Politechnika Wrocławska

Indukcyjne Metody Analizy Danych

Ćwiczenie 5 - Zespoły klasyfikatorów

Marcel Cielinski
Index: 236747

prowadzący
dr inż. Paweł Myszkowski

4 czerwca 2020

Wprowadzenie

1.1 Problem

Celem ćwiczenia było zapoznanie się z trzema metodami tworzenia zespołów klasyfikatorów: las losowy (ang. random forest), bagging oraz boosting. Do składowych zadania zalicza się implementację wymienionych algorytmów w jednym z dwóch języków programowania: *R* lub *Python*. Kolejno, należało przeprowadzić badania na trzech, ustalonych na potrzeby ćwiczenia, zbiorach danych (+ *iris* do sprawdzenia poprawności działania metod). Przedmiotem testów miło być zbadanie wpływu poszczególnych parametrów trzech algorytmów. Zadbać należało także o zastosowanie operacji na danych, które je ujednolicą (standaryzacja lub normalizacja) oraz zbadanie dwóch rodzajów walidacji krzyżowej do oceny skuteczności algorytmu poprzez obserwację metryk mówiących o jakości klasyfikatora. Ostatnim poleceniem jest porównanie otrzymanych rezultatów z wynikami standardowych klasyfikatorów z poprzednich ćwiczeń.

1.2 Zespoły klasyfikatorów

Zespół klasyfikatorów składa się (na co wskazuje sama nazwa) z określonej liczby podstawowych klasyfikatorów, nazywanych zwykle bazowymi. Mogą nimi być różne klasyfikatory proste, takie jak: naiwny klasyfikator Bayesowski, drzewo decyzyjne, SVM (Support Vector Machine), klasyfikator oparty o architekturę sieci neuronowej i inne. Główną myślą opisywanej metody jest budowa „dobrego" zespołu z „mniej zdolnych" członów. Do zalet podstawowych jednostek klasyfikujących zaliczyć należy nieskomplikowaną złożoność obliczeniową ich konstruowania oraz prostotę budowy. Z kolei ich złączenie do bardziej złożonego systemu decyzyjnego również wydaje się przynosić wiele korzyści. Oczekiwanym jest bowiem, dysponowanie większą siłą wyrażania, a co za tym idzie - być może zdolność do uczenia się trudniejszych pojęć. Można się też spodziewać, że taka struktura będzie mniej podatna na osobliwości danych i mniej podatna na overfitting.

W zależności od konkretnego algorytmu, bazowe klasyfikatory są uczone na pełnym lub wybranej (najczęściej losowo) części zbioru treningowego, dodatkowo uczenie może odbywać się sekwencyjnie lub równolegle. Ostateczna decyzja zespołu jest podejmowana na podstawie głosowania klasyfikatorów bazowych. Może się ono odbywać na kilka różnych sposobów, najczęściej głosowanie większościowe lub głosowanie ważone.

Bagging

Bagging, który oznacza agregację bootstrap (Bootstrap aggregating), jest jednym z najprostszych algorytmów intuicyjnych jak i opartych na zespołach z zaskakująco dobrą wydajnością. Różnorodność klasyfikatorów uzyskuje się za pomocą replik danych treningowych bootstrap. Oznacza to, że różne podzbiory danych treningowych są tworzone losowo ze zbioru głównego (losowanie ze zwracaniem).

Każdy podzbiór danych treningowych służy do szkolenia innego klasyfikatora bazowego. Poszczególne klasyfikatory są następnie łączone, podejmując decyzję zwykłą większością głosów. Dla każdej instancji klasa wybrana przez więk-

szość klasyfikatorów jest decyzją zespołu. Ponieważ zbiory danych szkoleniowych mogą się znacząco nakładać, można zastosować dodatkowe środki w celu zwiększenia różnorodności. Ich przykładem może być wykorzystanie podzbioru danych szkoleniowych do szkolenia każdego klasyfikatora lub użycie względnie słabych klasyfikatorów.

Boosting

Podobnie jak w przypadku baggingu, boosting tworzy również grupę klasyfikatorów poprzez transformację sygnału danych, które są następnie łączone przez głosowanie większościowe. Jednak w fazie zwiększania, resampling jest strategicznie dostosowywany w celu dostarczenia najbardziej pouczających danych treningowych dla każdego kolejnego klasyfikatora. Bowiem błędnie zaklasyfikowane instancje dostaną wyższe wagi (niż prawidłowo zaklasyfikowane) i przez to mają większą szansę na pojawienie się w nowym zbiorze uczącym kolejnego klasyfikatora bazowego (a to z koeli pośrednio zwiększona jest szansa poprawy ich klasyfikacji).

Warto nadmienić, iż algorytmy boostingowe (w tym najpopularniejszy *AdaBoost*) korzystają z bardzo prostych klasyfikatorów bazowych (ang. weak learners). W praktyce są to najczęściej drzewa o jednym rozgałęzieniu - pnie (ang. stumps)

Random forest

Zbiór drzew decyzyjnych to las losowy (ang. random forest). Lasy losowe wykonują wewnętrzny bagging. Tworzone jest od kilku do kilku tysięcy drzew (zależnie od zadanych parametrów). Następnie obliczany jest najlepszy możliwy model dla danego zestawu danych. Zamiast rozpatrywać wszystkie atrybuty podczas dzielenia węzła, algorytm wybiera najlepszy atrybut spośród losowego podzbioru wszystkich możliwych. To powoduje większe odchylenie w przypadku mniejszej wariancji, co daje znacznie lepszy model.

Implementacja

Wybranym przeze mnie środowiskiem jest język *Python*. Do implementacji algorytmów random forest, bagging oraz boosting, które były przedmiotem ćwiczenia, użyto biblioteki uczenia maszynowego *sklearn*. Konkretniej, wykorzystano klasy: `RandomForestClassifier`, `BaggingClassifier` oraz `AdaBoostClassifier` z pakietu *sklearn.ensemble*. Pozwalają one na ustawienie wszystkich interesujących nas parametrów. W kodzie wykorzystano także takie biblioteki jak: *pandas*, *matplotlib*, *numpy*, *scipy* oraz *seaborn*.

2.1 Ocena jakości

Do oceny jakości zespołu klasyfikatorów stosuje się (analogicznie jak w przypadku podstawowych klasyfikatorów) metryki związane, bądź wynikające z macierzy błędów (ang. Confusion matrix). Po wykonaniu zadania klasyfikacji, możliwe jest zbadanie prawidłowości dopasowań, analizując rzeczywiste etykiety.

Właśnie w tym celu posługujemy się tablicą pomyłek, która zestawia liczebność instancji prawidłowo i nieprawidłowo oznaczonych przez algorytm. Podstawowe takie metryki to:

- *Accuracy* - dokładność - jak dobre są ogólne predykcje klasyfikatora. Jest to stosunek dobrze oznaczonych klas do wszystkich oznaczeń.
- *Precission* - precyzja - mówi o precyzyjności klasyfikatora. Jest to zdolność klasyfikatora do nie oznaczania negatywnych próbek jako pozytywne.
- *Recall* - czułość - jak dobrze klasyfikator radzi sobie ze znajdywaniem pozytywnych próbek. Jest to stosunek próbek dobrze zaklasyfikowanych jako pozytywne przez wszystkie rzeczywiście pozytywne.
- *F1-score* - oznaczane jako *FSC* - jest to w istocie średnia harmoniczna czułości oraz precyzji.

Do badania większości testów posłuży głownie miara *FSC*. Jest to często wykorzystywana metryka, która pozwoli na wymiernie dobre porównanie różnych podejść. W implementacji wykorzystany został pakiet *sklearn.metrics*, który udostępnia wszystkie powyżej wymienione metryki.

2.2 Ujednolicenie danych

Ujednolicenie danych polega na zastosowaniu operacji takich jak standaryzacja czy normalizacja. Pierwsza z metod opiera się na przemapowaniu wartości atrybutów w taki sposób, by ich średnia skupiała się na wartości 0, a odchylenie standardowe wynosiło 1. Z kolei normalizacja zamienia wielkości atrybutów w taki sposób, aby instancje zachowały wzajemne proporcje między wartościami atrybutów, jednak znalazły się w zakresie $[0, 1]$. Ustalono, że na potrzeby wszystkich testów, atrybuty zbirów danych będą znormalizowane (za pomocą funkcji dostępnej w pakiecie *sklearn.preprocessing*).

2.3 Kroswalidacja

Kroswalidacja (sprawdzian krzyżowy) jest stosowana w celu lepszej oceny działania modeli, takich jak klasyfikator czy ich zespół. Procedura determinuje podział danych na treningowe i testowe. A jej wykorzystanie tłumaczy się unikaniem overfittingu.

W realizacja zadania wykorzystano dwie metody, dostępne w pakiecie *sklearn.model_selection*. Są nimi:

- `KFold` - kroswalidacja - gdzie całkowity zbiór danych jest dzielony na K równolicznych podzbiorów (parametr *n_splits*), z których kolejno każdy jest traktowany jako zbiór testowy, kiedy model trenowany jest na połączonych (*K-1*) pozostałych podzbiorach. Otrzymuje się wówczas K wyników, z których następnie liczona jest średnia.
- `StratifiedKFold` - kroswalidacja stratyfikowana - jej działanie jest analogiczne do powyżej opisanego, wariantu zwykłego. Różnica polega na tym, że całkowity zbiór jest dzielony (na K części, także parametrem *n_splits*) w miarę możliwości na podzbiory o proporcjonalnym rozkładzie klas - zgodnie z proporcją istniejącą w całościowym zbiorze.

W badaniach ustawiano parametr *shuffle = False* w celu wymiernego zbadania różnicy między kroswalidacją podstawową, a stratyfikowaną. Testowany był także wpływ ilości foldów (K) na otrzymywane wyniki, dla obu metod.

2.4 Badane parametry

W badaniach, przedstawionych w sekcji poniżej, uwzględniono następujące ustawienia poszczególnych parametrów:

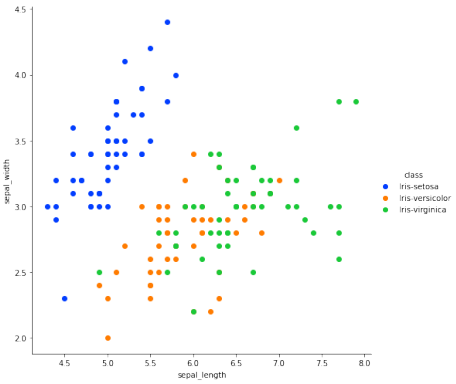
- `random forest`
  - *n_estimators* - [1, 5, 9, 25, 49, 75, 99]
  - *bootstrap* - [True, False]
  - *criterion* - [gini, entropy]
  - *max_features* - [sqrt, log2, None, 1.0]
- `bagging`
  - *n_estimators* - [1, 5, 9, 25, 49, 75, 99]
  - *bootstrap* - [True, False]
  - *max_samples* - [0.25, 0.5, 0.75, 1.0]
- `boosting`
  - *n_estimators* - [1, 5, 9, 25, 49, 75, 99]
  - *learning_rate* - [0.0001, 0.001, 0.01, 0.1, 1]
  - *algorithm* - [SAMME, SAMME.R]

2.5 Zbiory danych

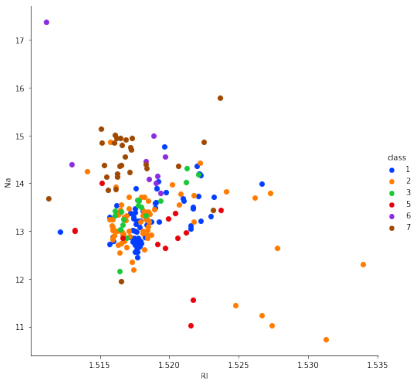
Do walidacji zaimplementowanych funkcjonalności wykorzystano łącznie cztery zbiory danych. Te zbiory, to:

- *iris* - gdzie liczba atrybutów to 4, a klas 3. Zbiór przedstawia odmiany kwiatów Iris, a atrybuty odpowiadają ich cechom.
- *glass* - gdzie liczba atrybutów to 9, a klas 7. Zbiór przedstawia typy szkła, gdzie atrybuty opisują skład chemiczny.
- *wine* - gdzie liczba atrybutów to 13, a klas 3. Zbiór ten zawiera dane o analizie chemicznej win z tego samego regionu, jednak z 3 różnych upraw.
- *seeds* - gdzie liczba atrybutów wynosi 7, a klas 3. Zbiór ten zawiera miary geometrycznych własności 3 różnych odmian ziaren pszenicy.

Na Rysunkach 1, 2, 3, 4 przedstawiono przykładowe rozkłady klas względem wybranych atrybutów. Z kolei w Tabelach 1, 2, 3, 4 zestawiono rozkład klas dla wszystkich badanych zbiorów.



Rysunek 1: Iris



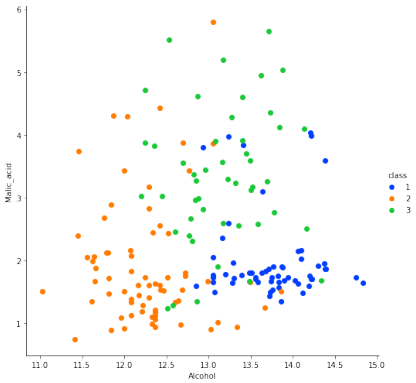
Rysunek 2: Glass

Klasa	Instancje	% rekordów
Setosa	50	33 (%)
Versicolor	50	33 (%)
Virginica	50	33 (%)

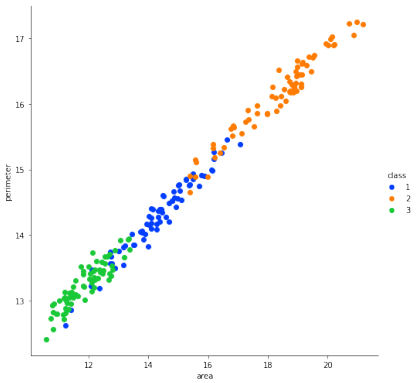
Tablica 1: Klasy zbioru *Iris*

Klasa	Instancje	% rekordów
1	70	33 (%)
2	76	36 (%)
3	17	8 (%)
4	0	0 (%)
5	13	6 (%)
6	9	4 (%)
7	29	13 (%)

Tablica 2: Klasy zbioru *Glass*



Rysunek 3: Wine



Rysunek 4: Seeds

Klasa	Instancje	% rekordów
1	59	33 (%)
2	71	40 (%)
3	48	27 (%)

Tablica 3: Klasy zbioru *Wine*

Klasa	Instancje	% rekordów
1	70	33 (%)
2	70	33 (%)
3	70	33 (%)

Tablica 4: Klasy zbioru *Seeds*

Wyniki eksperymentów

3.1 Zbiór Glass

Poniższe tabelki (Tabela 5, 6, 7, 8, 9, 10, 11, 12, 13) przedstawiają wszystkie przeprowadzone testy na zbiorze *Glass*, z podziałem na badane wartości parametru mówiącego o zastosowanej ilości estymatorów bazowych oraz na pozostałe testowane ustawienia poszczególnych parametrów każdej z zaimplementowanych metod zespołów klasyfikatorów. Wszystkie badania zrealizowane zostały przy pomocy sprawdzianów krzyżowych: kroswalidacja „zwykła" oraz stratyfikowana dla testowanego $n\_estimators$ oraz stratyfikowana dla pozostałych parametrów. W kolejnych kolumnach, każdej z wyżej wymienionych Tabel, umieszczone zostały odczyty metryki *FSC* (o której wspomniano w poprzedniej sekcji).

Dodatkowo na rysunku 5 przedstawione są wykresy prezentujące graficznie wpływ parametru $n\_estimators$ na osiągane wartości metryki *FSC* dla wszystkich badanych metod.

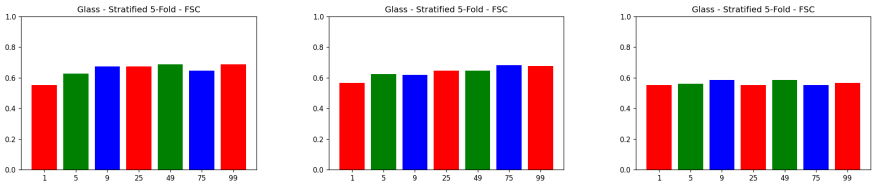
Dla zbioru *Glass*, przy wykorzystaniu algorytmów opartych o budowanie zespołów klasyfikatorów, zauważamy raczej przyrost jakości klasyfikacji wraz ze wzrostem ilości estymatorów bazowych. Wywnioskować można także, że dla lasów losowych wskazane jest stosowanie bootstrapu. Jeżeli chodzi o metodę bagging - również lepsze rezultaty dla bootstrapu oraz dla używanej części danych do budowy klasyfikatorów bazowych wynoszącej 0.5 - 0.75. W przypadku boostingu niskie wartości wskaźnika uczenia się oraz stosowanie algorytmu SAMME, dla tego zbioru, pozytywnie wpływa na wyniki.

n	RanFor	Bag	Boost
2-Fold			
1	0.085	0.077	0.073
5	0.072	0.074	0.085
9	0.073	0.075	0.08
25	0.071	0.06	0.08
49	0.067	0.067	0.074
75	0.077	0.071	0.08
99	0.075	0.07	0.065
5-Fold			
1	0.14	0.187	0.14
5	0.131	0.176	0.127
9	0.17	0.135	0.134
25	0.155	0.15	0.139
49	0.184	0.15	0.135
75	0.155	0.16	0.145
99	0.168	0.143	0.134
10-Fold			
1	0.348	0.349	0.369
5	0.327	0.337	0.362
9	0.46	0.354	0.37
25	0.405	0.397	0.364
49	0.424	0.417	0.361
75	0.377	0.406	0.355
99	0.396	0.414	0.367

Tablica 5: Wpływ ilości estymatorów bazowych na *FSC* - zbiór *Glass* - kroswalidacja „zwykła”

n	RanFor	Bag	Boost
Stratified 2-Fold			
1	0.418	0.49	0.444
5	0.46	0.587	0.473
9	0.578	0.525	0.44
25	0.488	0.569	0.451
49	0.533	0.57	0.467
75	0.541	0.567	0.444
99	0.544	0.539	0.441
Stratified 5-Fold			
1	0.446	0.557	0.549
5	0.643	0.525	0.585
9	0.603	0.643	0.533
25	0.622	0.628	0.527
49	0.621	0.643	0.491
75	0.645	0.657	0.579
99	0.659	0.594	0.561
Stratified 10-Fold			
1	0.532	0.64	0.55
5	0.673	0.624	0.55
9	0.654	0.623	0.563
25	0.687	0.683	0.554
49	0.733	0.694	0.54
75	0.729	0.675	0.577
99	0.732	0.715	0.545

Tablica 6: Wpływ ilości estymatorów bazowych na *FSC* - zbiór *Glass* - kroswalidacja stratyfikowana



(a) Random forest (b) Bagging (c) Boosting

Rysunek 5: Wpływ ilości estymatorów bazowych na *FSC* - zbiór *Glass*

Random forest

Bootstrap	*FSC*
Stratified 2-Fold	
True	0.545
False	0.589
Stratified 5-Fold	
True	0.662
False	0.65
Stratified 10-Fold	
True	0.737
False	0.718

Tablica 7: Wpływ stosowania bootstrapu - zbiór *Glass*

Max features	*FSC*
Stratified 2-Fold	
sqrt	0.576
log2	0.566
None	0.595
1.0	0.567
Stratified 5-Fold	
sqrt	0.654
log2	0.649
None	0.639
1.0	0.628
Stratified 10-Fold	
sqrt	0.726
log2	0.718
None	0.691
1.0	0.693

Tablica 8: Wpływ ograniczenia liczby badanych atrybutów podczas podziału - zbiór *Glass*

Criterion	*FSC*
Stratified 2-Fold	
gini	0.623
entropy	0.571
Stratified 5-Fold	
gini	0.637
entropy	0.619
Stratified 10-Fold	
gini	0.717
entropy	0.739

Tablica 9: Wpływ metody pomiaru jakości podziału - zbiór *Glass*

Bagging

Bootstrap	*FSC*
Stratified 2-Fold	
True	0.512
False	0.471
Stratified 5-Fold	
True	0.615
False	0.562
Stratified 10-Fold	
True	0.683
False	0.589

Tablica 10: Wpływ stosowania bootstrapu - zbiór *Glass*

Max samples	*FSC*
Stratified 2-Fold	
0.25	0.4
0.5	0.554
0.75	0.508
1.0	0.509
Stratified 5-Fold	
0.25	0.479
0.5	0.614
0.75	0.585
1.0	0.557
Stratified 10-Fold	
0.25	0.638
0.5	0.651
0.75	0.679
1.0	0.616

Tablica 11: Wpływ użytej części danych wykorzystywanych do budowy klasyfikatorów bazowych - zbiór *Glass*

Boosting

Learning rate	*FSC*
Stratified 2-Fold	
0.0001	0.466
0.001	0.447
0.01	0.442
0.1	0.453
1	0.443
Stratified 5-Fold	
0.0001	0.526
0.001	0.511
0.01	0.525
0.1	0.495
1	0.495
Stratified 10-Fold	
0.0001	0.571
0.001	0.565
0.01	0.567
0.1	0.551
1	0.564

Tablica 12: Wpływ wskaźnika uczenia się - zbiór *Glass*

Algorithm	*FSC*
Stratified 2-Fold	
SAMME	0.439
SAMME.R	0.454
Stratified 5-Fold	
SAMME	0.529
SAMME.R	0.504
Stratified 10-Fold	
SAMME	0.571
SAMME.R	0.543

Tablica 13: Wpływ stosowanego algorytmu - zbiór *Glass*

3.2 Zbiór Wine

Poniższe tabelki (Tabela 14, 15, 16, 17, 18, 19, 20, 21, 22) przedstawiają wszystkie przeprowadzone testy na zbiorze *Wine*, z podziałem na badane wartości parametru mówiącego o zastosowanej ilości estymatorów bazowych oraz na pozostałe testowane ustawienia poszczególnych parametrów każdej z zaimplementowanych metod zespołów klasyfikatorów. Wszystkie badania zrealizowane zostały przy pomocy sprawdzianów krzyżowych: kroswalidacja „zwykła" oraz stratyfikowana dla testowanego $n\_estimators$ oraz stratyfikowana dla pozostałych parametrów. W kolejnych kolumnach, każdej z wyżej wymienionych Tabel, umieszczone zostały odczyty metryki *FSC* (o której wspomniano w poprzedniej sekcji).

Dodatkowo na rysunku 6 przedstawione są wykresy prezentujące graficznie wpływ parametru $n\_estimators$ na osiągane wartości metryki *FSC* dla wszystkich badanych metod.

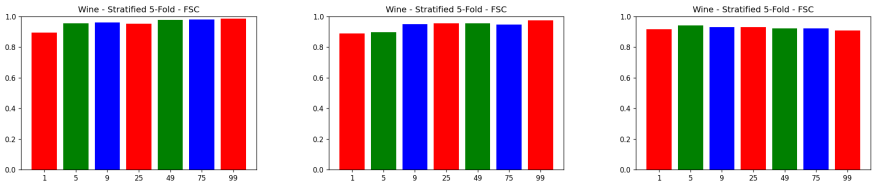
Dla zbioru *Wine*, przy wykorzystaniu algorytmów opartych o budowanie zespołów klasyfikatorów, również zauważamy raczej przyrost jakości klasyfikacji wraz ze wzrostem ilości estymatorów bazowych (poza przypadkiem kroswalidacji stratyfikowanej - 10 foldów). Wywnioskować można także, że dla lasów losowych, na zbiorze *Wine*, stosowanie bootstrapu nie jest wskazane. Jeżeli chodzi o metodę bagging - lepsze rezultaty dla bootstrapu oraz dla używanej części danych do budowy klasyfikatorów bazowych wynoszącej 0.5. W przypadku boostingu wskaźnik uczenia się oraz stosowany algorytmu zdaje się nie wpływać na wyniki dla tego zbioru (wszystkie są generalnie dobre).

n	RanFor	Bag	Boost
2-Fold			
1	0.198	0.228	0.183
5	0.223	0.184	0.183
9	0.176	0.19	0.183
25	0.185	0.193	0.183
49	0.178	0.192	0.183
75	0.186	0.189	0.183
99	0.184	0.191	0.183
5-Fold			
1	0.663	0.743	0.827
5	0.892	0.881	0.832
9	0.945	0.863	0.832
25	0.945	0.906	0.845
49	0.967	0.876	0.839
75	0.945	0.915	0.838
99	0.951	0.917	0.827
10-Fold			
1	0.832	0.848	0.85
5	0.9	0.882	0.856
9	0.929	0.906	0.856
25	0.962	0.903	0.861
49	0.955	0.91	0.851
75	0.973	0.932	0.856
99	0.957	0.915	0.861

Tablica 14: Wpływ ilości estymatorów bazowych na *FSC* - zbiór *Wine* - kroswalidacja „zwykła”

n	RanFor	Bag	Boost
Stratified 2-Fold			
1	0.873	0.796	0.934
5	0.935	0.928	0.927
9	0.926	0.893	0.921
25	0.95	0.922	0.927
49	0.966	0.933	0.923
75	0.977	0.933	0.927
99	0.961	0.939	0.927
Stratified 5-Fold			
1	0.951	0.874	0.876
5	0.905	0.946	0.888
9	0.951	0.95	0.869
25	0.966	0.939	0.882
49	0.961	0.962	0.865
75	0.967	0.945	0.875
99	0.977	0.956	0.89
Stratified 10-Fold			
1	0.917	0.905	0.867
5	0.957	0.922	0.873
9	0.955	0.939	0.868
25	0.966	0.946	0.884
49	0.972	0.968	0.868
75	0.978	0.967	0.868
99	0.983	0.962	0.856

Tablica 15: Wpływ ilości estymatorów bazowych na *FSC* - zbiór *Wine* - kroswalidacja stratyfikowana



(a) Random forest (b) Bagging (c) Boosting

Rysunek 6: Wpływ ilości estymatorów bazowych na *FSC* - zbiór *Wine*

Random forest

Bootstrap	*FSC*
Stratified 2-Fold	
True	0.961
False	0.972
Stratified 5-Fold	
True	0.961
False	0.967
Stratified 10-Fold	
True	0.978
False	0.984

Tablica 16: Wpływ stosowania bootstrapu - zbiór *Wine*

Max features	*FSC*
Stratified 2-Fold	
sqrt	0.966
log2	0.972
None	0.922
1.0	0.928
Stratified 5-Fold	
sqrt	0.983
log2	0.968
None	0.945
1.0	0.956
Stratified 10-Fold	
sqrt	0.984
log2	0.968
None	0.967
1.0	0.967

Tablica 17: Wpływ ograniczenia liczby badanych atrybutów podczas podziału - zbiór *Wine*

Criterion	*FSC*
Stratified 2-Fold	
gini	0.978
entropy	0.961
Stratified 5-Fold	
gini	0.967
entropy	0.973
Stratified 10-Fold	
gini	0.989
entropy	0.983

Tablica 18: Wpływ metody pomiaru jakości podziału - zbiór *Wine*

Bagging

Bootstrap	*FSC*
Stratified 2-Fold	
True	0.903
False	0.928
Stratified 5-Fold	
True	0.949
False	0.9
Stratified 10-Fold	
True	0.945
False	0.867

Tablica 19: Wpływ stosowania bootstrapu - zbiór *Wine*

Max samples	*FSC*
Stratified 2-Fold	
0.25	0.888
0.5	0.933
0.75	0.887
1.0	0.923
Stratified 5-Fold	
0.25	0.944
0.5	0.945
0.75	0.927
1.0	0.943
Stratified 10-Fold	
0.25	0.918
0.5	0.972
0.75	0.946
1.0	0.934

Tablica 20: Wpływ użytej części danych wykorzystywanych do budowy klasyfikatorów bazowych - zbiór *Wine*

Boosting

Learning rate	*FSC*
Stratified 2-Fold	
0.0001	0.934
0.001	0.921
0.01	0.923
0.1	0.934
1	0.934
Stratified 5-Fold	
0.0001	0.882
0.001	0.883
0.01	0.877
0.1	0.889
1	0.888
Stratified 10-Fold	
0.0001	0.866
0.001	0.861
0.01	0.867
0.1	0.861
1	0.878

Tablica 21: Wpływ wskaźnika uczenia się - zbiór *Wine*

Algorithm	*FSC*
Stratified 2-Fold	
SAMME	0.934
SAMME.R	0.921
Stratified 5-Fold	
SAMME	0.876
SAMME.R	0.895
Stratified 10-Fold	
SAMME	0.868
SAMME.R	0.879

Tablica 22: Wpływ stosowanego algorytmu - zbiór *Wine*

3.3 Zbiór Seeds

Poniższe tabelki (Tabela 23, 24, 25, 26, 27, 28, 29, 30, 31) przedstawiają wszystkie przeprowadzone testy na zbiorze *Seeds*, z podziałem na badane wartości parametru mówiącego o zastosowanej ilości estymatorów bazowych oraz na pozostałe testowane ustawienia poszczególnych parametrów każdej z zaimplementowanych metod zespołów klasyfikatorów. Wszystkie badania zrealizowane zostały przy pomocy sprawdzianów krzyżowych: kroswalidacja „zwykła" oraz stratyfikowana dla testowanego $n\_estimators$ oraz stratyfikowana dla pozostałych parametrów. W kolejnych kolumnach, każdej z wyżej wymienionych Tabel, umieszczone zostały odczyty metryki *FSC* (o której wspomniano w poprzedniej sekcji).

Dodatkowo na rysunku 7 przedstawione są wykresy prezentujące graficznie wpływ parametru $n\_estimators$ na osiągane wartości metryki *FSC* dla wszystkich badanych metod.

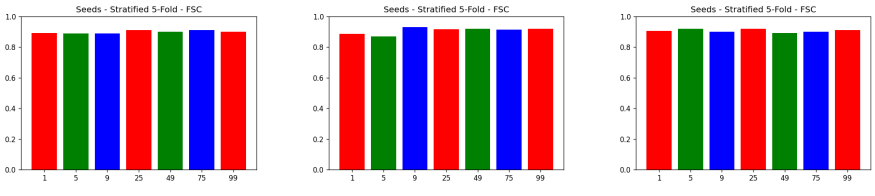
Dla zbioru *Seeds*, przy wykorzystaniu algorytmów opartych o budowanie zespołów klasyfikatorów, nie zauważamy tendencji w ocenie jakości klasyfikacji wraz ze zmianą ilości estymatorów bazowych. Wywnioskować można także, że dla lasów losowych, na zbiorze *Seeds*, stosowanie bootstrapu raczej nie jest wskazane. Jeżeli chodzi o metodę bagging - nie obserwujemy zależności wyników od użytego bootstrapu, natomiast - jak w każdym z poprzednich przypadków - dla używanej części danych do budowy klasyfikatorów bazowych wynoszącej 0.5 dostajemy najlepsze wyniki. W przypadku boostingu wskaźnik uczenia się zdaje się nie wpływać na wyniki dla tego zbioru (wszystkie są generalnie dobre). Raczej stosowany algorytm SAMME.R przekłada się na nieznacznie lepsze odczyty metryk.

n	RanFor	Bag	Boost
2-Fold			
1	0.262	0.246	0.262
5	0.256	0.26	0.257
9	0.246	0.251	0.249
25	0.253	0.264	0.257
49	0.246	0.264	0.251
75	0.251	0.263	0.262
99	0.248	0.265	0.263
5-Fold			
1	0.779	0.767	0.757
5	0.856	0.768	0.761
9	0.819	0.836	0.802
25	0.822	0.836	0.773
49	0.82	0.831	0.784
75	0.817	0.837	0.766
99	0.814	0.841	0.774
10-Fold			
1	0.816	0.839	0.867
5	0.88	0.876	0.855
9	0.91	0.857	0.857
25	0.886	0.895	0.856
49	0.863	0.895	0.862
75	0.882	0.9	0.861
99	0.859	0.895	0.861

Tablica 23: Wpływ ilości estymatorów bazowych na *FSC* - zbiór *Seeds* - kroswalidacja „zwykła”

n	RanFor	Bag	Boost
Stratified 2-Fold			
1	0.818	0.879	0.89
5	0.865	0.873	0.878
9	0.897	0.887	0.895
25	0.874	0.883	0.905
49	0.897	0.906	0.887
75	0.883	0.92	0.885
99	0.898	0.901	0.91
Stratified 5-Fold			
1	0.89	0.9	0.904
5	0.897	0.928	0.889
9	0.881	0.905	0.899
25	0.886	0.909	0.866
49	0.887	0.904	0.904
75	0.891	0.919	0.918
99	0.896	0.909	0.914
Stratified 10-Fold			
1	0.89	0.928	0.88
5	0.905	0.909	0.884
9	0.891	0.9	0.884
25	0.905	0.919	0.903
49	0.895	0.924	0.875
75	0.905	0.914	0.893
99	0.919	0.919	0.869

Tablica 24: Wpływ ilości estymatorów bazowych na *FSC* - zbiór *Seeds* - kroswalidacja stratyfikowana



(a) Random forest (b) Bagging (c) Boosting

Rysunek 7: Wpływ ilości estymatorów bazowych na *FSC* - zbiór *Seeds*

Random forest

Bootstrap	*FSC*
Stratified 2-Fold	
True	0.887
False	0.887
Stratified 5-Fold	
True	0.896
False	0.905
Stratified 10-Fold	
True	0.914
False	0.924

Tablica 25: Wpływ stosowania bootstrapu - zbiór *Seeds*

Max features	*FSC*
Stratified 2-Fold	
sqrt	0.897
log2	0.892
None	0.892
1.0	0.892
Stratified 5-Fold	
sqrt	0.886
log2	0.886
None	0.914
1.0	0.914
Stratified 10-Fold	
sqrt	0.919
log2	0.9
None	0.919
1.0	0.924

Tablica 26: Wpływ ograniczenia liczby badanych atrybutów podczas podziału - zbiór *Wine*

Criterion	*FSC*
Stratified 2-Fold	
gini	0.883
entropy	0.892
Stratified 5-Fold	
gini	0.896
entropy	0.891
Stratified 10-Fold	
gini	0.914
entropy	0.909

Tablica 27: Wpływ metody pomiaru jakości podziału - zbiór *Seeds*

Bagging

Bootstrap	*FSC*
Stratified 2-Fold	
True	0.874
False	0.881
Stratified 5-Fold	
True	0.881
False	0.895
Stratified 10-Fold	
True	0.914
False	0.904

Tablica 28: Wpływ stosowania bootstrapu - zbiór *Seeds*

Max samples	*FSC*
Stratified 2-Fold	
0.25	0.877
0.5	0.91
0.75	0.9
1.0	0.878
Stratified 5-Fold	
0.25	0.919
0.5	0.877
0.75	0.9
1.0	0.91
Stratified 10-Fold	
0.25	0.891
0.5	0.919
0.75	0.909
1.0	0.914

Tablica 29: Wpływ użytej części danych wykorzystywanych do budowy klasyfikatorów bazowych - zbiór *Seeds*

Boosting

Learning rate	*FSC*
Stratified 2-Fold	
0.0001	0.896
0.001	0.887
0.01	0.882
0.1	0.89
1	0.885
Stratified 5-Fold	
0.0001	0.904
0.001	0.894
0.01	0.913
0.1	0.894
1	0.909
Stratified 10-Fold	
0.0001	0.871
0.001	0.895
0.01	0.855
0.1	0.899
1	0.88

Tablica 30: Wpływ wskaźnika uczenia się - zbiór *Seeds*

Algorithm	*FSC*
Stratified 2-Fold	
SAMME	0.885
SAMME.R	0.892
Stratified 5-Fold	
SAMME	0.884
SAMME.R	0.899
Stratified 10-Fold	
SAMME	0.894
SAMME.R	0.904

Tablica 31: Wpływ stosowanego algorytmu - zbiór *Seeds*

Podsumowanie

Zadanie pozwoliło na zaznajomienie się z trzema metodami tworzenia zespołów klasyfikatorów (random forest, bagging, boosting) oraz również z kolejnymi narzędziami służącymi pracy na danych. Pokazane zostało jak wprowadzać dodatkowe funkcjonalności oraz rozszerzenia,które przekładają się na polepszenie oceny modelu. Wreszcie uświadomiło, że nie istnieje jeden sprawdzony przepis na wszystkie problemy, a tworzenie najlepszych rozwiązań jest związane z odszukaniem właściwego podejścia.

Na koniec, w Tabeli 32 zestawiono wyniki klasyfikatora opartego o algorytm k-nn z innymi, zaimplementowanymi w poprzednich zadaniach metodami, rozwiązującymi zagadnienie przyporządkowywania etykiet.

Classifier (*params*)	*ACC*	*PREC*	*REC*	*FSC*
Zbiór Glass				
MultinominalNB (smothing)	0.52	0.68	0.60	0.50
C4.5 ($C = 0.25$, $M = 2$)	0.68	0.68	0.80	0.74
***k*-nn** ($k = 5$, *distance*, *minkowski*)	0.70	0.64	0.65	0.63
Random forest ($n = 100$, *bootstrap*, *entropy*)	0.76	0.78	0.74	0.74
Bagging ($n = 99$, *bootstrap*)	0.72	0.71	0.72	0.72
Boosting ($n = 99$, *lr=1*, *SAMME.R*)	0.61	0.58	0.6	0.58
Zbiór Wine				
GaussianNB (smothing)	0.94	0.94	0.95	0.95
C4.5 ($C = 0.25$, $M = 2$)	0.94	0.97	0.97	0.97
***k*-nn** ($k = 20$, *uniform*, *minkowski*)	0.98	0.98	0.98	0.98
Random forest ($n = 100$, *bootstrap*, *gini*)	0.99	0.99	0.99	0.99
Bagging ($n = 49$, *bootstrap*)	0.97	0.97	0.97	0.97
Boosting ($n = 50$, *lr=1*, *SAMME*)	0.93	0.93	0.94	0.93
Zbiór Seeds				
GaussianNB (smothing)	0.89	0.90	0.90	0.90
C4.5 ($C = 0.10$, $M = 2$)	0.95	0.92	0.93	0.92
***k*-nn** ($k = 6$, *uniform*, *minkowski*)	0.93	0.93	0.93	0.93
Random forest ($n = 99$, *bootstrap*, *gini*)	0.92	0.92	0.92	0.92
Bagging ($n = 5$, *bootstrap*)	0.93	0.93	0.93	0.93
Boosting ($n = 75$, *lr=1*, *SAMME.R*)	0.92	0.92	0.92	0.92

Tablica 32: Porównanie klasyfikatorów dla zbiorów *Glass*, *Wine*, *Seeds*