Politechnika Warszawska
Wydział Matematyki i Nauk Informacyjnych

Metody sztucznej inteligencji 2

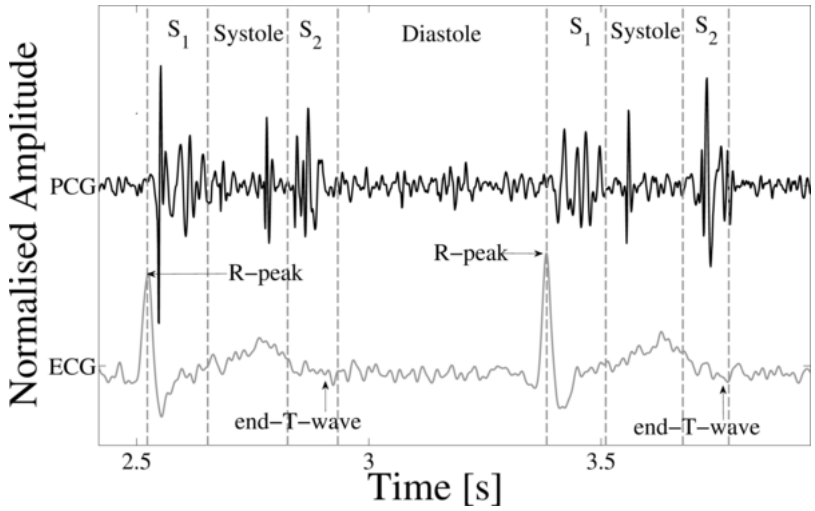
Raport z projektu

Anna Zawadzka & Piotr Waszkiewicz

10 czerwca 2016

0.1 Opis projektu

Projekt polega na realizacji zadania przedstawionego na stronie *https://www.physionet.org/challenge/2016*. Jego celem jest przetworzenie nagrań dźwięków serca i zidentyfikowanie, które z nich reprezentują zaburzoną pracę serca i wymagają diagnozy eksperta.
Fonokardiogram (*Phonocardiogram*,PCG) jest graficznym zapisem pracy serca. Poniższy rysunek przedstawia fragment zapisu PCG.



Rysunek 1: Zapis PCG połączony z zapisem ECG

Głównymi etapami zadania są:

- Segmentacja nagrań na poszczególne części
- Wyodrębnienie cech sygnału
- Klasyfikacja obiektów na poprawne i zaburzone

Wraz z informacjami zamieszczonymi na stronie, dotyczącymi tematyki projektu, dołączone zostały zbiory nagrań audio pracy serca, oraz wstępny program realizujący cel zadania. Program ten zawiera wszystkie z wymienionych wcześniej etapów, jednak wymaga modyfikacji.

Użyta w implementacji segmentacja opisana jest przez autorów konkursu jako *state of the art* i nie została przez nas zmieniona. Wyodrębniane cechy nie wyczerpują zakresu możliwych do wyekstrahowania informacji, dlatego też uzupełniliśmy istniejący zestaw o kilka nowych cech. W dostarczonym rozwiązaniu najgorzej jednak prezentuje się klasyfikacja, która wykorzystuje naiwne techniki statystyczne. To właśnie ona w pierwszym kroku została poddana usprawnieniom i poprawkom.

0.2 Ekstrahowane cechy

W ramach konkursu Physionet przygotowane zostały pliki przeprowadzające wstępną segmentację, ekstrakcję cech i prostą klasyfikację. Ponieważ segmentacja korzysta z najnowszych algorytmów i podejść, została ona pozostawiona i wykorzystana w ramach projektu. Do zestawu wyciąganych cech, na które składały się takie wartości jak odchylenia standardowe, wartości średnie oraz wariancje długości trwania poszczególnych wyróżnionych części cyklu PCG, dodane zostały trzy nowe. Były nimi wartości minimalne, maksymalne oraz kurtoza dla interwałów pomiędzy poszczególnymi elementami cyklu bicia serca. Ostatecznie ekstrahowany wektor cech miał długość 30 elementów.

0.3 Klasyfikatory

0.3.1 Opis

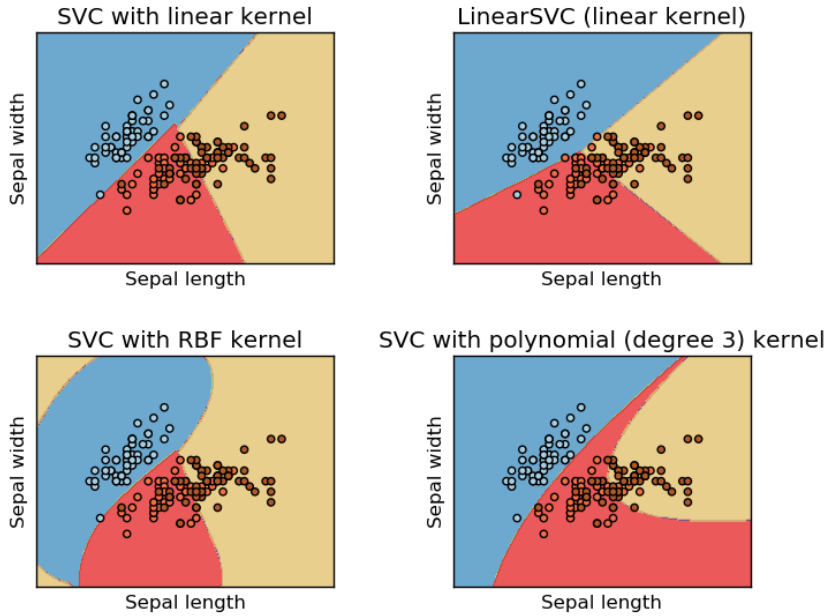
Przy konstruowaniu lepszych klasyfikatorów rozważyliśmy kilka znanych i sprawdzonych metod: SVM, KNN i Lasy losowe. Wszystkie z wymienionych podejść charakteryzują się bardzo dobrymi wynikami w dziedzinie klasyfikacji, jednak niekoniecznie sprawdzają się równie dobrze w tych samych warunkach. W naszej pracy chcieliśmy zbadać wpływ wyboru klasyfikatora na skuteczność rozpoznawania zaburzeń rytmu serca w nagraniach. Testy zostały przeprowadzone z użyciem gotowej implementacji dostępnej na stronie konkursu. Wszystkie implementacje wykorzystywanych klasyfikatorów zostały użyte w ramach pakietu MATLAB.

SVM

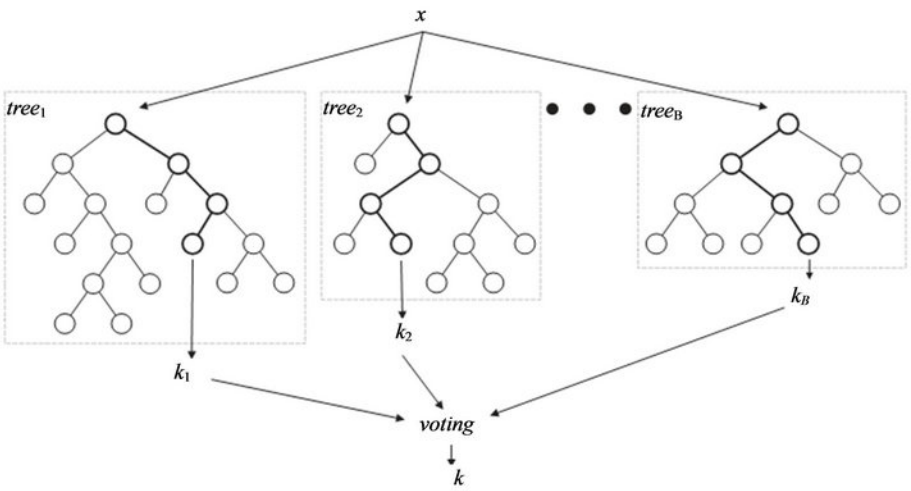
Maszyna wektorów nośnych (SVM) jest klasyfikatorem którego nauka ma na celu wyznaczenie hiperpłaszczyzny rozdzielającej z maksymalnym marginesem przykłady należące do klas. Najprostszym sposobem jest próba liniowego odseparowania punktów. Niestety, często okazuje się, że jest ona niewystarczająca lub wręcz niemożliwa. W tym celu stworzone zostały funkcje jądrowe które mapują dostarczone punkty na odpowiadające im elementy z wyższych wymiarów celem umożliwienia ich późniejszej separacji.

Las losowy

Klasyfikator korzystający ze zbioru drzew decyzyjnych, które dla każdego elementu głosują nad jego przynależnością do określonej klasy. Liczba drzew decyzyjnych wykorzystywanych w procesie uczenia i klasyfikacji jest parametrem którego wartość należy samemu określić.



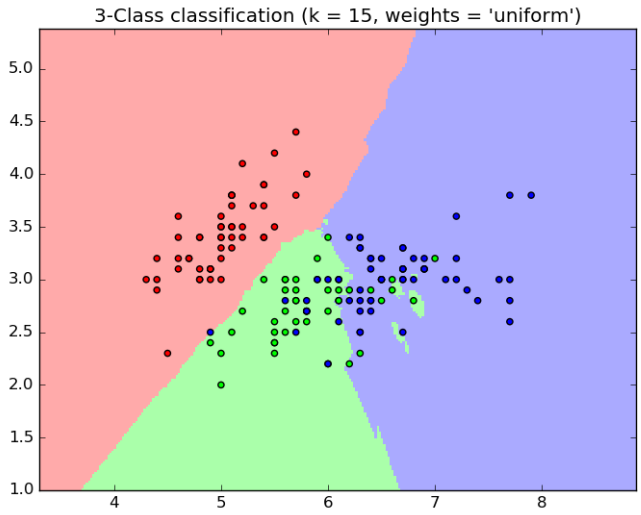
Rysunek 2: Przykład zachowania klasyfikatora SVM dla różnych funkcji jądrowych



Rysunek 3: Klasyfikator korzystający z lasów losowych buduje odpowiednią liczbę drzew które następnie głosują nad zaklasyfikowaniem podanego elementu

KNN

Klasyfikacja metodą k-najbliższych sąsiadów polega na sprawdzeniu otoczenia wybranego punktu i nadanie mu przynależności do tej klasy do której należy większość sąsiadów. Liczba rozpatrywanych sąsiadów jest wartością arbitralną i jej zmiany mogą znacząco wpływać na jakość wyników.



Rysunek 4: Klasyfikator KNN przydziela punktów do klas na podstawie ich odległości

0.3.2 Dobór parametrów

Każdy z wybranych przez nas klasyfikatorów ma pewne charakterystyczne parametry, które odpowiednio dobrane skutkują najlepszą skutecznością. Aby zapewnić, że wybrane parametry rzeczywiście są najlepsze ich dobór nastąpił przy użyciu metody grid search. Jest to prosta koncepcyjnie metoda przeszukująca punkty na siatce w zdefiniowanej, ograniczonej przestrzeni potencjalnych wartości. Po jej zastosowaniu otrzymane zostały następujące wartości:

- SVM: jądro: RBF (Gaussowskie)
- RF: liczba lasów = 100
- KNN: liczba sąsiadów branych pod uwagę: 5

0.4 Wykrywanie anomalii

Podczas obliczeń okazało się, że jednym z problemów z jakim zmierzyć się musiały klasyfikatory była pewna liczba zaszumionych nagrań, których wektor cech nie chciał łatwo poddawać się klasyfikacji. Nagrania te mogą doprowadzić do sytuacji w której klasyfikator udzieli błędnej odpowiedzi, co może mieć fatalne skutki w momencie gdy nagranie zawierało błędną pracę serca a uznane było za w pełni poprawne. Aby uniknąć takich problemów, organizatorzy konkursu zdecydowali się wprowadzić możliwość uznania nagrania za niemożliwe lub zbyt ryzykowne do klasyfikacji. Ponieważ jednak użyte w projekcie klasyfikatory charakteryzują się udzielaniem odpowiedzi w postaci binarnej (nagranie poprawne/niepoprawne) należało użyć nowego mechanizmu wzbogacającego istniejący już proces.

Wykorzystany w tym celu detektor anomalii w serii nagrań korzysta z testu Kołmogorowa-Smirnowa. Jest to test statystyczny zwracający dla każdego elementu wartość rzeczywistą z przedziału [0-1] oznaczającą prawdopodobieństwo, że wybrany punkt jest anomalią w serii.

0.5 Podsumowanie i wyniki

Wstępne testy przeprowadzone zostały na znormalizowanych cechach dostepnych w ramach dostarczonego przez konkurs oprogramowania. Zawierały one takie wartości jak odchylenia standardowe interwałów dla poszczególnych segmentów sygnału PCG, standardowe odchylenia oraz wartości średnie. Wyniki klasyfikacji zostały przedstawione w tabelce 1.

KNN	0.29
RF	0.23
SVM	0.27

Tablica 1: Error ratio dla każdej z metod klasyfikacji dla znormalizowanych danych

Próba uzyskania bardziej dokładnych klasyfikacji zakładała poszerzenie istniejącego wektora cech z 20 do 30 cech. Dodane zostały takie miary statyczne jak: różnica między wartością maksymalną a minimalną dla poszczególnych wartości z poprzedniego wektora cech oraz kurtozę tych pomiarów. Ponowne obliczenia doprowadziły do uzyskania takich wartości jakie zostały przedstawione w tabeli 2.

KNN	0.32
RF	0.24
SVM	0.24

Tablica 2: Error ratio dla każdej z metod klasyfikacji dla danych z dodanymi nowymi cechami

Oprócz standardowej klasyfikacji przy użyciu wspomnianych wcześniej narzędzi wykorzystane zostały narzędzia analizy statystycznej mającej na celu wykrycie anomalii. Ich celem było wykrycie najbardziej prawdopodobnych punktów leżących poza zbiorem "dobrych danych"które należy odrzucić. Te z punktów których przynależność do poszczególnych klas nie zgadzała się w przypadku klasyfikacji standardowymi narzędziami klasyfikującymi a wynikami z wykrycia anomalii uznawane były za zbyt zaszumione do poprawnej analizy (i liczące się w procesie zliczania błędów jako 0.5 błędu). Wyniki tego podejścia zaprezentowane zostały w tabelce 3.

KNN	0.24
RF	0.25
SVM	0.25

Tablica 3: Error ratio dla każdej z metod klasyfikacji z uwzględnieniem anomalii

Jak widać najbardziej na nowym podejściu skorzystała metoda KNNów. Wyniki dla SVMów i lasów losowych nieznaczenie się pogorszyły - jest to prawdopodobnie spowodowane sposobem liczenia punktów dla klasyfikacji. Nie są to jednak znaczne różnice i można powiedzieć, że metodę wykrywania anomalii warto wykorzystywać w połączeniu z KNNami.