania klasyfikatora [1, 2]. Selekcję stosuje się w celu zmniejszenia wymiarowości problemu oraz uproszczenia modelu.

* 1. **Klasyfikacja SVM**

Metoda wektorów nośnych (ang. *SVM – Support Vector Machines*) jest metodą uczenia maszynowego do rozwiązywania zagadnień klasyfikacji oraz regresji [3]. Klasyfikacja SVM jest realizowana przy pomocy zbioru uczącego, gdzie dane są podzielone na dwie klasy.

Algorytm trenujący znajduje hiperpłaszczyznę, która rozdziela dane z jak największym marginesem zaufania. W przypadku klasyfikatora liniowego hiperpłaszczyznę opisuje się wzorem:

gdzie:

– wektor zawierający wagi wszystkich cech

– macierz danych pochodzących z mikromacierzy

– wyraz wolny (przesunięcie)

W efekcie uczenia uzyskuje się wektor wag wszystkich cech, na podstawie którego można wybrać zbiór najbardziej istotnych cech.

* 1. **Algorytm RFE**

Wybierając jednorazowo zbiór najlepszych cech według określonego kryterium może się okazać nie najlepszym sposobem ze względu na wpływ cech odstających. Można temu zapobiegać poprzez rekurencyjne usuwanie cech o najmniejszym znaczeniu. Jest to tzw. algorytm RFE (ang. *Recursive Feature Elimination*), składający się z trzech kroków:

1. Wytrenowanie klasyfikatora
2. Wyznaczenie rankingu cech według kryterium
3. Usunięcie najmniej znaczącej cechy

Kroki te należy powtarzać aż do osiągnięcia zbioru najlepszych cech o określonym rozmiarze lub do momentu przekroczenia progu kryterium. W celu zredukowania czasu obliczeń możliwe jest usuwanie większej liczby cech w każdej pętli [4], jednak może się to wiązać ze zmniejszeniem dokładności metody.

* 1. **F-score**

Fisher score jest to algorytm powiązany z analizą wariancji. Dla każdej cechy wyliczana jest wartość średnia oraz wariancja. Otrzymane wartości są następnie podstawiane do wzoru.

gdzie:

– suma kwadratów pomiędzy grupami (ang. *Between Sum of Squares*)

– suma kwadratów wewnątrz grup (ang. *Within Sum of Squares*)

– liczba klas

– liczba przypadków

Fisher score używany jest jako kryterium istotności cech do zmniejszania wymiarowości [5]. Jednak algorytm ten często wybiera cechy zbędne jako istotne, co wpływa na wynik klasyfikacji.

1. **Wykonanie projektu**
   1. **Opis danych**

Do niniejszego projektu wykorzystano zbiór danych o nazwie ‘Peptidome2\_240’, pochodzących ze spektrometrii mas. Zbiór ten zawiera wartości odpowiadające ekspresji każdego z 240 peptydów (cech) dla każdego z 60 pacjentów (próbek). Próbki przyporządkowane są do 2 klas, określających stan pacjentów: chory, zdrowy.

* 1. **Działanie programu**

W ramach projektu napisano program, w którym kolejno dokonywana jest selekcja cech za pomocą algorytmu RFE oraz f-score, a następnie porównywane są otrzymane wyniki. W pierwszym kroku importowany jest plik zawierający zbiór danych oraz ustawiane są parametry selekcji, takie jak:

Z – końcowa liczba najważniejszych cech

X – liczba eliminowanych cech w każdej pętli algorytmu RFE

Następnie dokonywana jest selekcja cech za pomocą stworzonej funkcji ‘RFE’. Wykorzystanie algorytmu RFE polega na rekurencyjnym wytrenowaniu klasyfikatora SVM za pomocą gotowej funkcji ‘fitcsvm’, obliczeniu wektora wag oraz usunięciu X cech o najniższych wagach. Wektor wag jest wyznaczany poprzez podniesienie do kwadratu wartości wektora ‘Beta’ zawartego w modelu klasyfikatora SVM. Pętla algorytmu RFE powtarzana jest aż do osiągnięcia ustawionej liczby Z najbardziej istotnych peptydów.

Selekcja cech za pomocą f-score dokonywana jest w funkcji ‘fisher\_score’. Polega ona na obliczeniu miary *f* dla wszystkich cech, a następnie ich posortowaniu ze względu na wartość tej miary. Ostateczny zbiór peptydów wyznaczany jest poprzez wybranie Z cech o największych wartościach f-score.

Na koniec dokonywana jest walidacja krzyżowa obydwu metod funkcją ‘svm\_classifier’. Zbiór danych ograniczony do Z najbardziej istotnych cech jest dzielony za pomocą funkcji ‘cvpartition’ na 3 części o równej lub bardzo zbliżonej liczbie próbek, z zachowaniem równomiernego rozkładu klas wewnątrz podzbiorów. Następnie wymiennie klasyfikator SVM zostaje wytrenowany na dwóch podzbiorach oraz przetestowany na trzecim. Dla każdej iteracji obliczony zostaje błąd klasyfikacji zgodnie ze wzorem:

Na wyjściu funkcji ‘svm\_classifier’ otrzymywana jest wartość średnia z wektora , która odpowiada błędowi klasyfikacji danej metody oraz danemu zbiorowi wcześniej wybranych cech.

* 1. **Wyniki oraz wnioski**

W celu porównania algorytmu RFE oraz f-score wyznaczono błędy klasyfikacji dla kolejno 10, 20, 50, 100 najbardziej istotnych cech (zmienna Z). Wyniki przedstawiono w Tabeli 1.

Tabela 1 Porównanie metody RFE (X=1) oraz f-score

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Liczba wybranych cech (Z) | RFE | | f-score | |
| Błąd klasyfikacji [%] | AUC | Błąd klasyfikacji [%] | AUC |
| 10 | 0.00 | 1 | 5.00 | 1 |
| 20 | 0.00 | 1 | 5.00 | 0.99 |
| 50 | 0.00 | 1 | 1.67 | 1 |
| 100 | 0.00 | 1 | 1.67 | 1 |
| 200 | 0.00 | 1 | 0.00 | 1 |

Wartości błędów klasyfikacji dla metody RFE są znacznie niższe w porównaniu do metody f-score, co oznacza, że selekcja cech za pomocą metody RFE jest znacznie dokładniejsza.

Wyniki selekcji RFE przedstawione powyżej zostały otrzymane dla wartości X=1, czyli w każdej pętli usuwana była tylko 1 cecha. Aby sprawdzić działanie metody dla większej liczby usuwanych cech, ustawiano kolejno wartości 1, 2, 5, 10, 20, 50. Końcową liczbę cech (Z) ustawiono na 50. Otrzymane wyniki przedstawiono w Tabeli 2.

Tabela 2 Porównanie wyników RFE dla różnych wartości usuwanych cech (Z=50)

|  |  |
| --- | --- |
| Liczba usuwanych cech (X) | Błąd klasyfikacji RFE [%] |
| 1 | 0.00 |
| 2 | 0.00 |
| 5 | 0.00 |
| 10 | 0.00 |
| 20 | 0.00 |
| 50 | 0.00 |

Pomimo zwiększenia liczby usuwanych cech, wartości błędów klasyfikacji były zerowe. Świadczy to o bardzo dużej dokładności metody RFE. Jednak na tak dobre wyniki mogło mieć wpływ uprzednie ograniczenie zbioru cech. Dlatego sprawdzono działanie programu również dla rozszerzonego zbioru danych o nazwie ‘Peptidome2\_918’ (Tab. 3 i 4).

Tabela 3 Porównanie metody RFE (X=1) oraz f-score dla większego zbioru danych

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Liczba wybranych cech (Z) | RFE | | f-score | |
| Błąd klasyfikacji [%] | AUC | Błąd klasyfikacji [%] | AUC |
| 10 | 0.00 | 1 | 8.33 | 0.96 |
| 20 | 0.00 | 1 | 1.67 | 0.99 |
| 50 | 0.00 | 1 | 1.67 | 1 |
| 100 | 0.00 | 1 | 1.67 | 1 |
| 200 | 0.00 | 1 | 1.67 | 1 |

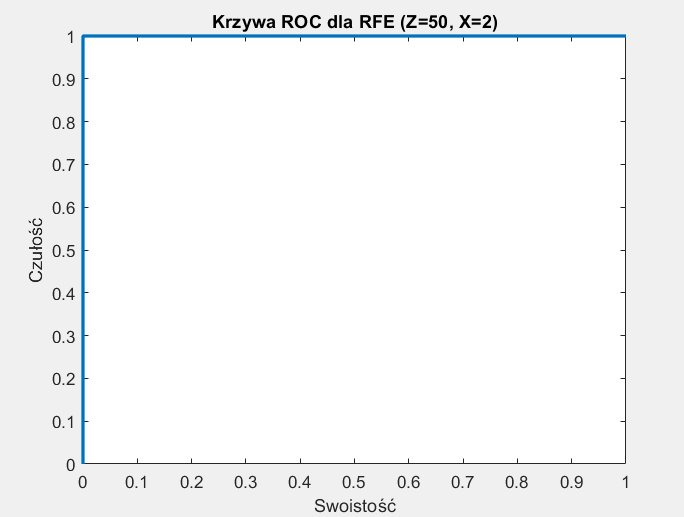
Na podstawie powyższych wyników można zauważyć, że dla rozszerzonego zbioru danych zmniejszyła się dokładność metody f-score, z największym błędem wynoszącym 8.33% dla 10 wybranych cech. Natomiast błędy klasyfikacji po selekcji algorytmem RFE pozostały zerowe.

Tabela 4 Porównanie wyników RFE dla większego zbioru danych (Z=50)

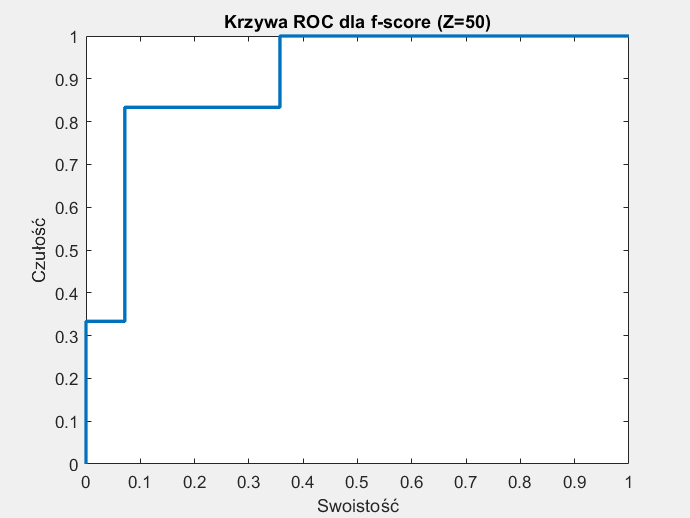
|  |  |
| --- | --- |
| Liczba usuwanych cech (X) | Błąd klasyfikacji RFE [%] |
| 1 | 0.00 |
| 2 | 0.00 |
| 5 | 0.00 |
| 10 | 0.00 |
| 20 | 0.00 |
| 50 | 0.00 |

Wszystkie błędy klasyfikacji wynosiły 0%, zatem również zwiększenie zbioru danych nie wpłynęło na wyniki selekcji cech metodą algorytmu RFE.

Dodatkowo narysowano krzywe ROC dla 50 wybranych cech w celu porównania metod. Otrzymane wykresy przedstawiono poniżej (Rys. 1 i 2).



Rysunek 1 Krzywa ROC dla algorytmu RFE



Rysunek 2 Krzywa ROC dla algorytmu f-score

Krzywa ROC dla algorytmu RFE ma idealny przebieg, co jest zgodne z zerowym błędem klasyfikacji. Natomiast dla algorytmu f-score krzywa ROC pokazuje, że nie jest aż tak dokładny.

Algorytm RFE jest metodą rekurencyjną wykorzystującą klasyfikator SVM. Jest bardziej zaawansowany niż f-score, stąd też tak mały błąd klasyfikacji.

1. **Bibliografia**

[1] Mahendran N, Durai Raj Vincent PM, Srinivasan K and Chang C-Y (2020) *Machine Learning Based Computational Gene Selection Models: A Survey, Performance Evaluation, Open Issues, and Future Research Directions*. *Front. Genet.* 11:603808. doi: 10.3389/fgene.2020.603808

[2] Rubel T., *Selekcja cech różnicujących*, Uczenie maszynowe w informatyce

[3] Goszczyński J., *Klasyfikacja Tekstur Za Pomocą Svm – Maszyny Wektorów Wspierających*, <https://ir.ptir.org/artykuly/pl/88/IR(88)_424_pl.pdf>

[4] Guyon I., Weston J., Barnhill S., Vapnik V., *Gene Selection for Cancer Classification using Support Vector Machines*, Machine Learning, 46, 389–422, 2002

[5] Sun L., Zhang XY., Qian YH. et al. *Joint neighborhood entropy-based gene selection method with fisher score for tumor classification*. Appl Intell 49, 1245–1259 (2019). <https://doi.org/10.1007/s10489-018-1320-1>