UMA - Projekt

Semestr 2024Z

- Mikołaj Garbowski
- Michał Pałasz

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.

Opis Rozwiązania

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.

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 - przypisujemy klasę pozytywną w miejsce klasy c, przypisujemy klasę negatywną w miejsce wszystkich pozostałych.

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.

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.

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*.

Prawidłowość implementacji

O poprawności zaimplementowanych algorytmów zapewniają testy jednostkowe w katalogu `tests`.

```
Name                                                 Stmts   Miss  Cover   Missing
----------------------------------------------------------------------------------
src/__init__.py                                          0      0   100%
src/classifiers/__init__.py                              0      0   100%
src/classifiers/classifier.py                           17      3    82%   12, 17, 26
src/classifiers/id3.py                                  77      0   100%
src/classifiers/one_vs_one.py                           31      0   100%
src/classifiers/one_vs_rest.py                          21      0   100%
src/dataset/__init__.py                                  0      0   100%
src/dataset/dataset.py                                 116     14    88%   48, 160-174
...
----------------------------------------------------------------------------------
TOTAL                                                  644     54    92%
```

Zbiory danych

Primary tumor

Zwitter, M. & Soklic, M. (1987). Primary Tumor [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5WK5Q.

Celem zbioru danych jest klasyfikacja rodzaju nowotworu na podstawie informacji o pacjencie i jego stanie zdrowia.

- `lung` (25%)
- `head & neck` (6%)
- `esophasus` (3%)
- `thyroid` (4%)
- `stomach` (12%)
- `duoden & sm.int` (0%)
- `colon` (4%)
- `rectum` (2%)
- `anus` (0%)
- `salivary glands` (1%)
- `pancreas` (8%)
- `gallblader` (5%)
- `liver` (2%)
- `kidney` (7%)
- `bladder` (1%)
- `testis` (0%)
- `prostate` (3%)
- `ovary` (9%)
- `corpus uteri` (2%)
- `cervix uteri` (1%)
- `vagina` (0%)
- `breast` (7%)

Charakterystyka zbioru danych

- liczba przykładów: 339
- liczba atrybutów: 17
- liczba klas: 22
- typy danych kategoryczne
- brakujące dane zastąpione wartościami `?`

Car evaluation

Bohanec, M. (1988). Car Evaluation [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5JP48.

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 przykładów: 1728
- liczba atrybutów: 6
- liczba klas: 4
- typy danych kategoryczne
- bez brakujących danych

Balance scale

Siegler, R. (1976). Balance Scale [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5488X.

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` (46%) - szalka przechyla się w lewo
- `B` (8%) - szalka jest w równowadze
- `R` (46%) - szalka przechyla się w prawo

Charakterystyka zbioru danych

- liczba przykładów: 625
- liczba atrybutów: 4
- liczba klas: 3
- typy danych kategoryczne
- bez brakujących danych

Wyniki

W ramach eksperymentu porównujemy działanie trzech klasyfikatorów: ID3, One vs Rest, One vs One na zbiorach danych opisanych powyżej.

Żaden z klasyfikatorów nie posiada konfigurowalnych parametrów.

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.

Stosujemy 5-krotną walidację krzyżową.

Dla każdej metryki podajemy wartość średnią i odchylenie standardowe uzyskane przy walidacji krzyżowej.

W każdym eksperymencie prezentujemy wyniki w 2 tabelach

- Wartości średnie metryk dla poszczególnych modeli przy uśrednianiu mikro i makro
- Wartości odchylenia standardowego metryk dla poszczególnych modeli przy uśrednianiu mikro i makro

W tabelach pogrubiono najlepsze wartości w każdej kolumnie dla uśredniania mikro i makro.

Zbiór Primary tumor

Model	Uśrednianie	Dokładność (avg)	Odzysk (avg)	Precyzja (avg)	Miara F (avg)	Specyficzność (avg)	TP rate (avg)	FP rate (avg)
ID3	macro	0,854	0,231	0,235	0,217	0,911	0,231	0,089
OvR	macro	0,822	0,207	0,215	0,194	0,896	0,207	0,104
OvO	macro	**0,863**	**0,267**	**0,263**	**0,258**	**0,917**	**0,267**	**0,083**
ID3	micro	0,843	0,392	0,392	0,392	0,909	0,392	0,091
OvR	micro	0,807	0,333	0,333	0,333	0,887	0,333	0,113
OvO	micro	**0,854**	**0,413**	**0,413**	**0,413**	**0,916**	**0,413**	**0,084**

Model	Uśrednianie	Dokładność (std)	Odzysk (std)	Precyzja (std)	Miara F (std)	Specyficzność (std)	TP rate (std)	FP rate (std)
ID3	macro	0,046	0,056	0,081	0,064	0,031	0,056	0,031
OvR	macro	0,040	0,049	0,056	0,050	0,022	0,049	0,022
OvO	macro	0,026	0,081	0,073	0,071	0,017	0,081	0,017
ID3	micro	0,052	0,090	0,090	0,090	0,032	0,090	0,032
OvR	micro	0,049	0,073	0,073	0,073	0,032	0,073	0,032
OvO	micro	0,029	0,063	0,063	0,063	0,018	0,063	0,018

Zbiór Primary tumor posiada znaczną liczbę klas - 22, część z nich ma niewielką liczebność, a samych przykładów jest stosunkowo niewiele (339).

Wyniki dla modeli One vs One według każdej miary są lepsze niż dla pozostałych modeli, zarówno dla uśredniania mikro, jak i makro. Najgorsze wyniki uzyskano dla modelu One vs Rest - gorsze od ID3 przy dodatkowym nakładzie obliczeniowym.

Przy tak dużej liczbie klas znacznie więcej modeli składowych wykorzysta model OvO (ponad 200) względem OvR (22). Natomiast wiele ze składowych modeli OvO będzie trenowanych na bardzo małych podzbiorach danych.

Wszystkie modele osiągają wysoką dokładność powyżej 80%, jednak niską precyzję i odzysk, wartości odzysku, precyzji i miary F są znacznie niższe przy makro-uśrednianiu ze względu na występowanie w zbiorze wielu mało licznych klas.

Zbiór Car evaluation

Model	Uśrednianie	Dokładność (avg)	Odzysk (avg)	Precyzja (avg)	Miara F (avg)	Specyficzność (avg)	TP rate (avg)	FP rate (avg)
ID3	macro	0,960	0,847	0,829	0,836	0,963	0,847	0,037
OvR	macro	0,967	0,855	0,868	0,859	0,967	0,855	0,033
OvO	macro	**0,978**	**0,953**	**0,946**	**0,948**	**0,977**	**0,953**	**0,023**
ID3	micro	0,960	0,922	0,922	0,922	0,973	0,922	0,027
OvR	micro	0,967	0,936	0,936	0,936	0,978	0,936	0,022
OvO	micro	**0,977**	**0,956**	**0,956**	**0,956**	**0,985**	**0,956**	**0,015**

Model	Uśrednianie	Dokładność (std)	Odzysk (std)	Precyzja (std)	Miara F (std)	Specyficzność (std)	TP rate (std)	FP rate (std)
ID3	macro	0,001	0,021	0,019	0,013	0,007	0,021	0,007
OvR	macro	0,005	0,046	0,036	0,037	0,004	0,046	0,004
OvO	macro	0,005	0,028	0,014	0,016	0,008	0,028	0,008
ID3	micro	0,001	0,003	0,003	0,003	0,001	0,003	0,001
OvR	micro	0,005	0,010	0,010	0,010	0,004	0,010	0,004
OvO	micro	0,005	0,009	0,009	0,009	0,003	0,009	0,003

W zbiorze Car evaluation wyraźnie dominuje klasa `unacc` (70%), a klasy `good` i `vgood` są mało liczne (po 4%).

Wszystkie modele osiągają bardzo wysoką dokładność powyżej 95%, pozostałe metryki również są na wysokim lub bardzo wysokim poziomie.

Wyniki dla modeli One vs One są najlepsze dla wszystkich miar, zarówno dla uśredniania mikro, jak i makro. Bardzo istotna przewaga modeli OvO nad pozostałymi modelami jest widoczna przy makro-uśrednianiu, gdzie odzysk, precyzja i miara F są wyższe o ok. 10 punktów procentowych względem ID3 i OvR. Przewaga modelu OvO jest mniej istotna przy mikro-uśrednianiu.

Model OvR jest tylko nieznacznie lepszy od ID3 przy większym nakładzie obliczeniowym.

Model OvO jest dla tego zadania najlepszym wyborem, szczególnie jeśli istotna jest precyzja dla mniej licznych klas `good` i `vgood`.

Zbiór Balance scale

Model	Uśrednianie	Dokładność (avg)	Odzysk (avg)	Precyzja (avg)	Miara F (avg)	Specyficzność (avg)	TP rate (avg)	FP rate (avg)
ID3	macro	**0,791**	**0,515**	**0,487**	**0,500**	**0,809**	**0,515**	**0,191**
OvR	macro	0,784	0,508	0,483	0,494	0,802	0,508	0,198
OvO	macro	0,751	0,481	0,478	0,477	0,778	0,481	0,222
ID3	micro	**0,786**	**0,710**	**0,710**	**0,710**	**0,830**	**0,710**	**0,170**
OvR	micro	0,779	0,702	0,702	0,702	0,825	0,702	0,175
OvO	micro	0,748	0,664	0,664	0,664	0,798	0,664	0,202

Model	Uśrednianie	Dokładność (std)	Odzysk (std)	Precyzja (std)	Miara F (std)	Specyficzność (std)	TP rate (std)	FP rate (std)
ID3	macro	0,018	0,013	0,029	0,018	0,020	0,013	0,020
OvR	macro	0,027	0,026	0,027	0,025	0,024	0,026	0,024
OvO	macro	0,026	0,016	0,020	0,016	0,020	0,016	0,020
ID3	micro	0,020	0,024	0,024	0,024	0,016	0,024	0,016
OvR	micro	0,029	0,036	0,036	0,036	0,024	0,036	0,024
OvO	micro	0,026	0,030	0,030	0,030	0,022	0,030	0,022

Zbiór balance scale zawiera trzy klasy, 2 z nich mają taką samą liczebność (46%), a jedna jest mniej liczna (8%). Ze względu na występowanie mało licznej klasy, wyniki dla uśredniania makro są znacznie niższe niż dla mikro.

Według wszystkich miar model ID3 osiąga najlepsze wyniki, zarówno dla uśredniania mikro, jak i makro. Model OvR osiąga zbliżone, ale nieco gorsze wyniki co ID3 (przy większym nakładzie obliczeniowym).

Model OvO jest dla tego zadania najgorszy, według każdej miary o ok. 4 punkty procentowe.

Inne przebadane zbiory

Nie zaobserwowano istotnych różnic w miarach jakości dla zbiorów danych

- Nursery
- NPHA

Wnioski zbiorcze

W obrębie jednego eksperymentu modele zachowywały taką samą relację według wszystkich miar jakości. W obrębie jednego eksperymentu, wartości precyzji, odzysku i miary F znacznie różniły się przy uśrednianiu makro i mikro. Dobór wariantu uśredniania miar jakości powinien brać pod uwagę cel postawionego zadania - czy poprawna klasyfikacja jest jednakowo ważna dla każdej klasy (w tym występujących rzadko).

W przeprowadzonych eksperymentach model One vs Rest osiągnął gorsze lub w najlepszym wypadku bardzo zbliżone wyniki do standardowego modelu ID3. Konieczność zbudowania wielu modeli drzewiastych (tyle samo co klas) i trenowanie każdego z nich na zbiorze danych o jednakowej liczbie przykładów wiąże się z większym kosztem obliczeniowym, a jak pokazały eksperymenty, nie przynosi znaczącej poprawy jakości klasyfikacji względem podstawowego algorytmu.

Natomiast model One vs One osiągnął lepsze wyniki niż podstawowy model ID3 na 2 z 3 badanych zbiorów. Poprawa jakości była szczególnie widoczna w przypadku zbioru Car evaluation, gdzie poprawa odzysku, precyzji i miary F (przy makro-uśrednianiu) wynosiła ok. 10 punktów procentowych.

Przewaga modelu OvO była bardziej istotna na zbiorze car evaluation o

- większej liczbie przykładów,
- mniejszej liczbie klas,
- mniejszej liczbie atrybutów,
- bez brakujących danych,

niż na zbiorze Primary tumor, gdzie przewaga nad modelem ID3 wynosiła 2-4 punkty procentowe (w zależności od przyjętej miary). Prawdopodobnie, podejście OvO mogłoby przynieść większe korzyści, gdyby nie fakt, że w zbiorze Primary tumor, dla niektórych klas występuje zaledwie po kilka przykładów, przez co wiele ze składowych drzew jest uczonych na bardzo małych zbiorach.

Wnioski końcowe

Na podstawie przedstawionych wyników, można stwierdzić, że warto rozważyć zastosowanie wariantu One vs One do zadań klasyfikacji wieloklasowej, szczególnie w przypadku, kiedy dla każdej klasy występuje wystarczająca liczba przykładów, tak, aby algorytm miał na czym trenować każde ze składowych drzew. Taka modyfikacja algorytmu ID3 w wybranych zadaniach może znacznie poprawić jakość klasyfikacji.

Dla wariantu One vs Rest przeprowadzone eksperymenty nie dostarczyły podstaw, żeby zalecać jego zastosowanie zamiast standardowego algorytmu ID3. Taka modyfikacja nie przynosi korzyści w jakości klasyfikacji, a wiąże się z większym nakładem obliczeniowym.

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 **Primary tumor**.

We wzorach matematycznych w algorytmie ID3 używany jest logarytm o podstawie 2, uzyskiwane wartości różnią się od tych z dokumentacji wstępnej (gdzie przedstawiono obliczenia z logarytmem naturalnym).