POLITECHNIKA WROCŁAWSKA
WYDZIAŁ INFORMATYKI I ZARZĄDZANIA

KIERUNEK: INFORMATYKA

Systemy uczące się
Laboratorium

Algorytm C4.5

AUTOR:
Bartosz Kardas

PROWADZĄCY:
dr inż. Paweł Myszkowski

OCENA PRACY:

WROCŁAW, 2017

Spis treści

Spis rysunków

Spis tabel

Rozdział 1

Opis algorytmu C4.5

1.1. Algorytm budowania drzewa

Algorytm C4.5 jest opisuje metodę budowy drzewa decyzyjnego. Jest to rozszerzenie algorytmu ID3, który również był autorstwa Rossa Quinlana. Stał się bardzo popularny po publikacji w której znalazł się w pierwszej 10 algorytmów związanych z data mining.

Algorytm rekurencyjnie buduje drzewo decyzyjne na podstawie zysku informacyjnego w danym węźle (analogicznie do algorytmu ID3). To znaczy, że dla danego węzła analizuje każdy z atrybutów i wybiera ten, którego współczynnik zysku informacyjnego jest największy. Następnie dokonywany jest podział podzbioru według wartości danego atrybutu. Warunkami stopu w rekurencji mogą być następujące przypadki:

- Wszystkie elementy należą do jednej klasy
- Przeanalizowano wszystkie atrybuty, a podzbiór jest nadal niepusty. W tym przypadku liść oznaczany jest klasą najczęściej występującego elementu w podzbiorze.
- Brak elementów dla danej wartości atrybutu. Liść jako najczęstszy element z podzbioru rodzica.

1.2. Usprawnienia względem ID3

Algorytm C4.5 posiada szereg usprawnień względem ID3. Pierwszym z nich jest akceptacja zarówno wartości kategorycznych jak i ciągłych wartości liczbowych. Realizowane jest to poprzez podział wartości na dwie części (mniejsze niż i większe lub równe niż).

Kolejną różnicą względem ID3 jest inna metoda wyboru atrybutu w danym węźle. W ID3 był to zysk informacyjny (information gain), w C4.5 jest to współczynnik zysku informacyjnego (ang. gain ratio). Zaletą współczynnika jest mniejszy redukcja biasu oraz większa odporność na dużą ilość wartości danego atrybutu.

W algorytmie C4.5 zastosowano również prunning. Polega on na modyfikacji struktury drzewa decyzyjnego tak, aby otrzymać najlepsze wyniki. Istnieją dwa etapy prunningu:

- postprunning - eliminacja niewiarygodnych części zbudowanego drzewa
- preprunning - zatrzymanie rozbudowy gałęzi, gdy informacja staje się niewiarygodna

Postprunning opiera się na dwóch operacjach:

- subtree replacement - wybór poddrzewa i zamiana go na liść
- subtree rising - usunięcie wierzchołka i redystrybucja instancji

Warto zauważyć, że replacement jest znacznie szybszą operacją niż rising.

Preprunning natomiast polega na zatrzymaniu rozbudowy danego poddrzewa w określonym momencie. Do wyboru dobrego momentu służą najczęściej odpowiednie testy statystyczne. W tym podejściu istnieje ryzyko, żę proces budowy poddrzewa zatrzyma się zbyt wcześnie ze względu na zbyt małe podobieństwo między wartością atrybutów a wartością klasy.

W przeciwieństwie do algorytmu ID3, C4.5 został wyposażony w mechanizm radzenia sobie z brakującymi danymi. Realizowane jest to poprzez oznaczenie dowolnej wartości atrybutu ('?') dla danego wierzchołka.

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

Wyniki eksperymentów

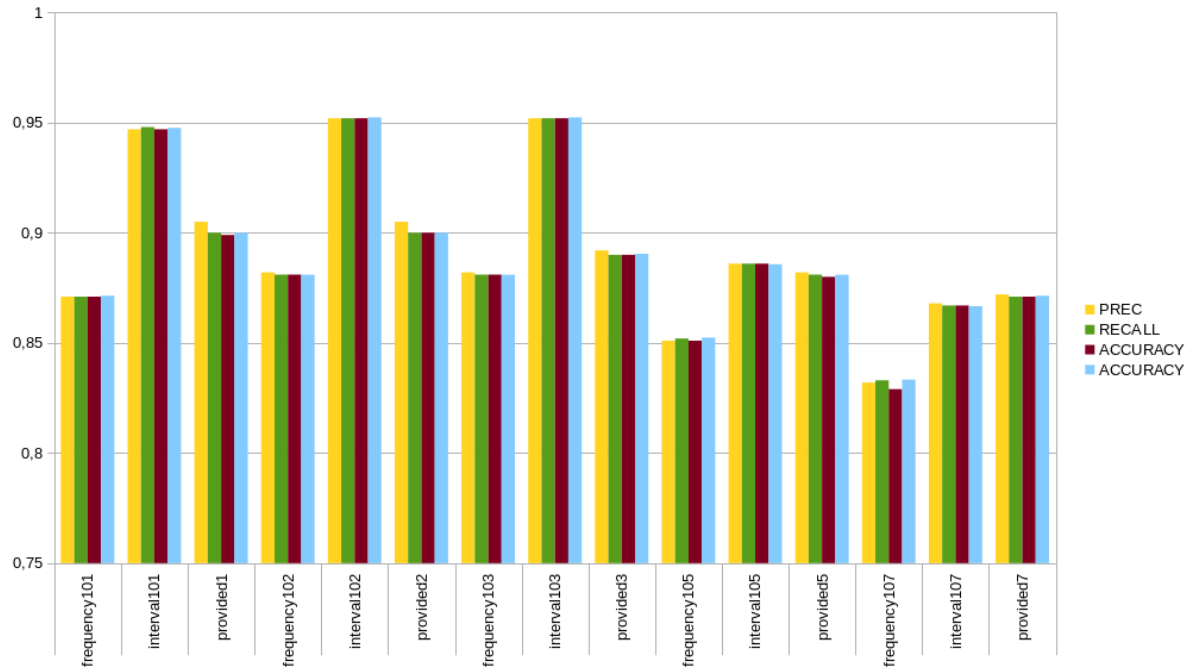
3.1. Porównanie metod dyskretyzacji

3.1.1. Zbiór seeds

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

Tab. 3.1: Tabela zbiór seeds.

METHOD	LEAF	PREC	RECALL	FSCORE	ACCURACY
frequency10	1	0,871	0,871	0,871	0,871
interval10	1	0,947	0,948	0,947	0,948
provided	1	0,905	0,900	0,899	0,900
frequency10	2	0,882	0,881	0,881	0,881
interval10	2	0,952	0,952	0,952	0,952
provided	2	0,905	0,900	0,900	0,900
frequency10	3	0,882	0,881	0,881	0,881
interval10	3	0,952	0,952	0,952	0,952
provided	3	0,892	0,890	0,890	0,890
frequency10	5	0,851	0,852	0,851	0,852
interval10	5	0,886	0,886	0,886	0,886
provided	5	0,882	0,881	0,880	0,881
frequency10	7	0,832	0,833	0,829	0,833
interval10	7	0,868	0,867	0,867	0,867
provided	7	0,872	0,871	0,871	0,871



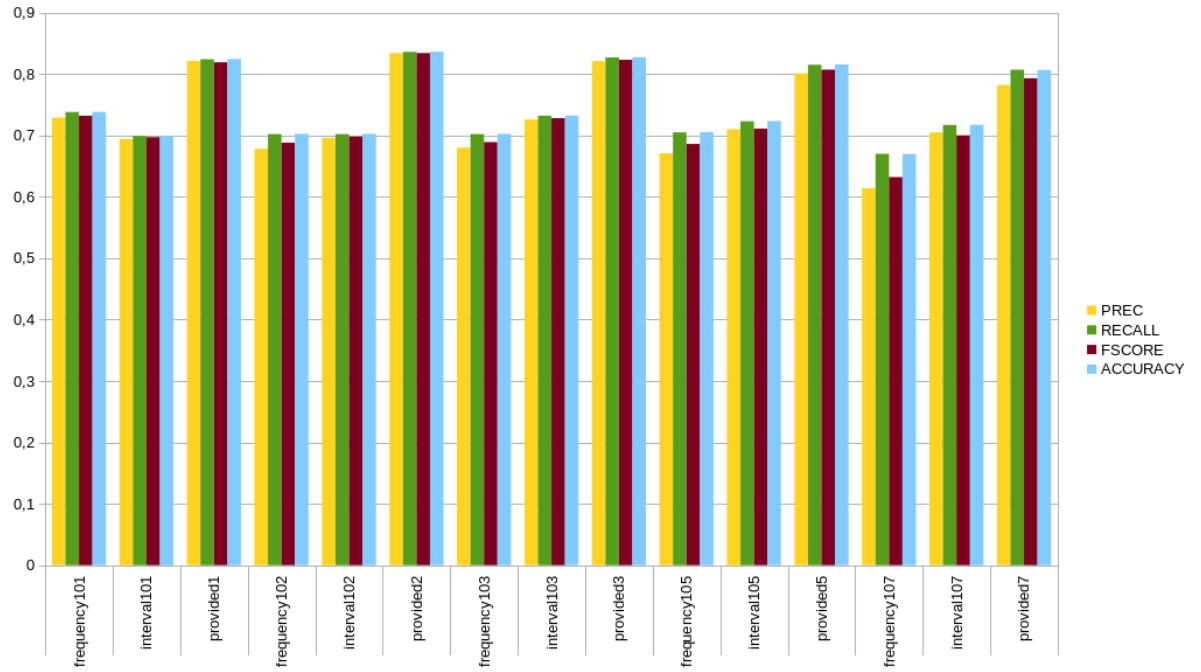
Rys. 3.1: Wyniki klasyfikatora dla różnych metod dyskretyzacji. Zbiór seeds.

3.1.2. Zbiór ecoli

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

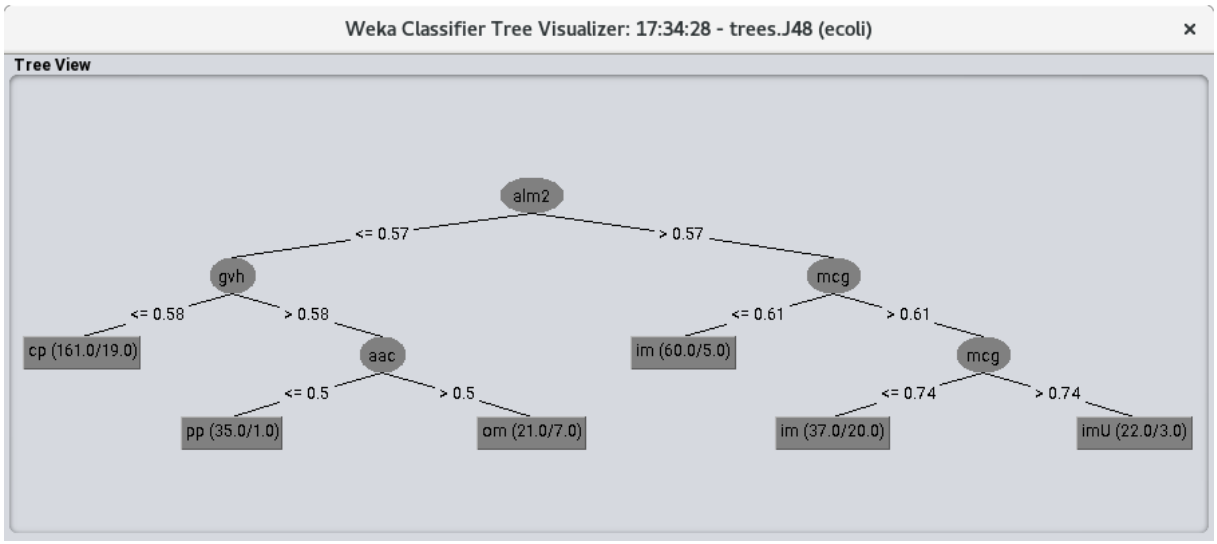
Tab. 3.2: Tabela zbiór ecoli.

METHOD	LEAF	PREC	RECALL	FSCORE	ACCURACY
frequency10	1	0,729	0,738	0,732	0,738
interval10	1	0,694	0,699	0,697	0,699
provided	1	0,821	0,824	0,819	0,824
frequency10	2	0,678	0,702	0,688	0,702
interval10	2	0,696	0,702	0,698	0,702
provided	2	0,834	0,836	0,834	0,836
frequency10	3	0,680	0,702	0,689	0,702
interval10	3	0,726	0,732	0,728	0,732
provided	3	0,821	0,827	0,823	0,827
frequency10	5	0,671	0,705	0,686	0,705
interval10	5	0,710	0,723	0,711	0,723
provided	5	0,801	0,815	0,807	0,815
frequency10	7	0,614	0,670	0,632	0,670
interval10	7	0,705	0,717	0,700	0,717
provided	7	0,782	0,807	0,793	0,807



Rys. 3.2: Wyniki klasyfikatora dla różnych metod dyskretyzacji. Zbiór ecoli.

Poniżej przedstawiono przykładowy wygląd drzewa C4.5 dla zbioru ecoli.



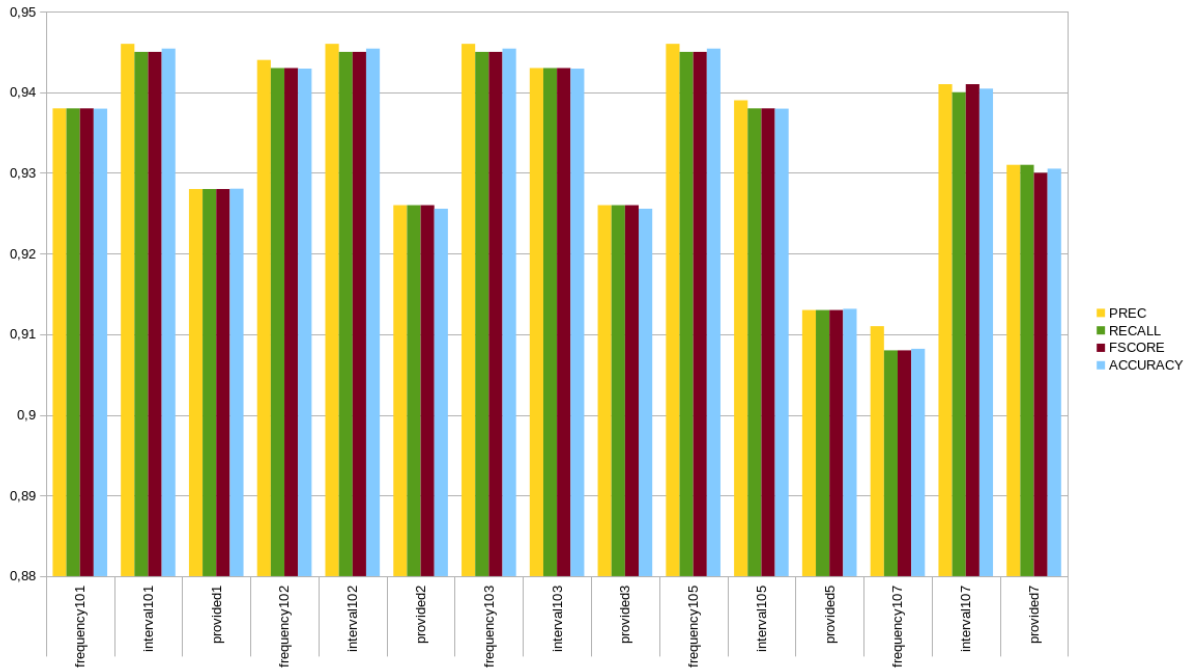
Rys. 3.3: Przykładowa instacja drzewa C4.5

3.1.3. Zbiór ukm

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

Tab. 3.3: Tabela zbiór ukm.

METHOD	LEAF	PREC	RECALL	FSCORE	ACCURACY
frequency10	1	0,938	0,938	0,938	0,938
interval10	1	0,946	0,945	0,945	0,945
provided	1	0,928	0,928	0,928	0,928
frequency10	2	0,944	0,943	0,943	0,943
interval10	2	0,946	0,945	0,945	0,945
provided	2	0,926	0,926	0,926	0,926
frequency10	3	0,946	0,945	0,945	0,945
interval10	3	0,943	0,943	0,943	0,943
provided	3	0,926	0,926	0,926	0,926
frequency10	5	0,946	0,945	0,945	0,945
interval10	5	0,939	0,938	0,938	0,938
provided	5	0,913	0,913	0,913	0,913
frequency10	7	0,911	0,908	0,908	0,908
interval10	7	0,941	0,940	0,941	0,940
provided	7	0,931	0,931	0,930	0,931



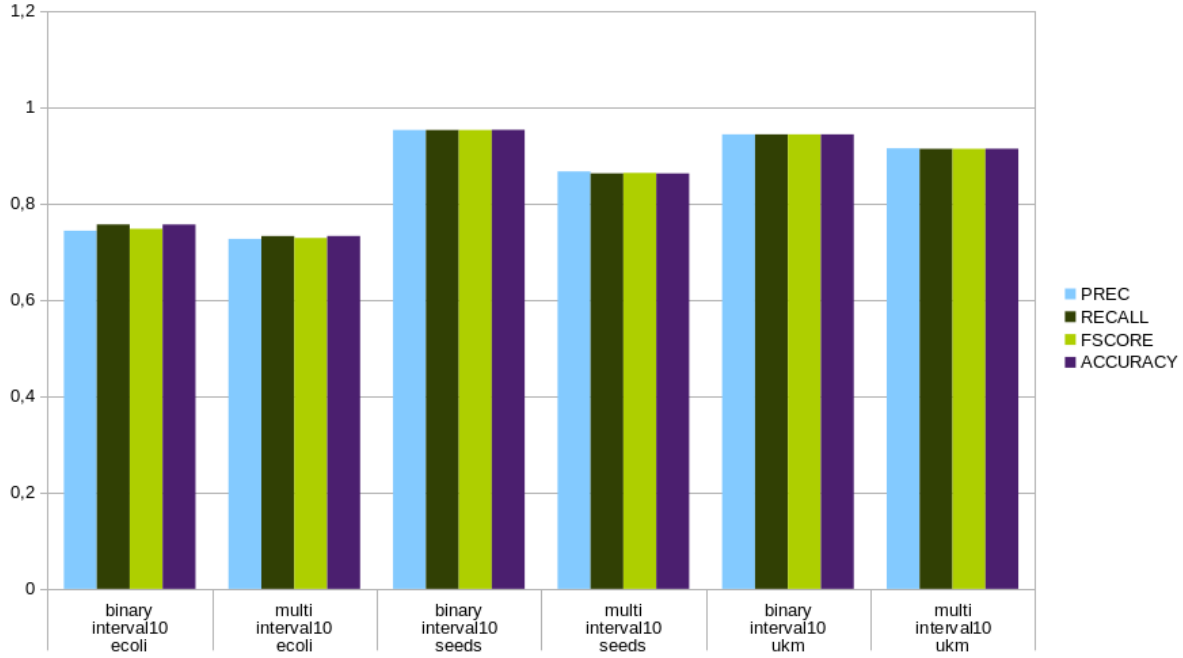
Rys. 3.4: Wyniki klasyfikatora dla różnych metod dyskretyzacji. Zbiór ukm.

3.2. Porównanie metod podziału drzewa

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

Tab. 3.4: Tabela podział drzewa.

NAME	BINARY	PREC	RECALL	FSCORE	ACCURACY
ecoli	binary	0,743	0,756	0,747	0,755952
ecoli	multi	0,726	0,732	0,728	0,732143
seeds	binary	0,952	0,952	0,952	0,952381
seeds	multi	0,866	0,862	0,863	0,861905
ukm	binary	0,943	0,943	0,943	0,942928
ukm	multi	0,914	0,913	0,913	0,913151



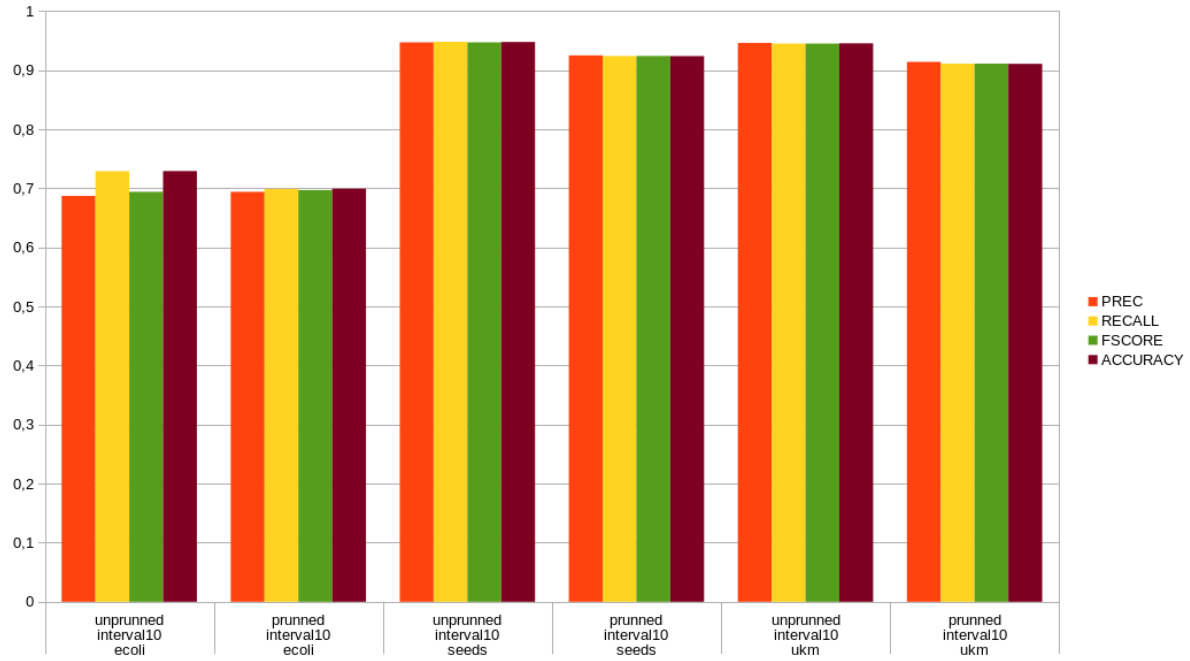
Rys. 3.5: Wyniki klasyfikatora dla różnych metod podziału drzewa.

3.3. Porównanie metod prunningu

Poniżej przedstawiono wyniki klasyfikatorów z włączonym i nie prunningiem. Wyniki klasyfikatorów uśredniono za pomocą średniej ważonej.

Tab. 3.5: Tabela prunningu.

NAME	UNPRUNNED	PREC	RECALL	FSCORE	ACCURACY
ecoli	unprunned	0,687	0,729	0,694	0,729167
ecoli	prunned	0,694	0,699	0,697	0,699405
seeds	unprunned	0,947	0,948	0,947	0,947619
seeds	prunned	0,925	0,924	0,924	0,92381
ukm	unprunned	0,946	0,945	0,945	0,945409
ukm	prunned	0,914	0,911	0,911	0,91067



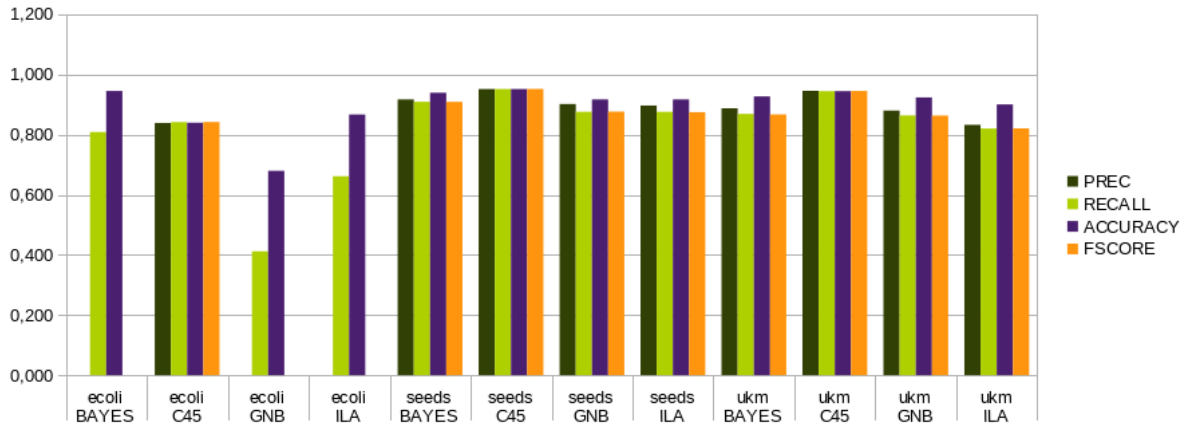
Rys. 3.6: Wyniki klasyfikatora dla różnych metod prunningu.

3.4. Porównanie z innymi klasyfikatorami

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, oraz algorytm C45. Dodatkowo wybrano jedynie najlepsze klasyfikatory spośród wszystkich wyników.

Tab. 3.6: Tabela porównawcza.

CLASSIFIER	NAME	PREC	RECALL	ACCURACY	FSCORE
BAYES	ecoli	nan	0,809	0,945	nan
C45	ecoli	0,839	0,842	0,84	0,842262
GNB	ecoli	nan	0,412	0,680	nan
ILA	ecoli	nan	0,662	0,867	nan
BAYES	seeds	0,917	0,910	0,940	0,909
C45	seeds	0,952	0,952	0,952	0,952381
GNB	seeds	0,902	0,876	0,917	0,877
ILA	seeds	0,897	0,876	0,917	0,875
BAYES	ukm	0,888	0,869	0,927	0,867
C45	ukm	0,946	0,945	0,945	0,945409
GNB	ukm	0,880	0,864	0,924	0,864
ILA	ukm	0,833	0,821	0,901	0,821



Rys. 3.7: Analiza porównawcza klasyfikatorów.

Rozdział 4

Wnioski

4.1. Dyskretyzacja

Dla zbiorów seeds oraz ukm, dokonanie ręcznej dyskretyzacji pomiędzy 10 wartości okazało się lepszym rozwiązaniem niż wbudowany mechanizm radzenia sobie z wartościami ciągłymi w C45. Dzięki takiemu zabiegowi uzyskano dokładność klasyfikacji powyżej 0.93. W przypadku zbioru ecoli zabieg ten wykazał pogorszeniem wyników. Dla tego zbioru najlepszą metodą okazał się wbudowany mechanizm, dzięki czemu wyniki wzrosły o około 0.1 niezależnie od rodzaju miary.

4.2. Podział poddrzewa

We wszystkich analizowanych zbiorach, rodzaj podziału węzła w drzewie metodą binarną okazał się lepszy niż podział wielokryterialny.

4.3. Prunning

Dla domyślnych parametrów prunningu, testy nie wykazały znaczących różnic przy stosowaniu prunningu. Dzieje się tak, ze względu na zbyt niski ćonfidence factor", który domyślnie wynosił 0.1. Współczynnik ten oznacza zaufanie do maksymalnego błędu jaki zbudowane drzewo może popełnić. Im większy, tym mniejszy prunning (większe zaufanie do popełnianego błędu).

4.4. Porównanie klasyfikatorów

Dla zbiorów seeds oraz ukm klasyfikator oparty na algorytmie C45 wykazał się lepszą skutecznością niż pozostałe klasyfikatory (bayesowskie oraz ILA). W zbiorze ecoli odnotowano natomiast pogorszenie dokładności klasyfikacji metodą C45 wględem metody bayesowskiej oraz ILA.