POLITECHNIKA WROCŁAWSKA
WYDZIAŁ INFORMATYKI I ZARZĄDZANIA

KIERUNEK: INFORMATYKA

Systemy uczące się
Laboratorium

Naiwny klasyfikator bayesowski.

AUTOR:
Bartosz Kardas

PROWADZĄCY:
dr inż. Paweł Myszkowski

OCENA PRACY:

WROCŁAW, 2017

Spis treści

Spis rysunków

Spis tabel

Rozdział 1

Naiwny klasyfikator bayesowski - opis

1.1. Algorytm

Naiwny klasyfikator bayesowski opiera się na twierdzeniu bayesa o prawdopodobieństwie warunkowym. Słowo 'naiwny' pochodzi od zaadaptowania warunku iż każdy z atrybutów jest od siebie niezależny. W rzeczywistości warunek ten jest ciężki do spełnienia, ale pomimo tego klasyfikator ten charakteryzuję się zaskakująco dobrą skutecznością.

Budowa modelu polega na wyznaczeniu cząstkowych prawdopodobieństw $P(X_i|C)$ wystąpienia danego atrybutu X_i pod warunkiem wystąpienia klasy C oraz na obliczeniu prawdopodobieństw $P(C)$ wystąpienia danej klasy w zbiorze. Następnie w wyniku zastosowania twierdzenia bayesa i opisanej wyżej 'naiwności' można zastosować poniższy wzór służący do klasyfikacji obiektu o atrybutach $x_1, x_2, ..., x_n$ za pomocą poniższego wzoru.

$$predict(x_1, x_2, ..., x_n) = argmax_c P(C = c) \prod_{i=1}^{n} P(X_i = x_i|C = c)$$

Innymi słowy, klasyfikacja odbywa się na zasadzie znalezienia wartości maksymalnej z iloczyny wcześniej wyliczonych prawdopodobieństw.

W implementacji algorytmu często stosuje się wersję z sumą logarytmów prawdopodobieństw, zapewniającą dokładniejsze wyniki dla małych prawdopodobieństw i dużej ilości atrybutów (w wyniku mnożenia i niedokładności liczb zmiennoprzecinkowych można otrzymać zerowe prawdopodobieństwo).

$$predict(x_1, x_2, ..., x_n) = argmax_c ln(P(C = c)) \sum_{i=1}^{n} ln(P(X_i = x_i|C = c))$$

1.2. Problem zerowego prawdopodobieństwa

Podczas nauki klasyfikatora może nastąpić problem zerowego prawdopodobieństwa, w przypadku gdy dana klasa atrybutu nie występuje w parze z klasą. Aby uniknąć tego typu sytuacji należy dodać 1 do każdego elementu macierzy liczności klasa-atrybut. Metoda ta nazywa się wygładzaniem Laplace'a (ang. Laplace smoothing). Poniżej przedstawiono przykład obliczeniowy ilustrujący działanie metody.

Tab. 1.1: Rozkład danych przed zastosowaniem metody

	$C = yes$	$C = no$
$A = sunny$	5	3
$A = windy$	0	0

Z powyższej tabeli wynika, że

$$P(A = sunny|C = yes) = \frac{5}{5} = 1$$

$$P(A = windy|C = yes) = \frac{0}{5} = 0$$

Po zastosowaniu powyżej opisanej metody otrzymujemy

Tab. 1.2: Rozkład danych przed zastosowaniem metody

	$C = yes$	$C = no$
$A = sunny$	6	4
$A = windy$	1	1

Z odpowiadającymi nie zerowymi prawdopodobieństwami.

$$P(A = sunny|C = yes) = \frac{6}{7}$$

$$P(A = windy|C = yes) = \frac{1}{7}$$

1.3. Dane ciągłe, rozkład normalny

Dla danych ciągłych można także zastosować rozkład Gaussa do obliczenia statystyki. Poniżej przedstawiono wzór na prawdopodobieństwo warunkowe wystąpienia wartości i-tego atrybutu x_i, pod warunkiem wystąpienia klasy c.

$$P(x_i|c) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

, gdzie
x_i - wartość i-tego atrybutu
x_i - klasa
μ - wartość średnia zbioru wartości i-tego atrybutu
σ - odchylenie standardowe zbioru wartości i-tego atrybutu

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 3.1. 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 3.1

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 1.1.

Jeżeli zostanie ustawiona odpowiednia opcja, poszczególne atrybuty rzeczywiste zostaną potraktowane jako wartości ciągłe o odpowiednim rozkładzie normalnym. W przeciwnym razie wartości ciągłe zostaną poddane dyskretyzacji zgodnie z wybraną metodą, opisaną w 3.1

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

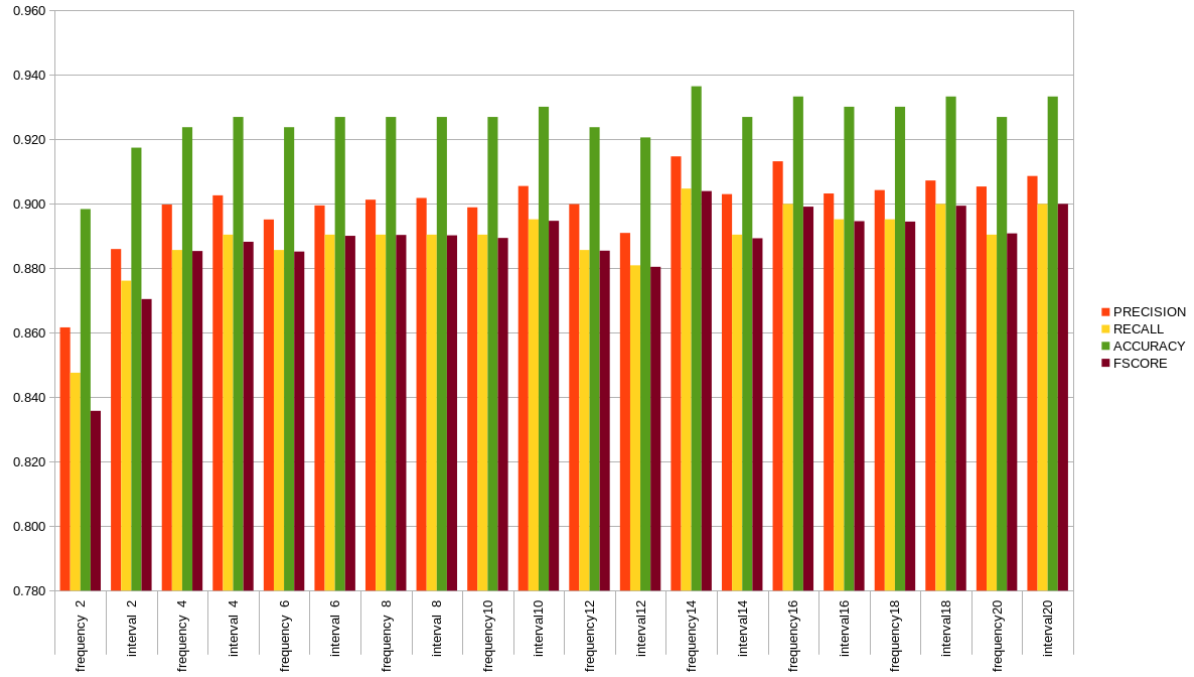
4.1. Porównanie metod dyskretyzacji

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	SIZE	PRECISION	RECALL	ACCURACY	FSCORE
frequency	2	0.862	0.848	0.898	0.836
interval	2	0.886	0.876	0.917	0.870
frequency	4	0.900	0.886	0.924	0.885
interval	4	0.903	0.890	0.927	0.888
frequency	6	0.895	0.886	0.924	0.885
interval	6	0.900	0.890	0.927	0.890
frequency	8	0.901	0.890	0.927	0.890
interval	8	0.902	0.890	0.927	0.890
frequency	10	0.899	0.890	0.927	0.889
interval	10	0.906	0.895	0.930	0.895
frequency	12	0.900	0.886	0.924	0.886
interval	12	0.891	0.881	0.921	0.880
frequency	14	0.915	0.905	0.937	0.904
interval	14	0.903	0.890	0.927	0.889
frequency	16	0.913	0.900	0.933	0.899
interval	16	0.903	0.895	0.930	0.895
frequency	18	0.904	0.895	0.930	0.895
interval	18	0.907	0.900	0.933	0.899
frequency	20	0.905	0.890	0.927	0.891
interval	20	0.909	0.900	0.933	0.900



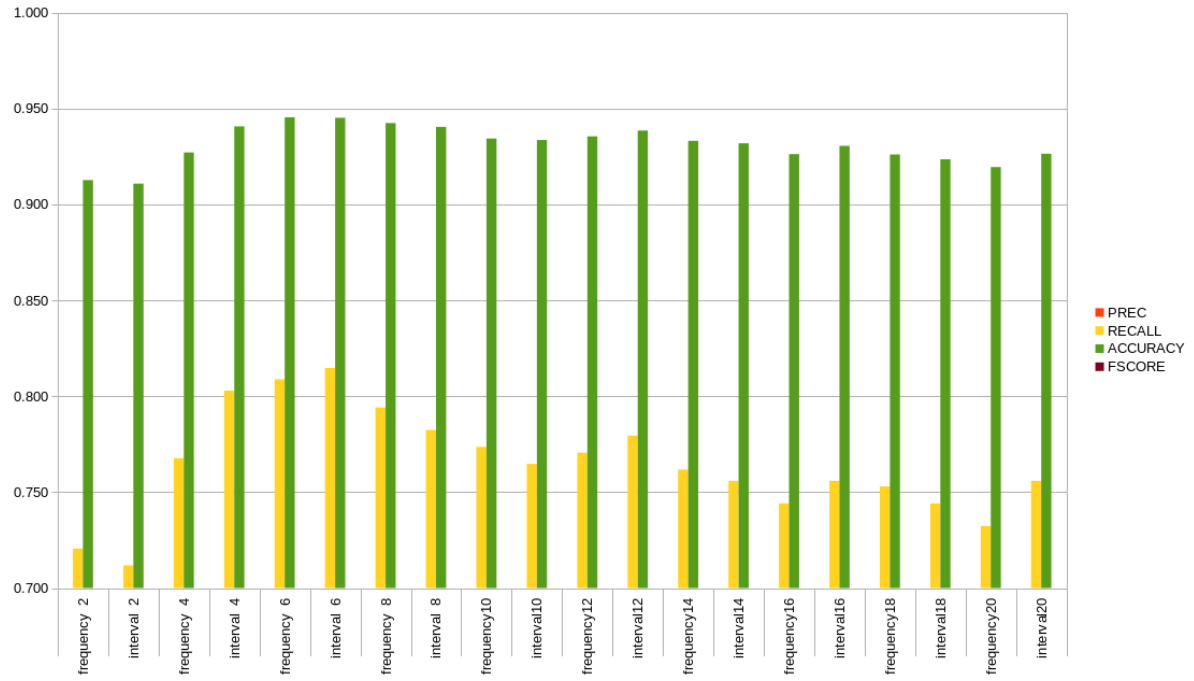
Rys. 4.1: Wyniki klasyfikatora dla różnych metod dyskretyzacji. Zbiór seeds.

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
frequency	2	nan	0.721	0.913	nan
interval	2	nan	0.712	0.911	nan
frequency	4	nan	0.768	0.927	nan
interval	4	nan	0.803	0.941	nan
frequency	6	nan	0.809	0.945	nan
interval	6	nan	0.815	0.945	nan
frequency	8	nan	0.794	0.942	nan
interval	8	nan	0.782	0.940	nan
frequency	10	nan	0.774	0.934	nan
interval	10	nan	0.765	0.934	nan
frequency	12	nan	0.771	0.935	nan
interval	12	nan	0.779	0.938	nan
frequency	14	nan	0.762	0.933	nan
interval	14	nan	0.756	0.932	nan
frequency	16	nan	0.744	0.926	nan
interval	16	nan	0.756	0.931	nan
frequency	18	nan	0.753	0.926	nan
interval	18	nan	0.744	0.924	nan
frequency	20	nan	0.732	0.919	nan
interval	20	nan	0.756	0.926	nan



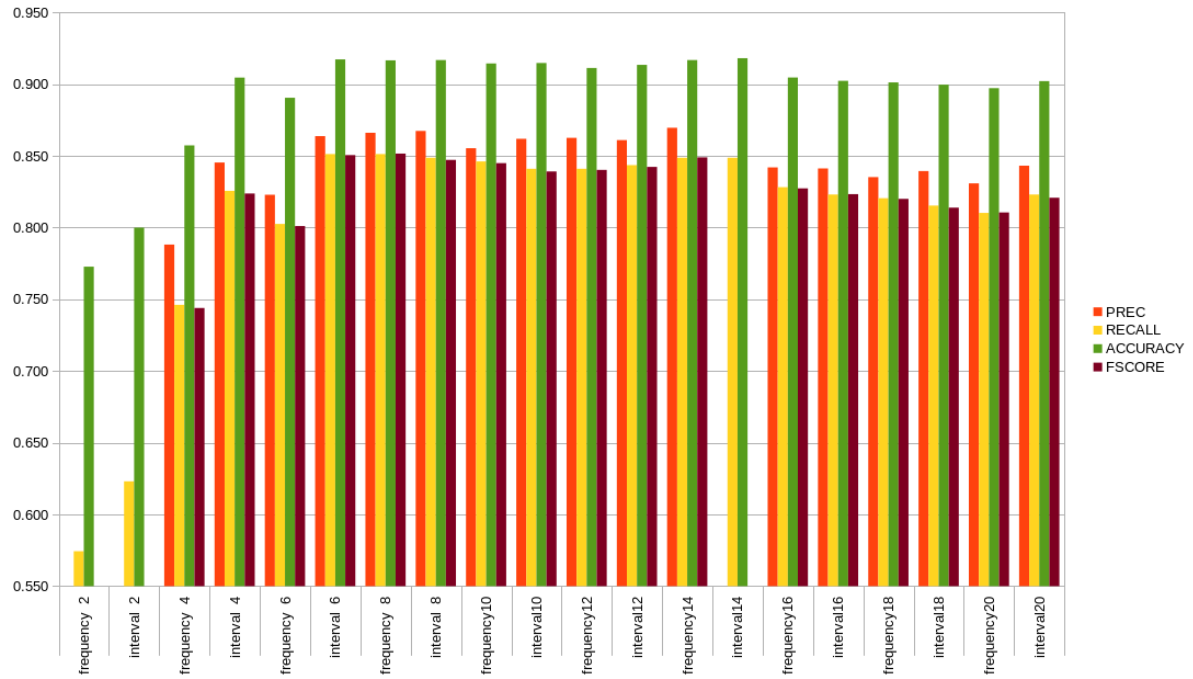
Rys. 4.2: Wyniki klasyfikatora dla różnych metod dyskretyzacji. Zbiór ecoli.

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
frequency	2	nan	0.574	0.773	nan
interval	2	nan	0.623	0.800	nan
frequency	4	0.788	0.746	0.857	0.744
interval	4	0.845	0.826	0.905	0.824
frequency	6	0.823	0.803	0.891	0.801
interval	6	0.864	0.851	0.917	0.851
frequency	8	0.866	0.851	0.917	0.852
interval	8	0.867	0.849	0.917	0.847
frequency	10	0.855	0.846	0.914	0.845
interval	10	0.862	0.841	0.915	0.839
frequency	12	0.863	0.841	0.911	0.840
interval	12	0.861	0.844	0.913	0.842
frequency	14	0.870	0.849	0.917	0.849
interval	14	nan	0.849	0.918	nan
frequency	16	0.842	0.828	0.905	0.827
interval	16	0.841	0.823	0.902	0.823
frequency	18	0.835	0.821	0.901	0.820
interval	18	0.839	0.815	0.900	0.814
frequency	20	0.831	0.810	0.897	0.811
interval	20	0.843	0.823	0.902	0.821



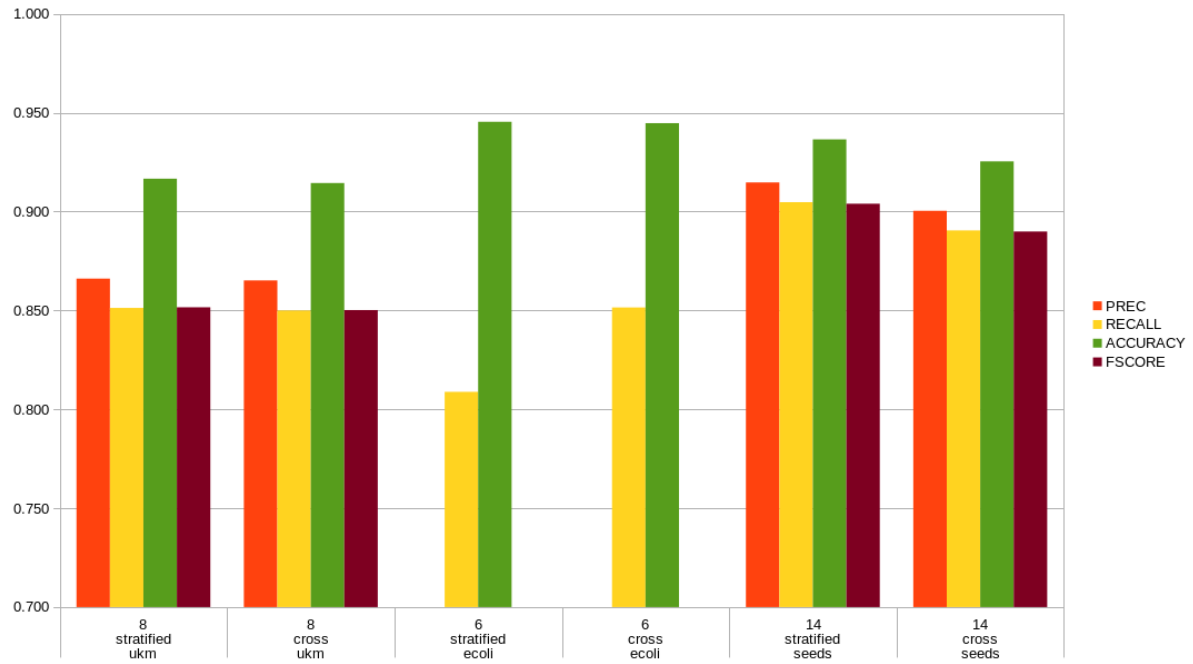
Rys. 4.3: Wyniki klasyfikatora dla różnych metod dyskretyzacji. Zbiór ukm.

4.2. Porównanie metod kroswalidacji

Poniżej przedstawiono wyniki klasyfikatorów obliczone różnymi typami kroswalidacji. Dodatkowo wybrano jedynie najlepsze klasyfikatory spośród wszystkich, które wykorzystywały częstotliwością metodę dyskretyzacji. Wyniki klasyfikatorów uśredniono za pomocą średniej ważonej.

Tab. 4.4: Tabela kroswalidacja.

NAME	DATASET	CROSS	AVG	SIZE	PREC	RECALL	ACCURACY	FSCORE
ukm	frequency	stratified	w	8	0.866	0.851	0.917	0.852
ukm	frequency	cross	w	8	0.865	0.850	0.914	0.850
ecoli	frequency	stratified	w	6	nan	0.809	0.945	nan
ecoli	frequency	cross	w	6	nan	0.852	0.945	nan
seeds	frequency	stratified	w	14	0.915	0.905	0.937	0.904
seeds	frequency	cross	w	14	0.900	0.890	0.925	0.890



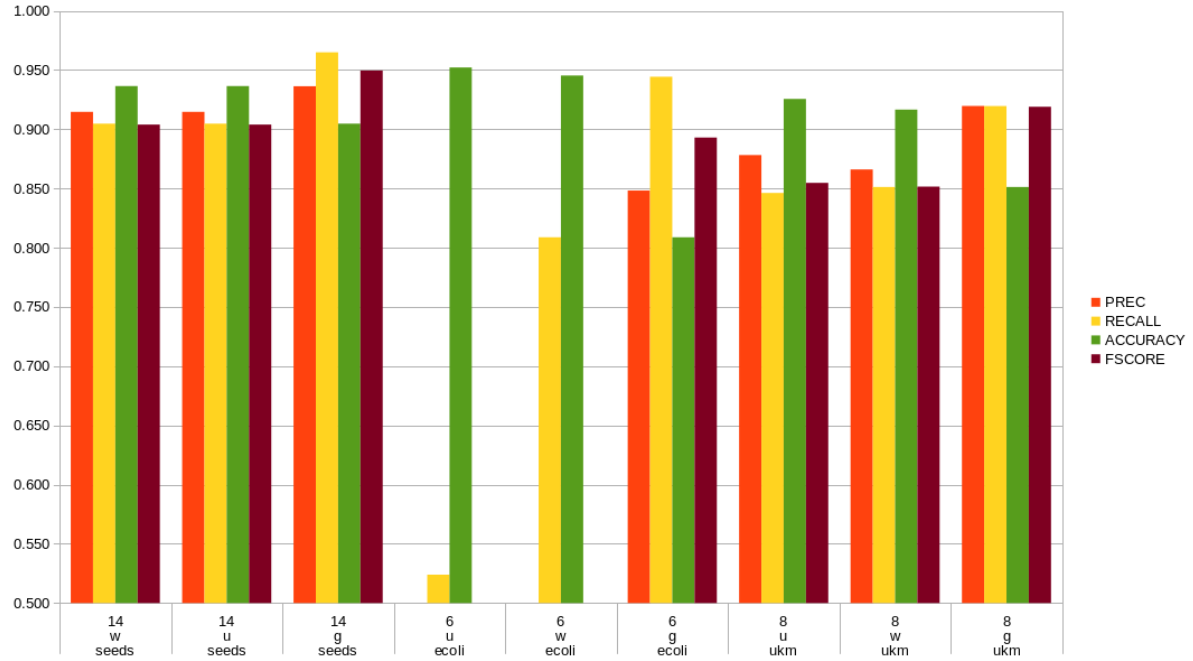
Rys. 4.4: Wyniki klasyfikatora dla różnych metod kroswalidacji.

4.3. Porównanie metod uśredniania statystyk

Poniżej przedstawiono wyniki klasyfikatorów obliczone różnymi metodami uśredniania statystyk. Dodatkowo wybrano jedynie najlepsze klasyfikatory spośród wszystkich, które wykorzystywały częstotliwością metodę dyskretyzacji oraz kroswalidację stratyfikowaną.

Tab. 4.5: Tabela statystyki.

NAME	DATASET	CROSS	AVG	SIZE	PREC	RECALL	ACCURACY	FSCORE
seeds	frequency	stratified	w	14	0.915	0.905	0.937	0.904
seeds	frequency	stratified	u	14	0.915	0.905	0.937	0.904
seeds	frequency	stratified	g	14	0.936	0.965	0.905	0.950
ecoli	frequency	stratified	u	6	nan	0.524	0.952	nan
ecoli	frequency	stratified	w	6	nan	0.809	0.945	nan
ecoli	frequency	stratified	g	6	0.848	0.944	0.809	0.893
ukm	frequency	stratified	u	8	0.878	0.846	0.926	0.855
ukm	frequency	stratified	w	8	0.866	0.851	0.917	0.852
ukm	frequency	stratified	g	8	0.920	0.920	0.851	0.919



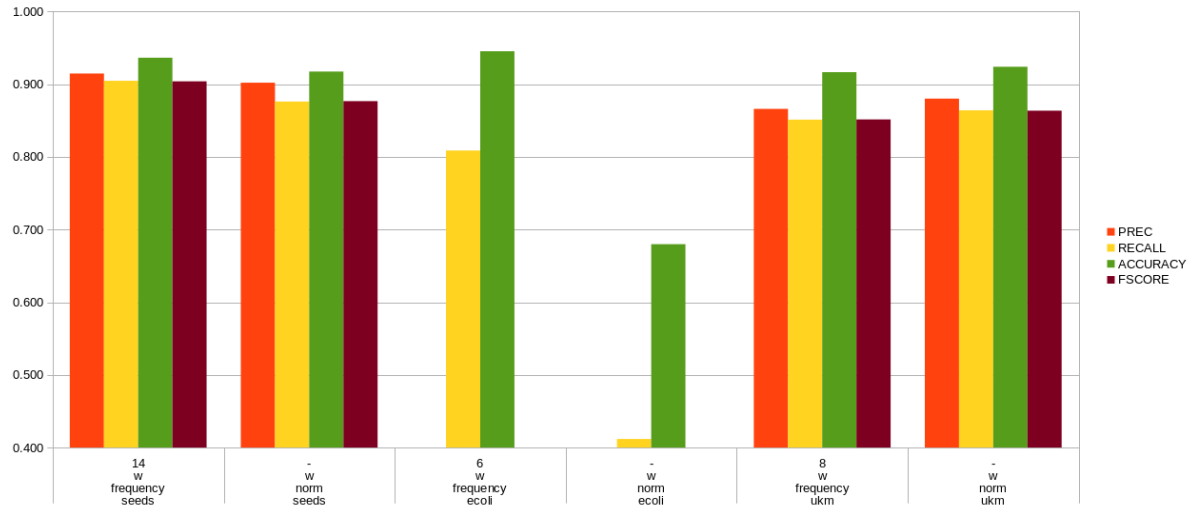
Rys. 4.5: Wyniki klasyfikatora dla różnych metod obliczania statystyki.

4.4. Rozkład normlany

Poniżej przedstawiono wyniki klasyfikatorów z zastosowaną dyskretyzacją w porównaniu do tych które używały prawdopodobieństwa z rozkładu Gaussa. Dodatkowo wybrano jedynie najlepsze klasyfikatory spośród wszystkich, które wykorzystywały częstotliwością metodę dyskretyzacji oraz kroswalidację stratyfikowaną. Wyniki klasyfikatorów uśredniono za pomocą średniej ważonej.

Tab. 4.6: Tabela dane ciągłe.

NAME	DATASET	CROSS	AVG	SIZE	PREC	RECALL	ACCURACY	FSCORE
seeds	frequency	stratified	w	14	0.915	0.905	0.937	0.904
seeds	norm	stratified	w	-	0.902	0.876	0.917	0.877
ecoli	frequency	stratified	w	6	nan	0.809	0.945	nan
ecoli	norm	stratified	w	-	nan	0.412	0.680	nan
ukm	frequency	stratified	w	8	0.866	0.851	0.917	0.852
ukm	norm	stratified	w	-	0.880	0.864	0.924	0.864



Rys. 4.6: Porównanie wyników dyskretnych z wartościami ciągłymi.

Rozdział 5

Wnioski

5.1. Dyskretyzacja

W przypadku zbioru seeds, dyskretyzacja obiema metodami nie przyniosła znacznych różnic w skuteczności klasyfikatorów. Obie metody wykazały się podobną skutecznością. Największą dokładność (accuracy) uzyskano dla dyskretyzacji wszystkich atrybutów pomiędzy 14 wartości. Wyniosła ona odpowiednio 0.937 dla metody częstotliwościowej oraz 0.927 dla metody interwałowej. W przypadku zbyt niskiej ilości przedziałów dyskretyzacji (np. 2,3) odnotowano znaczący spadek skuteczności klasyfikatorów, co jest zgodne z intuicją - klasyfikacja będzie skuteczniejsza gdy wartości atrybutów będą niosły ze sobą większą ilość informacji.

W zbiorze ecoli dla precyzji (precision) oraz fscore wielokrotnie wystąpiła wartość 'nan' oznaczająca dosłownie nie-liczbę. W przypadku precyzji wartość ta oznacza, że wszystkie wystąpienia danej klasy zostały zaklasyfikowane jako negatywne, co oznacza, że nie można w tym przypadku określić zachowania klasyfikatora dla pozytywnych wartości predykcji. W związku z tym, że fscore oblicza się na podstawie precision oraz recall, wartość ta nie mogła zostać obliczona. Największą dokładność (accuracy) uzyskano dla dyskretyzacji wszystkich atrybutów pomiędzy 14 wartości. Wyniosła ona 0.945 dla obu metod dyskretyzacji.

Dla ostatniego zbioru - ukm, również w kilku miejscach pojawiła się wartość 'nan'. Powód jej wystąpienia jest opisany powyżej. Największą wartość dokładności uzyskano dla 8 wartości dyskretnych. Wyniosła ona około 0.915 dla obu metod.

5.2. Kroswalidacja

Do porównania wyników kroswalidacji użyto najlepszych wyników klasyfikatorów dla danej metody dyskretyzacji i ilość przedziałów. Dla wszystkich wyników obie metody dały podobne rezultaty. Jednym z powodów może być dosyć równomierna dystrybucja klas w badanych zbiorach (w szczególności zbiór seeds). 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. Ocena klasyfikatorów

W przypadku trzech metod uśredniania statystyk, uśrednianie arytmetyczne i ważone zachowywało się podobnie dla zbiorów danych z dużą ilością elementów. W przypadku zbioru ecoli, gdzie kilka klas było małolicznych, widać wyraźne różnice we wskaźniku recall. Dla średniej arytmetycznej recall wyniósł 0.524, a dla średniej ważonej 0.809. Dzięki temu wyraźnie widać, jak duży wpływ na ten wskaźnik mają klasy o niewielkiej ilości elementów, czego powiedzieć nie można o wskaźniku accuracy, gdzie oba wyniki wyniosły około 0.95. Globalna metoda liczenia statystyk, zachowywała się odrobinę inaczej od pozostałych. Dla wskaźników precision oraz recall, fscore dawała wyższe wartości, a dla accuracy niższe.

5.4. Rozkład normalny

Klasyfikatory wykorzystujące aproksymację prawdopodobieństw z rozkładu normalnego charakteryzowały się niższą dokładnością dla zbiorów seeds oraz ecoli (o odpowiednio 2% oraz 28%). Różnice w wynikach mogą świadczyć o tym, że niektóre z atrybutów nie posiadały rozkładu normalnego. Dla zbioru ukm, wartości te były nieznacznie wyższe, zatem można przypuszczać, że atrybuty w zbiorze testowym zachowywały rozkład normalny.