Zastosowanie systemów wieloklasyfikatorowych do diagnostyki chorób wątroby

Autor: Michał Honc

Promotor: Prof. dr hab. inż. Marek Kurzyński

HR EXCELLENCE IN RESEARCH

Politechnika Wrocławska

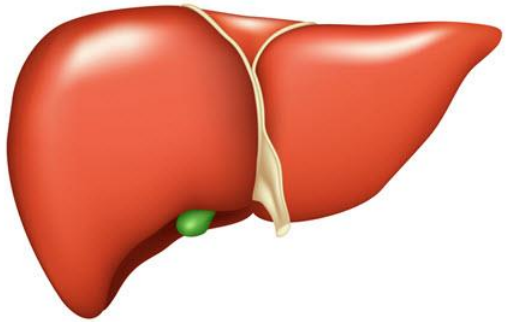
Plan prezentacji

- Cel pracy
- Opis problemu oraz zbioru danych
- Dotychczasowe postępy
- Harmonogram

Cel pracy

Stwierdzenie, czy dany pacjent ma chorą wątrobę możemy sprowadzić do problemu klasyfikacji. Klasyfikacja polega na przypisaniu rozpoznawanemu obiektowi odpowiedniej klasy na podstawie znajomości wartości wybranych cech.

W przypadku diagnozowania chorób wątroby, badanymi cechami mogą być dane dotyczące organizmu pacjenta takie jak np. wiek, płeć lub zawartość konkretnych substancji we krwi.



Zbiór danych

- https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient +Dataset)
- Dla badanego problemu klasyfikacji zostały zdefiniowane dwie klasy. Ich rozkład w rozpatrywanym zbiorze danych: – Pacjent ze zdrową wątrobą - 167 próbek (27,46%) – Pacjent z chorą wątrobą - 441 próbek (72,54%) - dane niezbalansowane.
- Łącznie 10 cech – wiek, płeć, oraz 8 cech wyrażających stężenia poszczególnych białek (np. Bilirubina całkowita)
- Do przeprowadzenia eksperymentów wykorzystano walidację krzyżową 5x5
- Zbiór został dzielony na zbiór uczący, walidacyjny oraz testowy w proporcjach 60/20/20%
- W eksperymentach porównano wyniki dla <5-10> cech, tworząc ranking cech za pomocą algorytmu ANOVA

Metody dynamicznej selekcji

Stworzenie algorytmu dynamicznej selekcji klasyfikatora **DCS-LA (Local accuracy)**, działa on następująco:

1. Dla każdego punktu x w zbiorze testowym X znajdujemy k najbliższych sąsiadów tego punktu.
2. Dla każdego klasyfikatora z puli dokonujemy predykcji lokalnej wykorzystując zbiór walidacyjny, dla k najbliższych sąsiadów punktu x.
3. Dla każdego punktu x zostaje wybrany klasyfikator, który uzyskał najlepszy lokalny wynik.
4. Mając przyporządkowany klasyfikator w każdym punkcie x, dokonujemy predykcji na zbiorze testowym.

Metody dynamicznej selekcji

Stworzenie algorytmu dynamicznej selekcji klasyfikatorów **k-Nearest Oracle-Eliminate (KNORA-E)**:

- Ta metoda wyszukuje lokalną wyrocznię, która jest klasyfikatorem bazowym, który poprawnie klasyfikuje wszystkie próbki należące do regionu kompetencji próbki testowej. Wybierane są wszystkie klasyfikatory o doskonałej wydajności w regionie kompetencji (czyli te które sklasyfikują poprawnie k najbliższych sąsiadów). W przypadku, gdy żaden klasyfikator nie osiąga doskonałej dokładności, rozmiar regionu kompetencji jest zmniejszany (poprzez usunięcie najdalszego sąsiada), a wydajność klasyfikatorów jest ponownie oceniana. Wyniki wybranego zespołu klasyfikatorów są łączone przy użyciu schematu głosowania większościowego. Jeśli nie zostanie wybrany żaden klasyfikator bazowy, do klasyfikacji używana jest cała pula.

Metody dynamicznej selekcji

Stworzenie algorytmu dynamicznej selekcji klasyfikatorów **Dynamic ensemble selection-Performance(DES-P)**:

- Metoda ta wybiera wszystkie klasyfikatory bazowe, które osiągają wydajność klasyfikacji w obszarze kompetencji (dla k najbliższych sąsiadów punktu) wyższą niż klasyfikator losowy (RC). Wydajność klasyfikatora losowego jest określona przez RC = 1/L, gdzie L jest liczbą klas w problemie (dla naszego problemu jest to więc ½). Jeśli nie zostanie wybrany żaden klasyfikator bazowy, do klasyfikacji używana jest cała pula.

Wyniki

- Pula heterogeniczna – 6 klasyfikatorów bazowych vs komitet złożony z tych klasyfikatorów

Mean scores:

	5	6	7	8	9	10
GaussianNB	0.6941	0.6940	0.6857	0.6869	0.6923	0.6845
kNN	0.6947	0.6827	0.7265	0.7182	0.6930	0.7122
SVC	0.6526	0.6454	0.6562	0.6677	0.6737	0.6725
LogisticRegression	0.6983	0.7145	0.7049	0.7079	0.7091	0.7200
MLP	0.6917	0.7097	0.6989	0.6881	0.6863	0.6905
DecisionTree	0.7193	0.7152	0.7157	0.7236	0.7019	0.7170
OLA	0.7115	0.7212	0.7308	0.7283	0.7236	0.7289
Knorae	0.7265	0.7332	0.7380	0.7440	0.7284	0.7374
DESP	0.7223	0.7199	0.7307	0.7530	0.7302	0.7428

Wyniki

- Pule homogeniczne – 50 klasyfikatorów uczonych na zbiorze uczącym za pomocą bootstrapingu vs pojedynczy klasyfikator uczony na zbiorze uczącym+walidacyjnym.

Mean scores:

	5	6	7	8	9	10
LogisticRegression	0.7128	0.7121	0.7134	0.7092	0.7133	0.7091
OLA	0.7170	0.7236	0.7248	0.7206	0.7200	0.7115
Knorae	0.7296	0.7283	0.7217	0.7308	0.7266	0.7265
DESP	0.7182	0.7212	0.7139	0.7134	0.7200	0.7223

STD scores:

	5	6	7	8	9	10
LogisticRegression	0.0367	0.0315	0.0343	0.0263	0.0262	0.0276
OLA	0.0396	0.0321	0.0268	0.0245	0.0262	0.0332
Knorae	0.0367	0.0274	0.0300	0.0188	0.0246	0.0277
DESP	0.0443	0.0297	0.0298	0.0241	0.0215	0.0227

Harmonogram

- Początek czerwca: dokończenie pracy magisterskiej.

Literatura

[1] Źródło internetowe danych medycznych https://archive.ics.uci.edu/ml/datasets/ILPD+Indian+Liver+Patient+Dataset)

[2] L. Kuncheva, Combining Pattern Classifiers, Methods and Algorithms, Wiley Interscience 2014

[3] K. Woods, W. Kegelmeyer, K. Bowyer, Combination of multiple classifiers using local accuracy estimates, IEEE Trans. on Pattern Analysis and Machine Intelligence 19, 405 – 410 (1997)