Dedykowane Algorytmy Diagnostyki Medycznej
k-NN + Extended NN

Olga Janiszewska & Katarzyna Kokosza

23 grudnia 2016

1 Wstęp teoretyczny

1.1 Elektrokardiografia

Elektrokardiografia to nieinwazyjne badanie diagnostyczne. Polega na rejestracji elektrycznej aktywności mięśnia sercowego za pomocą elektrod przymocowanych do klatki piersiowej. Rejestrowany sygnał wynika ze zjawisk zachodzących w układzie bodźcotwórczo-bodźcoprzewodzącym serca czyli *depolaryzacji i repolaryzacji* komórek.

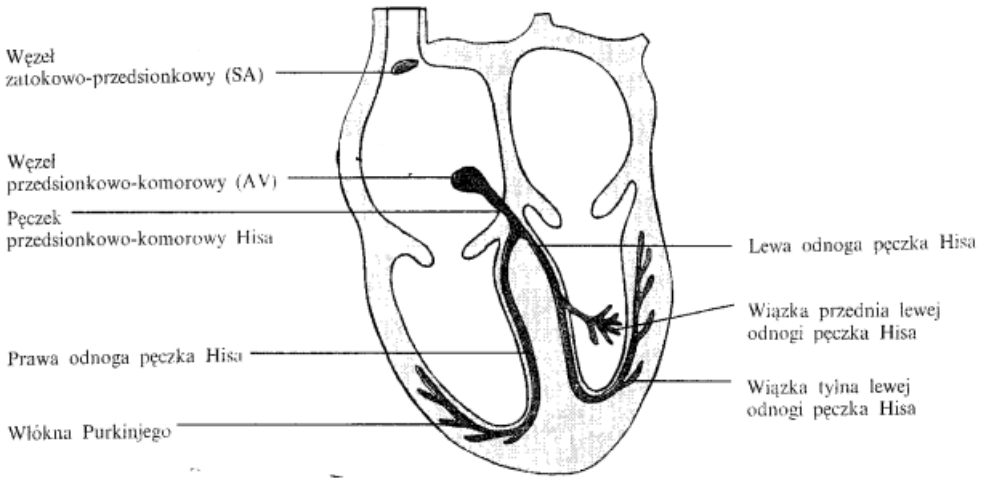
Wyróżnia się cztery podstawowe techniki badania EKG: [1]

- EKG spoczynkowe (12 odprowadzeń)
- Wektokardiografia (3 odprowadzenia ortogonalne)
- Elektrokardiografia próby wysiłkowej
- Monitoring EKG (1 lub 2 odprowadzenia)

Każde z tych badań cechuje się innymi właściwościami. Przykładowo diagnozowanie arytmii jest bardzo trudne podczas EKG spoczynkowego, ponieważ pomimo dokładności tego badania to jest ono zbyt krótkie, żeby epizod arytmii mógł wystąpić.

1.2 Zjawiska w cyklu pracy serca [2]

Anatomię układu bodźcotwórczo-przewodzącego serca przedstawia Rys.1



Rysunek 1: Anatomia układu bodźcotwórczo-przewodzącego serca [2]

Podstawowy generator rytmu serca to *węzeł zatokowo-przedsionkowy*, który sterowany jest przez sympatyczny i parasympatyczny układ nerwowy. Wysyła on w regularnych odstępach czasu impulsy generujące pobudzenie. Rytm serca spowodowany przez ten węzeł nazywany jest rytmem zatokowym.

Z węzła zatokowo-przedsionkowego pobudzenie rozprzestrzenia się na komórki prawego przedsionka, a także za pośrednictwem szlaku międzyprzedsionkowego na komórki lewego przedsionka. Skurcz przedsionków EKG reprezentowany jest przez załamek P

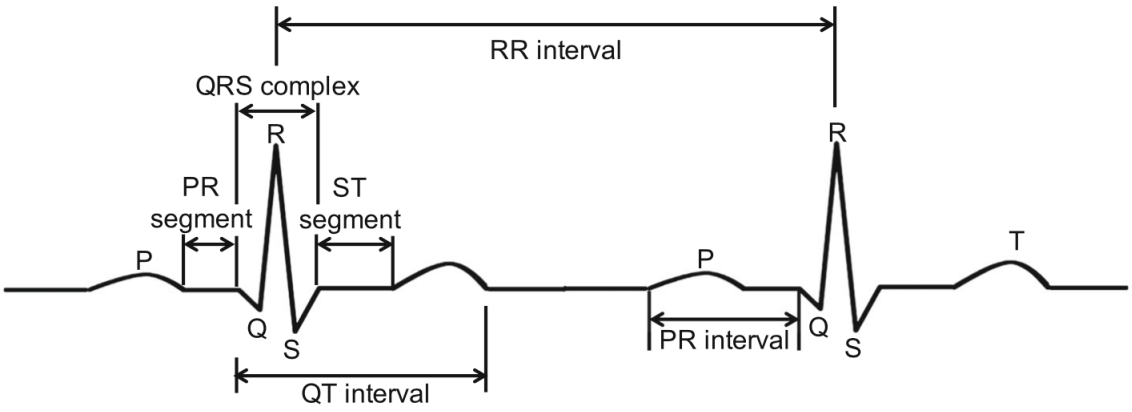
Pobudzenie przednim szlakiem międzywęzłowym dociera do węzła przedsionkowo-komorowego, gdzie napotyka na ośrodek o bardzo małej prędkości przewodzenia. Przez czas ok. 120ms żadne zjawiska nie są obserwowane (*linia izoelektryczna*).

Po przejściu pobudzenia przez pęczek Hisa rozpoczyna się depolaryzacja przegrody międzykomorowej i w niedługim czasie obejmuje koniuszek serca oraz dolne części komór. Po osiągnięciu nasycenia dochodzi do skurczu komór, co reprezentowane jest przez zespół QRS.

Wszystkie komórki, które uległy depolaryzacji podlegają repolaryzacji, czyli odbudowywana jest ich spoczynkowa różnica potencjałów, co na wykresie EKG przedstawia załamek T.

1.3 Sygnał EKG

Sygnał EKG (Rysunek 2) to ciąg próbek napięcia pozyskanych w równych odstępach czasu. Niesie on informacje o cyklicznej pracy serca. Wartość próbki jest różnicą potencjałów elektrycznych pomiędzy dwiema elektrodami. Na wykresie EKG można zaobserwować poszczególne fazy skurczu serca, przez co używany jest do diagnozowania chorób kardiologicznych. Do parametrów na podstawie których określa się prawidłowość rytmu serca należą parametry czasowe oraz morfologiczne. [3]



Rysunek 2: Przebieg sygnału elektrokardiograficznego u zdrowej osoby [4]

1.4 Automatyczna analiza EKG

W ostatnich latach metody automatycznej detekcji rozwijają się bardzo dynamicznie. Rośnie liczba klas oraz cech branych pod uwagę przez klasyfikator. Trwają prace nad wydajnymi algorytmami, które umożliwią automatyczną interpretację zapisów elektrokardiograficznych. Analiza dobowego zapisu pracy serca jest procesem żmudnym, dlatego użycie oprogramowania pozwalającego na zautomatyzowanie tej czynności i wychwycenie potencjalnych nieprawidłowości jest znacznym ułatwieniem pracy dla lekarza kardiologa. Ponadto nowoczesne holtery pozwalają na monitorowanie pacjenta w trybie rzeczywistym - dane są wysyłane bezprzewodowo do stacji monitorującej, co pozwala na natychmiastową reakcję służb medycznych.

2 k-Nearest Neighbours

2.1 Założenia wstępne metody

- Dany jest zbiór uczący wraz z wektorem zmiennych objaśniających oraz wartością zmiennej objaśnianej
- Dany jest zbiór testowy wraz z wektorem zmiennych objaśniających dla którego prognozuje się wartość zmiennej objaśnianej

Proces uczenia się klasyfikatora k-NN polega po prostu na doborze parametru k. Zbiór uczący i testowy to wektory cech wyodrębnionych na drodze przetwarzania sygnału EKG. Na podstawie zbioru testowego klasyfikator uczy się przyporządkowywania cech do odpowiednich klas, a już podczas analizy zbioru testowego klasyfikator samodzielnie przyporządkowuje cechy do odpowiedniej klasy. Proces klasyfikacji obrazuje Rysunek 3.

Rysunek 3: Proces klasyfikowania zespołu QRS do danej klasy

2.2 Opis metody [5]

Klasyfikator k-Najbliższych Sąsiadów to nieparametryczna metoda klasyfikacji oznaczana jako k-NN (*k-Nearest Neighbours*). Rozpoznawany obiekt zalicza się do tej klasy, do której należy większość z jego K najbliższych sąsiadów. Obiekty są analizowane w taki sposób, że oblicza się odległości pomiędzy nimi. Istnieją różne miary podobieństwa - w przypadku niniejszego projektu posłużono się *odległością euklidesową*:

$$d(x,y) = \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2}$$

Klasyfikator ten odnajduje K najbliższych sąsiadów bieżącej próbki i zlicza próbki przypisane do poszczególnych klas. Bieżąca próbka zostaje przydzielona do klasy, która występuje najczęściej wśród swoich najbliższych K sąsiadów. Odległość od danych próbek na tym etapie nie ma już znaczenia, chyba że liczba sąsiadów z poszczególnych klas będzie identyczna.

Na jakość klasyfikacji ma wpływ liczba K uwzględnionych sąsiadów oraz rozkład w przestrzeni wielowymiarowej poprzednich próbek. Wybór odpowiedniej liczby sąsiadów K jest podstawowym zadaniem podczas projektowania klasyfikatora. Zbyt duża liczba K będzie powodowała wysoką złożoność obliczeniową, a zbyt mała sprawi, że klasyfikator nie będzie odporny na szumy.

Zaletami tej metody jest prosty proces uczenia się i łatwość implementacji. Do wad można zaliczyć proces klasyfikacji, który musi przeanalizować cały zbiór uczący, żeby obliczyć wszystkie odległości. Powoduje to, że klasyfikator jest tym wolniejszy im obszerniejszy jest zbiór uczący.

2.3 Klasy

Podczas implementacji prototypu w środowisku MATLAB w wektorach danych testowych i treningowych wyróżniono pięć klas, do których klasyfikowano zespoły QRS. Są to klasy rekomendowane przez *Association for the Advancement of Medical Instrumentation*:

- **N** (*normal*) - normalne, fizjologiczne uderzenie serca
- **S** (*supraventricular*) - ektopowe pobudzenie nadkomorowe
- **V** (*ventricular*) - ektopowe pobudzenie komorowe

- **F** (*fusion*) - pobudzenie mieszane (jednoczesne pobudzenie komorowe i nadkomorowe)
- **Q** (*unknown beat*) - nierozpoznane pobudzenie

Podczas implementacji algorytmu w środowisku C++ do testowania algorytmu użyto danych, które klasyfikowały próbki do dwóch klas:

- 1 - uderzenie nadkomorowe (prawidłowe)
- -1 - uderzenie komorowe (nieprawidłowe)

2.4 Implementacja prototypu algorytmu w środowisku MATLAB

W projekcie zaimplementowano funkcję obliczającą k-NN, która po otrzymaniu pojedynczej próbki zawierającej cechy sygnału EKG klasyfikuje ją do odpowiedniej klasy. Algorytm przetestowano na próbkach, których klasa jest znana, po to aby móc zweryfikować poprawność jego działania. Implementacji prototypu dokonano w środowisku MATLAB.

Poniżej przedstawiono zaimplementowaną funkcję, która zrealizowana została dla każdej próbki ze zbioru testowego:

```
function [ out ] = knnClassification(testData, trainData, trainLabels, k)

dist = pdist2(testData, trainData);
  for i=1:k
   [val(i),idx] = min(dist);
   out(i) = trainLabels(idx);
   dist(idx) = max(dist);
  end
end
```

2.5 Walidacja algorytmu

Algorytm przetestowano na zbiorze 100 próbek testowych. Wynik walidacji wykazał, że 98 probek ze 100 zostało sklasyfikowanych poprawnie. Pokazuje to, że algorytm klasyfikuje skutecznie, jednak dalsze testy powinny być przeprowadzone na dużo większym zbiorze testowym.

3 Extended Nearest Neighbours

3.1 Ograniczenia metody k-NN [6]

Algorytm k-NN jest szeroko wykorzystywany w wielu różnych dziedzinach. Został nawet zakwalifikowany na konferencji *IEEE International Conference on Data Mining* do dziesięciu najlepszych algorytmów do zastosowań w obrębie *data-mining*. Jednak zastosowanie tego algorytmu wymaga odpowiednio zdefiniowanego parametru k oraz odpowiednio dobranej miary obliczania niepodobieństwa. Ponadto próbki, które nie pasują do żadnej klasy lub są reprezentowane przez klasę o niższej częstości występowania w zbiorze uczącym zostaną *zdominowane* przez klasy o najwyższej gęstości występowania. W takim przypadku liczba sąsiadów dla próbki z klasy drugiej może charakteryzować się większą ilością próbek z klasy pierwszej. Skłoniło to do poszukiwania rozwiązań, które umożliwiłyby jeszcze lepszą skuteczność omawianego algorytmu.

3.2 Opis metody eNN [7] [8]

Algorytm ENN (*extended nearest neighbor*) w uproszczeniu używa klasyfikatora KNN do znalezienia k-najbliższych sąsiadów próbki testowej w zbiorze treningowym, a następnie na podstawie obliczenia średniej ważonej kolejnych sąsiadów wybiera tę klasę, dla której suma średnich ważonych do niej przypisanych wśród najbliższych sąsiadów jest największa. Dużą różnicą w stosunku do algorytmu kNN jest brak konieczności wyboru k - algorytm iteruje po wszystkich k od 1 do $\sqrt{(}n$, gdzie n to liczba próbek w zbiorze testowym. Pod uwagę brane są tylko nieparzyste wartości k - pozwala to uniknąć sytuacji,

w której doszłoby do konfliktu oraz zmniejsza złożoność obliczeniową algorytmu, która ze względu na iteracje po k jest duża wyższa w porównaniu z eNN. Dodatkowo algorytm kNN nie tworzy żadnego modelu obliczeniowego - każda próbka testowa jest porównywana z całym zbiorem uczącym, co prowadzi do dłuższego czasu działania w porównaniu z innymi klasyfikatorami.

Średnia ważona i-sąsiada w zborze k-najbliższych sąsiadów:

$$w(i) = \frac{1}{log_2(1+i)}$$

Średnia ważona dla poszczególnej klasy jest obliczana w następujący sposób:

$$WSc = \sum_{k=1}^{\sqrt{n}} \sum_{i=1}^{k} \begin{cases} w(i), A_i = c \\ 0, otherwise \end{cases}$$

Następnie pod uwagę brana jest ta klasa, która występuje najczęściej:

$$class = argmax(cWSc)$$

3.3 Implementacja w środowisku MatLab

Danymi wejściowymi algorytmu są dane testowe oraz dane treningowe. Każdy zespół w ramach grupy posłużył się identycznymi danymi, aby wyniki można było porownać. Na początku znormalizowano dane, jednak i bez tej operacji algorytm działa poprawnie.

Zaimplementowaną funkcję liczącą skuteczność algorytmu. Algorytm jest przygotowany tak, aby działał dla różnych zbiorów danych. Funkcja *knnClassification* zwraca wektor k-najbliższych sąsiadów próbki testowej w kolejności od najbliższego do najdalszego sąsiada. Następnie dla każdego sąsiada obliczana jest jego waga. Wagi są sumowane w obrębie danej klasy. Wynikiem klasyfikacji jest ta klasa, której suma wag jest największa. Następnie, gdy zostanie obliczony rezultat dla każdej wartości k to spośród wektora klas wynikowych algorytm klasyfikuję próbkę testową do tej klasy, która występuje najczęściej.

```
%enn classification for every test sample
for testSample=1:length(testData)
    result = [];
    for k=1:2:sqrt(length(trainData))
        predictedClasses = knnClassification(testData(testSample,:), trainData, trainLabels, k);
        classes = unique(predictedClasses);
        weights = zeros(length(classes),1);
        for i=1:numel(predictedClasses)
            weight = 1/log2(i+1);
            index = find(classes == predictedClasses(i));
            weights(index) = weights(index) + weight;
        end
        [val, idx] = max(weights);
        result(size(result)+1) = classes(idx);
    end
    ennResult(testSample) = mode(result);
end
```

4 Podsumowanie

Algorytmy najpierw zaimplementowano w środowisku MatLab, a później w języku obiektowym C++ w standardzie C++11 używając środowiska programistycznego Eclipse. Podczas ostatecznej implementacji udoskonalono algorytm, zarówno w C++ jak i w MatLabie.

Przeprowadzono test działania programu. Sprawdzono czas realizacji algorytmu oraz skuteczność właściwego klasyfikowania próbek wyrażoną w procentach. Na zastosowanym zbiorze testowym nie wykryto różnicy w skuteczności dla algorytmu KNN oraz ENN - oba poprawnie zaklasyfikowały 98 próbek ze

100. Był to zbiór testowy wybrany przez całą grupę dla porównania wyników uzyskanych w obrębie poszczególnych zespołów. Oczekiwanym rezultatem była większa skuteczność algorytmu ENN, ponieważ uwzględnia on dalsze sąsiedztwo (a co za tym idzie więcej przypadków). Na innym zestawie danych (z bazy danych physionet) rzeczywiście te oczekiwania się potwierdzały. Jednak algorytm ENN ma dużo większą złożoność obliczeniową niż KNN (ponieważ kilkukrotnie wykorzystuje algorytm KNN do własnych obliczeń).

Literatura

[1] Tompkins W. J. *Biomedical Digital Signal Processing.* Prentice Hall, New Jersey, 2000.

[2] Tomasik T. Windak A. *Elektrokardiografia dla lekarza praktyka.* Uniwersyteckie Wydawnictwo Medyczne Vesalius, Kraków, 1998.

[3] Augustyniak P. *Elektrokardiografia dla Informatyka - Praktyka.* Wydawnictwo Studenckiego Towarzystwa Naukowego, Kraków, 2011.

[4] Bhuiyan M. Z. A Park J. Nearest neighbor search with locally weighted linear regression for heartbeat classification, 2016.

[5] Chmielnicki W. *Efektywne metody selekcji cech i rozwiązywania problemu wieloklasowego w nadzorowanej klasyfikacji danych.* Instytut Podstawowych Problemów Techniki Polskiej Akademii Nauk, Kraków, 2012.

[6] Jayalalitha S. Susan D. et al. *K-nearest Neighbour Method of Analysing the ECG Signal (To Find out the Different Disorders Related to Heart).* Journal of Applied Sciences, 2014.

[7] He H. Tang B. *Enn: Extended nearest neighbor Method for Pattern Recognition.* IEEE ComputatIonal Intelligence MagazIne, 2015.

[8] Altarawneh G. A. et al. Alkasassbeh M. *On enhancing the performance of nearest neighbour classifiers using hassanat distance metric*, volume 9. Canadian Journal of Pure and Applied Sciences, 2015.