

Zastosowanie Informatyki w Medycynie

Mateusz Król

June 13, 2019

Klasyfikacja chorób związanych z anemią przy użyciu maszyny wektorów nośnych

1 Wprowadzenie

Celem projektu było napisanie algorytmu, który posłuży do klasyfikacji osób do poszczególnych chorób. W tej części zbadano choroby związane z anemią. W początkowym etapie zapoznano się z algorytmem maszyny wektorów nośnych. Wskazano jego wady oraz zalety. Następnie dokładnie przeanalizowano zbiór danych dotyczący chorób tj. Wyodrębniono klasy, przeprowadzono selekcję cech a następnie powstał ranking cech. Podążając dalej zaimplementowano algorytm klasyfikujący i testowano jego skuteczność dla różnej liczby cech; poczynając od tej mającej największy wpływ na klasyfikację do danej choroby przez kolejne cechy aż do tej, która miała ówże wpływ najmniejszy. Ostatecznie wyniki zebrano w tabelach i omówiono.

Kolorowa wersja sprawozdania wraz ze wszystkimi plikami składowymi znajduje się w repozytorium pod adresem:

<https://github.com/mat kro96/ZIwM>

2 Opisy narzędzi i badań

2.1 Maszyna wektorów nośnych(Support Vector Machine -SVM)

Algorytm do klasyfikacji binarnej. Cechą wyróżniającą go od innych klasyfikatorów binarnych jest to, że wyszukuje on najlepszego podziału przy pomocy wektorów pomocniczych rozdzielają w ten sposób klasy.

Problem 1. SVM jest klasyfikatorem binarnym, a nasz problem jest wieloklasowy.

Rozwiązanie: W naszych badaniach problem nie był binarny. Postawiono przed nami zadanie klasyfikacji na podstawie 30 symptomów 20 klas. Problem musiał zostać rozwiązany implementacją wielu maszyn wektorów nośnych. Jako algorytm uczący wybrano system jeden-do-wielu. Polega on na tym że bierzemy jedną klasę i oznaczamy ją jako pozytywną a wszystkie pozostałe klasy oznaczamy jako negatywne. Doprowadza to do tego, że otrzymujemy 20 maszyn wektorów nośnych.

Problem 2. Co gdy symptomy pacjenta nie wskazują na żadną dolegliwość?

Rozwiązanie: Jest to zasadnicza wada klasyfikatorów binarnych stosowanych w problemach klasyfikacji wieloklasowej. Przykładowym rozwiązaniem jest przyjęcie progu ufności do naszego algorytmu i jeżeli wartości klasyfikacji są zbyt niskie dla którejkolwiek z klas możemy podejrzewać, że pacjent nie cierpi na żadną z chorób opisanych danymi symptomami. Rozwiązanie to nie zawsze jest bardzo wrażliwe na szumy.

2.2 Podział na cechy

Do podziału cech na te przydatne w klasyfikacji i na te mniej przydatne czy wręcz przeszkadzające posłużono się skojarzeniem wzrostu danej klasy i wzrostu/spadku wartości danej cechy. Zastosowano Macierz korelacji Pearsona, która bada oddziaływanie poszczególnych cech na siebie nawzajem. Wyniki poszczególnych elementów macierzy zawierają się w przedziale $<-1\ 1>$. Gdzie cechy wykazującą się korelacją zbliżoną do wartości -1 są cechami silnie skorelowanymi ujemnie, te blisko wartości 1 wykazują się silną korelacją dodatnią a te bliskie 0 nie wykazują żadnego oddziaływania na siebie nawzajem. Gdy mamy do czynienia z cechami A i B silnie skorelowanymi

ujemnie oznacza, że wzrost cechy A poniesie za sobą spadek cechy B. Gdy mówimy o cechach silnie skorelowanych dodatnio to wzrost cechy A poniesie za sobą wzrost cechy B. Gdy korelacja 2 cech nie wykazuje korelacji oznacza, że nie potrafimy określić jak się zachowa cech B dy cecha A ulegać będzie zwiększeniu,

Problem. Jak wybrać najważniejsze cechy, które będą miały największy wpływ na przynależność obiektu do klasy?

Dzięki macierzy korelacji Pearsona możemy stworzyć ranking cech klasyfikacyjnych. Gdy potraktujemy naszą klasę jako jedną z cech, to możemy badać jej korelację z innymi cechami. W praktyce oznacza to, że gdy nasz algorytm klasyfikuje binarnie, to obserwujemy jakie cechy wykazują silną korelację pozytywną albo negatywną względnie wzrostu cechy chory. Cecha ta przyjmie 2 wartości 0 - Jest chory na daną chorobę 1 - nie jest chory na daną chorobę, a jest chory na inną z chorób. Dla ułatwienia zestawienia cech w odpowiedniej kolejności posłużono się wartością bezwzględną poszczególnych elementów macierzy korelacji Pearsona.

2.3 Metodyka testów

W celu przeprowadzania testów klasyfikacyjnych przynależności obiektów do klas stworzono ranking cech. Powstał on na podstawie macierzy korelacji Pearsona. Następnie zaimplementowano algorytm klasyfikujący i przebadano jego działania na kilku przykładowych zbiorach dostępnych w sieci. Potem podjęto pracę nad klasyfikacją pacjentów na podstawie objawów chorób. W pierwszym etapie używano jedynie cechy najsilniej skorelowanej z naszą klasą, potem stopniowo dokładano kolejne cechy. Założeniem tego działania było wyznaczenie optymalnego doboru liczby cech potrzebnych do poprawnej klasyfikacji pacjenta. Testy były uśrednianie na podstawie 5 przeprowadzonych prób przy zadanej ilości cech. Badanie prowadzono na różnych proporcjach zbioru uczącego do zbioru testowego [0.9 0.1], [0.85 0.15] [0.8 0.2]

3 Dane uczące

3.1 Jakie informacje płyną z naszego zbioru danych

Zbiór danych dotyczący anemii składał się z 20.

nr klasy	nazwa klasy	liczba obiektów
1	Niedokrwistość normocytowa	24
2	niedokrwistość megaloblastyczna	19
3	Niedokrwistość z niedoboru żelaza	27
4	Pierwsza niedokrwistość aplastyczna	18
5	Wtórna niedokrwistość aplastyczna	22
6	Wrodzona sferocytoza	17
7	wrodzona elipocytoza	22
8	wrodzona stomatocytoza	20
9	Akantocytoza	26
10	Niedokrwistość wywołana niedoborem G-6-PD	16
11	Kinaza pirroniowa	24
12	Niedokrwistość śródziemnomorska	21
13	niedokrwistość sierpowatokrwinkowa	17
14	niedokrwistość autoimmunohemolityczna	26
15	polowiczna niedokrwistość immunohemolityczna	18
16	niedokrwistość jatrogenna	17
17	krwotoczna utrata krwi	16
18	Krwotok wywołany ankilostomoza	21
19	krwotok wywołany wrzodem jędra	23
20	krwotok wywołany nadżerką	16

Każda z klas miała 31 cech.

nr cechy	nazwa cechy	rodzaj	wartości przyjmowane
1	Koncentracja Hemoglobiny	wielowartościowa	1,2,3,4,5
2	Liczba erytrocytów	wielowartościowa	1,2,3,4,5
3	Średnia objętość Krwinki	wielowartościowa	1,2,3
4	Średnie stężenie HB w krwince	binarna	1,2
5	Wielkość erytrocytów	wielowartościowa	1,2,3,4,5,6,7
6	Rodzaj erytrocytów	wielowartościowa	1,2,3,4,5,6,7
7	Tkanka siateczkowa	wielowartościowa	1,2,3
8	Szpik kostny	wielowartościowa	1,2,3,4,5
9	Wielkość	binarna	1,2
10	Stosunek Jądro do cytoplazmatyczny	binarna	1,2
11	Rodzaj jądra	wielowartościowa	1,2,3
12	Struktura Chromatyny Jądrowej	wielowartościowa	1,2,3,4
13	Jaderko	binarna	1,2
14	Pasożyty	binarna	1,2
15	Ziarenka Seleza	wielowartościowa	1,2,3,4
16	Poziom żelaza	binarna	1,2
17	Poziom trwałych związków żelaza	binarna	1,2
18	Poziom witaminy B12	binarna	1,2
19	Poziom Kwasu Foliowego	binarna	1,2
20	NIE OPISANO TEJ CECHY		
21	Reakcja	binarna	1,2
22	Reakcja	binarna	1,2
23	Reakcja	binarna	1,2
24	Plec	binarna	1,2
25	Wiek	wielowartościowa	1,2,3,4,5,6
26	Gorączka	binarna	1,2
27	Krwawienie	binarna	1,2
28	Skóra	wielowartościowa	1,2,3,4
29	Wzrost chłonne	binarna	1,2
30	Szmery sercowe	binarna	1,2
31	Wątroba Sledziona	binarna	1,2

Co oznaczają poszczególne wartości przyjmowanych przez cechy można sprawdzić w pliku ane-mia049.pdf

W plikach tekstowych Klasy oraz Cechy zamieszczono opisy zawarte w tabelach. Warto mieć otwarte te pliki gdzieś na uboczu, dlatego że w dalszej części badań i sprawozdania będziemy się posługiwać jedynie cyframi cech i klas a nie ich nazwami

Co możemy stwierdzić patrząc na nasz zbiór danych?

1. Mamy do czynienia ze sporą liczbą klas oraz jeszcze większą liczbą cech.// 2. Jak na liczbę cech opisujących daną chorobę występuje mało obiektów w konkretnych klasach, co na pewno wpływa niekorzystnie na naukę i testy algorytmu, ponieważ w przypadku występowania 16 obiektów może się zdarzyć że walidacja klasyfikatora odbędzie się na podstawie jedynie 3 obiektów, co znacząco wpływa na procent poprawnej/niepoprawnej klasyfikacji.
3. W przypadku cech nie znaleziono opisu dotyczącego cechy numer 20. Postanowiono ją z badań wyłączyć.
4. Zbiory przyjmowanych wartości przez cechy są zarówno binarne jak i wielowartościowe

3.2 Sporządzenie rankinhu cech

Do stworzenia rankingu cech posłużono się macierzą korelacji Pearsona. W jaki sposób owamacierz jest wypełniana opisano krótko w sekcji 2.2 Podział na cechy. Na standardowym wyjściu pojawiała się macierz o rozmiarach 32x32 (równa liczbie cech + numer klasy). Warto wspomnieć, że przed generowaniem macierzy zawsze normowano wartości w przedziale $<0, 1>$. W celu zwizualizowania tej macierzy wykorzystano bibliotekę Seaborn i wykonano mapę ciepła cech.

UWAGA: W testach zakładano tylko, że znaczenie ma wielkość korelacji pomiędzy zmianą klasy z $0 \rightarrow 1$ oraz korelacją cech do tej zmiany. Nie ważne było czy korelacje są dodatnie, czy ujemne. W tym celu nałożono wartość bezwzględną na macierz Pearsona.

Poniżej przedstawiono przykładową macierz Pearsona dla klasy 2

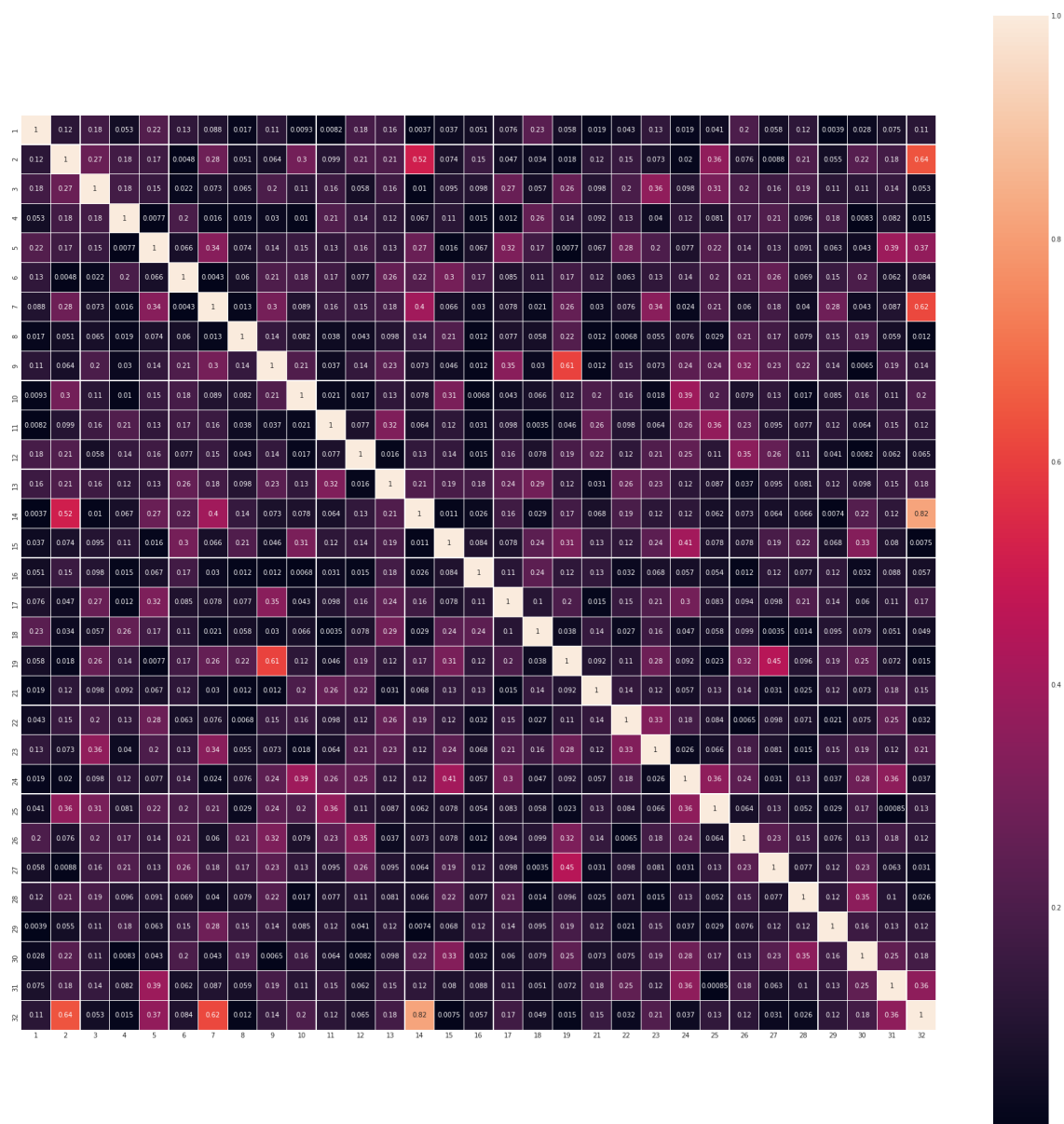


Figure 1: Korelacja poszczególnych cech dla klasy 2, czyli niedokrwistości megaloblastycznej

3.3 Ranking cech

Po przeprowadzeniu podobnych badań jak w sekcji wyżej dla 20 przypadków, które będą służyły jako wzorzec rankingu cech, do nauki i testowania klasyfikatora.

Uwaga: Współczynnik korelacji Pearsona jest doskonałym narzędziem do uwydatniania korelacji wzajemnej cech, lecz każdorazowo trzeba sprawdzać, czy wyniki przez niego wyprowadzane mają sens. Może to doprowadzić do różnego rodzaju nieporozumień. Przykładem takim mogłaby być korelacja między płcią a liczbą zajęć w ciąży (ten przypadek nie dotyczy naszego badania).

Nasz zbiór okazuje się być wiarygodny i moim okiem nie widzę złych powiązań między cechami, jednakże nie potrafię stwierdzić tego na pewno, gdyż nie jestem ekspertem chorób związanych z anemią, ani nie zaczerpnałem eksperckiej wiedzy.

Poniżej zamieszczono w tabelach rankingi poszczególnych cech dla kolejnych klas. Interpretacja wygląda następująco. Gdy korelacja jest wysoka oznacza to, że zmiana klasy z 0→1, czyli rozróżnienie obiektu jako chorego na daną chorobę (tutaj jako 0) oraz chorego na inną chorobą (tutaj jako 1) najlepiej widać po zmianach w cechach o najwyższej korelacji.

Zarówno nazwy cech jak i nazwy poszczególnych klas zebrano w sekcji 3.1 Jakże informacje płyną z naszego zbioru danych.

nr cechy	korelacja z chorobą
11	0.274729
14	0.235893
26	0.225936
3	0.224451
7	0.204664
2	0.178257
9	0.170596
8	0.16899
13	0.1614
6	0.144197
4	0.143305
22	0.112476
30	0.109733
25	0.100977
27	0.096328
5	0.091321
19	0.082164
21	0.069666
31	0.064513
12	0.061192
17	0.054092
15	0.051494
28	0.046513
16	0.045212
10	0.020779
29	0.019816
24	0.011057
23	0.008794
18	0.006348
1	0.000333

UWAGA: Z powodów dużych rozmiarów tabel oraz dużą liczbę tabel postanowiono nie zamieszczać wszystkich rankingów wszystkich cech a jedynie jeden przykładowy jak powyżej. Pozostałe rankingi cech znajdują się w pliku ranking.csv

Top 6 cech dla każdej choroby według współczynnika korelacji Pearsona

nazwa klasy	numery cech najbardziej skorelowanych z klasą
Niedokrwistosc normocytowa	6 7 2 15 26 1
niedokrwistosc megaloblastyczne	11 3 5 6 1 26
Niedokrwistosc z niedoboru zelaza	3 11 8 1 12 5
Pierwsza niedokrwistosc aplastyczna	5 8 11 3 9 6
Wtorna niedokrwistosc aplastyczna	11 14 26 3 7 2
Wrodzona sferocytoza	5 30 11 9 13 3
wrodzona elipocytoza	5 6 13 26 11 3
wrodzona stomatocytoza	11 4 23 3 26 8
Akantocytoza	6 17 11 5 3 16
Niedokrwistosc wywolana niedoborem G-6-PD	2 6 15 12 8 16
Kinaza pirgryoniowa	15 12 13 17 29 6
Niedokrwistosc srodziemnomorska	30 15 12 6 2 10
niedokrwistosc sierpowatokrwinkowa	8 12 15 2 6 4
niedokrwistosc autoimmunohemolityczna	12 21 17 19 18 2
polowiczna niedokrwistosz immunohemolityczna	12 17 24 27 18 10
niedokrwistosc jatrogenna	27 19 17 23 24 10
krwotoczna utrata krwi	23 4 12 26 19 10
Krwotok wywolany anklilostomoza	15 2 6 26 7 22
krwotok wywolany wrzodem jeita	15 6 16 26 19 25
krwotok wywolany nadzgerka	11 5 3 1 6 23

Obserwacje:

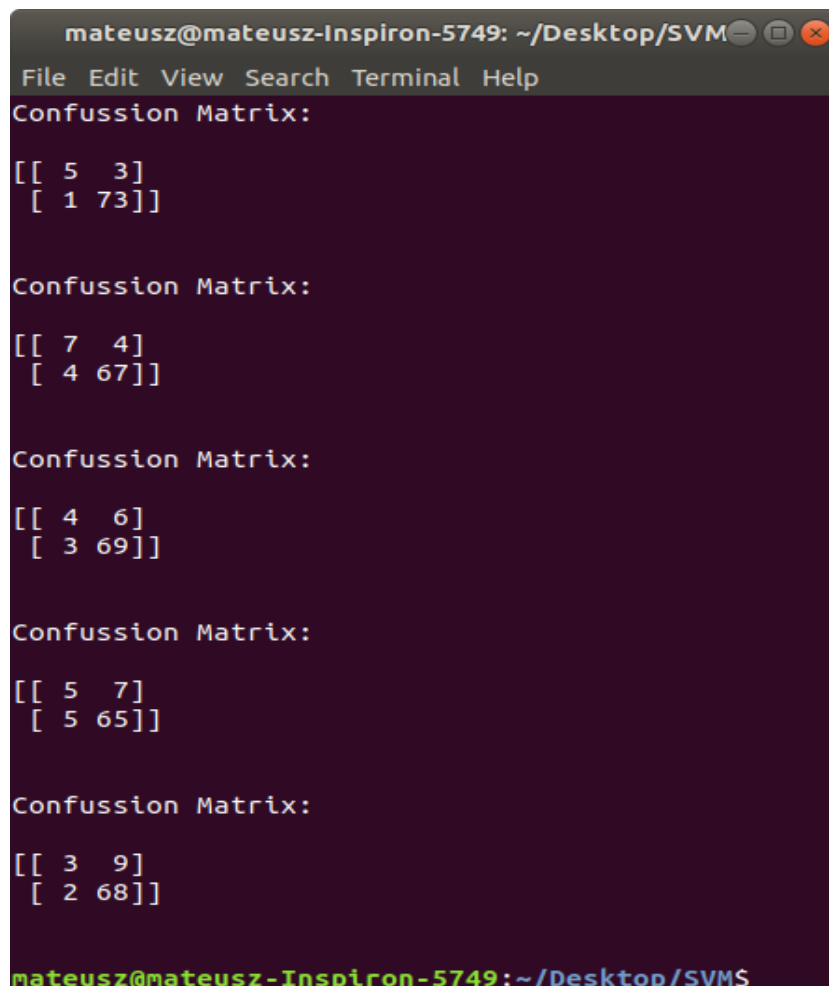
Widzimy na podstawie tabeli pierwszej, że korelacje przy tak wielkiej ilości cech nie są zbyt silne. W tabeli 2 pokazano najlepiej skorelowane cechy do danej choroby. Zbiór na szczęście okazał się być w miarę zróżnicowany i już po kilku cechach można tekie klasy separować

4 Testy

Testy Przeprowadzano w następującej kolejności:

1. Podziel zbiór danych na testujący/uczący 3 różnych proporcjach [0.9/0.1], [0.85/0.15], [0.8/0.2].
2. Wykonaj po pięć testów dla każdej wartości zbiorów uczących/testujących.
3. W pierwszej iteracji liczba cech = 1. Jest to cecha najlepiej skorelowana z daną klasą
4. Wróć do punktu pierwszego zwiększając liczbę cech biorących udział w klasyfikacji o jeden i powtórz algorytm.
5. Warunek wyjścia to wyraźny spadek efektywności algorytmu.

Znając wyniki testów algorytmu możemy stworzyć macierz Pomyłek (ang. Confusion Matrix). Macierz ta jest macierzą 2D o wymiarach $k \times k$, gdzie k to liczba klas. Wiersze macierzy odnoszą się do przewidywanych klas, zaś kolumny odnoszą się do faktycznych klas danych obiektów. Element $[w, k]$ macierzy zawiera liczbę próbek które przewidywano, że posiadają klasę w , ale w rzeczywistości miały klasę k . Celem dobrego klasyfikatora jest sprawienie, że macierz pomyłek będzie diagonalna, co poświadczy, że przewidywana klasa pokrywa się z rzeczywistą klasą dla danego obiektu. Poniżej przedstawiono macierz pomyłek przy nauce i testach taktyką one-vs-all dla podziału zbiorów uczącego i testowego [0.85, 0.15] w przypadku 3 cech.



```
mateusz@mateusz-Inspiron-5749: ~/Desktop/SVM
File Edit View Search Terminal Help
Confussion Matrix:
[[ 5  3]
 [ 1 73]]

Confussion Matrix:
[[ 7  4]
 [ 4 67]]

Confussion Matrix:
[[ 4  6]
 [ 3 69]]

Confussion Matrix:
[[ 5  7]
 [ 5 65]]

Confussion Matrix:
[[ 3  9]
 [ 2 68]]

mateusz@mateusz-Inspiron-5749:~/Desktop/SVM$
```

Figure 2: Wyjście programu, pokazujące macierz klasyfikacji obiektów do danych klas przez Maszynę Wektorów Nośnych na zbiorze testowym

Obserwacje:

Widzimy, że macierz pomyłek jest kwadratowa i niestety nie jest diagonalna, co świadczy, że algorytm nie klasyfikuje naszych obiektów jeszcze poprawnie

Dzięki bibliotece sklearn możemy dodatkowo skorzystać z bogatszego raportu klasyfikacji. Wygląda on następująco.

```
mateusz@mateusz-Inspiron-5749: ~/Desktop/SVM
File Edit View Search Terminal Help
Confussion Matrix:
[[ 6  5]
 [ 1 70]]

classification report:
              precision    recall  f1-score   support

     0       0.86      0.55      0.67        11
     1       0.93      0.99      0.96        71

 accuracy      0.93        82
  macro avg    0.90      0.77      0.81        82
weighted avg    0.92      0.93      0.92        82

Confussion Matrix:
[[ 3  4]
 [ 4 71]]

classification report:
              precision    recall  f1-score   support

     0       0.43      0.43      0.43         7
     1       0.95      0.95      0.95        75

 accuracy      0.90        82
  macro avg    0.69      0.69      0.69        82
weighted avg    0.90      0.90      0.90        82

mateusz@mateusz-Inspiron-5749:~/Desktop/SVM$
```

Figure 3: Wyjście programu, pokazujące pełny raport kwalifikacyjny obiektów do danych klas przez Maszynę Wektorów Nośnych na zbiorze testowym

Na zdjęciu 3 widzimy już dokładniejszy opis klasyfikatora. Do naszych badań i tak nadal najważniejsza będzie kwadratowa macierz pomyłek [2x2] jednak to właśnie z niej można wyciągnąć wiele informacji przedstawionych w pełnym raporcie klasyfikacji takich jak:

1. Precyzja (ang. Precision) to proporcja przypadków faktycznie pozytywnych wśród wszystkich zaklasyfikowanych przez klasyfikator jako pozytywne.
2. Pełność (ang, recall) to proporcja przypadków faktycznie pozytywnych i zaklasyfikowanych przez system jako pozytywne wśród wszystkich faktycznie pozytywnych.

4.1 Zestawienie testów

UWAGA:

W instrukcji postawiono przed nami zadanie odwracania zbiorów uczącego z testującym. Jednak wyniki te były bardzo mało wiarygodne, ponieważ wybrane proporcje nie pozwalały na osiągnięcie odpowiedniej liczby obiektów do nauki algorytmu. Posiadanie 4 obiektów uczących oraz 16 obiektów testowych klasy chory na daną chorobę okazało się niewystarczające. W przypadku gdy proporcje wynosiły "pół na pół" badania również były mało miarodajne.

Poniżej przedstawiono macierze pomyłek dla rosnącej liczby cech weryfikacyjnych. Pozostałe macierze znajdują się pliku maciezeklasyfikacji.csv

Liczba cech	Macierz pomyłek	tendencja pomyłek
1	$\begin{bmatrix} 0 & 15 \\ 0 & 67 \end{bmatrix}$	Brak
2	$\begin{bmatrix} 0 & 13 \\ 0 & 69 \end{bmatrix}$	lekko rosnąca
3	$\begin{bmatrix} 3 & 7 \\ 5 & 67 \end{bmatrix}$	rosnąca
4	$\begin{bmatrix} 4 & 6 \\ 2 & 70 \end{bmatrix}$	rosnąca
5	$\begin{bmatrix} 5 & 6 \\ 2 & 69 \end{bmatrix}$	rosnąca
6	$\begin{bmatrix} 3 & 11 \\ 1 & 77 \end{bmatrix}$	rosnąca
7	$\begin{bmatrix} 6 & 7 \\ 1 & 68 \end{bmatrix}$	rosnąca
8	$\begin{bmatrix} 8 & 4 \\ 0 & 70 \end{bmatrix}$	rosnąca
9	$\begin{bmatrix} 4 & 3 \\ 1 & 74 \end{bmatrix}$	znacznie malejąca
10	$\begin{bmatrix} 2 & 3 \\ 3 & 74 \end{bmatrix}$	rosnąca
11	$\begin{bmatrix} 3 & 3 \\ 2 & 74 \end{bmatrix}$	malejąca

Tabela przedstawia macierze pomyłek dla różnej liczby cech dla klasy Niedokrwistość śródziemnomorska

Liczba cech	Macierz pomyłek	tendencja pomyłek
1	$\begin{bmatrix} 0 & 12 \\ 0 & 68 \end{bmatrix}$	Brak
2	$\begin{bmatrix} 0 & 12 \\ 0 & 68 \end{bmatrix}$	Brak
3	$\begin{bmatrix} 3 & 9 \\ 1 & 67 \end{bmatrix}$	rosnąca
4	$\begin{bmatrix} 6 & 6 \\ 0 & 68 \end{bmatrix}$	rosnąca
5	$\begin{bmatrix} 11 & 1 \\ 0 & 68 \end{bmatrix}$	rosnąca
6	$\begin{bmatrix} 12 & 0 \\ 1 & 67 \end{bmatrix}$	Optymalna
7	$\begin{bmatrix} 6 & 6 \\ 1 & 68 \end{bmatrix}$	Znacznie malejąca
8	$\begin{bmatrix} 8 & 4 \\ 0 & 68 \end{bmatrix}$	malejąca

Tabela przedstawia macierze pomyłek dla różnej liczby cech dla klasy Niedokrwistość śródziemnomorska

5 Wnioski

5.1 Co udało się zrealizować?

1. Problem klasyfikacji objawów do chorób został rozwiązany z dobrymi rezultatami, lecz nie doskonałymi. Pośredni wpływ na badania miała mała liczba obiektów w danych klasach.

2. Cechy posegregowano wraz ze spadającą wartością merytoryczną dla klasyfikatora z dość dobrze. Po raz kolejny nie można było wyodrębnić mocnych korelacji pomiędzy cechami a klasami z powodu małej liczby próbek a dużej liczby cech. Najczęściej liczba cech znacznie przewyższała liczbę obiektów. Nawet po odseparowaniu tych cech, które nie miały dużego wpływu na inne cechy, korelacji nadal nie można było uznać za silną. W literaturze przyjmuje się różny próg korelacji uważanej za silną. Często były to wartości 0.6 albo 0.7

W praktyce może to oznaczać, że choroby poddane klasyfikacji nie są od siebie wcale aż tak różne i ciężko wyodrębnić znaczącą cechę, która pozwoli nam jasno klasyfikować chorego do danej klasy choroby.

3. Przeprowadzić testy dla różnych parametrów:

Zmiennej liczby podziału zbioru na uczący/testowy (3 wartości)

Zmiennej liczby cech znaczących przy klasyfikacji obiektu do klasy (od 7 do 11 cech)

Testy powtarzano pięciokrotnie a wyniki uśredniano

Łącznie wykonano 20 klasyfikatorów x 3 wartości podziału zbioru x zmienna liczba cech (założmy 7) x 5-krotność testów.

w rezultacie otrzymaliśmy ponad 2100 testów.

5.2 Czego nie udało się zrealizować

Przeprowadzić testy dla metody 2-krotnej krzyżowej walidacji danych.

Zaprzestano tych badań, gdyż rezultaty wyprowadzane przez algorytm klasyfikujący były bardzo niewiarygodne. Problemem okazała się zbyt mała liczba obiektów należących do danych klas.

5.3 Co można poprawić

1. Stosowanie algorytmu do klasyfikacji binarnej do problemów wieloklasowych nie zawsze daje dobre rezultaty. W tym przypadku problem ten starano się obejść przez wzięcie jednej choroby (klasy) i porównanie jej ze wszystkimi innymi. Dzięki wcześniej sporządzonemu rankingowi cech udało nam się stwierdzić które cechy są znaczące w odróżnieniu danej choroby (klasy) od pozostałych chorób (klas). W literaturze tę metodę nauki nazywamy nauką jeden-do-wielu (one-vs-all).

Niestety statystycznie okazuje się, że jest ona gorsza od metody jeden-do-jednego), która mówi, aby wytrenować każdą maszynę wektorów nośnych na podstawie poszczególnych innych klas. W praktyce te algorytmy można przedstawić następująco:

one-vs-all W sali mamy jednego mówcę, który wita się z każdym gościem uściskiem ręki. Każdy uścisk symbolizuje jedną maszynę wektorów nośnych.

one-vs-one W sali mamy tylko gości, z których każdy wita się z pozostałym uściskiem ręki. W tym przypadku uścisków jest znacznie więcej.