POLITECHNIKA WROCŁAWSKA
WYDZIAŁ INFORMATYKI I ZARZĄDZANIA

KIERUNEK: INFORMATYKA

Systemy uczące się
Laboratorium

Klasyfikator oparty o Inductive Learning
Algoritm

AUTOR:
Bartosz Kardas

PROWADZĄCY:
dr inż. Paweł Myszkowski

OCENA PRACY:

WROCŁAW, 2017

Spis treści

1. Inductive Learning Algorithm - opis 4
1.1. Algorytm wyznaczania reguł 4
1.2. Klasyfikacja 4

2. Dane testowe - charakterystyka 5
2.1. Seeds 5
2.1.1. Opis 5
2.1.2. Dystrybucja klas 5
2.1.3. Atrybuty 5
2.2. Ecoli 6
2.2.1. Opis 6
2.2.2. Dystrybucja klas 6
2.2.3. Atrybuty 6
2.3. User Knwoledge Modelling 7
2.3.1. Klasy 7
2.3.2. Dystrybucja klas 7
2.3.3. Atrybuty 7

3. Działanie klasyfikatora 8
3.1. Ładowanie i przetwarzanie danych 8
3.2. Budowanie klasyfikatora 8
3.3. Kroswalidacja 8
3.4. Obliczenie statystyk 9

4. Wyniki eksperymentów 10
4.1. Porównanie metod dyskretyzacji 10
4.1.1. Zbiór seeds 10
4.1.2. Zbiór ecoli 11
4.1.3. Zbiór ukm 12
4.2. Porównanie metod kroswalidacji 13
4.3. Porównanie z klasyfikatorem NB 14
4.4. Porównanie czasu wykonania algorytmów 15

5. Wnioski 16
5.1. Dyskretyzacja 16
5.2. Kroswalidacja 16
5.3. Porównanie klasyfikatorów 17
5.4. Porównanie czasu wykonania 17

Spis rysunków

4.1. Wyniki klasyfikatora dla różnych metod dyskretyzacji. Zbiór seeds. 11
4.2. Wyniki klasyfikatora dla różnych metod dyskretyzacji. Zbiór ecoli. 12
4.3. Wyniki klasyfikatora dla różnych metod dyskretyzacji. Zbiór ukm. 13
4.4. Wyniki klasyfikatora dla różnych metod kroswalidacji. 14
4.5. Analiza porównawcza klasyfikatorów. 14
4.6. Porównanie czasowe klasyfikatorów. 15

Spis tabel

2.1. Dystrybucja klas w zbiorze seeds. 5
2.2. Opis atrybutów w zbiorze seeds. 5
2.3. Dystrybucja klas w zbiorze ecoli. 6
2.4. Opis atrybutów w zbiorze seeds. 6
2.5. Dystrybucja klas w zbiorze seeds. 7
2.6. Opis atrybutów w zbiorze ukm. 7

3.1. Rodzaje dyskretyzacji. 8
3.2. Rodzaje uśredniania. 9

4.1. Tabela zbiór seeds. 10
4.2. Tabela zbiór ecoli. 11
4.3. Tabela zbiór ukm. 12
4.4. Tabela kroswalidacja. 13
4.5. Tabela porównawcza. 14
4.6. Tabela czasowa. 15

Rozdział 1

Inductive Learning Algorithm - opis

1.1. Algorytm wyznaczania reguł

Inductive Learning Algorithm (ILA) to algorytm dzięki któremu można wyznaczyć zbiór reguł (forma przedstawiona niżej) opisujących zbiór danych. Na ich podstawie można utworzyć klasyfikator.

$$IF \quad (X_1 = x_1) \wedge (X_2 = x_2) \wedge ... \wedge (X_i = x_i) \quad THEN \quad C = c$$

Znajdowanie reguł polega na podziale zbioru danych na podzbiory zawierające wyłącznie rekordy należące do danej klasy. Następnie w ramach danego podzbioru analizowane są reguły zawierające kolejno $1, 2, ..., n$ atrybutów w każdych z możliwych konfiguracji. Analiza polega na sprawdzeniu czy dana reguła aplikuje się wyłącznie do bieżącego podzbioru. Spośród wszystkich reguł znalezionych w ten sposób reguł wybieramy tę, która odnosi się do największej ilości rekordów i dodajemy ją do zbioru reguł wyjściowych. Warto zaznaczyć, że reguły o większej ilości atrybutów analizujemy, gdy reguły na danym poziomie, nie można zaaplikować do obecnego podzbioru, dzięki temu zapewniona jest możliwie największa prostota reguł.

1.2. Klasyfikacja

Na podstawie wyznaczonych reguł w 1.1 można stworzyć klasyfikator. Jego działanie polega na sprawdzeniu wszystkich reguł w zbiorze, które są prawdziwe dla elementu poddawanego predykcji. Następnie obliczana jest ilość prawdziwych reguł wskazujących na daną klasę. Klasa o największej ilości zgodnych reguł uznawana jest za wynik predykcji. W przypadku braku spełnionych reguł (tzw. unseen exmaples) przyjmowana jest najliczniejsza klasa ze zbioru uczącego.

Rozdział 2

Dane testowe - charakterystyka

2.1. Seeds

2.1.1. Opis

Zbiór zawiera dane dotyczące charakterystki ziaren. Zawiera 3 klasy ziaren: Kama, Rosa and Canadian, po 70 rekordów każdy. Każde z ziaren opisane jest przez 7 atrybutów rzeczywistych. Łącznie 210 rekordów. Każde z ziaren jest opisane, nie posiada brakujących wartości. Zbiór może być wykorzystywany do metod klasyfikacji oraz klasteryzacji.

2.1.2. Dystrybucja klas

Liczności poszczególnych klas zostały przedstawione w tabeli 2.1

Tab. 2.1: Dystrybucja klas w zbiorze seeds.

NAZWA	LICZNOŚĆ
Kama	70
Rosa	70
Canadian	70

2.1.3. Atrybuty

Poniżej 2.2 przedstawiono opis poszczególnych atrybutów.

Tab. 2.2: Opis atrybutów w zbiorze seeds.

NAZWA	RODZAJ	OPIS
Wielkość	REAL	area A
Perymetr	REAL	perimeter P
Kompaktowość	REAL	compactness C = $4 * pi * A/P^2$
Długość	REAL	length of kernel
Szerokość	REAL	width of kernel
Współczynnik asymetrii	REAL	asymmetry coefficient
Długość rowka	REAL	length of kernel groove

2.2. Ecoli

2.2.1. Opis

Zbiór zawiera dane dotyczące miejsca lokalizacji białek. Zawiera 8 klas lokalizacji przedstawione w tabeli 2.3. Każda z lokalizacji opisana jest przez 5 atrybutów rzeczywistych, 2 binarne, 1 kategoryczny, z czego ostatni jest nazwą konkretnej sekwencji (unikalna dla każdego rekordu). Łącznie 336 rekordów. Każdy z nich jest w pełni opisany, nie posiada brakujących wartości. Zbiór może być wykorzystywany do metod klasyfikacji.

2.2.2. Dystrybucja klas

Liczności poszczególnych klas zostały przedstawione w tabeli 2.3

Tab. 2.3: Dystrybucja klas w zbiorze ecoli.

NAZWA	LICZNOŚĆ	OPIS
cp	143	cytoplasm
im	77	inner membrane without signal sequence
pp	52	perisplasm
imU	35	inner membrane, uncleavable signal sequence
om	20	outer membrane
omL	5	outer membrane lipoprotein
imL	2	inner membrane lipoprotein
imS	2	inner membrane, cleavable signal sequence

2.2.3. Atrybuty

Poniżej 2.4 przedstawiono opis poszczególnych atrybutów.

Tab. 2.4: Opis atrybutów w zbiorze seeds.

NAZWA	RODZAJ
name	UNIQUE
mcg	REAL
gvh	REAL
lip	BINARY
chg	BINARY
aac	REAL
alm2	REAL
alm1	REAL

2.3. User Knwoledge Modelling

2.3.1. Klasy

Zbiór zawiera dane dotyczące koorelacji między wynikami testów badanych osób, a czasem spędzonym na naukę. Zawiera 4 klasy oznaczające wynik egzaminu, przedstawione w tabeli 2.5. Każda z lokalizacji opisana jest przez 5 atrybutów rzeczywistych. Łącznie 403 rekordów. Każdy z nich jest w pełni opisany, nie posiada brakujących wartości. Zbiór może być wykorzystywany do metod klasyfikacji oraz klasteryzacji.

2.3.2. Dystrybucja klas

Liczności poszczególnych klas zostały przedstawione w tabeli 2.5

Tab. 2.5: Dystrybucja klas w zbiorze seeds.

NAZWA	LICZNOŚĆ
very-low	50
low	129
middle	122
high	130

2.3.3. Atrybuty

Poniżej 2.6 przedstawiono opis poszczególnych atrybutów.

Tab. 2.6: Opis atrybutów w zbiorze ukm.

NAZWA	RODZAJ	OPIS
STG	REAL	Stopień poświęcenia czasu uczenia na główny cel
SCG	REAL	Stopień powtarzania informacji o głównym celu
STR	REAL	Stopień poświęcenia czasu uczenia na elementy powiązane z głównym celem
LPR	REAL	Wynik egzaminu powiązanego z głównym celem
PEG	REAL	Wynik egzaminu z głównym celem

Rozdział 3

Działanie klasyfikatora

3.1. Ładowanie i przetwarzanie danych

Dane do nauki klasyfikatora zapisane są w formacie tekstowym, z których każda wartość odseparowana jest od siebie przecinkiem. Po załadowaniu wszystkich elementów, rozmieszczane są one losowo w tablicy. Po załadowaniu danych programista może określić jaką metodę dyskretyzacji chciałby użyć. Poniżej przedstawiono typy zaimplementowanych dyskretyzacji.

Tab. 3.1: Rodzaje dyskretyzacji.

NAZWA	KOD	OPIS
Interwałowa	interval	Podział danych równomiernie pomiędzy odpowiednie zakresy liczbowe.
Częstotliwościowa	frequency	Podział danych równomiernie pomiędzy odpowiednie zakresy częstotliwościowe.

3.2. Budowanie klasyfikatora

Po załadowaniu danych następuje budowanie klasyfikatora. Do nauki zbioru danych wykorzystuje się metodę, która implementuje algorytm opisany w 1.2.

Wartości ciągłe przed budową klasyfikatora muszą zostanć poddane dyskretyzacji zgodnie z wybraną metodą, opisaną w 3.1. Ze względu na dużą złożoność obliczeniową, ilość przedziałów dyskretyzacji nie powinna być zbyt duża.

3.3. Kroswalidacja

Zaimplementowane zostały dwa sposoby wykonywania kroswalidacji:

- K-krotna walidacja krzyżowa
- Kroswalidacja stratyfikowana

Pierwsza z nich dzieli zbiór wejściowy na K części, w kolejnych iteracjach każda z nich jest zbiorem testowym, a pozostałe części zbiorem uczącym.

Druga z nich to połączenie powyższego z warunkiem zapewniającym, że w każdej części znajdzie się proporcjonalna ilość rekordów danej klasy co w zbiorze wejściowym.

3.4. Obliczenie statystyk

W celu porównania klasyfikatorów między sobą oraz oceny ogólnej charakterystyki dla danego zbioru danych, zaimplementowano cztery statystyki:

- precision
- recall
- accuracy
- fscore

Ich uśrednianie wynikające z zastosowania kroswalidacji zostało zaimplementowane na 3 sposoby przedstawione w tabeli 3.2

Tab. 3.2: Rodzaje uśredniania.

NAZWA	KOD	OPIS
Arytmetyczna	u	Średnia arytmetyczna wskaźników dla każdej klasy.
Ważona	w	Średnia ważona wskaźników dla każdej klasy. Wagi proporcjonalne do wystąpień klas w zbiorze.
Globalna	g	Globalne obliczanie statystyk z macierzy konfuzji.

Rozdział 4

Wyniki eksperymentów

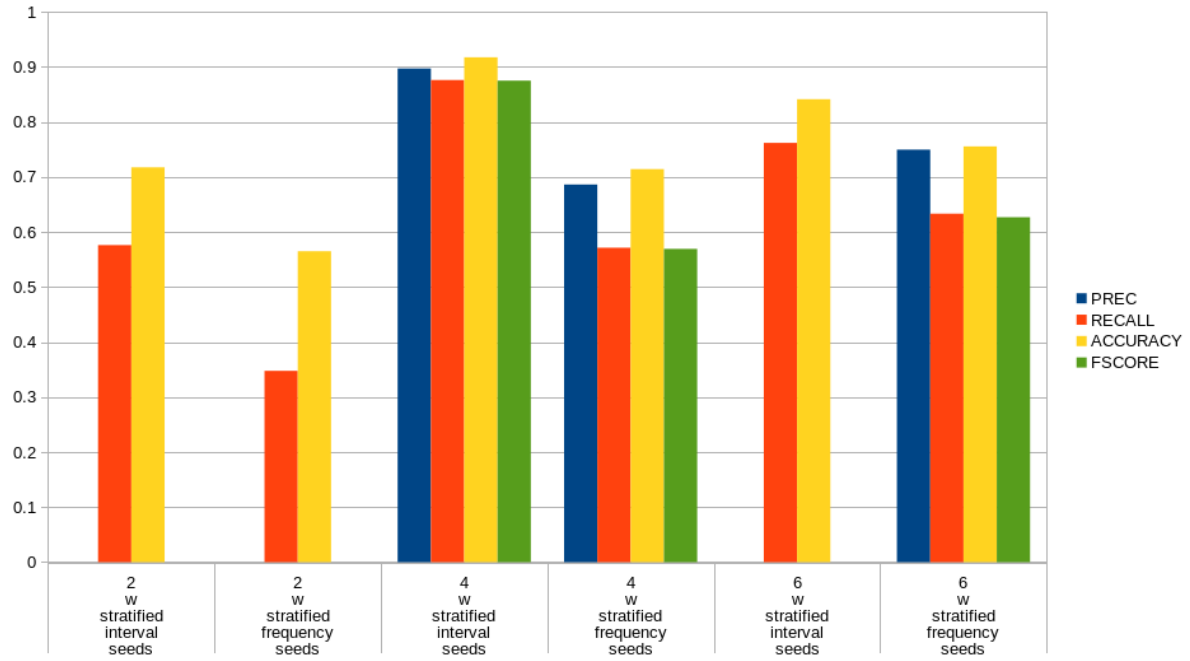
4.1. Porównanie metod dyskretyzacji

4.1.1. Zbiór seeds

Poniżej przedstawiono porównanie metod dyskretyzacji w wynikach poszczególnych klasyfikatorów. Wyniki przedstawiają dane obliczone przy użyciu kroswalidacji stratyfikowanej oraz uśredniania ważonego.

Tab. 4.1: Tabela zbiór seeds.

METHOD	CROSS	SIZE	PREC	RECALL	ACCURACY	FSCORE
interval	stratified	2	nan	0.576	0.717	nan
frequency	stratified	2	nan	0.347	0.565	nan
interval	stratified	4	0.897	0.876	0.917	0.875
frequency	stratified	4	0.686	0.571	0.714	0.569
interval	stratified	6	nan	0.761	0.841	nan
frequency	stratified	6	0.749	0.633	0.755	0.626



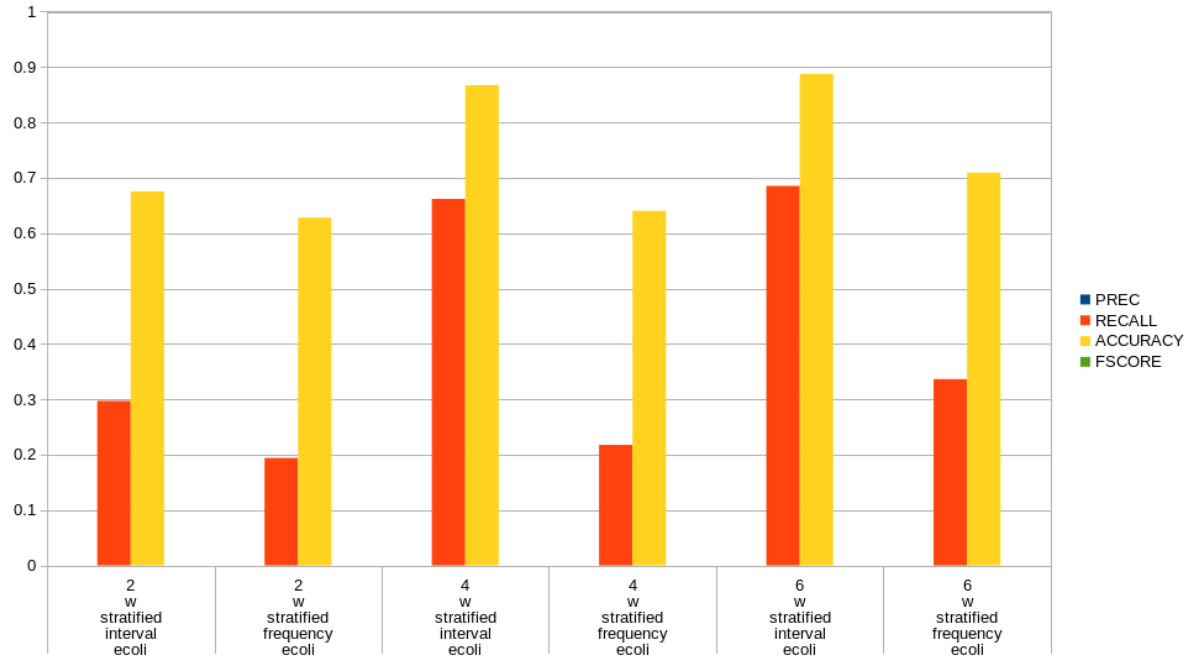
Rys. 4.1: Wyniki klasyfikatora dla różnych metod dyskretyzacji. Zbiór seeds.

4.1.2. Zbiór ecoli

Poniżej przedstawiono porównanie metod dyskretyzacji w wynikach poszczególnych klasyfikatorów. Wyniki przedstawiają dane obliczone przy użyciu kroswalidacji stratyfikowanej oraz uśredniania ważonego.

Tab. 4.2: Tabela zbiór ecoli.

METHOD	SIZE	PREC	RECALL	ACCURACY	FSCORE
interval	2	nan	0.297	0.675	nan
frequency	2	nan	0.194	0.628	nan
interval	4	nan	0.662	0.867	nan
frequency	4	nan	0.218	0.640	nan
interval	6	nan	0.685	0.887	nan
frequency	6	nan	0.336	0.709	nan



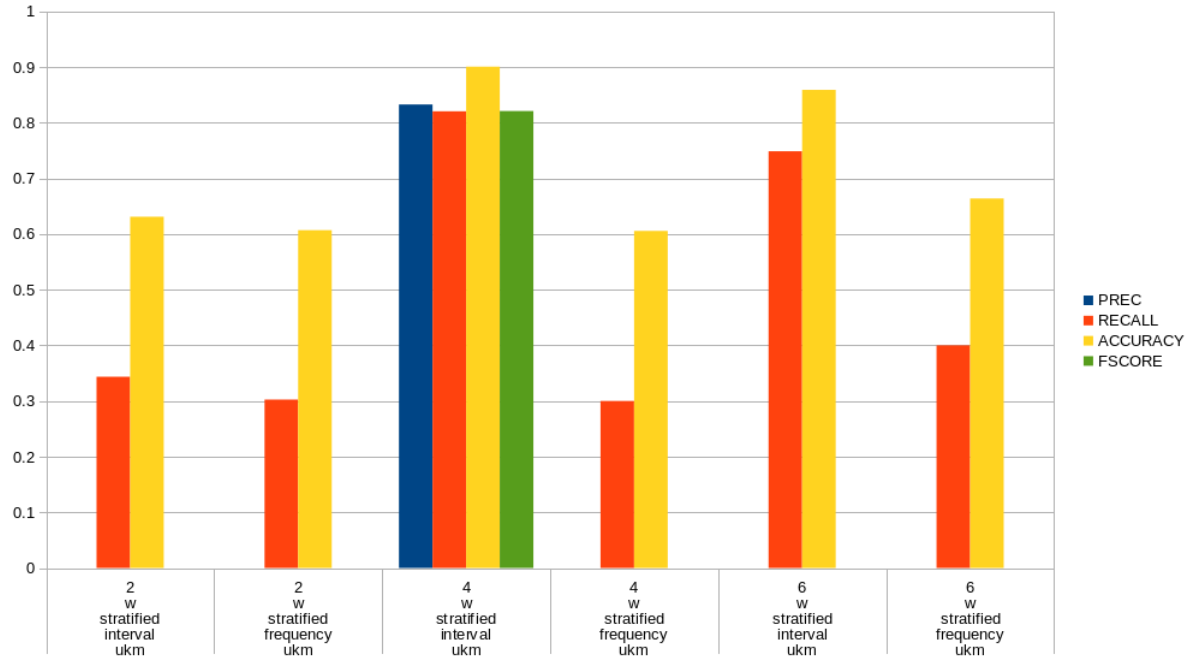
Rys. 4.2: Wyniki klasyfikatora dla różnych metod dyskretyzacji. Zbiór ecoli.

4.1.3. Zbiór ukm

Poniżej przedstawiono porównanie metod dyskretyzacji w wynikach poszczególnych klasyfikatorów. Wyniki przedstawiają dane obliczone przy użyciu kroswalidacji stratyfikowanej oraz uśredniania ważonego.

Tab. 4.3: Tabela zbiór ukm.

METHOD	SIZE	PREC	RECALL	ACCURACY	FSCORE
interval	2	nan	0.344	0.631	nan
frequency	2	nan	0.303	0.607	nan
interval	4	0.833	0.821	0.901	0.821
frequency	4	nan	0.300	0.606	nan
interval	6	nan	0.749	0.859	nan
frequency	6	nan	0.400	0.664	nan



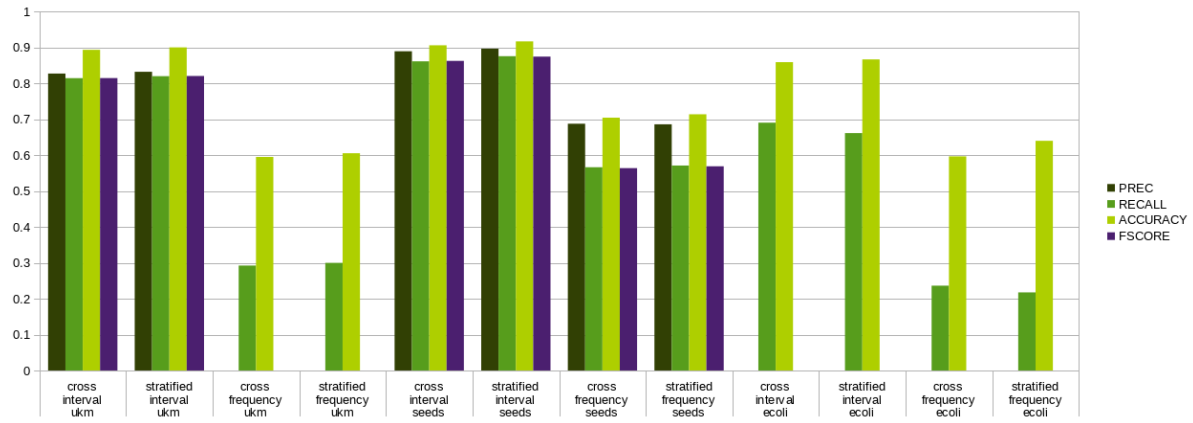
Rys. 4.3: Wyniki klasyfikatora dla różnych metod dyskretyzacji. Zbiór ukm.

4.2. Porównanie metod kroswalidacji

Poniżej przedstawiono wyniki klasyfikatorów obliczone różnymi typami kroswalidacji. Wyniki klasyfikatorów uśredniono za pomocą średniej ważonej.

Tab. 4.4: Tabela kroswalidacja.

NAME	METHOD	CROSS	SIZE	PREC	RECALL	ACCURACY	FSCORE
ukm	interval	cross	4	0.828	0.815	0.894	0.815
ukm	interval	stratified	4	0.833	0.821	0.901	0.821
ukm	frequency	cross	4	nan	0.293	0.595	nan
ukm	frequency	stratified	4	nan	0.300	0.606	nan
seeds	interval	cross	4	0.890	0.862	0.907	0.863
seeds	interval	stratified	4	0.897	0.876	0.917	0.875
seeds	frequency	cross	4	0.688	0.567	0.705	0.564
seeds	frequency	stratified	4	0.686	0.571	0.714	0.569
ecoli	interval	cross	4	nan	0.691	0.860	nan
ecoli	interval	stratified	4	nan	0.662	0.867	nan
ecoli	frequency	cross	4	nan	0.236	0.597	nan
ecoli	frequency	stratified	4	nan	0.218	0.640	nan



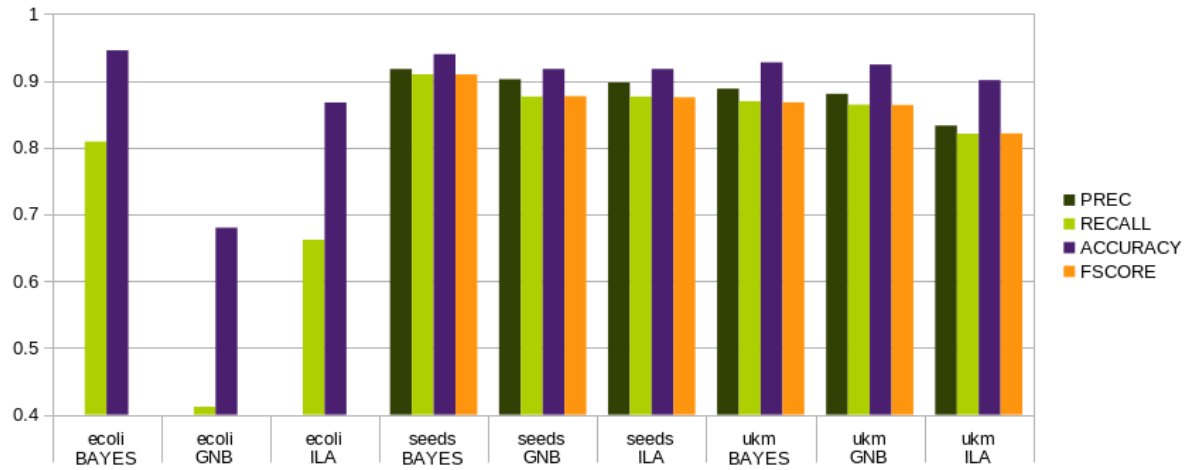
Rys. 4.4: Wyniki klasyfikatora dla różnych metod kroswalidacji.

4.3. Porównanie z klasyfikatorem NB

Poniżej przedstawiono wyniki różnych klasyfikatorów. BAYES - standardowa implementacja naiwnego klasyfikatora bayesowskiego. GNB - naiwny klasyfikator bayesowski opierający się o wyliczanie prawdopodobieństwa z rozkładu normalnego. ILA - klasyfikator oparty na regułach. Dodatkowo wybrano jedynie najlepsze klasyfikatory spośród wszystkich wyników.

Tab. 4.5: Tabela porównawcza.

C	NAME	METHOD	CROSS	A	S	PREC	RECALL	ACCURACY	FSCORE
BAYES	ecoli	frequency	stratified	w	6	nan	0.809	0.945	nan
GNB	ecoli	norm	stratified	w	-	nan	0.412	0.680	nan
ILA	ecoli	interval	stratified	w	4	nan	0.662	0.867	nan
BAYES	seeds	interval	stratified	w	11	0.917	0.910	0.940	0.909
GNB	seeds	norm	stratified	w	-	0.902	0.876	0.917	0.877
ILA	seeds	interval	stratified	w	4	0.897	0.876	0.917	0.875
BAYES	ukm	interval	stratified	w	7	0.888	0.869	0.927	0.867
GNB	ukm	norm	stratified	w	-	0.880	0.864	0.924	0.864
ILA	ukm	interval	stratified	w	4	0.833	0.821	0.901	0.821



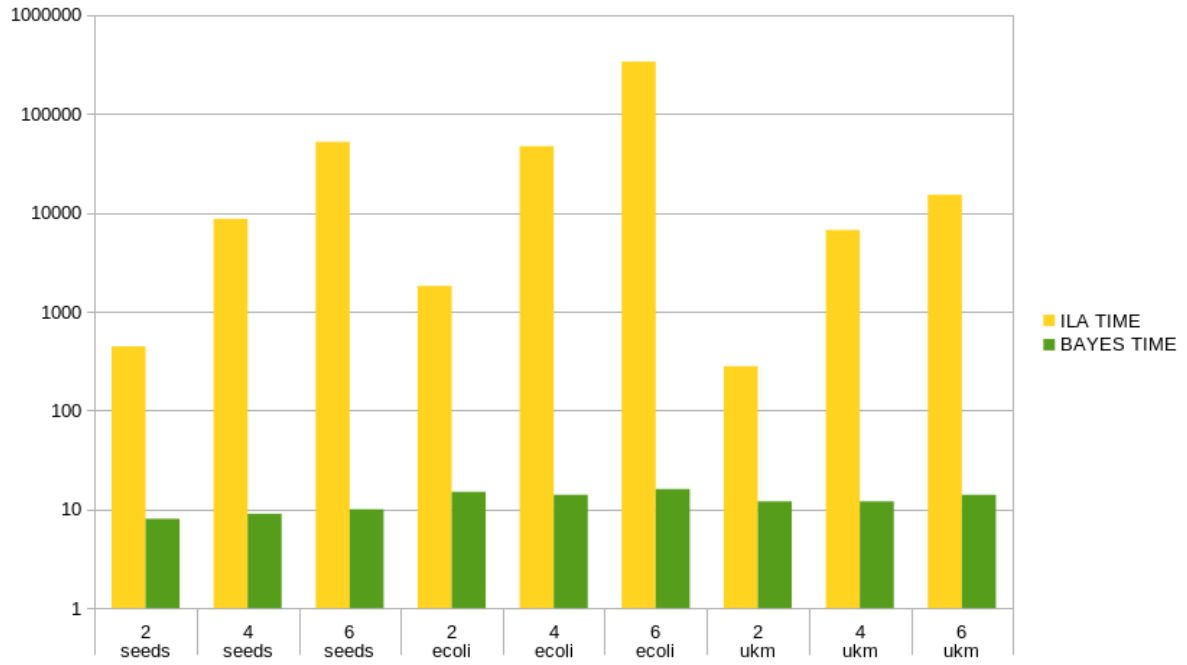
Rys. 4.5: Analiza porównawcza klasyfikatorów.

4.4. Porównanie czasu wykonania algorytmów

Poniżej przedstawiono czas wykonania dziesięciokrotnej kroswalidacji stratyfikowanej dla różnych klasyfikatorów. Wartości w tabeli podane są w milisekundach.

Tab. 4.6: Tabela czasowa.

NAME	SIZE	ILA TIME [ms]	BAYES TIME [ms]
seeds	2	445	8
seeds	4	8670	9
seeds	6	52153	10
ecoli	2	1818	15
ecoli	4	47064	14
ecoli	6	337903	16
ukm	2	280	12
ukm	4	6693	12
ukm	6	15191	14



Rys. 4.6: Porównanie czasowe klasyfikatorów.

Rozdział 5

Wnioski

5.1. Dyskretyzacja

Dla każdego analizowanego zbioru można zauważyć, że dyskretyzacja interwałowa charakteryzowała się znacznie lepszą skutecznością niż dyskretyzacja częstotliwościowa. Wynik ten wskazuje na fakt, że podział zbioru wartości na podzbioru o różnej ilości elementów (metoda interwałowa) pozwala na łatwiejsze znalezienie reguł opisujących zbiór danych co w konsekwencji prowadzi do zbudowania lepszego klasyfikatora. Dla zbioru seeds oraz ukm, najlepszym stopniem dyskretyzacji okazał się podział pomiędzy 4 przedziały, w wyniku czego osiągnięto dokładność powyżej 0.8. W przypadku zbioru ecoli lepszy wynik uzyskano przy podziale na 6 przedziałów (dokładność niespełna 0.7).

5.2. Kroswalidacja

Do porównania wyników kroswalidacji użyto najlepszych wyników klasyfikatorów dla danej metody dyskretyzacji i ilość przedziałów. Niezależnie od wyboru metody, klasyfikatory uzyskały bardzo zbliżone wyniki. Co świadczyć może o tym, że generacja reguł ze zbioru algorytmem ILA jest dosyć odporna na dystrybucję klas w zbiorze testowym.

Warto zauważyć, że w przypadku zwykłego rodzaju kroswalidacji możliwe jest wystąpienie wartości 'nan' dla recall, co dla kroswalidacji stratyfikowanej jest niemożliwe. Dzieje się tak dlatego, gdyż kroswalidacja stratyfikowana zapewnia równomierny rozkład atrybutów dla zbiorów testowych oraz przynajmniej 1 wystąpienie każdej z klas.

5.3. Porównanie klasyfikatorów

Dla zbiorów seeds oraz ukm klasyfikator oparty na algorytmie ILA wykazał się gorszą skutecznością niż naiwne klasyfikatory bayesowskie (nominalny oraz gaussowski). O ile w przypadku zbioru seeds różnica między gaussowskim klasyfikatorem, a klasyfikatorem ILA jest nieznaczna, to w przypadku zbioru ukm, różnice te są już wyraźnie większe, rzędu kilku punktów procentowych. W zbiorze ecoli odnotowano natomiast znaczą poprawę dokładności klasyfikacji metodą ILA niż dla metody gaussowskiej. Taki wynik świadczyć może o braku zachowania rozkładu gaussa w tym zbiorze testowym.

5.4. Porównanie czasu wykonania

Klasyfikatory wykorzystujące metody bayesowskie wykazały się bardzo małym czasem wykonania w porównaniu do klasyfikatora ILA. Warto również zauważyć, że ich czas wykonania jest w bardzo małym stopniu zależny od ilości przedziałów dyskretyzacji. Dla klasyfikatora ILA wyraźnie widać zależność wykładniczą względem ilości przedziałów dyskretyzacji. Warto również zauważyć, że złożoność obliczeniowa ILA wzrasta wraz ze wzrostem ilości atrybutów w zbiorze (15 sekund dla zbioru ukm, wobec 337 sekund w ecoli - odpowiednio 4 i 8 atrybutów).