Dobór klasyfikatora na podstawie metryk złożoności

Kacper Kołodyński, Szymon Witusiak

Politechnika Wrocławska

1 Przegląd literaturowy

1.1 Miary złożoności klasyfikacji

W najnowszej lieraturze miary złożoności klasyfikacji podzielone są na kilka grup, w zależności od cechy na jakiej są oparte. Ana C. Lorena, Luís P. F. Garcia, Jens Lehmann, Marcilio C. P. Souto, Tin K. Ho dokonali ich rozróżnienia i przeprowadzili badania pomiaru złożoności klasyfikacji [1]. W ich pracy przedstawiony został podział miar złożoności na sześć grup.

- Miary oparte na cechach(Feature-based measures), które opisują, jak informatywne są dostępne cechy do rozdzielenia klas. Miary te oceniają moc dyskryminacyjną cech. W wielu z nich każda cecha jest oceniana indywidualnie. Jeśli w zbiorze danych znajduje się przynajmniej jedna bardzo dyskryminująca cecha, problem można uznać za prostszy niż w przypadku, gdy nie ma takiego atrybutu. Wszystkie miary z tej kategorii wymagają, aby cechy miały wartości liczbowe. Większość z tych miar jest również zdefiniowana tylko dla problemów klasyfikacji binarnej.

- Miary liniowości(Linearity measures), które próbują ilościowo określić, czy klasy mogą być liniowo rozdzielone.Miary te próbują ilościowo określić, w jakim stopniu klasy są liniowo separowalne, czyli czy możliwe jest rozdzielenie klas za pomocą hiperpłaszczyzny. Motywowane są założeniem, że problem liniowo separowalny można uznać za prostszy niż problem wymagający nieliniowej granicy decyzyjnej.

- Miary sąsiedztwa(Neighborhood measures), które określają obecność i gęstość tych samych lub różnych klas w lokalnych sąsiedztwie. Miary te próbują uchwycić kształt granicy decyzyjnej i scharakteryzować nakładanie się klas poprzez analizę lokalnego sąsiedztwa punktów danych. Niektóre z nich wychwytują również wewnętrzną strukturę klas. Wszystkie pracują na macierzy odległości przechowującej odległości pomiędzy wszystkimi parami punktów w zbiorze danych.

- Miary sieciowe(Network measures), które wydobywają informacje strukturalne ze zbioru danych poprzez modelowanie go jako grafu. Aby wykorzystać

te miary, konieczne jest przedstawienie zbioru danych klasyfikacyjnych w postaci grafu. Otrzymany graf musi zachowywać podobieństwa lub odległości między przykładami w celu modelowania zależności między danymi. Każdy przykład z zestawu danych odpowiada węzłowi lub wierzchołkowi grafu, podczas gdy nieukierunkowane krawędzie łączą pary przykładów i są ważone przez odległości pomiędzy przykładami.

– Miary wymiarowości(Dimensionality measures), które oceniają rozproszenie danych w oparciu o liczbę próbek w stosunku do wymiarowości danych. Miary z tej kategorii dają orientacyjny obraz rozproszenia danych. Są one oparte na wymiarowości zbiorów danych, zarówno oryginalnych jak i zredukowanych. Idea jest taka, że wyodrębnienie dobrych modeli z rzadkich zbiorów danych może być trudniejsze, ze względu na prawdopodobną obecność regionów o niskiej gęstości, które będą arbitralnie klasyfikowane.

– Miary nierównowagi klas(Class imbalance measures), które uwzględniają stosunek liczby przykładów pomiędzy klasami. Miary te próbują uchwycić jeden aspekt, który może w dużym stopniu wpłynąć na predykcyjną wydajność technik ML podczas rozwiązywania problemów klasyfikacji danych: nierównowaga klas, czyli duża różnica w liczbie przykładów na klasę w zbiorze danych treningowych. W istocie, gdy różnice są poważne, większość technik klasyfikacji ML ma tendencję do faworyzowania klasy większościowej i występują problemy z generalizacją.

1.2 Analiza miar złożoności danych w kontekście klasyfikacji

Już wcześniej podjęto się analizy zależności między miarami złożoności danych a zachowaniem klasyfikatorów. José-RamónCano w swojej pracy [2] ocenia każdą z metryk w zakresie jej wartości i bada dokładność klasyfikatorów na tych wartościach. Wyniki dostarczają informacji o przydatności tych miar, oraz o tym, które z nich pozwalają na analizę charakteru zbioru danych wejściowych i pomagają w podjęciu decyzji, która metoda klasyfikacji może być najbardziej obiecująca. Autor bada miary złożoności biorąc pod uwagę ocenę przydatności metryki i jej tradycyjne zastosowanie. W tym celu, analizowana jest każda z miar złożoności zaproponowanych przez Ho i Basu (2002), obejmując jej zakres wartości, oraz badamy związek pomiędzy tymi miarami a wydajnością klasyfikatorów.

1.3 Metody zespołowe uczenia maszynowego

W kontekście tworzenia klasyfikatorów służących do testowania metrykami złożoności wykorzystywane są zespołowe metody uczenia maszynowego. Są to metody, które konstruują zestaw klasyfikatorów, a następnie klasyfikują nowe punkty danych poprzez (ważone) głosowanie ich predyktorów. Oryginalną metodą zespołu jest uśrednianie Bayesa, ale nowsze algorytmy obejmują kodowanie wyjściowe z korekcją błędów, Bagging i boosting. W większości przypadków podstawowe modele same w sobie nie działają tak dobrze, ponieważ mają duże obciążenie

(na przykład modele o niskim stopniu swobody) lub dlatego, że mają zbyt dużą wariancję, aby były solidne (na przykład modele o wysokim stopniu swobody). Zespół klasyfikatorów to taki w którym indywidualne decyzje pojedynczych klasyfikatorów są łączone w jakiś sposób, zazwyczaj poprzez ważone lub nieważone głosowanie, w celu sklasyfikowania nowych przykładów. Jednym z najbardziej aktywnych obszarów badań w zakresie uczenia zespołowego jest badanie metod konstruowania dobrych zespołów klasyfikatorów. Została udowodniona hipoteza że zespoły klsyfikatorów są dokładniejsze niż pojedyncze klasyfikatory, które je tworzą [3] [4].

1.4 Pomiar złożoności danych dla problemów klasyfikacji z niezrównoważonymi danymi

W literaturze znajduje się szereg różnych powiązanych ze sobą złożoności zestawów danych. W artykule [5], którego autorem jest Nafees Anwar, Geoff Jones oraz Siva Ganesh zostaje poruszona tematyka miar złożoności dla problemów klasyfikacji uwzględniającą pogorszenie klasyfikatora wynikające z nierównowagi klasowej. W problemach klasyfikacji dwugrupowej dane uczące są wykorzystywane do opracowania reguły przypisywania podmiotów lub przypadków do jednej z dwóch grup lub klas na podstawie informacji współzmiennych, tj. zmiennych lub cech zarejestrowanych dla każdego podmiotu. Nierównowaga klas występuje, gdy przypadki należące do jednej klasy w danych uczących znacznie przewyższają liczbę przypadków należących do drugiej klasy. Problem z niezrównoważonymi zbiorami danych polega na tym, że konwencjonalne klasyfikatory mają na celu maksymalizację ogólnej dokładności, co często osiąga się przez przypisanie wszystkich lub prawie wszystkich przypadków do klasy większościowej. Tak więc w sytuacjach nierównowagi klasowej występuje tendencja do uprzedzeń wobec klasy mniejszościowej. W wielu takich sytuacjach klasa mniejszości jest klasą pozytywną. Dlatego często lepiej jest podkreślać dobre wyniki w klasie mniejszości przy zachowaniu rozsądnej ogólnej dokładności.

1.5 Wpływ baggingu oraz boostingu na wydajność klasyfikacji w niezrównoważonej klasyfikacji binarnej

Klasyfikatory Bagging i Boosting w ostatnich czasach zyskały na popularności ze względu na ich odporność na niezrównoważone etykiety klas. Obie metody wykorzystują pojęcie zespołu do uogólnienia modelu i przewidywania na podstawie niewidocznych danych. Artykuł [7] opisuje badania mające na celu poprawienie wydajności klasyfikacji poprzez baggig oraz boosting klasyfikatorów na niezrównoważonych zestawach danych klasyfikacji binarnej. Poruszony zostaje w nim problem nieoznaczoności klas powodujących pogorszenie wydajności w problemie uczenia się i klasyfikacji. Przyczyną powstałego często bywają zestawy danych, które pochodzą ze świata rzeczywistego, w którym klasy pozytywne są często rzadkie lub nie w tej samej skali negatywności. Okazują się, że klasyfikatory bagging oraz boosting zwracają obiecujące rezultaty przeciwko nieoznakowanym klasom. Warto również dodać, że zarówno bagging jak i boosting dają

bardziej precyzyjne wyniki w przeważającej klasie pozytywnej niż standardowe algorytmy uczenia się.

1.6 Praktyczne zastosowanie metod uczenia zespołowego

Istnieją różne rodzaje uczenia maszynowego, takie jak uczenie nadzorowane, uczenie nienadzorowane, uczenie częściowo nadzorowane i uczenie wzmacniające. Każdy typ algorytmu ML jest używany do rozwiązywania określonego rodzaju problemów; niektóre algorytmy mogą być używane do klasyfikacji, inne do regresji, a niektóre do grupowania. Wybór odpowiedniego algorytmu zależy od rodzaju problemu i wielu innych czynników, takich jak parametryzacja, czas uczenia się, czas przewidywania, tendencja do nadmiernego dopasowania i wielkość pamięci. Metody zespołowego mają na celu poprawę predykcji wydajności dla zadanych algorytmów klasyfikacji. Przykładowymi aglorytmami nauczania zespołowego mogą być:
-**Adaboost -** Jednorodny uczeń, który tworzy serię klasyfikatorów mające na celu poprawę dokładności klasyfikatora. W zależności od wydajności każdego klasyfikatora, zestaw treningowy zostanie wybrany. Nieprawidłowo sklasyfikowana próbka będzie wybierane częściej niż prawidłowo sklasyfikowane próbki. W rezultacie nowy klasyfikator utworzony przez wzmocnienie algorytm, który działa dobrze na nowym zbiorze danych. Używając głosów większością ważoną, wzmocnienie wpłynie na klasyfikator.
- **Algorytm Bootstrap Agregating-Bagging -** Jednorodny słaby uczeń, który generuje próbkowanie instancji ze zbioru uczącego w celu uzyskania zagregowanego predyktora, który jest uzyskiwany przy użyciu reguły głosowania większościowego. Bagging sprawdza się bardzo dobrze w modelach 'overfit', ponieważ pracuje nad zmniejszeniem błędu średniokwadratowego wariancji dla danej operacji, takie jak drzewa decyzyjne lub inny algorytm, wybierając zmienną i układając je w model liniowy.

Algorytmy uczenia zespołowego przyczyniły się do przewidywania w wielu sektorach. Bagging i Boosting znacznie poprawiają przewidywanie rezygnacji, gdy są stosowane w bazie danych klientów amerykańskiej firmy telekomunikacyjnej Lemmens i Croux 2006. Badanie polegające na przewidywaniu produktywności pracowników w oparciu o metody uczenia maszynowego zostało opisane w [6], gdzie zostały wykorzystane metody wspomnianych w poprzednich sekcjach bagging oraz boosting.

1.7 Redundancja oraz złożoność metryk dla dużych zbiorów danych

Pomimo łatwości znajdowania oraz gromadzenia dużych ilości danych w wielu dziedzinach dane te wymagają wstępnego przetworzenia - odrzucenia tych próbek, które są destrukcyjne, i wybrać dane, które dostarczają wysokiej jakości informacji do uczenia maszynowego. Zgodnie z badaniem przeprowadzonym za pomocą pewnych metryk w [8], duże zbiory danych często pokazują informacje o redundancji w swoich próbkach. Ta wysoka redundancja pozwala zredukować rozmiar próbek do 25% bez drastycznego wpływu na dokładność uzyskiwaną przez klasyfikatory, osiągając znacznie szybsze czasy pracy. Pokazuje to, że liczba ich wystąpień w dużych zbiorach danych jest większa niż to konieczne, podkreśla również potrzebę nadania priorytetu technikom wstępnego przetwarzania w celu uzyskania inteligentnych danych. Należy podkreślić redundancję w wielu problemach klasyfikacji dużych zbiorów danych, gdzie przy znacznie mniejszym zbiorze danych o małej jakości możemy uzyskać podobne lub lepsze wyniki. W przypadku big data wyzwaniem jest uzyskanie inteligentnych danych o minimalnym wymaganym rozmiarze. Posiadanie większej ilości danych nie zawsze jest równoznaczne z uzyskaniem bardziej istotnych informacji, może za to spowodować niepotrzebnie większy koszt obliczeniowy.

2 Plan eksperymentu

Eksperyment ma na celu zbadanie wpływu uwzględnienia metryki złożoności klasyfikatorów (*complexity*) na jakość klasyfikacji. Ze względu na złożoność oraz na szeroki zakres projektu, jesteśmy w stanie go podzielić na dwa potencjalnie niezależne eksperymenty.

2.1 Research Questions

1. Czy złożoność metryk ma wpływ na jakość (balanced accuracy) klasyfikatorów? Jeśli tak to, jaka jest zależność pomiędzy nimi?
2. W jaki sposób powinny zostać wybrane zestawy danych, aby wyniki był miarodajne? Dlaczego?
3. Jaki jest wpływ wykorzystanej metryki złożoności na uzyskane wyniki?
4. Jaki wpływ mają zestawy danych na otrzymane w eksperymencie drugim wyniki?

2.2 Cel eksperymentu

Eksperyment 1.
Idea pierwszego z dwóch eksperymentów zakłada stworzenie syntetycznego zbioru danych, a następnie wykonanie na nim metody uczenia zespołowego, czyli "baggingu". Oznacza to, że na stworzonym zbiorze wylosowana zostałaby pewna ilość podprzestrzeni wspomnianego zbioru, po czym policzone zostałyby za pomocą odpowiednich metod złożoności danych podzbiorów. Na podstawie otrzymanych złożoności, zbadana zostałaby ich relacja z jakością klasyfikatora. Celem

tej części eksperymentu jest zbadanie wspomnianej wcześniej relacji oraz stworzenie pewnej reguły polegającej na wyznaczeniu zależności pomiędzy "balanced accuracy" na podstawie stworzonych charakterystyk.

Eksperyment 2.

Otrzymana z eksperymentu pierwszego reguła ma znaczący wpływ na eksperyment drugi, ponieważ wyznaczona zasada byłaby pomocna w nadawaniu wag klasyfikatorom podczas eksperymentu drugiego. Jednakże tak jak wcześniej zostało zaznaczone, dwa eksperymenty mogą być od siebie niezależne, pod warunkiem że podczas przeprowadzania drugiego badania reguła z eksperymentu pierwszego stałaby się założeniem, stworzonym wbrew intuicji zamiast obliczeń. Według intuicji można przypuszczać, że podprzestrzeń trudniejsza, czyli ta o większej złożoności powinna mieć niższą wagę, a ta, która jest łatwiejsza powinna mieć niższą. Podejście do problemu z ten sposób, może spowodować faworyzowanie klasyfikatorów, które niczego się nie nauczyły. Dlatego też powstałą dygresja czy nie jest lepszym rozwiązaniem założyć wbrew intuicji i zastosować podejście, w którym wyższa waga jest nadawana przy mniejszej złożoności i tym łatwiejszy klasyfikator. W ten sposób można zbadać, która metryka wpływa na otrzymanie lepszej puli.

Założenie eksperymentu pierwszego zakłada, odwrotną proporcjonalność złożoności podzbioru do jakości, oznacza to, że im większa złożoność podprzestrzeni tym większa jest jakość. Istnieje jednak możliwość, że założenie jest błędne i w takim przypadku gdy wyniki otrzymane z eksperymentu polegającego na tym założeniu byłyby nie satysfakcjonujące, przeprowadzone by zostały te same badania na podstawie założenia odwrotnego, które mówi że złożoność podprzestrzeni jest proporcjonalna do jakości. Celem eksperymentu drugiego jest porównanie wyników eksperymentu bazującego na dwóch odmiennych założeniach oraz przeanalizowanie otrzymanych wyników. Celem projektu jest przeprowadzenie drugiego z przedstawionych badań oraz w przypadku niejednoznaczności otrzymanych wyników wykonać eksperyment pierwszy w celu uzasadnienia rezultatów.

2.3 Zestawy danych użyte do eksperymentu

Zbiory danych wykorzystane w eksperymencie pochodzą ze strony KEEL zawierającej zestawy danych do uczenia maszynowego. Do eksperymentu wybrano zbiory danych o wyłącznie dwóch klasach, bez brakujących danych. Przy tych założeniach wybrane zostały następujące zbiory danych:

- **prima -** z Narodowego Instytutu Cukrzycy i Chorób Trawiennych i Nerek. Wszyscy pacjenci to kobiety w wieku co najmniej 21 lat, pochodzenia Indian Pima. Etykieta klasy wskazuje, czy dana osoba nie ma cukrzycy, czy też ma cukrzycę.

- **bupa -** zestaw danych zawiera dane na temat niektórych zaburzeń wątroby i ilości wypijanego alkoholu. Etykieta klasy wskazuje, czy dana osoba cierpi na alkoholizm czy nie.

Data set	Samples	Features
prima	768	8
bupa	345	6
monk-2	432	6
wisconsin	683	9
mammographic	830	5
bands	365	19
breast	277	9
heart	270	13
hepatitis	80	19
phoneme	5404	5
wdbc	569	30

– **monk-2 -** zbiór trzech binarnych problemów sztucznej klasyfikacji, które należą do klasy 0 lub klasy 1.

– **wisconsin -** zestaw danych zawiera przypadki z badania przeprowadzonego w szpitalu University of Wisconsin w Madison, dotyczącego pacjentów, którzy przeszli operację raka piersi. zbiór zawiera dwie klasy, guz jest łagodny lub złośliwy.

– **mammographic -** zestaw danych zawiera dwie klasy (łagodny lub złośliwy) guza mammograficznego oraz opis atrybutów BI-RADS i wieku pacjentki. Dane zostały zebrane w Instytucie Radiologii Uniwersytetu Erlangen-Norymberga w latach 2003-2006.

– **bands -** problem klasyfikacyjny z druku rotograwiurowego, gdzie zadaniem jest określenie czy dany fragment jest wstęgą cylindryczną.

– **breast -** zawiera 201 instancji jednej klasy i 85 instancji innej klasy. Instancje są opisane przez 9 atrybutów, z których część jest liniowa, a część nominalna.

– **heart -** baza danych zawiera informacje o pacjentach potencjalnie chorych na serce oraz przyporządkowanie ich do dwóch klas (chory lub zdrowy).

– **hepatitis -** zestaw danych zawiera zbiór atrybutów, z informacjami o pacjentach dotkniętych chorobą zapalenia wątroby, oraz przyporządkowanie ich do jednej z dwuch klas (umrą lub przeżyją).

– **phoneme -** baza zawiera opis próbki głosu i przyporządkowanie do jednej z dwóch klas (dźwięk nosowy lub ustny).

– **wdbc -** baza danych zawiera 30 cech obliczonych na podstawie cyfrowego obrazu aspiratu cienkoigłowego (FNA) masy piersi. Opisują one cechy jąder komórkowych obecnych na obrazie. Dane dzielą się na dwie klasy (guz łagodny lub guz złośliwy)

Niektóre z wyżej wymienionych zbiorów posiadają etykiety opisane tekstowo, przed przystąpieniem do eksperymentyu zostaną one zamienione w '0' lub '1', w zależności odf etykiety.

2.4 Wykonanie eksperymentu

Z 20 podprzestrzeni uzyskane zostanie 20 klasyfikatorów. Dla każdego z klasyfikatorów zostanie policzone *complexity* - wynik metryki złożoności na danym klasyfikatorze. Wykorzystana metryka złorzoności nosi nazwę Średnia gęstość sieci (Density). Dla każdego zbioru danych przeprowadone zostanie uczenie zbioru klasyfikatorów, a następnie jego testowanie na wcześniej wyodrębnionym zbiorze testowym. Jako protokół eksperymentalny wybrano walidację krzyżową 5×2. Dokonanie klasyfikacji polega na pomnożeniu wyniku każdego z dwudziestu klasyfikatorów przez odpowiadającą mu wagę (wyliczoną z metryki złożoności), a następnie metodą głosowania większościowego (majority voting) przy ważonych etykietach, wybrania jednej z dwóch klas.
Dla otrzymanych danych przeprowadzone zostaną testy rankingowe. Do porównania jakości działania klasyfikatorów użyte zostaną metryki takie jak dokładność zrównoważona (balanced accuracy). Jest ona używana do oceny klasyfikatorów w problemach z niezrównoważonymi danymi. Wszystkie Wyniki zostaną zweryfikowane pod kątem istotności statystycznej za pomocą testu t-Studenta przy poziomie istotności $p < 0.05$.

2.5 Środowisko eksperymentalne

Eksperyment zostanie przeprowadzony za pomocą języka Python, biblioteki scikit-learn służącej do predykcyjnej analizy danych. Ponadto ze względu na specyfikację i charakter przeprowadzanego projektu, zostanie w nim wykorzystana biblioteka "problexity" [1].

[1] https://github.com/w4k2/problexity

3 Wyniki badań i analizy statystycznej

Dla każdego zbioru danych, dla trzech metod -baggingu, baggingu ważonego metrykami, baggingu odwrotnie ważonego metrykami otrzymano po dziesięć wartości (po dwie dla każdego foldu) dokładności zrównoważonej (ballanced accuracy score).

3.1 Test t-Studenta

Hipoteza 0: Należnośc do wspólnych rozkładów.
Wartośc p - Prawdopodobieństwo przyjęcia hipotezy 0
Hipoteza 0 odrzucamy, oznacza to, że nie należą do wspólnych rozkładów, a wyniki są istotnie różne statystycznie. Jeśli wartość testu jest dodatnia, to znaczy, że pierwsza metoda jest lepsza od drugiej. Natomiast, jeśli zwrócona zostaje wartość ujemna, oznacza to, że druga metoda jest lepsza od pierwszej.
Jeśli wartość p jest mniejsze niż zadany próg, czyli 0.05, a wynik statystyki jest dodatni to oznacza, że pierwsza metoda jest istotnie statystycznie lepsza niż druga.

Za pomocą testu t-Studenta porównano parowo wymienione metody głosowania większościowego. Otrzymano następujące wyniki:

- WM - bagging ważony metrykami
- OWM - bagging odwrotnie ważony metrykami
- BAG - bagging

Tabela 1. bands - Statystyka

	WM	OWM	BAG
WM	0	-0,227	-0.141
OWM	0,227	0	0.064
BAG	0.141	-0.064	0

Tabela 2. bands - Warość p

	WM	OWM	BAG
WM	1	0.822	0.889
OWM	0.822	1	0.949
BAG	0.889	0.949	1

Tabela 3. ionosphere - Statystyka

	WM	OWM	BAG
WM	0	1,085	-0.242
OWM	-1.085	0	-1,227
BAG	0.242	1,227	0

Tabela 4. ionosphere - Wartość p

	WM	OWM	BAG
WM	1	0.292	0.811
OWM	0.292	1	0.235
BAG	0.8101	0.235	1

Tabela 5. bupa - Statystyka

	WM	OWM	BAG
WM	0	2,097	-0,739
OWM	-2,097	0	-3,031
BAG	1,739	3,031	0

Tabela 6. bupa - Wartość p

	WM	OWM	BAG
WM	1	0.050	0.098
OWM	0.050	1	0.007
BAG	0.098	0.007	1

Tabela 7. heart - Statystyka

	WM	OWM	BAG
WM	0	0.498	0.233
OWM	-0.498	0	-0.254
BAG	-0.233	0,254	0

Tabela 8. heart - Wartość p

	WM	OWM	BAG
WM	1	0,624	0.818
OWM	0.624	1	0.801
BAG	0.818	0.801	1

Tabela 9. hepatitis - Statystyka

	WM	OWM	BAG
WM	0	1.720	-0.065
OWM	-1.720	0	-1,806
BAG	0.065	1.806	0

Tabela 10. hepatitis - Wartość p

	WM	OWM	BAG
WM	1	0.102	0.948
OWM	0.102	1	0.087
BAG	0.948	0.087	1

Tabela 11. mammographic - Statystyka

	WM	OWM	BAG
WM	0	0,265	-0.602
OWM	-1,265	0	-2.091
BAG	0,602	2.091	0

Tabela 12. mammographic - Wartość p

	WM	OWM	BAG
WM	1	0.221	0.554
OWM	0.221	1	0.050
BAG	0.554	0.050	1

Tabela 13. prima - Statystyka

	WM	OWM	BAG
WM	0	-0,412	-2,230
OWM	0.412	0	-1.791
BAG	2,230	1.791	0

Tabela 14. prima - Wartość p

	WM	OWM	BAG
WM	1	0,685	0.038
OWM	0.685	1	0.090
BAG	0.038	0,090	1

Tabela 15. phoneme - Statystyka

	WM	OWM	BAG
WM	0	-0,082	-1.746
OWM	0,082	0	-1.824
BAG	1.746	1.824	0

Tabela 16. phoneme - Wartość p

	WM	OWM	BAG
WM	1	0.954	0.097
OWM	0.935	1	0.084
BAG	0.097	0.084	1

Tabela 17. wdbc - Statystyka

	WM	OWM	BAG
WM	0	1,638	0,240
OWM	-1,638	0	-1,317
BAG	-0,240	1,317	0

Tabela 18. wdbc - Wartość p

	WM	OWM	BAG
WM	1	0.118	0.812
OWM	0.118	1	0.204
BAG	0,812	0.204	1

Tabela 19. wisconsin - Statystyka

	WM	OWM	BAG
WM	0	4,302	-0.244
OWM	-4,302	0	-4,265
BAG	0.244	4,265	0

Tabela 20. wisconsin - Wartość p

	WM	OWM	BAG
WM	1	0.001	0.809
OWM	0.001	1	0.001
BAG	0.809	0.001	1

Tabela 21. appendicitis - Statystyka

	WM	OWM	BAG
WM	0	1.999	0.133
OWM	-1.999	0	-1,780
BAG	-0.133	1,780	0

Tabela 22. appendicitis - Wartość p

	WM	OWM	BAG
WM	1	0.060	0.895
OWM	0.060	1	0.091
BAG	0.895	0.091	1

Tabela 23. titanic - Statystyka

	WM	OWM	BAG
WM	0	-0,546	0,059
OWM	0,546	0	0,607
BAG	-0,059	-0,607	0

Tabela 24. titanic - Wartość p

	WM	OWM	BAG
WM	1	0.591	0.953
OWM	0.591	1	0.551
BAG	0.953	0.551	1

3.2 Tabela najlepszych metod dla poszczególnych zbiorów danych

Uwzględnione są tylko przypadki istotne statystycznie.

bands	WM	OWM	BAG
WM	0	0	0
OWM	0	0	0
BAG	0	0	0

ionosphere	WM	OWM	BAG
WM	0	0	0
OWM	0	0	0
BAG	0	0	0

bupa	WM	OWM	BAG
WM	0	1	0
OWM	0	0	0
BAG	0	1	0

heart	WM	OWM	BAG
WM	0	0	0
OWM	0	0	0
BAG	0	0	0

hepatitis	WM	OWM	BAG
WM	0	0	0
OWM	0	0	0
BAG	0	0	0

mammographic	WM	OWM	BAG
WM	0	0	0
OWM	0	0	0
BAG	0	1	0

prima	WM	OWM	BAG
WM	0	0	0
OWM	0	0	0
BAG	1	0	0

phoneme	WM	OWM	BAG
WM	0	0	0
OWM	0	0	0
BAG	0	0	0

wdbc	WM	OWM	BAG
WM	0	0	0
OWM	0	0	0
BAG	0	0	0

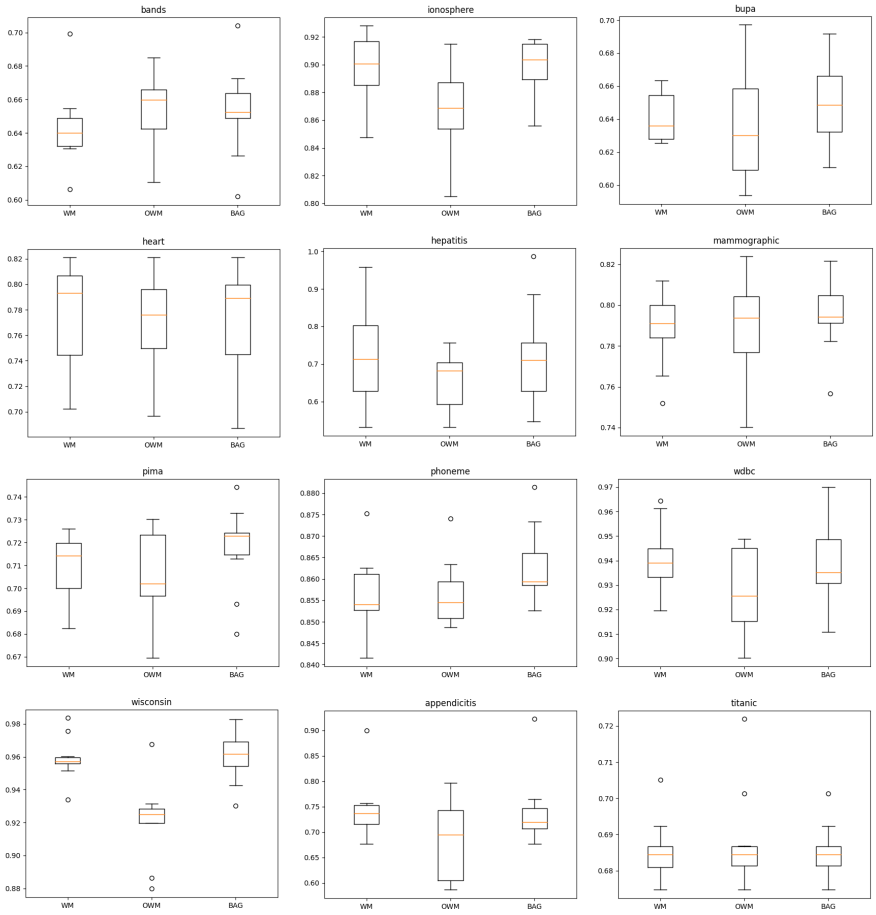
wisconsin	WM	OWM	BAG
WM	0	1	0
OWM	0	0	0
BAG	0	1	0

appendictis	WM	OWM	BAG
WM	0	0	0
OWM	0	0	0
BAG	0	0	0

titanic	WM	OWM	BAG
WM	0	0	0
OWM	0	0	0
BAG	0	0	0

sum (p<0.05)	WM	OWM	BAG
WM	0	2	0
OWM	0	0	0
BAG	1	3	0

sum (p<0.1)	WM	OWM	BAG
WM	0	3	0
OWM	0	0	0
BAG	3	7	0



Rysunek 1. Wykresy kubełkowe

4 Wnioski

Prosty bagging daje lepsze rezultaty niż ważony metrykami w 1 przypadku, a niż odwrotnie ważony metrykami w 3 przypadkach. Bagging ważony metrykami daje lepsze rezultaty niż odwrotnie ważony w 2 przypadkach. Dla poziomu istotności $p < 0.1$ sytuacja się pogłębia. W żadnym przypadku bagging ważony metrykami, lub odwrotnie ważony metrykami nie daje lepszych rezultatów niż zwykły bagging. Wynika z tego że wykorzystanie wag związanych z metryką Density w głosowaniu większościowym klasyfikatorów pogarsza wyniki. Złożoność zbioru danych (Density) nie ma wpływu na jakość klasyfikatora na nim wyuczonego. Zbiory danych na których zaobserwowano różnicę to wisconsin, bupa, prima i mammographic.

Literatura

1. Ana C. Lorena, Luís P. F. Garcia, Jens Lehmann, Marcilio C. P. Souto, Tin K. Ho, How Complex Is Your Classification Problem?: A Survey on Measuring Classification Complexity (2019)
2. José-RamónCano, Analysis of data complexity measures for classification (2013)
3. Thomas G Dietterich, Ensemble Methods in Machine Learning (2000)
4. Edouard Duchesnay, Ensemble learning: bagging, boosting and stacking (2020)
5. Nafees Anwar, Geoff Jones and Siva Ganesh, Measurement of Data Complexity for Classification Problems with Unbalanced Data (2014)
6. Ruba Obiedat, A Combined Approach for Predicting Employees' Productivity based on Ensemble Machine Learning Methods (2022)
7. Yash Singhal, Ayushi Jain, Shrey Batra, Yash Varshney, Megha Rathi, Review of Bagging and Boosting Classification Performance on Unbalanced Binary Classification (2018)
8. Isaac Triguero, Francisco Herrera, Jesus Maillo, Redundancy and Complexity Metrics for Big Data Classification: Towards Smart Data (2020)