UMA - Projekt

Mikołaj Garbowski, Michał Pałasz

Semestr 2024Z

1 Temat projektu

Implementacja drzewa decyzyjnego, porównanie sposobu radzenia sobie z problemami wieloklasowymi, czyli porównanie jakości wyników typowej implementacji ID3 z jakością wyników dwóch podejść:

1) tworzymy osobny model binarny dla każdej klasy (jedna klasa traktowana jako pozytywna, wszystkie pozostałe jako negatywne), predykcja przez wybór klasy o maksymalnej wartości funkcji decyzyjnej (wymaga posiadania przez każdy klasyfikator „stopnia pewności siebie", co można zdefiniować na wiele sposobów). 2) tworzymy osobny model binarny dla każdej pary klas (jedna klasa traktowana jako pozytywna, druga jako negatywna), predykcja przez głosowanie.

2 Opis rozwiązania

2.1 Algorytm ID3

Przy tworzeniu liścia, poza klasą większościową zapamiętujemy również częstość występowania przykładów klasy większościowej w zbiorze przykładów rozważanym w danym liściu. W przypadku jednakowej klasy dla wszystkich przykładów - wartość 1, w pozostałych przypadkach - wartość z przedziału $(0, 1)$. Częstość potraktujemy jako stopień pewności modelu co do decyzji.

2.2 Wariant One vs Rest

Dla problemu klasyfikacji z liczbą klas równą n powstanie n modeli klasyfikacji binarnej według algorytmu ID3.

Dla pojedynczego modelu dla klasy c, modyfikujemy etykiety w zbiorze danych - wszystkie etykiety klas innych niż c zamieniamy na klasę negatywną, c traktujemy jako klasę pozytywną.

Wynikiem predykcji zespołu modeli będzie ta klasa, którą model binarny zaklasyfikował pozytywnie z największą pewnością.

Przykładowo, dla klas A, B, C, D

Model binarny dla klasy	A	B	C	D
Predykcja (0/1)	1	0	1	0
Pewność	0,8	0,7	0,7	0,9

Predykcją zespołu modeli będzie klasa A, ponieważ zarówno model binarny dla klasy A i C dał pozytywny wynik klasyfikacji, ale model dla klasy A miał większą pewność.

Jeśli wszystkie modele binarne dadzą predykcję negatywną, predykcją zespołu modeli będzie ta klasa, która została zaklasyfikowana negatywnie z najmniejszą pewnością. Przykład:

Model binarny dla klasy	A	B	C	D
Predykcja (0/1)	0	0	0	0
Pewność	0,9	0,7	0,8	0,5

Predykcją zespołu modeli będzie klasa D, ponieważ wszystkie predykcje dają klasę negatywną, ale model dla klasy D ma najmniejszą pewność

2.3 Wariant One vs One

Dla problemu klasyfikacji z liczbą klas równą n powstanie $n(n-1)/2$ modeli klasyfikacji binarnej - po 1 dla każdej pary klas. Klasyfikator binarny rozstrzyga do której klasy z pary należy przykład.

Do trenowania klasyfikatora binarnego dla pary klas A i B użyjemy takiego podzbioru zbioru trenującego, który zawiera tylko przykłady klas A i B.

Predykcja zespołu klasyfikatorów będzie wyznaczana przez głosowanie. Ze względu na możliwość remisu przy zliczaniu głosów jako liczby modeli które przewidują daną klasę, proponujemy poniższy sposób obliczania głosów ważonych stopniem pewności predykcji (stopień predykcji definiowany jak w poprzednim wariancie).

Przykład dla klas A, B i C:

Model binarny dla pary	A vs B	B vs C	C vs A
Predykcja	A	B	C
Pewność	0,99	0,8	0,7

Klasa	Ważony głos
A	1,29 = 0,99 + (1-0,7)
B	0,81 = (1-0,99) + 0,8
C	0,9 = (1-0,8) + 0,7

Dla powyższego przykładu predykcją zespołu klasyfikatorów byłaby klasa A, ponieważ choć wszystkie uzyskały po 1 głosie, dla każdej klasy obliczamy sumę pewności głosów *za* i dopełnienia do 1 pewności głosów *przeciw*.

3 Zbiory danych

3.1 Car Evaluation

Celem zbioru jest klasyfikacja samochodów na podstawie ich cech, takich jak cena, koszty utrzymania czy pojemność pasażerska, do jednej z czterech kategorii akceptowalności

- `unacc` (70%)
- `acc` (22%)
- `good` (4%)
- `vgood` (4%)

Charakterystyka zbioru danych:

- **Liczba próbek:** 1728
- **Liczba cech:** 6 atrybutów wejściowych + 1 cecha docelowa
- **Typy danych:** kategoryczne

3.2 Balance Scale

Celem zbioru jest klasyfikacja stanu równowagi szalki wagi na podstawie masy i odległości obiektów umieszczonych na jej lewym i prawym ramieniu. Każda próbka jest przypisywana do jednej z trzech kategorii:

- `L` - szalka przechylona w lewo (46%)
- `B` - szalka w równowadze (8%)
- `R` - szalka przechylona w prawo (46%)

Charakterystyka zbioru danych:

- **Liczba próbek:** 625
- **Liczba cech:** 4 atrybuty wejściowe + 1 cecha docelowa
- **Typy danych:** kategoryczne

3.3 National Poll on Healthy Aging (NPHA)

Celem zbioru jest przewidywanie liczby odwiedzanych lekarzy przez Amerykanów powyżej 50 roku życia na podstawie odpowiedzi respondentów w ankiecie na temat ich zdrowia. Liczba lekarzy jest sprowadzona do 3 kategorii:

- `1` - 0 lub 1 (18%)
- `2` - 2 lub 3 (52%)
- `3` - 4 lub więcej (30%)

Charakterystyka zbioru danych:

- **Liczba próbek:** 714
- **Liczba cech:** 14 atrybutów wejściowych + 1 cecha docelowa
- **Typy danych:** kategoryczne

Żaden z powyższych zestawów danych nie posiada brakujących danych.

4 Wyniki

4.1 Macierze pomyłek

Przykładowe macierze pomyłek dla wszystkich zbiorów danych i klasyfikatorów. Prezentowane macierze pomyłek są otrzymane z jednokrotnego wykonania predykcji przy podejściu z oddzielnym zbiorem uczącym i testowym (podzielone w proporcji 0.6).

	B	L	R
B	0	13	9
L	8	92	20
R	2	25	81

(a) ID3: Balance Scale

	acc	good	vgood	unacc
acc	125	8	4	10
good	4	12	9	3
vgood	3	4	21	0
unacc	18	1	2	468

(b) ID3: Car

	1	2	3
1	15	22	10
2	31	82	43
3	17	41	25

(c) ID3: NPHA

	B	L	R
B	0	14	4
L	5	83	12
R	10	40	82

(d) One vs One: Balance Scale

	acc	good	vgood	unacc
acc	141	6	2	11
good	3	25	3	0
vgood	5	0	20	0
unacc	16	3	0	457

(e) One vs One: Car

	1	2	3
1	4	39	18
2	17	90	37
3	9	45	27

(f) One vs One: NPHA

	B	L	R
B	0	13	7
L	9	92	14
R	10	32	73

(g) One vs Rest: Balance Scale

	acc	good	vgood	unacc
acc	141	5	4	12
good	6	17	0	0
vgood	3	7	21	0
unacc	12	2	0	462

(h) One vs Rest: Car

	1	2	3
1	6	37	14
2	20	95	30
3	8	54	22

(i) One vs Rest: NPHA

Rysunek 1: Macierze pomyłek

4.2 Metryki

Do porównania jakości klasyfikacji modeli wykorzystujemy metryki: dokładność, odzysk, precyzja, miara F, specyficzność, TP rate, FP rate.

Dla zastosowania standardowych metryk klasyfikacji binarnej do problemu klasyfikacji wieloklasowej stosujemy podejście makro-uśredniania i mikro-uśredniania. Wyniki dla obu wariantów przedstawiono w tabelach.

Dla każdej metryki podajemy wartość średnią i odchylenie standardowe wyliczone przy k-krotnej walidacji krzyżowej z $k = 5$.

Makro-uśrednianie

Model	Zbiór danych	Dokładność (avg)	Dokładność (std)	Odzysk (avg)	Odzysk (std)	Precyzja (avg)	Precyzja (std)	Miara F (avg)	Miara F (std)
ID3	Car	0.964	0.005	0.856	0.033	0.825	0.041	0.835	0.035
One vs Rest	Car	0.965	0.008	0.845	0.064	0.841	0.015	0.837	0.04
One vs One	Car	0.979	0.006	0.935	0.026	0.926	0.043	0.927	0.033
ID3	Balance scale	0.775	0.036	0.499	0.029	0.471	0.03	0.484	0.028
One vs Rest	Balance scale	0.786	0.034	0.515	0.027	0.501	0.025	0.504	0.025
One vs One	Balance scale	0.738	0.048	0.471	0.038	0.473	0.048	0.469	0.04
ID3	NPHA	0.508	0.029	0.355	0.03	0.351	0.031	0.349	0.031
One vs Rest	NPHA	0.53	0.055	0.34	0.04	0.336	0.045	0.334	0.044
One vs One	NPHA	0.516	0.037	0.341	0.021	0.341	0.02	0.339	0.02

Model	Zbiór danych	Specyficzność (avg)	Specyficzność (std)	TP rate (avg)	TP rate (std)	FP rate (avg)	FP rate (std)
ID3	Car	0.967	0.004	0.856	0.033	0.033	0.004
One vs Rest	Car	0.972	0.004	0.845	0.064	0.028	0.004
One vs One	Car	0.983	0.007	0.935	0.026	0.017	0.007
ID3	Balance scale	0.794	0.031	0.499	0.029	0.206	0.031
One vs Rest	Balance scale	0.814	0.027	0.515	0.027	0.186	0.027
One vs One	Balance scale	0.767	0.047	0.471	0.038	0.233	0.047
ID3	NPHA	0.558	0.027	0.355	0.03	0.442	0.027
One vs Rest	NPHA	0.568	0.036	0.34	0.04	0.432	0.036
One vs One	NPHA	0.558	0.032	0.341	0.021	0.442	0.032

Mikro-uśrednianie

Model	Zbiór danych	Dokładność (avg)	Dokładność (std)	Odzysk (avg)	Odzysk (std)	Precyzja (avg)	Precyzja (std)	Miara F (avg)	Miara F (std)
ID3	Car	0.963	0.005	0.929	0.01	0.929	0.01	0.929	0.01
One vs Rest	Car	0.964	0.008	0.931	0.014	0.931	0.014	0.931	0.014
One vs One	Car	0.979	0.007	0.959	0.013	0.959	0.013	0.959	0.013
ID3	Balance scale	0.769	0.038	0.69	0.046	0.69	0.046	0.69	0.046
One vs Rest	Balance scale	0.783	0.035	0.707	0.043	0.707	0.043	0.707	0.043
One vs One	Balance scale	0.735	0.05	0.65	0.059	0.65	0.059	0.65	0.059
ID3	NPHA	0.503	0.03	0.403	0.029	0.403	0.029	0.403	0.029
One vs Rest	NPHA	0.524	0.056	0.424	0.056	0.424	0.056	0.424	0.056
One vs One	NPHA	0.512	0.037	0.412	0.036	0.412	0.036	0.412	0.036

Model	Zbiór danych	Specyficzność (avg)	Specyficzność (std)	TP rate (avg)	TP rate (std)	FP rate (avg)	FP rate (std)	Specyficzność (avg)	Specyficzność (std)
ID3	Car	0.975	0.004	0.929	0.01	0.025	0.004	0.975	0.004
One vs Rest	Car	0.976	0.005	0.931	0.014	0.024	0.005	0.976	0.005
One vs One	Car	0.986	0.004	0.959	0.013	0.014	0.004	0.986	0.004
ID3	Balance scale	0.816	0.031	0.69	0.046	0.184	0.031	0.816	0.031
One vs Rest	Balance scale	0.828	0.029	0.707	0.043	0.172	0.029	0.828	0.029
One vs One	Balance scale	0.786	0.043	0.65	0.059	0.214	0.043	0.786	0.043
ID3	NPHA	0.574	0.03	0.403	0.029	0.426	0.03	0.574	0.03
One vs Rest	NPHA	0.594	0.053	0.424	0.056	0.406	0.053	0.594	0.053
One vs One	NPHA	0.583	0.036	0.412	0.036	0.417	0.036	0.583	0.036

4.3 Wnioski

Zastosowanie różnych wariantów klasyfikatora daje różnice rzędu kilku punktów procentowych dla przyjętych metryk.

Na zbiorze danych **Car** wyraźne najlepiej sprawdza się podejście One vs One. Miara F przy makro-uśrednianiu jest większa aż o ok. 0.09 względem pozostałych wariantów. Bardziej widoczna jest przewaga podejścia OvO przy makro-uśrednianiu, gdzie jakość predykcji każdej z klas traktujemy jako jednakowo ważne. Przy mikro-uśrednianiu, gdzie predykcje klas są traktowane jako ważne proporcjonalnie do ich częstości, przewaga OvO pozostałymi wariantami jest mniejsza. Standardowy model i OvR osiągnęły zbliżone wyniki. Przykłady klasy **unacc** dominują w zbiorze (70% wszystkich przykładów).

Na zbiorze danych **Balance scale** najlepsze rezultaty osiąga wariant One vs Rest, a najgorsze One vs One. W tym zbiorze danych przykłady klasy **B** są znacznie rzadsze niż pozostałych klas, więc odzysk, precyzja i miara F są zdecydowanie niższe przy makro-uśrednianiu. Wartości wszystkich metryk zachowują podobne relacje i (OvR $>$ ID3 $>$ OvO) i różnice rzędu 0.03.

Na zbiorze danych **NPHA** rezultaty pod kątem miary F są zbliżone dla ID3 i OvR przy makro-uśrednianiu. Przy mikro-uśrednianiu OvR daje o ok. 0.02 lepsze rezultaty według wszystkich miar, a ID3 i OvO bardzo zbliżone.

Dla badanych zbiorów danych jeden z modeli zespołowych opartych na ID3 sprawdzał się lepiej niż standardowy klasyfikator.

Wszystkie z stosowanych miar jakości zachowują podobne relacje i proporcje różnic pomiędzy różnymi modelami, dla tego samego zbioru danych.

Wariant OvO, gdzie pojedyncze drzewa uczone są na mniejszych (znacznie mniejszych dla mniej licznych klas) podzbiorach zbioru uczącego, uzyskał wyraźną przewagę na zbiorze **Car**, który był bardziej liczny niż pozostałe, a w rozkładzie klas jedna wyraźnie dominuje i 2 są stosunkowo rzadkie (poniżej 4%).

Wariant OvR sprawdził się lepiej na zbiorach danych **Balance scale** i **NPHA** o podobnej liczności, ale różnej liczbie atrybutów. Dla tych zbiorów gdzie wariant OvR osiąga najwyższe miary jakości, wariant OvO osiąga gorsze lub porównywalne wyniki do standardowego klasyfikatora.

5 Odstępstwa od dokumentacji wstępnej

Uzupełnienie: przy obliczaniu miar jakości klasyfikacji stosujemy k-krotną walidację krzyżową z $k = 5$ (średnie i odchylenia standardowe).

Zrezygnowaliśmy z wykreślania krzywych ROC i wyliczania AUC ze względu na problem ze zdefiniowaniem progu odcięcia dla rozważanych modeli klasyfikacji.

Zbiór danych **nursery** zastąpiliśmy zbiorem **NPHA**, ponieważ wszystkie modele osiągały na nim dokładność zbliżoną do 100%