Hierarchical Classifier

Michał Kuźma

Prowadzący: dr inż. Jerzy Stefanowski

October 2, 2017

1 Wstęp

Algorytmy uczenia maszynowego zdają się dzisiaj być wszechobecne. Od rozpoznawania twarzy podczas robienia zdjęć, przez automatyczne etykietowanie zdjęć, a skończywszy na alertach lekarskich ostrzegających o ryzyku wystąpienia choroby. Ta dziedzina nauki rozwija się na naszych oczach i jesteśmy skłonni powierzyć automatycznym klasyfikatorom coraz bardziej złożone i odpowiedzialne zadania.

Jednak pomimo dużego postępu i znacznego wzrostu precyzji klasyfikatorów na przestrzeni lat, niektóre przypadki wciąż są bardzo trudne do rozwiązania. Takim przypadkiem jest między innymi predykcja przy użyciu klasyfikatora nauczonego na zbiorze niezrównoważonym - takim, w którym znaczna większość przykładów przypisanych jest do jednej z dwóch klas.

Wspomniany problem niezrównoważonych zbiorów danych uczących znany jest od długiego czasu i zostało zaproponowanych wiele rozwiązań mających mu zaradzić. Od wykorzystania innych miar do oceny skuteczności algorytmu, przez różne metody wstępnego przetwarzenia danych, po klasyfikatory odporne na zjawisko zbiorów niezrównoważonych.

Trudniejszym zadaniem jest nauczenie klasyfikatora na zbiorze niezrównoważonym, w którym klasa mniejszościowa nie jest jednorodna (występuje w wielu skupiskach, a niekiedy jako pojedyncze przykłady otoczone większościowymi). Uważamy, że potencjalnym rozwiązaniem problemu może być zaproponowany przez nas klasyfikator hierarchiczny.

2 Koncepcja projektu

W publikacji z 2015 roku [1] zaproponowano następujacy podział przykładów klasy mniejszościowej:

- Przykłady bezpieczne (*Safe*) – Przykłady otoczone klasą mniejszościową, które bezpiecznie można uznać za mniejszościowe

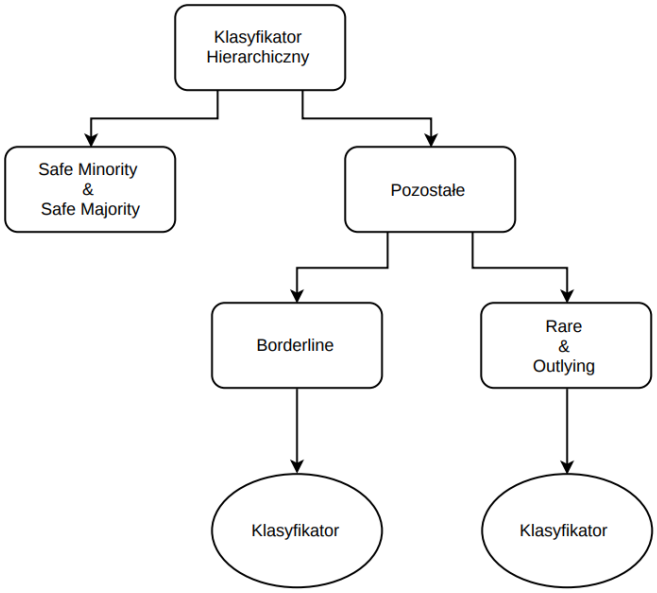

Figure 1: Schemat działania klasyfikatora hierarchicznego

- Przykłady graniczne (*Borderline*) – Przykłady na granicy grupy klasy mniejszościowej, sąsiadują zarówno z przykładami klasy mniejszościowej, jak i większościowej
- Przykłady rzadkie (*Rare*) – Przykłady klasy mniejszościowej stanowiące element małej grupy otoczonej większościowymi
- Przykłady odosobnione (*Outlying*) – Pojedyncze obiekty klasy mniejszościowej otoczone w całości przez przykłady większościowe

Zaproponowany przez nas klasyfikator uczy się i dokonuje predykcji w sposób hierarchiczny, jak przedstawia figura 1. Na pierwszej warstwie zbiór dzielony jest na przykłady bezpieczne (tak z klasy mniejszościowej, jak i z większościowej), które mogą być wprost zaetykietowane i pozostałe. Te są dalej dzielone na graniczne i "niebezpieczne" (*Rare* i *Outlying*). Oba zbiory stanowią zestawy uczące dla prostych klasyfikatorów.

Zgodnie z sugestią z [1] przykłady etykietowane są określonym typem w procesie analizy sąsiedztwa k najbliższych sąsiadów.

3 Wyniki eksperymentów

Celem eksperymentów było porównanie wyników czystego drzewa decyzyjnego C4.5 i opartego na nim klasyfikatora hierarchicznego. Główny nacisk położyliśmy

na czułość dla klasy mniejszościowej (stosunek poprawnie sklasyfikowanych przykładów mniejszościowych do wszystkich obiektów tej klasy).

Eksperymenty przeprowadzono na zbiorach rzeczywistych, które charakteryzują się występowaniem niejednorodnej klasy mniejszościowej. Wykorzystane zbiory danych:

- *Seismic bumps*
- *Haberman*
- *Solar flare*
- *CMC*

3.1 Zbiór *seismic bumps*

Oryginalna \Predykcja	**Mniejszościowa**	**Większościowa**
Mniejszościowa	11	159
Większościowa	39	2375

Ważona czułość	0.923375
Czułość mniejszościowej	0.064706
Czułość większościowej	0.983844

Table 1: Macież pomyłek i czułości klasyfikatora hierarchicznego

Oryginalna \Predykcja	**Mniejszościowa**	**Większościowa**
Mniejszościowa	0	170
Większościowa	2	2412

Ważona czułość	0.933437
Czułość mniejszościowej	0.000000
Czułość większościowej	0.999171

Table 2: Macież pomyłek i czułości drzewa decyzyjnego

3.2 Zbiór *haberman*

Oryginalna \Predykcja	**Mniejszościowa**	**Większościowa**
Mniejszościowa	19	62
Większościowa	15	210

Ważona czułość	0.748366
Czułość mniejszościowej	0.234568
Czułość większościowej	0.933333

Table 3: Macież pomyłek i czułości klasyfikatora hierarchicznego

Oryginalna \Predykcja	Mniejszościowa	Większościowa
Mniejszościowa	24	57
Większościowa	29	196

Ważona czułość	0.718954
Czułość mniejszościowej	0.296296
Czułość większościowej	0.871111

Table 4: Macież pomyłek i czułości drzewa decyzyjnego

3.3 Zbiór *solar flare*

Oryginalna \Predykcja	Mniejszościowa	Większościowa
Mniejszościowa	2	41
Większościowa	3	1020

Ważona czułość	0.958724
Czułość mniejszościowej	0.046512
Czułość większościowej	0.997067

Table 5: Macież pomyłek i czułości klasyfikatora hierarchicznego

Oryginalna \Predykcja	Mniejszościowa	Większościowa
Mniejszościowa	0	43
Większościowa	0	1023

Ważona czułość	0.959662
Czułość mniejszościowej	0.000000
Czułość większościowej	1.000000

Table 6: Macież pomyłek i czułości drzewa decyzyjnego

3.4 Zbiór *cmc*

Oryginalna \Predykcja	Mniejszościowa	Większościowa
Mniejszościowa	85	248
Większościowa	99	1041

Ważona czułość	0.764426
Czułość mniejszościowej	0.255255
Czułość większościowej	0.913158

Table 7: Macież pomyłek i czułości klasyfikatora hierarchicznego

Oryginalna \Predykcja	Mniejszościowa	Większościowa
Mniejszościowa	104	229
Większościowa	87	1053

Ważona czułość	0.785472
Czułość mniejszościowej	0.312312
Czułość większościowej	0.923684

Table 8: Macież pomyłek i czułości drzewa decyzyjnego

4 Wnioski

Na chwilę obecną klasyfikator w zależności od zbioru osiąga większą lub mniejszą dokładność od drzewa decyzyjnego. Dalsze prace są wskazane.

References

[1] Krystyna Napierala and Jerzy Stefanowski. Types of minority class examples and their influence on learning classifiers from imbalanced data. *Journal of Intelligent Information Systems*, 46:563–597, 2016.